COMPUTER MODELLING AND NEW TECHNOLOGIES

Computer Modelling
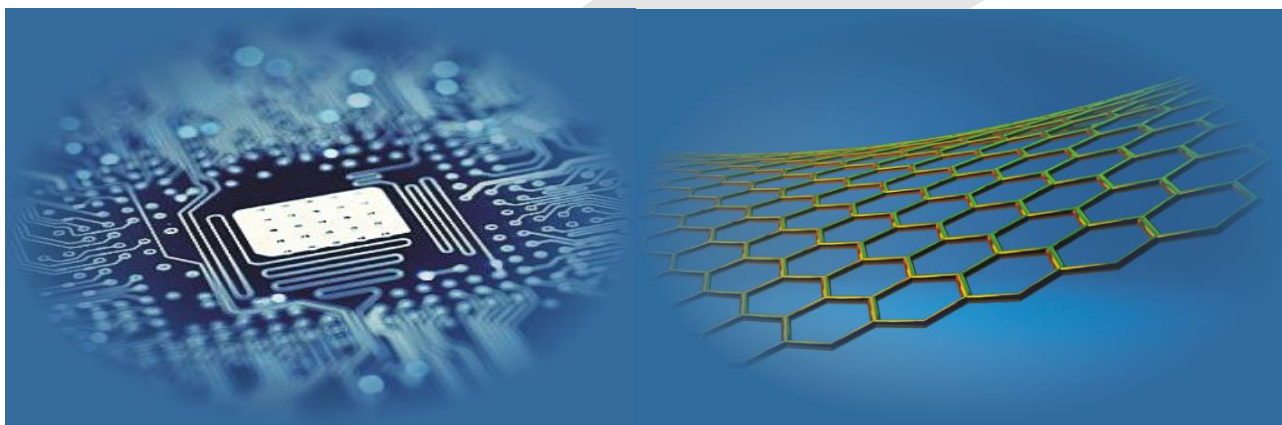# and
New Technologies

2014 Volume 18 No 10

Journal topics:	Publisher	Supporting Organizations
- mathematical and computer modelling - computer and information technologies - natural and engineering sciences - operation research and decision making - nanoscience and nanotechnologies - innovative education	Latvian Transport Development and Education Association	Latvian Academy of Sciences Latvian Operations Research Society Transport and Telecommunication Institute, Latvia Fraunhofer Institute for Factory Operation and Automation IFF, Germany International IT University, Kazakhstan

Articles should be submitted in **English**. All articles are reviewed.

Content

Operation Research and Decision Making

Nature Phenomena and Innovative Engineering

Editors' Remarks

****

A GENERAL SUMMARY

by Rudyard Kipling

We are very slightly changed
From the semi-apes who ranged
India's Prehistoric clay;
He that drew the longest bow
Ran his brother down, you know,
As we run men down to-day.

"Dowb," the first of all his race,
Met the Mammoth face to face
On the lake or in the cave:
Stole the steadiest canoe,
Ate the quarry others slew,
Died -- and took the finest grave.

When they scratched the reindeer-bone,
Some one made the sketch his own,
Filched it from the artist -- then,
Even in those early days,
Won a simple Viceroy's praise
Through the toil of other men.
Ere they hewed the Sphinx's visage
Favouritism governed kissage,
Even as it does in this age.
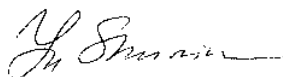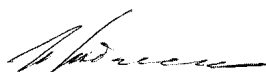
Who shall doubt "the secret hid
Under Cheops' pyramid"
Was that the contractor did
Cheops out of several millions?
Or that Joseph's sudden rise
To Comptroller of Supplies
Was a fraud of monstrous size
On King Pharaoh's swart Civilians?

Thus, the artless songs I sing
Do not deal with anything
New or never said before.
As it was in the beginning
Is to-day official sinning,
And shall be for evermore!

Rudyard Kipling (1809-1849)♣

****

This 18th volume No.10 presents actual papers on main topics of Journal specialization, namely, **Mathematical and Computer Modelling, Computer and Information Technologies, Operation Research and Decision Making and Nature Phenomena and Innovative Engineering.**

Our journal policy is directed on the fundamental and applied sciences researches, which are the basement of a full-scale modelling in practice. This edition is the continuation of our publishing activities. We hope our journal will be interesting for research community, and we are open for collaboration both in research and publishing. We hope that journal's contributors will consider the collaboration with the Editorial Board as useful and constructive.

EDITORS

Yuri Shunin

Igor Kabashkin

♣ **Joseph Rudyard Kipling** (30 December 1865 – 18 January 1936) was an English short-story writer, poet, and novelist. He is chiefly remembered for his tales and poems of British soldiers in India and his tales for children. He was born in Bombay, in the Bombay Presidency of British India, and was taken by his family to England when he was five years old. Kipling is best known for his works of fiction, including The Jungle Book (a collection of stories, which includes and his poems, including "Mandalay" (1890), "Gunga Din" (1890), "The Gods of the Copybook Headings" (1919), "The White Man's Burden" (1899), and "If—" (1910). He is regarded as a major "innovator in the art of the short story"; his children's books are enduring classics of children's literature; and his best works are said to exhibit "a versatile and luminous narrative gift".

The study on steel elliptical cup drawing based on finite element analysis

Jiansheng Xia*, Shasha Dou

Yancheng Institute of Technology, Yancheng City, Jiangsu Province, China

Nanjing University of Aeronautics & Astronautics, Nanjing City, Jiangsu Province, China

Received 1 September 2014, www.cmnt.lv

Abstract

The sheet metal elliptical cup drawing is a complex process, Under the assumption of Prandtl-Reuss flow rule and von Mises yield criterion, the incremental Elasto-plastic large deformation finite element model was established based on the Updated Lagrangian Formulation (ULF).The Elasto-plastic conversions of boundary and deformation are reduced with r_{min} rule. The friction phenomenon of slippage and viscosity at the boundary interface is revised with increment of revision Coulomb rule. The increment rules are led into the whole stiffness matrix, and derived out the stiffness equation. The studies show that the influence on steel elliptical cup drawing deformation is influenced by die structure and parameter. The dates show that finite element simulation and experimental result have a good consistency.

Keywords: elasto-plastic, FEM simulation, elliptical cup drawing

1 Introduction

Sheet metal forming is a common material processing method, which can be divided into stretching, bending, drawing and flanging, etc. The drawing has been widely applied in various applications [1].

Stainless steel is an alloy, which is mainly composed of iron, adding chromium (Cr), nickel (Ni) and other minor elements, and it has some excellent performance, such us: chemical resistance, heat resistance, corrosion. Stainless steel can be roughly divided into 300 and 400 series by chromium and nickel contents. Because of good formability, 300 belongs to Ni series, and drawing parts are used in building materials, kitchen, utensils, pipe fittings, etc. Because of good hardness, 400 belongs to Cr series, and drawing parts are used as decorative purposes or made into tableware, mechanical parts [2].

Zaky [3] 1998 studied the sheet deep drawing of low carbon steel and aluminous plate, found the ideal contour parameters which does not have the lug, and predicted the anisotropy of sheet in forming process.

Hah [4] 2000 studied the failure reasons in the process of drawing forming based on analysis of shell element of finite element, and found the local deformation in long axis, wrinkling in short axis. The reason is the location problem in forming, such us different drawing rates, wrong position between the sheet and die.

Kim [5] 2001 studied the wrinkles problems in the elliptical cup deep drawing based on finite element analysis, proposed the bifurcation-theory for the first time. He mentioned the reasons of causing wrinkles, such us the stress state, mechanical properties, geometry structure, and contact part of blank and die.

In this paper, the steel sheet 304 cup is analysed with the finite element method, some relationships are studied, such us: relationship between punch load and displacement, distribution of stress and strain, distribution of thickness, and verify by actual experiments. It used to reference for operation process and altered design. Ductile fracture criterion to predict zirconium sheet stretch forming limit. The results show that the extension of zirconium sheet is high, but the stretch is low.

2 Fundamental theory

2.1 VIRTUAL WORK PRINCIPLE

It descripts the elastic-plastic deformation with the updated Lagrangian formulation ULF [6], the Virtual work principle formulation can be shown as follows:

$$\int_{V^E}\left(\overset{\triangledown}{\sigma}_{ij} - e\sigma_{ik}\dot{\varepsilon}_{kj}\right)\delta\dot{\varepsilon}_{ij}dV + \int_{V^E}\sigma_{jk}L_{ik}\delta L_{ij}dV = \int_{S_f}\dot{f}\delta v_i dS \,, (1)$$

where, $\overset{\triangledown}{\sigma}_{ij}$ is the Cauchy stress tensor, $\dot{\varepsilon}_{kj}$ is the rate of stress tensor, $\dot{\varepsilon}_{ij}$ is the strain tensor, σ_{jk} is the rate of strain tensor, $\delta\dot{\varepsilon}_{ij}$ is the virtual strain tensor of the point, δL_{ij} is the virtual velocity gradient tensor of the point, δv_i is the velocity component, $\dot{f}$ is surface force component, L_{ij} is velocity gradient tensor, V is unit volume, S is unit surface area.

Xia Jiansheng, Dou Shasha

2.2 CONSTITUTIVE RELATION

In preparing the theory of elasto-plasticity, we have made certain assumptions [10]:

1) The material is homogeneous and isotropic;

2) There is no strain before manufacturing;

3) Temperature effect do not consider when manufacturing;

4) It obeys the laws of the Hooke's Law in elastic stage;

5) It obeys the von Mises yield rule and Prandtl-Reuss plastic flow rule;

6) It contains Isotropic strain hardening in constitutive equation;

7) There are elastic strain stage and plastic strain stage in material strain rate;

8) Punch, die and holder are steel structure;

9) The Bauschinger effect do not consider in reverse unloading.

After assuming above, the constitutive relation can be written as follow:

$$\overset{\circ}{\sigma}_{ij} = C_{ijmn}^{ep} \dot{\varepsilon}_{mn}, \tag{2}$$

$$C_{ijmn}^{ep} = C_{ijmn}^{e} - \frac{C_{ijkl}^{e} C_{uv}^{e} \dfrac{\partial f}{\partial \sigma_{kl}} \dfrac{\partial f}{\partial \sigma_{uv}}}{C_{kluv}^{e} \dfrac{\partial f}{\partial \sigma_{kl}} \dfrac{\partial f}{\partial \sigma_{uv}} + H' \dfrac{\sigma_{uv} \dfrac{\partial f}{\partial \sigma_{uv}}}{\bar{\sigma}}}, \tag{3}$$

where $\overset{\circ}{\sigma}_{ij}$ is Jaumann differential of σ_{ij}, C_{ijmn}^{ep} is the elastic-plastic module, C_{ijmn}^{e} is Elastic module, f is the initial yield function, H' is the strain hardening rate, $\bar{\sigma}$ is Von Mises yield function, so the Matrix form of C_{ijmn}^{ep} can be expressed as bellow:

$$\left[C^{ep} \right] = \left[C^{e} \right] -$$

$$\frac{1}{S} \begin{bmatrix} S_1^2 & S_1 S_2 & S_1 S_3 & S_1 S_4 & S_1 S_5 & S_1 S_6 \\ & S_2^2 & S_2 S_3 & S_2 S_4 & S_2 S_5 & S_2 S_6 \\ & & S_3^2 & S_3 S_4 & S_3 S_5 & S_3 S_6 \\ & & & S_4^2 & S_4 S_5 & S_4 S_5 \\ & & & & S_5^2 & S_5 S_6 \\ symm & & & & & S_6^2 \end{bmatrix}, \tag{4}$$

where:

$$S = \frac{4}{9} \bar{\sigma}^2 H' + S_1 \sigma'_{xx} + S_2 \sigma'_{yy} + S_3 \sigma'_{zz} +$$
$$2 S_4 \sigma'_{yz} + 2 S_5 \sigma'_{zx} + 2 S_6 \sigma'_{xy}, \tag{5}$$

$$S_1 = 2G \sigma'_{xx}, \ S_2 = 2G \sigma'_{yy} \ S_3 = 2G \sigma'_{zz}, \tag{6}$$

$$S_4 = 2G \sigma'_{yz}, \ S_5 = 2G \sigma'_{zx}, \ S_6 = 2G \sigma'_{xy}, \tag{7}$$

where σ'_{ij} is deviator of σ_{ij}, G is the friction flow potential, $G = \sigma_1^2 + \sigma_2^2$, $\left[C^e \right]$ is the equation in minimum strain, which can be expressed as bellow:

$$\left[C^e \right] = \frac{E}{1+v} \begin{bmatrix} \dfrac{1-v}{1-2v} & \dfrac{1-v}{1-2v} & \dfrac{1-v}{1-2v} & 0 & 0 & 0 \\ & \dfrac{1-v}{1-2v} & \dfrac{1-v}{1-2v} & 0 & 0 & 0 \\ & & \dfrac{1-v}{1-2v} & 0 & 0 & 0 \\ & & & \dfrac{1}{2} & 0 & 0 \\ & & & & \dfrac{1}{2} & 0 \\ symm & & & & & \dfrac{1}{2} \end{bmatrix}, \tag{8}$$

where E is modulus of elasticity, v is Poisson's ratio. If the material is homogeneous and isotropic, the Elasto-plastic rate equation can be written:

$$\overset{\circ}{\sigma}_{ij} = \frac{E}{1+v} \left[\delta_{ik} \delta_{jl} + \frac{v}{1-2v} \delta_{ij} \delta_{kl} - \frac{3\alpha \left(\dfrac{E}{1+v} \right) \sigma'_{ij} \sigma'_{kl}}{2\bar{\sigma}^2 \left(\dfrac{2}{3} H' + \dfrac{E}{1+v} \right)} \right] \dot{\varepsilon}_{kl} . \tag{9}$$

If $\alpha = 1$, it is a plastic stage; if $\alpha = 0$, it is an elastic stage or unloading stage.

Equivalent stress and equivalent plastic strain relations can express by n-power law equation:

$$\dot{\bar{\sigma}} = C \left(\varepsilon_0 + \dot{\varepsilon}_p \right)^n, \tag{10}$$

where C is material constant, n is strain hardening index; $\dot{\bar{\sigma}}$ is the equivalent stress, $\dot{\varepsilon}_p$ is the equivalent plastic strain, ε_0 is the initial strain.

2.3 THE FINITE ELEMENT EQUATION

Finite element analysis is the method that the structure is divided into many small units called discrete entity. Based on Large deformation stress and stress rate relation, the finite deformation of Update Lagrangian Formulation, material constitution relationship, the velocity distribution of each unit is show bellow:

$$\{v\} = \leq [N] \{\dot{d}\}, \tag{11}$$

$$\{\dot{\varepsilon}\} = \leq [B] \{\dot{d}\}, \tag{12}$$

$$\{L\} = \leq [M] \{\dot{d}\}, \tag{13}$$

8

where $[N]$ is shape function, $\{\dot{d}\}$ is nodal velocity, $[B]$ is strain rate-velocity matrix, $[M]$ is velocity gradient-velocity matrix.

The principle of virtual work equation and the constitutive equation based on update Lagrangian is linear equation. The formula can be written by the form of incremental representation.

After finite element discrimination, the large deformation rigid general equation is written as bellow:

$$[K]\{\Delta u\}=\{\Delta F\}, \qquad (14)$$

where:

$$Q = \begin{bmatrix} 2\sigma_{xx} & 0 & 0 & \sigma_{xy} & 0 & \sigma_{xz} \\ & 2\sigma_{yy} & 0 & \sigma_{xy} & \sigma_{zy} & 0 \\ & & 2\sigma_{zz} & 0 & \sigma_{zy} & \sigma_{xz} \\ & & & \frac{1}{2}(\sigma_{xx}+\sigma_{yy}) & \frac{1}{2}(\sigma_{zx}) & \frac{1}{2}(\sigma_{zy}) \\ & & & & \frac{1}{2}(\sigma_{yy}+\sigma_{zz}) & \frac{1}{2}(\sigma_{xy}) \\ symm & & & & & \frac{1}{2}(\sigma_{xx}+\sigma_{zz}) \end{bmatrix}, \qquad (17)$$

$$Z = \begin{bmatrix} \sigma_{xx} & 0 & 0 & \sigma_{xy} & \sigma_{xz} & 0 & 0 & 0 & 0 \\ & \sigma_{yy} & 0 & 0 & 0 & \sigma_{xy} & \sigma_{yz} & 0 & 0 \\ & & \sigma_{zz} & 0 & 0 & 0 & 0 & \sigma_{xz} & \sigma_{yz} \\ & & & \sigma_{yy} & \sigma_{yz} & 0 & 0 & 0 & 0 \\ & & & & \sigma_{zz} & 0 & 0 & 0 & 0 \\ & & & & & \sigma_{xx} & \sigma_{xz} & 0 & 0 \\ & & & & & & \sigma_{zz} & 0 & 0 \\ & & & & & & & \sigma_{xx} & \sigma_{xy} \\ symm & & & & & & & & \sigma_{yy} \end{bmatrix}. \qquad (18)$$

$$[K] = \sum_{\langle E\rangle} \int_{V^e} [B]^T \left([C^{ep}] - [Q] \right) [B] dV +$$
$$\sum_{\langle E\rangle} \int_{V^e} [E]^T [Z][E] dV, \qquad (15)$$

$$\{\Delta F\} = \sum_{\langle E\rangle} \int_{S^e} [N]^T \{\dot{f}\} dS\Delta t, \qquad (16)$$

$[K]$ is the overall Elasto-plastic stiffness matrix, $\{\Delta F\}$ is the nodal displacement increment, $\{\Delta u\}$ is the nodal forces incremental, $[Q]$ and $[Z]$ are stress correction matrix.

2.4 FRICTION PROCESSING

There is friction in sheet forming process, so we need to pay attention to materials and tools of the interface conditions [11]. When the material moves along the tool surface curve of the slide, the contact force can be expressed as:

$$F = F_l l + F_n n, \qquad (19)$$

where, F_l is radial force and F_n is normal force, and differential equation of F can be expressed as:

$$\dot{F} = \dot{F}_l l + F_l \dot{l} + \dot{F}_n n + F_n \dot{n}, \qquad (20)$$

where, differentials of $\dot{l}$ and $\dot{n}$ are expressed as:

$$\dot{l} = -\Delta u_l^{rel} / R, \qquad (21)$$

$$\dot{n} = \Delta u_l^{rel} / R, \qquad (22)$$

where, R is tool radius, Δu_l^{rel} is the local relative velocity between the tool and node, and the nodes relative speed can be expressed as:

$$\Delta u_l^{rel} = \Delta u_l - \dot{u}_{tool} \sin\theta, \qquad (23)$$

where, Δu_l is the contact tangent displacement increment of nodes, $\dot{u}_{tool}$ is the displacement increment of tooling, θ is the rotation angle.

The increment equation of $\dot{F}$ is expressed as follow:

$$\dot{F} = \left(\dot{F}_l - F_n \Delta u_l / R + F_n \dot{u}_{tool} \sin\theta / R \right) \cdot l +$$
$$\left(\dot{F}_n - F_l \Delta u_l / R - F_l \dot{u}_{tool} \sin\theta / R \right) \cdot n. \qquad (24)$$

Rigid matrix governing equation of the contact nodes is expressed bellow:

$$\begin{bmatrix} K & ... & \\ ... & K_{11} + F_n / R & K_{12} \\ ... & K_{21} + F_n / R & K_{22} \end{bmatrix} \begin{Bmatrix} ... \\ \Delta u_l \\ \Delta u_n \end{Bmatrix} = \begin{Bmatrix} ... \\ \dot{F}_l + F_n \dot{u}_{tool} \sin\theta / r \\ \dot{F}_l - F_l \dot{u}_{tool} \sin\theta / r \end{Bmatrix} . \quad (25)$$

2.5 INCREMENTAL STEPS OF r_{min} METHOD

Using the elastic plastic finite element method with large deformation method, also called the Yamada r_{min} method. Each incremental step value is equal to incremental displacement of initial deformation increment of the tooling. Adopting the method of updated Lagrangian formulation, calculating each increment of displacement, strain, stress, load, springback value after forming the final shape of sheet mental in unloading condition, the value of load incremental in each step is controlled by r_{min} equation, which is shown as bellow:

$$r_{min} = MIN(r_1, r_2, r_3, r_4, r_5) , \quad (26)$$

where, r_1 is The maximum allowable strain increment, r_2 is the maximum allowable rotation increment, r_3 is the minimum value in all elastic elements, r_4 is contact position between free node and tooling, r_5 is discontent position between free node and tooling.

3 Numerical analysis flow

Based on the finite deformation theory, ULF equation and r_{min} method, a set of effective analysis of sheet metal forming process is established. Firstly, a 3d part and mold is designed with the NX software, and then mesh them with NASTRAN software. Secondly, the meshed models are drawn into the data file and did finite element analysis. The simulation flow chart is shown in Figure 1.
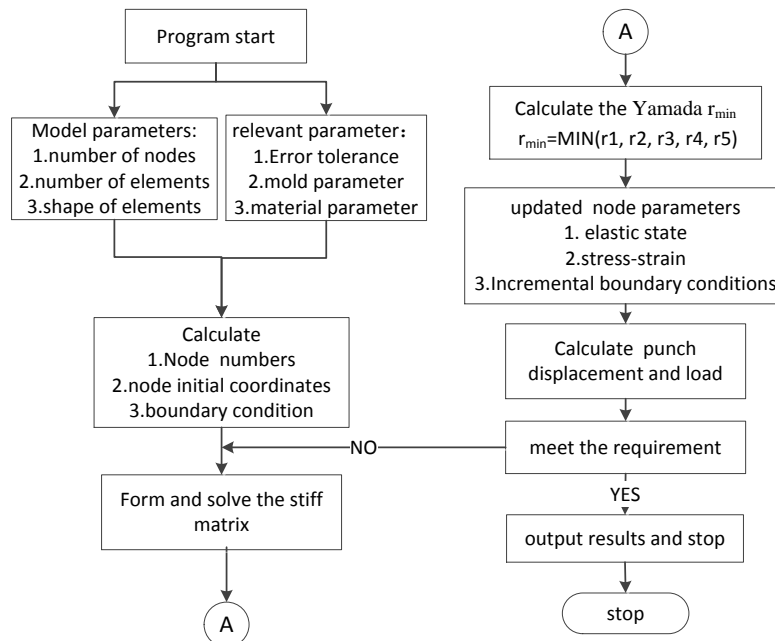
FIGURE 1 Numerical simulation of flow chart

Based on the theory upwards, the research of steel elliptical cup drawing is studied, including relationship between the punch load and displacements, stress and strain, thickness, spring-back and warpage. Simulation experimental parameters were carried out, which are friction coefficient (μ), punch radius (r_p), die radius (R_d). The parameters of warpage problems are verified by the experiment are optimized and served a reference for drawing designer.

The whole structure is composed of punch, die and blank holder. The model picture was shown as Figure 2.

The initial relation of part and die is shown in Figure 2a, also, the punch down a certain travel case is shown Figure 2b. It takes two coordinates to solve the problem, which are fixed coordinates (X, Y, Z) and local coordinates (ξ, η, ζ). It uses the fixed coordinates (X, Y, Z) when nodes do not contact with the tool, and uses the local coordinates (ξ, η, ζ) when nodes contact with the tooling. Using coordinates rule based on the right-hand rule. L-axis is the

tangential direction of contact line between the part and tools when n-axis is the normal direction.
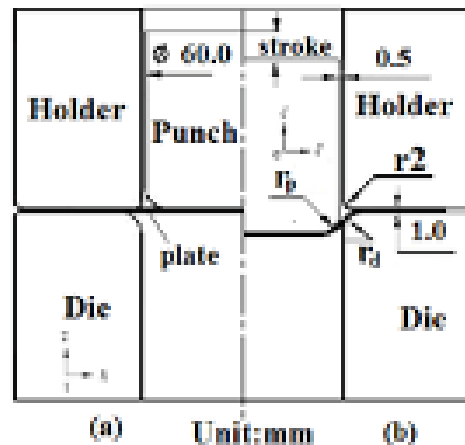
FIGURE 2 Sheet metal and die size chart: a) before deformation, b) after deformation

10

The contact condition of each node of plates will change depending on deformation in sheet metal forming. When the displacement increment is zero, the boundary conditions of increment displacement of the next node will changes to free node boundary conditions. When sheet contacts the tools, contact condition is changed to the boundary conditions, which bases on the generalized r_{min} method.

The SUS304 material stainless steels are provided by a steel Crop, of which the chemical and mechanical properties as shown in Table1 as bellow [9].

TABLE 1 Mechanical properties of SUS304

stress-strain relationship: $\bar{\sigma}=1400\left(0.024+\bar{\tau}_p\right)$	
initial thickness: $t = 0.8mm$	Poisson's ratio: $\upsilon = 0.3$
yield stress: $\sigma y = 315MPa$	coefficient of friction: $\mu = 0.1$
Yang coefficient:	fractured thickness:
$E = 2.1\times10^5MPa$	$tf = 0.536mm$

Because of the symmetrical sheet model, 1/4 model is taken to analysis.

It uses the quadrilateral segmentation of degenerated shell element in sheet metal meshing, when the die meshing uses the triangle segmentation (Table 2).

TABLE 2 The finite element mesh dates of punch, die and sheet

Name	Element type	Element number	Node number
Punch	Triangle	1356	569
Die	Triangle	1638	749
Holder	Triangle	813	426
Blank	Four-node shell element	1478	1321

4 Results

4.1 DISTRIBUTION OF STRESS AND STRAIN

The analysis in elliptical cup drawing process generally can be divided into three stages.

The first stage: Punch and sheet mental just begins to contact, when the stroke is 5mm. The stress distribution of elliptical cup drawing is as shown in Figure 3a. As can be seen from the figure, the normal stress occurs around the punch fillet, the concentration tensile stress appears in long axis centre.

The second stage: when the stroke is 30.0mm, the stress distribution of elliptical cup drawing is as shown in Figure 3b. As can be seen from the figure, the side wall of the sheet gradually forms to straight wall of elliptical cup along the punch. As the sheet flow into the die and deform plastically the punch load increases. The normal stress also occurs in the sheet between the holder and punch, and tensile stress occurs in the straight wall of elliptical cup which had flowed into the die.

The third stage: this is the finish stage in drawing. The stress distribution of elliptical cup drawing is as shown in Figure 3c, the stress concentration occurred in punch fillet, and the maximum value is in the top of the cup, which is 1560MPa.
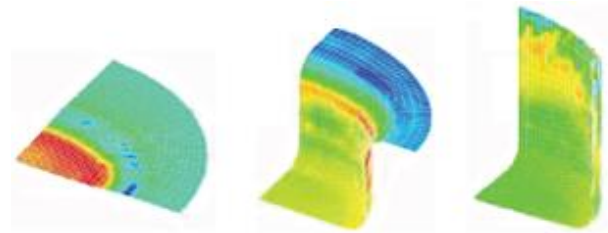
FIGURE 3 Stress distribution: a) stroke = 0.05mm, b) stroke = 30mm, c) finish stroke

4.2 PUNCH LOAD ANALUSIS ANALYSIS

When the elliptical cup forming, the sheet will be under a lot of stress, such us the tensile load of the punch moving down, the friction between the die and sheet, the compression force caused by sheet diameter reduced gradually. The punch force increases with the punch stroke firstly. The maximum value is 78kN, when the stroke is 28mm. Next, the value decreases with the increase of stroke. Numerical analysis and experimental results are as shown in Figure 4.
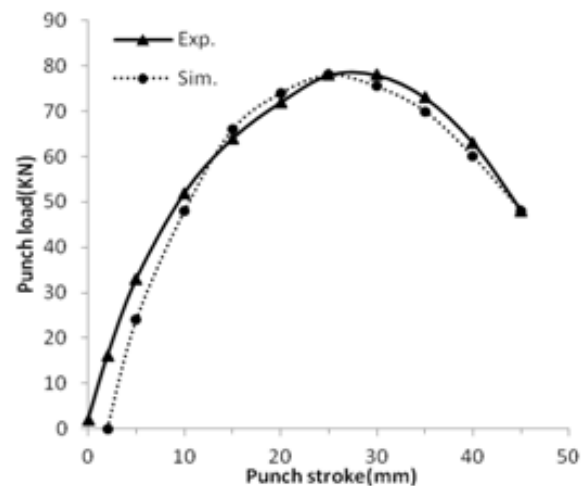
FIGURE 4 The relationship of punch load and stroke

4.3 THICKNESS CHANGE ANALYSIS

The sheet mental is used to do drawing experiment analysis, of which the diameter is 105mm. The thickness data are measure along the direction of long-axis and short-axis and as shown in Figure 5. As can be seen from the figure, the thinnest thickness is in the bottom area near the long-axis fillet, which is the 0.58mm and slightly larger than the rapture thickness, then the drawing part can be finished.

a) The thickness in long-axis

b) The thickness in short-axis
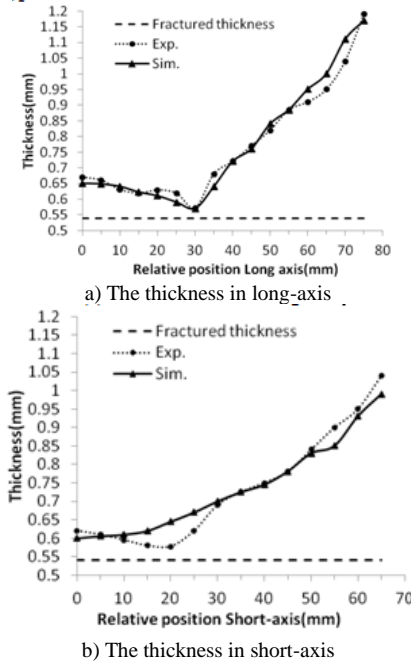
FIGURE 5 Thickness change analysis in long axis and short axis

4.4 FORMING LIMIT ANALYSIS

In order to find out the limit thickness of elliptical flange, some experiments are carried out with the different diameters of sheet, and set the value 0.41mm as the criterion of rupture. If the thickness is thinner, the program will judge that the sheet is in the limit thickness. FR (Forming ratio) LFR (Limit Forming Ratio), and EFR (Excessive Forming Ratio) are referred by Huang and Chen [9].

DR (Drawing Ratio), LDR (Limit Drawing Ratio) and EDR (Excessive Drawing Ratio) are used to calculate and judge, which are defined by Huang [9]:

$$FR = \frac{D_p}{D_h} = \frac{C_p}{C_h}, \qquad (27)$$

$$LFR = \frac{D_P}{D_{h,\min}} = \frac{C_P}{C_{h,\min}}, \qquad (28)$$

$$EDR = \frac{D_p}{D_{h,f}} = \frac{C_p}{C_{h,f}}, \qquad (29)$$

where, D_p is the diameter of punch; D_h is the diameter of sheet with necking or fracture; $D_{h,\min}$ is the maximum diameter of sheet without necking or fracture; D_p is the perimeter of the elliptical punch; $D_{h,f}$ is the perimeter of the sheet without necking or fracture; C_p is the circumference of punch, C_h is the circumference of initial inner elliptical hole with necking or fracture; $C_{h,\min}$ is the maximum the circumference of initial inner elliptical hole without necking or fracture; $C_{h,f}$ is circumference of initial inner elliptical hole with part necking or fracture.

The experimental arrangement of sheet blank diameter: φ105mm → φ106mm → φ107mm, and definition DR = 2.116 (blank diameter = 105mm), LDR=2.135 (blank diameter = 106mm) and EDR = 2.168 (blank diameter = 107mm).

The thickness distribution of long and short axis which the blank diameter is 107mm is shown in Figure3. As can be seen from the figure, because of pinching-out, there is necking and rupture at the relative position 29mm in the long-axis.

a) Long-axis direction

b) Short-axis direction
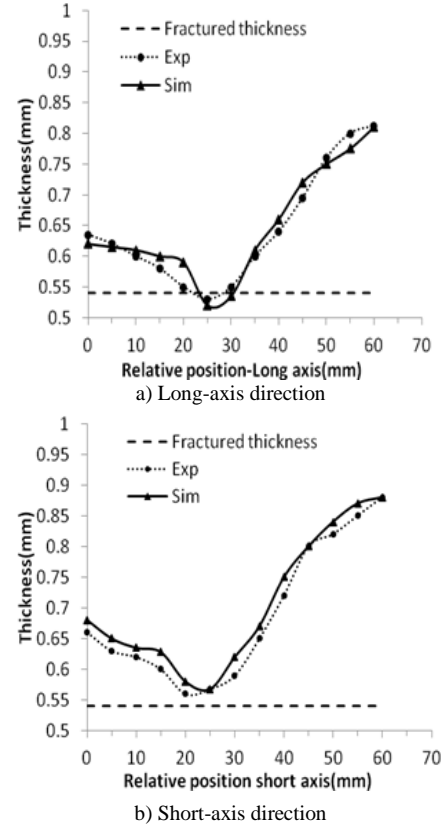
FIGURE 6 The distribution of thickness

The other simulation experiments are carried out. It is found that the minimum thickness is 0.55mm along the long-axis or short-axis with the blank diameter 106mm. The thickness is very close to the rupture, and reaches the limit.

4.5 LIMIT DRAWING COEFFICIENT ANALYSIS

In order to study the relationship between the limit drawing ratio and punch or die fillet radius, eight groups' experiments are carried out. The punch radius and die fillet are 3.0mm, 5.0mm, 7.0mm and 9.0mm, respectively. The results are as shown in Table3 as bellow.

As can be seen from the Figure 7, when punch radius increases from 3.0mm to 9.0mm, the limit drawing ratio also increases from LDR=2.1mm to LDR=2.157. When die radius increases from 3.0mm to 9.0mm, the limit drawing ratio LDR also increases from 2.07 mm to 2.181mm.

TABLE 3 Limit drawing ratio with the different radius of punch and die

Radius(mm)	3.0	5.0	7.0	9.0
Rp=5.0mm	2.12	2.135	2.148	2.156
Rd=5.0mm	2.06	22.135	2.163	2.18
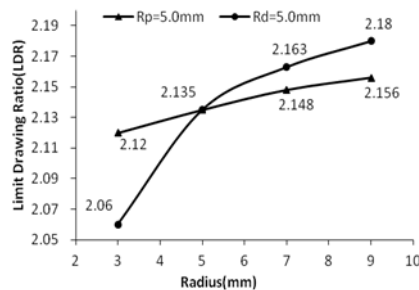
FIGURE 7 The relationship between the limit drawing ratio and radius of punch or die

5 Conclusion

Based on the numerical analysis and experimental results, combined with finite element method with the incremental Elasto-plastic theory, analysed warpage phenomenon in stainless steel elliptical cup drawing process, the following conclusions are obtained.

1) The punch load increase with the increase of punch stroke. When the stroke is 27.0mm, the maximum punch load is 78.0kN, which is close to actual experimental data.

2) The position of minimum thickness focuses on the contact of sheet and punch in long-axis, because there is the maximum tensile stress in this area. The thickness is little change in short-axis, because the larger radius and smaller tangential tensile stress.

3) The limit drawing ration has a relationship with the radius of fillet, which increases with the increase of radius of punch and die.

4) Maximum punch stroke increases with increasing punch radius and decreasing of initial elliptical hole diameter.

5) As the forming limit defined by the inner circumference and diameter of elliptical punch, the experimental flange forming limit ratio of 1.46.

Acknowledgment

References

[1] Xia J S, Dou S S 2013 Experimental Research on the Flanging Height Adjustment Based on Environmental Way *International Journal of Applied Environmental Sciences* **8**(20) 2470-89

[2] Huang Y M, Chen T C 2005 Influence of Blank Profile on the V-Die Bending Camber Process of Sheet Metal International *Journal of Advanced Manufacturing Technology* **25**(7) 668-77

[3] Zaky A M 1998 Optimum Blank Shape of Cylindrical Cups in Deep Drawing of Anisotropic Sheet Metals *Journal of Materials Processing Technology* **76**(1) 203-211

[4] Huh H, Kim S H, Kim S H 2000 Process Design For Multi-Stage Elliptic Cup Drawing with the Large Aspect Ratio *European Congress on Computational Methods in Applied Sciences and Engineering* **8**(11) 1-16

[5] Kim H B, Yoon J W, Yang D Y, Barlat F 2001 Investigation into Wrinkling Behavior in the Elliptical Cup Deep Drawing Process by Finite Element Analysis using Bifurcation-Theory *Journal of Materials Processing Technology* **111**(1) 170-4

[6] Huang Y M, Chien K H 2001 Influence of the Punch Profile on the Limitation of Formability in Hole-Flanging Proces *Journal of Materials Processing Technology* **113**(1) 720-4

[7] Takuda H, Hatta N 1998 Numerical Analysis of Formability of a Commercially Pure Zirconium Sheet in Some Forming Processes *Materials Science and Engineering* **242**(1) 15-21

[8] Leu D K1996 Finite-Element Simulation of Hole-Flanging Process of Circular Sheets of Anisotropic Materials *International Journal of Mechanical Sciences* **38**(8) 917-33

[9] Huang Y M, Chien K H 2002 Influence of Cone Semi-Angle on theFormability Limitation of the Hole-Flanging Process *International Journal of Advanced Manufacturing Technology* **19**(8) 597-606

[10] Sousa L C, Castro C F, Antonio C A C 2006 Optimal Design of V and U Bending Processes Using Genetic Algorithms *Journal of Material Processing Technology* **172**(1) 35-41

[11] Liu G, Lin Z, Xu W, Bao Y 2002 Variable Blankholder Force in U-Shaped Part Forming for Elimination Springback Error *Journal of Material Processing Technology* **120** 259-64

[12] Samuel M 2000 Experimental and numerical prediction of springback and side wall curl in u-bending of anisotropic sheet metals *Journal of Materials Processing Technology* **105** 382-93

Authors

Jiansheng Xia, born in September, 1980, Yancheng, Jiangsu Province, P.R. China

Current position, grades: lecturer at Yancheng Institute of Technology, China.
University studies: MSc at Northeast Agricultural University in China.
Scientific interests: mold design, CAD/CAE/CAM technologies.
Publications: more than 10 papers.
Experience: teaching experience of 8 years, 2 scientific research projects.

Shasha Dou, born in December, 1982, Yancheng, Jiangsu Province, P.R. China

Current position, grades: lecturer at Yancheng Institute of Technology, China.
University studies: MSc at Shandong University of Technology in China.
Scientific interests: plastic mold design, CAD/CAE/CAM technologies.
Publications: more than 10 papers.
Experience: teaching experience of 8 years, 1 scientific research project.

Dynamic modelling and small signal analysis of push-pull bidirectional DC-DC converter

Hongjuan Zhang, Long Quan*, Yan Gao

Key Lab of Advanced Transducers and Intelligent Control System, Ministry of Education, Taiyuan University of Technology, Taiyuan 030024, China

Abstract

Bidirectional DC-DC converter can not only act as a contact bridge between two different voltage levels systems, but also can achieve the energy flow in both directions. However, due to the operating characteristics of the active switch and diode in the converter, thereby the inverter becomes a strong nonlinear circuit. In light of this, the push-pull bidirectional DC-DC converter is designed and the method of state space small signal analysis is proposed. First of all, the state-space equations of the converter are established. Then each variable of inverter circuit is averaged in a switching cycle to eliminate the influence of the switching ripple. The respective average variable expressions are decomposed into the DC component and AC small signal component. After the DC component of the signal is eliminated, the expression of the AC small signal component is obtained, to achieve the purpose of separating the small signal. Finally, the expression of small signal component is linearized, thereby the nonlinear system is approximated the DC operating point. This study will provide the theoretical foundation ready for the analysis and design of converter controller.

Keywords: dynamic modelling, state space averaging method, small signal analysis, linearization

1 Introduction

As the development of the new energy technology, transmission technology, the power supply technology and motor driver technology, the bidirectional DC-DC (Direct Current to Direct Current) converter technology becomes a new branch of power electronic technology. Not only the bidirectional DC-DC converter can transform the voltage, but also can achieve two-way power flow. So it has been widely used in all kinds of occasions. At the same time, it also becomes a research hotspot [1].

The isolated bidirectional DC-DC converter is a kind indirect transform circuit, which consists of the inverter link, high-frequency transformer and rectifier. Because the push-pull converter has many advantages of simple structure, transformer winding bidirectional excitation, the small switch conduction loss and simple drive circuit, that it is widely used in small and medium power switching power supply [2]. In the DC converter working process, due to the active switching devices and diodes in converter is working in a wide range of its characteristic curve, so that the converter is a strong nonlinear circuit. In order to improve the control performance, work efficiency, and reduce loss, many scholars have done a lot of research in domestic and foreign countries [3]. The several typical topologies of isolated bidirectional DC-DC converters were presented. And the circuit structure and working principle were analysed [4]. The ZVS push-pull resonant converter has been proposed and the half bridge rectifier has been adopted in the second of transformer [5]. The bidirectional isolated DC-DC converter has been presented, combined push-pull-forward circuit, half-bridge circuit

with high-frequency transistors [6]. The state feedback control of push-pull DC-DC converter has been proposed by frequency-domain conditions based on the Hermite-Biehler theorem, and the robust stability has been analysed [7]. A resonant isolation push-pull converter was investigated. In order to improve the system efficiency, a switching loss analysis was performed [8].

In the above analysis and research, the full bridge rectifiers are adopted in the rectifier link of the push-pull bi-directional DC-DC converter, or the half bridge rectifiers are used. More switching devices are used in circuit. And the converters are mainly applied in DC power output voltage above 100V. In view of the energy needs of two-way conversion in the medium power applications, a push-pull bi-directional DC-DC converter is designed in the paper. The topology of each variable in a switch period of work modal is analysed. The analysis method of the state space small signal is adopted. The AC (Alternation Current) small signal state space expression of push-pull bidirectional DC-DC converter is established and is linearized. These works lay the foundation for the further design of control model.

2 Working principle and dynamic modelling

The push-pull isolated bidirectional DC-DC converter introduces a pair of symmetrical switches in the output terminal of common push-pull DC-DC converter. So the both ends of the circuit structure become symmetry, and the two-way flow of energy can be realized. Figure 1 below shows the push-pull bidirectional DC/DC converter.

---

* *Corresponding author* e-mail: quanlong@tyut.edu.cn
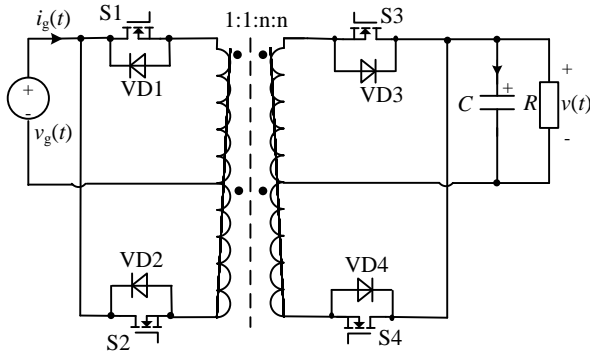
FIGURE 1 Push-pull bidirectional DC-DC converter

In order to realize the two-way flow of energy, the converter must operate in continuous mode. When the energy transmitted from the left side to the right side, the converter has four kinds of switch modes in a switch period. The switch S1 and S2 alternately turn on and respectively formed the opposite phase AC voltage in the side of two windings. When the switch S1 is turned on, the diode VD3 is in conducting state. When the switch S2 is turned on, the diode VD4 is in conducting state. When the two switches are turned off, the diode VD3 and VD4 are all in the on state, each shares a half of the current. When the energy transmitted from the right side to the left side, the working state is same with the above situation because of the circuit symmetry. When the switch S3 is turned on, the diode VD1 is in conducting state. When the switch S4 is turned on, the diode VD2 is in conducting state. When the two switches are turned off, the diode VD1 and VD2 are all in the on state, each shares a half of the current.

The magnetizing inductance of transformer is involved in energy transfer, so the actual transformer is replaced by the magnetizing inductance and the ideal transformer parallel model. Considering the internal resistance of magnetizing inductance, the analysis circuit of push-pull bidirectional DC-DC converter is shown in Figure 2.
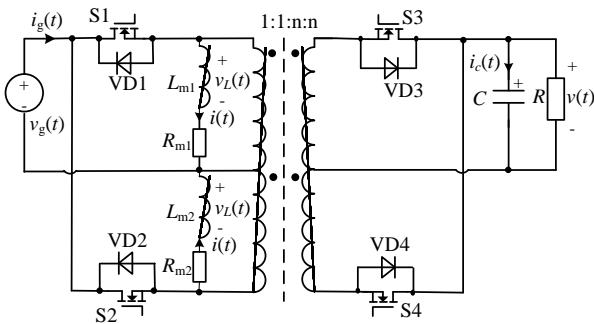
FIGURE 2 Analysis circuit of push-pull bidirectional DC-DC converter

The timing control rules of switch driving in the push-pull bidirectional DC-DC converter are shown in Figure 3. The duty ratio $d(t)$ is a controlled variable and is defined as the conduction time and the cycle ratio, $d(t)=t_{on}/T_s$. According to the control law of the switch, the state equations of the converter are calculated in four kinds of operating condition. The inductor current $i(t)$ and $v(t)$ capacitor voltage are selected as the state vector, and the state vector is expressed for $x(t)=[i(t), v(t)]^T$. The input

current $i_g(t)$ and the output voltage $v(t)$ are used as the output vector, and the output vector is expressed for $y(t)=[i_g(t), v(t)]^T$. The input voltage $v_g(t)$ is uses as the input vector, and the input vector is expressed for $u(t)= [v_g(t)]$.
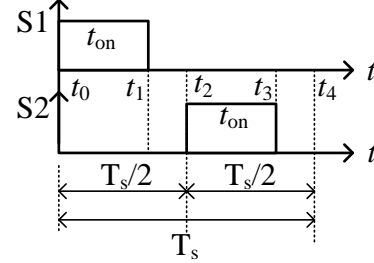
FIGURE 3 Switching sequence of S1 and S2

The working mode 1: $t_0 \sim t_1$ time interval is for $[0, d(t)T_s]$. In this period, the switch S1 is turned on, and the diode VD3 is on state. The current flows through the transformer, diode VD3, a filter capacitor C and load R. At the same time the current value increases. The topological structure is shown in Figure 4 in this period. Due to symmetry circuit, a magnetizing inductances are supposed for $L_{m1}=L_{m2}=L_m$, and the inductive resistances are supposed for $R_{m1}=R_{m2}=R_m$.

The state equation is expressed for:

$$\begin{cases} v_L(t) = L_m \dfrac{di(t)}{dt} = v_g(t) - i(t)R_m \\ i_c(t) = C \dfrac{dv(t)}{dt} = \dfrac{i(t)}{n} - \dfrac{v(t)}{R} \end{cases}. \qquad (1)$$

The output equation is expressed for:

$$\begin{cases} i_g(t) = i(t) \\ v(t) = v(t) \end{cases}. \qquad (2)$$

Can be derived for:

$$\begin{cases} \dot{x}(t) = \begin{bmatrix} \dot{i}(t) \\ \dot{v}(t) \end{bmatrix} = A_1 \begin{bmatrix} i(t) \\ v(t) \end{bmatrix} + B_1 \begin{bmatrix} v_g(t) \end{bmatrix} \\ y(t) = \begin{bmatrix} i_g(t) \\ v(t) \end{bmatrix} = C_1 \begin{bmatrix} i(t) \\ v(t) \end{bmatrix} + E_1 \begin{bmatrix} v_g(t) \end{bmatrix} \end{cases}, \qquad (3)$$

where $A_1 = \begin{bmatrix} -\dfrac{R_m}{L_m} & 0 \\ \dfrac{1}{nC} & -\dfrac{1}{RC} \end{bmatrix}$, $B_1 = \begin{bmatrix} \dfrac{1}{L_m} \\ 0 \end{bmatrix}$, $C_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$,

$E_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$.
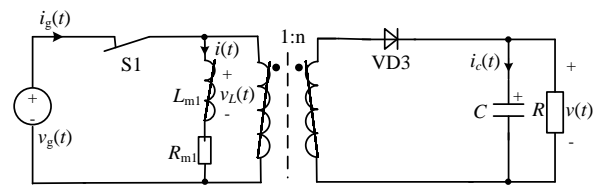
FIGURE 4 Topological structure in $t_0 \sim t_1$

The working mode 2: $t_1 \sim t_2$ time interval is for $[d(t)T_s, T_s/2]$. In this period, the switches S1 and S2 are turned off. The freewheeling current flows through the diode VD3 and the diode VD, The current value is respectively half of total current. At the same time the current value decreases. The topological structure is shown in Figure 5 in this period.

The state equation is expressed for:

$$\begin{cases} v_L(t) = L_m \dfrac{di(t)}{dt} = \dfrac{v(t)}{n} - i(t)R_m \\ i_c(t) = C \dfrac{dv(t)}{dt} = \dfrac{i(t)}{n} - \dfrac{v(t)}{R} \end{cases}. \tag{4}$$

The output equation is expressed for:

$$\begin{cases} i_g(t) = 0 \\ v(t) = v(t) \end{cases}. \tag{5}$$

Can be derived for:

$$\begin{cases} \dot{x}(t) = \begin{bmatrix} \dot{i}(t) \\ \dot{v}(t) \end{bmatrix} = A_2 \begin{bmatrix} i(t) \\ v(t) \end{bmatrix} + B_2 \begin{bmatrix} v_g(t) \end{bmatrix} \\ y(t) = \begin{bmatrix} i_g(t) \\ v(t) \end{bmatrix} = C_2 \begin{bmatrix} i(t) \\ v(t) \end{bmatrix} + E_2 \begin{bmatrix} v_g(t) \end{bmatrix} \end{cases}. \tag{6}$$

where $A_2 = \begin{bmatrix} -\dfrac{R_m}{L_m} & \dfrac{1}{nL_m} \\ \dfrac{1}{nC} & -\dfrac{1}{RC} \end{bmatrix}$, $B_2 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $C_2 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$,

$E_2 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$.
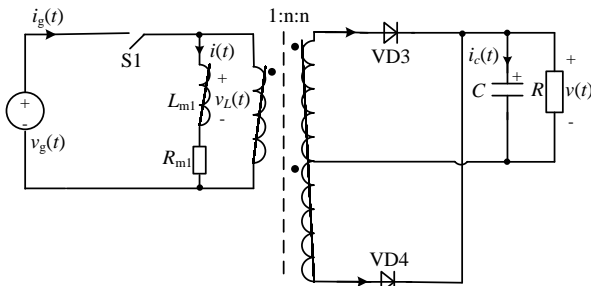
FIGURE 5 Topological structure in $t_1 \sim t_2$

The working mode 3: $t_2 \sim t_3$ time interval is for $[T_s/2, T_s/2+d(t)T_s]$. In this period, the switch S2 is turned on, and the diode VD4 is on state. The current flows through the transformer, diode VD4, a filter capacitor C and load R. At the same time the current value increases. The topological structure is shown in Figure 6 in this period.
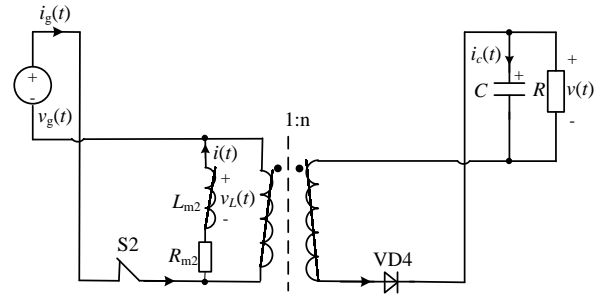
FIGURE 6 Topological structure in $t_2 \sim t_3$

Due to the symmetry of the circuit, the state equation and output equation are expressed for:

$$\begin{cases} \dot{x}(t) = \begin{bmatrix} \dot{i}(t) \\ \dot{v}(t) \end{bmatrix} = A_3 \begin{bmatrix} i(t) \\ v(t) \end{bmatrix} + B_3 \begin{bmatrix} v_g(t) \end{bmatrix} \\ y(t) = \begin{bmatrix} i_g(t) \\ v(t) \end{bmatrix} = C_3 \begin{bmatrix} i(t) \\ v(t) \end{bmatrix} + E_3 \begin{bmatrix} v_g(t) \end{bmatrix} \end{cases}, \tag{7}$$

where $A_3=A_1$, $B_3=B_1$, $C_3=C_1$, $E_3=E_1$.

The working mode 4: $t_2 \sim t_3$ time interval is for $[T_s/2+d(t)T_s, T_s]$. In this period, the work state is like the work mode 2. The topological structure is shown in Figure 7 in this period.

The state equation and output equation are expressed for:

$$\begin{cases} \dot{x}(t) = \begin{bmatrix} \dot{i}(t) \\ \dot{v}(t) \end{bmatrix} = A_4 \begin{bmatrix} i(t) \\ v(t) \end{bmatrix} + B_4 \begin{bmatrix} v_g(t) \end{bmatrix} \\ y(t) = \begin{bmatrix} i_g(t) \\ v(t) \end{bmatrix} = C_4 \begin{bmatrix} i(t) \\ v(t) \end{bmatrix} + E_4 \begin{bmatrix} v_g(t) \end{bmatrix} \end{cases}. \tag{8}$$

where $A_4=A_1$, $B_4=B_1$, $C_4=C_1$, $E_4=E_1$:
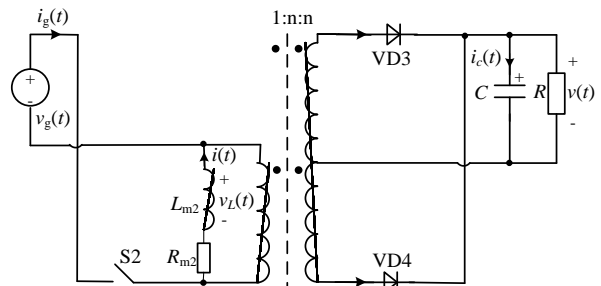
FIGURE 7 Topological structure in $t_3 \sim t_4$

3 Average state vector

In order to eliminate the influence of the switching ripple, it is necessary that the state variables are averaged in a switching period, and the average state variables equations are established. The average state vectors are defined as:

$$\langle \dot{x}(t) \rangle_{T_s} = \frac{d}{dt} \langle x(t) \rangle_{T_s} =$$

$$\frac{d}{dt} \left(\frac{1}{T_s} \int_t^{t+T_s} x(\tau) d\tau \right) = \frac{1}{T_s} \int_t^{t+T_s} \dot{x}(\tau) d\tau = \qquad (9)$$

$$\frac{1}{T_s} \left\{ \int_t^{t+d(t)T_s} \dot{x}(\tau) d\tau + \int_{t+d(t)T_s}^{t+T_s/2} \dot{x}(\tau) d\tau + \right.$$

$$\left. \int_{t+T_s/2}^{t+T_s/2+d(t)T_s} \dot{x}(\tau) d\tau + \int_{t+T_s/2+d(t)T_s}^{t+T_s} \dot{x}(\tau) d\tau \right\}$$

The Equations (3), (6)-(8) are substituted into equation (9) and then the following expression can be obtained:

$$\langle \dot{x}(t) \rangle_{T_s} = \frac{1}{T_s} \left\{ \int_t^{t+d(t)T_s} [A_1 x(\tau) d\tau + B_1 u(t)] d\tau + \right.$$

$$\int_{t+d(t)T_s}^{t+T_s/2} [A_2 x(\tau) d\tau + B_2 u(t)] d\tau +$$

$$\int_{t+T_s/2}^{t+T_s/2+d(t)T_s} [A_3 x(\tau) d\tau + B_3 u(t)] d\tau + \qquad (10)$$

$$\left. \int_{t+T_s/2+d(t)T_s}^{t+T_s} [A_4 x(\tau) d\tau + B_4 u(t)] d\tau \right\}$$

When the converter to meet the low frequency hypothesis and small ripple hypothesis, with respect to the state variables and input variables can be replaced with the instantaneous value of average value in a switching period, and they can be regarded as the average to maintain the constant value in a switching period. These treatments can not introduce large error analysis. So the state variables $<x(t)>_{Ts}$ and $<u(t)>_{Ts}$ can be regarded as constants in a switching period. Then the Equation (10) can be approximately simplified and inferred. The expressions of average state variable equation are obtained:

$$\langle \dot{x}(t) \rangle_{T_s} = [2d(t) \cdot A_1 + (1 - 2d(t) \cdot A_2] \langle x(t) \rangle_{T_s} +$$

$$[2d(t) \cdot B_1 + (1 - 2d(t) \cdot B_2] \langle u(\tau) \rangle_{T_s} \qquad (11)$$

The same analysis method is applied to the output average state vector. The expression of the average output vector can be inferred for:

$$\langle y(t) \rangle_{T_s} = [2d(t) \cdot C_1 + (1 - 2d(t) \cdot C_2] \langle x(t) \rangle_{T_s} +$$

$$[2d(t) \cdot E_1 + (1 - 2d(t) \cdot E_2] \langle u(\tau) \rangle_{T_s} \qquad (12)$$

4 Small signal separation and linearization

In order to further determine static working point of the push-pull bi-directional DC-DC converter, and analyse the working condition of the AC small signal on the static working point, the state average variables need to be decomposed. The state average variables are decomposed into DC and AC small signal component. So the state average variables $<x(t)>_{Ts}$, $<u(t)>_{Ts}$ and $<y(t)>_{Ts}$ can be resolved as follows:

$$\begin{cases} \langle x(t) \rangle_{T_s} = X + \hat{x}(t) \\ \langle u(t) \rangle_{T_s} = U + \hat{u}(t) , \\ \langle y(t) \rangle_{T_s} = Y + \hat{y}(t) \end{cases} \qquad (13)$$

where X is the DC component vector of the state vector. U is the DC component vector of the input vector. Y is the DC component vector of the output vector. $\hat{x}(t)$ is the AC small signal component vector of the state vector. $\hat{u}(t)$ is the AC small signal component vector of the input vector. $\hat{y}(t)$ is the AC small signal component vector of the output vector. At the same time, the control variable of the duty ratio $d(t)$ contains AC component, and it is decomposed of the DC component D and AC small signal component $\hat{d}(t)$. The expression is for:

$$d(t) = D + \hat{d}(t). \qquad (14)$$

The state average vectors of the push-pull bidirectional DC-DC converter meet the small signal hypothesis, namely the amplitudes of each AC small signal component are far less than the amplitudes of the DC component. After the merger of similar items, the expression is for:

$$\langle \dot{x}(t) \rangle_{T_s} = [2DA_1 + (1 - 2D)A_2]X +$$

$$[2DB_1 + (1 - 2D)B_2]U +$$

$$[2DA_1 + (1 - 2D)A_2]\hat{x}(t) +$$

$$[2DB_1 + (1 - 2D)B_2]\hat{u}(t) + \qquad (15)$$

$$[(2A_1 - 2A_2)X + (2B_1 - 2B_2)U]\hat{d}(t) +$$

$$(2A_1 - 2A_2)\hat{x}(t)\hat{d}(t) + (2B_1 - 2B_2)\hat{u}(t)\hat{d}(t),$$

$$\langle y(t) \rangle_{T_s} = [2DC_1 + (1 - 2D)C_2]X +$$

$$[2DE_1 + (1 - 2D)E_2]U +$$

$$[2DC_1 + (1 - 2D)C_2]\hat{x}(t) +$$

$$[2DE_1 + (1 - 2D)E_2]\hat{u}(t) + \qquad (16)$$

$$[(2C_1 - 2C_2)X + (2E_1 - 2E_2)U]\hat{d}(t) +$$

$$(2C_1 - 2C_2)\hat{x}(t)\hat{d}(t) + (2E_1 - 2E_2)\hat{u}(t)\hat{d}(t),$$

where the DC and AC components are corresponding equal on either side of the equals sign. When the DC items are corresponding equal, can be deduced:

$$\begin{cases} \dot{X} = [2DA_1 + (1 - 2D)A_2]X + [2DB_1 + (1 - 2D)B_2]U \\ Y = [2DC_1 + (1 - 2D)C_2]X + [2DE_1 + (1 - 2D)E_2]U \end{cases}. \qquad (17)$$

When the converter is on the steady state, the DC component X of the state vector is constant. So the static work point is solved in push-pull bidirectional DC-DC converter. The static work point is for:

$$\begin{cases} X = -A^{-1}BU \\ Y = (E - CA^{-1}B)U \end{cases}, \qquad (18)$$

where $A = 2DA_1 + (1-2D)A_2$, $B = 2DB_1 + (1-2D)B_2$, $C = 2DC_1 + (1-2D)C_2$, $E = 2DE_1 + (1-2D)E_2$.

When the AC items are corresponding equal, it can be deduced:

$$\dot{\hat{x}}(t) = A\hat{x}(t) + B\hat{u}(t) + [(2A_1 - 2A_2)X +$$
$$(2B_1 - 2B_2)U]\hat{d}(t) + (2A_1 - 2A_2)\hat{x}(t)\hat{d}(t) + \quad (19)$$
$$(2B_1 - 2B_2)\hat{u}(t)\hat{d}(t),$$

$$\hat{y}(t) = A \cdot \hat{x}(t) + B \cdot \hat{u}(t) + [(2C_1 - 2C_2) \cdot X +$$
$$(2E_1 - 2E_2) \cdot U] \cdot \hat{d}(t) + (2C_1 - 2C_2) \cdot \hat{x}(t) \cdot \hat{d}(t) + \quad (20)$$
$$(2E_1 - 2E_2) \cdot \hat{u}(t) \cdot \hat{d}(t),$$

Equation (19) is small signal state equation of push-pull bidirectional DC-DC converter. Equation (20) is small signal output equation of push-pull bidirectional DC-DC converter. But the above two equations are nonlinear equations. Therefore they are linearized in the vicinity of the static working point. The right side of the equal sign is the product of nonlinear small signal in Equations (19) and (20). When the push-pull bidirectional DC-DC converter meets the small signal assumption, the amplitude of small signal product term is far less than the rest of the equal right. Therefore the product terms of the equation can be omitted.

The method of linearization does not introduce large errors, in order to achieve the linearization of nonlinear small signal equations. Equations (19) and (20) are linearized. The linear small signal state equation and the linear small signal output equation are inferred.

$$\begin{cases} \dot{\hat{x}}(t) = A\hat{x}(t) + B\hat{u}(t) + [(2A_1 - 2A_2)X + (2B_1 - 2B_2)U]\hat{d}(t) \\ \hat{y}(t) = A\hat{x}(t) + B\hat{u}(t) + [(2C_1 - 2C_2)X + (2E_1 - 2E_2)U]\hat{d}(t) \end{cases} \quad .(21)$$

5 Conclusions

Aiming at the need to achieve two-way energy flow, a push-pull bidirectional DC-DC converter is designed. The topological structure of the converter circuit is analysed. The dynamic modelling of circuit is done by using the state space averaging method. The linearization method is implemented by the AC small signal analysis. The linear expressions of the state space representation are derived. The work lays the foundation for the design of the control model and the study of the converter characteristics in the future. The design of push-pull bidirectional DC converter has the advantages of simple structure and a small number switching devices. The circuit will be adapting to small and medium power applications.

Acknowledgements

References

[1] Cinar S M 2012 *Journal of Engineering Science and Technology Review* **5**(4) 30-4
[2] Hu X F, Gong C Y A 2012 *Proceedings of the Chinese Society for Electrical Engineering* **32**(15) 8-15
[3] Camara M B, Gualous H, Gustin F 2010 *IEEE Transactions on Industrial Electronics* **57**(2) 587-97
[4] Tong Y B, Wu T, Jin X M, Chen Y 2007 *Proceedings of the Chinese Society for Electrical Engineering* **27**(13) 81-6
[5] Chainarin E, Kamon J 2013 *Journal of Power Electronics* **13**(4) 626-35
[6] Zhang Z, Thomsen O C, Andersen M A E 2012 *IEEE Transactions on industrial electronics* **59**(7) 2761-71
[7] Hote Y V, Choudhury D R, Gupta J R P 2009 *IEEE transactions on power electronics* **24**(10) 2353-6
[8] Han D H, Lee Y J, Kwon W S, Bou-Rabee M A, Choe G-H 2012 *Journal of Power Electronics* **12**(3) 418-28

Authors

Hongjuan Zhang, born in September, 1974, Shanxi, China

Current position, grades: associate professor at Taiyuan University of Technology, Taiyuan City, China.
University studies: PhD degree in Mechanical and Electronic Engineering from Taiyuan University of Technology, Shanxi, China, in 2011.
Scientific interests: computer modelling, system computer modelling, software programming, energy saving control.
Publications: 10.

Long Quan, born in March, 1959, Shanxi, China

Current position, grades: Doctor, professor, doctoral supervisor, Institute Director of Mechanical and Electronic Engineering Research, Taiyuan University of Technology, Shanxi, China.
Scientific interest: computer modelling, system computer modelling and design, software programming, intelligent control on mechatronics engineering, fluid power transmission and control, energy saving control.
Publications: over 80.

Yan Gao, born in July, 1969, Shanxi, China

Current position, grades: lecturer at Taiyuan university of technology, Taiyuan City, China.
University studies: M.E. in automation and information engineering from Taiyuan University of Technology, Shanxi, China, in 1998.
Scientific interests: computer application, software programming, control theory and intelligent control
Publications: 10.

An inter-satellite dynamic ranging algorithm based on two-way time synchronization

Feijiang Huang[1, 2], Jun Yang[1*], Xiaochun Lu[3, 4], Qingxiao Shan[1],
Yongbin Zhou[1], Jianyun Chen[1], Zhuli Hu[1]

[1]*College of Mechatronic and Automation, National University of Defense Technology, Changsha 410073, China*

[2]*Department of Electronics and Communication Engineering, Changsha University, Changsha 410022, China*

[3]*National Time Service Center, Chinese Academy of Sciences, Xi'an 710600, China*

[4]*Key Laboratory of Precision Navigation and Timing Technology, Chinese Academy of Sciences, Xi'an 710600, China*

Received 01 August 2014, www.cmnt.lv

Abstract

Inter-satellite precise ranging is the foundation for all aerospace application systems in realizing autonomous navigation. To acquire a high-accuracy inter-satellite range, this study investigates an inter-satellite dynamic ranging algorithm. Referring to the simulation of inter-satellite range variation rules in constellation, this study analyzes the negative impact of satellite motion on inter-satellite ranging and proposes corresponding improved methods to eliminate the major error caused by satellite motion. This algorithm solves the minimal error in inter-satellite range using a combination of inter-satellite range fitting polynomial and inter-satellite clock-offset fitting polynomial, both of which are generated by two-way time synchronization data. Simulation calculation results show that the accuracies of inter-satellite ranging can be controlled within 3m provided that simulation error is considered. The algorithm can be used to improve the accuracy of inter-satellite dynamic ranging of various aerospace application systems.

Keywords: aerospace application systems, autonomous navigation; inter-satellite communication, two-way time synchronization, inter-satellite dynamic ranging

1 Introduction

With the rapid development of aerospace technology, constellation autonomous navigation and autonomous ranging have become a key direction in aerospace technology development [1-5]. As the key technology in constellation autonomous navigation, inter-satellite communication links establishment and maintenance technologies and inter-satellite ranging technologies has been studied widely. Long-term satellite ephemeris and clock reference are revised constantly from the ground station through the data exchange in inter-satellite ranging and the filtering process by satellite-borne processors. Furthermore, navigation messages are autonomously generated, and basic constellation configuration is thus maintained. As a result, the application demand for constellation autonomous navigation is met [5]. Choosing the appropriate inter-satellite ranging method and obtaining a precise inter-satellite range are the key issues that must be addressed to realize constellation autonomous navigation.

Many studies have investigated constellation autonomous navigation and inter-satellite ranging [5-8]. Reference [5] takes advantage of time synchronization and ephemeris in satellite constellation for updating the

Kalman filter to process two-way inter-satellite ranging data. This method is able to realize precise time synchronization for satellite constellation and high-accuracy satellite orbit determination. The maximum user ranging error after filter convergence of satellite radial orbit is less than 6 m. Reference [6] analyzes the constellation autonomous navigation method based on inter-satellite ranging information and deduces the location of inter-satellite ranging information as well as the model and condition of time decoupling. The work puts forward the solving process that involves updating the location and time consecutively. The work also deduces a distributed Kalman filtering algorithm based on the characteristics of state and measurement equations. This algorithm is applicable in solving constellation autonomous navigation. Reference [7] proposes an autonomous orbit determination algorithm using an improved Kalman filtering fusion dynamic information and inter-satellite ranging information. This algorithm could achieve on-time revision of integral initial value. Reference [8] proposes the use of little ground emission source to provide a ground base. The method combines inter-satellite ranging and ground emission source information to conduct and improve the accuracy of orbit determination for whole satellite constellation. The

**Huang Feijiang, Yang Jun, Lu Xiaochun, Shan Qingxiao,
Zhou Yongbin, Chen Jianyun, Hu Zhuli**

aforementioned studies focus on improving the accuracy of orbit determination and realizing autonomous navigation through inter-satellite ranging data. However, research on how to obtain high-accuracy inter-satellite range is rare. Referring to reference 9 and by simulating the changing characteristics of moving satellites, this study proposes a dynamic measurement algorithm for high-accuracy inter-satellite range gained from an inter-satellite pseudo-range fitting polynomial. This algorithm is capable of realizing high-accuracy inter-satellite ranging while achieving high-accuracy time synchronization.

2 Principle of two-way time synchronization for inter-satellite ranging

The principle of two-way time synchronization for inter-satellite ranging is shown in Figure 1 [9]. A radio transmitter and receiver are installed on satellites A and B. The following equations can be obtained when satellite A and B are sending time synchronization signal to each other at the same time and receiving each other's signal:

$$T_1 = \Delta t + t_2 + \tau_{BA} + r_1 + \delta_1 , \qquad (1)$$

$$T_2 = -\Delta t + t_1 + \tau_{AB} + r_2 + \delta_2 . \qquad (2)$$

In the previous equations, Δt is the clock correction of satellites A and B. T_1 is the time difference between the transmission of timing signal by satellite A and its receipt of timing signal transmitted by satellite B. t_2 is the transmission delay of satellite B, τ_{BA} is the propagation delay from satellite B to satellite A, r_1 is the receiving delay of satellite A, δ_1 is the other delay [10], T_2 is the time difference between the transmission of timing signal by satellite B and its receipt of timing signal transmitted by satellite A, t_1 is the transmission delay of satellite A, τ_{AB} is the propagation delay from satellite A to satellite B, r_2 is the receiving delay of satellite B, and δ_2 is the other delay.
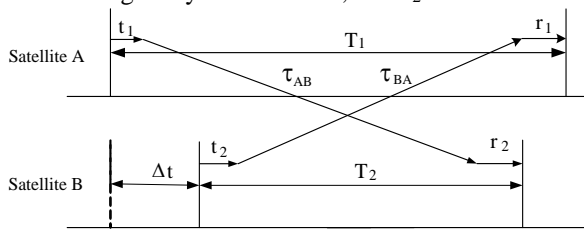
FIGURE 1 Schematic of two-way time synchronization for inter-satellite ranging

To determine the inter-satellite range and clock offset of the two satellites, Equations (1) and (2) can be arranged such that the following results are obtained:

$$(\tau_{BA} + \tau_{AB}) = (T_1 + T_2) - (t_1 + t_2) - (r_1 + r_2) - (\delta_1 + \delta_2) , \qquad (3)$$

$$\Delta t = \frac{T_1 - T_2}{2} + \frac{t_1 - t_2}{2} + \frac{r_2 - r_1}{2} + \frac{\tau_{AB} - \tau_{BA}}{2} + \frac{\delta_2 - \delta_1}{2} . \qquad (4)$$

In the previous equation, T_1 and T_2 can be obtained from the measurement of satellites A and B. t_1, t_2, r_1, and

r_2 can be pre-calibrated according to the transmission frequency of satellite signal. When satellites A and B have similar timing signal frequencies, linking is symmetrical, and transmission delay is approximately equal. These conditions indicate that $\tau_{AB} = \tau_{BA}$, and other delay impacts are ignored. Finally, the inter-satellite range and clock offset of the two satellites can be obtained.

Assuming the clock correction Δt of two satellites is kept the same in the time synchronization process, and the impacts of receiving and transmitting equipment delay and other delays are ignored, Equations (3) and (4) can be simplified as:

$$\tau_{AB}(\tau_{BA}) = \frac{(T_1 + T_2)}{2} , \qquad (5)$$

$$\Delta t = \frac{T_1 - T_2}{2} . \qquad (6)$$

Equation (5) involves the multiple speed of light c, thus indicating that the two-way time synchronization results in the following inter-satellite range formula:

$$\rho = c\tau_{AB} = c\frac{T_1 + T_2}{2} . \qquad (7)$$

3 Analysis of impact of satellite motion on the accuracy of two-way inter-satellite ranging

To analyze the impact of satellite motion on the accuracy of inter-satellite ranging, the rules on the variation of moving satellites A and B in a certain constellation for inter-satellite ranging are simulated using the Satellite Tool Kit (STK), a satellite simulation tool software. The simulation results are shown in Figure 2.
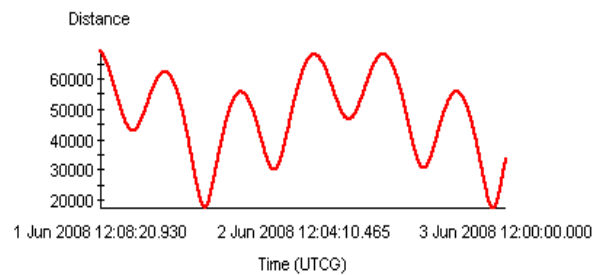
FIGURE 2 Simulation of inter-satellite range variation between satellites A and B in a constellation

As shown in Figure 2, the rules on the variation in inter-satellite range among mobile satellites in a constellation are as follows: with the moving of satellites. The phenomenon of inter-satellite range shifting from long to short and back to long occurs several times; this condition leads to regular changes in the transmission delay of the two-way inter-satellite time synchronization signal. Therefore, $\tau_{AB} = \tau_{BA}$ cannot be met normally if the two-way

Huang Feijiang, Yang Jun, Lu Xiaochun, Shan Qingxiao, Zhou Yongbin, Chen Jianyun, Hu Zhuli

time synchronization algorithm is applied to solve inter-satellite ranging. Hence, the inter-satellite range acquired by Equation (7) is not equal to an actual one.

Next, a situation involving two-way ranging between GEO and MEO satellites is considered. The ranging scheme is shown in Figure 3 [11].
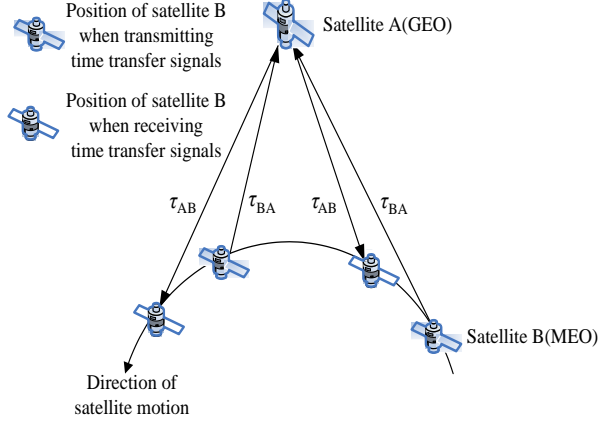
FIGURE 3 Scheme of variation of inter-satellite ranging results following inter-satellite two-way time synchronization signals propagation delay

Satellite B first moves toward satellite A then moves away. The impact of the motions of the MEO satellite on the propagation delay of two-way time synchronization signals and ranging in this situation is analyzed. Equations (1) and (2) are simplified as:

$$\tau_{AB}(\tau_{BA}) = \frac{(T_1 + T_2)}{2}, \tag{8}$$

$$\Delta t = \frac{T_1 - T_2}{2}. \tag{9}$$

When satellite B approaches satellite A, τ_{BA} becomes greater than τ_{AB} (Figure 3). Therefore, the time difference T_2 of satellite B obtained from Equation (9) is smaller than that measured when satellite B is stationary. If Equation (7) is still adopted to calculate range, then the result is smaller than the actual inter-satellite range. When satellite B keeps away from satellite A, τ_{BA} is smaller than τ_{AB} (Figure 3). Therefore, the time difference T_2 of satellite B obtained from Equation (9) is greater than that measured when satellite B is stationary. If Equation (7) is still adopted to calculate range, then the result is greater than the actual inter-satellite range.

4 An inter-satellite dynamic ranging algorithm

Considering both situations above as well as the increase and decrease in range between satellites A and B, the inter-satellite range obtained from the two-way satellite time synchronization algorithm accordingly varies from being smaller to being greater than the actual inter-satellite range. In this variation process, a certain moment occurs when the inter-satellite range obtained by the two-way inter-satellite time synchronization method is nearest the

actual inter-satellite range. At the moment when the range between satellites A and B is the smallest, the propagation delays of the two-way time synchronization signals of the satellites are closest to each other. Thus, the range obtained at this moment with two-way time synchronization bears the least difference from the inter-satellite range.

According to this analysis, the pseudo-range sequence and the clock offset sequence between satellites A and B obtained by two-way time synchronization ranging in this variation process can be expressed by the pseudo-range polynomial and the clock offset polynomial, respectively. Then, the corresponding time for the minimum inter-satellite pseudo-range can be acquired from the pseudo-range polynomial. Substituting this time in the pseudo-range polynomial obtains the inter-satellite range approximate to the actual one. This method is an inter-satellite dynamic ranging algorithm that considers the impact of satellite motion on two-way time synchronization ranging. Setting the inter-satellite pseudo-range polynomial after two-way time synchronization fitting as ρ and the inter-satellite clock offset polynomial as Δt yields the following polynomial:

$$\rho = f_1(t), \tag{10}$$

$$\Delta t = f_2(t). \tag{11}$$

In Equation (10), the following is set:

$$\frac{df_1(t)}{dt} = 0. \tag{12}$$

By solving Equation (12), t_3 that corresponds to minimum inter-satellite pseudo-range $\rho_{\min}$ can be obtained. Substituting $\rho_{\min}$ into pseudo-polynomial (10), the inter-satellite range approximate to the actual range can be obtained as:

$$\rho_{\min} = f_1(t_3). \tag{13}$$

5 Simulation results and analysis

Satellite A is a GEO satellite, whereas as satellite B is an MEO satellite. Assume that the clock offset of satellites A and B is maintained at 1 μs during dynamic two-way ranging, and that the equipment delay of the receiver and transmitter as well as other delays are ignored. Aided by STK, the simulations of inter-satellite range from long to short to long are performed four times. Durations for two-way time synchronization dynamic ranging data are 4, 8, 10, and 8 minutes. The last duration (i.e., 8 minutes) at an asymmetrical time is chosen to achieve a minimum inter-satellite range. Least squares fitting are carried out for the ranging data described above. The pseudo-range polynomial between satellites A and B is then obtained. With this polynomial, the inter-satellite range minimum can be calculated. Then, this minimum value is compared with the theoretical value of the minimum inter-satellite

Huang Feijiang, Yang Jun, Lu Xiaochun, Shan Qingxiao,
Zhou Yongbin, Chen Jianyun, Hu Zhuli

range acquired by STK simulation. In doing so, the correctness of the algorithm can be tested.

The least squares fitting curve of the pseudo-range at different ranging times is shown in Figure 4. The pseudo-range is two times fitting.
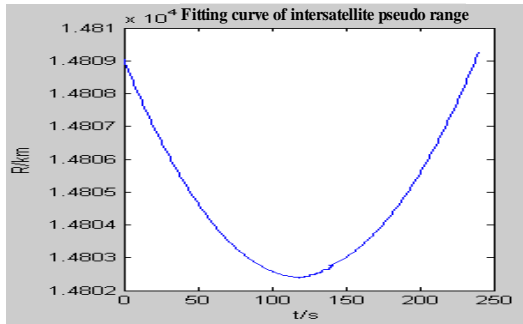
a) Results of ranging fitting for 4 minutes: 1 Jun 2008 23:35:50.000–1 Jun 2008 23:39:50.000
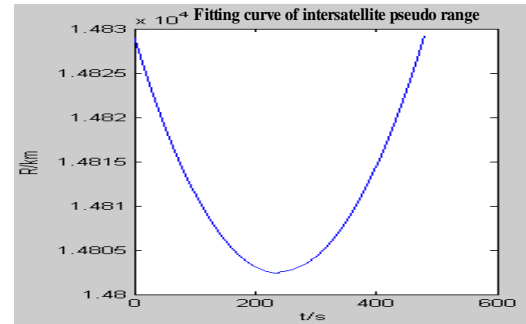
b) Results of ranging fitting for 8 minutes: 1 Jun 2008 23:33:50.000-1 Jun 2008 23:41:50.00
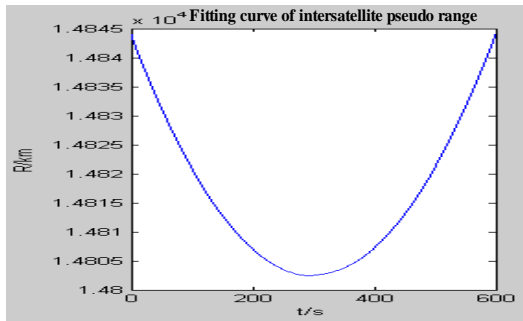
c) Results of ranging fitting for 10 minutes: 1 Jun 2008 23:32:50.000-1 Jun 2008 23:42:50.00
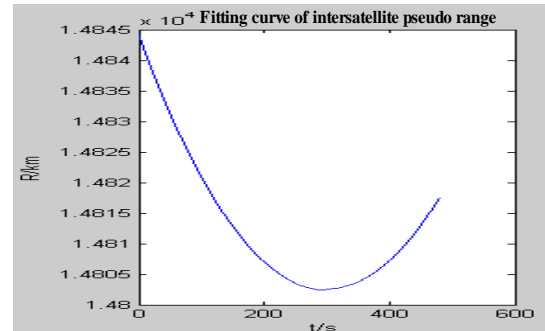
d) Results of ranging fitting for 8 minutes: 1 Jun 2008 23:32:50.000-1 Jun 2008 23:40:50.00

FIGURE 4 Pseudo-range fitting curve between satellites A and B

The comparison of the calculation results and the theoretical value is shown in Table 1.

TABLE 1 Comparison of the results of least squares fitting during different ranging periods (June 1, 2008, with actual clock offset of 1 μs)

Two-way time intervals for ranging	23:35:50.00-23:39:50.00	23:33:50.00-23:41:50.00	23:32:50.00-23:42:50.00	23:32:50.00-23:40:50.00
Pseudo fitting polynomial based on two-way time synchronization	$\rho=0.0004627328\,t^2$ $-0.1097262858\,t$ $+14808.9999915512$	$\rho=0.0004623313\,t^2$ $-0.2212624171\,t$ $+14828.9715935283$	$\rho=0.0004620897\,t^2$ $-0.2767866515\,t$ $+14843.9503933924$	$\rho=0.0004642350\,t^2$ $-0.2777046942\,t$ $+14844.0118630869$
Time corresponding to pseudo minimal ρ_{min}(s)	118.5633280881435	239.2898757962789	299.4944839741193	299.0992571156721
Minimum inter-satellite range determined by pseudo polynomial (km)	14802.49523448867	14802.49866317461	14802.50235286662	14802.48122818178
Theoretical value of minimum inter-satellite range by STK simulation (km)	14802.496873	14802.496873	14802.496873	14802.496873

The comparison of the inter-satellite pseudo-range fitting curve and the ranging results in Table 1 at different time show that the inter-satellite range obtained by the inter-satellite dynamic ranging algorithm based on two-way time synchronization is close to the theoretical value. Simulation results indicate that the proposed inter-satellite dynamic ranging model and the algorithm based on two-way time synchronization are correct.

As shown in Figures 4a, 4b, 4c, the fitting inter-satellite range polynomial has high accuracy, low fitting error, and high ranging accuracy when the two-way ranging duration and the time of minimum inter-satellite range are basically symmetrical. The minimal difference of the theoretical

22

Huang Feijiang, Yang Jun, Lu Xiaochun, Shan Qingxiao, Zhou Yongbin, Chen Jianyun, Hu Zhuli

value is 1.638511330389 m. As the ranging time duration increases, the polynomial fitting error grows, and ranging accuracy decreases.

As shown in Figure 4d, the polynomial fitting accuracy is low, fitting error is large, and ranging accuracy is low when the two-way ranging duration and the time of minimum inter-satellite range are asymmetrical. The ranging result error for an eight-minute simulation can reach -15.64481822061 m.

As indicated by the above analyses, ranging duration and the duration of minimum inter-satellite range are basically symmetrical when applying the inter-satellite dynamic ranging algorithm based on two-way time synchronization in the range measurement of mobile satellites in constellation. The algorithm can reduce polynomial fitting error and increase inter-satellite ranging accuracy.

6 Conclusions

Considering the need for precise inter-satellite ranging for all kinds of aerospace application systems, this study analyzes the negative impact of satellite motion in constellation on the accuracy of two-way inter-satellite ranging. An inter-satellite dynamic ranging algorithm is also proposed. This algorithm may be used to eliminate the negative impact of satellite motion on inter-satellite ranging. The actual satellite data simulation shows that for cases in which simulation error is included, the ranging accuracy of this algorithm covers 3m when the ranging period is basically symmetrical to the moment of minimum inter-satellite range and when the pseudo-range polynomial undergoes least squares fitting. Therefore, the proposed algorithm can be used to achieve high accuracy in inter-satellite precise ranging via the mobile satellite of an aerospace application system.

Acknowledgments

References

[1] Wang D, Xu B, Liu W, Sun G 2014 A novel navigation inter-satellite links ranging hierarchy and its orbit determination performance *Journal of National University of Defense Technology* **36**(1) 62-6 (*in Chinese*)

[2] Kang K, Li H, Wu Y, Zou Z, Xing L 2012 Demonstration on the design of filter indexes of inter-satellite high accuracy ranging system for gravity satellite *Chinese Journal of Geophysics* **55**(10) 3240-7 (*in Chinese*)

[3] Cui P, Xu R, Zhu S, Zhao F 2014 State of the art and development trends of on-board autonomy technology for deep space explorer *Acta Aeronautica et Astronautica Sinica* **35**(1) 13-28 (*in Chinese*)

[4] Qian Y, Jing W, Gao C, Wei W 2013 Autonomous orbit determination for quasi periodic orbit about the translunar libration point *Journal of Astronautics* **34**(5) 625-33 (*in Chinese*)

[5] Shuai P, Qu G, Chen Z 2006 Studies on autonomous navigation techniques for navigation constellations *Engineering Science* **8**(3) 22-30

[6] Liu J, Yu F, He L, Xiong Z 2009 Autonomous positioning and timing for navigation constellation *Journal of Astronautics* **30**(1) 215-9 (*in Chinese*)

[7] Zhang S, Han J, Liu G 2008 Autonomous orbit determination of navigation constellation based on cross-link range *GNSS World of China* (2) 9-12

[8] Zhu J, Liao Y, Wen Y 2009 The integrated autonomous orbit determination of the navigation constellation based on crosslink range and ground-based emitter *Journal of National University of Defense Technology* **31**(2) 15-9 30 (*in Chinese*)

[9] Huang F, Lu X, Wu H, Bian Y 2010 An algorithm of dynamic two-way time transfer based on intersatellite range variation *Geomatics and Information Science of Wuhan University* **35**(1) 13-6 (*in Chinese*)

[10] Li Z G, Li H X, Zhang H 2002 Reduction for the two-way satellite time and frequency transfer *Acta Astronomica Sinica* **43**(4) 422-31

[11] Huang F, Lu X, Wu H, Bian Y, Zhao H 2009 Algorithm of intersatellite dynamic two-way time transfer based on geo satellite *European Frequency & Time Forum & IEEE Int'l Frequency Control Symposium* 688-91

Authors

Huang Feijiang, born in 1972, Hunan, China

Current position, grades: post doctor in College of Mechatronic and Automation, National University of Defense Technology, associate professor in Department of Electronics and Communication Engineering of Changsha University.
University studies: Ph.D. degree in astrometry and celestial mechanics at Graduate University of Chinese Academy of Sciences in 2009.
Scientific interest: time synchronization, satellite navigation and inter-satellite links.

Yang Jun, born in 1972, Jiangsu, China

Current position, grades: professor in College of Mechatronic and Automation, National University of Defense Technology.
University studies: Ph.D. degree from National University of Defense Technology.
Scientific interest: space instrumentation.

Lu Xiaochun, born in 1970, Shaanxi, China

Current position, grades: research professor in National Time Service Center, Chinese Academy of Sciences.
University studies: Ph.D. degree in astrometry and celestial mechanics at Graduate University of Chinese Academy of Sciences in 2004.
Scientific interest: satellite navigation, time synchronization.

Huang Feijiang, Yang Jun, Lu Xiaochun, Shan Qingxiao,
Zhou Yongbin, Chen Jianyun, Hu Zhuli

	Shan Qingxiao, born in 1973, Hunan, China **Current position, grades:** associate professor in College of Mechatronic and Automation, National University of Defense Technology. **University studies:** Ph.D. degree at National University of Defense Technology in 2003. **Scientific interest:** techniques of measurement control in time and frequency.
	Zhou Yongbin, born in 1982, Henan, China **Current position, grades:** associate professor in College of Mechatronic and Automation, National University of Defense Technology. **University studies:** Ph.D. degree at National University of Defense Technology. **Scientific interest:** wireless signal processing, space instruments engineering.
	Chen Jianyun, born in 1975, Jiangsu, China **Current position, grades:** associate professor in College of Mechatronic and Automation, National University of Defense Technology. **University studies:** Ph.D. degree at National University of Defense Technology. **Scientific interest:** space instruments engineering.
	Hu Zhuli, born in 1957, Hunan, China **Current position, grades:** associate professor in College of Mechatronic and Automation, National University of Defense Technology. **University studies:** M.S. degree at National University of Defense Technology. **Scientific interest:** space instrumentation.

Multi-agent simulation of partner selection behaviour based on matching degree in collaborative product development process

Shuo Zhang*, Yingzi Li

Dongling School of Economics and Management, University of Science and Technology Beijing, Beijing, China

Abstract

Collaborative Product Development (CPD) process is characterized by autonomous task control, dynamic task sequence, and frequent team collaboration, which endow the process with high flexibility and uncertainty. To make the process predictable and improve process efficiency, it is essential to model, simulate, and analyse the process by considering all these characteristics. Our work focuses on studying the human working behaviours in CPD process by agent-based simulation, which we think is the main source of process uncertainty and flexibility. In this paper, the partner selection behaviours are studied under the frame of agent-based simulation. In the simulation, the design agent selects his partner according to matching degree including ability and character. The simulation results indicate that the proposed utility strategy can effectively shorten the project total time of the case.

Keywords: collaborative product development, partner selection, multi-agent simulation

1 Introduction

Scholars usually describe, predict, evaluate, and optimize the design process CPD process by modelling and simulation. The typical modelling methods are: DSM (design structure matrix), Petri Nets and activity-on-node graphs et al. DSM can be applied to describe the complex relationship and dynamical dependencies among tasks implicitly by information flow, which can predict and evaluate the cost and duration [1-3]; Petri nets and activity-on-node graphs can be used to simulate workflow and to predict lead times of product development projects by formulating task networks [4-6]. In the above models, the process is described as a set of interactive tasks or activities, while the designers are considered as passive "design resources" assigned to activities. As a result, the designer's autonomous and cooperative activities cannot be described and evaluated directly in the corresponding simulations, and it is difficult to anticipate the process and task execution time for the complex interaction relationship among designers in CPD process. To make the design process be able to quickly respond to the highly dynamic and distributed design environment, it is emphasized in CPD that team members are the most flexible and active elements. In this case, agent-based modelling and simulation (ABMS) has been recently considered as a valuable research approach as it supports active and collaborative process description of adaptive complex systems and reflects human's autonomous as well as cooperative behaviours [7, 8]. In this paper, we concentrate on cooperation behaviour of members by agent-based modelling, including local partner selection behaviour of design agents, negotiation behaviour between design agents and arbitration behaviour of manager agent in partner selection conflict resolution.

In CPD project, an initial plan should be worked out before the real executing of the process. And then, during the executing process, the plan has to be adjusted according to the dynamic uncertainty of the design environment. One important reason of dynamic uncertainty is caused by ability limitation of designers. In this case, designers need partner to collaborate, which make the plan is adjusted by agents' selection behaviours during the simulation. It is obvious that different selection behaviours of agents may play an important role to the efficiency of the process. Traditional partner selection often focuses on the macro fields such as supplier selection, enterprise alliance selection [9,10] and they provide efficient algorithm, systematic criterions et al. [11, 12], but the flexible human factors in the high dynamic environment are less motioned. In this paper, we emphasize the partner selection behaviours in relatively micro world and explore the behaviours how to influence the development efficiency.

To make the agents' behaviours more flexible, we designed a partner selection behaviour process for the agents, in which the design agent selects his partner according to matching degree. To make the process more efficient, the strategies are proposed with consideration of task priority and matching degree simultaneously.

The rest of the paper is organized as follows. In Section 2, the agent model and partner selection process are introduced; Section 3 proposes matching degree algorithm. Then, simulation experiments of different scenarios are designed and carried out based on a real CPD project in Section 4.

---

*Corresponding author e-mail: zhangshml@163.com

2 Simulation model and predicate definition of behaviours

In our approach, the simulation model is composed by agent model and environment model. The agent model maps organization members as intelligent agents and describes dynamic collaboration and iteration characteristics of agent's behaviour by defining a set of behaviour goals and knowledge (as shown in Figure 1). The environment model takes design tasks, product information, and design resource as environment objects, which can change their status by designers' behaviours. Components of the agent model include agent sensor, agent decision protocols, agent behaviours, and agent driver. The agent included 3 levels: react level, local programming level and group level. Agent will call different decision protocols under different environment change. The agent apperceives design environment objects, i.e. tasks, product information, and resource, by agent sensor. Then, the agent behaves according to the status of the environment and his decision protocols. Thereafter, the agent's behaviours act on the environment objects by agent driver. Based on this principle, the design process can be simulated as a continuous evolving process of the agents and the design environment objects. Hence, this simulation approach is expected to effectively support human behaviours analysis of the process.
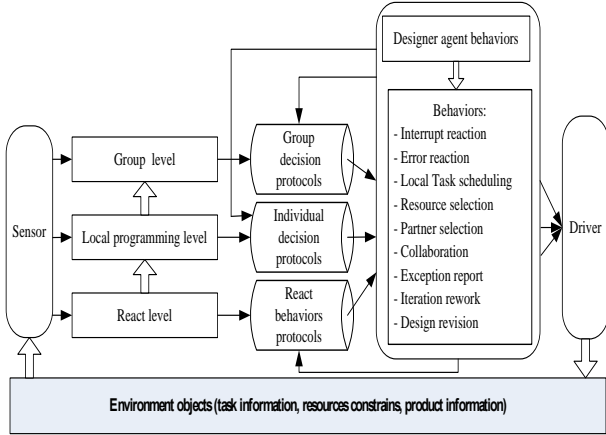
FIGURE 1 Agent-based simulation model of CPD

The detail definition and description of decision protocols of design agent can be found in our former researches [13, 14]. In this paper, the predicate and function definition are added shown as Table 1.

Based on the characteristics of CPD process, we have proposed the partner selection process shown as Figure 2. There are 3 phases in the process: 1) collaboration request, 2) collaboration response, 3) partner selection. Design agent firstly releases collaboration request to potential collaborative partners; potential agents respond to request according to their behaviour rules. Design agents who need partners select the best one from potential partners who accept request according to their status and matching degree between each other. The detailed steps are given as followings by using predicate and function definition.

TABLE 1 Predicate definition and function

Predicate definition	Description
DESIGNER(x)	Design agent x, domain of individuals is A={x₁, x₂... xᵢ}
TASK(y)	Design task y, domain of individuals is B={y₁, y₂... yᵢ}
EXECUTE(x,y)	Design agent x execute design task y, x∈A, y∈B
SEL_T(x,yᵢ)	Design agent x select task yᵢ, which belongs to his task box, x∈A, y∈B
SEL(x₁,xc)	Design agent x₁ send collaboration request to xc, x₁∈A, xc⊂A, x₁∉xc
ACCEPT(x₁,x₂)	Design agent x₂ response collaboration request from x₁, x₁,x₂∈A, x₁≠x₂
SEL_P(x₁,x₂)	Design agent x₁ select x₂ as the collaboration partner, x₁,x₂∈A, x₁≠x₂
function definition	Description
Match(x)	Compute matching degree
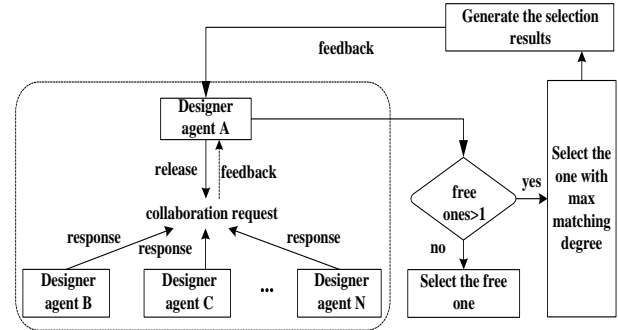
FIGURE 2 Partner selection process in CPD project

Step1.

$$(\forall x_1)(\forall x_2)(EXECUTE(x_1, y)) \wedge$$
$$E(collabrate(x), \alpha) \rightarrow SEL(x_1, x_2)$$

from Figure 2, design agent A is x_1, B, C, D can be x_2, and x_1 release the collaboration request.

Step2.

$(\forall x_1)(\forall x_2)(ACCEPT(x_1, x_2))$, other design agents respond to the request.

Step3.

$$(\forall x_1)(\forall x_2)(SEL\_P(x_1, x_2)) \wedge$$
$$E(match(x), max) \wedge E(status(x), FREE)$$, design agent A

selects the partner with "free" status and maximum matching degree.

3 Matching degree algorithm

Matching degree and task priority are 2 important indexes in partner selection process. Task priority algorithm is proposed in our paper [14], and matching degree algorithm will be introduced in the following. To compute matching degree between agents, we firstly studied the attributes of design agents in CPD process.

3.1 DESIGNERS' ATTRIBUTES ANALYSIS

As the most active element in the CPD process, designers play a significant role in CPD for their autonomy, initiative

and collaboration. Thus, designers' attributes are the most important basis to select partner. In this paper, we have summarized designers' attributes including ability attributes and character attributes according to the references. The specification is shown as Table 2.

TABLE 2 Designers' attributes

Category	Attributes name	Attributes description
Ability attributes	technique ability $T_d()$	The ability of designers to solve technical problems successfully in professional fields. It is the combination of theoretical knowledge gained by specific education with experience by practices. Thus, the technique ability can be measured by 2 aspects: theoretical knowledge level and level of solving the common technical problem successfully
	innovation ability $I_d()$	The ability of designers to solve creative problem based on their knowledge and experiences, introducing multi-industry and interdisciplinary knowledge, using innovation tools and approaches. There are 3 main parts: expertise, proficiency of innovation tools and approaches.
	collaboration ability $C_d()$	There are 2 kinds of collaboration in CPD: 1) based on task relationship; 2) collaboration autonomously. In the first situation, designers must do process and data collaboration for the task relationship; in the second one, designers collaborate for lack of ability, they should select partner to finish the task. Collaboration ability includes: collaboration experience level and communication ability.
Character attributes	decision character $D_p()$	There are 3 types: irresolute, conventional and resolute. Irresolute ones make decision speed slowly, and are easily affected by the rules; conventional ones are usually scholasticism; resolute ones have a quick wit and decisions are made quickly.
	cooperation character $C_p()$	There are 3 types: leading, obedient and cooperative. Leading ones get on well with others and emphasize the macroscopic; obedient ones emphasize the details and are weak in communication; cooperative ones also pay attention to details but are good at communication.

3.2 MATCHING DEGREE OF ABILITY ATTRIBUTES

Matching degrees between design agents are different for different ability requirements of tasks. In this case, algorithm of matching degree of ability is designed as follows while Agent i and Agent i' execute Task j.

$$PP_{ii'j}^A = \sum_k^3 \omega_k^j \cdot PP_{ii'k} . \tag{1}$$

In the above algorithm, $PP_{ii'k}$ means the matching degree of Agent i and Agent i' in ability k, while ω_k^j (

$\sum_k^3 \omega_k^j = 1$) is weight of requirement for ability k of Task j.

The algorithm of $PP_{ii'k}$ is as follows.

$$PP_{ii'k} = 1 - \frac{|y - x|}{Y - X}, x, y \in [X, Y] \tag{2}$$

x : value of Agent i in ability k, $x \in [a_1, b_1]$;

y : value of Agent i' in ability k, $y \in [a_2, b_2]$;

X : $X = \min(a_1, a_2)$;

Y : $Y = \max(b_1, b_2)$.

In above algorithms, a and b stand for values of upper and lower bounds of fuzzy interval respectively, which is corresponding to Table 3.

TABLE 3 Interval values corresponding to linguistic utility values of designers

	Value	[0, 0.2]	[0.2, 0.4]	[0.4, 0.6]	[0.6, 0.8]	[0.8, 1.0]
Linguistic Utility Value	technique ability	Lower	Low	Normal	High	Higher
	innovation ability	Lower	Low	Normal	High	Higher
	collaboration ability	Lower	Low	Normal	High	Higher

3.3 MATCHING DEGREE OF CHARACTER ATTRIBUTES

Based on Chen's research [15], we summarize the character attributes of designer in Table 1. According to their accomplishment, the identification of character attributes are shown as Table 4, Table 5 is the symbol illustration, and the matching degree of various attributes are shown in Table 10. The final value of character attributes $PP_{ii'j}^C$ can be got after normalization.

TABLE 4 Character attributes of designers and identifications

attributes	decision character			cooperation character		
	irresolute	conventional	resolute	leading	obedient	cooperative
identification	P+T	J+T	J+F	N+E	S+I	S+E

TABLE 5 Combination of character attributes and symbol illustration

	P+T	J+T	J+F		
P+T	+&o	-&o	-&+		
J+T	-&o	+&o	+&+		
J+F	-&+	+&+	+&o		
	N+E	S+I	S+E	symbol	value
N+E	o&+	+&o	+&+	+	9
S+I	+&o	o&-	o&o	-	-3
S+E	+&+	+&o	+&+	o	3

3.4 MATCHING DEGREE ALGORITHM

Based on ability and character attributes, matching degree algorithm of Agent i and Agent i' is shown as follows while executing Task j.

$$PP_{ii'j} = \lambda_1 PP_{ii'j}^A + \lambda_2 PP_{ii'j}^C . \qquad (3)$$

In above algorithm, λ_1 and λ_2 are the weights of ability and character attributes, $\lambda_1 + \lambda_2 = 1$.

4 Case study

4.1 SIMULATION RUNNING AND EXPERIMENTS DESIGN

In order to demonstrate the partner selection method, we design 2-group contrastive experiments as shown in Table:

TABLE 6 Experiment design

Item	Partner selection	Simulation times
Experiment 1	Select partner randomly	100
Experiment 2	Select partner based on higher matching degree	100
Aim	prove the importance of matching degree	

TABLE 7 Attributes of designers

Designer	Ability attributes			Character attributes		
	Technique	*Innovation*	*Collaboration*	*Decision*	*Cooperation*	*Identification*
A_1	0.8	0.6	0.7	conventional	cooperative	JTSE
A_2	0.7	0.8	0.7	resolute	leading	JFNE
A_3	0.7	0.6	0.6	conventional	obedient	JTSI
A_4	0.6	0.5	0.8	resolute	cooperative	JFSE
A_5	0.8	0.7	0.7	irresolute	obedient	PTSI
A_6	0.6	0.9	0.8	resolute	cooperative	JFSE
A_7	0.5	0.6	0.8	irresolute	obedient	PTSI
A_8	0.9	0.7	0.6	resolute	leading	JFNE
A_9	0.8	0.8	0.8	conventional	cooperative	JTSE

TABLE 8 Parameters

Initial status of product information $PI()$	PI_0= Known; PI_j= Unknown (j=1,2,3…16);
Learning effect factor *decrease*	redo_decrease=0.8, except_decrease=0.8
Collaboration rate: α	α=2%
Exception rate: β	β=2%
Error rate: γ	γ=2%
Rework rate: λ	$\lambda(TA_1)$=2%, $\lambda(TA_5)$=2%, $\lambda(TA_7)$=2%, $\lambda(TA_{10})$=2%, $\lambda(TA_{11})$=2%
Rejection rate: τ	$\tau(1)$=5%,$\tau(2)$=2%,$\tau(3)$=1%,
Response for selection partner:	Global response

4.3 MATCHING DEGREE RESULTS

Based on former matching degree algorithm, matching degrees of character attributes and ability attributes are calculated in detail respectively. Table 9 shows matching

4.2 ORIGINAL DATA AND INPUT PARAMETERS

There is some collaborative product development project with 15 tasks and 9 designers. The process chart with normal execution time of every task and the designer-task allocation is shown as Figure 3.
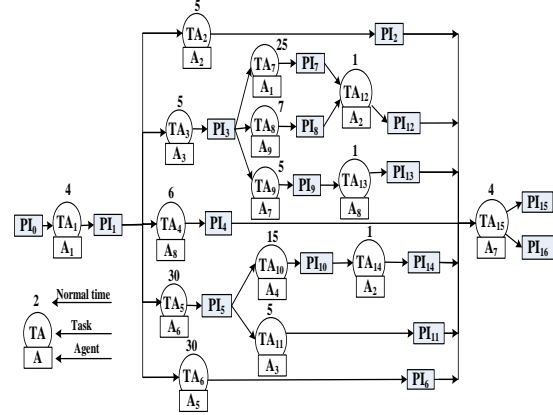
FIGURE 3 The process chart of some collaborative product development process with normal execution time of every task and the designer-task allocation

Table 7 and 8 are attributes of designers and input parameters respectively.

degrees of character attributes between each agent, and Table 10 contains matching degrees of ability attributes of A1 with other agents while executing TA_{18}, TA_{20}, TA_{23}, TA_{25} and TA_{27} which are temporary tasks generated in the process in some simulation.

TABLE 9 Matching degree of character attributes

	A_1 (JTSE)	A_2 (JFNE)	A_3 (JTSI)	A_4 (JFSE)	A_5 (PTSI)	A_6 (JFSE)	A_7 (PTSI)	A_8 (JFNE)	A_9 (JTSE)
A_1 (JTSE)	-								
A_2 (JFNE)	24	-							
A_3 (JTSI)	30	36	-						
A_4 (JFSE)	30	18	36	-					
A_5 (PTSI)	6	18	0	18	-				
A_6 (JFSE)	30	18	36	30	18	-			
A_7 (PTSI)	6	18	0	18	12	18	-		
A_8 (JFNE)	24	24	36	18	18	18	18	-	
A_9 (JTSE)	36	24	30	30	6	30	6	24	-

TABLE 10 Matching degree of ability attributes of A_1

	TA$_{18}$	TA$_{20}$	TA$_{23}$	TA$_{25}$	TA$_{27}$
A$_2$	0.701065	0.67282	0.714565	0.709795	0.71809
A$_3$	0.798395	0.82718	0.787235	0.790205	0.78191
A$_4$	0.719345	0.73031	0.735845	0.720005	0.7343
A$_5$	0.5653267	0.5255767	0.5626867	0.5728567	0.5668567
A$_6$	0.618005	0.58922	0.630665	0.626195	0.63449
A$_7$	0.43976	0.481565	0.47378	0.437315	0.468905
A$_8$	0.5668	0.5668	0.5668	0.5668	0.5668
A$_9$	0.863185	0.832345	0.846565	0.86629	0.848995

4.4 PARTNER SELECTION EXPERIMENT

Null hypothesis: *The project total time has no significant difference with that of the random selection, when designers select partner based on matching degree, i.e. $PTT_1 = PTT_2$.* Results of z test are shown in Table 11.

TABLE 11 z test results

Item	Sample 1	Sample 2
Mean	85.2	71.03
Covariance	121.5152	163.3223
N	100	100
Hypothesis of mean deviation	0	
z	8.395977529	
P(Z<=z) One-tailed	0	
z One-tailed	1.644853627	
P(Z<=z) Two-tailed	0	
z Two-tailed	1.959963985	

In addition, utility strategy is adopted in the 2 experiments to resolve conflict. After 100 runs respectively, $PTT_1 = 85.2$, $PTT_2 = 71.03$, Significance level $\alpha = 0.05$. This value exceeds the critical value and the null hypothesis $PTT_1 = PTT_2$ is rejected. This means the value of PTT in experiment 2 is significantly shorter than PPT in experiment 1.

Managerial insight from partner selection experiment: in collaborative product development process, matching degree influences the duration significantly. Random selection method lacks rigor without consideration of designer's attributes. In this case, simulation can help to find out the most suitable partner for designer who needs collaboration.

5 Conclusions

In this paper, firstly, designer attributes including ability attributes and character attributes are described as the basis to evaluate the matching degree between designers; secondly, partner selection behaviour of design agents are studied under the framework of agent-based simulation. To make the agents' behaviours more flexible, a selection process is developed for the agents by using prediction definition and function definition; lastly, According to the comparative simulation study, the proposed selection method is approved to be effective in shortening lead time of the project. The simulation results also show high accordance with the typical management rules in CPD projects.

Compared with existing references, the proposed approach is developed especially for CPD projects and has the following highlights: 1) designers' ability and character are considered simultaneously; 2) Besides task priority, matching degrees between agents are also integrated into the partner selection strategy, so that the designers can select suitable collaborative partner in time; 3) considering the partner selection behaviours impact on the efficiency of the implementation of the whole process is our main work.

However, the current case study is carried out only in single project environment. It should be extended to the multi-project environment by adding project priority into the partner selection strategy when agents come from different projects and distributed organizations. In addition, to reflect the CPD process explicitly, the resource and cost restraints should be added. Based on the current work, more experiments can be carried out in the near future with the consideration of project priority, resources and costs.

References

[1] Danilovic M, Browning T R 2007 Managing complex product development projects with design structure matrices and domain mapping matrices *International Journal of Project Management* **25** 300-14

[2] Wang T, et al. 2012 Process planning for collaborative product development with CD-DSM in optoelectronic enterprises *Advanced Engineering Informatics* **26**(2) 280-91

[3] Amira S, Dori D, De Weck O 2009 Model-based Design structure matrix: deriving a DSM from an object-process model *Second International Symposium on Engineering Systems* 1-12

[4] Yan H S, Zheng W, Jiao X C 2003 Modeling, scheduling and simulation of product development process by extended stochastic high-level evaluation Petri Nets *Robotics and Computer Integrated Manufacturing* **19** 329–49

[5] Zhang X, Sun Y, Wang L, Xu D 2010 Product model-based design process modeling in collaborative design *In Industrial Engineering*

and Engineering Management (IEEM) 2010 IEEE International Conference 315-9

[6] Arie K, Reich Y 2011 Process Modeling Using Workflow-Nets *Managing the Dynamics of New Product Development Processes Springer London* 75-95

[7] Martínez M J, Pavón J 2008 An agent-based simulation tool to support work teams formation *International Symposium on Distributed Computing and Artificial Intelligence Advances in Soft Computing Springer* 80-9

[8] Garcia R 2010 Uses of agent-based modeling in innovation/new product development research *Journal of Product Innovation Management* **22**(5) 380-98

[9] Wu C, Barnes D 2011 A literature review of decision-making models and approaches for partner selection in agile supply chains *Journal of Purchasing & Supply Management* **17**(4) 256–74

[10] Ye F, Li Y N 2009 Group multi-attribute decision model to partner selection in the formation of virtual enterprise under incomplete information *Expert Systems with Applications* **36**(5) 9350-57

[11] Huang bin, Gao Chenghui, Chen Liang 2010 Partner selection with fuzzy completion time and fuzzy due date in a virtual enterprise *Systems Engineering-Theory &Practice* **30**(6) 1085-91

[12] Feng B, Fan ZP, Ma J 2010 A method for partner selection of co-development alliances using individual and collaborative utilities *International Journal of Production Economics* **124**(1) 159-70

[13] Zhang XD, Luo L, Yang Y 2009 A simulation approach for evaluation and improvement of organizational planning in collaborative product development projects *International Journal of Production Research* **47**(13) 3471-501

[14] Zhang X 2012 Task scheduling behaviour in agent-based product development process simulation *International Journal of Computer Integrated Manufacturing* **25**(10) 914-23 (*in Chinese*)

[15] Chen S J, Li L 2004 *IEEE Transaction on Engineering Management* **51**(2) 111-24

Authors

Shuo Zhang, born on August 7, 1985, China

Current position, grades: Ph.D. student of University of Science and Technology Beijing, Beijing, China.
University studies: Master degree in major of Industrial Engineering in Chongqing University.
Scientific interest: agent-based modelling and simulation.
Publications: 10 papers.

Yingzi Li, born on July 1, 1985, China

Current position, grades: doctor of Management, lector of University of Science and Technology Beijing, Beijing, China.
University studies: Master degree in major of Management Science and Engineering in Chongqing University.
Scientific interest: agent-based modelling and simulation.
Publications: 30 papers.

Solving electrically-large objects RCS based on 3-D vector parabolic equation method

Meng Kong[1, 2], Zhong-xiang Zhang[2*], Xiao-jing Kuang[1, 2], Ming-sheng Chen[2], Xian-liang Wu[1, 2]

[1]*School of Electronics and Information Engineering, Anhui University, Hefei, China*

[2]*School of Electronics and Information Engineering, Hefei Normal University, Hefei, China*

Abstract

The vector parabolic equation (VPE) method is introduced to calculate bistatic RCS of three-dimensional (3-D) electrically-large objects and polarization effects are fully taken into account. According to an approximate form of the vector wave equation and divergence-free condition the VPE was derived in this paper. The numerical results conducted on the scattering from perfectly conducting cube show the VPE agree with the exact method, and computation time is acceptable compared with the traditional full wave method.

Keywords: vector parabolic equation, electrically-large objects, radar cross section

1 Introduction

The Parabolic Equation (PE) is an approximation of wave equation, and it first was introduced by Leontovich and Fock [1, 2] in 1940s to treat the problem of diffraction of radio waves around the Earth. Recently, it have been applied to radar cross section (RCS) calculations [3, 4]. The scalar PE method expands the pseudo-differential operator of the scalar wave equation, discretizes objects with a series of planar element, the solution is marched in that direction from one transverse plane to the next, thus reducing the full three-dimensional problem to a sequence of two-dimensional calculations, which enhance the efficiency of computing greatly. We can get RCS at all scattering angles by near-field/far-field transformations and rotating the paraxial direction. The PE techniques may bridging the gaps between rigorous numerical methods [5] and asymptotic methods [6]. Using the PE method one can avoid both the limits of CPU time and memory by the rigorous numerical methods and those of narrow applications by the asymptotic methods, especially for electrically-large objects scattering problem.

In order to treat polarization effects fully for electromagnetic scattering, the Vector PE is obtained by coupling component scalar parabolic equations via suitable boundary conditions on scatterers. In this paper the vector parabolic equation method was introduced to calculate bistatic RCS of 3-D electrically-large objects.

2 3-D Parabolic equation

2.1 3-D SCALAR PARABOLIC EQUATION

In this paper, the time dependence of the fields is assumed as exp(-jωt). We work in Cartesian coordinates (*x, y, z*) and start with 3-D scalar wave equation

$$\frac{\partial^2 \psi}{\partial x^2} + \frac{\partial^2 \psi}{\partial y^2} + \frac{\partial^2 \psi}{\partial z^2} + k^2 n^2 \psi = 0 , \tag{1}$$

where *k* is the wave number, and *n* is refractive index. Choosing the positive *x*-direction as the paraxial direction and defining the reduced field *u* by

$$u(x, y, z) = \exp(-ikx)\psi(x, y, z) . \tag{2}$$

The scalar wave equation in terms of *u* is

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} + 2ik \frac{\partial u}{\partial x} + k^2 (n^2 - 1)u = 0 \tag{3}$$

and it can be formally factored as

$$\left\{ \frac{\partial}{\partial x} + ik(1 - \sqrt{Q}) \right\} \left\{ \frac{\partial}{\partial x} + ik(1 + \sqrt{Q}) \right\} u = 0 \tag{4}$$

in which the pseudo-differential operator *Q* is given as

$$Q = \frac{1}{k^2} \frac{\partial^2}{\partial y^2} + \frac{1}{k^2} \frac{\partial^2}{\partial z^2} + n^2 . \tag{5}$$

Equation (4) can be reduced to

Kong Meng, Zhang Zhong-xiang, Kuang Xiao-jing,
Chen Ming-sheng, Wu Xian-liang

$$\frac{\partial u}{\partial x} = -ik\left(1 - \sqrt{Q}\right)u , \qquad (6)$$

$$\frac{\partial u}{\partial x} = -ik\left(1 + \sqrt{Q}\right)u . \qquad (7)$$

Equation (6) is the called forward parabolic equation, which is most importance in many cases. Using first-order Taylor expansions of the square root $\sqrt{Q}$:

$$\sqrt{Q} \approx 1 + \frac{Q-1}{2} . \qquad (8)$$

The 3-D scalar Parabolic Equation is given by

$$\frac{\partial u}{\partial x} = \frac{ik}{2}\left[\frac{1}{k^2}\left(\frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right) + n^2 - 1\right]u . \qquad (9)$$

2.2 3-D VECTOR PARABOLIC EQUATION

Outside the objects the electric and magnetic fields $\vec{E}$ and $\vec{H}$ satisfy the vector wave equation

$$\begin{cases} \nabla^2 \vec{E} + k^2 \vec{E} = 0 \\ \nabla^2 \vec{H} + k^2 \vec{H} = 0 \end{cases} . \qquad (10)$$

In this paper we solve in terms of electric field $\vec{E}$ and magnetic field $\vec{H}$ that can be obtained from $\nabla \times \vec{E} = -j\omega\mu_0\vec{H}$, if required. We write

$$\vec{E} = \vec{E}^i + \vec{E}^s , \qquad (11)$$

where $\vec{E}$, $\vec{E}^i$ and $\vec{E}^s$ are total, incident and scattered fields, respectively. The three fields all satisfy the vector wave Equation. The reduced scattered filed u in x-direction is

$$u^s = e^{-jkx}\vec{E}^s .$$

As scalar parabolic equation approximation, the vector wave equation can be factored three scalar parabolic equations for the components (u_x^s, u_y^s, u_z^s) of u^s:

$$\begin{cases} \frac{\partial u^s_x}{\partial x} = \frac{ik}{2}\left[\frac{1}{k^2}\left(\frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right) + n^2 - 1\right]u^s_x \\ \frac{\partial u^s_y}{\partial x} = \frac{ik}{2}\left[\frac{1}{k^2}\left(\frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right) + n^2 - 1\right]u^s_y \\ \frac{\partial u^s_z}{\partial x} = \frac{ik}{2}\left[\frac{1}{k^2}\left(\frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right) + n^2 - 1\right]u^s_z \end{cases} . \qquad (12)$$

In what follows, we consider perfectly conducting objects embedded in a vacuum ($n=1$), so the tangential electric field must be zero on the surface of scattering objects. This gives the following system of equation:

$$\begin{cases} n_z u_y^s(P) - n_y u_z^s(P) = -e^{-jkx}(n_z E_y^i(p) - n_y E_z^i(p)) \\ n_x u_z^s(P) - n_z u_x^s(P) = -e^{-jkx}(n_x E_z^i(p) - n_z E_x^i(p)) \\ n_y u_x^s(P) - n_x u_y^s(P) = -e^{-jkx}(n_y E_x^i(p) - n_x E_y^i(p)) \end{cases} , \qquad (13)$$

where P is a point on the surface of scattering objects and (n_x, n_y, n_z) is the outer normal to the surface at P.

In order to obtain a well-determined system we must introduce the divergence-free condition of Maxwell's equation because the three equations of in (13) are not independent. The divergence-free equation is:

$$\frac{\partial \vec{E}_x^s}{\partial x} + \frac{\partial \vec{E}_y^s}{\partial y} + \frac{\partial \vec{E}_z^s}{\partial z} = 0 . \qquad (14)$$

Combining the first Equation of (12) and (14), we can get the vector parabolic equation as

$$\frac{i}{2k}\left(\frac{\partial^2 u_x^s}{\partial y^2} + \frac{\partial^2 u_x^s}{\partial z^2}\right) + iku_x^s + \frac{\partial u_y^s}{\partial y} + \frac{\partial u_z^s}{\partial z} = 0 . \qquad (15)$$

3 Implementation aspects

3.1 FINITE-DIFFERENCE SCHEME

As scalar parabolic equation method [7], the vector parabolic equation can be solved using finite-difference scheme. To 3-D object scattering problem, we use a double-pass method [8], where the field is first propagated assuming the object is not present at the next range. For this first pass the equations are separable and the scheme can be factored into tridiagonal matrices, which can be inverted efficiently with Gauss pivot methods. In the second pass the field is recalculated taking the object into account, using the first pass results as boundary values for a small transverse region enclosing the scatterer. A sparse matrix formulation implementing the electromagnetic boundary conditions is used for this second pass, ensuring that polarization effects are fully taken into account pass the equations are separable.

3.2 DOMAIN TRUNCATION

For domain truncation in the transverse, we added Perfectly Matched Layer (PML) as the truncation boundary conditions. We construct 3-D PML by replacing coordinate y and z with complex coordinate $\hat{y}$ and $\hat{z}$ given by

$$\hat{y} = y - i\int_0^y \sigma(\zeta)d\zeta , \qquad (16)$$

$$\hat{z} = z - i\int_0^z \sigma(\xi)d\xi , \qquad (17)$$

where

Kong Meng, Zhang Zhong-xiang, Kuang Xiao-jing,
Chen Ming-sheng, Wu Xian-liang

$$\begin{cases} \sigma(\zeta) = 0, & y_a \leq y \leq y_b \\ \sigma(\zeta) > 0, & \text{other} \end{cases} \quad (18)$$

$$\begin{cases} \sigma(\xi) = 0, & z_a \leq z \leq z_b \\ \sigma(\xi) > 0, & \text{other} \end{cases}. \quad (19)$$

Integration domain with PML absorbing boundary condition is shown in Figure 1. It is composed by the area of $[y_a, y_b] \times [z_a, z_b]$.
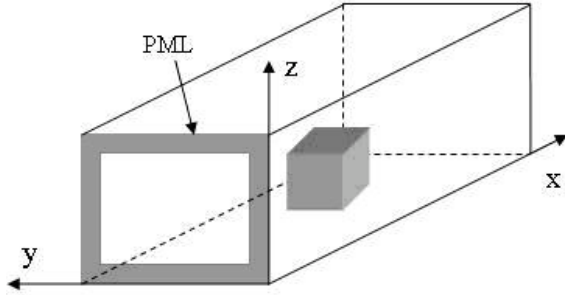
FIGURE 1 Integration domain with PML absorbing boundary condition

3.3 FAR-FIELD FORMULAS

When receiver polarization has to be taken into account, the total bistatic RCS is defined as

$$\sigma_{\vec{t}}(\theta, \varphi) = \lim_{r \to \infty} 4\pi r^2 \frac{\left|\vec{E}^s(x, y, z) \cdot \vec{t}\right|^2}{\left|\vec{E}^i(x, y, z)\right|^2}, \quad (20)$$

where $x = r\cos\theta$, $y = r\sin\theta\cos\varphi$, $z = r\sin\theta\sin\varphi$.

We assume the receiver is polarized along vector $\vec{t}$. If the incident field is a plane wave of unit amplitude, we can get the equation:

$$\sigma_{\vec{t}}(\theta, \varphi) =$$

$$\frac{k^2 \cos^2\theta}{\pi} \left| \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \vec{E}^s(x_0, y, z) \cdot \vec{t} e^{-ik\sin\theta(y\cos\varphi + z\sin\varphi)} dy dz \right|^2 \quad (21)$$

in which x_0 can usually be chosen as 10λ (λ is the wavelength of incident wave). The total RCS is obtained by summing x, y, z components of RCS.

4 Numerical results

In this numerical example the incident field is a plane wave with its amplitude set to be 1.0, and incident angle is 0°. The incident source with its wavelength of 1m (frequency 0.3 GHz) is vertically polarized and propagating in a vacuum (*n*=1), which illuminated a perfectly conducting cube 12λ on each side. We use PML absorbing boundary conditions at computation domain. To obtain the bistatic RCS (0°-180°), we need six rotated PE runs for VPE. Figure 2 and Figure 3 show bistatic RCS of perfect conducting cube 12λ on each side for E and H plane

patterns respectively. The results proved the validity of the present method. Using the VPE method, the CPU consumed by this example was under 10 min on desktop computer, but this case would be quite stressful for traditional Method of moments [8].
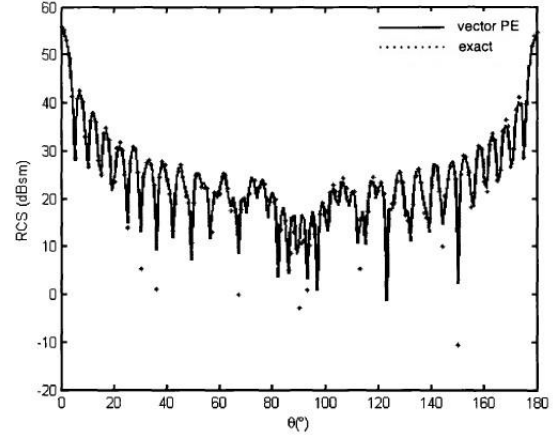
FIGURE 2 Bistatic RCS of perfect conducting cube 12λ on each side (E plane)
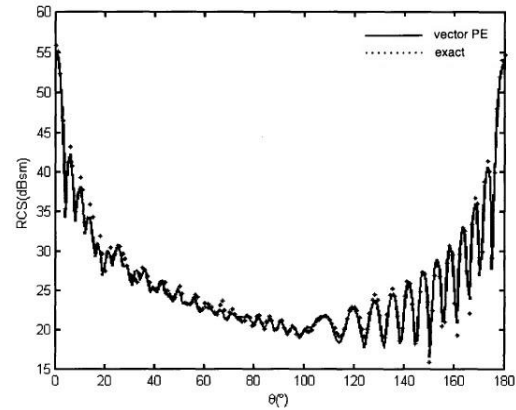
FIGURE 3 Bistatic RCS of perfect conduct cube 12λ on each side (H plane)

5 Conclusions

The vector parabolic equation allows accurate treatment of polarization effects within the paraxial constrains. The combination of VPE formulation with the rotated PE methods provides a powerful tool for electromagnetic scattering problem. The results of example proved the efficiency of the VPE method to calculate Electrically-large Objects RCS. The work presented here is limited to perfectly conducting objects embedded in a homogeneous background.

Mathematical and Computer Modelling

Kong Meng, Zhang Zhong-xiang, Kuang Xiao-jing,
Chen Ming-sheng, Wu Xian-liang

References

[1] Leontovich M A, Fock V A 1946 Solution of the problem of propagation of electromagnetic waves along the earth's surface by the method of parabolic equation *J Phus* **10**(1) 13-23

[2] Fock V A 1965 *Electromagnetic diffraction and propagation problems* Pergamon.Press New York

[3] Levy M F, Borsboom P P 1996 Radar cross-section computations using the parabolic equation method *Electron.Lett* **32**(13) 1234–6

[4] Mallahzadeh A R, Soleimani M 2006 RCS Computation of airplane using parabolic equation *Progress In Electromagnetics Research* **57** 265–76

[5] Pocock M D, Bluck M J, Walker S P 1998 *IEEE Trans Antennas Propagat* **46**(8) 1212-9

[6] Domingo M, Torres R P, Catedra M F 1994 *IEEE Trans Antennas Propagat* **42**(6) 885–8

[7] Zaporozhets A A, Levy M F 1998 Radar cross section calculation with marching methods *Electronics Letters* **34**(20) 1971-2

[8] Levy M F 2000 Parabolic equation methods for electromagnetic wave propagation *The Institution of Electrical Engineers Press London*

Authors

Meng Kong, born on March 31, 1983, Anhui province, China

Current position, grades: lecturer of Electronics and Information Engineering in Hefei Normal University.
University studies: Electronic and Information Engineering in Anhui University.
Scientific interest: computational electromagnetics.

Zhong-xiang Zhang, born on December 12, 1976, Anhui province, China

Current position, grades: doctor of Electromagnetic Field and Microwave Technology, associate professor in Hefei Normal University.
University studies: Electronic Engineering and Information Science in University of Science and Technology of China.
Scientific interest: computational electromagnetics, microwave and millimeter wave engineering.

Xiao-Jing Kuang, born on October 17, 1984, Anhui province, China

Current position, grades: lecturer in Hefei Normal University.
University studies: Electronic and Information Engineering in Hefei Normal University.
Scientific interest: time-domain numerical methods and metamaterials.

Ming-sheng Chen, born in October, 1981, Anhui province, China

Current position, grades: professor of Electronics and Information Engineering in Hefei Normal University.
University studies: Electronic and Information Engineering in Anhui University.
Scientific interest: computational electromagnetics.

Xian-liang Wu, born in August, 1955, Anhui province, China

Current position, grades: professor of Electronics and Information Engineering in Hefei Normal University.
Scientific interest: computational electromagnetics, microwave engineering.

Empirical research on the diffusion of home appliances to the rural areas in china based on Bayesian estimated bass model

Yanhong Guo*, Wei Liu

[1]*Faculty of Management and Economics, Dalian University of Technology, 116024, China*

Abstract

Quantitative research on the diffusion of Home Appliances to the Rural Areas in China is a very important topic both in theory and in application. Most of the research on diffusion focuses on the discussion and qualitative analysis of the policy. This paper expands the Bass model applications by leading an empirical research based on the history sales data of Home Appliances to the Rural Areas. The model has been estimated by both linear and nonlinear least squares method and Bayesian method separately, and predicts the maximum potential market in the end. This empirical research shows that Bayesian method performs best in predicting the diffusion of Home Appliances to the Rural Areas. Finally, the main factors influencing the policy implementation was analysed. It predicts the periodical sales and gives some suggestions to the government and enterprises accordingly.

Keywords: home appliances to the rural areas, diffusion of innovation, Bass model, Bayesian method

1 Introduction

Due to the financial crisis, Chinese government Appliances to the Rural Areas in China has boosted Chinese rural market. However, there is a series of unknown factors in the market: the prospect of the home appliances in the rural areas, the space of the rest of the market, the amount of the annually financial subsidy budget, the method of enterprises allocating resources, and so on. With these phenomena, there is a key factor, which is the forecast of the Home appliances market demand. It requires us to make precise predictions of the speed and scale of the diffusion of the home appliance products. The government, then, could make an accurate estimate of the governmental financial subsidy budget and the relevant formulation to guide the enterprises properly. The enterprises, then, could organize relevant production and development based on the estimation. For government, enterprises and the farmer consumers, it is clear that it has a practical significance to study the diffusion process of home appliance, predict the market capacity and the sales amount in each period and find important factors of the home appliance diffusion and the impact mechanism to the market.

On the other hand, Bass Model has been widely applied to the field of durable consumer goods and technology innovation diffusion. But there are several questions related to the bass model. Is it suitable to use it to analyse a variety of consumer durable goods, which are mixed together in the electrical home appliances especially with the effect of government subsidy? How does the brand effect with the mass media in the process of rural consumer accepting home appliances? What is the difference between the general consumer durable diffusion and

innovation diffusion, and so on? Few literature studies could be found on this topic. This paper will focus on the diffusion study in the market of the home appliance products in the rural market. We propose the Bayesian parameter estimation method to the Bass Model of the application to compensate for the random error of the prediction due to the small sample size in the original studies.

This paper will analyse the pattern of the dynamic development and its potential power. On the one hand, it will expand the application of Bass Model. On the other hand, it could provide scientific evidence for the formulation of the governmental policy and the decision of the production and management of the enterprises.

2 Diffusion research on the home appliance products in the rural areas in China

The policy of home appliance products in the rural areas in china is an important move to thoroughly implement the government's proposal to expand the domestic demand. The detail of the policy is to start a country-wide project with an emphasis on the rural consumers in the R&D, production, and sale of the home appliances with fixed orientation, with government subsidy to certain levels. We call it a government supported "rural electrical and electronics propulsion project."

The home appliances diffusion is different to those of the other single products because it mixes various types of goods: television, refrigerator, air condition, microwave oven, washing machine, computer, mobile phone and so on. Although it is a mixture, it is clear that these consumer goods all belong to the durable ones. Consumers tend to be more careful when they make a purchasing decision. They

---

Corresponding author e-mail: guoyh@dlut.edu.cn

shall always ask the opinions of the people around, and pay attention to the information on the television, broadcast, and newspapers and so on. These features are consistent with the ones of single products.

Fan Qiu Ju (2010) and Dunn Adam G (2012) proposed that the external influence is the promotion and implementation of national policy on home appliances. Until now, the U.S. subprime mortgage crisis has had a negative impact on China's home appliance exports. In order to mitigate the impact, the country was rolled out of pilot job of home appliances to the countryside which is directed by R&D, production, sales of designated home appliances for the rural consumer and capital subsidy given by the financial sector. After achieving excellent sales results, the second round of home appliances to countryside process was implemented fully and by stimulating domestic consumption and stimulating domestic demand would accomplish the goals of economic growth. With the coordination and propaganda of government and media, this policy has been fully understood by the farmers: they can buy products with reliable brand names with tangible subsidies. Then the farmers start to buy. This is the external factors of promoting the home appliance to countryside. These external factors promote innovators to proliferate to go to buy them firstly.

After obtaining satisfying outcomes, the farmers who have purchased the home appliances in the countryside would be willing to recommend this policy to more families to make a purchase. Their purchasing behaviour would inadvertently affect other farmers. Seeing the tangible benefits, other farmers will be stimulated to make their decisions to purchase. And they do. Wurui Juan (2010) proposed that this is the internal factors of home appliances to countryside – a process of the imitators following the innovators.

To sum up, home appliances in the countryside can be viewed as a consumer durable goods diffusion by Lu Dong Ning (2009). On one hand, it will promote the innovator in the farmer consumer groups to purchase appliance firstly under the influence of government, business, media and other external propaganda. On the other hand, as the key factor, the exemplary role and word of mouth will affect the amount and the rate of countryside home appliances' diffusion. The elements of diffusion and the factors of the Bass Model are very relevant in this sense. So we have chosen the Bass Model as countryside home appliance diffusion forecasting model to research the application of the Bass Model in the forecasting of countryside home appliances.

3 BASS model and its literature review

3.1 BASS BASIC MODEL

Bass Frank M (1969), D. Heckerman (1995) and Van den Bulte (2007) proposed that potential adopters are influenced by two factors: the mass media and word of mouth which are the two simple assumptions in Bass model. Innovators are influenced by the mass media, and imitators are influenced by the word of mouth. Bass's model is represented mathematically as:

$$h(t) = \frac{f(t)}{1-F(t)} = p + qF(t), \tag{1}$$

where: $h(t)$ is the chance of using new products, $f(t)$ is the time density function of adopter in the t period, $F(t)$ is the ratio of cumulative adopter between the totals, p represents the external influences (or Innovation coefficient or the influence of mass media), q represents the internal influence (or the imitation coefficient, or, word of mouth).

Then let us assume m be the maximum potentiality of the market or the maximum capacity of the market. The number of adopter $n(t)$ at t time is the formula two:

$$n(t) = mf(t). \tag{2}$$

The number of cumulative adopter is:

$$N(t) = mF(t). \tag{3}$$

By integrating Equation (2) we will get:

$$N(t) = m\left[\frac{1 - e^{-(p+q)t}}{1 + \frac{q}{p}e^{-(p+q)t}}\right]. \tag{4}$$

3.2 THE BAYESIAN PARAMETER ESTIMATION METHOD OF BASS MODEL

The accuracy of Bass Model parameter estimation is one of the key factors of the Diffusion Model. Because Bayesian parameter estimation method combines priori information with posteriori information, and it can effectively reduce random error because of few observations. This study selects Bayesian parameter estimation method as the basic method of the model, exploring the application of the Bayesian parameter estimation method in the Bass Model. In order to estimate the parameter m, p, and q in the Bass Model, we need to show the priori distribution of the parameter firstly, as follows:

Let $y(i)$ be the number of annual new adopters, and where $e_i \sim N(0, \sigma^2)$. Then

$$y(i) = m(F(t_i) - F(t_{i-1})) + e_i. \tag{5}$$

Assume as $\sigma = e^\alpha$ in Equation (5) and each parameter's prior distribution is:

$$P(p) = N(\mu_p, v_p), P(q) = N(\mu_q, v_q),$$
$$P(m) = N(\mu_m, v_m) \text{ and } P(\alpha) = N(\mu_\alpha, v_\alpha).$$

We will get the likelihood function:

$$\left(\frac{1}{\sqrt{2\pi e^{2\alpha}}}\right)^{n}\exp\left(-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{y(i)-m\left(F(t_i)-F(t_{i-1})\right)}{\exp(\alpha)}\right)^2\right), \quad (6)$$

where, n represents the period number of data.

Since the observation was produced from data, we can take the likelihood function as the conditional density function $P(y|p,q,m,\alpha)$ of that observation regarding $\theta=\{p,q,m,\alpha\}$.

Assume $h(\theta,y)(\theta,y)$ as the joint density function of, θ, y. Then $h(\theta,y)=P(y|\theta)\pi(\theta)$, where $\pi(\theta)$ is a parameter. Then from the prior density function and the Bayesian equation, we can get:

$$h(\theta,y)=P(y|\theta)\pi(\theta)=P(\theta|y)m(y), \quad (7)$$

Where $P(\theta|y)$ is the posterior density function of parameters and $m(y)$ is the marginal density function of y where:

$$m(y)=\int_{\Theta}h(\theta,y)d\theta. \quad (8)$$

From Equation (7), we get the posterior density function:

$$P(\theta|y)=\frac{P(y|\theta)\pi(\theta)}{\int_{\Theta}P(y|\theta)\pi(\theta)d\theta}. \quad (9)$$

It is not easy to carry out the multi-dimensional integral that contains a number of unknown parameters. By pulling in MCMC method and Gibbs sampling, it generates some steady-state distribution as the Markov chain of the parameters' posteriori distribution. The average of all the iteration generated from steady Markov chain is the Bayesian estimation.

4 The experiments and analysis

4.1 THE SCREENING AND THE METRIC OF THE DATA SETS

According to the announcement regarding the information management system of the home appliances going to the countryside on the official website, we got the data of the national monthly sales volume of home appliances in the countryside from January 2009 to December 2009. Each data point represents the sales in all provinces of all kinds of household appliances, which are reported monthly.

This paper uses the MAE method to measure the preciseness of prediction, and the smaller the MAE the better. In the test set, the future points are $y=\{y_i|i=1,2,\ldots,n\}$, the real sales data set is

$s=\{s_i|i=1,2,\ldots,n\}$, to each pair of "forecasted sales, real sales", to $\langle y_i,s_i\rangle$, there will be Equation:

$$MAE=\frac{\sum_{i\in N}(y_i-s_i)}{N}. \quad (10)$$

4.2 RESULTS OF THREE PARAMETERS ESTIMATION METHODS

TABLE 1 Estimation results of the linear least square method

	Estimation	*T* value	*P* Value
1	838088	1.484	0.171
x1	0.330	3.137	0.0119
x2	-6.10043E-09	-1.785	0.107
***R* squared**	0.763	***F* Ratio**	14.518
Adjusted *R* squared	0.710	***P* Value**	0.001

TABLE 2 Estimation results of the non-linear least square method

p	*0.00999058*	{0.002,0.0176}	
q	0.293087	{0.067,0.519}	
	DF	**Mean Squared**	***F* ratio**
Model	3	5.08494×10^{13}	47.918
Error	9	1.06117×10^{12}	
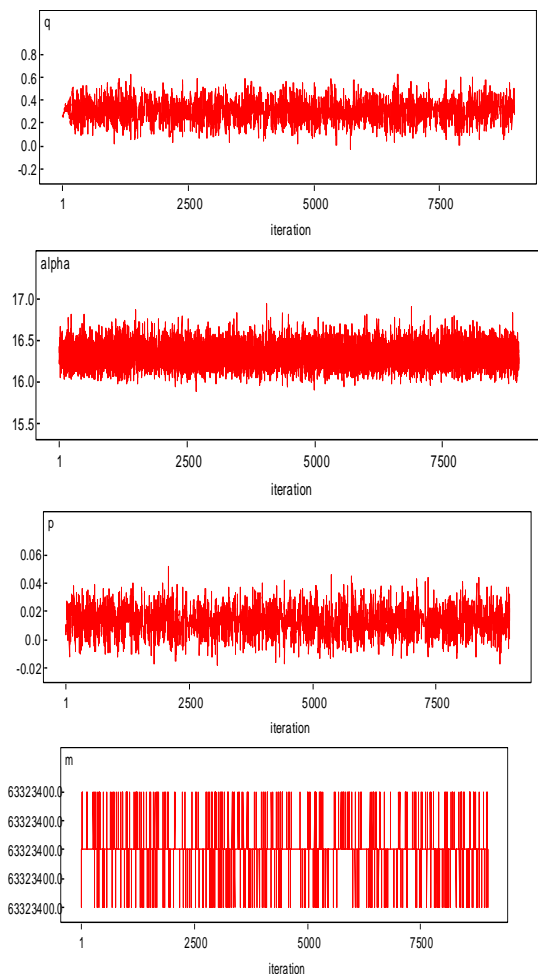
FIGURE 1 The convergence process of Bayesian parameters

Estimation results of least square method and non-linear least square method are presented in Table 1 and Table 2. Figure1 shows the convergence process of Bayesian parameters. The values given by the linear least square method are 56565700, 0.0148, 0.3451 correspondingly. The values given by the non-linear least square method are 56565700, 0.0148, 0.3451 correspondingly.

By the Bayesian parameters method we get the values of "m, p, q" are 63320000, 0.01293, 0.3044 correspondingly. Predicted values got by the three parameter estimation methods are presented in Table 3. We can see that the prediction results by the Bayesian parameters method are much better than those of the other two methods.

TABLE 3 Historical data and predictions (Unit: Tai)

Period	Historical data	Forecast data		
		Linear least square method	Non-linear least square method	Bayesian Estimated Bass method
1	338135	540330	649362	818727
2	875390	723418	835399	1054139
3	1484517	958139	1069371	1347749
4	1763304	1259188	1360065	1707577
5	2227725	1637793	1715493	2138392
6	2921096	2101288	2140993	2638363
7	4273119	2648019	2636329	3194646
8	3787773	3260276	3191807	778635
9	3159942	3896577	3783891	4342709
10	7046823	4486926	4371676	4821439
11	3871402	4937224	4896605	5140535
12	5930591	5148845	5288332	5234566
MAE		841149	805103	64040

5 Conclusion and future research

In summary, the growth mechanism of Bass model is applied to forecast National sales of home appliances in the countryside. And we can analyse and calculate its dynamic growth characteristics by the Bass model. Bayesian parameter estimation method has a good effect on the Bass model predictions. In addition, Because of declining growth rate of home appliances sales, diffusion of Home Appliances in the Rural Areas are dominated by the internal factors. According to the model, the government and the enterprise can estimate and predict the sales and market saturation time. Furthermore, it can make policy adjustment accordingly.

Ignoring the factors of the price of the home appliance product and the purchasing power of farmers, and so on, this paper is committed to investigating the impact of mass media and reputation on the policy of home appliances. In the future research, we will gradually add these variables in order to get deeper and more detailed investigation both quantitatively and qualitatively on the Diffusion of Home Appliances in the Rural Areas. On the other hand, the government continues to increase the varieties of home appliances in the countryside, and introduce a home appliance recycling policy. The future research will focus on the influence the policy on the diffusion of the appliances in the countryside.

References

[1] Baten M A, Kami A A 2013 Inventory-Production Control Systems with Gumbel Distributed Deterioration *International Journal of Applied Mathematics and Statistics* **34** 30-51
[2] Fouad R H, Mukattash A M 2012 Specifying Optimal Number of Cells Using a Modified Dissmilarity Measure Based on AVV Aproach *International Journal of Applied Mathematics and Statistics* **28** 96-100
[3] Kijek T, Kijek A 2011 Modelling of Innovation Diffusion. *Operations Research & Decisions* **3-4** 53-68
[4] Bass F M 1969 A New Product Growth Model for Consumer Durables. *Management Science* **15**(5) 215-27
[5] Boehner R, Gold S 2012 The Influence of the Marketing Mix on the Diffusion of Innovation: Bass Model Redux *Academy of Business Journal* **2** 40-65
[6] Heckerman D 1995 *A tutorial on learning with Bayesian networks. Technical Report MSR-TR-95-06* Microsoft Research revised November 1996
[7] Dunn A G, Braithwaite J 2012 Nation-scale adoption of new medicines by doctors: an application of the Bass diffusion model *BMC Health Services Research* **12**(1) 248-56
[8] Fan Q 2010 Investigation and analysis On the home appliance to the country of present situation and problems *Chinese Soft Science* **3** 186-92 *(in Chinese)*

[9] Jesen F V, Lauritzen S L, Olesen K G 1990 *Bayesian updating in recursive graphical models by local computation* Computational Statistics Quarterly **4** 269-282
[10] Lu D 2009 Thinking home appliance to the country policy on based on home appliance of demand forecasting *Rural economy* 22 *(in Chinese)*
[11] Mahajan V, Muller E, Bass F M 1990 New Product Diffusion Models in Marketing: A Review and Directing for Research *Journal of Marketing* **54** 1-26
[12] Van den Bulte C, Joshi Y V 2007 New Product Diffusion with Influentials and Imitators *Marketing Science* **26** 400-421
[13] Wurui J, Dong J, Wu B 2010 Chinese farmers consumers Purchase intention of the countryside home appliances *Soft Science* 1 40-52 *(in Chinese)*
[14] Xu X, Song Q 2007 Based on improved BASS model for short life cycle product demand forecast *Industrial engineering and management* **5** 27-31
[15] Xie J, Song M, Sirbu M, Wang Q 1997 Kalman Filter Estimation of New Product Diffusion Models *Journal of Marketing Research* **34** 378-93

Authors

Yanhong Guo, born in April, 1977, Jilin Province, China

Current position: lecturer.
University studies: PhD in management science and engineering.
Scientific interest: market engineering, financial marketing.
Publications: 3 papers.

Wei Liu, born in July, 1990,Shandong Province, China

Current position, grades: graduate student of Dalian University of Technology.
University studies: Bachelor degree in enterprise management at China agricultural University.
Scientific interest: marketing engineering.

Dynamic multi-species coevolution large-scale optimize based on the fuzzy clustering and trust region

Xuhui Zhang[1, 2*]

[1]*Department of Mechanical Engineering, Xi'an University of Science and Technology, No.58 Yatan Road, Xi'an 710054, China*

[2]*Xi'an Coal Mining Machinery Co., Ltd., Xi'an 710032, China*

Abstract

Based on the above-mentioned studies, this article put the modified Fuzzy Clustering method into the particle swarm optimization, which solved the curse of dimensionality existing in the conventional algorithms. The large-scale parameter optimization method applied the modified Fuzzy C-Means method to clustering the large-scale dimension under the coevolution frame and achieved the valid dimension grouping. Later on, the dynamic neighbourhood topology multi-species particle algorithms divided the whole species into packets and constructed the subspecies sharing neighbourhood information, which improve the searching efficiency. The bring-in trust region could have the self-adapt adjustment for the particle optimization range, accelerate the optimizing speed, and decrease the iterations in the dead space. We use 20 standard large-scale testing functions for simulation. Compared with the top-ranked tournament algorithm, the mentioned algorithm achieved a better optimizing result in most functions, which surely laded the foundation of the large-scale neural network parameter optimization and the application in the control system.

Keywords: multi-species coevolution, particle swarm optimization, guzzy clustering

1 Introduction

Particle Swarm Optimization doesn't rely on the specific objective function form and avoided the counting process of the Jacobi matrix and Hessian matrix in the classical mathematics methods. But its search capability would heavily decline facing the large-scale problem. The phenomenon was not only caused by the existence of the locally optimal solution, but also by the possible degradation of the particle's speed, which lead to the search in a continuous subspace of the whole searching hyperplane. Besides, there are many complicated high-dimensional optimization problems in the actual industrial production process. Therefore, it's highly necessary to find out a fast and efficient algorithm to solve the large-scale optimization problem. This article put forward the Dynamic Neighborhood Particle Swarm Optimization Algorithm with the Fuzzy Clustering, and mainly improved the three things as follows:

Adjust searching space in the self-adapt way and accelerate the optimization process. When the next iterative point could fully decrease the objective function, the trust region (Independent variable search space) would diminish and the particle could rapidly find out the optimal point, without wasting more iteration in the dead space.

The most important problem in the coevolution is the packets. How many dimensions should be there in one packet? And which dimensions should appear in the same packet? Hence, using the improved Fuzzy C-Means algorithm suggested in this article to analyse the

relationship between variables and divide the similar variables into one packet for optimizing.

Adopting the Dynamic Neighbourhood Particle Swarm Optimization Algorithm, this article introduced an optimize method in packets. The self-Adapting topological structure could learn the neighborhood position and self-organize and construct the subpopulation to share the information, which improve the continuous space search ability

2 The particle large-scale parameters optimization method based on the Fuzzy C-Means Clustering

2.1 FCM CLUSTERING ALGORITHM

Traditional clustering algorithms assign each object to one and only one cluster in accordance with strict standards. The clusters have distinct boundaries. Objects in a cluster are very dissimilar from objects in the other clusters. This is called hard clustering. Traditional clustering algorithms are all based on *"hard partition"* [1]. However, things in the objective world are closely connected to each other; nothing can exist in total isolation. In other words, there is not a thing that belongs to one and only one class; objects cannot be divided into classes that share no similarities. People have a better understanding of the world as it develops. But traditional hard clustering methods fail to solve practical problems well. *"Soft partition"*, a more realistic clustering method, was thus created in contrast to *"hard partition"*.

Fuzzy C-Means Clustering Algorithm (FCM clustering algorithm) is a fuzzy clustering algorithm based on objective functions, which is different from traditional clustering algorithms. FCM clustering algorithm, in essence, is rooted in Hard-means Algorithm (HCM) [2].

2.2 SOLVING PROCESS OF FCM CLUSTERING ALGORITHM

Based on the principle of least squares, FCM clustering algorithm aims to minimize its objective function $J(\mu, A)$ and realize convergence of the iteration process. The objective function $J(\mu, A)$ of FCM is defined as:

$$J(\mu, A) = \sum_{i=1}^{c} \sum_{j=1}^{n} \mu_{ij}^{m} \left\| X_j - A_i \right\|^2 , \qquad (1)$$

where $A_i (i = 1, 2, ..., c)$ is the cluster centre and m ($m>1$) is the fuzziness index. Below are the value formulas of μ and A during convergence of the iteration process:

$$\mu_{ij} = \begin{cases} \left(\sum_{k=1}^{c} \dfrac{\left\| X_j - A_i \right\|^{2/(m-1)}}{\left\| X_j - A_k \right\|^{2/(m-1)}} \right)^{-1}, \left\| X_j - A_i \right\| \neq 0; \\ \\ 1, \qquad \left\| X_j - A_k \right\| = 0 (k = j); \\ \\ 0, \qquad \left\| X_j - A_k \right\| = 0 (k \neq j), \end{cases} \qquad (2)$$

$$A_i = \sum_{j=1}^{n} \mu_{ij}^{m} X_j / \sum_{j=1}^{n} \mu_{ij}^{m} . \qquad (3)$$

Suppose the sample space $X = (x_1, x_2, ..., x_m)$ and c denotes all positive integers greater than 1. A fuzzy matrix $\mu = (\mu_{ij})$ is used to divide Z into c classes. μ_{ij} means the ith sample point belongs to the j-th membership. Obviously, μ_{ij} meets the following conditions:

$$\mu_{ij} \in [0,1], \qquad i = 1, 2, ..., n, \quad j = 1, 2, ..., c$$

$$\sum_{j=1}^{c} \mu_{ij} = 1, \forall i, \quad i = 1, 2, ..., n, \quad j = 1, 2, ..., c . \qquad (4)$$

$$\sum_{i=1}^{n} \mu_{ij} > 1, \forall j, \quad i = 1, 2, ..., n, \quad j = 1, 2, ..., c$$

The iteration process of FCM clustering algorithm is as follows [3, 4]:

Step 1: Start with a random initialization of $\mu^{(0)}$; initialize $A^{(0)}$ and calculate $\mu^{(0)}$; make the number of iterations $k=1$; select the number of cluster centres c and the fuzziness index m;

Step 2: Calculate the cluster centre, assign a value to $A^{(k)}$, and calculate $\mu^{(0)}$ according to Expression (2);

Step 3: Recount the membership, assign a value to $\mu^{(k)}$, and calculate $A^{(k)}$ according to Expression (3);

Step 4: If $\max | \mu_{ij}^{k} - \mu_{ij}^{k-1} | \leq \varepsilon$ and the algorithm convergence is achieved, stop the iteration process; otherwise, make $k=k+1$ and return to Step 2; ε is a given threshold value.

2.3 PARAMETER SETTING OF FCM CLUSTERING ALGORITHM

The fuzzy weighting exponent m is a parameter introduced by Bezdek in order to popularize the criterion functions in fuzzy clustering. m must lie between $0<m<\infty$. When m approaches to 1, FCM clustering algorithm will lose its fuzziness; when m approaches to $+\infty$, FCM clustering algorithm will lose the meaning of clustering. We usually assign 2 to m.

2.4 ADVANTAGES AND DISADVANTAGES OF FCM CLUSTERING ALGORITHM

FCM clustering algorithm is featured by simple design, fast convergence, wide application areas, easy implementation, and strong local search ability. The algorithm is rooted in traditional C-means clustering algorithms and in essence, it is a type of local search algorithm. So, it inevitably has the disadvantages of C-means clustering algorithm. If the initial cluster centre is next to a local optimum, the algorithm will converge to a local minimum. It is sensitive to the initial value and noisy data. For different initial values, the clustering results are dissimilar. As a result, the objective function may easily falls into local optima. Worse still, it may lead to a degenerate solution or no solutions at all [5].

3 PSO algorithms

3.1 DESCRIPTION OF PSO

Suppose the speed of a basic particle is v and its position is x. The iterative formula of the basic PSO algorithm is described as:

$$v_{t+1} = v_t + c_1 r_1 (pbest - x_t) + c_2 r_2 (gbest - x_t) , \qquad (5)$$

$$x_{t+1} = x_t + v_{t+1} . \qquad (6)$$

Equations (5) and (6) are respectively the iterative formulas of the speed and position. c_1 and c_2 are learning factors; r_1 and $r_2 \in U(0,1)$; $pbest$ and $gbest$ stand for the position of the optimal solution of the individual and the entire group, separately. The value of speed v is:

$$v_{t+1} = \begin{cases} v_{max}, v_{t+1} > v_{max} \\ v_{t+1}, -v_{max} < v_{t+1} < v_{max} \\ -v_{max}, v_{t+1} < -v_{max} \end{cases},\qquad (7)$$

where v_{max} is the value range of speed v. The iterative formula of speed v in Expression (6) is divided into three parts: v_t, $c_1 r_1$ ($pbest$-x_t), and $c_2 r_2$ ($gbest$-x_t). The first part stands for the speed of the particle; the second part represents its cognition level, meaning the particle should be considered based on its own conditions; the third part is the social part of the particle in group learning, which means that learning is conducted through information exchange with the *"society"*. The particle updates itself through its current state (first part), self-learning (second part), and the social information (third part). If the third part is not available, the particle will lack information exchange, and the probability of obtaining the optimal solution is thereby reduced. The behaviour of working behind closed doors and learning blindly will lead to a lack of exchanges and has a negative influence on the final results. If the second part is not available, the particle will easily fall into local optima, though there are enough exchanges and extended searching space. Hence, PSO is an intelligent model composed of both *"cognitive part"* and *"social part"* [6, 7].

However, the basic PSO algorithm ignores the first part – the effect of the particle speed on algorithm convergence. To solve the problem, Y. Shi and R. C. Eberhart improved the basic PSO algorithm in 1998 by adding an inertia weight factor into the iterative equation of the speed. The improved equation is:

$$v_{t+1} = w v_t + c_1 r_1 \left(pbest - x_t \right) + c_2 r_2 \left(gbest - x_t \right),\qquad (8)$$

where w is the inertia weight factor and the rest parameters are the same as before. At present, the standard PSO algorithm is the amended algorithm with an inertia weight factor [8].

3.2 ITERATIVE STEPS OF STANDARD PSO

Step 1: Initialize m particles, including their initial speed and positions; initialize all relevant parameters;
Step 2: Suppose the present position of the particle is the best position $pbest$; select the global best position $gbest$ according to the best position of all particles;
Step 3: Each particle updates its speed in line with Expression (8), keeps the speed within v_{max} based on Expression (7), and updates its position according to Expression (6);
Step 4: If the current position after the update is better than the previous best position of the particle, update $pbest$ and replace the previous best position with the current position;
Step 5: Update the global best position $gbest$ according to the best position of all particles;
Step 6: If it meets the requirements for termination, stop the iteration process; otherwise, return to Step 3.

4 SP-FCM algorithm design

4.1 BASIC IDEAS OF THE ALGORITHM

Rooted in HCM, FCM clustering algorithm inherits both its advantages and disadvantages. In nature, it is also a type of local search algorithm. Below are the disadvantages [9]:
1) Very sensitive to the initial values and noisy data; easy to fall into local optima;
2) The rate of convergence will be greatly reduced if the initial cluster centre is far away from the optimal point;
3) The algorithm will converge to a local optimum if the initial cluster centre is next to the local optimum.

To solve the above disadvantages, the thesis uses PSO, SFLA, and other intelligent optimization algorithms to find the optimal initial cluster centre and thus prevent FCM from being sensitive to initial values or falling into local optima. The specific theoretical basis is as follows [10]:
1) PSO has strong global search ability, fast rate of convergence, few parameters, and other advantages. PSO makes it easier to rapidly lock the range of solutions of the optimal initial cluster but also easier to fall into the local optimum;
2) SFLA has the advantages such as strong global optimization ability, strong ability to jump out of local optima, and ability to enhance the search ability of the optimal initial cluster centre. But it has poor local search ability and slow rate of convergence.

Therefore, the thesis puts forward a new SP-FCM algorithm on the basis of current literature. Experiments show the algorithm can effectively increase the searching ability and clustering results of fuzzy clustering algorithms. It combines SFLA and PSO by setting a searching granularity coefficient. Thus, it can find the optimal initial cluster centre and avoid the disadvantages of FCM by dint of the advantages of the two. After that, it adopts the basic FCM algorithm [11].

4.2 IMPROVED BASIC ALGORITHMIC PROCESS

The algorithmic process contains four modules: Init Module, SP Module, FCM Module, and Output Module. The specific functions of each module are as follows.

4.2.1 Init Module

The function of this module is to initialize Q cluster centres. Make the cluster sample set $X_t = (x_1, x_2, ..., x_i, ..., x_n)$, where x_i is the d-dimensional vector. Make an individual represent the set of cluster centres, namely, $C_t = (c_1, c_2, ..., c_i, ..., c_c)$; c_i is the d-dimensional code of the ith cluster centre. Real-number encoding is applied to each d-dimensional cluster centre. Count the fuzzy matrix and the fitness value. Arrange the Q cluster centres in descending order of their fitness values. The fitness function of individual c_i is defined as:

$$f(x_i) = \frac{1}{J(\mu, A) + 1} .\qquad (9)$$

4.2.2 SP Module

This module is aimed at finding the optimal cluster centre according to the fitness value. Combine PSO and SFLA through a searching granularity coefficient. Integrate PSO into SFLA and update the operators of SFLA. The basic procedure of the module is shown in Figure 1.
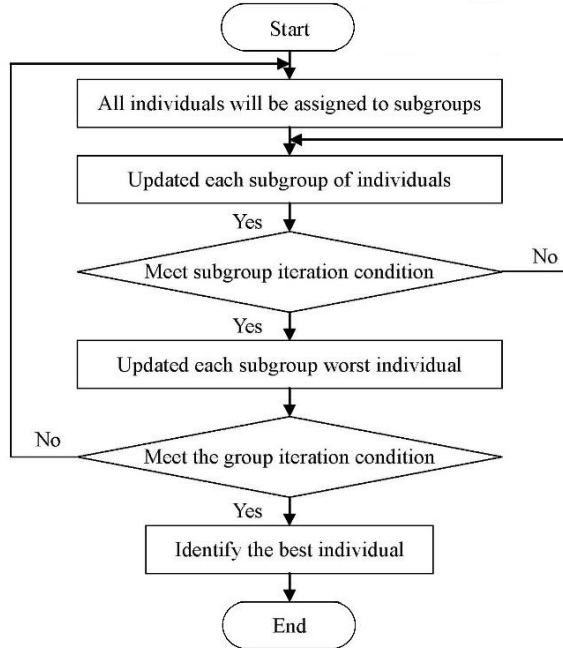
FIGURE 1 Flow Chart of SP Module

The algorithm needs to fully make use of the strong search ability of PSO in the earlier stage so as to narrow down the search range and accelerate convergence. During late stages of the iteration, it is necessary to make full use of the ability of SFLA to jump out of local optima so as to expand the search range and avoid falling into local optima. PSO and SFLA have their own advantages and disadvantages [12]. To put it simply, the combination of the two does not only inherits their advantages but also their disadvantages. The update strategy of PSO is used in each subgroup during the search. Meanwhile, mixed update operation is applied to each subgroup, which reduces the rate of algorithm convergence as well as the local search ability of PSO. Thus, the thesis proposes a searching granularity coefficient for the search based on the advantages of the two. So, the algorithm can make full use of the advantage of PSO in realizing rapid optimization in the earlier stage so as to find the basic range of the optimal solution rapidly; after that, it greatly increases the ability of SFLA to jump out of local optima and thus avoids falling to a local optimum. The searching granularity coefficient is defined as:

$$Granularity = \left[\sqrt[n]{\frac{Iter}{runIter + \varepsilon}} \right],\qquad (10)$$

where ε is the initial searching granularity; η is the global searching granularity; *Granularity* represents the searching granularity coefficient; *Iter* is the group's maximum number of iteration; *runIter* is the group's current number of iteration. We can control the requirements for searching precision by initializing initial searching granularity ε and global searching granularity η. Prior knowledge is required for each data set. Following a linear decrease, we assign the values of 0.9 to 0.1 to inertia weight factor w. The definition is as follows:

$$\omega = \omega_{max} - run \frac{(\omega_{max} - \omega_{min})}{runMax} ,\qquad (11)$$

where ω_{max}, ω_{min}, *run*, and *runMax* denote the maximum inertia weight coefficient, the minimum inertia weight coefficient, the current number of iterations, and the maximum number of iterations, respectively.

4.2.3 Fcm module and output module

The function of Fcm Module is to update the fuzzy matrix through FCM algorithm after obtaining the optimal cluster centre c. Output Module outputs the clustering results based on the fuzzy matrix [13].

4.3 ALGORITHM DESCRIPTION

Step 1: Initialize Q cluster centres; *runIter* is the counter of the group's number of iterations and *runsubIter* is the counter of the subgroup's number of iterations; initialize all relevant parameters;
Step 2: Count the fuzzy matrix of each individual with Equation (2) and count the fitness value $f(x_j)$ with Equation (9);
Step 3: Arrange the cluster centres in descending order of the fitness values and assign each individual in Q cluster centres to subgroup q by the allocation strategy of SFLA. Calculate the current optimal solution X_{best} and the current worst solution X_{worst} of each subgroup as well as the group's optimal solution in line with the fitness value X_{gbest};
Step 4: Use Equations (8) and (6) to update V_j and X_j;
Step 5: If *runsubIter<Granularity*, return to Step 4; otherwise, go to Step 6;
Step 6: Update the current worst solution X_{worst}; if successful, switch to Step 8; if not, switch to Step 7;
Step 7: Update the current worst solution X_{worst};
Step 8: Update the fitness value $f(x_j)$; if *runIter<Iter*, return to Step 3; otherwise, go to Step 9;
Step 9: Find the optimal fitness value X_{gbest}, which represents the fuzzy matrix set of the optimal cluster centres.

4.4 SELF-ADAPT ADJUSTMENT FOR SEARCHING SPACE BASED ON TRUST REGION

The trust region method is another important way to solve the non-restriction optimization, different with the linear searching. The most obvious advantage is the global convergence.

The trust region thought is define the model function $\max m_k(x)$ in the first on every iteration point x^k, and desire it could well approximate the object function f in the x^k's suitable neighbourhood. The neighbourhood is called Trust Region, TR, which defined as:

$$S_k = \left\{ x \in \Re^n \left\| \|x - x^k\|_k \leq \Delta_k \right. \right\},\qquad (12)$$

where Δ_k is the trust region radius, $\|\bullet\|$ is a norm rely on the iteration. Based on this trust region, seek for the optimal trail step s^k for $m_k(x)$ which satisfies $x^k + s^k$, and make $m_k(\bullet)$ decrease and $\|s^k\| \leq \Delta_k$. If $x^k + s^k$ make f decline enough, it will be accepted as the next iteration point which increases Δ_k, or keeps Δ_k unchanged; otherwise, make x^{k+1} as x^k, decrease Δ_k and shrink the trust region to a better similarity for the model function and object function on the new trust region. Part 1 algorithm flow is:

Step 1: Set the initial point $x^1 \in \Re^n$, trust region radius Δ_1, the constant η_1, η_2, r_1 and r_2, which satisfied $0 < \eta_1 \leq \eta_1 < 1, 0 < r_1 \leq r_2 < 1$, and make $k = 1$;

Step 2: Select the norm $\|\bullet\|_k$ and define the model $m_k(x)$ on the $S_k = \left\{ x \in \Re^n \left\| \|x - x^k\|_k \leq \Delta_k \right. \right\}$;

Step 3: Calculate s^k to make m^k declined enough and $x^k + s^k \in S_k$;

Step 4: $p_k \dfrac{f(x^k) - f(x^k + s^k)}{m_k(x^k) - m_k(x^k + s^k)}$, if $p_k \geq \eta_1$, set $x^{k+1} = x^k + s^k$; otherwise $x^{k+1} = x^k$;

Step 5: Make $\Delta_k \in \begin{cases} [\Delta_k, \infty] & p_k \geq \eta_2 \\ [r_2\Delta_k, \Delta_k] & p_k \in [\eta_1, \eta_2) \\ [r_1\Delta_k, r_2\Delta_k] & p_k < \eta_1 \end{cases}$

4.5 THE DYNAMIC NEIGHBORHOOD MULTI-SPECIES PARTICLE SWARM PARAMETERS OPTIMIZATION

Bring in the dynamic neighbourhood strategy into the particle swarm algorithms and divide the total swarm into several sub-species with self-adaptation, increase the information share between particles in the neighbourhood topology. First, we define the neighbourhood space distance R_1 as follows:

$$R_1 = \max\left(p_{bestjd}(k), x_{id}(k)\right) - \min\left(p_{bestjd}(k), x_{id}(k)\right),\qquad (13)$$

$p_{bestj}(k)$ is the *d-th* dimension of the *j-th* particle's best solution in the neighbourhood, x_{id} is the *d-th* dimension's position of the *i-th* particle. The Equation (13) could be divided into 3 situations:

Situation 1: If $x_i < \min(p_{bestj})$, that is the particle position is less than the neighbourhood particle historical optimal value, shown as Figure 2a. There into the circle mark is the current particle position, the pentagon is the swarm optimize position in the neighbourhood. The rhombus is the distance from unit historical optimal value P_{besti} to every optimal value in each neighbourhood, then

$$R_1 = \max(L_i) = L_s = \max\left(p_{bestj}\right) - x_i;$$

Situation 2: If $x_i > \min(p_{bestj})$ and $x_i < \max(p_{bestj})$, that is the particle is among all the optimal value, shown as the Figure 2b, then $R_1 = L_1 + L_s = \max\left(p_{bestj}\right) - \min\left(p_{besti}\right)$;

Situation 3: If the particle position is in the right of every optimal value, shown as the Figure 2c, then $R_1 = \max(L_i) = L_1 = x_i - \min\left(p_{bestj}\right)$, thus we can conclude Equation (13).

a) first situation

b) second situation
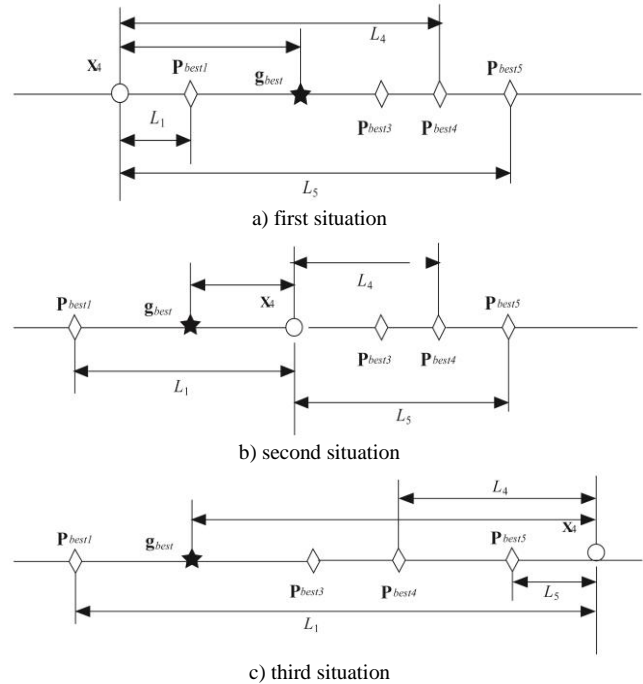
c) third situation
FIGURE 2 Diversity analysis diagram

Define the diversity of the potential searching space of the global particle swarm algorithms, the *j-th* neighbourhood, as:

$$diversity_j = \frac{1}{SL}\sum_{i=1}^{s}|R_1|,\qquad (14)$$

where S stands for the particle number in the neighbourhood, L stands for the linking length. There are three neighbourhood creation methods.

In the first method, the nearest neighbourhood in the searching space $N_i(k)$ is:

$$N_i(k) = \left\{ j, j \neq i \big\| \| x_i(k) - x_j(k) \| \leq \zeta_i \right\}. \qquad (15)$$

Make a judgment rely on the distance between different particle space positions $x_i(k)$. When it's less than the threshold value ζ_i, then create the topological relation between the two neighbourhoods;

The second method is nearest neighbourhood of the adaptive value function space:

$$N_i(k) = \left\{ j, j \neq i \big\| \| f(x_i(k)) - f(x_j(k)) \| \leq \zeta_i \right\}. \qquad (16)$$

Make the judgment rely on the adaptive function distance and divide the particles with different degree of evolution into different neighbourhood;

The third method is random neighbourhood topology:

$$N_i(k) = \left\{ j, j \neq i \big| \varepsilon_{ij} \leq \varepsilon \right\}. \qquad (17)$$

The global Particle Swarm Optimization with Dynamic Neighbourhood Topology, PSO-DNT, could be achieved with follow steps, shown as Figure 3.
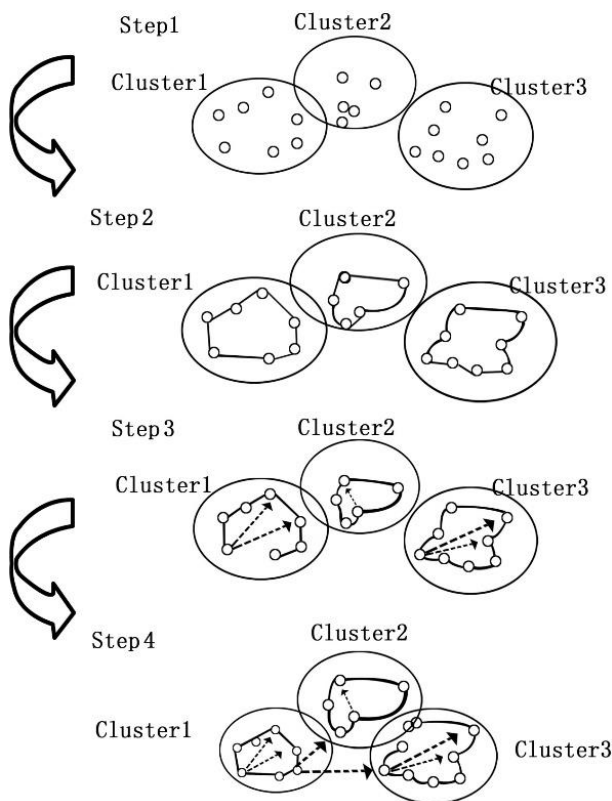
FIGURE 3 Dynamic neighbourhood topology schematic diagram

First divide the overall particle swarm into C sub-species with LA-FCM algorithms, as Step 1. It's noticed that the fussy clustering is aimed at particles, while Part 2 is aimed at all dimensions. The process objects are different. Then create circle topology structure and inner

random topology structure in step 2 and step 3 to guarantee the complete validity of the neighbourhood optimization information. Third, have the regular algorithms optimization, when the particle searching space decrease gradually, that is $diversity < d_{\min}$, create the external random topology structure between species to ensure the information sharing.

Thus, the speed evolution formula based on the neighbourhood topology is:

$$v_{id}(k+1) = w v_{id} + c_1 rand_{id}(\bullet) \left[p_{id}(k) - x_{id}(k) \right] + $$
$$c_2 rand_{id}(\bullet) \left[l_{gd}(k) - x_{id}(k) \right], \qquad (18)$$

where $l_{gd}(k)$ is the current position of the optimal particle in the neighbourhood. The method mainly modified the third part of the evolution formula. It abandons learning the swarm optimal value of the total algorithms and pays more attention to the mutual learning between neighbourhood particles.

5 The simulation analysis of large-scale parameter optimization

We selected 20 standard functions for testing. The large-scale optimization functions are two types: separable functions and inseparable function. When

$$\arg \min_{(x_1,...,x_n)} f(x_1,...,x_n) = $$
$$\left(\arg \min_{x_1} f(x_1,...), ..., \arg \min_{x_n} f(...,x_n) \right). \qquad (19)$$

Every independent variable $x_1,...,x_n$ is independent with each other, then its separable function and easily solved. When there are m variables to the most are dependent, its call m-inseparable function. When every random two variables are dependent with each other, it is called inseparable function.

The 20 standard functions are divided four class:

1 separable functions.

2 partially inseparable function, in which there are fewer parts are relevant.

3 partially inseparable function, in which there are large parts are relevant.

4 completely inseparable function.

All the experiments repeat for 20 times. The testing initial conditions: Testing Dimensions=1000, Cluster Number=10, Iteration Times=3000000, Sub-species Iteration Times=2000, Particle Dimensions=100, Acceleration Coefficient $c_1 = 1.49, c_2 = 1.49$, Inertia Weight $\omega = 0.8$. The convergence curve of mentioned SP-FCM is shown in Figure 4.

According to the Figure 4, the mentioned SP-FCM algorithms achieved better results. It is also much closed to the best result with stronger stability. In a word, the mentioned SP-FCM algorithm gets the most optimal value of the 20 testing functions and has a better optimization result.

(a) Functions 1-5

(b) Functions 6-10

(c) Functions 11-15

(d) Functions 15-20
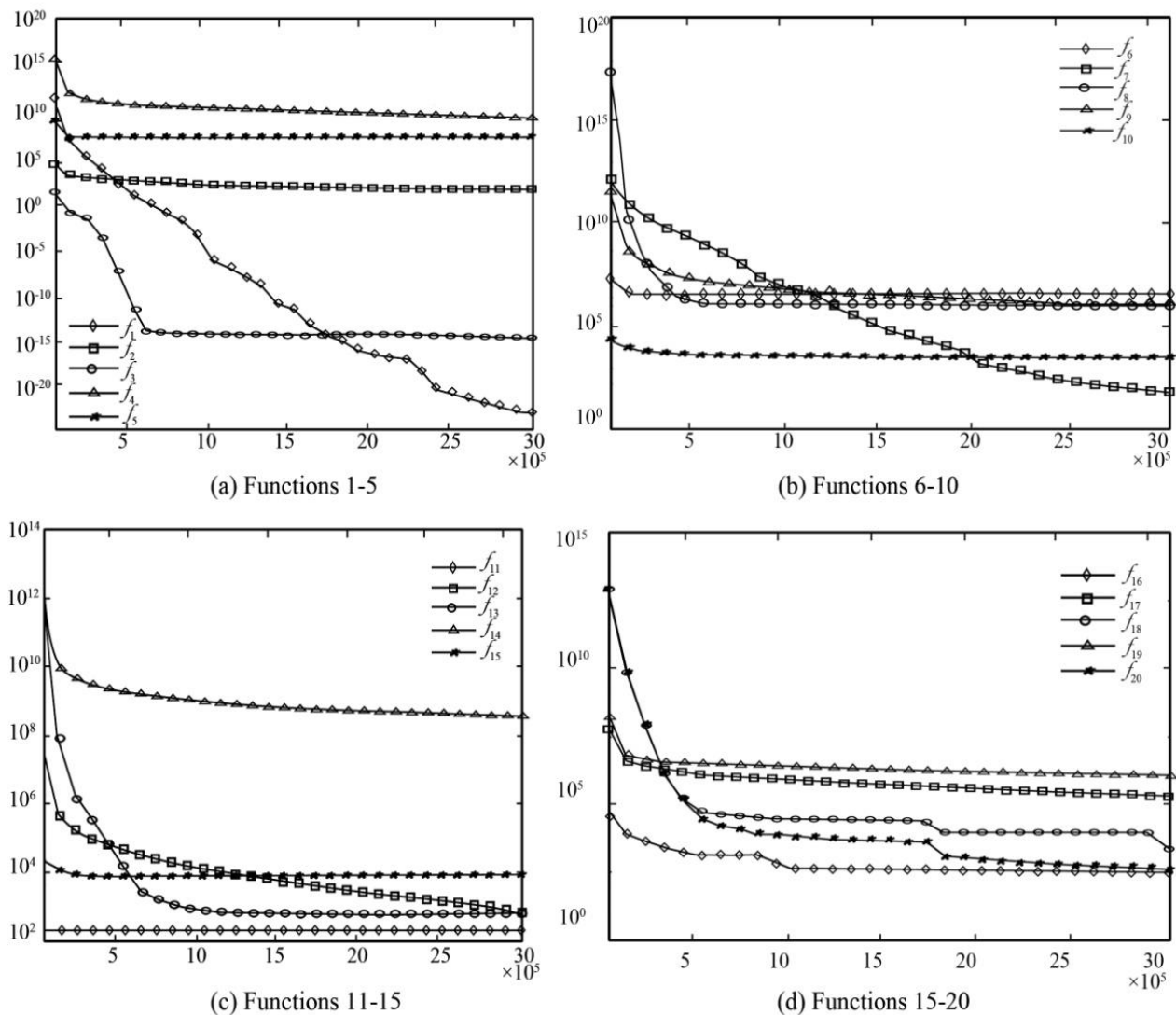
FIGURE 4 Convergence curves of the proposed SP-FCM algorithm

6 Conclusions

In actual industrial process, the control system is much more complicated, the parameters to optimize is more and more, in which case the normal optimization often lose the validity. This article came up with the particle swarm large-scale parameters optimization based on the Fussy C-Means Clustering to solve the curse of dimensions. But the Fuzzy C-Means clustering algorithm still has the sensitivity to the initial points. In this case, this article firstly put forward the Fussy C-Means two-phase clustering algorithms based on linear distribution to get better initial points by means of linear distribution strategy, and then substituted them into the global circulation. It has discovered that the mentioned algorithm had a better clustering result by simulation analysis. And then, this article puts forward a new SP-FCM algorithm on the basis of current literature.

Experiments show the algorithm can effectively increase the searching ability and clustering results of fuzzy clustering algorithms. It combines SFLA and PSO by setting a searching granularity coefficient. Thus, it can find the optimal initial cluster centre and avoid the disadvantages of FCM by dint of the advantages of the two. After that, it adopts the basic FCM algorithm, and it achieved the data property weight and initial cluster centre to guide the overall cluster progress, which efficiently improve the calculate speed and cluster precision.

Acknowledgments

References

[1] Kennedy J, Eberhart R C, Shi Y H 2009 Swarm Intelligence *Singapore: Elsevier*

[2] Yousfi C, Tournier R 2011 Steady state optimization inside model predictive control *Proceedings of American Control Conference* Boston USA 2011 1866-70

[3] Scokaert P O M, Mayne D Q 2009 *IEEE Transactions on Automatic Control* **43**(8) 1136-42

[4] Bemporad A, Borrelli F, Morari M 2012 *IEEE Transactions on Automatic Control* **47**(12) 1974-85

[5] Tang P L, de Silva C W 2012 *IEEE Transactions on Control Systems Technology* **14**(4) 707-18

[6] Zhang T J, Feng G, Lu J H 2011 *IEEE Transactions on Fuzzy Systems* **15**(4) 686-98

[7] Lu C H, Tsai C C 2010 *IEEE Transactions on Industrial Electronics* 2010 **55**(3) 1366-75

[8] Munoz de la Pena D, Christofides P D 2010 *IEEE Transactions on Automatic Control* **53**(9) 2076-89

[9] Torrico B C, Roca L, Normey-Rico J E, Guzman J L 2011 *IEEE Transactions on Control Systems Technology* **18**(6) 1430-9

[10] Qin S J, Badgwell T A 2009 A survey of industrial model predictive control technology *Control Engineering Practice* **11**(7) 733-64

[11] [Grosdidier P, Froisy J B, Hamann M D 2008 The IDCOM-M controller *Proceedings of IFAC Workrhop on Model Based Process* Atlanta USA 31-6

[12] Scokaert P O M, Mayne D Q 2008 *IEEE Transactions on Automatic Control* **43**(8) 1136-42

Authors

Xuhui Zhang, born in October, 1972, Shaanxi province, China

Current position, grades: professor and dean of Mechanical Engineering, Xi'an University of Science and Technology, China.
University studies: PhD in Instrument Science and Technology (2009) at Xi'an Jiaotong University.
Scientific interest: Intelligent detection and control, non-destructive testing and evaluation, condition monitoring and fault pattern recognition.
Publications: 70 papers, 5 books.
Experience: more than 20 years.

Analytical solution on the sensitivity of matric suction profile in soil layer which is under the condition of one-dimensional steady flow

Jian Li[1*], Xiyong Wu[1], Long Hou[2]

[1]*Department of Geosciences and Environmental Engineering, The Southwest Jiaotong University at Chengdu, Chengdu, Sichuan 610031, China*

[2]*Wanzhou District Commission of Urban-Rural Development in Chongqing Wanzhou, Chongqing 404000, China*

Abstract

To simulate the suction profile within unsaturated soil layer, which is under the condition of one-dimensional infiltration, the point that water potential energy is mainly composed of gravitational potential energy and suction potential energy was used. And this paper also takes the standpoint that the relationship between these two variables is reciprocal. Combined with the mass conservation law, *Darcy*'s law and *Gardner* empirical equation, an analytical solution was got. This solution could be used to describe the suction profile within unsaturated soil layer when the seepage field of this soil layer reaches steady stage and the rainfall infiltrates to the soil layer along with the vertical direction. Several conditions with different rainfall intensities q, different inverse values of air-entry pressure α and different ratio values of q/k_s were examined by using this analytical solution, respectively. Compared with other factors, the rainfall intensity will take more influence on the suction profile, namely, the suction profile within unsaturated soil layer is more sensitive to the change of rainfall. The results obtained in this paper could be used as one useful reference for the research work about rainfall-induce landslide and the corresponding computer simulation.

Keywords: analytical solution, unsaturated soil, matric suction, sensitivity, steady seepage

1 Introduction

The seepage field profile in one slope could be used as an important reference for the stability evaluation of slope. If the "hysteresis effect" could be ignored, thus the relationship between seepage field and suction stress field could be established by the SWCC (soil-water characteristic curve) [1]. Due to the suction is a vital factor in the growth of shear strength, the field profile of it has strong relation with the development of slope stability [2-5].

As for the soil, slope suction field profile within shallow soil layer will influence by the change of some factors such as rainfall intensity, soil porosity and saturated hydraulic conductivity. This phenomenon also acquires some researcher's attention; Li presented one type of variation of suction profile by taking the point of soil pore structure [6]. Chen completed some research on the relationship between total suction and soil saturation by considering the effect of infiltrate [7]. By using simulation method, Zhu obtained the stability development for one soil slope which under different rainfall durations [8]. But most research were conducted under the change of only one factor, and comparative study on the suction field within unsaturated soil and under the influence of factor variation are rare. For above reasons this paper will discuss the suction field profile under the condition of different rainfall intensities q, inverse value of air-entry pressure α and q/k_s (rainfall intensity/saturated hydraulic conductivity). Based on obtained results, the sensitivity of suction field to different influence factors also experienced.

2 Common suction field profile within unsaturated soil layer

Following the change of influence factor or soil property, the seepage field (i.e., water content field) will also express variation, and the field re-adjust is a typical dynamic process, which will convert into stable phase at the final, namely, the steady flow stage that the field profile is regardless of the time. Correspondingly, the suction field could be divided into two different forms, transient state and steady state [9, 10].

The steady suction field keeps at a constant state, which does not product any change as time elapse. And the profile of it will be controlled by the boundary condition of soil surface and the initial condition of the whole soil layer. In order to get the general features this paper takes study on the steady suction field within only one soil layer, which is under the condition of one-dimensional steady flow.

---
[*]***Corresponding author*** e-mail: jlicivil@gmail.com

3 Analytical analysis on the steady suction field

3.1 ASSUMPTIONS USED IN THIS STUDY

As for the soil layer which is in the stage of one-dimensional steady flow, the water flow in it could consider as along the vertical direction only, and there are two types of boundary conditions will be used in this section: rainfall q or evaporation presents the upper boundary condition (only rainfall considered in this paper) and the water table refers to the lower boundary condition. The water content profile or suction field within one soil layer with no upper boundary condition and steady water table will be used as initial condition for analytical analysis.

3.2 ANALYTICAL SOLUTION ON SUCTION FIELD BY CONSIDERING ENERGY POTENTIAL

According to the continuity equation of water flow, the water content of any soil unit within one soil layer keeps at a constant value, namely, the mass "break-even" in soil unit, which can be expressed as follow:

$$\frac{dq}{dz} = \frac{d(\theta)}{dt} = 0 , \qquad (1)$$

where z is the flow direction, θ is water content of soil and t presents flow time.

Commonly, the solution potential, velocity potential and power potential could be ignored due to they all take very small influence on slope stability. The total potential of soil pore water at any location is composed by gravity potential and suction potential, which can be expressed as:

$$E_t = E_g - E_m = m_w g z - V_w (u_a - u_w) , \qquad (2)$$

where E_t, E_g and E_m refers to total potential, location potential and suction potential, respectively. The negative sign in this equation presents these two potential are two shift terms when they contribute to the total potential. As the rise process of capillary water, that is, the suction potential is getting smaller and smaller with water location potential increase gradually, and the water always flow to the location with low potential.

If use h to present the total water head, the above equation can be rewritten to:

$$h = z - \frac{(u_a - u_w)}{\rho_w g} , \qquad (3)$$

where ρ_w is water density, $(u_a - u_w) = s$ refers to matric suction. Combined the *Gardner* empirical equation $k(s) = k_s e^{[-\alpha(u_a - u_w)]}$, *Darcy* unsaturated flow equation $q = -k(s)\frac{dh}{dz}$ and Equation (3), we can obtain a new equation as follow:

$$-\frac{k_s e^{-\alpha(u_a - u_w)}}{\rho_w g} \cdot \frac{d^2(u_a - u_w)}{dz^2} +$$

$$\frac{\alpha k_s e^{-\alpha(u_a - u_w)}}{\rho_w g} \cdot \left(\frac{d(u_a - u_w)}{dz}\right)^2 - , \qquad (4)$$

$$\alpha k_s e^{-\alpha(u_a - u_w)} \frac{d(u_a - u_w)}{dz} = 0$$

where α is the inverse value of air-entry pressure, k_s is a constant value which presents the saturated hydraulic conductivity, the above equation can be simplified as:

$$\frac{d^2(u_a - u_w)}{dz^2} - \alpha\left(\frac{d(u_a - u_w)}{dz}\right)^2 +$$

$$\alpha\rho_w g \frac{d(u_a - u_w)}{dz} = 0 . \qquad (5)$$

Namely:

$$\frac{d^2(u_a - u_w)}{dz^2} + \alpha \frac{d(u_a - u_w)}{dz}\left(\rho_w g - \frac{d(u_a - u_w)}{dz}\right) = 0 . \qquad (6)$$

Let $f = \frac{d(u_a - u_w)}{dz}$ by reduced order processing, then $\frac{d^2(u_a - u_w)}{dz^2} = \frac{df}{dz} = \frac{df}{d(u_a - u_w)} \cdot f$, after taking this new term into Equation (6), we can get a new equation:

$$f\frac{df}{d(u_a - u_w)} + \alpha f(\rho_w g - f) = 0 . \qquad (7)$$

Namely:

$$\frac{df}{d(u_a - u_w)} - \alpha f + \alpha\rho_w g = 0 . \qquad (8)$$

Solving Equation (8), an expression which describes the first derivative of suction within soil layer to the soil depth is as follow:

$$f = \frac{d(u_a - u_w)}{dz} = Ce^{\alpha(u_a - u_w)} + \rho_w g , \qquad (9)$$

where C is the undetermined coefficient and can be ensured by three factors: soil pore type, air-entry pressure and pore water retain type.

Combined with $q = -k(s)\frac{dh}{dz}$, $k(s) = k_s e^{[-\alpha(u_a - u_w)]}$ and $h = z - \frac{(u_a - u_w)}{\rho_w g}$, we can obtain the following:

$$q = -k_s e^{[-\alpha(u_a - u_w)]}\left(1 - \frac{1}{\rho_w g}\frac{d(u_a - u_w)}{dz}\right) =$$

$$-k_s e^{[-\alpha(u_a - u_w)]}\left(1 - \frac{1}{\rho_w g}(Ce^{\alpha(u_a - u_w)} + \rho_w g)\right) . \qquad (10)$$

Accordingly, when the flow amount in the soil surface, the matric suction, the soil density and the air-entry pressure are obtained, we can determine the undetermined coefficient by using the above analytical solution.

Some experimental results show us that the suction profile can be described by one smooth monotonic function curve if there is no mutation in soil properties of the whole soil layer. According to the method provide in literature [11], we can use obtained C value to determine the suction of soil at any location, furthermore, the related profile of suction field within one soil layer which is under different conditions also can be got.

4 Sensitivity analysis on the matric suction profile to influence factor

4.1 MAIN INFLUENCE FACTORS

There are many factors can take influence on the final profile of suction field within unsaturated soil layer and under steady flow state. Overall, more important factors include the boundary condition (rainfall or evaporation), soil property and the related hydraulic features. In order to experience the sensitivity of suction field to the above issues, the study with three different conditions will conduct in this section, that are:

1) Taking different rainfall intensity values but the remaining parameters at constant ($\alpha = 0.3$/kPa), namely, $q = 0$ m/s, -2.0×10^{-8} m/s, -4.0×10^{-8} m/s and -7.2×10^{-8} m/s (negative sign on behalf of rainfall infiltration);

2) The rainfall intensity and other parameters keep constant but the inverse value of air-entry pressure is different, namely, $\alpha = 1.0$/kPa, 0.5/kPa, 0.3/kPa and 0.1/kPa;

3) The inverse value of air-entry pressure $\alpha = 0.3$/kPa, but the ratio of rainfall intensity to saturated hydraulic conductivity is different: $q/k_s = -0.1$, -0.4, -1.0 and -1.3.

The other calculation parameters are: the homogeneous isotropic soil layer is composed with unsaturated soil and the thickness of it is 2.5m. The elevation of steady water table is 0m and the saturated hydraulic conductivity is $k_s = 7.2 \times 10^{-8}$ m/s. Whole soil layer can be divided into 50 sub-layer by 0.05m thickness.

4.2 SENSITIVITY ANALYSIS ON THE MATRIC SUCTION PROFILE

The obtained calculation results are summarized in Figures 1 to 3 and can be used to describe the suction field profile within soil layer under different conditions. From Figures 1 to 3 we can know that rainfall intensity take greatest influence on the suction field profile of soil layer, and the variation of suction value is most evident compared with other influence factors, namely, the suction field profile will be more sensitive to the rainfall intensity value.
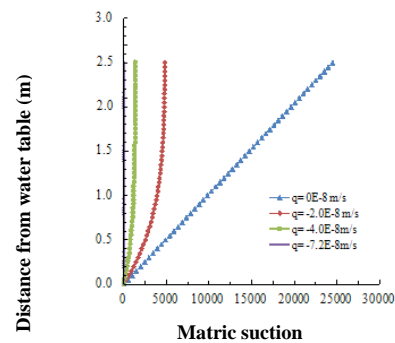
FIGURE 1 The suction profile within soil layer which is under the conditions with different rainfall intensities q

As show in Figure 1, even though at the same depth, the suction value of soil layer with no rainfall infiltration is largest, and it has linear proportional relation with the distance from water table. The suction field corresponding to this state can be considered as the initial state before rainfall starts. As the rainfall intensity increases and the seepage reaches to steady state, the suction value at the same location will be smaller and smaller, the suction field function curve will gradually move to left as well. If the rainfall intensity equal to the saturated hydraulic conductivity at the value (7.2×10^{-8} m/s), the whole soil layer will reach its saturated state when the seepage is steady, thus all the suction at any location within this soil layer are disappeared.
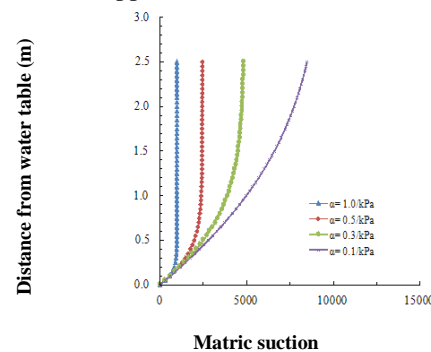
FIGURE 2 The suction profile within soil layer which is under the conditions with different inverse values of air-entry pressure
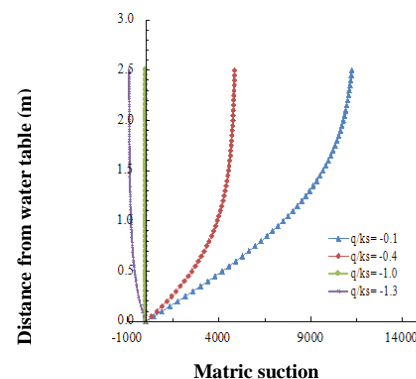
FIGURE 3 The suction profile within soil layer which is under the conditions with different ratio values of q/k_s

Compared to rainfall intensity, the change of inverse value of air-entry pressure only takes smaller influence on the suction field variation. Study on the Figure 2 we can find that even though at the same depth, the suction value related to soil with small air-entry pressure (i.e., Sandy soil) is always smaller than that of soil with large air-entry pressure (i.e., Clay soil).

Figure 3 presents the suction field profile with different ratios of rainfall intensity to saturated hydraulic conductivity. The ratio value less than 1 refers to the rainfall intensity is smaller than soil's capacity of water hold, thus all the rainfall will be accepted by the soil layer. When the seepage reaches steady state, the soil above water table also keep at saturated condition. And for the same depth, the suction value related to high ratio value (q/k_s) is smaller than other. The $q/k_s=1$ means that rainfall intensity is equal to soil's water hold capacity, and at this point the whole soil layer will just reach its saturated condition, correspondingly, all the suction value within it are disappeared. The soil layer cannot accept all the rainfall if the $q/k_s>1$, and water will occur at the surface of it. As show in Figure 3 the suction value with different depth is different for each other but all of them are less than zero.

5 Conclusions

From the perspective of potential theory, this paper considers the gravity potential and suction potential as the two major terms of pore water potential of unsaturated soil. Combined with mass conservation law, *Darcy*'s law and the *Gardner* empirical equation, an analytical solution that can be used to describe the suction field profile within unsaturated soil layer which is under one-dimensional steady flow is obtained.

With three different conditions, this paper gets the corresponding suction field profiles related to several influence factors by using the obtained. And the study on the results shows us that the suction profile within unsaturated soil layer is more sensitive to the change of rainfall than other influence factors.

Acknowledgments

References

[1] Liang T Z, NG Charles 2007 Effect of suction on shear strength and dilatancy of an unsaturated expansive clay *Chinese Journal of Geotechnical Engineering* **29**(1) 82−7 *(in Chinese)*

[2] Rui G W, Shu W Y, Wei D D 2004 Analysis of seepage stability of high-filled embankment slope due to rainfall infiltration *China Journal of Highway and Transport* **17**(4) 25−30 *(in Chinese)*

[3] H. Rahardjo, T.T. Lee, E.C. Leong, R.B. Rezaur 2005 Response of a residual soil slope to rainfall *Canadian Geotechnical Journal* **42** 340−51

[4] Dong Y L, Long H, Chuan S W 2012 Numerical analysis for unsaturated soil slope which locates on reservoir region and under the condition of water level fluctuation *Journal of Information and Computational Science* **9**(10) 2719−29

[5] Chun N S, Xiang W F, He W W 2012 Research on effects of suction, water content and dry density on shear strength of remolded unsaturated soils *Rock and Soil Mechanics* **30**(5) 1347−51 *(in Chinese)*

[6] Zhi Q L, Rui L H, Li C W 2006 Study on SWCC of unsaturated expansive soil *Rock and Soil Mechanics* **27**(5) 730−4 *(in Chinese)*

[7] Xiao B C, Wang G X, Xiao P L 2011 Effect of rainfall infiltration on additional settlement and stability debasement of granular soil fillings embankment *Journal of Central South University (Science and Technology)* **42**(3) 765−71 *(in Chinese)*

[8] Lei Z, Bao N H 2009 Stability analysis of coal measure soil slope under rainfall infiltration *Rock and Soil Mechanics* **30**(4) 1035−40 *(in Chinese)*

[9] Li Z W, Run Q H 2011 Analytical analysis of coupled seepage in unsaturated soils considering varying surface flux *Chinese Journal of Geotechnical Engineering* **33**(9) 1370−5 *(in Chinese)*

[10] Gardner, W. R 1956 Calculation of capillary conductivity from pressure plate outflow data *Soil Science Society of America Proceedings* **20** 317−20

[11] Shun Q L, Shao F Z, Xue Z W 2006 Matric suction of unsaturated soils in one-dimensional steady flow *Rock and Soil Mechanics* **27**(s) 90−4 *(in Chinese)*

Authors

Jian Li, born on September 23, 1974, Sichuan province, China

Current position, grades: Ph.D. Candidate, Department of Geosciences and Environmental Engineering, Southwest Jiaotong University.
University studies: Master degree in Guizhou University in 2008.
Scientific interest: computer simulation and software development in the field of geological hazard.
Publications: 5 papers, 1 patent.

Xiyong Wu, born on May 14, 1963, Sichuan province, China

Current position, grades: professor of Department of Geosciences and Environmental Engineering, Southwest Jiaotong University. **University studies**: Ph.D. degree in Southwest Jiaotong University in 1998.
Scientific interest: geological hazard and its control engineering, rock weathering and weakening evaluation of engineering characteristics.
Publications: 40 papers, 3 patents.

Long Hou, born on June 25, 1982, Henan province, China

Current position, grades: engineer in the field of unsaturated soil mechanics.
University studies: Ph.D. degree in Chongqing University in 2012.
Scientific interest: computer simulation of unsaturated soil mechanics.
Publications: 10 papers, 2 patents.

Improved particle swarm optimization algorithm with unidimensional search

Pu Han, Li Meng[*], Biao Wang, Dongfeng Wang

School of Control and Computer Engineering, North China Electric Power University, Baoding 071003, Hebei, China

Received 1 March 2014, www.cmnt.lv

Abstract

In this paper, a strategy of unidimensional search is introduced to particle swarm optimization (PSO). The global exploration capability of PSO is used to identify a promising region in search space. With the region as the starting point, a unidimensional local search is applied to search a more accuracy solution. The local search does not rely on the population information, which makes it can jump out of a local optimum when the population stagnates. With combination of global exploration and local exploitation, the algorithm can discover more favourable search area effectively and obtain a better solution. The improved PSO method is tested on eight benchmark functions. Experimental results show that the method can not only improve the accuracy of solution, but also reduce the influence of initial population distribution upon the algorithm performance. Finally, the influence of parameter variation on algorithm is analysed.

Keywords: particle swarm optimization, unidimensional local search, population initialization, population distribution

1 Introduction

Particle swarm optimization (PSO), which was proposed by Kennedy and Eberhart [1] in 1995, is one of the most effective intelligent optimization algorithms developed recent years. Although has some common with evolutionary algorithm, PSO does not use evolution operators such as crossover and mutation. It emulates the flocking behaviour of birds and fish. Each particle adjusts its position according to the success of itself and its neighbourhood. As PSO is simple and easy to implement, it has already been applied in many real-world problems.

However, PSO suffers from trapping in local minima point easily and slow convergence speed. Many researchers have worked on solving these problems in various ways. Shi and Eberhart [2] propose a linearly decreasing inertia weight in evolution course. In [3], a new variant of PSO with nonlinear inertia weight which relative to generation number is proposed. Parameters are adjusted automatically according to the population distribution information in [4]. In [5] orthogonal experiment design method is used to construct a better guidance exemplar by using information lies in a particle's best historical position and its neighbourhood's best position. In [6, 7] particles have several updating strategies. A choice mechanism related to the population information is used to steer different strategies in an adaptive and parallel way. To remain the population diversity, [8] proposes comprehensive learning particle swarm optimizer, which enables each dimension of a particle, has the opportunity to learn from a different exemplar. In [9] information drawn from the distances between the global best particle and every other particle is applied to updating particle

velocity. In [10] neighbourhood structure is changed dynamically during evolution. The topology structure in [11] is tree-like in which particles with better evaluations are placed in the upper nodes of the tree. Particles' positions in the tree are adjusted at each iteration. [12] uses global exploration capability of PSO algorithm to locate a good local minimum approximately, then quasi Newton-Raphson is done with the best solution as its starting point for accurate local exploration. [13] proposes a hybrid swarm optimizer which combines PSO with LM (Levenberg-Marquardt) algorithm. In [14] six discrete crossover operators are incorporated respectively into a global best particle swarm optimizer.

Although the existing variants of PSO have some exciting results, there still some effort to do. When the search space is high-dimensional, there exists such a phenomenon that some dimensions develop toward global optima, while some deviate from it. If the search can detail to each dimension, there will be a considerable improvement on algorithm performance. A new learning strategy is introduced in [15]. The whole position vector is divided into several subvectors. The subvectors are updated iteratively in the updating process of a position vector. In this paper this strategy is integrated into PSO for a more accurate single dimensional search based on the potential solution located by PSO. The search based on a dimension by dimension way can protect the dimensions which are in good positions and mitigate against premature convergence in a single dimension. Then the more favourable result obtained by local search is fed back to population evolution as a guide.

---

[*]*Corresponding author* e-mail: mengli2014@163.com

2 Standard PSO (SPSO)

In SPSO each particle represents a potential solution in D dimensional search space. $x_i = (x_{i1}, x_{i2}, ..., x_{iD})$ is the position of particle i, $v_i = (v_{i1}, v_{i2}, ..., v_{iD})$ is the velocity. $p_i = (p_{i1}, p_{i2}, ..., p_{iD})$ is the best previous position of particle i, $p_g = (p_{g1}, p_{g2}, ..., p_{gD})$ is the best solution found by the whole population so far. The d th dimension of position and velocity of particle i are updated as Equations (1) and (2):

$$v_{id}^{k+1} = wv_{id}^k + c_1 r_1 (p_{id} - x_{id}^k) + c_2 r_2 (p_{gd} - x_{id}^k), \qquad (1)$$

$$x_{id}^{k+1} = x_{id}^k + v_{id}^{k+1}, \qquad (2)$$

where k is current iteration number, r_1 and r_2 are random number between $[0,1]$, keeping the diversity of population; c_1 and c_2 are acceleration constants; w is inertia weight which used to balance between the local and global search abilities.

3 Improved particle swarm optimization algorithm (IPSO)

In SPSO a particle updates each dimension of its position vector at one time, there is no guarantee that all dimensions fly to global optima. In our method, the search strategy borrowed from literature [15] is applied to single-dimensional local search around the promising region found by PSO. Then the more precise solution is employed to guiding population evolution.

3.1 THE IMPROVED ALGORITHM

Here, a new variable named p_g' is introduced. Meanwhile p_g which stands for the best solution found by population (as defined in section 2) is reserved. To distinguish, we call p_g' is the best solution found by the new algorithm, not by population. The initial value of p_g' is set as equal as p_g. When the population evolves a certain (M) iterations, a single dimension local search for p_g' is proceeded, updating each dimension of p_g' circularly. If a better solution is yielded, we update p_g' as the better solution. In the next M iterations, when p_g is superior to p_g', p_g' renewal as p_g. Otherwise p_g' remains unchanged until next local search. In our algorithm, global search and local search proceed repeatedly. In order to avoid converging to p_g' too quickly, p_g' and p_g are all acting on the updating of particle velocity, see Equation (3):

$$v_{id}^{k+1} = wv_{id}^k + c_1 r_1 (p_{id} - x_{id}^k) + c_2 r_2 (r_3 (p_{gd} - x_{id}^k) + (1-r_3)(p'_{gd} - x_{id}^k)), \qquad (3)$$

x_i, v_i, p_i, p_g, w, c_1, c_2, r_1 and r_2 in Equation (3) are the same as in SPSO. r_3 is a random number with uniformly distributed between $[0,1]$. The update of position vector is as Equation (2). Figure 1 is a flow chart of the IPSO, where minimum optimization problem is taken as a example. The process of local search can be seen from the chart. Also the distinction between p_g and p_g' is more easy to understand.
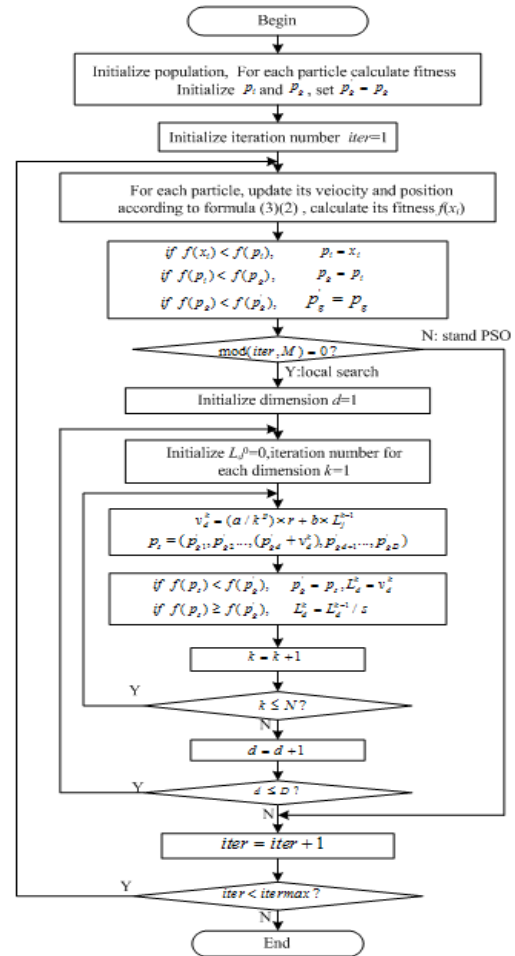
FIGURE 1 Flowchart of IPSO

In Figure 1, $mod()$ is remainder function. $itermax$ is maximum iterations of population. N is maximum iterations for single dimensional search. a, p, b, s and r are parameters for single dimensional search. Velocity vector consists of two parts [15]: diversity section $(a/k^p) \times r$ and the learning section $b \times L$. a is diversity factor, p is decreased factor, a and p are all positive constants. r is a random number with uniformly distributed between $[-0.5, 0.5]$. b and s are all constants greater than 1.

Han Pu, Meng Li, Wang Biao, Wang Dongfeng

3.2 COMPARED WITH SPSO

In our work eight benchmark functions are used to test the algorithm. The name, search space and search target accuracy of every function are shown in Table 1.

Generally it is difficult to determine the appropriate range of search space for real-life optimization problems. If the performance of the optimum algorithm has well relationship to the distribution of initial population, the results may not be satisfactory. Given this situation, the experiment is designed as follows: Assuming the search space range is $[-Space_{Max}, Space_{Max}]$, there are two methods to initialize the population: $P1$: the particles are uniformly distributed in $[0, Space_{Max}]$; $P2$: the particles are uniformly distributed in $[-Space_{Max}, Space_{Max}]$.

TABLE 1 Test function used in this paper

Fun No.	Fun Name	Expession	Search Space	Optimal Solution	Target Precision
f_1	sphere	$f_1(x) = \sum_{i=1}^{D} x_i^2$	$[-100,100]^D$	$f(0,0\ldots0)=0$	10^{-20}
f_2	rosenbrock	$f_2(x) = \sum_{i=1}^{D-1}(100(x_{i+1} - x_i^2)^2 + (x_i - 1)^2)$	$[-30,30]^D$	$f(1,1\ldots1)=0$	1
f_3	rastrigin	$f_3(x) = \sum_{i=1}^{D}(x_i^2 - 10\cos(2\pi x_i) + 10)$	$[-5.12,5.12]^D$	$f(0,0\ldots0)=0$	20
f_4	quadric	$f_4(x) = \sum_{i=1}^{D}(\sum_{j=1}^{i} x_j)^2$	$[-100,100]^D$	$f(0,0\ldots0)=0$	1
f_5	schwefel's P 2.22	$f_5(x) = \sum_{i=1}^{D}|x_i| + \prod_{i=1}^{d}|x_i|$	$[-10,10]^D$	$f(0,0\ldots0)=0$	10^{-2}
f_6	schwefel's P 2.21	$f_6(x) = \max_i\{|x_i|, 1 \le i \le D\}$	$[-100,100]^D$	$f(0,0\ldots0)=0$	10^{-2}
f_7	penalized	$f_7(x) = 0.1\{\sin^2(3\pi x_i) + \sum_{i=1}^{D-1}(x_i - 1)^2 \cdot$ $[1+\sin^2(3\pi x_{i+1})] + (x_D - 1)^2 \cdot$ $[1+\sin^2(2\pi x_D)]\} + \sum_{i=1}^{D} U(x_i,5,100,4)$ $U(x_i,a,k,m) = \begin{cases} k(x_i - a)^m, & x_i > a \\ 0, & -a \le x_i \le a \\ k(-x_i - a)^m, & x_i < -a \end{cases}$	$[-50,50]^D$	$f(-1,-1\ldots-1)=0$	10^{-5}
f_8	ackley	$f_8(x) = -20e^{-0.2\sqrt{\frac{1}{D}\sum_{i=1}^{D} x_i^2}} - e^{\frac{1}{D}\sum_{i=1}^{D}\cos(2\pi x_i)} + 20 + e$	$[-32,32]^D$	$f(0,0\ldots0)=0$	10^{-2}

All functions are set as 30 dimensions. Each algorithm runs 50 times independently. The algorithm parameters are set as: SPSO: population size is 20, the maximum iterations is 5000, $c_1 = c_2 = 1.49618$, $w = 0.72984$; IPSO: parameters for population are the same as SPSO, parameters for local search: diversity factor $a = 3$, decreased factor $p = 6$, $b = 2$, $s = 4$. For each algorithm, record mean value (Mean) and standard deviation (SD) of the solutions obtained in the 50 independent runs. Meanwhile record the success ratio (R) for target accuracy. The results are given in Table 2. IPSO achieves better results than SPSO on all test functions. IPSO shows a success rate of 100% except f_2 and f_6. SPSO only has a so perfect work on f_1 ($P2$), but a rather unsatisfactory work on f_3, f_4 ($P1$), f_5 (P1) and f_8. For the two initialization modes, IPSO has a fairly stable performance on all functions. While SPSO performs very differently on f_1, f_2, f_4 and f_7. On the whole, SPSO gives a poor performance when the initial population distributes as $P1$. The optimal solution of test functions used here are mostly distributed in the middle of the space search, particles are far from the optimal point when they initialized as $P1$. This may increase the difficulty to local the optimal region.

Figure 2 gives a more intuitive comparison between the two algorithms. As can be seen from Figure 2, IPSO converges faster than SPSO no matter how the distribution of the initial population. When initial population obtained by means of $P1$, SPSO falls into local optimum and stagnates too early on most of the tested functions except f_6. Figure 3 shows the results obtained in 50 runs, illustrating the solution distribution.

TABLE 2 Test functions and results

		sphere		rosenbrock		rastrigin		quadric	
		P1	*P2*	*P1*	*P2*	*P1*	*P2*	*P1*	*P2*
SPSO	**Mean**	2.72e+004	3.85e-043	8.85e+007	7.52e+003	2.93e+002	1.01e+002	1.24e+005	4.7e+003
	SD	1.23e+004	2.63e-042	6.68e+007	2.40e+004	5.41e+001	2.53e+001	9.81e+004	6.2e+003
	R(%)	4.0	100.0	2.0	18.0	0.0	0.0	0.0	44.0
IPSO	**Mean**	2.36e-063	1.43e-066	2.53e+000	2.19e+000	0.00e+000	0.00e+000	9.51e-013	3.66e-013
	SD	1.08e-062	5.80e-066	2.87e+000	3.05e+000	0.00e+000	0.00e+000	1.24e-012	4.41e-013
	R(%)	100.0	100.0	50.0	60.0	100.0	100.0	100.0	100.0
		schwefel's P 2.22		schwefel's P 2.21		penalized		ackley	
		P1	*P2*	*P1*	*P2*	*P1*	*P2*	*P1*	*P2*
SPSO	**Mean**	5.62e+001	4.17e+000	6.54e-002	2.66e-001	3.20e+008	2.26e-001	1.14e+001	4.16e+000
	SD	1.64e+001	7.56e+000	1.17e-001	4.95e-002	3.45e+008	5.72e-001	5.04e-002	2.88e+000
	R(%)	0.0	70.0	22.0	52.0	34.0	68.0	0.0	0.0
IPSO	**Mean**	1.08e-013	1.92e-013	2.43e-003	5.05e-004	2.07e-004	2.03e-004	3.57e-013	4.43e-013
	SD	5.22e-013	9.75e-013	2.91e-003	1.63e-003	1.95e-019	1.86e-019	1.51e-012	1.72e-012
	R(%)	100.0	100.0	98.0	98.0	100.0	100.0	100.0	100.0
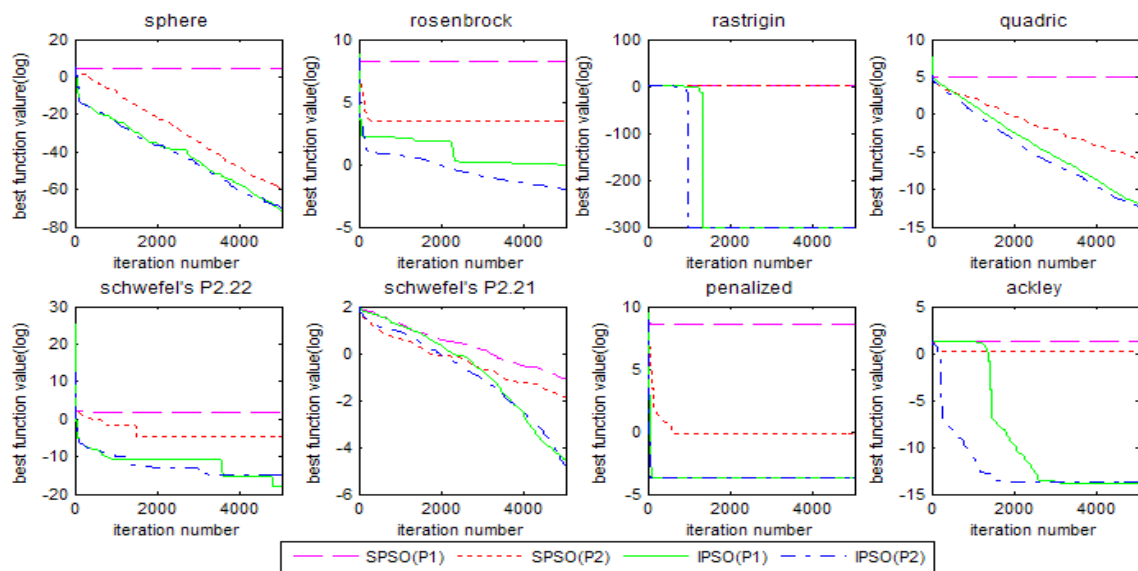
FIGURE 2 Best solution found by algorithm during evolution
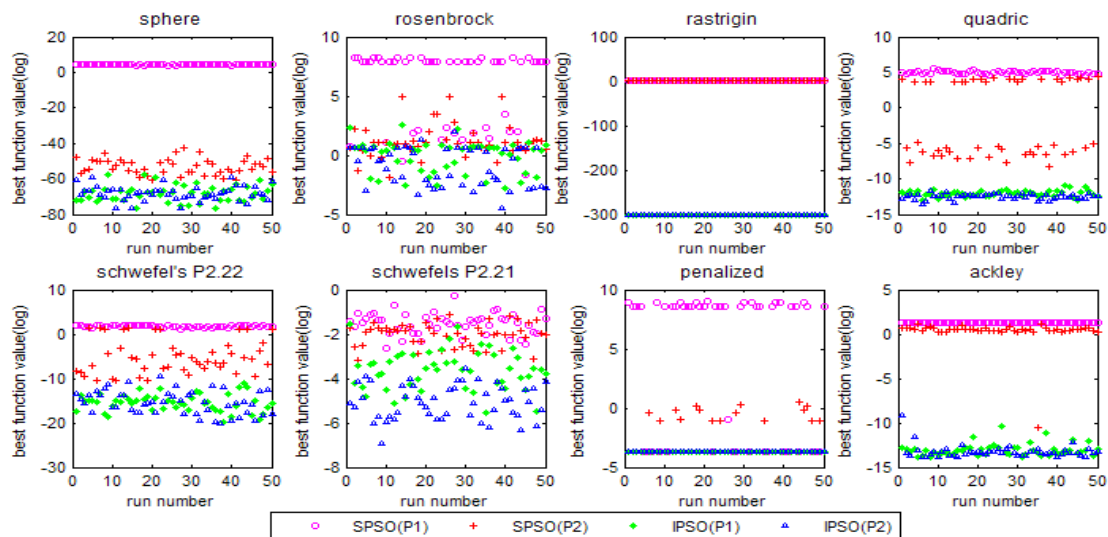
FIGURE 3 Solution distribution of 50 runs

Han Pu, Meng Li, Wang Biao, Wang Dongfeng

3.3 THE INFLUENCE OF PARAMETERS

Form section 3.2, we can see that with same control parameters, IPSO have good performances than SPSO on all tested functions. In this section, we study the performances of IPSO when inertia weight w and acceleration constants c_1, c_2 changes. Here, w varies from 0.6 to 0.8, with the variation amplitude of 0.02. Set

$c_1 = c_2 = c$, c varies from 1.0 to 2.0, with the variation amplitude of 0.1. The other parameters are set as the same as in section 3.2. The initial population distributes as $P2$, namely distributes uniformly within the whole search space. For each set of w and c, record the average value of results of 50 independent runs. Then a 3D graphics can be obtained, please see Figure 4.
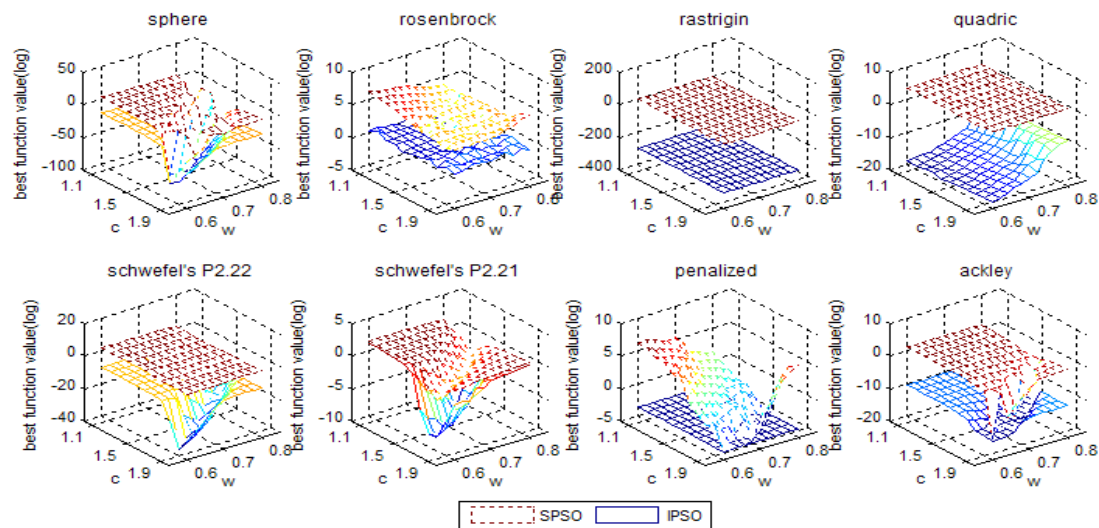
FIGURE 4 Influence of parameter changing

It can be seen from Figure 3, IPSO is always superior to SPSO when parameters vary. The affects of parameters variation on different functions are different. For f_1 and f_6, when c takes value of 1.0 to 1.2, the varying of w is basically no influences on the performances of the two algorithms; when c takes value of 1.3 to 2.0, the influences of w on the two algorithms are the same by the large: when c increases, w should decrease in order to get a better performance. For f_2 and f_8, the effect trends of parameters changing on the two algorithms are the same in the gross, except that SPSO has some mutation phenomenon, and IPSO is relatively stable. For f_7 SPSO is influenced by parameters greatly while IPSO is more stable. For f_5 SPSO is more stable when parameter varies, the change trend of performances of IPSO is the same as that for f_1. Moreover, when c is 1.8 and w is 0.6, IPSO achieve the best solutions on f_1 and f_5. For f_4, SPSO is stable; when c is fixed, the performance of IPSO steady decline with the increase of w, when w is fixed, the effect of c on the algorithm is relatively small. Generally, we can choose population parameters according to the principle which in SPSO.

4 Conclusions

When solving high-dimensional complex functions, SPSO is easy to prematurely fall into local optimum and

optimizing accuracy is not satisfactory. In addition to the loss of diversity, the updating mode of a particle is also responsible for the defects, in which all dimensions of the position are updating at one time. In this paper, on the basis of solution found by population global search, a single-dimension local search is applied for a more precise solution. Then this better solution is used to guide the population search. By iteratively repeated, the global and local search fully combined. Based on the experimental results, it can be concluded that IPSO is superior to the SPSO on convergence speed and solution precision on the chosen benchmark functions.

Another advantage of IPSO is that its performance is basically not affected by the mode of initial population distribution. This is splendid for practical problems. However, the performance of IPSO is influenced by the parameters, which is the same as SPSO. How to select appropriate parameters for different problems is also a problem worthy of studying.

Besides SPSO, the unidimensional search strategy can be combined with other PSO variants. Also the method for unidimensional updating can be replaced. And another more exciting algorithm may be obtained.

References

[1] Kennedy J, Eberhart R C 1995 Particle swarm optimization *Proc Conf on Neural Networks Piscataway USA* 1942-8

[2] Shi Y H, Eberhart R C 1998 A modified particle swarm optimizer *Proc. Conf. on Evolutionary Computation Anchorage AK* 69-73

[3] ChaReoee A, Siarry P 2006 Nonlinear inertia weight variation for dynamic adaptation in particle swarm optimization *Computers and Operations Research* **33**(3) 859- 71

[4] Zhan Z H, Zhang J, Li Y, Chung H S-H 2009 *IEEE Transactions On Systems Man and Cybernetics-part B: Cybernetics* **39**(6) 1362-81

[5] Zhan Z H, Zhang J, LI Y, Shi Y H 2011 *IEEE transactions on Evolutionary Computation* **15**(6) 832-47

[6] Wang Y, Li B, Thomas W, Wang J Y, Yuan B, Tian Q J 2011 Self-adaptive learning based particle swarm optimization *Information Science* **181**(20) 4514-38

[7] Li C H, Yang S X, Neuyen T T 2012 *IEEE Transactions On Systems Man and Cybernetics-part B: Cybernetics* **42**(3) 627-46

[8] Liang J J, Qin A K, Suganthan P N, Baskar S 2006 *IEEE Transactions on Evolutionary Computation* **10**(3) 281-95

[9] Martins A A, Oluyinka A A 2013 An adaptive velocity particle swarm optimization for high-dimensional function optimization *Proc. Conf. on Evolutionary Computation Cancun Mexico* 2352-9

[10] Suganthan P N 1999 Particle swarm optimizer with neighborhood operator *Proc. Conf. on Evolutionary Computation Piscataway NJ* 1958-62

[11] Janson S, Middendorf M 2005 *IEEE Transactions on Systems Man and Cybernetics Part B: Cybernetics* **35**(6) 1272-82

[12] Noel M M 2012 A new gradient based particle swarm optimization algorithm for accurate computation of global minimum *Applied Soft Computing* **12**(1) 353-9

[13] Katare S, Kalos A, West D 2004 A hybrid swarm optimizer for efficient parameter estimation *Proc Conf on Evolutionary Computation Portland US* 309-15

[14] Engelbrecht A P 2013 Particle swarm optimization with discrete crossover *Proc Conf on Evolutionary Computation Cancun Mexico* 2457-64

[15] Ji Z, Liao H L, Wang Y W, Wu Q H 2007 A novel intelligent particle optimizer for global optimization of multimodal functions *Proc Conf on Evolutionary Computation Singapore* 3272-5

Authors

Pu Han, born in September, 1959, China

Current position, grades: a professor and the Ph.D. supervisor with the School of Control and Computer Engineering, China.
University studies: B.Eng. degree in thermal measurement and automation from North China Electric Power University in 1982.
Scientific interest: modern control theory and its application, modeling and control of electric power process,.
Publications: 12 books.

Li Meng, born in July, 1985, China

Current position, grades: Ph.D. student in North China Electric Power University, China.
University studies: B.Eng. degree in automation from North China Electric Power University in 2008.
Scientific interest: swarm intelligence-based optimization and its application.

Biao Wang, born in October, 1987, China

Current position, grades: a laboratory teacher in North China Electric Power University, China.
University studies: M.Eng. degree in control engineering from North China Electric Power University in 2013.
Scientific interest: modern control theory and its application.

Dongfeng Wang, born in January, 1971, China

Current position, grades: a professor with the School of Control and Computer Engineering, China.
University studies: Ph.D. degree in thermal engineering from North China Electric Power University in 2001.
Scientific interest: swarm intelligence- based optimization and intelligent control.

SMS text similarity calculation based on topic model

Chengfang Tan[1, 2*], Caiyin Wang[2], Lin Cui[1, 2]

[1]*School of Information Engineering, Suzhou University, Suzhou 234000, Anhui, China*

[2]*Intelligent Information Processing Lab, Suzhou University, Suzhou 234000, Anhui, China*

Abstract

The traditional text similarity calculation is mainly based on the statistical method and the semantic method, it exists data sparse and high-dimensional problems and so on. In order to improve the ability of SMS text similarity calculation, this paper puts forward a kind of similarity calculation method based on topic model. By using LDA (Latent Dirichlet Allocation) to model SMS document set and inference parameter via Gibbs sampling algorithm. The topic-word probability distribution and document - topic probability distribution of the SMS document set are generated. Then use JS (Jensen-Shannon) distance formula to calculate SMS text similarity, finally perform the text clustering experiments on the similarity matrix by single-pass incremental clustering algorithms. Compared with traditional text similarity calculation method, experimental results show that this proposed method can obtain better F-measure, which proves the effectiveness and superiority of the proposed text similarity calculation method.

Keywords: SMS text, similarity calculation, topic model, text clustering, latent Dirichlet allocation

1 Introduction

Text similarity calculation plays an important role in the field of information processing, it is one of the important factors to achieve effective text clustering , and it has been widely used in the field of information retrieval, text copy detection, Q & A system and so on.

The existed text similarity calculation method can be divided into two categories: the statistical method based on TF-IDF and the semantic method based on dictionary [1]. The statistical method based on TF-IDF (Term Frequency - Inverse Document Frequency) represents the text as vector space model and calculates the cosine between the vectors to get the size of text similarity. Its disadvantage is that this method needs large-scale corpus support, ignores the existence of semantic relationships between words, and text representation model has high dimensionality and sparse. The semantic method based on dictionary often represents text as word frequency vector model, constructs semantic relations between words by using domain-specific knowledge base, such as WordNet, HowNet and other external dictionaries. Compared with the statistical method based on TF-IDF, this method does not require the support of large-scale corpus, and has higher accuracy, but the establishment of knowledge base is a complicated project, furthermore, it is difficult to solve the semantic problem of words which are not logged in the dictionary, and it does not take into account the dimension reduction, lacks the definition of text similarity measure.

In this paper, we propose the SMS text similarity calculation method based on topic model, which makes full use of modelling advantages of topic model, maps SMS text corpus to each topic space and mines relationships hidden in the text between different topics and words. We use LDA to model SMS text, and then obtain the topic-word probability distribution and document- topic probability distribution from the model, based on this calculate the SMS text similarity by using famous JS distance formula. Experimental results show that this method has better F-measure.

2 Theory of topic model

LDA model is the typical representative of statistical topic model. Because of its superiority in the text modelling, it becomes a research hotspot in recent years. LDA was put forward by Blei et al [2]. It is a three-tier structure that contains words, topics and documents. The structure of LDA is shown in Figure 1.
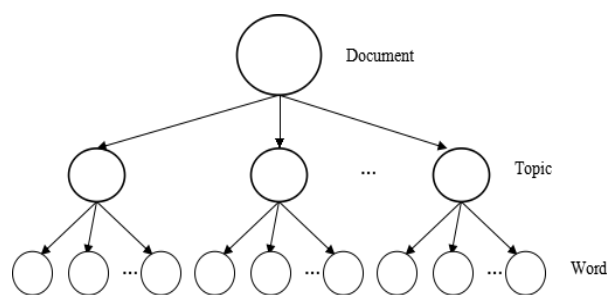
FIGURE 1 The structure of LDA

The main idea of LDA model is that document is represented as the probability distribution of topics, and each topic is represented as the probability distribution of words [3], which is shown in Figure 2.

---

*Corresponding author e-mail: 874036730@qq.com
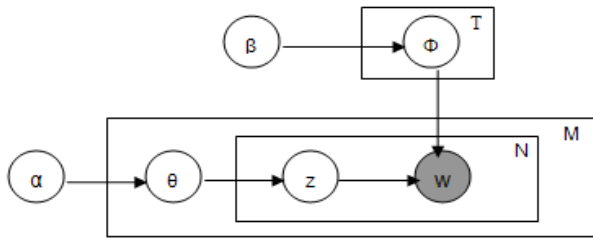
FIGURE 2 LDA model

The meaning of each symbol in LDA model diagram is shown in Table 1.

TABLE 1 The meaning of each symbol

Symbol	Meaning	Symbol	Meaning
α	Super parameter of θ	w	Word
β	Super parameter of Φ	N	The number of words
θ	Document-topic probability distribution	T	The number of topics
Φ	Topic-word probability distribution	M	The number of documents
z	Word probability distribution		

Assuming that the document set $D = \{d_1, d_2, ..., d_m\}$, one document of D is $d = \{w_1, w_2, ..., w_n\}$, the number of topics of D is T, the probabilistic generative process of LDA model is as follows [4]:

1) For each topic t, a word multinomial distribution $\phi^{(t)}$ is obtained from the *Dirichlet(β)* distribution.

2) For each document d, a topic multinomial distribution θ_d is obtained from the *Dirichlet(α)* distribution.

3) For each word w_i in each document, extract a topic t from topic multinomial distribution θ_d, and extract a word as w_i from word multinomial distribution $\phi^{(t)}$ of the topic

From the LDA topic model, we can see that the conditional probability of the word w in a document can be calculated as Equation (1):

$$p(w \mid d, \hat{\theta}, \hat{\varphi}) = \sum_{z=1}^{T} p(w \mid z, \varphi) p(z \mid \hat{\theta}, d), \qquad (1)$$

where z is the topic corresponding to w, T is the number of topics, $\hat{\theta}$ and $\hat{\varphi}$ is the prior estimate of parameters θ and φ respectively.

3 Steps of SMS text similarity analysis based on LDA

SMS text is consisted of a series of independent short text essentially, which contains some special symbols to represent the interaction between user's behaviour and topics. In this paper, we take Chinese SMS text as an example to introduce the process of SMS text similarity calculation.

The analysis process of Chinese SMS text similarity is realized by text pretreatment, feature selection, text representation, LDA modelling and semantic similarity calculation, all together 5 steps, the specific

implementation steps and key problems will be described in followed each section. The analysis steps are as shown in Figure 3.
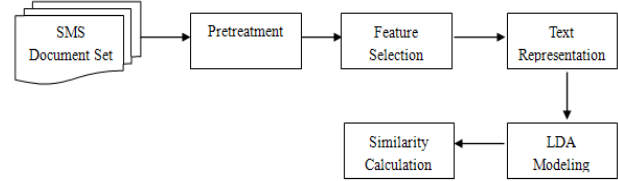
FIGURE 3 Steps of SMS text similarity analysis based on LDA

3.1 PRETREATMENT

Text pretreatment mainly refers to word segmentation, word tagging, and remove stop words, and so on. Chinese SMS text is a very short (less than 74 characters) and narrative essay information. We can remove the very small or irrelevant words with the topic, and use the named entity recognition technology to process personal name, place name, organization name, telephone number and other special information of SMS text, when performing feature selection, we can ignore these special information to avoid affecting the result of text clustering. In addition, the synonyms or near synonyms words are combined and expressed by the unified words.

3.2 FEATURE SELECTION

Feature selection can improve similarity precision for SMS text, and not all the words in SMS text play significant roles on similarity calculation. The words prevalent in all SMS text have no good distinguishing ability and have no great help on similarity calculation, even can affect the SMS text similarity precision, so these words should be removed.

In this paper, we select a number of words which is nouns or verbs in each SMS text as keywords to represent text, thus reduce the dimension of text feature vector as much as possible.

The ordinary feature selection methods include mutual information (MI), information gain (IG), the weight of evidence for text (WET), CHI statistic (CHI), and expected cross entropy (ECE), etc [5]. This paper uses MI method which is good at Chinese text to extract feature words. Assuming that there is a word w_i and category c, the mutual information between w_i and c can be defined as Equation (2):

$$MI(w_i, c) = \log \frac{p(w_i \cap c)}{p(w_i) p(c)}, \qquad (2)$$

where $p(w_i \cap c)$ represents the probability that the word w_i and the category c appears simultaneously, $p(w_i)$ represents the probability that w_i appears and $p(c)$ represents the probability that c appears.

3.3 TEXT REPRESENTATION

This paper uses vector space model (VSM) to represent Chinese SMS text, which can be expressed as Equation (3):

$$D_{mi} = \{w_1, w_2, ..., w_n\} . \tag{3}$$

The feature's weight is calculated by using the famous TF-IDF equation, which is expressed as Equation (4):

$$w_i(d) = \frac{tf_i(d)\log(N/n_i + 0.1)}{\sqrt{\sum_{i=1}^{n} t(f_i(d))^2 \times \log^2(N/n_i + 0.1)}}, \tag{4}$$

where $w_i(d)$ represents the weight of word i in document d, and $tf_i(d)$ represents the word frequency of i in document d, N represents the total number of training texts, n_i represents the number of text which contains i and i appears in training texts. The denominator is the normalization factor.

3.3 LDA MODELLING

In the process of LDA modelling, the parameters of model are required to be estimated, the common estimation method are variational Bayesian inference, expectation propagation algorithm and Collapsed Gibbs sampling. Gibbs sampling algorithm is the most common method, it is easy to understand and implementation, and extract topics efficiently from large scale document set. In this paper, the method used for parameter estimation is Gibbs sampling.

Two matrixes can be obtained from Gibbs sampling process: topic-word matrix and document-topic matrix, namely θ and φ [6], the calculation method is as shown follows Equations (5) and (6):

$$\theta_d = \frac{n_j^m + \alpha}{n_\bullet^d + T\alpha} , \tag{5}$$

$$\theta_d = \frac{n_j^w + \beta}{n_j^\bullet + W\beta}, \tag{6}$$

where n_j^m represents the number of words that document d_m assign to topic j, $n_\bullet^d$ represents the number of words which have been assigned to topic in d_m, n_j^w represents the frequency that word w assign to topic j, $n_j^\bullet$ represents the number of words which have been assigned to topic j.

In topic-word matrix, each row corresponds to feature word of the feature list, each column corresponds to each topic, matrix elements value represent the number of times that the feature word is assigned to corresponding topic. In document-topic matrix, each row corresponds to a SMS text message of data set, each column corresponds to each topic, matrix elements value represents the number of times that feature words of SMS text is assigned to a particular topic.

3.5 TEXT SIMILARITY CALCULATION

Different models usually correspond to different similarity calculation method. This paper uses JS (Jensen-Shannon) distance formula which can measure the distance of probability distribution to calculate the similarity between two documents [7]. We obtain document-topic probability distribution and topic-word probability distribution by constructing LDA model. Therefore, the similarity calculation of two documents can be realized by computing the corresponding topic probability distribution. The distance of vector $p = (p_1, p_2, ..., p_k)$ to $q = (q_1, q_2, ..., q_k)$ is computed as follows Equation (7):

$$D_{js}(p,q) = \frac{1}{2}\left[D_{KL}\left(p, \frac{p+q}{2}\right) + D_{KL}\left(q, \frac{p+q}{2}\right)\right], \tag{7}$$

where $D_{KL}(p,q) = \sum_{j=1}^{T} p_j \ln \frac{p_j}{q_j}$. p and q represents topic probability distribution respectively.

4 Experiments and analysis

To verify the proposed SMS text similarity calculation method in this paper, we use the thought of single-pass incremental clustering to perform experiment on text clustering. The basic idea of text clustering is [8]: preset a clustering threshold u, sequentially process the input SMS text, and calculate the similarity between the new text and identified topic clusters, if the similarity value is greater than the threshold u, the new text is added to its maximum similarity topic cluster, otherwise, the SMS text is created as a seed topic.

The experimental evaluation indexes adopt F-measure to measure the text similarity. F-measure is a balance index, which combine precision rate with recall rate in information retrieval, F-measure is greater, the clustering effect is better.

Precision rate and recall rate are calculated as follows Equations (8) and (9):

$$p(i,j) = \frac{n_{ij}}{n_j}, \tag{8}$$

$$r(i,j) = \frac{n_{ij}}{n_i}, \tag{9}$$

where n_i represents the number of documents contained in the category i, n_j represents the number of documents contained in the clustering j, n_{ij} represents the number of documents where the clustering j belongs to the category i.

F-measure is defined as follows Equation (10):

60

$$F(i,j) = \frac{2p(i,j)r(i,j)}{p(i,j)+r(i,j)} . \qquad (10)$$

The F-measure of global clustering is defined as follows Equation (11):

$$F = \sum_i \frac{n_i}{n} \max \{F(i,j)\} . \qquad (11)$$

4.1 EXPERIMENTAL DATA

This experiment selected five categories of Chinese junk messages as test data，they are lottery, advertisement, service charge, invoice and eroticism, a total of 1000 SMS text messages, we annotated these messages manually.

4.2 EXPERIMENTAL PROCEDURE

The specific experimental procedure is as follows:

1) SMS document set was pre-processed by using ITCLAS (Institute of Computing Technology, Chinese Lexical Analysis System) to execute word segmentation, remove stop words, and so on. We obtained the text vector matrix.

2) LDA was introduced to model the text vector, and used Gibbs sampling method to estimate parameter. We obtained two matrixes: the document-topic matrix and topic- word matrix.

3) The similarity matrix was obtained by using JS distance formula to compute the similarity of SMS text messages.

4) Clustered the SMS text based on the similarity matrix by using single-pass incremental clustering algorithm, and analyzed the clustering results.

4.3 EXPERIMENTAL ANALYSIS

According to the experience value, we set $\alpha = 50/T$ and $\beta = 0.01$, the number of topics T directly affects the precision of LDA model, which will affect the accuracy of clustering results, so we determined the value of T by experiment. The F-measure is higher, the clustering effect is better. Different number of topics generates different F-measure of the clustering, which is as shown in Figure 4.
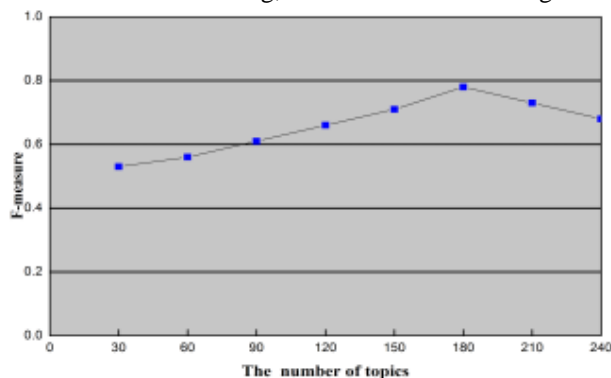
FIGURE 4 Relationship between the number of topics and F-measure

Here the abscissa represents the number of topics, the ordinate represents the size of the F-measure. The Figure 4 reflects the change of F-measure with the different number of topics. As can be seen from Figure 4, the F-measure is highest when $T = 180$, so the number of topics is determined as 180 in this paper.

The experimental statistic found that the similarity of SMS text under the same category of experimental data is commonly above 0.49, and the similarity of SMS under the different categories is generally below 0.35. Therefore, the experiment set the range of clustering threshold u, where $u \in [0.35, 0.49]$. Experiments were carried out 8 times, and used the TDT standard as evaluation index [9], system performance was evaluated by using the macro average of detection cost in 5 categories SMS, The normalized detection overhead changed along with the change of cluster threshold, which is as shown in Figure 5.
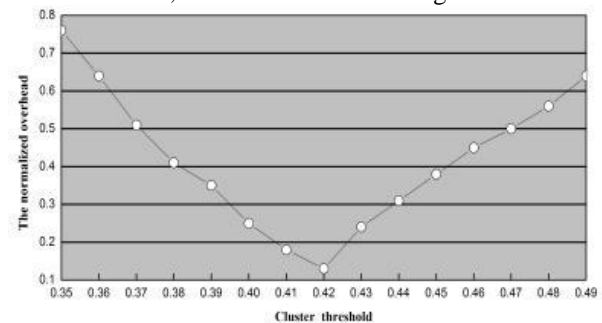
FIGURE 5 Determine the optimal cluster threshold

Here the abscissa represents the preset clustering threshold, the ordinate represents the system normalized overhead. As can be seen from Figure 5, when the threshold is taken 0.42, the system normalized overhead is the minimum, and its value is 0.125, so determine the cluster threshold u is 0.42 in this paper.

To verify the effectiveness of our proposed similarity calculation method based on LDA, we adopt the statistical method based on TF-IDF (referred to as statistics-based) and the semantic method based on Chinese HowNet dictionary (referred to as dictionary- based) as comparative experiments, and compare their clustering effects by single-pass incremental clustering algorithm, Experimental comparative results of three different methods are as shown in Table 2.

TABLE 2 Experimental results of three different methods

Experimental method	Precision (%)	Recall (%)	F-measure (%)
Statistics-based	81.35	81.68	81.51
Dictionary- based	84.24	85.19	84.71
LDA-based	87.87	88.37	88.12

From Table 2, we can see that this proposed method achieves better effects compare with other two methods.

5 Conclusions

In view of the defect of existed statistics-based method and dictionary-based method, this paper takes full advantage

61

of topic model to calculate the text similarity of Chinese SMS, uses LDA to construct the SMS text topic space, comprehensively considers the similarity from semantic similarity of words and word probability distribution in the text, enhances the vector representation of documents, at the same time, greatly reduces the dimension of the document, speeds the computing speed, thus improves the clustering effect. Experimental results show the effectiveness of this method.

In the next step work, we still need to improve the accuracy of document-topic probabilistic distribution and topic-word probabilistic distribution of LDA model, which make an important impact on text clustering.

Acknowledgements

References

[1] Liu J L, Song L Y, Fan Y H 2012 Study of Chinese SMS Text Similarity Based on Semantic Information *Computer Engineering* **38**(13) 58-62 *(in Chinese)*

[2] Blei D M, Jordan M 2003 Modelling annotated data *Proceedings of 26th Annual International SIGIR Conference on Research and Development in Information Retrieval SIGIR'03 ACM* 127-34

[3] Blei D M, Ng A Y, Jordan M I 2003 Latent Dirichlet Allocation *The Journal of Machine Learning Research* **3**(1) 993-1022

[4] Tan C F 2013 Short Text Classification Based on LDA and SVM *International Journal of Applied Mathematics and Statistics (IJAMS)* **51**(22) 205-14

[5] Yahya W B, Ulm K, Fahrmeir L, Hapfelmeier A 2011 A sequential feature selection and prediction method in microarray studies *International Journal of Artificial Intelligence* **6**(11) 19-47

[6] Quan X, Liu G, Lu Z, Ni X, Liu W 2010 Short Text Similarity Based on Probabilistic Topics *Knowledge and Information Systems* **25**(3) 473-91

[7] Griffiths T L, Steyvers M 2004 Finding scientific topics. *Proceedings of 3th National Academy of Sciences* **101** 5228-5235

[8] Zhao A H, Liu P Y, Zheng Y 2013 Subtopic Division in News Topic Based on Latent Dirichlet Allocation *Journal of Chinese Computer Systems* **34**(4) 732-7 *(in Chinese)*

[9] The 2004 topic detection and tracking (TDT2004) task definitionand evaluation plan 2004 *http://www.nist. gov*

Authors

Chengfang Tan, born in February, 1981, China

Current position grades: Researcher at Intelligent Information Processing Lab, Suzhou university, China.
University studies: Master degree in education technology from Nanjing Normal University, China in 2007.
Scientific interests: information retrieval, sentiment analysis and text mining.

Caiyin Wang, born in September, 1978, China

Current position grades: Researcher at Intelligent Information Processing Lab, Suzhou University, China.
University studies: Master degree in computer science and technology at Hefei University of Technology, China in 2009.
Scientific interests: P2P and information retrieval.

Lin Cui, born in August, 1979, China

Current position grades: Researcher at Intelligent Information Processing Lab, Suzhou university, China.
University studies: Master degree in computer science and technology from Hefei University of Technology, China in 2008.
Scientific interests: information retrieval and Semantic Web.

An anti-collision algorithm for adaptive search matrix of cotton seed traceability system based on RFID

Qinglin Huang, Lixin Zhang*, Chaoyang Sun, Xiang Zhang

College of Mechanical and Electrical Engineering, Shihezi University, Shihezi 832003, China

Abstract

Focusing on the tags collision problem of traceability management system for the cotton seed quality and safety based on RFID, we proposed an adaptive search matrix for anti-collision algorithm based on ABS in this paper. Meanwhile, the concept of collision stack is put forward to effectively lower the time of request and transmission of redundant information in the proposed algorithm. The theoretical analysis indicates that this new algorithm performs superiority to the ABS.

Keywords: anti-collision algorithm, adaptive search matrix, RFID, traceability

1 Introduction

With the emergence of seed quality and safety issues, consumers pay more and more attention to the safety of the seed industry, thus controlling quality and safety of the seed is becoming more and more important. Through the whole process of the cotton seed production monitoring based on RFID Technology, complete security control system is established which can quickly traced back cotton seed quality and safety information and warn the major safety and quality hazard of the cotton seed early [1].

For the communication mechanism between electronic tag and reader is the most important module in the system, it is urgent for the reader to identify a large number of tags rapidly and efficiently in the limited time. Anti-collision algorithm is divided into reader and tag anti-collision algorithms, in which tag anti-collision algorithm is to solve the problem of multiplying tags simultaneously communicate with the reader within the effective communication range of the reader. In the high-frequency (HF) bands, ALOHA and related algorithms are applied in tags anti-collision algorithm [2]. In the ultra-high frequency (UHF) band, the widely used algorithm method is binary search algorithm to avoid collisions. Commonly used anti-collision algorithm needs a long time identification and cannot meet the requirement of high-speed movement tags. And although most of the new algorithms have the advantage of short term identification, they require a higher tag design, such as the need to increase the random number generator, remember bits or delays, etc. It is difficult to meet the requirement of low-cost system design. Therefore, maintaining a certain complexity and cost conditions, to minimize the search time and improve the recognition rate is the trend for anti-collision algorithm research.

2 ALOHA algorithm and related algorithms

2.1 ALOHA Algorithm

ALOHA is designed to transmit data used for interactive computer based on a time division multiple access method. Multiple access used for ALOHA [14-15] system is essentially a random time division multiple access, or called random multiple access. Algorithm of ALOHA is based on TDMA (Time Division Multiple Access) thoughts, and therefore it is not for real-time high occasions, but for read-only tags which only store some of the tag data or so.

The basic idea of ALOHA random access mode is that users can send data frame at any time, and we call it the pure ALOHA. The data frame sent by different users may cause conflictions and it will be destroyed, leading to transmission failure, and the sender will not get a confirmation response. If the sender find the frame is destroyed, he will wait for a random length of time before re-transmitting the data frame. When the tag is not enough, it can work well, but in the process of data frame transmission, the probability of collision is greatly increased.

2.2 TIMESLOTTED ALOHA ALGORITHM [3]

In the Aloha algorithm, the tag transmits data through the loop sequence. The time of tag data transmission is only a small part of cycle time. The time of tag data transmission is only a small part of cycle time. After all data transfer to the reader, the tag will wait for a longer time before re-transmitting the data again. Every tag needs a very short waiting time.

---
* *Corresponding author* e-mail: zhlx2001730@126.com
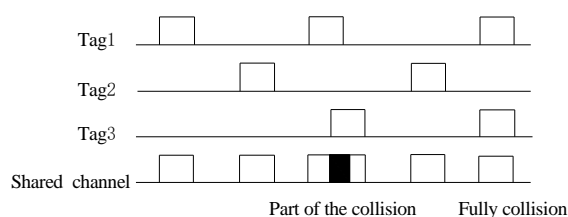
FIGURE 1 ALOHA algorithm model

By this way, after all the tags transfer data to the reader, the repeat process will be ended. By analysing the operating mechanism of Aloha algorithm, it is easy to find that when a tag sent data to the reader, another tag also began sending data to the reader, and the data collision is inevitably occurred.

Given these shortcomings, timeslotted Aloha algorithm have been proposed by researchers. In this algorithm, the tag can only transfer data in the beginning of timeslots. The timeslots for the transmission of data is controlled by the reader. Only when the reader is allocating all of the timeslots can tag transmit data through the timeslots. Therefore, different from Pure Aloha algorithm, timeslotted ALOHA algorithm is TDMA anti-collision algorithm based on random asking drivers.

Since the data is only transmitted in the determinate timeslots, the frequency of the occurrence of the collision is only half of the pure Aloha algorithm but the data throughput performance of the system is doubled. The disadvantage is that the clock needs to be synchronized, and the tag can calculate timeslots.
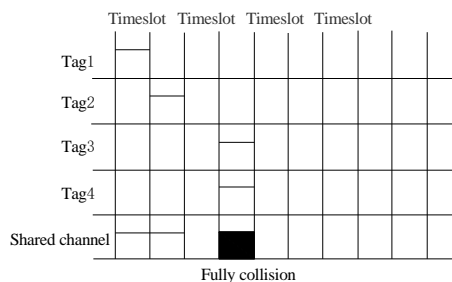
FIGURE 2 Timeslot ALOHA algorithm model

2.3 FRAME TIMESLOTTED ALOHA ALGORITHM [4]

Frame timeslotted ALOHA algorithm is improved based on the timeslotted ALOHA algorithm. The algorithm divided the communication channel into a number of frames, and each frame is divided into many different timeslots. Tag in a data frame can only select a timeslot for data transmission, which can obtain ideal average time for reading the desired tag. The disadvantage is that the frame length is fixed, as a result, it is difficult to achieve better throughput in the actual application.

3 Binary search algorithm [5]

ALOHA algorithm and timeslotted ALOHA algorithm for optimal utilization of the channel is only 18.4% and 36.8%. And with an increasing in the number of tags, its performance shows a sharp deterioration. Binary search algorithm is an algorithm without memory, which is the tag does not have to store the previous check cases so that the cost can be reduced. In this algorithm, a bit prefix is queried in the reader, only these with the serial numbers matched with prefix of the query tags can response to reader commands and send the serial numbers . The tag can be successfully identified when there is only one tag reader response. But the collision can be occurred when there are many tags simultaneously respond. Manchester coding is used in the algorithm to identify the characteristics of collision bits. The prefixes of the query should be modified in the next searching time and the tag eligible to the query should be reduced until there is only one tag left to be identified. The algorithm is simple and easy to implement, but its recognition time is longer with lower search efficiency.

4 An anti-collision algorithm based on adaptive search matrix

4.1 ADAPTIVE ANTI-COLLISION ALGORITHM

Adaptive anti-collision algorithm is generally based on the response of the tag [6], adaptively adjusting the search strategy and the search paths. Algorithm, according to the tag's response, will generally divide timeslots into three states:

1. Collision timeslot: there are two or more electronic tags in the region in response to a reader query command. Due to the collision, the reader cannot identify any of the electronic tag. Therefore, the algorithm needs further subdivision to tag under the timeslot, making one part of the tags continue to response and another part of the electronic tags in a waiting state until there is no or only one tag response through continuous segmentation, .

2. Idle timeslot: There is no tags in the region in response to reader inquiries command. Algorithm returned to the appropriate timeslot node to continue the search according to the end of searching for the timeslot.

3. Readable timeslot: There is only one tag in the region in response to a reader query command. The reader can complete the tag identification and carry on correspond read and write operations. At the same time, Algorithm returned to the appropriate timeslot node to continue the search according to the end of searching for the timeslot. For adaptive search algorithm can save the related search information in the process of searching, and thus unlike binary search algorithm, it does not need the algorithms from scratch after completion of a tag identification, therefore it is more efficient in searching.

Huang Qinglin, Zhang Lixin, Sun Chaoyang, Zhang Xiang

4.2 THE IDEAS OF ALGORITHM DESIGNING

The algorithm in the search process, the search of prefix is not a complete EPC of tag, nor is part of the EPC, but is one bit of EPC, namely to bit by bit to search the EPC. This requires that each tag has a counter, for comparing the current recording bit of EPC i ($1 \leq i \leq n$), where n is the length of the tag's EPC. Reader only sends the i-th bit of EPC, and the tag compare itself to i-th bit of EPC whether consistent with the reader queries, and if not, the tag is in a wait state. If they are consistent, the tag send to reader (i+1)-th bit of EPC, while the tag counter is automatically updated, i= i+1. Reader accord to the response of the tag, judging the slot is idle or collision, readable, and adaptively search on the tags.

4.3 PRINCIPLES AND STEPS OF THE ALGORITHM

Since Manchester [7] coding can effectively recognize the position of the bit of EPC collision occurs. Therefore, according to the collision, the EPC segment. When the EPC have k ($0 \leq k \leq n$)bit in a collision, the collision according to the number of k bits , which can be divided into k / 2 or (k +1)/2 segments, when k is even, the EPC divided into k/ 2 segments; when k is an odd number, the EPC divided into (k +1)/2 segments. For example, there are four tags, they are Tag1: 10011001, Tag2: 10011000, Tag3: 11011001, Tag4: 11010000. Reader identify tag response to 1X01X00X, wherein X represents a bit of collision, So the EPC can be divided into 2 sections, respectively: 1X01X and 00X. Thus, each time searching paragraph by paragraph, rather than the binary collision algorithm bit by bit, in order to reduce the number of searches and improve the search efficiency. The new algorithm for quadtree splitting, but it is not bit by bit, which is paragraph by paragraph. For example, 1X01X bifurcation was 10010,10011,11010,11011; 00X bifurcation was 000 and 001.Thus, there will be a problem, the first paragraph may constitute a fourth-order matrix, second only constitute a second-order matrix, so two can not form a unified matrix, To solve this problem, the use of complementary bit way constitute a fourth-order matrix, thus, k EPC segment can be decomposed into a 4 x k/2 order search matrix, which was R. For example, in the example of Figure 3, the corresponding search matrix can be constructed as follows:

$$R = \begin{bmatrix} r_{1,1} & r_{2,1} \\ r_{1,2} & r_{2,2} \\ r_{1,3} & r_{2,3} \\ r_{1,4} & r_{2,4} \end{bmatrix} = \begin{bmatrix} 10010 & 000 \\ 10011 & 000 \\ 11010 & 000 \\ 11011 & 001 \end{bmatrix}$$ the algorithm identifies Tag1

and Tag2 only 2 depth to complete the search.

Tag1:10011001
Tag2:10011000
Tag3:11011001
Tag4:11010000
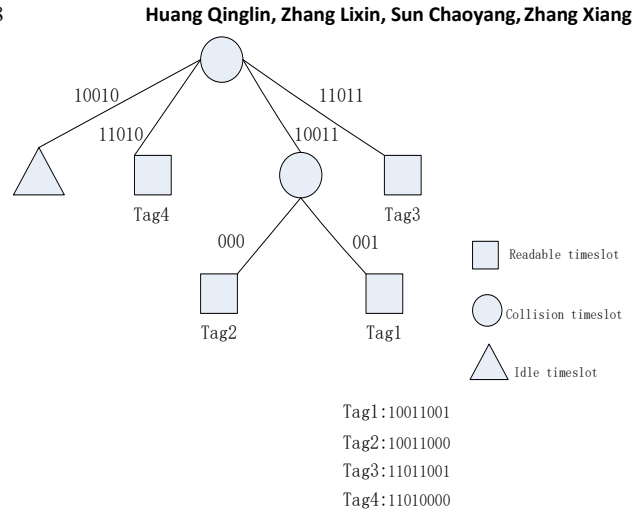
FIGURE 3 Anti-collision algorithm based on adaptive search matrix

Meanwhile, the new algorithm has been improved search strategy also introduces the concept of a collision stack [8]. It is worth noting that, unlike other search algorithms, where the collision stack only stored the search depth values, instead of storing the full search prefix, in the case of the EPC of tag was very long; the reader can save a lot of memory. Since the new algorithm in the search process, need adaptive to determine the search matrix according to the collision bits, so it called anti-collision algorithm based on adaptive search matrix, it is described in detail below:

Step one, Initialization: first of all, the reader send a search command, all tags response, the reader according to the number and location of the collision bit appears, and determining the number of segments k/2 and 4 x k/2 order search matrix. And so i = 1, j = 1; i corresponding column matrix represents the search depth, j represents the number of rows of the matrix corresponding to the search of the ary tree [9].

Step two, depending on the state of the timeslot adaptively adjusted search path.

Collision timeslot: Analysing whether the current search was full Ary Tree (j = 2), and if not, the current value of search depth i stored into collision stack, and, due to the node is greater than or equal to 2 tags exist, so the node needs to be quadtree split, required to make i = i +1, j = 1; reader according to the current i, j, is determined to continue the search EPC.

Idle timeslot: that there is no tag in the node, without having to continue to quadtree split. To determine whether the current search ary tree was full, and if not, then j = 2, continue to search for the other ary tree. If so, you need to return to search. The substance of the binary search algorithm is to return to the initial state of the algorithm, and the algorithm strategy is returned to the first depth of the collision stack. Meanwhile, remove the contents of the first stack, the reader based on the current i, j, determine the EPC to continue the search.

Readable timeslot: Show that within the node has a unique tag, then the reader identify the tag and get on the

corresponding read-write operation. Return to search strategy, and the same with idle timeslot.

Step 3, end of the algorithm: if the stack is empty, when the depth of the search is full ary tree and time slot

state is non collision, the algorithm is over. The above algorithm can be described by a block flow chart of Figure 4:
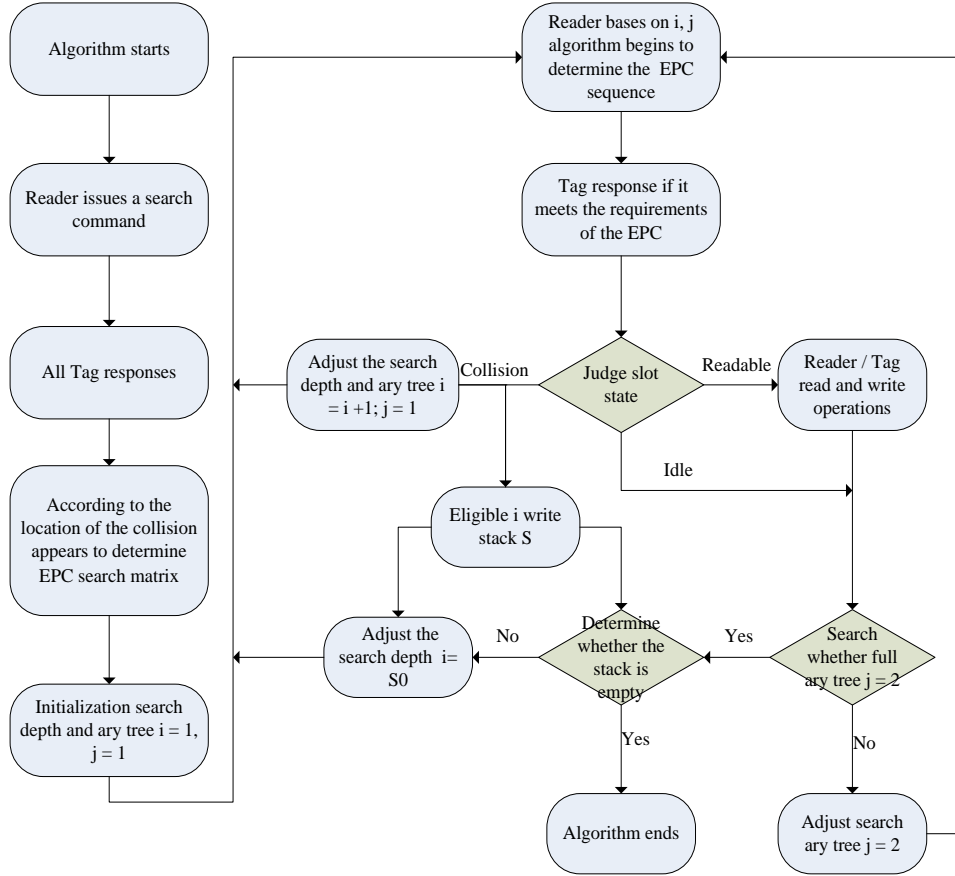
FIGURE 4 Flow chart of the algorithm

4.4 PERFORMANCE ANALYSIS OF ALGORITHMS

Here, though the number of idle timeslots and the collision bit amount calculation and transmission, to analyse the performance of the algorithm.

We use the method of literature [5] to analyse the algorithm, the total number of timeslots required:

$$T = T_c + T_r + T_e.$$

Wherein, the number of time slots T_c, T_r, T_e respectively represent the collision, readable and idle. And The total number of slots is the sum of the number of slots corresponding to the different depth i.

$$T(k) = \sum_{i=0}^{k} T(N, i), \tag{1}$$

$$T_c(k) = \sum_{i=0}^{k} T_c(N, i), \tag{2}$$

$$T_e(k) = \sum_{i=0}^{k} T_e(N, i), \tag{3}$$

where, k represents the maximum search depth, N for the number of tags to be identified within the scope of he reader.

The i depth search can be determined up to 4^i EPC timeslot. for the systems based on RFID, which the length of the tag is n and the total number of tag is N:

1) Estimate the number of idle timeslots

$$T_e(N, i) = 4^i (1 - \frac{1}{4^i})^n.$$

2) Number of readable timeslots.

$$T_r(N, i) = N(1 - \frac{1}{4^i})^{N-1}. \tag{4}$$

3) Number of collision timeslots.

$$T_c(N, i) = 4^i \left[1 - T_e(N, i) - T_r(N, i) \right]$$
$$= 4^i \left[1 - 4^i (1 - \frac{1}{4^i})^n - N \left(1 - \frac{1}{4^i} \right)^{N-1} \right]. \tag{5}$$

Therefore, the formula (5) into formula (2), the total number of collision timeslots.

$$
\begin{aligned}
T_c(k) &= \sum_{i=0}^{k} T_c(N,i) \\
&= \sum_{i=0}^{k} 4^i \left[1 - 4^i (1 - \frac{1}{4^i})^N - N \left(1 - \frac{1}{4^i}\right)^{N-1} \right]
\end{aligned} \quad (6)
$$

The formula (4) into formula (3), the total number of idle timeslots.

$$
T_e(k) = \sum_{i=0}^{k} T_e(N,i) = \sum_{i=0}^{k} 4^i \left(1 - \frac{1}{4^i}\right)^N . \quad (7)
$$

Since the algorithm selected number of segments is based on the median of collision bits, the probability of a bit without collision is $1/4^{N-1}$, the number of stages $k/2$ satisfy the binomial distribution:

$$
p\left(\frac{k}{2}\right) = C_n^{n-\frac{k}{2}} \left(1 - \frac{1}{2^{N-1}}\right)^{\frac{k}{2}} \left(\frac{1}{2^{N-1}}\right)^{n-\frac{k}{2}} . \quad (8)
$$

Formula (6), (7) into formula (8), the algorithm can be obtained the number of collision and idle timeslots:

$$
\begin{aligned}
T_c &= \sum_{k=1}^{n} T_c(k) p(k) \\
&= \sum_{k=1}^{n} \sum_{i=0}^{k} 4^i \left[1 - 4^i (1 - \frac{1}{4^i})^n - N \left(1 - \frac{1}{4^i}\right)^{N-1} \right] \bullet ,
\end{aligned} \quad (9)
$$

$$
C_n^{n-\frac{k}{2}} \left(1 - \frac{1}{2^{N-1}}\right)^{\frac{k}{2}} \left(\frac{1}{2^{N-1}}\right)^{n-\frac{k}{2}} .
$$

$$
\begin{aligned}
T_e &= \sum_{k=1}^{n} T_e(k) p(k) \\
&= \sum_{k=1}^{n} \sum_{i=1}^{k} 2^i (1 - \frac{1}{2^i})^N C_n^{n-\frac{k}{2}} \left(1 - \frac{1}{2^{N-1}}\right)^{\frac{k}{2}} \left(\frac{1}{2^{N-1}}\right)^{n-\frac{k}{2}} .
\end{aligned}
$$

Since binary segmentation algorithm satisfies: $T = 4T_c + 1$, so the formula (9) into the above equation, you can get the total number of timeslots .

Notably, when the N >> n, $p(k=n) \to 1, p(k<n) \to 0$, that formula (9) simplifies

to: $T_c = \sum_{i=1}^{n} 4^i \left[1 - 4^i (1 - \frac{1}{4^i})^N - N \left(1 - \frac{1}{4^i}\right)^{N-1} \right]$

This is consistent with the conclusion of literature [5], when the number of tags are much greater than the

sequence length of EPC, EPC collision occurs almost in every bit, so that the search algorithm only bit by bit, the performance of algorithm is same with ABS.

Tags identification time except related to the number of timeslot, also related to the amount of data communication. ABS [10-13] algorithm for each timeslot transmission only two bits of data, so the identification tag required for transmission of data bits is $N_{ABS} = 2T_{ABS}$.

For this algorithm, data bits to be transmitted in each time timeslot an average of 4n / k, then:

$$
N = 4 \sum_{k=1}^{n} \frac{n}{k} T(k) p(k) .
$$

For BT algorithm, the number of timeslots required is the $T_{BT} = N(\log_2 + 1)$, each timeslot requires 2n-bit data, then: $N_{BT} = 2nN(\log_2 N + 1)$.

From the realization of the algorithm, this algorithm like with the ABS algorithm, each tag also needs a counter for recording the current search depth. Difference is that the reader need to add a register, used as a collision stack, the ABS algorithm requires each tag generates a random number of 0 or 1. So, for a RFID systems with a large number of tags, this algorithm increasing the cost of the reader is more effective and practical than the ABS algorithm increase in cost of tags.

5 Conclusion

In this paper, by constructing a EPC adaptive search matrix and concept of collision stack, and adaptive adjustment of the search path according to the time slot status, so as to put forward a kind of collision algorithm based on adaptive search matrix. Theoretical analysis shows that this algorithm is simple and effective, compared with the other algorithms, by reducing the collision and idle timeslot, and within the timeslot data traffic, which can effectively reduce the identification time, improve the search efficiency of the algorithm.

Reference

[1] Wang Xicheng, Zhen Haifeng 2013 Study of traceability system for agricultural products *Research on Agricultural Issues* (3) 36-42
[2] Finkenzeller K 2003 *RFID Handbook: Fundamentals and Applications in Contact less Smart Cards and Identification* New York, USA: John Wiley&Sons
[3] Hwang T W, Lee B G, Kim Y S, et al. 2006 Improved Anti-Collision Scheme for High Speed Identification in RFID System *Proc. of the 1st International Conference on Innovative Computing,*

information and Control, Beijing, China **11** 449-52
[4] Cha J R, Kim J H 2005 Novel Anti-Collision Algorithms for Fast object Identification in RFID System *Proc. of the 11th International Conference on Parallel and Distributed System. Fukuoka, Japan* **11** 63-7
[5] Myung J, Lee W 2005 Adaptive Binary splitting A RFID Tag Collision Arbitration Protocol for Tag Identification *Proc. of the 2nd International Conference on Broadband New Yorks. Boston,*

USA **I** 347-55

[6] Myung J, Lee W, Shih T K 2006 An Adaptive Memory less Protocol for RFID Tag Collision Arbitration *IEEE Trans on Multimedia* **8**(5) 1096-101

[7] Du Haitao, Xu Kunliang, Xu Weilian 2006 An Anti-Collision Algorithm Based on Binar-Tree Searching of Backtracking *Journal of Yunnan University: Natural Sciences Edition* **28**(SI) 133-6 *(in Chinese)*

[8] Yu Songsen, Zhang Yiju, Peng Weidong, et al. 2004 An Anti-Collision Algorihm Based on Binary-Tree Searching of Regressive Index and Its Practice *Computer Engineering and Applications* **26**(16) 26-8 *(in Chinese)*

[9] ISO/IEC. ISO18000-6-2003(E) Parameters for Air Interface Communications at 860-960 MHz[S] 2003

[10] EPC Radio-Frequency Identity Protocols Class-1 Generation-2 UHF RFID Protocol for communication at 860-960 MHz, 2005

[11] Park J, Chung M Y, Lee T 2007 Identification of RFID tags in framed-slotted Aloha with robust estimation and binary selection *IEEE Commun. Lett.* **11**(5) 452-4

[12] Vogt H 2002 Multiple object identification with passive RFID tags *IEEE International Conference on Systems, Man and Cybernetics*

[13] Schoute F C 1983 Dynamic frame length ALOHA *IEEE Transactions on Communications* **COM-31**(4) 565-8

[14] Xu Lixiang, Lan Yunwei 2006 Implement of RFID Anti-collision of a Binary Search *Microcontroller and Embedded Systems Applications* **6**(5) 33-5

[15] Lee S, Joo S, Lee C 2005 An Enhance Dynamic Framed Slotted ALOHA Algorithm for RFID Tag Identification *Proc. of Mobile and Ubiquitous Systems: Networking and Services. San Diego, California, USA*

Authors

Qinglin Huang

Current position, grades: Master, second
University studies: Shihezi University
Scientific interests: key technologies of radio frequency identification
Publications: none
Experience: Graduated from the Hubei university of Automotive and technology, Now studying in Shihezi University as a master.

Lixin Zhang

Current position, grades: Professor , Ph.D supervisor
University studies: Xi'an science and engineering university
Scientific interests: CNC machine and accurate measurement
Publications: 9 EI, 3 books
Experience: Graduated from Xi'an science and engineering university as a PH.D. Professor, Ph.D supervisor in the Department of Mechanical and Electrical engineering, University of Shehezi of China.

Chaoyang Sun

Current position, grades: Master
University studies: Shihezi University
Scientific interests: key technologies of radio frequency identification
Publications: none
Experience: Shihezi University

Xiang Zhang

Current position, grades: Master, second
University studies: Shihezi University
Scientific interests: Mechanical design and manufacturing automation
Publications: none
Experience: 2009-2013,Shihezi University, 2013-now,Shihezi University

Default assumption reasoning based on fuzzy description logics

Min Zhang*

Department of Computer Engineering JiMei University, Xiamen, China

Abstract

Fuzzy description logics (DLs for short) provide a convenient tool for dealing with inconsistency and uncertainty. People can infer with uncertain and incomplete information. According to the characteristics and requirement of the knowledge representation, fuzzy DLs can play an important role in the commonsense reasoning. Default rules express concise pieces of knowledge having implicit exceptions, which is appropriate for reasoning under incomplete information. Default assumption reasoning based on fuzzy DLs is proposed. Possibility theory is used for representing both uncertainty and defeasibility. Inference service is considered in the logic and algorithms are provided for it.

Keywords: fuzzy description logics, default assumption reasoning, fuzzy reasoning

1 Introduction

Description Logic provides a logical reconstruction of the frame-based knowledge representation languages [1]. It is a significant and expressive representation and based on sound and complete constraint propagation calculi for reasoning in it. It provides a logic foundation [2] of knowledge representation and reasoning for Semantic Web [3]. DL is one of the leading formalisms for storing and manipulating knowledge in the Semantic Web. DLs allow the representation of sophisticated relations between concepts and roles; the sophistication of these relations varies, depending on the DL at hand and determines the expressive power as well as the (algorithm) complexity of reasoning in this DL.

Dealing with uncertainty has been recognized as an important problem in the recent decades. Two important classes of languages for representing uncertainty are probabilistic logic and possibilistic logic. Arguably, another important class of language for representing uncertainty is possibilistic theory [4]. Some approaches have been proposed to extend description logics with uncertainty reasoning such as reported in [5].

Typically, description logic is limited to dealing with crisp concepts. However, many useful concepts that are needed by an intelligent system do not have well defined boundaries. The need of expressing and reasoning with imprecise knowledge and the difficulties arising in classifying individuals with respect to an existing terminology is motivating research on nonclassical DL semantics, suited to these purposes. To cope with this problem, fuzzy description logics have been proposed that allow for imprecise concept description by using fuzzy sets and fuzzy relations. Umberto Straccia [6, 7] extends description logic to the fuzzy case. Fuzzy logic directly deals with the notion of vagueness and imprecision using fuzzy predicates. Therefore, it offers an appealing foundation for a generalization of description logic in order to dealing with such vague concepts.

The process of human is reasoning decision is dynamic and uncertain. We can get the conclusion under the condition of fuzzy and incomplete information. Handling exceptions in a knowledge-based system has been considered as an important issue in many domains of applications. Reiter's default logic [8] is one of the most popular formalisms for describing non-monotonic reasoning and has been extensively investigated by the community working on logical foundations of artificial intelligence. Default rules express concise pieces of knowledge having implicit exceptions, which is appropriate for reasoning under incomplete information. Handling uncertainty in a given complete information context is a need in various situations. For example, high level descriptions of dynamical systems often requires both the use of default rules expressing persistence and the processing of uncertainty due to the limitation of the available information.

Reasoning under incomplete information by means of rules having exceptions, and reasoning under uncertainty are two important type of reasoning that artificial intelligence has studied at length and formalized in different ways in order to design inference systems able to draw conclusions from available information as it is. As already said, reasoning with default rules and under uncertainty are two important research trends that have been developed quite independently from each other in AI. They indeed address two distinct problems, respectively using symbolic and numerical approaches in general.

In this paper, we extend fuzzy DLs by providing a framework for the default assumption reasoning. This paper outlines a joint handling of defaults and fuzzy DLs.

---

* *Corresponding author* e-mail: flyinskyzhang@126.com

Default assumption reasoning based on fuzzy DLs is proposed. Inference service is considered in the logic and algorithm is provided for it. This is a continuous process of adjustment to the fuzzy DLs reasoning. However, the main novelty is that these adjustments are made to non-monotonic reasoning, as defaults describe assumption which guide the reasoning process.

The rest of this paper proceeds as follows. Preliminaries on fuzzy description logics are given in Section 2. Default assumption reasoning based on Fuzzy DLs is provided in Section 3. The inference services are also given. After that, we provide algorithms for implementing reasoning problems. And the last one is the conclusion and the future work.

2 Preliminaries

In this section, we introduce some background knowledge about fuzzy description logics.

Description Logics are a well-known family of knowledge representation formalisms. They are based on the notions of concepts (unary predicates, classes) and roles (binary relations), and are mainly characterized by constructors that allow complex concepts and roles to be built from atomic ones. The expressive power of a DL system is determined by the constructs available for building concept descriptions, and by the way these descriptions can be used in the terminological (*TBox*) and assertional (*ABox*) components of the system.

Straccia extends description logic to fuzzy description logic with fuzzy capabilities. Due to the limitation of space, we do not provide a detailed introduction of fuzzy DLs, but rather point the reader to [6].

2.1 SYNTAX AND SEMANTICS OF FUZZY DLS

Similarly to crisp DL languages, fuzzy DLs concepts are defined by the following syntax rule:

$$C, D \rightarrow \mathrm{T} \mid \perp \mid A \mid \neg C \mid C \cup D \mid C \cap D$$
$$\exists R.C \mid \forall R.C$$

A fuzzy DL knowledge base consists of two finite and mutually disjoint sets. A *TBox,* which introduces the terminology, and *ABox,* which contains facts about particular objects in the application domain.

A terminology, or *TBox*, is defined by a finite set of fuzzy concept inclusion axioms of the form $A \subseteq C$ and fuzzy concept equalities of the form $A \equiv C$.

Objects in the ABox are referred to by a finite number of individual names and these names may be used in two types of assertional statements: concept assertions of the type $a : C$ and role assertions of the type $(a,b): R$, where C is a concept description and R is a role name, and a,b are individual names.

Let $I = \{a,b,c...\}$ be a set of individual names. A fuzzy assertion is of the form $\langle a : C\theta n \rangle$ or $\langle (a,b) : R\theta n \rangle$, where θ stands for $\geq, \leq, >, <$. Intuitively, a fuzzy assertion of the form $\langle a : C \geq n \rangle$ means that the membership degree of a to the concept C is at least equal to n. A finite set of fuzzy assertions defines a fuzzy *ABox* A. The concept of conjugated pairs of fuzzy assertions has been introduced, in order to represent pairs of assertions that form a contradiction. The possible conjugated pairs are defined in table 1, where ϕ represents a concept expression.

TABLE 1 Conjugated pairs of fuzzy assertions

	$\langle \phi < m \rangle$	$\langle \phi \leq m \rangle$
$\langle \phi \geq n \rangle$	$n \geq m$	$n > m$
$\langle \phi > n \rangle$	$n \geq m$	$n \geq m$

TABLE 2 SEMANTICS OF FUZZY CONCEPTS

$$\mathrm{T}^I(a) = 1$$
$$\perp^I(a) = 0$$
$$(\neg C)^I(a) = 1 - C^I(a)$$
$$(C \cup D)^I(a) = \max\left(C^I(a), D^I(a)\right)$$
$$(C \cap D)^I(a) = \min\left(C^I(a), D^I(a)\right)$$
$$(\forall R.C)^I(a) = Inf_{b \in \Delta^I}\left\{\max\left(1 - R^I(a,b), C^I(b)\right)\right\}$$
$$(\exists R.C)^I(a) = Sup_{b \in \Delta^I}\left\{\min\left(R^I(a,b), C^I(b)\right)\right\}$$

A fuzzy set $C \subseteq X$ is defined by its membership function (μ_C), which given an object of the universal set X it returns the membership degree of that object to the fuzzy set. By using membership functions, we can extend the notion of an interpretation function to that of a fuzzy interpretation. A fuzzy interpretation I consists of a pair $(\Delta^I, \cdot^I)$, where Δ^I is the domain of interpretation, as in the classical case, and $\cdot^I$ is an interpretation function which maps concepts (roles) to a membership function $\Delta^I \rightarrow [0,1]$ ($\Delta^I \times \Delta^I \rightarrow [0,1]$), which defines the fuzzy subset C^I (R^I). The semantics of fuzzy DL are depicted in table 2.

A fuzzy concept C is satisfiable iff there exists some fuzzy interpretation I for which there is some $a \in \Delta^I$ such that $C^I(a) = n$, and $n \in [0,1]$. A fuzzy interpretation I satisfies a *TBox T* iff $\forall a \in \Delta^I, A^I(a) \leq D^I(a)$, for each $A \subseteq D$, and $\forall a \in \Delta^I, A^I(a) = D^I(a)$, for each $A \equiv D$. Fuzzy interpretations are also extended to interpret individual and assertions that appear in an *ABox*. For a fuzzy *ABox*, an interpretation maps, additionally, each individual $a \in I$ to some element $a^I \in \Delta^I$. An

interpretation I satisfies a fuzzy assertion

$\langle a : C \geq n \rangle$ iff $C^I(a) \geq n$,

$\langle (a,b) : R \geq n \rangle$ iff $R^I(a^I, b^I) \geq n$.

The satisfiability of fuzzy assertions with $\leq, >, <$ is defined analogously.

A fuzzy *ABox* A is consistent iff there exists an interpretation I that satisfies each fuzzy assertion in the fuzzy *ABox*. We then say that I is a model of A.

2.2 A FUZZY TABLEAU FOR FUZZY DLs

In the following, the fuzzy entailment problem is reduced to the unsatisfiability problem of a set of fuzzy assertions. A tableau algorithm is used to construct a fuzzy tableau for a fuzzy KB. Given a fuzzy $KB \Sigma = \langle T, A \rangle$, let

$S_\Sigma = \{\langle a : C \geq n \rangle\} \cup \{\langle (a,b) : R \geq n \rangle\}$.

It follows that $\Sigma \Rightarrow \langle a : C \geq n \rangle$ iff $S_\Sigma \cup \{\langle a : C < n \rangle\}$ not satisfiable.

The calculus, determining whether a set S of fuzzy assertions is satisfiable or not, is based on a set of fuzzy constraint propagation rules transforming a set S of fuzzy assertions into "simpler" model preserving sets S_i untile either all S_i contains a clash.

A set of fuzzy assertions S contains a clash iff it contains either $\langle w : \perp \geq n \rangle$ with $n > 0$ or $\langle w : \perp > n \rangle$ or $\langle w : \perp < 0 \rangle$ or $\langle w : \top \leq n \rangle$ with $n < 1$, or $\langle w : \top < n \rangle$ or $\langle w : \top > 1 \rangle$, or S contains a conjugated pair of fuzzy assertions.

Given a fuzzy assertion σ, with σ^c we indicate a conjugate of σ. Concerning the rules, for each connective $\bigcap, \bigcup, \neg, \forall, \exists$, there is a rule for each relation $rel \in \{\geq, \leq, >, <\}$. The rules for the case $rel \in \{>, <\}$ are quite similar. The rules are the following:

$(\neg_\geq)$ $\langle w : \neg C \geq n \rangle \rightarrow \langle w : C \leq 1 - n \rangle$

$(\neg_\leq)$ $\langle w : \neg C \leq n \rangle \rightarrow \langle w : C \geq 1 - n \rangle$

$(\bigcap_\geq)$ $\langle w : C \bigcap D \geq n \rangle \rightarrow \langle w : C \geq n \rangle,$
$\langle w : D \geq n \rangle$

$(\bigcup_\leq)$ $\langle w : C \bigcup D \leq n \rangle \rightarrow \langle w : C \leq n \rangle,$
$\langle w : D \leq n \rangle$

$(\bigcup_\geq)$ $\langle w : C \bigcup D \geq n \rangle \rightarrow \langle w : C \geq n \rangle |$
$\langle w : D \geq n \rangle$

$(\bigcap_\leq)$ $\langle w : C \bigcap D \leq n \rangle \rightarrow \langle w : C \leq n \rangle |$
$\langle w : D \leq n \rangle$

$(\forall_\geq)$ $\langle w_1 : \forall R.C \geq n \rangle, \sigma^c \Rightarrow \langle w_2 : C \geq n \rangle$
if σ *is* $\langle (w_1, w_2) : R \leq 1 - n \rangle$

$(\exists_\leq)$ $\langle w_1 : \exists R.C \leq n \rangle, \sigma^c \Rightarrow \langle w_2 : C \leq n \rangle$
if σ *is* $\langle (w_1, w_2) : R \leq n \rangle$

$(\exists_\geq)$ $\langle w : \exists R.C \geq n \rangle \rightarrow \langle (w, x) : R \geq n \rangle,$
$\langle x : C \geq n \rangle$ *if* x *is a new* variable

$(\forall_\leq)$ $\langle w : \forall R.C \leq n \rangle \rightarrow \langle (w, x) : R \geq 1 - n \rangle,$
$\langle x : C \leq n \rangle$ *if* x *is a new* variable

A set of fuzzy assertions S is said to be complete if no rule is applicable to it. These rules are called monotonic rules.

3 Default assumption reasoning

Reasoning under incomplete information by means of rules having exceptions is a very important type of reasoning that artificial intelligence has studied at length and formalized in different ways in order to design inference system able to draw conclusions from available information as it is. It is useful especially for the knowledge representation, which is popular in framework reasoning, diagnostic reasoning, and natural language processing and so on.

However, well-known results from the literature show that DLs have limitations: they do not allow for expressing default knowledge due to their inherent monotonic semantics. One needs nontrivial extensions to the semantics of description logics to express exceptional knowledge.

Example 1. Take, as an example, a bird ontology expressed in the fuzzy DL knowledge base:

$KB = \langle T, A \rangle$:

$T = \begin{cases} Flier \subseteq_1 \neg NonFlier, Pengin \subseteq_1 Bird, \\ Pengin \subseteq_1 NonFlier \end{cases}$

$A = \{Bird(tweety)\}$

Intuitively, KB distinguishes between flying and non-flying objects. We know that penguins, which are birds do not fly. Nevertheless, we cannot simply add the axiom $Bird \subseteq_1 Flier$ to KB to specify the common view that "birds normally fly", as this update will make KB inconsistent. From our bird ontology, we would like to conclude that tweety flies; and if we learn that tweety is a penguin, the opposite conclusion would be expected.

Hence, the simple ontology KB from above cannot express exceptional knowledge, an extension of the semantics of terminological knowledge was given in [9], which is an early attempt to support default logic in the domain of description logics.

Several other attempts to extend DLs with nonmonotonic features have been made based on default logics [10, 11].

In fuzzy DLs, the concept descriptions are interpreted as universal statements, which means, unlike frame languages, they do not allow for exceptions. One needs a formalism that can handle default assumptions, but does

not destroy the definitional character of concept description. Default rules are useful in order to express general behaviours concisely, without referring to exceptional cases. Moreover, they only require general information to be fired, which agrees with the situations of incomplete information. With the aim to offer a user-friendly reasoned over ontologies, we consider default reasoning on top of ontologies based on fuzzy DLs, which integrate default rules and ontologies.

The process of human is reasoning decision under incomplete information can be abstracted as follows: firstly, make an assumption; secondly, reason with the assumptions and draw the conclusions in the case of default assumptions; thirdly, make an evaluation of the result. If the conclusion based on default assumptions is satisfied or consistent with current knowledge base, it is accepted and continues to make further decision reasoning. Otherwise, the assumption is denied, either given up or reassumed.
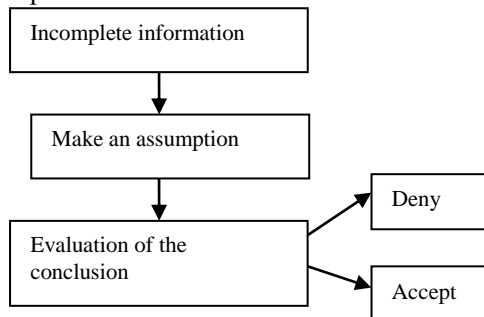
FIGURE 1 The process of default assumption reasoning

According to the process, we consider the default assumption reasoning of fuzzy DLs. First of all, the default assumption rule base is made under the incomplete information from which we can draw conclusions. Secondly, evaluation system is established. The conclusion is evaluated according to the decision goal. Finally, if the conclusion is inconsistent with the growing information, we can go back to the previous hypothesis.

3.1 THE FRAMEWORK OF DEFAULT ASSUMPTION REASONING

In the inference system based on fuzzy DLs, the rule base is divided into two parts: one is the general rule base and the other one is default assumption rule base. Both rule bases including knowledge base work together to draw a conclusion. The framework of reasoning is shown in figure 2.
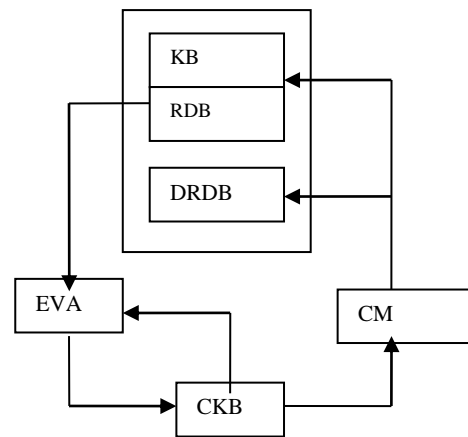
FIGURE 1 The framework of default assumption reasoning

Each part of figure 2 is defined as follows:

KB: a knowledge base of the form $\langle T, A \rangle$.

RDB: a rule base including the reasoning rules based on fuzzy DLs, which are described in section 2.2

DRDB: a default assumption rule base

CKB: a current knowledge base contains current information, which is used during the reasoning process

EVA: the evaluation mechanism, which needs a trigger

CM: the inference control mechanism, which controls the reasoning process

The basic idea and the process are described as follows:

(1) The current knowledge base CKB submits to the inference control mechanism CM the current knowledge (fuzzy assertions);

(2) Inference control mechanism CM will search for the information in the current knowledge base CKB and the rule base RDB. If there are rules matched, the inference is triggered by the reasoning mechanism. Otherwise, if there are no rules or knowledge matched in CKB and RDB, we will check the default assumption rule base DRDB. If there are some default assumption rules matched in DRDB, we will go on the hypothesis reasoning and get the conclusion. At the same time, we mark the conclusion and prepare for the tracing back.

(3) The conclusion directly reasoned by the rule base RDB will be put in the current knowledge base CKB. If it is drawn by the default assumption rule base DRDB, it needs the evaluation mechanism EVA to check it. If it is inconsistent with the current knowledge of CKB, the conclusion reasoned by default assumption is considered to be not plausible and abandon, and we will go back to the previous hypothesis state.

(4) For the fuzzy assertion of the form $\langle x : C \geq n \rangle$, if we cannot check the satisfiability of its negative form $\langle x : \neg C < 1 - n \rangle$, then the default assumption rule base DRDB is called. We assume that $\langle x : C \geq n \rangle$ be satisfiable and reason with it. Following the hypothesis we get the conclusion. After that, we call the evaluation mechanism EVA to check it.

3.2 UNCERTAIN DEFAULT RULES

A default rule is an expression $a \sim> b$ where a and b are formulas and $\sim>$ is a new symbol. $a \sim> b$ Translates, in the possibility theory framework, into the constraint $\Pi(a \wedge b) > \Pi(a \wedge \neg b)$, which expresses that having b true is strictly more possible than having it false when a is true.

The use of default rules has two main interests. First, it simplifies the writing: it allows us to express a rule without mentioning every exception to it. Second, it allows us to reason with incomplete descriptions of the world: if nothing is known about the exceptional character of the situation, it is assumed to be normal, and reasoning can be completed.

In order to have more expressive representation formalism, we now introduce the notion of uncertain default rule.

Definition 1 An uncertain default rule is a pair $(a \sim> b, \alpha)$ where a and b are concepts or roles of DLs, and α is the certainty level of the rule, the symbol $\sim>$ is a non classical connective encoding a non-monotonic consequence relation between a and b.

The intuitive meaning of $(a \sim> b, \alpha)$ is "by default" if is true then has a certainty level at least equal to α. For instance, $(bird \sim> flies, \alpha_1)$ means that a bird generally flies with certainty α_1. It is a default rule since it admits exceptions mentioned in other rules: for instance, $(bird \wedge young \sim> \neg flies, \alpha_2)$: young birds generally do not fly. But it is also an uncertain rule since all we know is that we are in presence of a bird, the certainty level α_1 is attached to the provisional conclusion that it flies. Thus, the α's provide an additional information with respect to the default rule.

The core of the treatment of uncertain default rules proposed in this paper is based on the idea of translating them into a set of uncertain rules.

Roughly speaking, default reasoning amounts to apply a set of default rules to the knowledge base describing a context.

The core of the treatment of uncertain default rules proposed in this paper is based on the idea of translating them into a set of uncertain (non defeasible) rules.

Definition 2 Let $D = \{(a_i \sim> b_i, \alpha_i), i = 1, 2..n\}$ be an uncertain default rules set. Suppose $\beta_j, j = 1, 2..k$ are all distinct weights appearing in D such that $\beta_1 > \beta_2 > ... > \beta_k$. Let $\Sigma_D = (S_1, S_2, ... S_k)$, where $S_i = \{a_l \rightarrow b_l : (a_l \sim> b_l, \alpha_l) \in D, \alpha_l = \beta_l\}$.

Since possibilistic inference suffers from the drowning problem, we consider a drowning-free variant of possibilistic inference, called linear order inference. Uncertain default rules are stratified by the weights. Reasoning from a set of exception-tolerant default rules in presence of incomplete knowledge first amounts to

select default rules. The selected set of rules should focus on the current context describing the particular incomplete information situation that is considered, and then this set of rules can be applied to this information situation in order to draw plausible conclusions. When new information is available on the current situation, these conclusions may be revised at the light of more appropriate default rules. The selection problem is solved in practices by rank-ordering the default rules in such a way that most specific rules whose conclusion may conflict with the conclusion of more general defaults, receive a higher level of priority. Clearly, the level of priority of a particular rule depends on the whole set of default rules which are considered.

3.3 ALGORITHMS FOR INFERENCE IN THE FRAMEWORK

We give algorithms for the inference in the framework based on fuzzy DLs.

(1) Algorithm 1 computes the fuzzy membership degree n of the fuzzy assertion $\langle x : C \geq n \rangle$. $i.\alpha$ of the algorithm 1 is the fuzzy membership degree of the prequisite of the default rule.

Algorithm 1. Function FMA

Data: $KB = \langle T, A \rangle$; a DL concept $\langle a : C \geq n \rangle$

Result: The membership degree n associated with a query $C(a)$

begin

Foreach r in RDB and DRDB do

　　　if $\langle a : C \geq n \rangle$ matching r.rule

　　　　if r.rule is default rule in DRDB

　　　then flag=1

else flag=0

　　　　if flag==1 then

　　　　　if $r.\alpha < 1$ then $n = r.\alpha$

　　　　　else $n = 1$

else according to the RDB to compute n

end

(2) Algorithm 2 of the inference control mechanism FCM

μ of the algorithm 2 is the threshold.

Algorithm 2 Function FCM

Data: $CKB = \{\langle \phi_i, \alpha_i \rangle : \alpha_i \in [0,1], i = 1,...,k\}$, where k is the number of fuzzy assertions in the current knowledge base CKB; a fuzzy DL concept $\langle a : C \geq n \rangle$

Result: The update of knowledge base KB

begin

if $\langle a : C \geq m \rangle$ exists in CKB then

　　add $\langle a : C \geq \min(m, n) \rangle$ to knowledge base KB

else FMA($\langle a : C \geq n \rangle$)

if $n > \mu$ then

if flag==1 then EVA($\langle a : C \geq n \rangle$)

else add $\langle a : C \geq n \rangle$ to knowledge base KB

end

    (3) Algorithm 3 of the evaluation mechanism EVA

    Alogrithm 3 Function EVA($\langle a : C \geq n \rangle$)

    Data: $KB = \langle T, A \rangle$; a fuzzy DL concept $\langle a : C \geq n \rangle$

    Result: The update of knowledge base KB

begin

check $\langle a : C \geq n \rangle$ in knowledge base KB

if it is consistent then add it to knowledge base
else give up the result
end

3.4 DISSUTIONS ON THE CONSISTENCY

If the conclusion is drawn from the default assumption rule or part 4 of section 3.1, we will mark a sign in the algorithm FMA and FCM. When the conclusions tagged are inconsistent with ones from EVA, we will delete the tagged conclusions to make assure the consistency of the knowledge base.

    To get a sense of how representations work within this formalism consider the following example.

    Example 2. Suppose we have a knowledge base $KB = \langle T, A \rangle$

$$T = \begin{cases} Penguin \subseteq_{0.8} \neg fly, Ostrich \subseteq_{0.8} \neg fly, \\ Sparrow \subseteq_1 bird, Penguin \subseteq_1 bird \end{cases}$$

$$A = \varnothing$$

$$DRDB = \begin{cases} (bird \wedge Penguin \sim> \neg fly, 0.8), \\ (bird \wedge Ostrich \sim> \neg fly, 0.8), \\ (bird \sim> fly, 0.6) \end{cases}$$

    Now the query is "Can Sparrow fly?", which means that we should judge the entailment problem of the concept "Sparrow" and "fly", namely $Sparrow \subseteq_n fly$, $n \in [0,1]$. At first, we change T and A into the form of fuzzy constraint system:

$$S = \begin{cases} x : \neg Penguin \cup fly \geq 0.8, \\ x : \neg Ostrich \cup fly \geq 0.8, \\ x : \neg Sparrow \cup bird \geq 1, \\ x : \neg Penguin \cup bird \geq 1 \end{cases};$$

    By the definition 2, the uncertain default rules of $DRDB$ are stratified: $DRDB = (S_1, S_2)$, where

$$S'_1 = \begin{cases} (bird \wedge Penguin \sim> \neg fly, 0.8), \\ (bird \wedge Ostrich \sim> \neg fly, 0.8) \end{cases}$$

$$S'_2 = \{(bird \sim> fly, 0.6)\}$$

    For simplicity, we translate them directly into a set of uncertain rules:

$$S_1 = \begin{cases} (bird \wedge Penguin \to \neg fly, 0.8), \\ (bird \wedge Ostrich \to \neg fly, 0.8) \end{cases}$$

$$S_2 = \{(bird \to fly, 0.6)\}$$

Secondly, the problem is changed into the fuzzy assertion satisfiability or default assumption satisfiability check. We add $\{x : Sparrow \cup \neg fly < n\}$ to the fuzzy constraint system S, the propagation rules of RDB are applied as follows:

(1) $\langle x : Sparrow \cup \neg fly < n \rangle$

(2) $\langle x : Sparrow < n \rangle$ $\cup_<$

(3) $\langle x : \neg fly < n \rangle$ $\cup_<$

(4) $\langle x : \neg Sparrow \cup bird \geq 1 \rangle$

(5) $\langle x : \neg Sparrow \geq 1 \rangle$ $\cup_\geq$

(6) clash   (2) and (5)

(7) $\langle x : bird \geq 1 \rangle$ $\cup_\geq$

(8) $\langle x : fly \geq 0.6 \rangle$ the rule of S_2

(9) clash if $n < 0.6$ (3) and (8)

It follows that $Sparrow \subseteq_n fly$ if $n \geq 0.6$. Therefore, if the threshold of μ is set 0.5, we can draw a conclusion that "Sparrow can fly in general".

4 Related work

There have been some works in fuzzy DLs, such as the work reported in [12-16]. But there has been very few works handling both defeasibility and uncertainty. Our main goal is to provide a framework for fuzzy default reasoning. The formalism used in this paper is an extension of fuzzy DLs as introduced by Straccia.

    In [17] a technique is described for assigning a preference semantics for defaults in terminological logics, which uses exceptions and therefore has some similarities to our work. They draw a distinction between strict inclusions (TBox statements of the form) and defaults, which is interpreted as "soft" inclusions. We also make use of the notion of the stratification of default theories studied in detail by Choleuinski [18]. In our framework, the uncertain default rules are stratified by membership degree.

    Using an uncertain framework in order to describe an evolving system has been done by many authors, for instance in a probabilistic setting. But reasoning in this setting implies to dispose of many priori probabilities; this is why using defeasibility may help to reduce the size of information for representing the system. In this sense our framework is more expressive. However, the use of default rules requires more complex approach. The computation of extensions is difficult and we do not have to take account for it.

    An altogether different approach is the explicit introduction of nonmonotonicity into DLs, usually some variant of default logic. See [19] for an overview. But this extension is limited to classical description logics, fuzzy characteristic is not considered.

    Nicolas, Garcia and Stephan [20] also present an approach that deals with defeasibility and uncertainty in a possibilistic framework. But, they combine possibilistic

logic with Answer Set Programming rather than using the same setting for default and uncertainty handling. Dupin [21] has introduced a new rewriting algorithm in applications to handle uncertain default rules. This work is different from the work [22] on extending fuzzy DLs with default rules, where classical default rules are directly attached with the tableau algorithm. The classical default rules are more complex while the use of default assumption rule allows us to express a rule without mentioning every exception to it.

5 Conclusions and future work

Based on the fuzzy description logic with default reasoning, we construct a knowledge base system that incorporates *TBox*, *ABox* and default rules. We have proposed a default assumption extension for fuzzy description logics and provided corresponding algorithms in this paper. We provide a theoretical framework allowing us to study the feasibility of applying default assumption in fuzzy DLs. The model described here shows an efficient way to use uncertain default reasoning as a tool for fuzzy DLs. This topic is interesting because the problem of dealing with uncertainty is closely related to the problem of nonmonotonic reasoning and possibility

theory has been widely accepted as an important formalism to deal with uncertainty.

We have shown how the structure of fuzzy DL languages can be exploited to define basic knowledge base. Our focus was on the formal semantics of knowledge representation, although we also provided high-level decision procedures. The next step is the development of tableaux-based algorithms for implementing the strategies outlined in the paper. The combination of default assumption rules and fuzzy DLs enable us to capture knowledge about vagueness and defeasibility. Furthermore, the incorporation of default assumption allows us to encode and reason under incomplete information. Stratification strategy is considered in the uncertain default rule set which is more convenient for default reasoning. We have not considered the complexity analysis of algorithms but leave them as future work.

An obvious question to consider is whether any additional structure, such as the specification of role hierarchies, can be exploited further to modify the knowledge representation in appropriate ways. Finally, the management of other notions of incoherence, such as concept unsatisfiability, is currently the topic of further investigation.

References

[1] Baader F, Nutt W 2003 *Basic description logic* The Description Logic Handbook: Theory, Implementation and Applications. Cambridge: Cambridge University Press 47-100

[2] Horrocks I, Patel Schneider P F 2003 Reducing OWL entailment to description logic satisfiability Fensel D, Sycara K, Mylopoulos J. *Proc of te 2003 International Semantic Web Conferences(ISWC 2003). Berlin: Springer* 2003 17-29

[3] Berners Lee T, Hendler J, Lassila O 2001 The semantic Web *Scientific American* **284**(5) 34-43

[4] Dubois D, Lang J, Prade H 1994 *Possibilistic logic* In: Handbook of Logic in Artificial Intelligence and Logic Programming 439-513 Oxford University Press, Oxford

[5] Guilin Qi, Jeff Z, Ji Qiu 2007 *Extending description logics with uncertainty reasoning in possibilistic logic* ECSQARU 2007, LNAI 4724 828-39

[6] Straccia U 1998 A fuzzy description logic, Proceedings of the Fifteenth National *Tenth Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence, Madison, Wisconsin, United States* 594-9

[7] Straccia U 2001 Reasoning within fuzzy description logics *Journal of Artificial Intelligence Research* **14** 137-66

[8] Reiter R 1980 A Logic for Default Reasoning *Artificial Intelligence* **13** 81-132

[9] Baader F, Holunder B 1995 Embedding defaults into terminologyical knowledge representation formalisms *J. Autom. Reasoning* **14**(1) 149-80

[10] Straccia U 1993 Default inheritance reasoning in hybrid KL-ONE-style logics *In: IJCAI 1993* 676-81 Morgan Kaufmann, San Francisco

[11] Padgham L, Zhang T 1993 A terminological logic with defaults: a definition and an negation as failure *In: IJCAI 1993* 662-8 Morgan Kaufmann, San Francisco

[12] Straccia U 2009 Multi Criteria Decision Making in Fuzzy Description Logics: A First Step *In KES* 79-87

[13] Bobillo F, Delgado M 2008 A reasoner for fuzzy OWL 1.1. *In URSW 2008. CEUR Workshop Proceedings* **423** 10

[14] Bobillo F, Straccia U 2008 FuzzyDL: An expressive fuzzy description logic reasoner *In: International Conference on Fuzzy Systems, pp. 923-930. IEEE Computer Society, Los Alamitos*

[15] Lukasiewicz T, Straccia U 2008 Managing uncertainty and vagueness in description logics for the semantic web *Journal of Web Semantics* **6** 291-308

[16] Stoilos G, Stamou G, Tzouvaras V 2007 Reasoning with very expressive fuzzy description logics *Journal of Artificial Intelligence Research* **30**(5) 273-320

[17] Quantz J, Royer V 1992 A preference semantics for defaults in terminological logics *In proceedings of KR'92* 294-305

[18] Choleuinski P 1993 *Seminormal stratified default theories* Tech.Rep., Univ. of Kentueky, Lexington

[19] Baader F, Kusters R, Wolter F 2003 *Extensions to description logics* In The Description Logic Handbook

[20] Nicolas P, Garcia L, Stephan I 2005 A possibilistic inconsisteny handling in answer set programming *In Proc. of ESCQARU'05* 402-14 Springer

[21] Dupin F, Prade H 2006 Possibilistic Handling of Uncertain Default Rules with Applications to Persistence Modelling and Fuzzy Default Reasoning *American Association for Artificial Intelligence* 440-50

[22] Zhang Min 2013 A Fuzzy Description Logic with Statistical Default Inference *JDCTA* **7**(1) 50 - 7

Authors

Zhang Min, born in May, 1980, China

Current position, grades: lecturer of Jimei university
University studies: master in Xiamen university
Scientific interest: AI, logic reasoning
Publications: Statistical Default Inference Based on DFL, CSSE
Experience: I was graduated from Xiamen university and got master degree of computer application technology. Now I am working in Jimei university as a teacher. Main research interests are inconsistent information processing, imprecise reasoning, artificial intelligence and so on.

Case-based reasoning adaptive optimization algorithm for power transformer fault diagnosis

Wei Zhang, Qiu-li Wu, Yu-rong Deng, Ze-cheng Lv*

Guangxi Electric Power Research Institute, 530023 Nanning Guangxi Zhuang Autonomous Region, China

Abstract

The adaptive learning rate for the introduction of case-based reasoning transformer fault type identification. The adaptive learning rate theory, through improved data normalization, typicality and best filtering diversity to extract the original example and optimal neural network. In the sample processing and analysis process to be solved according to the type of fault feature automatically adjusts the data processing methods, processes, boundary conditions and constraints to adapt statistical distribution, the probability characteristics. Examples show that this method can overcome the DGA data ambiguity and dispersion problems in the recognition accuracy and convergence speed advantage.

Keywords: Adaptive, Case-Based reasoning, neural network, fault type, normalization, data filtering

1 Introduction

Dissolved gas analysis in oil technology has proven to be simply and effectively on transformer fault diagnosis. In recent years, scholars have put forward comprehensive diagnostic methods, which is mainly of dissolved gas analysis, combined with other electrical test results, such as pteri [1-2], information fusion network [3-4], decision tree [5-6], Expert System [7], etc. At present transformer faults diagnosis is by building more between fault symptoms and the mechanism of the mathematical model for fault type recognition, but because of the complex relationship between fault symptoms and mechanism, it is difficult to extract the input feature effectively according to the fault type [9]. Making fault recognition through the deterministic model is difficult.

In recent years, based on the DGA, some more efficient ways of dealing with the uncertain fault rough set theory and Bayesian network. But the methods need lots of statistical information and prior knowledge, which make calculating and training complex; By the way, the complexity of the actual transformer faults and the running environment that cause failure to grasp whether the model in the learning process fully or excessive[9], when type key information incomplete, fault fuzzy, rough knowledge intensive rule base too large, which affect the diagnosis of applicability.

In view of the need to overcome ambiguity problem in transformer fault type identification, and accurately reveal the key fault information; In this article, through improved typical and sample screening classification method and the network training process, set up Case-Based reasoning model based on vector optimization of adaptive adjustment. Case analysis indicates that the

reasoning model can reduce redundant diagnostic information, avoid the tedious calculating deduction, overcome data discreteness and fuzziness of gases dissolved in transformer oil, which make fault recognition neural network model more practical.

2 Case-Based matching method of fault type

2.1 CASE-BASED REASONING PROCESS

Case-Based reasoning make the problem to be solved as the goal, and the existing problems of samples is called the source example. Case-Based reasoning, in a nutshell, is by the target sample tips to known source paradigm, and directed by source sample solving target Case [10], the process is shown in figure 1.
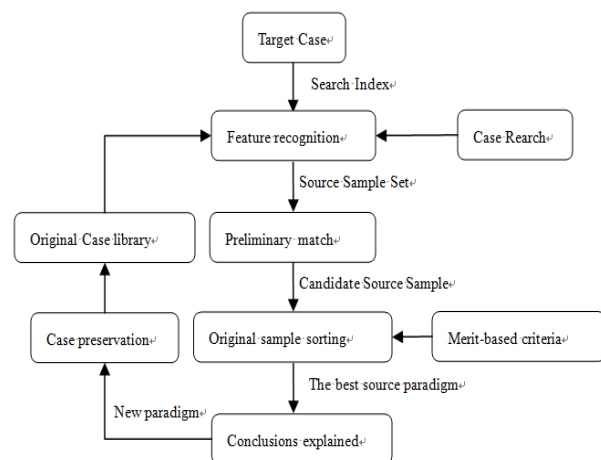
FIGURE 1 the basic principle of case-based reasoning

---

* *Corresponding author* e-mail: wu_ql.sy@gx.csg.cn

Zhang Wei, Wu Qiu-li, Deng Yu-rong, Lv Ze-cheng

Process, the sample retrieval decided to source sample quality of solving the key problem. Sample retrieval contains characteristics identification, best source example, preliminary matches the selected three sub-processes:

(1) feature recognition is a new problem in the process of feature extraction.

(2) Initial matching is based on the new problems from the source of case base retrieval related candidate source example.

(3) Selected from the best refers to the selected candidate for example a and failure type of the current source example of the best match.

As a result of the example reasoning knowledge basis is sample, and sample acquisition relatively easy; At the same time, will be the best source sample results and answer for new solution of the fault type, improve the efficiency of solving the new problems. It is hard to according to the procedures for identification of fault types, using example reasoning may have better effect.

2.2 HIERARCHICAL DIAGNOSTIC CLASSIFICATION AND CASE BASE

Through hierarchical diagnosis for fault feature data necessary to simplify and reasonable assumptions, can make the fault diagnosis process linearization, homogenization. The hierarchical diagnosis can step by step a thorough and detailed strategies, achieve a detailed and accurate fault type recognition [11]. It step by step by step a course assigned to the segment to 11 class using the neural network algorithm to calculate the possible fault types and similar cases. In the hierarchical diagnosis combined with the prior data of the sample reasoning, can improve diagnosis effectiveness; its premise is to must classify source case base.

2.3 SAMPLE CLASSIFICATION RETRIEVAL

Model retrieval based on hierarchical diagnostic process, coarse assigned to the segment classification retrieval method. The method has the characteristics of heuristic algorithm, can quickly find the optimal solution; And avoid the classification carefully caused by different fault types of interference.

Based on heuristic rules to retrieve, not all conditions are necessary, may be only part of the sample to infer the conclusion. Retrieval process is the process of identification of target sample, in the first level recognition classification using three ratio method to coarse points, the second and third level by using BP neural network (BPNN) segmentation step by step. The basic process is shown in figure 2.
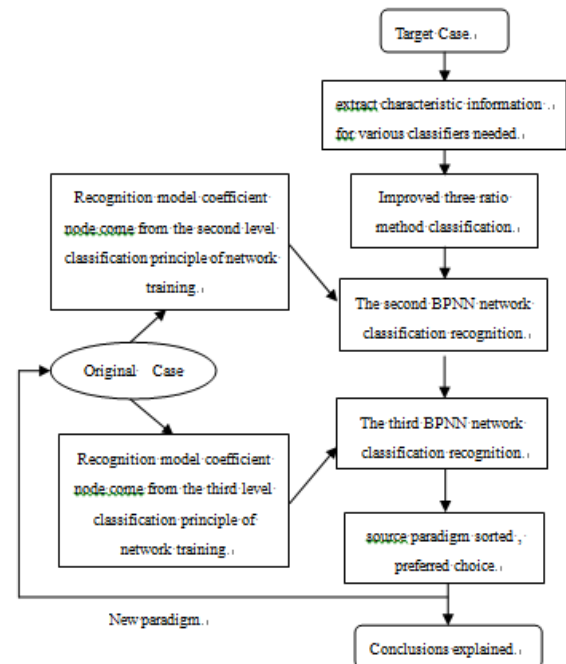
FIGURE 2 case retrieval process

2.4 IDENTIFICATION MODEL OF NETWORK STRUCTURE

BPNN is according to the error backward propagation algorithm training of the multilayer feedforward network, is currently one of the most widely used neural network model [12]. BPNN can learn and store a lot of input - output model mapping, without prior mathematical equation describing the mapping relation, for example reasoning. According to example reasoning of causality, structure to solve the fault type and source of case base relationships between each vector equation model, can map the complex nonlinear relationship, make it simple and applicable. The network structure as shown in figure 3.
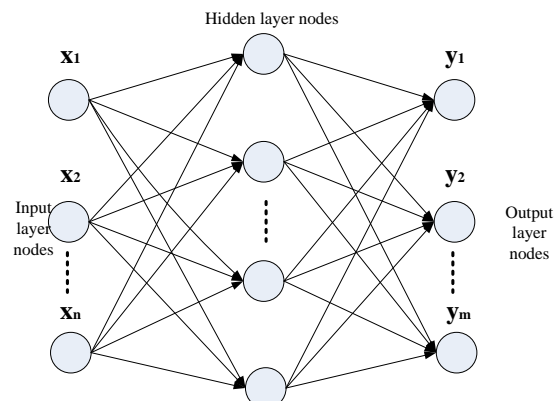
FIGURE 3 Schematic diagram of BPNN

3 Adaptive adjustment of vector model building

BPNN model was constructed based on sample classification retrieval algorithm, and the coefficient of

77

BPNN trained by analytic network nodes in each layer, to determine the network identification model. BPNN network training contains positive transmission output and back propagation adjustment of two parts, in the process of its positive transmission output, the input value by the input layer through the hidden layer nodes and send to the output layer; If the output values reach the desired effect, the output value of error back propagation along the original connection path to the input layer, automatic correction the connection weights between nerve cells, reduce the error gradually narrowed.

3.1 ADAPTIVE ADJUSTMENT VECTOR ALGORITHM

Network identification model first need to training the model, the training of the BPNN model steps as follows:

(1) the beginning of the weight coefficient of each layer of Wij, I said layer index, j said node number index.

(2) From the standard input value and initialized weights coefficient, network positive propagation path is used to calculate the output.

(3) According to the difference of the output and the expected output as a feedback factor adjustment coefficient of node weights Wij.

(4) Repeat steps 2 and 3 until the calculation error of the output and the expected value to achieve the ideal range.

(5) Set the input vector X = (x1, x2… xn), toward the output of Y = (y1, y2…ym) and expected output for T = (t1, t2…tm), node weights coefficient is W = (wi1, wi2…wis). The first k times back propagation error is: $E(k) = T - Y(k) = T - XW(k)$.

In order to improve the convergence of the network training algorithm, using the adaptive vector adjustment of network learning algorithm, its weight adjustment process can be represented as:

$$\begin{cases} w(k+1) = w(k) + \Delta w(k+1) \\ \Delta w(k+1) = mc \times \Delta w(k) + (1 - mc) \times \alpha(k) \times D(k) \\ D(k) = -\dfrac{\partial E}{\partial W} \end{cases} \quad (1)$$

Type, was the weight variation; The MC for momentum factor MC (0;1); (k) is a vector of k time; D(k) for the gradient values of k time.

Vector among them, the adjustment has the adaptive ability, when training network output error is greater than the last time in the process of the output of the error, namely E E (k) (k - 1), the vector is automatically reduced, namely(k + 1) = (k) by dm (constant) of dm for less than 1. < E and E (k) (k - 1), the vector will increase, namely (k + 1) = (k) by im (im constant of greater than 1).

Network training and network to identify the difference between network training input layer and output layer parameters as known variables, and the coefficient of network nodes as unknown variables, should be calculated through the network training; Network to identify the parameters of input layer and the network node coefficient is known variables, output layer output parameters as unknown variables, to identify calculated through the network. Network training, first of all, to do the normalized processing data, ensure the different characteristics of the measuring data consistency; And then to the training sample selection, remove the borderline sample; Finally, the training sample data after processing.

2.2 DATA NORMALIZATION PROCESS

Due to the different characteristics of the components of the content of dissolved gas in transformer oil and its sensitive reflect the fault degree vary; So directly to the characteristics of the component gases neural network data input, will lost some small volume contains the key information [13, 14]. At the same time, if the source of BPNN model concentrated too discrete sample data will lead to the neural network convergence difficulties, so the concept of cumulative frequency of the data of gas dissolved in transformer oil are normalized processing, the process is:

(1) collecting and confirmed by the verification, and the conclusion is clear transformer accident before chromatography detection data of the training sample set.

(2) In view of the seven characteristics of gas (H2, CH4, C2H6, C2H4, C2H2, CO, CO2) content size sorting, respectively, and size are grouped according to its content.

(3) Statistics of seven characteristics of gas in the frequency of each packet rj and frequency wj = rj/n, where n is number of observations for gas.

(4) Using the concept of cumulative frequency calculation after each group of data, such as in the first set of data the cumulative frequency I to

$$F_i = \sum_{j=1}^{i} w_j = \sum_{j=1}^{i} \frac{\Delta r_j}{n} = \frac{r_i}{n}.$$

Of them by the end of the first group I to the ri cumulative frequency. The following will have to calculate the cumulative frequency value of Fi instead of the group characteristics of the gas content as the input of the neural network.

(5) By the training sample set of the seven characteristics of the gas content is replaced by their corresponding cumulative frequency, of the original sample set.

Using this method in 699 cases of chromatography test sample data normalization processing, H2 and cumulative frequency change relations as shown in figure 4.

Zhang Wei, Wu Qiu-li, Deng Yu-rong, Lv Ze-cheng
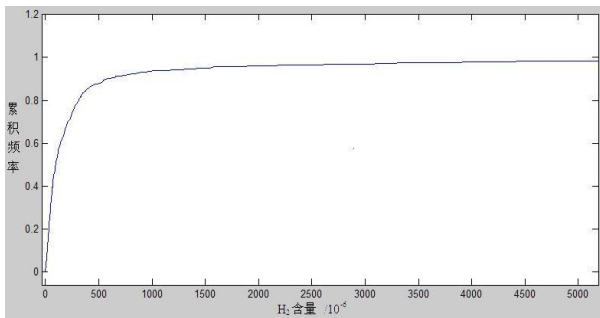
FIGURE 4 the relationship between H2 content and cumulative frequency

2.3 DATA SCREENING CLASSIFICATION

Because it is difficult to describe professional disciplines in a precise numerical definition and complex expert experience a lot of words in [15]; Examples so as to ensure that the reasoning can effectively identify the fault type, a source of case base training sample should satisfy certain boundary conditions or constraints, namely has typicality and differences. Identification model of typicality to guarantee training effectively identify the unknown sample data; Difference is to reduce the redundancy of the training database.

Define a generic function is the key to the typical samples selected, N training samples respectively under K fault type of solution, the evaluation of the individual sample generic function can be represented as:

$$F_{class} = \max(M_1/l, M_2/l, \cdots M_i/l \cdots M_k/l) . \qquad (2)$$

Type of l for the nearest distance is evaluating individuals training sample, the amount can be set according to the size of the existing data artificially, I said the failure mode of the subscript, Mi is l study samples belong to the first class I fault sample.

Visible, according to the evaluated samples in specific degree of fault type of the relative merits of the calculation results, can be assessed individual sample belongs to the barrier type of possibility. Therefore, a generic function values can be as the measure is to assess individuals sample representativeness index; In all of the source case base sample after a generic function value calculation, preferable one of the biggest generic function values as a representative of the combination of the training sample data sample.

On the basis of the typical filter, to eliminate differences in borderline example of small samples. The measure of diversity index, through correlation analysis to solve the phase relationship between numerical implementation. Numerical calculation of phase relationship can be represented as:

$$r = \sum_{i=1}^{7}(x_i - \bar{x})(y_i - \bar{y}) / [(\sum_{i=1}^{7}(x_i - \bar{x})^2) \times (\sum_{i=1}^{7}(y_i - \bar{y})^2)]^{0.5}$$

Type in the x, y, respectively in the oil dissolved gas analysis results of two different training sample; Whereas, respectively mean {xi} and {yi}.

After the normalization processing of the original sample set after screening, available capacity is not big but have typical source example of training sample set.

2.4 NETWORK TRAINING PROCESS

According to the classification of fault type recognition process (figure 2), in the second classification in overheating fault and discharge two types, and each fault type contains two branch fault types. Therefore, the second neural network training, the need to build corresponding overheating fault, discharge two branch network model. At the same time, by back propagation to constantly adjust the network weights and threshold, reduce the error sum of squares of the network. Network model training and the corresponding fault type source case base.

Network training input layer data for gas concentration data after normalization processing. If conditions permit, the transformer can be electrical test data as input parameters, also facilitate neural network identification. The network training process as shown in figure 5.
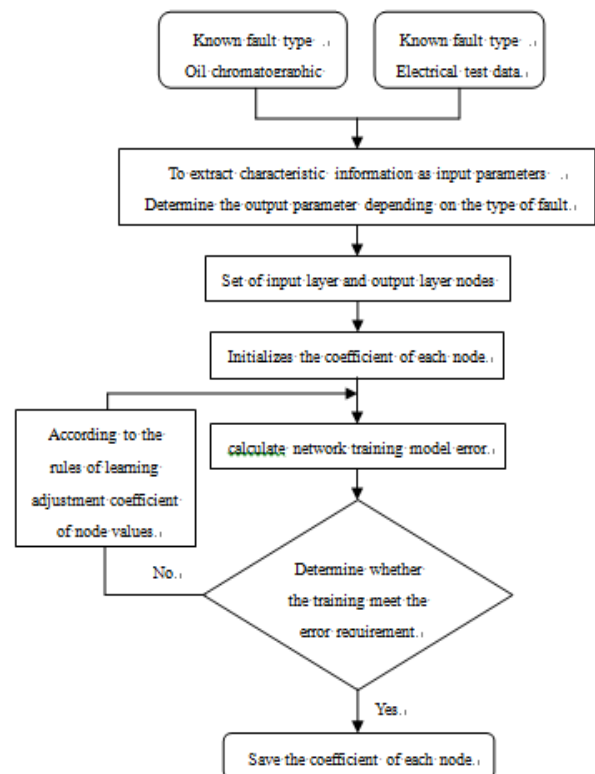
FIGURE 5 the training process of neural network

Will get the network node coefficient were input to the training BPNN (figure 4) nodes in each layer, can establish the corresponding fault type recognition model. Its output is type matching similarity, paradigms and specific fault types of target is obtained matching

79

Zhang Wei, Wu Qiu-li, Deng Yu-rong, Lv Ze-cheng

similarity, the higher the similarity is to determine the accuracy of the higher. If sample retrieval using a variety of retrieval way, all kinds of retrieval methods can get a classification as a result, the final target is obtained by using the weighted average of the sample and the failure types of comprehensive matching similarity, ensure that the minimum error sum of squares of the network. The comprehensive evaluation index as:

$$syn = \sum Rig_{ij} \times Sim_{ik} / 2 . \tag{3}$$

Type, Rigij for a large number of failure data (DGA) calculation for each sample retrieval algorithm for each source sample correct classification rate index of the search space, it is using the known failure data to judge each classification retrieval algorithm is sentenced to rate, Simik is the i retrieval algorithm retrievals

4 The example analysis

A 110 kv good feng varying oil chromatographic data as the target of the 2012-11-23 07:48 time paradigm, monitoring data as shown in table 1,

TABLE 1 Oil chromatographic monitoring data of liangfeng station

Monitoring time	H2	CH4	C2H2	C2H4	C2H6	CO
2012.11.23 07:48:00	19.87	13.75	1.23	45.67	4.11	3016.26

Step 1: sample retrieval

Known from the analysis of three ratio encoding C2H2 / C2H4, CH4 / H2, C2H4, C2H6 corresponding code of 0, 0, 2. Therefore, through the example to retrieve the first level classification target sample can be classified as overheating fault.

Sample retrieval in the second classification and the third level is used to identify the BPNN network in the classification, so need to source and examples of typical network training samples, to obtain network node coefficient. Source at the same time, the case base from 699 cases after selecting the type of known fault cases of original samples, so the second layer neural network training samples for 383 cases of thermal fault samples. Among them, the thermal circuit fault samples of 171 cases of magnetic circuit thermal failure of 212 cases.

The source case base in the magnetic circuit fault and heat circuit fault data samples respectively do normalization processing and typical sample data filtering, with typical samples as training BPNN network learning samples. Figure 7 for the gas concentration and the mapping relationship between cumulative frequency, by the mapping relationship between the sample data normalization.
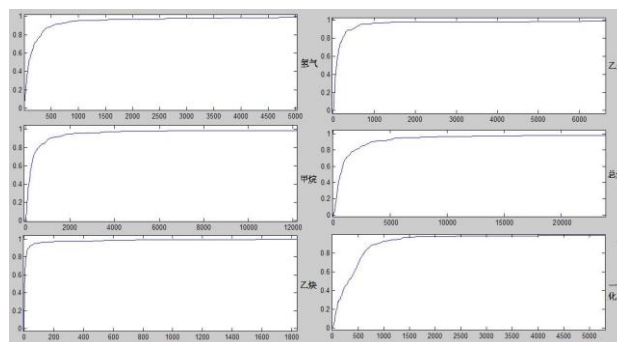
FIGURE 6 The mapping relationship between gas content and cumulative frequency

Data normalization typicality after screening, screening after 50 cases received circuit fault and 50 cases of magnetic circuit hot fault. The second layer network training input data, a total of 100 groups, namely the $X_i=\{x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}, x_{i6}, x_{i7}\}$, $Y_i=\{y_{i1}, y_{i2}\}$, $i \in [1,100]$.

Type in the Xi for the normalized oil chromatographic data, Yi for fault classification results (i.e. $y_{i1}=1$, $y_{i2}=0$ means heat circuit fault; $y_{i1}=0$, $y_{i2}=1$said magnetic circuit thermal failure), the subscript i said the first group of data, I said the size of the y value and the fault type of compatibility. Network model are obtained by training, identification of target model, the calculation results as follows: $y_{11}=0.999999439228727$, $y_{12}=3.30988595178941*10^{-7}$.

Similarity matching with heat circuit fault close to 100%, thus can be considered as heat circuit malfunction.

Carried out in accordance with the same method in the third layer neural network recognition, the result is: $y_{11}=0.99999980962573$, $y_{12}=4.57039595983834*10^{-5}$, $y_{13}=1.92874984796392*10^{-22}$.

Similarity matching with tap-changer fault also close to 100%, thus can be considered a fault tap-changer.

Step 2: source sample ordering, the preferred choice of fault type

According to the sample, matching similarity retrieval failure is calculated, sample selection and source library failure types match the highest failure type as the target sample identification results, namely tap-changer of failure.

Step 3: keep sample.

Examples given recognition as a result, the target as a source of known fault types to keep new paradigm case base, namely source similar fault samples increase 1 case in case base.

5 Conclusions

Case reasoning is a kind of artificial intelligence methods based on prior knowledge reasoning, it according to the problem requirements or features, matching sample source, and get the optimal original sample under its guidance. By comparing the fault identification probability, select from course to fine hierarchical structure, so establish the diagnosis mode must give full

Zhang Wei, Wu Qiu-li, Deng Yu-rong, Lv Ze-cheng

consideration to the edge the effect on the stability of the sample of the model.

In this paper, the adaptive adjustment algorithm have great inhibition effect to solve unstructured edge data influence. At the same time to the associated classification pattern extraction and classification of input vector, the adaptive adjustment of vector optimization is also has strong flexibility, makes the establishment of the diagnosis model is more accurate. From the hierarchical diagnosis, the case of adaptive adjustment vector reasoning algorithm achieved ideal result.

References

[1] Wang Jianyuan, Ji YanChao 2003 Fuzzy Petri network knowledge representation method and application in transformer fault diagnosis *Proceedings of the CSEE* **23**(1) 121-5

[2] Wang Nan, L Fangcheng, Liu Yunpeng, etc. 2003 Based on rough set theory and fuzzy Petri network comprehensive oil-immersed power transformer fault diagnosis *Proceedings of the CSEE* **23**(12) 128-32

[3] Peng Jjian, Luo An, Zhou Ke, etc. 2007 The application of information fusion technology in the transformer fault diagnosis *High voltage technology* (3) 144-7

[4] Shang Yong, Rui Chunjiang, Yan Zhang, etc. 2002 Based on information fusion of large oil-immersed power transformer fault diagnosis *Proceedings of the CSEE* **22**(7) 115-8

[5] Sun Hui, Li Weidong, Sun Qizhong 2001 Decision tree method is used for transformer fault diagnosis research *Proceedings of the CSEE* **2**(2) 50-5

[6] Wang Yingying, Luo Yi, Tu GuangYu 2008 Distribution network fault diagnosis based on rough set and decision tree method *High voltage technology* (4) 794-8

[7] Mo Juan, Zhang Yan, Li Hua, etc. 2005 Comprehensive transformer fault diagnosis model based on multiple intelligence method *Automation of electric power systems* **29**(18) 85-9

[8] Dong Xiucheng, Tao JiaGui, Wang Haibin, etc. 2010 The adaptive fuzzy support vector machine increment algorithm in the application of transformer fault diagnosis *Electric power automation equipment* **30**(11) 48-52

[9] Li Hua, Mo Juan, Zhang Rui, etc. 2005 The transformer insulation fault diagnosis system based on Web mode *High voltage technology* **31**(6) 12-4

[10] Shi Haojie, Xing QingHua, Liu Fuxian, etc. 2009 Based on the introduction of probability sample matching new method *Computer engineering and application* (10) 62-4

[11] Chen Jiangbo, Wen XiShan, Lan Lei, etc. 2007 Based on a new radial basis function networks *Journal of transformer fault diagnosis of high voltage technology* (3) 140-3

[12] Zhang Quanming, Liu Huijin 2008 Least squares support vector machine in the application of power quality disturbance classification *Proceedings of the CSEE* **28**(1) 106-10

[13] Zhang Weizheng, Wang Xiaoming, Wu Xiaohui, etc. 2007 Combination of neural network in the transformer fault diagnosis application *High voltage electrical appliances. Lancet* (5) 364-7

[14] Lan Hua, Zheng Guangrui, Liao Zhimin *Based on the fuzzy TOPSIS and the BP network* 15:26 2014/4/13

Authors

Wei Zhang

Current position, grades: Senior Engineer
University studies: High voltage technology
Scientific interests: Online monitoring
Publications: none
Experience: Be engaged in On-line monitoring of electric equipment of radio frequency identification technology, Excellence at warning centre maintenance management, condition evaluation, reliability analysis.

Qiu-li Wu

Current position, grades: Engineer
University studies: High voltage technology
Scientific interests: Online monitoring
Publications: none
Experience: Be engaged in On-line monitoring of electric equipment of radio frequency identification technology, Excellence at warning centre maintenance management, condition evaluation, reliability analysis.

Yu-rong Deng

Current position, grades: Professor level senior engineer
University studies: High voltage technology
Scientific interests: Online monitoring
Publications: none
Experience: Be engaged in On-line monitoring of electric equipment of radio frequency identification technology, Excellence at warning centre maintenance management, condition evaluation, reliability analysis.

Ze-cheng Lv

Current position, grades: Senior Engineer
University studies: High voltage technology
Scientific interests: Online monitoring
Publications: none
Experience: Be engaged in On-line monitoring of electric equipment of radio frequency identification technology, Excellence at warning centre maintenance management, condition evaluation, reliability analysis.

The model construction and implementation of discrete physical system in industrial CPS

Huaping Zhou¹*, Xutong Zhang²

¹*Faculty of Computer Science & Engineering Anhui University of Science and Technology, Huainan, China*

²*Faculty of Computer Science & Engineering Anhui University of Science and Technology, Huainan, China*

Abstract

Cyber physical system referred as CPS, or information physical system, achieves the integrated coordination of the cyber world and the physical world. The collaborative system is more reliable and efficient. By analyzing the cyber physical system of integrated sense and control, the constructing approach of the discrete physical model is proposed and applied to mine surface production systems for CPS of the combination of discrete-time input and output and control functions. By modelling, analyzing and verifying the system, the cyber system and the physical system achieve deep fusion. On this basis, trusted software design, formal description methods and reasoning theory are established in the cyber physical system. It provides the authentication methods for the trusted software analysis and modelling of discrete physical system in the cyber physical system. The coal surface production system model is established by the model construction method of discrete physical systems in CPS proposed in this paper. The technology of computer, computer network and embedded systems is adopted to build the information world and achieve the integrated design and collaborative control of the subsystem in CPS. It is applied in a coal mine in the Huaibei Coal Mine Shares Limited Company and achieves good results. To speed up the wide applications to industries and enterprises of the CPS, a useful exploration is carried out in the paper.

Keywords: information world, physical world, discrete physical system, information fusion, credible design

1 Introduction

Cyber physical system is a complex system integrated the information processing technology and physical environment. It is an integrated system of the physical process and the calculation process. Cyber physical system achieves wisdom perception, real-time control and interactive services in the industrial and mining engineering system through the further integration of computing, communications and controlling. Information physical fusion system achieves the integrated design of the information world and the physical world and makes the cooperative system more reliable, efficient [1]. In recent years, information physical fusion system has become an important direction of research and development in the academic and scientific fields, has also become a priority development industry in industrial mining field. It is of great significance for accelerating China's industrialization and information fusion to carry out the research and application of physical information fusion system [2].

Due to security and other requirements, application technology of coal mine enterprises is very advanced in our country. Accordingly, information physical fusion system is more suitable for large-scale enterprises such as mine. It has multiple productions, transportation and other systems. Each system includes sensors, actuators, control unit, etc. and has many parameters detected and controlled and has also many corresponding system status parameters. Thus, mine information physical fusion system is very large. It is difficult to analyse it from the whole system. As we all know, any a complex system, is a combination of multiple simple systems [3]. Mine information physical fusion systems is also true. Therefore, build and combine the coal production system, transportation system and each other subsystems to complete the entire mine information physical fusion system [4].

In this paper, on the base of the analysis of construction method of information physical fusion system model having the discrete-time input, output combination and control, select a ground mine production system as an example. Through an analysis of the modelling method and verification method of the physical system, fuse it with the information world and establish a kind of software credible design method, research method, formal description method and reasoning theory for the coal information physical fusion system. It provides mathematical models and verification means for trusted software system analysis and model study of the coal information physical fusion system and constructs the software physical theories and methods for the application of information to ensure that information physical fusion system fidelity and faithful implementation. It utilizes industrial control computer, embedded system and computer network to control the

---

ground production system of a mine in Huaibei Coal Mine Company.

2 Discrete physical system model building approach in CPS

In industrial and mine enterprises, so many physical systems have the integration function with discrete time input and output combination, sensing and controlling.

As a discrete event occurs in an instant, rather than continuously changes over time. Thus, the signal in discrete system has a functional form as follows: signals: R {absent, present}, t∈R [5].

For discrete physical system, we use a five element finite state to represent it as follows.

(States, Inputs, Outputs, Updates/Possible updates, Initial State) of which:

States: a finite set of the stations in the system.

Inputs: input signal set in the system.

If a discrete system has N input ports P = {P1, P2, ... ,PN}, and each port p∈P has a corresponding type Vp, the system input signal set Inputi: P VpiU…..UVpNU{absent}, for each p∈P, Inputi(P)∈VpU{absent}.

Outputs: system output value set.

Updates: an upgrade function, thought it a state and an input value are mapped to the next state and the output value set. States × Inputs States × Outputs, Possible updates map the state and an input value to the next possible state and the output value set. Namely, States×Inputs 2States×Outputs.

Assume S: N States, is a function. at the same time, assume S(0)= initialState0,(x0,y0) to be initial input and output. The system trajectories (x, si, yi) (i=0, 1, 2, 3…) including the state is all the state trajectories tracked of the discrete system [6-7].

In the case of the input x: N Inputs, status S: N States, output y: N Outputs, get the discrete-time system model:

$$(S(n+1), y(n)) = \\ Update(S(n), x(n)) / Possible\ Update(S(n), x(n))$$ (1)

3 Mine ground production system components, operation mode and function

Mine ground production system is composed of a multi-belt conveyor, multiple coal feeders, vibrating screen, pumps and other components. It is a kind of key transport equipment in the course of coal production. Generally, it is set a few hundred meters in length. In addition to maintenance, one day it runs for at least 22 hours. It is the vital lifeline of mine. Normal production of the entire coal mine is related to safe operation.

The main function of surface production system is to control the four coal feeders, vibrating screen, the main belt, hand-selected belt, something blending belt, floor belt, back to the coal belt and water pump. The system does not affect the original manual control operation. It

realizes remote monitoring through information physical fusion system and the computers in the dispatching control center. The remote computers real-time display the motion state of coal feeders, vibrating screen, the main belt, hand-selected belt, something blending belt, floor belt, back to the coal belt, water pump, can achieve the remote control.

The system's control is in three ways, respectively: linkage mode, independent manner and maintenance mode. At the same time, there are emergency stop and a variety of protection function in the system.

A. Linkage Mode

By linkage mode, computer operation can start in sequence No.1,2 coal feeder, vibrating screen, the main belt, hand-selected belt, something blending belt, belt landing. Startup sequence is: 1) Start stuff blending belts, floor belt; 2) start the main belt; 3) Start the belt hand-selected and shaker; 4) North and South coal feeder. There is a ten second delay between each step. And before starting, detect the situation of heap coal and motor current. Start operation will fail if heap coal and overcurrent occurs.

By the way, of linkage mode, No. 1, 2 coal feeder, vibrating screen, the main belt, hand-selected leather belt, something blending belt, belt landing can be stopped in sequence. Stop order is: 1) Stop the coal feeder; 2) Stop the belt hand-selected and shaker; 3) Stop the main belt; 4) Stop things blending belts, floor belt. There is a ten seconds delay between each step.

When needed, add the control to back to the coal belt, No.1, 2 floor coal feeder. Back to the coal belt, No.1, 2 floor coal feeder, pumps, etc. can also be controlled independently.

B. Independent Manner

By the way of independent manner, console and computer can both control the start and stop of a single device, and can be any combination of the start and stop, in other words, console controls the start-up. Computer controls hut-down. Computer controls start-up. Console controls hut-down. Computer controls start-up and hut-down.

By this manner, two coal feeders, vibrating screen, the main belt, hand-selected belt, something blending belt, the floor belt, back coal belt, water pump and North and South floor coal feeder are controlled by the computer independently. Before starting, detect the situation of heap coal and motor current. Start operation will fail if heap coal and overcurrent occurs.

C. Maintenance Mode

By maintenance mode, operation and independent control is basically the same way. The difference is that heap coal and motor overcurrent condition need not be detected before driving by the way of maintenance mode. If heap coal or overcurrent occurs, the operation also can be carried out and drive.

D. Emergency Stop

If system accidents happen in running state, press the "stop" button (emergency stop) (It lies in both on-site and

computer control interface of control center). At this time no matter what state each equipment of the system is, will be forcibly stopped. System immediately stops.

E. System Protection Function

Ground production system has function of heap coal protection. When coal accumulation exceeds a certain height, the system sends heap coal signal, and system is stopped and processed. The system is with overcurrent protection. When the motor current exceeds the set value, the system sends an overcurrent signal and stops to reliably protect the motor.

System also has the function of belt deviation protection, interlock protection, back-tested protection, blocking protection, under voltage protection.

Ground production system structure is shown in Figure 1.

FIGURE 1 Ground production system

4 Model construction of ground production system model

Through analysis of the ground production system, get its input, output, and status signals as follows:

1) Input signal
Independent linkage control I0
Emergency stop (stop) I1
Overhaul state detection I2
Overcurrent detection I3
Heap coal state detection I4
Main belt running monitoring I5
Coal gangue belt running monitoring I6
No.1 coal feeder running monitoring I7
No.2 coal feeder running monitoring I8
No.1 coal blending belt running monitoring I9
No.2 coal blending belt running condition monitoring I10
Floor belt running monitoring I11
Clean water pumps running detection I12
Sewage pumps running state detection I13
No.1 return coal feeder running monitoring I14
No.2 return coal feeder running monitoring I15
Vibrating screens running monitoring I16
Vibrating screens with or without coal monitoring I17
Stop I18
Bunker full of coal detection I19
Back coal belt running detection I20
Main belt start signal detection I25
Coal gangue belt start signal detection I26

No.1 coal feeder start signal detection I27
No.2 coal feeder start signal detection I28
No.1coal blending belt start signal detection I29
No.2 coal blending belt start signal detection I30
Floor Belt start signal detection I31
Clean water pumps start signal detection I32
Sewage pumps start signal detection I33
No.1 back coal feeder start signal detection I34
No.2 back coal feeder start signal detection I35
Vibrating screens start signal detection I36
Back coal belt start signal detection I40
2) Output Signal
Main belt output control Q0
Coal gangue belt output control Q1
Vibrating screens output control Q2
No.1 coal feeder output control Q3
No.2 coal feeder output control Q4
No.1 coal blending belt output control Q5
No.2 coal blending belt output control Q6
Floor Belt Output Control Q7
No.1 return coal feeder output control Q8
No.2 return coal feeder output control Q9
Clean water pumps output control Q10
Sewage pumps output control Q11
Return coal belt Q12 output control
3) Status Signal
S0, S1, S2, S3, S4, S5, S6-1, s6-2, S7, S8-1, S8-2
Therefore,
Input signals= {I0, I1… I40}
Output signals= {Q0, Q1…Q12}
State signals= {Ss, S0, S1…S8-2, Se}

Based on the system composition, operation principle and technological process get the running state transformation diagram shown in figure 2.

The figure 2 shows that in the initial state press the linkage Run button, system will run by the way of computer remote control and the state changes from Ss to state S0. Under the state S0, if the system meets the condition of the linkage mode, not pressing emergency stop (stop) button, not in the way of maintenance mode, not detecting an overcurrent condition and heap coal conditions and other conditions, the system will enter the S1 state after the 10-second delay. The system will also change to the S1 state in the case of not the linkage mode, detecting No. 1 coal blending belt start signal and the NO.2 coal blending belt start signal. After the system changes to the S1 state, start No.1 and No.2 blending coal belt. And According to system operating mode and p process, turn into the corresponding state and control the corresponding controlled object.
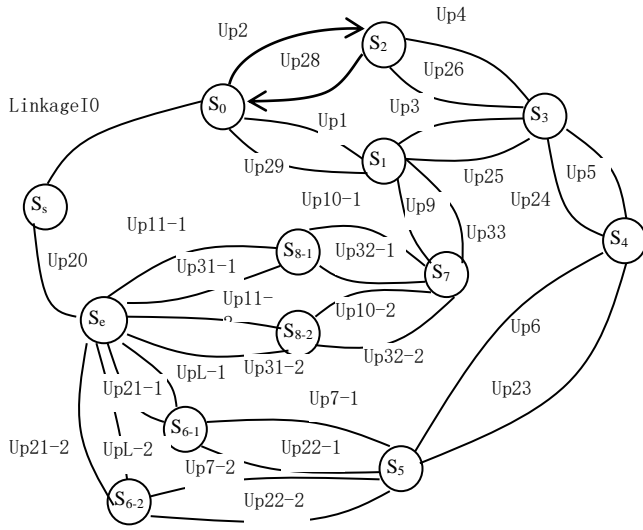
FIGURE 2 Transformation state of system

Thereby the system state transition relations and output control model are as follows:

$$Updates1\_inputs = [I0(not)I1(not)I2(not)$$
$$I3(not)I4(not)I19F(t+10)]$$
$$[(not)I0I29I30](S0,\ Updates1\_inputs)$$
$$-->(S1,Q5,Q6)$$
(2)

$$Updates2\_inputs =$$
$$[I0(not)I1(not)I2(not)$$
$$I3(not)I4(not)I19F(t+10)]$$
$$[(not)I0I31](S0,\ Updates2\_inputs)$$
$$-->(S2,Q7)$$
(3)

$$Updates3\_inputs = [I0(not)I1(not)I2(not)$$
$$I3(not)I4F(t+10)I9I10][(not)I0I25]$$
$$Updates4\_inputs = [I0(not)I1(not)$$
$$I2(not)I3(not)I4F(t+11)I10]$$
$$[(not)I0I25](S1,\ Updates3\_inputs)$$
$$(S2,\ Updates4\_inputs)-->(S3,Q0)$$
(4)

$$Updates5\_inputs = [I0(not)I1(not)I2(not)$$
$$I3(not)I4(not)I19F(t+10)I5]$$
$$[(not)I0I26](S3,\ Updates5\_inputs)$$
$$-->(S4,Q1)$$
(5)

$$Updates6\_inputs =$$
$$[I0(not)I1(not)I2(not)I3(not)$$
$$I4(not)I19F(t+10)I6][(not)I0I36]$$
$$(S4,\ Updates6\_inputs)-->(S5,Q2)$$
(6)

$$Updates7-1\_inputs =$$
$$[I0(not)I1(not)I2(not)$$
$$I3(not)I4(not)I1F(t+10)I6]$$
$$[(not)I0I34](S5,\ Updates7\_inputs)$$
$$-->(S6,Q3)$$
(7)

$$Updates7-2\_inputs =$$
$$[I0(not)I1(not)I2(not)$$
$$I3(not)I4(not)I19F(t+10)I6]$$
$$[(not)I0I35](S5,\ Updates7\_inputs)$$
$$-->(S6,Q4)$$
(8)

Updates8-1_inputs=const
Updates8-2_inputs=const

$$(S6-1,\ Updates8-1\_inputs)$$
$$(S6-2,\ Updates8-2\_inputs)-->(Se,none)$$
(9)

$$Updates9\_inputs =$$
$$[I0(not)I1(not)I2(not)$$
$$I3(not)I4(not)I19F(t+10)(not)I1I10]$$
$$[(not)I0I40](S1,\ Updates9\_inputs)$$
$$-->(S7,Q12)$$
(10)

$$Updates10-1\_inputs =$$
$$[I0(not)I1(not)I2(not)$$
$$I3(not)I4(not)I19F(t+10)(not)I1]$$
$$[(not)I0I34](S7,\ Updates10-1\_inputs)$$
$$-->(S8,Q8)$$
(11)

$$Updates10-2\_inputs =$$
$$[I0(not)I1(not)I2(not)$$
$$I3(not)I4(not)I19F(t+10)(not)I17]I20$$
$$[(not)I0I35](S7,\ Updates10-2\_inputs)$$
$$-->(S8,\ Q9)$$
(12)

Updates11-1_inputs=const
Updates11-2_inputs=const

$$(S8-1,\ Updates11-1\_inputs)$$
$$(S8-1,\ Updates11-2\_inputs)$$
$$-->(Se,none)$$
(13)

$$Updates20\_inputs =$$
$$I0I18(Ss,\ Updates20\_inputs)-->(Se,none)$$
(14)

$Updates21-1\_inputs =$
$[I0I18(not)I17F(t+10)]I1$
$[(not)I0(not)I27](Se,\ Updates21-1\_inputs)$ ，$\qquad$ (15)
$—> \big(S6-1,(not)Q3\big)$

$Updates21-2\_inputs =$
$[I0I18(not)I17F(t+10)]I1$
$[(not)I0(not)I28](Se,\ Updates21-2\_inputs)$ ，$\qquad$ (16)
$—> \big(S6-2,(not)Q4\big)$

$Updates22-1\_inputs =$
$[I0I18(not)I7F(t+10)]I1[(not)I0(not)I36]$
$Updates22-2\_inputs =[I0I18(not)$
$I8F(t+10)]I1[(not)I0(not)I36]$ ，$\qquad$ (17)
$\big(S6-1,\ Updates22-1\_inputs\big)$
$\big(S6-2,\ Updates22-2\_inputs\big)$
$—> \big(S5,(not)Q2\big)$

$Updates23\_inputs =$
$[I0I18(not)I16F(t+10)]I1$
$[(not)I0(not)I26](S5,\ Updates23\_inputs)$ ，$\qquad$ (18)
$—> \big(S4,(not)Q1\big)$

$Updates24\_inputs =$
$[I0I18(not)I6F(t+10)]I1$
$[(not)I0(not)I25](S4,\ Updates24\_inputs)$ ，$\qquad$ (19)
$—> \big(S3,(not)Q0\big)$

$Updates25\_inputs =$
$[I0I18(not)I5F(t+10)]I1$
$[(not)I0(not)I9(not)I20]$
$Updates31-1\_inputs =$
$[I0I18F(t+10)]I1[(not)I0(not)I34]$ ，$\qquad$ (20)
$\big(Se,\ Updates31-1\_inputs\big)$
$—> \big(S8-1,(not)Q8\big)$

$Updates31-2\_inputs =$
$[I0I18F(t+10)]I1[(not)I0(not)I35]$
$\big(Se,\ Updates31-2\_inputs\big)$ ，$\qquad$ (21)
$—> \big(S8-2,(not)Q9\big)$

$Updates32-1\_inputs =$
$[I0I18F(t+10)(not)I14]I1$
$[(not)I0(not)I40](Se,\ Updates31-1\_inputs)$ ，$\qquad$ (22)
$—> \big(S8-1,(not)Q8\big)$

$Updates32-2\_inputs =$
$[I0I18F(t+10)(not)I15]I1$
$[(not)I0(not)I40]\ \big(S8-1,\ Updates32-1\_inputs\big)$ ，(23)
$\big(S8-1,\ Updates32-1\_inputs\big)$
$—> \big(S7,(not)Q12\big)$

$Updates33\_inputs =$
$[I0I18F(t+10)I1$
$[(not)I0(not)I9(not)I10]$
$\big(S3,\ Updates25\_inputs\big)\big(S7,\ Updates33\_inputs\big)$ ，(24)
$—> \big(S1,(not)Q5,(not)Q6\big)$

$Updates26\_inputs =$
$[I0I18F(t+10)(not)I5]I1$
$[(not)I0(not)I11]\big(S3,\ Updates26\_inputs\big)$ ，$\qquad$ (25)
$—> \big(S2,\ (not)Q7\big)$

Updates27_inputs=const
Updates28_inputs=const

$\big(S1,\ Updates27\_inputs\big)$
$\big(S2,\ Updates28\_inputs\big)--> \big(S0,\ none\big)$ ．$\qquad$ (26)

The above model is easy to realize through the remote computer and embedded microcontroller processor.

5 Analysis of the reachability and model verification of the system state

A. Linear Temporal Logic

For the accessibility and accuracy verification of the above system model, we introduce the linear temporal logic to analyse it. Linear temporal logic is able to clearly express the occurrence of an event and its properties, the causal relationship between events and their sequence [8]. First we introduce the execution trace Et. Execution trace is a sequence. It reflects the execution process, which is expressed as: Et_0, Et1, Et2 …

Among them, $Et_i = (x_i, s_i, y_i)$. X_i, s_i and y_i respectively represent input, state, and the output value of the ith.

A linear temporal logic formula applies to all execution traces Et_0, Et_1, Et_2...of the whole system [9]. We use Fm_0, Fm_1, Fm_2 to represent the linear temporal logic formula and use Pp_0, Pp_1 and Pp_2, etc. to represent the thesis in the model.

For the state model Ms and the corresponding linear temporal logic formula Fm, if Fm is right for all the possible execution trace Et of the system model, Fm is right for the Ms.

To better reasoning out the whole trace of a state model of, then we introduce some temporal operator in the linear temporal logic.

Global operator Gl (Fm)

If the linear temporal logic formula Fm is true for a track and all the following tracks, then the property Gl (Fm) is true for this track.

Final operator Fi (Fm)

If the linear temporal logic formula Fm is true for a following track, then the property Gl (Fm) is true for this track.

Later state operator Nx (Fm)

If the linear temporal logic formula Fm is true for the path Et0, Et1, Et2 ... in state model, then the property Nx (Fm) is true for this track.

Until operator (Fm1) Un (Fm2)

If the linear temporal logic formula Fm1 is true for a track after a track and Fm1 maintains until Fm2 is true, then the property (Fm1) Un (Fm2) is true for this track.

With the above definition, we can have an analysis of an infinite number of events of state model Gl (Fi) (Fm), state stability Fi (Gl (Fm)) and request response properties Gl (Pp=>Fi (Et)) and so on to.

B. Accessibility Analysis

To verify and analyse this system is to verify system model and the enclosed model composed of the environment model of the system [10-11]. Therefore, the verification process has three input signals. They are respectively validation system model S, environment model E and the properties Fm to be verified. The validator will create a Yes/No output result. Model validation process is shown in Figure 3.
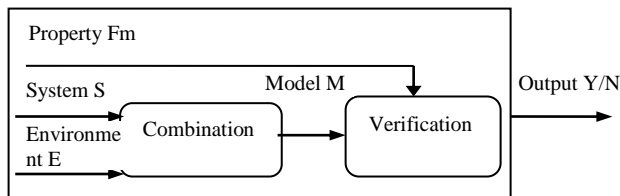
FIGURE 3 Model validation process

Through the analysis of the ground production system environment, the ground production system environment model is shown in figure 4. The environmental information includes mainly man-machine dialogue, hoisting system, transportation system, storage system etc.. Due to limited space, the paper has no its in-depth analysis.

The relationship between the ground production system environment model and the ground production system model is shown in figure 5.

For a state model Ms and the attribute Fm is the linear temporal logic formula in the form of Gl (Pp) (Pp is a proposition). When p is true for each state in a track, Gl (Pp) is a linear temporal logic formula in this trajectory.

To make the system state models Ms meet it, each listed track of the system state model Ms must meet the Gl (Pp).

FIGURE 4. Environment model of ground production system.

Next we prove the accessibility of surface production system model through enumeration method.
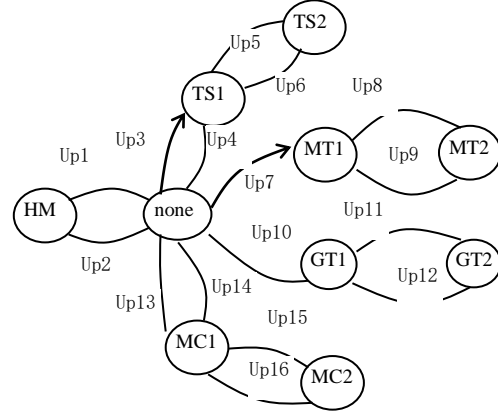
FIGURE 4 Environment model of ground production system.

To be verified closed system Mm composed of the system model Ms and environment model Me are an uncertain system with finite state and no input. The next state sets are only related to the current state. This relationship is described by the function denoted by Tr. So Tr is the next set of possible states of the state s of Mm.
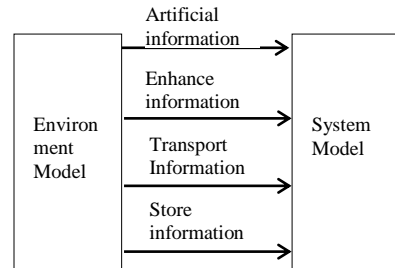
FIGURE 5 The ground production system environment model and the ground production system model

Adopting depth-first traversal, by constantly using Tr to dynamically generate state diagram for the access state. Computing the state set that Mm can achieve. The algorithm is as follows:

```
    Input
    Output
    Initialize
    DepthFirstTraversal()
    {
    while Stack Sall is not empty do
pop the state s at the top of Sall
compute Tr(s)
for each ss ∈ Tr(s) do
            if ss not ∈ R then
                    R:=R ∪ {ss}
                    Push ss onto Sall
                    DepthFirstTraversal()
```

End

    End

  End

}

Program calls the algorithm and completes, the set R is all accessible state set of Mm.

The storage space and running time of the above algorithm is proportional to the state number of the Mm. The number of state in the system Mm is index times of the size of the set Ms and set Me. For the system (figure), it is mainly the discrete system, if we take time t as the change quantity and seconds as unit to calculate the number of state, the Ms alone has 121 Boolean state variables, coupled with nine of Me, the Mm has 130 Boolean state variables total and 2130 states. So, it has a large amount of data far more than the general computer storage capacity. For this, the above algorithm need be improved.

The above algorithm detects the accessibility for each individual state which produces a large amount of data. If we adopt the special data structures to test a state set only, it can greatly reduce the calculation. The improved algorithm is as follows:

```
Input
Output
Initialize
StateSetSearch ()
{
Rnew=R
while Rnew != 0 do
    Rnew = {s| ∃ss ∈ Rs.t.s ∈ Tr( ss )}
    R: =R ∪ Rnew
end
}
```

In the algorithm StateSetSearch, the set R is the whole set of all the state which can reach and be searched. Rnew represents the new state generated by those points. When there is no new status, the algorithm stop. At this time, all of the states which can reach from S0 are put in the set R.

Use algorithm StateSetSearch to analyse the reachability of the ground production system model, get the reachability state set. The correctness of the model is verified by the set.

6 Conclusion

Information physical system is not a simple combination of information system and physical system, but the all-round and multi-level depth fusion and integration. It is a kind of self-sensing system. It is a kind of new system, which consists of a large number of sensing devices, computing devices and communication networks. These characteristics make it more suitable for the production and management of industrial and mining enterprises. Information physical system is huge and it is very huge for the information physical system of an industrial and mining enterprise. It contains a number of technology and rich resources, which has not been well studied and applied. In the course of researching and developing the ground production monitoring system in coal mine, the author try to adopt the integration idea, information physical system, to achieve wisdom awareness, real-time control and interactive services. As an actual subsystem of information physical system applied in the field of industrial and mining enterprises, this article has an analysis of the model building method of the common discrete physical system in information physical systems in industrial and mining enterprises firstly. Then, according to the input, output, and the technology in the mine surface production system the system model is constructed. In order to ensure the accessibility and the correctness of the model, the author has established the discrete physical system model, its reliable software design, the formal description method and the reasoning theory in information physical fusion system. It provides validation method for the reliable software analysis and modelling in the discrete physical system of information physical fusion system. The study will accelerate the development of industrial and mining enterprises information physical system. The next step we will put more system in it.

Lastly, the mine surface production system model is established by the model building method of discrete physical system in the industrial CPS proposed in the paper. And computer network and embedded system are used to construct the information world and implement the CPS subsystem design and the cooperative control as an organic whole. The system has been applied to Yang Zhuang coal mine in Huaibei Coal Mining Shares Limited Liability Company. Good results have been achieved.

References

[1] Sazia Parvin, Farookh Khadeer Hussain, Omar Khadeer Hussain, etc 2013 Multi-cyber framework for availability enhancement of cyber physical systems *Computing* **95**(10-11) 927-48

[2] Grit Denker, Nikil Dutt, Sharad Mehrotra, Mark-Oliver Stehr, Carolyn Talcott, Nalini Venkatasubramanian 2012 Resilient dependable cyber-physical systems: a middleware perspective *Journal of Internet Services and Applications* **3**(1) 41-9

[3] Sang Oh Park, Jong Hyuk Park, Young-Sik Jeong 2013 An Efficient Dynamic Integration Middleware for Cyber-Physical

Systems in Mobile Environments *Mobile Networks and Applications* **18**(1) 110-5

[4] Fatos Xhafa, Leonard Barolli, Omar Hussain.Computing 2013 *Special issue on cyber physical systems* **95**(10-11) 923-6

[5] Anis Koubâa, Maissa Ben Jamâa 2013 Taxonomy of Fundamental Concepts of Localization in Cyber-Physical and Sensor Networks *Wireless Personal Communications* 72(1) 461-507

[6] Seppo Sierla, Bryan M. O'Halloran, Tommi Karhela, Nikolaos Papakonstantinou, Irem Y. Tumer 2013 Common cause failure analysis of cyber–physical systems situated in constructed environments *Research in Engineering Design* **24**(4) 375-94

[7] Patrick Martin, Magnus B. Egerstedt 2012 Hybrid systems tools for compiling controllers for cyber-physical systems *Discrete Event Dynamic Systems* **22**(1) 101-19

[8] Gabriela Magureanu, Madalin Gavrilescu, Dan Pescaru 2013 Validation of static properties in unified modeling language models for cyber physical systems *Journal of Zhejiang University SCIENCE C* **14**(5) 332-46

[9] Yong-hwan Kim, Chan-Myung Kim, Youn-Hee Han, Young-Sik Jeong, Doo-Soon Park 2013 An efficient strategy of nonuniform sensor deployment in cyber physical systems *The Journal of Supercomputing* **66**(1) 70-80

[10] Robert Mitchell, Ing-Ray Chen 2013 On Survivability of Mobile Cyber Physical Systems with Intrusion Detection *Wireless Personal Communications* **68**(4) 1377-91

[11] Young-Sik Jeong, Hyun-Woo Kim, Haeng Jin Jang 2013 Adaptive resource management scheme for monitoring of CPS *The Journal of Supercomputing* **66**(1) 57-69

Authors

Huaping Zhou, born on February 13, 1979, Nanyang, Henan, China

Current position, grades: School of Computer Science and Engineering in Anhui University of Science and Technology, Teacher
University studies: Computer technology
Scientific interest: Information fusion, data mining
Publications:10
Experience: Huaping Zhou is born in Henan province in China in 1979. She is graduated from Anhui University of Science and Technology. My major is computer, and I got her master degree after her graduation in the year of 2004. I am graduated from China University of Mining & Technology (Beijing) in 2009, and my major is computer application technology, and got Ph.D. Now, the main research areas are information fusion, data mining, pattern recognition and other aspects of the research.

Xutong Zhang, born in January, 1991, Bengbu, Anhui, China

Current position, grades: School of Computer Science and Engineering in Anhui University of Science and Technology, graduate student in grade one
University studies: Computer technology
Scientific interest: Information fusion
Experience: From September 2009 to June 2013, I majored in Computer Science and Technology from School of Computer Science and Engineering in Anhui University of Science and Technology.
From September 2013 to now, I majored in Computer Technology from School of Computer Science and Engineering in Anhui University of Science and Technology.
In the summer vacation of the junior year, I went to be an trainee in Kunshan ambow training base.

Research on adaptive H-Infinity tracking for inhibition fluttering of picking robot arm

Zhiyong Zhang[1], Hongli Guo[2*], Lvwen Huang[1], Xiaoting Zhang[1]

[1]*College of information Engineering, Northwest A&F University, Yangling, Shaanxi, 712100, China*

[2]*College of Mechanical & Electronic Engineering,Northwest A&F University, Yangling, Shaanxi, 712100, China*

Abstract

This paper aim at solve issue that conventional frequency domain theory is unsuitable for MIMO system and LOG theory is unsuitable for model perturbation, we commit to the research on H-infinity stability and focus on discuss design principles of H-Infinity stability system and methods on Riccati equation or inequality solution. As for the uncertainty and external interference, we discuss two-output equation or an algebraic Riccati state feedback equation using character on Riccati equation. According to the state space theory, we derive the controller to make structural equation to meet the requirement of state feedback and observer, and then draw suboptimal solutions in the form of engineering management experience. To reduce the impact on interference to control stability by selecting the appropriate interference attenuation coefficient γ, so that its stability could be casted to meet harvest scene.

Keywords: H-infinity, Riccati equation, Interference attenuation coefficient, Domain theory

1 Introduction

Since the state space has been greatly developed, LQG feedback design methods based on quadratic optimal regulator and the Kalman-Bucy filter theory were emerged, these method required control system precise mathematical model to ensure its robustness. If the system model unfit with uncertain external environment, it is necessary for statistical properties of the interference signal known or assumed to be white noise with if-then condition, or robustness design LQG will become poor. However, the interference signal often change with picking environment variability in the actual picking environment, it is difficult to predict the interference statistics in advance. For above reasons, LQG algorithm on the application of the picking robot arm is persuasive, but actually not ideal [1]. As for the interference signal LQG unreasonable restrictions problem, Zames specially designed a famous H-Infinity. He firstly put the interference signal of SISO system into a finite set known signal energy, then deserved feedback controller meeting the closed-loop system, screening a minimal impact on the system disturbance that the minimum H-infinity norm of the sensitivity function.

Douat L. R. and Queinnec [2] obtained the H-infinity optimal controller by solving two Riccati equation, and then proposed a simplified equation for the state space H-Infinity controller, which is equal to the order for the generalized object. Later state feedback $H\infty$ was proposed by solving an algebraic Riccati equation, but mathematics and computation is very powerless because of using the state space method, and later the rapid

development of computer technology and software toolbox control theory appeared to overcome complex calculations shortcomings, so that it could better solve two problems in design stability which are unsuitable MIMO system design and model perturbation, in recent years, it has become a hot research in control theory field and has made some practical application results. Related literature [3-4] indicated to build accurate mathematical model from external interference has been unrealistic, but it must be thought illumination, shelter as the uncertainty case, the picking robot arm control system can meet the requirements of the desired stability. Penne and R, Smet E [5] established one simplified dynamic equations of piezoelectric flexible manipulator and its control system model to design active vibration control system for active vibration control of flexible manipulator by using a piezoelectric element analysis model methods. Later they also established a linear system dynamics equation using time domain precise integration method on solving state equations and algebraic Riccati equation, therefore by using the control disturbance attenuation theory for control strategy of power system set to study. Qiu Q. et al [6] applied H-infinity control strategies aiming at 2-DOF manipulator unstructured uncertainties, the design global tracking controller of two arms inhibit fluttering. These analyzes provided inspiration & reference on continuous research and development for robust control theory uncertain systems. Therefore, this paper describe the maximum gain from the limited input energy to the output energy according to transfer function H-infinity norm of the picking robot arm, and Sahli N, Moulin B [8] was optimized through the norm of the transfer function

* *Corresponding author* e-mail:hongliguo1975@126.com

as the objective function, composed by criterion of Riccati inequality to determine the solvability of control problems [9-10], the interference with the limited power spectrum has minimal impact on the stability of the desired output.

2 Picking robot arm controller design

As for picking robot arm control systems based H-infinity theory, whether interference control problem or robustness stability problems, can be summarized the closed-loop stability system and made the closed-loop transfer function matrix H-infinity minimum or less than a given value from the feedback controller problem [11],it can to design the controller in this way. Using time domain design theory method based on Riccati equation to solve a state feedback or 2 output feedback algebraic equation, derive the controller based on the state space representation with observer plus state feedback form of controller[12]. Considering the time domain is due to comprehensive rigorous theoretical system design owning optimal control theory LQR / LQG approaching the characteristics concept. So it is ape to achieve the optimal or suboptimal robustness in many fields for most researchers to enforce the stability of controller design.

2.1 THE STANDARD DESIGN ON H-INFINITY

In the study of the various H-infinity control optimization problems ,we found H-Infinity controller is divided into state feedback controller and output feedback controller according to the source measured signal, but in anyway, it need to reduce to the standard design problems in the way of actual controlled object model and design requirements. The related study [13-15] has been found that regardless of robust stabilization problem, the sensitivity minimization problem, mixed sensitivity optimization problem, three freedom degrees problems, tracking problems, the model matching problem, filtering problem and many other control problems are all attributed to the H-Infinity control problem, which will produce a far-reaching impact on control theory. The Following is the general H**-**Infinity tracking control method for picking robot arm
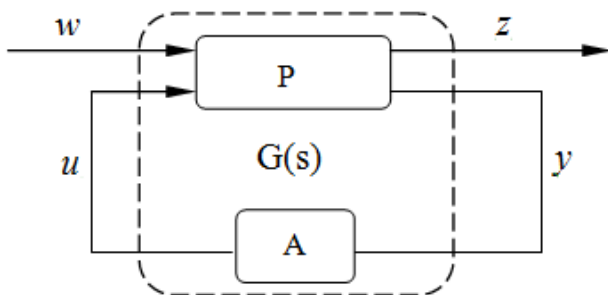
FIGURE1 The general H**-**Infinity tracking control method for picking-arm

Figure1 stands on the general problem of H-Infinity control block diagram. Where P includes three drive arm

joints and its general object weighting function is defined as

$$
\begin{bmatrix} z \\ y \end{bmatrix} = P \begin{bmatrix} w \\ u \end{bmatrix} = \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} \begin{bmatrix} w \\ u \end{bmatrix},
\tag{1}
$$

where $w \in R^r$ represents the external input signal, $z \in R^m$ is controlled output signal, $u \in R^p$ denotes a control signal, $y \in R^q$ represents table measurement signal.

Considering equation (1) shown as the closed-loop system transfer function G(s), also known as augmentation controlled object, actually determined by the input signal from u , w to output signal z , y . Transfer martix including weighting function and the actual controlled object. The state space can be illustrated as following

$$
\begin{cases} x = Ax + B_1 w + B_2 u \\ z = C_1 x + D_{11} w + D_{12} u \\ y = C_2 x + D_{21} w + D_{22} u \end{cases},
\tag{2}
$$

where x denotes n dimensional state variables, u is p dimensional state variable, w is r dimensional state variable, y is q dimensional state variable, z is m dimensional state variable, the equation (1) can also be written as

$$
G(s) = \begin{bmatrix} w \\ u \end{bmatrix} = \begin{bmatrix} G_{11}(s) & G_{12}(s) \\ G_{21}(s) & G_{22}(s) \end{bmatrix} = \begin{bmatrix} A & B_1 & B_2 \\ C_1 & D_{11} & D_{12} \\ C_2 & D_{21} & D_{22} \end{bmatrix},
\tag{3}
$$

wherein, H-Infinity control is equal to the performance on positive definite controller

$$
U = Ky
\tag{4}
$$

where K stands on controller, considering closed loop control system for the picking arm will be stable, the norm of the transfer function from w to z is minimum

$$
T_{zw} = G_{11} + G_{12} K (1 - G_{22})^{-1} G_{21}
\tag{5}
$$

Namely, to solve the following equation

$$
Min \| T_{zw} \| \infty .
\tag{6}
$$

Equation (5) denotes H-Infinity control problem, if given γ>0, Equation (6) will be written as

$$
\| T_{zw} \| \infty < \gamma .
\tag{7}
$$

Therefore, the above equation indicate sub-optimal control problem.

If H-Infinity norm of equation (5) and (6) is rewritten H_z norm, namely $Min \| T_{zw} \|_z$ or $\| T_{zw} \|_z < \gamma$, the algorithm then is called as H_z algorithm, LQG algorithms. As for the known G(s), H-Infinity suboptimal will be solvability through the seek tentatively suboptimal solution, repeated

descending γ, making γ close to γ_0, explore the optimal solution approximation solution.

Another equation (7) is equivalent to

$$\left\| \frac{1}{\gamma} T_{zw} \right\|_\infty < 1 \cdot \tag{8}$$

However, $(1/\gamma)T_{zw}$ is equal to the augmented controlled object actually

$$G_\gamma(s) = \begin{bmatrix} \gamma^{-1}G_{11}(s) & G_{12}(s) \\ \gamma^{-1}G_{21}(s) & G_{22}(s) \end{bmatrix}. \tag{9}$$

Also, with the negative feedback controller G(s) consist of close-loop transfer function shown as Figure2, the case $\gamma=1$ will be only considered on the controller design.
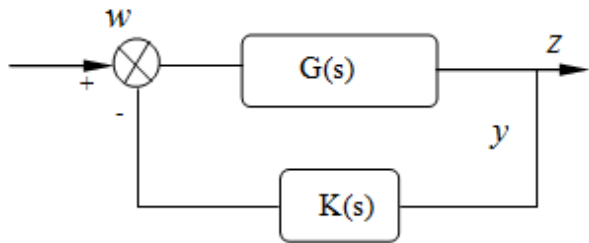
FIGURE2 The schematic diagram of H-Infinity close-loop control for picking-arm

2.2 GENERAL SOLUTION OF OUTPUT FEEDBACK

Assuming augmented control object matrix meet the following conditions [18]:
(1) (A, B_1) and (A, B_2) are stable pair,
   (2) (C_1, A) and (C_2, A) are detectable pair
   (3) $G_{12}^T[C_1 \; C_{12}] = [0 \; I]$
   (4) $D_{21}[D_{21}^T \; B_1^T] = [I \; 0]$
   (5) $G_{12}(s)$ and $G_{12}(s)$ have no zeros points on the imaginary axis.
   (6) $D_{11}=0$ and $D_{22}=0$
The following can be obtained from the equation (3)

$$G(s) = \begin{bmatrix} A & B_1 & B_2 \\ C_1 & 0 & D_{12} \\ C_2 & D_{21} & 0 \end{bmatrix}. \tag{9}$$

Theorem 1 Assuming the augmented control object (8) meet the conditions (1)-(6). There exits controller $K(s)$ shown as Figure 2, which close-loop system stabilize, moreover the sufficient and necessary condition for $\|T_{zw}\|\infty < 1$ is set up

(1) Riccati equation has a semi-positive definite solution, making the array stable

$$A^T X + XA + X(B_1 B_1^T - B_2 B_2^T)X + C_1^T C_1 = 0. \tag{11}$$

(2) Riccati equation has positive semi-definite, so the following array is stable

$$AY + YA^T + Y(C_1^T C_1 - C_2^T C_2)Y + B_1 B_1^T = 0. \tag{12}$$

(3) Riccati equation

$$\gamma_{max}(XY) < 1. \tag{13}$$

If the condition (1) - (3) established for the augmented controlled object (10), the H-Infinity standard solution is given by the following equation

$$K(s) = \begin{bmatrix} A + B_1 B_1^T - Z^{-1}LC_2 + B_2F & -Z^{-1}L \\ -F & 0 \end{bmatrix}, \tag{14}$$

where $F = B_1^T X$, $L = YC_2^T$, $Z = I-XY$

2.3 PICKING ROBOT ARM TRACKING CONTROLLER DESIGN

Seen from the H-Infinity tracking controller shown as Figure3, the output y tracks reference input r, the controlled object $P(s)$ is known, the controlling input of the object u namely r, y is generated by the forward channel controller C_1, the feedback controller C_2, respectively.
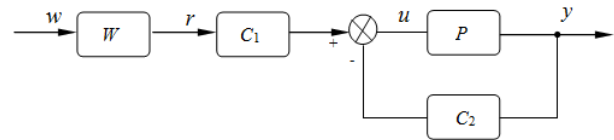
FIGURE 3 The schematic diagram of H-Infinity tracking control for picking-arm

Where r denotes a finite energy instead of an known input signal, r denotes a signal in the equation of $\{r | r = Ww, w \in H_2, \|w\|_2 \leq 1\}$, H_2 indicates the set components of all finite energy signal

$$H_2 = \left\{ w(t) \Bigg| \left[\int_0^\infty w^*(t)w(t)dt \right]^{\frac{1}{2}} \leq +\infty \right\}. \tag{15}$$

In the course of tracking control, the controller P and W are known, but C_1 and C_2 remains to be further designed, the tracking error r and y are the actual controlled variables pursuit smallest error tracking system, taking $\|r-y\| + \|\rho u\|$ into the following equation objective function of the system, which corresponds to the signal.

$$z = \begin{bmatrix} r-y \\ \rho u \end{bmatrix}, \tag{16}$$

where ρ denotes the weighting factor greater than 0 Therefore, regarding the objective function minimization problem as the key issue on picking arm tracking

$$\sup_{w \in BH_2} \|z\|, \tag{17}$$

where BH$_2$ denotes H$_2$ unit sphere. If we transfer tracking control issue on picking robot arm to standard H-Infinity control issue, we may obtain

$$z = \begin{bmatrix} r\text{-}y \\ \rho u \end{bmatrix}, \quad v = \begin{bmatrix} r \\ y \end{bmatrix}, \tag{18}$$

u and v as control signals and external input signals, respectively. The equation on controller and the generalized controlled device is shown as following

$$\begin{bmatrix} z \\ v \end{bmatrix} = \begin{bmatrix} r\text{-}y \\ \rho u \\ r \\ y \end{bmatrix} = \begin{bmatrix} W & \text{-}P \\ 0 & \rho I \\ W & 0 \\ 0 & P \end{bmatrix} \begin{bmatrix} w \\ v \end{bmatrix}, \tag{19}$$

$$u = \begin{bmatrix} C_1 & C_2 \end{bmatrix} \begin{bmatrix} r \\ y \end{bmatrix}. \tag{20}$$

K and G are correspondingly shown as

$$G = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix}, \tag{21}$$

where $G_{11} = \begin{bmatrix} W \\ 0 \end{bmatrix}, G_{12} = \begin{bmatrix} \text{-}P \\ \rho I \end{bmatrix}, G_{21} = \begin{bmatrix} W \\ 0 \end{bmatrix}, G_{22} = \begin{bmatrix} 0 \\ P \end{bmatrix}$,

$K = \begin{bmatrix} C_1 & C_2 \end{bmatrix}$, now, Figure3 shown as the tracking problem can be converted to standard problem.
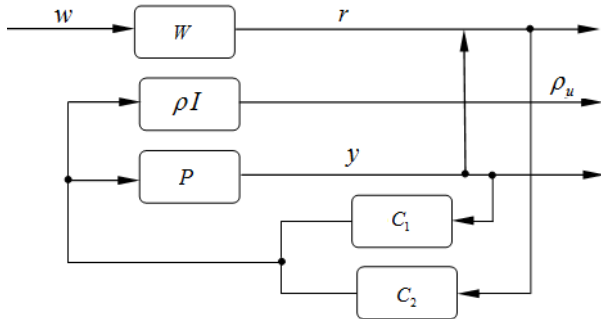
FIGURE 4 The schematic diagram of standard H-Infinity tracking control method

3 Design on H-Infinity controller picking robot arm

The Picking robot arm is regarded as a highly nonlinear control system, according to [15] description of the transfer matrix H$_2$ norm that is the gain of control system showing that the existence of a positive definite solution $H_2 < \gamma$ of the Riccati equation on the necessary and sufficient conditions. As described the above, if the gain of non-linear system is greater than $\gamma (\gamma > 0)$, the control system (11) is the dissipation. Actually, on the certain conditions, stable system can be detected, no zero, which meet the necessary conditions that gain is less than γ. The necessary conditions and the range of linear systems is consistent. By analysing on nonlinear systems (10), if

$r^2 I\text{-}d^T(x)d(x) \geq 0, \forall x \in R^n$ is established, then $w(x)$ can be taken as

$$\omega(x) = \frac{\sqrt{2}}{2} \left[r^2 I\text{-}d^T(x)d(x) \right]^{1/2}. \tag{22}$$

Therefore, $w(x)$ is invertible matrix, according to Equation (12), $l(x)$ can be represented as

$$l(x) = -\frac{1}{2}\omega^{-1}(x) \left[L_g V(x) + h^T(x)d(x) \right]^T, \tag{23}$$

Substitute the equation (25) into the first relationship of equation (8), $V(x)$ meets the inequality condition

$$L_f V(x) + \frac{1}{2}h^T(x)h(x) + \frac{1}{2}[L_g V(x) + h^T(x)d(x)]$$
$$\times [r^2 I\text{-}d^T(x)d(x)]^{-1}[l_g V(x) + h^T xd(x)]^T \leq 0 \tag{24}$$

Therefore, we can deduce Theorem 1, there exit the following theorem.

Theorem 2 If the optimal solution on the equation (7) is established, for nonlinear picking arm system (25) meet the semi-definite storage system consisting of a smooth function differentiable conditions make it necessary and sufficient dissipation conditions for the inequality (26), has the positive semi-definite solution. If $d(x) = 0$, then inequality (26) can be expressed via the following inequality:

$$L_f V(x) + \frac{1}{2\gamma^2}L_g V(x)L_g^T V(x) + \frac{1}{2}h^T(x)d(x) \leq 0, \tag{25}$$

equivalently:

$$\frac{\partial V}{\partial x}f(x) + \frac{1}{2r^2}\frac{\partial V}{\partial x}g(x)g_g^T(x)(\frac{\partial V}{\partial x})^T + \frac{1}{2}h^T(x)h(x) \leq 0. \tag{26}$$

If the system dissipative coefficient $\gamma > 0$, on the nonlinear system (25) $d(x) = 0$ conditions, storage function $V(x)$ is smooth- differentiable making the dissipation γ necessary and sufficient condition for inequality (27) meet semi-positive definite function $V(x)$. If the system gain L_2 is less than γ, so the system is dissipated. Therefore the system (27) is greater than picking arm gain γ sufficient condition still exists positive semi-definite function meeting the inequality (27). To discuss the nonlinear picking arm system, there exits the following equation

$$\begin{cases} \dot{x} = f(x) + \Delta f(x) + [g(x) + \Delta g(x)]u + [k(x) + \Delta k(x)]d \\ y = h(x) \end{cases} \tag{27}$$

As equation (27) have shown strong anti-jamming attenuation performance, while meeting performance index of H-Infinity, then there must be a smooth differentiable positive definite function $V(x)=0$ be up to the control law $u=u(x)$,the following inequality holds

$$HJI = V_x[f(x) + \Delta f(x)] + \frac{1}{2} V_x[k(x) + \Delta k(x)][k(x) + \Delta k(x)]^T V_x^T$$
$$+V_x[g(x) + \Delta g(x)]u + \frac{1}{2} h^T(x)h(x) \le 0 \qquad (28)$$

Namely, from the system disturbance d to the output y are less than or equal gain γ, where $V_x = \partial V / \partial x$. According to related description of the mathematical model 3-DOF picking robot arm, we can obtain the uncertain environment dynamic model picking robot arm

$$u = M(q)\ddot{q} + C(q,\dot{q})\dot{q} + G(q) + F(\dot{q}) + d, \qquad (29)$$

where M denotes $M(q) \in \mathrm{R}^{3 \times 3}$, C denotes $C(q,\dot{q}) \in \mathrm{R}^{3 \times 3}$, G denotes $G(q) \in \mathrm{R}^3$, F denotes $F(\dot{q}) \in \mathrm{R}^{3 \times 3}$ related physically significant have be illustrated before. u denotes the input moment vector for 3 joints, as for $\forall T > 0$, interference variables $d \in [0, T_2]$.

Taking into account parameters uncertainty for picking robot arm control system including the external operating disturbances environment, moreover different harvesting conditions allocated with different requirements on 3 arm joints, we regards the external interference as one part of the uncertainty via H-Infinity stability control strategy.

3.1 UNCERTAINTY AND ANTI-INTERFERENCE STRATEGY

Referenced by the literature [13-17], we can regard the uncertainty and external interference as one the inseparable whole. The system interference is defined as

$$u = M_0(q)\ddot{q} + C_0(q,\dot{q})\dot{q} + G_0(q) + F_0(\dot{q}) + \omega, \qquad (30)$$

where $f(x) = \begin{bmatrix} \dot{q} \\ -M_0^{-1}(C_0\dot{q} + G_0 + F_0) \end{bmatrix}$, $g(x) = \begin{bmatrix} 0 \\ M_0^{-1} \end{bmatrix}$, $\omega' = \begin{bmatrix} 0 \\ M_0^{-1}\omega \end{bmatrix}$.

Learned from equation (30), anti-interference controller can be obtained

$$u_1 = g(x)^{-1}[-f(x) + \ddot{x}_d + k_1\dot{e} + k_2 e], \qquad (31)$$

where x_d denotes expected trajectory for picking robot arm, k_1 and k_2 represent constants. So system errors can be expressed as

$$\ddot{e} + k_1\dot{e} + k_2 e = -\omega'. \qquad (32)$$

If $\omega'=0$ $=0$, then there exit no interference and uncertainty for picking robot arm, now $\lim_{t \to \infty} e(t) = 0$,

showing that the tracking errors restrain gradually. If $\omega' \ne 0$, the compensation control can be considered to add $u_2(\mathrm{t})$, then, we may obtain the close–loop system expression as

$$\ddot{e} + k_1\dot{e} + k_2 e = u_2(t) - \omega'. \qquad (33)$$

If the equation (33) is adapted to state space equation, it can be expresses as

$$\begin{cases} \dot{e} = Ae + Bu_2 - B\omega' \\ y = e \end{cases}, \qquad (34)$$

where $A = \begin{bmatrix} 0 & 1 \\ -k_1 & -k_2 \end{bmatrix}$, $B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$.

Assumming the Lyapunov function $V(t)=0.5\ e^T pe$, where $V(t) > 0$, as for $\forall > 0, V(0)=0$ p is a positive definite solution for the Riccati equation

$$A^T P + PA - \frac{2}{\beta} pBB^T p + \frac{1}{\gamma^2} pBB^T = -Q. \qquad (35)$$

Equation (35) with the premise of positive definite solution is $\beta \le 2\gamma^2$, where γ denotes the interference attenuation, assuming $u_2 = -\frac{1}{\beta} B^T Pe$, where $\beta > 0$, now, we may testify compensate controller u_2 meets the stability index of H-Infinity system for equation (35).

Firstly, according to the equation (35) of close-loop err system, derivate Lyapunov function $V(x)$ and compare equation (36), we can gain the equation (37), Then integrate both side of equation (37) and gain the inequality as following.

$$\int_0^T e^T Qedt \le \gamma^2 \int_0^T \omega^T \dot{\omega} dt, \qquad (36)$$

using $y = e$, therefore $\max_{\omega \in L_2} \frac{\|y(t)\|_{L_2}}{\|\dot{\omega}(t)\|_{L_2}} \le \gamma^2$, it can be proved that equation (33) to be robust disturbance rejection performance, namely, we can gain L_2 is less than the given positive γ from interference ω' to system output y, furthermore, if the less γ is, the quickly interference attenuate, the better H-Infinity stability is, therefore we can take full advantage of attenuation rate γ to gain the optimal control stability.

$$\dot{V}(t) = \frac{1}{2}\dot{e}^T pe + \frac{1}{2}e^T p\dot{e} =$$
$$-\frac{1}{2}e^T Qe - \frac{1}{2}(\frac{1}{\gamma}B^T pe - r\dot{\omega})^T(\frac{1}{\gamma}B^T pe - r\dot{\omega}) + \dot{e}^T pe + \frac{1}{2}\gamma^2 \omega^T \dot{\omega} \le -\frac{1}{2}e^T Qe \frac{1}{2}\gamma^2 \omega^T \dot{\omega}. \qquad (37)$$

Zhang Zhiyong, Guo Hongli, Huang Lvwen, Zhang Xiaoting

4 The experimental simulation and analysis

4.1 EXPERIMENTAL DESCRIPTION

Picking robot arm model diagram is shown as Figure7, where m_1, l_1 and q_1 respectively represent mass, length and the angle of upper arm, m_2, l_2 and q_2 denote mass, length and angle of forearm, m_3, l_3 and q_3 respectively denote mass, length and angle of picking arm, J_1, J_2, J_3 represent the moment of inertia of above 3 arms. By the simulation, it is assumed q=[q_1,q_2], $\tau = [\tau_1, \tau_1]^T$ the model of picking robot arm can be represented by the equation (2), where $G(q)$, $M(q)$ and $C(q,\dot{q})$ can be represented as

$$M(q) = \begin{bmatrix} \theta_1 + \theta_3 + 2\theta_2 \cos q_2 & \theta_3 + \theta_2 \cos q_2 & \cos q_2 \\ \theta_3 + \theta_2 \cos q_2 & \theta_3 & \theta_2 \cos q_2 \\ \theta_1 + \theta_3 + \theta_2 \cos q_2 & \theta_1 + \theta_3 os q_3 & \theta_3 \end{bmatrix}. \quad (38)$$

$$C(q,\dot{q}) = \begin{bmatrix} -\theta_2\dot{q}_2 \sin q_2 & -\theta_2(\dot{q}_2 + \dot{q}_1)\sin q_2 & \dot{q}_2\dot{q}_3 + \dot{q}_1\dot{q}_2 \sin q_2 \cos q_3 \\ \theta_2\dot{q}_1 \sin q_2 & 0 & (\dot{q}_2\dot{q}_1)\cos q_2 \\ \dot{q}_2 + \dot{q}_3 \cos q_2 & \dot{q}_1 + \dot{q}_2\dot{q}_3 \cos q_3 & 0 \end{bmatrix}. \quad (39)$$

Assuming G(q)=0, then define the parameter vector

$$\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix} = \begin{bmatrix} (0.42m_1 + m_2 + 0.58m_3)l_1^2 + J_1 \\ 0.68m_2l_1l_2 + 0.22m_3l_3^2 \\ 0.25m_2l_2^2 + J_3 \end{bmatrix}. \quad (40)$$

Derivative the inertia matrix of the matrix

$$\dot{M} = \begin{bmatrix} -2\theta_2\dot{q}_2 \sin q_2 & -\theta_2\dot{q}_2 \sin q_2 & \dot{q}_2\dot{q}_3 + \dot{q}_1\dot{q}_2 \cos q_2 \\ -\theta_2\dot{q}_2 \sin q_2 & 0 & -\dot{q}_2\dot{q}_1 \cos q_2 \\ \dot{q}_2 + \dot{q}_3 \sin q_2 & \dot{q}_1 + \dot{q}_2\dot{q}_3 \sin q_3 & 0 \end{bmatrix}. \quad (41)$$

Select equation (41) as the picking robot arm the system's physical parameter vector, then the regression matrix can be gained

$$\Phi = \begin{bmatrix} \ddot{q}_1 & 2\ddot{q}_1 \cos q_2 - 2\dot{q}_1\dot{q}_2\dot{q}_3 \sin q_2 + \ddot{q}_2 \cos q_2 & \ddot{q}_1 + \ddot{q}_2 \\ \ddot{q}_2 & \ddot{q}_1\ddot{q}_3 \cos q_2 - \dot{q}_1^2 \sin q_2 \cos q_3 & \ddot{q}_2 + \ddot{q}_3 \\ \ddot{q}_2 + \ddot{q}_3 & \dot{q}_1 + \dot{q}_2\dot{q}_3 \sin q_3 & \ddot{q}_1 + \ddot{q}_3 \end{bmatrix}. \quad (42)$$

5.2 EXPERIMENTAL METHODS

Using experimental data for picking robot arm to validate stability and effectiveness of H-Infinity algorithm, the specific physical experimental parameters for simulation experiments extracted from [17, 18] were taken different disturbance attenuation coefficient to observe 3 picking arm joint trajectory tracking situation, two different interference attenuation coefficient were taken and analysis simulation comparison, other simulation parameters are chosen as $\beta = 3\gamma^2$, $p = \begin{bmatrix} 6 & 2 \\ 2 & 6 \end{bmatrix}$, the control effect are revealed in Figure 8, Figure 9 and Figure 10 shows. And in order to verify the stability of the control results, the three drives joints are respectively applied the same sinusoidal disturbance signal, and the 3 joints are made with and without H-Infinity control simulation comparison, the control results are shown as Figure 11, Figure 12 and Figure 13.
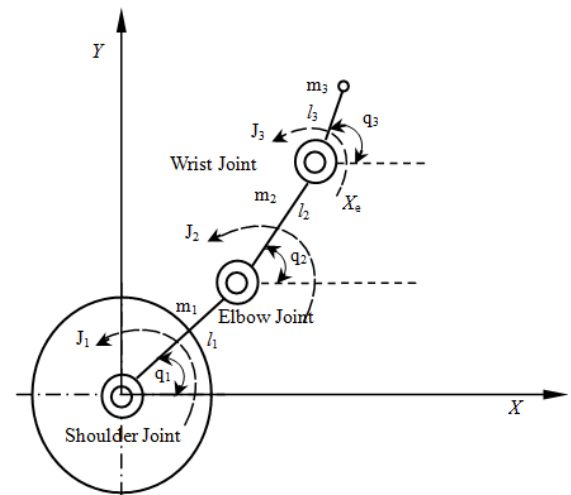
FIGURE 5 The model schematic diagram of picking-arm

Zhang Zhiyong, Guo Hongli, Huang Lvwen, Zhang Xiaoting
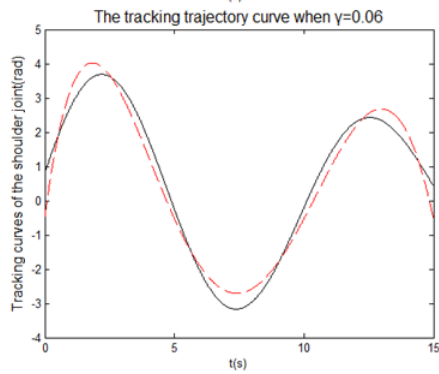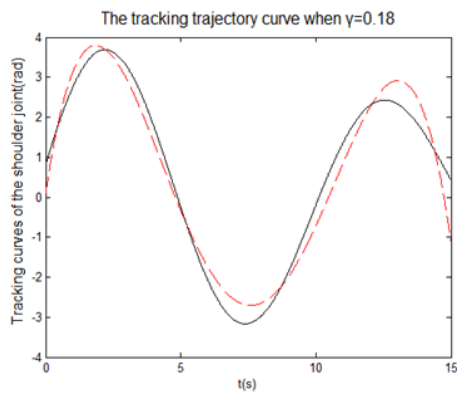
FIGURE 6 Tracking curves of the elbow joint
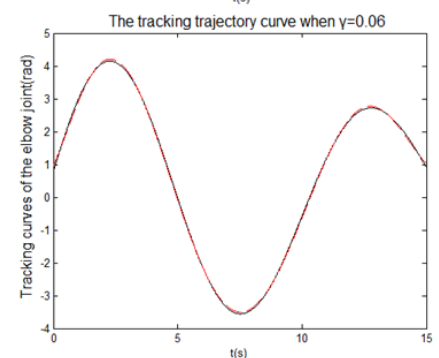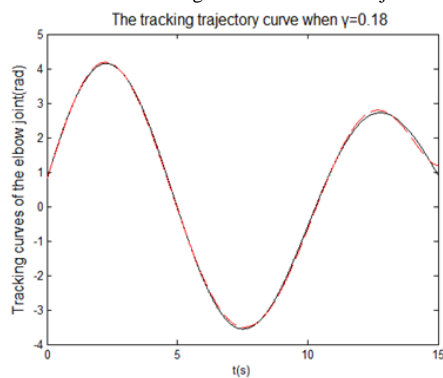
FIGURE 7 Tracking curves of the elbow joint
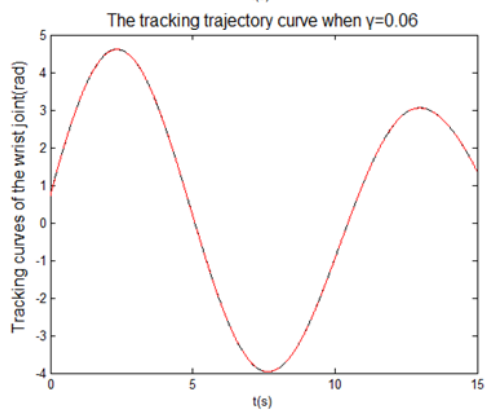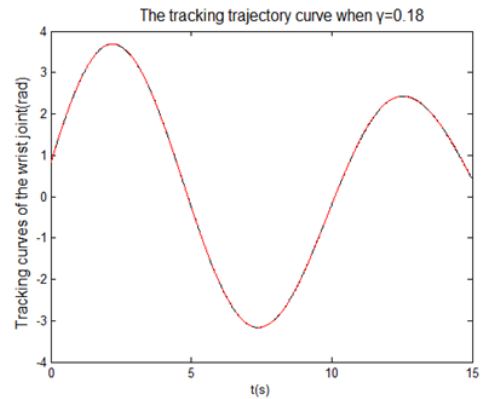
FIGURE 8 Tracking curves of the wrist joint
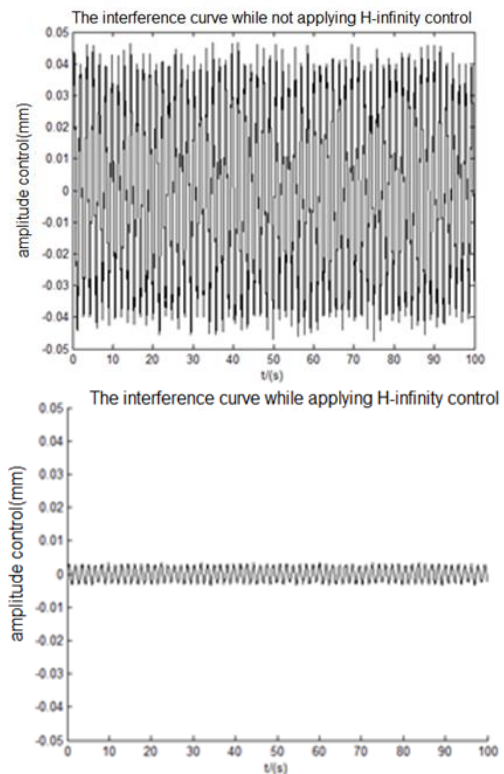
FIGURE 9 Response curves of the shoulder joint for disturbance

The interference curve while not applying H-infinity control

The interference curve while applying H-infinity control
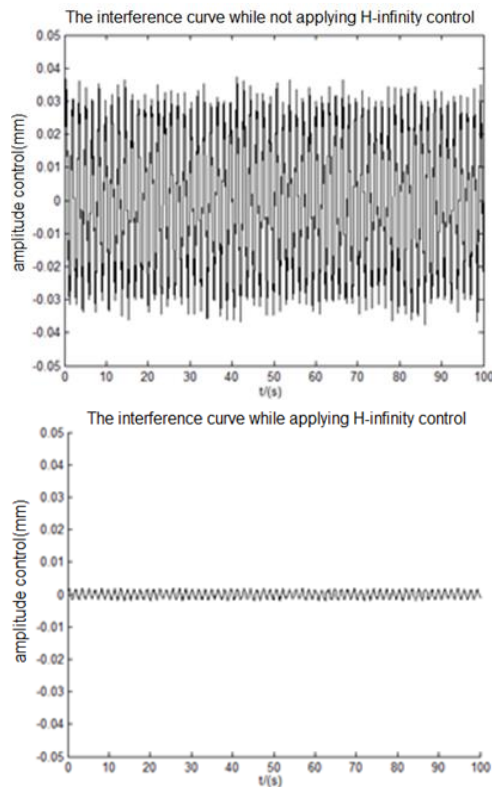
FIGURE 10 Response curves of the elbow joint for disturbance

The interference curve while not applying H-infinity control

The interference curve while applying H-infinity control
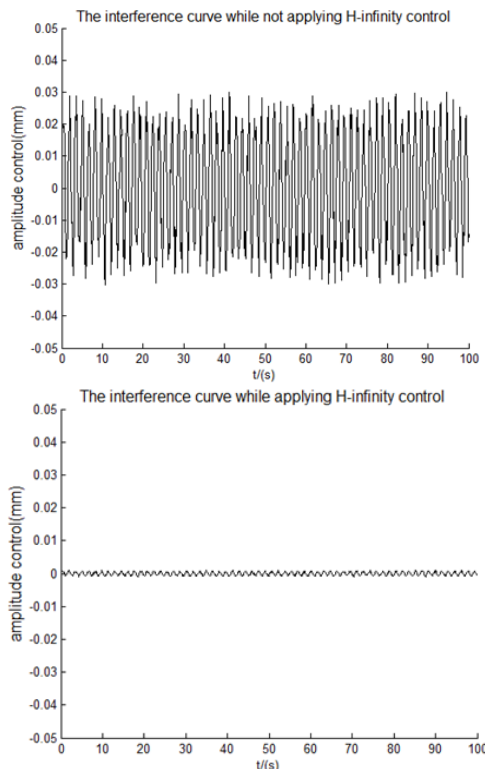
FIGURE 11 Response curves of the wrist joint for disturbance

Figure 8, Figure 9 and Figure 10 represent respectively the trace err for the shoulder, elbow and wrist. Simulation results showed that the joint stability is related to disturbance attenuation coefficient γ, the smaller the disturbance attenuation, faster disturbance attenuate, better system track, stronger its stability is, but γcannot be taken a unlimited less. Therefore, we need to choose a more suitable small disturbance attenuation coefficient to satisfy the final desired requirement. Meanwhile, the best wrist tracking, the tracking error is almost zero to better guarantee crawling and picking stability, enable to avoid fruit injury, elbow joint followed, which tracking error is less than 1%, maximum error emerge in shoulder joint, for which lie the bottom to maintain the other two joints steadily. Figure 11, Figure12 and Figure13 respectively represent interference output response curve of shoulder, elbow and wrist joints, it can be seen by contrasting the presence or absence from H-Infinity control, if H∞ control is applied to the system, the transition time for each arm is a large shorten and the steady state amplitude decrease significantly, showing a good static quality. the system vibration is reduced within 1.5%-3% of the original sinusoidal signal. The comprehensive FIGURE8-FIGURE13 shows that H-Infinity controller can guarantee high dynamic and static quality, strong robustness to model error interference suppression for picking arm closed-loop system. It is great theoretical and practical significance to improve the system stability.

5 Conclusion

This paper aims to do with external interference and uncertainty problem for picking robot arm by solving the two-output Riccati equation, and adjust the appropriate interference control to reduce interference attenuation effect on its stability and design the adaptive H-Infinity controllers. The system can satisfy specific control stability requirements. Simulation results show that the adaptive controller has strong interference adaptablity to environmental disturbances uncertain within 1.5% -3%. The controller on the fluttering inhibition for picking robot can be extended to other similar stability control applications.

Acknowledgements

References

[1] Fang R W, Chen J S 2002 A cross-coupling controller using an H-infinity scheme and its application to a two-axis direct-drive robot *Journal of Robotic Systems* **19**(10) 483-97

[2] Theesar S J S, Balasubramaniam P 2011 H infinity Fuzzy Control of Markovian Jump Nonlinear Systems with Time Varying Delay *Control, Computation and Information Systems*. Gandhigram, INDIA 35-46

Mathematical and Computer Modelling

[3] Douat L R, Queinnec I, et al. 2011 H infinity Control Applied to the Vibration Minimization of the Parallel Robot Par2. *2011 Ieee International Conference on Control Applications* Denver, CO: 947-52

[4] Pedamallu C, Ozdamar L 2011 Solving kinematics problems by efficient interval partitioning *Optimization and Engineering* **12**(3) 459-568

[5] Penne R, Smet E, Klosiewicz P 2011 A Short Note on Point Singularities for Robot Manipulators *Journal of Intelligent & Robotic Systems* **62**(2) 205-12

[6] Qiu Q, Han J 2011 2.5-dimensional angle potential field algorithm for the real-time autonomous navigation of outdoor mobile robots *SCIENCE CHINA Information Sciences* **54**(10) 2100-9

[7] Sahli N, Moulin B 2009 EKEMAS, an agent-based geo-simulation framework to support continual planning in the real-word *Applied Intelligence* **31**(2) 188-97

[8] Sakai S, Iida M, Osuka K, Umeda M 2008 Design and control of a heavy material handling manipulator for agricultural robots *Autonomous Robots* **25**(3) 189-98

[9] Stentz A, Dima C, Wellington C, Herman H, Stager D 2002 A System for Semi-Autonomous Tractor Operations *Autonomous Robots* **13**(1) 87-98

[10] Umeda M, Kubota S, Iida M 1999 Development of "STORK", a watermelon-harvesting robot *Artificial Life and Robotics* **3**(3) 143-9

[11] Viola R, Walsh J, Melka A, Womack W, Murphy S, Riboldi-Tunnicliffe A, Rupp B 2011 First experiences with semi-autonomous robotic harvesting of protein crystals *Journal of Structural and Functional Genomics* **12**(2) 77-82

[12] Vougioukas S, Fountas S, Blackmore S, Tang L 2005 Combining reactive and deterministic behaviours for mobile agricultural robots *Operational Research* **5**(1) 153-9

[13] Wu Y L, Sun F C, Zheng J C, Song Q 2010 A robust training algorithm of discrete-time MIMO RNN and application in fault tolerant control of robotic system *Neural Computing & Applications* **19**(7) 1013-9

[14] Wu Y X, Feng Y, Hu Y M 2007 Output tracking control of mobile manipulators based on dynamical sliding-mode control *Frontiers of Mechanical Engineering in China* **2**(1) 110-8

[15] Xie S Q, Cheng D, Wong S, Haemmerle E 2008 Three-dimensional object recognition system for enhancing the intelligence of a KUKA robot *The International Journal of Advanced Manufacturing Technology* **38**(7) 822-39

[16] Yin Y G, Hosoe S, Luo Z W 2007 A mixed logic dynamical modeling formulation and optimal control of intelligent robots *Optimization and Engineering* **8**(3) 321-30

[17] Yahya H Zweiri 2008 Identification schemes for unmanned excavator arm parameters *International Journal of Automation and Computing* **5**(2) 185-92

[18] Zou X J, Zou H X, Lu J 2012 Virtual manipulator-based binocular stereo vision positioning system and errors modelling *Machine Vision and Applications* **23**(1) 43-63

Authors

Zhiyong Zhang

Current position, grades: Lecture
University studies: Fundamental Digital Circuit, Fundamentals of Robot technology, Controlling Technique for Programming
Scientific interests: Intelligent detecting and control, agricultural robot, and agricultural information technology
Publications:
Experience: The B.Sc.degree in agricultural engineering in 1999, the M.Sc. degree in agricultural Engineering in 2004, and the Ph.D. degree in agricultural engineering from Northwest A&F University, Yangling, Shaanxi, China, in 2012.

Hongli Guo

Current position, grades: Lecture
University studies: Theoretical Mechanics, Fundamentals of Robot technology, Controlling Technique for Programming
Scientific interests: Mechanical design and manufacturing and electromechanical integration design.
Publications:
Experience: The B.Sc.degree in agricultural Engineering in 1999, the M.Sc. degree in Processing and storage engineering of agricultural products in 2004

Lvwen Huang

Current position, grades: Lecture
University studies: Computer Control System, Fundamentals of Robot technology, Controlling Technique for Programming.
Scientific interests: Intelligent system and image processing.
Publications:
Experience: The B.Sc. degree in automation engineering in 1999, the M.Sc. degree in signal and information processing from the Xi'an University of Technology, Xi'an, Shaanxi, China in 2005, and the Ph.D. degree in Agricultural Electrification and Automation from Northwest A&F University, Yangling, Shaanxi, China, in 2013.

Xiaoting Zhang

Current position, grades: Lecture
University studies: Fundamental Digital Circuit, Fundamentals of Robot technology, Reliability Engineering
Scientific interests: Intelligent system, database and Management Information System.
Publications:
Experience: The B.Sc.degree in Economic Information Management of Xi'an Petroleum Institute, Shaanxi, China in 2001, the M.Sc. degree in agricultural economic management of Northwest A&F University, China in 2009.

Spectrum allocation algorithm based on user requirements under the circumstance of advanced user existence

Yuliang Cong¹, Shuyang Zhang²*, Limin Xu, Lili Sun

¹ College of Communication Engineering, Jilin University, Changchun, 130012, China

² College of Communication Engineering, Jilin University, Changchun, 130012, China

Abstract

Since wireless spectrum is a non-renewable resource, how to improve the spectrum utilization is always the problem to be resolved by wireless communication technology. With the development of wireless communication technology, the contradiction between supply and demand for spectrum resource has been more and more intense. In this case, cognitive radio technology emerged. The traditional list-colouring algorithm aims at maximizing the number of allocated bandwidth; CSGC algorithm (Colour Sensitive Graph Colouring algorithm) is to achieve the maximum benefits of bandwidth for the cognitive users; local bargaining algorithm is an improved algorithm on time complexity based on CSGC. However, the three algorithms do not take the bandwidth demand of cognitive users into consideration. Even with the proportion allocation of CSGC algorithm, the problem of ill-considered for the bandwidth demands of cognitive users also exists, which results in irrational allocation of spectrum resources. To solve this problem, this article proposes a priority order with the consideration of cognitive users in spectrum allocation based on advanced users.

Keywords: Cognitive Radio, Spectrum Allocation, Graph Colouring, User Needs

1 Introduction

Cognitive radio is an effective means to alleviate the current tension radio spectrum resources in the international and domestic, more and more research organizations consider that cognitive radio could perfectly play an important role in the next generation wireless communication system; it is a landmark in new wireless technology [1-4].

Dynamic spectrum allocation is one of the key technologies of cognitive radio. A useful dynamic spectrum algorithm can improve the efficient use of spectrum resources, the flexibility use of the spectrum and the throughput of the system effectively [5, 6]. Therefore, dynamic spectrum algorithm has become a hot research spot of cognitive radio.

2 The priority of users participating in spectrum allocation

The purposes of communication are different for different users. Therefore, the demands for bandwidth are not the same [7]. In order to solve the problem of the bandwidth demands for individual cognitive user, the cognitive users involved in spectrum allocation will be divided into distribution levels, and the cognitive users with high level (advanced users) will allocate spectrum in priority. Meanwhile, in order to avoid the case that the users with low level will be allocated with little spectrum after labelled level, small number of cognitive users will be

given priority, and the rest of the users do not calibrate the priority or give the same priority level [8, 9]. There are 10 cognitive users, for example, of which there are three advanced users, and then decide priorities based on the three high-level users' information (such as bandwidth demands, communication mode, etc.), then execute the following assignment. It not only can meet the advanced users' demands, but also take the needs of other users into account, and ensure the fairness of spectrum allocation, so that spectrum allocation is more reasonable. Taking the fairness of the distribution, and in order to avoid the case that a small number of users take up too much spectrum resources, the proportion of advanced users is very low [10, 11].

3 Mathematical description of the algorithm

In cognitive radio systems, the wireless environment change. The appearance of authorized users, as well as other factors, changes the channel quality. These changes will affect the users' available spectrum [12, 13]. Dynamic spectrum allocation algorithm studies in a cycle (very short), which allocates the detected available spectrum to cognitive users, and then the cognitive users access and use the spectrum [14]. The proposed spectrum's demand is a short-term demand in one cycle.

Assume that the bandwidth demand of a cognitive user n participating in spectrum allocation in a cycle is λ_n, after one distribution, the unmet bandwidth demands of cognitive user *n* is

---
* *Corresponding author* e-mail 1182744430@qq.com

Cong Yuliang, Zhang Shuyang, Xu Limin, Sun Lili

$$US_n = \left(\lambda_n - \sum_{m=1}^{U} a_{n,m} \cdot b_{n,m} \right)^+ , \quad (1)$$

where $(x)^+ = \max(x, 0)$. So the sum of all the cognitive users' unmet bandwidth is

$$US = \sum_{n}^{V} US_n = \sum_{n}^{V} \left(\lambda_n - \sum_{m=1}^{U} a_{n,m} \cdot b_{n,m} \right)^+ . \quad (2)$$

The goal of the proposed algorithm is to take the minimum of the sum of the unmet bandwidth in single cycle,

$$\min \sum_{n}^{V} \left(\lambda_n - \sum_{m=1}^{U} a_{n,m} \cdot b_{n,m} \right)^+ . \quad (3)$$

4 The design of allocation algorithm

After taking the advanced users' participation in the distribution into account, the proposed algorithm makes an adjustment on the traditional CSGC algorithm. In the traditional CSGC algorithm, all nodes simultaneously participate in the distribution, and then the band label corresponding to each node in the figure will be calculated. The user corresponding to the largest label will be assigned to the band, until all the bands are allocated. This article considers the advanced users' involvement in spectrum allocation, which is different from CSGC algorithm. First of all, the advanced users participate in the spectrum distribution in accordance with the priority order. For example, when the advanced user s participates in spectrum allocation, make a label according to its colour (spectrum) in the list based on the corresponding criteria of CSGC algorithm, and allocate the spectrum to the user s with the largest label. And then calculate its benefits. If it does not meet the demands, then continue to allocate spectrum until meet the minimum bandwidth requirements. Delete the nodes interfering with the user s and the colour (spectrum) has been assigned to the user s in the list. The following step is to allocate spectrum to the advanced user with a lower priority than user s. After the assignment for the advanced users, the ordinary users take part in the spectrum allocation. The method is to allocate spectrum corresponding to label criterion among the ordinary users. To be fair, the advanced users are not allowed to participate at this time. Calculate the unmet basic bandwidth demands of a related user after assigned a spectrum. Until all the bandwidth demands meet the basic needs of the users, the rest of the spectrums are allocated to all users (for advanced users and ordinary users), then the CSGC algorithm will be conducted till the rest of the spectrums are allocated.

The allocation steps of the algorithm are:

Step 1: According to the users' information, the cognitive users are divided into advanced users and ordinary users;

Step 2: The advanced user are calibrated priorities, in an order of priority , set a node as a unit and then select the colour that can bring maximum benefit to the user in the list of available colours, and assign to the user;

Step 3: Update topology and delete the colour, which is corresponding to the node interference with the assigned node in the colour list, and delete the nodes whose colour is empty;

Step 4: Make sure whether the advanced users meet the bandwidth requirements or not, if yes, then go to Step 5; if not, judge if the chart is empty. If it is empty, then go to the end; if it's not empty, judge whether the list of available spectrum is empty or not, if it is empty ,then go to step 5; if it's not empty, return to step 3;

Step 5: Allocate spectrum for ordinary users, and the advanced users do not take part in the allocation. Calculate the colour label of each node (ordinary user). The corresponding colour (spectrum) is assigned to the maximum value of the node labels. Then delete the corresponding colour and nodes whose colour list is empty. Determine whether it meets, the user's needs, the users whose demands are meted exit the distribution and wait to allocate with the senior users;

Step 6: If the graph is not empty, then return to step 5; if it is empty and all ordinary users are satisfied, then go to step 7. If not satisfied, then go to step 10;

Step 7: All cognitive users (including advanced users and ordinary users) participate in spectrum allocation accordance with the corresponding algorithm;

Step 8: Update topology, delete the colour, which is corresponding to the node interference with the assigned node in the colour list, and delete the nodes whose colour is empty;

Step 9: If the graph is not empty, then return to step 7; if it is empty, then go to step 10

Step10; The spectrum allocation ends up.

5 Analysis of algorithm simulation

To test the performance of the improved algorithm when advanced users exist, simulation and analysis between the improved algorithm and CSGC algorithm will be taken. The paper mainly discussed the bandwidth demand for advanced users and ordinary users, as well as the total bandwidth of improved algorithm and CSGC algorithm when advanced users exist with or without user priority.

This algorithm mainly considers the spectrum allocation under the condition that advanced users exist, and the simulation parameters are shown in Table 1. The bandwidth need of users mainly depends on its type of business. Different business types are corresponding to different bandwidth requirements. According to IEEE802.22, 7 kinds of bandwidth demand levels can be obtained, as shown in Table 3.

TABLE 1 Simulation Parameters Settings

The total number of users (Number of nodes)	10
The number of advanced users	3
The number of channels (The number of colors)	20
Available matrix **L**	Randomly generated 0, 1 binary matrix
Efficiency matrix **B**	total of 6 level, as shown in table 4-4
Interference matrix collection **C**	Each matrix is the randomly generated 0, 1 binary symmetric matrix
The number of trials	10000

TABLE 2 Business Source Parameters

Name		business	Packet size (bit)	packet rate (packets/s)	The average data rate (kbps)
UGS		VOIP	528	17.561	9.3
rtPS	down	MPEG	1504	1263	1900
	up			126.3	190
nrtPS	do wn	FTP	1536	6.5104	10
	up			1.3021	2

TABLE 3 Parameters of Bandwidth Requirement

Level	Business combination type	Bandwidth requirements per cycle（bit/period）
1	VIOP	9.3×20=186
2	MPEG Upside	190×20=3800
3	VIOP+MPEG Upside	199.3×20=3986
4	MPEG Downside	1900×20=38000
5	VIOP+MPEG Downside	1909.3×20=38186
6	MPEG Upside + Downside	2090×20=41800
7	VIOP+MPEG Upside + Downside	2099.3×20=41986

Make a simulation with the unmet demands of advanced users and the ordinary users, as well as the total bandwidth of the system based on the rule of CMSB.
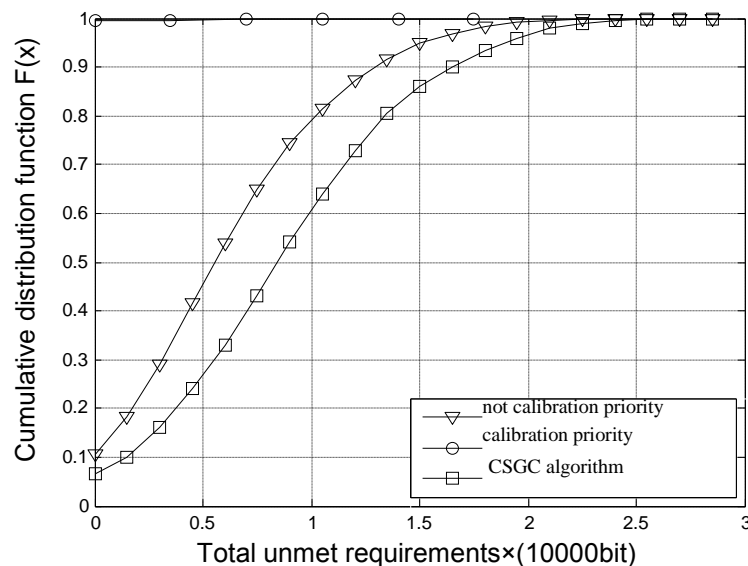
FIGURE 1 Bandwidth unmet for advanced users under the rule of Collaborative Max Sum Bandwidth (CMSB)

Figure 1 shows that the distribution situation that the bandwidth is not satisfied for advanced users under the rule of Collaborative Max Sum Bandwidth (CMSB). It can be seen from the figure that the function value generated by the algorithm based on demand is smaller than the CSGC algorithm without differing ordinary users and advanced users. It indicates that a demand-based algorithm satisfies the needs of ordinary users. After calibration advanced users priority, the needs of advanced users get better satisfied. One of the basic purposes of the calculation is achieved: ensure that the bandwidth requirements for advanced users.

Figure2 shows distribution situation that the bandwidth is not satisfied for ordinary users by Collaborative Max.

Sum Bandwidth, CMSB. It can be seen from the figure that the algorithm based on demand can content the requirement of the users well. After calibration advanced

user priority, the total demand of the users has increased. It follows that if bandwidth demand for advanced users is guaranteed, it has to sacrifice the benefits of ordinary users. Therefore, this article proposed the algorithm controls the number of advanced users. It ensures that the

distribution is reasonable. Although the improvement of total demand is unmet, it is less than CSGC algorithm. In accordance with the purpose of the proposed algorithm, it meets the need of ordinary users.
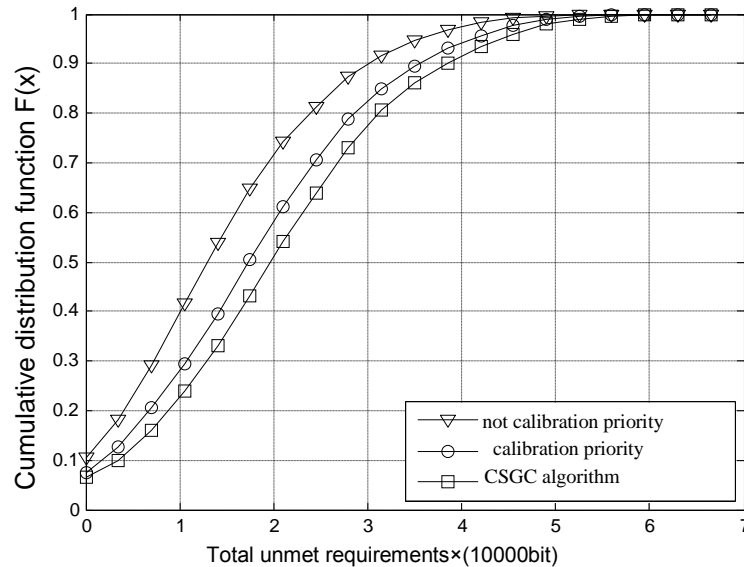
FIGURE 2 Unmet bandwidth for ordinary users under the rule of the total bandwidth under the rule of Collaborative Max Sum Bandwidth, CMSB
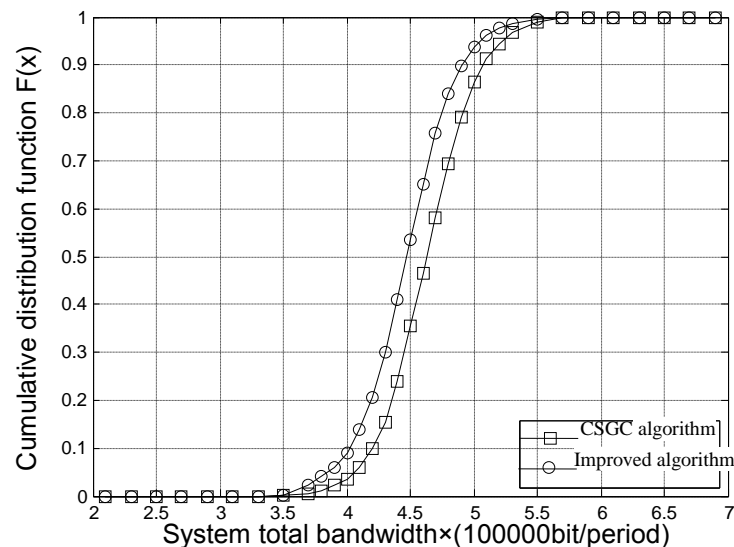
FIGURE 3 Total bandwidth under the rule of Collaborative Max Sum Bandwidth

Figure 3 shows the results of the two algorithms. It can be seen from the figure that the total system bandwidth of the improved algorithm is less than the CSGC algorithm. It suggests that it has to sacrifice the benefits of total system bandwidth to fulfil the needs of users better by the proposed algorithm.

6 Summary

In this paper, the traditional spectrum allocation algorithms had been analysed. The traditional algorithms consider the total system bandwidth or fairness. This paper gives the improved algorithm of spectrum

allocation taking the different communication purposes of different users with the present of demand of advanced users into account. The target of the algorithm is to minimize the unmet demand of the total bandwidth for cognitive user when the advanced users are in normal communication (meet the demand of minimum bandwidth for advanced users in normal communication).

Simulation results show that it can meet the bandwidth requirements for advanced users better. The unmet total demand for bandwidth of the domestic users increased than not prioritized ones. But it is still has a better performance than CSGC algorithm.

References

[1] Mitola J, Gerald Q 1999 Cognitive Radio: Making Software Radios More Personal *IEEE Personal Communications* **1**(6)

[2] Mitola J 2000 *Cognitive radio: An integrated agent architecture for software defined radio* Stockholm, Sweden: Royallnst Technol (KTH)

[3] FCC.03.322. Facilitating Opportunities for Flexible, Efficient, and Reliable Spectrum Use Employing Cognitive Radio Technologies

[4] Haykin Simon 2005 Cognitive Radio: Brain.Empowered Wireless Communications *IEEE Journal Selected Areas in Communications* **23**(2)

[5] IEEE 802.16's License.Exempt (LE) TaskGroup [EB/OL] http://www.ieee802.org/16/le

[6] IEEE 802 LAN/MAN Standards Committee802.22 WG[EB/OL].http://www.ieee802.org/22/

[7] IEEE P1901.1^{TM}/D01.IEEE P1900 Working Group[EB/OL] http://grouper.ieee.org/groups/ eme/eme/1900/index.html

[8] IEEE 802.11h.2003 Part 11. Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specification Amendment 5: Spectrum and Transmit Power Management Extensions in the 5 GHz band in Europe[S], 2003

[9] IEEE P802.11.Task Group Y.Meeting Update: Status of project IEEE 802.11y[EB/OL] http://grouper//.ieee.org/groups/802/11/Reports/thy _update.html

[10] The XG Vision.V2.0 [EB/OL]/ http://www.ir.bbn.com/projects/xmac/vison/html.

[11] End.to.End Reconfiguration, E2R_WP5_D5.3_50727[EB/OL] http://e2r.motlabs.com/Deliv- errables/E2R_WP5_D5.3_50727.pdf

[12] Weissata Jondral F 2004 Spectrum pooling: an innovative strategy for the enhancement of spectrum efficiency *Communication Magazine* **42**(3) 8-14

[13] Rwbroderson Wolisz D.CORVUS: A cognitive radio approach for usage of virtual unlicensed spectrum [EB/OL] http://bwrc.eecs.berkeley.edu/Research/MCMA/CR White paper nall.pdf

[14] Ifakyildiz Y Li 2006 OCRA: OFDM.based cognitive radio networks, Broadband and Wireless Networking Laboratory Technical Report

Authors

Yuliang Cong, born in February, 1966

Current position, grades: professor of College of Communication Engineering, Jilin University.
Experience: was director of National High Technology Research and Development Program 863. She has co-authored Digital signal processing theory and Matlab implementation, which is used as the teaching material in Jilin University Her research area are signal processing and Intelligent navigation.

Shuyang Zhang, born in April, 1989

Current position, grades: Graduate student of, College of Communication Engineering, Jilin University. Mainly research in Spectrum sensing

Limin Xu

Current position, grades: graduate student
University studies: Jilin University
Scientific interests: Spectrum sensing
Experience: Natural Science Fund Project of Jilin: Research on the key technology of spectrum sensing under multiple cognitive users

Lili Sun

Current position, grades: graduate student
University studies: Jilin University
Scientific interests: signal processing and Spectrum sensing
Publications: Cooperative spectrum sensing technology based on the maximum and minimum Eigen value of multi cognitive user Journal of Jilin University Engineering Edition
Experience: Natural Science Fund Project of Jilin: Research on the key technology of spectrum sensing under multiple cognitive users

Approximate completed trace equivalence of real-time linear algebraic Hybrid Automata

Zhi Liu*

Sichuan College of Architectural Technology, China

Abstract

In allusion to design simpler software system, the paper proposes approximate completed trace equivalence of real-time linear algebraic Hybrid Automata. Firstly, it pulls real-time linear algebraic program into Hybrid Automaton and establishes real-time linear algebraic Hybrid Automaton. Next, it uses matrix Frobenius norm to analyse approximation of real-time linear algebraic Hybrid Automata. Afterwards, it gets approximate completed trace equivalence of real-time linear algebraic Hybrid Automata. Finally, the Email virus spreading automata example shows that approximate completed trace equivalence of real-time algebraic Hybrid Automata can simplify automaton.

Keywords: Hybrid Automata, approximate, completed trace equivalence, algebraic program

1 Introduction

Hybrid system [1-4] has continuous variables processes and discrete event processes. They exchange information with each other. Most existing control system can be seen as Hybrid Systems. For example, in the level, control system water level is a continuous quantity and valve switch is discrete quantity. In the room temperature control system, room temperature is a continuous quantity and air conditioning switch is discrete quantity. Completed trace equivalence [5] can effectively reduce states of automaton. Approximate completed trace equivalence theory analysis the approximation of automata and gets completed trace equivalence automaton of approximate automaton. For real-time linear algebraic Hybrid Automata, there is no corresponding approximate completed trace equivalence theory.

This paper is organized as follows. In section 2, it establishes real-time linear algebraic Hybrid Automaton which uses real-time linear algebraic programs to describe actions. In section 3, it uses Frobenius norm to analyse approximation of real-time linear algebraic Hybrid Automata. In section 4, it gets approximate completed trace equivalence of real-time linear algebraic Hybrid Automata. In section 5, the Email virus spreading automata example shows that approximate completed trace equivalence of real-time linear algebraic Hybrid Automata can optimize real-time linear algebraic Hybrid Automaton.

2 Real-time linear algebraic Hybrid Automaton

Definition 1 (Real-Time Linear Algebraic Program): let R be the set of real numbers, $x_i (i = 1, \cdots, n)$ and

$x_i' (i = 1, \cdots, n)$ be the real variables, $t \in \mathrm{R}$ be the time variable. A real-time linear algebraic program is an algebraic program likes $X' = X + (AX + b)t$, where $X' = (x_1', \cdots, x_n')^{\mathrm{T}}$ is the post-state value of real-time linear algebraic program transition, $X = (x_1, \cdots, x_n)^{\mathrm{T}}$ is the pre-state value of real-time linear algebraic program transition.

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} \text{ and } A \neq 0, \ b = (b_1, b_2, \cdots, b_n)^{\mathrm{T}},$$

is n dimensional column vector.

In the real-time linear algebraic program $X' = X + (AX + b)t$, when $X = (x_1, \cdots, x_n)^{\mathrm{T}}$ is given, $AX + b$ is a n dimensional column vector. $X' = (x_1', \cdots, x_n')^{\mathrm{T}}$ changed monotonically along with the time t.

Definition 2 (Real-Time Linear Algebraic Hybrid Automaton): a real-time linear algebraic Hybrid Automaton is a tuple $H = <Q, V, \ HX, Init, Lab, E, Inv, F, R >$, where:

1) Q is a set of system discrete locations.

2) V is a set of system continuous variables.

3) HX is a set of system continuous variables values.

4) $Init \in Q \times HX$ is a set of system initial states. $Init = \{< q_0, X_0 >\}$. It has only one initial state in the real-time linear algebraic Hybrid Automaton.

Corresponding author's e-mail: 1157713577@qq.com

5) *Lab* is a set of discrete transition programs. Discrete transition programs are real number assignment programs $X' = C$.

6) *E* is a set of discrete transitions.

7) *Inv* is a set of continuous variables invariants.

8) *F* is a set of real-time linear algebraic programs $X' = X + (AX + b)t$ which describe system continuous variables dynamic processes.

9) *R* is a set of discrete location transition conditions. In the each discrete location of real-time linear algebraic Hybrid Automata, every element of *X* changed monotonically along with the time *t*. Because each discrete transition program is a real number assignment program, continuous variables value after discrete transition is independent with continuous variables value before discrete transition.

Example 1 (Cattle and sheep breeding law automaton): let x_1, x_2 respectively be cattle number and sheep number, vector $X = (x_1, x_2)^{\mathrm{T}}$ be total number.
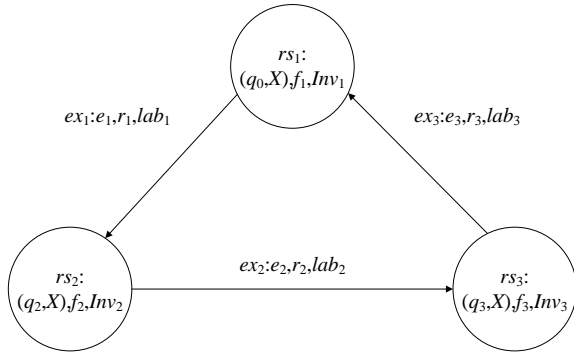
FIGURE 1 Cattle and sheep breeding law automaton

In Figure 1, the system initial state is $< q_0, X_0 >$, where $X_0 = (50, 50)^{\mathrm{T}}$. In the discrete locations q_0, q_2, q_3 , time $t \in [0, 3]$. In the discrete location q_0 the fodder is short, cattle and sheep breeding capacity is weak, real-time linear algebraic program f_1 is $\begin{cases} x_1' = x_1 + (0.2x_1 + 0.1x_2)t \\ x_2' = x_2 + (0.2x_2 + 2)t \end{cases}$. In the discrete location q_2 the fodder is moderate, cattle and sheep breeding capacity is moderate, real-time linear algebraic program f_2 is $\begin{cases} x_1' = x_1 + (0.5x_1 + 0.2x_2 + 1)t \\ x_2' = x_2 + (0.6x_2 + 3)t \end{cases}$. In the discrete location q_3 the fodder is abundant, cattle and sheep breeding capacity is strong, real-time linear algebraic program f_3 is $\begin{cases} x_1' = x_1 + (x_1 + 0.3x_2 + 2)t \\ x_2' = x_2 + (x_2 + 4)t \end{cases}$. Discrete transition programs $lab_i (i = 1, 2, 3)$ are $\begin{cases} x_1' = 50 \\ x_2' = 50 \end{cases}$.

3 Approximation of real-time linear algebraic Hybrid Automata

In the real life, matrix *A* and vector *b* are the result of tool measure. They often have some inevitable errors. Let *A* be the approximate matrix, *b* be the approximate vector, $A + \delta A$ be the actual matrix, $b + \delta b$ be the actual vector, $X' = X + [(A + \delta A)X + (b + \delta b)]t$ be the actual real-time linear algebraic program, $X' = X + (AX + b)t$ be the approximate real-time linear algebraic program. We use Frobenius norm to study approximation of real-time linear algebraic Hybrid Automata.

Matrix $A \in \mathrm{R}^{m \times n}$, the Frobenius norm of matrix *A* is

$$\|A\|_F = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2 \right)^{\frac{1}{2}} .$$

For the actual real-time linear algebraic program $X' = X + [(A + \delta A)X + (b + \delta b)]t$ and approximate real-time linear algebraic program $X' = X + (AX + b)t$, they can be respectively written as $X' = X + \begin{pmatrix} A + \delta A & b + \delta b \end{pmatrix} \begin{pmatrix} X \\ 1 \end{pmatrix} t$ and

$X' = X + \begin{pmatrix} A & b \end{pmatrix} \begin{pmatrix} X \\ 1 \end{pmatrix} t$. Approximation of actual real-time linear algebraic program and approximate real-time linear algebraic program can be converted to approximation of matrix $\begin{pmatrix} A + \delta A & b + \delta b \end{pmatrix}$ and matrix $\begin{pmatrix} A & b \end{pmatrix}$.

Let $B + \delta B = \begin{pmatrix} A + \delta A & b + \delta b \end{pmatrix}$, $B = \begin{pmatrix} A & b \end{pmatrix}$, $\delta B = \begin{pmatrix} \delta A & \delta b \end{pmatrix}$. We use matrix Frobenius norm to analysis approximation of matrix $B + \delta B$ and matrix *B*.

$$\frac{\|(B + \delta B) - B\|_F}{\|B\|_F} = \frac{\|\delta B\|_F}{\|B\|_F} = \frac{\left(\sum_{i=1}^{m} \sum_{j=1}^{n} |\delta b_{ij}|^2 \right)^{\frac{1}{2}}}{\left(\sum_{i=1}^{m} \sum_{j=1}^{n} |b_{ij}|^2 \right)^{\frac{1}{2}}} = W .$$

For a given positive number ε , if it has $W < \varepsilon$, then matrix $\begin{pmatrix} A + \delta A & b + \delta b \end{pmatrix}$ and matrix $\begin{pmatrix} A & b \end{pmatrix}$ are approximate about ε . For a given positive number ε , if it has $W \geq \varepsilon$, then matrix $\begin{pmatrix} A + \delta A & b + \delta b \end{pmatrix}$ and matrix $\begin{pmatrix} A & b \end{pmatrix}$ are not approximate about ε .

Definition 3: for a given positive number ε , if matrix $\begin{pmatrix} A + \delta A & b + \delta b \end{pmatrix}$ and matrix $\begin{pmatrix} A & b \end{pmatrix}$ are approximate about ε , then actual real-time linear algebraic program $X' = X + [(A + \delta A)X + (b + \delta b)]t$ and approximate real-time linear algebraic program $X' = X + (AX + b)t$ are approximate. For a given positive number ε , if matrix

$(A+\delta A \quad b+\delta b)$ and matrix $(A \quad b)$ are not approximate about ε, then actual real-time linear algebraic program $X' = X + [(A+\delta A)X + (b+\delta b)]t$ and approximate real-time linear algebraic program $X' = X + (AX+b)t$ are not approximate.

Definition 4: if all corresponding real-time linear algebraic programs in two corresponding traces of two real-time linear algebraic Hybrid Automata are approximate, then two traces are approximate.

Definition 5: if all corresponding traces of two real-time linear algebraic Hybrid Automata are approximate, then two real-time linear algebraic Hybrid Automata are approximate.

When real-time linear algebraic programs of real-time linear algebraic Hybrid Automata are approximate, continuous variables values are also accordingly changed. The continuous variables invariants and discrete transition condition in each discrete location also can occurrence change. Approximation of real-time linear algebraic Hybrid Automata can optimize real-time linear algebraic programs and simplify real-time linear algebraic Hybrid Automaton. It also can increase mathematical computation speed. It gets the approximate automaton of actual real-time linear algebraic Hybrid Automaton by approximation of real-time linear algebraic Hybrid Automata.

4 Approximate completed trace equivalence of real-time linear algebraic Hybrid Automata

It is known that an actual real-time linear algebraic Hybrid Automaton H_1. We can get approximate real-time linear algebraic Hybrid Automaton H_2 of actual real-time linear algebraic Hybrid Automaton H_1 by approximation of real-time linear algebraic Hybrid Automata. And then, we get completed trace equivalence Automaton H_3 of approximate real-time linear algebraic Hybrid Automaton H_2 by completed trace equivalence theory. H_1 and H_3 are approximate completed trace equivalence. It uses H_3 to replace H_1.
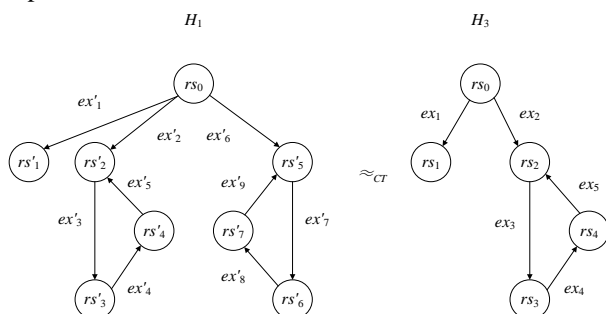
FIGURE 2 An approximate completed trace equivalence of real-time linear algebraic Hybrid Automata example

In Figure 2, we can use approximate method and completed trace equivalence theory to get approximate completed trace equivalence of real-time linear algebraic Hybrid Automata. Approximate completed trace equivalence of real-time linear algebraic Hybrid Automata can optimize real-time linear algebraic programs and

reduce states of real-time linear algebraic Hybrid Automaton.

5 Experiments

In the Email of computer Internet, some Emails often have viruses. They infect other computers and increases virus number. Let Email group have four computer, where three computers send Email with each other, one computer sends Email to other three computer and it can not receive Email. We can only consider the influence of communication frequency high, medium, low three cases and open Email number. Three computers Email virus number respectively be x_1, x_2, x_3. It uses $X = (x_1, x_2, x_3)^T$ to express three computers Email virus total number. The other one computer Email virus number is always invariable. It sets 15 days as a cycle, previous 5 days communication frequency is low, middle 5 days communication frequency is medium, latter 5 days communication frequency is high. In the end of each 5 days disinfect three computers and the leaving virus number is the initial virus number. In the Email virus spreading automaton open Email number respectively are four times and five times.
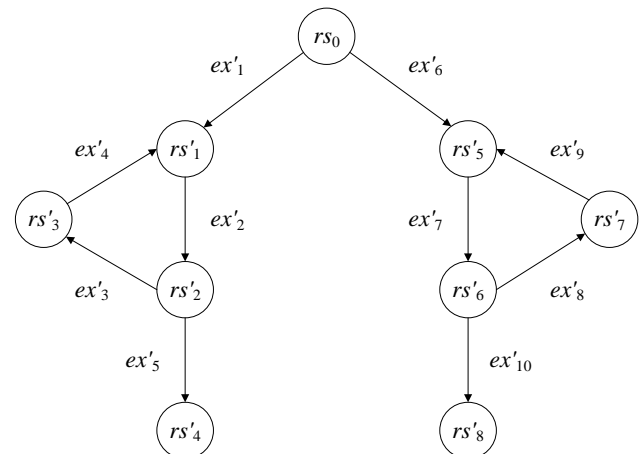
FIGURE 3 Actual Email virus spreading automaton

In Figure 3, left branch is open Email four times and right branch is open Email five times. Initial Email virus number of Email virus spreading automaton is $X_0 = (2, 2, 2)^T$.

The actual real-time linear algebraic program, which describe system variables dynamic process

f'_1 is $X' = X + \left[\begin{pmatrix} 1.95 & 0.97 & 0.98 \\ 0.97 & 1.96 & 0.98 \\ 0.99 & 0.97 & 1.98 \end{pmatrix} X + \begin{pmatrix} 0.97 \\ 0.97 \\ 0.97 \end{pmatrix} \right] t$,

f'_2 is $X' = X + \left[\begin{pmatrix} 2.96 & 1.97 & 1.96 \\ 1.99 & 2.98 & 1.98 \\ 1.94 & 1.99 & 2.97 \end{pmatrix} X + \begin{pmatrix} 1.98 \\ 1.98 \\ 1.98 \end{pmatrix} \right] t$,

f'_3 and f'_4 are

$$X' = X + \left[\begin{pmatrix} 3.94 & 2.97 & 2.99 \\ 2.98 & 3.96 & 2.96 \\ 2.97 & 2.98 & 3.97 \end{pmatrix} X + \begin{pmatrix} 2.98 \\ 2.98 \\ 2.98 \end{pmatrix} \right] t ,$$

f'_5 is $X' = X + \left[\begin{pmatrix} 2.02 & 1.02 & 1.02 \\ 1.03 & 2.01 & 1.02 \\ 1.04 & 1.01 & 2.03 \end{pmatrix} X + \begin{pmatrix} 1.01 \\ 1.01 \\ 1.01 \end{pmatrix} \right] t ,$

f'_6 is $X' = X + \left[\begin{pmatrix} 3.02 & 2.01 & 2.02 \\ 2.03 & 3.05 & 2.02 \\ 2.04 & 2.01 & 3.04 \end{pmatrix} X + \begin{pmatrix} 2.03 \\ 2.03 \\ 2.03 \end{pmatrix} \right] t ,$

f'_7 and f'_8 are

$$X' = X + \left[\begin{pmatrix} 4.04 & 3.02 & 3.06 \\ 3.02 & 4.01 & 3.03 \\ 3.03 & 3.03 & 4.02 \end{pmatrix} X + \begin{pmatrix} 3.03 \\ 3.03 \\ 3.03 \end{pmatrix} \right] t .$$

Discrete transition programs $lab'_i (i = 1, 2, \cdots, 10)$ is $X' = (2, 2, 2)^{\mathrm{T}}$.

The approximate real-time linear algebraic program which describe system variables dynamic process

f_1 and f_5 are $X' = X + \left[\begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix} X + \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \right] t ,$

f_2 and f_6 are $X' = X + \left[\begin{pmatrix} 3 & 2 & 2 \\ 2 & 3 & 2 \\ 2 & 2 & 3 \end{pmatrix} X + \begin{pmatrix} 2 \\ 2 \\ 2 \end{pmatrix} \right] t ,$

f_3, f_4, f_7 and f_8 are $X' = X + \left[\begin{pmatrix} 4 & 3 & 3 \\ 3 & 4 & 3 \\ 3 & 3 & 4 \end{pmatrix} X + \begin{pmatrix} 3 \\ 3 \\ 3 \end{pmatrix} \right] t .$

Discrete transition programs $lab_i (i = 1, 2, \cdots, 10)$ is $X' = (2, 2, 2)^{\mathrm{T}}$.

For the given allowable error $\varepsilon = 0.03$, we get approximate matrix Frobenius norm max relative error $W_{\max} \approx 0.02267 < \varepsilon$. Because corresponding traces of H_1 and H_2 are approximate, H_1 and H_2 are approximate. In the error allowable range, it can use H_2 to replace H_1. Approximation of Email virus spreading automata can optimize real-time linear algebraic programs.

It gets the approximate completed trace equivalence automaton of actual Email virus spreading automaton by approximate completed trace equivalence model of actual Email virus spreading automaton state transition model (Figure 4).
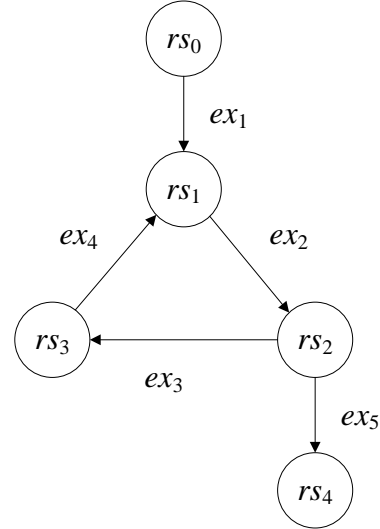
FIGURE 4 Approximate completed trace equivalence automaton of actual Email virus spreading automaton

In the error allowable range, approximate completed trace equivalence automaton of actual Email virus spreading automaton can replace actual Email virus spreading automaton to study virus number change of Email virus spreading model. Approximate completed trace equivalence of real-time linear algebraic Hybrid Automata can optimize real-time linear algebraic programs of actual Email virus spreading automaton and eliminate states of actual Email virus spreading automaton.

6 Conclusion

In this paper, approximate completed trace equivalence of real-time linear algebraic Hybrid Automata is proposed. It can simplify real-time linear algebraic Hybrid Automaton. In the future work, we will use approximate completed trace equivalence of real-time linear algebraic Hybrid Automata to develop software.

Acknowledgments

References

[1] Mansour C, Clodic D 2012 Dynamic modeling of the electro-mechanical configuration of the Toyota Hybrid System series/parallel power train *International Journal of Automotive Technology* **13**(1) 143-66

[2] Suzuki T 2010 Advanced motion as a hybrid system *Electronics and Communications in Japan* **93**(12) 35-43

[3] Allibert B 1999 Analytic controllability of a linear hybrid system *SIAM Journal on Control and Optimization* **37**(3) 844-68

[4] Lauer F, Bloch G, Vidal R 2011 A continuous optimization framework for hybrid system identification *Automatica* **47**(3) 608-13

[5] van Glabbeek R J 1990 The Linear Time – Branching Time Spectrum *CONCUR '90 Theories of Concurrency: Unification and Extension Lectures Notes in Computer Science* Springer-Verlag **458** 278-97

[6] Bauer S S, Fahrenberg U, Juhl L, Larsen K G, Kegay A, Thrane C 2013 Weighted modal transition systems *Formal Methods in System Design* **42**(2) 193-220

[7] Shen M 2013 H∞ filtering of continuous Markov jump linear system with partly known Markov modes and transition probabilities *Journal of the Franklin Institute* **350**(10) 3384-99

[8] Fu J, Tanner H, Heinz J and Chandlee J 2014 *IEEE Transactions on Automatic Control* **59**(2) 505-11

[9] Bernardo M, Nicola N D and Loreti M 2014 Revisiting trace and testing equivalences for nondeterministic and probabilistic processes *Logical Methods in Computer Science* **10**(1:16) 1-42

[10] Cheval V, Cortier V and Delaune S 2013 Deciding equivalence-based properties using constraint solving *Theoretical Computer Science* **492** 1-39

[11] André P 2010 Logical Analysis of Hybrid Systems: Proving Theorems for Complex Dynamics *Springer*

Authors

Zhi Liu, China

Current position, grades: Master's degree.
Scientific interests: symbolic computation.

Completed trace equivalence of inhomogeneous linear algebraic Hybrid Automata

Lei Huang[1, 2*], Zhi Liu[2]

[1]*Chengdu Institute of Computer Application, Chinese Academy of Sciences, China*

[2]*Sichuan College of Architectural Technology, China*

Abstract

In order to reduce states of inhomogeneous linear algebraic Hybrid Automaton, the paper proposes completed trace equivalence of inhomogeneous linear algebraic Hybrid Automata. Firstly, it introduces inhomogeneous linear algebraic programs into Hybrid Automata and establishes inhomogeneous linear algebraic Hybrid Automata. And then, it uses mathematical computation and completed trace equivalence to get completed trace equivalence of inhomogeneous linear algebraic Hybrid Automata. Finally, the travel queue automata example shows that completed trace equivalence of inhomogeneous algebraic Hybrid Automata can reduce states.

Keywords: Hybrid Automata, completed trace equivalence, algebraic program

1 Introduction

The main study of Hybrid Systems [1-4] is a class of dynamical systems, which are constituted by continuity subsystems and discrete subsystems. Dynamic characteristics of continuity subsystems evolve with time. Discrete subsystems dynamic evolution is driven by event. Continuity subsystems and discrete subsystems affect mutually. The movement path of Hybrid Systems on the whole shows the transitions of discrete locations. The movement path of Hybrid Systems on the partial shows continual condition approach evolution. Based on the evolution of continuous variable dynamic systems and discrete event systems, it displays a more complex dynamic behaviour. Hybrid Automata are the most commonly used model in Hybrid Systems modelling. They come by finite state machine in theoretical computer science. Hybrid automata models are suitable for the formal verification of Hybrid Systems and security analysis. The formal verification of Hybrid Systems main uses Hybrid Automata models. Hybrid Automata are finite automata with real continuous variables and they can use graphics to describe Hybrid Systems.

Equivalence model is the system simplified model under certain equivalence criteria. The model, which describes system behaviour sometimes, is very complex. In order to understand the interdependences among the main factor or to express and compute model, people always simplify the original model under certain criteria. The equivalence between thing A and thing B generally refers to A and B have common properties in some areas. When people study these common properties, thing A and thing B are the same thing. Completed trace equivalence [5, 6] is a very common equivalence relation and it can optimize model.

This paper is organized as follows. In section 2, it introduces inhomogeneous linear algebraic programs to Hybrid Automata and establishes inhomogeneous linear algebraic Hybrid Automata. In section 3, it uses numerical calculation method and structure of inhomogeneous linear algebraic Hybrid Automata to analyze equivalence of inhomogeneous linear algebraic Hybrid Automata. It gets the completed trace equivalence of inhomogeneous linear algebraic Hybrid Automata by completed trace equivalence theory. In section 4, the travel queue automaton example shows that completed trace equivalence of inhomogeneous linear algebraic Hybrid Automata can optimize inhomogeneous linear algebraic Hybrid Automata.

2 Inhomogeneous linear algebraic Hybrid Automata

If a Hybrid System is divided by time, then discrete event process and continuous variable dynamic process have two different time mode. They are discrete time mode and continuous time mode. The set of discrete times is presented as $t_b \in T_b$:$\{t_i | i=0,1,\dots\}$. The set of continuous times is presented as $t_c \in T_c$:$[t_0,\infty] \subset R$, R is the set of real numbers. The set of Hybrid Systems times is presented as $T_a = T_b \cup T_c = [t_0,\infty] \subset R$.

Definition (Inhomogeneous Linear Algebraic Hybrid Automata [7]): an inhomogeneous linear algebraic Hybrid Automaton is a tuple $H=<Q, V, HX, Init, Lab, E, Inv, F, R>$, where:

1) Q is a set of system discrete locations.
2) V is a set of system continuous variables.

---

*Corresponding author e-mail: mathhl@sina.com

3) *HX* is a set of system continuous variables values.

4) *Init* $\in Q \times HX$ is a set of system initial states. $Init = \{< q_0, X_0 >\}$. It has only one initial state in the inhomogeneous linear algebraic Hybrid Automata.

5) *Lab* is a set of discrete transition programs. All discrete transition programs are $X' = X$.

6) *E* is a set of discrete transitions.

7) *Inv* is a set of continuous variables invariants.

8) *F* is a set of inhomogeneous linear algebraic programs $X' = AX + b$ which describe system continuous variables dynamic processes.

9) *R* is a set of discrete location transition conditions.

According to the above definition of inhomogeneous linear algebraic Hybrid Automata, the movement process of inhomogeneous linear algebraic Hybrid Automata is as follows:

a) Inhomogeneous linear algebraic Hybrid Automata run the initial state $< q_0, X_0 > \in Init$ as the starting point. The transition processes of inhomogeneous linear algebraic Hybrid Automata include continuous variables dynamic processes and discrete event processes.

b) When the system discrete location is *q*, if system continuous variables value *X* is in the invariant set *Inv*(*q*) of discrete location *q*, then the system carries on continuous evolution. The evolution of continuous variables value *X* follows corresponding inhomogeneous linear algebraic programs. If system continuous variables value *X* is not in the invariant set *Inv*(*q*) of discrete location *q*, then system occurs transition of discrete location.

c) After the discrete location transition, the evolution of continuous variables value *X* follows the new inhomogeneous linear algebraic program. The discrete location is invariable until continuous variables beyond the range of the invariant set.

We can discrete the continuous variables dynamic process of inhomogeneous linear algebraic Hybrid Automata. Continuous time of each discrete location is divided into several same time periods. In each time period, the evolution of continuous variables value *X* follows corresponding inhomogeneous linear algebraic program and the inhomogeneous linear algebraic programs in the same location are same. The elements of continuous variables value *X* in each discrete location are monotonic. Example 1 (Motor speed control automaton): there is a motor speed control automaton and it controls the car acceleration and deceleration by the speed controller. It is designed to control motor speed in 50 kilometers per hour. The rule of speed controller control speed is motor acceleration when motor speed is below 48 kilometers per hour and motor deceleration when speed is above 48 kilometers per hour.
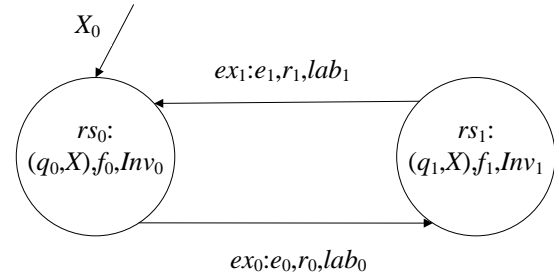
FIGURE 1 Motor speed control automaton

In Figure 1, $X_0 = 48$, initial state is $< q_0, X_0 >$, in the discrete location q_0 motor acceleration, the inhomogeneous linear algebraic program f_0 is $X' = X + 2$ and it expresses motor speed per second increase 2 kilometers per hour, invariant set Inv_0 is $\{X \leq 52\}$, discrete transition condition r_0 is X=52. In the discrete location q_1 motor deceleration, the inhomogeneous linear algebraic program f_1 is $X' = X - 2$ and it expresses motor speed per second decrease 2 kilometers per hour, invariant set Inv_1 is $\{X \geq 48\}$, discrete transition condition r_1 is X=48. The discrete transition program lab_0 and lab_1 are X'=X.

In the state transition model of inhomogeneous linear algebraic Hybrid Automata, a trace is an action sequence $l_1, l_2, \ldots, l_n$ and satisfies an execute sequence $\eta = s_0 l_1, s_1 l_2, \ldots, s_n l_n$. $trace(\eta) = l_1, l_2, \ldots, l_n$. An execute sequence $\eta = s_0 l_1, s_1 l_2, \cdots, s_n l_n$, if $s_0 = s_n$, then this execute sequence constitute a loop. We can be write program $X' = X$ as inhomogeneous linear algebraic program $X' = AX + b$. Let $l_i (i = 1, 2, \cdots, n)$ be $X' = A_i X + b_i$, we can receive:

$$A_1 A_2 \cdots A_n X + A_1 A_2 \cdots A_{n-1} b_n + A_1 A_2 \cdots A_{n-2} b_{n-1} + A_1 b_2 + b_1 = X,$$
$$A_1 A_2 \cdots A_n = E,$$

$$A_1 A_2 \cdots A_{n-1} b_n + A_1 A_2 \cdots A_{n-2} b_{n-1} + A_1 b_2 + b_1 = 0.$$
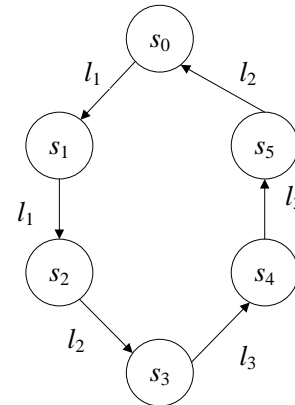
$(l_1 l_1 l_2 l_3 l_3 l_2)^*$

FIGURE 2 State transition model of motor speed control automaton

In Figure 2, it is the state transition model of motor speed control automaton.

3 Completed trace equivalence of inhomogeneous linear algebraic Hybrid Automata

The inhomogeneous linear algebraic Hybrid Automaton has two traces $l_1, l_2, \cdots, l_m$ and $h_1, h_2, \cdots, h_n$, where $l_i (1 \le i \le m)$ and $h_j (1 \le j \le n)$ respectively be $X' = B_i X + b_i$ and $X' = C_j X + c_j$.

As $l_1, l_2, \cdots, l_m$ and $h_1, h_2, \cdots, h_n$ have discrete transition relation and continuous transition relation. The equivalence of $l_1, l_2, \cdots, l_m$ and $h_1, h_2, \cdots, h_n$ not only consider two algebraic programs:

$$X' = B_m B_{m-1} \cdots B_1 X + B_m B_{m-1} \cdots B_2 b_1 + \cdots + B_m b_{m-1} + b_m$$

and

$$X' = C_n C_{n-1} \cdots C_1 X + C_n C_{n-1} \cdots C_2 c_1 + \cdots + C_n c_{n-1} + c_n$$

satisfied mathematical operation condition, but also consider two traces satisfied structure condition. Only when $l_1, l_2, \cdots, l_m$ and $h_1, h_2, \cdots, h_n$ satisfied mathematical operation condition and structure condition, $l_1, l_2, \cdots, l_m$ and $h_1, h_2, \cdots, h_n$ are equivalence.

If $B_m B_{m-1} \cdots B_1 = C_n C_{n-1} \cdots C_1$ and $B_m B_{m-1} \cdots B_2 b_1 + \cdots +$ $B_m b_{m-1} + b_m = C_n C_{n-1} \cdots C_2 c_1 + \cdots + C_n c_{n-1} + c_n$, then $l_1, l_2, \cdots, l_m$ and $h_1, h_2, \cdots, h_n$ satisfied mathematical operation condition.

Two traces equivalence structure condition can be divided into 5 types:
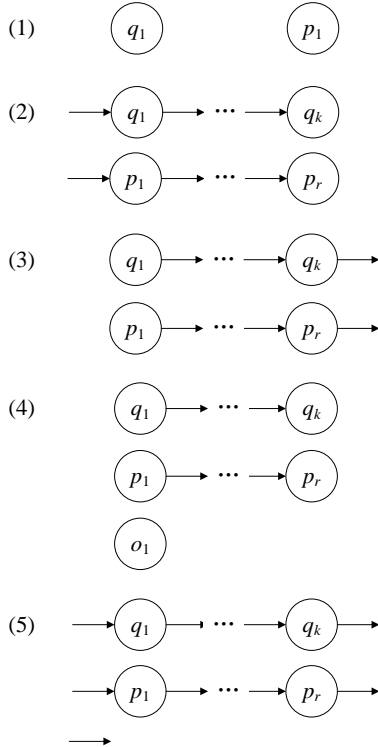
FIGURE 3 Two traces equivalence structure condition

In Figure 3, circles represent continuous variables dynamic process in the discrete locations. Arrows represent discrete transition. In the cases (1-3), two forms satisfied structure condition. In the cases (4) and (5), three forms satisfied structure condition.

We can get the completed trace equivalence automaton of inhomogeneous linear algebraic Hybrid Automata.
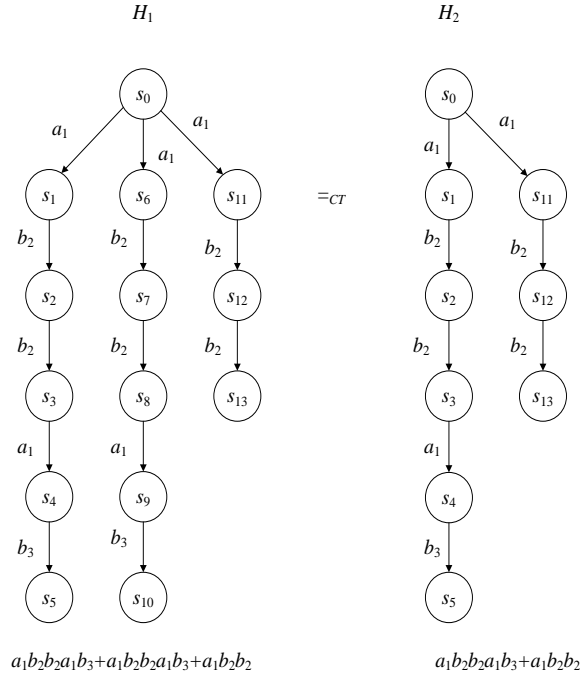
FIGURE 4 An example of completed trace equivalence of inhomogeneous linear algebraic Hybrid Automata state transition model

In Figure 4, H_1 has 14 states and 3 different algebraic programs, H_2 has 9 states and 3 different algebraic programs. H_1 and H_2 have the same function.

Completed trace equivalence of inhomogeneous linear Hybrid Automata can reduce inhomogeneous linear Hybrid Automaton states.

4 Experiments

As the development of the economy, people like travel. The travel queue problem has emerged. We use travel queue automaton example to show the efficient of completed trace equivalence of inhomogeneous linear algebraic Hybrid automata.

The travel queue number is divided into adult number and children number. Many factors affect the travel queue number. We can only consider discrete location duration and weather affect the number. Time period is from 7 am to 10 am. From 7 am to 8 am and from 9 am to 10 am are flat peak period. From 8 am to 9 am is rush peak period.

111
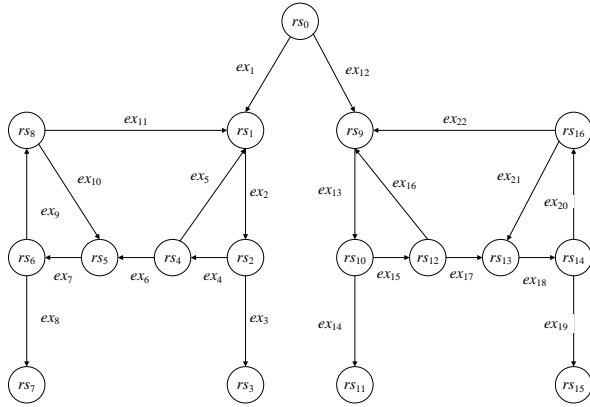
FIGURE 5 Travel queue automaton

The time at 7 am is 0 second. The time from 7 am to 10 am is $t \in [0,10800]$. Adult number and children number respectively are x_1, x_2. $X = (x_1, x_2)^T$ expresses the travel queue number. Let $X_0 = (5,5)^T$. Left side weather is sunny and right side weather is rain. The duration of rs_1, rs_6, rs_9 and rs_{14} is 10 seconds, the duration of rs_2, rs_3, rs_{10} and rs_{11} is 20 seconds, the duration of rs_4, rs_7, rs_{12} and rs_{15} is 30 seconds, the duration of rs_5 and rs_{13} is 5 seconds, the duration of rs_8 and rs_{16} is 45 seconds. One continuous transition time in the $rs_1, rs_2, rs_3, rs_4, rs_5, rs_6, rs_9, rs_{10}, rs_{11},$ $rs_{12}, rs_{13}, rs_{14}$ is 5 seconds, One continuous transition time in the $rs_7, rs_8, rs_{15}, rs_{16}$ is 15 seconds. The inhomogeneous linear algebraic program:

$$f_1, f_2, f_9, f_{10} \text{ is } X' = \begin{pmatrix} \dfrac{5}{2} & -1 \\ \dfrac{1}{2} & 1 \end{pmatrix} X + \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

$$f_3, f_4, f_{11}, f_{12} \text{ is } X' = \begin{pmatrix} \dfrac{1}{3} & \dfrac{1}{3} \\ -\dfrac{1}{6} & \dfrac{5}{6} \end{pmatrix} X + \begin{pmatrix} -\dfrac{2}{3} \\ -\dfrac{2}{3} \end{pmatrix},$$

$$f_5, f_6, f_{13}, f_{14} \text{ is } X' = \begin{pmatrix} \dfrac{23}{4} & -\dfrac{7}{2} \\ \dfrac{7}{4} & \dfrac{1}{2} \end{pmatrix} X + \begin{pmatrix} \dfrac{5}{2} \\ \dfrac{5}{2} \end{pmatrix},$$

$$f_7, f_8, f_{15}, f_{16} \text{ is } X' = \begin{pmatrix} \dfrac{1}{18} & \dfrac{7}{18} \\ -\dfrac{7}{36} & \dfrac{23}{36} \end{pmatrix} X + \begin{pmatrix} -\dfrac{10}{9} \\ -\dfrac{10}{9} \end{pmatrix}.$$

Invariant set $Inv_1, Inv_5, Inv_9, Inv_{13}$ is:

$$\left\{ (5 \le x_1 \le 13.75) \wedge (5 \le x_2 \le 13.75) \right\},$$

$Inv_2, Inv_3, Inv_6, Inv_7, Inv_{10}, Inv_{11}, Inv_{14}, Inv_{15}$ is:

$$\left\{ (13.75 \le x_1 \le 77.734375) \wedge (13.75 \le x_2 \le 77.734375) \right\},$$

$Inv_4, Inv_8, Inv_{12}, Inv_{16}$ is:

$$\left\{ (5 \le x_1 \le 77.734375) \wedge (5 \le x_2 \le 77.734375) \right\}.$$

$$f_1 f_1, f_2 f_2, f_9 f_9, f_{10} f_{10}, f_5, f_6, f_{13}, f_{14}:$$

are equivalent,

$$f_3 f_3, f_4 f_4, f_{11} f_{11}, f_{12} f_{12}, f_7, f_8, f_{15}, f_{16}:$$

are equivalent.

$$f_1 f_1 lab_2 f_2 f_2 f_2 f_2, f_5 lab_7 f_6 f_6, f_9 f_9 lab_{13} f_{10} f_{10} f_{10} f_{10},$$
$$f_{13} lab_{18} f_{14} f_{14}, f_5 f_6 f_6: \text{ are equivalent,}$$

$$lab_3 f_3 f_3 f_3 f_3, lab_8 f_7 f_7, lab_{14} f_{11} f_{11} f_{11} f_{11}, lab_{19} f_{15} f_{15}:$$

are equivalent,

$$f_4 f_4 f_4 f_4 f_4 f_4, f_8 f_8 f_8, f_{12} f_{12} f_{12} f_{12} f_{12} f_{12}, f_{16} f_{16} f_{16}:$$

are equivalent.
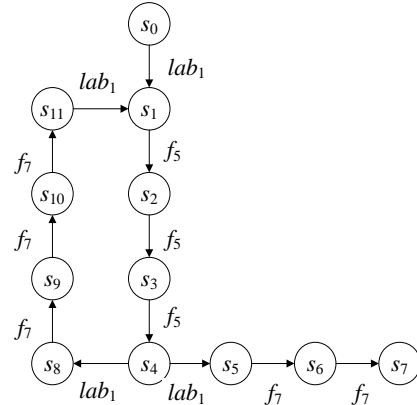
FIGURE 6 Completed trace equivalence state transition model

Using the completed trace equivalence theory, we can get the completed trace equivalence state transition model. The completed trace equivalence automaton of travel queue automaton is shown as Figure 7.
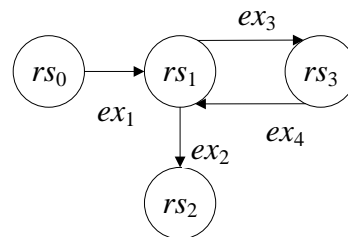
FIGURE 7 Completed trace equivalence automaton of travel queue automaton

From what has been discussion above, completed trace equivalence of inhomogeneous linear algebraic Hybrid Automata can reduce automaton states and optimize inhomogeneous linear algebraic Hybrid Automaton.

5 Conclusion

In this paper, completed trace equivalence of inhomogeneous linear algebraic Hybrid Automata is

proposed. It can optimize inhomogeneous linear algebraic Hybrid Automaton and reduce states. In the future work, we will study completed trace equivalence of nonlinear algebraic Hybrid Automata.

References

[1] Mansour C, Clodic D 2012 Dynamic modeling of the electro-mechanical configuration of the Toyota Hybrid System series/parallel power train *International Journal of Automotive Technology* **13**(1) 143-66

[2] Suzuki T 2010 Advanced motion as a hybrid system. *Electronics and Communications in Japan* **93**(12) 35-43

[3] Allibert B 1999 Analytic controllability of a linear hybrid system *SIAM Journal on Control and Optimization* **37**(3) 844-68

[4] Lauer F, Bloch G, Vidal R 2011 A continuous optimization framework for hybrid system identification *Automatica* **47**(3) 608-13

[5] de Frutos Escrig D, Rodríguez C G, Palomino M 2008 Coinductive Characterisations Reveal Nice Relations Between Preorders and Equivalences *Electronic Notes in Theoretical Computer Science* **212** 149-62

[6] Cheval V, Comon-Lundh H, Delaune S 2011 Trace equivalence decision:Negative tests and non-determinism *18th ACM Conference on Computer and Communications Security (CCS'11)* 21-30

[7] Yang H, Wu J, Zhang Z 2012 Approximate Completed Trace Equivalence of Inhomogeneous Linear Transition Systems *International Journal of Advancements in Computing Technology* **4**(8) 58-66

[8] Fu J, Tanner H, Heinz J and Chandlee J 2014 *IEEE Transactions on Automatic Control* **59**(2) 505-511

[9] Cheval V, Cortier V, Delaune S 2013 Deciding equivalence-based properties using constraint solving *Theoretical Computer Science* **492** 1-39

[10] André P 2010 Logical Analysis of Hybrid Systems: Proving Theorems for Complex Dynamics *Springer*

[11] Bonchi F, Bonsangue M, Caltais G, Rutten J, Silva A 2012 Final semantics for decorated traces *Electronic Notes in Theoretical Computer Science* **286** 73-86

[12] Weidich M, Mendling J 2012 Perceived consistency between process models *Information Systems* **37**(2) 80-98

Acknowledgments

Authors

Huang Lei, China

Current position, grades: doctoral student.
University studies: Chengdu Institute of Computer Application, Chinese Academy of Sciences.
Scientific interests: formal verification, symbolic computation.

Zhi Liu, China

Current position, grades: Master's degree.
Scientific interests: symbolic computation.

Local reconstruction and local fisher discriminant based semi-supervised dimensionality reduction algorithm

Xiangyang Chen*

School of Electronic Engineering, University of Electronic Science and Technology, Chengdu, Sichuan Province, China, 611731

School of Mathematics and Computing Sciences, Guilin University of Electronic Technology, Guilin city, Guangxi Province, China, 541004

Abstract

Local reconstruction and global preserving based semi-supervised dimensionality reduction (LRGPSSDR) algorithm gives no consideration to data locality when processing intra-class relationship and class relationship. Enhanced semi-supervised local fisher discriminant analysis algorithm (ESELF) also neglects locality of data manifold structure when maintaining data manifold structure. To address these problems, the local reconstruction and local fisher discriminant based semi-supervised dimensionality reduction (LRLFSDR) algorithm was proposed in this paper. It depicts significance of sample distance with an improved thermonuclear weight. In this way, intra-class relationship and class relationship of the same cluster attracts more attentions, thus enabling to shorten or widen intra-class distance or class distance firstly. Moreover, it uses idea of LLE algorithm to make neighbourhood linear reconstruction relationship of each point in low-dimensional space to be similar with that in high-dimensional space, which takes locality of data manifold structure into account. Test result confirmed that the proposed LRLFSDR algorithm is superior to other semi-supervised dimensionality reduction algorithms in classifying standard libraries like COIL20, Extended YaleB and CMU PIE.

Keywords: local fisher discriminant, local reconstruction, semi-supervised learning, dimensionality reduction

1 Introduction

"Curse of dimensionality" caused by "small data sample but high dimensionality" is a common problem in machine learning, data mining and pattern recognition. Dimensionality reduction is an effective solution to the "curse of dimensionality" [1-3]. It includes supervised dimensionality reduction and unsupervised dimensionality reduction according to whether training data provides dimensionality reduction algorithm on label. Supervised dimensionality reduction algorithm can make full use of data category structure to apply dimensionality reduced data better in data classification. Linear discriminant analysis (LDA) [4] is a typical supervised dimensionality reduction algorithm. Unsupervised dimensionality reduction algorithm reduces data dimensionality mainly based on data manifold structure or sample relationship. It is mainly used as pre-processing of clustering or supervised dimensionality reduction algorithm. Principal component analysis (PCA) [4] and local linear embedding (LLE) [5] are typical unsupervised dimensionality reduction algorithms.

In most practical application cases (e.g. face recognition and speaker recognition), supervised dimensionality reduction algorithm requires many labelled samples, which is very difficult. On the contrary, it is very easy to acquire unlabelled samples. Under this background, many scholars began to study how to use unlabelled samples to increase effect of supervised dimensionality

reduction algorithm, such as semi-supervised discriminant analysis (SDA) [6] and LRGPSSDA [8].

Advantages and disadvantages of these algorithms are:

1) SDA [6] can embody discriminant information in labelled samples and maintain manifold structure of all samples. However, it is sensitive to neighbourhood parameter setting and limits dimensions of the projection subspace within sample categories.

2) LRGPSSDA [8] has advantages of SDA and overcomes two disadvantages of SDA. It moves faraway sample points farther so that dimensions of the projection subspace are not limited within sample categories basically. However, this is not overcome completely because faraway samples are less significant to classification result in many classification algorithms.

3) LRGPSSDR [7] has advantages of SSDR and overcomes some disadvantages of SSDR. However, it uses global covariance structure to describe negative and positive constraints, but neglects local structure. To address these problems, LRLFSDR algorithm was proposed in this paper. Improved from ESELF and LRGPSSDR, it has two advantages. Firstly, LRLFSDR algorithm requires neighbour linear reconstruction relationship of each point in low-dimensional space to be similar with that in high-dimensional space to maintain data manifold structure. Different from ESELF, it takes locality of data manifold structure into account. Secondly, LRLFSDR algorithm uses weighted covariance matrix instead of intra-class scattering matrix, thus promising

---
Corresponding author e-mail: xiangyangchen0428@163.com

high effect on objective function optimization even under small intra-class sample size. This is different from LRGPSSDR algorithm.

2 Related algorithms

The general form of semi-supervised dimensionality reduction algorithm is given in this section because LRLFSDR algorithm is a semi-supervised dimensionality reduction algorithm. Later, two related algorithms related to the LRLFSDR algorithm are introduced briefly: LRGPSSDR and ESELF.

2.1 GENERAL FORM OF SUPERVISED DIMENSIONLITY REDUCTION ALGORITHM

Most of existing semi-supervised dimensionality reduction algorithms can be depicted by the following general mathematical description. Given n labelled samples $\{x_1, x_2, ... x_n\} \in R^m$ and their labels $\{l_1, l_2, ... l_n\}$ as well as p unlabelled samples $\{x_1, x_2, ... x_p\} \in R^m$, semi-supervised dimensionality reduction algorithm implements mapping of $(X \in R^m) \rightarrow (X \in R^d)$, $d \ll m$ using labelled samples and unlabelled samples. To achieve mapping, most of existing semi-supervised dimensionality reduction algorithms have the following or similar objective function [1-3, 6-8]:

$$W^* = \arg \ \max \frac{W^T SbW}{W^T SwW + \lambda W^T SmW} , \qquad (1)$$

where W is the mapping matrix of data from high-dimensional space to low-dimensional space. $W^T SbW$ is to make data points of different classes in the low-dimensional space as far as possible. $W^T SwW$ is to make data points of same data class in the low-dimensional space as close as possible. $W^T SmW$ is to make data points close to each other in high-dimensional space still close to each other in low-dimensional space. Different semi-supervised dimensionality reduction algorithms have different definitions to above three items or add some new items in Equation (1).

2.2 LRGPSSDR ALGORITHM

The objective function of LRGPSSDR algorithm [7] is:

$$W^* = \arg \ \max_{W^T W - 1} \frac{W^T SbW + \lambda_1 W^T SfW}{W^T SwW + \lambda_2 W^T SmW} , \qquad (2)$$

where $W^T SfW$ is to make data points away from each other in high-dimensional space still far away from each other in low-dimensional space. Rest items are for similar purpose with those in Equation (1). To be more specific:

$$W^T SbW = \sum_{l_i \neq l_j} \left(W^T x_i - W^T x_j \right)^2 , \qquad (3)$$

$$W^T SwW = \sum_{l_i = l_j} \left(W^T x_i - W^T x_j \right)^2 , \qquad (4)$$

$$W^T SfW = \sum_{x_i \notin N(x_j) \ and \ x_i \notin N(x_j)} \left(W^T x_i - W^T x_j \right)^2 , \qquad (5)$$

where $N(x_i)$ is the domain of x_i . It can be known from Equations (3)-(5) that $W^T SbW$, $W^T SwW$ and $W^T SfW$ are the distance sum of different point sets in low-dimensional space. Then, $W^T SmW$ can be defined as:

$$W^T SmW = \sum_{x_i \in N(x_j) \ and \ x_i \in N(x_j)} \left(W^T x_i - W^T x_j \right)^2 , \qquad (6)$$

However, LRGPSSDR algorithm [7] doesn't use Equation (6). This is because Wei Jia stated in Reference [7] that in LLE algorithm [5], Roweis et al. assume that the neighbourhood of input space is locally linear. In other words, each point in high-dimensional space can be reconstructed through linear combination of the points in its neighbourhood and requires similar linear reconstruction relationship in low-dimensional space. Under this hypothesis, LLE algorithm achieved satisfying test result. As a result, Wei Jia [7] redefined $W^T SmW$ based on this hypothesis:

$$W^T SmW = W^T XMX^T W , \qquad (7)$$

where $M = (I - A)^T (I - A)$. A can be calculated from:

$$\varepsilon_1(A) = \sum_i \left\| x_i - \sum_{j:x_j \in N(x_i)} A_{ij} x_j \right\| . \qquad (8)$$

If $\sum_{j:x_j \in N(x_i)} A_{ij} = 1$, the linear relationship between local neighbourhoods (A) can be calculated by least square method.

Based on difference among Equations (3)-(5) and (7), it fails to achieve good dimensionality reduction effect by defining items in Equation (2) with distance sum of different point sets in low-dimensional space. Equations (3)-(6) reflect that distances between samples are equally significant. This will cause evident shortcomings in many cases.

1) For class scattering matrix in low-dimensional space defined by maximum Equation (3), further increase of the distance between two faraway samples is less significant to classification.

2) For intra-class scattering matrix in low-dimensional space defined by minimum Equation (4), further decrease of the distance between two faraway samples is also less significant to classification. When data of same class has many distribution intervals in high-dimensional space, decreasing distance between these data in low-dimensional space will affect optimization of Equation (2) significantly, thus making the mapping matrix not always beneficial to

classification. Similar disadvantages exist in Equations (5) and (6).

The LRGPSSDR algorithm still has most of abovementioned disadvantages although it has overcome some by using Equation (7) instead of Equation (6), most are remained.

2.3 ESELF ALGORTHM

ESELF noticed that distances between different samples shall be given with different weights, which can overcome some disadvantages of LRGPSSDR algorithm. A brief introduction of ESELF algorithm is made in this section. Similarly, the objective function of ESELF algorithm is given:

$$W^* = \arg \max_{W^T W - 1} \frac{W^T SbW - W^T SwW}{W^T StW}. \qquad (9)$$

Reference demonstrates that $W^T SbW$ in (9) is to maximize distance between data points of different classes within certain high-dimensional neighbourhood in low-dimensional distance. $W^T SwW$ is to minimize distance between data points of same class within certain high-dimensional neighbourhood in low-dimensional distance. $W^T StW$ is to make relationships of data points in low-dimensional space similar with those in high-dimensional space. To be more specific:

$$W^T SbW = \sum_{l_i \neq l_j} (W^T x_i - W^T x_j)^2 S_{ij}^b, \qquad (10)$$

$$W^T SwW = \sum_{l_i = l_j} (W^T x_i - W^T x_j)^2 S_{ij}^w, \qquad (11)$$

$$W^T StW = \sum (W^T x_i - W^T x_j)^2, \qquad (12)$$

where S_{ij}^b and S_{ij}^w in Equations (10) and (11) are weight of relationship between i and j. Generally speaking, the further the i and j are, the smaller the S_{ij}^b and S_{ij}^w will be. Equations (10) and (11) give certain consideration to data locality when defining $W^T SbW$ and $W^T SmW$. However, ESELF algorithm still neglects locality of data manifold structure and its data manifold structure in high-dimensional space goes against local linear data.

3 LRLFSDR algorithm

This section introduces the proposed LRLFSDR algorithm which can overcome disadvantages of both ESELF and LRGPSSDR algorithms.

3.1 OBJECTIVE FUNCTION

Objective function of the proposed LRLFSDR algorithm is:

$$W^* = \arg \max_{W^T W - 1} \frac{W^T SbW}{W^T StW + \lambda_2 W^T SmW}, \qquad (13)$$

where $W^T SbW$ is defined in (10) and $W^T SmW$ is defined in (7). $W^T StW$ is defined as:

$$W^T StW = \sum (W^T x_i - W^T x_j)^2 s_{ij}^t, \qquad (14)$$

where $s_{ij}^t = B_{ij}$ and,

$$B_{ij} = \exp\left(\frac{-d^2(x_i, x_j)}{\sigma_i^2 \sigma_j^2}\right), \qquad (15)$$

where σ_i and σ_j are distances from x_i and x_j to their k_c neighbouring samples. Different from SELF algorithm, the proposed LRLFSDR algorithm defines S_{ij}^b in (10) as:

$$S_{ij}^b = \begin{cases} B_{ij} & if \quad , l_i \neq l_j \\ 0, & others \end{cases}. \qquad (16)$$

3.2 THEORETICAL BASIS

a) Equation (13) uses $W^T StW$ instead of $W^T SwW$ in Equation (2). This is because $W^T SwW$ influences denominator less when there are few intra-class training samples. To overcome disadvantages of $W^T SbW$ is added to both numerator and denominator of Equation (1) [6]:

$$W^* = \arg \max_{W^T W - 1} \frac{W^T SbW + W^T SbW}{W^T SwW + W^T SbW + \lambda W^T SmW}. \qquad (17)$$

Equation (17) can be simplified into Equation (13). Therefore, when the numerator is $W^T SbW$, $W^T SwW$ can be minimized by minimizing $W^T StW$.

b) Numerator of Equation (13) involves no $W^T SfW$. This is because $W^T SbW$ and $W^T SfW$ which represent sample distance repeat for many times when there's a small intra-class sample size. When there are abundant intra-class samples, $W^T SfW$ may have to be far away from the sample which belongs to the sample class but is beyond one point domain. This disagrees with the objective of $W^T StW$.

c) Equation (13) keeps data manifold structure by using Equation (7), which can overcome two disadvantages of ESELF algorithm, which uses covariance of all samples to keep data manifold structure.

d) Advantages of $W^T StW$ and $W^T SbW$ definitions in the proposed LRLFSDR algorithm are analysed. Weight of samples B_{ij} has various definitions [6-8]

$$B_{ij} = 1, \qquad (18)$$

$$B_{ij} = \exp\left(\frac{-d^2(x_i, x_j)}{\sigma}\right). \tag{19}$$

Equation (18) is 0-1 weight, while Equation (19) is thermonuclear weight and σ is the variance of thermonuclear.

If B_{ij} adopts 0-1 weight, Equation (10) is equal to (4). In other words, B_{ij} in LRGPSSDR algorithm uses 0-1 weight, which is the cause of its disadvantages according to analysis in Section 2.2.

SDA algorithm can represent sample relationship with thermonuclear weight defined in Equation (19). However, thermonuclear weight has to set σ. Each dataset has only one σ, which is inadequate to describe all sample relationships in a dataset because many datasets have different data distribution patterns in different neighbourhoods.

To address this problem, Lihi Zelnik-Mano represent sample relationship by using Equation (15). It doesn't need to preset parameters and has two advantages. Firstly, weight of sample relationship changes with data distribution. Secondly, data clustering structure is more obvious. When calculating $W^T S t W$ and $W^T S b W$ through such weight of sample relationship, priori attention will be paid to increase distance of samples belonging to different classes of same cluster during optimizing Equation (13). Moreover, attention also will be paid to distance between samples belonging to sample class in same cluster. As a result, samples of same class can form clusters in low-dimensional space rather than accumulate all samples together when optimizing Equation (13), thus making Equation (13) more conducive to data classification. The proposed LRLFSDR algorithm can overcome disadvantages of LRGPSSDR algorithm in Section 2.2.

Based on previous analysis, we can conclude that $W^T S b W$ in Equation (13) is mainly to maximize distance of data points belonging to different classes within certain high-dimensional cluster in low-dimensional space. $W^T S t W$ is mainly to minimize distance of data points belonging to different classes within certain high-dimensional cluster in low-dimensional space. $W^T S m W$ is mainly to maintain the local linearity in high-dimensional space to low-dimensional space. Additionally, weights of sample relationship within certain cluster in $W^T S t W$ and $W^T S b W$ are set zero, which maintains a stable global data structure. Apparently, objective of the proposed LRLFSDR algorithm is different from that of LRGPSSDR and ESELF algorithms. It overcomes disadvantages of both LRGPSSDR and ESELF algorithms.

3.3 OPTIMIZATION OF OBJECTIVE FUNCTION

Optimization of Equation (13) is a generalized rayleigh quotient problem. If $S_t + \lambda_2 S_m$ is non-singular, solution of

Equation (13) is the eigenvector corresponding to the maximum generalized eigenvalue of Equation (20):

$$S_b w = \lambda(S_t + \lambda_2 S_m)w. \tag{20}$$

According to Equation (7),

$$Sm = XMX^T. \tag{21}$$

Then S_b and S_t can be calculated from Equations (10) and (14). For Equation (10),

$$W^T SbW = \sum_{l_i \neq l_j} \left(W^T x_i - W^T x_j\right)^2 S_{ij}^b = $$
$$2\sum_i W^T x_i D_{ii}^b x_i W - 2\sum_{ij} W^T x_j D_{ij}^b x_j W = $$
$$2W^T X \left(D^b - S^b\right) X^T W = 2W^T X L^b X^T W, $$
$$D_{ii}^b = \sum_j S_{ij}^b, L^b = D^b - S^b. \tag{22}$$

S_b can be calculated:

$$S^b = XL^b X^T. \tag{23}$$

For Equation (14),

$$W^T StW = \sum \left(W^T x_i - W^T x_j\right)^2 S_{ij}^b = 2W^T X L^t X^T W, $$
$$D_{ii}^t = \sum_j S_{ij}^t, L^t = D^t - S^t. \tag{24}$$

S_t can be calculated:

$$S^t = XL^t X^T. \tag{25}$$

Therefore, steps of the proposed LRLFSDR algorithm can be concluded:

Algorithm 1 LRLFSDR:

Input: n labelled samples $\{x_1, x_2, ...x_n\} \in R^m$ and their labels $\{l_1, l_2, ...l_n\}$, p unlabelled samples $\{x_1, x_2, ...x_p\} \in R^m$, neighbourhood parameter k_m of Equation (8), and neighbourhood parameter k_c of equation (15).

Output: Mapping matrix W.

Calculate weight of sample relationship (B) from equation (15).

Calculate S_m, S_b and S_t from Equations (21), (23) and (25).

Calculate eigenvalues and eigenvectors in Equation (20) and rank eigenvectors from high eigenvalue to small. Then, the mapping matrix W can be gained.

4 Simulation experiment

Data classification is the main goal of supervised or semi-supervised dimensionality reduction algorithm. In this part, several classification experiments will be conducted to verify the effectiveness of the proposed LRLFSDR algorithm. All experiments use 1-NN classifier. The

proposed LRLFSDR algorithm is compared with SDA [6], PCA [4], LDA [4], LRGPSSDR [7] and ESELF algorithms. Only SDA, LRGPSSDR and the proposed LRLFSDR algorithms have to set parameters. The SDA algorithm can jump over parameter setting by using Equation (15) to represent weight of sample relationship. LRGPSSDR algorithm applies parameters recommended by Reference [7]: $\lambda_1 = \lambda_2 = 0.1$, $k_m = 1$ (the neighbourhood parameter of Equation (8)) and $k_f = 5$ (the neighbourhood parameter of Equation (5)). Parameters of the proposed LRLFSDR algorithm include $\lambda_2 = 0.1$, $k_m = 1$ and $k_c = 5$.

4.1 EXPERIMENTAL DATASET

Three actual datasets are used in this paper.

1) COIL20 image library. It includes 20 objects and each object has 72 images, totally 1,440 images. These images are zoomed into 32×32 in this paper. Therefore, the data dimension is 1024.

2) Extended YaleB. It is a face dataset, including 38*64 256-gray images with different illuminations. Each image is cut into 32×32.

3) CMU PIE. It is also a face dataset, including 41,368 face images of 68 persons. These images are collected under different postures, illuminations and expressions. Each image is cut into 32×32 and identified as 256-gray image.

4.2 EXPERIMENT SETTINGS

The experiment uses three algorithms, namely, unsupervised dimensionality reduction algorithm (e.g. PCA), supervised dimensionality reduction algorithm (e.g. LDA) and semi-supervised dimensionality reduction algorithm (e.g. SDA, LRGPSSDR, ESELF and the proposed LRLFSDR).

During the training process, unsupervised and semi-supervised dimensionality reduction algorithms can use both labelled and unlabelled training samples, while supervised dimensionality reduction algorithms can only use labelled ones. n labelled training samples and unlabelled training samples are selected randomly from each class for each experiment (Table 1). To analyse effect of the amount of unlabelled samples on the proposed LRLFSDR algorithm and other semi-supervised dimensionality reduction algorithms, experiments 1-3 have same amount of labelled samples, but their unlabelled samples increase successively. Experiments 4-6 have more labelled samples in order to discuss effect of the amount of labelled samples on the proposed LRLFSDR algorithm and other semi-supervised dimensionality reduction algorithms. Moreover, the amount of unlabelled samples in experiments 4-6 also changes to maintain same proportion of unlabelled samples with that in experiments 1-3. Each experiment is repeated for 50 times, taking the mean as the final result.

TABLE 1 Amount of labelled and unlabelled samples in different experiments

n	Experiment 1	Experiment 2	Experiment 3	Experiment 4	Experiment 5	Experiment 6
n_l	5	5	5	10	10	10
n_{ul}	3	6	9	5	15	20

4.3 CLASSIFICATION RESULT

Classification result is represented by accuracy in this paper [7]. The calculation formula of accuracy is:

$$precision = \frac{\#(ecorrectly\_recognized\_recordings)}{\#(all\ recognized\_recordings)} . (26)$$

Classification accuracies of Experiments 1-3 under different dimensions are shown in Figure 1. Maximum classification accuracies under different dimensions are listed in Table 2. Figure 1a is Experiment 1's classification accuracy of COIL20 dataset. Figure 1b is Experiment 2 classification accuracy of COIL20 dataset. Figure 1c is Experiment 3 classification accuracy of COIL20 dataset. Figure 1d is Experiment 1 classification accuracy of CMU PIE dataset. Figure 1e is Experiment 2 classification accuracy of CMU PIE dataset. Figure 1f is Experiment 3 classification accuracy of CMU PIE dataset. Figure 1g is Experiment 1's classification accuracy of Extended YaleB dataset. Figure 1h is Experiment 2 classification accuracy of Extended Yale B dataset. Figure 1i is Experiment 3

classification accuracy of Extended Yale B dataset.

Experimental results reveal that the proposed LRLFSDR algorithm achieves better classification accuracy of all datasets compared to other algorithms. This proves effectiveness of the proposed LRLFSDR algorithm. Based on analysis of Experiments 1-3, the proposed LRLFSDR algorithm and LRGPSSDR algorithm improve the semi-supervised dimensionality reduction effect significantly by using certain unlabelled samples. However, when there are more unlabelled samples than labelled ones, further increase of unlabelled samples will greatly lower the classification accuracy of semi-supervised dimensionality reduction algorithms. Hence, the proposed LRLFSDR algorithm and LRGPSSDR algorithm shall choose reasonable amount of unlabelled samples. Moreover, the proposed LRLFSDR algorithm can achieve stable recognition rate at lower dimensions than other algorithms. This indicates that the proposed LRLFSDR algorithm can provide classifier lower-dimensional features, which can accelerate operation of the classifier significantly.
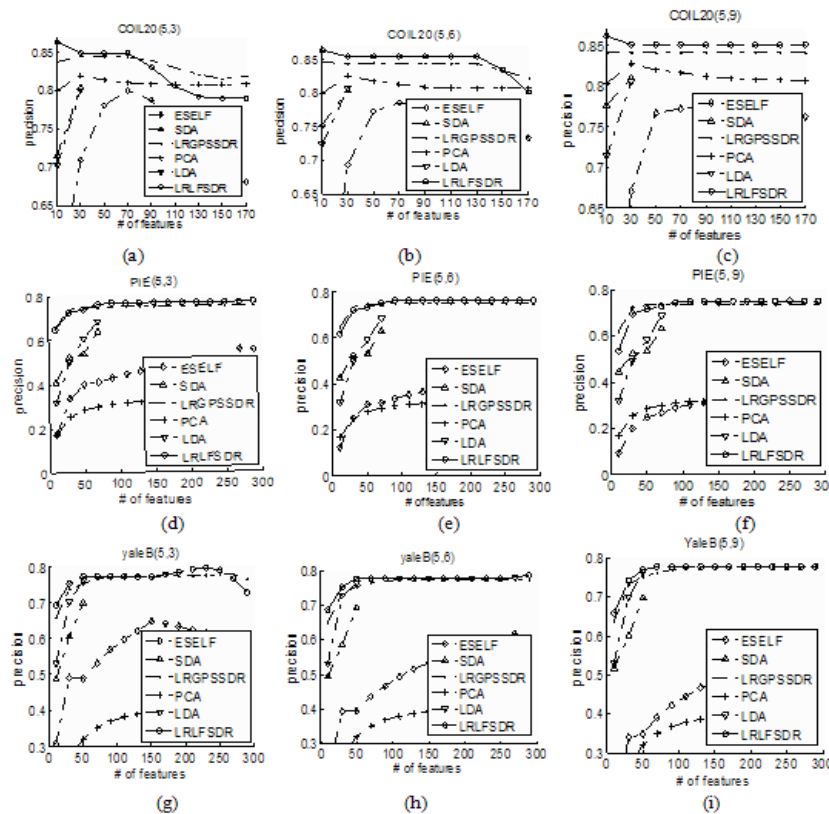
FIGURE 1 Experiments 1-3 classification accuracies of three datasets under different dimensions

TABLE 2 Experiments 1-3's maximum classification accuracies of three datasets

	ESELF	SDA	LRGPSSDR	PCA	LDA	LRLFSDR
COIL20(1)	80.60	80.41	84.44	81.94	80.06	86.32
COIL20(2)	78.66	80.41	84.63	82.61	80.84	86.43
COIL20(3)	77.28	81.04	84.12	82.72	80.52	86.16
PIE(1)	55.17	63.84	75.15	33.08	68.24	76.79
PIE(2)	40.83	63.08	75.15	32.71	68.29	76.45
PIE(3)	34.38	63.21	73.82	32.98	68.48	74.95
YaleB(1)	64.83	70.04	77.96	40.44	76.30	79.66
YaleB(2)	61.74	69.31	77.35	40.48	76.35	78.66
YaleB(3)	54.28	69.74	77.72	40.18	76.31	77.72

Maximum classification accuracies of Experiments 4-6 under different dimensions are listed in Table 3. It can be seen from Table 2 that the proposed LRLFSDR algorithm can offer optimal dimensionality reduction effect under most cases, indicating that it is superior to other algorithms under different amounts of labelled samples.

TABLE 3 Experiments 4-6's maximum classification accuracies of three datasets

	ESELF	SDA	LRGPSSDR	PCA	LDA	LRLFSDR
COIL20(4)	88.36	87.90	92.03	90.00	87.54	92.63
COIL20(5)	86.22	88.93	92.17	89.82	86.86	92.69
COIL20(6)	85.52	89.89	92.07	90.46	87.21	92.43
PIE(4)	61.73	79.94	85.21	47.27	74.01	87.43
PIE(5)	58.84	80.12	84.58	47.18	74.00	86.65
PIE(6)	57.13	80.65	85.26	47.30	74.03	86.32
YaleB(4)	79.57	86.28	90.43	56.19	87.23	90.35
YaleB(5)	70.17	86.57	90.07	56.37	87.05	90.50
YaleB(6)	63.99	86.95	89.92	55.84	87.44	90.25

5 Conclusions

Experimental results demonstrate that the proposed LRLFSDR algorithm can achieve good dimensionality reduction effect under low dimensions. However, it neglects samples beyond the neighbourhood domain when calculating local reconstruction coefficient. Meanwhile, it sets only one neighbourhood parameter to the whole dataset, which neglects locality of data distribution to a certain extent. Therefore, how to involve data locality into the building of reconstruction coefficient requires further research.

References

[1] Song De, Yao L 2013 Semi-supervised Dimensionality Reduction Based on Manifold Structure Preserving and Discriminative Locality Aligment *Microcomputer Applications* **30**(5) 17-20 *(in Chinese)*

[2] Du C, Sun J, Zhou S, Wang L, Zhao J 2013 Dimensionality reduction based on sparse representation and nonparametric discriminant analysis. *Journal of National University of Defense Technology* **35**(2) 143-7 *(in Chinese)*

[3] Liu H, Zhou C 2013 Dimensionality reduction based on sparse representation and nonparametric discriminant analysis *Computer engineering and design* **33**(2) 228-33 *(in Chinese)*

[4] Martinez A M, Kak A C 2001 PCA Versus LDA *IEEE Transactions on Pattern Analysis and Machine Intelligence* **23**(2) 228-33

[5] Roweis S T, Saul L K 2000 Nonlinear dimensionality reduction by locally linear embedding *Science* **290**(5500) 2323-6

[6] Cai D, He X, Han J 2007 Semi-Supervised Discriminant Analysis *IEEE 11th International Conference on Computer Vision 2007 (ICCV 2007)* 1-7

[7] Wei J, Peng H 2008 Local and global preserving based semi-supervised simensionality reduction method *Journal of software* **19**(11) 2833-42

[8] Wei J, Yang C, Ma Q et al. 2010 Semi-supervised siscriminant analysis method based on local reconstruction and global preserving. *Journal of South China University of Technology* **38**(7) 45-9 *(in Chinese)*

Author

Xiangyang Chen, born in January 1972, Guilin, Guangxi Province, P.R. China

Current position, grades: lecturer at Guilin University of Electronic and Technology, Guangxi, China.
Scientific interests: multi-source information fusion and the optimization of sensor management.
Publications: more than 11 papers.
Experience: teaching experience of 16 years, 2 scientific research projects.

DDoS attacks defence strategies based on nonparametric CUSUM algorithm

Changhong Yan[1*], Qin Dong[2], Hong Wang[3]

[1]*School of Information Engineering, Yancheng Institute of Technology, No.9 XiWang Avenue Road, Yancheng, China*

[2]*School of Information Engineering, Yancheng Institute of Technology, No.9 XiWang Avenue Road, Yancheng, China*

[3]*YanCheng junior high school, No.199 The liberation of south Road, Yancheng, China*

Abstract

In the Internet network attacks, distributed denial of service (DDoS) has aroused world attention because of its destructive power. It seems particularly difficult to defend against DDoS attacks for they have characteristics such as abrupt attacks, attacking host computer in a very wide distribution, and so on. To guard against network security and defend distributed denial of service attacks (DDoS), research should begin from the detection of DDos attacks. On the basis of deep research of DDoS attacks, the thesis summarizes and analyses the mechanism and principles of intrusion detection firstly. This paper starts with the analysis of the principle of DDoS attacks. Followed by inquiry and analysis of data packet of DDoS attacks detection, the thesis gives out the computation method for detecting DDos attacks based on Flow Connection density and presents a defending model against DDos attacks based on the temporal series of Flow Connection Condensity (Density). With the defending module based on the temporal series of Flow Connection Condensity (Density), data packet can be effectively filtered so that DDos attacks can be effectively defended and prevented. Finally, experiments prove that the module can effectively filter data packet from network.

Keywords: network security, distributed denial of service, flow connection density, time series, defence strategies

1 Introduction

With the network coming into the Internet era, network security problems appear, of which distributed denial of service attack has great impact on the network security. Distributed Denial of Service, referred to as DDoS, is a kind of denial of service attacks which is offensive and destructive. At present, the Internet is everywhere, and DDoS attacks launched by hackers are everywhere unless we disconnect the network. Distributed denial of service mainly target host node, switch, routers and other network equipment. Tools used by hackers for distributed denial of service are easy to develop. Hackers conduct diversified attacks secretly and the attacking techniques improve day after day which can cause devastating damage even immeasurable losses.

DDoS is widely applied by hackers because it is easy to implement, difficult to prevent and of great harm. In February 2000, unidentified hackers launched a huge-scale distributed denial of service, attacking a series of world-renowned sites such as eBay, yahoo, Microsoft, MSN, Amazon, and so on, and causing numerous system paralyses for several days and significant social economy loss which mounted to billions of dollars. In early 2003, hackers, who will do whatever they can to cause damage, used a new technology of distributed denial of service attacks to damage the Internet in a wide range including North America, Europe and Asia. This not only caused

hundreds of thousands of computers nearly paralyzed and nearly one hundred thousand network servers unable to run, but also resulted in incalculable economic losses and adverse social impact. As for China, on May 19, 2009, DNS resolution system of China Telecom was attacked by a large flow of DDoS, causing a massive network paralysis in telecommunications network of six southern provinces and thousands of web services termination [1].

Distributed denial of service, referred to as DDos, is of destructive power. For an analogy, how can you get through when 10000 people call you at the same time? And DDos attacks are like that. DDos mainly take advantage of the loopholes and shortcomings of network transport protocol - TCP/IP protocol. It chooses computers with scattered network locations as its attacking host to send large amounts of data to the target host. This will not only cause the resources or network bandwidth of attacked hosts consume a lot, but also that the attacked host is so overloaded and paralyzed that it will stop providing normal network services. As a result, legitimate users cannot get access to or use the resources, nor can the attacked host provide any services. Schematic diagram for DDos attacking principle is shown in Figure 1.

---

[*]***Corresponding author*** e-mail: hycit@ycit.cn
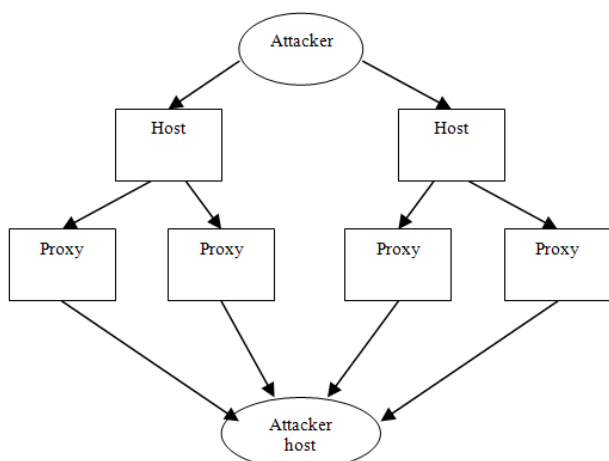
FIGURE1 DDoS attack principle

2 Common DDoS defence strategies analysis

The DDoS attacks on the Internet are becoming increasingly fierce and will continually intensify. Only in 2013, DDoS attacks have emerged in an endless stream in the world, and it seems difficult to measure the amount and size by using statistics. A new IDC study found that prevention solution market for DDoS attacks and DoS attacks is expected to grow 18.2 percent from 2012 to 2017 and related spending will reach $870 million [2].

The substantial harm of DDOS attacks is forcing people to defend against DDOS attacks to minimize the loss. To defend against DDOS attacks, people have explored a variety of DDOS defence strategies from various prospective based on the principle of DDoS attacks. From the DDoS attack principle in Figure1, we can see that core router of the attacked host is the proxy host and it forwards data through the intermediate network routers. Thus, when analysing distributed denial of service, we can divide the whole network into three parts, involving the attacking end of the network, the middle layer and the attacked end. Correspondingly, DDoS attack defence strategies are also divided into the attacking end defence, the middle layer defence and the attacked end defence.

2.1 THE NET WORK DEFENCE IN THE ATTACKING END

In this process, defence node is deployed on the ingress router of the network, then the node counts and analyses the flow based on packet information, which is monitored by the ingress router. Finally, through repeated comparisons between statistics and the normal flow model, dangerous abnormal packets will be filtered out. In this way, we cannot only track information about the attacking end, but also to avoid further damage from outside to the network. There are of course some shortcomings in the attacking end, for instance, if the attacking flow in the attacking end does not converge, it will be difficult to establish normal flow model, which will cause

misjudgement rate of packet increase, resulting in a loss of legitimate data information. Currently, the most typical source-side defence strategy is D-WARD model proposed by Jelena Mirkovic and others [3].

2.2 THE NETWORK DEFENCE IN THE MIDDLE LAYER

Currently, the network defence in the middle layer mainly depends on intrusion detection systems on the network. Intrusion detection systems detect attacks by capturing and analysing network packets. If the network is attacked, it will take measures to correspondingly limit the rate of the attacking data flow. The benefit of this defence strategy is that once an attack is detected, you can quickly suppress the traffic, thus greatly reducing the harm to the attacked end. Disadvantage of network defence strategies in the middle layer is that the data flow on the middle tier network router is large which will not only consume more resources, but also make it difficult to decide whether the data flow is legitimate, ultimately causing damage to legitimate network traffic, and even to the performance of the whole network.

2.3 THE NETWORK DEFENCE IN THE ATTACKED END

The attacked end is direct victim of the DDoS attacks, and it is most immediate, most accurate and most effective to deploy defence system in the attacked end. Thus, it is effective to deploy defence system in the attacked end to defend against DDoS attacks, which is the outstanding advantage of attack end defence. Of course, there are some shortcomings in the attacked end defence strategy. For example, the attacked end is the main attacking target. If the attack is fierce, resulting in paralysis in storage and processing system on the defence node, it may not be able to respond to the defence system deployment to locate the position of the fiercest attack, which will lead to limited response to attacks. To balance its advantages and disadvantages, it is a good choice to deploy the defence system in the end network. The challenge is to find a good defence strategy, making the DDoS attacks dysfunctional so as to defend against DDoS attacks and to secure the attacked host.

3 DDoS attacks defence module and strategies based on nonparametric CUSUM algorithm

To design DDoS defence module based on nonparametric CUSUM algorithm and then make a better defensive strategies to effectively defend against DDoS attacks on the victim port. It is more timely and accurate to detect DDoS attacks by counting and testing traffic of attacked port, so as to defend against DDoS attacks.

3.1 THE DEFENCE BASIS OF NETWORK TRANSMISSION

The data transmission is carried out in the form of data packets on the network. In the Internet, in order to overcome the heterogeneity of the network, and to ensure the correct data transmission, IP protocol defines a unified packet format, which is called an IP datagram. The structure is shown in Figure 2.
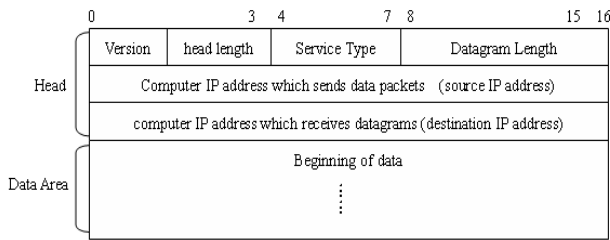
FIGURE 2 IP datagram structure

In the Internet network, any packet transmission, including that is attacked by DDoS, must be organized in the form of IP datagrams. IP datagram must indicate the specified data to be transmitted, the IP address of the computer that sends datagrams and that receives the datagram.

According to the principle of DDoS attacks, DDoS attacks are carried out by a large number of hosts that are geographically dispersed on the network, which send plenty of packets to attack the victim host. Therefore, we can analyse the source IP address, destination IP addresses and port numbers of IP datagram that reaches computers. If large amounts of datagram are from several hosts that are in geographically dispersed network, then it is DDoS attack, which we should guard against.

3.2 FLOW CONNECTION DENSITY

Firstly, in a certain time period, we intercept a collection of identical network data package that contains the same port number of the source address, destination port, and destination port number, and this is called a flow connection, and the amount of the flow connection in this period is called flow connection density. It can be showed in the following form.

We assume that in unit time, the collection of data packets within network traffic is $R = \{p_1,\ldots, p_i,\ldots,p_M\}$, of which, collection of related data packets is $\{R_1,R_2,\ldots,R_N\}$, then we define that in this network traffic, the flow connection density is the amount of related data packets collection, that is N [4-5].

Then we take the following table as an example. Table 1 indicates the obtained IP data packets collection within a unit of time on a certain network.

TABLE 1 Information collection of IP data packets within a unit time on a certain network

Data packets	Source IP address	Destination IP address	Destination port number
P0	s1	D1	port1
P1	s1	D1	port1
P3	s3	D2	port1
A	s3	D1	port1
P4	s2	D2	port2
C	s2	D2	port2
B	s3	D2	port2
R	s3	D2	port1

Through analysis, we conclude that by comparing the same set of IP packets within this network flow per unit time, relevant set of data packets are $\{p_0, p_1\}$, $\{p_4, C\}$, $\{p_3, R\}$. According to the definition, in a time unit, the flow connection density is 3, because there are 3 relevant data packets collection within the time unit of the network. Then, we can decide whether the network traffic is normal based on the flow connection density, so as to determine whether there is distributed denial of service attack.

3.3 CALCULATION FOR IMPROVING FLOW CONNECTION DENSITY BASED ON NON-PARAMETRIC CUSUM ALGORITHM

Through the study of non-parametric CUSUM algorithm [6-8], we can better improve the calculation method of flow connection density. We no longer count the collection of identical network data packages in the unit time, but we calculate the added source IP address in a Δt time to get flow connection density. Then we get a time series $\{Z\}$, which consists of flow connection density sequences within multiple Δt time. The basic idea based on nonparametric CUSUM algorithm to improve flow connection density can be showed by the following Equations (1)-(3):

$$Y_0 = 0,$$
$$Y_n = (Y_{n-1} + Z_n)^+, \tag{1}$$

$$x^+ = \begin{cases} 0, & x \le 0 \\ x, & x > 0 \end{cases}, \tag{2}$$

Y_n is the cumulatively positive value of Z_n, and the decision functions are:

$$d_N(Y_n) = \begin{cases} 1, & Y_n > M \\ 0, & Y_n \le M \end{cases}, \tag{3}$$

where the constant value of detection threshold of DDoS is M, the function $d_N(Y_n)$ represents the judgment result for M in a certain generating time. Through the judgment result, we define that, when the result of function $d_N(Y_n)$ is 1, there are a large number DDoS attacks; when the result is 0, there are no DDoS attacks and the network is normal.

3.4 CALCULATION DDoS ATTACKS DEFENCE MODEL BASED ON NON-PARAMETRIC CUSUM ALGORITHM TO IMPROVE FLOW CONNECTION DENSITY

The DDoS attacks defence model based on nonparametric CUSUM algorithm to improve flow connection density consists of three modules, respectively are acquisition module, time series module and filter module.

- Acquisition module: In this module, IP data packets received by the host are collected in every certain time of Δt, then it comes to data format conversion according to required data format in time series module and filter module and then the data is transmitted to the time series module and filter module in different levels.
- Time series module: In this module, we use the additional received IP data packets within a certain time of Δt to calculate the abstract flow connection density. Then send the final judgment result of function $d_N(Y_n)$, which is based on the judgment result of Equations (1), (2) and (3) to the filter module by combining with the time series of $\{Zn\}$ composed of flow connection density data received in previous time periods of $(n-1)$ Δt.
- Filter module: whether to receive or discard the data packets based on the data information results from processing a data packet and judgment results of function $d_N(Y_n)$ passed from time series module.

DDoS attacks defence model based on nonparametric CUSUM algorithm to improve flow connection density is shown as follows (Figure 3):
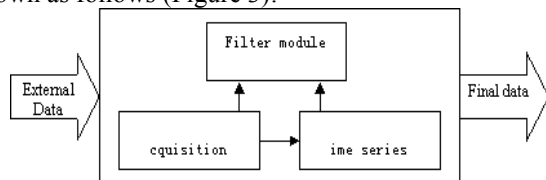
FIGURE 3 DDoS attacks defence model based on nonparametric CUSUM algorithm to improve flow connection density

The mechanism of DDoS attacks defence model based on nonparametric CUSUM algorithm to improve flow connection density is that, the system firstly obtains data within a specified time in the acquisition module, then the data packets received is analysed, including source address, destination address and specific port number of the data packets. Then it comes to data format conversion according to required data format in time series module and filter module, and then the data is transmitted to the time series module and IP packets filter module in different

levels. We use the data in time series module and additional received IP data packets to calculate the specific flow connection density. Then we use the additional collected source IP data packets within a certain time of Δt in time series module to calculate the abstract flow connection density, and calculate its cumulative positive value by combining with the time series of $\{Z_n\}$ composed of flow connection density data received in previous time periods of $(n-1)$ Δt. Then it will send the final judgment result of function $d_N(Y_n)$, which is based on the judgment result of Equations (1)-(3) to the filter module. The filter module will compare the cumulative positive value with the threshold value, if the cumulative positive value is greater than the threshold value, then it will decide to receive or discard the data packets based on the data packets information in acquisition module.

We have conducted a series of experiments and obtained evidence for DDoS attacks defence model based on nonparametric CUSUM algorithm to improve flow connection density. According to the analysis of theoretical and experimental value and comparison between them, the error rate of the experimental values is only 3.618%. It is found out that the wrong results are mainly caused by the external factors such as identification delays and network noise. Of course, these external factors can be resolved by a series of measures. As for identification delay, we can take proper means of improving the sensitivity of the system to reduce recognition delays, thereby reducing the risk of errors. To sum up the experimental results, we can see that DDoS attacks defence model, based on nonparametric CUSUM algorithm to improve flow connection density, can effectively detect DDoS attacks and properly filter out malicious attacking data packets, securing the normal operation of the host or network.

4 Conclusions

This paper gives an intensive study of the operation mode and principle of DDoS attacks, discusses the simple defence strategies against DDoS attacks. It proposes algorithm to detect DDoS attacks based on flow connection density according to characteristics of network data transmission and DDoS attacking principle. Finally, it gives DDoS attacks defence strategy based on the time series analysis of flow connection density. Based on nonparametric CUSUM algorithm, modules are decided to prevent DDoS, which can lead to good defence strategies, defending against DDoS attacks to a great extent.

References

[1] Xu C 2012 Research and implementation of DDoS detecting algorithm in application layer *Master thesis Chongqing university:* Chongqing China *(in Chinese)*

[2] Yang Y 2014 Worries of DDoS attacks *Journal of China education network* **2014**(01) 21-2 *(in Chinese)*

[3] Qi Y, Tang M, Zhang M 2014 Mass customization in flat organization: The mediating role of supply chain planning and corporation coordination *Journal of Applied Research and Technology* **12**(2) 171-81

[4] Yan C 2009 Research and implementation of DDoS attacking detection and defence strategies based on flow connection density *Master thesis* Suzhou University Suzhou China *(in Chinese)*

[5] Zhang C, Huang L, Zhao Z 2013 Research on combination forecast of port cargo throughput based on time series and causality analysis *Journal of Industrial Engineering and Management* **6**(1) 124-34

[6] Takada H, Hofmann U 2004 Application and Analyses of Cumulative Sum to Detect Highly Distributed Denial of Service Attacks using Different Attack Traffic Patterns.Inter-domain *QoS Newsletter* **2004**(7) 414-8

[7] Xiong K, Zhang Y, Zhang Z, Wang S, Zhong Z 2014 PA-NEMO: Proxy mobile IPv6-aided network mobility management scheme for 6LoWPAN *Elektronika ir Elektrotechnika* **20**(3) 98-103

[8] Cai X 2006 Based on the network data acquisition and sequence analysis of the environment of the DDoS attack *Master thesis*, Nanjing university of posts and telecommunications Nanjin China *(in Chinese)*

Authors

Changhong Yan, born in May, 1980, Yancheng, Jiangsu Province, P.R. China

Current position, grades: lecturer.
University studies: MSc in Computer Science and Technology at Suzhou University.
Scientific interests: intermediate lecturer, intelligent control, network security, information processing, remote sensing remote sensing.
Publications: more than 19 papers.
Experience: teaching experience of 12 years, 3 scientific research projects.

Qin Dong, born in October, 1974, Yancheng, Jiangsu Province, P.R. China

Current position, grades: associate professor.
University studies: MSc in Computer Science and Technology at Nanjing University of Science and Technology.
Scientific interests: intermediate lecturer, intelligent control, network security, information processing.
Publications: more than 20 papers.
Experience: teaching experience of 18 years, 10 scientific research projects.

Hong Wang, born in October, 1979, Yancheng, Jiangsu Province, P.R. China

Current position, grades: lecturer at the School of Yancheng, China.
University studies: BSc in Mathematics at Yunnan University in China. MSc at Qinghai Normal University in China.
Scientific interests: education management, mathematics education, network security, information processing.
Publications: 10 papers.
Experience: Teaching experience of 14 years, 2 scientific research projects.

Optimization model of power system unit commitment allocation problem considering the value-point effect and its simulation analysis

Kai Zhang¹, Tingsong Du¹, ²*, Tianbo Wang¹, Wenqing Liu¹

¹Institute of Nonlinear and Complex Systems, China Three Gorges University, Yichang 443002, China

²Hubei Province Key Laboratory of System Science in Metallurgical Process (Wuhan University of Science and Technology), Wuhan 430065, China

Abstract

Based on the studies of the allocation problem of large-scale unit commitment in the power system, a mathematical optimization model is established involving the valve point effect of unit commitment. The optimal solution obtained from the method that the standard artificial fish swarm algorithm (AFSA) is applied to the commitment allocation problem of three-units improves the result recently reported in literature. Considering the visual selection of AFSA affects foraging, huddling and other activities and convergence performance much when increasing the unit size, the proposed improved artificial fish swarm algorithm(IAFSA) in this paper uses the linear decreasing vision function instead of the fixed vision. It can speed up the convergence, jump out of local convergence effectively, and obtain a global optimal solution. Finally, the simulation comparing experiment is conducted for commitment allocation problem of ten-units. The simulation result shows that the IAFSA not only improves the convergence but also enhances the global search capability.

Keywords: unit commitment, valve point effect, linear decreasing

1 Introduction

Large-scale unit combination problem (UC) in the power system is a high-dimensional, discrete, nonlinear engineering optimization problems [1, 2]. It belongs to the category of NP - Hard problem completely. For a long time, these kind of combinatorial optimization problems are widely concerned. As is known to all, when the size of the system is large, it's very difficult to obtain optimum solution accurately in theory. It is very necessary to study the large scale unit combination problem. Because the traditional optimization algorithm in solving the problem of UC all has some defects more or less, it often can not get ideal global optimum solution. The modern intelligent optimization algorithm for handling constraints is convenient and has strong global search ability, so its application in the optimal combination problem has showed strong advantage. Some progress has been made in this field currently. Some modern intelligent algorithms, such as particle swarm optimization (PSO) [3, 4], genetic algorithm (GA) [5], artificial neural network (NN), tabu search (TS), simulated annealing (SA) etc. have been successfully applied to this kind of problem.

However, for the above modern intelligent optimization, the original algorithm also exists some shortage, such as complicated calculation, easy to fall into local optimum and premature convergence. But AFSA applied to solving large scale UC, its research is still rare. AFSA belongs to a class of swarm intelligence

optimization algorithm based on animal behaviour [6, 7]. The AFSA's structure is simple, and it is easy to implement, has fast convergence rate and higher computational efficiency. The tracing behaviour inside is easy to find the global optimal solution of the problem, has a strong ability to avoid falling into local extremum and the search space also have certain adaptive ability. AFSA is not sensitive to initial value and parameter selection, has stronger robustness and better convergence performance. At present, AFSA has been successfully applied to solving a large number of nonlinear, non-differentiable, high-dimensional complex optimization problems, such as optimization of forward neural network, signal processing, TSP, and achieved some progress.

In this paper, AFSA is applied to solve UC. Inspired by the idea to UC in literature [8], if its field of vision in AFSA is too big, convergence speed will be slow and if the vision is too small, artificial fish algorithm easy to fall into local optimal solution. Therefore, IAFSA is proposed in this paper, the main improvement is that the algorithm uses the linear decreasing vision function instead of the fixed vision. The advantage of IAFSA is that early vision of artificial fish is larger, the easier it is to find the global optimal value. With the deepening degree of evolution, vision continuously decreased, and the smaller the field of vision, it is more conductive to get local optimal solution. Therefore, to select the appropriate value of vision can balance the global and local search ability of AFSA, get better approximate

---
*Corresponding author e-mail: tingsongdu@ctgu.edu.cn

solutions. Finally, the standard of AFSA and IAFSA applied to UC of the same problem, the experimental results showed that IAFSA in solving UC has better accuracy and stronger global search ability.

2 Mathematical model of UC considering valve point effect

UC is also known as short-term planning problem, which arrange the units' on and off reasonably to make the system to minimize the total operation cost while meeting all kinds of unit operation conditions in a certain period.

In the actual situation, the turbine inlet valve produces the valve point effect when it is opened suddenly, which affects obviously the solution for the optimal allocation. So, the aim of this paper is a static UC optimization problem considering the valve-point effect. UC optimization model is established with restraint to load equalization and units' output etc., to obtain minimal generation costs.

2.1. OBJECTIVE FUNCTION

$$\min F = \min \sum_{i=1}^{N} F_i(P_i) , \qquad (1)$$

where F is the total electricity costs, N is the total number of generators of the system, P_i is the meritorious power of the i generator and F_i is the electricity cost of the i generator respectively.

Generally, $F_i(P_i)$ can be described as follows:

$$F_i(P_i) = a_i P_i^2 + b_i P_i + c_i , \qquad (2)$$

where a_i, b_i and c_i stand for the i generator running character parameters.

If considering the valve point effect, then

$$F_i(P_i) = a_i P_i^2 + b_i P_i + c_i + E_i , \qquad (3)$$

$$E_i = \left| e_i \times \sin(f_i \times (P_{i\min} - P_i)) \right| , \qquad (4)$$

where e_i and f_i stand for the valve point effect coefficient of the i generator respectively, and $P_{i\min}$ is the minimum output of the i generator.

Research shows that the valve point effect influences the optimal distribution scheme obviously [9].

2.2. CONSTRAINTS

When the system is running, it subjects to the power balance constraints and the power generation unit operation constraints. The power balance constraint is that the active power generator is equal to the sum of the system total network loss and the system total load.

The power balance constraint

$$\sum_{i=1}^{N} P_i = P_L + P_D , \qquad (5)$$

where P_L is the system loading and P_D is the system network loss.

When the electric system network cover is dense we can ignore network loss. According to the thought of the literature [9, 10] and combining with the practical situation, we ignore network loss in this paper. Then, the power balance constraint is simplified as follows:

$$\sum_{i=1}^{N} P_i = P_L . \qquad (6)$$

The power generation unit output constraint

$$P_{i\min} \leq P_i \leq P_{i\max} , \qquad (7)$$

where $P_{i\min}$ and $P_{i\max}$ are the output minimum value and the output maximum value of UC respectively.

In summary, the optimization model of UC considering the valve point effect as follows:

$$\min F = \min \sum_{i=1}^{N} F_i(P_i)$$

$$S.T. \begin{cases} F_i(P_i) = a_i P_i^2 + b_i P_i + c_i + E_i \\ E_i = \left| e_i \times \sin(f_i \times (P_{i\min} - P_i)) \right| . \\ \sum_{i=1}^{N} P_i = P_L \\ P_{i\min} \leq P_i \leq P_{i\max} \end{cases} \qquad (8)$$

3 Description of AFSA

In this paper, the artificial fish are defined four basic behaviours, namely foraging, huddling, following and random behaviour [11, 12]. We simulate four kinds of activities of fish to let them live in the environment.

3.1 RELATED DEFINITIONS

$X = (x_1, x_2, ..., x_n)$ is the state of artificial individual fish, where x_i $(i = 1, 2, ..., n)$ is the variable. $Y = f(X)$ is the food concentration of artificial fish at its current location, where Y is the objective function value. $d_{ij} = \left\| X_i - X_j \right\|$ is the distance of i and j. $step$ is the maximum step

Zhang Kai, Du Tingsong, Wang Tianbo, Liu Wenqing

size of artificial fish moving. *visual* is the vision of artificial fish. *try_number* is the number of attempts. δ is the congestion factor. N is the total number of artificial fish.

3.2 BEHAVIOUR DESCRIPTION

Foraging. Foraging is a kind of the basic behaviour of artificial fish, which is an activity that tends to food. X_i is the current state of artificial fish i. X_j is a state that selects randomly within the artificial fish's sensing.

$$X_j = X_i + Visual \cdot Rand() , \tag{9}$$

where $Rand()$ is a random number between 0 and 1. If Y_i is bigger than Y_j, make a step forward to this direction.

$$X_i^{t+1} = X_i^t + \frac{X_j - X_i^t}{\| X_j - X_i^t \|} \cdot Step \cdot Rand() . \tag{10}$$

Otherwise, select a new state X_j and judge if it meets the conditions of moving forward.

If it still does not meet the conditions after trying *try_number* numbers, make a step forward randomly.

$$X_i^{t+1} = X_i^t + Visual \cdot Rand() . \tag{11}$$

Bunching. X_i is the current state of artificial fish i. n_f is the number of partners and X_c is the centre position in the current field ($d_{ij} < Visual$). If $\frac{Y_c}{n_f} > \delta Y_i$, it shows that there are more food around the partner and it's not too crowded. Then move a step forward to the centre position of this partner

$$X_i^{t+1} = X_i^t + \frac{X_c - X_i^t}{\| X_c - X_i^t \|} \cdot Step \cdot Rand() . \tag{12}$$

Otherwise, perform foraging behaviour.

Tailgating. X_i is the current state of artificial fish i. Y_j is the greatest partner in the current field ($d_{ij} < Visual$).

If $\frac{Y_j}{n_f} > \delta Y_i$, it shows that there are more food around

X_j and it's not too crowded. Then move a step forward to the direction of X_j.

$$X_i^{t+1} = X_i^t + \frac{X_j - X_i^t}{\| X_j - X_i^t \|} \cdot Step \cdot Rand() . \tag{13}$$

Otherwise, perform foraging behaviour.

Random. Random behaviour is to select a state in the field of vision, and then move to the direction. In fact, it is a default behaviour of foraging. The next position of X_i is as follows:

$$X_{i|next} = X_i + Visual \cdot Rand() . \tag{14}$$

4 Three-Unit combination experiments

In order to test AFSA has better convergence speed and higher precision. Comparing the improved particle swarm algorithm and AFSA as given in literature [8] and [13], the result is as follows:

TABLE 1 Comparing the results of three-unit combination

	overall load P (MW)	all-in cost C($)
IPSO[13]	500	5266.48
LIPSO[8]	500	5250.77
AFSA	500	5095.81

where, P is the total load for generators and C is the generating cost. Table 1 shows that solving three-unit UC with AFSA can gain higher accuracy of the approximate optimal solution, which it can balance, the global search ability and local search ability of the algorithm well, and improve the adaptive ability of the algorithm and accuracy.
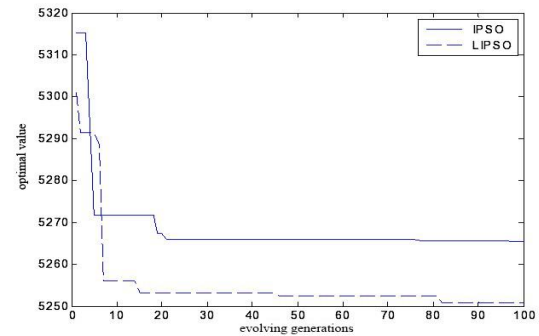
FIGURE 1 Objective function convergence curves of IPSO and LIPSO
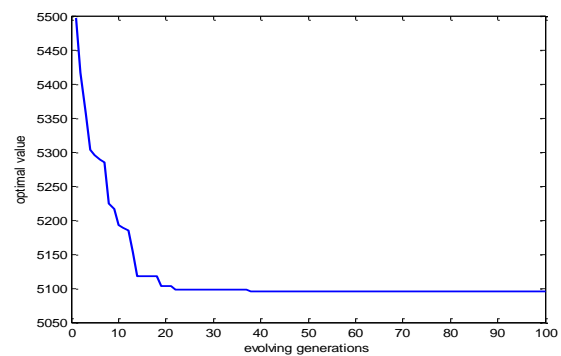
FIGURE 2 Objective function convergence curve of AFSA

Zhang Kai, Du Tingsong, Wang Tianbo, Liu Wenqing

Figure 1 and figure 2 shows that AFSA can obtain objective function value with better accuracy, namely the total electricity cost is less, and especially it can jump out of local convergence more effectively and obtain the global approximate optimal value.

5 AFSA based on improved vision

In the foraging behaviour of the standard AFSA, the vision is fixed. Due to the vision has great influence on each behaviour of the algorithm [14, 15, 16], its change affects the convergence performance complicatedly. Generally speaking, when the field of vision is small, the foraging and random behaviour are more frequent. On the opposite, the following and huddling become more frequent. It is not conducive to find the global extreme point for the artificial fish near the position of the global extreme value when the foraging and random behaviour happened in a large range. In general, greater the field of vision is, easier it is to make artificial fish finding global extreme value and converged. So changing the vision of artificial fish is an efficient way to improve the performance of AFSA.

If the field is too large, the convergence speed will be very slow. On the other hand, if the field of view is too small, AFSA may lead to local optimal solution. In order to overcome these disadvantages, we propose the following improvement strategies.

In the initial stages of AFSA, each artificial fish looks for solutions with a large field of vision, which will expand the scope of optimization. With the generation of algorithm increasing, the vision of fish has appropriately reduction to accelerate the rate of convergence. Therefore, in the initial stages of algorithm, to enhance the global search ability and convergence speed, the artificial fish search in the broader vision hastily with a bigger field of vision. As the search progressing, the field of vision gradually decreases and the algorithm gradually evolved into local search. The algorithm locates at the nearby area of the optimal solution and does fine search so as to improve the local search ability and the accuracy of the optimal value.

So, the vision function defined as follows:

$$f(gen) = V_{\max} - \frac{(V_{\max} - V_{\min})}{MAXGEN} \cdot gen, \quad (15)$$

where, $V_{\max}$, $V_{\min}$ are the upper and lower horizons respectively, gen is the evolution generation currently, $MAXGEN$ is the maximum evolution generation.

The flow chart of IAFSA as follows:
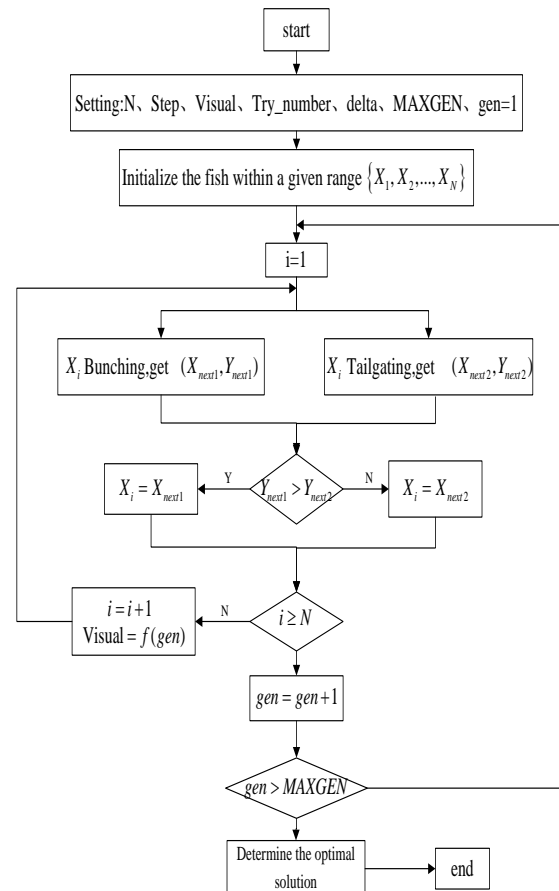
FIGURE 3 The flow chart of IAFSA

Remark 1: After the initialization of the fish, X_i is the current state of artificial fish. When fish simultaneously conduct huddling and following, comparing the objective function values of huddling and following, the algorithm chooses the bigger one.

Remark 2: When the evolution of artificial fish is complete, the algorithm uses the linear decreasing vision function instead of the fixed vision.

6 Simulation experiment

In order to test the effectiveness of the proposed algorithm (IAFSA), the simulation comparing experiment is conducted for the combination experiments of ten-unit. To make the calculation more convenient, this paper ignores system network loss. The characteristic constants of the energy dissipation, the constants of valve point effect, the upper and lower limit of the active power of ten-unit refer to the data of literature 9.

In this experiment, *fishnum* =100, *MAXGEN* =300, *try_number* =100, *delta* =12, *visual* =[6,15], *step* =10, *PL* =2500MW, where *fishnum* is the number of artificial fish, *MAXGEN* is the evolution generation of fish, *try_number* the maximum number of trying, *visual* is the scope of perceived distance, and *PL* is the

Zhang Kai, Du Tingsong, Wang Tianbo, Liu Wenqing

total load of generators respectively.

Take the best optimization calculation results of 30 times as the value of the objective function. The simulation calculation results are shown in table 2.

TABLE 1 Comparison of simulation results of 10-unit

	overall load P (MW)	*all-in cost* C($)
AFSA	2500	6921.91
IAFSA	2500	6844.81

The objective function convergence curve is shown in figure 4.
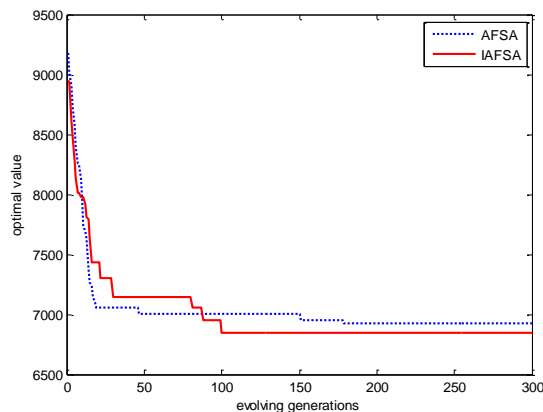
FIGURE 4 Convergence comparison charts of AFSA and IAFSA

The experimental results show that AFSA with a linear decreasing vision is better than the one with a fixed vision, and its iteration speed of IAFSA is faster. It gets the approximate optimal value near the 150th generation, while AFSA gets the approximate optimal value near the 260th generation or so. Compared with AFSA, IAFSA has the better approximate optimal value and faster rate of convergence.

7 Conclusion

This paper reviewed the rules and characteristics of AFSA, IAFSA is proposed, and it was used for solving UC problems. AFSA is a new kind of random search optimization algorithm and this algorithm adopts a bottom-up design pattern. Between the various behaviours there is relative independence and complementarily, so the convergence of the algorithm is stable. The artificial fish's vision function was improved based on AFSA. Through the analysis of the linear decrease vision function, we know that IAFSA can guarantee the algorithm to jump out of local minima areas quickly and speed up the convergence of the algorithm. Finally, the simulation comparison between different scales of UC shows that IAFSA has faster rate of convergence and higher precision while it is applied to solve large-scale UC problems, and the actual effect is satisfactory. It is easy to implement and convenient to be used in electric power systems engineering.

Acknowledgements

References

[1] Yi Sh J, Ma Y, Zhang J N 2007 Overview on Model and Algorithm for Unit Commitment *Northeast Electric Power Technology* **28**(7) 40-2

[2] Chen H Y, Wang X F 1999 A Survey of Optimization Based Methods for Unit Commitment *Automation of Electric Power Systems* **23**(4) 51-6

[3] Trelea I C 2003 The Particle Swarm Optimization Algorithm: Convergence Analysis and Parameter Selection *Information Processing Letters* **85**(6) 317-25

[4] Du T S, Fei P Sh, Shen Y J 2007 A Modified Niche Genetic Algorithm Based on Evolution Gradient and Its Simulation Analysis *IEEE International Conference on Natural Computation, IEEE Press, Haikou, China, August 24-27, 2007* **4** 35-9

[5] Parsopoulos K E, Vrahatis M N 2004 On the Computation of All Global Minimizers through Particle Swarm Optimization *IEEE Transactions on Evolutionary Computation* **8**(3) 211-24

[6] Rocha A M A C, Martins T F M C, Fernandes E M G P 2011 An Augmented Lagrangian Fish Swarm Based Method for Global Optimization *Journal of Computational and Applied Mathematics* **235**(16) 4611-20

[7] Tsai H C, Lin Y H 2011 Modification of the Fish Swarm Algorithm With Particle Swarm Optimization Formulation and Communication Behaviour *Applied Soft Computing* **11**(8) 5367-74

[8] L D Q, Du T S 2012 Improvement on Particle Swarm Optimization Method Solving Unit Commitment Dispatch in the Power System Considering the Valve-point Effect *Proceedings of 2012 the 3rd International Conference on Mechanic Automation*

and Control Engineering, *IEEE Press, Baotou, Inner Mongolia, China, July 27-29,2012* **2** 323-6

[9] Jiang X J, Gong X H, Li C 2008 Using MATLAB to Solve Economic Dispatch Considering the Valve-Point Effect *Journal of Electric Power* **23**(6) 467-9

[10] Xie H D, Gu F H 2013 Research on Particle Swarm Optimization with Application to Power System Unit Commitment *Science Technology and Engineering* **13**(7) 1965-9

[11] Yao X G, Zhou Y Q, Li Y M 2010 Hybrid Algorithm with Artificial Fish Swarm Algorithm and PSO *Application Research of Computers* **27**(6) 2084-6

[12] Li X L, Xue Y C, Lu F, Tian G H 2004 Parameter Estimation Method Based-on Artificial Fish School Algorithm *Journal of Shandong University (Engineering Science)* **34**(3) 84-7

[13] Park J B, Lee K S, Shin J R 2003 Economic Load Dispatch for Non-smooth Cost Functions Using Particle Swarm Optimization *IEEE Power Engineering Society General Meeting, IEEE Press, Toronto, Ontario, Canada, July 13-17, 2003* **2** 938-43

[14] Liu Y J, Jiang M Y 2009 Improved Artificial Fish Swarm Algorithm Based on Adaptive Visual and Step Length *Computer Engineering and Applications* **45**25 35-47

[15] Zhu W L, Jiang J Q, S Ch Y, Bao L Y 2012 Clustering Algorithm Based on Fuzzy C-means and Artificial Fish Swarm *Procedia Engineering* **29** 3307-11

[16] Wu Y P, Du Y 2012 Parameters Analysis of Improved Artificial Fish Swarm Algorithm *Computer Engineering and Applications* **48**(13) 48-52

Authors

Kai Zhang, born on November 28, 1989, Hebei province of China

Current position, grades: Master of computer science and technology, mainly engaged in numerical analysis and computer simulation.
University studies: China Three Gorges University, in the School of Information and Computing Sciences.
Scientific interest: data mining technology and intelligent computing.
Publications: 2 papers.

Tingsong Du, born on September 13, 1969, Hebei province of China

Current position, grades: professor of the Institute of Nonlinear and Complex Systems, China Three Gorges University.
University studies: Department of mathematics Wuhan University, Wuhan, China(1987-1991), B.S. degree in applied mathematics; College of mathematics and computer science Wuhan University, Wuhan, China (1998-2001), M.S. degree in computational mathematics.
Scientific interest: applied mathematics, optimization theory and intelligent computing.
Publications: 48 scientific papers.

Tianbo Wang, born on December 18, 1991, Hebei province of China

Current position, grades: undergraduate student in the China Three Gorges University.
University studies: Bachelor degree will be earned in major of information and computing sciences, China Three Gorges University in 2014.
Scientific interest: numerical analysis and intelligent algorithm.

Wenqing Liu, born on August 26, 1992, Hubei province of China

Current position, grades: undergraduate student in the China Three Gorges University
University studies: Bachelor degree will be earned in major of Information and Computing Sciences, China Three Gorges University in 2014.
Scientific interest: numerical analysis and intelligent algorithm.

Software development for water quality's monitoring centre of wireless sensor network

Liang Li*

Jiangsu Institute of Commerce, Nanjing, Jiangsu, 211168, China

Received 12 May, 2014, www.cmnt.lv

Abstract

Water quality's monitoring centre software of wireless sensor network is designed through applying C#.net and SQL database technology. The monitoring and querying of sensor node data is realized through adopting C/S (Client/Server) pattern. Standard structured query language (SQL) and ADO.NET database access technology are adopted to realize the rapid operation and efficient management on database. Graphical interfaces could display the topology and node status of sensor network, as well as the real-time and history parameters collected by each sensor node and so on. The practice has proved that this software could satisfy the data collection, as well as monitoring and management requirements of wireless sensor network monitoring system.

Keywords: wireless sensor network, water quality parameters, database, c#.net

1 Introduction

Traditional methods of water quality monitoring mainly contain following 2 points: (1) Record analysis is performed through artificial sampling laboratory or data collection is made through adopting portable monitor. (2) Automatic monitoring station of water quality is constructed .The former has low sampling frequency and intense automatic monitoring station; it could not reflect the continuous dynamic changes of water environment. The latter has high investment cost and long construction period. Laying cable and land acquisition for constructing station have influence on the environment of tested waters. In recent years, as the rapid developed technology of wireless sensor network has advantages of flexible node configuration, rapid ad-hoc network, low cost and low power consumption etc, they are very suitable for monitoring outdoor environment parameters [1, 3]. Through aiming at the requirement of wireless sensor network's water quality monitoring system, this paper has developed water quality monitoring centre software on the basis of wireless sensor network technology. C#.net and SQL database have developed water quality monitoring centre software, which is linked with wireless sensor network, and realize network status monitoring, storage of water quality parameters measured by each node and real-time dynamic display. Meanwhile the alarm level of water quality status is set up according to the anomaly degree of data .Control strategy is collected in accordance with corresponding parameters ,by which to realize the requirements of intelligent sampling and security control.

2 Introduction of wireless sensor network

Wireless sensor network [1] (Wireless Sensor Network, WSN) is composed of sensor nodes in different function and quantity, and it has wireless self-organizing ability. The sensor node has wireless communication module, sensor unit, data processing unit and power module etc. Each acquisition node which is dispersed in examination area of wireless sensor network obtains corresponding environmental monitoring data through sensor, such as temperature and humidity, light intensity, and pressure etc. Then corresponding computing process is performed on collected data, and these sensors have the ability of communicating information with their neighbour nodes at the same time [2, 3]. A typical system architecture in wireless sensor network is shown as Figure 1. We could see clearly from figure that it is composed of sensor node, PC computer, aggregation node, Internet or satellite network and task management node.
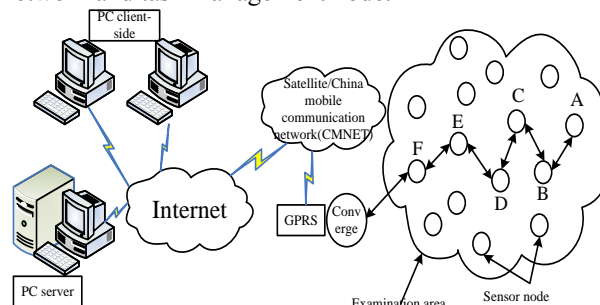
FIGURE 1 Typical architecture of sensor network system

Generally speaking, sensor nodes have relatively weak ability of processing and storing collected data due to its characteristics of structure miniaturization, and

---

* *Corresponding author* e-mail: llliliang_1@163.com

communication capability is also restricted by that[4]. Its main function includes (1) taking charge of obtaining the data information in physical environment. (2) Routing function. It stores the data forwarded by other nodes, performs integration, management and processing. (3) Accomplishing task with other task collaboratively [5]. Integrating node ability of processing, storage and communication is stronger relative to the sensor nodes; it does not needs the ability of physical environment data as sensor node. Its responsible task contains (1) connecting with sensor network through wireless approach, the collected sensor data would be passed on to external network (PC computer) through wireless network cable or network with high reliable communication quality (PC computer). (2) The detection task of management station equipment (it is generally PC computer) would be sent to sensor nodes .As the scale and the covering scope of wireless sensor network is not restricted, intensively distributed sensor could be composed of various sensor, and it has characteristics of miniaturization and micromation. The overall network is self-organizing and reliable, and it could be used in unattended, extremely hostile environment due to its advantage of taking the data as the core.

Wireless sensor network has changed the traditional interactive mode between human and environment. As a kind of comprehensively new method of information acquisition and processing method, it has been widely applied in human activities, such as intelligent furniture and intelligent living environment [6], medical health monitoring and treatment [7], urban transportation and so on in various fields. For example, corresponding number of wireless sensors could be deployed in the place which is easy to occur forest fire for detecting the real-time information in this region, the feedback information could effectively avoid and reduce the loss of human being. Wireless sensor network could also monitor the change of climate and soil environment in farmland, farmers could ensure the crop harvest through timely taking appropriate measures according to this information.

3 The water quality monitoring system construction of wireless sensor network

Water environment monitoring system of wireless sensor network is shown as Figure 2, it includes wireless sensor networks and data monitoring centre. Wireless sensor network installed on the monitoring waters is composed of various water parameters' sensor nodes (sensor node), wireless router nodes (routing node) and aggregation node (sink node) with the capacity of sensing, computing and transmitting wireless data, IEEE802.15.4/ZigBee protocol is adopted to realize self-organizing network between nodes, the quantity of sensor nodes and routing nodes could be set up according to the actual size and monitoring requirements of measured area. Sensor nodes adopt the construction approach of embedded technology,

which could configure sensors like temperature, PH value, DO (Dissolved Oxygen) and so on. Local processing is first performed on the data collected by sensor node, distributed measurement data of various sensor node would connect with data monitoring centre through the approach of asynchronous serial mode or GPRS remote communication way after simple processing at aggregation node. The monitoring centre software receives the wireless sensor network packets sent by aggregation node, and the data would be stored in database after analysis and processing. Analysis result will be represented in the form of graphic interface through analysing data. At the same time, comparison among the data received this time, history data and water quality standard data stored in database are performed to determine whether the data is reasonable and abnormal change would occur or not. Abnormal warning level of water quality status is set up according to the abnormal degree, and control strategy would be gathered according to corresponding parameters [8].
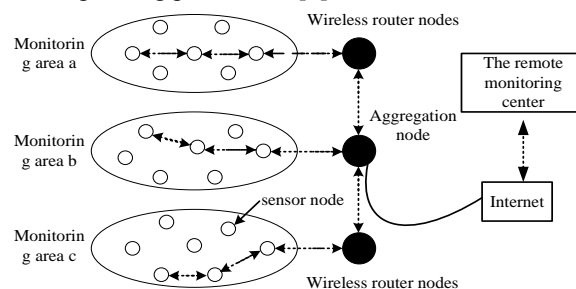
FIGURE 2 Structure of water quality monitoring system in wireless sensor network

4 Function modules of data monitoring center software

As the system shown in Figure 2, the sensor nodes send the collected data through adopting ZigBee protocol based on the triggering mode of combing timing and events. Sink nodes send the data to local monitoring centre in real time through RS232 serial port trunk, the monitoring centre receives and makes analytical processing on the data sent from Sink nodes, send it to local database for storage or remote client-side through Internet, and answer the query from users or application. Compared with the information collection system constructed by B/S (Browser/Server) [9], such C/S (Client/Server) query mode has gotten rid of the frequent confirmation time of sensor network node during customer search, and avoid network congestion etc when simultaneously sending a large number of sensing information, it has the characteristic of fast response speed and strong real-time property.

Data monitoring centre software requires to realize following function: (1) monitoring the serial port, node acquisition parameters, node state, node residual energy, network topology and other information would be obtained from Sink nodes. (2) Check, parse, store and chart shows will be performed on the received data. (3)

Updating, management and maintenance of database (4) Management function of client user: querying and displaying network status, querying and printing node data or history data in certain period of time etc. (5) Analysis, judgment and evaluation would be performed on the water environment of monitoring area through combing with water quality standard database. (6) Parameter settings and system synchronization: such as sampling interval, alarming threshold, releasing synchronization information of system time etc. When performing system's functional division, data would be taken as the main line, modular design method would be adopted to divide the whole procedure into serial port communication module, data processing module, database module, human-computer interaction module and so on, each module would complete the corresponding function. The monitoring centre software adopts C# language and SQL DBD (Database Design) under Visual Studi environment. The procedure flow charts of wireless sensor nodes are shown as Figure 3.
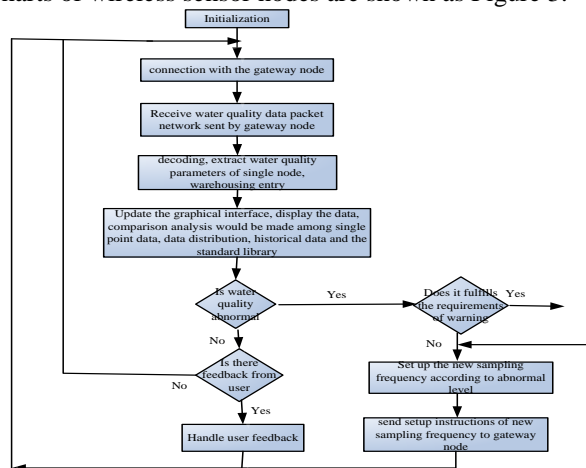
FIGURE 3 Program flow chart of wireless sensor node

5 Implementation technique of key function module

5.1 SERIAL PORT COMMUNICATION MODULE AND DATA PROTOCOL

Communication connection between monitoring centre software and wireless sensor network would be constructed first. Its process is as following: link order would be sent to sink nodes from monitoring centre software. Sink contains the topology information of wireless sensor network. After constructing communication connection, monitoring centre would send data acquisition order according to set acquisition interval. Sink nodes would send confirmation information to monitoring centre after receiving the connection instructions; such information contains the topology information of wireless sensor network. After constructing communication connection, monitoring centre would send the data acquisition order to aggregation node according to set sampling interval. After Sink nodes' receiving data collection order, it

would send such order to the node broadcasting of wireless sensor networks, then wait to receive the node packets of each sensor node. After collecting the data package of all nodes, the data would be converged and integrated into network monitoring data packets and sent to monitoring centre. The application layer protocol data could be formulated according to the actual size and monitoring parameter of wireless sensor network.

The main frame format:

(1) Information data frame of wireless sensor network topology is composed of start code(1 byte), information source address (2 bytes), the number of nodes passing through the information (1byte), the passed routing number of information (10 bytes), check code (1 byte), epilogue code (1 byte).

(2) The data collection command frame sent by monitoring centre is composed of starting byte (1 byte), slave computer address (1 byte), identifier (1 byte), sample time interval (2 bytes), check code (1 byte) and end byte (1 byte).

(3) Parameter data frame of monitoring centre sent by Sink nodes is composed of start code (1 byte), sampling time (6 bytes), node numbering (1 byte), and PH value (2 bytes), dissolved oxygen (4 bytes), temperature (3 bytes), and check code (1 byte), epilogue code (1 byte).

The design of serial port communication program applies serial port control, and it is composed of serial port setup, open port, sending command of serial port as well as reading data of serial port. The reading of serial data adopts interrupt mode. After monitoring centre has received data, check and parsing would be made on data package, useful data would be extracted and turned into displaying data at the end.

5.2 DATABASE MODULE

The design objective of database module is security (database storage structure would be of foreign shielding), universality (all kinds of data services would be provided externally through port), high efficiency (efficiency of data storage and query is enhanced internally). The main module has database structure design module, which mainly takes charge of building up backend database. The database management module takes charge of managing daily database, including establishment of the log, maintenance of the database etc. Database query and analysis module is used to optimize the query of big data table and enhance access efficiency etc.

(1) 1Database structure design

The design of database structure is the key of database module, it focuses on the design table structure and the relationship between tables, which is to store the possibly involved data and related information .In our system, database storage information could be divided into two types: data sheet and the information table. The data sheet is used to store collected data. Table of information is

used to store some basic information, such as temperature, PH, solubility.

(2) The storage and query of data.

This system adopts the technology of ADO. NET database access [10].ADO.NET has two core components: ADO.NET Dataset and ADO.NET Data Provider, namely the provider class (Provider) and the user class (Consumer). The provider class would accomplish the data source reading and write in etc, after the data is read to storage medium, function like access and manipulating of database would be completed by the user class.

The main code of reading data source and write in is as following:

StringContr="DataSoure=CHINA-D3CD2701A\\SQLEX-PRESS; Initial Catalog=db1; Integrated Security=True";//declare a string, it acts as linking data base 'auditing the password'

SqlConnection conn=new SqlConnection(Contr);

conn.Open();//Open the connection channel;

System.Data.SqlClient.SqlCommand cmd=new System.Data.SqlClient.SqlCommand();

cmd.CommandType=System.Data.CommandType.Text;

cmd.CommandText="insert into sensor(ph temperature,temperature,time) values('"+SteA+"','"+StrB+"','"+StrC+"','"+Time+"');cmd.Connection=conn;

cmd.ExecuteNonQuery();//data would be placed into data base through performing SQL interposition predict

5.3 SOFTWARE FUNCTION INTRODUCTION

The main function of the software is briefly introduced as follows:

(1) Warning function: When there is some abnormity detected among temperature, PH, DO (Dissolved Oxygen), the system software will start alarm function immediately and send out a warning message to remind the staff of detection software centre, who could also start off-limit alarm button for sending out warning notice.

(2) Set sampling rate: changing sampling rate is to adjust the time scale of observation data. Under normal status, in order to reduce energy consumption, wireless sensor network could perform data acquisition with low sampling rate. When there is abnormality in parameters, the system would publish anomalous event, sensor node would increase sampling frequency, and then the resolution ratio of data would be enhanced.

(3) Synchronizing function: as the sampling frequency of system is not high, sampling interval is 1 minute at least, thus local computer clock is taken as master clock, and time synchronization information would be sent to other nodes in WSN through aggregation node at regular intervals, the transmission delaying between two nodes would be not more than 100ms and could satisfy the requirement of synchronization.

The monitoring centre software structure is shown as figure 4.
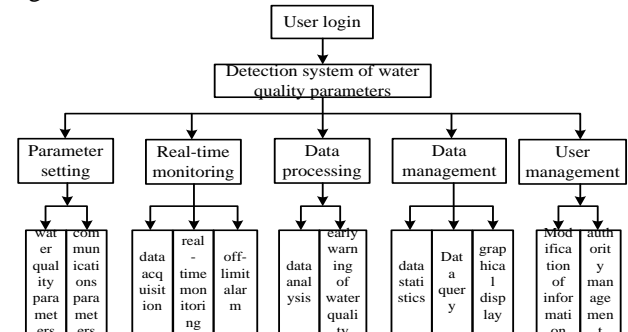
FIGURE 4 Structured flowchart of the monitoring centre software

5.4 THE TEST OF EXPERIMENTAL RESULT

PH value, temperature, dissolved oxygen degree in experimental design is shown as Table 1, the scope is within the range of pond water quality, the data collection interval is 1h. Specific design of experimental design is shown as Table 1, the results of wireless sensor network monitoring is shown as Table 2. The real time monitoring interface of monitoring centre is shown as Figure 5.
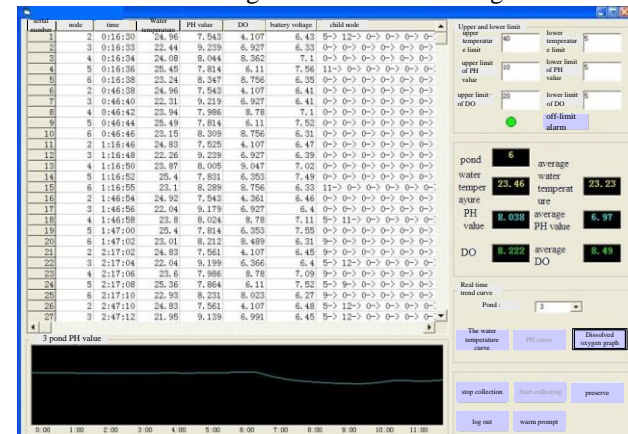
FIGURE 5 Real-time monitoring interfaces

TABLE 1 Artificial simulation test results

Number	1	2	3	4	5	6	7
PH	7.5	7.7	7.2	6.7	8.5	7.9	6.4
Temperature ℃	21	25	23	18	26	27	24
Dissolved oxygen degree %	19	16	27	21	23	14	17

TABLE 2 the actual test results

Number	1	2	3	4	5	6	7
PH	7.1	7.5	6.8	6.9	8.9	7.6	7.2
Temperature (℃)	23.5	26.8	24	19.4	25.4	28.6	23.7
Dissolved oxygen degree (%)	20.5	17.3	26.6	20.3	24.3	15.7	16.2

Comparison would be performed on the measured actual data and data from monitoring system, the relative error is limited within 5 %, thus the system has good quantitative detection performance.

6 Conclusion

This software has realized the collection of sensor node data through adopting the technology of wireless sensor network, serial port programming techniques and ADO.NET database access based on c# language. Graphical interface display is realized through Cl Chart module. Integrating this software and wireless sensors network at the front end could be applied in aquaculture water quality detection, and has obtained better detection effect. The water quality inspection system based on wireless sensor network could realize the online collection and safety inspection of real time data in two-dimensional space and time within certain water areas under low cost condition. This would provide a kind of reliable security technology for water quality analysis and management.

References

[1] Pattichis C S, Kyriacou E, Voskarides S, et.al. (2002) Wireless telemedicine systems: An overview *IEEE Antennas and Propagation Magazine* **44**(02) 143-5
[2] David E C, Hong W. et al 2004 Wireless sensor networks *Communication of ACM* **47**(6) 30-3
[3] Jiang Peng, Xia Hongbo, He Zhiye, et al 2009 Design of a Water Environment Monitoring System Based on Wireless Sensor Network *Sensors* **9**(8) 6411-34
[4] Liming Sun, Jianzhong Li et al 2005 *Wireless sensor network* Beijing: Tsinghua university press *(in Chinese)*
[5] Boqi Mai, Jiapin Chen et al 2009 Test platform of wireless sensor node's localization function *Computer Engineering and Design* **13** *(in Chinese)*
[6] Li J, Lazarou G, et al 2004 A bit-map-assisted energy-efficient MAC scheme for wireless sensor networks *The 3rd Int'l Symp. On Information Processing in Sensor Networks (IPSN04)*
[7] Xu Y, Bien S, Mori Y, et al 2003 *Topology control protocols to conserve energy in wireless ad hocnetworks*
[8] Sai Zou, Changming Liu, Faping Li, et al 2010 Water environment monitoring system based on wireless sensor network *Sensor and micro system* **29**(9) 104-9 *(in Chinese)*
[9] JieZhang, Weiren Shi, Qiaoling Tu et al 2009 Information detection system design based on wireless sensor network *Journal of transducer technology* **22**(6) 861-4 *(in Chinese)*
[10] Renpeng Sun 2010 Research for the application of ADO.NET under multi-layer mode *Computer Engineering and Design* **31**(16) 3621-4 *(in Chinese)*

Authors

Liang Li, born in 1976, Anhui Province of China

Current position, grades: lecturer
University studies: Master degree was earned in major of computer application technology, Southeast University in 2004.
Scientific interest: software engineering, data mining

Automation control and design application of factory sewage disposal system

Zhicheng Chen*

Henan University of Technology, Henan Zhengzhou 450001

Received 12 May, 2014, www.cmnt.lv

Abstract

Intellectualization and automatization become the key technology of sewage disposal for effective realization of sewage disposal technology, continuous and stable operation of system and water quality standard. This paper took a full consideration on the function of online instrumentation in autonomous system and confirmed monitoring system scheme of three-in-one network (data, video and voice) and PLC control scheme adopting means of centralized management and decentralized control combining with sewage disposal technology. Factory adopts Ethernet looped network of optical fibre industry as backbone network for communication and located monitoring system at the central control room of comprehensive office building in sewage disposal factory. It realizes collection and monitoring of procedure parameter, analysis and processing of data, remote monitoring of core equipment. Coarse screen system subprogram, fine screen system screen, immersible pump subprogram, return sludge system program and return sludge system subprogram were detailed described according to technology requirement and design principle of PLC substation. It also discussed test quality guarantee of autonomous system.

Keywords: Sewage disposal, autonomous system, data communication, PLC system

1 Introduction

Urban sewage usually enters drainage system by urban sewage collection pipe. These sewage contain pathogenic microorganism such as bacteria and virus [1]. Therefore, urban sewage must be disposed. Sewage disposal factory can play fatal function for local environment protection. However, sewage disposal technology environment in current stage is relatively serious, which requires sewage disposal system to have high automatization control level [2]. Sewage disposal factory in our country started late. Although current autonomous system basically realizes domestication [3], part of key equipment and component still need to be imported. Operating cost of disposal factory relies on the financial situation of government, which is easy to cause lack of cost consciousness, overstaff and old-fashioned management means [4]. Therefore, it is urgent need to do optimal design on automation control of sewage disposal system.

Domestic sewage disposal factories adopt a few automatic control systems. They can be divided into SCADA system, DCS system, IPC+ PLC system and system composed of bus-based industrial computer [5]. Li Yan [6] from Tianjin University applied control technology of PLC and PROFIBUS fieldbus in automation system design of sewage disposal factory, adopted DCS form of IPC+PLC structure and selected communication scheme of industrial Ethernet+ PROFIBUS fieldbus. Chen Zhengjing [7] adopts star optical fibre Ethernet structure as computer monitoring

system and established Client/ Serber structure based on TCP/IP networking protocol. Shan Liang [8] composed control system based on PLC monitoring technology and open network integrated technology and realized optimizing of automatic control process through optimal design and selection of PLC and instrument of automatic control system. This paper adopted computer monitoring (SCADA) main system of central control room to monitor water input and output and water quality index. Relative data will be stored for at least one year as basis of assisted decision for production and operation in future. It has significant function on flexible, efficient, and reliable of sewage disposal process and lower the cost of early construction and capital construction.

2 Overview of sewage disposal autonomous system technology

Control substation realizes unattended operation. Mainstream product in automation control equipment industry is selected and expansion space is reserved on design. Monitoring system adopts sharing network of 100 Mbps optical fibre Ethernet with self-healing capacity, automatic monitoring system and video monitoring system, that is, three-in-one network of data, video and voice. The selected software is modularization and structuralization software for edit, debugging, modification and update of user program [9]. Control means is central control, workshop control and site equipment control, which combines automation and

---

* *Corresponding author* e-mail chen_1958@163.com

manual operation. Local control level is prior [10]. It is shown in Fig.1.
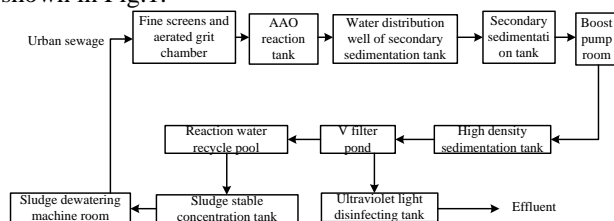
FIGURE 1 diagram of technology flow in sewage factory

Automatization control system of sewage disposal factory is required to realize running mode of centralized management and decentralized control for dispersing breakdown. System allocation needs flexibility as well as compatibility, openness, expansibility for realizing standardization and modularization construction [11]. Software interface standard OPC is regarded as data standard because it can realize data share of soft hardware of different brands.

3 Structure design of sewage disposal autonomous system

3.1 SYSTEM COMPOSITION

This paper adopts mode of "centralized management and decentralized control". Autonomous system can be divided into central control, workshop control and site equipment control. Automation and mutual control are combined together. And local control level is prior. Main control system adopts S7-200 and S7-300 series PLC produced by Germany SIEMENS corporation [12]. And dewatering room and technology section of oil desilting field adopt moisture proof and anticorrosive PACR2 control system produced by OPTO Corporation.

Autonomous system of sewage disposal factory is composed of PLC 1 (including coarse screen room, influent room and primary power distribution room), PLC 2 (including fine screen room, grit chamber, drain tank and pump room), PLC 3 (biological pool and sludge dewatering room), PLC 4 (blower room and power distribution room), PLC 5 (secondary sedimentation tank, water distribution well and chlorine dosing room) and PLC 6 (boost pump room, filter pond, backwash pump room and power distribution room). Every substation is made up of PLC system, HIMI man-machine interface, industrial Ethernet switching, uninterruptible power supply (UPS) and lighting proof protection device. Substation selects SIEMENS S7-300 series PLC and CPU317-2DP/PN PLC controller. MP277 colour touch screen of SEMENS Corporation is equipped in every substation as man-machine interface of PLC control station. PLC 6 control substation includes seven control subsites PLC6.1-PLC6.7. Every substation monitors two filter. These seven subsites select SEMENS S7-200 series PLC and CPU226 controller. Every substation is equipped one 5.7 inches colour touch screen. Mater

station and substation, substation and subsite are connected by industrial Ethernet as shown in Fig.2.
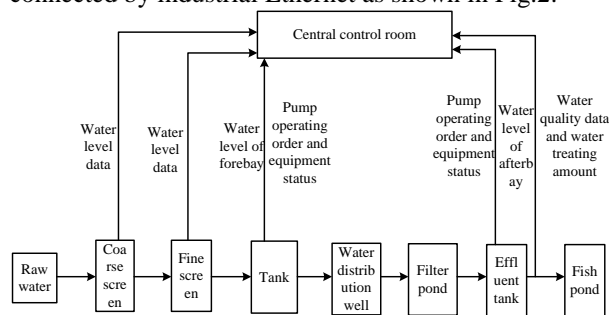
FIGURE 2 structure diagram of autonomous system of sewage disposal factory

3.2 NETWORK COMPOSITION SCHEME

Computer monitoring system of sewage disposal factory adopts 100 Mbps optical fibre Ethernet with self healing function which is sharing network of video monitoring system and autonomous system. That is three-in-one network of data, voice and video. Communication backbone adopts optical fibre Ethernet. Autonomous system design adopts PLC integrated control system.

4 Model Construction

4.1 SETTING OF ENGINEERING CONTROL MODE

Engineering control mode can be divided into manual mode and automatic mode.

(1) Manual control mode: equipment can operate independently to avoid effect on other equipment's when one equipment is broken down. Mud scraper, grillage machine, reverse dosing equipment, washing equipment, water pump and valve are individually controlled by site control button.

(2) Autonomous control pattern: equipment operation does not need manual intervention. It is start-stop controlled by technology parameter and PLC according to working condition of sewage factory.

And flow chart is shown in Fig.3.
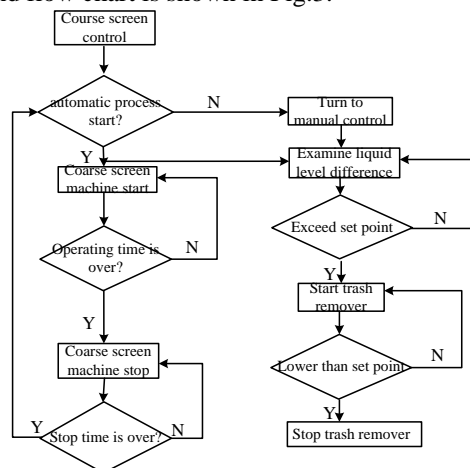
FIGURE 3 working flow diagram of coarse screen

The function of coarse screen system subprogram is to control the operating of coarse screen and trash remover. In automatic mode, coarse screen starts and is timed. Coarse screen stop operating after 20 min and is timing 2 h. coarse screen continues to operate for 20 min after 2 h. It moves in such circles. Meanwhile, liquid level difference is examined. If liquid level difference exceeds set point, then trash remover starts. If liquid differences lower than set point, then trash remover stops. Fine screen system subprogram is similar to coarse screen. The only difference is that fine screen controls the operating of fine course and trash remover.
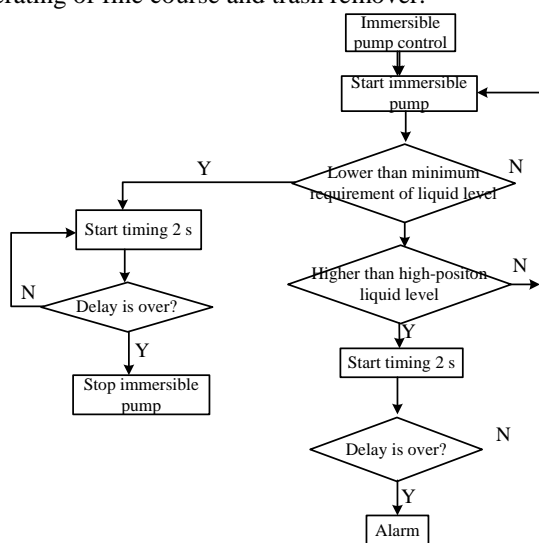
FIGURE 4 work flow of immersible pump

Function of immersible pump subprogram is to control the operating mode of immersible pump. In automatic mode, immersible pump starts. At that moment, liquid level is examined whether it is lower that lowest-position sensor. It it is, and then timing starts for preventing misjudgment. If liquid level is still lower than the lowest-position sensor when the timing is over, then immersible pump stops operating. Otherwise, immersible pump continues to operating. If liquid level is between meso-position and high-position sensor, then timing starts for preventing misjudgment. If liquid level is still in high-position sensor when timing is over, then alerting signal is output.
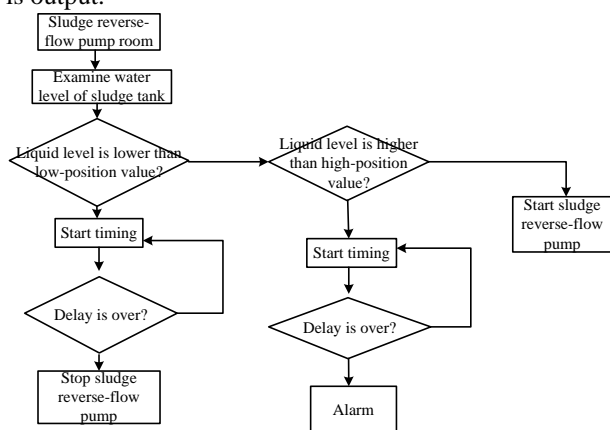
FIGURE 5 workflow of sludge reverse-flow system

The function of sludge reverse-flow system program is to control start and stop of sludge reverse-flow pump. In automatic mode, liquid level is monitored. If it is lower than lowest liquid value, then timer starts. If liquid level is still lower than lowest liquid level sensor when timing is over, then the reverse-flow pump automatically stop operating. If the liquid level is between highest-position and lowest-position, sludge reverse-flow pump starts. If it is higher than highest-position, then timer starts. If liquid level is still in the highest-position when timing is over, then alerting signal of accidents is output.

Function of sludge dewatering system program is to control centrifugal extractor. Centrifugal extractor and timer starts. When timing of timer is over, then polymer pump starts. At this moment, timer starts again. When the timing is over, then cutting machine and sludge reverse-flow pump starts operating.

4.2 REALIZATION OF SYSTEM FUNCTION

Automatic control function of sewage disposal factory is realized by main system of computer monitoring (SCADA) in central control room. Four monitor workstations in central control room operate configuration monitoring program. And configuration software realizes quantum hot standby. SCADA can adjust the production process of the whole factory, monitor operating of subsystems and equipment's intensively and realize information management function.

4.3 VIDEO DISPLAY OF CENTRAL CONTROL ROOM

Video display frames provided by superior system have following kinds: overall display picture of overview map of whole factory, overall display picture of technological process of whole factory, main picture display of monitoring system, summary picture of overall signal, summary picture of alarming, production curve picture, picture of system menu and detailed monitoring picture of equipment or facilities.

4.5 QUALITY CONTROL MEASURE OF AUTOMATIC CONTROL SYSTEM FUNCTION TEST OF SEWAGE FACTORY

System fault caused by current software failure has become a problem that cannot be neglected in industrial automatic control system. Software failures have two characteristics compared to hardware failure. One is that software failure will not change with time. Second is that software failure is mostly caused by inherited error in procedure code. Interface error between software and hardware is also an important factor of configuration software failure [8]. Therefore, key point of software is the test before leaving factory and in early period of use, especially the system test after integration of software and hardware.

System test environment. We need to establish a system test environment for the operating of software in real hardware environment, adding test case into tested software by testers and collecting test data, as shown in Fig.6. It is a test platform composed of software and hardware
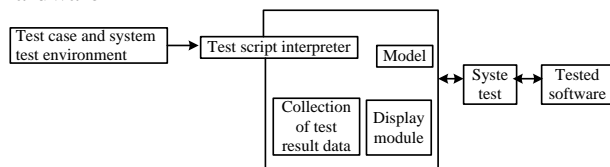
FIGURE 6 diagram of software system test environment

System test process. After establishing test environment, we can use it to do system test. System test can be divided into following procedures: design test case, operating test, analysis of test results and test conclusion.

TABLE 1 classification of design test case

Design test case		
normal function test	*boundary situation test*	*abnormal operation test*
Simulate some representative operation according to the normal use situation of tested engine	boundary test on analogy quantity input in tested system	For inspecting abnormal condition processing ability of industrial automatization control system software

It can be considered from normal function, boundary situation, abnormal operating according to the different emphasis of test while designing test case. Content or objective is shown in Table 1. In graphical editing environment provided by test environment, test case that have been designed transforms into test script form which can be identified by test environment by drawing picture

or text. A test case transforms into a test script [14]. After test, test environment will automatically collect operating data of tested equipment. These data and expected result in test case are compared and analysed to judge whether this test equipment is broken down.

5 Conclusion

This design adopts mode of centralized management and decentralized control, monitoring system scheme of three-in-one network of data, video and voice and PLC control scheme according to technology requirement and design principle of PLC substation. Factory communication applies optical fibre industrial Ethernet as backbone network. Autonomous system can be divided into central control, workshop control and site equipment control. Automatic control and manual control are combined. And local control level is prior. Autonomous system of sewage disposal factory in this design is composed of central monitor station and six site control substations. And PLC system is made model selection. Dewatering machine room and oil desilting field adopt moisture proof and anticorrosive PACR2 control system produced by OPTO Corporation. And substation selects SIEMENS S 7-300 series PLC. Subsite selects SIEMENS S 7-200 series PLC. Software should be tested before leaving factory and in early period of use for ensuring autonomous system function, especially the system test after integration of software and hardware.

6 Acknowledgment

References

[1] Yin L Q, Yin C X, Zhou X, et al 2013 Monitoring On-Line and Control of Sludge Concentration in Sewage Treatment System *Applied Mechanics and Materials* (373) 897-901
[2] Tang L 2012 Application of DCS system in automation control of sewage disposal equipment *Shandong University*
[3] Ma L Z 2012 Research on urban sewage disposal control system based on modified form A~ 2/O technology *Shandong University*
[4] An H X, Jiang J J, Xu R 2012 Analysis and design of autonomous control project engineering system of sewage disposal factory *Collected Papers of International Security Science &Technology Academic Conference* 284-9
[5] Zhu B K, Lin X S, Qi Q S 2012 Form, structure and function of autonomous system of sewage disposal factory *Automation Panorama* (2) 66-9
[6] Li Y 2011 Engineering autonomous system design of Qinglan sewage disposal factory in Hainan *Tianjin: Tianjin University*
[7] Chen Z J 2009 Optimal design of autonomous system of urban sewage disposal factory *Fujian Light Textile* (008) 35-8

[8] Shan L 2013 Simple analysis of autonomous system design of some sewage disposal factory *Electronic Technology and Software Engineering* (23) 263-4
[9] Binbin Y, Xiafu L, Yong C, et al. 2011 Optimized Design of Automatic Control for Dissolved Oxygen in A/A/O Processes of Urban Sewage Treatment *Water Purification Technology* **30**(6) 70-3
[10] You B W, Wei W, Sheng X F, et al 2011 Design of Automatic Control System in South Sewage Treatment Plant of Nanjing *China Water & Wastewater* **27**(8) 68-70
[11] Huidong G, Yanru W 2012 Requirement of Sewage Treatment Techniques of Small-size Plant for Automatic Control System *Value Engineering* **31**(3) 22
[12] Hong-wei L, Zhou-jian Y, Wu W, et al 2011 The application of PLC in SBMBR sewage treatment automatic control of the system *Networking, Sensing and Control (ICNSC), 2011 IEEE International Conference on. IEEE* 422-6
[13] Liu M 2014 On Application of Automatic Control in Sewage Treatment *3rd International Conference on Science and Social Research (ICSSR 2014) Atlantis Press*

Authors

Chen Zhicheng, born in February, 1960, Henan Province of China

Current position, grades: associate professor
University studies: Master degree was earned in major of vegetable protein and comprehensive utilization, Jiangnan University in 2001.
Scientific interest: computer and factory automation

Research on computer information integration based on some wireless sensor network model

Chenxiang Zhang*

SuZhou Industrial Park Institute of Services Outsourcing, Suzhou, Jiangsu 223003, China

Abstract

Network model in wireless sensor integrates wireless communication technology, sensor technology and embedded computer technology. This new computer pattern is the joint elements in current network hot technology. Self-organizing feature Map is also termed as SOM network. It can solve conception problem, which can only be fulfilled by human brain nervous tissue. We integrated computer information applying SOM and wireless sensor technology to realize information integration model. We made a conclusion of the characteristics of wireless sensor network model and studied its protocol architecture. We induced information integration system into software engineering and developed it into application in all industries and areas in the perspective of definition, structure, classification and calculation of information integration. This paper integrated information's, reduce redundancy of information, decrease energy consumption and lengthen the service time of network by SOM wireless sensor network model. It also made a analytical research of case on information integration.

Keywords: wireless sensor, SOM network, information integration, network model, evaluation

1 Introduction

Network node in wireless sensor network is powered by batteries. It is usually arranged in places of bad environment. Its special application area determines that battery charge and battery replacement are not allowed in the process of using. Once the battery is out of energy, the node will lose its function. Therefore, application of any technology and protocol [1] should take energy conservation as premise in the process of wireless sensor network design. Wireless sensor network has lots of nodes. Energy is consumed when network node achieve data collection, calculation or transmission. The energy for consumption is related to data size, sampling frequency, sensor types and application requirements. In addition, many scholars have proved that information transmission by sensor will consume more electric energy compared with calculation performance. For instance, electric energy needed in transmitting one bit of information is enough for carrying out 3000 of calculation command. Meanwhile, large data volume produced in short time will lead to network congestion and shorten the life time of network. This paper tried to solve the problem of large amount of transmission and high energy consumption adopting efficient processing method of information integration.

This paper constructed information integration based on SOM neural network clustering method. This method is widely applied in medical area. We introduced it to construct network model to better stimulate reasoning and stimulation ability for information integration. The

Fourth Military Medical University in Shanxi studied clinical inspection item and establishes SOM model based on SOM neural network clustering method [2]. Yanshan University of Hebei effectively avoided drawback and fulfil solution [3] by integrating neural network on high-dimensional data visualization area based on self-organizing feature map. Liao Guanglan et al calculated the recognition of dimensional feature on input pattern clustering in neuron weights of SOM network competition layer. The research found that SOM can effectively recognize character subset and realize pattern clustering between different data [4]. The above research area effectively solves relative problems applying SOM technology. It shows the strong processing ability of SOM technology. Introduction of SOM provides indispensable assistance for network information integration of this paper. As an emerging subject, information integration technology increasingly shows its superiority on multi-source information processing from birth to today. Wireless sensor network adopting information integration technology can reduce network burden and improve instantaneity to some extent. It has obvious advantages compared with those without information integration technology. It is of deep theoretical and practical meaning to fulfil analysis and processing of perceptual information of sensor node by relative technology in information integration.

---

* *Corresponding author* e-mail: zcxcxz@163.com

Information and Computer Technologies

2 Overview of sensor network

Sensor network largely extend specific internet. It can monitor the objects that is moving, growing and emitting heat. Its appearance will affect our life all the time. Wireless sensor network is a new kind of information processing system facing to physical world and also a new calculation model. Integrated micro sensor terminals real time monitors centre of sensing. Then the information is transmitted to remote user by wireless network. The objective of this network is realization of ubiquitous calculation.

2.1 CONCEPTION OF WIRELESS SENSOR NETWORK

Wireless sensor network combines sensor technology, embedded computer technology, distributed information processing technology and wireless communication technology. It can collaboratively real time monitor, perceive and collect information of different environment or monitoring object, process them and transmit them to users who needed [5]. The process of the work is to provide service for users. The typical working mode is as follows: lots of sensor nodes are thrown into area of interest by aircraft and nodes form a wireless network by self-organizing feature map. Node is the collector and issuer of information and also routing of information. The collected information reaches base station by single hop or multi-hop. Base station is special node. It can communicate with monitoring centre by Internet, mobile communication network, satellite, etc. A typical wireless sensor network is shown in Figure 1. Direction of the arrow shows how information is transmitted to users.
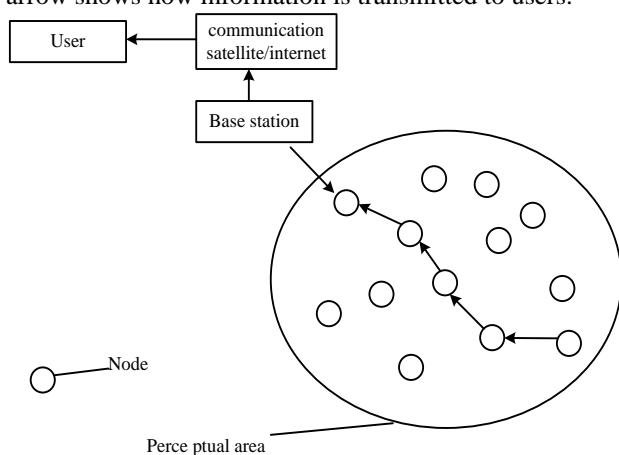
FIGURE 1 Wireless sensor network

2.2 SENSOR NODE

Wireless sensor network is composed of lots of sensor nodes. Therefore, we propose figure of sensor node structure (Figure 2). Every node has its processor. It can perceive data and make a simple processing in node. Then the detailed information is obtained by processing of useful information. And the information is transmitted to users who need information.
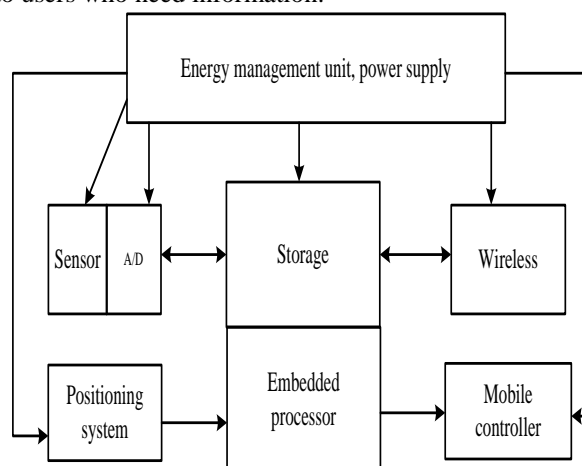
FIGURE 2 Figure of sensor node structure

Node is composed by calculation subsystem composed by micro processor and storage, communication subsystem; subsystems composed by sensor and stimulate device and energy supply subsystem.

3 Information integration and SOM method

Information integration is a process of dealing data or information coming from one or more sources. The process carries out association, interconnection and combination to obtain accurate location and identity estimation and makes a timing and full evaluation on battlefield situation, threat and the important degree [6].

Information integration in wireless sensor network should process perceptual information of different levels adopting mathematical tool and other technologies in the perspective of users. It uses optimal routing to reduce communication consumption, lower energy consumption, lengthen life time of network and satisfy users to the largest extent.

3.1 ADVANTAGE OF INFORMATION INTEGRATION

Information integration is a process of synthetic processing of data in sensor on multiple levels. Every processing level reflects different degree of abstract on original data and produces new meaningful information. In addition, this new information cannot be obtained by any single sensor. Its advantage is enlarging space search scope and instantaneous search scope, lowering reasoning fuzzy degree, reducing cost and time of information obtainment, improving the fault-tolerant ability and self-adaptivity and increasing dimensions of target feature vector.

3.2 INFORMATION INTEGRATION ARCHITECTURE

Information integration can be divided into centralization, distribution and mixing according to the location of integration node in information flow.

(1) Centralization architecture is transmitting original data from multiple sensor nodes to central integration processing centre and fulfilling tasks of data matching and correlation functions and classification of target tracking. This method can process data with high accuracy. But size of data transmission and processing is large. It requires highly on communication line and processor. As shown in Figure 3, F is expressed as integration centre, C is expressed as information user and S is expressed as information source.

(2) Every sensor makes processing on own monitoring data and produces state vector and attribute parameter. Then the processing result is transmitted to integration centre for integration processing. That is the distributed architecture. Corresponding to centralized architecture, distributed architecture have little load but its integration accuracy is lower than that of centralized architecture. Common distributed architecture includes architecture of parallel distribution, serial distribution and feedback.

(3) Mixing architecture is the combination of centralization and distribution. It is complicated. It requires pre-processing sensor of distributed architecture and high-speed communication link that transmitted original information to integration centre. And the calculation amount of processor is large. It is commonly used for large integration system.
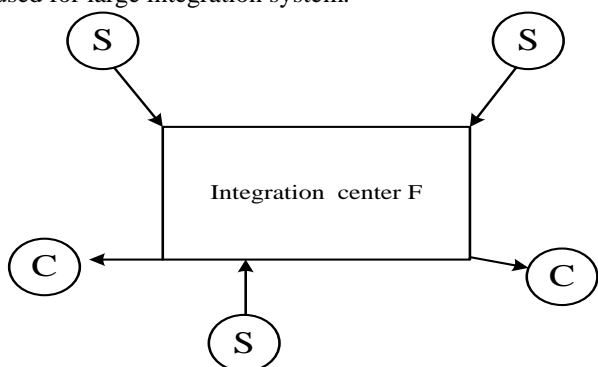
FIGURE 3 Centralization integration architecture
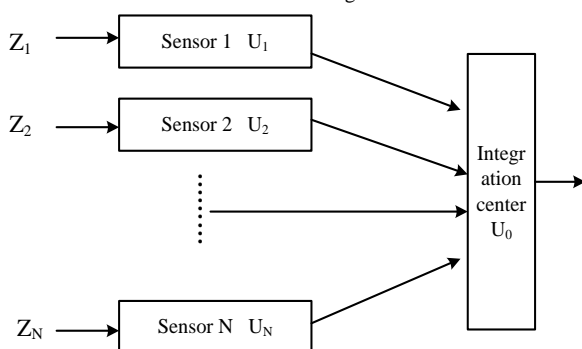
FIGURE 4 Parallel distributed architecture
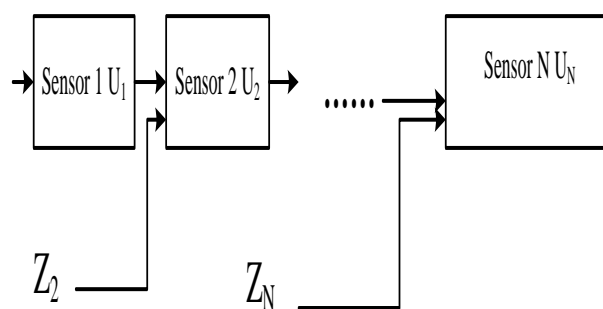
FIGURE 5 Parallel distributed integration architecture

3.3 INFORMATION INTEGRATION BY SOM METHOD

Self-organizing feature map at the earliest is an effective algorithm of data compression and clustering proposed by neural network expert TeuvoKohonen from University of Helsinki in Finland in 1981. In the past 20 years, SOM was widely studied and applied in robot vision, mechanical control, speech recognition, and vector quantization machine and pattern recognition. In addition, SOM can be widely used in sample order, sample classification and sample detection as a sample characteristics detector.

3.3.1 SOM network model

Output that have similar function will be close and output that have different function will be apart after competition learning of output sample [7] by SOM. It automatically lines some ruleless output. In the process of adjustment of joint weight, distribution of weight and probability density of input sample are made similarly.

SOM is a two-layer network composed of input layer and output layer. Figure 6 gives a hidden-free SOM model composed of input layer and output layer. Input layer is used for receiving input pattern. It is one dimensional sequence composed of N neurons. Junction points of output layer are relative to space dimensions of output pattern. Output layer is two dimension planar array composed of M*N=H of neurons. Its junction points are relative to pattern space dimension after map. It can be dot matrix of one or two dimensions. Junction points of input and output layer are full of joint. Junction points within input layer and output layer have no connected relation. Every neuron in output layer is connected with neighbourhood. This connection is a relationship of manual encouragement. Different junction points of output layer represents different classification pattern after training.

Output layer

Input layer

X_1 $X_2 \cdots X_n$
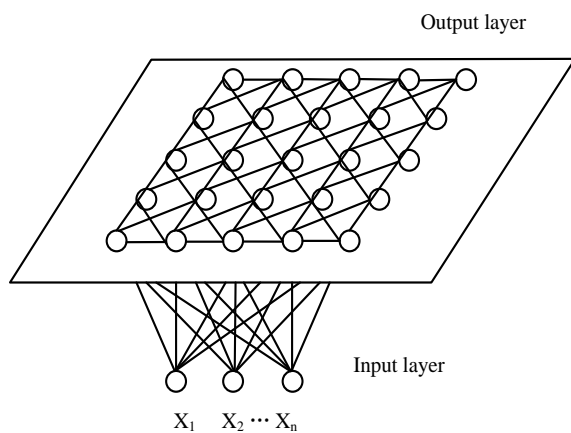
FIGURE 6 Topological graph of SOM network

SOM network model is made up of processing element array, compared selection mechanism, partial interconnected function and self-adaptative process. Model of Figure 2.4 can process the distributed topological graph with input signal on cell array. However, in initial state, there is no distributed topological graph with signal characteristics on cell array. SOM can extract characteristics from internal environment according to some measure or topological space.

3.3.2 SOM algorithm

Procedure 1, initialization. Continuous weights of N input neuron and output neuron is given the small value. J of adjacent neurons in output neuron is selected to gather Sj. Where Sj (0) is expressed as set of adjacent neurons of Neuron j at time t = 0 and Sj (t) is expressed as set of adjacent neurons of neuron j at time t. Area Sj (t) will decrease with the growth of time.

Procedure 2, new output pattern X is provided.

Procedure 3, Euclidean distance dj is calculated. That is, the distance between output sample and every output neuron j:

$$d_j = \left\| X - W_j \right\| = \sqrt{\sum_{i=1}^{N} \left[x_i(t) - w_{ij}(t) \right]^2} \ . \tag{1}$$

A neuron with minimum distance is calculated. Select a unit k that makes any j

$$d_k = \min_j (d_j) \ . \tag{2}$$

Procedure 4, neighbourhood Sk (t) is given.

Procedure 5, out [put neuron j^* and its weight is revised according to Formula (3).

$$w_{ij}(t+1) = w_{ij}(t) + \eta(t) \left\lfloor x_i(t) - w_{ij}(t) \right\rfloor , \tag{3}$$

where η is gain item that is decreased with the change of time.

$$\eta(t) = \frac{1}{t} \ . \tag{4}$$

Procedure 6, calculate output o_k

$$o_k = f(\min_j \left\| X - W_j \right\|) \ . \tag{5}$$

The formula (5) is generally function between 0-1 or nonlinear.

To sum up the above procedures, SOM can be simply described as only part of weight vector can be adjusted to close the latter and away from input vector for every network input. All weight vectors are separated from each other within input vector space. And pattern that represent input respectively are formed. That is the clustering function of SOM characteristics automatic identification.

4 Case analysis

Safety production of coal mine is the key point in current coal mine development industry. Monitoring of gas density and temperature is the basic of safety production. The detected environmental parameter of coal mine is mostly fulfilled by single sensors. When a sensor is broken down, the environmental parameter data of that location can not be detected, which will lead to false-alarm and even no alarm. In addition, the processing of environmental parameter is simple. Therefore, the evaluation of real time safety situation is not accurate and effective measurement can be not carried out. It will cause collapse of coal mine and fire disaster. These situations will bring large hidden danger for safety production of coal mine.

4.1 PARAMETER ANALYSIS

For the prevention of gas coal fire to harm personal and equipment safety, working face, heading end, main roadway and mechanical and electrical bully chamber should be equipped with sensors of gas, temperature, carbon dioxide and wind speed to real time monitor environment [8].

(1) For temperature, when the temperature under the coal mine well reaches certain condition, oxidation of coal will be promoted and lead to spontaneous combustion or gas combustion. Temperature of working face in coal well digging should not exceed 26℃. Air temperature of mechanical and electrical equipment room should not exceed 30℃. Working time of workers in high temperature place must be shortened and they should be provided high temperature care treatment when the temperature exceeds the value. If air temperature of digging working face exceeds 30 ℃ and that of mechanical and electrical room exceed 34 ℃, the operation should be stopped.

(2) Wind speed can directly affect ventilation quantity. Unreasonable ventilation quantity may increase

or induce the oxygen supply factor for natural fire. And it will also have certain effect on heat dissipation and increase the risk degree of natural fire. Different roadway have different requirement for wind speed. For instance, minimum wind speed in roadway of electric locomotive stringing is 1.0 and the maximum is 8. Minimum and maximum of wind speed in stone drift under digging is 0.15 and 4. Minimum and minimum of wind speed in conveyor lane, digging area and air return way is 0.25 and 6.

(3) Gas is formed from coal or other C class material in mine well. Methane is the major ingredient. It is lighter than air and easy to burning and blast. When gas concentration of air current in return air lane and digging working face exceed 1.0%, the operation should be stopped and workers should leave. And relative measurement must be carried out.

In addition, there are the effect of concentration of carbon monoxide and dust [9]. Maximum value of parameter is given in Table 1.

TABLE 1 Maximum reference value of major parameter in coal well

Name	Maximum parameter
Temperature	26
Wind speed	1.0/8
Gas	1.0
Carbon monoxide	0.0024
Dirt	10

4.2 CONCRETE ANALYSIS

Facing with single node, size of transmission amount is taken into account to determine whether compression is needed. Compression can be divided into lossless compression and lossy compression according to whether loss is allowed. For instance, searching for the temperature of some node at some time can be directly relative to the node. When we search for the temperature of some node during some period, compression can be considered because of the large size of data. Another reason is that the change of temperature is small. For instance, for a temperature signal of 20 ℃, in most data that record environment, only the data after radix point change while the integer part do not change. That is, the integer part is the same. If few records derivation is recorded adopting method of known value derivation, then it can effectively lower the storage space of data.

Whether the information needed by users comes from homogeneous sensor node should be considered facing with area. If the information comes from homogeneous sensor node, the purpose of information integration is how to make a synthetic processing on this information to obtain more reasonable index value of characteristics of research object. Integration algorithms that can be adopted are weighted average, Bayesian method, Dempster-Shafer, etc. For instance, if the average temperature of some area is needed, then the node in the area is made weighted average according to reliability. At last, only the needed average temperature value is

transmitted to users rather than all node data. It saves node energy, improved efficiency and lengthens the life time of network. Output of n sensor nodes for detection of parameter x is written as xi, I = 1, 2,…n. Information integration adopt weighting algorithm. And the formula of algorithm is:

$$x^+ = \sum_{i=1}^{n} a_i x_i \, , \ \sum_{i=1}^{n} a_i = 1 \, . \tag{6}$$

The key is to determine coefficient a_i to make accuracy of integration algorithm highest. Determination of coefficient becomes a critical factor because the node location is different and the interference suffered by sensor node is also different. If the information come form non-homogeneous sensor node, then the purpose of information integration is to synthesize information implied in different indexes to estimate the global property of research object. For example, neural network method can be adopted. For example, in some area, there are various sensors such as temperature sensor, humidity sensor, pressure sensor, light sensor, etc. After synthesis of all perceptual information, the judgment of the air quality levels can be obtained. Facing to network also need distinction of homogeneous information and non-homogeneous information and application of relative algorithm. The procedure is shown in Figure 7 and Figure 8.
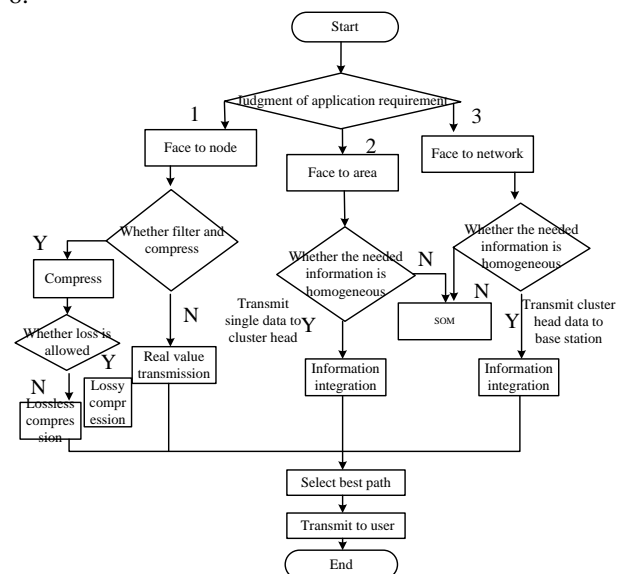
FIGURE 7 Flow diagram of information integration

Property facing to single node in Figure 7 includes type code, location, energy, alarm threshold value, working attitude and sample frequency. Node can be temperature sensor, humidity sensor or the node code of some location. The major task is data connection and calculation. Property facing to area includes size and type. The major task is calculation and evaluation of state. Property facing to network includes types of various requirements.

Analysis of information integration should choose proper integration architecture, integration algorithm and

database support needed in integration establishment according to the requirements of system. In environment monitoring of coal mine, the first step is alarm threshold data setting. That is the setting of alarm threshold of harmful gas for alarm judgment. In advance, alarm threshold is set on node and is expressed as constant. Sensor node makes a comparison on perceptual information and constant. If perceptual information is larger than constant, then alarm will be given. For instance, when the gas parameter exceed threshold, relative equipment can be carried out real time outage. If the parameter is within the normal range, then real time monitoring should be going on.
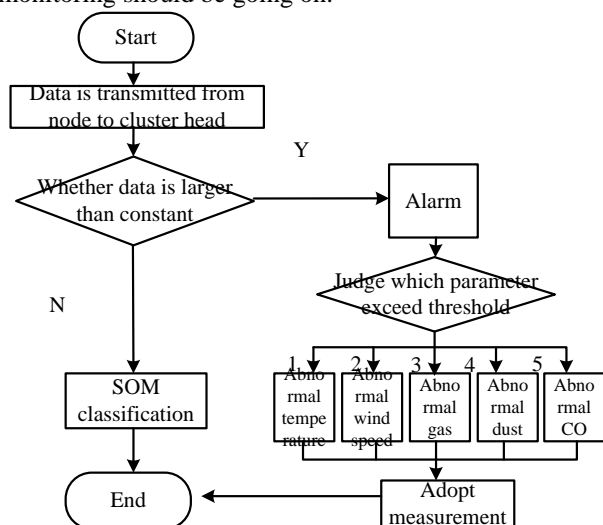
FIGURE 8 Flow diagram of environment monitoring of coal mine

5 Conclusion

Wireless sensor network will change the interactive mode between human and nature, which is the hot point in the current research area. Monitoring data of wireless sensor network have characteristics of large size, redundancy and repetition. In addition, the power of sensor node is supplied by battery. And the energy consumption should also be reduced. These factors make information integration not to be neglected. The conception of integration at the earliest come from military area and then penetrates into other areas. Single uses of one method have certain limitation because the application area of information integration is quite wide. We make advantages integration of various methods, which gradually become the key point of research of information integration algorithm. Application of algorithm and architecture in information integration can save energy and lengthen the life time of network. We construct network model on the basis of SOM algorithm for achieving expected effect in application.

This paper considers the user demand in the perspective of software engineering. Class is regarded as base class in application and extended based on the property with good definition. At first, we analyse the problem of current monitoring system if coal mine safety in the background of coal mine safety. Then we discuss the advantages of equipping wireless sensor network according to the actual situation, which can increase the confidence level of system monitoring and improve monitoring performance. Afterwards, we analyse some critical parameters such as effect of temperature, gas concentration and dust concentration. The current coal mine system is mainly to monitoring the major parameter applying sensor and transmits the result it monitoring centre. Analysis and processing data by integration data is superior to traditional monitoring system of coal mine safety. With the assistance of SOM and its strong functions of analysis and display, this research vividly reflects data characteristics and processes single parameter and synthesis judgment in the form of flow diagram.

References

[1] Yu Haibin, Zeng Peng 2006 Intelligent wireless sensor network system *Beijing Science Press* 17-20
[2] Shang Lei, Zhang Yuhai 2003 *Applicability research neural network clustering methid in clinical test knowledge discovery based on SOM* Shangxi: the Firth Military Medical University
[3] Shi Lihong 2013 *High-dimensional data visualization based on SOM algorithm* Hebei: Yanshan University
[4] Liao Guanglan, Shi Tielin, Jiang Nan, et al 2005 Research on characteristics selection technology based on SOM network *Journal of Mechanical Engineering* **41**(2) 46-50
[5] Wei Ruijuan, Li Xuejun, Ren Weicheng, et al 2010 Research on the construction and application of single base station *Bulletin of Surveying and Mapping* **06** 23-26
[6] Pan Quan, Wang Zengfu, Liang Yan, et al. 2012 Basic method and progress of information integration theory (II) *Control Theory and Application* **29**(10) 1233-44
[7] Chen MR, Deng FQ 2004 Data clustering Based on SOM Computer knowledge and Teehnie 59-61
[8] Chen Tianxiang, Fu Lin 2009 Discussion on monitoring performance evaluation system of government coal mine safety production-From the investigation and survey of City J Shanxi *Journal of Xiangtan University* **33**(5) 18-30
[9] Hong Feng, Chu Hongwei, Jin Zongke 2010 Review of latest development of wireless sensor network application system *Computer Research Development* **47** 81-7

Authors

Chenxiang Zhang, born in 1980, Henan Province of China

Current position, grades: lecturer
University studies: Bachelor's degree was earned in major of computer science and technology, Zhengzhou University in 2003; master's degree was earned in major of control engineering and domain engineering, Jiangnan University in 2009.
Scientific interest: software engineering

Example-based geometric texture synthesis: a survey

Lingqiang Ran[*], Xiangxu Meng

School of Computer Science and Technology, Shandong University, 1500 Shunhua Road, Jinan, China

Abstract

3D object modelling is a key step in computer animation industry. How to generate models rich in high quality geometric details is still a challenging task and affects the final visual effects of animation. Example-based geometric texture synthesis (EGTS) is a powerful tool to automatically build models with rich geometric details. Given a small patch of geometric texture as an example and an arbitrary model as a target model, using EGTS, we can change the surface of the target model to the new geometric texture style. EGTS has been studied a lot by the literature and many papers have been published. In this paper, we will summarize the influential papers and present our thoughts on this topic.

Keywords: example-based, geometric texture synthesis, survey

1 Introduction

In the real world the majority of three-dimensional objects, especially those found in nature exhibit some form of texture. Plant surfaces like veins, thorns or needles, and animal surfaces like feathers or fur are the examples. There are two ways to add texture to an object, with colour texture or geometry texture. Adding colour texture means setting an object's surface colours. And by altering an object's surface geometry, geometry texture can be added.

In order to minimize the geometric complexity while creating visual complexity, texture mapping is developed by the early computer graphics researchers to balance the gap between the hardware limit of polygon processing and the need forever richer computer-generated scenes. To reduce the artefacts of image texturing, more general forms of textures are introduced, such as volumetric textures and bump mapping while modelling and rendering the 3D details of a surface is still excluded. But with the fast development of hardware, nowadays, commodity video cards are equipped with extremely flexible and powerful graphics processor. Displaying tens of millions of triangles and texture mapping in real-time have been both achieved. So the image-based textures' well-documented visual artefact's such as inaccurate shadows, smoothed contours and lack of parallax can be all overcome by enabling exquisite details to be directly rendered and purely geometrically modelled. As it does not suffer from most of the traditional limitations of animation, editing and modelling, this purely geometric detail representation offers an unrestrictive and versatile tool for artistic creation and turns out to be very desirable.

However, compared to image texture synthesis, it is still tedious to model complex geometric details such as weaves, ivies, chain mails or veins. So the problem of synthesizing 3D geometric textures remains challenging, while the basic techniques for image texture synthesis are mostly worked out. One way is to apply image texture synthesis ideas to geometric texture synthesis. Firstly, an example geometry is provided. Then, similarity matching can be used by a synthesis algorithm to generate more geometry. This is the basic idea of EGTS.

A number of approaches have been proposed for EGTS, but only some of them are mentioned by [1, 2]. Here we will give a more overall description on the topic covering all the influential papers published. A different interpretation of this idea is come up with as each research group looks at this problem.

And as the choice of geometric representation often guides the algorithm's style, we classify the papers into three categories according to the representation used: volumetric model [3-5], polygon mesh [6-12] and height field [13]. The section that follows gives a brief description of these approaches, and the interested readers are referred to the original papers for details.

2 Example-based geometric texture synthesis methods

The idea of texture synthesis is extended by some researchers to the creation of geometric texture from example geometry. For performing geometry synthesis from example, a number of different approaches have been invented. In this section the papers are firstly sorted into three categories (as shown in Table 1). Then we will introduce the representative works in detail by category.

---

[*] *Corresponding author* e-mail: rlq163@163.com

TABLE 1 Taxonomy of EGTS approaches

Geometric representation	Representative works
Volumetric model	Volumetric Grids [3], Search Tree [4], Model Pieces [5]
Polygon mesh	Geometry Images [6], Trivariate Mapping [7], Shell Maps [8], Cellular Texture [9], Mesh Quilting [10], Laplacian Texture [11], Discrete Element [12]
Height field	Terrain Synthesis [13]

2.1 VOLUMETRIC MODEL METHODS

Bhat et al. [3] generalize image texture synthesis to geometry by drawing an analogy between 2D pixels and 3D voxels. Begin with fine volumetric details such as holes, grooves, pits and spikes, their approach can place these detailed features on the surface of a model. The main idea of the approach is to extend Image Analogies [14] to the volumetric domain. Firstly, an object is provided, onto which additional geometric details are to be added. Then a pair of similar volumetric geometric objects input to the algorithm, one with and one without geometric features, as geometric example. For example, a cylinder with and without grooves. An example of the three input models and the synthesis result are shown in Figure 1. Besides the three input models, a vector field for orienting 3D voxel neighbourhoods needs to be built on the model to be modified.
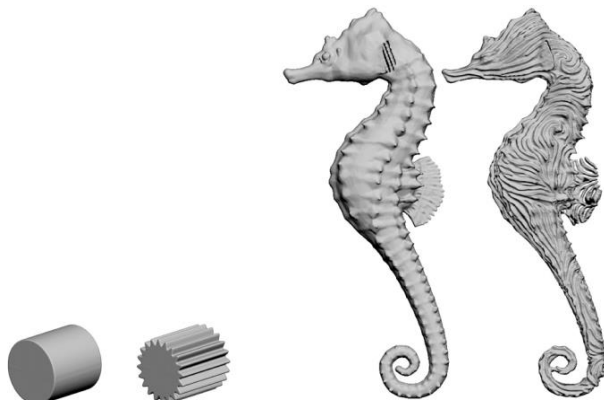
FIGURE 1 The grooved cylinder and the flat cylinder form the input geometric example (left). The seahorse model (middle right) is the input model to be modified. The last model is the resulting volumetric geometry (far right)

Lagae et al. [4] develop a similar approach using voxel matching to synthesis geometry. The input example geometry is assumed to satisfy the constraints of a Markov Random Field model. And a hierarchical distance field is used to represent the input geometry. Their method is a two-phase process. In the first analysis phase, on the input example geometry, the regularly sample distance field is computed, and its hierarchical representation is organized in a tree. In the second synthesis phase, on the distance field, for a partially synthesized neighbourhood, a tree search algorithm is used to find its best match. The un-synthesized samples are replaced with those of the best match. They construct

the final synthesized geometry in parallel. A result of their approach is shown in Figure 2.
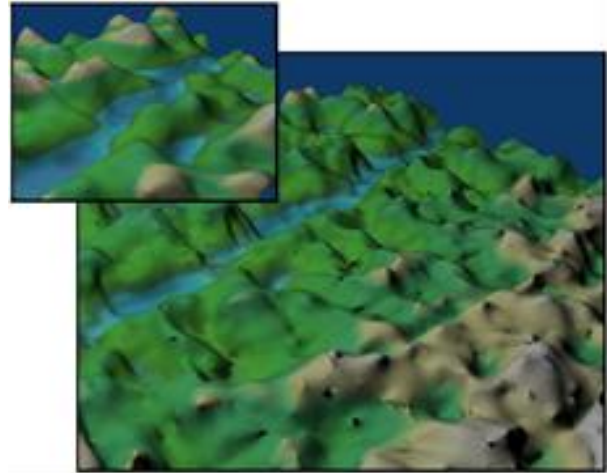
FIGURE 2 The example geometry in the upper left is used to synthesize the terrain geometry in the large image

Another way to synthesis 3D models based on example geometry is demonstrated by Merrell [5]. First, the model is partitioned into uniformly-sized blocks and then these blocks are reassembled in a consistent manner. Given the example geometry, the user specifies the block size, which is used to partition the geometry. After the size of the desired output model is given by the user, then a pattern of blocks is found by the system that is consistent block adjacency. The relative positions of adjacent blocks from the example model is defined as consistency. An iterative algorithm is used to assemble blocks in a consistent manner. First, a simple pattern of block ID's (such as all zeroes) is used to fill a 3D lattice. Then the sub-regions of the lattice are analysed and the block ID's is modified in a constrained random fashion to maintain correct block adjacency. Then the process is repeated several times for more sub-regions. A model of a pillar divided into four pieces is shown in Figure 3.
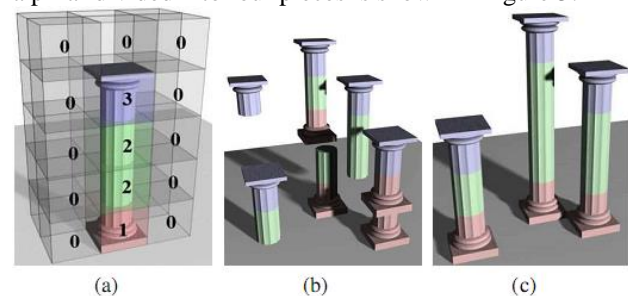
FIGURE 3 Four model pieces compose a model (a), An Inconsistent Model (b) and A Consistent Model (c)

2.2 POLYGON MESH METHODS

As polygon mesh is the most popular representation for 3D model, EGTS methods of this category is far ahead over the other two in quantity. And the generality of polygon mesh makes these methods more feasible in application.

Lai et al. [6] present a method that can synthesis a geometric texture which is designed manually to a model,

and can automatically transfer geometric texture between different objects. Their method is based on geometry images proposed by Gu et al. [15]. They first represent the geometric example and target model in geometry images. Then in the 2D geometry images domain, they use image synthesis method to modify the target model's geometry image. A vector field is built to guide sequence of the image texture synthesis. Finally, the result geometric model is reconstructed from the geometry image (as shown in Figure 4).
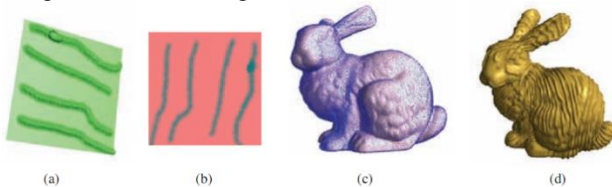
FIGURE 4 A result of Lai et al. [6]. A manually created texture (a) is the converted to a geometry image (b). The vector field (c) is built based on a given orientation. And (d) is the synthesized result

For geometric detail modelling, Elber et al. [7] and Porumbescu et al. [8] use mesh-based details as a versatile representation. In both the two methods, textures are simply tiled over the plane and then mapped to 3D surfaces. And the process is only suitable for periodic textures. Most importantly, as a global parameterization over the plane is not supported by arbitrary surfaces, texture discontinuity will be produced across chart boundaries. Fleischer et al. [9] propose a method to create geometric textures on arbitrary meshes based on mesh details. But they are restricted mostly to the synthesis of simple texture elements like scale or thorns over the surface, the design of complex textures is not allowed, for instance woven materials.

To create complex geometric details in the manner guided by a user-supplied base mesh such as vines, chain mail and weave, Zhou et al. [10] introduce Mesh Quilting, an algorithm depending on having a correspondence between patches in 2D and the base mesh. The algorithm repeatedly uses portions of the texture swatch proceeds to fill empty regions until geometric detail covers the base mesh. In detail, firstly, an empty region of the surface that already has texture detail surrounding it is identified. For filling the empty region, these adjacent parts of the texture work as start constraints. Then in order to fill the empty region with texture, they try portions of the texture swatch. For each part of the swatch, they test several offsets of the texture. Finally, with retaining of the best match, they find a correspondence between the existing pieces and the triangles of the new piece. Figure 5 shows their result.

Ran et al. [11] present a method to synthesize geometric texture details from a sample texture patch onto an arbitrary surface. Their key idea is to represent geometric texture details using Laplacian texture images, which enables flexible geometry texture editing and facilitates effective and simple geometry texture synthesis. First, a sample model and a target model are given. Then a patch from the sample model is selected to extract the geometric texture details.
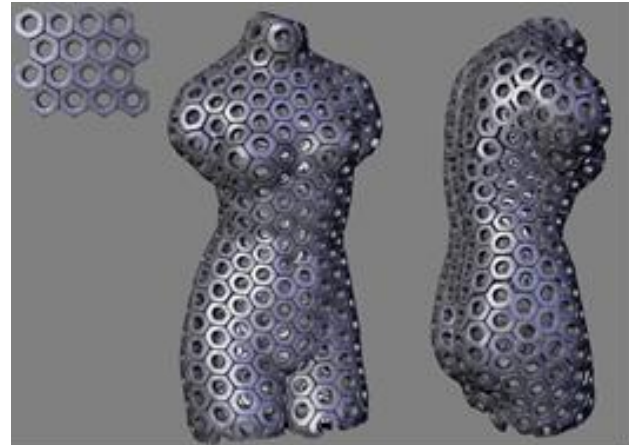
FIGURE 5 Mesh Quilting result by Zhou et al. [10]. The geometric example is shown in the upper left

Next, a Laplacian texture image of the extracted texture is constructed. An image texture synthesis technique is used to synthesize the Laplacian texture image to the target model. Finally, the textured target model is reconstructed based on adjusted Laplacian coordinates of the target model. Some of their results are shown in Figure 6.

FIGURE 6 Laplacian texture method by Ran et al. [11]. From left to right: example texture, target model and the synthesis result

For synthesizing repetitive elements from a small input exemplar to a large output domain, Ma et al. [12] present a data-driven method, discrete element textures (as shown in Figure 7). Individual element aggregate distributions and their properties can be both preserved by the method. Due to the generality, it can be applied to different effects including both physically realistic and artistic ones, and a variety of phenomena, including different distributions, different element properties and dimensionalities. Each element is represented by one or multiple samples. Relevant element attributes are encoded by samples positions including orientation, shape, size, and position. An energy optimization solver and a sample-based neighbourhood similarity metric are proposed to synthesize desired outputs. Both the input exemplars and output domains are observed. Optional constraints such as boundary conditions, orientation fields and physics are all solved. Existing element distributions can also be edited using their method.

FIGURE 7 Synthesis results of Ma et al. [12]. The smaller images are the example textures. The larger ones are the synthesis results

2.3 HEIGHT FIELD METHOD

With land forms guided by a user's sketch, Zhou et al. [13] assemble real terrain data represented by height field to create terrain geometry. First, the user's sketch is segmented into small curvilinear features including branch points, curves and straight segments. Then matching features are found from patches of the input height field data. A combination of 2D warping, graph cut merging and Poisson blending is used to blend together these height field patches. The patches are typically 80×80 height samples in size. They reduce the terrain features such as valleys, mountains and rivers into simplified curvilinear features to speedup this method. Figure 8 shows a result from this approach.
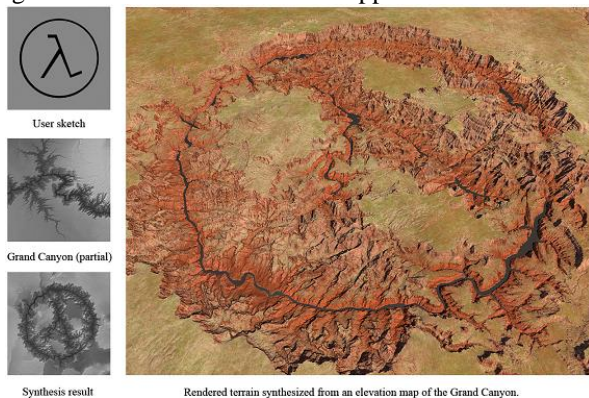
FIGURE 8 Terrain synthesis (right) by Zhou et al. [13] based on example texture (middle left) a user input sketch (upper left).

3 Conclusions

For providing end-users with powerful modelling tools, EGTS is one of the most promising ideas. By combining examples, almost anyone can learn to create results with rich geometric texture. Also, this intuitive approach requires minimal technical skills and experience. We believe that EGTS provides insights and key tools towards efficient rich-detail modelling.

In this paper, we make the first comprehensive survey of the EGTS method, which is a powerful tool to synthesis geometric texture on arbitrary model given an example texture. Since 2004, many influential papers have been published on the topic. Depending on the type of geometric representation (volumetric model, polygon mesh and height field), we classify the previous works into three categories. Of all the three types, with the wide application of mesh model, the literature also mainly focus on EGTS algorithm for polygon mesh, and reaches convincing results.

Though great achievements have been made in the last decade, there are still challenges left for us to solve. For example, the synthesized geometric texture are too uniformly distributed, but sometimes we need a more flexible arrangement of the texture's density. And after the successfully automatic synthesis of geometric texture on an arbitrary model, there are still room for improvement on the complexity of the example texture. We hope this survey will help and encourage researchers to direct towards these challenges.

References

[1] Wei L Y, Lefebvre S, Kwatra V, Turk G 2009 State of the art in example-based texture synthesis *Eurographics 2009 State of the Art Report EG-STAR* 93-117

[2] Turk G 2007 Part II: Texturing surfaces and geometry creation *ACM SIGGRAPH 2007 courses* 2

[3] Bhat P, Ingram S, Turk G 2004 Geometric texture synthesis by example *Proceedings of the 2004 Eurographics/ACM SIGGRAPH symposium on Geometry processing* 41-4

[4] Lagae A, Dumont O, Dutre P 2005 Geometry synthesis by example *Shape Modeling and Applications 2005 International Conference IEEE* 174-83

[5] Merrell P 2007 Example-based model synthesis *Proceedings of the 2007 symposium on Interactive 3D graphics and games* 105-12

[6] Lai Y K, Hu S M, Gu D X,Martin R R 2005 Geometric texture synthesis and transfer via geometry images *Proceedings of the 2005 ACM symposium on Solid and physical modeling* 15-26

[7] Elber G 2005 Geometric texture modeling *Computer Graphics and Applications IEEE* **25**(4) 66-76

[8] Porumbescu S D, Budge B, Feng L, Joy K I 2005 Shell maps *ACM Transactions on Graphics (TOG)* **24**(3) 626-33

[9] Fleischer K W, Laidlaw D H, Currin B L, Barr A H 1995 Cellular texture generation *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques* 239-48

[10] Zhou K, Huang X, Wang X, Tong Y, Desbrun M, Guo B, Shum H Y 2006 Mesh quilting for geometric texture synthesis *ACM Transactions on Graphics (TOG)* **25**(3) 690-7

[11] Ran L Q, Meng X X 2010 Geometry texture synthesis based on Laplacian texture image *Journal of Computer Science and Technology* **25**(3) 606-13

[12] Ma C, Wei L Y, Tong X 2011 Discrete element textures *ACM Transactions on Graphics (TOG)* **30**(4) 62

[13] Zhou H, Sun J, Turk G, Rehg J M 2007 Terrain synthesis from digital elevation models *Visualization and Computer Graphics IEEE Transactions on* **13**(4) 834-48

[14] Hertzmann A, Jacobs C E, Oliver N, Curless B, Salesin D H 2001 Image analogies *Proceedings of the 28th annual conference on Computer graphics and interactive techniques* 327-40

[15] Gu X, Gortler S J, Hoppe H 2002 Geometry images *ACM Transactions on Graphics (TOG)* **21**(3) 355-61

Authors

Lingqiang Ran, born on November 17, 1986, Linyi, China

Current position: 4th year PhD candidate in the School of Computer Science and Technology, Shandong University, Jinan, China.
University studies: B.S. degree in Software College from Shandong University in 2008.
Scientific interest: computer graphics and digital geometry processing.
Publications: 4 papers.

Xiangxu Meng, born on November 20, 1962, Longkou, China

Current position: professor and doctoral supervisor in Shandong University, a senior member of China Computer Federation.
University studies: B.Sc and M.Eng Degrees in Dept. of Computer Science of Shandong University in 1982 and 1985 respectively, PhD degree from the Institute of Computing Technology, Chinese Academy of Science in 1998.
Scientific interest: human-computer interaction, virtual reality, computer graphics, CAD/CAM/CIMS, grid computing, visualization and scientific computing.
Publications: 100 papers.

Two duality problems for a class of multi-objective fractional programming

Xiaoyan Gao*

College of Science, Xi'an University of Science and Technology, Xi'an 710054, China

Abstract

In this article, we investigate the duality results for a class of non-differentiable multi-objective fractional programming problems. The parametric dual models and Wolfe dual models are formulated for this fractional programming. Weak, strong and strict converse duality theorems are established and proved based on the generalized invexity assumptions. Some previous duality results for differentiable multi-objective programming problems turn out to be special cases for the results described in the paper.

Keywords: invexity, multi-objective fractional programming, parametric dual, Wolfe dual

1 Introduction

The term multi-objective programming is an extension of mathematical programming where a scalar valued objective function is replaced by a vector function. Many approaches for multi-objective programming problems have been explored in considerable details, see for example [1,2]. The multi-objective fractional programming refers to a multi-objective problem where the objective functions are quotients of two functions. Recently, the studies about the optimizations of the multi-objective fractional programming problems have also been a focal issue due to in many practical optimization problems the objective functions are quotients. For example, in [3-5] some necessary and sufficient optimality conditions for a feasible point to be an efficient solution for non-differentiable multi-objective fractional programming problems were obtained in the framework of generalized convexity.

Furthermore, duality plays a fundamental role in mathematics, especially in optimization. It has not only used in many theoretical and computational developments in mathematical programming itself but also used in economics, control theory, business problems and other diverse fields. It is not surprising that duality is one of the important topics in multi-objective optimization. A large literature was developed around the duality in multi-objective fractional optimization under the generalized convexity assumption. For example, the results in [6-11] have weakened the convexity hypothesis and made important contribution in duality theorems.

In this paper, motivated by the above work, the duality results are obtained for a class of multi-objective fractional programming problem under the assumptions of $(b,\alpha) - \rho - (\eta,\theta)$ – invexity.

2 Definitions and preliminaries

Definition 1. Let $X \subseteq R^n$. The function $f : X \to R$ is locally Lipschitz on X, if there exists a positive constant k, such that:

$$\left| f(x_1) - f(x_2) \right| \le k \left\| x_1 - x_2 \right\|, \ \forall x_1, x_2 \in X . \tag{1}$$

Definition 2. If $f : X \to R$ is locally Lipschitz on X, the generalized Clarke directional derivative of f at $x \in X$ in the direction $d \in R^n$ is defined by:

$$f^0(x;d) = \lim_{(y,t) \to (x,0^+)} \sup \frac{f(y+td) - f(y)}{t} . \tag{2}$$

The generalized sub-gradient of a locally Lipschitz function f at $x \in X$ is defined by:

$$\partial f(x) = \{ \xi \in R^n : f^0(x;d) \ge \langle \xi, d \rangle, \forall d \in R^n \} . \tag{3}$$

Throughout this paper, we will use the following the definitions given in [11], we always assume that X is an open subset of R^n, $b : X \times X \times [0,1] \to R^+$, $b(x_1, x_2) = \lim_{\lambda \to 0^+} b(x_1, x_2, \lambda) \ge 0, \alpha : X \times X \to R^+ \setminus \{0\}, \eta : X \times X \to R^n$, $\theta : X \times X \to R^n, \rho \in R$.

Definition 3. A Lipschitz function $f : X \to R$ is said to be $(b,\alpha) - \rho - (\eta,\theta)$ – invex at $u \in X$, if there exists b, α, η, θ and ρ, such that:

$$b(x,u)[f(x) - f(u)] \ge \langle \alpha(x,u)\xi, \eta(x,u) \rangle + \rho \left\| \theta(x,u) \right\|^2 . \tag{4}$$

Remark 1. If in the above definition, we have strict inequality for any $x \ne u$, then we say that f is $(b,\alpha) - \rho - (\eta,\theta)$ – strictly invex at $u \in X$.

Corresponding author e-mail: angang21@126.com

Definition 4. A Lipschitz function $f : X \to R$ is said to be $(b, \alpha) - \rho - (\eta, \theta)$ – pseudo-invex at $u \in X$, if there exists b, α, η, θ and ρ, such that:

$$\langle \alpha(x, u)\xi, \eta(x, u) \rangle + \rho \|\theta(x, u)\|^2 \geq 0 \Rightarrow \\ b(x, u)[f(x) - f(u)] \geq 0, \forall x \in X, \xi \in \partial f(u), \tag{5}$$

Definition 5. A Lipschitz function $f : X \to R$ is said to be $(b, \alpha) - \rho - (\eta, \theta)$ – quasi-invex at $u \in X$, if there exists b, α, η, θ and ρ, such that:

$$b(x, u)[f(x) - f(u)] \leq 0 \Rightarrow \\ \langle \alpha(x, u)\xi, \eta(x, u) \rangle + \rho \|\theta(x, u)\|^2 \leq 0, \forall x \in X, \xi \in \partial f(u). \tag{6}$$

Definition 6. A Lipschitz function $f : X \to R$ is said to be $(b, \alpha) - \rho - (\eta, \theta)$ – strictly quasi-invex at $u \in X$, if there exists b, α, η, θ and ρ, such that:

$$b(x, u)[f(x) - f(u)] \leq 0 \Rightarrow \langle \alpha(x, u)\xi, \eta(x, u) \rangle + \\ \rho \|\theta(x, u)\|^2 < 0, \forall x \neq u \in X, \xi \in \partial f(u). \tag{7}$$

3 Parametric duality

In this section, we consider the parametric dual model for (MFP). The dual can be formulated as follows:

$$\max v = (v_1, v_2, \cdots, v_p)$$

$$s.t. \ 0 \in \sum_{i=1}^{p} \lambda_i [\partial f_i(u) - v_i \partial g_i(u)] + \sum_{j=1}^{m} \mu_j \partial h_j(u),$$

(MFD_1) $f_i(u) - v_i g_i(u) \geqq 0, \text{for all } i = 1, 2, \cdots, p,$

$$\sum_{j=1}^{m} \mu_j h_j(u) \geqq 0,$$

$$\lambda \in \Lambda^{++}, \mu \in R_+^m, v \in R_+^p.$$

Let:

$$D^0 = \{(u, \lambda, \mu, v) \in X \times \Lambda^{++} \times R_+^m \times R_+^p, |$$

$$0 \in \sum_{i=1}^{p} \lambda_i [\partial f_i(u) - v_i \partial g_i(u)] + \sum_{j=1}^{m} \mu_j \partial h_j(u), f_i(u) - v_i g_i(u) \geqq$$

$$0 \text{ with } i = 1, 2, \cdots, p, \sum_{j=1}^{m} \mu_j h_j(u) \geqq 0\}$$

denote the feasible set of (MFD_1).

Theorem 1. (Weak Duality) Let $x^0 \in X^0$ and $(\bar{u}, \bar{\lambda}, \bar{\mu}, \bar{v}) \in D^0$. Suppose that:

(i) f_i is $(b_i, \alpha) - \rho_i - (\eta, \theta)$ – invex at $\bar{u}$, $-g_i$ is $(b_i, \alpha) - \bar{\rho}_i - (\eta, \theta)$ – invex and regular at $\bar{u}$, $i = 1, 2, \cdots, p$;

(ii) h_j is $(c, \alpha) - \gamma_j - (\eta, \theta)$ – invex at $\bar{u}$, $j = 1, 2, \cdots, m$;

(iii) $b_i(\bar{x}, \bar{u}) > 0, i = 1, 2, \cdots, p, c(\bar{x}, \bar{u}) \geqq 0$;

(iv) $\sum_{i=1}^{p} \bar{\lambda}_i (\rho_i + \bar{v}_i \bar{\rho}_i) + \sum_{j=1}^{m} \bar{\mu}_j \gamma_j \geqq 0.$

Then $F(\bar{x}) \nleq \bar{v}$.

Proof: Suppose contrary to the result that $F(\bar{x}) \leq \bar{v}$. That implies:

$$\frac{f_i(\bar{x})}{g_i(\bar{x})} \leqq \bar{v}_i, \text{for } i = 1, 2, \cdots, p, \tag{8}$$

and:

$$\frac{f_k(\bar{x})}{g_k(\bar{x})} < \bar{v}_k, \text{for some } k \in \{1, 2, \cdots, p\}. \tag{9}$$

That is:

$$f_i(\bar{x}) - \bar{v}_i g_i(\bar{x}) \leqq f_i(\bar{u}) - \bar{v}_i g_i(\bar{u}), \text{for } i = 1, 2, \cdots, p, \tag{10}$$

and:

$$f_k(\bar{x}) - \bar{v}_k g_k(\bar{x}) < f_k(\bar{u}) - \bar{v}_k g_k(\bar{u}), \\ \text{for some } k \in \{1, 2, \cdots, p\}. \tag{11}$$

Since $(\bar{u}, \bar{\lambda}, \bar{\mu}, \bar{v}) \in D^0$ and $b_i(\bar{x}, \bar{u}) > 0, i = 1, 2, \cdots, p$, the above inequalities yield:

$$\sum_{i=1}^{p} b_i(\bar{x}, \bar{u}) \bar{\lambda}_i [f_i(\bar{x}) - \bar{v}_i g_i(\bar{x})] < \\ \sum_{i=1}^{p} b_i(\bar{x}, \bar{u}) \bar{\lambda}_i [f_i(\bar{u}) - \bar{v}_i g_i(\bar{u})]. \tag{12}$$

Also, using $\bar{x} \in X^0$ and $c(\bar{x}, \bar{u}) > 0$, we have:

$$c(\bar{x}, \bar{u}) \sum_{j=1}^{m} \bar{\mu}_j h_j(\bar{x}) \leqq c(\bar{x}, \bar{u}) \sum_{j=1}^{m} \bar{\mu}_j h_j(\bar{u}). \tag{13}$$

Adding the above two inequalities together, we get:

$$\sum_{i=1}^{p} b_i(\bar{x}, \bar{u}) \bar{\lambda}_i (f_i(\bar{x}) - \bar{v}_i g_i(\bar{x})) + c(\bar{x}, \bar{u}) \sum_{j=1}^{m} \bar{\mu}_j h_j(\bar{x}) < \\ \sum_{i=1}^{p} b_i(\bar{x}, \bar{u}) \bar{\lambda}_i (f_i(\bar{u}) - \bar{v}_i g_i(\bar{u})) + c(\bar{x}, \bar{u}) \sum_{j=1}^{m} \bar{\mu}_j h_j(\bar{u}), \tag{14}$$

By the hypothesis (i) and (ii), we have:

$$b_i(\bar{x}, \bar{u})(f_i(\bar{x}) - f_i(\bar{u})) \geqq \\ \langle \alpha(\bar{x}, \bar{u})\xi_i, \eta(\bar{x}, \bar{u}) \rangle + \rho_i \|\theta(\bar{x}, \bar{u})\|^2, \forall \xi_i \in \partial f_i(\bar{u}), \tag{15}$$

$$b_i(\bar{x}, \bar{u})(-g_i(\bar{x}) + g_i(\bar{u})) \geqq \\ \langle \alpha(\bar{x}, \bar{u})(-\zeta_i), \eta(\bar{x}, \bar{u}) \rangle + \bar{\rho}_i \|\theta(\bar{x}, \bar{u})\|^2, \forall \zeta_i \in \partial g_i(\bar{u}), \tag{16}$$

$$c(\bar{x}, \bar{u})(h_j(\bar{x}) - h_j(\bar{u})) \geqq \\ \langle \alpha(\bar{x}, \bar{u})\tau_j, \eta(\bar{x}, \bar{u}) \rangle + \gamma_j \|\theta(\bar{x}, \bar{u})\|^2, \forall \tau_j \in \partial h_j(\bar{u}). \tag{17}$$

Multiplying Equation (15) with $\bar{\lambda}_i$ and Equation (16) with $\bar{\lambda}_i \bar{v}_i$, then summing up these equations, we obtain:

$$\sum_{i=1}^{p} b_i(\bar{x},\bar{u})\bar{\lambda}_i (f_i(\bar{x}) - f_i(\bar{u}) - \bar{v}_i g_i(\bar{x}) + \bar{v}_i g_i(\bar{u})) +$$

$$c(\bar{x},\bar{u})\sum_{j=1}^{m} \bar{\mu}_j \left[h_j(\bar{x}) - h_j(\bar{u}) \right] \geqq \quad (18)$$

$$\left\langle \alpha(\bar{x},\bar{u})[\sum_{i=1}^{p} \bar{\lambda}_i(\xi_i - \bar{v}_i\zeta_i) + \sum_{j=1}^{m} \bar{\mu}_j\tau_j], \eta(\bar{x},\bar{u}) \right\rangle +$$

$$\left[\sum_{i=1}^{p} \bar{\lambda}_i(\rho_i + \bar{v}_i \bar{\rho}_i) + \sum_{j=1}^{m} \bar{\mu}_j\gamma_j \right] \left\| \theta(\bar{x},\bar{u}) \right\|^2.$$

According to the constraint condition of (MFD₁) and the hypothesis (iv), we can conclude that

$$\sum_{i=1}^{p} b_i(\bar{x},\bar{u})\bar{\lambda}_i (f_i(\bar{x}) - f_i(\bar{u}) - \bar{v}_i g_i(\bar{x}) + \bar{v}_i g_i(\bar{u})) +$$

$$c(\bar{x},\bar{u})\sum_{j=1}^{m} \bar{\mu}_j [h_j(\bar{x}) - h_j(\bar{u})] \geqq 0. \quad (19)$$

We have a contradiction. Hence, the result is true.

Theorem 2. (Weak Duality) Let $\bar{x} \in X^0$ and $(\bar{u},\bar{\lambda},\bar{\mu},\bar{v}) \in D^0$. Suppose that:

(i) $f_i - \bar{v}_i g_i$ is $(b_i, \alpha) - \rho_i - (\eta, \theta) -$ strictly quasi-invex and regular at $\bar{u}, i = 1, 2, \cdots, p$;

(ii) h_j is $(c, \alpha) - \gamma_j - (\eta, \theta) -$ invex at $\bar{u}$, $j = 1, 2, \cdots, m$;

(iii) $b_i(\bar{x},\bar{u}) \geqq 0, i = 1, 2, \cdots, p, c(\bar{x},\bar{u}) > 0$;

(iv) $\sum_{i=1}^{p} \bar{\lambda}_i\rho_i + \sum_{j=1}^{m} \bar{\mu}_j\gamma_j \geqq 0$;

Then $F(\bar{x}) \nleq \bar{v}$.

Proof: Suppose contrary to the result that $F(\bar{x}) \leq \bar{v}$. That implies:

$$\frac{f_i(\bar{x})}{g_i(\bar{x})} \leqq \bar{v}_i, \text{for } i = 1, 2, \cdots, p, \quad (20)$$

and:

$$\frac{f_k(\bar{x})}{g_k(\bar{x})} < \bar{v}_k, \text{for some } k \in \{1, 2, \cdots, p\}. \quad (21)$$

That is:

$$f_i(\bar{x}) - \bar{v}_i g_i(\bar{x}) \leqq f_i(\bar{u}) - \bar{v}_i g_i(\bar{u}), \text{for } i = 1, 2, \cdots, p, \quad (22)$$

and:

$$f_k(\bar{x}) - \bar{v}_k g_k(\bar{x}) < f_k(\bar{u}) - \bar{v}_k g_k(\bar{u}), \text{for some } k \in \{1, 2, \cdots, p\}. (23)$$

Since $b_i(\bar{x},\bar{u}) \geqq 0$, the above inequalities yield:

$$b_i(\bar{x},\bar{u})[(f_i(\bar{x}) - \bar{v}_i g_i(\bar{x})) - (f_i(\bar{u}) - \bar{v}_i g_i(\bar{u}))] \leqq 0. \quad (24)$$

By the hypothesis (i), we have:

$$\left\langle \alpha(\bar{x},\bar{u})(\xi_i - \bar{v}_i\zeta_i), \eta(\bar{x},\bar{u}) \right\rangle + \rho_i \left\| \theta(\bar{x},\bar{u}) \right\|^2 < 0,$$

$$\forall \xi_i \in \partial f_i(\bar{u}), \zeta_i \in \partial g_i(\bar{u}), i = 1, 2, \cdots, p. \quad (25)$$

Since $\bar{\lambda}_i > 0$, we obtain:

$$\left\langle \alpha(\bar{x},\bar{u})\sum_{i=1}^{p} \bar{\lambda}_i(\xi_i - \bar{v}_i\zeta_i), \eta(\bar{x},\bar{u}) \right\rangle + \sum_{i=1}^{p} \bar{\lambda}_i\rho_i \left\| \theta(\bar{x},\bar{u}) \right\|^2 < 0. (26)$$

According to the hypothesis (ii), we get:

$$c(\bar{x},\bar{u})(h_j(\bar{x}) - h_j(\bar{u})) \geqq \left\langle \alpha(\bar{x},\bar{u})\tau_j, \eta(\bar{x},\bar{u}) \right\rangle +$$

$$\gamma_j \left\| \theta(\bar{x},\bar{u}) \right\|^2, \forall \tau_j \in \partial h_j(\bar{u}), j = 1, 2, \cdots, m. \quad (27)$$

Since $\bar{x} \in X^0$ and $(\bar{u},\bar{\lambda},\bar{\mu},\bar{v}) \in D^0$, it follows that:

$$c(\bar{x},\bar{u})(\sum_{j=1}^{m} \bar{\mu}_j h_j(\bar{x}) - \sum_{j=1}^{m} \bar{\mu}_j h_j(\bar{u})) \geqq$$

$$\left\langle \alpha(\bar{x},\bar{u})\sum_{j=1}^{m} \bar{\mu}_j\tau_j, \eta(\bar{x},\bar{u}) \right\rangle + \sum_{j=1}^{m} \bar{\mu}_j\gamma_j \left\| \theta(\bar{x},\bar{u}) \right\|^2 \geqq 0. \quad (28)$$

We have:

$$\left\langle \alpha(\bar{x},\bar{u})\left[\sum_{i=1}^{p} \bar{\lambda}_i(\xi_i - \bar{v}_i\zeta_i) + \sum_{j=1}^{m} \bar{\mu}_j\tau_j \right], \eta(\bar{x},\bar{u}) \right\rangle +$$

$$\left[\sum_{i=1}^{p} \bar{\lambda}_i\rho_i + \sum_{j=1}^{m} \bar{\mu}_j\gamma_j \right] \left\| \theta(\bar{x},\bar{u}) \right\|^2 < 0. \quad (29)$$

By the hypothesis (i) and (ii), we have:

$$b_i(\bar{x},\bar{u})(f_i(\bar{x}) - f_i(\bar{u})) \geqq \left\langle \alpha(\bar{x},\bar{u})\xi_i, \eta(\bar{x},\bar{u}) \right\rangle +$$

$$\rho_i \left\| \theta(\bar{x},\bar{u}) \right\|^2, \forall \xi_i \in \partial f_i(\bar{u}), \quad (30)$$

$$b_i(\bar{x},\bar{u})(-g_i(\bar{x}) + g_i(\bar{u})) \geqq \left\langle \alpha(\bar{x},\bar{u})(-\zeta_i), \eta(\bar{x},\bar{u}) \right\rangle,$$

$$\bar{\rho}_i \left\| \theta(\bar{x},\bar{u}) \right\|^2, \forall \zeta_i \in \partial g_i(\bar{u}), \quad (31)$$

$$c(\bar{x},\bar{u})(h_j(\bar{x}) - h_j(\bar{u})) \geqq \left\langle \alpha(\bar{x},\bar{u})\tau_j, \eta(\bar{x},\bar{u}) \right\rangle +$$

$$\gamma_j \left\| \theta(\bar{x},\bar{u}) \right\|^2, \forall \tau_j \in \partial h_j(\bar{u}). \quad (32)$$

Using the constraint condition of (MFD₁), we obtain:

$$\sum_{i=1}^{p} \bar{\lambda}_i\rho_i + \sum_{j=1}^{m} \bar{\mu}_j\gamma_j < 0. \quad (33)$$

We have a contradiction to hypothesis (iv). Hence, the proof is complete.

Theorem 3 (Strong Duality). Let $\bar{x}$ be an efficient solution of (MFP) at which the generalized Kuhn-Tucker constraint qualification is satisfied. Then there exists $\bar{\lambda} \in \Lambda^{++}, \bar{\mu} \in R_+^m$ and $\bar{v} \in R_+^p$, such that $(\bar{x},\bar{\lambda},\bar{\mu},\bar{v})$ is a feasible solution of (MFD₁), and the values of the objective

functions for (MFP) and (MFD$_1$) are equal at $\bar{x}$ and $(\bar{x}, \bar{\lambda}, \bar{\mu}, \bar{v})$.

Furthermore, if the assumptions of Theorem 4 or Theorem 5 hold for all feasible solutions of (MFP) and (MFD$_1$), then $(\bar{x}, \bar{\lambda}, \bar{\mu}, \bar{v})$ is an efficient solution of (MFD$_1$).

Proof: Since is an efficient solution of (MFP) for which the generalized Kuhn-Tucker constraint qualification is satisfied, then there exists $\bar{\lambda} \in \Lambda^{++}, \bar{\mu} \in R_+^m$ and $\bar{v} \in R_+^p$, such that $(\bar{x}, \bar{\lambda}, \bar{\mu}, \bar{v})$ satisfies the conditions of (GKT). The conditions imply that $\bar{v}_i = \dfrac{f_i(\bar{x})}{g_i(\bar{x})}$, for all $i = 1, 2, \cdots, p$.

It is clear that the values of the objective functions for (MFP) and (MFD$_1$) are equal at $\bar{x}$ and $(\bar{x}, \bar{\lambda}, \bar{\mu}, \bar{v})$.

In addition, from the weak duality Theorem 4 or Theorem 5, for any feasible solution $(x, \lambda, \mu, v) \in D^0$, the following cannot hold:

$$\frac{f_i(\bar{x})}{g_i(\bar{x})} \leqq v_i, \text{for } i = 1, 2, \cdots, p, \tag{34}$$

and:

$$\frac{f_k(\bar{x})}{g_k(\bar{x})} < v_k, \text{for some } k \in \{1, 2, \cdots, p\}. \tag{35}$$

Hence, we conclude that $(\bar{u}, \bar{\lambda}, \bar{\mu}, \bar{v})$ is an efficient solution of (MFD$_1$).

Theorem 4 (Strict Converse Duality): let $\bar{x}$ and $(\bar{u}, \bar{\lambda}, \bar{\mu}, \bar{v})$ be an efficient solution for (MFP) and (MFD$_1$), respectively with $\bar{v}_i = \dfrac{f_i(\bar{x})}{g_i(\bar{x})}$, for all $i = 1, 2, \cdots, p$. Suppose that:

(i) $f_i - \bar{v}_i g_i$ is $(b_i, \alpha) - \rho_i - (\eta, \theta)$ – strictly invex and regular at $\bar{u}, i = 1, 2, \cdots, p$;

(ii) h_j is $(c, \alpha) - \gamma_j - (\eta, \theta)$ – invex at $\bar{u}$, $j = 1, 2, \cdots, m$;

(iii) $b_i(\bar{x}, \bar{u}) \geqq 0 (i = 1, 2, \cdots, p), c(\bar{x}, \bar{u}) \geqq 0$;

(iv) $\sum_{i=1}^{p} \bar{\lambda}_i \rho_i + \sum_{j=1}^{m} \bar{\mu}_j \gamma_j \geqq 0$.

Then $\bar{x} = \bar{u}$, $\bar{u}$ is an efficient solution for (MFP).

Proof: Suppose contrary to the result that $\bar{x} \neq \bar{u}$.

By the hypothesis (i), we have:

$$b_i(\bar{x}, \bar{u})[(f_i(\bar{x}) - \bar{v}_i g_i(\bar{x})) - (f_i(\bar{u}) - \bar{v}_i g_i(\bar{u}))] >$$
$$\langle \alpha(\bar{x}, \bar{u})(\xi_i - \bar{v}_i \zeta_i), \eta(\bar{x}, \bar{u}) \rangle + \rho_i \|\theta(\bar{x}, \bar{u})\|^2, \tag{36}$$
$$\forall \xi_i \in \partial f_i(\bar{u}), \zeta_i \in \partial g_i(\bar{u}), i = 1, 2, \cdots, p.$$

From the constraint condition of (MFD$_1$) and $b_i(\bar{x}, \bar{u}) \geqq 0$, we get:

$$\langle \alpha(\bar{x}, \bar{u})(\xi_i - \bar{v}_i \zeta_i), \eta(\bar{x}, \bar{u}) \rangle + \rho_i \|\theta(\bar{x}, \bar{u})\|^2 < 0. \tag{37}$$

Since $\bar{\lambda}_i > 0$, the above inequality yields:

$$\left\langle \alpha(\bar{x}, \bar{u}) \sum_{i=1}^{p} \bar{\lambda}_i (\xi_i - \bar{v}_i \zeta_i), \eta(\bar{x}, \bar{u}) \right\rangle +$$
$$\sum_{i=1}^{p} \bar{\lambda}_i \rho_i \|\theta(\bar{x}, \bar{u})\|^2 < 0 \tag{38}$$

Using the hypothesis (ii), we have:

$$c(\bar{x}, \bar{u})(h_j(\bar{x}) - h_j(\bar{u})) \geqq \langle \alpha(\bar{x}, \bar{u}) \tau_j, \eta(\bar{x}, \bar{u}) \rangle +$$
$$\gamma_j \|\theta(\bar{x}, \bar{u})\|^2, \forall \tau_j \in \partial h_j(\bar{u}), j = 1, 2, \cdots, m, \tag{39}$$

Since $\bar{\mu}_j \geqq 0, j = 1, 2, \cdots, m$, the inequality follows:

$$c(\bar{x}, \bar{u})(\sum_{j=1}^{m} \bar{\mu}_j h_j(\bar{x}) - \sum_{j=1}^{m} \bar{\mu}_j h_j(\bar{u})) \geqq$$
$$\left\langle \alpha(\bar{x}, \bar{u}) \sum_{j=1}^{m} \bar{\mu}_j \tau_j, \eta(\bar{x}, \bar{u}) \right\rangle + \sum_{j=1}^{m} \bar{\mu}_j \gamma_j \|\theta(\bar{x}, \bar{u})\|^2 \geqq 0, \tag{40}$$

From the constraint condition of (MFD$_1$) and the above inequality, we obtain

$$\left\langle \alpha(\bar{x}, \bar{u}) \sum_{j=1}^{m} \bar{\mu}_j \tau_j, \eta(\bar{x}, \bar{u}) \right\rangle + \sum_{j=1}^{m} \bar{\mu}_j \gamma_j \|\theta(\bar{x}, \bar{u})\|^2 \leqq 0, \tag{41}$$

Summing up Equations (38) and (41), we concludes that

$$\left\langle \alpha(\bar{x}, \bar{u}) \left[\sum_{i=1}^{p} \bar{\lambda}_i (\xi_i - \bar{v}_i \zeta_i) + \sum_{j=1}^{m} \bar{\mu}_j \tau_j \right], \eta(\bar{x}, \bar{u}) \right\rangle +$$
$$\left[\sum_{i=1}^{p} \bar{\lambda}_i \rho_i + \sum_{j=1}^{m} \bar{\mu}_j \gamma_j \right] \|\theta(\bar{x}, \bar{u})\|^2 < 0 \tag{42}$$

The above inequality together with the hypothesis (iv) implies

$$\left\langle \alpha(\bar{x}, \bar{u}) \left[\sum_{i=1}^{p} \bar{\lambda}_i (\xi_i - \bar{v}_i \zeta_i) + \sum_{i=1}^{p} \bar{\mu}_j \tau_j \right], \eta(\bar{x}, \bar{u}) \right\rangle < 0. \tag{43}$$

We have a contradiction. Hence $\bar{x} = \bar{u}$.

4 Wolfe duality

In this section, we consider the Wolfe type dual model for (MFP). The Wolfe type dual can be formulated as follows: (MFD$_2$):

$$\max G(u,\mu) = (\frac{f_1(u) + \sum_{j=1}^{m}\mu_j h_j(u)}{g_1(u)}, \cdots, \frac{f_p(u) + \sum_{j=1}^{m}\mu_j h_j(u)}{g_p(u)})$$

$$s.t. \quad 0 \in \sum_{i=1}^{p}\lambda_i g_i(u)[\partial f_i(u) + \sum_{j=1}^{m}\mu_j \partial h_j(u)]$$

$$- \sum_{i=1}^{p}\lambda_i[f_i(u) + \sum_{j=1}^{m}\mu_j h_j(u)]\partial g_i(u),$$

$$\lambda \in \Lambda^{++}, \mu \in R_+^m, u \in X.$$

Let:

$$E^0 = \{(u,\lambda,\mu) \in X \times \Lambda^{++} \times R_+^m$$

$$\Big| 0 \in \sum_{i=1}^{p}\lambda_i g_i(u)[\partial f_i(u) - v_i \partial g_i(u)]\sum_{j=1}^{m}\mu_j \partial h_j(u),$$

$$f_i(u) - v_i g_i(u) \geqq 0 \text{ with } i = 1,2,\cdots,p, \sum_{j=1}^{m}\mu_j h_j(u) \geqq 0\}$$

denote the feasible set of (MFD$_2$).

Theorem 5. (Weak Duality)

Let $\bar{x} \in X^0$ and $(\bar{u}, \bar{\lambda}, \bar{\mu}) \in E^0$. Suppose that:

(i) f_i is $(b_i, \alpha) - \rho_i - (\eta, \theta)$ – invex at $\bar{u}$, $-g_i$ is $(b_i, \alpha) - \bar{\rho}_i - (\eta, \theta)$ – invex and regular at $\bar{u}, i = 1,2,\cdots,p$;

(ii) h_j is $(c, \alpha) - \gamma_j - (\eta, \theta)$ – invex at $\bar{u}$, $j = 1,2,\cdots,m$;

(iii) $b_i(\bar{x}, \bar{u}) > 0, i = 1,2,\cdots,p, 0 < c(\bar{x}, \bar{u}) \leqq 1$;

(iv) $\rho_i + b_i(\bar{x}, \bar{u})\sum_{j=1}^{m}\bar{\mu}_j \gamma_j \geqq 0, \bar{\rho}_i \geqq 0, i = 1,2,\cdots,p$.

Then $F(\bar{x}) \nleqq G(\bar{u}, \bar{\mu})$.

Proof: since $(\bar{u}, \bar{\lambda}, \bar{\mu},) \in E^0$, it follows that there exists $\xi_i \in \partial f_i(\bar{u}), \zeta_i \in \partial g_i(\bar{u}), \tau_j \in \partial h_j(\bar{u}), i = 1,2,\cdots,p$, $j = 1,2,\cdots,m$, such that:

$$\sum_{i=1}^{p}\bar{\lambda}_i g_i(\bar{u})\Big[\xi_i + \sum_{j=1}^{m}\bar{\mu}_j \tau_j\Big] - \sum_{i=1}^{p}\bar{\lambda}_i[f_i(\bar{u}) - \sum_{j=1}^{m}\bar{\mu}_j h_j(\bar{u})]\zeta_i = 0$$

Suppose contrary to the result of the theorem that $F(\bar{x}) \nleqq G(\bar{u}, \bar{\mu})$. That implies:

$$\frac{f_i(\bar{x})}{g_i(\bar{x})} \leqq \frac{f_i(\bar{u}) + \sum_{j=1}^{m}\bar{\mu}_j h_j(\bar{u})}{g_i(\bar{u})}, \text{for } i = 1,2,\cdots,p, \quad (44)$$

and

$$\frac{f_k(\bar{x})}{g_k(\bar{x})} < \frac{f_k(\bar{u}) + \sum_{j=1}^{m}\bar{\mu}_j h_j(\bar{u})}{g_k(\bar{u})}, \text{for some } k \in \{1,2,\cdots,p\}. \quad (45)$$

Therefore:

$$f_i(\bar{x})g_i(\bar{u}) - g_i(\bar{x})[f_i(\bar{u}) + \sum_{j=1}^{m}\bar{\mu}_j h_j(\bar{u})], \quad (46)$$
$$\leqq 0, \text{for } i = 1,2,\cdots,p$$

and:

$$f_k(\bar{x})g_k(\bar{u}) - g_k(\bar{x})\Big[f_k(\bar{u}) + \sum_{j=1}^{m}\bar{\mu}_j h_j(\bar{u})\Big] < 0 \quad (47)$$

for some $k \in \{1,2,\cdots,p\}$.

Since:

$$b_i(\bar{x}, \bar{u}) > 0, g_i(\bar{u}) > 0 \text{ and } \bar{\mu}_j h_j(\bar{x}) \leqq 0, (i = 1,2,\cdots,p),$$
$(j = 1,2,\cdots,m)$, the above inequalities yield

$$\sum_{i=1}^{p}\bar{\lambda}_i g_i(\bar{u})b_i(\bar{x}, \bar{u})[f_i(\bar{x}) + \sum_{j=1}^{m}\bar{\mu}_j h_j(\bar{x})] -$$
$$\sum_{i=1}^{p}\bar{\lambda}_i g_i(\bar{x})b_i(\bar{x}, \bar{u})[f_i(\bar{u}) + \sum_{j=1}^{m}\bar{\mu}_j h_j(\bar{u})] < 0. \quad (48)$$

Thus:

$$\sum_{i=1}^{p}\bar{\lambda}_i g_i(\bar{u})b_i(\bar{x}, \bar{u})\Big[(f_i(\bar{x}) - f_i(\bar{u})) +$$
$$\Big(\sum_{j=1}^{m}\bar{\mu}_j h_j(\bar{x}) - \sum_{j=1}^{m}\bar{\mu}_j h_j(\bar{u})\Big)\Big] - \sum_{i=1}^{p}\bar{\lambda}_i\Big[f_i(\bar{u}) + \sum_{j=1}^{m}\bar{\mu}_j h_j(\bar{u})\Big]$$
$$b_i(\bar{x}, \bar{u})(g_i(\bar{x}) - g_i(\bar{u})) < 0. \quad (49)$$

By the hypothesis (i) and (ii), we have:

$$b_i(\bar{x}, \bar{u})(f_i(\bar{x}) - f_i(\bar{u})) \geqq \langle \alpha(\bar{x}, \bar{u})\xi_i, \eta(\bar{x}, \bar{u}) \rangle +$$
$$\rho_i \|\theta(\bar{x}, \bar{u})\|^2, \forall \xi_i \in \partial f_i(\bar{u}). \quad (50)$$

$$b_i(\bar{x}, \bar{u})(-g_i(\bar{x}) + g_i(\bar{u})) \geqq$$
$$\langle \alpha(\bar{x}, \bar{u})(-\zeta_i), \eta(\bar{x}, \bar{u}) \rangle + \bar{\rho}_i \|\theta(\bar{x}, \bar{u})\|^2, \forall \zeta_i \in \partial g_i(\bar{u}). \quad (51)$$

$$c(\bar{x}, \bar{u})(h_j(\bar{x}) - h_j(\bar{u})) \geqq$$
$$\langle \alpha(\bar{x}, \bar{u})\tau_j, \eta(\bar{x}, \bar{u}) \rangle + \gamma_j \|\theta(\bar{x}, \bar{u})\|^2, \forall \tau_j \in \partial h_j(\bar{u}). \quad (52)$$

Since $0 < c(\bar{x}, \bar{u}) \leqq 1$, Equation (52) implies:

$$h_j(\bar{x}) - h_j(\bar{u}) \geqq \langle \alpha(\bar{x}, \bar{u})\tau_j, \eta(\bar{x}, \bar{u}) \rangle + \gamma_j \|\theta(\bar{x}, \bar{u})\|^2. \quad (53)$$

Using $\bar{\mu}_j \geqq 0(j = 1,2,\cdots,m)$, Equation (53) yields:

155

$$\sum_{j=1}^{m}\overline{\mu}_j h_j(\overline{x}) - \sum_{j=1}^{m}\overline{\mu}_j h_j(\overline{u}) \geqq$$
$$\left\langle \alpha(\overline{x},\overline{u})\sum_{j=1}^{m}\overline{\mu}_j\tau_j, \eta(\overline{x},\overline{u})\right\rangle + \sum_{j=1}^{m}\overline{\mu}_j\gamma_j \left\|\theta(\overline{x},\overline{u})\right\|^2 . \tag{54}$$

Multiplying Equations (50) and (54) with $\overline{\lambda}_i g_i(\overline{u})$ and $\overline{\lambda}_i g_i(\overline{u})b_i(\overline{x},\overline{u})$, respectively, and multiplying Equation (51) with $\overline{\lambda}_i(f_i(\overline{u}) + \sum_{j=1}^{m}\overline{\mu}_j h_j(\overline{u}))$, then summing these inequalities together, we obtain:

$$\sum_{i=1}^{p}\overline{\lambda}_i g_i(\overline{u})b_i(\overline{x},\overline{u})\left[\left(f_i(\overline{x})-f_i(\overline{u})\right)+\right.$$
$$\left.\left(\sum_{j=1}^{m}\overline{\mu}_j h_j(\overline{x})-\sum_{j=1}^{m}\overline{\mu}_j h_j(\overline{u})\right)\right] -$$
$$\sum_{i=1}^{p}\overline{\lambda}_i\left[f_i(\overline{u})+\sum_{j=1}^{m}\overline{\mu}_j h_j(\overline{u})\right]b_i(\overline{x},\overline{u})(g_i(\overline{x})-g_i(\overline{u})) \geqq$$
$$\left\langle \alpha(\overline{x},\overline{u})\left[\sum_{i=1}^{p}\overline{\lambda}_i g_i(\overline{u})\left(\xi_i+\sum_{j=1}^{m}\overline{\mu}_j\tau_j\right)-\sum_{i=1}^{p}\overline{\lambda}_i(f_i(\overline{u})+\right.\right.$$
$$\left.\left.\sum_{j=1}^{m}\overline{\mu}_j h_j(\overline{u}))\zeta\right]_i, \eta(\overline{x},\overline{u})\right\rangle + \left[\sum_{i=1}^{p}\overline{\lambda}_i g_i(\overline{u})(\rho_i+b_i(\overline{x},\overline{u})\right.$$
$$\left.\sum_{j=1}^{m}\overline{\mu}_j\gamma_j)+\overline{\rho}_i(f_i(\overline{u})+\sum_{j=1}^{m}\overline{\mu}_j h_j(\overline{u}))\right]\left\|\theta(\overline{x},\overline{u})\right\|^2 . \tag{55}$$

The above inequality together with the constraint conditions and hypothesis (iv), implies

$$\sum_{i=1}^{p}\overline{\lambda}_i g_i(\overline{u})b_i(\overline{x},\overline{u})\left[(f_i(\overline{x})-f_i(\overline{u}))+\right.$$
$$\left.\left(\sum_{j=1}^{m}\overline{\mu}_j h_j(\overline{x})-\sum_{j=1}^{m}\overline{\mu}_j h_j(\overline{u})\right)\right] -$$
$$\sum_{i=1}^{p}\overline{\lambda}_i\left[f_i(\overline{u})+\sum_{j=1}^{m}\overline{\mu}_j h_j(\overline{u})\right],$$
$$b_i(\overline{x},\overline{u})(g_i(\overline{x})-g_i(\overline{u})) \geqq 0, \tag{56}$$

which contradicts Equation (49). This completes the proof.

Theorem 6: (Weak Duality). Let $\overline{x} \in X^0$ and $(\overline{u},\overline{\lambda},\overline{\mu}) \in E^0$. Suppose that:

(i) $f_i(\cdot)+\sum_{j=1}^{m}\overline{\mu}_j h_j(\cdot)$ is $(b_i,\alpha)-\rho_i-(\eta,\theta)-$ invex and regular at $\overline{u}$, $-g_i(\cdot)$ is $(b_i,\alpha)-\overline{\rho}_i-(\eta,\theta)-$ invex and regular at $\overline{u}, i=1,2,\cdots,p$;

(ii) $b_i(\overline{x},\overline{u}) > 0, i=1,2,\cdots,p$;

(iii) $\rho_i \geqq 0, \overline{\rho}_i \geqq 0, i=1,2,\cdots,p$.

Then $F(\overline{x}) \not\leqq G(\overline{u},\overline{\mu})$.

Proof: the proof is similar to the Theorem 5.

Theorem 7. (Strong Duality) Let $\overline{x}$ be an efficient solution of (MFP). Suppose that there exists $\overline{\lambda} \in \Lambda^{++}$ and $\overline{\mu} \in R_+^m$, such that $\overline{\mu}_j h_j(\overline{x}) = 0 (j=1,2,\cdots,m)$ and $(\overline{u},\overline{\lambda},\overline{\mu})$ is a feasible solution of (MFD₂). Then the objective function values of (MFP) and (MFD₂) are equal at $\overline{x}$ and $(\overline{u},\overline{\lambda},\overline{\mu})$.

Furthermore, if the assumptions of Theorem 8 or Theorem 9 hold for all feasible solutions of (MFP) and (MFD₂), then $(\overline{u},\overline{\lambda},\overline{\mu})$ is an efficient solution of (MFD₂).

Proof: Based on the weak duality theorem, we can obtain the result that $(\overline{u},\overline{\lambda},\overline{\mu})$ is an efficient solution of (MFD₂).

Theorem 8. (Strict Converse Duality) Let $\overline{x}$ and $(\overline{u},\overline{\lambda},\overline{\mu})$ be an efficient solutions of (MFP) and (MFD₂), respectively. Suppose that:

(i) $\dfrac{f_i(\overline{x})}{g_i(\overline{x})} \leqq \dfrac{f_i(\overline{u})+\sum_{j=1}^{m}\overline{\mu}_j h_j(\overline{u})}{g_i(\overline{u})}$, $i=1,2,\cdots,p$;

(i) $f_i(\cdot)+\sum_{j=1}^{m}\overline{\mu}_j h_j(\cdot)$ is $(b_i,\alpha)-\rho_i-(\eta,\theta)-$ invex and regular at $\overline{u}$, $-g_i(\cdot)$ is $(b_i,\alpha)-\overline{\rho}_i-(\eta,\theta)-$ invex and regular at $\overline{u}, i=1,2,\cdots,p$;

(iii) $b_i(\overline{x},\overline{u}) > 0, \rho_i \geqq 0, \overline{\rho}_i \geqq 0, i=1,2,\cdots,p$;

Then $\overline{x} = \overline{u}$, that is, $\overline{u}$ is an efficient solution of (MFP).

5 Discussion and conclusion

Throughout this paper, we have established two dual models namely parametric duality models and Wolfe duality models for the multiobjective fractional programming problem. Several duality results were derived and proved with the help of $(b,\alpha)-\rho-(\eta,\theta)-$ Invexity assumption. The results should be further opportunities for exploiting this structure of the multiobjective fractional programming problem.

Acknowledgments

References

[1] Jayswal A 2010 *J Glob Optim* **46**(2) 207-16

[2] Gao X Y 2013 *J Netw* **8**(3) 739-46

[3] Gao 2012 X Y *J theor Appl Inf Technol* **46**(1) 347-54

[4] Chen X H 2002 *J Math Anal Appl* **273**(1) 190-205

[5] Nobakhtian S 2008 *J Glob Optim* **41**(1) 103-15

[6] Nahak C, Mohapatra R N 2012 *Optim Lett* **6**(2) 253-60

[7] Kuk H, Lee G M, Tanina T 2001 *J Math Anal Appl* **262**(1) 365-75

[8] Bae K D, Kang Y M, Kim D S 2010 *Ann Oper Res* **154**(2) 133-47

[9] Long X J, Huang N J, Liu Z B 2008 *J Ind Manag Optim* **4**(2) 287-98

[10] Gao X Y 2012 *Int. Rev. Comput. Softw* **7**(5) 2574-2581

[11] An G, Gao X Y 2013 *J Comput Theor Nanosc* **10**(12) 2943-8

Author

Xiaoyan Gao, born in August 1979, Shanxi, China

University studies: Master's degree in optimization theory and applications at Yan'an University, China in 2005.
Scientific interests: generalized convexity, the optimization theory and applications for semi-infinite and multi-objective programming.

Research of hand gesture using Kinect based on finger recognition

Zhuo Shi[1, 2, 3], Yinghui Li[2], Ke Yu[2], Yuanquan Cheng[2], Changshao Zhou[2*]

[1]*South China University of Technology, Guangzhou, Guangdong 510275, China*

[2]*Computer Science & Engineering of the Guilin University of Electronic Technology, Guilin, Guangxi 541004, China*

[3]*Guangxi Key Laboratory of Trusted Software (Guilin University of Electronic Technology)*

Abstract

Using the depth image from Kinect, we present a novel method to archive the Hand Gesture Recognition (HGR). Firstly we overview the major technical components of the complete method. Then we elaborate several key challenges such as palm recognition, contour analysis and fingertip detection. We present the Biggest concave point detect schemes to contour analysis which is the key to achieving the HGR. Finally, we demonstrate the feasibility and effectiveness of the method by running the system.

Keywords: human-computer interaction, hand gesture recognition, finger recognition, Kinect

1 Introduction

In our life, we sometimes have to communicate with others by using hand gesture or sign language. For example, the place where quiet is required, and the cooperation between doctor and nurse in the operating room, etc. But most of us are not familiar with some special gesture, let alone sign language. In order to solve this problem, we need a way to help us get the meaning of some gesture or even sign language. In this paper, we get the depth data from Kinect, then we focus on the hand area, after a group of processes, we get finger recognition, finally meaning of hand gesture is display on the screen.

In order to archive a believable HGR system, several challenges have to address. First, the skeletal tracking in Kinect for Windows SDK cannot provide the finger recognition, so we need to focus on the research of finger tracked. Second, we are only interested in the hand region, thus it must be segmented from the background. Third, a believable finger recognition algorithm is needed. After read a lot of paper, we opt to use an improved Convex and Concave Angle Recognition Algorithm [1].

We will first discuss related previous work in Section 2. We then give an overview of our HGR system in Section 3 followed by details of the above mentioned key components. Lastly, we present our HGR system, and discuss the limitations of the current system for potential future improvements.

2 Related Work

There are about two main methods for dynamic hand gesture recognition: one is based on hardware, for

example, data glove; the other one is based on the computer vision. At first, almost all the HGR are based on the wearable device [2-4]. By using data glove, we can directly get details of the hand joints. When the computer gets the data, after some analysis and process, we can quickly get feedback from computer, Human-Computer Interaction (HCI) complete. In China, there is already some research on Chinese Sign Language by using data glove or some other devices [5, 6]. For a period time, it is really useful for HGR or HCI by using extra hardware. But then people found it was not convenient, not only for researchers, but also for users. Because this hardware is expensive, the common user cannot afford one. What's worse is that users should be wearing the device all through the recognizing. It seems inconvenient indeed. In order to solve this problem, researchers noticed our natural hands, and then they focused on the recognition of natural hands only by computer vision. HGR based on computer vision is mainly divided into two aspects: coloured markers at the tip of fingers [7] and natural hand recognition [8, 9]. The former one requires coloured markers at tip of users' fingers; it is still complicated for us. Nowadays HGR with non-invasive, natural and convenient features is the trend of the research. According to the different technologies, it can be divided into 2D images and 3D images. With common camera, we can quickly get 2D images and then analyse what we are interested in. It seem to be simple and will not bother users, but there are lots of difficulties in distinguishing the target region from the complicated background, let alone recognizing gestures, and it's not easy to make computer get the meaning of any movement. The recognition rate that based on 2D images cannot satisfy the researchers

---

[*]***Corresponding author*** e-mail: shzh.cn@gmail.com

Shi Zhuo, Li Yinghui, Yu Ke, Cheng Yuanquan, Zhou Changshao

despite of growing amount of excellent algorithms and methods. It begins a new era to see the existing problems in the area of HGR, as the sensors which can give depth information, are available. Microsoft Kinect, which is released in 2010, is one of the examples and it is able to detect individual motion. The Kinect cannot provide the finger recognition, but we focus on the depth data, combining with the traditional algorithms on the 2D images, then we optimize the algorithm and enhance believability and feasibility of the system.

3 System overview

3.1 GETTING DEPTH IMAGE FROM KINECT

The Kinect contains infrared camera and PrimeSense sensor to compute the depth of the object while the RGB camera is used to capture the images. The depth images and RGB image of the object could be getting at the same time. This 3D scanner system called Light Coding, which employs a variant of image-based 3D reconstruction. The depth output of Kinect is 11 bit with 2048 levels of sensitivity [10]. Supposing that the dimension in *x-y* coordinate system of image is $w \times h$, the real coordinate system is like Figure 1.
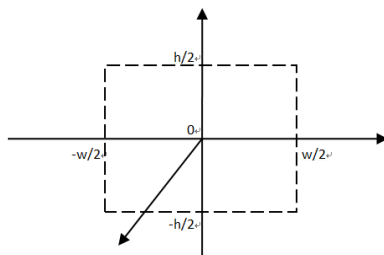
FIGURE 1 The Kinect coordinate diagram

The depth value of a point in 3D can be defined as calibration procedure [11].

$$d = K \tan\left(\frac{d_{raw}}{2842.5} + 1.1863\right) - O ,$$

where d is the depth of that point in m, $K = 0.1236$m and $O = 0.037$m.

According to the depth of the point, we can get the location (i, j, d) of the pixel in depth image from Kinect, then we transform (i, j, d) into (x, y, z), which is the location in three-dimensional system of coordinate in reality. The transform equation is as follows:

$$\begin{cases} x = \left(i - \dfrac{w}{2}\right)(d - 10)s\dfrac{w}{h} \\ y = \left(j - \dfrac{h}{2}\right)(d - 10)s \\ z = d \end{cases} .$$

3.2 HAND TRACKING AND SEGMENTING

From above, we can get the height of user, the schematic shows in Figure 2.
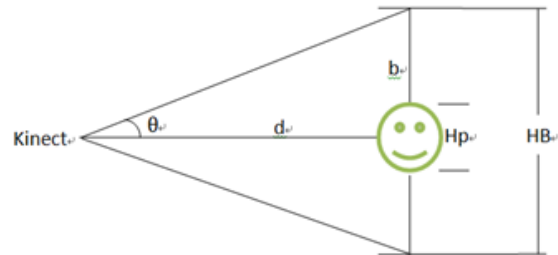
FIGURE 2 The principle diagram of the height measurement

The equation is as follows:

$$b = d \tan(\theta) ,$$

$$\frac{Hp}{HB} = \frac{Hr}{2b} \underset{equals}{\Rightarrow} Hr = \frac{2bHp}{HB} ,$$

d is the distance from the point to Kinect, θ is the vertical angle above the horizontal line, Hp is the height of user's pixels, HB is the height of background pixels and Hr is the height of user's body, what we want to get. According to the related research [12], we know that there are some statistical relations between the height of body and length of hand. From [12] we can get the following equation:

$$HL = \frac{Hr - 75.70178}{5.082380} ,$$

where HL is the length of hand, Hr is height of user.

As we know, Kinect can track hand but finger detection. We now get the hand length, the hand data should be in the spherical area, which radius is half of the hand length, and the centre is hand point, the equation is as follows:

$$HandPoint = \left\{ p \,\middle\|\, p(x, y, z) - p(x_0, y_0, z_0) \,\middle\|\, \leq \frac{HL}{2} \right\} ,$$

where $p(x_0, y_0, z_0)$ is hand point, HL is hand length. After this step, we can get the hand region what we want, and make convenient for hand processing.

FIGURE 3 Hand tracking and segmenting

Shi Zhuo, Li Yinghui, Yu Ke, Cheng Yuanquan, Zhou Changshao

3.3 PALMCONTOUR RECOGNITION

In last part, we have got the hand region. And now we need to simplify the coordinate system, so we put them on plane projection. We define the distance between two points like this:

$$D(P_1, P_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \ ,$$

$P_1(x_1,y_1)$, $P_2(x_2,y_2)$ are the two points.

Clustering is the task of grouping a set of objects in such a way that objects in the same group. We use K-means clustering algorithm [13]. We put N pixels into K pixel clusters, then we get. Every time before the new pixel gets into one cluster, it should be chosen the one by the value, which is mean value of all the points' range in the cluster. After the iteration we get the minimum value in different cluster

After the K-means clustering, we process these pixels by setting the threshold value of pixels in a cluster. In this way, we can eliminate the interference of noise pixels, and then we get the mask image like in Figure 4.

FIGURE 4 The mask images of hand after K-means clustering algorithm

Next we use the graham scan algorithm to detect the convex hull sets of the hand image above, and we choose Moore neighbourhood algorithm to detect the contour of the mask of hand. The two algorithms processes show in Figures 5 and 6. After contour recognition we define centre of palm circle as the centre of hand instead of the centre of hand cluster. Because centre point will be used to compute the direction of each finger, we need a more stable point. The differences are showed in Figure 7.
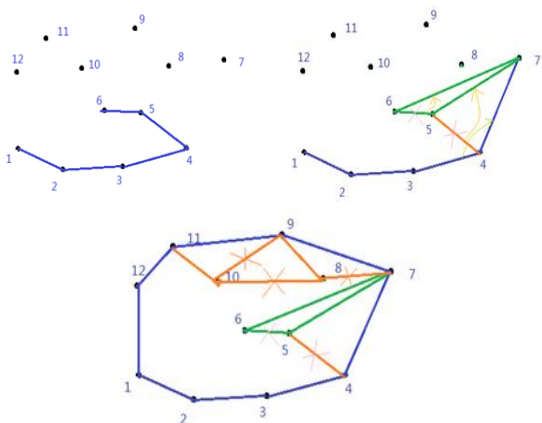
FIGURE 5 Graham scan algorithm process sample

(1) (2) (3) (4)

(5) (6) (7) (8)
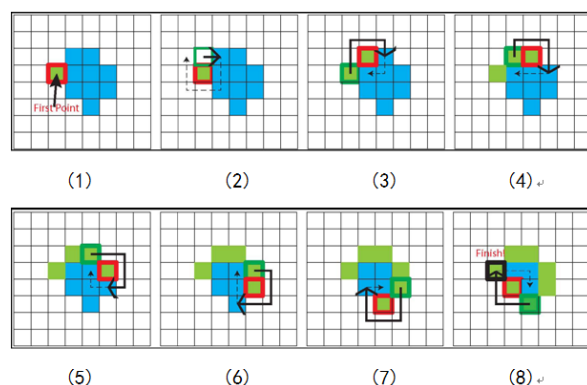
FIGURE 6 Moore neighbourhood algorithm process sample
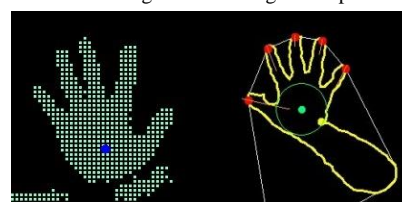
FIGURE 7 The differences of the centre
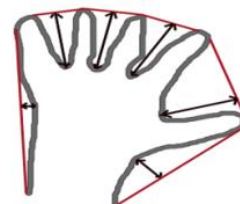
FIGURE 8 Biggest concave point detect schemes

3.4 FINGERTIPS DETECTIONS

Now we present recognition algorithm of concave and convex angle to detect fingertips. Firstly, we need to find the farthest concave point, which locates between two convex points. In order to get the points, we need the hand contour point chain table and concave point set chain table, both of the data structure are generated before. Secondly, we connect the one concave point to the adjacency convex point and connect it to the concave point on the other side, then get the angle of the two lines. If the angle is less than the threshold value, abandon it, or else the convex point is the fingertip we detect. The principle is showed in Figures 8 and 9, and the end of fingertip detection is showed in Figure 10.
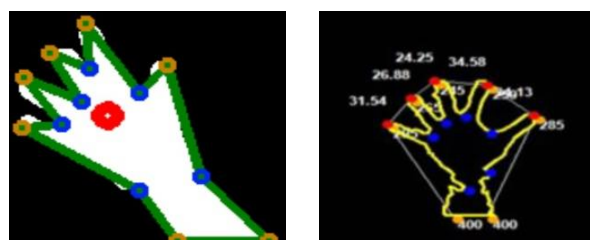
FIGURE 9 Algorithm of fingertip recognition by angle diagram sketch
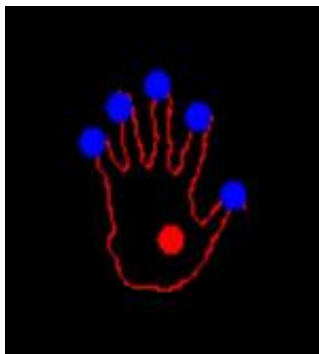
FIGURE 10 Fingertip Detection

3.5 HAND GESTURE RECOGNITION

In this part, firstly we recognize the different fingers, and then confirm the direction of each one. The direction is start from palm centre to fingertip. We use a vector to represent the direction. Suggesting where is the palm centre point, where is a fingertip, then is the vector. Secondly, we require the user to stretch their hands to make computer compute the distance of each fingertip, then with four steps showing in Figure 11, we can get the particular finger and display them on screen, like in Figure 12.
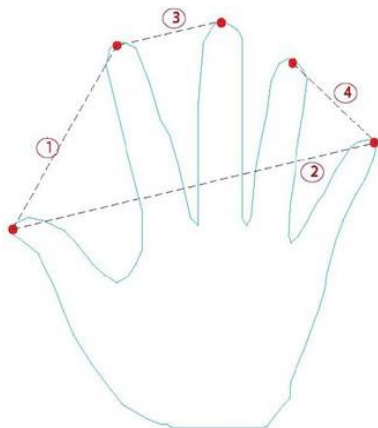
FIGURE 11 Four steps of identify fingers diagram
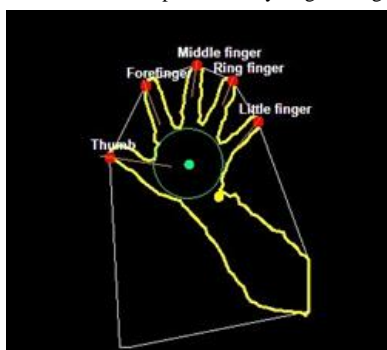
FIGURE 12 Finger recognition result

Thirdly we use a decision tree to classify hand gesture. The decision tree has three layers, the chance nodes of first layer are the finger counters, the second ones are names of finger, and the last are the angles of vectors. After the process above, we get HGR shown in Figures 13-15.
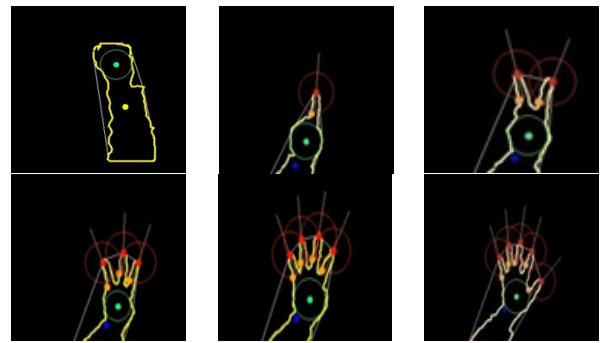
FIGURE 13 Number gesture recognition: zero to five
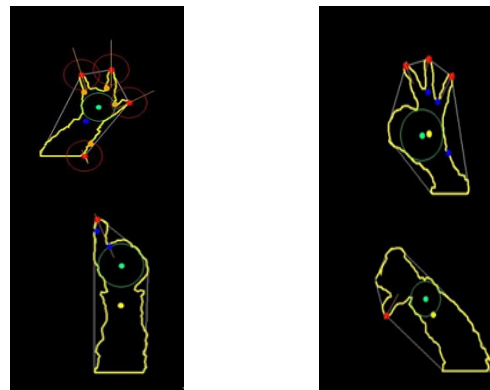
FIGURE 14 International gesture recognition: "I love you", "Okay", "Good", "Bad"

FIGURE 15 Chinese number sign recognition: six, seven and eight

4 Conclusions

In this research we analyse and design the processing flow of gesture recognition system, using a means clustering algorithm for image pre-processing, contour recognition with Moore neighbourhood algorithm, Graham scanning algorithm and related algorithms for convex point detection, and we opt to three-point alignment algorithm and the bump maximum angle for fingertip recognition. According to the geometrical relationship of the feature points, we get gesture recognized. Finally, we complete the whole HGR prototype system.

After testing the prototype system, we can find that it is proved to be high-performance. The system can correctly get the geometrical relationship of feature points on fingertips and palm, and then give a reliable feedback displayed on the screen. But the system also has some shortcomings: the method how Kinect capture depth images is the speckle, which reflected back with the infrared, is not good enough. The more distance between user and Kinect, the less reliability it will reach. Because part of palm cannot entirely reflect the infrared speckle, the

161

Kinect cannot get the accurate depth data of the palm; it will result in glinting slimly. Therefore when recognize fingertip in the depth stream again and again, the fingertip will also glint at a certain small range. The problem will be solved in Kinect 2.

References

[1] Mei P 2010 Fingertips Recognition based on Contour and curvature *Anhui: a dissertation for Master's Degree* University of Science and Technology of China

[2] Liang R, Ouhyoung M 1998 A Real-Time Continuous Gesture Recognition System for Sign Language *The Third International Conference on Automatic Face and Gesture Recognition* 558-65

[3] Nakamura T, Takahashi S, Tanaka J 2008 Double-Crossing: A New Interaction Technique for Hand Gesture Interfaces *Lecture Notes in Computer Science* **5068** 292-300

[4] Hu Y 2005 Gesture Recognition Technology Review *China Science and Technology Information* 41-2

[5] Wu J, Gao W, Chen X 1999 A System Recognizing Chinese Finger-Spelling Alphabets Based On Data-Glove Input PR&AI **12** 74-8

[6] Li Y, Chen X, Zhang X, Wang K, Wang Z J 2012 *IEEE Transactions On Bio-Medical Engineering* **59**(10) 2695-704

[7] Maes P, Mistry P 2009 Unveiling the "Sixth Sense" game-changing wearable tech *TED 2009* Long Beach CA USA

[8] Guo C, Zhang L, Sun L 2008 Sign Language Recognition and Classification Based on Computer Vision *Science & Technology Information* **17** 248-9

[9] Yan Y 2012 The study of Gesture Recognition Based on Video *Wuhan: a dissertation for Master's Degree* Central China Normal University.

[10] *http://en.wikipedia.org/wiki/Kinect.*

[11] Raheja J L, Chaudhary A, Singal K 2011 Tracking of Fingertips and Centres of Palm using Kinect *The Third International Conference on Computational Intelligence, Modelling and Simulation (CIMSiM)* 248-52

[12] Wang X, Wang S 1990 Applied Multivariate Statistical Analysis *Shanghai Science and Technology Press*

[13] Yan P, Kong Y, Kong D 2008 Sign Language Recognition based on K-means/SCHMM Multi-stage Classification *Microcomputer Information Press* 187-9

Authors

Zhuo Shi, born on August 25, 1978, Changzhou, China

Current position: Doctor of Computer Software and Theory, vice-professor and Master supervisor in Guilin University of Electronic Theology.
University studies: Mechanical and Electronic Engineering in Shandong University
Scientific interest: digital media, virtual reality and natural interaction.
Publications: 3 Patents, 30 Papers

Yinghui Li, born on August 28, 1988, Guilin, China

Current position: Master of Computer Technology in Guilin University of Electronic Technology.
University studies: Computer Science and Technology in Wuhan Polytechnic University.
Scientific interest: graphic algorithm and virtual reality.
Publications: 1 patents.

Ke Yu, born on April 22, 1979, Liuzhou, China

Current position: Guilin University of Electronic Theology.
University studies: Guangxi Arts Institute.
Scientific interest: virtual reality.
Publications: 3 papers.
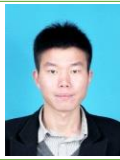
Yuanquan Cheng, born on February 24, 1989, Guilin, China

Current position: Master of Mathematics in Guilin University of Electronic Technology.
University studies: Mathematics and Applied Mathematics in Fuyang Teachers College.
Scientific interest: graphic algorithm and virtual reality.
Publications: 1 patent.

ChangShao Zhou, born on November 21, 1978, GuangZhou, China

Current position: Teacher of Guangzhou College of Technology and Business.
University studies: Master of Software Engineering in Guilin University of Electronic Technology,.
Scientific interest: big data, graphic algorithm and virtual reality.
Publications: 1 patents, 2 papers.

Test method of railway video surveillance system

Zhenyu Zhang*, Yong Qin

[1]*State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing 100044, China*

[2]*School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, China*

Abstract

Test method of railway video surveillance system was mainly explored in this paper. In order to evaluate accurately railway video surveillance system, based on railway video surveillance system test platform, an objective, direct and effective test system was established for testing equipment requirements, system functionality and performance. In addition, it provides standard of railway video surveillance system and makes field of video surveillance standardization.

Keywords: railway, video surveillance, test method

1 Introduction

As a means of intuitive, real-time and accurate remote monitoring and management, video surveillance plays an important role in social safety, economic development and the construction of all industries. With the development of Chinese railway, the video surveillance system become more and more important in transport scheduling, public security, organization, security monitoring and other aspects [1]. All kinds of railway information system need the video surveillance system urgently. The video surveillance system has been widely used in key areas in railway station, the signal room, rail overpass, it provides better protect for the safety of railway. However, these systems have some problems, such as limited application, small-scale networks, not having the conditions for networking and sharing resources, maintenance management is not standardized. In order to meet Chinese railway leapfrog development of the new requirements, establishing full coverage, three-dimensional, highly reliable railway video surveillance system with height information sharing has become the focus of system and the only way [2].

2 Test purpose and test system

Railway video surveillance system has entered the standardized operation and management stage currently, but as time goes on, the shortcomings in system application process gradually come out, multi-point with long line, multi-equipment model, network technology is more complex, and different technical standard. Testing and acceptance of equipment's' quality and functional performance in railway video surveillance system is not perfect, and lack scientific and effective methods of testing and evaluating system. For these reasons, an objective,

direct and effective test system is an urgent need to test the railway equipment quality and overall functionality of railway video surveillance system and accurately evaluate the system.

Based on "Railway video surveillance system technical specifications" and "Video security monitoring system design specification", learn from the previous research [3-8] on railway video surveillance system, the following system was proposed to establish for testing and evaluating railway video surveillance system, shown in Figure 1.
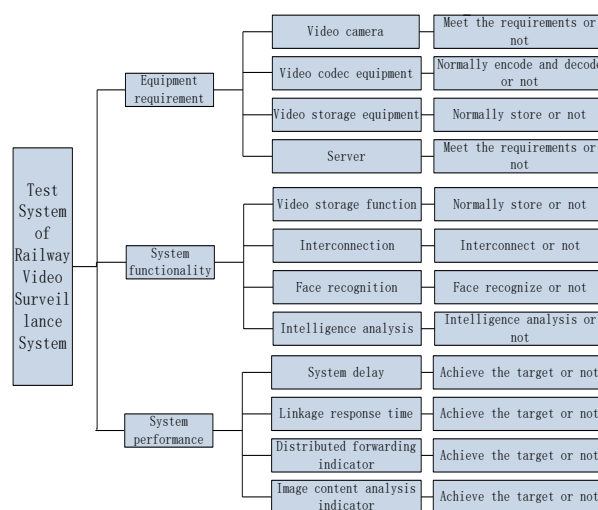
FIGURE 1 Test system diagram of railway video surveillance system

The system mainly focuses on the analysis of equipment requirements, system functionality and system performance in railway video surveillance system.

Corresponding author e-mail: zhangzhenyu7@foxmail.com

3 Railway video surveillance system platform test

3.1 TEST ENVIRONMENT

Railway video surveillance system platform test environment topology has been shown in Figure 2.
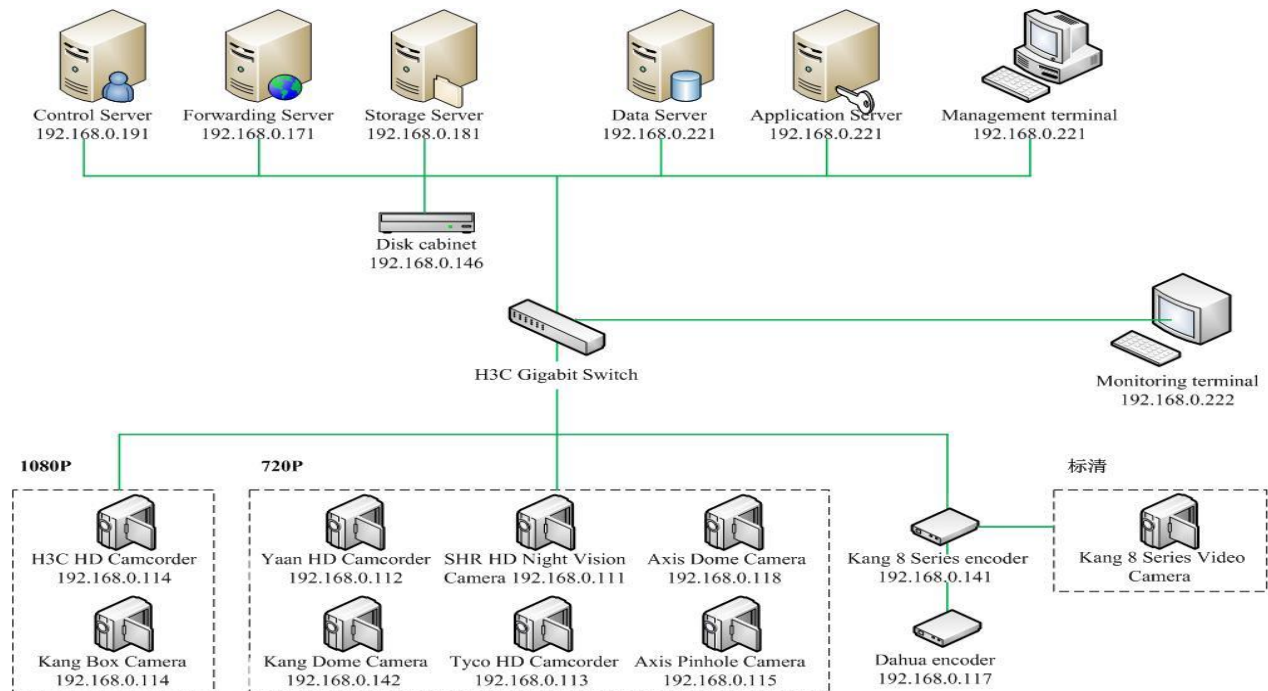
FIGURE 2 Connection diagram of test environment

3.2 EQUIPMENT REQUIREMENTS

The equipment of railway video surveillance system has been no strict requirements of the standard equipment. Different companies have different equipment's and standards, and the equipment is not compatible with each other, which makes the whole system run by a very good big impact. Therefore, in order to make the whole railway video surveillance system to function, to reduce the possibility of error, the railway sector makes hard targets to the equipment of the requirements. These equipment's include cameras, codec's equipment's, storage equipment, servers, etc. Railway video surveillance system platform hardware parameters are shown in Table 1.

TABLE 1 Hardware configuration table of railway comprehensive video surveillance system platform

N	Equipment Name	Equipment Type	Specification	Remark
1	Server	BCECR-N8000	CPU: Intel Core 2 Duo E7500 Memory: 4G	3
2	Lenovo Desktop Terminal	Think Centre	CPU: Intel Core 2 Duo E7500 Memory: 4G Graphics: NVIDIA GeForce G100	2
3	Kang encoder	Kang	Kang 8 Series	1
4	SHR HD Night Vision Camera	SHR	SHR-HLV400HK	1
5	Yaan HD Camcorder	Yaan	HDH5407E-H65-R21	1
6	Tyco HD Camcorder	Tyco	AV1305DN	1
7	Kang Dome Camera	Kang	DS-2CD763PF-E 2.7-9mm	1
8	Kang Box Camera	Kang	DS-2CD877DF	1
9	Axis Pinhole Camera	Axis	Unknown	1
10	Axis Dome Camera	Axis	AXIS P3344 6MM Fixed	1
11	H3C HD Camcorder	H3C	IPC-HIC-5421	1
12	Dahua encoder	Dahua	DH/DVR1604HE	1
13	Disk cabinet	SkySan	SkySan SS1500	1

Seen from Table 1, railway video surveillance system in accordance with the equipment requirements, cameras, video encoding and decoding equipment, video storage equipments, servers and other equipment are essential to meet specified technical requirements. On this basis, the railway video surveillance system built the platform preparing for subsequent system functional and performance testing.

3.3 FUNCTIONALITY TEST

Railway video surveillance system functions include video acquisition, processing, real-time monitoring, storage, playback, PTZ control, video distribution / forwarding feature, the video content analysis, alarm, linkage, system interconnection, system management and user management.

Because the system function contains the complex and multiple contents, real-time video viewing, PTZ control, video playback and video storage were selected as test objects, and then evaluated the railway video surveillance system.

1. Real-time video viewing. All access equipment's can normally browse images (single-screen, multi-screen). Through the HD camera's configuration, the terminal can display text, time and other OSD overlay information.

2. PTZ control. All access cameras only Yaan HD cameras, SHR HD night vision cameras and Kang 8 Series SD PTZ camera support PTZ control, the above three kinds of equipment's are able to achieve PTZ call, zoom and focus functions.

3. Video playback. The platform used storage server for video. The video stored on disk cabinet and can playback history through monitoring terminal.

4. Video storage. The merits of storage equipment's and storage mode are important to video surveillance system. Good storage equipment's are helpful to improve operational efficiency and stability of the video surveillance system and greatly enhance its overall performance. Here Iometer software was used to test the performance of storage equipment's.

Iometer is a tool used to measure and describe the I/O subsystem work in single and clustered systems. It is responsible for the determination of the control system performance under load.

In order to obtain the performance value of the storage equipment, the storage performance under different conditions were tested, such as the load, sequential random percentage, read and write percentage.

TABLE 2 Performance value of storage equipment under different loads

Project Load	IPOS	MBS	Average Response Time	Maximum Response Time	CPU Utilization
4K	22362.95	87.36	0.72	18.48	31.26%
8K	11775.55	92	1.36	33.43	22.38%
16K	5982.97	93.48	2.67	63.8	13.75%
32K	3346.82	104.59	4.78	57.7	9.42%
64K	1750.95	109.43	9.14	61.22	7.73%
128K	878.62	109.83	18.21	90.92	7.38%
256K	447.46	111.86	35.76	110	6.61%
512K	224.28	112.14	71.35	127.14	6.05%
1024K	111.6	111.6	143.27	168.15	6.34%

IOPS is capabilities of read and write per second, MBS is the amount of data transmitted per second.

Table 2, respectively, describes the 5 performance indicators of the storage equipment's, showed that read and write per second and CPU utilization decreased as the load increased, the average IO response time and the maximum IO response time increased significantly as the load increased and the amount of data transmitted per second change little while load changed. It can be seen that the data transmit capability of the storage equipment's is strong. In the condition of high load, the read and write capacity per second become weak and both the average IO response time and the maximum IO response time are longer, which are limited to the equipment only in the low load operation.

TABLE 3 The performance value of read and write percentage of different storage equipment

Project Read and write	IPOS	MBS	Average Response Time	Maximum Response Time	CPU Utilization
80%	111.53	111.53	143.38	274.03	9.06%
60%	111.07	111.07	143.94	316.1	11.17%
40%	112.66	112.66	141.92	301.66	12.92%
20%	115.49	115.49	138.48	468.86	16.54%

As is shown in Table 3, the read and write percentage have few impacts on read and write capacity, the amount of data transmitted per second and the average response time of IO.

TABLE 4 Performance value of storage equipment's under different sequential random percentage

Project Sequence	IPOS	MBS	Average Response Time	Maximum Response Time	CPU Utilization
80%	110.52	110.52	144.66	283.32	6.55%
60%	109.82	109.82	145.58	235.26	6.5%
40%	109.47	109.47	146.19	247.71	7.08%
20%	109.89	109.89	145.6	262.29	6.6%

In the 5 performance indicators of the storage equipment described in Table 4, the sequential random percentage changes a lot, the indexes of storage equipment's changed little, it is stable within a certain range. The equipment can work efficiently under different sequential random percentage.

Seen from the test results, the performance values meet all the normal operating needs and storage equipment's performance is good. But the equipment's in a high load condition, the reading and writing literacy per second decline, the average IO response time, maximum IO response time and the CPU utilization all increase. This shows that there are certain problems of the stability of equipment when it works under high load. The read and write percentage and the sequential random percentage have few impacts on performance of storage equipment's. The equipment can work effectively under different read and write percentage and sequential random percentage, the disruption is small. As the equipment's' limited, this test cannot test a variety of equipment's in contrast, so only one single equipment to be tested.

3.4 PERFORMANCE TEST

System performance tests include video network performance, network bandwidth, response latency, distribution and forwarding capabilities, stability and reliability, quality and assessment of system images and other aspects.

1. System delay. System delay is the time from the monitoring platform issuing commands to the time front-end equipment executing the command. By monitoring and analysing network signalling processes, the time of terminal issuing the command and the time of the recipient response to the confirmation message are easy to obtain, the relative difference between the two is test delay. The results show that, the case of automated testing tools simulate 200 users simultaneously request to video, first packet video data is received taking 3.728 seconds maximum, minimum taking 0.035 seconds, average time is 0.438 seconds.

2. The forwarding server. As Gigabit Ethernet performance is not good, the server configuration is low. In 110-way store and 200 users simultaneously request to single-channel video, then forwarding server CPU utilization is 38% and bandwidth is 294.1Mbits/s.

3. Storage server. Because the test conditions cannot afford to multiple users simultaneously requesting and browsing video, so this test use automated testing tools simulate multiple (200) users for real-time video viewing. Iometer was used to simulate the 100-way video source, and the storage server videoed and stored. Because of lower server configuration, the CPU utilization is 40% and bandwidth is 245.1Mbits/s when the storage server video 110-way (HD + simulation, where HD 108-way, SD 2-way).

Note: The total bandwidth is low, due to Axis HD camera with a better coding algorithm. Without reducing the image quality to lower image stream, Axis is helpful to transmit image data.

4. Network bandwidth. Video network bandwidth was decided by the number of user terminals, real-time video images large ones that each user terminal allows simultaneous transfer to see, network flow control mechanisms, storage mechanism, image quality requirements of each real-time video, historical image quality requirements.

Based on the platform, six types of cameras were tested for single-screen, multi-screen (dual-screen, three screens, four screens) real-time image viewing, analysis its monitoring terminal static and dynamic image occupy computer performance (CPU, bandwidth). Test configuration has been shown in Figure 3.
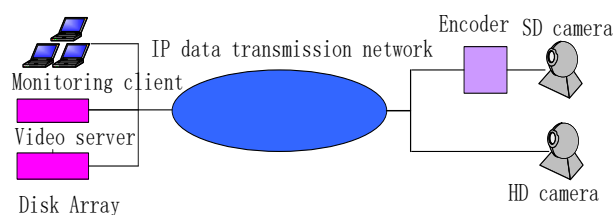
FIGURE 3 Test diagram of network bandwidth configuration

Table 5 shows the CPU and bandwidth:

TABLE 5 CPU, bandwidth consumption of single-screen, multi-screen static and dynamic images

	Single screen	Dual screen	Three screens	Four screen	Explanation
Video camera1	(10%, 4M)	(21%, 6.9M)	(25%, 10.4M)	(38%, 13.7M)	Static images
	12%, 4.8M)	(26%, 9.1M)	(36%, 13.3M)	(51%, 17.5M)	Dynamic images
Video camera 2	(9%, 4.5M)	(16%, 9.0M)	(34%, 13.4M)	(44%, 18.0M)	Static images
	15%, 4.4M)	(27%, 8.5M)	(49%, 15.8M)	(65%, 22.3M)	Dynamic images
Video camera 3	10%, 4.5M)	(25%, 9.2M)	(39%, 13.5M)	(55%, 18.2M)	Static images
	11%, 4.4M)	(33%, 9.7M)	(43%, 14.3M)	(56%, 16.8M)	Dynamic images
Video camera4	(11%, 18.3M)	(23%, 43M)	(36%, 63.4M)	(53%, 85M)	Static images
	(12%, 20.9M)	(26%, 39.5M)	(40%, 58M)	(47%, 79M)	Dynamic images
Video camera5	(18%, 180K)	(39%, 336.1K)	(55%, 524.5K)	(78%, 700.5K)	Static images
	(21%,1.3M)	(43%, 3.3M)	(61%, 3.2M)	(85%, 5.6M)	Dynamic images
Video camera6	(19%, 700.6K)	(41%, 1.4M)	(56%, 2.1M)	(80%, 2.8M)	Static images
	19%, 1.3M)	(55%, 6.8M)	(73%, 9.6M)	(92%, .1M)	Dynamic images

After the system performance test, system latency, forwarding servers, storage servers, network bandwidth meet the requirements. It has good platform compatibility and can access 8 types of equipment's, including six types of high-definition and two types of standard definition. Terminal interface on the platform can browse all types of real-time image and playback the history images normally.

In the terminal configuration, the monitoring terminal maximum decoding is 4-way 720p, CPU utilization is about 38%, while 1-way 1080p, CPU utilization is about 48%. In the condition of the server configuration, forwarding server maximum support 310 channels video data forwarding, storage server supports 110 channels storing. The server CPU utilization reached 40% in the moment, far exceeding the requirements of engineering applications (10% -20%),it is recommended for servers and terminal hardware configuration upgrade (Note: engineering applications generally require the terminal CPU utilization below 40%, the server CPU utilization below 20%). Delay indicator of asking video showed no significant impact on HD mixing with SD test.

4 Conclusions

In this paper, mainly using Imoeter software tested on equipment requirements, system functionality and performance, including video storage equipment's, network bandwidth in a variety of states. The results show that this method can be applied to the various functional and performance testing of the railway video surveillance system, and can achieve better results.

Through research and test, a set of railway video surveillance system testing methods is initially formed to accurately evaluate the railway video surveillance system to ensure that the system can stably and efficiently run.

Acknowledgments

References

[1] Qin Y, Peng H J, Zhu L X 2010 Design and implementation of interoperable platform of railway integrated video monitoring system *Railway Signalling & Communication Engineering* **7** China Railway Publishing House Beijing 17-20 (*in Chinese*)

[2] Huang Z H 2012 Railway video surveillance engineering testing technical analysis *Railway Signalling & Communication Engineering* **9** China Railway Publishing House Beijing 26-9 (*in Chinese*)

[3] Lin H W, Tu D, Li G H 2003 Based on cutting off background technology moving target detection method *Journal of the National Defense University* **25** National Defense University Press Changsha 66-9 (*in Chinese*)

[4] Liu X, Liu H, Qiang Z P, Geng X T 2008 Adaptive background modeling based on mixture gaussian model and frame subtraction *Journal of image and graphics* **13** Science Press Beijing 729-34 (*in Chinese*)

[5] Wang J 2010 Research on detection technology of railway video surveillance system *Railway Signalling & Communication Engineering* **7** China Railway Publishing House Beijing 10-2 (*in Chinese*)

[6] Chang C-c, Lin L-k, Tsai C-h 2010 A real-time surveillance system of electric multiple units (EMUs) in Taipei MRT *Proceeding 2010 Sixth International Conference on Natural Computation IEEE Piscataway* 4425-9 (*in Chinese*)

[7] Bocchetti G, Flammini F, Pragliola C, Pappalardo A 2009 Dependable integrated surveillance systems for the physical security of metro railways *Third ACM/IEEE International Conference on Distributed Smart Cameras(ICDSC) IEEE Computer Society Piscataway* 7-10

[8] Xie Z Y, Dong B T, Chen Y L, Yang Y F 2011 Comparative study on moving detection method of video surveillance system in China high-speed railway transport hub *Key Engineering Materials* v467-469 *Trans Tech Publications Elsevier Switzerland* 503-8

Authors

Zhenyu Zhang, born in 1991, Nanchang, China

Current position: Master student of Beijing Jiaotong University, China.
Scientific interest:: transportation safety, intelligent transportation, reliability of urban rail vehicle.

Yong Qin, born in 1971, Xuzhou, China

Current position: Doctor of Traffic information engineering and control, professor and doctoral supervisor in Beijing Jiaotong University, China.
University studies: Transportation signal and control in Shanghai Tiedao University.
Scientific interest: intelligent transportation system, transportation safety, transportation information system, intelligent control.

Information and Computer Technologies

An Improved model of product design case reuse based on extension theory

Wang Li*

Department of materials science and Engineering, Luoyang Institute of Science and Technology

Abstract

This article studies the question of product design case reuse based on extension theory for complex product with various information, categories and attributes. It also proposes to use case space elementary system and case subject index elementary model based on elementary model of product design case research to describe complex design case. Besides, it also builds design case research and reuse model based on extension correlation function to acquire the most similar design result. The application of this method can acquire the design case closest to the design objective rapidly, thus improving the efficiency of product configuration, and shortening design period of complex product. Finally, it testifies the effectiveness of the model with application cases.

Keywords: Product Design, CBR, Design Reuse, Extension Theory, Artificial Intelligence

1 Introduction

Case based research (CBR) technology is an important design method for the rapid intelligent design of large complex products. It is an analogical reasoning method by simulating human beings' thoughts to seek intelligent and rapid solution and decision of the design question, search for similar design case according to design requirements, and adapting to the new design configuration by adjusting to related design parameters. [1-4] CBR technology is a rapidly developing research method in the field of artificial intelligent research. But traditional CBR technology has its own limitations that influence its deeper application in large complex product designs. For example, the attributes of complex product design are usually complicated, yet traditional case model cannot clearly describe the case reasoning nor combine quantitative and qualitative attributes, which influences the modelling of design case with various attributes, categories and information. Secondly, traditional CBR system mainly uses qualitative case indexing research, such as inductive indexing method, approximate indexing method and knowledge guidance method. But the design of large complex product is always fuzzy and uncertain and traditional CBR can hardly deal with fuzzy research questions that combined with qualitative and quantitative methods. At present, scholars have carried out some research upon these questions. Zhao Yanwei studied the product configuration design method and applied the calculation of extension distance in chain saw configuration design [5]. Dou Zengfa studied the hybrid expert system based on extension theory and case based research. [6] Wang Tichun and his colleagues studied the multiple-level case research model for product design

based on knowledge reuse [7]. Therefore, based on past research, this article will describe complex design case through case space elementary system and case subject index elementary model and build design case research and reuse model based on extension correlation function, in order to study the product design case reuse based on extension theory and testify the calculation method.

2 Extension theory and extension cases

Extension theory was first proposed by Chinese scholar, Professor Cai Wen, in 1983 as a new trans-disciplinary subject. It aims to study the transformation rules and solutions of contradictory problems. It uses elementary model (as the basic logical unit to understand and analyze problems) to carry out formalized and modeled presentation of the study object and objective. Meanwhile, in order to quantify the solving process and realize the computerization of the transformation of contradictory problems, extension theory establishes corresponding extension set, correlation function and other mathematical tools, to describe the object's attribute and transformation.[8-10] The extension unit describes the design question's extension affiliates and the essence of extension transformation. Correlation functions are concepts used to describe the correlation between objects within the extension unit. At present, the theoretical framework and technical method of extension theory are based on elementary theory, extension unit theory and extension logic. Their respective application technologies in each field are called extension engineering and have generated related engineering results. [11-15]

Elementary is one of logic cells of extension theory. It presents the design object in a formalized and modelled

---

* *Corresponding author* e-mail: luoyangwangli@yeah.net

way. Elementary is expressed by a sequential triad $J = (N, c, v)$, among which N represents the name of the design object, c its features and v means the object N's magnitude about the design feature c. One design object can have several features. The design object N's features $c_1, c_2, ..., c_n$ and corresponding magnitudes $v_1, v_2, ..., v_n$ can be expressed as:

$$J = \begin{bmatrix} N & c_1 & v_1 \\ & c_2 & v_2 \\ & ... & ... \\ & c_3 & v_3 \end{bmatrix} = \begin{bmatrix} J_1 \\ J_2 \\ ... \\ J_n \end{bmatrix},$$ (1)

where J is called n-dimensional elementary. The introduction of multi-dimensional elementary can describe the design object in a more formalized and modelled way, thus providing a new thought for the design of large complex products. No matter how many design layers or how complicated the design problem, it can always be expressed by multi-dimensional elementary and examined in case space.

Definition 1 Case space elementary system J_S Gather all design cases and related design information, build a model based on the elementary to form a design case space elementary system J_S. Every elementary of the design case is one element in the case space elementary system. The case space elementary system represents design results and conditions corresponding the past design requirements.

$$J_S = (J_D, J_R, J_O, J_{CON}).$$ (2)

In the formula, J_D represents the identification of design case elementary space system, which is the attribute of design case elementary. J_R represents design case elementary, J_O the design objective and J_{CON} the design restrictive information in the system.

As can be seen, the design case space elementary proposed by this article extends the presentation of design case. The features of design objects can describe the design result of the case, namely case information including both the static basic attributes and the design behaviours during the process of developing a design. They can also express the design objectives, which determine the essence of design case in different design cases. Furthermore, they can express the initial conditions that give rise to the case and the restrictive information, which the case should also stick to when satisfying the design objectives. It is beneficial for the rapid design of products with huge information, various categorization and complicated attributes.

Definition 2 Design case subject index elementary

J_{IND} is the concentration and representation of design case elementary and also an important signal that distinguish it from other design case elementary.

$$J_{IND} = \begin{bmatrix} N(J_{IND}) & cx_1 & v_1 | f(cx_1) \\ & cx_2 & v_2 | f(cx_2) \\ & ... & ... \\ & cx_n & v_n | f(cx_n) \end{bmatrix},$$ (3)

where $N(J_{IND})$ represents the name of design case subject index, cx the features of subject index and $v | f(cx)$ the magnitude function of subject index.

Definition 3 Design case elementary base J_B. It is the physical embodiment of design case space elementary system in a well-organized and managed way. The base should include design elementary related information, subject index information and information correlation.

$$J_B = (J_s, J_{IND}, J_C),$$ (4)

where J_C represents the correlated elementary in design case.

3 Model and calculation method of complex product design based on extension theory

During the process of rapid design of large complex products, the CBR technology, based on design case space, needs to divide the product design into subspaces, establish allocation platform for the design space, describe the design case space with elementary, establish the design needs with elementary and transform them into objective elementary. Then use design case subject index elementary to carry out matching research in the design case space. If a certain design case subspace satisfies the needs of the design objective elementary, then the design case in this design case subspace can be used as reusing objects. Thus, the objective of rapid design for large complex products is achieved.

3.1 ELEMENTARY MODELS IN CLASSICAL DOMAIN AND JOINT DOMAIN IN DESIGN CASE SPACE

Assume there are m design case sub-spaces in a design case space. The No. i sub-space J_{Si} has n subject index, then the classical domain elementary model can be expressed a follows:

$$
\boldsymbol{J}_{Si} = \begin{bmatrix} N_i, C, V \end{bmatrix} = \begin{bmatrix} N_i & C_{i1} & V_{i1} \\ & C_{i2} & V_{i2} \\ & \vdots & \vdots \\ & C_{in} & V_{in} \end{bmatrix} = \begin{bmatrix} N_i & C_{i1} & v_{i1} \mid f(C_{i1}) \\ & C_{i2} & v_{i2} \mid f(C_{i2}) \\ & \vdots & \vdots \\ & C_{in} & v_{in} \mid f(C_{in}) \end{bmatrix} = \begin{bmatrix} N_i & C_{i1} & \langle v_{ai1}, v_{bi1} \rangle \mid f(C_{i1}) \\ & C_{i2} & \langle v_{ai2}, v_{bi2} \rangle \mid f(C_{i2}) \\ & \vdots & \vdots \\ & C_{in} & \langle v_{ain}, v_{bin} \rangle \mid f(C_{in}) \end{bmatrix} , \tag{5}
$$

where $C_{i1}, C_{i2}, ..., C_{in}$ are the different subject indexes for N_i. $V_{i1}, V_{i2}, ..., V_{in}$ are the classical domains for N_i corresponding the different subject indexes $C_{i1}, C_{i2}, ..., C_{in}$ respectively. $v_{i1}, v_{i2}, ..., v_{in}$ show the range of the subject indexes $C_{i1}, C_{i2}, ..., C_{in}$, namely the classical domain of N_i, which satisfies

$V_{ij} = \langle v_{aij}, v_{bij} \rangle (i = 1, 2, ..., m, j = 1, 2, ..., n)$.

$f(C_{i1}), f(C_{i2}), ..., f(C_{in})$ are the matching functions for the different subject indexes $C_{i1}, C_{i2}, ..., C_{in}$ for N_i.

Based on the classical domain of design case space elementary model, there can be another joint domain J_P:

$$
\boldsymbol{J}_P = (N_0, C, V) = \begin{bmatrix} N_0 & C_{01} & V_{01} \\ & C_{02} & V_{02} \\ & \vdots & \vdots \\ & C_{0n} & V_{0n} \end{bmatrix} = \begin{bmatrix} N_0 & C_{01} & v_{01} \mid f(C_{01}) \\ & C_{02} & v_{02} \mid f(C_{02}) \\ & \vdots & \vdots \\ & C_{0n} & v_{0n} \mid f(C_{0n}) \end{bmatrix} = \begin{bmatrix} N_0 & C_{01} & \langle v_{a01}, v_{b01} \rangle \mid f(C_{01}) \\ & C_{02} & \langle v_{a02}, v_{b02} \rangle \mid f(C_{02}) \\ & \vdots & \vdots \\ & C_{0n} & \langle v_{a0n}, v_{b0n} \rangle \mid f(C_{0n}) \end{bmatrix} , \tag{6}
$$

$C_{01}, C_{02}, ..., C_{0n}$ are the different subject indexes for N_0, and $V_{01}, V_{02}, ..., V_{0n}$ are the classical domains for indexes. $v_{i1}, v_{i2}, ..., v_{in}$ are the corresponding range of $C_{01}, C_{02}, ..., C_{0n}$, namely the joint domain of N_0, which satisfies $V_{oj} = \langle v_{a0j}, v_{b0j} \rangle = (j = 1, 2, ..., n)$. $f(C_{01}), f(C_{02}), ..., f(C_{0n})$ are the matching functions of the subject indexes $C_{01}, C_{02}, ..., C_{0n}$. Apparently, $V_{ij} \subseteq V_{oj} (i = 1, 2, ..., m, j = 1, 2, ..., n)$.

3.2 THE BUILDING OF EXTENSION-RELATED FUNCTION BASED ON EXTENSION DISTANCE IN DESIGN CASE SPACE

Assume the corresponding the design objective is G. Building an elementary model for G, as follows:

$$
\boldsymbol{J}_G = \begin{bmatrix} N_G & C_{g1} & V_{g1} \\ & C_{g2} & V_{g2} \\ & \vdots & \vdots \\ & C_{gn} & V_{gn} \end{bmatrix} . \tag{7}
$$

If the magnitude V_{gj} of design case subject index C_{gj}

concerning the design object J_G is a point value, namely $V_{gj} = v_{gj}$, then the extension distance between C_{gj} of the design object J_G and the No. i design case subspace J_{Si} is:

$$
\rho(V_{gj}, V_{ij}) = \left| v_{gj} - \frac{v_{aij} + v_{bij}}{2} \right| - \frac{1}{2}(v_{bij} - v_{aij}) . \tag{8}
$$

Correspondingly, the extension distance between the design case subject index C_{gj} of the design object J_G and the design case space J_P is:

$$
\rho(V_{gj}, V_{0j}) = \left| v_{gj} - \frac{v_{a0j} + v_{b0j}}{2} \right| - \frac{1}{2}(v_{b0j} - v_{a0j}) . \tag{9}
$$

If the magnitude V_{gj} of the design case subject index C_{gj} of the design object J_G is a fussy value, namely $V_{gi} = \langle v_{agj}, v_{bgj} \rangle$, the extension distance between the design case subject index C_{gj} of the design object J_G and the No. i design case subspace J_{Si} is:

$$
\rho(V_{gj}, V_{ij}) = \frac{1}{2} \left[\rho(V_{agj}, V_{ij}) + \rho(V_{bgj}, V_{ij}) \right]
$$
$$
= \frac{1}{2} \left[\left(\left| v_{agj} - \frac{v_{aij} + v_{bij}}{2} \right| - \frac{1}{2}(v_{bij} - v_{aij}) \right) + \left(\left| v_{bgj} - \frac{v_{aij} + v_{bij}}{2} \right| - \frac{1}{2}(v_{bij} - v_{aij}) \right) \right] . \tag{10}
$$

170

Correspondingly, the extension distance between the design case subject index C_{gj} of the design object J_G and

the design case space J_P is:

$$\rho\left(V_{gj},V_{0j}\right)=\frac{1}{2}\left[\rho\left(V_{agj},V_{0j}\right)+\rho\left(V_{bgj},V_{0j}\right)\right]$$

$$=\frac{1}{2}\left[\left(\left|v_{agj}-\frac{v_{a0j}+v_{b0j}}{2}\right|-\frac{1}{2}\left(v_{b0j}-v_{a0j}\right)\right)+\left(\left|v_{bgj}-\frac{v_{a0j}+v_{b0j}}{2}\right|-\frac{1}{2}\left(v_{b0j}-v_{a0j}\right)\right)\right] \qquad (11)$$

Therefore, the extension correlation function value $K_i\left(V_{gj}\right)$ between the design case subject index C_{gj} of the

design object J_G and the No.i design case space J_{Si} is:

$$K_i\left(V_{gj}\right)=\begin{cases}\dfrac{-\rho\left(V_{gj},V_{ij}\right)}{\left|V_{ij}\right|} & V_{gj}\in V_{ij}\\[4mm]\dfrac{\rho\left(V_{gj},V_{ij}\right)}{\rho\left(V_{gj},V_{0j}\right)-\rho\left(V_{gj},V_{ij}\right)} & V_{gj}\notin V_{ij}\end{cases} \qquad (12)$$

When considering the matching function $f\left(C_{ij}\right)$ of different design case subject index, the extension-related level $K_i\left(G\right)$ between all the design case subject indexes of the design object J_G and the No. i design case subspace J_{Si} is:

$$K_i\left(G\right)=\sum_{j=1}^{n}\left(f(C_{ij})*K_i\left(V_{gj}\right)\right) \qquad (13)$$

3.3 THE REUSE MODEL AND CALCULATION OF DESIGN CASE BASED ON EXTENSION THEORY

With the extension correlations between design object J_G and all design case subspaces, we can get an extension correlation array $K(G)=\left(K_1(G),K_2(G),...,K_n(G)\right)$. Therefore, the design case subspace of the best reusing object should satisfy:

$$K_0=MAX\left(K_1\left(G\right),K_2\left(G\right),\cdots,K_n\left(G\right)\right) \qquad (14)$$

We can get the reusing object from the design case subspace of the best reusing object.

Meanwhile, the design case subspace of the reusing object should satisfy:

$$K^*=\left(K_i(G)\,|\,\left(K_1(G),K_2(G),\cdots,K_s(G)\right)\wedge K_i(G)\geq 0\right),\quad 1\leq s\leq n \qquad (15)$$

Based on these design case subspaces, the transformation design or improvement design of the design objectives can be carried out. Due to word limit, this article will focus on the acquisition research of the best design case subspace. The analysis of the transformation design or improvement design of the design objectives based on design case subspaces will be explained in another article.

Thus, when dealing with complex product design with various information, categories and attributes, we can build a design case subspace and design case subject index and carry out elementary description of it. Then based on the design case subject index elementary, we can carry out matching research for the design subspace in the design case space in order to acquire the design

case subspace belonging to the best reusing design case. This can provide support for the realization of rapid design of complex products. Graph 1 shows the flow chart of the product design case reuse model based on the extension theory.
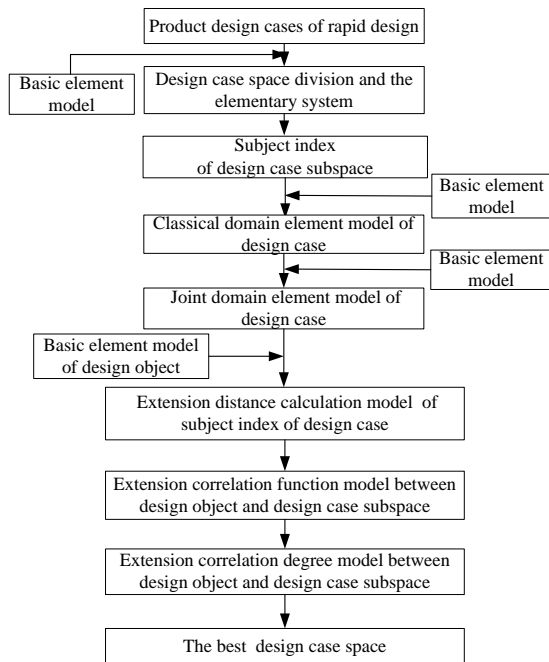
FIGURE 1 The product design case reuse model based on the extension theory

To conclude, the realization of the product design case reuse model based on the extension theory shown in graph 1 can be calculated as follows:

Step1 Divide the design case subspace based on the data about the product rapid design case. Meanwhile, build a design case subspace elementary system based on definition 1.

Step 2 Build corresponding design case subject index for the established design case subspace and also build a design case subject index elementary model based on definition 2.

Step 3 Build corresponding classical domain elementary model and joint domain elementary model for the established design case subspace and subject index elementary.

Step 4 If the magnitude of the design case subject index concerning the design object is a point value, calculate the extension distances between the different design case subject indexes of the design object and each design case subspace, according to formula (8) and (9) respectively.

Step 5 If the magnitude of the design case subject index concerning the design object is a fuzzy value, calculate the extension distances between the different design case subject indexes of the design object and each design case subspace, according to formula (10) and (11) respectively.

Step 6 Calculate the extension correlation function values between the design case subject indexes of the design object and each design case subspace according to formula (12). Calculate the comprehensive extension correlation levels between design case subject indexes of design object and each design case subspace according to formula (13).

Step 7 According to the acquired comprehensive extension correlation order, acquire the design case subspace of the best reusing object with formula (14) and the design case subspace with the reusing object according to formula (15).

Step 8 Carry out matching for best reusing case within the design case subspace of the best reusing objects and reuse. Acquire reusing objects within the design case subspace of the reusing objects and provide support for product design.

4 Applications

I will use the design of a power wheel, a core component of a certain power machine to testify and illustrate the abovementioned points. According to the suggestions and experience of design experts and the installation form of the power wheel, the design case subspace of it can be divided into 3 design spaces, namely the vertical power wheel design space, horizontal power wheel design space and side power wheel design space. Meanwhile taking into consideration the performance parameters, design structure attributes and other design parameters of the power wheel, establish the design case subject index attributes as power range, efficiency range, carrying capacity, diameter range and weight range. Thus, we can build the classical domain elementary model and joint domain elementary model corresponding the design case subspaces. The classical domain elementary model of the vertical power wheel design space J_Z, of the horizontal power wheel design space J_W and of the side power wheel design space $J_{Z\text{-}W}$ are shown respectively as follows:

$$J_Z = \begin{bmatrix} N_Z & \text{power (kW)} & \langle 15.00, 20.00 \rangle | 0.25 \\ & \text{efficiency (\%)} & \langle 85.00, 90.00 \rangle | 0.25 \\ & \text{bearing capacity (kN)} & \langle 0.00, 18.00 \rangle | 0.25 \\ & \text{diameter (m)} & \langle 0.00, 1.00 \rangle | 0.15 \\ & \text{weight (kg)} & \langle 0.00, 500.00 \rangle | 0.10 \end{bmatrix}, \quad (16)$$

$$J_{Z\text{-}W} = \begin{bmatrix} N_{Z\text{-}W} & \text{power (kW)} & \langle 20.00, 35.00 \rangle | 0.25 \\ & \text{efficiency (\%)} & \langle 90.00, 92.00 \rangle | 0.25 \\ & \text{bearing capacity (kN)} & \langle 18.00, 30.00 \rangle | 0.25 \\ & \text{diameter (m)} & \langle 1.00, 1.60 \rangle | 0.15 \\ & \text{weight (kg)} & \langle 500.00, 1000.00 \rangle | 0.10 \end{bmatrix}, \quad (17)$$

$$J_W = \begin{bmatrix} N_W & \text{power (kW)} & \langle 35.00, 46.00 \rangle | 0.25 \\ & \text{efficiency (\%)} & \langle 90.00, 92.00 \rangle | 0.25 \\ & \text{bearing capacity (kN)} & \langle 30.00, 42.00 \rangle | 0.25 \\ & \text{diameter (m)} & \langle 1.60, 2.20 \rangle | 0.15 \\ & \text{weight (kg)} & \langle 1000.00, 1800.00 \rangle | 0.10 \end{bmatrix}. \quad (18)$$

According to the classical domain elementary models of the established design case subspaces, the corresponding joint elementary model J_P is expressed as:

$$J_P = \begin{bmatrix} N_P & \text{power (kW)} & \langle 15.00, 46.00 \rangle | 0.25 \\ & \text{efficiency (\%)} & \langle 85.00, 92.00 \rangle | 0.25 \\ & \text{bearing capacity (kN)} & \langle 0.00, 42.00 \rangle | 0.25 \\ & \text{diameter (m)} & \langle 0.00, 2.20 \rangle | 0.15 \\ & \text{weight (kg)} & \langle 0.00, 1800.00 \rangle | 0.10 \end{bmatrix}. \quad (19)$$

Convert it to the design project objective elementary J_G according to the design parameters as follows:

$$J_G = \begin{bmatrix} N_G & \text{power（kW）} & \langle 19.00, 24.00 \rangle \\ & \text{efficiency（%）} & \langle 88.00, 90.00 \rangle \\ & \text{bearing capacity（kN）} & \langle 26.00, 32.00 \rangle \\ & \text{diameter（m）} & \langle 1.50, 1.80 \rangle \\ & \text{weight（kg）} & \langle 900.00, 1200.00 \rangle \end{bmatrix} \quad . \quad (20)$$

With the extension distance calculation formulas (8) and (10), acquire the extension distance sequence $\rho(J_G \to J_z)$, $\rho(J_G \to J_{z-w})$, $\rho(J_G \to J_w)$ between the design object objective elementary J_G and the classical domain elementary J_Z, J_{Z-W} and J_W of the design case subspaces which satisfy the following:

$$\rho(J_G \to J_Z) = \left(\rho(V_{g1}, V_{11}), \rho(V_{g2}, V_{12}), \rho(V_{g3}, V_{13}), \rho(V_{g4}, V_{14}), \rho(V_{g5}, V_{15}) \right) \\ = (3.00, -2.00, 22.00, 1.30, 1100.00) \quad , \quad (20)$$

$$\rho(J_G \to J_{Z-W}) = \left(\rho(V_{g1}, V_{21}), \rho(V_{g2}, V_{22}), \rho(V_{g3}, V_{23}), \rho(V_{g4}, V_{24}), \rho(V_{g5}, V_{25}) \right) \\ = (7.00, 2.00, -2.00, 0.10, 100.00) \quad , \quad (21)$$

$$\rho(J_G \to J_W) = \left(\rho(V_{g1}, V_{01}), \rho(V_{g2}, V_{02}), \rho(V_{g3}, V_{03}), \rho(V_{g4}, V_{04}), \rho(V_{g5}, V_{05}) \right) \\ = (27.00, 2.00, 2.00, -0.10, -100.00) \quad . \quad (22)$$

With the extension distance calculation formulas (9) and (11), acquire the extension distance sequence $\rho(J_G \to J_P)$ between the design object elementary J_G and the joint domain elementary J_P of the design case subspaces.

$$\rho(J_G \to J_P) = \left(\rho(V_{g1}, V_{01}), \rho(V_{g2}, V_{02}), \rho(V_{g3}, V_{03}), \rho(V_{g4}, V_{04}), \rho(V_{g5}, V_{05}) \right) \\ = (-13.00, -5.00, -26.00, -1.10, -1500.00) \quad . \quad (23)$$

With the extension distance calculation formula (12), acquire the extension correlation function value sequences $K(J_G \to J_z)$, $K(J_G \to J_{z-w})$ and $K(J_G \to J_w)$ between the design object objective elementary J_G and the classical domain elementary J_Z, J_W and J_{Z-W} of design case subspaces, as follows:

$$K(J_G \to J_Z) = \left(K_1(V_{g1}), K_1(V_{g2}), K_1(V_{g3}) \right) = (-0.1875, 0.4000, -0.4583, -0.5417, -0.4231) \quad , \quad 24$$

$$K(J_G \to J_{Z-W}) = \left(K_2(V_{g1}), K_2(V_{g2}), K_2(V_{g3}) \right) = (-0.3500, -0.2587, 0.1667, -0.0833, -0.0625) \quad , \quad 25$$

$$K(J_G \to J_W) = \left(K_3(V_{g1}), K_3(V_{g2}), K_3(V_{g3}) \right) = (-0.6750, -0.2857, -0.0714, 0.1000, 0.0714) \quad . \quad 26$$

According to formula (13) when introducing matching function for the design case subject indexes, the extension correlation sequence $K(G)$ between design object objective elementary J_G and design case subspace should be:

$$K(G) = (-0.1850, -0.1360, -0.2359) \quad . \quad (27)$$

According to formulas (14) and (15), the best design case subspace is J_{Z-W}. Therefore, we can make a choice of the best design case in this design case subspace and apply this design case in the development of new products.

5 Conclusion

This article proposes a design case model for complex products based on extension theory. This model clarifies the layers of the design for complex products by building design case subspace and design case subspace subject index for design product and making formalized presentation of elementary model. Meanwhile, this model achieves the rapid selection of design reusing case by processing the extension distance and extension correlation function of different design case subspace and acquiring the case space attribute of the design objective elementary based on the classical domain and joint domain of the design case space. Thus, the effectiveness of the complex product case reuse is enhanced and the product design period is shortened, providing support for the implementation of product rapid design.

References

[1] Yang Yanhua, Zhu Zuping, Yao Ligang 2009 The application of case based research in the design of mechanical products *Chinese Journal of Construction Machinery* **7**(3) 312-6

[2] Wei Feng, Wang Zongyan, Wu Shufang 2010 The intelligent design platform for mechanical products based on case based research *Machinery Design and Manufacture* **11** 253-5

[3] Madhusudan Therani, Zhao J Leon, Marshall B 2004 A case-based reasoning framework for workflow model management *Data & Knowledge Engineering* **50**(1) 87-115

[4] Wu Muh-Cherng, Lo Ying-Fu, Hsu Shang-Hwa 2008 A fuzzy CBR technique for generating product ideas *Expert Systems with Applications* **34**(1) 530-40

[5] Zhao Yanwei, Su Nan, Zhang Feng, et al. 2010 Product family allocation design method based on extension case reasoning *Chinese Journal of Mechanical Engineering* **46**(15) 146-54

[6] Dou Zengfa 2007 *Hybrid expert system based on extension theory and case based research* Xi'an: Xi'an Electronic and Engineering University

[7] Wang Tichu, Bo Liangfeng, Wang Wei 2011 Multiple-level case research model of product design based on knowledge reusing *Computer Integrated Manufacturing System* **17**(3) 571-6

[8] Yang Chunyan, Cai Wen 2007 *Extension engineering* Beijing: Science Press 162-9

[9] Zhao Yanwei, Liu Haisheng, Zhang Guoxian 2003 Design case research method based on extension theory *Chinese Engineering Science* **5**(5) 63-9

[10] Wang Tichun, Huang Xiang 2012 Extension case design for complex products based on knowledge reusing *Journal of Nanjing University of Aeronautics and Astronautics* **44**(4) 548-52

[11] Qiu Dongdong, Zhang Jielin 2012 The evaluation of city bus station layout based on multiple-level extension method *Journal of Wuhan University of Technology (Transportation Science and Engineering)* **36**(3) 519-22+7

[12] Zhang Xiong, Zhai Jingchun, Zhang Zongming 2011 The application of extension theory in the evaluation of helicopter maintenance ability *System Simulation Technology* **7**(2) 163-7

[13] Zhao Y W, Zhang G X 2012 A New Integrated Design Method Based On Fuzzy Matter-Element Optimization *Journal of Materials Processing Technology* 129(1-3) 612-8

[14] Wang Tichun, Zhao Sai, Chen Bingfa 2012 Association Rule Extension Mining and Reuse in Scheme Design of Large-scale Hydraulic Turbines *Information: An International Interdisciplinary Journal* **15**(6) 2403-9

[15] Wang Meng-Hui, Tseng Yi-Feng 2011 A novel analytic method of power quality using extension genetic algorithm and wavelet transform *Expert Systems with Applications* **38**(10) 12491-6

Author

Wang Li, born in March, 1968, Heqing City, Yunnan Province, China

Current position, grades: Department of materials science and Engineering, Luoyang Institute of Science and Technology
University studies: doctoral degree at Xi'an university of Architecture and Technology in 2013.
Scientific interest: material hardness analysis and simulation, product model and toughness calculation.
Publications: Published 30 academic papers in various journals
Experience: Multi-award winning papers, presided over a number of provincial projects

Definability of concept in incomplete information systems

Shunliang Huang*

Business School, Shandong University of Technology, No.88, Gongqingtuan Rd., Zibo 255000, China

Received 1 August 2014, www.cmnt.lv

Abstract

An incomplete information table (a set) can be expressed as a family of complete information tables (sets). The family of complete information sets maybe constructs an interval set. A concept in incomplete information situation, called partially known concept, is said to be definable if its extension can be expressed as an interval set. The new definition of definability proposed in this paper, named interval definable, is different from its usual meaning in the rough sets theory where a concept is definable means that its extension is a definable set, which is the union of some equivalence classes. The new definition of definability not only provides a new interpretation of interval sets, but also endows more general meaning and deeper understanding of definability.

Keywords: incomplete information, definability, concept representation, interval sets, uncertainty

1 Introduction

Definability, in rough sets theory [1, 2], is an important notion. In many studies, the notion of definability is introduced in two ways through equivalence classes and approximations [3], respectively. In general, definability and approximation can be defined each other. A set is said to be definable if it is the union of some equivalence classes. Alternatively, some authors considered the definability of a set based on its approximations. A set is said to be definable if its lower and upper approximations are equal [2, 4-7]. The two definitions of definability are equivalent. Yao [3] has investigated explicitly the notion of definability in the framework of rough sets theory and given a new interpretation. In the above mentioned studies, for an incomplete information table, the subsets of the incomplete information table are usually indefinable. In order to describe these subsets, many new theories are introduced such as rough sets theory, interval sets theory [8-10], etc. Interval sets, as a new set theory introduced, is used to describe partially known concept. In practice, partially known concept can be described by incomplete information table. By Lipski's model [11, 12], an incomplete information table can be expressed as a family of complete tables. This means that a subset of an incomplete information table can be expressed as a family of sets, which are the subsets of those complete tables, respectively. A family of sets maybe constructs an interval set. From this point of view, the notion of definability can be defined in a different way, namely, if an incomplete information table can be expressed as an interval set, then the information table is definable. As a more primitive notion, this paper not only provides a different viewpoint with the interpretation of interval sets, but also endows more general meaning and deeper understanding the notion of definability.

The rest of this paper is organized as follows. In section 2, definability of concept in complete information is examined. The notion of meaning triangle [13] is introduced to interpret the relation between the intension and the extension of a concept. A logic language, which is used by Pawlak [2], is adopted for the representation of the intension of a concept. In section 3, a new definition of definability is given. A sub-language of the above mentioned logic language is defined to describe the incomplete information. Based on the logic language, a constructive method of interval sets is introduced, and it is proved that an incomplete information table can be expressed as an interval set, namely, is definable in sense of new definition of definability. Finally, some concluding remarks are given.

2 Definability of concept in complete information

In the classical view, every concept is understood as a unit of thought that consists of two parts, namely, the intension and the extension of the concept [14, 15]. The intension of a concept specifies the necessary and sufficient conditions for a thing being a member of a specific set. The extension of a concept is a list naming every object that is a member of a specific set. The name, intension, and extension of a concept, as three vertexes, construct a triangle just like meaning triangle proposed by Ogden and Richards [13]. The classical view of concepts enables us to study concepts in a logic setting in terms of intension and also in a set-theoretic setting in terms of extension [16]. In a logic language, the intension of a concept can be denoted by a logic formula, and the extension is a set corresponding with the logic formula. By logic language, we can describe an information table and study the representation, interpretation and processing of information.

Corresponding author e-mail: huangshunliang@163.com

Definition 1 Let U be a finite and nonempty set, called the universe, an information table T can be expressed as a quadruple:

$$T = (U, AT, \{V_a \mid a \in AT\}, \{f_a \mid a \in AT\}),\tag{1}$$

where AT is a finite set of attributes, V_a is a nonempty set of values for an attribute $a \in AT$, and $f_a : U \to V_a$ is an information or a description function. It is assumed that the mapping f_a is single-valued. In this case, an information table is called complete information table.

One can define a logic language to express the intension of a concept as a equation of the language. We adopt the decision logic language L used and studied by Pawlak [2]. Equations of L are recursively constructed based on a set of atomic equations corresponding to some basic concepts.

(i) $a = v$ is an atomic equation, where $a \in AT$, $v \in V_a$.

(ii) If p and q are equations in L, respectively, then $\neg p, p \wedge q, p \vee q$ are also equations in L.

Given an information table $T = (U, AT, V_a, f_a)$, $f_a(x) = v$, written as $p = (a = v)$, is an equation. If object $x \in U$ satisfies p, then it can be denoted by $x \models p$. For two equations p and q, the semantics of the logic language can be defined as follows:

(i) $x \models p$, iif $f_a(x) = v$; (ii) $x \models p \wedge q$, iff $x \models p$ and $x \models q$; (iii) $x \models p \vee q$, iff $x \models p$ or $x \models q$; (iv) $x \models \neg p$, iff not $x \models p$; (v) $x \models p \to q$, iff $x \models \neg p$ and $x \models q$; (vi) $x \models p \leftrightarrow q$, iff $x \models p \to q$ and $x \models q \to p$.

Definition 2: let p be an equation of L. With respect to an information table $T = (U, AT, V_a, f_a)$, the set of all the objects that satisfy the equation p, denoted by $m(p)$, is called meaning set of the logic equation.

The meaning set of an equation is a subset of U, and can be denoted as follows:

$$m(p) = \{x \in U \mid x \models p\}.\tag{2}$$

The meaning set $m(p)$ includes all objects expressed by the formula p. From the point of view of concept, p represents the intension of a concept, and $m(p)$ represents the extension of a concept. Thus, a pair $(p, m(p))$ represents a concept. Obviously, the following proposition holds [16].

Proposition 1: Let p and q be two atomic formulas of the logic language L in a complete information table, the logic connectives and set-theoretic operators can be expressed by each other as follows:

$$(i) m(\neg p) = U - m(p);$$
$$(ii) m(p \wedge q) = m(p) \bigcap m(q);$$
$$(iii) m(p \vee q) = m(p) \bigcup m(q);\tag{3}$$
$$(iv) m(p \to q) = m(\neg p) \bigcup m(q);$$
$$(v) m(p \leftrightarrow q) = m(p \to q) \bigcap m(q \to p).$$

With the introduction of language L, we can discuss the definability of sets with a formal description of concepts.

Definition 3 Let $T = (U, AT, V_a, f_a)$ be an information table. A subset of $X \subseteq U$, representing the extension of a concept, is called a definable set if and only if there is a formula p of L such that:

$$X = m(p).\tag{4}$$

Otherwise, it is indefinable.

A definable set is something that can be described precisely by using the properties of definable set.

Definition 4 Let $T = (U, AT, V_a, f_a)$ be an information table. 2^U is the power set of U. p is a formula of L. The set, consisted of all definable sets, is called the systems of definable sets. It can be denoted by:

$$DEF(U, L) = \{X \in 2^U \mid X = m(p), p \in L\}.\tag{5}$$

A concept, the pair $(p, m(p))$ in an information table T, where $p \in L$, is definable if its extension $m(p)$ is definable. Table 1 is an example of a complete information table.

TABLE 1 An complete information table

	a_1	a_2
x_1	0	0
x_2	0	0
x_3	1	1
x_4	0	1

The example of logic equations and their meaning sets can be computed as follows:

$$m(a_1 = 0) = \{x_1, x_2, x_4\},$$

$$m(a_2 = 0) = \{x_1, x_2\};$$

$$m((a_1 = 0) \wedge (a_2 = 0)) = m(a_1 = 0) \bigcap m(a_2 = 0) = \{x_1, x_2\}.$$

3 Definability of concept in incomplete information

3.1 INCOMPLETE INFORMATION TABLE

For various reasons, the information contained in an information table is usually incomplete. There are many situations about incomplete information, but in our discussion, incompleteness means that instead of having a single value of an attribute, we have a subset of the attribute domain, which represents our knowledge that the actual value, though unknown, is one of the values in this subset [11]. This extends the idea of Codd's null value [17],

corresponding to the case where this subset is the whole attribute domain. Mathematically, an incomplete information table can be defined as follows.

Definition 5 Let $T = (U, AT, V_a, f_a)$ be an information table. If each object $x \in U$ is mapped into a nonempty subset of V_a such that $\varnothing \neq F_a(x) \subseteq V_a$, where $F_a : U \to 2^{V_a} - \{\varnothing\}$, then the information table is called incomplete information table.

An example of incomplete information table is given as follows:

TABLE 2 An incomplete information table

	a_1	a_2
x_1	{0}	{0}
x_2	{0}	{1}
x_3	{0,1}	{1}
x_4	{1,2}	{1}

3.2 REPRESENTATION OF INCOMPLETE INFORMATION

For an incomplete information table, Lipski proposed a reasonable semantic interpretation that an incomplete information table can be represented by multiple complete information tables, namely, an incomplete information table is equivalent to a family of complete information tables.

Definition 6: Let $T' = (U, AT, \{V_a \mid a \in AT\}, \{f'_a \mid a \in AT\})$ be a complete information table and $T = (U, AT, \{V_a \mid a \in AT\}, \{F_a \mid a \in AT\})$, be an incomplete information table. For each $x \in U$, $a \in AT$, if $f'_a(x)$ has single element and $f'_a(x) \subseteq F_a(x)$, then we say that T_1 is a completion of T. The set, consisted of all the completions of an incomplete information table T, is equivalent to the information table T, written as:

$$CT(T) = \{T' \mid T' \text{ is a completion of } T\}. \tag{6}$$

All the completions of Table 2 are listed in Table 3.

TABLE 3 Completions of incomplete information TABLE 2

	T1			**T2**			**T3**			**T4**	
	a_1	a_2		a_1	a_2		a_1	a_2	a_2	a_1	a_2
x_1	0	0	x_1	0	0	x_1	0	0	x_1	0	0
x_2	0	0	x_2	0	0	x_2	0	0	x_2	0	0
x_3	0	0	x_3	0	1	x_3	1	1	x_3	1	1
x_4	1	1	x_4	2	0	x_4	1	1	x_4	2	1

According to Lipski's model, we can solve many problems about incomplete information table by solving those problems in the family of complete information tables.

3.3 DEFINABILITY OF CONCEPT IN INCOMPLETE INFORMATION

Given an equation p of L, there is a semantic interpretation about p in each completion of an incomplete information table. In the family of semantic interpretations about p, there are two sets can well define the semantic of the given formula which is called lower bound and upper bound, respectively.

$$\|p\|_* = \{x \in U \mid \forall T' \in CT(T), x \in \|p\|^{T'}\} = \bigcap_{T' \in CT(T)} \|p\|^{T'}$$

$$\|p\|^* = \{x \in U \mid \exists T' \in CT(T), x \in \|p\|^{T'}\} = \bigcap_{T' \in CT(T)} \|p\|^{T'}, \tag{7}$$

where $\|p\|^{T'}$ denotes the set of semantic of the formula p in complete information table T'. Apparently, for arbitrary formula p of L, $\|p\|_* \subseteq \|p\|^*$ always holds. According to the definition of interval sets [8], it can be proved that $\|p\|_*, \|p\|^*$ construct an interval set, written as:

$$[\|p\|_*, \|p\|^*] = \{A \subseteq U \mid \|p\|_* \subseteq A \subseteq \|p\|^*\}. \tag{8}$$

Formally, the semantic interval set of an atomic formula $a = v$ of L in an incomplete information table can be computed as follows:

$$\|a = v\|_* = \{x \in U \mid F_a(x) = \{v\}\};$$
$$\|a = v\|^* = \{x \in U \mid v \in F_a(x)\} \tag{9}$$

For complicate logic formula, semantic interval sets have the properties as follows:

Theorem 1 Let p and q be two atomic formulas of the logic language L in an incomplete information table, the following equations hold.

$$\|p \wedge q\|_* = \|p\|_* \cap \|q\|_*;$$
$$\|p \wedge q\|^* \subseteq \|p\|^* \cap \|q\|^*;$$
$$\|p \vee q\|_* \supseteq \|p\|_* \cup \|q\|_*; \tag{10}$$
$$\|p \vee q\|^* = \|p\|^* \cup \|q\|^*.$$

Proof: we only prove $\|p \wedge q\|_* = \|p\|_* \cap \|q\|_*$. Suppose $p = (a_1 = v_1)$ and $q = (a_2 = v_2)$. First, we prove $\|p \wedge q\|_* \subseteq \|p\|_* \cap \|q\|_*$. Suppose $x \in \|p \wedge q\|_*$, this means x satisfies formulas p and q at the same time. According to Equation (9), $x \in \{x \in U \mid F_{a_1}(x) = \{v_1\}\}$ and $x \in \{x \in U \mid F_{a_2}(x) = \{v_2\}\}$ hold. Thus we have $x \in \|p\|_*$ and $x \in \|q\|_*$, namely $x \in \|p\|_* \cap \|q\|_*$. Therefore, $\|p \wedge q\|_* \subseteq \|p\|_* \cap \|q\|_*$. Secondly, we prove $\|p \wedge q\|_* \supseteq \|p\|_* \cap \|q\|_*$. Suppose $x \in \|p\|_* \cap \|q\|_*$, namely $x \in \|p\|_*$ and $x \in \|q\|_*$. According to Equation (9), we have $x \in \{x \in U \mid \quad F_{a_1}(x) = \{v_1\}\}$ and $x \in \{x \in U \mid F_{a_2}(x) = \{v_2\}\}$. This means $x \in \{x \in U \mid F_{a_1}(x) = \{v_1\} \wedge F_{a_2}(x) = \{v_2\}\}$, i.e.,

$x \in \|p \wedge q\|_*$. Then we have $\|p \wedge q\|_* \supseteq \|p\|_* \cap \|q\|_*$. Therefore, $\|p \wedge q\|_* = \|p\|_* \cap \|q\|_*$.

The properties above mentioned show that the operations of semantic interval sets may be used to interpret the logic operations. But the properties given in Theorem 1 cannot work directly for the task because there is inequality sign (inclusion). The existence of equality is due to that the atomic formulas involved in the complicated formula have the same attributes. For example, if $p = (a_1 = 0)$, $q = (a_1 = 1)$, $p \vee q = ((a_1 = 0) \vee (a_1 = 1))$ then in Table 2, the semantic set of these formulas can be computed as follows: $\|p\|_* = \|a_1 = 0\|_* = \{x_1, x_2\}$, $\|q\|_* = \|a_1 = 1\|_* = \varnothing$,

$\|p \vee q\|_* = \|a_1 = 0 \vee a_1 = 1\|_* = \{x_1, x_2, x_3\}$. So, we can get $\|p \vee q\|_* \supseteq \|p\|_* \cup \|q\|_*$, where the equation $\|p\|_* \cup \|q\|_*$ can be interpreted as "known to be $a_1 = 0$ or known to be $a_1 = 1$", but the equation $\|p \vee q\|_*$ is interpreted as "known to be $a_1 = 0$ or 1". If we assume that p and q have no common attributes, then the formulas of Theorem 1 would be all equations. Therefore, we can define a type of logic language L_0, which is a subset of L, investigated by Lipski [11], as the description language of incomplete information.

(i) $a = v$ is an atomic equation, where $a \in AT$, $v \in V_a$.

(ii) If p and q are equations in L_0, respectively, so as $p \wedge q, p \vee q$ are also equations in L_0.

Definition 7: Let $T = (U, AT, \{V_a \mid a \in AT\}, \{F_a \mid a \in AT\})$ be an incomplete information table, p be a formula of L_0. For $X \subseteq U$, we say that X is interval-definable if there is an interval set $M(p)$ such that:

$$X \propto M(p),\qquad(11)$$

where $M(p) = [m_*(p), m^*(p)]$, $\|p\|_* = m_*(p), \|p\|^* = m^*(p)$.

A concept in incomplete information is interval-definable if its extension is an interval-definable set, written as $(p, M(p))$. p is the description formula corresponding with the interval set $M(p)$. For a concept $(p, M(p))$, if $\|p\|_* = m_*(p) = \|p\|^* = m^*(p)$, then the concept is a precise concept, and the definability of the concept is same as usual definition. It means that the information about this concept is complete. All the interval-definable concepts construct a family, which is a family of interval-definable sets in an incomplete information table, or the universe U . We can give the following definition to describe this sets family.

Definition 8: Let $T = (U, AT, \{V_a \mid a \in AT\}, \{F_a \mid a \in AT\})$ be an incomplete information table. 2^U is the power set of U . p is a formula of L_0. The set, consisted of all interval-definable sets, is called the interval-definable concept system. It can be denoted by:

$$IDEF(U, L_0) = \{X \in 2^U \mid X \propto M(p), p \in L_0\}.\qquad(12)$$

Two concepts included in the interval-definable concept system, which are inequality, include different information, or they have different uncertainty. In order to study the properties of interval-definable concepts, the knowledge ordering [9] is introduced to describe the difference of two concepts.

3.4 PROPERTIES OF INTERVAL-DEINABLE CONCEPTS

Let A and B be two interval-definable concepts in the system $IDEF(U, L_0)$. Assume that $[\varnothing, \varnothing] \in IDEF(U, L_0)$. Suppose that $A = [A_l, A_u]$ and $B = [B_l, B_u]$. We have more knowledge about the concept A than the concept B, as we are more sure about the former than the latter, if $A \subseteq B$. This suggests that the standard set inclusion provides a knowledge ordering on interval sets.

Definition 9 A knowledge ordering $\preceq_k$ [9] on interval sets can be defined by:

$$[B_l, B_u] \preceq_k [A_l, A_u] \Leftrightarrow [A_l, A_u] \subseteq [B_l, B_u] \Leftrightarrow$$
$$B_l \subseteq A_l \subseteq A_u \subseteq B_u.\qquad(13)$$

In some sense, the knowledge ordering reflects the fact that A is tighter than B. Generally, we use the standard set inclusion $\subseteq$ as a knowledge ordering.

The set intersection of two interval sets is an interval set, namely, for $A = [A_l, A_u]$ and $B = [B_l, B_u]$,

$$A \cap B = \begin{cases} [A_l \cup B_l, A_u \cap B_u], when\ A_l \cup B_l \subseteq A_u \cap B_u, \\ [\varnothing, \varnothing] \qquad, otherwise. \end{cases}\qquad(14)$$

The following properties hold for the relation $\subseteq$ on interval-definable concept system:

Theorem 2 For $A, B, C, D \in IDEF(U, L_0)$, the following equations hold.

$$A \subseteq B \Leftrightarrow A \cap B = A; A \subseteq B \Leftrightarrow A \cup B = B,\qquad(15)$$

$$A \subseteq B\ and\ C \subseteq D \Rightarrow A \cap C \subseteq B \cap D,\qquad(16)$$

$$A \cap B \subseteq A, A \cap B \subseteq B.\qquad(17)$$

Proof We only prove $A \subseteq B \Leftrightarrow A \cap B = A$. Suppose $A = [A_l, A_u]$ and $B = [B_l, B_u]$. According to the definition of knowledge ordering, $[A_l, A_u] \subseteq [B_l, B_u] \Leftrightarrow B_l \subseteq A_l \subseteq A_u \subseteq B_u$. On the other hand, $A \cap B = A \Leftrightarrow [A_l \cup B_l, A_u \cap B_u] = [A_l, A_u]$. This means that

$A_l \cup B_l = A_l$, $A_u \cap B_u = A_u$. In the sense of general set-theoretic operations, this is equivalent to $B_l \subseteq A_l, A_u \subseteq B_u$, thus we have $B_l \subseteq A_l \subseteq A_u \subseteq B_u$. Therefore, $A \subseteq B \Leftrightarrow A \cup B = A$ holds.

The relation $\subseteq$ on $IDEF(U, L_0)$ is a reflexive and transitive relation. It is an ordering relation for defining the semi-lattice operation $\cap$. With respect to the knowledge ordering $\subseteq$, we can say that the interval-definable concept system $IDEF(U, L_0)$ is a semi-lattice $IDEF(U, L_0), \cap$, or $(IDEF(U, L_0), \subseteq)$.

4 Concluding remarks

The notion of definability in complete information table is examined. For studying the incomplete information table, a new definition of definability, named interval-definable, is proposed. The foundation of this notion is that an incomplete information table can be represented by an interval set and expressed as a family of complete information tables. Based on the definition of interval-definable, the interval-definable concept system is defined. By introducing a knowledge ordering, the properties of interval-definable concept system are investigated. The conclusion indicates that the interval-definable concept system is a semi-lattice with respect to the operation $\cap$, or relation $\subseteq$. By studying definability in incomplete information table and interval sets, we hope that we can gain more insights into the representation and processing of imprecise or partially known concepts, and into the approximations of indefinable or complex concepts. More development and applications of interval sets in incomplete information will be studied in the near future.

Acknowledgements

References

[1] Pawlak Z 1982 Rough sets *Int J of Computer and Inf Sci* **11**(5) 341-56

[2] Pawlak Z 1991 Rough sets-Theoretical Aspects of Reasoning about Data *Kluwer Academic Publishers:* Dordrecht

[3] Yao Y Y 2007 A note on definability and approximations *LNCS Trans. on Rough Sets VII* **4400** 274-82

[4] Buszkowski W 1998 Approximation spaces and definability for incomplete information systems *Rough Sets and Current Trends in Computing* LNAI **1424** 115-22

[5] Grzymala-Busse J W 2005 Incomplete data and generalization of indiscernibility relation, definability, and approximations *The 10th Int Conf on Rough Sets, Fuzzy sets, Data Mining, and Granular Computing* **1** Springer-Verlag: Heidelberg 244-53

[6] Järvinen J, Kortelainen J 2004 A note on definability in rough sets theory *Current Issues in Data and Knowledge Engineering, Akademicka Oficyna Wydawnicza EXIT* Warsaw 272-7

[7] Wasilewska A 1987 Definable sets in knowledge representation systems *Bulletin of the Polish Academy of Sci: Math* **35** 629-35

[8] Yao Y Y 1993 Interval-set algebra for qualitative knowledge representation *ICCI '93 Proceedings of the Fifth International Conference on Computing and Information* 370-74

[9] Yao Y Y 2009 Interval sets and interval-set algebra *Proceedings of The 8th IEEE International Conference on Cognitive Informatics* 307-14

[10] Miao D, Wang G, Yao Y, Liang J, Wu W, Zhang Y 2012 The Cloud Model and Granular Computing *Science Press*

[11] Lipski W J R 1979 On semantic issues connected with incomplete information databases *ACM Transactions on Database Systems* **4**(3) 262-96

[12] Lipski W J R 1981 On database with incomplete information *J of the ACM* **28**(1) 41-70

[13] Ogden C K, Richards I A 1946 The Meaning of Meaning: A Study of the Influence of Language Upon Thought and of the Science of Symbolism (8th ed) *Harcourt Brace Jovanovich:* New York

[14] Smith E E 1989 Concepts and induction in Posner M I (Ed.) Foundations of Cognitive Science *Cambridge MA MIT* 501-26

[15] Sowa J F 1984 Conceptual Structures: Information Processing in Mind and Machine *Addison-Wesley Longman Publishin Co Inc Boston MA USA*

[16] Yao Y Y 2009 Interpreting concept learning in cognitive informatics and granular computing *IEEE Transactions on Systems, Man, and Cybernetics Part B: Cybernetics* **39**(4) 855-66

[17] Codd E F 1975 Understanding relations (installment #7) *FDT Bulletin of ACM-SIGMOD* **7**(3-4) 23-8

Authors

Shunliang Huang, born in November, 1977, Shandong, China

Current position, grades: Doctor of Engineering, Associate professor in Shandong University of Technology.
University studies: Control Theory and Control Engineering in Shandong University.
Scientific interest: rough sets theory, data mining.
Publications: 22 papers.

Research on multi-virtual queue based on flow estimation

Wenqing Huang[1, 2*]

[1]*Department of Computer and Information Engineering, Heze University, Heze 274015, Shandong, China*

[2]*Key Laboratory of computer Information Processing, Heze University, Heze 274015, Shandong, China*

Received 1 August 2014, www.cmnt.lv

Abstract

To solve the equity issues in the network bandwidth competition of TCP/UDP mixed flow, we have raised an AQM algorithm framework based on "granularity", which is aimed to realize the divide-and-rule goal of the conservative TCP and the greedy UDP so as to alleviate the equity issues of the network resources caused by different protocols by designing a common framework. Furthermore, under the AQM framework of "granularity", we have brought forth Multi-Virtual Queue based on Flow Estimation (VFQ) and we have also given the detailed implementation schemes for different modules of VFQ algorithm. In order to further verify the feasibility and validity of VFQ algorithm, we have designed 5 computer network congestion scenarios in AQM network simulation platform and made simulation comparisons with the classical ARED, PI and Blue algorithms. Numerous simulation results demonstrate that in contrast with other AQM algorithms, VFQ can better adapt to various sudden network congestion scenarios and it has better response speed and queue management performance.

Keywords: multi-virtual queue, flow estimation, TCP/UDP mixed flow

1 Introduction

Network burstiness is the main factor which causes the instable performances of AQM algorithm. However, most of the current AQM algorithms handle the network traffic based on "packet granularity". In face of the network with the coexistence of TCP and UDP flows, these algorithms can't guarantee the equity of the bandwidth allocation. In addition, since UDP flow is an open-loop non-response flow, it ignores ECN congestion feedback signal of AQM algorithm, making it impossible for AQM algorithm to adjust the resource occupancy of UDP flow through the feedback signal; therefore, the traditional AQM fails to control the tip nodes by only ECN feedback mechanism [1].

In order to solve these problems, this paper has first designed an AQM algorithm framework based on "granularity", based on which, it has raised Multi-Virtual Queue based Flow Estimation (VFQ). What this algorithm is different from the traditional AQM algorithms include:

Ⅰ. VFQ introduces the idea of "granularity" and differentiates and maintains the traffic information passing through every bottleneck route;

Ⅱ. In accordance with the different characteristics of UDP and TCP flows, it has constructed 2 logically-isolated virtual queues (VQ) respectively and designed different congestion control strategies for these 2 VQ;

Ⅲ. Through the network load Perceive information of TCP/UDP flow, the system will adjust and control the relative sizes of these two VQ automatically.

2 Network congestion control mechanisms and active queue management (AQM)

When the demand of users for network resource (including CPU processing capability, the size of memory space and information channel transmission rate) exceeds the intrinsic network resource amount, the network will be in the state of "overloading" continuously. Under such circumstance, network congestion will occur, as shown in Equation (1). As for the computer network system, the occurrence of congestion exists objectively.

$$\sum the\ demand\ of\ all\ users\ for\ network\ resource > \atop network\ resource \quad . \quad (1)$$

2.1 ACTIVE QUEUE MANAGEMENT (AQM) MECHANISM

The source end network congestion control measure based on TCP protocol is essential and indispensable. However, there are certain limitations for such network congestion control measure. This measure is not always effective in successfully achieving congestion control for all networks [2]. Thus, it is necessary to add other network congestion control mechanisms to the central router so as to reach satisfactory effectiveness in either preventing network congestion in advance or remove network congestion after it occurs. In order to avoid the disadvantages of TCP source end network congestion control measure, IETF (Internet Engineering Task Force) firstly put forward to

---

Corresponding author email: Huangwenqing8686@126.com

Active Queue Management (AQM). AQM mechanism has the following three advantages:

1) Low loss in router node network packets: the transmission of data packets in computer network is featured with strong abruptness. When the free buffer memory space approaches to "0" or is just equal to "0", the router will fail to allocate certain buffer memory space for those abrupt network packets. When just a small part of the buffer memory space of a router is occupied, the router will have enough space to store those abrupt packets without needing to discard these packets. If AQM mechanism is not adopted, many packets will be discarded when the buffer memory space of a router is full. Such atmosphere is not good for the operation of computer network [3]. Three major reasons are: firstly, drop tail rule will make network to generate "overall synchronization", which may lead to users cannot fully utilize information channel bandwidth and meanwhile the network throughput will also decrease. Secondly, compared with the drop tail methods of simply discarding network packets, it is difficult for TCP to recover from the state of network congestion. Thirdly, meaninglessly discarding the packets is an expression of the fact transmission links cannot be effectively utilized [4].

2) Shortening the delay time of network service: if the instant queue size is made to fluctuate between small ranges. AQM mechanism will reduce the delay time of packet processing, which will make the instant data service of some internet better.

3) Removing "deadlock" behaviour: if AQM mechanism can always provide a memory space of a certain space to accept instant packet, it can avoid the occurrence of "deadlock" atmosphere.

2.2 SEVERAL TYPICAL AQM ALGORITHMS

2.2.1 ARED algorithms

The basic concept of ARED algorithm is to perceive whether RED should be aggressive or conserved through checking average changes of queue. It puts forward to an automatic regulation mechanism and set packet loss probability according to the changes of flow amount on the bottleneck link. Larger packet loss probability is suitable for more cases; vice versa. If the queue size is near *minth*, it indicates that the congestion control is too aggressive and the packet loss probability *maxp* should be reduced. The reducing degree is signified by "a", hence, *maxp=a·maxp*; If the queue size is near *maxth*, it indicates that the congestion control is too conserved and the packet loss probability *maxp* should be increased. The increasing degree is signified by "b", hence, *maxp=b·maxp*, where 0<a<1<b.

2.2.2 PI algorithm

PI algorithm is a kind of AQM algorithm based on typical control theory. This algorithm effectively improves the performance of network system and can be applicable for the changes of network flow. PI controller simply considers the network as a linear time invariant system and takes the packet loss probability as the measurement of current network congestion. The discrete version of PI control algorithm is:

$$p(k) = p(k-1) + a(q(k) - q_{ref}) - b(q(k-1) - q_{ref}), \quad (2)$$

where, *p(k)* refers to the packet loss probability at the time *k*; *a* and *b* respectively refers to the current queue size and the control parameter of the difference between the queue size at the previous moment and expected queue size; q_{ref} is the expected queue size. The algorithm control block diagram is as shown in Figure 1:
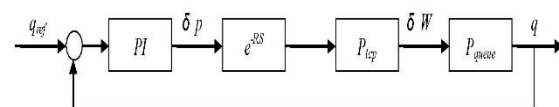
FIGURE 1 Systematic block diagram of PI algorithm

Although PI controller algorithm is simple and easy to achieve, its response time is relatively long. This disadvantage can be improved by introducing differentiation element to accelerate the response speed of the system [5].

2.2.3 Blue algorithm

Blue algorithm manages the queue by using queue overflow and the event of idle link, which is improvement on RED algorithm. It maintains a probability Pm to mark the drop probability. Pm will be increased if the event of packet loss happens caused by queue overflow in order to pick up the speed of sending congestion feedback information to the transmit end; on the contrary, Pm will be decreased if the queue is idle caused by idle link in order to slow down the speed of sending congestion feedback information to the transmit end. So the Blue algorithm can effectively control the speed of sending congestion feedback information [6].

Blue algorithm implements the congestion control according to the packet loss and link service condition directly. Compared with RED algorithm, it decreases packet loss and the demand for buffer memory space of the router, but there are still disadvantages for Blue algorithm as following [7]:

1) Blue algorithm increases the packet loss when queue overflow, and decreases the packet loss when queue is idle. But queue overflow and idle are two extreme, so the queue length increases or decreases rapidly and the amplitude of fluctuation is very large. It will result in the decrease of throughput when the queue length fluctuates heavily. Moreover, the oscillation of queue length results in greatly end-to-end delay jitter so that it is difficult to meet the requirements of QOS.

2) It lacks adaptability in practical application and cannot make relevant adjustment according to the virtual

condition of the network, which is one of the reasons for causing great fluctuation of throughput and queue length.

2.3 AQM IMPLEMENTATION SCHEME

The common AQM Implementation Scheme is to append complex AQM algorithm to the Router, shown as Figure 2. In the Figure, Sourcer means sending end, Receiver means receiving end, Qs means queue scheduling, Queue means the queue. When new data packet arrives, the AQM algorithm will be run in the Router; and then it will decide whether to drop the data packet or queue it up according to the running result of the algorithm; at last, it will be trans ponded by look up routing list according to the routing information carried at the top of rejected packets and at the same time the router will save the intermediate variable of the algorithm for the use of operation next moment [8].
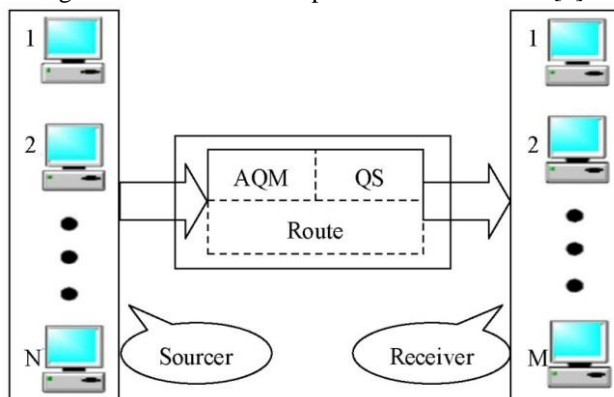
FIGURE 2 The implementation scheme of AQM algorithm

3 The design AQM algorithm and overall structure of performance detection platform

3.1 THE IMPLEMENTATION OF AQM ALGORITHM DETECTION PLATFORM

The Linux-based NS2 Platform can simulates the performance of different scale network algorithms, and the result is more close to practical effects, so this thesis designs a detection and performance evaluation platform for AQM a

Algorithm. The Figure 3 shows the AQM Mechanism Design and Detection Platform Framework based on NS2. As seen in the Figure 3, the reality of the platform can be mainly divided into several parts as follows:

1) AQM algorithm Engine Module: the AQM mechanism deployed at bottleneck router, which is executed frequently, and the execution efficiency will influence the control performance of the whole system's network congestion. The algorithm realized the functions as follows: it samples the link congestion condition periodically (based on timer module) and collects latent congestion variable in time via the congestion detection module of AQM network, and then updates packet loss dynamically via specific queue management engine; when every data packet reaches the router, if queue buffer

memory overflows, then physical packet loss will be carried out; if buffer memory is not full, then packet loss event in probability of AQM algorithm will be triggered [9].

2) Network Topology Module: the module is mainly designed to achieve the layout of network topology scene. Topology includes single-bottleneck mode and multi-bottleneck mode. Single-bottleneck adopts "dumbbell-shaped" structure. Although each packet needs to go through multiple links from the source end to the terminal, as for each linking flow, among many links only one is the bottleneck. Hence, it can simplify the network as single-bottleneck link; multi-bottleneck topology adopts "parking lot" structure. When the algorithm is detected through single-bottleneck topology, the network topology will be complicated and multi-link mode will be introduced, which will make the algorithm results are more close to real network performance. Network topology module is a simulated network macroscopic module, which is defined during the simulation initialization process. Specific network topology will be generated to define network bandwidth for each link and define the link transmission delay time [10].
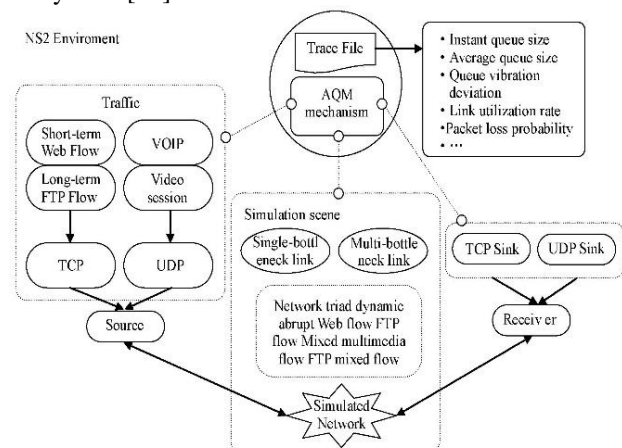
FIGURE 3 AQM algorithm detection platform based on NS2

3) Module for Abrupt Network: the module mainly manages the flow types and abrupt time. The flow types of internet are very abundant, but they are mainly consisted of long-term TCP flow (FTP), short-term TCP flow (HTTP) and response UDP flow (loading multimedia service). Network flow possesses inherent abruptness, which leads to that the network congestion is unforeseeable and difficult to control. Therefore, in order to verify the algorithmic comprehensive performance of AGM algorithm in facing complex network flows, the detection platform concentrates on detection the single abrupt case and mixed abrupt case of above flows so as to ensure that the simulation results can comprehensively reflect the performance of AQM algorithm and verify the disadvantages of this algorithm [11].

4) Module for Performance Analysis: the module mainly to achieve the function of intelligently analysing simulation results and extracting data relevant to AQM algorithm performance indicators according to simulation

results files (Trace File). Such data include instant queue size curve varying with time, average queue size, queue vibration variance, link utilization rate and packet loss probability. By analysing and comparing simulation results, it can reach relevant conclusions [12].

The following parts of the present research will develops by focusing on the specific designing of these modules.
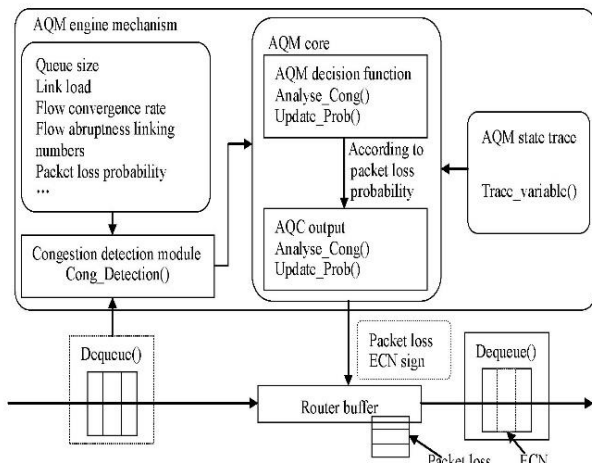
FIGURE 4 The Design of AQM Algorithm Engine Modules

3.2 THE DESIGN OF AQM ALGORITHM ENGINE MODULES

The unbalance of network flow distribution and the abruptness of data lead to the occurrence of network congestion. The AQM router placed in the scene, where congestion occurs, perceives the real-time network congestion through network flow changes and then utilizes AQM decision function to dynamically control packet loss probability. The router also makes the sending end learn the features of network congestions in advance through packet loss or ECN means and further achieve the reasonable utilization of network resource. AQM mechanism is the core module of congestion control. It needs to implement packet control for each arriving packets. Hence, it has high requirements for algorithm efficiency. As a result, here we adopt C++ language of NS2 to meet the requirement. Corresponding mapping is indicated as Queue/XXX. XXX is represented as the definition means of AQM. According to the above analysis, it can be known that a complete AQM mechanism (as shown in Figure 4) consists of five components: congestion detection module, AQM decision function module, AQM output control module, AQM state trace module and AQM engine message event management module.

1) Congestion detection module: the module mainly detects the congestion situation of current links. Different AQM mechanisms acquire different quantity of state, which mainly includes the following detection scales: instant queue size/average queue size (e.g. RED algorithm and PI algorithm), link load (e.g. Blue algorithm and Yellow algorithm) flow convergence rate (e.g. REM

algorithm and VRC algorithm), active flow abrupt linking numbers (e.g. FRED algorithm and SBF algorithm) and packet loss probability (e.g. LRED algorithm). Many algorithms adopt mixed scales to comprehensive detect the network congestion situation. Among the above scales, apart from instant queue size and link load which can be directly acquired from the links, all other scale need to make periodical sampling congestion situation to conduct a statistics about the comprehensive information of network congestion during a certain period so as to filter the transient interference of sampling information.

2) AQM Decision Module: the module is the core of AQM engine. It has two major functions: one is the function of analyzing and acquiring potential congestion cases according to network congestion parameters; another is the function of designing specific queue control function according to analysis results and adjusting the packet loss value according to network congestion state.

3) AQM Output Control Module: the functional module mainly dynamically makes congestion marking and control on the packets it goes through according to the control output (packet loss value) of AQM decision function. On the premise of the support of AQM router and TCP, when the buffer of AQM router overflows, mandatory physical packet loss will be achieved; otherwise, according to packet loss probability value, the probability will set the CEN of IP head for the passing packets as 1 so as to realize the logical packet loss. If TCP does not support ECN mechanism, the transmission and delivery of congestion feedback information will be achieved by conducting physical packet loss in probability.

4) AQM State Trace Module: the module mainly functions to make file records for certain variables during simulation process and provide original data for AQM algorithm performance analysis module. The variable values that AQM algorithm needs to trace are defined in accordance with the demands of algorithm design. Generally speaking, variable values needed to record include instant queue size, average queue size, packet loss value, link load and other customized variables.

5) AQM Engine Message Event Management Module: the module mainly consists of several timers and is used to manage the triggering of periodical events. For instance, the active flow estimation and packet loss estimation of congestion control module require periodical sampling. Queue control function periodically updates the packet loss probability function value.

3.3 VQUDP PACKET LOSS STRATEGY MODULE

Firstly, based on the results of UDP active flow perceives by APF, we will calculate and weight the convergence rate (r_{id}) of every UDP in unit time. Therefore, the total convergence rate (R_{ater}) of UDP is:

$$R_{ater} = \sum_{jd \in L} r_{id} .$$

Among it, *L* is UDP flow collection (the BFV maintained by UDP flow). Calculate the current UDP rate perceive factor (λ) through Equation (3) and *C* is the actual bandwidth of the current bottleneck link (with its unit as Mpbs).

$$\lambda = \frac{R_{ater}}{C}. \tag{3}$$

UDP is mostly used in the transmission of network multi-media audio/video and certain network packet loss won't affect people's understanding of the information. At the same time, UDP is an open-loop control flow, which means that it cannot respond to the packet loss/ECN feedback signal caused by network congestion. Therefore, the UDP flow data packet is usually abandoned in the forced manner.

VQudp packet loss strategy uses the similar packet loss mechanism to Blue algorithm, which can be seen in Equation (4). In this Equation, *kudp* is the probability packet loss gain value; $qlen_{upd}$ is the instantaneous virtual queue length of UDP and VQ_{upd} is the logic size of the current UDP virtual queue buffer.

$$p = \max\left\{\frac{qlen_{upd}}{VQ_{upd}} \cdot k_{udp}, 1\right\}. \tag{4}$$

In order to make VQudp packet loss strategy can adjust its packet loss probability in face of different UDP loads, we have divided the network congestion caused by UDP into the following scenarios via UDP rate perceive factor (λ): when $\lambda \leq 0.3$, UDP will lead to slight network congestion and when $\lambda > 0.7$, UDP will greatly deteriorate the network performance and cause network congestion. The above scenarios will adjust the packet loss gain according to the load value (λ) via Equation (5). When $\lambda > 0.3$, VQudp strategy uses the original packet loss gain (k_{upd}) and when $\lambda > 0.3$ We will adjust the packet loss gain (k_{udp}) dynamically according to the load (λ).

$$k_{upd}\begin{cases} k_{upd}\lambda \leq 0.3 \\ \frac{k_{upd}}{1-\lambda}, \lambda > 0.3 \end{cases}. \tag{5}$$

3.4 VQTCP PACKET LOSS STRATEGY MODULE

The VQtcp packet loss strategy in this paper is based on the recently-raised Blue algorithm. In this paper, we have made some alterations to Blue algorithm and come up with a VBlue based on virtual queue to adapt to the VQtcp. Since TCP is the response flow of congestion control, it can adjust and control its sending rate according to the network congestion feedback signal. Therefore, when we design VQtcp packet loss strategy, we will use packet loss/ECN strategy instead of the forced physical packet loss, which is to say, when the virtual queue logic of VQtcp overflows, forced packet loss will be used; otherwise, ECN

strategy will be used. VBlue algorithm has made the following changes to the original Blue algorithm:

1) On the basis of the original Blue algorithm, add the periodical update strategy of the size of virtual queue, namely after every period (ΔT), make BuffsizeVBlue=VQtcp via AFP algorithm;

2) Input the congestion control of the original Blue algorithm to make the instantaneous queue length (*qlen*) into (*qletcp*). Configure the target queue length of VBlue as $Q_{ref} = VQ_{tcp}/2$. VBlue algorithm takes the deviation (Δ*e*) of the instantaneous queue length (*qletcp*) and the target queue length $Q_{ref} = VQ_{tcp}/2$ as the algorithm input. Δ*e* represents the deviation degree between the instantaneous queue length and the target value.

3) In order to make the algorithm more flexible, VBlue algorithm first normalizes Δ*e* to the range of [1-11] through the normalizing factor (*Gq*) and then normalize Δ*e* to obtain the pack loss probability via Blue congestion control algorithm. The normalizing factor (*Gq*) here can be obtained from Equation (6).

$$G_q = \frac{2}{VQ_{tcp}}. \tag{6}$$

4 VFQ algorithm simulation analysis based on NS2 platform

This paper will design the following 5 congestion simulation scenarios (connection N mutations, dynamic link bandwidth C changes, dynamic round-trip time RTT, dynamic changes of non-response load λ & dynamic changes of Web loads P). To investigate the performance of these algorithms in the dynamic network status triad (N, RTT, C) and use the average queue length (*Avgq*), the queue oscillation variance (STD), link utilization (Utiliz) and packet loss rate (*LossR*) as the performance indexes.

4.1 THE ALGORITHM CHARACTERISTICS OF DIFFERENT CONNECTIONS

The continuous traffic change is the outstanding characteristics of computer network; therefore, the experiment is aimed to study the algorithm characteristics of VFQ algorithm under the ever-changing network congestion. Here, we only use TCP as the traffic of the experiments. In order to represent the dynamic changes of full-load network congestion, the active flow connections N will adopt the form of segmentation with time changes {300,600,200,700,300,500,200}, maintains every N for 50s and last the simulation for 350s.

The experimental results demonstrate that in most cases, ARED algorithm can control the instantaneous queue near the target value Q_{ref} =150pkt; however, when suffered from heavy load (N=700)0 queue oscillation is too serious to converge; the queue management performance of PI algorithm is the worst among 4 algorithms and it is impossible for it to converge qlen to the target value Q_{ref} =150pkt within 50s; the convergence

rate of Blue algorithm is relatively slow and it wobbles seriously with qlen; VFQ algorithm we come up with can capture the real-time network sudden changes through AFP active flow perceive algorithm. The actual value of VFQ active flow connections and the perceive value curve: AFP can converge the transient-state perceive value to the actual value 5s after every active flow change. Compared with other algorithms, VFQ algorithm has faster queue convergence rate and better queue control effects.

4.2 THE ALGORITHM CHARACTERISTICS OF DIFFERENT ROUND-TRIP TIME RTT

The scholars and engineers have verified the time-lag RTT parameter has a far-reaching influence on AQM algorithm from a great number of experiments. Our existing TCP/AQM system is in fact a time-lag RTT closed-loop control system with ECN/packet loss congestion. The reflected time lag is the packet RTT. Therefore, in order to verify that the robustness of VFQ algorithm in different RTT scenarios, this group of experiments have made repeated simulation to the network congestion scenarios when RTT = [60×i]ms (i =1, 2, ..., 7), N = 200 with other configuration as the default values.

ARED algorithm suffers most influence from RTT. When RTT≥240ms, its *qlen* is not stable, causing serious wobble; RTT plays little impact on *Avgq* and STD of PI and VFQ algorithms; the queue management of Blue algorithm is placed in the middle, but when RTT≥300m, its STD is also big. When RTT > 180ms, the link utilization Utiliz of ARED algorithm falls sharply, while the Utiliz performances of other various algorithm are almost the same. From the relationship between the packet loss probability *LossR* of various AQM algorithms and RTT, it is clear that the *LossR* of ARED and PI algorithms are at least one quantative grade higher than that of Blue and VFQ algorithms. To summarize, VFQ algorithm has better stability and robustness in the scenarios with RTT changes.

4.3 THE ALGORITHM CHRACTERISTICS OF DIFFERENT LINK BANDWIDTH C

The bandwidth *C* determines the available resources of the bottleneck link. In general, *C* in the wired network is usually fixed while it is instable in the wireless network or satellite link due to the environmental influence. Therefore, this part will verify the performances of various AQM algorithms in different bandwidth *C* on NS2 simulation platform in the network topology of Figure 1. Repeated simulation is made to the congestion network scenario when *C*={5,10,15,25,35,45,55} Mbps, RTT=80ms, N=300.
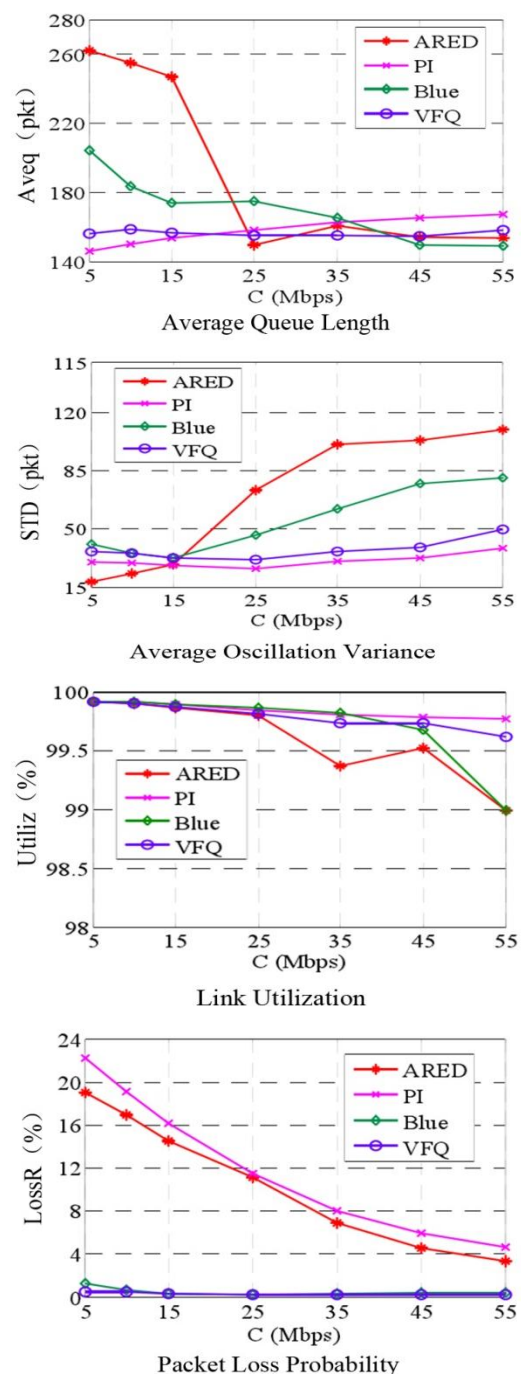
FIGURE 5 The comparisons of different algorithm performances under the changes of dynamic bandwidth C

Figure 5 describes the changes of various characteristic indexes of AQM in different network bandwidth *C*. On the whole, with the continuous changes of bandwidth *C*, the queue oscillation STD will increase gradually and the network utilization Utiliz decreases. So do *Avgq* and Loss R.

When C≤25Mbps, the *Avgq* of ARED algorithm is very big; however, with the gradual increase of the bandwidth *C*, *Avgq* gets closed to the target value *Qref* (150pkt). It is clear in Figure 5b that when *C* > 35Mbps, STD increases, meaning that ARED algorithm oscillates dramatically.

Both VFQ and PI algorithms have made impressing results in the queue management and bandwidth resource utilization; however, in Figure 5d, we can see that the *LossR* of PI algorithm and ARED algorithm are one quantative grade than that of VFQ algorithm. In one word, VFQ algorithm has better performance in the scenarios with dynamic changes in bandwidth C.

4.4 THE ALGORITHM CHARATERISTICS OF THE LOADS λ IN DIFFERENT NON-RESPONSE FLOW

If we want to verify the performance of the above-mentioned AQM algorithms in the computer network scenarios with mixed flow of TCP/UDP through NS2 simulation platform, we need to examine the UDP flow load λ to every AQM algorithm characteristic. Repeated simulation will be made to various AQM algorithms in TCP connections N = 400, λ={0.05,0.15,0.25,0.35,0.45,0.55,0.6} and RTT = 80ms.This is the relationship between UDP proportion factor λ and the performance indexes of various AQM algorithms. ARED algorithm has similar *Avgq* curve to Blue algorithm; however, STD curve of ARED algorithm wobbles little in contract with that of Blue algorithm. When λ=0.05, *Avgq* of PI algorithm is big and the queue becomes instable (*Avgq* deviates from Qref; STD increases) gradually with the increase of (λ). VFQ algorithm can stabilize *qlen* near the target area, although STD increases with the increase of UDP proportion factor λ. Λ has placed the biggest influence on Utiliz of PI algorithm, while the Utiliz of other algorithms can almost reach full-load 100% status. From the perspective of the packet loss probability, the *LossR* of ARED and PI algorithms have been over 13%, while the *LossR* of Blue and VFQ algorithms increases greatly with the increase of the load λ. The *LossR* increases 8~9 times when λ=0.05 and λ=0.6. To sum up, non-response UDP load λ has a big influence on the performance indexes of various AQM algorithms; however, compared with other algorithms, although the *LossR* of VFQ algorithm also suffers big impact, it can still maintain *qlen* in the target area Qref.

4.5 THE ALGORITHM CHARACTERISTICS OF DIFFERENT WEB LOADS P

This part verifies the above AQM algorithms performance in the long-time TCP/Web traffic mixed-flow network congestion scenario through NS2 simulation platform. It adopts PackMime model by Bell Laboratory as the generative model of Web traffic and uses HTTP/1.1 version protocol as the data communication protocol carrier of short-time Web flow, among which, *ρ* is the requests of Web traffic per second (connections). The simulation will be made to the AQM algorithms in the network scenarios when *ρ*={10,20,30,40,50,60,70}

(connection/seconds), RTT=80ms and the long-time TCP connections N=300.

The variation trends of different performance indexes of various AQM algorithms in different Web traffic loads. Although the *Avgq* of ARED and Blue algorithms keep stable with the changes of load ρ, it is clear from the STD curve that the queue wobbles (around 80pkt), meaning that ARED and Blue algorithms can't get satisfactory queue management performance in dealing with TCP/Web mixed-flow scenario. Although PI algorithm has a relatively stable STD curve, its *Avgq* is around 250pkt most of the time and it cannot converge *qlen* to near Q *ref*=150pkt. The VFQ algorithm proposed by this paper adopts AFP active flow perceive algorithm, making it better adapt to the changes of Web flow load ρ and its overall queue performances suffer little influence. Apart from PI algorithm, the Utiliz of other algorithms fall with the increase of Web load *ρ*. With the increase of Web load *ρ*, the *LossR* of various algorithms increase while the *LossR* of PI algorithm demonstrates an instable status. To sum up, the addition of Web traffic makes the network congestion scenario more complicated and causes bigger influence on the traditional AQM algorithms, which only take TCP into consideration. n the meanwhile, it can be seen from this group of simulations that, in the network congestion scenario of TCP/Web mixed flow, although various performances of VFQ algorithm fall, it receives the least influence compared with other AQM algorithm.

5 Conclusions

The present research conducts a detailed research on AQM algorithm. Based on Linux system, the present research establishes an AQM algorithm detection platform of NS2, which can simulate the performance of network algorithms with different scales. Moreover, the present research also designs an AQM algorithm engine module, on the basis of which, it raises Multi-Virtual Queue based on Flow Estimation (VFQ). Finally, this paper has also conducted plenty of simulation to VFQ algorithm, made some simulation comparisons with some classical AQM algorithms such as ARED algorithm, PI algorithm and Blue algorithm and analysed the performance of VFQ algorithm in order to verify the feasibility and validity of the new algorithm. It is demonstrated in the simulation results that in face of the dynamic load changes, VFQ algorithm has a rapid convergence rate and high-efficient congestion queue processing ability in contrast with other algorithms, suggesting that VFQ algorithm can better deal with the network congestion.

References

[1] Li F, Sun J, Zukerman M, Liu Z, Xu Q, Chan S, Chen G, Ko K-T 2014 A comparative simulation study of TCP/AQM systems for evaluating the potential of neuron-based AQM schemes *Journal of Network and Computer Applications* **41** 274-99

[2] Hwang L-C 2013 M-GREEN: An active queue management mechanism for multi-QoS classes *Computer Standards & Interfaces* **36**(1) 122-31

[3] Divakaran D M 2012 A spike-detecting AQM to deal with elephants *Computer Networks***56**(13) 3087-98

[4] Guesmi H, Djemal R 2013 Design of priority-based active queue management for a high-performance IP switch *Computers & Electrical Engineering* **39**(2) 246-60

[5] Zhang W, Tan L, Yuan C, Chen G, Ge F 2013 Internet primal–dual congestion control: Stability and applications *Control Engineering Practice* **21**(1) 87-95

[6] Zhan Z-q, Zhu J, Xu D 2012 Stability analysis in an AVQ model of Internet congestion control algorithm *The Journal of China Universities of Posts and Telecommunications* **19**(4) 22-8

[7] Avrachenkov K, Ayesta U, Doncel J, Jacko P 2013 Congestion control of TCP flows in Internet routers by means of index policy *Computer Networks* **57**(7) 3463-78

[8] Zhou H, Hu C, He L 2013 Improving the efficiency and fairness of eXplicit Control Protocol in multi-bottleneck networks *Computer Communications* **36**(10–11) 1193-208

[9] Zhang C, Cai Z, Chen W, Luo X, Yin J 2012 Flow level detection and filtering of low-rate DDoS *Computer Networks* 56(15) 3417-31

[10] Yilmaz M, Ansari N 2014 Achieving destination differentiation in ingress aggregated fairness for resilient packet rings by weighted destination based fair dropping *Computer Networks* **62** 43-54

[11] Giaccone P, Leonardi E, Neri F 2013 On the interaction between TCP-like sources and throughput-efficient scheduling policies *Performance Evaluation* **70**(4) 251-70

[12] Ju Y, Wang A, Liu X 2012 Evaluating emergency response capacity by fuzzy AHP and 2-tuple fuzzy linguistic approach *Expert Systems with Applications* **39**(8) 6972-81

Author

Wenqing Huang, born in 1981, Heze, Shandong, China

Current position, grades: an assistant of the Department of Computer and Information Engineering, Heze University.
University studies: Master of Engineering in Computer Science from the Liaoning University of Science and Technology in 2011.
Scientific interest: computer network, network security.

Information and Computer Technologies

Application of federated particle filter to SINS-GPS/BDS integrated navigation system

Yang Shi*, Huipeng Li, Xianmu Li, Lu Wang

School of Instrumentation Science and Opto-electronics Engineering, Beihang University, 100191, China

Abstract

When integrated navigation information is filtered, there may be non-linear sub-filter. The paper proposes federated particle filter. It is based on the framework of federated Kalman filter and uses the method of particle filter to process non-linear sub-filter, which enhances adaptability of federated Kalman filtering model. The paper applies federated particle filter to SINS-GPS/BDS integrated navigation system to establish filtering model. The simulation is made to verify the effectiveness of federated particle filter.

Keywords: federated Kalman filter, particle filter, integrated navigation

1 Introduction

In recent years, multi-sensor integrated navigation system with inertial navigation as core has developed rapidly. For example, based on inertial navigation system, GPS and BDS are introduced. Compared with single system, the advantage of integrated navigation is that navigation information of various sub-systems can be integrated [1]. Each system observing the same information sources makes redundancy of measured value increase, which enhances reliability and stability of navigation system. In order to develop the advantages of various navigation systems and enhance stability and accuracy of the system, the high-accuracy and adaptable data integration technique must be applied.

The common data integration methods for navigation system are Kalman filter and some improved algorithms [2]. There are two ways applying Kalman filter technique to data integration for navigation system, centralized Kalman filter and decentralized Kalman filter. For centralized Kalman filter, information fusion is full and filtering accuracy is high. But it includes error conditions of all sub-systems, which not only makes state dimension high and makes calculation great, but also is not good for real-time filter and is bad for fault detection and isolation. Federated Kalman filter is a decentralized Kalman filter. It uses the principle of information distribution. And it can achieve good compromise under different performance requirements. It has the advantages of parallel data process, flexible design, simple calculation and good fault tolerance, which makes it widely applied to integrated navigation system.

The conventional FKF uses the filtering framework of Kalman Filter. It requires that the senior filter and sub0-filters should meet linear gauss assumptions. But it is very difficult to meet the requirement in navigation system.

When there is non-linear element in the system, the conventional KF needs to linearize non-linear element, and the structure of filter needs to be redesigned, which has an influence on filter performance of system.

In order to solve the problem that conventional Kalman filter appears non-linear sub-filter, the paper proposes a new FPF algorithm. The algorithm uses particle filter method to process non-linear sub-filters, and integrates the information of sub-filters in senior filter, which not only achieves good filter effect, but also improves the performance of integration navigation system. The paper takes integration navigation multi-integration problem as an example, and establishes federated particle filter model. And the simulation is made to verify the effectiveness of the algorithm.

2 Design of federated particle filter

2.1 PARTICLE FILTER

Particle filter is a non-linear filtering algorithm. It uses a group of discrete weighted particles to simulate posterior probability. It completes the filtering process by predicting state, updating weight and resampling [3].

The random-state space model is: $\begin{cases} x_{\theta+1} = f(x_\theta) + \omega_\theta \\ y_\theta = h(x_\theta) + v_\theta \end{cases}$,

x_θ and y_θ are state and measurement information of the system. Map $f(\bullet)$ and $g(\bullet)$ are transference model function and measurement model function of system state. ω_θ and v_θ are process noise and measurement noise.

The ultimate goal of filter is to be based on observation sequence $y_{1:\theta} = \{y_i, i = 1, \dots, \theta\}$ to estimate posterior probability distribution function $p(x_\theta / y_\theta)$ recursively. From Bayesian framework and Monte Carlo

---
Corresponding author e-mail: sy481753@163.com

implementation method, we can summarize the algorithm flow of particle filter.

1) Initializing particles and weights. When $\theta = 0$, the particles are extracted from prior density functions to constitute sample set: $\left\{ x_0^i \sim p(x_0), d_0^i = 1/N; i = 1, ..., N \right\}$

2) Importance sampling. When $\theta \geq 1$,

a) Extracting N samples $\left\{ x_\theta^i; i = 1, ..., N \right\}$ from importance density functions.

b) Computing weight of each particle, $d_\theta^i = d_{\theta-1}^i p(y_\theta / x_\theta^i)$.

c) Normalizing weights. $\tilde{d}_\theta^i = d_\theta^i / \sum_{i=1}^{N} d_\theta^i$

3) Resample. Computing the number of effective samples $N_{eff} = round\left(1/\sum_{i=1}^{N}(\tilde{d}_\theta^i)^2 \right)$, where round (•) is the rounding operation. When $N_{eff} < N_{thr}$, the particles are resampled, and *N-th* is the given threshold, which can get the new sample set: $\left\{ x_\theta^i \sim N(x_\theta^i; \hat{x}_\theta^i, P_\theta^i), i = 1, ..., N \right\}$.

4) Output. State and variance is estimated by using the following Equation: $\begin{cases} \hat{x}_\theta = E(x_\theta / y_\theta) \approx \sum_{i=1}^{N} \tilde{d}_\theta^i x_\theta^i \\ \hat{P}_\theta = \sum_{i=1}^{N} \tilde{d}_\theta^i (\hat{x}_\theta - x_\theta^i)(\hat{x}_\theta - x_\theta^i)^T \end{cases}$,

$\theta = \theta + 1$, and returning to the second step.

2.2 DESIGN OF FEDERATED KALMAN FILTER

The architecture of federated Kalman filter is shown in Figure 1. Compared with conventional federated Kalman filter, the sub-filter of it is nonlinear filter consisting of particle sub-filter. As the common reference system, the output information of it and GPS, SINS composes sub-filter. The local estimation $\hat{X}_i$ and covariance matrix K_i of each sub-filter is loaded into senior filter. It is integrated with the estimated value of senior filter, which can get the optimal estimation. From the algorithm principle and flow of particle filter, we can see that particle filtering process achieves posterior distribution of state parameter, which can get the mean and variance information of the state parameter [4]. So the nonlinear sub-filter can be integrated into the framework of federated Kalman filter.

Federated Kalman filter only makes global filter on common state parameter of each system [5]. Applying the method of information distribution makes estimate status of local filters processed by irrelevant ways, which simplifies the flow of filtering algorithm. By combining particle filter and conventional federated Kalman filter, the paper gives the algorithm flow of federated Kalman filter.
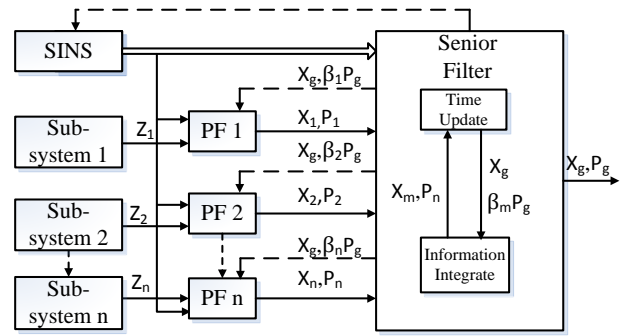
FIGURE 1 Structure of federated Kalman filter

1) Initializing particles and weights. When $\theta = 0$, the particles are extracted from prior density functions to constitute sample set: $\left\{ x_{0j}^i \sim p(x_0), d_0^i = 1/N; i = 1, ..., N, j = 1, ..., M \right\}$, where N is the number of particles, and M is the number of sub-filters.

2) Information distribution process. According to the following formula, federated Kalman filter distributes the information of initial value of the combined system to local filters: $\begin{cases} \hat{X}_i = \hat{X}_g \\ K_i = \gamma_i^{-1} K_g \\ R_i = \gamma_i^{-1} R_g \end{cases}$, where γ_i is information distribution coefficient, and meets $\sum_{i=1}^{N} \gamma_i + \gamma_m = 1$. Estimation error covariance matrix K meets $\sum_{i=1}^{N} K_i^{-1} + K_m^{-1} = K_g^{-1}$. System noise covariance matrix R meets $\sum_{i=1}^{N} R_i^{-1} + R_m^{-1} = R_g^{-1}$.

3) According to state equation, each sub-filter makes filter. When $\theta \geq 1$,

a) Extracting N samples $\left\{ x_\theta^i, i = 1, ..., N \right\}$ from importance density function: $q(x_\theta^i | x_{\theta-1}^i, y_\theta) = N(x_\theta^i; \hat{x}_\theta^i, K_\theta^i)$.

b) Computing weight of each particle, $d_\theta^i = d_{\theta-1}^i p(y_\theta | x_\theta^i)$.

c) Normalizing weights. $\tilde{d}_\theta^i = d_\theta^i / \sum_{i=1}^{N} d_\theta^i$.

d) Computing the number of effective particles: $N_{eff} = round\left(1/\sum_{i=1}^{N}(\tilde{d}_\theta^i)^2 \right)$.

e) Resample. If $N_{thr} \leq N_{eff} \leq N$ (N is the given threshold), particle filter algorithm is used for estimation [6]. If $N_{eff} \leq N_{thr}$ it indicates that particle degeneracy is serious, and the particles need to be resampled, which can get the new sample set $\left\{ x_\theta^i \sim N(x_\theta^i; \hat{x}_\theta^i, K_\theta^i) i = 1, ..., N \right\}$.

f) Status update. The particles and weights are used to compute output state estimation and variance, as follows:

$$\begin{cases} \hat{x}_\theta = \sum_{i=1}^N \tilde{d}_\theta^i x_\theta^i \\ \hat{K}_\theta = \sum_{i=1}^N \tilde{d}_\theta^i (\hat{x}_\theta - x_\theta^i)(\hat{x}_\theta - x_\theta^i)^T \end{cases}.$$

4) Integrating global information. After getting local estimation and estimation of senior filter of each sub-filter, the following formula is used for integration, which can get global state filter and estimator of variance.

$$\begin{cases} \hat{x}_g(\theta) = K_g(\theta)\sum_{i=1}^N (K_i^{-1}(\theta)\hat{x}_i(\theta)) \\ K_g(\theta) = (\sum_{i=1}^N K_i^{-1}(\theta))^{-1} \end{cases}.$$

5) After obtaining global state and estimator of variance, the formula in the second step is used to allocate and reset local filters according to information distribution principles.

6) $\theta = \theta + 1$ and it returns to the third step for repeating the above steps.

3 INS-GPS/BDS integration navigation system filter model

The structure diagram of federated particle filter of SINS-GPS/BDS integration navigation system is shown in Figure 2. SINS/GPS and SINS/BDS composes nonlinear PE sub-filter by the combination mode of position and velocity.
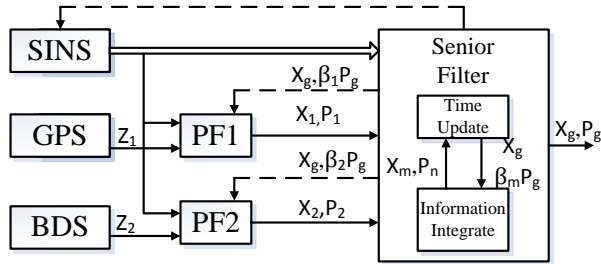
FIGURE 2 Structure of integration navigation filter

3.1 STATE EQUATION

The state variable of the selected system is:
$X(t) = [\varphi_e, \varphi_n, \varphi_u, \delta v_{ie}, \delta v_{in}, \delta v_{iu}, \delta L_i, \delta \lambda_i, \delta h_i,$
$\varepsilon_{bx}, \varepsilon_{by}, \varepsilon_{bz}, \nabla_{bx}, \nabla_{by}, \nabla_{bz}, \delta t_u, \delta t_{ru}]^T$.

The sensor error model and inertial navigation equation is used to establish state equation. ϕ_e, ϕ_n, ϕ_u means, that attitude angle error, $\delta v_{ie}, \delta v_{in}, \delta v_{iu}$ means velocity error, $\delta L_i, \delta \lambda_i, \delta h_i$ represents position error, $\varepsilon_{bx}, \varepsilon_{by}, \varepsilon_{bz}$ means constant drift of gyroscope, $\nabla_{bx}, \nabla_{by}, \nabla_{bz}$ means accelerometer zero offset, δtu is equivalent distance error caused by clock error, and δtru is equivalent distance error

caused by clock error frequency. The noise of state equation is: $W = [\omega_{rbx}, \omega_{rby}, \omega_{rbz}, \omega_{abx}, \omega_{aby}, \omega_{abz}]^T$.

The variable in the formula means the noise of gyroscope and accelerometer on coordinate system of carrier [7]. According to the description, the state equation of the system is $\dot{X}(t) = F(t)X(t) + G(t)W$ in which F(t) is state-transition matrix, X(t) is state parameter, G(t) is noise driving matrix and W is system noise matrix.

3.2 MEASUREMENT EQUATION

The measurement equation of SINS/GPS sub-filter is the same to that of SINS/BDS sub-filter model. The paper takes SINS/BDS as an example for analysis. SINS/BDS integration navigation uses the combination mode of pseudo-range and pseudo-range rate [8].

In earth coordinate system $Ox_e y_e z_e$, pseudo-range calculated by inertial navigation location is:
$\rho_{Ii} = \sqrt{(x_I - x_{si})^2 + (y_I - y_{si})^2 + (z_I - z_{si})^2}$, where x_I, y_I, z_I is the location of $Ox_e y_e z_e$ system output by SINS, which is achieved by calculating longitude, latitude and height output by SINS. x_{si}, y_{si}, z_{si} is the location of the i satellite [9].

The pseudo-range measured by the i satellite for BDS is: $\rho_{Bi} = R_i + \delta t_u + v_{\rho i}$.

From the equation above, we can get pseudorange measurement value is: $\delta \rho_i = \rho_{Ii} - \rho_{Bi}$.

The measurement equation of pseudo range difference is: $Z_\rho(t) = [\delta\rho_1 .. \delta\rho_n]^T = H_\rho(t)X(t) + V_\rho(t)$, where $V_\rho(t)$ is the pseudo range measurement noise of system [10] and n is the number of visible stars.

4 Simulation of federated particle filter algorithm

The structure of SINS-GPS/BDS integrated navigation and federated particle filter is used to simulate the system. And the conventional federated Kalman filtering method is used to simulate the system. The simulation condition requires that the output frequency of inertial navigation is 50Hz, the output frequency of GPS and BDS is 1Hz, zero bias stability of gyroscope is 0.01°/h, and zero bias stability of accelerometer is the flight path consists of acceleration, climb and level flight. The initial position of plane is 39 degrees north latitude and 116 degrees eastern longitude. The height is 500m. The initial velocity is 350m/s, the number of particles is 2000 and the simulation time is 1200s.

Figure 3 is the velocity error of conventional federated Kalman filtering method and federated particle filtering method. Figure 4 is position error of two methods. We can see that the filtering effect of federated particle filter is evidently better than that of conventional federated Kalman filter. The reason for which is that conventional

federated Kalman filter generally uses expanded Kalman filter to process nonlinear subfilter, and expanded Kalman filter uses the method of Taylor expansion to process nonlinear link, which ignores higher order term, linearizes nonlinear model and introduces linear error. But particle

filter directly processes nonlinear model without introducing new error, so the accuracy of federated particle filter is better than that of conventional federated Kalman filter for processing integrated navigation system with nonlinear links.
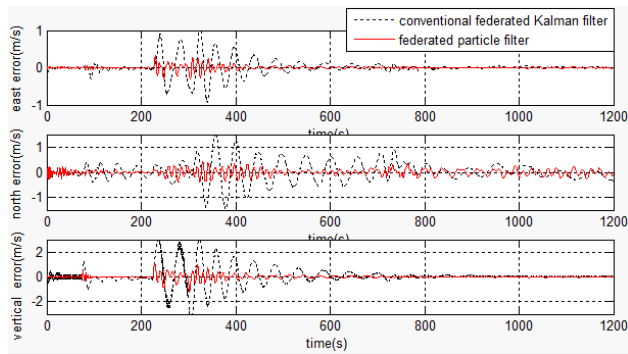
FIGURE 3 Velocity error curve of federated particle filter and conventional federated Kalman filter

FIGURE 4 Position error of conventional federated Kalman filter and federated particle filter

5 Conclusions

In integrated navigation multi-information data processing system, federated Kalman filter has the characteristics of flexible design and good error resilience, which can solve multi-information data integration of integrated navigation system. The paper introduces particle filter into the structure of federated Kalman filter, which expands

conventional federated Kalman filter to solve nonlinear system state and parameter estimation under complicated environment. Federated particle filter is applied to combination navigation, which provides a reference method for processing nonlinear problems in integrated navigation multi-information data processing system. The simulation verifies the effectiveness of algorithm.

References

[1] Gao S, Zhong Y, Zhang X, Shirinzadeh B 2009 *Aerospace Science and Technology* **5** 1-6
[2] Julier S J, Laviola J J 2007 *IEEE Transactions on Signal Processing* **55**(6) 2774-84
[3] Nicolo M, Morclli C, Rampa V 2008 *IEEE Transactions on Signal Processing* **56**(8) 3801-9
[4] Hiliuta A, Landry R, Gagnon F 2004 *IEEE Transaction on Aerospace and Electronic Systems* **4**(40) 591-600
[5] Shi H, Wu Z, Liu B 2006 *Computational Engineering in Systems Applications* **10**(4-6) 651-3
[6] Lee J G, Park C G, Park H W 1993 *IEEE Transaction on Aerospace and Electronic System* **29**(4) 121-130
[7] Sarkka S, Vehtari A, Lampinen J 2007 *Information Fusion* **8**(1) 2-15
[8] Giremus A, Tourneret J Y, Calmettes V 2007 *Signal Processing* **55**(4) 1275-85
[9] Toledo-Moreo R, Zamora-Izquierdo M A 2007 *IEEE Transactions on intelligent Transportion Systems* **8**(3) 491-511
[10] Bevly D M, Ryu J, Gerdes J C 2006 *IEEE Transactions on Intelligent Transportation Systems* **7**(4) 483-93

Authors

Yang Shi, born in 1989, Inner Mongolia, China

Current position, grades: Master's degree student in optical engineering at Beihang University.
University studies: BS degree at Beijing Union University, Beijing, China, in 2012.
Scientific interest: optical engineering, signal analysis and processing.

Huipeng Li, born in 1975, Inner Mongolia, China

Current position, grades: associate professor, School of Instrumentation Science and Opto-electronics Engineering, Beihang University.
University studies: PhD degree in precision instrument and machinery from Harbin Institute of Technology, China, in 2006.
Scientific interest: sensor technology, opto-electronics testing.
Publications: 8 papers, 2 patents.

Xianmu Li, born in 1990, Shandong, China

Current position, grades: PhD student in optical engineering at Beihang University.
Scientific interest: optical engineering, signal analysis and processing.

Lu Wang, born in 1988, Henan, China

Current position, grades: PhD student in optical engineering at Beihang University.
Scientific interest: optical engineering

A compressed-domain audio fingerprint algorithm for resisting linear speed change

Liming Wu[1], Wei Han[1, 2*], Songbin Zhou[2], Xin Luo[1], Yaohua Deng[1]

[1]*School of Information Engineering, Guangdong University of Technology, Guangzhou 510006, China*

[2]*Guangdong Institute of Automation, Guangzhou 510070, China*

Abstract

Existing compressed-domain audio fingerprint algorithms have been able to be used to recognize audio information effectively according to hearing content, and are robust to common time-frequency domain distortion, including echo, noise, band-pass filtering, 32Kbps@MP3 and so on. However, they are poor in resisting linear speed change, which is a very common method for audio processing. In this paper, we propose a novel compressed-domain audio fingerprint algorithm. It is robust to large linear speed change via using auto-correlation function to reduce unaligned degree of MDCT spectrum sub-bands' energy. Besides, it is similar with existing compression-domain audio fingerprint algorithms on the other aspects.

Keywords: compressed-domain audio recognition, audio fingerprint, linear speed change, robustness

1 Introduction

As compressed format have become the main form in storage and transmission of audio files, extract fingerprint directly from compressed-domain audio for audio recognition appears more practical significance. Existing compressed-domain audio fingerprint algorithms [1-4] have owned good robustness to common time-frequency domain distortion, such as echo, noise, band-pass filtering, equalization, volume changing, 32Kbps@MP3, etc. However, they are insufficient for Linear Speed Change (LSC), which is an idiomatic mean of audio processing. Especially radio stations often play music with little of acceleration, the business reason is to shorten the time of playing music can bring more advertisements or other commercial purposes. On the other hand, most audiences maybe prefer the faster rhythm [5].

Nevertheless, it has not been seen in researching compressed-domain audio recognition after dealing with LSC. Only a handful of uncompressed-domain audio fingerprint algorithms [6-9] do it, and obtain good robustness. But this isn't conform to the status that compressed format has been mainstream.

In the remainder of this paper, Section 2 discusses the trouble caused by LSC, and introduces implementation process of the proposed compressed-domain audio algorithm. Section 3 expounds why auto-correlation function can be used to resist LSC. Section 4 utilizes experiments to test the performance of novel algorithm. Section 5 summarizes the full text.

2 Audio fingerprint scheme for resisting LSC

2.1 THE DIFFICULTY IN RECOGNITION DERIVED FROM LSC

Assume that original audio signal is $x(t)$, and its corresponding Fourier spectrum is $X(w)$. After dealing with LSC, the time-domain signal and frequency-domain signal will become $x'(t)$ and $X'(w)$ respectively.

$$x'(t) = x(t / \rho + t_0) \,, \tag{1}$$

$$X'(w) = \int x'(t)e^{-jwt}dt = \int x(\rho / t + t_0)e^{-jwt}dt \,. \tag{2}$$

ρ is the scalability factor of LSC, and t_0 is the translational time.

After a series of calculation, the energy of $X'(w)$ can be expressed as follows:

$$\left|X'(w)\right| = \rho\left|X(\rho w)\right| \,. \tag{3}$$

To a large extent, MDCT spectrum is a linear approximation of Fourier spectrum, especially when only consider its energy [10]. Exactly, **MDCT** spectrum energy is selected to extract fingerprint in the proposed compressed-domain audio fingerprint algorithm. So hypothesize that δ indicates the linear relationship between MDCT spectrum and Fourier spectrum. Use $M(w)$ and $M'(w)$ individually denote the MDCT frequency-domain signal of original audio and distorted version leaded by *LSC*.

$$\left|M(w)\right| = \delta\left|X(w)\right| \,, \tag{4}$$

---

* *Corresponding author* e-mail: w.han@gia.ac.cn

$$\left|M\,'(w)\right| = \delta\left|X\,'(w)\right| = \delta\rho\left|X(\rho w)\right|. \tag{5}$$

Synthesize the above analysis, due to LSC will make the playing speed of audio faster or slower, which result in the change of MDCT spectrum energy. Worse, energy migration also happens as the variation of frequency of audio signal. However, most compressed-domain audio fingerprint algorithms extract audio fingerprint based on MDCT spectrum. This will induce that extracted audio fingerprint will have biggish difference before and after the distortion, which causes to reduce recognition rate.

Although all compressed-domain audio fingerprint algorithms have a large overlap degree between adjacent MDCT blocks [1-4], which will ensure that, even in the worst scenario, the unaligned extent of relevant blocks' border is very small. In other words, the sub-fingerprints of unknown audio clip waited to be identified are still very similar to the sub-fingerprints of the same clip in the database. Thus, algorithms can withstand a certain range of LSC. However, LSC will cause that MDCT coefficients

are not aligned along time-axis and frequency-axis, and the unaligned degree would have a cumulative effect as times goes on. So that audio recognition will become more difficult with the increase of LSC degree.

2.2 THE PROPOSED COMPRESSED-DOMAIN AUDIO FINGERPRINT ALGORITHM

As shown in Figure 1 is the overview of the proposed compressed-domain audio fingerprint algorithm. In fact, for almost all compressed-domain audio fingerprint algorithms, most of their steps are similar, in addition to the fingerprint feature, which is derived from MDCT spectrum energy. In general, the eventual audio fingerprint is directly obtained by fingerprint feature with a chain of simple calculation. $F(n,m)$ denotes one bit fingerprint, n generally refers to a sub-band. And the fingerprint extracted from a sub-band is called a sub-fingerprint, whose length is m.
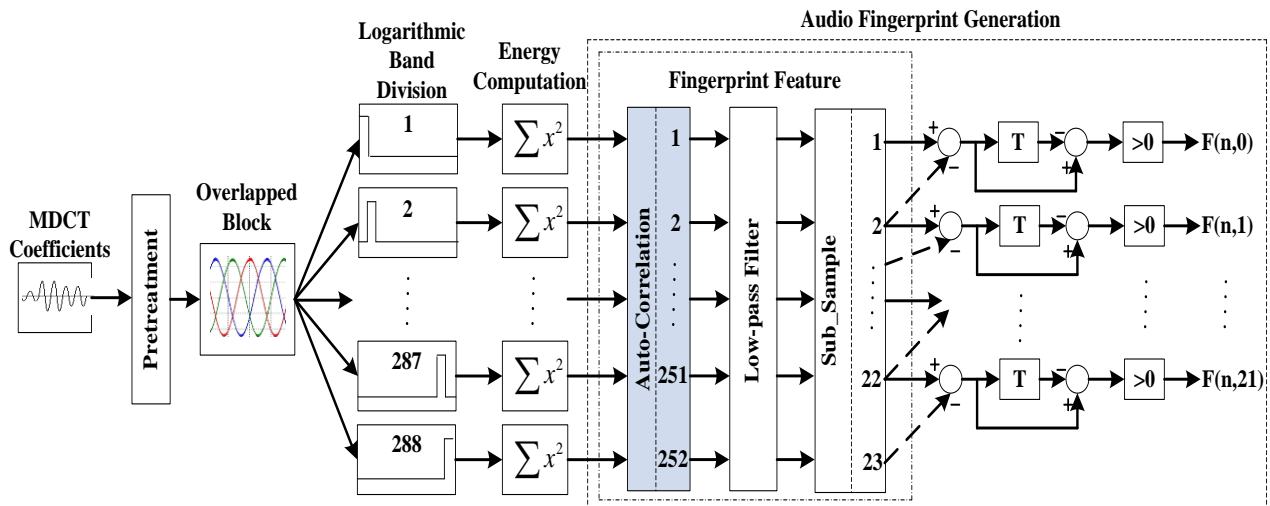
FIGURE 1 Overview of the proposed compressed-domain audio fingerprint algorithm

The proposed compressed-domain audio fingerprint algorithm will be started to introduce after MDCT coefficients originated from audio codec which can be accomplished according to reference [11]. At first, 12 frames MDCT coefficients acquired from the audio with MP3 format or *wave* format consist of one block with the length is about 0.31 seconds (when sampling rate is 44.1 KHz). The overlap degree is 95.83% between adjacent blocks, i.e., there is a section of hop distance because a MP3 frame contains two sections. Secondly, every section of MDCT spectrum (contains 576 coefficients) is divided into 288 sub-bands on the basis of logarithmic scale in the range of 300~2000Hz (the most relevant frequency range with the proposed fingerprint algorithm), and calculate the energy of each of them. Then sum these sub-bands' energy which locate in the same frequency range and pertain to an identical block. Assume that $SEN(i, j)$ represents the energy of the j-th sub-band which is belong to the i-th block, $S(m,n)$ indicates the energy of the n-th MDCT coefficient in the m-th section, $MDCT_p$ and $MDCT_q$ respectively represents the upper and lower bounds of

MDCT coefficients which belong to the same sub-band. Therefore, the $SEN(i, j)$ can be calculated by Equation (6). Thus, in any block, an energy sequence composed of 288 sub-band's energy.

$$SEN(i, j) = \sum_{m=0.5(i+1)}^{12+0.5(i+1)} \sum_{n=MDCT_p}^{MDCT_q} \left|s(m,n)\right|^2 . \tag{6}$$

As has been analyzed in Section 2.1, LSC will lead to shift for the energy of MDCT sub-band. It is well known that auto-correlation function has the invariance to shift. Thus, it can be used to handle sub-band energy sequence of each block in order to resist LSC. And after this, the length of energy sequence reduced to 252. In order to improve the robustness, the results from auto-correction operation should be processed by a low-pass filter. Finally, the 252 energy values will be decreased to 23 energy data by down-sampling process. The purpose to shorten the length of energy sequence is to simplify the computational complexity at the time of fingerprint matching.

Suppose that *Ena(i,m)* (*i* is the number of MDCT block, *m*=0, 2,..., 21) represents the final 22 energy value. The

$$F(i,m) = \begin{cases} 1 & if \quad Ena(i,m) - Ena(i,m+1) - Ena(i-1,m) + Ena(i-1,m+1) > 0 \\ 0 & if \quad Ena(i,m) - Ena(i,m+1) - Ena(i-1,m) + Ena(i-1,m+1) \le 0 \end{cases} \tag{7}$$

$$\rho_{ee}(x) = \sum_{j=1}^{M} e_1(k+j) e_2(x+j) \qquad 1 \le x \le N-M . \tag{8}$$

As a result, a sub-fingerprint contains 22 bits can be extracted from a block. It does not own enough information to identify the corresponding complete audio clip. But fingerprint sequence, which is often referred to as a query fingerprint block, can do it. In the compressed-domain audio fingerprint algorithm of this paper, the length of audio fragment corresponds to a query fingerprint block is about 3s (114 MP3 frames). And a fingerprint block includes 204 sub-fingerprints, which has a total of 4488 bits. For example, if there is an audio clip with the length of 5 minutes, the length of unknown audio needed to index it in the fingerprint database only is 3s. Of course the unknown audio must be a part of original audio clip.

The Bit Error Rate (BER) is used to estimate the similarity between two audio clips. If the BER between query fingerprint block and one fingerprint block stored in the database beforehand is lower than the threshold *T*, it is considered to be a reliable match. A large number of experiments have proved that when the BER is less than *T*=0.35, it is seen that matching result is effective. Detailed index and match process can be implemented according to reference [12].

3 The principle of resisting LSC

Compared with the existing compressed-domain audio fingerprint algorithms, novel algorithm is added the auto-correlation processing. This is because auto-correlation function has invariance to shift, which can be proved as follows.

Suppose that *f(t)* represents audio signal, its auto-correlation coefficient $\rho_{ff}(x)$ is:

$$\rho_{ff}(x) = \int f(t) f(t+x) dt . \tag{9}$$

g(t)=f(t+a) generated by shift of *f(t)*, its auto-correlation coefficient is $\rho_{gg}(x)$. $\rho_{gg}(x) = \rho_{ff}(x)$ can be inferred from the

fingerprint of an audio clip is actually a set of binary bit stream, and every bit of it defined by the Equation (7).

characteristic of auto-correlation function.

So, the auto-correlation coefficient of continuous function owns invariance to shift. But now is to deal with a discrete energy sequence. In order to approximate this characteristic of auto-correlation function, select a fixed sub-sequence e_1 from the complete energy sequence $e(n)$ of each block (contains 288 energy values) and utilize it to do correlation calculation with any sub-sequence e_2 in the identical energy sequence. The auto-correlation coefficients $\rho_{gg}(x)$ of $e(n)$ can be computed by Equation (8).

N stands for the length of entire energy sequence of each block, *M* represents the length of sub-sequence, and *k* denotes the starting position of a sub-sequence in the full sequence. In the proposed algorithms, *M*=36, *k*=54.

4 The effect of proposed algorithm

4.1 RESISTING LSC

Randomly choose 1000 audio clips (MP3 or *wav* format, stereo, 16 bit quantification, 44.1 KHz sampling rate) which belong to 10 different types of music, including DJ, electronic, classical, blues, jazz, folk, light music, hip-hop, country, rock and so on. The length of each clip is 20 seconds. Use various degree of LSC to distort each clip. Then severally extract fingerprint from the first 3 seconds in the distorted version and original version with existing algorithms and novel algorithm, and calculate each BER value between the fingerprints of an original version and its anamorphic version. Thus, 1000 BER data can be acquired for each treatment of LSC. The average value of these BER data embodies the ability of fingerprint algorithm to resist LSC. Figure 2 shows the relation between LSC and BER. It clearly shows that the proposed algorithm can resist LSC from −7% to +7%, which is obviously higher than the existing algorithms. In Figure 2 and Figure 3, the curve marked by "+" expresses the result of novel algorithm; the other curves demonstrate the results of exiting algorithms.
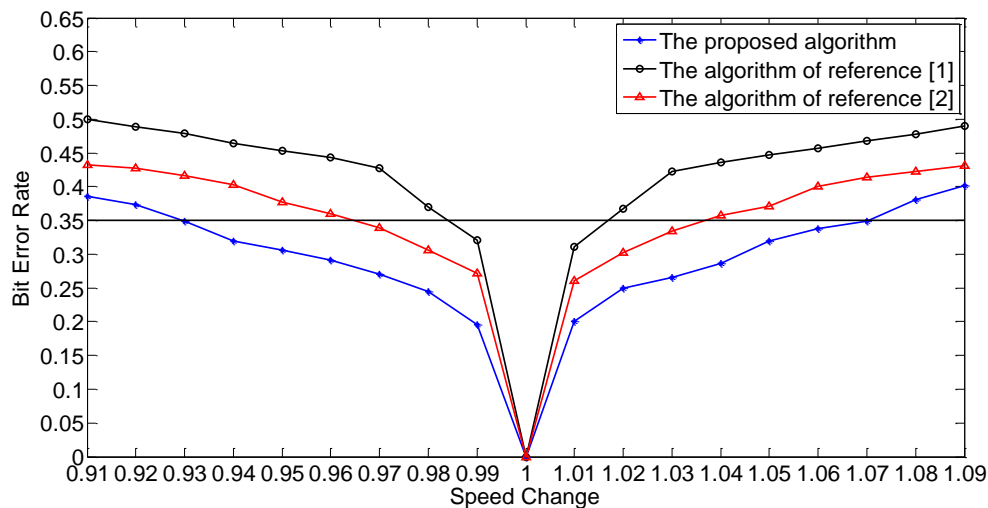
FIGURE 2 The robustness to LSC
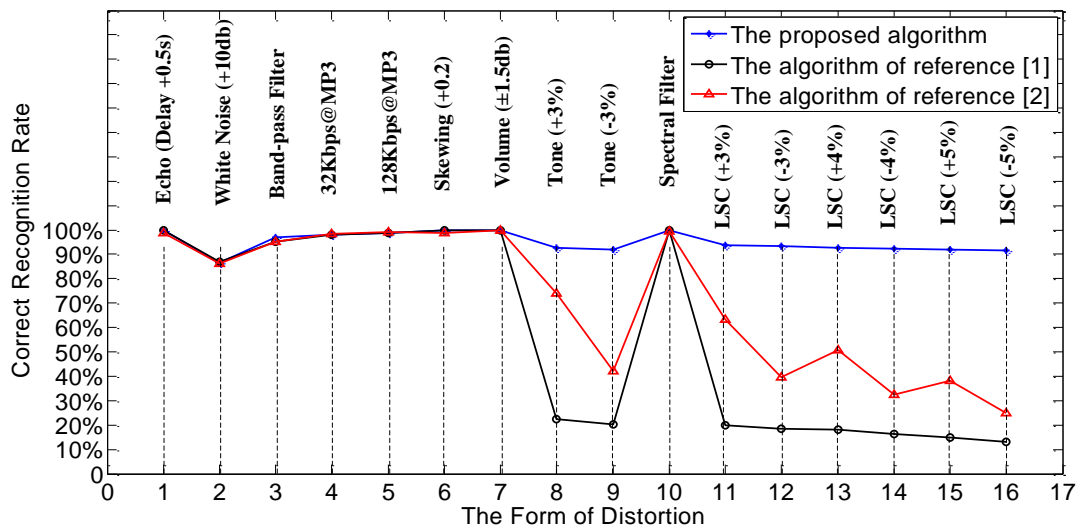
FIGURE 3 The comparison about recognition rate

4.2 THE ROBUSTNESS TO OTHER DISTORTION PROCESSING

Recognition rate, which reflects the reliability of algorithm, is the most important indicator to evaluate an audio fingerprint algorithm. Apply the following experiment to compare the recognition ability among novel algorithm, algorithm [1] and [2]. Experimental samples are the same as Section 4.1. Do 17 different distortion operations for each clip, including echo (delay 0.5s), noise (+10db), band-pass filter (300Hz~300Hz), 32Kbps@MP3, 128Kbps@MP3, offset (+0.2), volume (+1.5db), volume (−1.5 db), pitch (+3%), tone (-3%), spectrum filter (+1.0db), LSC (+3%), LSC (–3%), LSC (+4%), LSC (–4%), LSC (+5%), LSC (–5%). Randomly intercept a fragment with a length of 3s (or 114 MP3 frames) from each distorted audio clip. Thus there are 17000 unknown audio clips. Construct the fingerprint database according to the method of literature [12] for the fingerprints of original 1000 samples. And use the fingerprint of these 17000 unknown audio clips to search the homologous audio information in the database, the statistics of

identification results are shown in Figure 3. The results clearly show that the novel algorithm is similar to or better than the existing algorithms for most of time-frequency-domain distortion. Besides LSC, novel algorithm also has better robustness to Tone change. This is due to Tone change is just a special case of LSC.

5 Conclusions

This paper proposes a novel compressed-domain audio fingerprint scheme. Experimental results show that the robustness of novel algorithm is similar to exiting algorithms for most audio distortion. But novel algorithm has better performance to deal with LSC, can resist it at the range from -7% to 7%. Generally, this is enough to handle the LSC in radio. Follow-up work of this paper will increase the resistance range to LSC, and improve the robustness for other common distortion.

References

[1] Wu L M, Han W, Deng U H 2013 *International Journal of Advancements in Computing Technology* **5**(9) 291-8

[2] Li W, Liu Y D, Xue X Y 2010 *Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval* Geneva Switzerland 627-34

[3] Chih-Chin Liu, Po-Feng Chang 2011 An efficient audio fingerprint design for MP3 music *The 9th International Conference on Advances in Mobile Computing & Multimedia* 190-3

[4] Sáenz-Lechón N, Osma-Ruiz V, Godino-Llorente J I, Blanco-Velasco M, Cruz-Roldán F, Arias-Londono J D 2008 *IEEE Transactions on Biomedical Engineering* **55**(12) 2831-5

[5] Cano P, Batlle E, Kalker T, Haitsma J 2005 A review of audio fingerprinting *Journal of VLSI Signal Processing Systems for Signal, Image and Video Technology* **41**(3) 271-84

[6] Liu J X, Zhang T X 2011 *International Conference on Computing, Information and Control* Wuhan China 360-68

[7] Zhu B, Li W, Wang Z, Xue X 2010 *18th ACM International Conference on Multimedia ACM Multimedia* Firenze, Italy 987-90

[8] Jin Soo Seo, Jaap Haitsma, Ton Kalker 2002 *Proceedints of the 1st IEEE Benelux Workshop on Model based Processing and Coding of Audio* Leuven, Belgium 45-8

[9] Sun W, Lu Z M, Yu F X, Shen R J 2012 *International Journal of Digital Crime and Forensics* **4**(2) 49-69

[10] Wang Y, Yaroslavsky L, Vilermo M 2000 *Proceedings of the 5th International Conference on Signal Processing* Beijing China 44-47

[11] International Organization for Standardization 1993 *ISO/IEC 11172-3*

[12] Haitsma J, Kalker R 2002 *Proceedings of International Symposium on Music Information Retrieval* Paris, France 107-15

Authors

Liming Wu, born in January, 1962, Jieyang city, P.R. China

Current position, grades: professor of Guangdong University of Technology, Master Tutor.
University studies: graduated from the South China University of Technology in 2004 in China.
Scientific interests: speech processing, machine vision, optical, mechanical and electrical integration.
Publications: more than 80 papers.
Experience: teaching experience of 33 years, more than 20 scientific research projects.

Wei Han, born in July, 1987, Jingmen city, P.R. China

Current position, grades: assistant researcher of Guangdong Institute of Automation.
University studies: M.E. graduated from Guangdong University of Technology in 2013 in China.
Scientific interests: audio recognition, automatic control technology.
Publications: 8 papers.
Experience: 2 scientific research projects.

Songbin Zhou, born in July, 1978, Chaozhou city, P.R. China

Current position, grades: Doctor, Researcher of Guangdong Institute of Automation.
University studies: Dr.E. graduated from the South China University of Technology in 2008 in China.
Scientific interests: pattern recognition, intelligent sensing technology.
Publications: 15 papers.
Experience: 10 scientific research projects.

Xin Luo, born in April, 1989, Yichang city, P.R. China

University studies: Master's degree student of Guangdong University of Technology.
Scientific interests: Digital Image-processing, pattern recognition.
Publications: 3 papers.

Yaohua Deng, born in October, 1978, Huizhou city, P.R. China

Current position, grades: doctor, associate professor of Guangdong Institute of Automation.
University studies: Dr.E. graduated from in South China University of Technology in 2012 in China.
Scientific interest: intelligent measurement and control system, detection in flexible material processing, pattern recognition.
Publications: more than 20 papers.
Experience: teaching experience of 10 years, 6 scientific research projects.

System model of reconfigurable embedded motion control system based on IEC61499

Shoushan Liu[1]*, Maoyong Cao[1], Yan Chen[2]

[1]*College of Electronic, Communication and Physics, Shandong University of Science and Technology, Qingdao, 266590, China*

[2]*College of Mechanical and Electronic Engineering, Shandong University of Science and Technology, Qingdao, 266590, China*

Abstract

Based on the function block (FB), and distributed system model of IEC61499 standard, the application and system model of embedded and reconfigurable motion control system are modelled. According the motion control system architecture built by embedded microprocessor and reconfigurable logic devices, the function block of IEC61499 can modified to describe the function unit which being used as the basic element to construct the application. Based on the modified function units, in combining with the distribute resources of reconfigurable logic devices, the distributed application model of the motion control system can be configured. To testify the application model, the interpolation process based on the digital differential analyser (DDA) of computer numerical control (CNC) system is carried out, and the results show that the application model is well performed to describe the embedded reconfigurable motion control system.

Keywords: IEC61499, function block, embedded reconfigurable system, motion control, function unit

1 Introduction

Embedded reconfigurable computer systems (ERCS) are generally described as platforms consisting of programmable elements (such as microprocessors) and reconfigurable logics (Such as the FPGA: Field Programmable Gate Array). To an application, the ERCS can provide full flexibility, adaptability and better performance, comparing to architectures based only on embedded microprocessor or on application specific integrated circuit (ASIC).

Researchers have made great efforts to the methodologies of how to map an application into the ERCS. The relating research subjects of the mapping methodologies include task scheduling, software/hardware partition, reconfiguring methods, operation system configuration, memory management et al. [1-3]. Generally, before the mapping subjects having been implemented, the functional units (defined as the instances of tasks being implemented on the ERCS in this paper) interconnection diagram is always necessary. Although varies functional units interconnection diagrams (or system model) have been carried out, there have no one standard model for all kinds of application construction, the main reason is that the system models are always based on different ERCS platforms or to different applications. In addition, transplanting, or updating, an existing system model to new platform with different hardware framework always means more costs.

In robot or computerized numerical control (CNC) system, the motion control is always afforded by the controller system, which is in charge of the interpolation, path planning, tracking, obstacle avoidance, communication and other control tasks. To the controller of robot or NCS that have all the control tasks implemented in software, it would almost be a "mission impossible" to carry all the tasks out under real-time constraints, especially with computation-intensive tasks such as interpolation, path tracking with video information, path planning et al. The expensive ASIC can fulfil the speed criteria; however it is a complicated and expensive procedure if a slight changes occurs [4]. To the flexibility, real-time requirements of robot or NCS, a controller system based on the architecture of ERCS obviously would be a feasible solution, as the ERCS can provide both software and ASIC with reconfiguration.

To configure a motion control system based on ERCS, a distributed functional unit's interconnection diagram (system model) should be confirmed. In consideration of the computation-intensive characteristic of the algorithms in motion control system of robot or CNC, and the extensible, distributable requirements of system, a coarse-grained functional unit model is presented in this paper. To set up a uniform functional unit model for both software and reconfigurable implementations in the system, the function block of IEC61499 standard is introduced.

The IEC61499 standard defines a set of implementation-independent basic concepts for modular, reusable, distributed industrial process measurement and control system (IPMCS). The function Block (FB) is the basic unit, which can perform a specific algorithm or a kind of operation. The packaged function block (FB) can

be interconnected or called through the interfaces, in respond to different applications.

In order to building a motion control system for robot or CNC based the architecture of ERCS, the main job of this paper will focus on the functional units interconnection diagram (system model) based on function block (FB) of IEC61499. The rest of this paper is organized in 5 sections. Section 2 describes the function block (FB) and distributed system model of IEC61499. Section 3 presents the analysis of application model of motion control system for robot or CNC based on ERCS. In Section 4, the distributed model of the motion control system based on ERCS is described and confirmed. Experiments and results are described in Section 5. Section 6 draws the conclusions.

2 The distributed model of IEC61499 based on function block

The IEC61499 is a standard, proposed by International Electro-Technical Commission, for the development of distributed industrial process measurement and control system (IPMCS). The standard also describe the methodology that utilizes the Function block (FB) as the main building block and defines the way that FBs can be used to define robust, re-usable software components that constitute complex IPMCSs. Complete control applications can be built from networks of FBs that are formed by interconnecting their inputs and outputs [5].

2.1 FUNCTION BLOCK

The FB can also be described as a class that encapsulates a functionality, which could be one kind of algorithm or operation. One FB instance is described as the event-driven object of a defined FB class. During the construction of distributed application, the FB selection would depend on different FB libraries (such as the control algorithms, alarming manager, communication), and the selected FB eventually would be turned out as the FB implemented instance.

In IEC61499, basic FBs are the basic building blocks. The basic function block consists of a body with interface for data inputs and data outputs and of a header with interface for event inputs and outputs (Figure 1). The Execution Control Chart (ECC) takes charge of the execution of algorithms depending on input data, input events, inner variables and produces output events and output data. Composite function blocks may contain multiple FBs-basic FBs or/and composite FBs.
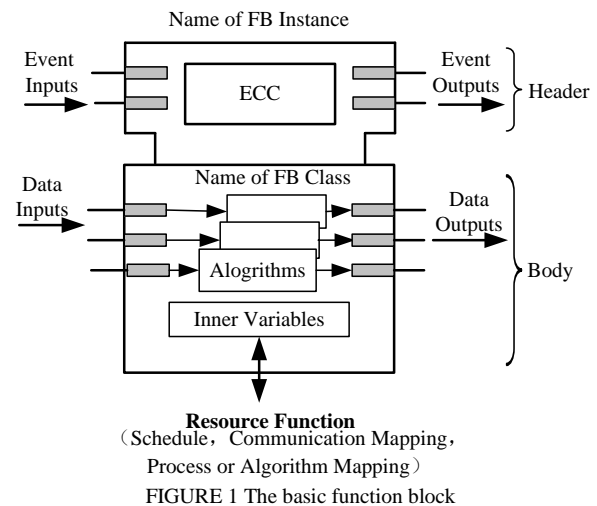
FIGURE 1 The basic function block

2.2 APPLICATION MODEL AND DISTRIBUTED SYSTEM

By interconnecting the basic FBs and composite FBs, a so-called sub-application can be built. The complete application model can be configured through interconnections of the sub-application and different FBs. The interconnections generally are afforded by event streams and data streams.

The distributed system (Figure 2) of IEC61499 is primarily described as the set of interconnected devices or of devices that can communicate through one kind, or more kinds, of communication network.
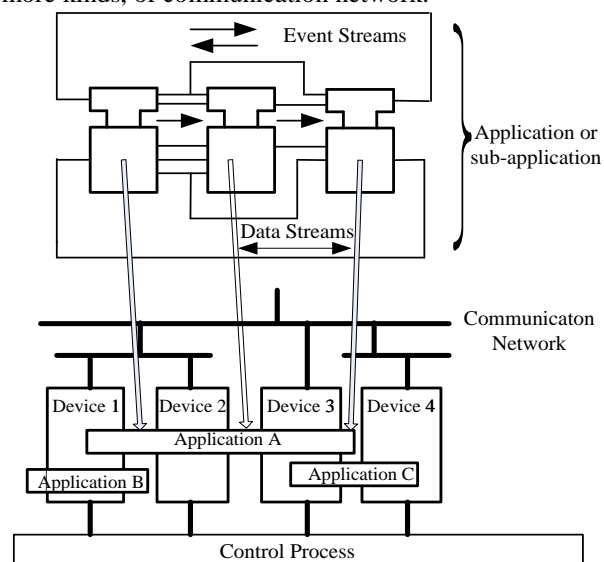
FIGURE 2 Distributed system model of IEC61499

The performances of the distributed system are always described by the complete application model and sub-application models. The application or sub-application would be distributed, as the FBs instances can be distributed, or mapped, to the resources (hardware or software resources being provided for the FBs instances) of one or more devices.

3 Analysis of application model of embedded reconfigurable motion control system

In the motion control system of robot or CNC, involved tasks include path planning, interpolation, tracking, obstacle avoidance, PIC control, communication and other control or signal processing tasks. Corresponding to variations of control objects and working environments, each kind of tasks may be realized by different algorithms. For example, algorithms relating to path planning task include simulate annealing algorithm, artificial potential field method, fuzzy logic algorithm, neural network, particle swarm algorithm and et al.; algorithms involved in interpolation task include pulse increment method (point-to-point comparison method, digital integration method), data sampling, the B spline curve interpolation method etc. In the application design, generally, the algorithms of the tasks are always implemented as executable functional units (or blocks) which can be called by the main procedure of the system. Obviously, the executable functional units (or blocks) are always coarse-grained. This kind of application design method is just object-oriented and justly meets the reusable and interchangeable requirements of system design. Just like the distributed system model in IEC61499, the motion control system can be described by an application model with functional units (of algorithms) interconnected by control and data streams.

To a motion control system based on ERCS, the control tasks or operations are implemented in microprocessors and reconfigurable logics. The available resources (such as the arithmetic unit in CPU, the registers and the memories) in microprocessor, to its software tasks in time domain, can be regarded as infinite, while in the space domain is finite. To algorithms implemented only in software, the only method to realize the parallel structure of one algorithm is to add microprocessors. The reconfigurable logics, otherwise, not only provide the infinite resources in time domain by reconfiguration, but also provide the platform on which the algorithms can be implemented in parallel structure. Especially, many kinds of commercialized reconfigurable logics can provide distributed resources (many FPGAs can provide multi-core of arithmetic units, different scales of memories and programmable bus etc.), which would benefit the applications configuration.

4 The distributed model of controller system based on ERCS

4.1 APPLICATION MODEL BASED ON FUNCTIONAL UNITS AND MANAGER UNITS

As mentioned in Section 3, the controller system of robot or CNC can be described by an application model with functional units (of algorithms) interconnections depending control and data streams. Referring to the basic FB, one functional unit (i.e. one implementation of an algorithm) in controller system can be shown in Figure 3. The interface of the functional unit is a set of data inputs interface, data outputs interface, control streams interface and status streams interface. By the interface (or interfaces), one functional unit can be interconnected to other functional units, or can be called by other procedures or functions. Two or more functional unit can be interconnected into a sub-application or application.
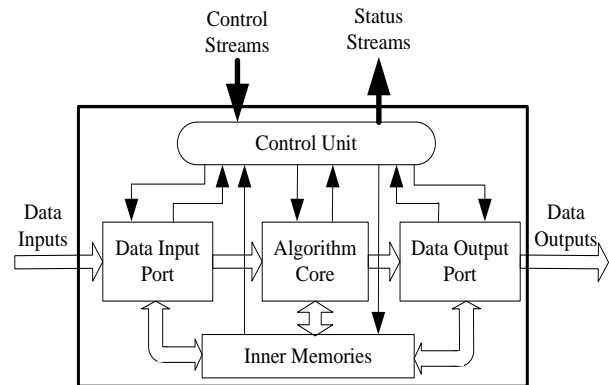
FIGURE 3 The functional unit

During the work of system, the functional unit will be activated by the change of control streams (control events), and begin to receive the input data through the data input port. The algorithm core also be activated and start to perform the data processing: receiving input data and outputting the data being produced during operation. The status streams, through the status port, reflect the operation states (e.g. the idle, busy or suspend) of functional unit and can be read by other functional units or the manager unit (the module bridging the functional units and the task scheduling).

In IEC61499, all the behaviours of the FB are controlled by the Execution Control Chart (ECC). According IEC61499, the behaviours being reflected from ECC include Execution Control status, Execution Control transform and Execution Control action. Thus the behaviours of the ECC can be described by the Finite State Machine (FSM). In this article, the role of the control unit in functional unit of the controller system based on ERCS is as almost same as the role of ECC in FB of IEC61499. The FSMs about the control unit in functional unit are always varied in correspond to the variation of the algorithm. The jobs of the control unit in functional unit include the control to algorithm behaviours, the communications between functional units and responding to the manager unit.

Unlike the functional unit, the role of the manager unit is to bridge different functional units, and to provide management information to upper manager units. The control mechanism of the manager unit can also be described by the FSM. The communication between functional unit and manager unit is performed through the events ports (control stream port and status stream port). By the event communication, the manager can control and monitor the behaviour of the functional units. Two or more manager units can also be combined into a new manager unit by combining the FSMs of the manager units into a new FSM [6]. In addition to the controls to the functional

units, the manager units are also interconnected to the task management and scheduling units, which always belong to layer of operation system (OS). The application model based on the functional units and manager units can be described in Figure 4.
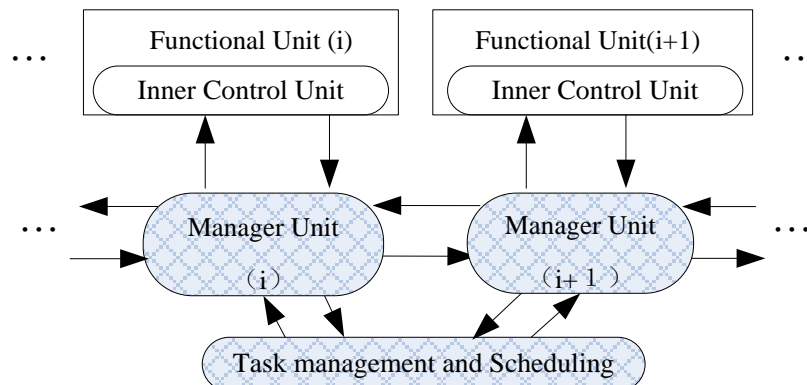
FIGURE 4 Application model based on functional units and manager units

4.2 DISTRIBUTED MODEL OF EMBEDDED RECONFIGURABLE MOTION CONTROL SYSTEM

Based on the devices and application model as shown in Figure 4, the distributed system model of the motion control system can be configured, as shown in Figure 5.
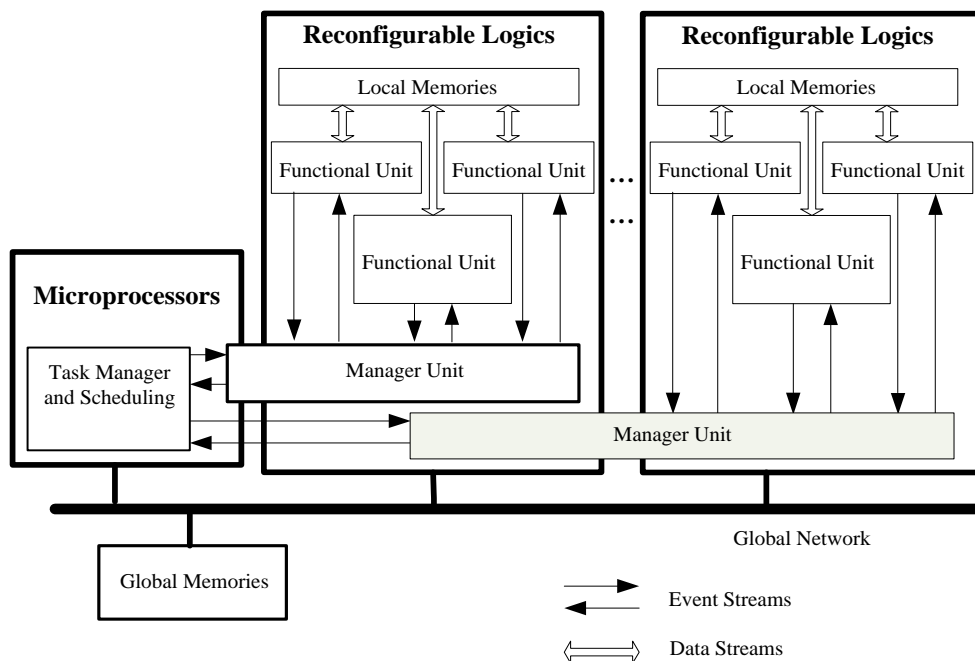
FIGURE 5 The distributed system model of motion control system

Since the devices of the motion control system based on ERCS are microprocessors and reconfigurable logics, the resources (e.g. the arithmetic units, the memories and networks) for the operation of functional units and manager units would be distributed, with temporally and spatially, in the microprocessors and reconfigurable logics (see Figure 5).

5 Experiments

According the proposed design ideas and system model based on IEC61499, an application (Figure 6) about interpolation process of motion control system (for CNC), under the architecture of ERCS, is designed and implemented.
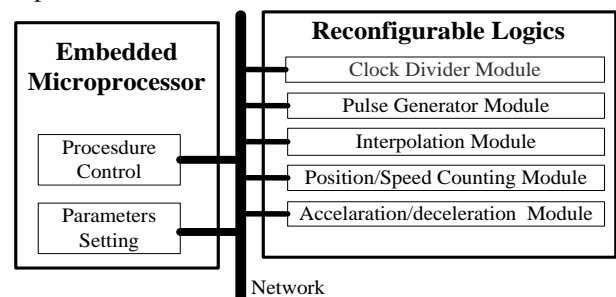
FIGURE 6 Diagram of interpolation

The interpolation process can be considered to be composed of the following modules: clock divider module,

the pulse generator module, the interpolation module, position/speed pulse counting module, acceleration and deceleration control module. Each module of the interpolation is designed as a functional unit. The interpolation application is composed of functional units.

Figure 6 shows the procedure control and parameters setting are performed in software, while the modules of interpolation are implemented in reconfigurable logics. The procedure control is manager unit that responds to the task manager of operation system and takes charge of the interpolation scheduling. Parameters setting provide parameters needed for the interpolation before the state

changing from idle to activation. The realization of the interpolation in reconfigurable logic is shown in Figure 7, where the clock divider module (**clock_div**) is designed in FSM. The position/speed pulse counting module (**capture**) is designed in FSM by combining the filter FSM and the direction discrimination FSM to the two signals of the incremental optical encoder. The implementation of interpolation module (**interpolate**) is a FSM, which is combined by two FSMs: one is the FSM of the interpolation algorithm (the Digital Differential Analyser is adopted), another is the FSM of acceleration and deceleration control module.
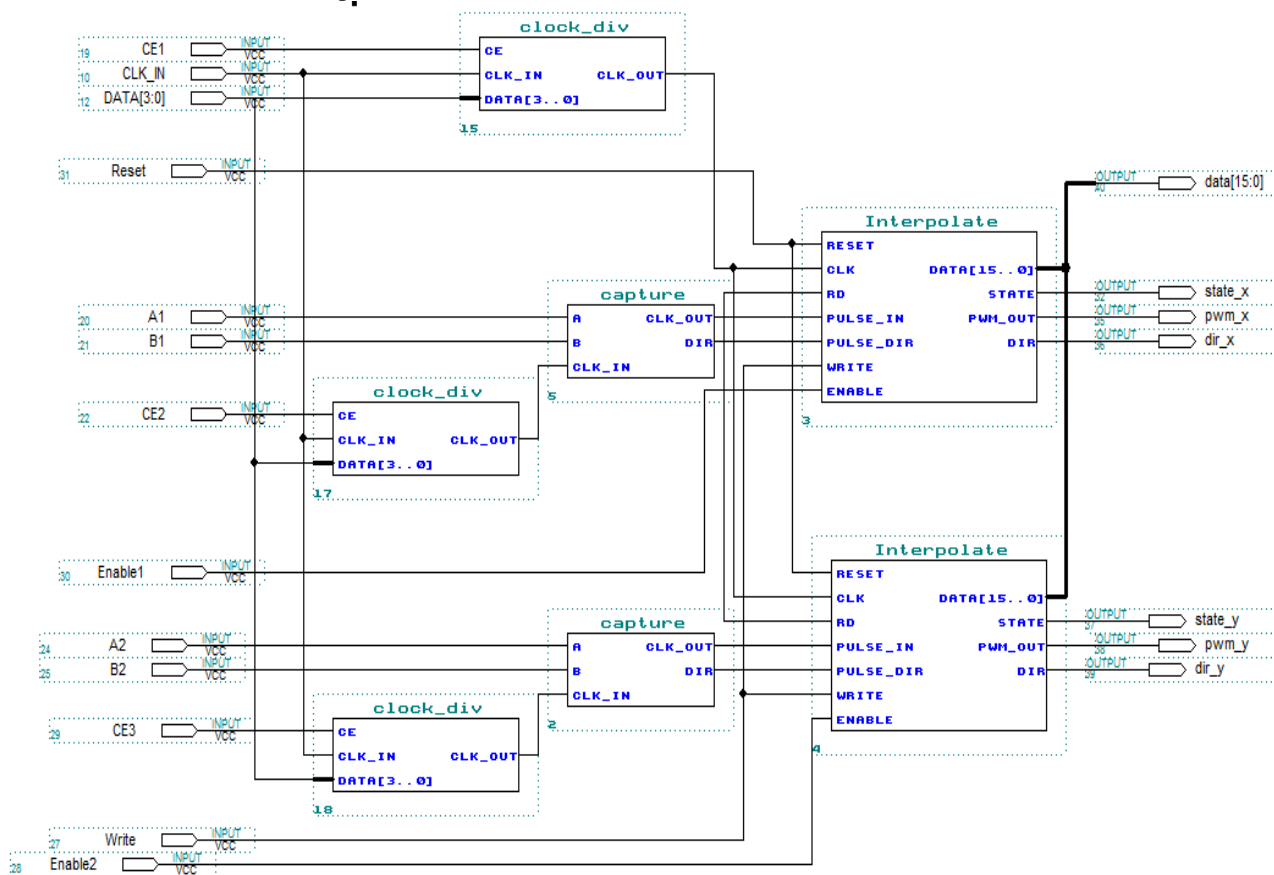
FIGURE 7 The realization of the interpolation in reconfigurable logic
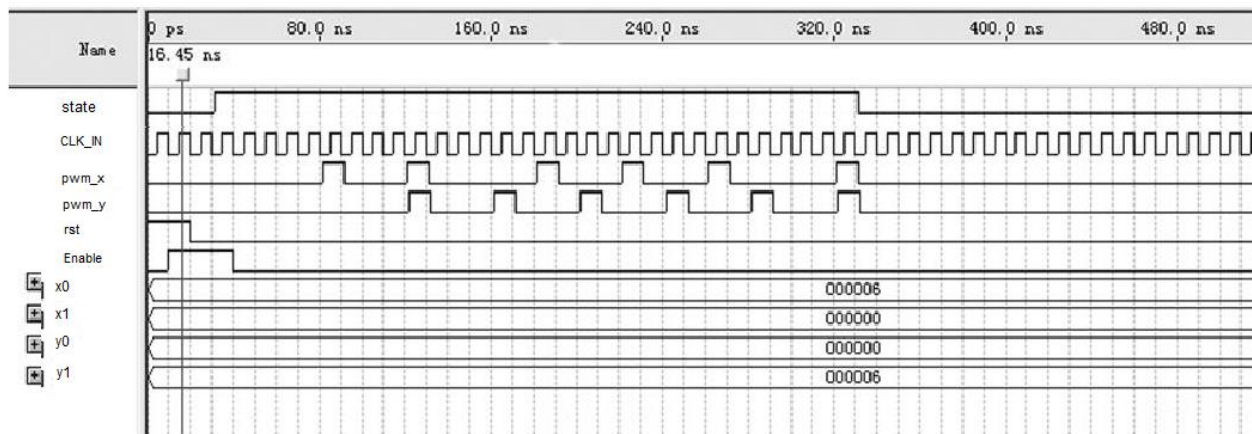
FIGURE 8 The simulation result of interpolation

201

The interpolation process is synthesized and simulated in Quartus II, and main results are shown in Figure 8. The results show that the outputs of the control pulse are in compliance with the Digital Differential Analyser (DDA).

6 Conclusions

In this paper, the concepts of the function block (FB) and the distribute architecture of IPMCS in IEC61499 are introduced to the configuration of motion control application based on embedded reconfigurable computing system (ERCS). The experiments results show that the introduction and application are feasible. Nowadays, the researches on applications of function block (FB) and the distributed system model of IEC61499 have been from the initial definition of the semantic model to the configuration of applicable model. Some researchers believe that the Unified Modelling Language (UML) would be very suitable to describe the behaviours of the FB and the distributed system architecture of IEC61499 for application construction [7]. At same time, the UML has been used to describe the embedded reconfigurable computing system [8, 9]. Out next research object will concentrate on the application of UML to the embedded reconfigurable motion control system.

Acknowledgements

References

[1] Haubelt C, Otto S, Grabbe C, Teich J 2005 *Proceedings of the ASP-DAC 2005 Asia and South Pacific Design Automation Conference* 2005 **1** 298-301

[2] Zhou X, Lu S, Wang F, Qi J 2007 A data-driven uniform programming model for reconfigurable computing *Acta Electronica Sinica* **35**(11) 2123-8 *(in Chinese)*

[3] Eskinazi R, Lima M E, Maciel P R M, Valderrama C A, Filho A G S, Nascimento P S B 2005 A Timed Petri Net Approach for Pre-runtime Scheduling in Partial and Dynamic Reconfigurable Systems *Proceedings of the 19th International Parallel and Distributed Processing Symposium 2005* 154a

[4] Y. Meng 2005 A Dynamic Self-reconfigurable Mobile Robot Navigation System *Proceedings of 2005 IEEE/ASME International Conference on Advanced Intelligent Mechatronics* 1541-6

[5] Thramboulidis K 2006 IEC61499 in Factory Application *Advances in Computer, Information, and Systems Science, and Engineering* 115-24

[6] Liu S, Xu W, Chen Y, Zheng J 2010 The Control Infrastructure for Reconfigurable DSP System *2010 International Conference on Computer and Communication Technologies in Agriculture Engineering (CCTAE)* 294-7

[7] Li Yang, Huang F, Li H, Wang Z 2004 Modeling of the IEC61499 Function Block Control System with UML *Control and Instruments in Chemical Industry* **31**(5) 44-8 *(in Chinese)*

[8] Preuße S, Missal D, Gerber C, Hirsch M, Hanisch H M 2010 On the Use of Model-Based IEC61499 Controller Design *International Journal of Discrete Event Control Systems (IJDECS)* **1**(1) 1-13

[9] Lecomte S, Guillouard S, Moy C, Leray P, Soulard P 2011 A Co-design Methodology Based on Model Driven Architecture for Real Time Embedded Systems *Mathematical and Computer Modelling* **53**(3-4) 471-4

Authors

Shoushan Liu, born in February, 1970, Qingdao, China

Current position, grades: doctor of Mechanical Engineering, Lecture at Shandong University of Science and Technology.
University studies: PhD in Mechanical Engineering (2007, Zhejiang University).
Scientific interests: application of optical and electronic engineering, embedded and reconfigurable applications in signal processing and motion control.
Publications: more than 20 papers.

Maoyong Cao, born in November, 1964, Qingdao, China

Current position, grades: professor, director of Academic Affairs Office at Shandong University of Science and Technology.
University studies: PhD in Optical Engineering (2002, Tianjin University).
Scientific interests: measuring and control system, imaging processing, pattern recognition.
Publications: more than 70 scientific papers, 12 patents and 3 teaching books.

Yan Chen, born in August, 1972, Qingdao, China

Current position, grades: PhD student, engineer at Shandong University of Science and Technology.
University studies: M.D. in Control Engineering (2012, Shandong University of Science and Technology).
Scientific interests: measuring and control system.
Publications: more than 10 papers.

Identification of crop weed based on image texture features

Jie Zhang¹, Fuli Song², Jiali Tang¹*

¹*School of Computer Engineering, Jiangsu University of Technology, Changzhou 213001, Jiangsu, China*

²*Personnel Division, Henan Radio & Television University, Zhengzhou 450008, Henan, China*

Abstract

By using computer image processing technology, texture features of weed in the corn seedling field are analysed, and then we present an algorithm combining Support Vector Machine (SVM) to form a classifier and promote an improved method of RBF network. The experimental results show that the proposed method is effective.

Keywords: crop, weed identification, image texture features, support vector machine (SVM)

1 Introduction

We know that weeds cause great impact on the growth and yield of crops. Along with the development of agricultural information, a new method called image-based machine recognition pattern was introduced for weed identification. Meanwhile the new method can realize directed drug spray, protect the environment and save money.

To separate the weeds from other crops precisely, we must be able to extract the characteristics of various types of weeds accurately. Many studies have confirmed that if we extract the colour feature, the shape feature and the texture feature of weeds, then put them together, we can achieve more accurate identification. However, most studies are just a combination of several types of features, so it takes a large amount of computation. To some extent, it affects the efficiency of identification. However the research results of single feature-based identification are less [1, 2].

In this article we will use single feature-based identification mode to only extract texture feature of weeds and then pre-process the images collected by us. On this basis we will use SVM classifier to identify them effectively and combine the RBF kernel function with the low-order polynomial kernel function to achieve optimization. Experimental results show that the proposed method is effective, and worth promotion.

2 Image acquisition and pre-processing

2.1 IMAGE ACQUISITION

In this study weed images will be collected from an experimental field. And we will use Nikon digital camera to capture images of the sample at three different times of sunny day (6:00, 10:00, and 16:00), and then store the original images. The common kinds of weeds include: cephalanoplos, pigweed, crabgrass and filed bindweed.

2.2 IMAGE PRE-PROCESSING

According to the requirement of image processing and analysis, there are three steps of image pre-processing operation: greying, enhancement and denoising and image segmentation.

2.2.1 Greying

There is quite a big difference between crop and weed images with soil images. The original images are of 24 bit true colour bitmap, which contains three components: R(red), G(green) and B(blue). And the green component in plant-growing period is most obvious and it has the largest difference with soil. In the processing of greying the original sample images, extra-green character is introduced. By combining the R, G, B three colour components, the greyscale image is generated. The specific approach [3] is shown in Equation (1).

$$E_G(\text{x,y}) = 2G(x, y) - R(x, y) - B(x, y) . \qquad (1)$$

In the above, $G(x, y)$, $R(x, y)$ and $B(x, y)$ stand for the R, G, B three colour components of the original image, and $E_G(x, y)$ is the extra-green value. Through this algorithm, we can obtain the grey image of the sample processing results.

2.2.2 Enhancement and denoising

In this processing the greyscale images can be improved by using filtering. In this study, neighbour average filtering and median filtering will be compared. There are some studies having shown that the two filtering methods have similar effects and most of the noise can be filtered [4]. In consideration of the smaller amount of calculation of median filtering and large number of image samples to be processed, we choose median filtering method to promote

efficiency. At the same time, it can not only ensure the quality of processing, but also reduce the use of system resources. We divide the images to be processed into different S-windows, and take the pixel grey value in the windows as the grey value of each window. It is expressed in Equation (2).

$$g(x, y) = median\{f(x-i,y-i)\},(i,j) \in S \qquad (2)$$

In the above the grey values of pixels in window are represented as $g(x, y)$, and $median()$ is the median value function. Take cephalanoplos for example, the following are the original image and processed image. Please see Figure 1.
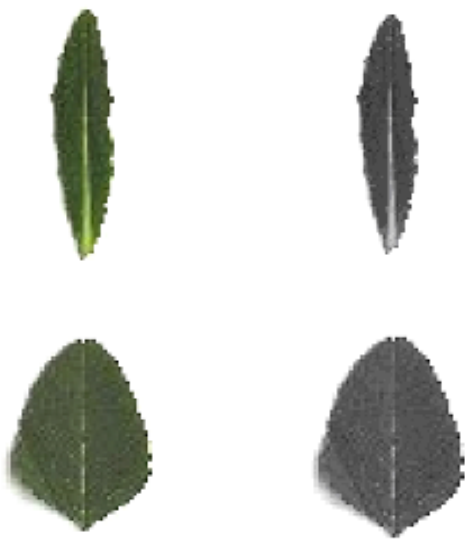
FIGURE 1 Comparison of the original image and processed image

2.2.3 Image segmentation

The image segmentation is implemented with threshold, which is selected with the iterative method [5, 6].

Firstly, traverse the pixel images to be processed to obtain the maximum and minimum values of the grey image, which are represented as T_{max} and T_{min} respectively;

Secondly, initialize the grey threshold value, and take the mid-value of maximum and minimum values, i.e. Mid-value $T0=(T_{max}+T_{min})/2$;

Then, take $T0$ as the threshold value, to divide the images to be processed into foreground and background regions, and get $M1$ the average grey value of foreground, and $M2$ the average grey value of background;

Lastly, combine $M1$ and $M2$ to obtain the new grey threshold value: $Tk+1=(M1+M2)/2$. And compare $Tk+1$ with last grey threshold value. If they are equal, $Tk+1$ is the final threshold value. If not, go on the iteration of step 2. Finally, we can obtain the binary image after segmentation.

3 Extraction of Texture Feature

In micro regularity, texture feature is often difficult to describe in quantitative way or qualitative way, but it has obvious macroscopic rules and the directional distribution regularity. The extracted texture features described herein are:

Average Value: $m = \sum_{i=0}^{L-1} z_i p(z_i)$;

Standard Deviation: $\sigma = \sqrt{\mu_2}$;

Energy: $E(\theta,d) = \sum_{i}^{n} \sum_{j}^{n} M_{(\theta,d)}(i, j)^2$;

Entropy: $H(\theta,d) = -\sum_{i}^{n} \sum_{j}^{n} M(\theta,d)(i, j) \log_2 M_{(\theta,d)}(i, j)$;

Moment of Inertia: $I(\theta,d) = \sum_{i}^{n} \sum_{j}^{n} (i - j)^2 M_{(\theta,d)}(i, j)$;

Dependency: $C(\theta,d) = \dfrac{\sum_{i}^{n} \sum_{j}^{n} i^* j^* M_{(\theta,d)}(i, j) - \mu_x \mu_y}{\sigma_x \sigma_y}$;

Seven invariants of Hu invariant moment are introduced: $\phi 1 — \phi 7$.

Part of the texture feature data is shown in the Table 1.

TABLE 1 Table of texture feature information

Weed species	Cephalanoplos	Pigweed	Crabgrass	Filed bindweed
Average value	91.939	96.547	93.612	112.334
Standard deviation	67.6944	66.6184	64.1488	76.97856
Moment of inertia	7.696	7.464	6.968	8.3616
Energy	0.1392	0.5512	0.4336	0.52032
Dependency	0.0848	0.0144	0.0144	0.01728
Entropy	5.0752	5.2128	5.212	6.2544

Part of the data from invariant moment is shown in the Table 2.

TABLE 2 Table of invariant moment information

Invariant moment	Cephalanoplos	Pigweed	Crabgrass	Filed bindweed
1	2.619	2.470	2.492	2.991
2	2.177	2.275	2.440	2.928
3	5.289	4.528	4.427	5.313
4	1.403	0.944	1.134	1.361
5	5.736	4.287	4.199	5.039
6	3.515	2.989	2.650	3.180
7	4.881	3.692	3.671	4.406

There are too many texture parameters of weeds, and to enhance efficiency, we select the SVM based on RBF kernel function:

Make $U \in R^n, V \in R^n, g \in R^+$, and the RBF kernel function is defined in Equation (3).

$$K(U,V) = \exp(-g \| U - V \|^2) \qquad (3)$$

In the above, R represents a vector space, and g is the parameter of kernel function.

4 Weed identification

4.1 SVM-BASED METHOD

In this study SVM (support vector machine) model is used for recognition. When using SVM into various recognition patterns, it will be faced with the problem of determination of the specific function and method parameters. At present, there is no highly-common and efficient mode, and many studies are based on trial and error method or empirical method to select the parameters. In order to identify the various types of weeds accurately, the Stprtool toolbox in MATLAB will be introduced in this study, and the way of one-to-many will be used.

4.2 DATA STANDARDIZATION

A lot of weed image data in this study should be initialized before being imported into MATLAB. That means we should convert initial data format supported by the Stprtool toolbox of MATLAB. Standard format of various types of image data can be expressed as:

<value1><value2> … < valuen>

The training data and test data of image for SVM are all of $m \times 1$ matrix structure. By extracting the specific values of the training set and test set from weed images, the identification results (expressed as integer) is generated. This study focuses on common four classes among the corn crop weeds, so the target values fall in the interval of [1,4]. In consideration of the six texture features (Average Value, Standard Deviation, Moment of Inertia, Energy, Dependency, Entropy) and seven invariant moment (Invariant Moments 1-7), the value of <value> is determined to be a single-precision positive real number.

4.3 DATA NORMALIZATION

To improve the efficiency of operation, and avoid the impact from a too-large or too-small characteristic parameters, all of the data should be normalized. Another reason for normalization is to enhance the processing efficiency of the system. Due to that, the SVM algorithm involves core computing, and it will call EXP algorithm, only the normalized data will not consume too much system resources in these operations.

4.4 EXPERIMENTAL PROCESS AND RESULTS

There are 20 images of each weed, 80 experimental samples in total. 40 weed images are selected to build SVM test set randomly and the rest 40 images used as test set. 13 characteristics are used as SVM input. In order to

select the best method, the present study compares the RBF kernel function, Polynomial kernel function and Linear kernel function, in which training and testing are conducted respectively.

4.4.1 Several kinds of kernel functions performance analysis

For the 80 image samples, we use RBF, Linear and Polynomial kernel functions to test them respectively. For Polynomial kernel function, the tests of from second order to fifth order are conducted. The Table 3 shows the results.

TABLE 3 Results of several kernel functions

Function Name	Parameter	Support vector	Identification rate
RBF kernel function	Penalty factor=32 σ =0.5	189	90%
Linear kernel function	Penalty factor =1000	178	78.6%
Polynomial kernel function (Second order)	Penalty factor =1000	178	85%
Polynomial kernel function (Third order)	Penalty factor =1000	177	76.3%
Polynomial kernel function (Fourth order)	Penalty factor =1000	174	75%
Polynomial kernel function (Fifth order)	Penalty factor =1000	175	73.8%

The test data shows that the best performance is RBF (radial basis function), and the rate of weed identification is 90%; the next is Polynomial kernel function (second-order), and the identification rate is 85%. The others are under 80%. It shows that the lower-order Polynomial kernel function is more appropriate in weed identification. The reason is that, although the learning ability of the system increases with the order, the complexity of the system also increases exponentially, and the "over learning" situation may happen.

4.4.2 Optimization of the classifier

In the process of SVM identification, each kernel function has their own characteristics and is suitable for different situations. From the perspective of bigger categories, kernel function includes global function and local function. The polynomial kernel function is a standard global function. It has a higher adaptability. Meanwhile, RBF belongs to standard local function, and its influence is limited to the test sample data, which makes its interpolation ability high and better than polynomial function.

In this study, we make use of the advantages of the above two types of kernel functions in an innovative way. They can be combined to achieve better identification results. This thinking also proves that the idea of SVM-based learning performance and generalization performance is great.

Based on the above research, RBF has a better interpolation ability and the Polynomial kernel function

(second order) has a better extrapolating ability. The combined kernel function is constructed as follows. See Equation (4).

$$K(x,y) = \lambda_1 \cdot (xy+1)^q + \lambda_2 \cdot \exp\{-\frac{\|\text{x-y}\|^2}{2\sigma^2}\} \qquad (4)$$

In the above, λ_1 and λ_2 stand for the weight of RBF and Polynomial kernel function (second order) in the combination kernel function, and they should be $\lambda_1 + \lambda_2 = 1$. K(x,y) is kernel function, $x \in R^d$ and $y \in [-1,+1]$ are class identifiers; $(xy+1)^q$ is RBF; $\exp\{-\frac{\|x-y\|^2}{2\sigma^2}\}$ is Polynomial kernel function. Take different λ_1 and λ_2, and the simulation data is shown in the Table 4.

TABLE 4 Simulation data of optimized classifier

λ1	λ2	Identification rate
0.9	0.1	81.25%
0.8	0.2	81.25%
0.7	0.3	80%
0.6	0.4	78.75%
0.5	0.5	78.75%
0.4	0.6	81.25%
0.3	0.7	78.75%
0.2	0.8	92.5%
0.1	0.9	78.75%

The data shows that the identification rate of the system vary greatly with the different λ1 and λ2 values. When λ1=0.2 and λ2=0.8, the combined function generates the highest identification rate 92.5%, higher than the conventional SVM kernel function result. Therefore, the performance of this kind of optimized classifier is better than single kernel function.

5 Conclusions

In this paper, a method of image processing is used to identify texture features of crop weeds. On the basis of traditional SVM algorithm, a combined identification mode is constructed. The experiment confirmed a better recognition result. The improved method helps reduce the amount of herbicide spraying, and realize intelligent management and operation of the weeding system.

Acknowledgments

References

[1] Drăgut L, Tiede D, Levick S R 2010 ESP: a tool to estimate scale parameter for multi group resolution image segmentation of remotely sensed data *International Journal of Geographical Information* **24**(6) 859-71

[2] Han N, Wang K, Yu L, Zhang X 2012 Integration of texture and landscape features into object-based classification for delineating torreya using ikonos imagery *International Journal of Remote Sensing* **33**(7) 2003-33

[3] Jin H, Li P, Cheng T 2012 Land cover classification using CHRIS/PROBA images and multi-temporal texture *International Journal of Remote Sensing* **33**(1) 101-19

[4] Kabir S, He D C, Samusi MA 2010 Texture analysis of IKONOS satellite imagery for urban land use and land cover classification *The Imaging Science Journal* **58** (3) 163-70

[5] Ota T, Mizoue N, Yoshida S 2011 Influence of using texture information in remote sensed data on the accuracy of forest type classification at different levels of spatial resolution *Journal of Forest Research* **16**(6) 432-7

[6] Shen J, Liu LP, Lin X 2011 Cropland extraction from very high spatial resolution satellite imagery by object-based classification using improved mean shift and one-class support vector machines *Sensor Letters* **9**(3) 997-1005

Authors

Jie Zhang, born on March 18, 1983, China

Current position: researcher at Jiangsu University of Technology, China.
University studies: Msc. degree in computer engineering from Yangzhou University, in 2012.
Scientific interest: software development and image processing.
Publications: 10 papers.

Fuli Song, born on July 3, 1982, China

Current position: lecture of the personnel division at Henan Radio & Television University, China.
University studies: Msc. Degree in software engineering from University of Electronic Science and Technology, in 2013.
Scientific interest: software engineering and image processing.

Jiali Tang, born on May 10, 1980, China

Current position: associate professor at Jiangsu University of Technology, China.
University studies: Msc. Degree in Material Processing Engineering from Nanjing University of Aeronautics and Astronautics, in 2005.
Scientific interest: industrial control software development and image super-resolution reconstruction.
Publications: 20 papers.

Rapid detection of bedding boundaries based on borehole images

Xiang Zhang[1]*, Wei Zhang[2], Xiaoling Xiao[1]

[1]*Key Laboratory of Exploration Technologies for Oil and Gas Resources, Ministry of Education, Yangtze University, Wuhan 430010, Hubei, China*

[2]*School of Geophysics and Oil Resources, Yangtze University, Wuhan 430010, Hubei, China*

Abstract

The bedding is an important sedimentary structure phenomenon. The rock bedding structure, the direction of sedimentary transportation and the ancient sedimentary environment analysis can be studied by extracting the bedding boundaries and dips. Electric imaging logging can provide rich information of a borehole wall and circumference, which reflects formation resistivity variations. The bedding boundaries are detected by using the electrical imaging logging data based on an image recognition method in this paper. On an oriented, unwrapped image of a cylindrical borehole, the trace of a planar-bedding boundary appears as a sine wave. The bedding boundaries are detected by the recognition of the sine curves in borehole image. The influence problems of bedding boundary detection caused by fractures and other geological events are solved by statistical analysis technology. Through the techniques of the slope fitting, the speed and accuracy problems of bedding boundary detection are solved, which has good anti-interference performance. The processed results of the theoretical models and the measured borehole images at the varied dip segment indicate that the detected bedding boundaries reflect the real situation, which are identical to those derived by the Autodip.

Keywords: imaging logging, bedding boundary, object detection, image recognition

1 Introduction

Sedimentary bedding is an important sedimentary structure along the vertical evolution of sedimentary structures, through mineral composition, structure, colour or textures, which is a reflection of the fluid flow direction and hydrodynamic mechanisms and biologic activities at the depositional moment. The extracted bedding boundaries and dip trends can study the direction of sedimentary transportation, ancient sedimentary environment and the favourable exploration object [1].

Electrical imaging logging is an advanced logging technology providing a lot of information of a borehole wall and circumference, in which the conductivity differences of various geological characteristics are the physical basis of electrical imaging logging. Electrical intensity measurements reflecting formation resistivity variations are converted into variable-intensity (gray-scale) images. By convention, black is low and white is high resistivity. Some formation boundaries, fractures and structure are identified from electric imaging logging. Borehole images can be quantitatively processed through image processing techniques to extract parameters of the bedding and dips [2]. Currently, there are some electrical imaging logging tools, such as Schlumberger FMI, Atlas StarII and Halliburton EMI, which all have the image mode and the dip mode. For example, FMI have three working modes: the full well wall mode, the four-pad mode and the dip mode [3]. The Full well wall mode can obtain greater coverage and high-resolution images of the borehole. Compared to the full well wall mode, the speed of the four-pad mode measurement is higher, the cost is lower and the coverage is the half of the full well wall mode. Eight electrodes are only used in the dip mode to collect information of borehole formation, in which the effect is the same that in the dip meter.

Accurate identification of bedding boundaries and their dips is very important to the exploration of oil-gas resources, electrical imaging logging is an effective means of extracting the bedding boundaries and their dips. Due to the influence of the heterogeneity of rock structures, fracture and hole structures, sutures, fault and other complex geological background factors, the detected bedding boundaries are often distorted. So, three factors, the interference, the speed and the accuracy, need to be considered in the detection of bedding boundaries.

The bedding boundaries are detected by using the electrical imaging logging data based on an image recognition method in this paper, which is similar to the method proposed by S-J Ye. The algorithm consists of four main steps:

1) determining the direction of the bedding by statistical analysis technology;

2) fitting the sinusoid curves by using the bedding orientations based on the Least-Mean-Squares method;

3) optimizing the location of the sine curves based on the interactive shift technology;

4) defining the reliability criterion of the detected bedding boundaries based on the comprehensive quality index. Through statistical analysis technology, the influence problems of bedding boundary detection caused by fractures and other geological events are solved.

Corresponding author e-mail: zx_jr_xl@163.com

Through the techniques of the slope fitting, the speed and accuracy problems of bedding boundary detection are solved.

2 The previous methods on the bedding dip detection

2.1 THE CORRELATION METHOD

When electric imaging logging is worked in the dip modes, some curves, such as conductivity, azimuth, radius or diameter, are obtained to calculate the formation dip and azimuth. The bedding dips are calculated by synthesizing one curve from multiple conductivity curves, combined with diameter and azimuth curves in the image mode. The morphology of the conductivity curves is similar in the same structure layer, and the formation dip can be calculated on the basis of the depth differences of the same bedding boundary. For example, the bedding dips are calculated by four six-arm bedding dip processing methods: the vector multiply method, the least square method, the clustering method and spherical imaging method, after taking out the dead electrodes and depth shift in Liu Yaowei paper [4]. Liu Yingming analysed the processing procedure of imaging logging data of STAR, and directly calculated the bedding dip using the imaging logging data by adopting the correlation analysis [5].

Different types of the bedding boundaries can fast extracted from the borehole images based on the side by side and pad to pad correlation. However, this method is very difficult to eliminate the affection of complicated geological background interference, such as fractures, and it cannot accurately determine the bedding boundaries, thus affects the precision of the bedding boundaries detection.

2.2 THE IMAGE RECOGNITION METHOD

The bedding boundaries are usually a set of parallel or nearly-parallel conductivity anomaly, and the abnormal width is narrow and uniform, which are often characterized by linear features. The bedding boundaries are continuous, similar and regular among different pads, which are a grey level change between the up and down images. A bedding boundaries perpendicular to the well presents a horizontal line, but a bedding boundaries inclined to the well presents a sine curve in a unfolded image [6].

Compared with fractures, the bedding boundaries in the imaging logging are numerous, therefore, the interactive processing method, in which three or more than three points on a bed boundary are selected to calculate its dip by fitting a sine curve according to the sine parameters, cannot be adopted [7]. Neither the time-consuming Hough transformation nor the Radon transformation method could be adopted to automatically extract dip parameters [8-10]. The fractures boundaries are accurately detected, and fast identified by the image processing analysis technology [11]. Shi Yongqian defined the rock structural bedding dip using 3D geometry mathematics methods

according to the imaging logging, deviation well parameter [12]. Jean Pierre Leduc analysed the bedding boundaries automatically based on the image processing method by FMI data, with the help of the lithology and texture changing features [13]. Zhu Qiangjun et al. derived the bedding boundaries of the core image by the edge detection algorithm based on the fuzzy edge detection [14]. Jean-Noel Antoine et al. detected the bedding boundaries and dips in electric imaging logging by the edge tracing and matching method [15]. Firstly, the gradient field is extracted to get stream lines based on the boundaries tracing method, and derive boundaries from stream lines; then, the bedding boundary traces of the different pads are connected by the matching algorithm. However under the condition of complex background, accurate stream lines and traces of the bedding boundary are hardly gotten by the edge tracing method. T.Quiniou detected the bedding boundaries and derived the dips in the core images by using the edge detection and the adaptive Hough transformation method [16]. Although the Hough transformation method has a good robustness, it is very time-consuming. Thus, this method can be used to detect the relatively small number of fractures, but cannot derive the bedding boundaries in the imaging logging. S.Ye and Naamen Keskes et al. directly derived the bedding boundaries from the original electric imaging on each pad, then performed the edge matching and statistical analysis for the bedding boundaries on each pad [17, 18]. David J.Rossi derived the bedding dips using the edge detection and multi-target tracking method from imaging logging [19]. Compared with the Hough transformation method, these methods have a fast speed. But these methods directly derive the bedding boundaries based on the original electric imaging, which are influenced by the blank areas between each pad, the precision of the dip calculation is affected.

3 The rapid detection method of the bedding boundary

In the image of electrical logging, a plane which inclines with the well is shown as a sinusoidal line. Compared with the fractures, on the one hand, the bedding planes in the borehole image are often a set of parallel or nearly parallel conductivity abnormities, the abnormal widths are narrow, uniform, regular, which occur in group, the characteristics of adjacent bedding planes are similar. On the other hand, the number of bedding planes is large, therefore, the method of fitting the slope of the bedding plane is used to detect the bedding boundaries.

3.1 DETERMINATION OF THE BEDDING DIRECTION

The bedding direction is the local orientation of an image. In a given window, the pixels are gotten at the corresponding direction, and the grey variance is calculated on all straight lines at the same direction on each

pad image. The minimum variance direction in all the straight lines is the bedding direction, which is shown in Figure 1. The calculation of grey variance is defined as follows:

$$Var_\theta = \sum_{i=1}^{M}\sum_{j=1}^{N} \frac{G_{ij} - \overline{G_i}}{M*N} , \qquad (1)$$

where, Var_θ represents the grey variance of all the straight lines in the direction θ ; $\overline{G_i}$ represents the grey mean value of all pixels of the straight line i in the direction θ ; G_{ij} represents the grey value in the straight line i and pixel j in the direction θ .
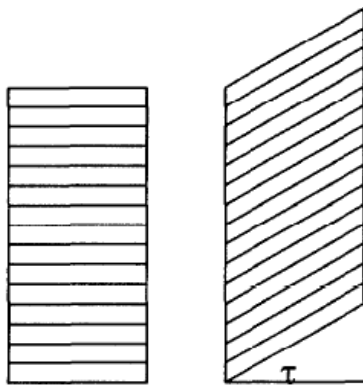
FIGURE 1 Estimation of the predetermined direction

3.2 THE FITTING OFSINUSOID CURVES USING THE BEDDING ORIENTATIONS

Considered the influence of fractures, caves and other electrical heterogeneity factors, it is hard to detect the trajectory coordinates of the bedding boundaries, therefore, the effective way of automatic calculation of the sine curves is the fitting of the sine curves based on the slope of the sine curves, not based on the trajectory coordinates of the sine curves. The tangents of the sinusoid can be approximated by the tangents of the bedding orientations. The bedding sine curves are then fitted by the tangents of slope.

In the four parameters of the standard sinusoid curve, two parameters are already determined, respectively, the angular frequency $\omega = \dfrac{2\pi}{T}$ and the offsets y_o , which is the height of the center line of the local window. Therefore, the fitting of the four-parameter sine curves is in fact that of the two parameters A and ϕ , the standard sinusoid curve is defined as follows:

$$y = ASin(\omega x + \phi) + y_o . \qquad (2)$$

Taking the derivative of the Equation (2), the slope is obtained as following:

$$T = \frac{dy}{dx} = A\omega \cos(\omega x + \phi) . \qquad (3)$$

Two parameters A and ϕ , are fitted by the Least-Mean-Squares method.

3.3 THE LOCATION OPTIMIZATION OF THE SINE CURVES BASED ON THE INTERACTIVE SHIFT TECHNOLOGY

In the process of defining the sine curve, because the local direction is the trend direction of the bedding, the sine curve based on the fitting of the curve slope is located in the center of the given window, not in the bedding boundaries. There is a mismatch deviation between the sine curve and the bedding boundary, but the sine curve can be shifted to the bedding boundaries by searching the maximum image grayscale contrast. In a general way, the extreme position of the differential curve is corresponding to the bedding boundary. Therefore, the differential curve is gotten by the filtering processing of the gray value in the vertical direction, then, the detected sine curve is moved to the extreme value position of the differential curve

3.4 THE RELIABILITY CRITERION OF THE DETECTED BEDDING BOUNDARIES BASED ON THE COMPREHENSIVE QUALITY INDEX

Considered the influence problem of some geological events in the bedding boundaries detection, the reliability measure of the detected sine curves is a very important problem.

A comprehensive confidence measurable parameter Q, which includes the sine fitting error, the taken pads number and the contrast of sine curve, is defined as following:

$$Q = Q_C \times Q_E \times Q_N , \qquad (4)$$

where, Q is the comprehensive quality index, which reflects the reliability of the detected sine curve; Q_C is the contrast of the sine curve, which the sum of first-order differential value along the sine curve; Q_E is the fitting error of the least mean squares; Q_N is the taken pads number in the fitting of the sine curve.

4 The processing results analysis

4.1 ANALYSIS ON THE THEORETICAL MODEL

In order to evaluate the effectiveness of the bedding boundaries detection method, a kind of formation theoretical model is designed, which includes four types of dip vector modes: the green mode (the dips do not change

with the depth), the red mode (the dips increase with the increasing depth), the blue mode (the dips decrease with the increasing depth) and the Chaotic mode (the dips change irregularly). The theoretical model is shown in Figure 2, and the red curves in Figure 3 are the detected bedding boundaries using the method described above. Compared Figure 2 with Figure 3, it can be seen that other detected boundaries are consistent with that in the designed model except that the first (top) detected bedding boundary is affected by the blank formation.
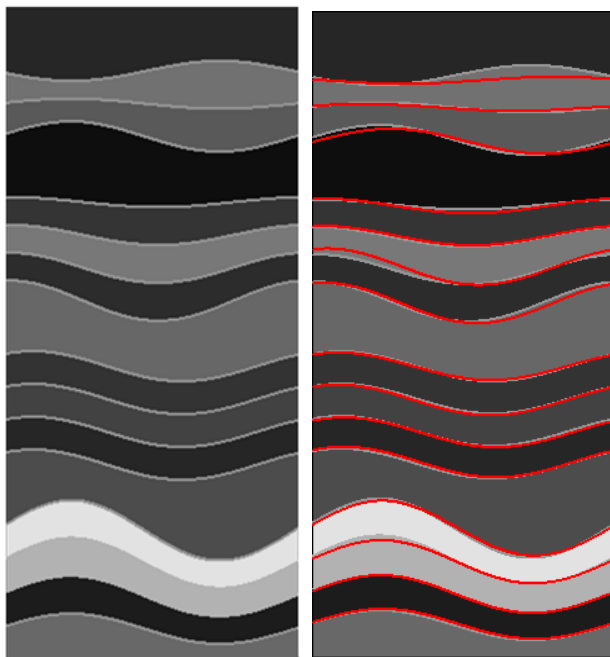
FIGURE 2 The theoretical formation model

FIGURE 3 The detected formation boundaries

4.2 ANALYSIS ON THE REAL BEDDING BOUNDARIES

In order to verify the validity and practicability of the above algorithm, the bedding boundaries detection are processed in borehole images in the following situations, which results are compared with that obtained by the AutoDip.

4.2.1 Analysis at different window lengths

Figure 4a, 4b and 4c are the detected bedding boundaries at different window lengths of 90, 120, and 150, respectively. From Figure 4, it can be seen that there are similar results in above three window lengths. That is to say, the detection results will be not affected by the window length.

a) Window Length=90

b) Window length=120
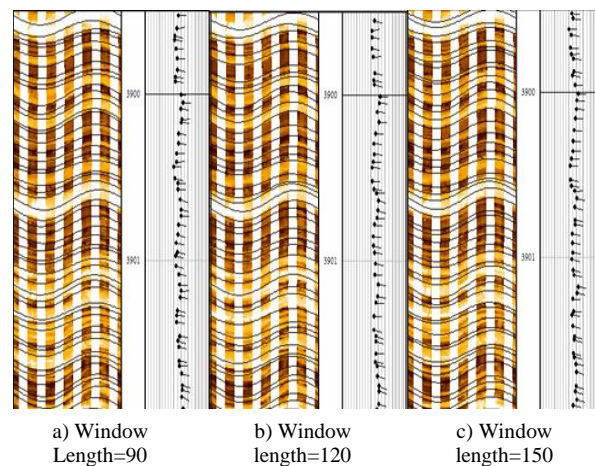
c) Window length=150

FIGURE 4 Results at different window length

4.2.2 Analysis at the varied dip segment

Figure 5 shows the processed results at the varied dip segment. There are nearly 5 dip modes in this segment, which are the blue mode, the red mode, the green mode, the blue mode and the green mode in turn from the top to the bottom. The detected results of this algorithm are shown in the left of Figure 5, the results of the AutoDip are in the right. It can be seen from Figure 5, the dip modes basically reflect the real situation, which are consistent with those obtained by the AutoDip.
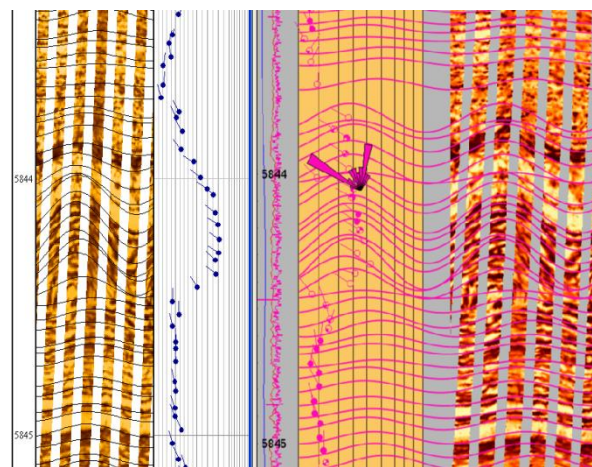
FIGURE 5 Comparison of the processed results at the varied dip segment

5 Conclusions

1) The bedding is an important sedimentary structure phenomenon in sedimentary rocks. The rock bedding structure, the direction of sedimentary transportation and the ancient sedimentary environment analysis can be studied by extracting the bedding boundaries and dips, which would be helpful in searching for favorable exploration direction.

2) Electric imaging logging is one of the advanced logging technique, which can provides rich information of a borehole wall and circumference, not only can

Zhang Xiang, Zhang Wei, Xiao Xiaoling

qualitatively identifies the various boundaries, fractures and structural configurations from the borehole images, but can derive the bedding boundaries and dips by quantitatively image processing and analysis using image processing technology.

3) The bedding boundaries are detected by using the electrical imaging logging data based on an image recognition method. The influence problems of bedding boundary detection caused by fractures and other geological events are solved by the statistical analysis technology. Through the slope fitting technique, the speed and accuracy of the bedding boundary detection method are solved, which has good anti-interference performance. The processed results of the theoretical models and the

measured borehole images at the varied dip segment indicate that the detected bedding boundaries reflect the real situation, which are identical to those derived by the Autodip.

Acknowledgements

References

[1] Chen Q, Wang W, Ge H 2007 Image well logging technology and its advance *World Well Logging Technology* **22**(3) 8-11
[2] Fu J W, Xiao L ZH, Zhang Y ZH 2004 Status and developing trend of acoustical and electrical imaging well logging tools *Progress in Geophysics* **19**(4) 730-8
[3] Zhang Sh Q, Gu Ch X, Cao G H 1997 The imaging logging technology and application *Petroleum Industry Press Beijing (in Chinese)*
[4] Liu Y W 2007 Electrical image log formation dip processing method study and software developing *Master's Degree Thesis China University Of Petroleum (in Chinese)*
[5] Liu Y M 2008 Micoscanner logging datum processing and interpretation method research *Master's Degree Thesis China University of Geosciences*
[6] Tian F, Zhang CH G 2010 Fracture identification of electric imaging logging data and its application *Journal of Engineering Geophysics* **7**(6) 723-7 (*in Chinese*)
[7] Lai F Q 2011 Micro-resistivity image logging processing and interpretation methods research *Doctor's Degree Thesis China University of Petroleum (in Chinese)*
[8] Kherroubi J 2008 Automatic extraction of natural fracture traces from borehole images *Proc. Conf. on Pattern Recognition Tampa Florida USA* 1-4
[9] Zhang X, Xiao X L 2009 Detection of Fractures in Borehole Image *Proc. Conf. on Multispectral Image Processing and Pattern Recognition Yichang China* 7495 11-6
[10] Van Ginkel M, Van Vliet LJ, Verbeek P W 2001 Robust curve detection using a radon transform in orientation space applied to

fracture detection in borehole images *Proc. Conf. on the Advanced School for Computing and Imaging Heijen Netherlands* 84-91
[11] Xiao X L, Zhang X 2010 Detection of fractures in borehole image based on orientation filtering *Journal of Yangtze University***7**(4) 69-72
[12] Shi Y Q 2010 The determ in action of the attitude of rock structural plane based on well-logging maging data *Geophysical & Geochem Ical Exploration* **34**(3) 358-61
[13] Leduc J P, Vincent D P 2002 FMI based sedimentary facies modelling *Proc. Conf. on CSPG Annual Convention Calgary Alberta Canada* 1-10
[14] Zhu Q J, Wu X H, Teng Q-ZH 2008 Feature extraction of core image based on fuzzy algorithm *Electronic Measurement Technology* **31**(3) 23-9
[15] Antoine J N 1993 A method to derive dips from bedding boundaries in borehole images *SPE Formation Evaluation* 96-102
[16] Quiniou T, Selmaoui N, Magoni C L, Allenbach M 2007 Calculation of bedding angles inclination from drill core digital images *Proc. Conf. on Machine Vision Applications Tokyo JAPAN* 252-5
[17] Ye S J, Rabiller P, Keskes N 1997 Automatic high resolution sedimentary dip detection on borehole imagery *Proc Conf on 38th Annual Logging Symposium* 1-13
[18] Keskes N, Ye S J 2000 Method for automatically determining stratification beds in a site *United States Patent (6125203)*
[19] Rossi D J, Neuilly sur, et al. 2001 Method and apparatus using multi-target tracking to analyze borehole images and produce sets of tracks and dip data *United States Patent (6226595)*

Authors

Xiang Zhang, born on July 14, 1969, China

Current position, grades: a professor at Yangtzeu University, China.
University studies: Ph.D. degree in pattern recognition and artificial intelligence from Huazhong University of Science and Technology, in 2001.
Scientific interest: reservoir evaluation, electricity imaging logging and image processing.

Wei Zhang, born on Octobrer 2, 1990, China

Current position, grades: current master students in Yangtze University of earth exploration and information technology, China.
University studies: Bachelor degree in exploration technology and engineering from Yangtze University, in 2012.
Scientific interest: reservoir evaluation and electricity imaging logging.

Xiaoling Xiao, born on April 12, 1973, China

Current position, grades: a professor at Yangtzeu University, China.
University studies: Ph.D. degree in computer application and technology from Wuhan University of Technology, in 2008.
Scientific interest: information processing and well log analysis.

Moving vehicle detection algorithm based on motion edge extractor

Xiaowei Hu, Jiexin Wang, Xuemiao Xu*, Biao Zhou

School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, Guangdong, China

Abstract

In this letter, we propose a moving vehicle detection method based on motion edge extractor (MEE). In the course of the vehicle detection, the motion and contour information are prominent, so we combine them together to extract the moving objects with complete contours. We first modify the Gaussian Mixture Model (GMM) to estimate the background more precisely. Then an object extraction method in static image is proposed. The original image and the background image go through the object extraction method and a series of logical operations to get the moving areas. At last we apply a simple filling method to refine the result and accurately extract the vehicle areas. The experiment result shows that our algorithm is not sensitive to illumination and can detect the vehicles with similar colour to the road robustly.

Keywords: vehicle detection, motion edge extractor, video monitoring

1 Introduction

With the rapid development of technology, the Intelligent Transportation System (ITS) becomes increasingly important, and the vehicle detection in image sequence becomes a hot subject of research. The traditional object detection methods have been studied for a long time. The frame differencing method is simple but has a high false detection rate. The optical flow method is able to deal with the movement of the camera but requires special hardware and large computational complexity. The background subtraction method is able to get the moving objects but sensitive to the sudden change of movement and illumination.

Among the traditional methods, the Gaussian Mixture Model method [1] is widely used to extract the foreground. This method works when the camera is fixed and it is robust to slight disturbance of the camera. But the GMM method has low convergence speed, which will affects the accuracy of the background modelling. It is also sensitive to global change. Many researchers have tried to modify the GMM method [2-4], Zoran [4] combines the GMM and frame differencing method together to solve the problem of multi-modal distribution of pixels and overcomes the influence of the illumination changes and other factors such as shaking branches. However, due to the low update rate, an object moving slowly will be considered as background and produce "ghosting" in the detection result. For the object with similar colour to the background, all of the above methods cannot extract the complete objects.

Recently some new methods have been proposed. Ye Li [5] presents a vehicle detection algorithm based on template matching which can effactually detect the cars in congested traffic, but this method is slow due to the matching procedure. Mithun [6] applies a method based on multiple virtual detection lines (MVDL) and time-spatial images (TSI) to detect the vehicles, this method is fast and accurate, but it is dependent on the placement of the camera and it will fail if a vehicle changes its lane. Bingfei Wu [7] applies different kinds of background modelling algorithm according to the traffic condition. And they apply sobel detector to optimize the detection result to get accurate result in the heavy traffic. This method relies on the placement of the camera and does not work fine when it is dim. Vargas [8] tries a Sigma-Delta background estimation method based on confidence measurement to sense the traffic condition and adjust the update rate of the model. This algorithm is efficient and is able to deal with the congested traffic. But when there are vehicles with similar colour to the road, the algorithm cannot extract the vehicle correctly. Jie Zhou [9] applies prior knowledge and SVM classifier to modify the moving objects detection. The algorithm is robust to different weather and illumination conditions, but for each image sequence, the user has to set many parameters manually. Zezhi Chen [10] proposes a self-adapted Gaussian Mixture Model to extract the moving vehicles. The level set method is applied to refine the result. This algorithm works well in different weather and illumination conditions but it can only deal with the vehicles heading for the same direction. Ye Li [11] presents a detection algorithm based on And-Or Graph to detect vehicles in congested traffic. But this algorithm requires each kind of vehicle to be trained and it cannot detect the side-view vehicles.

The contour and movement information are the most prominent features of the vehicle in the video. Considering

---

**Corresponding author* e-mail: xuemx@scut.edu.cn

only the movement of the vehicle, current algorithms require different parameters according to different conditions and they cannot get the complete contours of the moving vehicles. If we only consider the contour, it is hard to distinguish between the vehicles and other static objects. So we combine these two features together and propose a novel moving vehicle detection method based on motion edge extractor (MEE). We first modify the Gaussian Mixture Model (GMM) to model the background more precisely. And SVM is applied to cope with the "ghosting" problem. Then we find the contour profile of the connected area from the edge graph extracted by the phase congruency method. And we expand the internal and external contour to get the complete target area based on static image. Both of the original image and the background image go through the complete object extraction method and a series of logical operation to get the moving areas. We then eliminate the shadow in the YCrCb colour space [15]. At last we apply median filter and several steps of morphological operations to extract the accurate and complete moving objects. Our algorithm is able to cope with the vehicles that have similar colour to the background and it is not sensitive to illumination.

2 The proposed motion edge extractor algorithm

The Motion edge extractor (MEE) algorithm includes four parts: the improvement of the GMM, the complete object extraction from static image, the moving area extraction and patching the undetected area. Please see Figure 1.
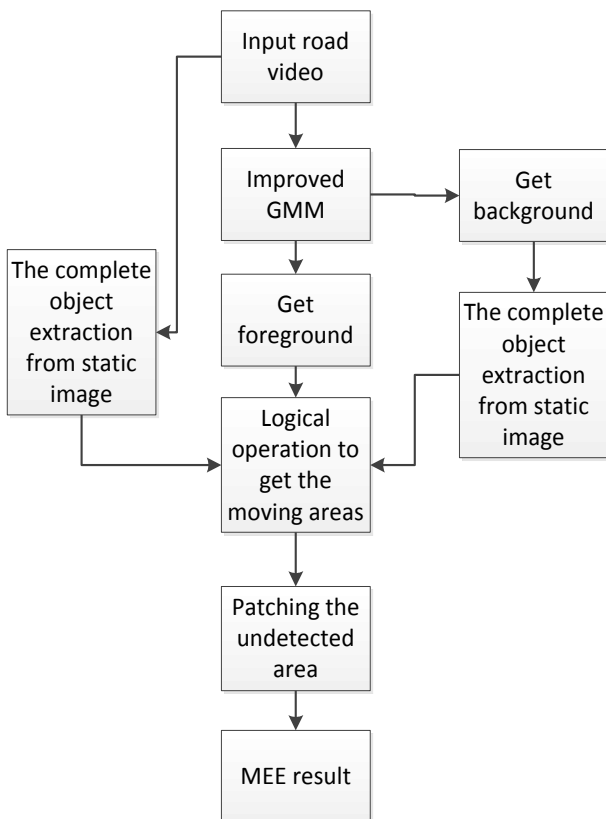
FIGURE 1 MEE flowchart

2.1 IMPROVED GAUSSIAN MIXTYRE MODEL TO GET MORE ACCURATE BACKGROUND

The traditional GMM method uses K Gaussian distributions to model each pixel. For a new frame, the matching with the K Gaussian distributions is done pixel by pixel, and the parameters of the first matched distribution are updated as [1].

$$\Omega_{n,t+1} = (1-\alpha) \cdot \omega_{n,t} + \alpha , \qquad (1)$$

$$\mu_{n,t+1} = \left(1 - \frac{\alpha}{\omega_{n,t+1}}\right) \cdot \mu_{n,t} + \frac{\alpha}{\omega_{n,t+1}} \cdot X_{t+1} , \qquad (2)$$

$$\sigma_{n,t+1}^2 = \left(1 - \frac{\alpha}{\omega_{n,t+1}}\right) \cdot \sigma_{n,t}^2 + \frac{\alpha}{\omega_{n,t+1}} \cdot \beta , \qquad (3)$$

$$\beta = \left(X_{t+1} - \mu_{n,t+1}\right)^{\mathrm{T}} \cdot \left(X_{t+1} - \mu_{n,t+1}\right)^{\mathrm{T}} . \qquad (4)$$

Parameters of the unmatched distribution are updated as Equation (5).

$$\omega_{n,t+1} = (1-\alpha) \cdot \omega_{n,t} , \qquad (5)$$

$\omega_{i,t}$ denotes the weight of the *nth* Gaussian model at time *t*, and it satisfies $0 \le \omega_{i,t} \le 1$, $\sum_{i=1}^{K} \omega_{i,t} = 1$. $\mu_{i,t}$ and $\sum_{i,t}$ denotes the median vector and covariance matrix of the ith Gaussian model at time *t* respectively, we have $\sum_{i,t} = \mu_{i,t}^2 I$.

I is an identity matrix and α is the update rate defined by the user, it satisfies $0 \le \alpha \le 1$. The update rate increases when α gets larger. Since the Gaussian Mixture Model updates the whole image, if the update rate is too small, it will produce "ghosting" and it will become sensitive to the change of illumination and the shake of camera. If the update rate is too large, the slowly moving vehicles will be treated as the background.

In order to solve this problem, we introduce an update parameter N and the update rate becomes Equation (6) and Equation (7).

$$\alpha = \alpha \cdot N , \qquad (6)$$

$$N = \begin{cases} 1 & \text{when the pixel is background} \\ 0 & \text{when the pixel is foreground} \end{cases} . \qquad (7)$$

From the above equations we know that according to the detection result of the previous frame, the background update rate in the moving region is set to zero. So the algorithm is able to detect the slowly moving vehicles when the update rate is very large.

a. Input Image　　b. Extracted Background　c. Extracted Background
Based On Tradition GMM Based On Improved GMM
FIGURE 2 Improved GMM

The method in Figure 2c is able get a more precise background image. But when there are vehicles in the first frame of the video, the GMM will set the whole image with the vehicles as the background. It will have large impact on the moving area detection and lead to failure. So when we deal with the first frame of the video, we apply HOG and support vector machine to scan the frame [12]. The detector windows tile with a grid of overlapping blocks in which HOG feature vectors are extracted. And it scans across the image at all positions and scales. The result is noted by *S*. For the background, $S = 1$, and for the vehicle, $S = 0$. Since the gradient directions of the road and vehicle are quite different, the result is very accurate. We use SVM to classify one hundred images and the road detection rate is up to 100%, the vehicle detection rate is 93%, just only a few trees are taken as vehicles. Figure 3 shows some of the SVM training set.

Positive class

Negative class
FIGURE 3 SVM training set

We modify the Equation (6) according to the result of SVM.

$$\alpha = \begin{cases} \beta & S = 1\ \&\ \text{frame} < 50 \\ \alpha & S = 0\ \&\ \text{frame} < 50 \\ \alpha \cdot N & \text{others} \end{cases} \quad (8)$$

In the first frame where there are vehicles, the first 50 frames update rate is set to β ($\beta \gg \alpha$) while the update rate of the rest keeps unchanged. With a high update rate we can eliminate the "ghosting" fast.

Figure 4 shows the background extracted at the 20th frame. Figure 4a and 4b denote the traditional GMM's result and the modified GMM's result respectively. The result shows that applying SVM to modify the update rate of the GMM is able to cope with the "ghosting" problem.

a Result Before Improvement　　b Result After Changing Update Rate
According to SVM
FIGURE 4 Comparison figure of the results for the modified GMM and traditional GMM

2.2 COMPLETE OBJECT EXTRACTION FROM STATIC IMAGE

The phase congruency (PC) uses the most consistent phase in Fourier components as the feature point. It fits the human visual system and it is robust to the change of illumination [13].

Phase congruency does not make any assumptions of the waveform. It finds feature points according to the consistency of the phase. Peter Kovesi [14] developed a modified measure consisting of the cosine minus the magnitude of the sine of the phase deviation and it incorporates noise compensation. The modified function will provide more local responses. Please see Equation (9) and Equation (10).

$$PC(x) = \frac{\sum_n W(x) \lfloor A_n(x)(\cos\delta - |\sin\delta|) - T \rfloor}{\sum_n A_n(x) + \varepsilon}. \quad (9)$$

$$\delta = \Phi_n(x) - \overline{\Phi}(x). \quad (10)$$

A_n denotes the amplitude of the scale n, $\Phi_n(x)$ is the phase value of the nth Fourier component at x, The weighted mean phase angle is given by $\overline{\Phi}(x)$. The term $W(x)$ is a factor that weights for frequency spread. A small constant ε is incorporated to avoid division by zero. Only energy values that exceed T, the estimated noise influence, are counted in the result. The symbols $\lfloor \ \rfloor$ denote that the enclosed quantity is equal to itself when its value is positive, and zero otherwise.

Then combine phase congruency information over many orientations by using Gabor wavelets [8]. We can obtain the edge information of the image. Please see Figure 5.
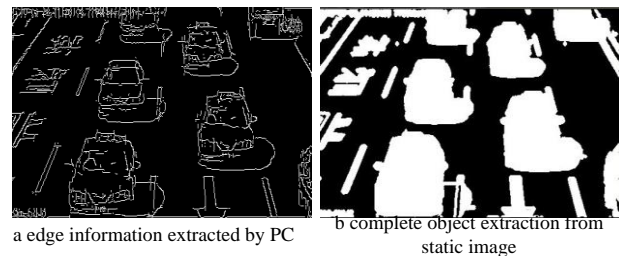
a edge information extracted by PC　　b complete object extraction from static image
FIGURE 5 The results of complete object extraction from static image

Hu Xiaowei, Wang Jiexin, Xu Xuemiao, Zhou Biao

Due to the discontinuity of the vehicle contour extracted by PC, it is hard to calculate the final motion area. We scan the image row by row to find the first unmarked point and mark it, then check if its eight neighbour points are connected and unmarked, if a neighbour is unmarked and forms a connected area with it, mark that point as seed point. This process is done recursively until the whole connected area is marked. Using this method we can mark all the connected area in the image.

Next the inner and outer contours of the connected areas were redrawn, the dilation operation is done. Thus the connected areas with continuous contours are obtained. Please see Figure 6. At last, all the connected area goes through the connecting and filling operation to produce the complete area. Please see Figure 5b.

FIGURE 6 Sketch map of discontinuous contour connection

This method is able to extract complete object area from a static image and it is robust to illumination change.

2.3 MOVING AREA EXTRACTION

Since the complete object extraction algorithm is able to extract complete object area and it's robust to the illumination and contrast change, we combine it with the GMM to effectual extract the motion area. The original frame (Figure 2a) and the background produced by the improved GMM (Figure 2c) go through step B. Then we combine the result and the foreground extracted by GMM to go through a series of logical operations to obtain the complete motion area. Please see Figure 7.

We apply the method in step B to process the original frame and the background image and obtain two images as Figure 8a and Figure 8b.
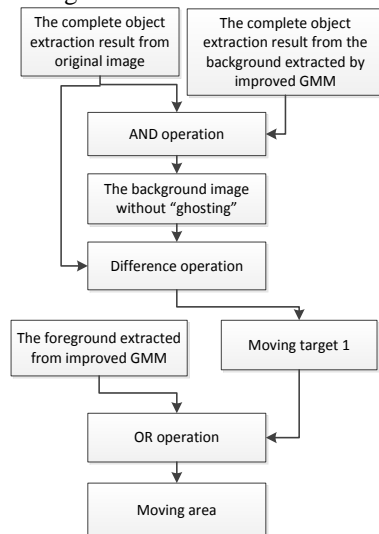
FIGURE 7 Flowchart of the motion area extraction algorithm

Let $f_1(x,y)$ denote the original frame processed by step B, $f_2(x,y)$ denote the background image processed by step B. The final motion area $f_3(x,y)$ can be calculated as follow.

$$f_3(x,y) = f_1(x,y) - \left[f_1(x,y) \cap f_2(x,y) \right] \qquad (11)$$

$f_3(x,y)$ is the moving object areas showed in Figure 8c. In order to avoid the influence of the "ghosting" in $f_2(x,y)$, we do not just apply difference operation between $f_1(x,y)$ and $f_2(x,y)$. If there exists "ghosting" in background image, part of the foreground in $f_1(x,y)$ will be mistaken as the background. We apply difference operation between $f_1(x,y)$ and the sum of $f_1(x,y)$ and $f_2(x,y)$ to improve the accuracy of detection.

Since there is an overlap between the pavement line and the vehicle in the image, the result of the difference operation may have some incomplete areas. Since the foreground image extracted by the improved GMM will not be influenced by the pavement marking line, we apply OR operation to the foreground from improved GMM (denoted as $f_4(x,y)$) and $f_3(x,y)$ to obtain the more complete moving areas denoted as $f_5(x,y)$.

$$f_5(x,y) = f_3(x,y) \cup f_4(x,y) \qquad (12)$$

The final result is shown in Figure 8d.

a The foreground image after step B

b The background image after step B

c The foreground after several logic operations

d OR operation between c and the foreground of GMM
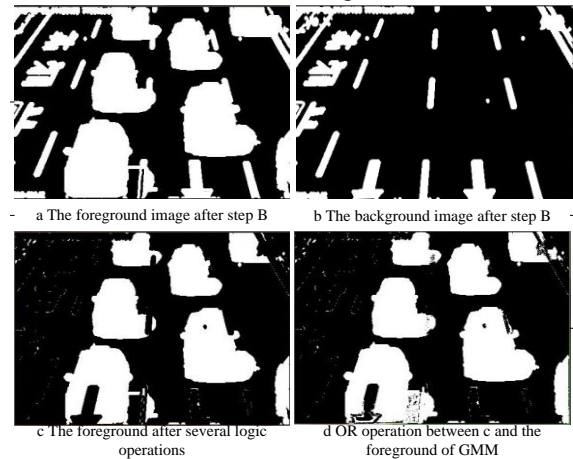
FIGURE 8 The extraction of the moving area

2.4 PATCHING THE UNDETECTED AREA

Figure 8c, 8d show that when the moving target overlaps the marking line, the final moving areas extracted will have some holes similar to the marking line. Although the OR operation between the moving object and the foreground of GMM will patch part of them, we still have to patch the undetected areas since it's hard to get full object from GMM model. So we should do the patching the undetected area operation.

Undetected region patching algorithm is as follows.

Firstly, we traverse the image with a rectangular window (10·8 in our experiment), if more than 35% of the pixels in the window are valid, go to next step, else keep traversal. In the process of moving the window, there is overlap between two consecutive windows to help patch the holes. Please see Figure 9a.

Secondly, divide that region into four parts, and for each part we find 2 outer points. Then connect these eight points to obtain a closed contour and set the area within the contour as valid. It can guarantee not extend the contour. Please see Figure 9b.

a The foreground image after step B

b Find and fill the contour
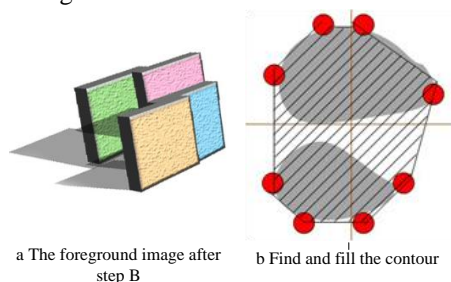
FIGURE 9 Patching the undetected area sketch map

This method is able to not only patch the undetected area but also keep the shape of the area. After patching the undetected area, the moving area (Figure 8d) is shown in Figure 10.

FIGURE 10 Result of the patching

3 Post processing

There exist shadows in the video due to the illumination, the shadows move with the vehicles and they are detected as the moving object too. The shadows affect the accuracy of the vehicle detection.

In YCrCb colour space, luminance and chrominance signals are independent. The YCrCb colour space is robust to brightly light backgrounds, Kumar [15] also confirms that YCrCb is the best colour space for shadow detection, so in this paper YCrCb colour space is used to eliminate the shadows.

After eliminating the shadows, we apply median filter and several morphological operations to obtain the accurate and complete moving objects.

FIGURE 11 The accurate and complete moving objects

4 Experiment results

The performance of the proposed method will be demonstrated in this section. Experiments are conducted on a computer with Intel(R) Core(TM)i5-2450M CPU @ 2.50 GHz 4.00GB RAM. The coding is finished in Microsoft visual studio 2010 and opencv2.3.1.

In the experiments, the image sequences are extracted from different videos under the different conditions.

We test the algorithm in a cloudy weather (the first row of Figure 12), in situations where there are many dark-colour cars (the second row of Figure 12) and in situations where there are cars with similar colour to the road (the last row of Figure 12). The experiment result shows that under these three kinds of conditions, the MEE algorithm is able to extract the complete and accurate moving vehicles.
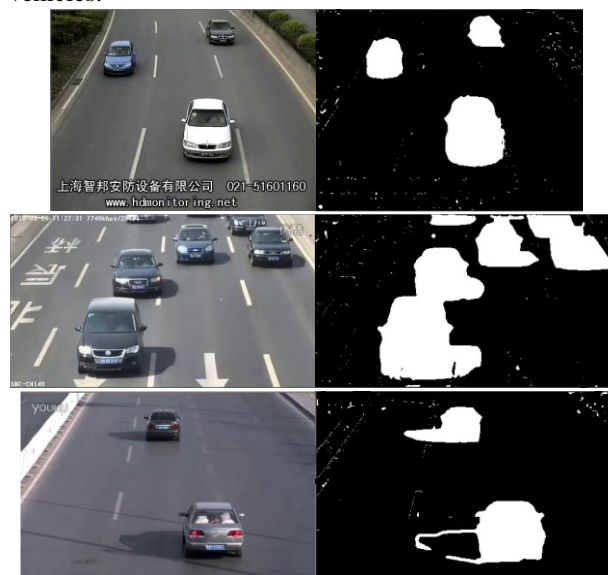
FIGURE 12 Vehicle detection results of the MEE algorithm proposed in this paper

a video image

b result image of our algorithm

c result image one of [16] algorithm

d result image two of [16] algorithm
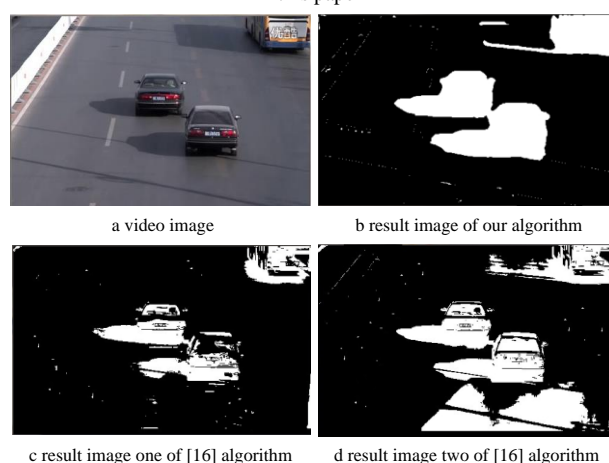
FIGURE 13 Some comparison results of moving vehicle detection in videos where there exist "ghosting" and dark-colour cars

In Figure 13, we compared our method with the method in [16]. Li [16] uses frame difference method and GMM to extract the foreground image. This method can model the

background fast and deal with the "ghosting". Please see Figure 13c. However this method still heavily relies on the parameters. When the background matching threshold is large, it is able to eliminate the "ghosting" quickly but it cannot correctly extract the vehicles with similar colour to the road. When the background matching threshold is small, it can extract the vehicles correctly, but it introduces in noises at the same time. When the "ghosting" exists, the disadvantage is more clearly. Please see Figure 13d. Our algorithm makes full use of the edge and motion information of the vehicles. It can cope with the "ghosting" and the cars with similar colour to the background. Please see Figure 13b.

At the start of the testing video, the illumination condition is good, but after about fifty frames, the illumination condition starts to get dark. Wu [7] applies different kinds of background modelling algorithm according to the traffic condition and uses sobel detector to optimize the detection result to get accurate result in the heavy traffic. Vargas [8] tries a Sigma-Delta background estimation method based on confidence measurement to sense the traffic condition and adjust the update rate of the model. This algorithm is efficient and is able to deal with the congested traffic. The above two algorithms cannot extract complete contours of the vehicles which have similar colour with the road when the illumination condition changes. Our algorithm makes full use of the edge and motion information to extract the complete vehicle contours and it's robust to Illumination change. Please see Figure 14.

a video image
b result image of our algorithm

c result image of [7] algorithm
d result image of [8] algorithm
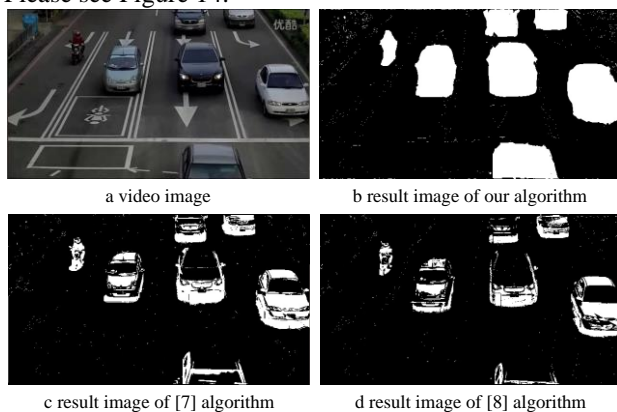
FIGURE 14 Some comparison results of moving vehicle detection in real traffic video captured

5 Conclusions

In this paper motion edge extractor algorithm has been proposed. In this algorithm we have stimulated the human visual system and detected the moving objects by the movement of the objects and their edge information. The algorithm has been divided into four steps: First, we have improved the GMM to obtain a more accurate background. Second, we have expanded the object contour extracted by phase congruency strategy to obtain the complete targets. Then we have extracted the moving area using the background information. Finally, the missed areas have been repaired. Our algorithm has made use of the edge and motion information of the vehicles to accurately extract the vehicles. The experiment result has shown that the proposed algorithm is able to cope with the "ghosting" and the cars with similar colour to the background and it is robust to the illumination change.

References

[1] Stauffer C, Grimson W E L 1999 Adaptive background mixture models for real-time tracking *IEEE Computer Society Conference on Computer Vision and Pattern Recognition F. Collins CO USA* 2246-52

[2] Shimada A, Arita D, Taniguchi R-i 2006 Dynamic control of adaptive mixture-of-gaussians background model *IEEE International Conference on Video and Signal Based Surveillance Sydney Australia*

[3] Hasan B A S, Gan J Q 2009 Unsupervised adaptive GMM for BCI *4th International IEEE/EMBS Conference on Neural Engineering* 295-8

[4] Zivkovic Z 2004 Improved adaptive gaussian mixture model for background subtraction *Proceedings of the 17th International Conference on Pattern Recognition (ICPR)* 28-31

[5] Li Y, Tian B, Li B, Xiong G, Zhu F, Wang K 2013 Vehicle detection with a part-based model for complex traffic conditions *IEEE International Conference on Vehicular Electronics and Safety (ICVES) Dongguan Guangdong Province China* 110-3

[6] Mithun N C, Rashid N U, Rahman S M M 2012 *IEEE Transactions on Intelligent Transportation Systems* **13**(3) 1215-25

[7] Wu B-F, Kao C-C, Juang J-H 2013 *IEEE Transactions on Intelligent Transportation Systems* **14**(1) 485-91

[8]   Vargas M, Milla J M, Toral S L, Barrero F 2010 *IEEE Transactions on Vehicular Technology* **59**(8) 3694-709

[9]   Zhou J, Gao D, Zhang D 2007 *IEEE Transactions on Vehicular Technology* **56**(1) 51-9

[10] Chen Z, Ellis T, Velastin S A 2012 Vehicle detection, tracking and classification in urban traffic *15th International IEEE Conference on Intelligent Transportation Systems (ITSC) Anchorage Alaska USA* 951-6

[11] Li Y, Li B, Tian B, Yao Q 2013 *IEEE Transactions on Intelligent Transportation Systems* **14**(2) 984-93

[12] Dalal N, Triggs B 2005 Histograms of oriented gradients for human detection *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR) San Diego CA USA* 886-93

[13] Kovesi P 1999 Phase congruency detects corners and edges *The Australian Pattern Recognition Society Conference Sydney Australia* 309-18

[14] Kovesi P, Robotics and Vision Research Group 1995 Image features from phase congruency *The University of Western Australia Technical Report Australia* 1-30

[15] Kumar P, Sengupta K, Lee A 2002 A comparative study of different color spaces for foreground and shadow detection for traffic monitoring system *The IEEE 5th International Conference on Intelligent Transportation Systems Singapore* 100-5

[16] Yao L, Ling M 2014 An improved mixture-of-gaussians background model with frame difference and blob tracking in video stream *The Scientific World Journal* **2014**(2014)

Authors
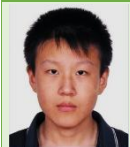
Xiaowei Hu, born on December 19, 1993, China

Current position: an undergraduate student in the South China University of Technology.
University studies: School of Computer Science and Engineering, South China University of Technology.
Scientific interest: the intelligent transportation based on image analysis and machine learning.

Jiexin Wang, born on February 23, 1994, China

Current position: an undergraduate student in the South China University of Technology.
University studies: School of Computer Science and Engineering, South China University of Technology.
Scientific interest: image processing and artificial intelligence.

Xuemiao Xu, born on December 20, 1979, China

Current position: an associate professor in the South China University of Technology.
Scientific interest: the image similarity measurement, intelligent transportation based on image analysis, digital Manga/Cartoon, biometric recognition.

Biao Zhou, born on July 8, 1992, China

Current position: a master student in the South China University of Technology.
Scientific interest: the intelligent transportation based on image analysis and machine learning.

A resource schedule method for cloud computing based on chaos particle swarm optimization algorithm

Lei Zheng[1, 2]*, Defa Hu[3]

[1]*School of Information Engineering, Shandong Youth University of Political Science, Jinan 250103, Shandong, China*

[2]*Key Laboratory of Information Security and Intelligent Control in Universities of Shandong, Jinan 250103, Shandong, China*

[3]*School of Computer and Information Engineering, Hunan University of Commerce, Changsha 410205, Hunan, China*

Abstract

In order to improve the cloud computing resource scheduling efficiency, this paper proposed a method for cloud computing resource schedule based on chaos particle swarm optimization algorithm. Firstly, the resource scheduling options were taken as the position of the particle, and resource load balancing was taken as the objective function. Then the optimal resource scheduling solution was obtained by sharing information and the exchange of particles, while chaos mechanism was introduced to guarantee the diversity of particle swarm to prevent appearing premature convergence and local optimal solution. The simulation test was carried out in CloudSim platform, and the results show that the proposed method can quickly find the optimal scheduling solution for cloud computing resources and improve the efficiency of resource, and the method has better practicability and feasibility.

Keywords: cloud computing, resource scheduling, chaotic particle swarm optimization algorithm

1 Introduction

Cloud computing covers the virtualization, grid computing, network storage, and distributed computing technology, Using virtualization technology to virtualize the whole server as a resource pool to provide various services for the users [1]. The cloud resource scheduling is one of the core technologies in cloud computing, its efficiency affects Cloud computing systems performance directly, so it is crucial how to find an efficient and reasonable computing resource scheduling model. It determines the performance of cloud computing systems and develops an important research direction in the current [1, 2].

In 2009 based on greedy algorithm NimBus and others put forward cloud computing resource scheduling model. This model is simple but it is difficult to meet the increasing scale of cloud computing [3]. Then at the university of Melbourne Alexander put forward intergrid cloud computing resource scheduling model; in domestic, TianGuanhua, LvLianggan and others put forward dynamic resource scheduling model. These models improved the certain degree of efficiency for the cloud computing resource scheduling [4-6]. Following the development on the bionic intelligent algorithm many scholars put forward based on the genetic algorithm, particle swarm optimization (PSO) algorithm, immune evolutionary algorithm, ant colony algorithm, Leapfrog algorithm etc. cloud computing resource scheduling model. These bionic intelligent algorithms have a lot of advantages: stronger search ability, parallelism, etc.; finding the optimal scheduling scheme on Cloud computing resources quickly. These elements can effectively improve the efficiency for using cloud computing resources [7-9]. However, in actual application bionic intelligent algorithms are easily trapped some defects: local optimal solution, convergence speed etc. Sometimes it cannot obtain the optimal resource scheduling scheme, so it is necessary to improve and perfect the bionic intelligent algorithm, improve the efficiency on cloud computing resources.

In order to obtain better cloud computing resource scheduling scheme, improve the efficiency on computing resources, the paper puts forward a kind of cloud computing resource scheduling scheme that base on chaos particle swarm optimization (chaotic particle swarm optimization algorithm, CPSO), and verify its performance through the simulation.

2 The description for cloud computing resource scheduling problem

Compared with traditional distributed computing platform the advantage of cloud computing is virtualization technology. In cloud computing systems resources have been abstracted into the same virtual resource. Due to different virtual information demand of users and the difference of physical resources, how virtual resources are evenly schedule to the physical resource is a difficult problem in the study of cloud computing system. By certain strategies, resource scheduling is to reasonably schedule the virtual resources to physical resources. In a

---

Corresponding author e-mail: hdf666@163.com

real cloud computing environment, a virtual resource can be abstracted to a certain attribute node. For example, a virtual resource with the properties of internal storage, CPU, or band width can be abstracted as $v(m,c,b)$, in which m represents the properties of internal storage, c represents the properties of CPU, b and represents the properties of broadband. A physical resource can also be abstracted as $p(m,c,b)$.

Cloud computing resource scheduling is a scheme, which refers to schedule the virtual resources in cloud computing environment to physical resources. Corresponding to the virtual resource m_{vi}, c_{vi}, b_{vi} $v_i=1,2,...,m$ and physical resources $p_i(m_{pi}, c_{pi}, b_{pi})$ $p_i=1,2,...n,$ a cloud computing resource scheduling can be represented by the permutation $L(p_{i1}, p_{i2}, ... p_{im})$, which consisted by m resources. That the virtual resources c_{v1} be scheduled to physical resources p_{i1}, virtual resources c_{v2} be scheduled to physical resources p_{i2} and so on. A large number of studies have shown that cloud computing resource scheduling is a NP problem, if using traditional exhaustive algorithm to solve all possible cloud computing resources scheduling problem, the algorithm's time complexity is $0(n^m)$ [10].

3 Cloud computing resource scheduling solution of CPSO

3.1 THE PARTICLE SWARM OPTIMIZATION ALGORITHM

In the particle swarm optimization algorithm (PSO), this paper regarded the particles as a feasible solution in the problem space. Through itself and company information to guide particles to fly in the multidimensional space of the solution, finally find the optimal position of particles. The updated formula of particle's velocity and position formula is as the following:

$$V_{id}(i+1) = \omega \times v_{id}(i) + c_1 \times rand() \times (P_{best} - X_{id}(i) + c_2 \times rand() \times (g_{best} - X_{id}(i)), \quad (1)$$

$$X_{id}(i+1) = X_{id}(i) + V_{id}(i+1). \quad (2)$$

Among them, both of c_1 and c_2 are accelerating factor, $v_{id}(i)$ and $V_{id}(i+1)$ is the particle current speed and the updated speed respectively; $x_{id}(i)$ and $X_{id}(i+1)$ is the particle current position and the updated position respectively, ω represents the inertia weight; $rand()$ is the random number function of (0,1).

3.2 THE CHAOS PARTICLE SWARM OPTIMIZATION ALGORITHM

A large number of studies have shown that when a specific particle found a local optimal solution in PSO, it will produce attract to other particles. In this way the other particles will soon flying next to the particle, this will appear some defects: the particle swarm prematurity and local optimal solution etc. In order to avoid disadvantages

of PSO, using the uncertainty and randomness of chaotic theory and improving PSO by Chaotic mutation operator to generate the Chaos Particle Swarm Optimization algorithm (CPSO). The work thought of CPSO as follows: in the process of particle's flying, to execute chaotic mutation for the history optimal solution of the particle swarm g_{best} in order to prevent the particle position convergence and make the other particles to jump out of local optimal solutions as soon as possible. Specific as follows:

The Logistic mapping Equation represents as

$$Z_{n+1} = 4Z_n(1 - Z_n), \quad (3)$$

where, Z_n represents chaotic variables.

According to the principle of chaos, to add chaos disturbance to that is:

$$Z'_k = (1 - \alpha)Z^* + \alpha Z_k, \quad (4)$$

where, $\alpha \in [0,1]$, it represent the strength of disturbance, Z_k and Z'_k respectively represent chaos vector at the iteration k times and after adding disturbance.

Adopting the strength of the disturbance to execute α adaptive value, at the beginning of the search, its value is bigger and strengthen the disturbance to solution vectors, then slowly decreases, and the specific changes as follows:

$$\alpha(k) = 1 - \left(\frac{k-1}{k}\right)^n. \quad (5)$$

3.3 THE DESIGN FOR CLOUD RESOURCE SCHEDULING MODEL OF CPSO

3.3.1 Particle coding

Cloud computing resource scheduling refers to search all possible scheduling along the column to find an optimal scheduling scheme in the end. Set the serial number of cloud computing resources as $v_i=1,2,...,m$. The physical resource Numbers in Cloud system is $p_i=1,2,...n$. In order to facilitate calculation and to improve the effect of particle swarm search, this research uses the sequence numbers as particle coding plan. The first step is to sort the cloud computing resources needed scheduling, set as $(v_1, v_2, ..., v_m)$, then according to the sort to execute scheduling for cloud computing resources scheduling. Set the scheduling sequence as $(v_{a1}, v_{a2}, ..., v_{am})$, among them, $v_{ai} \in p_i$, then the scheduling model corresponds to v_i, $i=1,2,...,m$, as for the mapping physical resources, the encoding style in the v_{ai} algorithm is $(v_{a1}, v_{a2}, ..., v_{am})$.

3.3.2 The original particle swarm

The population initialization is the first step on the particle swarm optimization (pso) algorithm, which is also the most critical step; the fine particle swarm can save more computing time. However, basic particle population usually adopts random way to produce the initial particle swarm. In this case, the probability of larger particles to

focus on a local area will become larger; as a result the feasible solution will distribute unevenly and easy to appear local optimal solution, therefore, the study adopts the homogenization method to produce initial particle swarm.

3.3.3 The moderate function design

In cloud computing systems, the goal about resource scheduling optimization is to make the load balancing on all kinds of resources, so to evaluate the performance of a scheduling scheme, the main method is to evaluate whether the scheme can meet the demand of load balancing. For a particle, that is, a feasible cloud computing resource scheduling scheme $(v_{a1}, v_{a2}, ..., v_{am})$, the balancing degree of CPU, memory, bandwidth respectively defined as follows:

1) The equilibrium degree of CPU properties. Set

$\Gamma_{ctotal} = \sum_{i=1}^{m} c_{vi}/\sum_{i=1}^{n} p_{vi}$,

$\Gamma_{ci} = \sum_{i=1}^{m} c_{vai}/c_{pi}$.

Then the balance of degree about CPU properties is:

$$p_c = \sqrt{\sum_{i=1}^{m}(\Gamma_{ci} - \Gamma_{ctotal})^2}. \qquad (6)$$

Among them, $c_{vai} = \begin{cases} c_{vi}, \alpha_i = 1 \\ 0, \text{ otherwise} \end{cases}$.

2) The balance of degree about memory properties

$$p_m = \sqrt{\sum_{i=1}^{m}(\Gamma_{mi} - \Gamma_{mtotal})^2}. \qquad (7)$$

3) The balance of degree about bandwidth properties

$$p_b = \sqrt{\sum_{i=1}^{m}(\Gamma_{bi} - \Gamma_{btotal})^2}. \qquad (8)$$

In this way the optimal cloud computing resource scheduling optimal solution is to make the minimum attribute with CPU, memory and bandwidth resources change, so the fitness Fit(i) is defined as:

$$Fit(i) = a \times p_c + b \times p_b + m \times p_m \qquad (9)$$

In the Equation, $a + b + m = 1$ c, b, m indicates the importance of CPU, memory, bandwidth, respectively.

3.3.4 The cloud computing resource scheduling process of CPSO

The CMP task scheduling process of CPSO is shown in Figure 1.
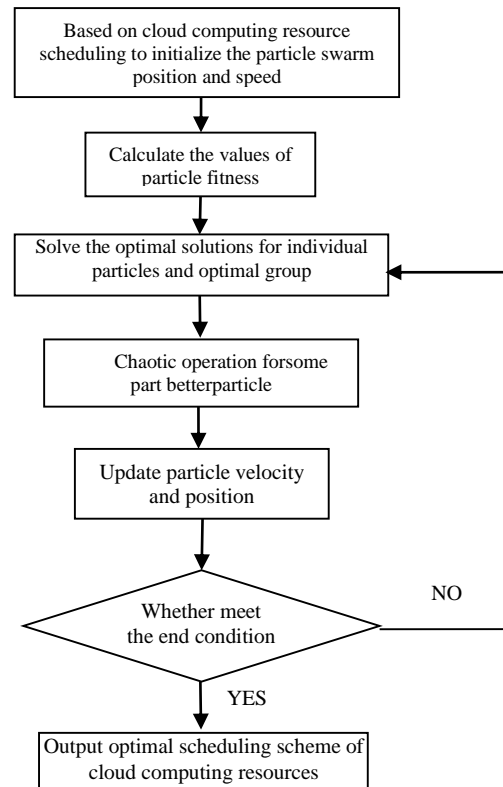
FIGURE 1 CMP task scheduling process of CPSO

1) According to the cloud computing system resources to initialize the particle swarm, the initial position and velocity respectively are: $v_{ai} = (v_{a1}, v_{a2}, ..., v_{am})$, $x_{ai} = (x_{a1}, x_{a2}, ..., x_{am})$.

2) According to the Equation (8) to calculate the fitness of each particle value, and choose the optimal particle as personal optimal solution P_{best} and group optimal solution g_{best}.

3) To determine whether the particle swarm appear the precocious phenomena. If there is a precocious phenomenon, the following step is to operate the chaotic conditions, the following step is to decode the position of the optimal particle in the group to get the Cloud computing resources optimal scheduling scheme, if do not meet the conditions, continue the iteration.

4 Simulation study

4.1 SIMULATION ENVIRONMENT

To test the validity of the cloud computing resource scheduling model based on CPSO, Gridsim software is used to simulate the local domain of a cloud computing. In order to measure the scheduling scheme of CPSO, using genetic algorithm (GA), the basic particle swarm optimization (PSO) as contrast model. To set the parameter of GA as: population size is 20, crossover and mutation probability were respectively 0.7 and 0.05; to set the parameter of PSO as: the size of particle swarm size is 20, c1 = c2 = 2; to set parameter of CPSO as: α = 0.5, other PSO is the same.

4.2 DISTRIBUTION OF CLOUD COMPUTING RESOURCE

In Cloud computing system the scope of CPU is [200 2000], the scope of memory scope is [10 1500], the scope of bandwidth scope is [5 200], randomly to generate 1000 physical resources and then virtualize them to guarantee the complexity of physical resource and the diversity of virtual task in cloud environment, part of virtual and physical resources after the abstraction of cloud computing environment as shown in Table 1.

TABLE 1 Virtual and physical resource after the abstraction

Physical resource			Virtual resource		
Memory	CPU	Band	Memory	CPU	Band
200	200	20	300	1000	100
400	500	10	200	2000	60
1500	1000	60	1000	700	40
100	50	5	800	600	20
10	60	30	1500	200	10

4.3 RESULTS AND ANALYSIS

To Run 100 times for each model, then to take the average as the final result of the algorithm. The balance of the CPU, memory and bandwidth for the three kinds of model is shown in Figures 2-4.
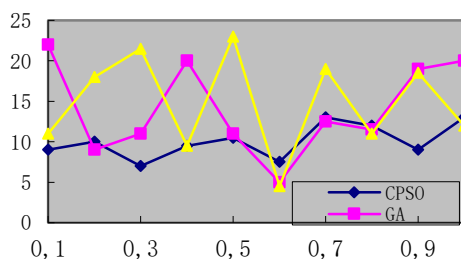
FIGURE 2 The change relation between the equilibrium degree and resources
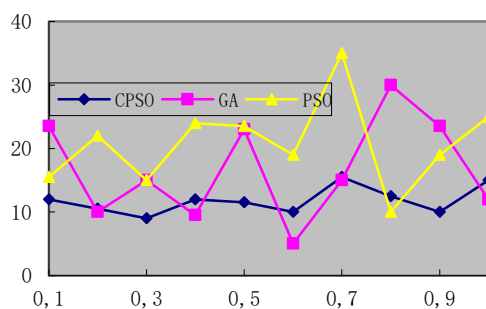
FIGURE 3 The change relationship between memory balancing degree and resources
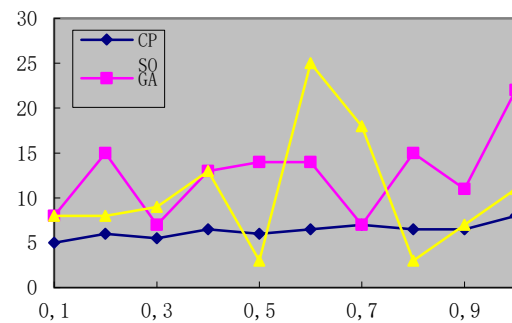
FIGURE 4 The change relationship between bandwidth balancing degree and resources

To analyse and compare the results of three scheduling model in Figures 2-4, we can obtain the following results:

1) In all models, the average performance of CPSO is the optimal; the performance of GA is the worst; it shows that CPSO can well reflect the large-scale cloud computing resources, sharing and dynamic characteristics, also can well decompose task and distribute it on computing resources, so CPSO is a kind of effective computing resource scheduling model.

2) CPSO is very stable relative to the particle swarm optimization (PSO), it can obtain better cloud computing resource scheduling scheme, this is mainly due to the CPSO introducing chaos mechanism, using the advantages of ergodicity and randomness of chaotic motion, ensures the diversity of particles and prevent the phenomenon of search in the late fall into local optimum, which suggesting that CPSO is effective for the improvement of the particle swarm optimization algorithm, and it further improves the convergence of algorithm c and enhances the Resource utilization of cloud computing.

5 Conclusion

According to the characteristics of cloud computing resources scheduling, this paper puts forward a kind of resource scheduling model based on chaotic particle swarm optimization (pso) algorithm, and the performance of the model is verified by simulation. The experimental results show that the CPSO can complete resource scheduling very well in cloud computing environment and has strong practical application value.

Acknowledgments

References

[1] Wright P, Sun Y L, Harmer T, Keenan A, Stewart A, Perrott R 2012 A constraints-based resource discovery model for multi-provider cloud environments *Journal of cloud computing: advances, systems and applications* **1**(1) 6

[2] Qiang Z, Ren-Bing X 2007 The virtual enterprise information sharing of the game theory and mechanism design *Computer integrated manufacturing system* **13**(8) 1566-71

[3] Truong H L, Dustdar S 2010 Composable cost estimation and monitoring for computational applications in cloud computing environments *Procedia Computer Science* **1**(1) 2175-84

[4] Harmer T, Wright P, Cunningham C, Hawkins J, Perrott R 2010 An application-centric model for cloud management *Proceedings of the 6th World Congress on Services Miami* 439-46

Information and Computer Technologies

[5] Maji P K, Bismas R, Roy A R 2010 Soft set theory *Computers & Mathematics with Applications* **45** 555-62

[6] Qi G, Ji Q, Pan J Z, Du J 2011 Extending description logics with uncertainty reasoning in possibilistic logic *International Journal of Intelligent Systems* **26** 353-81

[7] Aktas H, Cagman N 2007 Soft sets and soft groups *Information Sciences* **177** 2726-35

[8] Sun Y L, Perrott R, Harmer T, Cunningham C, Wright P 2010 An SLA focused financial services infrastructure *Proceedings of the 1st International Conference on Cloud Computing Virtualization (CCV 2010) Singapore*

[9] Rudolph S 2011 Foundations of Description Logics. In: Polleres A, D'Amato C, Arenas M, Handschuh S, Kroner P, Ossowski S, PatelSchneider P F Eds *Reasoning Web Semantic Technologies for the Web of Data LectureNotes in Computer Science Springer Berlin Heidelberg* 76-136

[10] Rochwerger B, Breitgand D, Levy E, Galis A, et al. 2009 The Reservoir model and architecture for open federated cloud computing *IBM Journal of Research and Development* **53**(4) 535-45

Authors
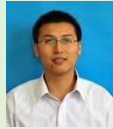
Lei Zheng, born on August 3, 1980, China

Current position, grades: a researcher at Shandong Youth University of Political Science, China.
University studies: master's degree in Computer Software and Theory from Shandong Normal University, China in 2006.
Scientific interest: cloud computing and distributed computing.

Defa Hu, born on October 5, 1980, China

Current position, grades: a researcher at Hunan University of Commerce, China.
University studies: Ph.D. degree in computer science and technology from Hunan University, China in 2010.
Scientific interest: information security and image processing.

CUBPT: Lock-free bulk insertions to B+ tree on GPU architecture

Yulong Huang[1*], Benyue Su[1], Jianqing Xi[2]

[1]*School of computer & information, Anqing Normal University, Anqing 246401, China*

[2]*School of Software Engineering, South China University of Technology, Guangzhou 510006, China*

Abstract

B+-tree is one of the most widely-used index structures. To improve insertion process, several batch algorithms are proposed, which all use one thread to complete one node insertion and cannot make full use of GPU's parallel throughput. So, a batch building and insertion method on GPU named CUBPT is proposed in this paper. During the process of bulk building and insertion, CUBPT use one thread to insert one key, which can maximize the performance by GPU. The experimental results show that when build a 10M tree, the overall performance of CUBPT improved 25.03 times compare with four threads PBI. When insert 10M uniform keys into a 10M tree, the overall performance of CUBPT improved 13.38 times compare with four threads PALM; when insert 10M highly skewed keys into tree with same size, the overall performance of CUBPT improved 15.23 times compare with four threads PALM.

Keywords: in-memory B+-tree, bulk build, Lock-free batch insertion, GPGPU

1 Introduction

Recently, the performance of single-core CPU is suffering a bottleneck and traditional architecture cannot promote the performance of multi-core CPU rapidly. Existing research result [1] shows that eight-core or above CPU cannot obtain any breakthrough on computing performance. Meanwhile, with the rapid development of GPU technology, it is ideal for high-performance computing task. Especially for the task which handles large-scale single-instruction and multiple data streams, the performance of GPU far beyond multi-core CPU. Hence, there has been a growing trend in leveraging the high parallel throughput of GPU for general purpose computing in parallel computing field. CUDA [2] is a programming tool, which allows programmers to write programs run on GPU rapidly. So, a large number of researchers use GPU to accelerate database operations such as sort, scan and achieved great results [3].

As memory capacity has increased dramatically, many database tables and related indices can reside in main memory completely now. So, more and more researchers devoted to improve operational performance of in-memory indices [4, 5]. The B+ tree is one of the most widely-used indexes. To improve its insertion, several batch algorithms are proposed [6, 7, 8]. These algorithms attempt to utilize different architecture processor to optimize the insertion process. Meanwhile, they cannot utilize parallel throughput of GPU. So, an experimental method on GPU is proposed [9]. However, this method just inserts one record at one time, which cannot make full use of the parallel throughput too. For

this reason, a novel lock-free batch insertion algorithm on GPU called CUBPT is proposed in this paper. It utilizes one thread to handle one key's insertion, which can take full advantage of the parallel throughput of GPU.

2 Related Works

In this section, the structure of B+-tree and serial insertion method are described firstly. On this basis, several batch algorithms are reviewed.

2.1 THE STRUCTURE AND SERIAL INSERTION OF B+ TREE

B^+ tree is a balanced search tree consists of internal and leaf nodes. To improve efficiency, a modified structure is proposed [10]. Here, each internal node contains the maximum key of its sub-trees and the pointers to these sub-trees. All leaf nodes are located on same layer and contain the keys and pointers to corresponding records. Leaf nodes are linked by order of the keys which makes it convenience to retrieve. To an internal node in B^+ tree with order m, its structure is as follows: $(P_0,K_0,P_1,K_1,\ldots,P_i,K_i)$ $0{\le}i{<}n$, $w$here, K_i represents the ith key and $K_{i-1}{<}K_i$. P_i is the pointer to the root node of ith sub-tree in which all the keys are lower than K_i. n is the number of keys stored in the node. For root node, the range of n is $[2,m)$. For internal nodes, the range of n is $[\lceil m/2 \rceil, m)$. The structure of leaf node is similar with internal node. It contains $n(\lceil m/2 \rceil{\le}n{\le}m)$ keys and pointers to corresponding records and also a pointer to adjacent node. A B^+ tree with order 3 is as follows.

---

[*] *Corresponding author* e-mail p3vsea2002@126.com
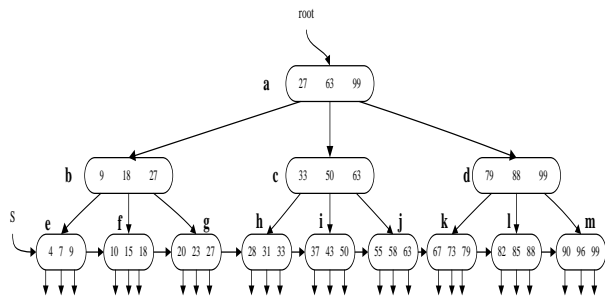
FIGURE 1 The structure of a 3-order B+ tree

As we all know, traditional B+ tree is constructed with one by one insertion, which start from an empty tree. So, only serial insertion algorithm with above B+ tree is described as an example. The insertion process of key 16 is as follows: First, compare 16 with *a*. Because 16<27, then compare 16 with *b*. Because 9<16<18, 16 should insert into *f*. Because there is no enough space, *f* should split into *f* and *f'*, where *f* contains keys {10, 15} and *f'* contains keys {16, 18}. Meanwhile, key 15 should insert into *b*. Similarly, *b* should split. Repeat such process until *a* split.

2.1 THE STRUCTURE AND SERIAL INSERTION OF B+ TREE

In traditional algorithm, the insertion process of each record needs to visit a path from root to leaf. If the size of records is very large, this process needs to execute frequently. So, a batch construction algorithm on single-core CPU is proposed [6]. Here, sort the records firstly and then build them bottom-up, once a layer from leaves to root. Because the path traversal times are reduced largely, the performance has improved greatly. However, this method cannot utilize the parallel throughput of multi-core CPU. So, another batch construction and insertion algorithm for B-link tree is proposed [7]. Similar with above algorithm, B-link tree is constructed with same way. But, this algorithm uses one thread to handle the construction process of one node. In batch insertion process, it also uses one thread complete one node insertion because the data contention. To leverage the parallel performance of many-core CPU, a new latch-free modifications algorithm called PALM is proposed [8]. Combined with BSP model [11] and auxiliary structures, PALM divides the insertion process into three stages. In every stage, it uses one thread to handle the insertion process of one node. To reduce synchronization cost, a point-to-point strategy is applied. For leveraging parallel throughput of GPU, an experimental insertion method is proposed [9], which can improve insertion efficiency with thread block. However, it only inserts one record by one time. In this basis, A B+ tree batch insertion algorithm on GPU is further proposed [14].Due to the requirement of this algorithm, it use many arrays to store the leaf nodes and internal nodes. So, the management of data structure is very complex. At the same time, since the space cost is very expensive, it

cannot complete batch insertion process in GPU when the scale of records is very large.

3 CUBPT lock-free batch insertion algorithm

3.1 THE STORAGE STRUCTURE OF CUBPT

As we all know, the design of data structure is one of the key problems for GPGPU programming and the array of structures is most suitable for GPU. So, we use array of nodes to store B+ tree. To optimize tree storage, we observe that all nodes have similar structure. Let *m* denote the order of tree. For every node, the size of array *keys* is *m* and array *ptrs* is *m*+1. In *ptrs*, the first *m* elements store the index of relevant nodes and the last has different usages. For leaf nodes, it is used to store the index of next node. For other nodes, it is used to store the first address of new storage space which generated in node split that can accelerate key insertion. All nodes contain a *type* to identify different type node. For our algorithm, a *parent* for parent node's location and a *kNum* for keys number in node are required. So, the size of every node is 2*(*m*+2) in device memory. The device memory structure of B+ tree described in Figure 1 is as follows.
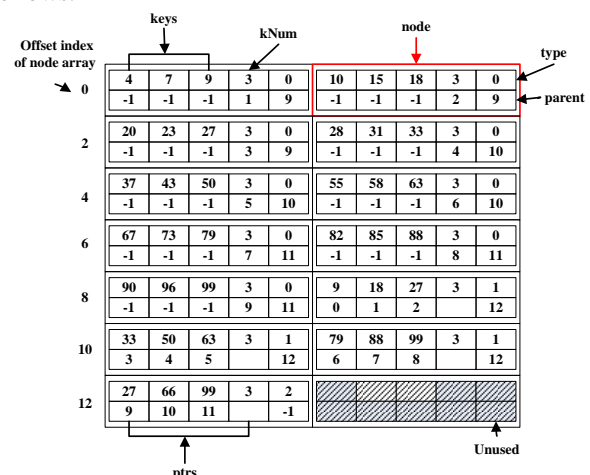
FIGURE 2 The storage structure of B+ tree in GPU

For B+ tree, the number of nodes increases by record insertion, which makes the expansion of nodes array necessary. However, CUDA does not support dynamic expansion. So, we adopt the following two-level structure to implement dynamic expansion.
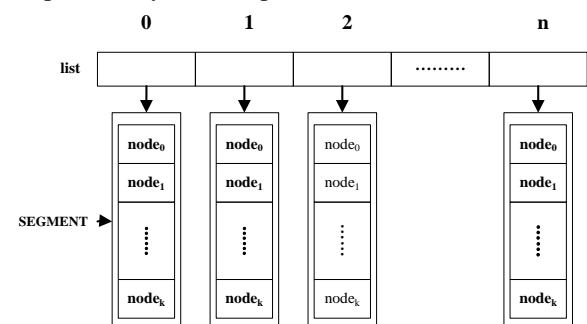
FIGURE 3 Two-level structure of nodes array

225

Here, nodes array is divided into several *segments* with fixed size and an array *list* is used to store every *segment's* address. Hence, only *list* needs to be expanded statically. So, the cost is very small. Expansion needs three stages: (i) create new *segments*; (ii) expand *list*; (iii) add address of new *segments* into *list*.

3.2 THE DESCRIPTION OF CUBPT ALGORITHM

To leverage parallel throughput of GPU to accelerate insertion process, a lock-free batch algorithm on GPU called CUBPT is proposed in here, which can improve this process very largely. The main idea is as follows: Firstly, allocate a buffer in main memory to store records to be inserted. When the number of records reaches threshold α, copy them from host to device. Then, use parallel primitive [12] to sort them. At last, batch insert these records into B$^+$ tree in device memory. The details are shown in Figure 4.
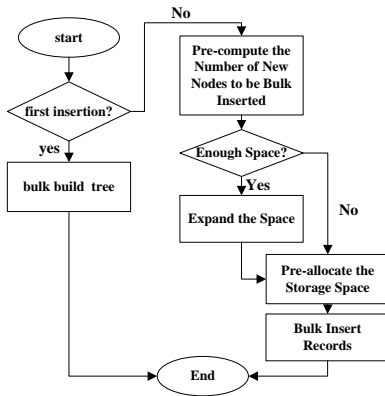
FIGURE 4 The flow chart of CUBPT

As depicted in Figure 4, when insert first batch records, we need to batch construct tree because B+ tree is null. After that, we may continue batch insertion process. So, our method consist of two parts which are batch construction and batch insertion. The detail of these two parts is as follows.

3.2.1 The process of batch construction

For an ordered keys set K, there are two stages to construct tree. Firstly, use $|K|$ threads to divide K into n leaf nodes. Then, use n threads to insert maximum key of every leaf node into its parent. Repeat this process until root node constructed. To distribute the keys more uniformly, we utilize the following Formula 1 and 2 to compute the number of nodes to build in every layer and the number of keys to stored in every node.

$$node\_num \leftarrow (|K|+m\text{-}1)/m, \tag{1}$$

$$\begin{cases} ave\_keynum \leftarrow (|K|+node\_num\text{-}1)/node\_num \\ h \leftarrow |K|\text{-}node\_num*(ave\_keynum\text{-}1) \end{cases}. \tag{2}$$

With above formulas, K is inserted into $node\_num$ leaf nodes, in which the first h leaf nodes contain

ave_keynum keys and the others contain ave_keynum-1 keys. The internal nodes are constructed by the same way. The detail of batch build process of CUBPT is as follows.

Batch_build: bulk building b+ tree in global memory
Input: an ordered key set K
Output: b+ tree in global memory T_{dm}
Begin //starting from the leaf layer and construct every layer of T_{dm} iteratively
1. $nodeNum\leftarrow(|K|+m\text{-}1)/m$;//compute the number of leaf nodes. m is the order of T_{dm}
2. **For each** $tid\in[0, nodeNum)$ **parallel do**
3.   set the value of $kNum$ and $prts[m]$of leaf node T_{dm}->node[tid];
4. **End for**
5. **For each** $tid\in[0,|K|)$ **parallel do** //bulk inserts the keys into the leaf nodes
6.   calculate index $node\_index$ and insert location $node\_loc$ of the leaf node that $K[tid]$ should be inserted;
7. T_{dm}->node[$node\_index$].keys[$insert\_loc$]$\leftarrow K[tid]$;
8. **End for**
9. **If**($nodeNum==1$) **return** T_{dm};
10.**While**($nodeNum>1$)
11. $cur\_num\leftarrow(nodeNum+m\text{-}1)$;//compute the number of nodes to be constructed in current layer
12.   Parallel set $kNum$ of the nodes in current layer by using $|cur\_num|$ threads;
13.   **For each** $tid\in[0,nodeNum)$ **parallel do**
14.   get the maximum key max_key of node[tid] in preceding layer;
15.   with formula 1 and 2, get the index $node\_index$ of the node that max_key should be inserted;
16.   insert max_key into T_{dm}->node[$node\_index$] and set the related pointer of child and parent node;
17.   **End for**
18. $nodeNum\leftarrow cur\_num$;
19.**End while**
20.Set the root pointer of T_{dm} and **return**;
End

As mentioned above, the difference between our algorithm and existing algorithms is that our algorithm uses one thread to handle one key's insertion, which can leverage the parallel throughput of GPU fully.

3.2.2 The process of batch insertion

In above process, CUBPT can maximally accelerate the construction process of b+ tree by using the parallel throughput of GPU. Similarly, to accelerate insertion process, a large-scale batch insert algorithm on GPU is proposed in here. It consists of the following stages:

(i) **Search**: find out the index of leaf nodes to be inserted.

(ii) **Batch insert leaf nodes layer:** According to the search results, batch insert the records into leaf nodes. If overflow, then split it.

(iii) **Batch insert internal nodes layer**: In previous stage, if leaf nodes split, then the maximum keys of newly increase nodes need to be inserted into parent nodes layer. During this process, if the nodes also split, continue insert into the next layer. If necessary, perform this process iteratively until the root node split.

(iv) **Create root node**: In 3rd stage, if root node split, then create a new one and insert related keys into it.

The details of these stages are as follows:

Search: according to the ordered keys set K, search the index of leaf nodes to be inserted by $|K|$ threads.

Starting from the root node, every thread uses binary search method to match the key to be inserted with current node, and then the next search node is obtained. Repeat this process until the index of leaf nodes to be inserted is found. Different with existing algorithms, this stage does not need to search leaf nodes. Hence, this reduced somewhat search cost.

Before batch insert K, we should ensure that there has enough space to accomplish current insertion, the total number of increase nodes in current batch insertion should be pre-counted. If no enough space, expand it with the way described in 3.1.The pre-count process consists of two steps: Firstly, parallel reduce the search result to array *mNodeIndex* and *mNode_keynum*. Then, use the following algorithm to count the total number of newly increase nodes.

CountAddNodeNum: Parallel count the number of newly increase nodes in current batch insertion

Input: index array *mNodeIndex* and increase keys number array *mNode_keyNum* of the leaf nodes to be inserted

Output: The total number of newly increase nodes *totalAddNum*

Begin

1. *mNode_keyNum*← get total keys number in related node after inserted by *mNodeIndex* and *mNode_keyNum*;
2. parallelly count the newly increase nodes number by formula 1 and reduce the result to *addNodeNum*;
3. if the newly increase nodes number in leaf layer is greater than 0,then iteratively count every internal layer with similar way.
4. If root node splits, then count the number of newly increased nodes
5. Reduce the results of above steps to *totalAddNum*

End

After the nodes array expanded, the following describe how to batch insert K into leaf nodes layer. The detail of this process is as follows:

Batch insert leaf nodes: If there has enough space, then batch insert K into the leaf nodes layer. This stage contains five steps as follows.

Step 1: According to the search result *mNodeIndex*, obtain original keys and indices of related leaf nodes with braid parallel method [5]. Then store them in array *ori_keys* and *ori_index*. Here, every thread block contains m threads. So, we can use one thread to get one key.

Step 2: Use following *merge_by_key* algorithm to merge *ori_keys, ori_index*, K and *mNodeIndex* into array *tmp_keys* and *tmp_index* by order of keys. For simplicity, we use Thrust [11] to implement which as follows.

---

merge_by_key: parallel merge four ordered arrays into two arrays

Input: K; index array *mNodeIndex*, original keys *ori_keys* and indices *ori_index* of leaf nodes to be inserted

Output: temporary keys array *tmp_index* and index array *tmp_index*

Begins

1. construct a virtual <*key, index*> pair array *part1* with K and *mNodeIndex*;
2. use similar method to construct a virtual <*key, index*> pair array *part2* with *ori_keys* and *ori_index*;
3. parallel merge *part1* and *part2* into *tmp_keys* and *tmp_index* respectively;

End

Step 3: according to *tmp_index*, use above allocation strategy to get the increase nodes number of every leaf

node, and store them in *addnode_num*. Figure 5 shows the details of step3 with a 3-order B$^+$ tree.
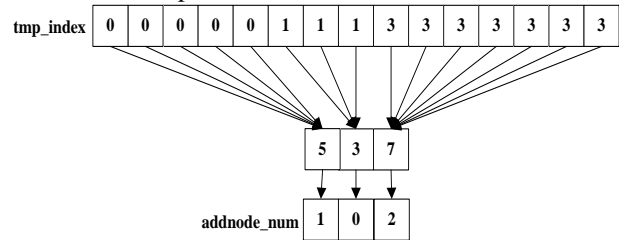
FIGURE 5 The details of Step 3

Step 4: with *addnode_num*, use $\sum addnode\_num[i]$ ($0\leq i<|addnode\_num|$) threads to allocate continuous space for newly increase nodes. Here, every thread need to set some values such as *ptrs[m]*, *kNum* and *parent*.

Step5: After that, according to *tmp_index*, use |*tmp_keys*| threads to parallel insert *tmp_keys* into the leaf nodes .In here, use *tmp_index* to split *tmp_keys* and store the result in array *partition_index*. Then, assign a thread for one key in *tmp_keys* to calculate index (*insert_index*) and insert location (*insert_loc*) of leaf node that the key to be inserted. Finally, batch inserts *tmp_keys* into the leaf nodes. The detail of perform this step on a 3 order b+ tree is as Figure 6.

Now, K is inserted into leaf nodes. The following shows how to batch inserts into internal nodes layer.

Batch insert internal nodes: Similar with previous stage, this stage also consists of five steps as follows:

Step 1: Obtain index and maximum key of newly increase nodes in previous level and Store them in array *addnode_index* and *addnode_maxkey*.

Step 2: According to *addnode_index*, the index of corresponding parent nodes are parallel obtained and store in array *parent_index* .On this basis, braid parallel method [5] is used to obtain index and maximum key of child nodes and store in array *ori_index* and *ori_maxkey*.

Step 3: Use the same way with step 2 in previous stage to merge *ori_maxkey*, *ori_index*, *addnode_index* and *addnode_maxkey* into array *tmp_keys* and *tmp_keys*.

Step 4 & Step 5: The process of these two steps is similar with the last two steps in previous stage. The only difference is that every thread in here also needs to modify the related child and parent node pointer.

At this point, all internal nodes split completely. If necessary, a new root node needs to be created.

Create a new root node: obtain the maximum key and index of newly increase nodes in the root node level and store in array *ori_maxkey* and *ori_index*. Then, use *merge_by_key* algorithm *to* merge *ori_maxkey, ori_index*, the maximum key and index of original root node into array *tmp_maxkey* and *tmp_index*. Finally use the same way with our batch construction algorithm to create a new root node.

In summary, according to the split situation of current nodes layer, our algorithm can use different number of threads to handle batch insertion process, which can make full use of the high parallel throughout to further accelerate the insertion process of B+-tree.
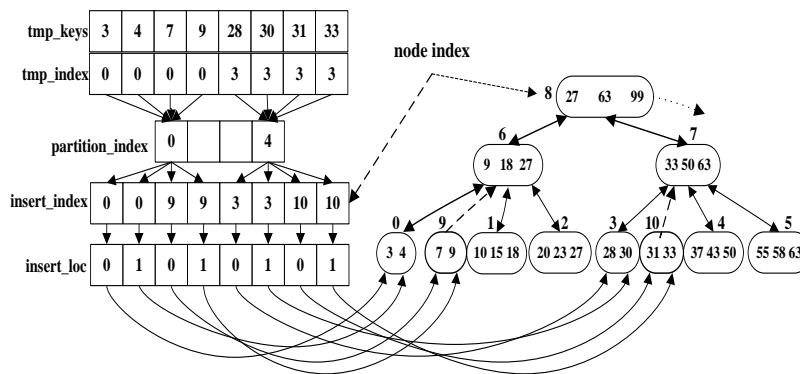
FIGURE 6 The details of Step 5 in stage of batch insert leaf nodes

4 Experimental results and analysis

To prove the effectiveness of our algorithm, we have implemented and tested on a PC with a NVIDIA GTX570 GPU and a recently-released Intel Corei7 Quad-core processor. The hardware configuration is shown in Table1 and software configuration is as follows: The operation system is WindowsXP professional sp3 and IDE is VS2008 with CUDA4.0; the detailed analysis of our algorithm is as follows.

TABLE 1 Hardware configuration

Hardware	GPU	CPU(quad-core)
Processors	780MHz *15 * 32	2.9GHz x*4
DRAM(GB)	1.25	6

4.1 BATCH CONSTRUCTION PROCESS ANALYSIS OF CUBPT

Here, randomly generate a group of datasets which size are 5M, 10M, 15M, and 20M respectively. To prove the effectiveness of our algorithm, three batch construction algorithms are selected to compare:(1) Kim's single-core CPU algorithm *SingleBuild*;(2)Liao's multi-core CPU algorithm *PBI*;(3)our GPU algorithm *CUBPTBuild*; For analysing the impact of different number of threads on *PBI* ,we use two (*PBITwo*) and four threads(*PBIFour*) to perform. As mentioned before, these algorithms all consist of sort and batch insertion stage. *CUBPTBuild* also contains data transfer stage. In sort stage, *singleBuild* use the sort function in STL[15], *PBI* use parallel_sort in TBB [13] and *CUBPTBuild* use a sort primitive in Thrust [12]. Meanwhile, the transfer stage is optimized by pinned memory. Figure 7 shows the elapsed time of above algorithms in batch construction stage when the order of B+-tree is 512.

From Figure 7 we know that, with the size of keys set increases, the performance speedup of *PBITwo* and *PBIFour* almost no changes, which are approximately 1.68 and 2.16. Meanwhile, our algorithm's speedup increase from 26.79 to 29.66. The reason is that *PBI* use one thread to complete one node's construction. When the size of keys set is very large, it needs more iterations. On the contrary, our algorithm uses one thread to handle one key's insertion and construct one layer by one time.

So, with increasing scale of keys set, our advantage will further increase.
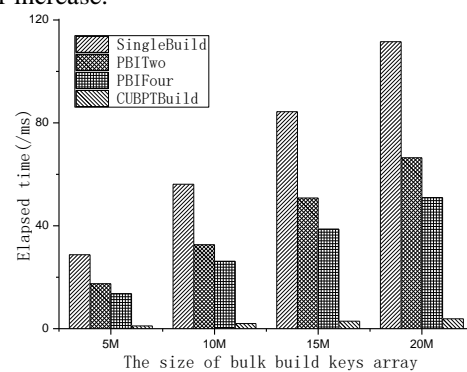
FIGURE 7. The build performance of above comparison algorithms

As above analysis, these comparison algorithms also contain other stages such as sort and transfer stage etc. So, we need to compare the overall performance which shows as Figure 8.
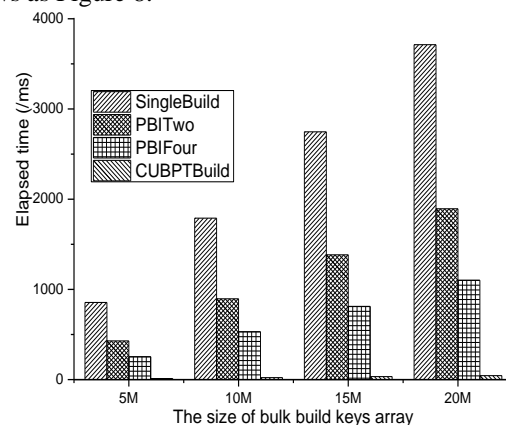
FIGURE 8 The overall performance of above comparison algorithms

As we can see from Figure 8, with the size of keys set increases, the overall speedup of *PBITwo* and *PBIFour* also no change, which are approximately 1.99 and 3.37. The overall speedup of our algorithm increases from 73.34 to 84.36. The reason is that GPU is more suitable for sorting than multi-core CPU, which brings a significant gain for our algorithm.

In summary, compare with *singleBuild* and *PBI*, our algorithm has distinct advantage in construction stage and overall performance. With the size of keys set increases, our algorithm's advantage continues to expand.

228

4.2 BATCH INSERTION PROCESS ANALYSIS OF CUBPT

Similarly, to prove effectiveness of the insertion process of our algorithm, three algorithms are selected to compare:(1) Liao's batch insertion algorithm on single-core CPU *PBIInsert*;(2) Jason's batch modification algorithm *PALM*;(3) our batch insertion algorithm *CUBPTInsert*. To analyse impact of the number of threads on *PALM*, two (*PALMTwo*) and four threads (*PALMFour*) are used to perform. These algorithms also consist of sort and insertion stage. In addition, our algorithm also contains transfer stage. In sort stage, these algorithms use the same method with above. We also optimize transfer with pinned memory. From section 3.2.2, we know that the impact of keys distribution is very large on our algorithm. So, we compare performance with different distribution.

4.2.1 Uniform distribution

Here, generate four uniform data sets. In them, use 5M, 10M, 15M and 20M to build tree respectively. On this basis, batch insert 2M, 4M, 6M, 8M and 10M into the tree, which constructed in front respectively. In this process, all leaf nodes need to be inserted. For above algorithms, we also compare the insertion performance and overall performance. When batch insert keys with different size into a 5M B+-tree, the elapsed time of above comparison algorithms in insertion stage shows as Figure 9.
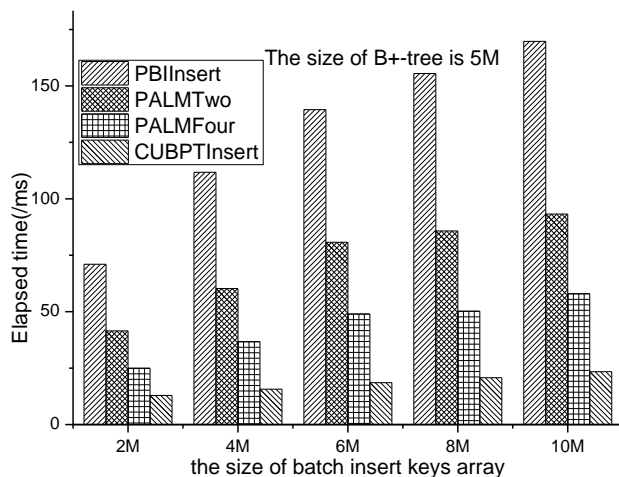
FIRURE 9 The insertion stage performance of above algorithms when insert uniform keys into a 5M B+-tree

When batch insert keys with different size into a10M B+-tree, the elapsed time of above comparison algorithms in insertion stage shows as Figure 10.
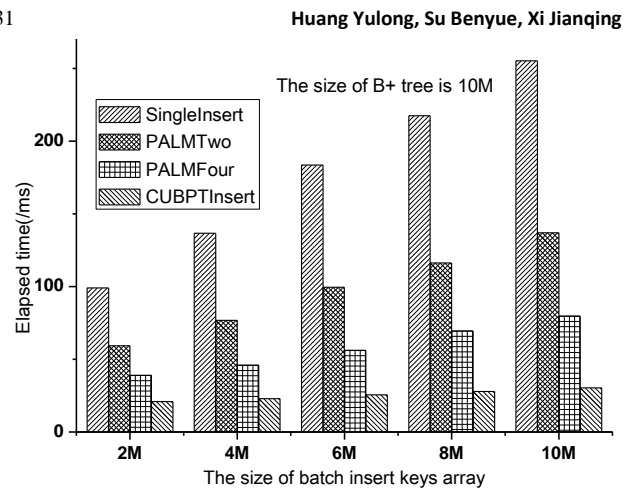
FIRURE 10 The insertion stage performance of above algorithms when insert uniform keys into a 10M B+-tree

In here, the elapsed time of all algorithm do not contain the search time. From figure 9 and figure 10 , we know that when the tree size is 5M, with the size of keys set increases, the speedup of *PALMTwo* and *PALMFour* in insertion stage are almost no change which are approximately 1.8 and 2.9. Meanwhile, our algorithm's speedup increases from 5.5 to 7.5.When the tree size is 10M, the speedup of *PALMTwo* and *PALMFour* almost no change too which are approximately 1.8 and 3.1. Meanwhile, the speedup of our algorithm increases from 4.76 to 8.43. According to this, we know that with the size of B+ tree increases, our algorithm's speedup increases by a small margin. The reason is that our algorithm use one thread to complete one key's insertion. So, it is more suitable to batch insert larger scale keys. Here, we also compared the overall performance of above comparison algorithms. When batch insert keys with different size into a 5M B+-tree, the overall performance of above comparison algorithms shows as Figure 11.
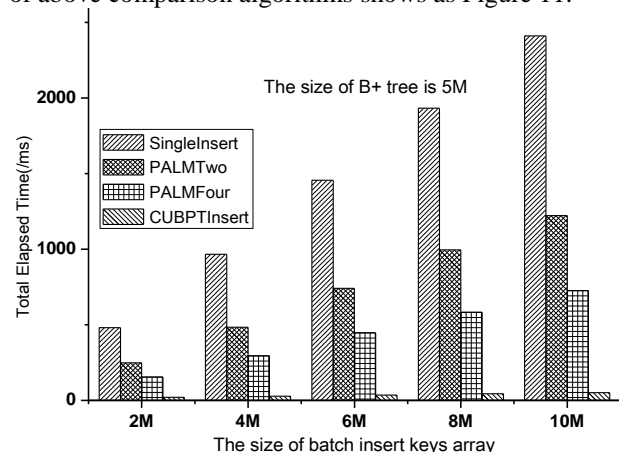
FIGURE 11 The overall performance of above algorithms when insert uniform keys into a 5M B+_tree

When batch insert keys with different size into a 10M B+-tree, the overall performance of above comparison algorithms shows as Figure 12.

From figure 11 and figure 12, we know that when tree size is 5M, with the size of keys set increases, the overall speedup of *PALMTwo* and *PALMFour* are almost no

change which approximately 1.95 and 3.2.However, our overall speedup increases from 24.4 to 47.4; when tree is 10M,the overall speedup of *PALMTwo* and *PALMFour* are almost no change too which approximately equals to above. Meanwhile, our speedup increases from 18.5 to 42.8. By analyzing the insertion and overall performance,we know that overall speedup is far more than the speedup of insertion stage for our algorithm. This is because performance gains in search and sorting stage far more than transfer cost.
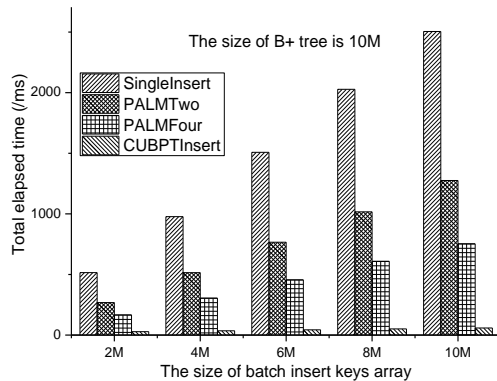
FIGURE 12 The overall performance of above algorithms when insert uniform keys into a 10M B+_tree

4.2.2 Highly skewed

Here, the size and usage of data sets is similar with uniform distribution. The difference is that the number of leaf nodes which to be inserted are less than 5% and 80% or more keys are inserted to one leaf node. In the same way, we also focus on insertion stage and overall performance of above algorithms. Next, we analyse the elapsed time of above algorithms in insertion stage firstly. When batch insert highly skewed keys with different size into a 5M B+-tree, the elapsed time of above comparison algorithms in insertion stage shows as Figure 13.
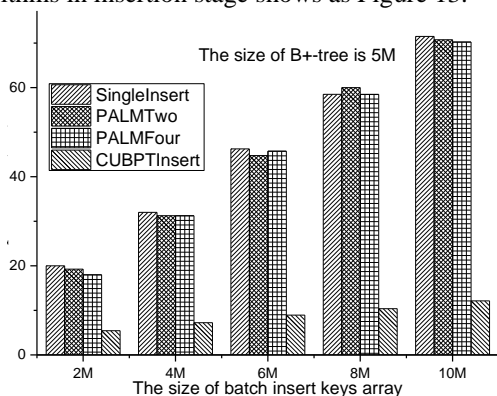
FIRURE 13 The insertion stage performance of above algorithms when insert highly skewed keys into a 5M B+-tree

When batch insert different scale highly skewed keys into a 10M B+-tree, the elapsed time of above comparison algorithms in insertion stage shows as Figure 14.

Here, all of the above comparison algorithms do not contains the search time too. With the size of keys set

increases, the performance of *PALM Two* and *PALMFour* have received almost no improvement. The reason is that Performance gains, which obtained from multiple threads execution are offset by cost of threads creation and synchronization. On the contrary, our speedup increases very obvious; when tree size is 5M, the performance speedup of our algorithm increases from 3.8 to 5.91. When tree size is 10M, the performance speedup of our algorithm increases from 3.7 to 6.1.The reason is same with uniform distribution. However, the performance promotion effect of our algorithm has a little decline. This is because the amount of data which processed by GPU is reduced slightly.
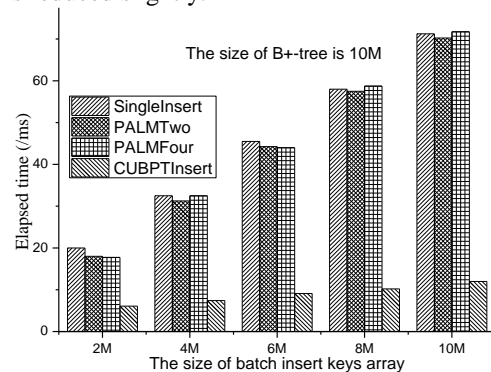
FIRURE 14 The insertion stage performance of above algorithms when insert highly skewed keys into a 10M B+-tree

To compare the overall performance of above algorithms, the total elapsed time when batch insert highly skewed keys with different size into a 5M B+-tree is described in Figure 15.
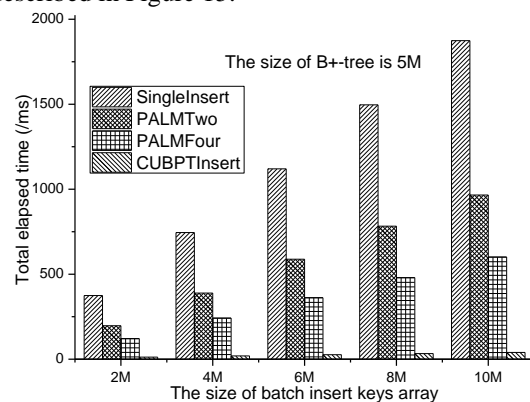
FIGURE 15 The overall performance of above algorithms when insert highly skewed keys into a 5M B+_tree

When batch insert highly skewed keys with different size into a 10M B+-tree, the total elapsed time of above comparison algorithms is described in Figure 16.

In figure 15 and figure 16, our algorithm contains transfer time. From them, we know that when the size of keys set increases, the overall performance speedup of *PALMTwo* and *PALMFour* remain unchanged which approximately 1.9 and 3.1. Meanwhile, our overall speedup continues to increase. If tree size is 5M, our overall speedup increases from 30.5 to 47.1. If the tree size is 10M, our overall speedup increases from 28.9 to 47.2. However, from above analysis, PALM can not get

any performance gains in insertion stage. Here,the overall speedup is mainly obtained by sort and search stages.For the same reason, our overall speedup is far more than the speedup of insertion stage.
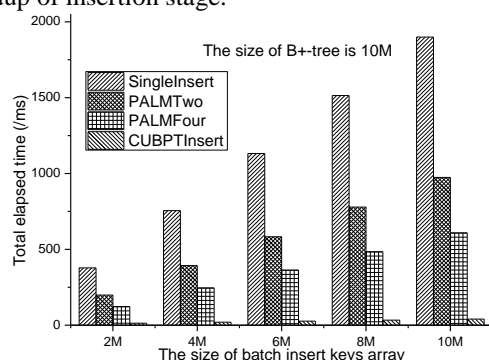
FIGURE 16 The overall performance of above algorithms when insert highly skewed keys into a 5M B+_tree

All in all, our algorithm has obvious performance advantage in both stages. Meanwhile, our speedup increases with the size of keys set which different with *PALM*. Especially, if keys distribution is highly skewed, our speedup is more obvious. So, our algorithm is more suitable for highly skewed keys insertion.

References

[1] Moore S K 2008 Multi-core is bad news for supercomputer *IEEE Spectrum* **15**
[2] *NVIDIA Corporation. NVIDIA CUDA Programming Guide, version 2.3* 2009 7-15
[3] BingSheng He, Ke Yang, Rui Fang, etc. 2008 Relational Joins on Graphics Processors *Proceedings of the 2008 ACM SIGMOD international conference on Management of data* 511-24
[4] Changkyu Kim, etc. 2010 FAST: Fast Architecture Sensitive Tree Search on Modern CPUs and GPUs *Proceedings of the 2010 international conference on Management of data* 339-50
[5] Fix J, Wilkes A, Skadron K 2011 Accelerating Braided B+ Tree Searches on a GPU with CUDA *Proceedings of the 2nd Workshop on Applications for Multi and Many Core Processors: Analysis, Implementation, and Performance*
[6] Sang-Wook Kim, Hee-Sun Won 2001 Batch Construction of B+-Trees *Proceedings of the 2001 ACM symposium on Applied computing* 231-5
[7] Jiangmiao Liao, Hu Chen, Yixia Yuan, et al. 2010 Parallel Batch B+-tree Insertion on Multi-core Architectures *Fifth International Conference on Frontier of Computer Science and Technology* 30-5
[8] Sewall J, Jatin Chhugani, et al. 2011 PALM: Parallel Architecture Friendly Latch-Free Modifications to B+ Trees on Many Core Processors *37th International Conference onVLDB* 795-806
[9] Kaczmarski K 2011 Experimental B+-tree for GPU *ADBIS2011 Research Communications, Austrian Computer Society* 232-40
[10] Yan Weiming, Wu Weimin 2007 *Data Structures* Tsinghua University Press 238-47
[11] Vliant L G 1990 A bridging model for parallel computation *Communications of the ACM* **33**(8) 103-11
[12] Hoverock J, Bell N 2011 *Thrust: A parallel template library*
[13] Reinders J 2007 *Intel Threading Building Blocks: Outfitting C++ for Multi-core Processor Parallelism* O'Reilly Media 78-9
[14] Kaczmarski K 2012 B-Tree Optimized for GPGPU *In Proceeding of OTM Conferences* 843-54
[15] Josuttis N M 2011 *The C++ Standard Library* Addison-Wesley

5 Conclusion and future research

B+-tree is one of the widely used index structures. To improve the process of keys insertion, several batch algorithms are proposed. Meanwhile, they cannot make full use of the parallel throughout of current GPU. So, a lock-free batch Insertion algorithm on GPU is proposed in here, which consists of two stages. The experimental results show that when batch construct tree, the total speedup of our algorithm can achieve 73.34 ~ 84.36. When batch insert keys, the total speedup of our algorithm can achieve 18.5 ~ 47.4. Here, we only research on batch construction and insertion algorithm. In the future, we will focus on batch search and deletion algorithm on GPU.

Acknowledgments

Authors

Yulong Huang, born in 1979, in Jian city, Jiangxi Province, China

Current position, grades: a lecturer of Computer Science and technology in School of Computer and information, Anqing Normal University.
University studies: He has received B.S degree in Computer Scicence and Technology from Wuhan University in 2002, received MSc and PhD in computer application technology from Guizhou University and South China University in 2005, 2013 respectivly.
Scientific interest: Parallel Computing, Internet of Things, Database Technology and Distributed System.

Benyue Su

Current position, grades: Professor and the Chair of the School of Computer and Information, Anqing Normal University in China. He is the senior member of China Computer Federation (CCF) and council member of Technical Committee of Geometric Design and Computing, CSIAM. He is also a senior member of the ACM.
University studies: Ph.D. in Computer Science from the Hefei University of Technology (HFUT), China, in 2007.
Scientific interest: visual computing, data analysis techniques, computer aided geometric design, computer graphics and digital image processing.
Publications: more than 60 refereed journal and conference papers in these areas.

Jianqing Xi

Current position, grades: a professor of software engineering in South China University of Technology.
University studies: MSc in software engineering and PhD in computer architecture from National University of Defense Technology in 1988 and 1992 respectively.
Scientific interest: Database and Data Warehouse, Distributed system based on P2P technology, Chinese information process, Data management and Software Development technology.

Probabilistic XML functional dependencies based on possible world model

Ping Yan[1], Teng Lv[2*], Weimin He[3]

[1]*School of Science, Anhui Agricultural University, Hefei 230036, China*

[2]*Teaching and Research Section of Computer, Army Officer Academy, Hefei 230031, China*

[3]*Department of Computing and New Media Technologies, University of Wisconsin-Stevens Point, 2100 Main Street, Stevens Point, WI 54481*

Abstract

With the increase of uncertain data in many new applications, such as sensor network, data integration, web extraction, etc., uncertainty both in relational databases and XML datasets has attracted more and more research interests in recent years. As functional dependencies (FDs) are critical and necessary to schema design and data rectification in relational databases and XML datasets, it is also significant to study FDs in uncertain XML datasets. This paper first proposed XML functional dependencies (XFDs) of deterministic XML dataset based on tree tuple models. Then two new kinds of functional dependencies based on possible worlds model for probabilistic XML dataset are introduced: probabilistic XML functional dependencies (pXFDs) and probabilistic approximate XML functional dependencies (pAXFDs). pXFDs extend the concept of XFDs of deterministic XML dataset by considering the probability of each possible world of probabilistic XML dataset, and pAXFDs extend the concept of probabilistic XML functional dependencies of probabilistic XML dataset by considering the degree of truth of tree tuples in each possible world of probabilistic XML dataset.

Keywords: uncertain XML, functional dependency, inference rule, closed set

1 Introduction

XML (Extensible Markup Language) has become the de facto standard of data exchange and is widely used in many fields. With the increase in applications such as data integration, web extraction, sensor networks, etc., XML datasets may be obtained from heterogeneous data sources and are not always deterministic. In such cases, XML datasets may contain uncertian data for the same attribute or element due to different data sources, information extraction, approximate query, and data measurement. Although functional dependencies (FDs) of uncertain XML datasets are much more complicated than the counterparts of traditional relational databases and deterministic XML datasets, it is possible and necessary to study them in uncertain XML datasets as shown in the paper.

Related work. Although there has been a lot of significant work in functional dependencies for relational databases and XML datasets, none of them can be directly applied to uncertain XML datasets. We analyse the related work in the following three aspects:

(1) For traditional relational databases, functional dependencies are thoroughly studied for several decades [1, 2]. Ref. [3] proposed a concept of functional dependencies in relational databases, which can deal with slight variations of data values. Ref. [4] proposed the conditional functional dependencies to detect and correct data inconsistency. It is obviously that the techniques of traditional relational databases cannot be directly applied to XML due to the significant difference in structure between XML documents and relational databases.

(2) For traditional deterministic XML datasets, functional dependencies are also thoroughly studied for some years. There are two major approaches to define functional dependencies in XML research community, i.e. path-based approach and sub-tree/sub-graph-based approach. In path-based approach [5-11], XML datasets are represented by a tree structure, and some paths of the tree with their values are used in defining XML functional dependencies. In Sub-tree/Sub-graph-based approach [12, 13], functional dependencies of XML datasets are defined by sub-graph or sub-tree in XML datasets. A sub-graph or a sub-tree is a set of paths of XML datasets. As an improvement over Sub-tree/Sub-graph-based approach functional dependencies, Refs. [14, 15] deal with XML functional dependencies with some constraint condition such that there exists a sub-tree is equal. The above XML functional dependencies cannot deal with uncertainty in XML datasets, which is the research topic of the paper. Ref. [16] proposes an approach to discover a set of minimal XML Conditional Functional Dependencies (XCFDs) from a given XML instance to improve data consistency. The XCFDs extends XML Functional Dependencies (XFDs) by incorporating conditions into XFD specifications. It is

easy to see that all the functional dependencies defined above cannot deal with uncertainty in XML datasets.

(3) For uncertain relational databases, Ref. [17] proposes the probabilistic functional dependency for probabilistic relational databases which associated with a likelihood of the traditional functional dependency is satisfied. Ref.[18] proposes some kinds of functional dependencies for probabilistic relational databases, such as Probabilistic Approximate Functional Dependencies (pAFD), Conditional Probabilistic Functional Dependencies (CpFD), and Conditional Probabilistic Approximate Functional Dependencies (CpAFD), which combine approximate, conditional, and approximate/conditional characteristics into traditional functional dependencies to defined corresponding functional dependencies for probabilistic relational databases. Ref. [19] proposes horizontal functional dependencies and vertical functional dependencies for uncertain relational databases, which extends the traditional relational functional dependencies into the uncertain relational databases. Although these work of uncertain relational databases are meaningful and sinificant, they can not directly applied in uncertain XML datasets, as XML are more complicated in structure then relational databases.

Contributions. In this paper, we will extend the concept of traditional deterministic XML functional dependencies (XFDs) to study the FDs of probabilistic XML datasets by considering the probability of each possible world and the degree of truth of tuples in each possible world. The main contributions of the paper are detailed as followings:

(1) We first proposed XML functional dependencies (XFDs) of deterministic XML dataset based on tree tuple models.

(2) Then a new kind of FDs called probabilistic XML functional dependencies (pXFDs) based on possible worlds model for probabilistic XML dataset is introduced, which extends the concept of XFDs of deterministic XML dataset by considering the probability of each possible world of probabilistic XML dataset.. pXFDs, and pAXFDs A sound and complete inference rules are given for the three types of uncertain XML functional dependencies.

(3) Finally another new kind of FDs called probabilistic approximate XML functional dependencies (pAXFDs) is introduced, which extend the concept of pXFDs of probabilistic XML dataset by considering the degree of truth of tree tuples in each possible world of probabilistic XML dataset.

(4) We also analyze the relationship among the tree kinds of FDs: XFDs, pXFDs and pAXFDs. pXFDs extend XFDs by considering the probability of each possible world of probabilistic XML dataset, and pAXFDs extend pXFDs by considering the degree of truth of tree tuples in each possible world of probabilistic XML dataset. More generally speaking, pXFDs are more

general than XFDs, and pAXFDs are more general than pXFDs.

Organizations. The rest of the paper is organized as following: Section 2 gives an example as our research motivation and demonstration of the concepts throughout the paper. Three types of functional dependencies, including XFDs, pXFDs, and pAXFDs, of probabilistic XML datasets are given in Section 3. Finally, Section 4 concludes the paper and points out the future directions of the paper.

2 A Motivating example

Suppose we want to know the people impressions of relationship between a man's salary and his diploma/height. In terms of functional dependencies, if it is the case that [diploma, height]→salary? We design a questionnaire and obtain some data as the following:

TABLE 1 Four interviewees' impressions about the relationship between a man's salary and his diploma/height

Record	Interviewee ID	diploma	height	salary	probability
1	1	High	high	high	0.9
2	1	High	low	high	0.8
3	1	Low	high	low	0.5
4	1	Low	low	low	0.7
5	2	High	high	high	0.9
6	2	High	low	high	0.7
7	2	Low	high	high	0.5
8	2	Low	low	low	0.8
9	3	High	high	high	0.6
10	3	Low	low	low	0.7
11	4	High	low	high	0.7
12	4	Low	high	high	0.5

For Record 1 in Table 1, interviewee 1 thinks (with confidence of 0.9) that if a man's diploma is high and height is high, then his salary is high. We can store Table 1 as an XML file (survey.xml) as following (attribute "Prob" stands for the confidence of an interviewee's answer):

```
<survey>
    <interviewee>
        <ID>1</ID>
        <answer Prob='0.9'>
            <diploma>high</diplomas>
            <height>high</height>
            <salary>high</salary>
        </answer>
        <answer Prob='0.8'>
            <diploma>high</diplomas>
            <height>low</height>
            <salary>high</salary>
        </answer>
        … …
    </interviewee>
    <interviewee>
        <ID>2</ID>
```

```
    <answer Prob='0.9'>
        <diploma>high</diplomas>
        <height>high</height>
        <salary>high</salary>
    </answer>
        … …
    </interviewee>
    <interviewee>
        <ID>3</ID>
        <answer Prob='0.7'>
            <diploma>high</diplomas>
            <height>high</height>
            <salary>high</salary>
        </answer>
        … …
    </interviewee>
    <interviewee>
        <ID>4</ID>
        <answer Prob='0.7'>
            <diploma>high</diplomas>
            <height>low</height>
            <salary>high</salary>
        </answer>
        … …
    </interviewee>
</survey>
```

We will use the above XML file as demonstration of our proposed concepts throughout the paper.

3 Functional dependencies of probabilistic XML dataset

We first give some preliminary definitions such as DTD (Document Type Definition), XML tree, tree tuple, etc.:

Definition 1 (DTD). A DTD[20] is defined to be D=(E, A, P, R, r), where (1) E is a finite set of element types; (2) A is a finite set of attributes; (3) P is a mapping from E to element type definitions. For each $\tau \in E$, $P(\tau)$ is a regular expression α defined as $\alpha ::= S \mid \varepsilon \mid \tau' \mid \alpha \mid \alpha, \alpha \mid \alpha^*$, where S denotes string types , ε is the empty sequence , $\tau' \in E$, "|", " , " and " * " denote union, concatenation and Kleene closure respectively; (4) R is a mapping from E to the power set of A : $P(A)$; (5) $r \in E$ is called the element type of the root.

A path p in D=(E, A, P, R, r) is defined to be $p = \omega_1 \cdots \omega_n$, where (1) $\omega_1 \in r$; (2) $\omega_i \in P(\omega_{i-1})$, $i \in [2, n-1]$; (3) $\omega_n \in P(\omega_{n-1})$ if $\omega_n \in E$ and $P(\omega_n) \neq \Phi$, or $\omega_n = S$ if $\omega_n \in E$, and $P(\omega_n) = \Phi$, or $\omega_n \in R(\omega_{n-1})$ if $\omega_n \in A$. Let $paths(D) = \{ p \mid p \in D \}$.

Definition 2 (XML tree). Let D=(E, A, P, R, r). An XML tree T conforming to D (denoted by $T \models D$) is defined to be T=(V, lab, ele, att, val, root), where (1) V is a finite set of nodes; (2) lab is a mapping from V to $E \cup A$; (3) ele is a partial function from V to V^* such that for any $v \in V$, $ele(v) = [v_1, \ldots, v_n]$ if $lab(v_1) \cdots lab(v_n)$ is defined in $P(lab(v))$; (4) att is a partial function from V to A such that for any $v \in V$, $att(v) = R(lab(v))$ if $lab(v) \in E$ and $R(lab(v))$ is defined in D ; (5) val is a partial function from V to S such that for any $v \in V$, $val(v)$ is defined if $P(lab(v)) = S$ or $lab(v) \in A$; (6) $lab(root) = r$ is called the root of T .

Given a DTD D and an XML tree T|=D, a path p in T is defined to be $p = v_1 \cdots v_n$, where (1) $v_1 \in root$; (2) $v_i \in ele(v_{i-1}), i \in [2, n-1]$; (3) $v_n \in ele(v_{n-1})$ if $lab(v_n) \in E$, or $v_n \in att(v_{n-1})$ if $lab(v_n) \in A$, or $v_n = S$ if $P(lab(v_{n-1})) = S$. Let $paths(T) = \{ p \mid p \in T \}$.

Fig.1 is a part of an XML tree corresponding to XML file survey.xml in Seciton 2, in which Interview 1 thinks with confidence 0.9 that if a man's diploma is high and height is high, then his salary is high.
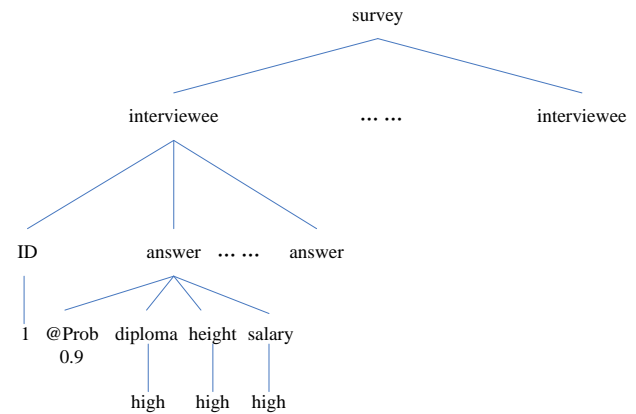
FIGURE 1 A part of an XML tree corresponding to XML file survey.xml

Definition 3 (Tree tuple). Given a DTD $D=(E, A, P, R, r)$ and an XML tree $T=(V, lab, ele, att, val, root)$ conforming to D, a tree tuple t for a node v in T is a tree rooted on node v with all of its decedent nodes.

In XML file survey.xml, there are 4 "interviewee" tree tuples, Fig 1 shows the first "interviewee" tree tuple which is the tree rooted on the first "interviewee" node (ID=1). There are also 12 "answer" tree tuples in the XML file, and Fig 1 shows the first "answer" tree tuple, which is the tree, rooted on the first "answer" node with probability of 0.9.

Based on the above concepts, we give the definition of XML functional dependencies (XML FDs) of deterministic XML dataset, which is based on tree tuple model.

Definition 4 (XML FDs). Given a DTD D and an XML tree $T \models D$, an XML Functional Dependency (XFD) has the form $\{s_h, S_L \to S_R\}$, where s_h is the head path, $S_L = \{s_{L1}, s_{L2}, \cdots, s_{Lm}\}$ is the left path set (i.e. determinant path set), and $S_R = \{s_{R1}, s_{R2}, \cdots, s_{Rn}\}$ is the right path set (i.e. determined path set). For $\forall s_{Li} \in S_L, s_{Rj} \in S_R$, $s_h, s_h.s_{Li}, s_h.s_{Rj} \in paths(D)$, where $i = 1, \cdots, m, j = 1, \cdots, n$. If T satisfies XFD $\{s_h, S_L \to S_R\}$ (denoted by $T \models \{s_h, S_L \to S_R\}$), then it implies that for each tree tuple t rooted on node $last(s_h)$, $t(s_h.s_{L1}, s_h.s_{L2}, \cdots, s_h.s_{Lm})$ can uniquely determines $t(s_h.s_{R1}, s_h.s_{R2}, \cdots, s_h.s_{Rn})$, where $last(s_h)$ denotes the last symbol of head path s_h.

By considering the probability of each possible world of probabilistic XML dataset, we give the definition of probabilistic XML FDs of deterministic XML dataset based on possible world model by extending the above concept of XFDs as following:

Definition 5 (Probabilistic XML FDs). Given a DTD D and an uncertain XML tree $T \models D$, a Probabilistic XML FD (pXFD) has the form $\{s_h, S_L \to_{pXFD} S_R\}$, where s_h is the head path, $S_L = \{s_{L1}, s_{L2}, \cdots, s_{Lm}\}$ is the left path set (i.e. determinant path set), and $S_R = \{s_{R1}, s_{R2}, \cdots, s_{Rn}\}$ is the right path set (i.e. determined path set). For $\forall s_{Li} \in S_L, s_{Rj} \in S_R$, $s_h, s_h.s_{Li}, s_h.s_{Rj} \in paths(D)$, where $i = 1, \cdots, m, j = 1, \cdots, n$. T satisfies pXFD $\{s_h, S_L \to_{pXFD} S_R\}$ (denoted by $T \models \{s_h, S_L \to_{pXFD} S_R\}$) with a confidence c if and only if the sum of probability of the fraction of possible worlds in which the corresponding XFD $\{s_h, S_L \to S_R\}$ holds is c,

$$c = \frac{\sum p(PW_i)}{\sum_{j=1}^{n} p(PW_j)}$$, where $p(PW_i)$ stands for the probability of possible world PW_i, $\sum p(PW_i)$ stands for the sum of probability of all those possible worlds which satisfy XFD $\{s_h, S_L \to S_R\}$, and n is number of all possible worlds of T.

Example 1. XML file survey.xml has 64 possible worlds (PW) w.r.t. element "interviewee", the first possible world is PW$_1$ with the following 4 tree tuples:

T_1^1 = (survey.interviewee.ID=1, survey.interviewee.answer.@Prob=0.9, survey.interviewee.answer.diploma=high, survey.interviewee.answer.height=high, survey.interviewee.answer.salary=high),

T_2^1 = (survey.interviewee.ID=2, survey.interviewee.answer.@Prob=0.9, survey.interviewee.answer.diploma=high, survey.interviewee.answer.height=high, survey.interviewee.answer.salary=high),

T_3^1 = (survey.interviewee.ID=3, survey.interviewee.answer.@Prob=0.6, survey.interviewee.answer.diploma=high, survey.interviewee.answer.height=high, survey.interviewee.answer.salary=high),

T_4^1 = (survey.interviewee.ID=4, survey.interviewee.answer.@Prob=0.7, survey.interviewee.answer.diploma=high, survey.interviewee.answer.height=low, survey.interviewee.answer.salary=high).

It is easy to see that PW$_1$ satisfies the XFD with confidence 0.9*0.9*0.6*0.7=0.3402, i.e., PW$_1$ contribute to the final confidence of the entire XML file survey.xml with confidence 0.3402. We omit other possible worlds considering the space here and we can see that the number of possible worlds which satisfy XFD {survey.interviewee.answer, [diploma, height]→salary} is 48, so XML file survey.xml satisfies pXFD {survey.interviewee.answer, [diploma, height]→$_{pXFD}$ salary} with confidence $c = \frac{\sum p(PW_i)}{\sum_{j=1}^{n} p(PW_j)} = \frac{11.9416}{13.1116} \approx 0.91$,

where $p(PW_i)$ stands for the probability of possible world PW_i (for example, $p(PW_1) = 0.9*0.9*0.6*0.7=0.3402$), $\sum p(PW_i)$ stands for the sum of probability of all those possible worlds which satisfy the XFD {survey.interviewee.answer, [diploma, height]→salary}, and n is the number of all possible worlds of survey.xml file. The meaning of the pXFD {survey.interviewee.answer, [diploma, height]→$_{pXFD}$ salary} is that the four interviewees in the survey think (with confidence c=0.91) that a person's diploma and height can determine his/her salary.

The relationship betweenXFDs and pXFDs are given in the following theorem:

Theorem 1 (Relationship between XFD and pXFDs). If each possible world satisfies a pXFD with probability of 1 (i.e. the XML dataset are deterministic), then the pXFD is equal to the corresponding XFD.

Proof. From the definitions of pXFDs and XFDs, we can see that pXFDs are natural extensions of corresponding traditional deterministic XFDs by considering the probability of each possible world of probabilistic XML dataset. So if all possible world satisfy a pXFD with probability of 1, then the entire XML dataset must satisfy the corresponding XFD.

It should be noted that pXFDs suffer from the same kind of flaws that XFDs do. More specifically, if the XML dataset is dirty, just a tree tuple in a possible world that does not conform to the XFDs can cause the entire possible world to satisfy the XFDs. For example, consider possible world PW$_{34}$ in XML file survey.xml, which has the following 4 tree tuples:

T_1^{34} = (survey.interviewee.ID=1, survey.interviewee.answer.@Prob=0.5, survey.interviewee.answer.diploma=low,

survey.interviewee.answer.height=high,
survey.interviewee.answer.salary=low),

T_2^{34} = (survey.interviewee.ID=2,
survey.interviewee.answer.@Prob=0.9,
survey.interviewee.answer.diploma=high,
survey.interviewee.answer.height=high,
survey.interviewee.answer.salary=high),

T_3^{34} = (survey.interviewee.ID=3,
survey.interviewee.answer.@Prob=0.6,
survey.interviewee.answer.diploma=high,
survey.interviewee.answer.height=high,
survey.interviewee.answer.salary=high),

T_4^{34} = (survey.interviewee.ID=4,
survey.interviewee.answer.@Prob=0.5,
survey.interviewee.answer.diploma=low,
survey.interviewee.answer.height=high,
survey.interviewee.answer.salary=high).

It is easy to see that PW$_{34}$ does not satisfy pXFD {survey.interviewee.answer, [diploma, height]$\rightarrow_{pXFD}$ salary} as tree tuple T_1^{34} and tree tuple T_4^{34} are contradict to each other.

To solve the above problem, we propose the concept of probabilistic approximate XML FDs as following:

Definition 6 (Probabilistic Approximate XML FDs). Given a DTD *D* and an uncertain XML tree $T \models D$, a Probabilistic Approximate XML FD (pAXFD) has the form $\{s_h, S_L \rightarrow_{pXAFD} S_R\}$, where s_h is the head path, $S_L = \{s_{L1}, s_{L2}, \cdots, s_{Lm}\}$ is the left path set (determinant path set), and $S_R = \{s_{R1}, s_{R2}, \cdots, s_{Rn}\}$ is the right path set (determined path set). For $\forall s_{Li} \in S_L, s_{Rj} \in S_R$, $s_h.s_h.s_{Li}, s_h.s_{Rj} \in paths(D)$. *T* satisfies pAXFD $\{s_h, S_L \rightarrow_{pXAFD} S_R\}$ (denoted by $T \models \{s_h, S_L \rightarrow_{pXAFD} S_R\}$) with a confidence *c* if and only if the sum of probability of each possible world multiplied by the maximal fraction of the tree tuples of the possible world, in which the corresponding pXFD $\{s_h, S_L \rightarrow S_R\}$ holds is *c*, $c = \dfrac{\sum (p(PW_i) \times \alpha_i)}{\sum_{j=1}^{n} p(PW_j)}$, where $p(PW_i)$ stands for the probability of possible world PW_i, and α_i stands for the maximal tree tuples fraction in possible world PW_i, which satisfies the corresponding pXFD $\{s_h, S_L \rightarrow_{pXFD} S_R\}$, and *n* is the number of all possible worlds of T.

Example 2. As noted before, possible world PW$_{34}$ in XML file survey.xml does not satisfy pXFD {survey.interviewee.answer, [diploma, height]$\rightarrow_{pXFD}$ salary} as tree tuple T_1^{34} and tree tuple T_4^{34} are contradict to each other. If we just remove any one of them, then PW$_{34}$ satisfies the pXFD {survey.interviewee.answer, [diploma, height]$\rightarrow_{pXFD}$ salary}. It is easy to see that for pXFD {survey.interviewee.answer, [diploma,

height]$\rightarrow_{pXFD}$ salary} to be satisfied by PW$_{34}$, the removed minimal fraction of tree tuples is 1/4, so PW$_{34}$ satisfies the pAXFD {survey.interviewee.answer, [diploma, height]$\rightarrow_{pAXFD}$ salary} with confidence p(PW$_{34}$)*(1-1/4)=0.5*0.9*0.6*0.5*3/4=0.10125. For the entire XML file survey.xml, it satisfies the pAXFD {survey.interviewee.answer, [diploma, height]$\rightarrow_{pXFD}$ salary} with the following confidence $c = \dfrac{\sum (p(PW_i) \times \alpha_i)}{\sum_{j=1}^{n} p(PW_j)} = \dfrac{12.8191}{13.1116} \approx 0.98$, where $p(PW_i)$ stands for the probability of possible world PW_i, and α_i stands for the maximal tree tuples fraction in possible world PW_i, which satisfies the corresponding pXFD {survey.interviewee.answer, [diploma, height]$\rightarrow_{pXFD}$salary}, and *n* is the number of all possible worlds. The meaning of the pAXFD {survey.interviewee.answer, [diploma, height]$\rightarrow_{pXAFD}$ salary} is that the four interviewees in the survey approximately think (with confidence *c*=0.98) that a man's diploma and height can determine his salary. We can see that the pAXFD's confidence is generally higher that that of pXFD (the former is 0.98, and the latter is 0.91).

The relationship between pXFDs and pAXFDs are given in the following theorem:

Theorem 2 (Relationship between pXFDs and pAXFDs). If all tree tuples of each possible world either satisfy a pXFD or not at all, then the entire probabilistic XML dataset satisfies the corresponding pAXFD with the same confidence. In general, the confidence of a pAXFD is always greater than the confidence of the corresponding pXFD.

Proof. (1) From the definitions of pXFDs and pAXFDs, we can see that a pAXFD is a natural extension of the corresponding pXFD by considering the degree of truth of tree tuples in each possible world of a probabilistic XML dataset. For a pAXFD, if each tree tuple of each possible world either satisfies the corresponding pXFD or not at all, then the degree of truth of tree tuples in each possible world is either 1 or 0 w.r.t. the pAXFD. So the pAXFD is equal to the corresponding pXFD with the same confidence. (2) In each possible world that the pXFD holds, the degree of the truth (α_i in Definition 6) of the corresponding pAXFD is 1. In each possible world that the pXFD does not hold, the degree of the truth of the corresponding pAXFD is always greater than 0. So the final confidence of the pAXFD, which is the weighted sum of the confidences of all possible worlds multiplied by the corresponding degree of truth, is always greater than the confidence of the corresponding pXFDs.

4 Conclusions

This paper studies the functional dependencies (FDs) of probabilistic XML datasets, which extends the notions of

functional dependencies of uncertain relational databases and traditional functional dependencies of XML datasets based on tree-tuple model. Three kinds of functional dependencies such as XFDs, pXFDs, and pAXFDs are given in the paper: XFDs are traditional FDs to capture the relationship between XML element in a deterministic XML dataset; pXFDs are natural extensions of XFDs by considering the probability of each possible world of probabilistic XML dataset; and pAXFDs are natural extensions of pXFDs by considering the degree of truth of each possible world of probabilistic XML dataset, which is very useful in the case that there are noisy or dirty data in probabilistic XML dataset. The relationship among XFDs, pXFDs, and pAXFDs are also studied in the paper. Generally speaking, pXFDs are more general than XFDs, and pAXFDs are more general that pXFDs.

An interesting work in the future is to assess the confidence of pXFDs and pAXFDs. As we know that both pXFDs and pAXFDs are defined on possible world model and the number of possible world is very large generally, it is a not trivial work to access the confidence of pXFDs and pAXFDs. Another interesting work is to mine the pXFDs and pAXFDs in a given probabilistic XML dataset. Specifically speaking, given a probabilistic XML dataset, find a minimal set of pXFDs or pAXFDs that are equivalent to or more general than any set of pXFDs or pAXFDs that hold over the XML dataset with a confidence higher than a user specified threshold.

Acknowledgments

References

[1] Abiteboul S, Hull R, Vianu V 1995 *Foundations of Databases. Addison-Wesley* Boston

[2] Ullman J D 1988 *Principles of Database and Knowledge-Base Systems* **1** Computer Science Press, New York

[3] Koudas N, Saha A, Srivastava D 2009 Venkatasubramanian, S.: Metric functional dependencies *Proc. of the 25th International Conference on Data Engineering* 1275-8 IEEE Computer Society Press, New York

[4] Bohanno P, Fan W, Geerts F, Jia X, Kementsietsidis A 2007 Conditional functional dependencies for data cleaning *Proc. of the 23rd International Conference on Data Engineering* 746-55 IEEE Computer Society Press, New York

[5] Janosi-Rancz K T, Varga V, Nagy T 2010 Detecting XML functional dependencies through formal concept analysis *Proc. of the 14th East European Conference on Advances in Databases and Information Systems* 595-8 Springer, Heidelberg

[6] Lee M L, Ling T W, Low W L 2002 Designing functional dependencies for XML *Proc. of the 8th International Conference on Extending Database Technology: Advances in Database Technology* 124-41 Springer-Verlag, London

[7] Liu J, Vincent M, Liu C 2003 Functional dependencies, from relational to XML *Proc. of 5th International Andrei Ershov Memorial Conference* 531-8 Springer, Heidelberg

[8] Liu J, Vincent M, Liu C 2003 Local XML functional dependencies *Proc. of 5th ACM CIKM International Workshop on Web Information and Data Management* 23-8 ACM, New York

[9] Vincent M, Liu L 2003 Functional dependencies for XML *Proc. of 5th Asian-Pacific Web Conference* 22-34 Springer, Heidelberg

[10] Vincent M, Liu J, Liu C 2004 Strong functional dependencies and their application to normal forms in XML *TODS* **29** 445-62

[11] Yan P, Lv T 2006 Functional dependencies in XML documents *Proc. of the 8th Asia Pacific Web Conference Workshop* 29-37 Springer, Heidelberg

[12] Hartmann S, Link S 2003 More functional dependencies for XML *Proc. of the 7th East European Conference on Advances in Databases and Information Systems* 355-69 Springer, Heidelberg

[13] Hartmann S, Link S, Trinh T 2010 Solving the implication problem for XML functional dependencies with properties *Proc. of the 18th Workshop on Logic, Language, Information and Computation* 161-75 Springer, Heidelberg

[14] Lv T, Yan P 2008 Removing XML data redundancies by constraint-tree-based functional dependencies *Proc. of ISECS International Colloquium on Computing, Communication, Control, and Management* 595-9 IEEE Computer Society, Washington, DC

[15] Lv T, Yan P 2007 XML normal forms based on constraint-tree-based functional dependencies *Proc. of Joint 9th Asia-Pacific Web Conference and 8th International Conference on Web-Age Information Management Workshops* 348-57 Springer, Heidelberg

[16] Vo L T H, Cao J, Rahayu W 2010 Discovering conditional functional dependencies in XML data *Proc. of the 22nd Australasian Database Conference* 143-52 Australian Computer Society, Sydney

[17] Wang D Z, Dong L, Sarma A D, Franklin M J, Halevy A Y Functional dependency generation and applications in pay-as-you-go data integration systems *Proc. of the 12th International Workshop on the Web and Databases*, http://www.cs.berkeley.edu/~daisyw/webdb09.pdf

[18] De S, Kambhampati S *Defining and mining functional dependencies in probabilistic databases* http://arxiv.org/pdf/1005.4714v2

[19] Sarma A D, Ullman J, Widom J Schema design for uncertain databases *Proc. of Alberto Mendelzon Workshop on Foundations of Data Management* http://ilpubs.stanford.edu:8090/820/

[20] Fan W, Libkin L 2002 On XML integrity constraints in the presence of DTDs *JACM* **49** 368-406

Authors

Ping Yan, born in December, 1972, Urumqi, Xinjiang Uygur Autonomous Region, China

Current position, grades: a professor in Anhui Agricultural University, PhD.
University studies: BSc degree in Applied Mathematics from Xinjiang University (1994), MSc degree in Applied Mathematics from Xinjiang University (1999), Ph.D degree in Applied Mathematics from Fudan University (2002)
Scientific interest: Her research interest fields include Neural Networks and Data management
Publications: more than 50 papers published in various journals and referenced conferences
Experience: She has teaching experience of 20 years, has completed 6 scientific research projects

Teng Lv, born in April, 1975, Datong, Shanxi Province, China

Current position, grades: an associate professor in Army Office Academy, PhD.
University studies: BSc degree in Computer Science from Artillery Academy (1997), MSc degree in Computer Science from Artillery Academy (2000), Ph.D degree in Computer Science from Fudan University (2003)
Scientific interest: Her research interest fields include Data management
Publications: more than 70 papers published in various journals and referenced conferences
Experience: He has teaching experience of 9 years, has completed 4 scientific research projects

Weimin He, born in December, 1973, Kunmin, Yunnan Province, China

Current position, grades: an assistant professor in UWSP, PhD.
University studies: BSc degree in Computer Science from Yunnan University (1995), MSc degree in Computer Science from Yunnan University (2000), Ph.D degree in Computer Science from University of Texas at Arlington (2008)
Scientific interest: Her research interest fields include XML Data Management, Information Retrieval and Peer-to-Peer Computing Data management
Publications: more than 30 papers published in various journals and referenced conferences
Experience: He has teaching experience of 9 years, has completed 5 scientific research projects

Information and Computer Technologies

Security evaluation model for the enterprise cloud services based on grey fuzzy AHP

Yu Bengong[1, 2], Wang Liu[1]*, Guo Fengyi[1]

[1]*School of Management, Hefei University of Technology, Hefei 230009*

[2]*Key Laboratory of Process Optimization and Intelligent Decision-making, Ministry of Education, Hefei 230009*

Abstract

This paper analyses the application status of cloud services to identify four factors that affect the security of enterprise cloud services (ECSs), including platform facilities, operational safety, operations management, and legal factors. Based on the four factors, the grey fuzzy analytic hierarchy process (GFAHP) is used to construct an evaluation model for the security of ECSs. An example is investigated to demonstrate the proposed model.

Keywords: enterprise cloud services, cloud safety, grey fuzzy AHP, evaluation model

1 Introduction

Inspired by the commercial operations of international network companies, the cloud service, as a new business model has been continuously refining and developing. Resources can be accessed anywhere on any device via the Internet. With more and more cloud platforms being applied, the security issues of the ECSs have arisen.

Recently many famous companies, such as Amazon [1], Google [2], and Sony, have paid much attention to the security of ECSs because of a large number of repeated security incidents. Specific security issues for the ECSs mainly include the technology and management in virtualized environment, the security issues of cloud service platforms, and etc. Those issues not only challenge the security technology and collaboration between the companies and service providers, but also pose threat to judicial supervision and privacy protection. Therefore, it is practically important to construct the security evaluation system of the ECSs.

The international research firm Gartner [4] enumerated seven major risks in the cloud computing which include the data access of the privileged users, auditability, data location, data isolation, data recovery, investigation support, long-term survival. FENG Deng-guo et al. [5] analysed the information security challenges of technologies, standards, regulatory and other aspects in the cloud computing. They also proposed the reference framework and investigated the key technology for cloud security.

FU Sha et al. [6] studied the information security risk assessment and presented a fuzzy analytic hierarchy process (AHP) methodology whose feasibility and effectiveness were demonstrated by an example. On the same issue, HUANG Jianxiong et al. [7] proposed a different evaluation index system and used a new model, which combined fuzzy AHP with grey evaluation method to evaluate the risk of information systems.

Cloud service has played a strategic role in the technical innovation of companies. A lot of research has been conducted, but there is little research about the security evaluation of the ECSs. This paper studies the security evaluation of the ECSs by which the security levels of the cloud service providers are evaluated. To avoid the subjectivity of experts, the grey theory and fuzzy AHP are combined to build the security evaluation model of the ECSs, which improve the accuracy of security evaluation.

2 Security evaluation model of the ECSs construction

There are many factors in cloud security. Huang Ying et al. [8] carried out research on security of cloud infrastructure. It observed the state-of-the-art of cloud infrastructure security, analysed security problems existing in cloud infrastructure from a wide perspective. In combination with a technical framework of cloud infrastructure security, it discussed principle key techniques of cloud infrastructure security. Liu Xiehua et al. [9] introduced the basic concept, and analyse the challenges of security model and policy in cloud computing. Ma Xiaoting et al. [10] described the characteristics and principles for the digital library based on cloud computing, discussed the factors affecting digital libraries securities, studied the threats and countermeasures of data security, application and

---

* *Corresponding author* e-mail wangliuatu@gmail.com

virtualization security for the digital library under cloud computing.

With a view to solving the practical problems in the application of cloud services, this paper construct an evaluation index model for the ECSs based on analysing and summarizing the affective factors.

2.1 AFFECTIVE FACTORS FOR THE SECURITY OF THE ECSS

The awareness and acceptance of users about the cloud service need to be improved because the applied models are diverse. The enterprises should reasonably analyse, actively participate even for the security risk, if they want to make full use of the cloud services to enhance their

competitiveness. The affective factors for the security of the ECSs are presented as follows:

2.1.1 Platform facilities

When enterprises use the cloud platform, the security of facilities provides the basic guarantees, including the security of the server, network and user identity management. Server security mainly refers to the security of the infrastructure, such as the keys being lost, the port listening to hijack accounts, and etc. Network security refers to the existing Internet DOS attack, hacker malicious behaviours such as network intrusions. And security of the user identity management includes the authenticity and reliability of user identity. It is possible that the users' information was stolen or false users can freely access the cloud.
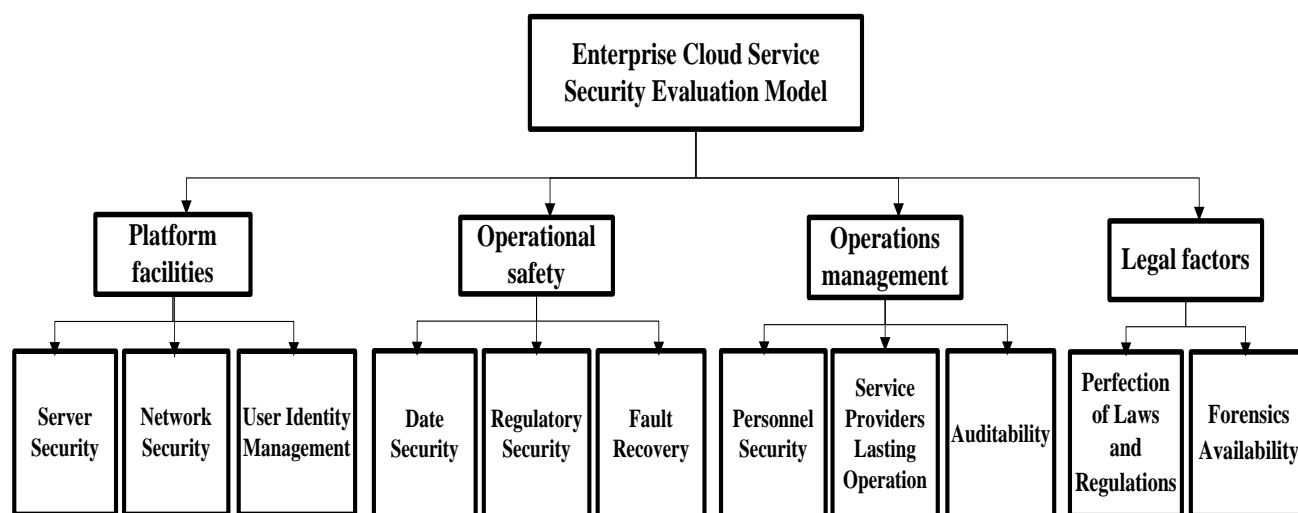
FIGURE 1 Security evaluation model of the ECSs

2.1.2 Operational safety

The security of the data resources is the most important in the cloud security. The regulation and recovery also need to be considered during the operation of services. The data security includes the security of the storage, migration, integrity, and etc. But there are still some security threats that the data isn't transferred completely or left history list. The security threats of regulatory mainly reflect that whether the security reports and the early warning system are submitted regularly. And the security threats of fault repair refer to the difficulty of restoring, retrieving or locating effectively when unknown security threats occurred.

2.1.3 Operations management

The specific model of cloud service strengthens the collaboration between users and providers. Nevertheless, the uncertainty of human resource and management increases the risk. In the operating process, the durable operation and auditability also need to be considered.

2.1.4 Legal factors

Cloud services make it complex to abide the laws and regulations. Different places have different rules. For example, some regulations require that certain data cannot be mixed with other data in a shared server or database; in some countries and regions, the citizens' data must be retained locally. As a large-scale resource centre, the cloud makes the data closer and breaks the geographical restrictions, but it brings the legal problems about data isolation and privacy protection at the same time. Besides, there are more security threats such as incomplete laws and regulations as well as the more difficultly forensic processes.

2.2 BUILDING SECURITY EVALUATION MODEL OF THE ECSS

Base on the mentioned factors and the principle of AHP [11], the model is established as shown in figure 1.

3 Security evaluation of the ECSs base on GFAHP

In order to reduce the large subjective errors in the evaluation and decision, this paper uses a new algorithm to comprehensively evaluate the security of the ECSs which combines grey theory with fuzzy AHP to form an integrated algorithm called the grey fuzzy AHP (GFAHP).

3.1 DETERMINE THE FUZZY CONSISTENT MATRIX AND THE WEIGHT SET

3.1.1 Determine the assembly of evaluation factors

The model is divided into 4 first-level indexes and 11 second-level indexes, respectively recorded as B_i, C_{ij}, index set $B = (B_1, B_2, \ldots, B_m)$, which, m is the total number of categories for the indicator, $B_i = (C_{i1}, C_{i2}, \ldots C_{in})$, $(i = 1, 2, \ldots n)$, where n is the sub-index number of categories of the category i.

$B = (B_1,\ B_2,\ B_3,\ B_4)$,
$B_1 = (C_{11},\ C_{12},\ C_{13})$,
$B_2 = (C_{21},\ C_{22},\ C_{23})$,
$B_3 = (C_{31},\ C_{32},\ C_{33})$,
$B_4 = (C_{41},\ C_{42})$.

3.1.2 Determine the weight set

First, we establish the fuzzy consistent matrix. Let the experts to adopt 0.1-0.9 scale method [12] on the various indicators to judge by the statistics and analysis, then we establish the fuzzy judgment matrix. Respectively obtain the upper layers of fuzzy consistent matrixes R, R_1, R_2, R_3, R_4 by the consistency of conversion.

Second, we do the consistency check by using fuzzy consistent judgment matrix properties on the judgment matrix.

Finally, we calculate the weight of each factor in fuzzy consistent matrix set. Known for each element of the fuzzy consistent judgment matrix $R = (a_1, a_2, \ldots, a_n)$, the weights were set as $w_1, w_2, \ldots w_n$, the weight calculation formula [13] is:

$$W_i = \frac{1}{n} - \frac{1}{2a} + \frac{1}{na}\sum_{k=1}^{n} r_{ik}, (i = 1, 2, \ldots, n)$$

And a is the parameter, here we take $a = \dfrac{n-1}{2}$. It can significantly reflect the difference and identification among the indicator elements.

After expert scoring and the calculation of the above steps, get the index system of weights set are shown in table 1 as follows:

TABLE 1 Index system of the total weight value

A	First level index	First level index weight	Second level index	Second level index weight	
				Single weight	Total weight
Security evaluation of the ECSs	B_1	0.2850	C_{11}	0.3134	0.0893
			C_{12}	0.3633	0.1035
			C_{13}	0.3233	0.0921
	B_2	0.3050	C_{21}	0.4334	0.1322
			C_{22}	0.3133	0.0956
			C_{23}	0.2533	0.0773
	B_3	0.2183	C_{31}	0.3233	0.0706
			C_{32}	0.3867	0.0844
			C_{33}	0.2900	0.0633
	B_4	0.1917	C_{41}	0.5300	0.1016
			C_{42}	0.4700	0.0901

3.2 GREY FUZZY COMPREHENSIVE EVALUATION

3.2.1 Determine the evaluation grade and sample evaluation matrix

According to evaluation requirements, the review set is divided into five grades. $V = (V1, V2, V3, V4, V5) =$ (Very safe, relatively safe, generally safe, not safe, unsafe). Suppose there are p evaluators, according to the range 0 to 10 score, get the evaluation sample matrix X:

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix},\ x_{ij} \text{ - the } j \text{ evaluator to } C_i$$

index of evaluation results.

3.2.2 Determine the evaluation of grey type

Determine the evaluation of grey type is to determine the rating number of the grey type evaluation, a grey degree and whitening weight function of grey number. Grey type must be defined by evaluation grade and qualitative

analysis. Divide evaluation of grey type into five categories and serial set of grey type evaluation is $k = \{1,2,3,4,5\}$. Then, make the whitening weight function $f_k (k \in \{1,2,3,4,5\})$ as follows [14]:

$$f_1(x) = \begin{cases} 0; x_{ij} < 0 \\ \dfrac{1}{9} x_{ij}; x_{ij} \in [0,9] \\ 1; x_{ij} \in [9, +\infty] \end{cases}, \quad f_2(x) = \begin{cases} 0; x_{ij} \notin [0,14] \\ \dfrac{1}{7} x_{ij}; x_{ij} \in [0,7] \\ -\dfrac{1}{7} x_{ij} + 2; x_{ij} \in [7,14] \end{cases},$$

$$f_3(x) = \begin{cases} 0; x_{ij} \notin [0,10] \\ \dfrac{1}{5} x_{ij}; x_{ij} \in [0,5] \\ -\dfrac{1}{5} x_{ij} + 2; x_{ij} \in [5,10] \end{cases}, \quad f_4(x) = \begin{cases} 0; x_{ij} \notin [0,6] \\ \dfrac{1}{3} x_{ij}; x_{ij} \in [0,3] \\ -\dfrac{1}{3} x_{ij} + 2; x_{ij} \in [3,6] \end{cases},$$

$$f_5(x) = \begin{cases} 0; x_{ij} \notin [0,5] \\ 1; x_{ij} \in [0,1] \\ -\dfrac{1}{4} x_{ij} + \dfrac{5}{4}; x_{ij} \in [1,5] \end{cases}.$$

3.2.3 Determine the grey number

The evaluation grey number of the index C_{ij} that belongs to the kth evaluation grey type is $n_{ijk} = \sum_{n=1}^{m} f_k(x_n)$

The overall evaluation grey number of C_{ij} that belongs to the all grey type is $n_{ij} = \sum_{k=1}^{5} n_{ijk}$.

3.2.4 Determine the grey evaluation weight

The kth grey evaluation weight of the jth evaluation index in the ith project is $y_{ijk} = n_{ijk} / n_{ij}$.

So we can structure the grey fuzzy weight matrix Y_i.

3.2.5 Calculate the grey fuzzy comprehensive evaluation matrix

Calculate the grey fuzzy comprehensive evaluation matrix of the ith project according to weight set and grey and fuzzy weight matrix. $Z_i = W_i \cdot Y_i; i\{1,2,3,4\}$.

3.2.6 Calculate the evaluation result

$Z = W * (Z_1, Z_2, Z_3, Z_4)^T$, Z is the result of the grey fuzzy comprehensive evaluation, according to the maximum membership degree principle to evaluate enterprise cloud service security.

4 Data and computation

This paper investigates the evaluation of application project of cloud email in a large manufacturing enterprise. It aims to comprehensively evaluate the security level of enterprise cloud mailbox project.

For various factors of security, the evaluation team adopts the form of scoring table. The five team members

Include the senior leadership of the manufacturing companies, IT department members, cloud computing experts. Scores as shown in Table 2.

TABLE 2 Enterprise cloud services security index mark sheet

	C_{11}	C_{12}	C_{13}	C_{21}	C_{22}	C_{23}	C_{31}	C_{32}	C_{33}	C_{41}	C_{42}
1	8	8	7	7	6	7	6	8	5	5	4
2	7	6	4	7	7	6	9	9	8	7	2
3	8	7	7	8	10	7	5	6	7	5	5
4	8	6	8	7	6	8	6	9	4	6	6
5	6	8	7	7	9	3	5	5	6	7	3

4.1 CALCULATE THE GREY NUMBER

The sample matrix is calculated to get the grey numbers of C_{11} index according to the formula

$$n_{111} = \sum_{n=1}^{5} f_1(x_{n1})$$
$$= f_1(8) + f_1(7) + f_1(8) + f_1(8) + f_1(6) = 4.111$$

Similarly $n_{112} = 4.429, n_{113} = 2.600, n_{114} = 0, n_{115} = 0$.

Finally $n_{11} == \sum_{j=1}^{5} n_{11j} = 11.14$.

4.2 CALCULATE THE GREY FUZZY WEIGHT MATRIX

To C11 index, the grey evaluation weights are:
$y_{111} = n_{111} / n_{11} = 4.111 / 11.14 = 0.369,$
$y_{112} = 0.398, y_{113} = 0.233, y_{114} = 0, y_{115} = 0$.

The grey evaluation weight vector is:
$y_{11} = (y_{111}, y_{112}, y_{113}, y_{114}, y_{115})$
$= (0.369, 0.398, 0.233, 0, 0)$

Similarly $y_{12} = (0.344, 0.391, 0.265, 0, 0)$,
$y_{13} = (0.305, 0.369, 0.250, 0.056, 0.021)$

Finally

$$Y_1 = \begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \end{bmatrix} = \begin{bmatrix} 0.369 & 0.398 & 0.233 & 0 & 0 \\ 0.344 & 0.391 & 0.265 & 0 & 0 \\ 0.305 & 0.369 & 0.250 & 0.056 & 0.021 \end{bmatrix},$$

$$Y_2 = \begin{bmatrix} 0.343 & 0.417 & 0.240 & 0 & 0 \\ 0.391 & 0.381 & 0.228 & 0 & 0 \\ 0.339 & 0.317 & 0.229 & 0.076 & 0.038 \end{bmatrix},$$

$$Y_3 = \begin{bmatrix} 0.293 & 0.328 & 0.323 & 0.057 & 0 \\ 0.351 & 0.330 & 0.291 & 0.028 & 0 \\ 0.274 & 0.328 & 0.295 & 0.082 & 0.021 \end{bmatrix},$$

$$Y_4 = \begin{bmatrix} 0.271 & 0.349 & 0.326 & 0.054 & 0 \\ 0.173 & 0.222 & 0.280 & 0.208 & 0.117 \end{bmatrix}.$$

4.3 CALCULATE THE SECOND LAYER OF GREY FUZZY

The comprehensive evaluation vectors can be calculated by the formula: $Z_i = W_i \cdot Y_i$.

The results are:

$Z_1 = (0.3292 \ 0.3861 \ 0.2501 \ 0.0181 \ 0.0068)$

$Z_2 = (0.3570 \ 0.3804 \ 0.2335 \ 0.0193 \ 0.0096)$

$Z_3 = (0.3099 \ 0.3288 \ 0.3025 \ 0.0530 \ 0.0061)$

$Z_4 = (0.2249 \ 0.2893 \ 0.3044 \ 0.1264 \ 0.0550)$

4.4 CALCULATE THE FIRST LAYER OF GREY FUZZY COMPREHENSIVE EVALUATION VECTOR

$$Z = W \cdot \begin{bmatrix} Z_1 \\ Z_2 \\ Z_3 \\ Z_4 \end{bmatrix}.$$

$= (0.3135 \ 0.3533 \ 0.2669 \ 0.0502 \ 0.0167)$

The results show that the maximum of Z is 0.3533, according to the maximum membership degree principle, the cloud email security level is identified as "relatively safe".

5 Conclusion

This paper investigates the assessment of the security of ECSs.

First, the influence factors of ECSs are analysed to construct an evaluation index system.

Second, based on the index system, the paper give the specific modelling steps:

(1) Establish a fuzzy consistent matrix and the value of the index weighting factor Wi;

(2) Create a score sample matrix and make it be whitened;

(3) Use the grey fuzzy comprehensive evaluation.

Finally, the security evaluation of cloud services in a large manufacturing enterprise is investigated to demonstrate the validity and effectiveness of the proposed model.

With the development of ECSs, their security evaluation index system will be correspondingly developed to stimulate new security evaluation models. Therefore, further researches will be devoted to a further and deeper understanding of the index system to construct the more comprehensive models.

References

[1] Amazon Security Bulletins [EB/OL] http://aws.amazon.com/security/security-bulletins/2009
[2] Google [EB/OL] http://googledocs.Blogspot.Com/2009/03/just-to-clarify.html 2009
[3] Microsoft. [EB/OL] http://www.Microsoft.Com/technet/security/2010
[4] Gartner, Teleworking in the cloud: Security risks and remedies 2010 [2010-11-11] http://www.gartner.com/resources/167600/137661/teleworking_in_the_cloud_sec_167661.pdf
[5] Feng D G, Zhang M, etc. 2011 Study on Cloud Computing Security *Journal of Software* **22** 71-83 2011
[6] Fu S, Liao M H, etc. 2012 Research and Exploration on Fuzzy AHP in the Field of Information Security *Journal of The China Society For Scientific and Technical Information* **31**(10) 1105-9
[7] Huang J X, Ding J L 201 Evaluation model for information system risk based on fuzzy analysis *Computer engineering and design* **33**(4) 1285-9

[8] Huang Ying, Shi Wenchang 2011 Survey of Research on Cloud Infrastructure Security *Computer Science* **38**(7) 24-9
[9] Liu Kaihua, Li Xiong 2011 Analyse the Security Model and Policy of Cloud Computing *Computer Knowledge and Technology* **7**(8) 1150-1
[10] Ma X T, Chen C 2011 On the Threat to the Security of Digital Library Information Resource and Its Tactics: Under the Environment of Cloud Computing *Information and Documentation Services* **2** 55-9
[11] Xu S B 1988 *The principle of analytic hierarchy process* Tianjin: Tianjin University Publisher
[12] Zhang J J 2000 Fuzzy analytic hierarchy process *Fuzzy Systems and Mathematics* **6** 80-8
[13] Lv Y J 2002 Fuzzy AHP Sort based on Fuzzy Consistent Matrix *Fuzzy Systems and Mathematics* **2** 81-2
[14] Liu S F, Xie N M 2011 New grey evaluation method based on reformative triangular whitenization weight function *Journal of Systems engineering* **26**(2) 244-50

[15] Yang X K, Wang Y Z 2012 Government Information Sharing Capability Evaluation based on Grey Fuzzy Theory *Journal of Dalian University of Technology* **52**(2) 297-303

[16] Wang J S, Fu Y, etc. 2011 Information System Security Risk Assessment based on Fuzzy Neural Network *Wuhan University of Technology (Transportation Science & Engineering Edition)* **35**(1) 51-4, 58

[17] Chen Y, Katz VPARH 2010 What's New About Cloud Computing Security? *Electrical Engineering and Computer Sciences*

[18] Carlin S, Curran K 2011 Cloud computing Security *International Journal of Ambient Computing and Intelligence (IJACI)* **3**(1) 14-9

[19] Jamil D, Zaki H 2001 Cloud computing security *International Journal of Engineering Science and Technology* **3**(4) 3478-83

[20] Sriram I, KhajehHosseini A 2010 Research Agenda in Cloud Technologies *Technical Report*

[21] Armbrust M, Fox A, Griffith R, et.al. 2009 Above the clouds: A Berkeley view of cloud computing *Communication Magazine*

[22] Chang D Y 1996 Applications of the extent analysis method on fuzzy AHP *European journal of operational research* **95**(3) 649-55

[23] Gumus A T 2009 Evaluation of hazardous waste transportation firms by using a two step fuzzy- AHP and TOPSIS methodology *Expert Systems with Applications* **36**(2) 4067-74

[24] Chiou H K, Tzeng G H 2001 Fuzzy hierarchical evaluation with grey relation model of green engineering for industry *International Journal of Fuzzy Systems* **3**(3) 466-75

[25] Ho W, Xu X, Dey P K 2010 Multi-criteria decision making approaches for supplier evaluation and selection: A literature review *European Journal of Operational Research* **202**(1) 16-24

[26] Marciano F A, Vitetta A 2011 Risk analysis in road safety: An individual risk model for drivers and pedestrians to support decision planning processes *International Journal of Safety and Security Engineering* **1**(3) 265-82

[27] De Wrachien D, Mambretti S 2011 Mathematical models for flood hazard assessment *International Journal of Safety and Security Engineering* **1**(4) 353-62

[28] Bertino E, Paci F, Ferrini R 2009 Privacy - preserving digital identity management for cloud computing *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering* **32**(1) 21-7

Authors

Yu Bengong, born on December 29, 1971, Hefei, Anhui, China

Current position, grades: Director of the Department of information management.
University studies: Hefei University of Technology
Scientific interest: decision-making science and technology, information system.
Experience: Yu Bengong received the B.Sc. degree (1994) in Computer Science, the M.Sc. degree (2005) and Ph.D. degree (2011) in Management Science and Engineering from Hefei University of Technology, Hefei, China. He is currently working as an professor for School of Management at Hefei University of Technology, Hefei, Anhui, China.
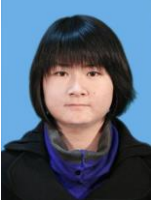
Wang Liu, born on December 14, 1989, Huanggang, Hubei, China

Current position, grades: Master of Hefei University of Technology, grade two
University studies: Hefei University of Technology
Scientific interest: information management, information system
Experience: Wang Liu is currently studying for the master degree, and received the B.Sc degree (2012) in Information Management and Information System from Hefei University of Technology.

Fengyi Guo, born on December 23, 1987, Luoyang, Henan, China

Current position, grades: Master of Hefei University of Technology, graduate
University studies: Hefei University of Technology
Scientific interest: information management, information system
Experience: Guo Fengyi received the master degree (2013) in Information Management and Information System from Hefei University of Technology, Hefei, Anhui, China.

A novel distributed network database mapping scheme analysis

Liangong Song*, Ke Han

North China University of Water Resources and Electric Power, Zhengzhou, Henan, China

Abstract

This research is based on distributed client-server model, and intends to add lacking fields in the target database by mapping in order to complete the database for subsequent analysis and tab the maximum economic value of a single database. This research takes China Industry, Commerce and Service Census Database (TTICSCDB) as the target database, and China Technology Innovation Survey Database (TTTISDB) as the assistant database. Then this research will employ techniques to mapping common fields from the assistant database to the target database and evaluate the accuracy of mapping. We will prove from the statistical perspective the simulation and accuracy of extension field of the network database proposed in this research. This network database will present its advantages in many ways.

Keywords: database, mapping, statistics, expanding

1 Introduction

In the age of knowledge economy, it is important to make sure of the accuracy of the information from the database [1-3]. In the course of data mining, one cannot get accurate or useful knowledge without correct or accurate fields in the database [27, 28]. Therefore, the extensibility and reliability of a database are significant [1, 4-7]. With the development of the internet, database has ever become independent from the computer system and developed into a giant and meticulous system group. Thus, this research is based on distributed client-server model [8-11], and intends to add lacking fields in the object database by mapping in order to complete the database. This system architecture includes five parts. First, distributed network database. Second, automation pre-processing mechanism. Third, create module relation of two network database. Fourth, map the module. Fifth, threshold statistics and testing mechanism. The first part will present the whole distributed network database. The second part will interpolate missing values and correct error values, and unify field and format. The third part will do statistics of the index field to verify the reliability of the network database. The fourth part will base relation on the index field and apply it to four disaggregated model namely, linear network, probabilistic neural network, radial basis function network and back-propagation neural network. The fifth part will create a statistics and testing mechanism and set up its threshold and select the best approach and the mapping module with the high accuracy from the four disaggregated model, and map the field to the object database.

The rest of the paper is organized as: Section 2 will introduce the whole network database, its extensibility and limitations. Section 3 will introduce the background information of two database. Section 4 will account details of the automation pre-processing mechanism. Section 5 will account practical method of establishing module relation. Section 6 will compare four ways of mapping and analyse their practical result which includes accuracy and related efficacy analysis. Section 7 will present the conclusion of the re-search and discuss further improvement. Section 8 will conclude the contribution of this research.

2 An overview on system architecture

2.1 LIMITATIONS OF THE DATABASE

Field for mapping should be equipped with effective index. Otherwise, the disaggregated model will lack its accuracy and the extension field will lack its reliability. Matrix of the object database and the assistant database should be the same as this research adopts Mann-Whitney U test [12-14] to judge their similarity. If the matrix are not the same, it may result in a disordered corresponding relationship that the mapping module cannot apply to the object database.

2.2 DISTRIBUTED NETWORK DATABASE

Distributed network database is more and more important with the popularity of the internet. The distributed network database designed by us is easily established on other software database. Object can be taken as a basis of data processing and data exchange but real data exchange is operated through communication hardware in the way of bit stream transmission in which the object and Bit stream switch to each other under the coordination of Agent, DNDM and Operation System.

Corresponding author e-mail: slg@ncwu.edu.cn

2.3 AUTOMATION PRE-PROCESSING MECHANISM

Processing missing values: This research will employ interpolating method to compensate missing fields. If there is too much missing in the data or the field (this research define it as 30%), then the data or the field with missing values will be deleted.

Processing error values: As database cannot avoid error values when transferring documents or inputting data, error values should be judged while taking definition of the domain value into consideration. Regularity and other related characteristics are of error values are expected to find out to return to true values. If error values cannot be returned to true values, then they shall be considered as missing values.

Switch of field and format: as the investors, purpose, time and other factors of two database vary, the domain name or the definition of domain value vary as well. As a result, the switch of domain value is needed to reach the same definition of two database.

Switch of variables: as this research targets at categorical variables, continuous variables in the assistant database should be switched to categorical variables to meet the basis assumption.

2.4 CREATE MODULE RELATION

To avoid the complicated computing of large number of data by way of Mann - Whitney U test and the deviation resulted from sampling, draw from the target database m set of samples so that the number of samples is close to that of the assistant database. Then select the pre-determined indices (refer to $S'_I(i)$ and $S_I(i)$ respectively) when pre-processing the target database and the assistant database by way of comparing fields, and apply Mann-Whitney U test of virtual-index to samples and data of the assistant database to see if their data distribution are the same and if they are qualified to be taken as real virtual-index.

To avoid deviation of sampling, we should consider every set of samples [29]. Thus, this research designs a variable $Temp_m$, representing the different distribution between $S'_I(i)$ and $S'_I(i)$ in set m of samples. When P-value>α,$Temp_m=1$ representing zero difference in distribution between $S'_I(i)$ and $S_I(i)$; otherwise, $Temp_m=0$ representing different distribution between $S'_I(i)$ and $S_I(i)$. Finally, calculate $Temp_t = \sum_{m=1}^{k} Temp_m$, and k represents the number of set of samples. Set θ, if $Temp_t \geq \theta, 0 \leq \theta \leq k$, and θ represents the minimum customized number of set of samples that accept H_0 , then $S'_I(i)$ and $S_I(i)$ can be taken as virtual-index. The purpose of setting θ lies in the effort to make sure that when the number of samples is smaller than the matrix, the error judgment results from the failure of representing matrix by samples can be lowered by imp roving the value of θ.

2.5 MAPPING MODULE AND THRESHOLD STATISTICS AND TESTING MECHANISM

Then, apply virtual-index of the module relation to mapping module. A mapping module includes selecting target filed, operating classification model, and analysing accuracy and executing mapping.

In the course of processing, select mapping field from the assistant database and input it into the target database, then take the selected mapping field as the output layer and the tested virtual-index VI of the assistant database as the input layer, establish classification model by way of linear network, probabilistic neural network, radial basis function network and back-propagation neural network. Then select the classification model with the highest accuracy as the mapping model and test its accuracy [30, 31].

If it has enough accuracy, then replace VI in the mapping model by virtual-index VI$'$ of the target database, and take VI$'$ as the output layer of the mapping model, the result of which shall be taken as the value of the extension field $MI'(i)$ for the purpose of extension. After abovementioned operation, every time when going through the mapping module, one mapping field can be mapped and a new extension field in the target database can be produced so that the purpose of database extension is reached.

The database will be based on China Industry, Commerce and Service Census Database (TTICSCDB) as the target database with the data capacity of 186,550 and China Technology Innovation Survey Database as the assistant database with the data capacity of 3,356. As for efficacy measurement index, we will take common field in two database as mapping fields and compare the accuracy of true value with that of mapping value as a measurement of efficacy. As TTICSCDB and TTTISDB only share common in "number of employers", "occupation", "whether to conduct research and development" and "turnover in 1990", thus this research will take two of the four fields as index and the other two to test simulation. Finally, this research will take different fields as mapping fields to conduct the experiment twice to test the accuracy, and analyse simulation of extension fields produced therefrom to see advantages and disadvantages of four classification models.

The target database is based on TTICSCDB in 1990. TTICSCDB has a data capacity of 186,550 and 99 field variables. The assistant database is based on TTTISDB in 1990. TTTISDB has a data capacity of 3,365 and 189 field variables. By pre-processing, there are four common fields, namely, "number of employers", "occupation", "whether to conduct research and development" and

"turnover in 1990". The background of two databases is shown in Table 1:

TABLE 1 Experimental data

	The target database	The assistant database
Name	TTICSCDB	TTTISDB
Data capacity	186,550	3,356
Number of field	99	189
Common fields	Number of employers Occupation Turnover in 1990 Whether to conduct research and development	

3 Automation pre-processing mechanism

To begin with, deal with missing values and error values. Interpolate fields with a missing value of less than 30% by Hot-deck imputation [15-18] and correct error fields. If error fields cannot be corrected, then deal with them as missing values. After this operation, the number of fields is reduced to 102 from 189, and 87 of them are deleted because of over-missing.

Next, deal with common fields. As common fields can be taken as virtual-index in this research for measuring the simulation of the fields, process each of the four common fields, erase deviation and increase the accuracy of data so that the data type and domain value of the two database are the same. The processing method is shown in Table 2:

TABLE 2 Processing common fields

	Type of domain value		Processing method
	TTICSCDB	*TTTISDB*	
Number of employers	Record the actual number of employers.	Divide the number of employers into five classes, and use 1-5 to represent.	Reclassify the number of employers in TTICSCDB on the basis of five classes of TTTISDB and use 1-5 to represent.
Occupation	Divide all enterprises in China into 540 classes, and record them respectively.	Divide all enterprises in China into 18 classes, and record them by specific numbers.	Reclassify 540 enterprises in TTICSCDB into 18 classes defined in TTTISDB.
Turnover in 1990	Record an actual turnover in 1990 with the unit of thousand Yuan.	Record an actual turnover in 1990 with the unit of thousand Yuan.	According to the standard of turn-over in Standards for small and medium enterprises by Ministry of Economics, enterprises with a turnover below 100 million Yuan are small and medium enterprises, and will be recorded a "1", enterprises with a turnover of above 100 million Yuan will be recorded a "2".
Whether to conduct research and development	Record an actual expenses for R&D in 1990 with the unit of thousand Yuan.	If it has, record a "1", otherwise, record a "0".	If expenses for R&D in TICSCDB is above 0, it means R&D is conducted in this enterprise, and record "1", otherwise, record "0".

4 Setting up the modular

Only four fields are commonly shared by TICSCDB and TTISDB. To evaluate simulations of the expanding fields mapped by the schema of this research, numerical value and the number of types of the four fields are measured. Two fields, "number of staff" and "occupation" which have most types respectively, are prepositioned as index fields. The other two fields, "turnovers in 1990" and "whether to develop" are positioned as searching fields to measure the closeness of mapping value and true value and whether the simulation is effective.

Researching the TTISDB used the measurement of stratified sampling based on number of staff of the company. Thus in order to avoid the impact of this measurement, the research process is designed based on the proportion of the staff sampled in TTISDB, with 5 times stratified sampling of staffs in TICSCDB. Five sets of samples are collected with about 3,300 samples in each set. This research assumes that the prepositioned fields ("number of staff" and "occupation") are index fields, thus, these two fields of the two databases will be tested through Mann-Whitney U to prove whether they can be deemed as objective index fields. However, every set of samples of TICSDB contains 3,300 samples, which only

accounts for 2% of the total number. θ is set as 4 to resolve the problem that the deficiency of sample number may result in the failure of representing the whole and the misinterpretation of the index. The outcome of the test is that the Temp_t numbers of the two fields are 5, which demonstrates that these two fields can be regarded as indexes.

4.1 RESEARCH STEPS

In the following research process, "turnovers in 1990" and "whether to develop" are positioned as searching fields, while "number of staff" and "occupation" are prepositioned as index fields. The research focuses on the classification model accuracy, sample accuracy and TICSDB accuracy. Measuring the accuracy will be used to determine whether the simulation of the schema is of high rate and whether there exist the absolute advantages and disadvantages of the accuracies of these four fields. In the Experiment 1, "whether to develop" in the assistant database is applied as the mapping fields, while "number of staff" and "occupation" is used as indexes. The input layer of classification model is "number of staff" and "occupation", while the output layer is "whether to

develop". Four kinds of classification models, linear networks (LN), radial basis function networks (RBF) [20, 21], probabilistic neural networks (PNN) [22-24], back propagation neural networks (BNN) [25,26] are measured to test the model accuracy $P_C(X)$, sample accuracy $P_S(X)$ and mapping accuracy $P_M(X)$. The data of accuracy P(X) measured will be compared to the mapping data with the equations below:

$$P_i(X) = \begin{cases} 1, & if \quad X_i^{true} = X_i^{mapping} \\ 0, & if \quad X_i^{true} \neq X_i^{mapping} \end{cases}, \qquad (1)$$

$$P(X) = \sum_{i=1}^{n} P_i(X)/n, \qquad (2)$$

where X_i^{true} represents the true value of data i in the field X. $X_i^{mapping}$ represents the mapping value of data i in the field X. $P_i(X)$ represents the consistency of true value and mapping value of data i in the field X. If the outcome is 1, they are consistent. 0, not; P(X) represents the ration of true value and mapping value of data i in the field X.

Thus, the accuracy of the classification models $P_C(X)$ represents the accuracy of the model, which are tested. The accuracy of the true value and estimating value of tested data. The sample accuracy $P_S(X)$ represents the consistency of true value and map-ping value after the model is mapped using classification models, which have been tested. Mapping accuracy $P_M(X)$ is the consistency of true value and mapping value after the mapping of target database TICSCDB based on the tested classification models as map-ping models. It is also the measurement to tell whether simulation is effective.

In Experiment 2, "turnover in 1990" in the assistant database is applied as the mapping fields, while "number of staff" and "occupation" is used as indexes. The input layer of classification model is "number of staff" and "occupation", while the output layer is "turnover in 1990". Four kinds of classification models, linear networks (LN), radial basis function networks (RBF) [20, 21], probabilistic neural networks (PNN) [22-24], back propagation neural networks (BNN) [25,26] are measured to test the model accuracy $P_C(X)$, sample accuracy $P_S(X)$ and mapping accuracy $P_M(X)$. The definitions of, and are the same as they are defined in experiment 1.

4.2 EXPERIMENT 1: SET "TO DEVELOP OR NOT" AS THE FIELD FOR MAPPING

Set "To develop or not" as MI, and "The number of the staff" and "category of employment" as VI, and randomly draw 50% bases from the database as training data, the other 50% as test data, and respectively apply four

methods, namely PNN, LN, BNN, RBF for classification model training.

According to the diagrams showed above, each model has its own numbers of input layers, hidden layers and output layers. The data are tabulated for clearer compare into Table 3 below:

TABLE 3 Number of Neurons of the Classified Model for Experiment 1

	Number of Input Layer	Number of Hidden layer 1	Number of Hidden Layer 2	Number of Hidden Layer 3
PNN	23	1678	2	2
BNN	23	6	-	1
LN	23	-	-	1
RBF	23	27	-	1

Then map the four models according to the sample and TICSCDB and calculate the accuracy $P_S(X)$ of the sample, mapping accuracy $P_M(X)$ and overall accuracy by comparing the databases with the true values in the original database. The consequence is showed in the Figures 1 and Figure 2.
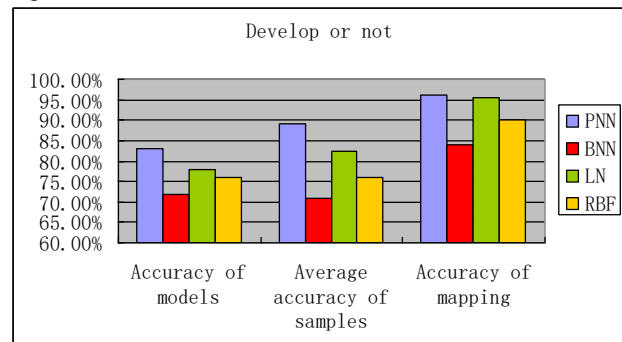
FIGURE 1 Bar chart for the consequence of Experiment 1
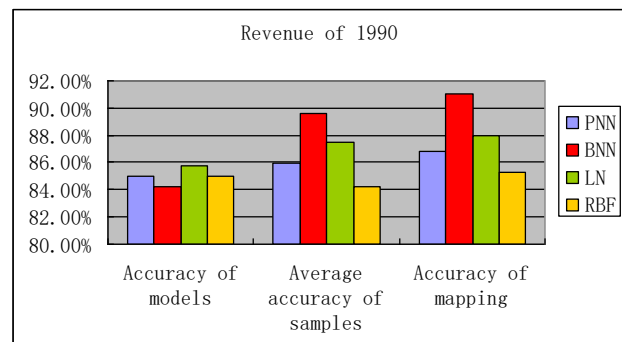
FIGURE 2 Bar chart for the result of Experiment 2

Conclusions of Experiment 1:

1) The ranking consequences of the accuracy $P_C(X)$ and mapping accuracy $P_M(X)$ of a classification model of the four classification models are the same, thus it is feasible to use the accuracy $P_C(X)$ of a classification model to assess whether the classification model can be used for judging which classification model is a mapping model.

2) According to the consequence, all the mapping accuracy are over 84.6%, which indicates that the similarity between the expanded field and the true values

is very high and also proves that the expanded fields produced under the structure of this system possesses a very high simulation.

3) In the four classification models, PNN fits the Experiment 1 for the most, while BNN does the least.

4.3 EXPERIMENT 2: SET "THE REVENUE OF 1990" AS THE FIELD FOR MAPPING

Set "the Revenue of 1990" as MI, and "The number of the staff" and "category of employment" as VI, and respectively apply four methods, namely PNN, LN, BNN and RBF for classification model training. When conducting the classification model training, in order to avoid the influence on the consequence of the accuracy $P_C(X)$ of the classification model caused by dealing with lost data, as well as to allow the material to wait for the computation so as to avoid inaccuracy caused by interpolations, the missing data of the field "the revenue of 1990" are ignored. And finally use the 2984 data of TTISDB for the training of classification models, and randomly draws 50% bases from the auxiliary database as training data, the other 50% of the auxiliary database as test data for the training of classification methods, and map the four models according to the sample and TICSCDB (this process is similar to that of Experiment 1. The accuracy is showed in Figure 3.
Conclusions of Experiment 2:

1) In Experiment 2, the difference of accuracy $P_C(X)$ between each two classification models is not as big as that of the Experiment 1. The difference between the lowest accuracy and the highest accuracy is only 0.4%. Thus, the best classification model for mapping model cannot be determined through Experiment.

2) The rank of the accuracy $P_S(X)$ and mapping accuracy $P_M(X)$ of the samples of the four classification models are the same. Thus it proves that the distribution of the sample is similar to that of parent population, and what caused the different ranking of mapping accuracy and the classification model accuracy of the different classification methods is the error between the target database and auxiliary database.

3) The difference in the rank of accuracy $P_C(X)$ and mapping accuracy $P_M(X)$ of classification model of the four classification models is a problem deserves further study and discuss.

5 Discussion

5.1 DISCUSSION ON ACCURACY

According to the two experiments above, the mapping accuracy for the four classification methods are showed in Figure 3.

No matter apply, which classification model as the mapping model, the accuracy is all very high (more than

84.6%), which proves that the system put forward in this study is a good structure for database expansion, and the expanded fields estimated by the system structure, which is put forward in this study, possess a very high simulation.

If order the four models in the two experiments by the mapping accuracy from the highest to the lowest.
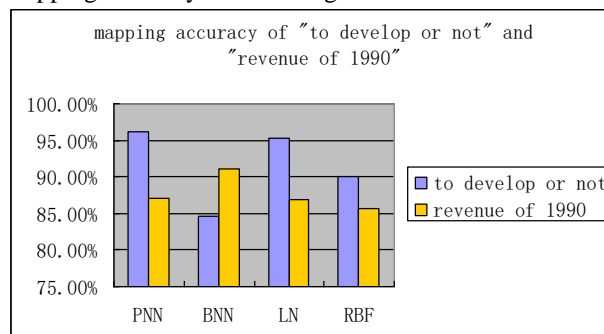
FIGURE 3 Bar chart of mapping accuracy

For Experiment 1, the classification model with the highest accuracy is PNN, up to 96.1%. But in Experiment 2, the accuracy of PNN is only 86.8%, ranked the last but one. While in Experiment 2, the classification model with the highest accuracy is BNN, up to 90.9%. But in the Experiment 1, the accuracy of BNN is only 84.6%, ranked last in accuracy of the four classification models.

Thus, there is no absolute inferior or superior in the four methods and there are different databases fit for each. So the process for the highest mapping accuracy cannot be simplified into a single model. And building several classification models and choosing the fittest model for mapping is a good method.

It's also found in the study that when the accuracy of classification model of different classification models are similar, the fittest classification model for mapping, because the error between two databases might be bigger than the difference between classification models. Thus, how to choose the mapping model effectively is also a problem deserves to study and discuss.

5.2 LIMITATIONS OF THE STUDY

The model does not fit for non-category data: as mentioned repeatedly in this study, the system structure and the methods applied in this study are data aiming at category field and does not fit non-category data.

The difference of indexes between two databases cannot be valued: the index between two databases is verified by Mann-Whitney U. Although the testing power can be controlled by adjusting level of significance α, the influence caused by the difference of P-value cannot be measured after being judged as an index. Thus, it is not sure to ignore the difference between index or not.

6 Conclusion

The four conclusions of this study are as followed:

1. To use the available resources for mutual expansion: by expansion of available databases, the data of an

enterprise can be effectively used and to avoid unsatisfactory subsequent analysis caused by missing of some field.

2. To reduce the cost of reinvestigation of database: by the method put forward in the study, the needed field can be expanded from another database of the enterprise, which can spare the time and cost of reinvestigation.

3. Databases with the same parent population can interflow so to make the subsequent analysis more efficient. By the technique in this study, the databases with the same parent population can interflow. So it will be easier to apply several databases in one analysis, and the inadequacy of relative fields needed frequently happened before in single database analysis can be avoided.

4. It enables the enterprise to know the situation of rivalries by their own databases: this study is feasible for all the databases with the same parent population. The custom group of competitive enterprises is similar. If an enterprise needs to know the information of its rivalry, it can draw some of the characteristics of customs of the rivalry. Then map the field of its own database into the database gained by investigation. Thus the overall characteristic of custom group and the direction of management and other.

Information of the rivalry can be gained effectively, and the plan to cope with it will be easy to make.

In the future, it will improve according to the following directions:

1. To expand its applied range: the method put forward in this study only aims at expansion of field of category data and cannot be applied to assist the field of continues data, text data, time series data, etc. Thus, in order to make is applicable for expansion of fields of different kinds of database, studies on other methods of fields expansion is an important direction in the future.

2. To increase the accuracy of classification model: how to increase the accuracy of classification model is also an important direction of study in the future. It can be improved by modifying the existing theory or citing the improved theories of other researchers, or even creating new methods for classification only if it can increase the overall accuracy of the system.

3. To increase the complexity of the index between two databases: if significance level can be tripped and do not need to factitiously examine whether the significances of the two fields are the same beforehand. Then one-to-many indexes can be built according to the distribution of values among several fields, thus to build between two databases a one-to-many index, for example: $f(\mathrm{x}_3) = f(\mathrm{y}_1) + f(\mathrm{y}_7) + c$, x_i indicates the field of the auxiliary database, y_i indicates the fields of the target database, c is a constant.

4. Add reference value for accuracy: it might be feasible to indicate the accuracy of classification model of the mapping model applied in each mapping field. Although there can be errors among indexed and the conformity between two databases, it offers the user an objective accuracy for subsequent analysis so as to judge whether this system deserves to apply.

References

[1] Ramakrishnan R, Gehrke J 2000 Database management systems *Osborne/McGraw-Hill*

[2] Gollub J, Ball C A, Binkley G, Demeter J, Funkelstein D B, Herbert J M, Hernandez-Boussard T, Jin H, Kaloper M, Matese J C, Schroeder M, Brown P O, Botstein D, Sherclock G 2003 The Stanford Microarray Database: data access and quality assessment tools *Nucleic Acids Research* **31**(1) 94-6

[3] Miles M B, Huberman A M 1994 Data management and analysis methods *Sage Publications*

[4] Agrawal R, Imielinski T, Swami A 1993 *IEEE Transactions on Knowledge and Data Engineering* **5**(6) 914-25

[5] O'Leary D E 1993 The impact of data accuracy on system learning. *Journal of Management Information Systems* **9**(4) 83-98

[6] Palmer D C 1998 Method for determining database accuracy *US Patent* No 5 787 443

[7] Dierickx I, Cool K 1989 Asset stock accumulation and sustainability of competitive advantage *Management science* **35**(12) 1504-11

[8] Özsu M T, Valduriez P 2011 Principles of distributed database systems *Springer*

[9] Bernstein P A, Goodman N 1981 Concurrency control in distributed database systems *ACM Computing Surveys (CSUR)* **13**(2) 185-221

[10] Hector G C 1983 Using semantic knowledge for transaction processing in a distributed database *ACM Transactions on Database Systems (TODS)* **8**(2) 186-213

[11] Bell D, Grimson J 1992 Distributed database systems *Addison-Wesley Longman Publishing Co Inc.* Boston MA USA

[12] Ruxton G D 2006 The unequal variance t-test is an underused alternative to Student's t-test and the Mann–Whitney U test *Behavioral Ecology* **17**(4) 688-90

[13] Hart H L A 1963 Law, liberty and morality Stanford CA USA *Stanford University Press*

[14] Rosner B, Grove B 1999 Use of the Mann-Whitney U-test for clustered data *Statistics in Medicine* **18**(11) 1387-400

[15] Rao J N K, Shao J 1992 Jackknife variance estimation with survey data under hot deck imputation. Biometrika **79**(4) 811-22

[16] Sande I G 1983 Hot-deck imputation procedures *Incomplete data in sample surveys* **3** 334-50

[17] Andridge R R, Little R J A 2010 A Review of Hot Deck Imputa-tion for Survey Non-response *International Statistical Review* **78**(1) 40-64

[18] Kim J K, Fuller W 2004 Fractional hot deck imputation *Biometrika* **91**(3) 559-78

[19] Chen W-K 1993 Linear Networks and Systems: Algorithms and Computer-Aided Implementations *World Scientific*

[20] Chen S, Cowan C F N, Grant P M 1991 *IEEE Transactions on Neural Networks* **2**(2) 302-9

[21] Park J, Sandberg I W 1991 Universal approximation using radial-basis-function networks *Neural computation* **3**(2) 246-57

[22] Specht D F 1990 Probabilistic neural networks *Neural networks* **3**(1) 109-18

[23] Specht D F 1990 *IEEE Transactions on Neural Networks* **1**(1) 111-21

[24] Specht D F 1992 Enhancements to probabilistic neural networks *International Joint Conference on Neural Networks* **1** 761-8

[25] Hecht-Nielsen R 1989 Theory of the back propagation neural network *International Joint Conference on Neural Networks* **1** 593-605

[26] Wythoff B J 1993 Backpropagation neural networks: a tutorial *Chemometrics and Intelligent Laboratory Systems* **18**(2) 115-55

[27] Frawley W J, Piatetsky-Shapiro G, Matheus C J 1992 Knowledge discovery in databases: An overview *AI magazine* **13**(3) 57-70

[28] Fayyad U M, Piatetsky-Shapiro G, Smyth P 1996 Advances in knowledge discovery and data mining *MIT Press* Cambridge MA 1-34

[29] Rubin D B, N Schenker N 1986 Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association* **81**(394) 366-74

[30] Streit R L, Luginbuhl T E 1994 *IEEE Transactions on Neural Networks* **5**(5) 764-83

[31] Topchy A P, Lebedko O A, Miagkikh V V, Kasabov N K 1997 Adaptive training of radial basis function networks based on cooperative evolution and evolutionary programming *International Conference on Neural Information Processing and Intelligent Information Systems* 253-8

Authors

Liangong Song, born in May, 1975, Xinjiang, China

Current position, grades: MSc, lecturer at North China University of Water Resources and Electric Power in China.
University studies: Bachelor of engineering at China Agricultural University in 1999, MSc in Information Management Study at Beijing Forestry University in 2005
Scientific interests: theory and application of computer, principle and application of database, information management
Publications: 8.
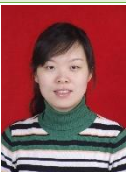
Ke Han, born in November, 11 1980, Henan, China

Current position, grades: MSc, lecturer at North China University of Water Resources and Electric Power in China.
University studies: Bachelor of engineering at Henan Normal l University in 2002, MSc in Information Management Study at Chinese Academy of Sciences in 2005.
Scientific interests: information management system, theory and application of computer
Publications: 11.

An improved boundary extraction method of STL model based on edge curvature estimation

Jingbin Hao[1*], Liang Fang[2], Haifeng Yang[1]

[1]*College of Mechanical and Electrical Engineering, China University of Mining and Technology, Xuzhou, 221116, China*

[2]*State Key Laboratory of Mechanical Behaviour of Metals, Xi'an Jiaotong University, Xi'an, 710049, China*

Abstract

To efficiently extract feature boundaries of the STL model, an improved method is proposed based on edge curvature estimation. Three curvature parameters (dihedral angle, perimeter ration and convexity) are used to estimate the surface curvature information of the STL model. Genetic Algorithm (GA) is used to determinate the threshold of feature edges. The extracted feature edges are grouped and filtered using the best-fit plane (BFP), which is calculated by Least Square Method (LSM). The Dijkstra's algorithm is used to close the incomplete feature boundaries. Several experimental results demonstrate that the amount of feature edges is significantly reduced, and useful feature edges are reserved to construct feature boundaries. The improved boundary extraction method has important significance in decomposing large complex STL models meaningfully.

Keywords: boundary extraction, curvature estimation, STL model, genetic algorithm, least square method

1 Introduction

One common strategy for dealing with a large complex model is to decompose it into smaller and simpler sub-models. As STL files use triangle meshes to represent the surface of a solid model, the core issue is how to decompose and reconstruct triangle meshes. To optimally partition the STL model, we first need to extract boundaries between the meaningful sub-models [1, 2]. There are three main approaches that can be used for the feature boundary extraction of an STL model.

1) Vertex-based method: The definition of a sub-model is the one in which regions consist of connected vertices which have the same curvature value (within a tolerance) [3, 4].

2) Edge-based method: A feature edge is an edge shared by two faces whose normal vectors make an angle greater than a certain threshold [5].

3) Face-based method: The face clusters, which are connected sets of faces, represent the aggregate properties of the original surface at different scales rather than providing geometric approximations of varying complexity [6, 7].

Comparing these three methods, the vertex-based method and the face-based method are costly for the large-scale model, because of multiple clustering of vertices and merging of triangle meshes. In contrast, the edge-based method is the most effective way to find the feature boundaries of the large-scale model, where the model can be partitioned meaningfully. The main challenges of the edge-based method include extracting and grouping the feature edges, selecting the best-fit loop and constructing the cutting contour [8].

In this paper, we propose an improve boundary extraction method based on edge curvature estimation. First, each edge of the STL model is estimated using three curvature parameters (dihedral angle, perimeter ration and convexity). Feature edges are extracted by Genetic Algorithm (GA). Then, the discrete feature edges are grouped and filtered by Least Square Method (LSM). Finally, incomplete feature boundaries are closed by the Dijkstra's algorithm. Several experimental results are given, and the efficiency and correctness of the proposed algorithm are analysed.

2 Topology reconstruction based on Hash-table

Since the STL model has no topology information, the topology reconstruction is necessary for the boundary extraction. The vertex table of triangle facets is built to check and delete repeated vertices. As the same time, the index table of facets is constructed to save the coordinate value index of three vertices.

Set the vertex collection of a STL model is V, the edge collection is E, and the facet collection is F. According to the Euler equation, the relationship of these collections is [9]:

$$V + F - E = 2 - 2H, \tag{1}$$

where, H denotes the number of holes on the model surface. Normally, the STL model is the closed mesh surface, $H=0$.

Corresponding author e-mail: jingbinhao@cumt.edu.cn

Each facet has three edges. Meanwhile, each edge is belong to two adjacent facets. The relationship is:

$$E = 1.5F . \tag{2}$$

The Equation (1) is simplified to be:

$$V = E - F + 2 = 0.5F + 2 , \tag{3}$$

It can be seen that the data size of the vertex table is smaller than that of the facet table. The vertex collection is merged and saved in the Hash-table. Then, the vertex data of the edge collection is conversed to be the index address of the vertex Hash-table.

The common methods of Hash-table construction are [10]: the directly addressing method, the square method, the numerical analysis method, the remainder method, etc.

Set the coordination value of a vertex is x, y, z , the corresponding address of Hash-table can be expressed as:

$$Index = (\text{int})((ax + by + gz)C + 0.5) \& T , \tag{4}$$

where, α, β, γ is the coefficients of Hash function, C is the proportionality coefficient, T is the length of Hash-table, 0.5 is the additional constant which is used to rationalise the address distribution of Hash-table. In dealing with a large data volume, the chain address method is used to handle the data conflict. Assuming the address range of Hash function is $[0, T-1]$, the pointer vector is set as *Chain-Hash[T]*. All vertices, whose address are *i*, are saved in the chain table of *Chain-Hash[i]* (As shown in Figure 1).
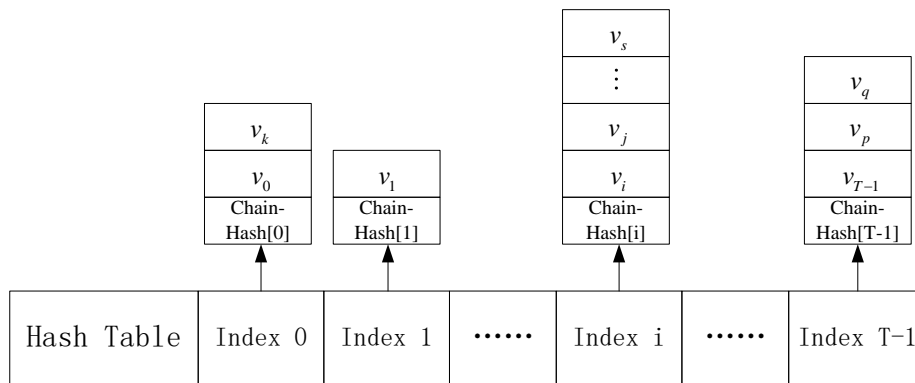
FIGURE 1 Hash-table with the chain-hash
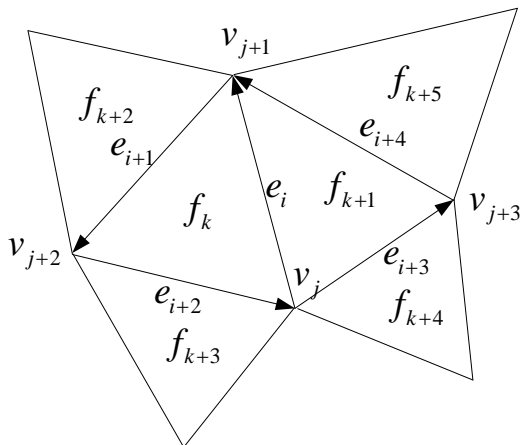
FIGURE 2 Forward-edge structure.

Considering the needs of the follow-up work, the STL model is then reconstructed based on the forward-edge structure. Set an edge e_i has four pointers: the first pointer $p_i[0]$ points to v_j, and the second pointer $p_i[1]$ points to v_{j+1}. The vector direction of $\overrightarrow{v_j v_{j+1}}$ is set as the forward direction. According to the right-hand rule, the third pointer $p_i[2]$ points to f_k, and the fourth pointer $p_i[3]$ points to f_{k+1}. Figure 2 shows the forward-edge structure of triangle facets.

3 Feature edge extracting based on GA

To extract the feature boundaries, there curvature parameters have be used to estimate the model surface.

1) Dihedral Angle: for the regular shapes and the obvious curvature changes, the feature boundaries can be directly distinguished by the dihedral angle [11].

2) Perimeter Ratio: in some regions, which have gradual curvature changes, the perimeter ratio is a sensitive criterion to find the feature boundaries.

3) Convexity: the convexity of feature boundaries is the useful parameter to determine the partition scheme[12]. Using these criteria, the feature edges with greater value than a certain threshold can be extracted.

The selection of feature thresholds is the key issue for extracting feature edges. Generally, the feature thresholds were preset based on experience or experimental analysis [13]. These traditional methods have common problems of poor universal application, low efficiency and accuracy. Both the density of facets and the curvature variation of edges should be considered when feature thresholds are selected. So we use GA to select reasonable threshold value automatically. GA is a stochastic global searching and optimizing algorithm which is suitable to solve complicated problems with large scale [14]. Figure 3 illustrates the working flow of the threshold selection based on GA.

253

1) Coding method: binary system is used to code individuals. The value range of the dihedral angle threshold ε is $(0, \pi/2)$. The value range of individuals is $(0, 15708) \subset [0, 2^{14}]$. So the size of coding scheme is 14. E.g. a binary string (10000111010010) represents an individual $x = 0.8658$.
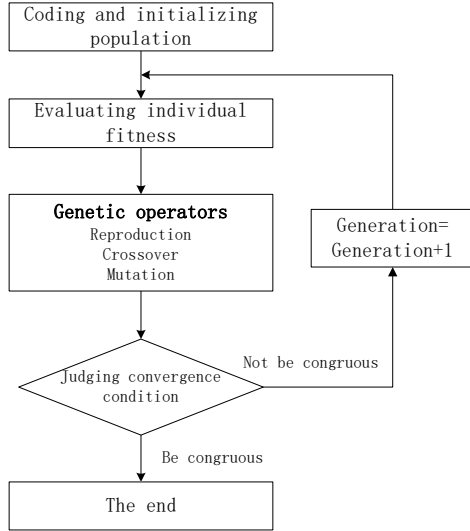
FIGURE 3 The working flow of GA

2) Initial population: the size of population is set $N = 20$. First generation population $P(t)$ is created with twenty individuals $A_1 - A_{20}$, which are initialized randomly between 0 and 15708.

3) Evaluation: in each generation, for which the GA is run, each individual in the population is evaluated against the unknown environment. The fitness values are associated with the values of objective function. We choose OTSU function to calculate the fitness values.

$$\sigma(x)^2 = w_1(x) \times w_2(x) \times (u_1(x) - u_2(x))^2, \qquad (5)$$

where, $w_1(x)$ denotes the number of edges whose dihedral angle is less than x. $u_1(x)$ denotes the average dihedral angle deviation of these edges. $w_2(x)$ denotes the number of edges whose dihedral angle is greater than x. $u_2(x)$ denotes the average dihedral angle deviation of these edges whose dihedral angle is less than x. The aim is to find out the individual whose OTSU value $\sigma(x)^2$ is the greatest, which can be selected as the feature threshold.

4) Genetic operators: Genetic operators drive the evolutionary process of a population in GA, after the Darwinian principle of survival of the fittest and naturally occurring genetic operations. The most widely used genetic operators are reproduction, crossover and mutation.

The selection strategy is chiefly based on the fitness level of the individuals actually presented in the population. The sum of the individual fitness is calculated as:

$$F = \sum_{k=1}^{N} \sigma(A_k)^2. \qquad (6)$$

The selection ratio of the individual k is:

$$p_k = \frac{\sigma(A_k)^2}{F}, \qquad (7)$$

As an individual is selected, reproduction operator only copy it from the current population into the new population without alternation.

The crossover operator starts with two selected individuals and then the crossover point (an integer between 1 and L-1, where L is the length of strings) is selected randomly. Assuming the two parental individuals are A_1 and A_2, and the crossover point is 5 (L=14). If

$A_1 = (01001|101010001)$, $A_2 = (11010|011000100)$

Then the two resulting offspring are:

$A_1' = (01001|011000100)$, $A_2' = (11010|101010001)$

The third genetic operator, mutation, introduces random changes in structures in the population, and it may occasionally have beneficial results: escaping from a local optimum. In our GA, mutation is just to negate every bit of the strings, i.e., changes a 1 to 0 and vice versa, with probability.

After a new population is formed by these three operators, the convergence condition will be judged. If the new population is not congruous, the genetic operators of the new population will be started. Otherwise, the genetic operation is ended. The greatest individual of the new population will be selected as the feature threshold.

4 Feature edge grouping and filtering based on LSM

The extracted feature edges are discrete in the edge array, the best-fit plane (BFP) was proposed to group and link feature edges. The similar and adjacent feature edges were firstly grouped to be feature edge sets. The BFP of a feature edge set was calculated by LSM. The isolated feature edges can be grouped or deleted by BFP.

The method of LSM assumes that BFP of a feature edge set is the plane that has the minimal sum of the deviations squared (least square error) from a given set of vertices [15]. For a set of vertices in a feature edge set: $\{V_i = (x_i, y_i, z_i), i = 0, 1, \cdots m - 1\}$, the function $S^*(V)$ represents the fitting plane that comes closest to pass through all of the vertices. According to LSM, BFP has the property that:

$$\|\delta\| = \sum_{i=0}^{m-1} \delta_i^2 = \sum_{i=0}^{m-1} [S^*(V_i) - V_i]^2 = \min, \qquad (8)$$

for each vertex, the least square error is: $S^*(V_i) - V_i = a_0 x_i + a_1 y_i + a_2 - z_i$, and the sum of the squares of the errors is:

$$S(a_k) = \|\delta\| = \sum_{i=0}^{m-1} [a_0 x_i + a_1 y_i + a_2 - z_i]^2, \qquad (9)$$

to make $S(a_k) = \min$, the unknown coefficients $a_k (k = 0,1,2)$ must yield zero first derivatives.

$$\begin{cases} \dfrac{\partial S}{\partial a_0} = 2\sum_{i=0}^{m-1}(a_0 x_i + a_1 y_i + a_2 - z_i)x_i = 0 \\ \dfrac{\partial S}{\partial a_1} = 2\sum_{i=0}^{m-1}(a_0 x_i + a_1 y_i + a_2 - z_i)y_i = 0, \\ \dfrac{\partial S}{\partial a_2} = 2\sum_{i=0}^{m-1}(a_0 x_i + a_1 y_i + a_2 - z_i) = 0 \end{cases} \qquad (10)$$

The conversion of Equation (10) is:

$$\begin{cases} a_0 \sum x_i^2 + a_1 \sum x_i y_i + a_2 \sum x_i = \sum x_i z_i \\ a_0 \sum x_i y_i + a_1 \sum y_i^2 + a_2 \sum y_i = \sum y_i z_i, \\ a_0 \sum x_i + a_1 \sum y_i + a_2 n = \sum z_i \end{cases} \qquad (11)$$

The matrix of Equation (11) is:

$$\begin{bmatrix} \sum x_i^2 & \sum x_i y_i & \sum x_i \\ \sum x_i y_i & \sum y_i^2 & \sum y_i \\ \sum x_i & \sum y_i & n \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} \sum x_i z_i \\ \sum y_i z_i \\ \sum z_i \end{bmatrix}, \qquad (12)$$

Equation (12) can be solved to quire the value of a_k, which will be used to determine BFP.

$$f_b : a_0 x + a_1 y - z + a_2 = 0. \qquad (13)$$

Some of the isolated feature edges are data noise or surface feature edges which are useless for our approach. But others may be the disconnected feature edges in certain BFP which need to be grouped for constructing feature boundaries. Checking each isolated feature edge, if this edge is almost in the nearest BFP with a tolerance, it will be marked the nearest BFP as the index. Repeat until all isolated edges have been processed. Finally, all useful adjacent and isolated feature edges have been filtered and grouped by BFP.

5 Feature boundary construction based on Dijkstra's algorithms

In most cases, the feature boundaries can be easily constructed by linking feature edges in the same BFP with the doubly linked structure which points to the neighbourhood edges. The case that requires more attention is the incomplete boundaries.

As shown in Figure 4, there are three major conditions of incomplete boundaries. For the conditions (a) and (b), the incomplete boundaries and unconnected edges are in the same BFP. Our method is to search for the shortest path of two endpoints based on the Dijkstra's algorithm, and then connect them with adding several new edges. For the condition (c), the hybrid features are in different BFPs. Our method is to close these loops separately based on the

Dijkstra's algorithm. In Figure 4(c), the complete boundary belongs to four BFPs. Each incomplete loop can be closed by the above method, and then four feature loops will be obtained.

(a) single incomplete boundary in one BFP

(b) several incomplete boundaries in one BFP

(c) incomplete boundaries in different BFPs
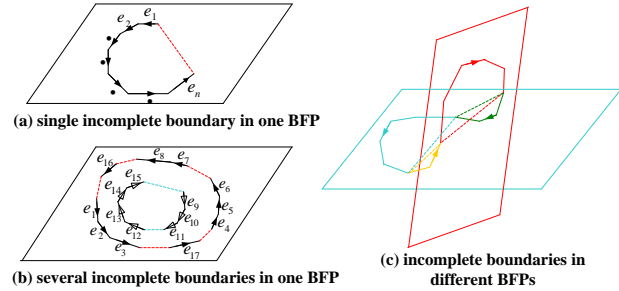
FIGURE 4 Different conditions of incomplete boundaries

The classical Dijkstra's algorithm is the theoretical foundation for solving the problem about the shortest path. However, this algorithm is the greedy algorithm. The optimal solution can be achieved efficiently by the constraint function [16]. We use the max tolerance plane of BFP as the region constraint condition.

Set f_b^+ and f_b^- as the two max tolerance planes of BFP, the normal of f_b^+ and f_b^- is the same as that of BFP. The distance D_b^+ and D_b^- are determinate by V_{max}^+ and V_{max}^-, which are the farthest vertices from BFP (as shown in Figure 5).

$$f_b^+ : a_0 x + a_1 y + a_2 z + D_b^+ = 0,$$

$$f_b^- : a_0 x + a_1 y + a_2 z + D_b^- = 0,$$

where, $D_b^+ = 2 \times Dist(v_{max}^+, f_b)$. The shortest path searching can be efficiently converged by the constraint region, which is constructed by f_b^+ and f_b^-.
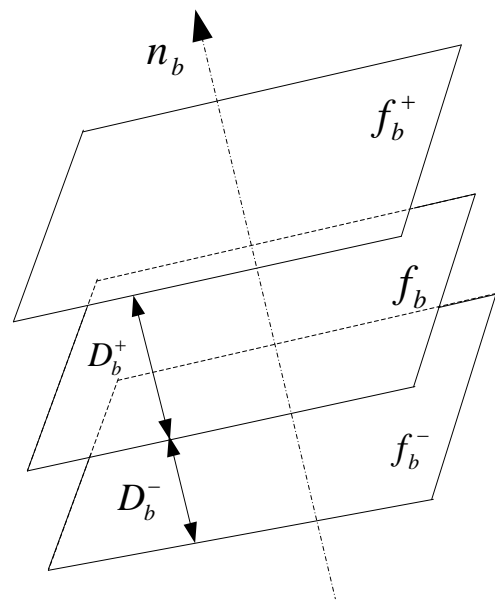
FIGURE 5 Max tolerant planes of BFP

6 Experimental results and analysis

In this section, four experimental STL models with different sizes and complexities are selected to be processed by our proposed method. The proposed

algorithms have been successfully implemented in the 3DEPS system developed by the present authors in China University of Mining and Technology. A summary of results of these models is shown in Figure 6:
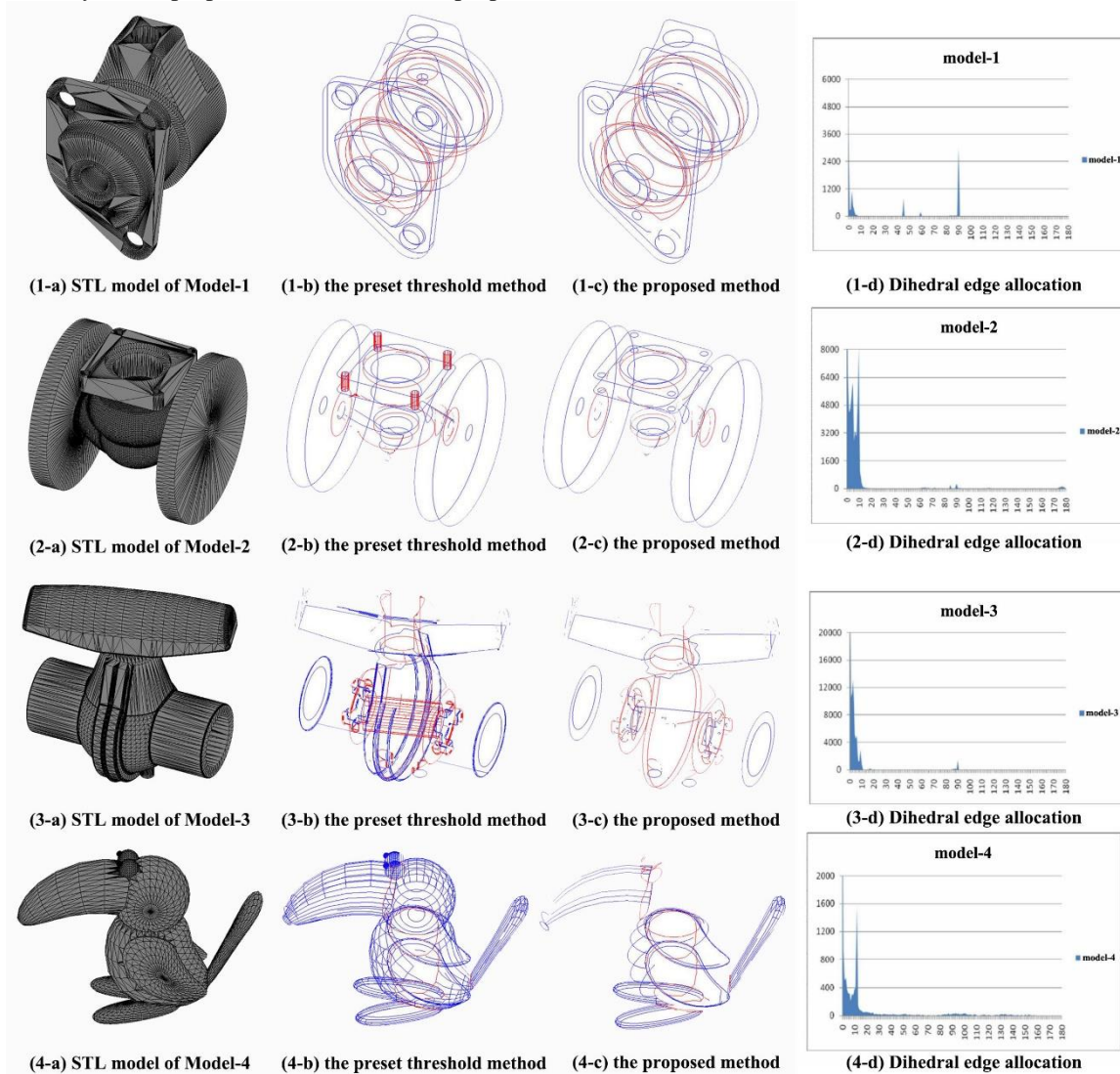
FIGURE 6 Feature boundaries extracting of experimental models

There are two types of feature boundaries:

1) the component boundary (red color ones). These kinds of boundaries are on the junction of components, which can be used to decompose the large complex model.

2) the contour boundary (blue color ones). These kinds of boundaries are on the ends of the model surface, which can be used to represent the outer contour.

For simple structure models (Model-1), the feature boundaries are less and complete. The result of the proposed method is similar to that of preset threshold method. For complex structure models (Model-2, 3, 4), the result of the proposed method is more efficient than that of preset threshold method. The amount of feature edges is about half of the preset threshold method, and useful feature edges were reserved. Take model-4 for example,

the amount of feature edges that extracted by the preset threshold is 4385; the amount of feature boundaries is 136. Using the proposed method, the amount of feature edges is 1845; the amount of feature boundaries is 53. Appendix A shows the detail steps of feature edge extraction based on GA. The experimental results of other models are as shown in Table 1.

By testing the feature boundary extraction of experimental models, we can see from Figure 6 and Table 1 that our proposed method has an obvious advantage on efficiency and accuracy than the preset threshold method. Moreover, when the STL model has a large quantity of facets (more than 100,000), the computing time of our algorithm is absolutely shorter than that of vertex-based method and face-based method.

TABLE 1 Experimental results of feature boundary extraction

MODEL	Model-1	Model-2	Model-3	Model-4
Size (mm) X/Y/Z	589 911 775	723 702 702	427 437 263	402 274 166
Triangle	8206	57574	48362	8052
Point	4093	28777	24180	4028
Feature edge	3146	2879	4373	1845
Threshold 1 Dihedral angle	87.89°	87.28°	14.89°	15.73°
Threshold 2 Perimeter ratio	9.245	5.772	10.425	6.292
Feature boundaries	28	26	36	53

7 Conclusions

In this paper, the edge-based method is chose to extract feature boundaries. Three curvature parameters of each edge are estimated, and the feature edges with greater value than the threshold are extracted. The feature thresholds are automatically selected based on GA. Then, the discrete feature edges are grouped and filtered to be feature boundaries by BFP, which is calculated based on LSM. Finally, Dijkstra's algorithm is used to close incomplete boundaries. The extracted component boundaries can be used to decompose STL model meaningfully, and the contour boundaries can be used to represent the outer contour of STL model. The experiment results show that the proposed method has reasonable execution efficiency, which is suitable to extracting feature boundaries of large complex models. The further enhanced algorithms can be met other file formats of large complex 3D models.

Acknowledgments

Appendix: A Feature edge extraction of Model-4

1) Coding method
The coding method is twelve binary code. The crossover probability p_c is set 0.6. The Mutation probability p_m is set 0.01. The stop condition is reaching the preset maximum number of iterations (the experimental number of iterations is 50 times) or the average of fitness of new generation is similar with the last generation [1,1.01].

2) Initial population
According to the coding method and population size, the initial population can be generated randomly from feasible solutions $(0, 2^{14})$. The size of initial population is $N = 20$. The randomly generated initial population $P(t)$ is as follows:

A1=(0000000 0101001)	A2=(0010101 1000111)	A3=(0110001 0111110)	A4=(1010100 0101000)
A5=(0011011 0000101)	A6=(0000000 0010000)	A7=(1011001 1010110)	A8=(1101010 1010010)
A9=(1010111 1110110)	A10=(100010 00110100)	A11=(010110 01001001)	A12=(110000 10010101)
A13=(011101 10010001)	A14=(000100 01011111)	A15=(100110 11101001)	A16=(000001 11101011)
A17=(001011 10110011)	A18=(101110 10100110)	A19=(010010 11011011)	A20=(010101 00111100)

3) Evaluation
The fitness value of an individual A_i is calculated by OTSU function $\sigma(A_i)^2$ as follows:

$\sigma(A_1)^2 =$ 4.49299e+06	$\sigma(A_2)^2 =$ 10.0996e+06	$\sigma(A_3)^2 =$ 2.47609e+06	$\sigma(A_4)^2 =$ 0.34640e+06
$\sigma(A_5)^2 =$ 8.40729e+06	$\sigma(A_6)^2 =$ 4.61561e+006	$\sigma(A_7)^2 =$ 0.88808e+06	$\sigma(A_8)^2 =$ 4.57586e+06
$\sigma(A_9)^2 =$ 0.68584e+06	$\sigma(A_{10})^2 =$ 0.11533e+06	$\sigma(A_{11})^2 =$ 3.58673e+06	$\sigma(A_{12})^2 =$ 2.12929e+06
$\sigma(A_{13})^2 =$ 0.88607e+06	$\sigma(A_{14})^2 =$ 3.98791e+06	$\sigma(A_{15})^2 =$ 0.030298e+06	$\sigma(A_{16})^2 =$ 4.42444e+06
$\sigma(A_{17})^2 =$ 9.53484e+06	$\sigma(A_{18})^2 =$ 1.40471e+06	$\sigma(A_{19})^2 =$ 5.54469e+06	$\sigma(A_{20})^2 =$ 4.20482e+06

Based on OTSU, the individual A_2 is the best, the individual A_{15} is the worst.

4) Genetic operators-roulette wheel selection
In this population, the crossover probability is as follows:

$p_1 = 0.062026$	$p_2 = 0.139426$	$p_3 = 0.034183$	$p_4 = 0.004782$
$p_5 = 0.116064$	$p_6 = 0.063719$	$p_7 = 0.012260$	$p_8 = 0.063170$
$p_9 = 0.009468$	$p_{10} = 0.001592$	$p_{11} = 0.049515$	$p_{12} = 0.029395$
$p_{13} = 0.012232$	$p_{14} = 0.055054$	$p_{15} = 0.000418$	$p_{16} = 0.061080$
$p_{17} = 0.131630$	$p_{18} = 0.019392$	$p_{19} = 0.076545$	$p_{20} = 0.058048$

After 20 times of roulette wheel selection, the new population p'_t is constructed as follows:

A'1=(0100101 1011010)	A'2=(0000000 0101001)	A'3=(0000000 0101001)	A'4=(0101100 1001001)
A'5=(0100101 1011010)	A'6=(0010111 0110010)	A'7=(0110001 0111110)	A'8=(0010101 1000111)
A'9=(0010101 1000111)	A'10=(010101 00111100)	A'11=(110101 01010010)	A'12=(001101 10000101)
A'13=(000001 11101011)	A'14=(000000 00001111)	A'15=(000001 11101011)	A'16=(001010 11000111)
A'17=(000001 11101011)	A'18=(010101 00111100)	A'19=(001101 10000101)	A'20=(010010 11011010)

5) Genetic operators-crossover and mutation
Since the crossover probability p_c is 0.6, N of the probabilities r_k are generated randomly from [0,1]. If $r_k < p_c$, the k chromosome of $P'(t)$ is selected. 16 individuals are chose to be paired and crossover randomly. The crossover location is selected randomly from [1,13]. The Mutation probability p_m is 0.01, and the total number of genes is 280. For each generation, 2.8 genes are mutated equally. The new population $P''(t)$ is obtained as follows,

A''_1=(010001 00101001)	A''_2=(000000 11011010)	A''_3=(000000 00101001)	A''_4=(010110 01001001)
A''_5=(010010 11011101)	A''_6=(001011 10110010)	A''_7=(011000 10100111)	A''_8=(001010 11011110)
A''_9=(000101 01010010)	A''_{10}=(010101 00111100)	A''_{11}=(111010 11000111)	A''_{12}=(001101 10000111)
A''_{13}=(000001 10101011)	A''_{14}=(000000 00001101)	A''_{15}=(000000 11000111)	A''_{16}=(001011 11101011)
A''_{17}=(000001 11101100)	A''_{18}=(010101 00111011)	A''_{19}=(001101 10000010)	A''_{20}=(010010 11011010)

The average fitness value of $P''(t)$ is 5.76753e+006, which is larger than that of $P(t)$ (3.62184e+006). Since the ratio is greater than 1.01, the next evolution is began. The fitness of A_2 is the largest in $P(t)$, while the fitness of A''_7 is the smallest in $P''(t)$. Therefore, A''_7 is replaced

by A_2 to generate the next initial population $P(t+1)$. Repeat the process until the stop condition. The final population is $P(t+8)$, which is iterated eight times.

A_1=(00101010 111110)	A_2=(00111011 000101)	A_3=(11101011 011100)	A_4=(00101100 111010)
A_5=(00110110 000010)	A_6=(00110110 000010)	A_7=(00101010 111010)	A_8=(00101111 101001)
A_9=(11101010 011101)	A_{10}=(1110101 1011010)	A_{11}=(0010101 1000100)	A_{12}=(0010110 0111101)
A_{13}=(0011001 1001001)	A_{14}=(0010101 0111001)	A_{15}=(0011011 0000010)	A_{16}=(0101011 1000110)
A_{17}=(1110101 1000111)	A_{18}=(0011011 0000010)	A_{19}=(0010101 0111110)	A_{20}=(1110101 1101010)

The fitness of the best individual A_{14} is 10.165e+006. So the corresponding angle of A_{14}(15.73°) is selected as the dihedral angle threshold of Model-4.

References

[1] Kou X Y, Tan S T 2009 *Rapid Prototyping Journal* **15**(1) 5–18
[2] Luo L J, Baran I, Rusinkiewicz S, Matusik W 2012 *ACM Transactions on Graphics* **31**(6) 9–18
[3] Mao Z H, Cao G, Zhao M X 2009 *The Visual Computer* **25**(3) 289–95
[4] Razdan A, Bae M 2003 *Computer-Aided Design* **35**(9) 783–9
[5] Lien J M, Amato N M 2004 *Proceedings of the 20th Annual ACM Symposium on Computational Geometry* ACM Press New-York NY USA 17-26
[6] Lau M, Ohgawar A, Mitani J, Igrashi T 2011 *ACM Transactions on Graphics (SIGGRAPH)* **30**(4) Article No 85
[7] Hilderand K, Bickel B, Alexa M 2012 *Computer Graphics Forum (Proc. Euro graphics)* **31** 583–92
[8] Hao J, Fang L, Williams R E 2011 *Rapid Prototyping Journal* **17**(2) 116-27
[9] Lu Y, Garboczi E J 2014 *Journal of computer in civil engineering* **28**(3) 503-22
[10] Huang K, Xie G G, Li R, Xiong, S 2013 *Journal of Network and Computer Applications* **36**(2) 657-66
[11] Kadry S, Abdallah A, Joumaa C 2011 *Lecture Notes in Electrical Engineering* **133**(2) 393-7
[12] Wang M, Feng J Q, Chen W 2014 *Computers & Graphics* **38** 212-21
[13] Michikawa T, Suzuki H 2013 *Computer-Aided Design* **45**(4) 822–8
[14] Lee J. Son H, Kim C, Kim C 2013 *Automation in Construction* **35** 199–207
[15] Vigo M, Pla N, Ayala D, Martinez J 2012 *Graphical Models* **74**(3) 61-74
[16] Fu L T, Kara L B, Shimada K 2014 *Computer-Aided Design* **46** 263–8

Authors

Jingbin Hao, born in May, 1982, Xuzhou, Jiangsu Province, China

Current position, grades: PhD, Lecturer at the College of Mechanical and Electrical Engineering, China University of Mining and Technology.
University studies: PhD in Mechanical Manufacture and Automation from the University of Mining and Technology, China in 2011.
Scientific interests: rapid prototyping, laser manufacturing, computer graphics.
Publications: 15 papers.

Liang Fang, born in January, 1957, Xi'an, Shaanxi Province, China

Current position, grades: Ph. D, Professor in the State Key Laboratory of Mechanical Behaviour of Metals, Xi'an Jiaotong University, China.
University studies: PhD in Materials Science and Engineering from the Xi'an Jiaotong University, China in 1997.
Scientific interests: rapid prototyping of polystyrene, materials of tribology, advanced materials processing
Publications: over 70 papers.

Haifeng Yang, born in May, 1981, Peixian, Jiangsu Province, China

Current position, grades: PhD, associate professor at the College of Mechanical and Electrical Engineering, China University of Mining and Technology.
University studies: PhD in Materials Science from the Jiansu University, China in 2009.
Scientific interest: laser nano-manufacturing, unconventional machining technology.
Publications: 23 papers.

Application of TV image compression technology based on neural network

Tao Guo*, Zhengqi Liu

School of information engineering, Longdong University, 745000, China

Abstract

Aiming at the disadvantages of digital TV, including a large amount of information and redundant information, a method of TV image compression technology based on neural network combining neural network with image compression technology is proposed in the work. Firstly, the TV image is divided into blocks as the input of neural network to build the network; secondly, the blocks are rebuilt to realize image compression recovery. The simulations show that the neural network algorithm can achieve the TV image compression effectively and the number of neurons of the hidden layer based on the neural network algorithm has great influence on the building and training of the network by contrast. When the number of neurons of the hidden layer is less, the image compression ratio will be higher and the image compression quality will be lower.

Keywords: digital TV technology, image compression, neural network, hidden layer, compression ratio

1 Introduction

With the rapid development of digital TV technology, the transmission quantity of the data becomes more and more large and the requirement for the software and hardware is more and more high. The traditional transmission and image compression technology cannot meet the requirement of the actual technology. The BP neural network has the advantage of good nonlinearity, parallel calculating of large data and self-organized learning. In the work, a TV image compression technology based on BP neural network, combining the neural network technology with the digital image compression technology, is proposed and applied to the process of TV image compression transmission. The simulations show that the method used in this work has the advantage of fast compression speed and high compression quality and it is more efficient.

2 BP neural network

BP neural network, an error back propagation network, was proposed by Rumelhart and McCelland in 1986 [1-3]. The structure model, showed in Figure 1, consists of the input layer, the hidden layer and the output layer.
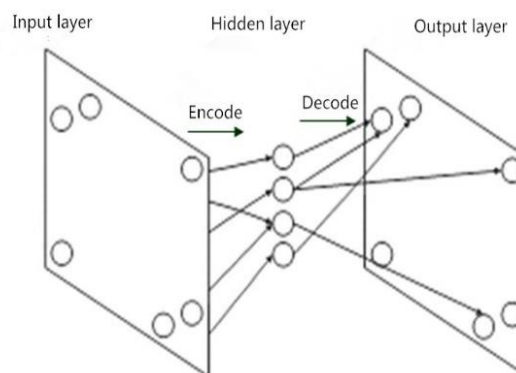
FIGURE 1 Structure model of BP neural network

3 Image compression model of BP neural network

3.1 BUILDING THE MODEL

Figure 2 shows that in view of the digital image with a pixel of N×N, the original image was divided by K×K into many small blocks, as the inputs of neural network [4-6].
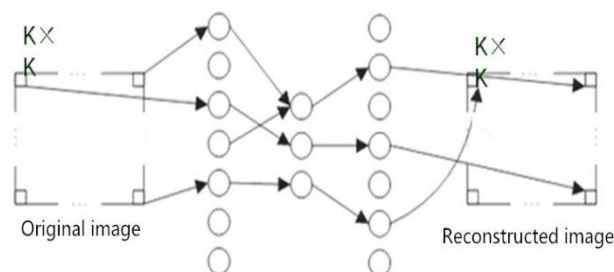
FIGURE 2 BP image compression model

The image compression algorithm of BP neural network is as follows:

*Corresponding author e-mail: yzz_hyyf@163.com

Step 1: Building the training sample. Divide the original image into 4×4 non-overlapped pixel blocks, transform each pixel block into 16×1 column vector and change the original data into 16×1024 matrix.
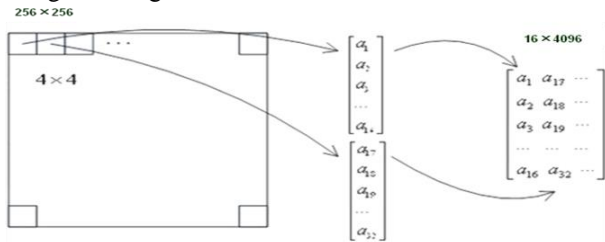
FIGURE 3 Dividing the image into blocks

Step 2: Building BP neural network model.

Train BP network considering the image data set as the input and the ideal signal. If the compression ratio of the network image is S, the node number of the network input layer is N_i, the node number of the hidden layer is N_h, then the compression ratio is [6-9]: $S = \dfrac{N_i}{N_h}$.

Step 3: Training the network. Use the Levenberg-Marquardt method, with a fast training speed.

Step 4: Coding the network.

Step 5: Compressing and rebuilding the image. Encode the compressed and coded data, and alternate into the size of original data value: multiply each matrix element by 255; recover the pixel value from interval [0, 1] to [0, 255]; transform the column vectors into the image blocks to form a complete image; finish the image rebuilt.

Figure 3 shows the calculating model as follows:
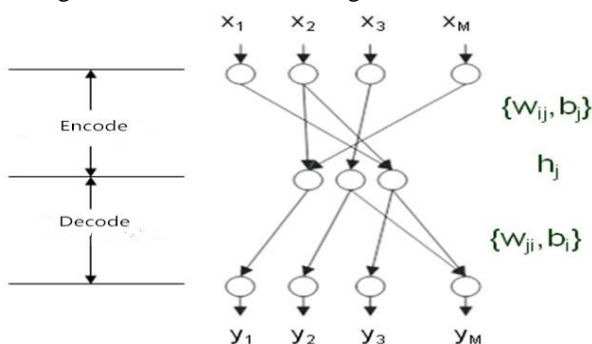
FIGURE 4 Calculating model of the image compression

3.2 FLOW CHART OF THE ALGORITHM

As mentioned, Figure 5 shows the flow chart of BP TV image compression [10].
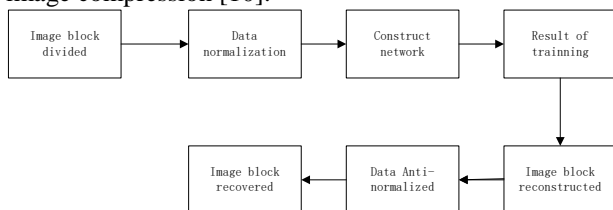
FIGURE 5 Flow chart of BP TV image compression algorithm

3.3 EVALUATING THE COMPRESSION QUALITY

Peak signal to noise ratio (PSNR) is a widely-used evaluation standard for the image quality evaluation. It can be expressed as the logarithm of the ratio of mean square error (MSE) between the original and the compressed image and $(2^{n-1})^2$ multiplied by 10. The calculating formula is indicated by [11-13]: $PSNR = 10 * log\left(255^2 / MSE\right)$.

4 Algorithm simulation

In order to test and verify effectiveness and reliability of the algorithm, the two test images including Figure 6 and Figure 7 were verified based on MATLAB [9]: The parameters of the neural network were set as: the training goal=0.01, the training times epoch=1000.
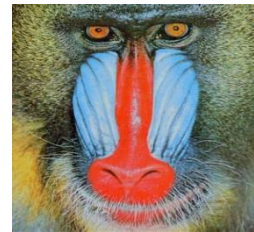
FIGURE 6 The 1st test image   FIGURE 7 The 2nd test image

If the compression ratio S is 16, the compression results will be as follows:

a) The 1st test image

b) Compression result of the 1st test image

c) The 2rd test image

d) Compression result of the 2rd test image

FIGURE 8 Compression results while the compression ratio S=16

If the compression ratio S is 8, the compression result will be as follows:

a) The 1st test image

b) Compression result of the 1st test image

c) The 2rd test image

d) Compression result of the 2rd test image

FIGURE 9 Compression results while the compression ratio S=8

If the compression ratio S is 4, the compression result will be as follows:
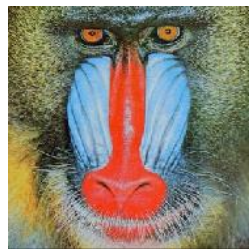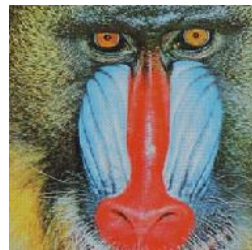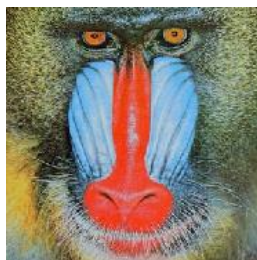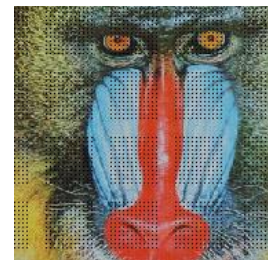
a) The 1st test image

b) Compression result of the 1st test image

c) The 2nd test image

d) Compression result of the 2nd test image

FIGURE 10 Compression results while the compression ratio S=4

Table 1 shows the contrast results of time compression.

TABLE 1 Time compression contrast results of the test images

Image	Compression Ratio	Time Compression/s
	16	10.5408
The 1st Test Image	8	7.5647
	4	8.3245
	16	4.3710
The 2nd Test Image	8	1.9802
	4	4.3010

Table 2 shows the PSNR contrast results of the test images.

TABLE 2 PSNR contrast results of the test images

Image	PSNR	Compression Ratio
	30.8796	0.0507
	31.8181	0.0897
	32.4118	0.1465
The 1st Test Image	31.9662	0.2030
	31.1015	0.2512
	30.7471	0.3158
	28.2056	0.3219
	31.9063	0.4148
	31.7869	0.0507
	32.1832	0.0897
	32.6738	0.1465
The 2nd Test Image	32.9882	0.2030
	33.1213	0.2512
	33.7521	0.3158
	33.9625	0.3219
	34.2213	0.4148
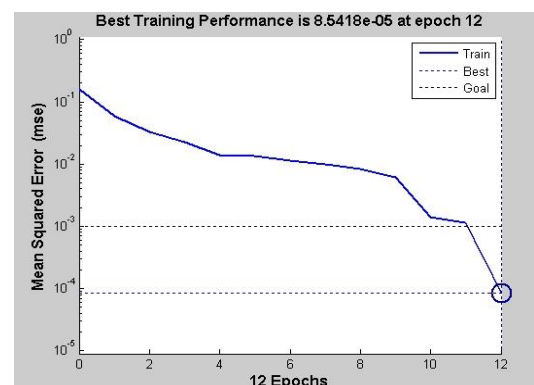
a) Error convergence graph while compression ratio is 4
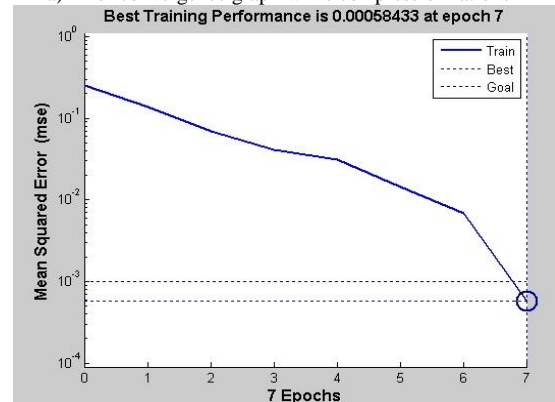
b) Error convergence graph while compression ratio is 8

c) Error convergence graph while compression ratio is 16

FIGURE 11 Error convergence graphs with different compression ratio

Table 2 shows that the PSNR decreases along with the increase of the compression ratio, that is, the image

compression quality decreases. Besides, Figure 11 indicates that the error training times increase along with the increase of the compression ratio.

5 Conclusions

In the work, a method of TV image compression technology based on BP neural network combining neural network with image compression technology is proposed and applied to the process of TV image compression transmission. The simulations show that the method has the advantage of fast compression speed and high compression quality and it is more efficient. Firstly, the TV image is divided into blocks as the input of neural network to build the network; secondly, the blocks are rebuilt to realize image compression recovery. The simulations

show that the neural network algorithm can achieve the TV image compression effectively and the number of neurons of the hidden layer based on the neural network algorithm has great influence on the building and training of the network by contrast. When the number of neurons of the hidden layer is less, the image compression ratio will be higher and the image compression quality will be lower.

References

[1] Gijsenij A, Lu R, Gevers T 2012 Color *IEEE Transactions on Image Processing* **21**(2) 697-707
[2] Gijsenij A, Gevers T, van de Weijer J 2011 *IEEE Transactions on Image Process* **20**(9) 2475-89
[3] Gijsenij A, Gevers T, van de Weijer J 2010 Generalized gamut mapping using image derivative structures for color constancy *IJCV* **86**(2-3) 127-39
[4] Gijsenij A, Lu R, Gevers T 2012 *IEEE Transactions on Image Processing* **21**(2) 697-707
[5] Wang P, Shi S, Hong X 2008 Simultaneous Location of Moving robots and Map Establishment Based on SIFT Algorithm *Journal of Ningbo University (science and technology)* **21**(1) 68-71
[6] Yang Q, Xiao X 2011 SIFT Algorithm Based on Improvement of Canny Feature Point *Computer Engineering and Design* **32**(7) 2428-31 *(in Chinese)*
[7] Wu Y, Yu J 2012 JPEG2000 Image Compression Algorithm and Its Application in Network *Computer Engineering* **29**(3) 7-10 *(in Chinese)*

[8] Zhang H, Wang D 2011 Imagine Compression Technique *Journal of System Simulation* **14**(7) 831-5 *(in Chinese)*
[9] Zhang Z, Liu G 2010 Development of research on video and imagine compression based on wavelet *Chinese Journal of Electronics* **30**(6) 883-9 *(in Chinese)*
[10] Ababneh J I, Bataineh M H 2008 Linear phase FIR filter design using particle swarm optimization and genetic algorithms *Digital Signal Processing* **18**(4) 657-68
[11] Luitel B, Venayagamoorthy G K 2008 Differential evolution particle swarm optimization for digital filter design *IEEE Congress on Evolutionary Computation CEC 2008 (IEEE World Congress on Computational Intelligence)* 3954-61
[12] Boo J 2007 Stock Price forecasting using PSO-trained neural networks *IEEE Congress on Evolutionary Computation CEC 2007* 2879-85
[13] Lin C J, Liu Y C 2009 Image backlight compensation using neuron-fuzzy networks with immune particle swarm optimization *Expert Systems with Applications* **36**(3) 5212-20

Authors

Tao Guo, born in December, 1979, Qingyang, Gansu Province, China

Current position, grades: Master Degree. Associate Professor, at School of information engineering, Longdong University, China.
University studies: Computer science and technology.
Scientific interests: Computer software and computer graphics.
Publications: 20 papers published in the international or national journals.

Zhengqi Liu, born in December, 1979, Fufeng, Shaanxi Province, China

Current position, grades: Professor in School of information engineering, Longdong University, China.
University studies: Computer science and technology.
Scientific interests: Computer software and computer graphics.
Publications: 30 papers.

Fabric defect detection system based on digital image processing

Zhaozhun Zhong[1]*, Pengjie Qi[2], Miao Guan[1], Yuedong Xia[2], Yuanhui Fu[3]

[1]*School of Iron and Steel Soochow University, Suzhou City, Jiangsu Province, PR China, 215021*

[2]*School of Mechanical and Electrical Engineering Soochow University, Suzhou City, Jiangsu Province, PR China, 215021*

[3]*Humanetics Innovative Solutions, Inc., Plymouth, USA, 48170*

Abstract

A fabric defect detection system based on digital image processing for textile fabric is proposed in this paper. The approach for the classification and identification of three commonly encountered classes of fabric defects (holes sto, missing end and mispick) is studied. The developments of both the hardware and software structures are presented. Firstly, median filter preprocessing and image segmentation based on Otsu threshold are applied to localize the fabric defects. Secondly, the features based on grey-level histogram and geometry are extracted. Thirdly, the classification and identification are accomplished by the method of artificial neural network based on the extracted features. Finally, a variety of textile images with different defects are tested to evaluate the performance of the proposed defect detection system. The experiment results indicate that the proposed system works efficiently with high accuracy, which can meet the requirements of the textile industry.

Keywords: fabric defect detection system, digital image processing, image segmentation, defects classification

1 Introduction

The textile industry plays an important role in the national economy. The throughput of a modern textile process is quite impressive, expensive mechanical equipment and modern control strategies are necessary in the process. The major specifications of the product that should be satisfied are quality and stable operation of the process. As for quality control, the fabric defect detection is a major concern because fabric defects may reduce the price of a product by 45% to 65%. The traditional fabric defect detection is generally manipulated by manual inspection, which suffers from the problems associate with human fatigue and boredom, inconsistency and high inspection cost. In general, only about 60%~70% of fabric defects can be detected by the most skilled inspectors under the conditions that the width of the fabric is less than 2m and the speed is less than 30m/min [1]. In order to overcome the shortages of human inspection, automated visual inspection system could be a possible way to detect possible defects in fabric which provides a more reliable and consistent quality control process [2, 3].

Automated fabric defect detection system works based on the fact that textile faults normally have features which are different from the features of the original fabric. As one of the most intriguing problems in visual inspection, fabric defect detection has attracted great attention in recent years. The detection of fabric defects can be considered as a segmentation and identification problem. A variety of algorithms have been proposed to solve the problem. For a plain fabric, Zheng et al. [4] reported that 92% of the defects could be detected simply by distance calculation among images. For complicated fabrics such as denim and twill fabrics, some effective schemes have also been proposed [5, 6]. The proposed algorithms can be categorized into five major classes: spectral, statistical, geometrical, structural and model based. Among all the five major classes, the spectral approach, which is based on the periodicity of the texture feature, is regarded as the most effective one for woven fabrics in literature. For spectral approach commonly used techniques are wavelet transform [7, 8], Fourier transform [9, 10], Gabor analysis [11, 12] and so on. As for real time fabric defect detection, Mak and Peng [13] proposed a new scheme and apparatus using odd and even symmetric real-valued Gabor filters, which are designed based on the optimally extracted texture features from a non-defective fabric image using Gabor wavelet network. A real time fabric defect detection system based on an embedded DSP (Digital Signal Processor) platform is proposed by Raheja [14, 15], in which textural features of fabric images are extracted based on grey level co-occurrence matrix and a sliding window technique is used for defect detection. A neural network approach is proposed for defect classification and identification on leather fabric in [16, 17].

Along the line of this research trend, an automatic fabric defect detection system based on machine vision is proposed in this paper. The fundamental hardware and

---

software structures of the proposed system are introduced. Median filter is selected for image pre-processing and Ostu threshold is chosen as the image segmentation method according to the comparison studies. Seven features extracted from the grey-level histogram and six extracted geometric features are used for fabric defect classification and identification. The classification and identification are accomplished by a multi-layered neural network which is appropriate for pattern classification. The proposed neural network is trained by fabric defect samples and tested by on-line detected defects. The experiment results verify the effectiveness of the proposed method.

The rest of the paper is organized as follows: The architecture of the fabric defect detection system is introduced in Section 2, image pre-processing and segmentation of fabric defects are discussed in Section 3. Feature extraction and defect classification as well as the experiment results are given in Sections 4 and 5 respectively. Finally, the paper is concluded in section 6.

2 Architecture of the fabric defect detection system

2.1 SYSTEM DESIGN

As illustrated in Figure 1, the fabric defect detection system consists of an image acquisition module, a defect detection module, a feature extraction module and a defect classification module. The system can be divided into three parts: the first part acquires the fabric image, the second part pre-processes the obtained images and detects the images with defects, the third part accomplishes the task of segmentation and localization of the defect area for feature extraction, and finally classifies the detected defects.
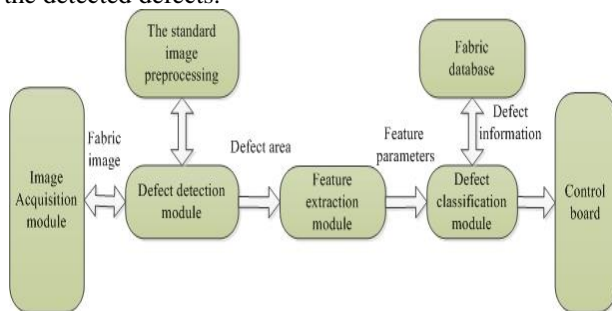
FIGURE 1 Architecture of the fabric defect detection system

2.2 HARDWARE STRUCTURE

The hardware part of the fabric defect detection system is mainly composed of image acquisition system and computer control system. Image acquisition system contains lighting equipment (LED light source), CCD array cameras, image capture card and so on. The hardware structure diagram of the system is shown in Figure 2. The roller system and guide rail of the inspection system are designed to ensure the smooth and even movement of the fabric. The effects of mechanical

vibration and wrinkles can be reduced by adjusting the working tension. LED light source is equipped as back lighting source which provides illumination to minimize the effects of shadows and glare. CCD array cameras are installed on the inspection bridge, which provides three degrees-of-freedom movement for the camera to guarantee the quality of captured images. The acquired images are sent to the parallel computers by an interface called camera link, which is of high data transfer rate. The synchronization of the fabric and camera is ensured by the means of a disc encoder.
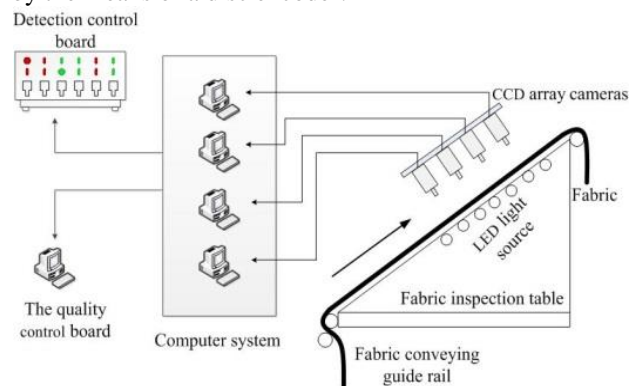
FIGURE 2 The hardware structure of the system

2.3 SOFTWARE STRUCTURE

An effective defect detection schema is regarded as the core part of the automated inspection system. In general, it can be divided into four steps: image pre-processing, image segmentation and localization, feature extraction, defect classification and identification. The specific methods will be introduced later in the following paragraphs. Figure 3 gives the basic procedure of the image processing schema.
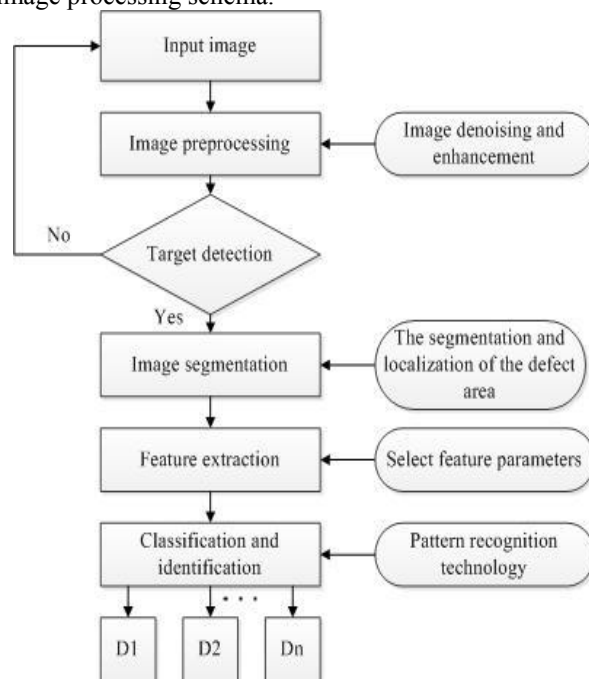
FIGURE 3 The basic procedure of the image processing schema

3 Image pre-processing and segmentation

Due to the technology and equipment faults in the process, a variety of defects are frequently encountered in fabrics, in terms of texture pattern violation, yarn misplacements, weaving defects and so on. The defects can be categorized into different types according to different classification methods. For example, they can be divided into point defect, line defect and planar defect according to the geometrical characteristics of the defects; according to the texture feature of defects, there will have statistical distortion defect, direction distortion defect and unstructured distortion defect. It is difficult to identify all the variety of defects encountered in the textile industry even by the most advanced automated visual inspection system. This paper limits the variety of fabric defects into three kinds of the most frequently encountered ones, that is, holes sto, missing end and mispick. For the sake of clarity, the images of these defects along with the normal fabric are given in Figure 4.

4a) Normal fabric	4b) Holes sto
4c) Missing end	4d) Mispick
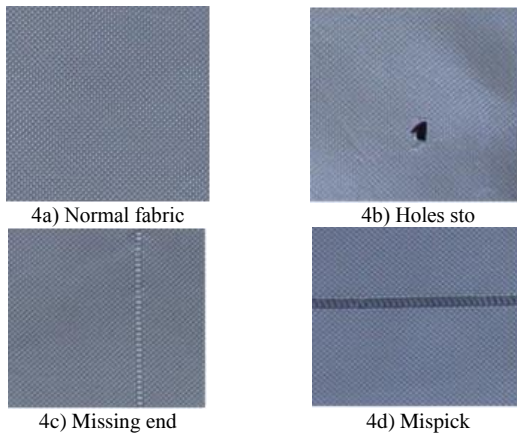
FIGURE 4a-Normal fabric, 4b, 4c, 4d-Defect fabrics

3.1 DEFECT IMAGE PRE-PROCESSING

During the process of image acquisition, the fabric image will get different degrees of noises due to the influence of environmental factors and the system itself. The noises will have a significant impact on the subsequent processing of the image. In order to reduce the effects of the noises, the acquired images should be pre-processed before further processing. During the process of image pre-processing, image denoising is the most crucial step because it directly affects the final effect of image processing. In the literature, mean filter and median filter are the most commonly used spatial filter methods.

Mean filter is an intuitive way of filter, it can effectively suppress noises to smooth the image, but it also has a negative impact of making the image edge blur. Median filter is a commonly used nonlinear spatial filter. It is particularly effective in the presence of impulse noises and it can effectively overcome the blur effect of the details brought by the linear filter. But median filter is not quite qualified for disposing the image with lots of

details. According to the comparison experiments between the two filters, the median filter is more suitable for the characteristics of the fabric image and is chosen as the filter for image pre-processing.

3.2 FABRIC DEFECT IMAGE SEGMENTATION

After pre-processing, the first step should be the judgement of whether there are suspected defects in the image area, which is known the target detection. The algorithms of target detection should be as simple as possible, because all collected images should be processed in a given short time interval. Complex algorithms will increase the amount of calculation and reduce the speed of processing. Most of the images captured by cameras are normal (without defects) which do not need further processing. For the images with defects, segmentations are needed to separate the defective area from the background for further processing. Image segmentation is a key part of fabric defect detection technique, the effect of segmentation directly affects the image feature extraction and identification of defects. Generally speaking, image segmentation methods can be divided into four categories: threshold-based segmentation, edge-based segmentation, region-based segmentation and specific theory-based segmentation. According to the characteristics of the selected fabric, edge-based segmentation and Otsu threshold-based segmentation are chosen for the comparison experiments.

FIGURE 5 Sobel operator templates (Z is the grey value)

1) Edge-based segmentation: Sobel operator and Roberts operator Sobel operator and Roberts operator are the commonly used edge detection operators, in Figure 5, z_5 is defined as the centre point $f(x, y)$, z_1 is $f(x-1, y-1)$, and so on.

The Sobel operator uses 3×3 templates for gradient calculation and increases the weights of the four neighbouring pixels of the centre pixel. The Sobel operator can be formulated as:

$$g_x = \frac{\partial f}{\partial x} = (z_7 + 2z_8 + z_9) - (z_1 + 2z_2 + z_3) \qquad (1)$$

$$g_y = \frac{\partial f}{\partial y} = (z_3 + 2z_6 + z_9) - (z_1 + 2z_4 + z_7) \qquad (2)$$

$$|g(x, y)| = \sqrt{(g_x^2 + g_y^2)} \approx |g_x| + |g_y| \qquad (3)$$

The Roberts operator uses 2×2 templates and the effect is not so good due to the lack of explicit central point. The templates of Roberts operator are given in Figure 6.

a)   b)
FIGURE 6 Roberts operator templates

The Roberts operator can be formulated as:

$$g_x = \frac{\partial f}{\partial x} = (z_9 - z_5) \tag{4}$$

$$g_y = \frac{\partial f}{\partial y} = (z_8 - z_6) \tag{5}$$

$$|g(x, y)| = [(z_9 - z_5)^2 + (z_8 - z_6)^2]^{\frac{1}{2}} \tag{6}$$

The three typical kinds of fabric defect images (holes sto, missing end and mispick) are processed using the edge-based segmentation (Sobel operator and Roberts operator). The results are given in Figure 7.

a) Holes sto after pre-processing   a1) Sobel operator   a2) Roberts operator

b) Missing end after pre-processing   b1) Sobel operator   b2) Roberts operator

c) Misspick after pre-processing   c1) Sobel operator   c2) Roberts operator
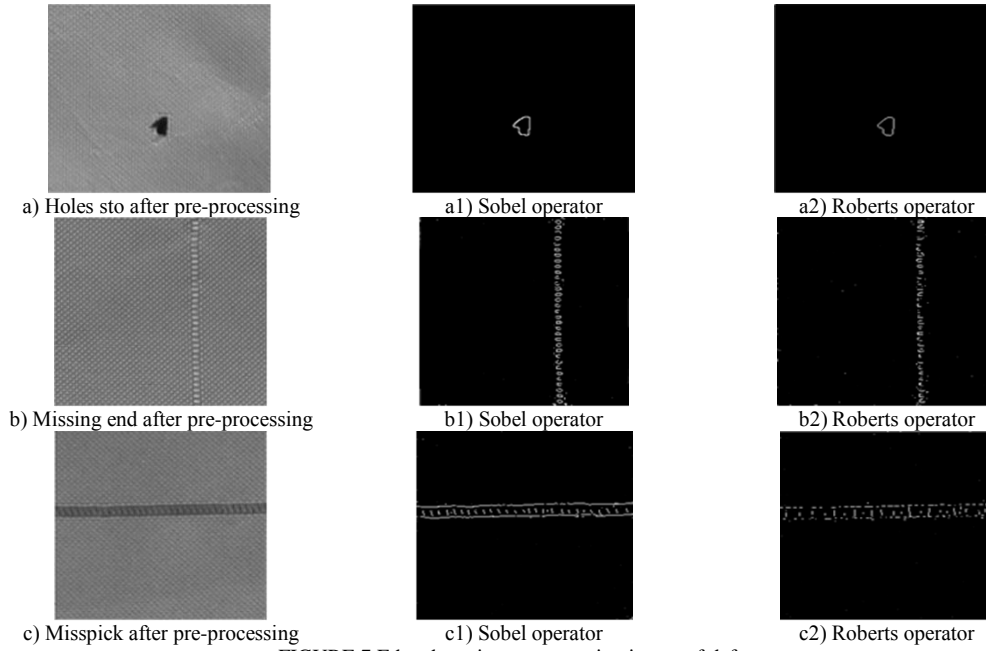
FIGURE 7 Edge detection segmentation image of defects

The experimental results show that the Sobel operator works better in the treatment of the two kinds of defects (missing end and mispick). Their edge is much clearer and the closure is better. The Sobel operator has the smooth effect on image noises and is less affected by noises. The bigger the template is, the better the effect of edge extraction is. But the boundary extracted by Sobel operator is not so good, which makes the target region not obvious.

2) Image segmentation based on Ostu threshold: Threshold segmentation is the method that divides the image pixels into several parts by setting one or more threshold to achieve the aim of image segmentation. This paper uses single threshold segmentation, which only selects a threshold T to divide the image into two parts (target and background) by comparing the grey value of pixels with the threshold T. The segmentation based on threshold is the commonly used method thanks to its simplicity and calculation speed. Otsu threshold segmentation determines the threshold by calculating the between-class variance. The basic idea is as follows: suppose the pixel number of the image is N, the grey-level range is [0, L-1], the number of pixels for each grey-level is N_i, probability is $P(i)$. Using the threshold T to divide the pixels of the image into two classes C_1 and, the grey-level range of C_1 is [0, T] and C_2 is [T+1, L-1], that is, C_1 consists of all the pixels in the image with intensity values in the range [0,T] and C_2 consists of the pixels with intensity values in the range [T+1, L-1].

As a result the average intensity value of the entire image is given by

$$u_T = \sum_{i=0}^{L-1} iP(i) \tag{7}$$

the mean intensity value of C_1 is

$$u_1 = \frac{\sum_{i=0}^{T} iP(i)}{\sum_{i=0}^{T} P(i)} = \frac{\sum_{i=0}^{T} iP(i)}{\omega_1} \tag{8}$$

and the mean intensity value of C_2 is

$$u_2 = \frac{\sum_{i=T+1}^{L-1} iP(i)}{1 - \sum_{i=0}^{T} P(i)} = \frac{\sum_{i=T+1}^{L-1} iP(i)}{1 - \omega_1} = \frac{\sum_{i=T+1}^{L-1} iP(i)}{\omega_2} \tag{9}$$

266

according to the above formulas and

$$u_T = \omega_1 u_1 + \omega_2 u_2 \qquad (10)$$

the between-class variance is given as:

$$\sigma_B^2 = \omega_1 \left(u_1 - u_T\right)^2 + \omega_2 \left(u_2 - u_T\right)^2. \qquad (11)$$

The threshold T is usually set as the one which maximizes the between-class variance σ_B^2.

Three typical kinds of fabric defect images (holes sto, missing end and mispick) are processed utilizing the Otsu threshold with the best thresholds (0.4, 0.56 and 0.42, respectively). The images after processing are shown in Figure 8.

| a) Holes sto | b) Missing end | c) Mispick |
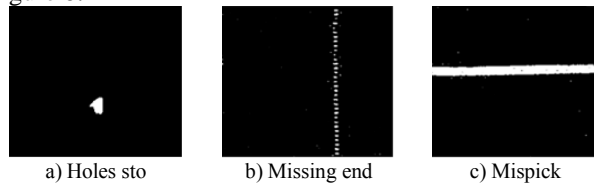
FIGURE 8 Threshold segmentation image of defects

The experimental results show that the Otsu threshold segmentation method can effectively segment the defect image to separate the target and background. Compared with the Sobel operator, the target segmented by the threshold is much more obvious, which facilitates the subsequent feature extraction process. Therefore, the threshold segmentation method is adopted in this paper.

3.3 FABRIC DEFECT LOCALIZATION

The fabric defect region still needs further processing before feature extraction. On one hand, the image after segmentation still has a certain amount of noises which needs further processing. On the other hand, the segmentation region needs to be localized, which is to find out the smallest rectangle that can completely contain the defect area. For each pixel, collecting the numbers n_0 and n_1 of its 0 and 1 valued neighbours, if the maximum number of n_0 and n_1 is less than a given threshold, the pixel point is regarded as noise and is removed. Then finding out the minimum and maximum values of the abscissa and ordinate of the 1 valued pixels,

which determine a rectangle. The determined rectangle is the required localization area. For example, the holes sto images after segmentation and localization are illuminated in Figure 9.

| a) Image after segmentation | b) Original defect image | c) Defect localization image |
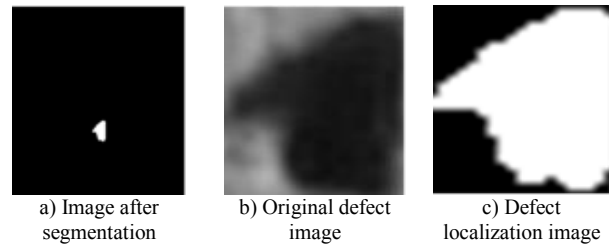
FIGURE 9 Holes sto localization image

4 Feature extraction

1) Feature extraction based on grey-level histogram: the grey-level histogram is about the statistical distribution of image grey value. Feature extraction method based on grey-level histogram is easy to achieve, but suffers from the disadvantages associated with large amount of calculation and low resolution. Seven feature parameters including grey average $\bar{i}$, variance σ^2, energy H_p, entropy H_e, the maximum grey value G_m, the minimum grey value G_n and the maximum difference of grey value G are extracted from grey-level histogram in this paper.

For the sake of clarity, the typical grey-level histogram images of localized holes sto, missing end and mispick are shown in Figure 10. And the feature parameters extracted from the histograms are given in Table 1.

TABLE 1 Histogram feature extraction

Feature parameters	Holes sto	Missing end	Mispick
Average	70.1956	127.3415	93.6002
Variance	48.0994	23.0919	19.1240
Energy	2.075e-3	5.067e-5	2.133e-4
Entropy	6.8279	4.3393	10.9437
Maximum grey value	178	199	245
Minimum grey value	14	59	54
Maximum difference of grey value	164	140	199

| a) Holes sto histogram | b) Missing end histogram | c) Mispick histogram |

FIGURE 10 Histogram of detects

2) Feature extraction based on geometry: geometric characteristic is the basic feature of the defect area, which includes it's shape information. For geometric feature

extraction, the first step is the binarization of the defect image. Then six feature parameters are selected:

- length L, which represents the horizon length of the defect area and is calculated through the horizontal projection,

- width W, which represents the vertical width of the defect area and is calculated through the vertical projection,

- the ratio of length and width B, which general describes the degree of warp or weft of the defect,

- area S, the area of the defect S is the sum of all pixels whose grey value is 1 in the binary image,

- perimeter C, which is the sum of all the pixels in the region R of the segmentation edge,

- dispersion A, it is a measure of the defect shape. For graphics with the same area, the smaller the perimeter is, the compacter the image is.

Geometric features extracted from typical defect images (localized holes sto, missing end and mispick) are shown in Table 2.

TABLE 2 Geometry feature extraction

Feature parameters	Holes sto	Missing end	Mispick
Length L	21	13	256
Width W	28	252	29
Ratio of length and width B	0.75	0.0515	12.8
Area S	353	536	3487
Perimeter C	94	415	784
Dispersion A	25.031	321.31	176.271

5 The fabric defect classification and identification

Classification and identification is the last step of fabric defect detection, which is also the objective of the defect detection system. The final classification and identification are based on the extracted features of the fabric defects. The classification methods can be summarized into three categories: statistical, syntactic and fuzzy logic [18]. Classification based on BP (Back Propagation) neural network is one of the most important statistical methods and is also the most commonly used one.

5.1 BP NEURAL NETWORK: A BRIEF REVIEW

Artificial neural networks, which are inspired by the animal's central nervous systems, are presented as systems of interconnected neurons and used to approximate unknown functions. Due to the adaptive nature of neural networks, they are capable of machine learning and pattern recognition. There are mainly two connection forms of artificial neural networks: hierarchical and all connection. Hierarchical neural network is made up of several layers, the number of neurons in each layer can be selected according to the requirements of the problem. Error back propagation neural network is one of the typical hierarchical neural networks, which is widely used in classification, function approximation and so on.

In general, BP neural network uses the feed-forward network structure, which consists of three kinds of layers: input layer, hidden layers (middle layers) and output layer. The design of BP neural network is to determine the number of layers, the number of neurons in each layer and various parameters. The structure is illustrated in Figure 11.
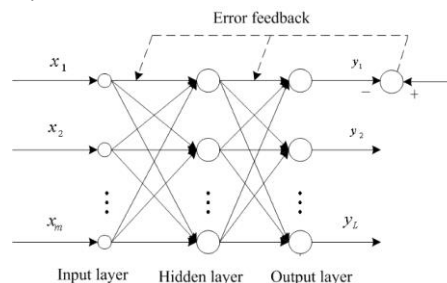
FIGURE 11 The BP neural network structure

The working process of the BP neural network can be divided into two stages. The first stage is the training of the neural network model by known samples, which essentially means selecting one model from the set of allowed models to minimize the cost criterion. Most of the algorithms used in training artificial BP neural network adjust the connection weights in the network in each step to minimize the mean square error between the desired target value and the actual output value of the network. The adjustments of the connection weights are generally based on the gradient descent method. The second stage is the application of the neural network for the classification and identification of fabric defects. The specific working process is shown in Figure 12.
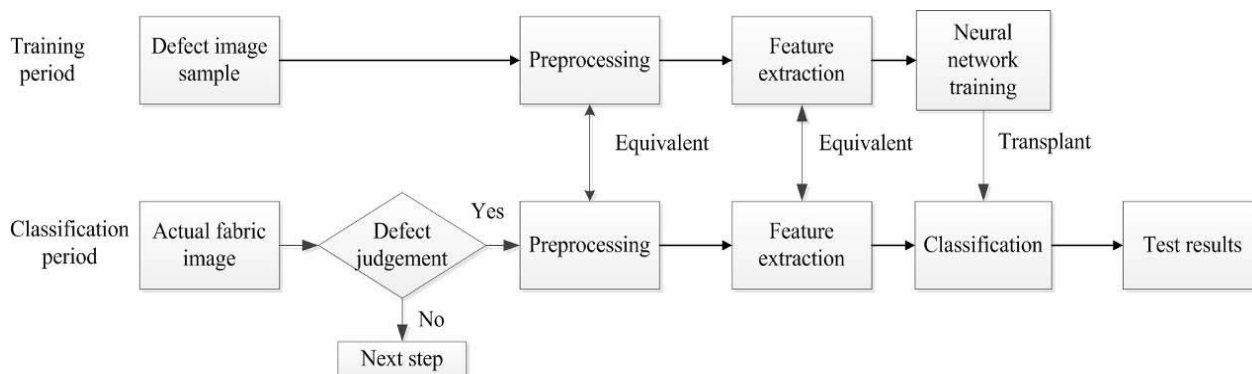
FIGURE 12 The working process of the BP neural network

5.2 THE DESIGN AND ANALYSIS OF THE FABRIC DEFECT CLASSIFIER

The multi-layered perceptron, which is regarded as the most appropriate neural network for pattern classification, is adopted for the fabric defect classifier. According to the characteristic of the fabric defects and the comparison study of network structures, a three-layered network (input layer, output layer and a hidden layer) is selected. The number of neurons in the input layer is set as the number of the input features 13. Since three typical classes of fabric defects (holes sto, missing end and mispick) are investigated, the number of neurons in the output layer is set as the number of the defined defect classes 3. As for the hidden layer, the number of the neurons is determined by experimentation which results in 8.

The three-layered network is then trained by supervised learning with defect samples which have been classified into three classes by human review in advance. All the initial weights are set to small random values at the beginning. All the input features are normalized to [0,1]. The learning rate is set as 0.3. Totally 100 defect samples are used for training and the convergence trajectory of the mean square error of the neural network is given in Figure 13.
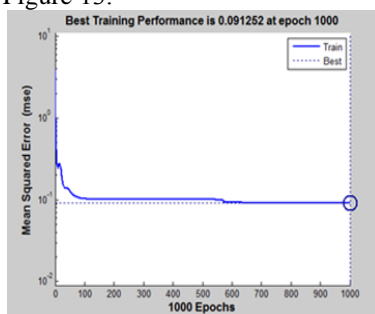
FIGURE 13 The convergence trajectory of the mean square error

After the training of the BP neural network, classification will be performed and each defect will be assigned to one of the specific defect classes. The process of the classifier for defects classification is shown in Figure 14.
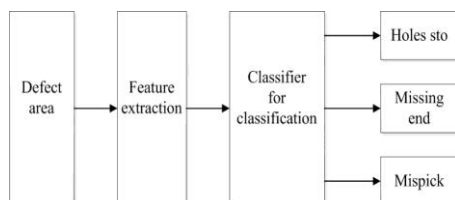
FIGURE 14 The classification and identification process of fabric defects

Totally 60 defect images are detected and used for the classification test: 20 defect samples for each of the three defect classes (holes sto, missing end and mispick). The classification result of the proposed classifier is given in Table 3.

TABLE 3 The classification result

Defect types	Test samples	Indentified	Accuracy
Holes sto	20	20	100%
Missing end	20	18	90%
Mispick	20	19	95%

The experimental results show that the identification rate of the classifier for holes sto is 100%, and the identification rates for missing end and mispick are both over 90%, which verify the effectiveness of the proposed classifier.

6 Conclusions

The development of an automated vision system to identify and classify defects on fabrics is presented in this paper. The hardware and software structures of the system are introduced. The defect detection process consists of image acquisition, image pre-processing, segmentation, localization and finally classification. A three-layered neural network is proposed for the classifier which results in high classification accuracy. Experiment results verify the effectiveness of the proposed method.

Future research work could be the issues associating with the industrial application, such as on-line computation burden, inspection cost, optimization of the algorithms and so on.

References

[1] Karayiannis Y A, Stojanovic R, Mitropoulos P, et al. 1999 Defect detection and classification on web textile fabric using multiresolution decomposition and neural networks *Proceedings of 6th IEEE International Conference on Electronics Circuits and Systems(Cat. No.99EX357)* **2** 765-8

[2] Mahajan PM, Kolhe SR, Patil PM 2009 A review of automatic fabric defect detection techniques *Advances in Computational Research* **1**(2) 18-29

[3] Stojanovic R, Mitropulis P, Koulamas C, et al. 2000 An approach for automated defect detection and neural classification of web textile fabric *Machine Graphics and Vision* **9**(3) 587-607

[4] Zheng G, Wang J 2007 Fabric defect detection method based on image distance difference *2007 8th International Conference on Electronic Measurement and Instruments ICEMI* 2822-5

[5] Kumar A, Pang G 2000 Fabric defect segmentation using multichannel blob defectors *Optical Engineering* **39**(12) 3176-90

[6] Kumar A, Pang G 2001 Identification of surface defects in textured materials using wavelet packets *Conference Record - IAS Annual Meeting (IEEE Industry Applications Society)* **1** 247-51

[7] Sari-Sarraf H, Goddard J S 1999 *IEEE Transactions on Industry Applications* **35**(6) 1252-9

[8] Hu M C, Tsai I S 2000 Fabric inspection based on best wavelet packet bases *Textile Research Journal* **70**(8) 662-70

[9] Chan C H, Pang G 2000 *IEEE Transactions on Industry Applications* **36**(12) 1267-76

[10] Tsai D M, Heish C Y 1999 Automated surface inspection for directional textures *Image and Vision Computing* **18**(1) 49-62

[11] Bodnarova A, Bennamoun M, Latham S 2002 Optimal Gabor filters for textile flaw detection *Pattern Recognition* **35**(12) 2973-91

[12] Kumar A, Pang G 2002 *IEEE Transactions on Industry Applications* **38**(2) 425-40

[13] Mak K L, Peng P 2008 An automated inspection system for textile fabrics based on Gabor filters *Robotics and Computer-Integrated Manufacturing* **24**(3) 359-69

[14] Stepien K 2014 Research on a surface texture analysis by digital signal processing methods *Tehnicki Vjesnik-Technical Gazette* **21**(4) 689-95

[15] Raheja J L, Ajay B, Chaudhary A 2013 Real time fabric defect detection system on an embedded DSP platform *Optik* **124**(21) 5280-4

[16] Kwak C, Ventura J A, Tofang-sazi K 2000 A neural network approach for defect identification and classification on leather fabric *Journal of Intelligent Manufacturing* **11**(5) 485-99

[17] Zhang J-H, Liu Y-Q, Tian D 2013 The short-term wind power forecast based on phase-space reconstruction and neural networks *Journal of Digital Information Management* **11**(1)40-5

[18] Zhou Q Q, Zhou Q F, Ning Y P, Yang F, Lei J Y, Tian R, Liu M D 2012 A comparison of pattern classification approaches for structural damage identification *Journal of Digital Information Management* **10**(2) 126-30

Authors

Zhaozhun Zhong, born in September, 1980, Suzhou City, Jiangsu Province, China

Current position, grades: the lecturer of School of Iron and Steel, Soochow University, China.
University studies: PhD in Automation from Shanghai Jiaotong University in China.
Scientific interests: digital image processing, real-time control and simulation.
Publications: 15 papers.
Experience: teaching experience of 3 years, completed 3 scientific research projects.

Pengjie Qi, born in May, 1989, Suzhou City, Jiangsu Province, China

Current position, grades: the graduate student of School of Mechanical and Electrical Engineering, Soochow University, China.
University studies: B.Sc. in Mechanical Engineering from Jiujiang University of Jiangxi in China.
Scientific interest: digital image processing, pattern recognition.
Publications: 3 papers.
Experience: completed 1 scientific research projects.

Miao Guan, born in February, 1979, Suzhou City, Jiangsu Province, China

Current position, grades: the researcher of School of Iron and Steel, Soochow University, China.
University studies: M.Sc. in Computer Science from Soochow University in China.
Scientific interest: digital image processing.
Publications: 10 papers.
Experience: completed 2 scientific research projects.

Yuedong Xia, born in November, 1992, Suzhou City, Jiangsu Province, China

Current position, grades: the graduate student of School of Mechanical and Electrical Engineering, Soochow University, China.
University studies: B.Sc. in Automation from Soochow University in China.
Scientific interest: real-time control.
Publications: 3 papers.
Experience: completed 1 scientific research projects.

Yanhui Fu, born in October, 1978, Plymouth, USA

Current position, grades: senior project engineer of Humanetics Innovative Solutions, Inc., Plymouth, USA.
University studies: M.Sc. in Mechanical Engineering from Nanjing University of Science and Technology in China.
Scientific interest: mechanical system design.
Publications: 5 papers.
Experience: completed 2 scientific research projects.

Joint resource allocation based on Nash bargaining game for wireless cooperative networks

Shuanglin Huang[*], Jianjun Tan

School of Information Engineering, Hubei University for Nationalities, 39 Xueyuan Road, Enshi, Hubei Province, China

Received 27 October 2014, www.cmnt.lv

Abstract

This paper considers the problem of resource sharing among selfish nodes in wireless cooperative networks. In the system, each node can be acted as a source as well as a potential relay, and both nodes are willing to achieve an extra rate increase by jointly adjusting their channel bandwidth and power levels for cooperative relaying. Nash bargaining solution (NBS) is applied to formulate the JBPA problem to guarantee fairness. Simulation results indicate the NBS resource sharing is fair and the fairness of resource allocation only depends on how much contribution its partner can make to its rate increase.

Keywords: Nash bargaining solution, bandwidth allocation, power control

1 Introduction

Cooperative diversity has been proposed for wireless network applications to enhance system coverage, link reliability and data transmission, and to decrease bit error rate (BER) [1] in recent years. Generally, all nodes in a non-commercial wireless network are assumed cooperative. In our daily life, the market often serves as a central platform where buyers and sellers gather together, negotiate transactions and exchange goods so that they can be satisfied immediately through bargaining and buying or selling. Similarly, the cooperation game theory just provides a flexible and natural tool to explore how the selfish nodes bargain with each other and mutual aid.

The pioneering work can be found in the following references. In [2], based on the NBS, the authors proposed a novel two-tier quality of service (QoS) framework and a scheduling scheme for QoS provisioning in worldwide interoperability for microwave access networks. A resource allocating scheme based on the NBS for downlink orthogonal frequency-division multiple access (OFDMA) wireless networks was proposed in [3]. The authors in [4] proposed a cooperation bandwidth allocation strategy for the throughput per unit power increase. In [5] the authors considered a bandwidth exchange incentive mechanism as a means of providing incentive for forwarding data. However, only bandwidth allocation problem was considered to encourage cooperation in [3-5]. In [6-8] the power allocation problem was considered to encourage cooperation. The authors in [6] considered fair power sharing between a user and its partner for an optimal signal-to-noise ratio (SNR) increase. From an energy-efficiency, perspective based no NBS, the authors in [7] studied a cellular framework including two mobile users

desiring to communicate with a common base station. In order to obtain both user fairness and network efficiency, a cooperative power-control game model based on Nash bargaining was formulated in [8]. In order to improve the fairness of virtual bandwidth allocation for multi tenants, especially for virtual links, the authors in [9] proposed a utility-maximization model for bandwidth sharing between virtual networks in the same substrate network.

However, the bandwidth only or power only allocation problem was studied in previous work, ignoring the JBPA in wireless network communication. Motivated by the aforementioned works, we constructed a symmetric wireless system model consisting of two user nodes and two destination nodes, which is shown in Figure 1.
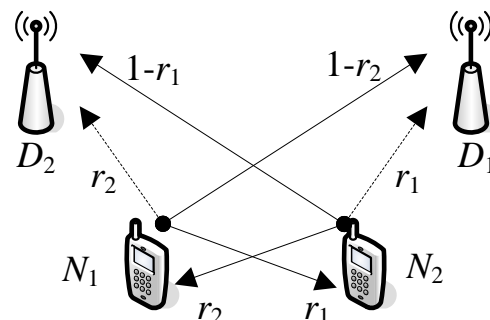
FIGURE 1 The system model for cooperative transmission

In the model, it is assumed that each user acts as a source as well as a potential relay. Furthermore, the proposed model represents a more general scenario, comparing to previous work. By bandwidth and power exchanging, each user has the opportunity to share the other's resources (e.g., bandwidth and power) and seek other user's help to relay its data to obtain the cooperative diversity, and vice versa. The cooperation degree between

---

[*]*Corresponding author* e-mail: huang-shuanglin@163.com

partners depends on two factors: one is the bandwidth and power contribution of each node on the cooperative rate increment; and another is their channel condition of each node on its cooperation benefit in terms of cooperative rate increment.

2 System model and cooperative schemes

As is shown in Figure 1, there is a symmetric cooperative communication system model [1] consists of two source nodes, N_1 and N_2, and their corresponding destination nodes, D_1 and D_2 (in particular, $D_1=D_2$). The two cooperating nodes communicate independent information over the orthogonal channels to the destinations.

The AF cooperation protocol is used in the model. The AF cooperative transmission between the two nodes occurs in two time slots. The system model is based on frequency division multiple access and each user occupies W hertz bandwidth for transmission. The total power consumptions of each user in the two time slots are the same.

As illustrated in Figure 2, N_1 and N_2 work independently to transmit their own data to D_1 and D_2 at time slot 1 with power P_1 and time slot 2 with power P_2 respectively. However, each user may seek other user's cooperation to relay its data to obtain the cooperative diversity. As shown in Figure 3, in time slot 1, N_1 transmits its own data independently with a part of its power, sharing a part of the common bandwidth, and N_1 simultaneously relays the data originating from node N_2 with another part of its power and another part of the common bandwidth. And vice versa for N_2 in the next time slot.
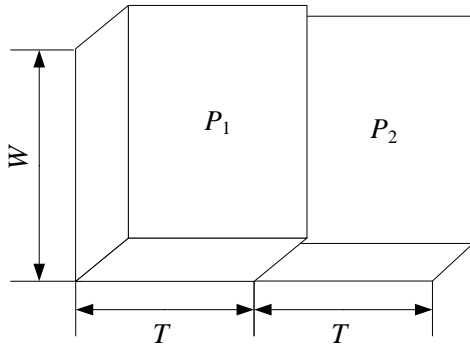
FIGURE 2 Direct transmission with interference
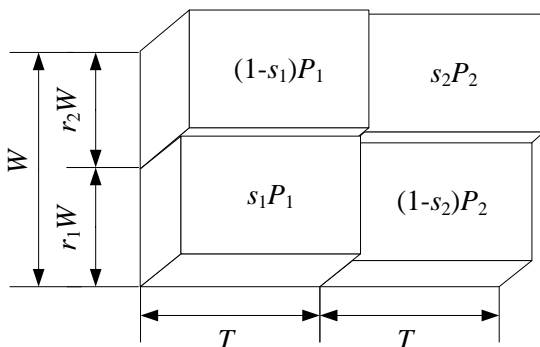
FIGURE 3 Time-division and frequency-division channel for JBPA

The details of cooperation between two nodes are illustrated in Figure 3. Specifically, in time slot 1, node N_1 allocates r_2 fraction ($r_2 \in (0, 1)$) of its bandwidth and $1-s_1$ fraction ($s_1 \in (0, 1)$) of its power P_1 to relay r_2 fraction of the data from node N_2, and it uses the r_1 fraction ($r_1 \in (0, 1)$) of the bandwidth and s_1 fraction ($s_1 \in (0, 1)$) of its power for its own data transmission. In time slot 2, node N_2 uses r_1 fraction ($r_1 \in (0,1)$) of the bandwidth and $1-s_2$ fraction ($s_2 \in (0, 1)$) of its power P_2 to forward the data originating from node N_1, and it uses the r_2 fraction of the bandwidth and s_2 fraction of its power for its own data transmission.

According to the cooperation details described above, a relay can forward no more than the amount of data as that originating from the source itself. There is $r_1=1-r_2$, which came from the result of [4]. Obviously, both r_1 and r_2 should be nonnegative for a meaningful cooperation. Then, we have

$$r_1 + r_2 = 1, \ r_1 > 0, \ r_2 > 0 . \tag{1}$$

Suppose that subscript denotes source node and superscript denotes destination node. Let G_i^j represents the channel gain between node N_i and node D_j ($i \neq j$), and let G_{ij} denote the channel gain from node N_i to node N_j. We assume that the noise power spectral density at different receivers is i.i.d. with the N_0. The cooperative transmission consists of two phases. In time slot 1, assumed that x_1 is the broadcasted signal from N_1 to N_2 and destination D_1, then, the achieved SNR helped by N_2 for N_1 to D_1 is given by [1]

$$\gamma_{12}^1 = \frac{s_1(1-s_2)P_2 G_2^1 P_1 G_{12}}{\sigma_1^2 \left[(1-s_1)P_2 G_2^1 + s_1 P_1 G_{12} + \sigma_1^2 \right]} . \tag{2}$$

And the effective rate of node N_1 at the D_1 is

$$r_1 R_{12}^{AF} = r_1 W \log\left(1 + \gamma_1^1 + \gamma_{12}^1\right) \tag{3}$$

where $\sigma_1^2 = r_1 W N_0$ and $\gamma_1^1 = s_1 P_1 G_1^1 / \sigma_1^2$ is the SNR that results from the direct transmission from node N_1 to the D_1 in the first time slot.

Similarly, the relayed SNR helped by N_1 for N_2 to D_2 is given by [1]

$$\gamma_{21}^2 = \frac{s_2(1-s_1)P_1 G_1^2 P_2 G_{12}}{\sigma_2^2 \left[(1-s_1)P_1 G_1^2 + s_2 P_2 G_{12} + \sigma_2^2 \right]} . \tag{4}$$

The effective rate of node N_2 at the D_2 is

$$r_2 R_{21}^{AF} = r_2 W \log\left(1 + \gamma_2^2 + \gamma_{21}^2\right) \tag{5}$$

where $\sigma_2^2 = r_2 W N_0$ and $\gamma_2^2 = s_2 P_2 G_2^2 / \sigma_2^2$ is the SNR that results from the direct transmission from node N_2 to the D_2 in the first time slot.

However, N_1 and N_2 may prefer working independently to transmit their own data, if it could make up the opportunity cost of cooperative transmission by

direct transmission. As illustrated in Figure 2. Then the direct rate at the D_1 is $R_1^D = W\log\left(1 + P_1 G_1^1 / \sigma_0^2\right)$, and the direct rate at the D_2 is $R_2^D = W\log\left(1 + P_2 G_2^2 / \sigma_0^2\right)$, where σ_0^2 is the AWGN received at the destination D_1 and D_2 on the condition of no partner for cooperation.

From the above introduction, it's clear that the resource allocation variables r_2 and s_1 reflect the N_1's rational decisions while r_1 and s_2 reflect the decisions of N_2, i.e., N_1 determines r_2 and N_2 determines r_1, and that the decisions of one user will affect the choices of its partner. Their pay out and payoff should be traded off and both users expect an optimal trade-off. The following sections will focus on in particular this problem's solution that can bring about win-win results.

3 Utility function and problem formulation

In this section, the utility functions of the source nodes are given and the model is analysed via the cooperative game theory. For node N_1 and node N_2, their utility functions U_1 and U_2 can be defined as $U_1 = r_1 R_{12}^{AF}$ and $U_2 = r_2 R_{21}^{AF}$, respectively. It is obvious that node N_1 will quit cooperation when its payoff is less than R_1^D, which ensures that a node would participate in cooperative transmission only if its effective rate is better than that of direct transmission. So the minimal values of U_1 and U_2 must be $U_1^{\min} = R_1^D$ and $U_2^{\min} = R_2^D$, respectively.

The bargaining problem of cooperative game theory can be described as follows. Let $K=\{1,2\}$ be the set of players. Let S be a closed and convex subset of R^K to represent the set of feasible payoff allocations that the players can get if they all work together. Let U_k^{min} be the minimal payoff that the k-th player would expect; otherwise, it will not cooperate. Suppose $\left\{U_k \in S \mid U_k > U_k^{\min}, \forall k \in K\right\}$ is a nonempty bounded set. Define $U^{min}=(U_1^{min}, U_2^{min})$, then the pair (S, U^{min}) is called a two-person bargaining problem.

As discussed above, each user has U_i as its objective function, where is bounded above and has a nonempty, closed, and convex support. The goal is to maximize all U_i simultaneously. U^{min} represents the minimal performance, and U^{min} is called the initial agreement point. The problem, then, is to find a simple way to choose the operating point in S for all users, such that this point is optimal and fair.

4 Joint resource allocation algorithm

Now, observe that the close form solution of (20) is impossibly to be got. So, we develop a numerical search algorithm, by which a global maximum not a local maximum will be obtained.

According to decomposition optimization theory [10], the optimization problem can be equivalently decomposed into the following two problems. Firstly, the maximal bandwidth solve

$$\tilde{U}^*(\tilde{r}_1*,\tilde{r}_2*,s_1,s_2) = \underset{r_1,r_2\in(0,1),r_1+r_2=1}{Arg\max}\ U(r_1,r_2,s_1,s_2), \forall s_1,s_2 \in (0,1)\ , (6)$$

where $\tilde{U}^*(\tilde{r}_1*,\tilde{r}_2*,s_1,s_2)$ is the maximal solution for given s_1 and s_2, not the optimal solution. $\tilde{r}_1*$ and $\tilde{r}_2*$ are the corresponding bandwidth allocation ratios.

Secondly, the optimal power allocation ratios solve

$$U^*(r_1*,r_2*,s_1*,s_2*) = \underset{s_1,s_2\in(0,1)}{Arg\max}\ \tilde{U}^*(\tilde{r}_1*,\tilde{r}_2*,s_1,s_2) \tag{7}$$

In this phase, we compare all $\tilde{U}^*(\tilde{r}_1*,\tilde{r}_2*,s_1,s_2)$ and choose the maximal one for all $\tilde{r}_1*$ and $\tilde{r}_2*$. This way, we can obtain the optimal solution.

In the following, Theorem 1 will be proved, which show that (6) and (7) all have a unique Nash equilibrium solution.

Theorem 1: For $\forall s_1,s_2 \in (0,1)$, the two-user bargaining game admits a unique Nash equilibrium solution $r=(r_1, r_2)$. For $\forall r_1,r_2 \in (0,1)$, the two-user bargaining game admits a unique Nash equilibrium solution $s=(s_1, s_2)$.

Proof: See Appendix A

For given s_1 and s_2, there exist the corresponding bandwidth allocation ratios $\tilde{r}_1*$ and $\tilde{r}_2*$. Substituting $\tilde{r}_1*$ and $\tilde{r}_2*$ into R_{12}^{AF} and R_{21}^{AF} respectively. This way, R_{12}^{AF} and R_{21}^{AF} will not include variables r_1 and r_2. So we have

$$\tilde{U}^*(\tilde{r}_1*,\tilde{r}_2*,s_1,s_2) = \underset{\substack{r_1,r_2\in(0,1),r_1+r_2=1\\,\forall s_1,s_2\in(0,1)}}{Arg\max}\ \left(r_1 R_{12}^{AF} - R_1^D\right)\left(r_2 R_{21}^{AF} - R_2^D\right) \tag{8}$$

For problem (8), by taking the derivative to r_1 and r_2 respectively, and equating them to zero, we get

$$\tilde{r}_1* = I_1\left(\tilde{r}_1*,\tilde{r}_2\right) = 0.5\left[1 + R_1^D\left(R_{12}^{AF}\right)^{-1} - R_2^D\left(R_{21}^{AF}\right)^{-1}\right], \tag{9}$$

$$\tilde{r}_2* = I_2\left(\tilde{r}_1*,\tilde{r}_2\right) = 0.5\left[1 - R_1^D\left(R_{12}^{AF}\right)^{-1} + R_2^D\left(R_{21}^{AF}\right)^{-1}\right]. \tag{10}$$

It is obvious that $\tilde{r}_1* + \tilde{r}_2* = 1$, which means that there is one variable only between $\tilde{r}_1*$ and $\tilde{r}_2*$. So the iterations of the bandwidth allocation ratio updating can be expressed as follows

$$\tilde{r}_i(t+1) = I_i\left(\tilde{r}_i(t)\right), i=1, 2\ . \tag{11}$$

We show next the convergence of the iterations in (11) by proving that the bandwidth allocation ratio updating function $I_i\left(\tilde{r}_i(t)\right)$ is a standard function [11].

Definition 1. A function $I_i\left(\tilde{r}_i\right)$ is standard if for all $\tilde{r}_i > 0$, the following properties are satisfied [11]:

1. Positivity. $I_i(\tilde{r}_i) > 0$;

2. Monotonicity. If $\tilde{r}_i > \tilde{r}_i'$, then $I_i(\tilde{r}_i) \geq I_i(\tilde{r}_i')$;

3. Scalability. For all $\alpha > 1$, $\alpha I_i(\tilde{r}_i) > I_i(\alpha\tilde{r}_i)$.

Proposition 1. The function $I_i(\tilde{r}_i)$ is standard.

Proof. See Appendix B.

In [11], a proof has been given. Starting from any feasible initial bandwidth allocation ratios r_1 and r_2, the bandwidth allocation ratios produced after several iterations of the standard bandwidth allocation algorithm gradually converges to a unique fixed point.

The problem (7) is a combinatorial problem which involves two continuous variables, s_1 and s_2. The optimal power allocation ratio s_1* and s_2* can be acquired by using gradient descent method.

5 Simulation results

To evaluate the performance of the proposed scheme, in what follows, the simulation results for JBPA are to be shown. A two-source and two-destination simulated system is conducted. Assumed that both nodes have the same initial transmission power with $P1 = P2 = 0.1$W, continuous strategy spaces with r_1, r_2, s_1, $s_2 \in (0, 1)$ and channel bandwidth with $W = 10^6$Hz. The path gain is set to $(7.75 \times 10^{-3})/d^{3.6}$, where d is the distance between the transmitter and the receiver (in meters). The noise level is 5×10^{-14}W. We locate the N_2 at $(0, 800)$，D_2 at $(0, 0)$, and D_1 at $(0, 1200)$, and fix the y coordinate of N_1 at 400 while increase its x coordinate from 0 to 500.

Let $X1$ denotes the x coordinate of node N_1. Figure 4 shows the NBS strategies, i.e., the optimal cooperation ratios (r_1, r_2, s_1, s_2) of both nodes when node N_1 moves. With the movement of N_1 from the (0,400) to (500,400), N_1's bandwidth allocation ratio is decreasing and that of N_2 is increasing correspondingly. The relation of cooperation bandwidth is $r_1 > r_2$, because N_2's channel condition is better than N_1's within this region, and thus, N_1 is willing to take out more bandwidth for cooperation to exchange for N_2's relaying. However, N_1 and N_2 are consuming most own power for themselves and small part of power for its partner. In N_1's moving process, N_2 is increasing more power for itself than N_1, because N_2's channel condition is better than N_1's within this region, and thus, N_1 is willing to take out more power for cooperation to exchange for N_2's relaying.

Figure 5 shows the performance comparisons of the proposed JBPA approach with the DT scheme, BA scheme and PA scheme. As shown in the Figure 5, the optimal rate obtained by BPJA approach is always bigger than the rate of BA scheme and the optimal rate obtained by BA scheme or by PA scheme is bigger than the rate of DT scheme, if the cooperation condition is satisfied and occurred. A node will choose not to cooperate when it can not get any benefit in the game.

From Figures 4 and 5, it's concluded that the proposed JBPA scheme could optimize the system performance while keeping the NBS fairness. The NBS fairness is embodied by the fact that the cooperative rate of each node is fundamentally determined by its channel condition, and that the cooperative rate increment of each node depends on its bandwidth and power contribution to maintain the cooperative transmission.
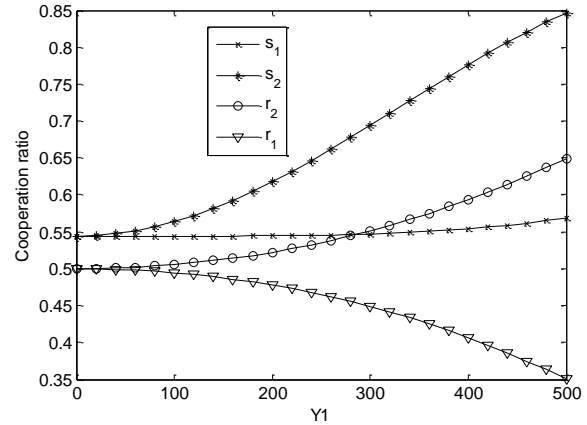
FIGURE 4 The optimal bandwidth and power cooperation ratios
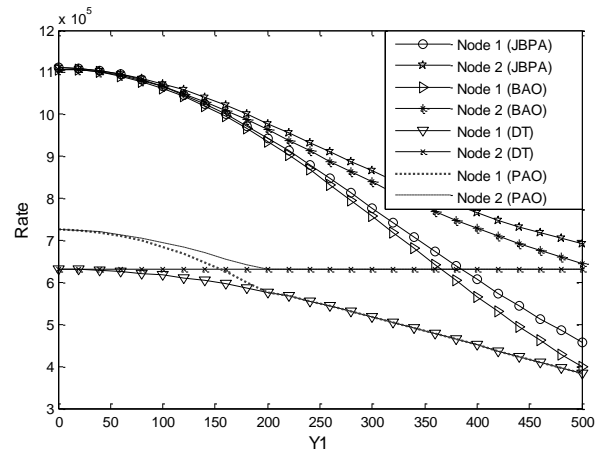
FIGURE 5 Three kinds of rates comparison

6 Conclusions

This paper has analysed the cooperation behaviour of selfish nodes in cooperative communication networks. We formulate the JBPA problem between two cooperating nodes as a cooperative game, and use the NBS function to obtain the solution of the game. Simulation results show that the resulting JBPA has the NBS fairness in that the degree of cooperation of a node only depends on how much contribution its partner can make to improve its performance. Furthermore, it is also shown that the proposed JBPA approach can achieve a comparable performance to that of the DT scheme, the BA scheme or PA scheme.

7 Acknowledgements

8 Appendices

8.1 APPENDIX A: PROOF OF THEOREM 1

Observe that the constraint set is convex. So if $U_i(r_i)$ is proved to be concave for r_i, Theorem 1 will be proved. For simplified representation, we define

$$A_i = 1 + \frac{a_i s_i}{r_i} + \frac{b_i c_i s_i (1 - s_j)}{r_i \left[b_i s_i + c_i (1 - s_j) + r_i \right]}.$$

So we have

$$\frac{\partial U_i}{\partial r_i} = W \log A_i - \frac{WA_i^{-1}}{\ln 2} \left(\frac{a_i s_i}{r_i} + \frac{b_i c_i s_i (1 - s_j) \left[b_i s_i + c_i (1 - s_j) + 2 r_i \right]}{r_i \left[b_i s_i + c_i (1 - s_j) + r_i \right]^2} \right) \tag{A12}$$

$$\frac{\partial^2 U_i}{\partial r_i^2} = \frac{WA_i^{-1}}{\ln 2} \frac{2 b_i c_i s_i (1 - s_j)}{(b_i s_i + c_i (1 - s_j) + r_i)^3} - \frac{WA_i^{-2}}{\ln 2} [A_i - 1 + \frac{b_i c_i s_i (1 - s_j)}{(b_i s_i + c_i (1 - s_j) + r_i)^2}]^2 \tag{A13}$$

then $\frac{\partial^2 U_i}{\partial r_i^2} < 0$. Therefore, $U_i(r_i)$ is concave for r_i.

Observe that the constraint set is convex. So if $U_1(s_1, s_2)$ is proved to be concave for s_1 and s_2, Theorem 1 will be proved. For simplified representation, we define $f_1(s_1) = 1/s_1$ and $f_2(s_2) = 1/(1 - s_2)$. For $f_1(s_1)$ and $f_2(s_2)$ are convex function. So $f_3(s_1, s_2) = b_1 f_2(s_2) + c_1 f_1(s_1) + r_1 f_1(s_1) f_2(s_2)$ is convex and $f_3(s_1, s_2)^{-1}$ is concave. So we have

$$U_1(s_1, s_2) = r_1 W \log \left(1 + \frac{a_1 s_1}{r_1} + \frac{b_1 c_1}{r_1 f_3(s_1, s_2)} \right) \tag{A14}$$

Furthermore, $f(u) = \log(1 + u), \forall u > 0$ is monotone increasing concave function. Since the compound function $f(u) = \log(1 + u(x, y))$ is concave if $u(x, y)$ is concave and $u(x, y) > 0$. Considering a positive linear combination of concave functions is concave. This way, $U_1(s_1, s_2)$ is proved to be concave for s_1 and s_2.

8.2 APPENDIX B: PROOF OF PROPOSITION 1

1. Positivity. It is obvious that $I_i(\tilde{r}_i) > 0$.

2. Monotonicity. If $\tilde{r}_i > \tilde{r}_i'$, then $I_i(\tilde{r}_i) \geq I_i(\tilde{r}_i')$.

For R_{ij}^{AF} and R_{ji}^{AF}, by taking the derivative to r_i respectively, we have

$$\frac{\partial R_{ij}^{AF}}{\partial r_i} = -\frac{WA_i^{-1}}{\ln 2} \left(\frac{a_i s_i}{r_i^2} + \frac{b_i c_i s_i (1 - s_j) \left[b_i s_i + c_i (1 - s_j) + 2 r_i \right]}{r_i^2 \left[b_i s_i + c_i (1 - s_j) + r_i \right]^2} \right) \tag{B15}$$

and

$$\frac{\partial R_{ji}^{AF}}{\partial r_i} = \frac{WA_j^{-1}}{\ln 2} \left(\frac{a_j s_j}{(1 - r_i)^2} + \frac{b_j c_j s_j (1 - s_i) \left[b_j s_j + c_j (1 - s_i) + 2(1 - r_i) \right]}{(1 - r_i)^2 \left[b_j s_j + c_j (1 - s_i) + 1 - r_i \right]^2} \right) \tag{B16}$$

So R_{ij}^{AF} is monotone decreasing function and R_{ji}^{AF} is monotone increasing function for r_i. Then, is $R_i^D \left(R_{ij}^{AF} \right)^{-1} - R_j^D \left(R_{ji}^{AF} \right)^{-1}$ monotone increasing function for r_i. Therefore, $I_i(\tilde{r}_i)$ is monotone increasing function for $\tilde{r}_i$.

3. Scalability.

For all $\alpha > 1$, let $\Delta I = \alpha I_i(\tilde{r}_i) - I_i(\alpha \tilde{r}_i)$.

Since $U_i(r_i)$ is monotone increasing function for r_i. $U_j(1 - r_i)$ is monotone decreasing function. So we have

$$\alpha R_{ij}^{AF}(\alpha \tilde{r}_i) = \frac{\alpha \tilde{r}_i R_{ij}^{AF}(\alpha \tilde{r}_i)}{\tilde{r}_i} > R_{ij}^{AF}(\tilde{r}_i) \tag{B17}$$

and

$$(1-\alpha\tilde{r}_i)R_{ji}^{AF}\left(1-\alpha\tilde{r}_i\right) < (1-\tilde{r}_i)R_{ji}^{AF}\left(1-\tilde{r}_i\right) \tag{B18}$$

Then, we have

$$\Delta I = \frac{\alpha-1}{2} + R_i^D \frac{\alpha R_{ij}^{AF}\left(\alpha\tilde{r}_i\right) - R_{ij}^{AF}\left(\tilde{r}_i\right)}{2R_{ij}^{AF}\left(\tilde{r}_i\right)R_{ij}^{AF}\left(\alpha\tilde{r}_i\right)} + R_j^D \frac{R_{ji}^{AF}\left(\tilde{r}_i\right) - \alpha R_{ji}^{AF}\left(\alpha\tilde{r}_i\right)}{2R_{ji}^{AF}\left(\alpha\tilde{r}_i\right)R_{ji}^{AF}\left(\tilde{r}_i\right)} \tag{B19}$$

$$\Delta I > \frac{\alpha-1}{2} + R_j^D \frac{R_{ji}^{AF}\left(\tilde{r}_i\right) - \alpha R_{ji}^{AF}\left(\alpha\tilde{r}_i\right)}{2R_{ji}^{AF}\left(\alpha\tilde{r}_i\right)R_{ji}^{AF}\left(\tilde{r}_i\right)} \tag{B20}$$

$$\Delta I > \frac{\alpha-1}{2}\left[1 - \frac{R_j^D}{(1-\alpha\tilde{r}_i)R_{ji}^{AF}\left(\alpha\tilde{r}_i\right)}\right] > 0 \tag{B21}$$

For $(1-\alpha\tilde{r}_i)R_{ji}^{AF}\left(\alpha\tilde{r}_i\right) > R_j^D$, which is the cooperation condition. Therefore, we can claim that $\alpha I_i\left(\tilde{r}_i\right) > I_i\left(\alpha\tilde{r}_i\right)$.

References

[1] Laneman J N, Tse D N C, Wornell G W 2004 *IEEE Transactions on Information Theory* **50**(12) 3062-80

[2] Jiao Y, Ma M, Yu Q, et al. 2011 Quality of service provisioning in worldwide interoperability for microwave access networks based on cooperative game theory *IET communications* **5**(3) 284-95

[3] Zhang G P, Zhang H L 2008 Adaptive resource allocation for downlink ofdma networks using cooperative game theory *11th IEEE SINGAPORE International Conference on Communications Systems* 98-103

[4] Zhang Z, Shi J, Chen H-H, et al. 2008 A cooperation strategy based on Nash bargaining solution in cooperative relay networks *IEEE Trans. Veh. Technol* **57**(4) 2570-77

[5] Qi Y, Tang M, Zhang M 2014 Mass customization in flat organization: The mediating role of supply chain planning and corporation coordination. *Journal of Applied Research and Technology* **12**(2) 171-81

[6] Zhang G P, Zhang H L, Zhao L Q, et al. 2009 Fair resource sharing for cooperative relay networks using nash bargaining solutions *IEEE Communications Letters***13**(6) 381-3

[7] Xiong K, Zhang Y, Zhang Z, Wang S, Zhong Z 2014 PA-NEMO: Proxy mobile IPv6-aided network mobility management scheme for 6LoWPAN *Elektronika ir Elektrotechnika* **20** (3) 98-103

[8] Yang C G, Li J D, Tian Z 2010 *IEEE Trans. Veh. Technol* **59**(4) 1696-706

[9] Zhang C, Huang L, Zhao Z 2013 Research on combination forecast of port cargo throughput based on time series and causality analysis *Journal of Industrial Engineering and Management* **6**(1) 124-34

[10] Boyd S, Vandenberghe L 2004 Convex optimization *Cambridge Univ. Press*

[11] Yates R 1995 *IEEE Selected Areas in Comm* **13**(7) 1341-48

Authors

Shuanglin Huang, born in August, 1975, Enshi City, Hubei Province, P.R. China

Current position, grades: the assistant professor of School of Information Engineering, Hubei University for Nationalities, China.
University studies: MSc from Taiyuan University of Technology, PSc from Huazhong University of Science and Technology in China.
Scientific interest: wireless networks, wireless communication system and game theory.
Publications: 10 papers.
Experience: teaching experience of 3 years, scientific research project.

Jianjun Tan, born in August, 1960, Enshi City, Hubei Province, P.R. China

Current position, grades: the professor of School of Information Engineering, Hubei University for Nationalities, China.
University studies: BSc in Electronic Engineering from Tsinghua University in China.
Scientific interest: wireless sensor network and wireless communication system.
Publications: 20 papers.
Experience: teaching experience of 30 years, 10 scientific research projects.

Adhesive image segmentation based on watershed algorithm

Xiaoyu Zhang*, Ning Liu

Department of Computer Science and Application, Zhengzhou Institute of Aeronautical Industry Management, Zhengzhou, Henan, 450015, China

Received 1 March 2014, www.cmnt.lv

Abstract

Since adhesive image has a deficiency of over-segmentation, the work applied watershed algorithm, whose merits include quick computing speed, closed outline and accurate location, to adhesive image segmentation. In the first few sessions, the basic principle and arithmetic steps of watershed algorithm were illustrated in detail. On this basis, following sessions correspondingly introduced the MATLAB simulation analysis and verification. Through simulation and contrast of the segmentation results by different algorithms, it could be concluded that watershed algorithm, the most effective method, could help split adhesive objects into single ones, while greatly reducing or even eliminating the over-segmentation phenomenon.

Keywords: Image Segmentation, Edge Detection, Threshold Segmentation, Watershed Algorithm, Image Processing, MATLAB Software

1 Introduction

With development of new theories and methods, image segmentation algorithms combined with feature theory are proposed. Thereinto, the image segmentation algorithm based on morphology is widely used at present [1-2]. Adhesive image, with disadvantage of over-segmentation, is segmented using watershed algorithm with advantages such as quick computing speed, closed outline and accurate location. Simulation shows that watershed algorithm, the most effective method, could help split adhesive objects into single ones, while greatly reducing or even eliminating the over-segmentation phenomenon.

2 Principle of watershed algorithm

Watershed algorithm, as an image segmentation method based on mathematical morphology, has been widely used, with advantages such as quick computing speed, closed outline and accurate location. It was firstly introduced into simple binary image processing by Digabel [3] and Lantuejoul [4]. After that, watershed theory was developed by Beucher [5], Vincent [6], et al. to get a general mathematical model for the application of grey-level image segmentation.

The basic thought of watershed algorithm derives from topography. A grey-level image can be seen as a topographic relief, where the grey level of a pixel is interpreted as its altitude in the relief. In the image, the local minimum and adjacent influence region correspond to a catchment basin, with its limits called watershed.

Watershed transformation can be described by the raindrop method. Namely, a drop of water falling on a topographic relief flows along a path to finally reach a local minimum, thus forming a connected region called catchment basin.

By watershed transformation, the original image can be transformed into the marker image using the same marker in a catchment basin and a special marker in watershed in image segmentation field.

3 Mathematical description of watershed algorithm

There is a grey-level image with the maximal and minimal grey levels of h_max and h_min, respectively. A recursive process is defined as the water level h varies from maximal grey level to minimum. In the recursive process, each catchment basin relevant to different local minimums constantly expands. $X(h)$ is a union of sets of catchment basins while water level is h. If water level is $h+1$, then a connected component $T(h+1)$ will be a new local minimum or a basin expansion of an existent $X(h)$. For the latter, the distances, between each point with an altitude of $h+1$ and catchment basins, are calculated according to adjacency relation. If one point is the same distance from more than two basins, then it will not belong to any basin; or else it will belong to the nearest basin, thus producing new $X(h+1)$.

$MIN(h)$ is defined as the local minimum at the altitude of h, and $Y(h+1,X(h))$ is marked as the set of points, which belong to $X(h)$ and have the altitude of $h+1$ [7-8].

$$X(h\_\min) = \{ p \in D \mid f(p) = h\_\min \} = T(h\_\min) , \qquad (1)$$

$$X(h+1) = MIN(h+1) \cup X(h) \cup Y(h+1, X(h)) . \qquad (2)$$

Watershed transformation $Watershed(f)$ is the complementary set of $X(h\_max)$, expressed as follows:

$$Watershed(f) = D \setminus (h\_\max) . \qquad (3)$$

* *Corresponding author* e-mail zzialn@139.com

Information and Computer Technologies

4 Procedure of watershed algorithm

(1) Reading grey-level image
(2) Creating structure elements
(3) Enhancing contrast of image
(4) Increasing distances between objects
(5) Transforming goal objects
(6) Valley detection
(7) Watershed transformation
(8) Extracting feature from label matrix

5 Simulation of adhesive image segmentation

Taking image segmentation as research object, the adhesive objects can be segmented into single ones using watershed algorithm.
(1)   Reading grey-level image
         afm=imread('afmsurf.tiff');
         imshow(afm)
Figure 1 shows the original grey-level image.
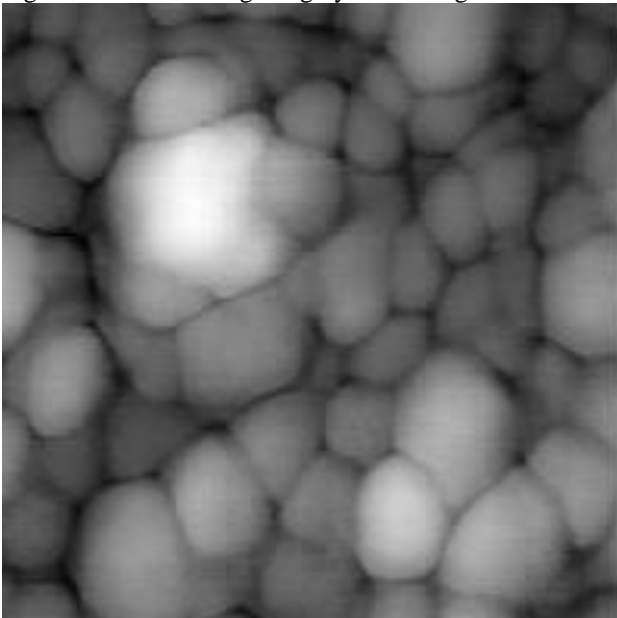
FIGURE 1 Original grey-level image

In Figure 1, adhesive objects with different scales derive that the original image has uneven surface. The valley points of the image can be found using watershed algorithm to achieve effective segmentation of different objects [9-10].

For best segmentation, the goal objects in the image are enhanced to minimizing the number of valley points. Meanwhile, the contrast of the image is increased using top-hat and bottom transformations.
(2)   Creating structure elements Se=strel('disk',15);
(3)   Enhancing contrast of image

Itop=imtophat(afm,se); % Enhancing contrast by top-hat transformation

Ibot=imbothat(afm,se); % Enhancing contrast by bottom-hat transformation

figure,imshow(Itop,[]),title('tophat result')

  figure,imshow(Ibot,[]),title('bottomhat result')

Figures 2 and 3 are the top-hat and bottom-hat images, respectively. The purpose is to enhance contrast of the images.
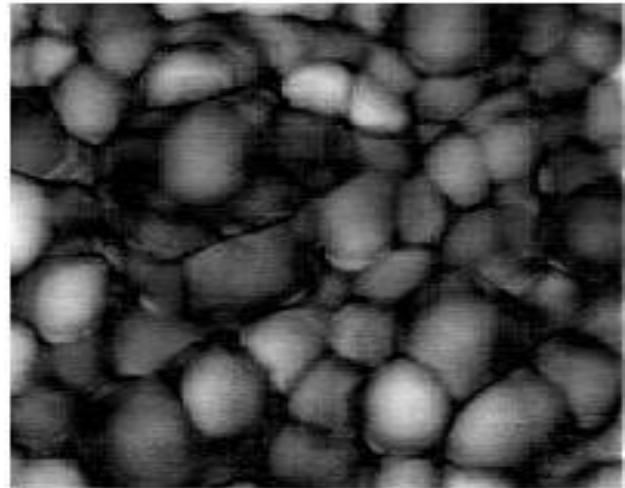
FIGURE 2 Top-hat image
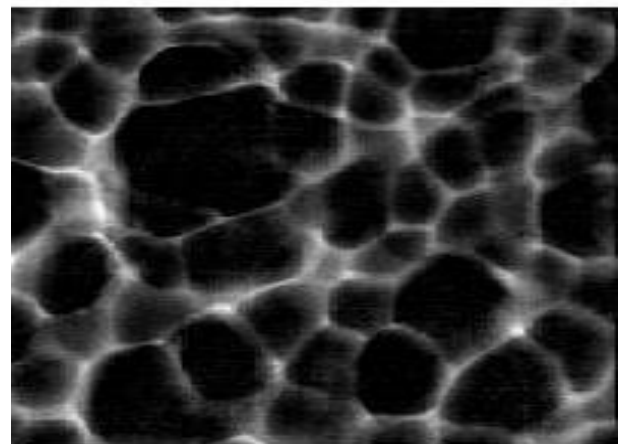
FIGURE 3 Bottom-hat image

(4) Increasing distances between objects

Figure 2 presents that the distances between objects should be increased because of close contact of objects in the image. The original image plus the top-hat transformation image subtracts the bottom-hat transformation image to achieve gap increase between objects [11].

Ienhance=imsubtract(imadd(Itop,afm),Ibot);

 %(original+tophat-bottomhat)Increasing contrast of gaps and objects

figure,imshow(Ienhance),title('original+tophat-bottomhat')

Figure 4 shows the image of increased gap.
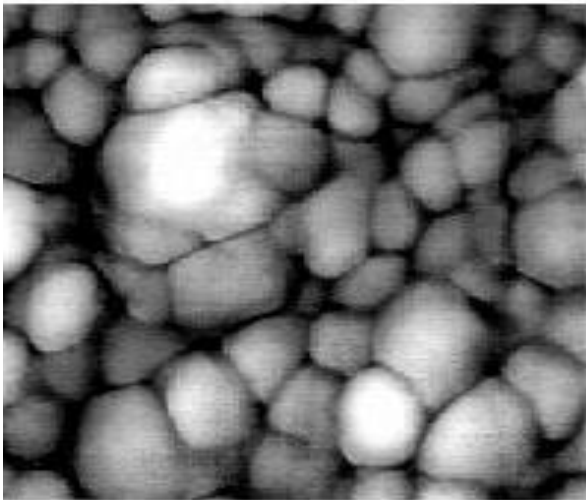
FIGURE 4 Image of increased gap

(5) Transforming goal objects

Before effective detection, valley points of the image are enhanced using imcomplement function.

Iec=imcomplement(Ienhance); %Enhancing valley points of the image

figure,imshow(Iec),title('complement of enhanced image')

Figure 5 shows the image of enhanced valley points.
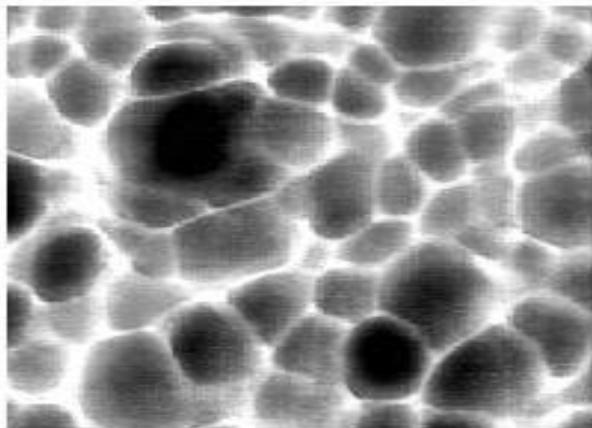
FIGURE 5 Image of enhanced valley points

(6) Valley detection

Figure 6 shows the valley detection images after enhancing valley points by Step 5.

Iemin=imextendedmin(Iec,22); %Detecting valley points

Iimpose=imimposemin(Iec,Iemin); % Detecting valley points

figure,imshow(Iemin),title('extended minima image')

figure,imshow(Iimpose),title('imposed minima image')
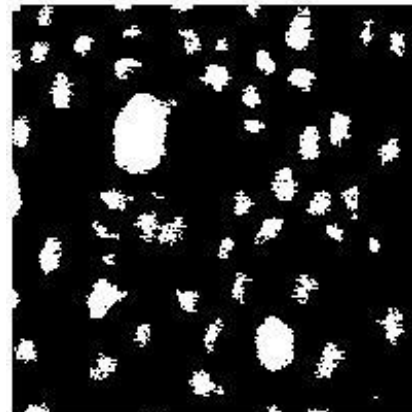
(a)Imextendedmin detection image
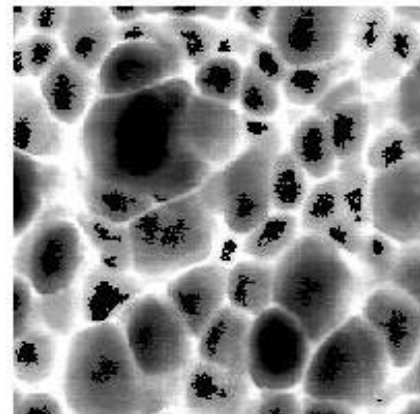
(b)Imimposemin detection image
FIGURE 6 Valley detection images

(7) Watershed transformation

After pretreatment of the original image, the image after valley detection can be conducted with watershed transformation. Figure 7 shows watershed segmented image, expressed as the visual label matrix.

wat=watershed(Iimpose); % Label matrix derived from watershed transformation

rgb=label2rgb(wat); % Transforming label matrix into color matrix to get watershed segmented image

figure,imshow(rgb),title('watershed segmented image')
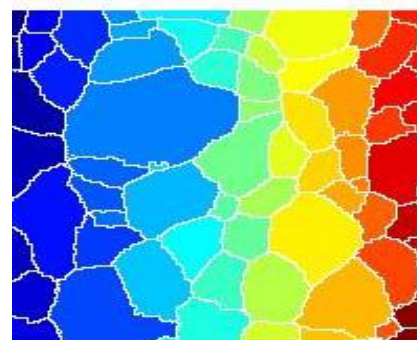
FIGURE 7 Watershed segmented image

279

(8) Extracting features from label matrix

The features of the image are extracted using regionprops function based on two observed orientations. Figure 8 shows the diagram of extracted features.

stats=regionprops(wat,'Area','Orientation');   % Extracting features from label matrix

area=[stats(:).Area]; %Area

orient=[stats(:).Orientation]; %Orientation

figure,plot(area,orient,'b*'),title('Relationship   of Particle Orientation to Area');

xlabel('particle   area   (pixels)'),ylabel('particle orientation (degrees)')

Figure 7 shows that the adhesive objects are transformed into single ones, with over-segmentation reduced. Simulation shows that watershed algorithm, the most effective method, could help split adhesive objects into single ones, while greatly reducing or even eliminating the over-segmentation phenomenon.

6 Comparisons of segmentation results using different algorithms

Segmentation results using different algorithms are contrasted to prove the efficiency of segmentation results using watershed algorithm (See in Figure 9).

FIGURE 8 Diagram of extracted features
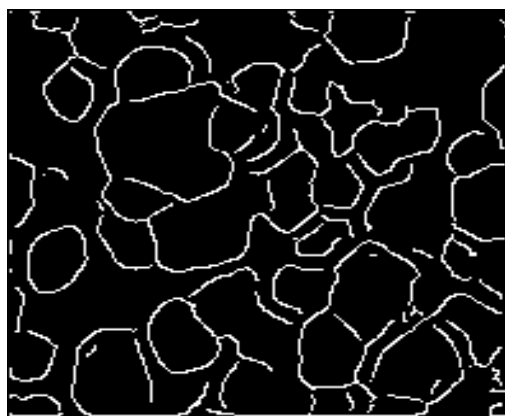
(a) Canny operator
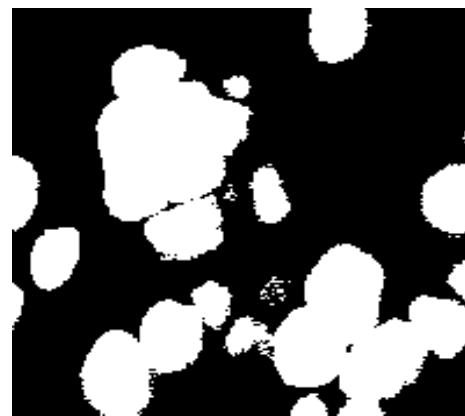
(b) Iterative threshold segmentation
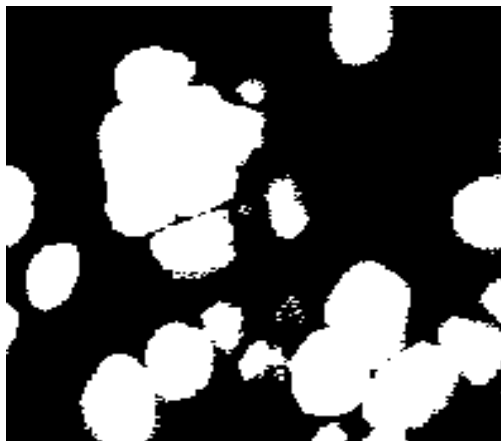
(c) Ostu segmentation
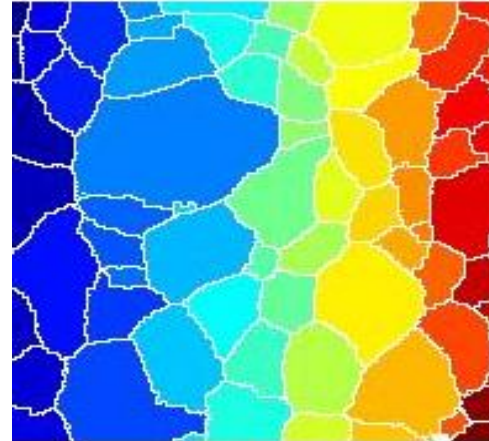
(d) Watershed algorithm

FIGURE 9 Comparison images of segmentation result using different algorithms

280

Figure 9 shows that watershed algorithm, the most effective method, could help split adhesive objects into single ones, while greatly reducing or even eliminating the over-segmentation phenomenon.

7 Conclusions

Watershed algorithm is applied in the work aiming at slow calculation and low segmentation quality of traditional image segmentation methods, especially for poor performance in adhesive image segmentation. Simulation shows that watershed algorithm, the most effective method, could help split adhesive objects into single ones, while greatly reducing or even eliminating the over-segmentation phenomenon.

References

[1] Song Huansheng, Liu Chunyang, Wu Chengke 2009 Multi-scale Ridge Edges and Their Applications in Object Segmentation *Acta Automatic Sinica* **25**(6) 12-5
[2] Zhangjing, Wang Honggang, Wang Yongtian 2010 Image Segmentation based on Fringe Pick-up *Optical Technique* **27**(5) 24-6
[3] Lin Feng, Liu Zhengkai 2000 Eye Edge Extraction Based on Image Segmentation *Application Research of Computers* **8**(3) 100-3
[4] Yin Dekui, Zhang Baomin, Bai Lianfa 2009 Multi-template Edge Detection Method Based on Thermal Image *Journal of Nanjing University of Science and Technology* **23**(1) 16-20
[5] Zhang Bin, Zhu Zhengzhong 2011 Multi-source Remote Sensing Image Segmentation Based on Edge and Contour Information *Journal of Image and Graphics* 3(5) 830-5
[6] Wang Yusheng 2012 Edge Detection Algorithm Based on Integral Transformation *Journal of Image and Graphics* **7**(2) 145-9
[7] Wang Guangjun, Tian Jinwen, Liujian 2008 Image Segmentation Technique Based on Quadtree Structure *Journal of Huazhong University of Science and Technology* (2) 12-6
[8] Liu Ningning, Tian Jie 2009 Interactive Image Segmentation Method and Its Application Based on Regional Characteristics *Journal of Software* **19**(3) 10-8
[9] Qian Xiaofeng 1999 An Algorithm for Color Image Region Segmentation and Vectorization Based on Contour Tracing in Anticlockwise Order *Journal of Data Acquisition & Processing* **16**(1) 52-7
[10] Cheng Jie 1999 Segmentation Based on Histogram *Journal of Huazhong University of Science and Technology* **27**(1) 20-3
[11] Zhao Xuesong, Chen Shuzheng 2001 Image Segmentation Based on Global Binarization and Edge Detection *Journal of Computer-Aided Design & Computer Graphics* **13**(2) 118-21

Authors

Xiaoyu Zhang, born in April, 1973, Luoyang, Henan Province, China

Current position, grades: Associate Professor, Master
University studies: Image Processing
Scientific interest: Image Processing, Pattern Recognition
Publications: 20 papers published in the international or national journals
Experience: She is an associate professor of Zhengzhou Institute of Aeronautical Industry Management, China. Her research interests include Image Processing and Pattern Recognition.

Ning Liu, born in July, 1980, Xuchang, Henan Province, China

Current position, grades: Lecturer, Master, Member of China Computer Federation
University studies: Image Processing
Scientific interest: Image Processing, Pattern Recognition
Publications: 12 papers published in the international or national journals
Experience: She is a lecturer of Zhengzhou Institute of Aeronautical Industry Management, China. Her research interests include Image Processing and Pattern Recognition.

Information and Computer Technologies

An improved sift image matching detection

Ning Liu*, Xiaoyu Zhang

Department of Computer Science and Applications, Zhengzhou Institute of Aeronautical Industry Management, China

Received 1 May 2014, www.cmnt.lv

Abstract

Aiming at the problems - edge response in the traditional SIFT descriptor and the insufficient correct matching feature points, the work proposed a kind of improved SIFT Image Matching Detection Algorithm. The candidate key point was firstly detected by the SIFT algorithm; Canny edge detection algorithm was used to detect image edge points; it was judged whether the candidate key point needed to be eradicated by comparing whether the candidate key point equals to the coordinates of edge points; K-means clustering pattern, which is combined by the vector space cosine similarity and vector Euclidean distance similarity, was adopted to perform global image similarity matching. Finally, RANSAC algorithm was used to further get rid of the wrong matching. The experimental result indicates the improved method greatly enhances the stability of SIFT algorithmic and the accuracy rate of matching.

Keywords: SIFT operator, Canny operator, K-means cluster, accuracy rate, key point, edge detection

1 Introduction

Image matching technique has become the research focus in the fields of computer vision and digital image processing. At present, image matching methods are mainly based on two methods [1] of pixel and features. Especially the SIFT algorithm that is presented by David G. Lowe [2-3] in 1999, with favourable invariance property on the conditions of image rotation, measure scaling, affine transformation and change of view. Based on that the identification and matching of SIFT feature are used in many fields, such as target identification [4-5], image stitching [6], and mobile robot localization and mapping [7-8].

Even though SIFT algorithm is successfully applied in many fields, SIFT algorithm itself also has obvious defects, because the DOG operator adopted in SIFT algorithm has strong edge response; therefore lots of unstable edge response points exist in the produced feature points, and the Canny edge detection algorithm in this work removes unstable edge response points. In the field of engineering mathematics, the most common vector similarity measurement method is that the cosine similarity determines its degree of vector direction difference by comparing the included angle of the vector space between two vectors. The more consistent the direction is, the greater the similarity is; on the contrary, the similarity is less in the opposite direction. This similarity evaluation method only applies the direction information of the vector; therefore, this work proposes a kind of better similarity, the combination of included angle cosine similarity and the Euclidean distance. Finally, we use RANSAC to remove the wrong matching and mismatching. The experimental result indicates that this algorithm keeps the robustness of the previous algorithm, and meanwhile improves the accuracy rate of the algorithm stability and matching.

2 Feature extraction of the SIFT and Canny combination

Because the DOG operator used in SIFT algorithm will produce strong edge response, the generated feature points contain lots of unstable edge response points. Therefore, this work adopts Canny edge detection algorithm on the basis of the classic SIFT features to extract more stable features aiming at the edge response, providing a good foundation for the later feature matching. The feature extraction mainly includes following five steps.

1. Construct DOG scale space

In order to effectively detect the stable key point in scale space, Lowe [3] puts forward the DOG scale-space, which is produced by different scales of DOG and image convolution.

$$D(x,y,\sigma) = \left[G(x,y,k\sigma) - G(x,y,\sigma) \right] * I(x,y),$$
$$= L(x,y,k\sigma) - L(x,y,\sigma)$$

(1)

where k is the fixed parameter, $G(x,y,\sigma) = \frac{1}{2\pi\sigma^2} e^{\frac{-(x^2+y^2)}{2\sigma^2}}$,

$I(x,y)$ is the original image, and σ is scale coordinate.

2. Detecting the extreme point of the scale space

In the process of detecting the scale space extreme value, the current detected pixel needs to compare with the 8 adjacent pixels at the same scale and the 9*2 pixel at the closely corresponding position of the 2 high and low scale.

3. Removing the unstable edge point

---

* *Corresponding author* e-mail zzialn@139.com

For all of the candidate feature points detected by extreme value detection, it is needed to eradicate the unstable edge point by means of the 2 following steps.

The first step is to eradicate the unstable edge response point by means of 2*2 Hessian matrixes. The form of H matrix is

$$H = \begin{bmatrix} D_{xx} & D_{xy} \\ D_{xy} & D_{yy} \end{bmatrix} . \tag{2}$$

The stability measurement η is

$$\eta = \frac{\left(D_{xx} + D_{yy}\right)^2}{D_{xx}D_{yy} - D_{xy}^2} < \frac{\left(\gamma + 1\right)^2}{\gamma} , \tag{3}$$

where γ is the specific value between the maximum feature value and the minimum feature value, which is applied to control the stability of the feature point.

The second step is to eradicate the unstable edge point by means of Canny edge detection algorithm.

Firstly, we need to calculate the position of each feature point F_1 on the original image after eradicating unstable edge response by means of principal curvatures, and then use Canny algorithm to detect the edge point. For each edge point F_2, we use Canny to calculate the point set F_3 within the 3*3 field. Finally, we check if the coordinates are equal by performing comparison on the candidate key point F_1 and edge point F_2. The key point F_1 is abandoned if it's equal. If it's unequal, F_1 continues to compare with F_3. If something in point set F_3 is equal to the F_1 coordinate, the key point F_1 is abandoned. The key point F_1 is preserved if no equal points exist.

4. Calculating the direction of the key point

The DOG convolution image is used to determine the only direction of the feature point in order to make the feature vector satisfy the rotational invariance. The formula of calculating the gradient magnitude m and the direction θ is [11-12]:

$$m = \sqrt{\left(L_{x+1,y} - L_{x-1,y}\right)^2 + \left(L_{x,y+1} - L_{x,y-1}\right)^2} , \tag{4}$$

$$\theta = \arctan \frac{L_{x,y+1} - L_{x,y-1}}{L_{x+1,y} - L_{x-1,y}} . \tag{5}$$

In order to improve the stability of the gradient calculation, the neighbourhood gradient histogram with the feature point as the centre is calculated, and the peak value of the histogram represents the direction of the feature point.

5. Constructing the descriptor of the feature vector

After choosing 16*16 field with feature point as the centre, it is divided into 4*4 image subblocks, 8 directions of vector information of each pixel is identified, and thus the $4 \times 4 \times 8 = 128$ dimension feature vector can be generated. The illumination invariance of the vector can be improved by means of the uniformization on the feature vector.

3 Matching strategy improvement of feature vector

The classic SIFT feature matching is realized by calculating the Euclidean distance between all the feature points in an image and those in another image. For example, a feature point a_1 in the Image A wants to find an accurate matching point b_1 from the Image B, the Euclidean distance between the feature point a_1 and all the feature points in the Image B, and then b_1 is obtained after comparison. Because the Euclidean distance of all the features in the Image B should be calculated when matching is performed on each feature point, the calculated amount is large, and the information of feature points cannot be sufficiently used.

1. The following aspects are improved, and the feature points are separated into several kinds for performing K-means cluster on the stable SIFT feature points, the basic principle is that higher similarity exists between the same kind, but the difference is larger between different kinds.

2. The measurement form of the feature descriptor similarity is improved, and uses the combination of the Euclidean distance and included angle cosine to replace the Euclidean distance, thus sufficiently applies the size and direction of the feature vector, enhancing the accuracy rate.

3. Using RANSAC to cancel the wrong matching, and improve the accuracy rate.

3.1 FEATURE POINTS CLUSTER BY K-MEANS

The K-means plan is adopted in this work in order to perform matching on the feature points with similar descriptors during the matching. The process of calculation:

1. After inputting A and B images, the stable SIFT features are calculated, which are feature group des1 and des2, respectively.

2. Use K-means cluster feature to separate the feature group des 1 and des 2 into 5 groups, the 5 feature groups of the first group are des11, des12, des13, des14 and des15; and the 5 feature groups in the second group are des21, des22, des23, des24 and des25.

3. The cluster centre in the 5 feature groups of the first group are separately c11, c12, c13, c14 and c15, the cluster centre in the 5 feature groups of the second group are c21, c22, c23, c24 and c25, respectively. The closest cluster centre from the 5 cluster centres are found in the second group, and all the feature points with the c11 are used as the cluster centre to be matched with all the feature points of the closest cluster centre to the c11. Matching can be performed on the other feature groups in the first group accordingly.

3.2 USING THE EUCLIDEAN DISTANCE AND THE INCLUDED ANGLE COSINE AS THE SIMILARITY

In the classic SIFT algorithm, the Euclidean distance of SIFT feature vector is adopted to calculate the similarity between a pair of SIFT features in two images, thus only the size of the feature vector is used, but the direction of feature vectors are not sufficiently used, so this work uses both Euclidean distance and the included angle cosine. The distance between vectors satisfying the Euclidean distance and the direction between the vectors should satisfy the included angle cosine, thus the size and direction of the feature vector can be sufficiently used.

The Euclidean distance between two n dimension points $x = (x_1, x_2, x_3, ... x_n)$ and $y = (y_1, y_2, y_3, ... y_n)$ is identified as:

$$D_o(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \ . \tag{6}$$

The included angle cosine is identified as:

$$\cos\theta = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} \ . \tag{7}$$

The data range for the included angle cosine is [-1,1]. The included angle of two vectors is smaller if the included angle cosine is larger. The maximum value 1 of the included angle cosine shall be performed when the direction of the two vectors is overlapped, when the direction of two vectors is completed opposite, the minimum value 1 of the included angle cosine shall be adopted.

3.3 USE RANSAC TO CANCEL THE WRONG MATCHING

This work adopts RANSAC to cancel the wrong matching. The input of the RANSAC algorithm contains a group of observed data and a parameter model

appropriate for observing the data. The steps of RANSAC algorithm are as follows [13-15]:

1. There is a model, which is appropriate for the assumed inside point, which means all of the unknown parameters can be calculated from the assumed inside points.

2. The model zone obtained from 1. is used to test all of the other data, if some point is appropriate for the estimated model, which is recognized as the inside point.

3. If sufficient pints are classified into the assumed inside points, then the estimated model is sufficiently reasonable.

4. Use all the assumed inside points to reevaluate the model.

5. Finally, the model is evaluated by assessing the error rate of the inside points and the model.

4 Experimental result and analysis

The original SIFT feature matching algorithm and the improved algorithm are separately used to perform rectification on the image and analyse the result. The experiment is performed under the hardware environment of Intel(R) Core(TM)2 Duo CPU E6550 2.33GHz and 2 GB computer memory, and the simulated experimental environment is the MATLAB 7.8 of Windows XP.

(1) In the feature extraction phase, several experiments and data indicate, the feature extracted by the SIFT algorithm after being optimized by the Canny algorithm is more stable than the original SIFT feature, which meanwhile provides good basis for the later matching. 30 pairs of images are selected in the experiment; each pair is captured from different angles, the images of the same scene with different scaling, different rotation angles and view change are used as the required data of the experiment. Limited by the length of this work, one group of images will be mainly analysed in image 1; the size of the image is 640*480. The two images of this group are marked by the original reference image and the original image without rectification, respectively.

(a)Original reference image　　　　　　　　　(b)Original image without rectification

FIGURE 1 Original image

The unstable edge point is not favourably eradicated in the classic SIFT feature extraction, so the feature

extracted by the algorithm in this work is more stable. The following image is the feature extraction result and

the original one. Under the same conditions, the experimental result indicates that the SIFT feature presented in the first image is 1229, and the SIFT feature presented in the second image 2012. In the algorithm of this work, the feature extracted from the first image is 951, and the feature extracted from the second image 1551. 278 unstable edge points are removed in the first image feature, and 461 unstable edge points are

eradicated in the second image feature. 23% edge points are averagely removed. These edge points are the wrong results which are generated when the image is affected by the noise, the feature point extracted by the algorithm of this work greatly enhances the robustness of the SIFT algorithm. The experimental results of 5 images are presented in the following Table 1.

(a)SIFT feature extraction

(b)SIFT feature extraction

FIGURE 2 Original SIFT feature

TABLE 1 Quantity comparison of feature points between the classic SIFT algorithm and the algorithm

	The 1st image	The 2nd image	The 3rd image	The 4th image	The 5th image
Feature points of classic SIFT	847	479	1535	936	596
Feature points extracted by the algorithm in this paper	604	299	1397	711	427
Unstable edge points eradicated by the Canny	243	180	132	225	169

(2) In the phase of rectification, the original SIFT matching algorithm and the matching algorithm are separately used to perform feature matching on the Figure 1 in order to testify the validity of the algorithm. The final matching feature points of the original SIFT matching algorithm are totally 65. The feature matching in the matching algorithm of this work is used, the

matching points of the first group is 70, the matching points of the second group 51, the matching points of the third group 29, the matching points of the fourth group 83, the matching points of the fifth group 50, and finally the matching points 283, which is 4 times of the previous algorithm. The experimental result is shown as the following figures.

Original SIFT feature matching result

Matching result of the 1st group in the algorithm of this paper

Matching result of the 2nd group in the algorithm of this paper

Matching result of the 3rd group in the algorithm of this paper

Matching result of the 4th group in the algorithm of this paper

Matching result of the 5th group in the algorithm of this paper

FIGURE 4 Feature matching

TABLE 2 Matching quantity comparison of feature points between the classic SIFT algorithm and the algorithm

	The 1st pair of image	The 2nd pair of image	The 3rd pair of image	The 4th pair of image
Matching feature points of the original SIFT matching algorithm	19	31	54	72
Matching feature points of the matching algorithm in this paper	49	102	224	280

The experimental results indicate that the matching feature points of the algorithm increase obviously compared to the original algorithm, which enhances the robustness of the feature and the stability of the matching.

5 Conclusions

On the basis of analysing the SIFT feature matching algorithm, this work puts forward that the improved accurate image matching algorithm based on the stable SIFT reduces the wrong matching rate, which improves the accuracy rate of the image matching. The work adopts Canny to remove unstable edge points on the classic SIFT feature, extracts the SIFT feature which has strong anti-noise capacity, and meanwhile performs K-means cluster on the extracted feature points, performs feature matching on the corresponding cluster area, Euclidean distance and the included angle cosine are adopted to be as similarity during matching, and completely apply the size and direction of the feature vector. Finally, the RANSAC algorithm is used to remove the mismatching. The algorithm in this work improves the stability and accuracy rate of the matching.

Acknowledgement

References

[1] Stein A, Hebert M 2005 Incorporating Background Invariance Into Feature-based Object Recognition *IEEE Workshops on Application of Computer Vision,Breckenridge,CO,USA*

[2] Lowe D G 2009 Object Recognition from Local Scale-invariant Features *Proceeding of the Seventh IEEE International Conference on Computer Vision.Kerkyra,Greece* 1150-7

[3] Lowe D G 2004 Distinctive Image Features from Scale-invariant Keypoints *International Journal of Computer Vision* **60**(2) 91-110

[4] Zhang Shuzhen, Song Hailong, Xiang Xiaoyan, Zhao Yingnan 2010 Fast SIFT Algorithm Adopted to Realize the Target Identification *Computer System Application* **19**(6)

[5] Xiong Ying, Ma Huimin 2010 The Extraction and Application of the SIFT Feature for 3-D Objects *China Image Journal* **15**(5) 814-9

[6] Zhang Chaowei, Zhou Yan, Wu Sili 2008 Monitoring Image Automatic Combination Base on the SIFT Feature Matching *Computer Application* **28**(1) 191-4

[7] Stephen S, Lowe D G, Little J J 2005 Vision – based Global Localization and Mapping for Mobile Robots *IEEE Transactions on Robots* **21**(3) 364-75

[8] Ting T O, Rao M V C, Loo C K 2006 A novel approach for unit commitment problem via an effective hybrid particle swarm optimization *IEEE Trans, Power Syst.* **21**(1) 411-8

[9] Song Y, Chen Z, Yuan Z 2007 New chaotic PSO-based neural network predictive control for nonlinear process *IEEE Trans. Neural Netw.* **18**(2) 595-601

[10] Tao W B, Jin H, Liu L M 2008 Object segmentation using ant colony optimization algorithm and fuzzy entropy *Pattern Recognition Letters* **28**(7) 788-96

[11] Yin P Y, Chen L H 1997 A fast iterative scheme for multilevel thresholding methods *Signal Process* **60**(3) 305-13

[12] Cao L, Bao P, Shi Z K 2008 The strongest schema learning GA and its application to multilevel thresholding *Image and Vision Computing* **26**(5) 716-24

[13] Yin P P 2007 Multilevel minimum cross entropy threshold selection based on particle swarm optimization *Applied Mathematics and Computation* **184**(2) 503-13

[14] Zahara E, Fan S K S, Tsai D M 2005 Optimal multi-thresholding using a hybrid optimization approach *Pattern Recognition Letters* 2005 **26**(8) 1082-95

[15] Maitra M, Chatterjee A 2008 A hybrid cooperative-comprehensive learning based PSO algorithm for image segmentation using multilevel thresholding *Expert Systems with Applications* **34**(2) 1341-50

Authors

Xiaoyu Zhang, born in April, 1973, Luoyang, Henan Province, China

Current position, grades: Associate Professor, Master
University studies: Image Processing
Scientific interest: Image Processing, Pattern Recognition
Publications: 20 papers published in the international or national journals
Experience: She is an associate professor of Zhengzhou Institute of Aeronautical Industry Management, China. Her research interests include Image Processing and Pattern Recognition.

Ning Liu, born in July, 1980, Xuchang, Henan Province, China

Current position, grades: Lecturer, Master, Member of China Computer Federation
University studies: Image Processing
Scientific interest: Image Processing, Pattern Recognition
Publications: 12 papers published in the international or national journals
Experience: She is a lecturer of Zhengzhou Institute of Aeronautical Industry Management, China. Her research interests include Image Processing and Pattern Recognition.

Uncertain random fault tree analysis based on cloud security protection framework

Changyou Guo[1, 2, 3*], Xuefeng Zheng[1, 2], Jianjun Liu[3]

[1]*School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing, China*

[2]*Beijing Key Laboratory of Knowledge Engineering for Materials Science, Beijing, China*

[3]*Department of Computer Science, Dezhou University, Dezhou, China*

Abstract

Based on uncertainty theory and chance theory, this paper proposes a method that constructs and analyses fault tree. The fault tree is constructed based on logical relations between bottom events. Fault rate of bottom event would be characterized as random variable if it is obtained from historical data, it would be characterized as uncertain variable if it has no statistical data but is obtained from expert's subjective judgment. The chance that top event occurs is an uncertain random variable. The minimal cut set of fault tree is obtained by Boolean algebra method and at same time, the simplest standard disjunction expression of top event is obtained. This paper also constructs hybrid simulation algorithm to calculate the chance that top event occurs. At last validity of this method is confirmed by taking cloud security protection framework risk fault tree as example.

Keywords: Fault tree, Chance theory, Uncertain random variable, Cloud Computing, Cloud security

1 Introduction

Fault tree is a special inverted treelike logic graph, it consists of logical gate symbols and event symbols. Logical gates mean relationships between events. A fault tree describes an accident model and interprets the relations between malfunctions of components and observed symptoms with gate. Fault tree analysis is important to predict reliability for complex and large scaled system. It is a logical and diagrammatic method to evaluate the probability of an accident resulting from sequences and combinations of faults and failure events. Since fault tree analysis was developed in 1962 at Bell Telephone Laboratories in USA [1], it has been extensively used in many fields such as semi conductor industry, man-machine system, flexible manufacturing systems, nuclear power plants transmission pipelines, chemical industries and LNG terminal emergency shut down systems [3-7].

The conventional fault tree analysis based on probabilistic approach has been used extensively in the past, but still has the following problems.

(1) Fault tree analysis is carried on at early designed stage of system, so it is difficult to estimate precise failure rates or failure probabilities of individual components or failure events.

(2) New components are usually used or new events occur in practical system, historical data that is used in probabilistic approach cannot be obtained.

(3) In a highly automated system, people are still the key component. Human component is responsible for 20%-90% of the failures in many systems. It is very difficult to be characterized as probability [2].

Since fuzzy set theory (FST) developed by Zadeh [8] in 1965, many researchers tried to apply it to solve the above problems. Paper [1] applied the fuzzy sets theory to model the fuzzy system structure, proposed the new procedure to calculate the system reliability and a new importance index of basic events. It overcame shortcoming, which probability was difficult to be evaluated precisely in traditional fault tree analysis, it was a simple and effective fault tree analysis method. Paper [2] was suitable for situations which both probabilistic and fuzzy evaluations were necessary. Instead of directly estimating failure probability, the fuzzy failure rate was used to characterize the failure occurrence of system events involving imprecise information such as human errors. Paper [9, 10] analysed that how probability of top event is calculated by using triangle fuzzy number, trapezoidal fuzzy number, LR fuzzy number, normal fuzzy number as chance that bottom event occurs. Paper [11-13] applied fuzzy fault tree analysis to many fields such as nuclear reactor, aerospace, petrochemical industry, pipelines and so on.

Because fuzzy measure does not obey the law of truth conservation and is inconsistent with the law of excluded middle and the law of contradiction [14], the above fuzzy fault tree analysis methods lost their correct theory foundation.

---

[*] *Corresponding author* e-mail guochangyouustb@139.com

When historical data are not available to estimate a probability distribution, we have to invite some domain experts to evaluate their belief degree that each event will occur. Since human beings usually overweight unlikely events, the belief degree may have much larger variance than the real frequency. Perhaps some people think that the belief degree is subjective probability. However, paper [20] showed that it is inappropriate because probability theory may lead to counterintuitive results in this case. In order to deal with this phenomena, uncertainty theory was founded by [15] in 2007 and refined by [18] in 2010. Nowadays uncertainty theory has become a branch of mathematics for modelling human uncertainty, and have been developed and applied widely to operational research, risk analysis, reliability, comprehensive evaluation, portfolio selection, etc. [16, 17, 19, 14, 23-25].

In many cases, uncertainty and randomness simultaneously appear in a complex system. For example, some quantities have no samples while others have samples enough to determine probability distributions. In order to describe this phenomenon, the concepts of uncertain random variable and chance measure were pioneered by Liu in 2012 [14]. Chance theory begins with uncertain variable and chance measure, and is a mathematical methodology composed of uncertainty theory and probability theory.

Wen [26] recently proposed uncertain random fault tree analysis, she led uncertain random measure into Boolean system. But in this paper, fault rate of bottom event would be characterized as random variable indirectly if it is obtained from reliable historical data otherwise it would be characterized as uncertain variable. The chance that top event occurs is an uncertain random variable and is presented with simple formula.

The chance that overall system's top event occurs is calculated by using hybrid simulation technology. Finally, feasibility and validity of this method is confirmed by taking the cloud security protection framework risk fault tree as example.

Cloud Computing is a new concept in recent years, and a newly computing model is proposed. Cloud computing is the development of distributed computing, parallel computing and grid computing [27]. The goal of cloud computing is to simplify the computing and storage for like public water and electricity, the user can be convenient to use these resources only could be connected to network, and to pay by the volume that they used. Cloud computing is usually have a distributed infrastructure, and can carry on real-time monitoring of the distributed system, in order to achieve the efficient usage of it [28]. The computing make computers act on the cloud and the computer makes parallel computing technology into people's life [29]. Users service themselves relying on some internet information resources, which lie on some nodes, such as computing resources, software resources, data resources and management resources. This service model emphasize the

demand driven, user dominant, on-demand services, no centralized control and users don't care where the server. The parallel computing and virtualization technology has become the core support technology after the concept of cloud computing was put forward. There are existing two means of cloud computing [30]: one aspect is describes the infrastructure, which used to construct applications and the role equivalents to the PC operating system: the other aspect is describes cloud computing applications based on the infrastructure.

The main concern of cloud computing is the safety issue according to latest survey of IDC. Therefore, the security of user data will be the key factor decided cloud computing for enterprise applications [31]. Survey of Gartner in 2009 showed that more than 70% of the respondents in the actual deployment of cloud computing is security and privacy issues [32]. Cloud computing characterized by dynamic services as the main technical, flexible "service contract" as the core business characteristics, is undergoing significant changes. This change has brought huge impact for the information security field.

There have been more and more foreign standards organization started to develop cloud computing and safety standards, in order to enhance the interoperability and security, reduce duplication of investment or reinvent, some organizations such as ITU-TSG17 team [33] launched the cloud computing standard work. In addition, some specially group, such as cloud computing security alliance also has made certain progress in cloud computing security standardization. In the domestic IT industry, all kinds of cloud computing security products and solutions appear. For example, Sun Microsystems released open source cloud computing security tools for Amazon's EC2, S3, and virtual private cloud platform to provide security protection.

2 Preliminary

2.1 UNCERTAINTY THEORY

Definition 1 ([15]) Let Γ be a nonempty set, τ a σ-algebra over Γ, and M an uncertain measure, M meets the conditions: (1)$M\{\Gamma\} = 1$; (2)$M\{\Lambda\} + M\{\Lambda^c\} = 1$ for any event Λ; (3)$M\{\bigcup_{i=1}^{\infty} \Lambda_i\} \leq \sum_{i=1}^{\infty} M\{\Lambda_i\}$ for every countable sequence of events $\{\Lambda i\}$. Then the triplet (Γ, τ, M) is called an uncertainty space.

Definition 2 ([15]) An uncertain variable is a measurable function from an uncertainty space (Γ, τ, M) to the set of real numbers, i.e., for any Borel set B of real numbers, the set

$$\{\xi \in B\} = \{r \in \Gamma | \xi(r) \in B\} \tag{1}$$

is an event.

Definition 3 ([15]) The uncertainty distribution Φ of an uncertain variable ξ is defined by

$$\Phi(x) = M\{\xi \le x\} \qquad (2)$$

for any real number x.

Definition 4 ([15]) An uncertain variable ξ is called linear if it has a linear uncertainty distribution

$$\Phi(x) = \begin{cases} 0, if x \le a; \\ (x-a)/(b-a), if a \le x \le b \\ 1, if x \le b \end{cases} \qquad (3)$$

denoted by L(a,b) where a and b are real numbers with a < b(figure 1).

Definition 5 ([15]) An uncertain variable ξ is called zigzag if it has a zigzag uncertainty distribution

$$\Phi(x) = \begin{cases} 0, if x \le a; \\ (x-a)/2(b-a), if a \le x \le b \\ (x+c\text{-}2b)/2(c-b), if b \le x \le c \\ 1, if x \ge c \end{cases} \qquad (4)$$

denoted by Z(a, b,c) where a, b, c are real numbers with a < b < c (figure 2).
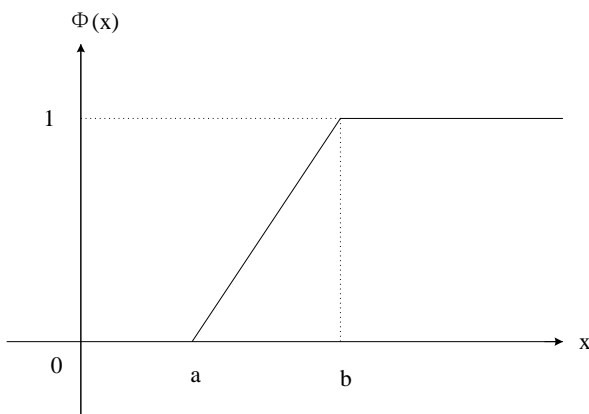
FIGURE 1 Linear Uncertainty Distribution
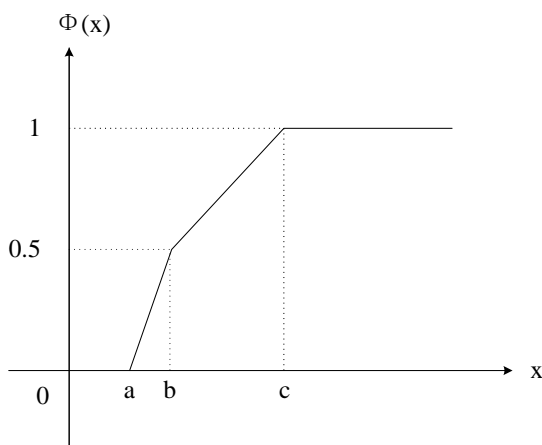
FIGURE 2 Zigzag Uncertainty Distribution

Definition 6 ([15]) An uncertain variable ξ is called normal if it has a normal uncertainty distribution

$$\Phi(x) = (1 + \exp(\frac{\pi(e-x)}{\sigma\sqrt{3}}))\text{-}1 , \quad x \in R \qquad (5)$$

denoted by N(e, σ) where e and σ are real numbers with $\sigma > 0$(figure 3).

Definition 7 ([15]) Let ξ be an uncertain variable. Then the expected value of ξ is defined by

$$E[\xi] = \int_0^{+\infty} M\{\xi \ge r\} \, dr - \int_{-\infty}^0 M\{\xi \le r\} dr$$

provided that at least one of the two integrals is finite.
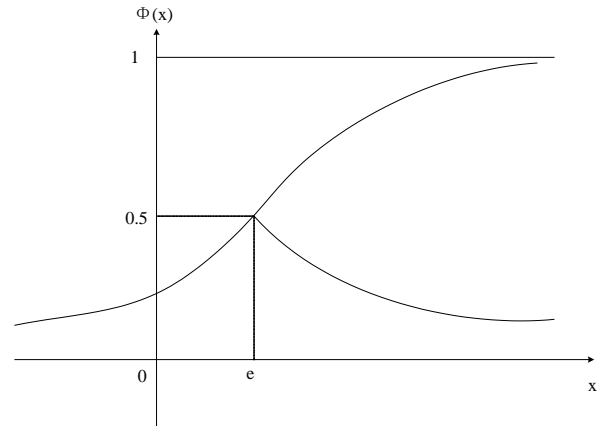
FIGURE 3 Normal Uncertainty Distribution

2.2 CHANCE THEORY

Definition 8 ([14]) An uncertain random variable is a function ξ from a probability space (Ω,A,Pr) to a collection of uncertain variables such that

$$M\{\xi(\omega) \in B\} \qquad (6)$$

is a measurable function of ω for any Borel set B of real numbers.

Definition 9 ([14]) Let f: $\Re$ n $\to$ $\Re$ be a measurable function, and $\xi1$, $\xi2$,…, ξn are uncertain random variables on the probability space (Ω,A,Pr). Then $\xi = f(\xi1, \xi2,…, \xi n)$ is an uncertain random variable defined by

$$\xi(\omega) = f(\xi1(\omega), \xi2(\omega), …, \xi n(\omega)). \qquad (7)$$

Definition 10 ([14]) Let ξ be an uncertain random variable, and let B be a borel set of real numbers. then the chance measure of uncertain random event $\xi \in B$ is defined by

$$Ch\{\xi B\} = \int_0^1 Pr\{\omega \in \Omega | M\{\xi(\omega) \in B\} \ge r\} dr. \qquad (8)$$

Definition 11 ([14]) Let ξ be an uncertain random variable. Then its chance distribution is defined by

$$\Phi(x) = Ch\{\xi \le x\} \qquad (9)$$

for any $x \in \Re$.

Theorem 1 ([14]) Let η_1, η_2,…, η_m be independent random variable with probability distributions ψ_1, ψ_2,…, ψ_m, and let τ_1, τ_2,…, τ_n be independent uncertain variables with uncertainty distributions γ_1, γ_2,…, γ_n, respectively.

Then the uncertain random variable $\xi = f(\eta_1, \eta_2, \ldots, \eta_m, \tau_1, \tau_2, \ldots, \tau_n)$ has a chance distribution ϕ (x)= $\displaystyle\int_{\Re_m} F(x;y_1,y_2,\ldots y_m), \quad d\Psi_1(y_1)d\Psi_2(y_2),\ldots,d\Psi_m(y_m)$, where $F(x; y_1, y_2, \ldots, y_m)$ is determined by its inverse function $F^{-1}(\alpha; y_1, y_2,\ldots, y_m) = f(y_1, y_2,\ldots, y_m, \gamma_1^{-1}(\alpha), \gamma_2^{-1}(\alpha)\ldots, \gamma_n^{-1}(\alpha))$ provide that $f(\eta_1, \eta_2, \ldots, \eta_m, \tau_1, \tau_2, \ldots, \tau_n)$ is a strictly increasing function with respect to $\tau_1, \tau_2, \ldots, \tau_n$.

Example 1 ([14]) Let $\eta_1, \eta_2, \ldots, \eta_m$ be independent random variables with probability distributions $\Psi_1, \Psi_2, \ldots, \Psi_m$, and let $\tau_1, \tau_2, \ldots, \tau_n$ be independent uncertain variables with uncertainty distributions $\gamma_1, \gamma_2, \ldots, \gamma_n$ respectively.

Then the minimum $\xi = \eta_1 \wedge \eta_2 \wedge \ldots \eta_m \wedge \tau_1 \wedge \tau_2 \wedge \ldots \wedge \tau_n$ is an uncertain random variable whose chance distribution is

$$\Phi(x) = \Psi(x) + \gamma(x) - \Psi(x)\gamma(x), \tag{10}$$

where ψ is the probability distribution of $\eta_1, \eta_2, \ldots, \eta_m$ determined by

$$\Psi(x) = 1 - (1 - \Psi_1(x))(1 - \Psi_2(x))\ldots(1 - \Psi_m(x)) \tag{11}$$

and γ is the uncertainty distribution of $\tau_1 \wedge \tau_1 \wedge \ldots \tau_n$ determined by

$$\gamma(x) = \gamma_1(x) \vee \gamma_2(x) \vee \ldots \gamma_n(x). \tag{12}$$

Example 2 ([14]) Let $\eta_1, \eta_2, \ldots, \eta_m$ be independent random variables with probability distributions $\Psi_1, \Psi_2, \ldots, \Psi_m$, $\tau_1, \tau_2, \ldots, \tau_n$ be independent uncertain variables with uncertainty distributions $\gamma_1, \gamma_2, \ldots, \gamma_n$ respectively. Then the maximum $\xi = \eta_1 \vee \eta_2 \vee \ldots \eta_m \vee \tau_1 \vee \tau_2 \vee \ldots \vee \tau_n$ is an uncertain random variable whose chance distribution is

$$\Phi(x) = \Psi(x)\gamma(x), \tag{13}$$

where Ψ is the probability distribution of $\eta_1 \vee \eta_2 \vee \ldots \eta_m$ determined by

$$\Psi(x) = \Psi_1(x)\Psi_2(x)\ldots\Psi_m(x) \tag{14}$$

and γ is the uncertainty distribution of $\tau_1 \vee \tau_2 \vee \ldots \vee \tau_n$ determined by

$$\gamma(x) = \gamma_1(x) \wedge \gamma_2(x) \wedge \ldots \wedge \gamma_n(x). \tag{15}$$

2.3 BASIC KNOWLEDGE ABOUT FAULT TREE

Cut set: a set of components whose failure interrupts all connections between input and output ends and thus causes an entire system to fail.

Minimal cut set: the smallest combination of components, which will cause the systems failure if they all fail.

Path set: a set of components whose success will make the system successful.

Minimal path set: the smallest combination of components whose success will make the system successful.

There are three methods to obtain minimal cut set of fault tree, Boolean algebra method, ling-row method and structural method. Because Boolean algebra method is simple and effective, it is used in the following analysis.

There are usually three steps to calculate the minimum cut sets by Boolean algebra method.

Step 1. Boolean expressions of top event about fault tree is established. Each layer of events is instead of last layer of events from top event. At last, the top event is instead of all bottom events.

Step 2. The Boolean expression is converted into standard disjunction expression.

Step 3. The standard disjunction expression is simplified to the simplest standard disjunction expression by logic operation rules in Boolean algebra.

3 Uncertain random fault tree analysis

We shall analyse system risk by using fault tree, so three risk definitions will be illustrated as follows.

Definition 12 ([14]) Assume that a system contains uncertain factors $\xi_1, \xi_2, \ldots, \xi_n$, and has a loss function f. Then the risk index is

$$risk = M\{f(\xi_1, \xi_2, \ldots, \xi_n) \geq 0\}.$$

Definition 13 ([14]) Assume that a system contains random factors $\xi_1, \xi_2, \ldots, \xi_n$, and has a loss function f. Then the risk index is

$$risk = \Pr\{f(\xi_1, \xi_2, \ldots, \xi_n) \geq 0\}.$$

Definition 14 ([14]) Assume that a system contains uncertain random factors $\xi_1, \xi_2, \ldots, \xi_n$, and has a loss function f. Then the risk index is

$$risk = \mathrm{Ch}\{f(\xi_1, \xi_2, \ldots, \xi_n) \geq 0\}.$$

For the implementation of uncertain random fault tree analysis for complex system (risk analysis), a systematic methodology is developed and given as follows:

Step 1. System modelling and planning. Identify the problem, carry out the preliminary analysis, gather the data and plan for the solution steps.

Step 2. Risk identification. Identify the top event and the sub-events as well as potential failure-consequence scenarios.

Step 3. Fault tree construction. Find failure logic and build up a fault tree using the bottom events.

Step 4. Bottom events classification. Fault rate of bottom event would be characterized as random variable

if it is obtained from historical data, otherwise it would be characterized as uncertain variable.

Step 5. Expression of top event simplification and obtainment. The expression is converted and simplified into the simplest standard disjunction expression and minimal cut set is obtained by Boolean algebra.

Step 6. Expression of top event conversion. The simplest standard disjunction expression is converted into another one based on variable type and Theorem 1.

Step 7. System risk calculation. The chance that top event (system risk) occurs is calculated by hybrid simulation algorithm.

Step 8. Risk management and control. analyse the results, monitor and review the process and propose countermeasures.

4 Uncertain random fault tree analysis application

The priority problem of cloud computing security needed to solve are these: establish a comprehensive cloud computing security model according threats, and actively carry out various key technology research. But in the long run, the security and privacy protection of user data needs is the core problem that is unable to avoid. The important security objectives of cloud users are the data security and privacy protection services.

A cloud security model is put forward through the research on data security in cloud computing security mechanism. As shown in Figure 4:
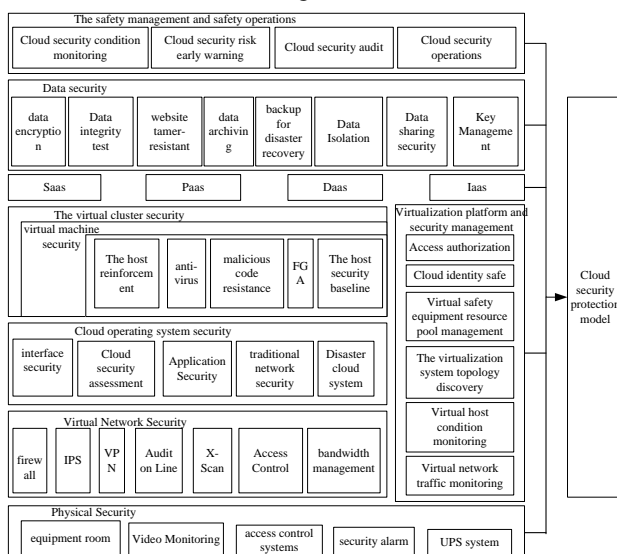
FIGURE 4 A Cloud Security Framework

Figure 4 mainly includes physical security, virtual network security, cloud operating system security, virtual cluster security, SaaS/PaaS/IaaS security, data security, safety management and safety operational and so on. In fact, cloud computing is introduced into the virtualization technology, and changed the service way, but did not overthrow the traditional safe mode.

Many risk factors that affect cloud security protection system include trusted access control, cipher text retrieval and processing, data exists and reusability, data privacy, virtual security technology and so on [34-37]. The

chances that some factors occur are obtained from historical data, so they are random, others have no historical statistical data but are obtained from questionnaire to experts, so they are uncertain. Therefore, this paper applies UR-FTA to cloud security protection framework risk analysis.

Step 1. System modelling and planning. This paper analyses cloud security protection framework from two aspects of internal risk and external risk, internal risk includes identity and security, data security and host security; external risk includes security audit, risk early warning, security condition monitoring and security resource management. It also analyses the characters of all factors and gathers the data and plans for the solution steps.

Step 2. Risk identification. Identify the top event and the sub-events as well as potential failure-consequence scenarios.

Step 3. Fault tree construction. The logical relationship between events is analysed and the fault tree is constructed, it is shown in Figure 5. In the figure, the top event E represents the entire cloud security protection framework risk, intermediate events B and C represent respectively external risks and internal risks in the framework, the bottom events 1, 2, 3 and 4 respectively security audit, risk early warning, security condition monitoring and security resource management, the bottom events 5, 6, 7 represent respectively identity and security, data security and host security. The logical relationships between them are shown in figure 5.

Step 4. Bottom events classification. Fault rate of bottom event would be characterized as random variable if it is obtained from historical data, otherwise it would be characterized as uncertain variable. It is shown in Table 1. Security audit, risk early warning, security condition monitoring and security resource management have historical statistical data, so η_1, η_2, η_3, η_4 are random variables with random distributions Ψ_1, Ψ_2, Ψ_3, Ψ_4; and identity and security, data security and host security have no statistical data, so η_5, η_6, η_7 are uncertain variables with uncertain distributions γ_5, γ_6, γ_7.
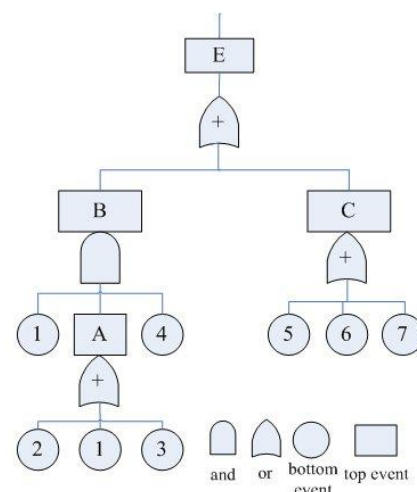
FIGURE 5 Cloud Security Protection Framework Risk Fault Tree

TABLE 1 Chance that Bottom Event Occurs

code	bottom event	chance
η_1	security audit	Ψ_1
η_2	risk early warning	Ψ_2
η_3	security condition monitoring	$\Psi3$
η_4	security resource management	$\Psi4$
η_5	identity and security	$\gamma5$
η_6	data security	$\gamma6$
η_7	host security	$\gamma7$

Step 5. Expression of top event obtainment. The risk formula of event B is obtained based on logical relations of bottom events in fault tree

$$B = event1 \wedge (event2 \vee event1 \vee event3) \wedge event4. \quad (16)$$

The risk formula of event C is obtained based on logical relations of bottom events in fault tree

$$C = event5 \vee event6 \vee event7. \quad (17)$$

So the risk formula of top event E is obtained

E = event1∧ (event2∨event1∨event3) ∧event4∨event5∨event6∨event7 = event1∧event4∧event5∨event6∨event7.

The simplest disjunction expression is

$$E = event1 \wedge event4 \vee event5 \vee event6 \vee event7. \quad (18)$$

So the minimal cut set of fault tree is {1,4,5,6,7}.The key bottom events of fault tree are 1,4,5,6,7. The equivalent fault tree of original fault tree is shown in figure 6.
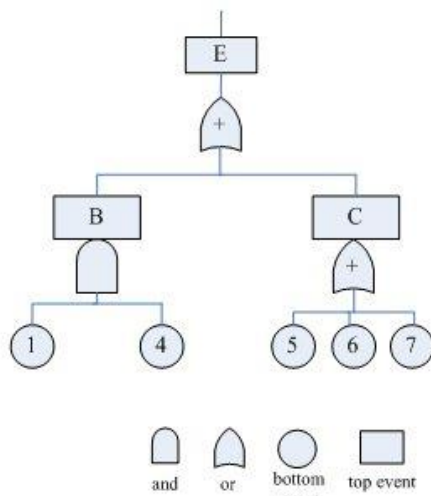
FIGURE 6 Equivalent Fault Tree

Step6. Expression of top event conversion. Because η_1, η_4 are probability variables, the risk formula of event B is obtained based on logical relations of bottom events in equivalent fault tree and two examples of chance theory

$$\Psi_B = 1 - (1 - \Psi_1)(1 - \Psi_4). \quad (19)$$

Because η_5, η_6, η_7 are uncertain variables, the risk formula of event C is obtained based on logical relations of bottom events in equivalent fault tree and two examples of chance theory

$$\gamma_C = \gamma_5 \wedge \gamma_6 \wedge \gamma_7. \quad (20)$$

So the risk formula of top event E is obtained

$$\Phi_E = (1 - (1 - \Psi_1)(1 - \Psi_4))(\gamma_5 \wedge \gamma_6 \wedge \gamma_7). \quad (21)$$

Step 7. System risk calculation. Suppose that Ψ_1, Ψ_2, Ψ_3, Ψ_4 are random exponential distribution, where $\Psi_1 = \exp(0.8)$, $\Psi_2 = \exp(0.5)$, $\Psi_3 = \exp(0.2)$, $\Psi_4 = \exp(0.4)$; γ_5, γ_6, γ_7 are uncertain normal distribution, where $\gamma_5 = N(0.1, 0.01)$, $\gamma_6 = N(0.4, 0.02)$, $\gamma_7 = N(0.5, 0.01)$. The risk chance can be obtained by hybrid simulation algorithm. The flow chart of algorithm is shown in figure 7.

Step 8. Risk management and control. The risk chance of top event 0.0132 is calculated by hybrid simulation algorithm. The risk chance of this system is small, it illustrates that risk control of this system is perfect.

5 Conclusions

This paper proposes a method that constructs and analyses fault tree based on uncertainty theory and chance theory.
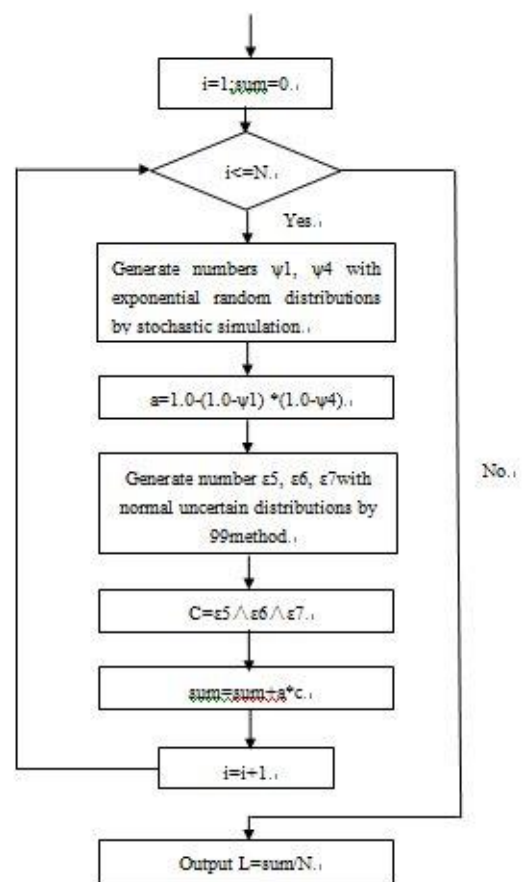
FIGURE 7 Flow Chart of Hybrid Simulation Algorithm

It also constructs hybrid simulation algorithm to calculate the chance that top event occurs. The URFTA methodology has several advantages compared to strictly deterministic FTA method and these are listed as follows.

(1) The proposed method is suitable for situations where probabilistic risk or uncertain risk exists.

(2) The methodology can be properly handled in a consistent manner in a situation with ill-defined, ambiguous information as well as probabilistic data.

(3) By using uncertainty theory and probability theory, the state of each bottom event can be described in a realistic form. The use of chance theory provides more precise descriptions and accurate solutions.

(4) This UR-FTA model is a simple and useful tool to estimate and evaluate risk in complex system.

(5) The methodology has been very versatile and flexible in applications. Therefore, it can easily be applied in other fields related risk analysis problems.

The future work of this study will focus on two directions: from the purely technical point of view, if operational rules between uncertain random variable and uncertain variable or random variable is correctly defined, the methodology will be applied to more complex system in all kinds of fields; from the practical point of view, dependent failures and common cause failures should be analysed in fault tree, the new general hybrid simulation algorithm should be developed to apply in all kinds of fault tree analysis.

Acknowledgments

References

[1] HanSuk Pan, WonYoung Yun 1997 Fault tree analysis with Fuzzy Gates *Computers ind Engng* **33**(3) 569-72

[2] Ching-torng Lin, Mao-jiun J.Wang 1997 Hybrid fault tree analysis using fuzzy sets *Reliability Engineering and system Safety* **58** 205-13

[3] Khan F I, Abbasi S A 2000 Analytical simulation and PROFAT II: a new methodology and a computer automated tool for fault tree analysis in chemical process industries *J.Hazard. Mater* **75** 1-27

[4] Hu W, Starr A G, Leung A Y T 2003 Operational fault diagnosis of manufacturing systems *J. Mater. Process Technol* **133** 108-17

[5] Li H X, Zuo M J 1999 A hybrid approach for identification of root causes and reliability improvement of a die bonding process a case study *Reliab. Eng. Syst.Saf.* **6** 43-8

[6] Sohn S D, Seong P H 2004 Quantitative evaluation of safety critical software testability based on fault tree analysis and entropy *J. Syst. Softw* **73** 351-60

[7] Roy P K, Arti B, Chitra R 2003 Quantitative risk assessment for accidental release of titanium tetrachloride in a titanium sponge production plant *J. Hazard. Mater* **102** 167-86

[8] Zadeh L A 1965 Fuzzy sets *Inform. and Control* **8** 338-53

[9] Hua Song, Hong-Yue Zhang, Chan C W 2008 Fuzzy fault tree analysis based on T-S model with application to INS/GPS navigation system *Soft Computing* **3**(9) 21-30

[10] Khan Faisal, Sadiq Rehan, Amyotte Paul, Veitch Brian 2011 Fault and Event Tree Analyses for Process Systems Risk Analysis Uncertainty Handling Formulations *Risk Analysis* **31**(1) 86-107

[11] Chanda R S, Bhattacharjee P K 1998 A reliability approach to transmission expansion planning using fuzzy fault tree model *Electric Power System Research* **45** 101-8

[12] Dong Yuhua, Yu Datao 2005 Estimation of failure probability of oil and gas transmission pipelines by fuzzy fault tree analysis *Journal of Loss Prevention in the Process Industries* **18** 83-8

[13] Ayhan mentes, Ismail H.Helvacioglu 2011 An application of fuzzy fault tree analysis for spread mooring systems *Ocean Engineering* **38** 285-94

[14] Liu B *Uncertainty Theory* 4th ed. http://orsc.edu.cn/liu/ut.pdf

[15] Liu B 2007 *Uncertainty Theory* 2nd ed. Springer-Verlag,Berlin

[16] Liu B 2009 Some research problems in uncertainty theory *Journal of Uncertain Systems* **3**(1) 3-10

[17] Liu B 2009 *Theory and Practice of Uncertain Programming* 2nd ed., Springer-Verlag, Berlin

[18] Liu B 2010 U*ncertainty Theory: A Branch of Mathematics for Modeling Human Uncertainty* Springer-Verlag, Berlin

[19] Liu B 2010 Uncertain risk analysis and uncertain reliability analysis *Journal of Uncertain Systems* **4**(3) 163-70

[20] Liu B 2012 Why is there a need for uncertainty theory? *Journal of Uncertain Systems* **6**(1) 3-10

[21] Liu B 2001 Fuzzy random chance-constrained programming *IEEE Transactions on Fuzzy Systems* **9**(5) 713-20

[22] Liu B 2001 Fuzzy random dependent-chance programming *IEEE Transactions on Fuzzy Systems* **9**(5) 721-6

[23] Liu J J 2011 Uncertain comprehensive evaluation method *Journal of Information Computational Science* **8**(2) 336-44

[24] Rong L X 2011 Two new uncertainty programming models of inventory with uncertain costs *Journal of Information &Computational Science* **8**(2) 280-8

[25] Yan L M 2009 Optimal portfolio selection models with uncertain returns *Modern Applied Science* **3**(8) 76-81

[26] Meilin Wen, Rui Kang Reliability Analysis in Uncertain Random System http://orsc.edu.cn/online/120419.pdf

[27] Greg B, Padma M, Dennis Q, et al. *Cloud Computing [EB/OL]* http://www.Cloud computing-china.cn

[28] Wang P 2010 *The key technology of cloud computing and application* Beijing: Posts and Telecom Press

[29] Chen K, Zheng W M 2009 Cloud computing: System instances and current research *Journal of Software* **20**(5) 1337−48

[30] John R, James R 2009 *Cloud Computing: Implementation, Management, and Security*

[31] Xia T Z, Li Z 2009 Research on Cloud Computing Based on Deep Analysis to Typical Platforms *cloudcom 2009,beijing, China* 601-8

[32] ITU http://www.itu.int/en/pages/default.aspx

[33] Azanza M P V 2006 HACCP certi_cation of food services in Philippine inter-island passenger vessel *Food Control* **17** 93-101

[34] Goldbach S G, Alban L 2006 A costCbene_t analysis of Salmonella control strategies in Danish pork *production Preventive Veterinary Medicine* **77**(1) 1-14

[35] Fraser R, Souza D Monteiro 2009 A conceptual framework for evaluating the most cost-e_ective intervention along the supply chain to improve food safety *Food Policy* **34** 477-81

[36] Okezie I, Aruoma 2006 The impact of food regulation on the food supply chain *Toxicology* **221** 119-27

Authors

Guo Changyou, born in 1976

Current position, grades: PhD candidate at the School of Computer and Communication Engineering, University of Science and Technology, Beijing.
Scientific interests: Access Control, Network Security, Information Security and Cloud Computing Security.

Zheng Xuefeng, born in 1951

Current position, grades: professor and doctoral supervisor in the School of Computer and Communication Engineering, University of Science and Technology Beijing.
His research interest: Computer Control Systems Development, Computer System Security Analysis, Network Security, Information Security and Distributed Systems Security. He is the senior member of the computer society.

Liu Jianjun, born in 1976

Current position, grades: PhD candidate at the School of Computer and Communication Engineering, University of Science and Technology, Beijing
Scientific interests: Access Control, Network Security, Information Security and Cloud Computing Security.

Design and analysis of hotel management system based on information technology

Guirong Guo*, Yan Lu

Qingdao Hotel Management College, Qingdao, Shandong, 266100, China

Abstract

With the rapid development of China's economic and tourism, competition in the hotel industry has become intense. Only by improving the means and methods of their management and improve their level of service constantly, can they receive adequate and healthy development. Therefore, the hotel operation computer management has become a priority task. These articles give a brief introduction to hotel management system design, in accordance with means of software engineering to do feasibility analysis, requirements analysis and design. The whole system divided into modules separately and introduce the function, while gives a set of criteria and the logical structure of the database of database management. These analyses and the guidelines basically meet the design requirements of hotel management system based on information technology, so as to improve the service quality and efficiency of the management of the hotel.

Keywords: hotel management, feasibility analysis, requirements analysis, database, logical structure

1 Introduction

While global economic integration make the hotel industry market wider, it also make the hotel industry facing increasingly fierce competition and rising customer expectations, forcing the insider to expand their customers constantly, improve service quality, lower administrative costs and improve customer satisfaction with new ways to enhance the hotel's core competence. Which is one of the effective means of mass application of advanced information technology, especially with the constant deepening of information, change hospitality industry competition and management models which in the traditional sense, in order to win new competitive advantages. Leading hotels in the world have been falling over themselves to explore, implement and promote integrated hotel management information. Modern hotel is the integration of consumer sites including rooms, catering facilities, business culture, and a variety of other services. Due to a huge membership and service projects, so naturally, it's the amount of carried in its management, in order to improve work efficiency, reduce operating costs, improve service quality and management levels, promote economic efficiency, we must use the computers to progress modern information management. The great development of modern computer technology, provide a favourable opportunity for hotel management to change.

2 Software architecture technology

Hotel management system have great development both at home and abroad, there are more sophisticated commercial systems, and has good application in medium to large level hotels. With the popularity and development of Internet, hotel management software based on b/s structure must occur. With the further development and enhancement of domestic software, internal software replace imported software to become the mainstream management software in high-star hotel will become inevitable. Much content in commercial systems do not suitable for hotel, and one-time investment for these systems is high. Therefore, developed an applicability hotel management software is extremely meaningful; this also is the focus of this article.

Software architecture technology based on b/s, system installation, modification, and maintenance can be solved on the server-side, which brings convenience to the hotel management, also save the cost of the hotel's technology and maintenance of the system. When an administrator uses the system, only requiring a browser could run all modules to reach "zero clients" functionality, it is easy to upgrade automatically at run time [1].

3 Analysis on management information system of hotel

3.1 FEASIBILITY ANALYSIS

The analysis of structure of hotels management information systems mainly from the use of feasibility, technical and economic aspects [3].

(1) Feasibility of using

At present, the computer-aided automation, convenient and efficient system of hotel management are

---
Corresponding author e-mail: ggrluyan@163.com

only used by some of the star hotel management, and each hotel has different needs, but most of the existing management system's definition of a project cannot be applied to general hotel management, so identifying these specific needs is the content should be made clear first before build a management information system.

Hotel management information system using feasibility analysis mainly focused on the operator's actual situation and environmental aspects. Operation situation analysis focuses on the hotel specific section's attitudes and awareness toward management information system specific functions, such as its system or part of systems capabilities considered not needed or are not willing to use, system construction and use will appear a series of problems. Therefore, when carrying out feasibility analysis, first focused on specific sectors, particularly mission critical jobs such as hotel General Manager, Department head's awareness and orientation on the system. In addition, when system is using feasibility analysis, it should consider the environmental conditions, particularly special circumstances in business activities, such as quickly to make the guest registration and organization when receive group guest.

(2)Technical feasibility

Management information system for technical feasibility involves information processing technology (such as information processing, input / output, and response time), reliability, scalability, security, maintainability, and environmental adaptability and so on many aspects, mainly analysis whether technical conditions can smoothly carry out development work, can meet the needs of developers of software and hardware, and so on.

B/s mode may be used for development of this system, combining hItem Intranet technologies, in line with the trend of technology development. Database server using MySQL database, it is able to handle large amounts of data, while maintaining the integrity of the data and provides many advanced management capabilities. Its flexibility, security, and easy of use provides good conditions for database programming. Therefore, systems' software platforms are already mature and feasible, it is technically feasible.

(3)Economic feasibility

Economic feasibility analysis mainly analysis and assessment the construction project of the funding needs and capital efficiency, including budget estimates, the input-output ratio, payback period of investment, investment benefit analysis, and so on. The development and maintenance of this hotel management information system requires hotel put into expense, but as an important aid to the hotel management, after it put into use, it can dramatically improve the department efficiency, its business volume, improved quality of service will far outweigh the upfront investment. Therefore, hotel's selection of management information systems has high investment returns and strong economic feasibility.

3.2 REQUIREMENTS ANALYSIS

Requirements analysis is the analysis definition process of requirements, is the start of the planning and development period of the project. Needs analysis task is to thoroughly describe the function and performance of the software, identify limits of software design and software interface details with other elements of the system, defining the effectiveness of software requirements. Requirements analysis includes business requirements, functional requirements and development requirements [4].

3.2.1 Business requirements analysis

(1) Room booking capability can handle customer bookings by various means, such as phone book, online bookings, and reservations at the front desk.

(2) Rooms of the hotel have different grades, and require the system to be able to categorize the room management, and according to book different types of rooms to offer available prices, booking discount timing function settings such as lowest price, easy to fit individual traveller and group reservations.

(3) Check function

(4) Keep abreast of the current state of all whether it is available, such as clean the room, which was checked out, to track whether the guest stay. Based on client occupancy information, you need to provide a guest information management function, in order to achieve unified management to the guests of the hotel information, such as ID number for customer inquiries, queries according to the checking dates, change the guest contact information and so on.

(5) A hotel management system should have recognition module in order to distinguish each business (accommodation, registration, cashier, hospitality, and other services) to which staff action is, that means the system should set login module.

3.2.2 Functional requirements analysis

Hotel management system consists of both background and foreground parts, front office is responsible for booking, reception and cashier services, the background used for administrators to manage systems, such as setting room type, room settings, operator settings, financial management and warehouse management.

Cash register function requirements. An integrated cash register system should have the following functional requirements: cash register charge of project settings, various payment methods (cash, accounting and business summary report).

Room settings requirements: Room is the hotel's main product; room management plays an important role in the hotel business. The most important tasks in room management is to modify the status of rooms, provided the information whether room is free, for rent, clean

rooms, dirty rooms and other information, so that bring convenience to receptionist assigned room reservation. Room management demands specific functions, including modified status of rooms, room management, post management, room supplies supplements article management, room history search, to check availability with the front desk, room maintenance, housekeeping, the internal administration of print-related housekeeping report.

Financial management requirements: Finance is the core of hotel Admin, it has the following functional requirements: set up account, initial processing, bookkeeping, certificate, voucher budget, other management review post, enter detailed ledger, inquire function, processing capabilities, a variety of reports and books of output.

Warehouse management requirements: Warehouse management functions including construction management, changes in the new warehouse, adjusting prices, to deal the daily account, including records and statistics code, name, unit price, price and so on of inbound and outbound items every day or every period. Report processing, includes articles, of inbound and outbound inventory processing, goods schedules, statements processing and so on.

Operator set up module: this module is used to manage the hotel operator information, add new operator, edit existing operator information, and delete operator information functions.
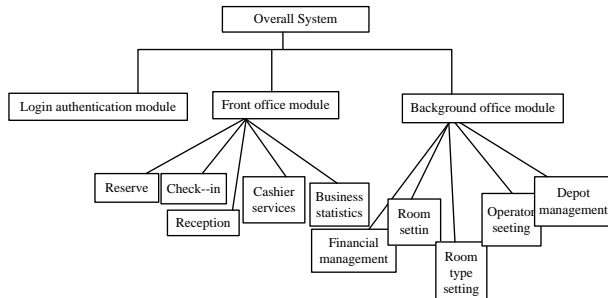
FIGURE 1 Systems function diagram

3.2.3 Development requirements analysis

Regardless of hotel scale, its future direction will be to network and information development, its main business needs mainly show in the network business process requirements and hotel information process requirements [5]. It mainly reflects in: remote query processing capabilities with the business requirements, cooperation with other relevant units of demand, as with other businesses such as hotels, travel agencies, the Ministry of public security, hotel information needs, such as electronic door locks, program-controlled exchange equipment, magnetic card consumer and other business process requirements.

4 Hotel management system design

4.1 PRINCIPLE OF SYSTEM DESIGN

Hotel management functional requirements determine the hotel management system, designed to not only take into account of conforming to the theoretical system of management information systems, but also consider the hotel industry's high level specificity required for reliability, real time management, health and safety and so on [6]. Management information system of hotel consists of front office systems and back-office system, systems are designed to do nothing but follow the general principles of design of management information system, and according to the particularity of the hotel management should also follow certain design principles.

4.1.1 Front office system design principle

(1) Concurrency considerations. In front Office system application, multiple users simultaneous access to a library that is very frequent thing, if not handled well, can cause a system deadlock, data loss, recovery difficult, seriously, it will cause incalculable consequences for the hotel business. Therefore, during system design, set a process accessing shared data or critical section, using PV operation to control the process, simultaneously in order to reduce the share of conflict, enhancing system security, and will often require exclusive use of the database into multiple, such as log file system is divided into 10: log 0~ log 9.

(2) A good user interfaces and eases of operation. Computer user request the system to complete a function through the user interface, also respond to user requests or reports through the user interface implementation. Therefore, the user interface is the embodiment of system functions in front of the user, which is to provide users with the means to use the system. Friendly interface will have: an attractive format, compatible with the screen user operation, the operation is simple and easy to learn , users can get help at any time, it mainly express through the window, menu design, screen design and so on.

(3) Safety, reliable design. With the development of networks, and in recent years many companies has exposed to disclose customer information, this also became problems to be faced by the hotel management information system. In order to ensure the safety and reliability of the data correct and effective, the system must provide a set of data protection subsystem, including security, reliability, integrity, and concurrency control and so on.

(4) Suitability. The system should develop toward the application and the future, therefore be able to adapt to different hotels in different management models, different settings, different product types, but also must be able to adapt to related equipment and system. To meet these requirements, you should open parameters as much as possible in the design of system, and when connect with

other associated equipment, facilities and services, stay well in advance as far as possible, and be able to adapt to the new trends in technological development, easy version of hardware replacement and software upgrades.

4.1.2 Back Office Design Principle

Hotel back office system design follow the general principles of software design, whose design is based on financial transaction, involves accounting (financial transaction, fixed assets, payroll, material warehouses, cash bank, procurement, etc.), cost analysis (financial analysis, cost analysis, profitability analysis, and so on), decision support, and so on.

(1) Principle based on financial accounting. Financial accounting managed by financial transactional, during the design process, it follows the following several points:

1. Uniformity and standardization. Bookkeeping voucher should use unified fill specification, for easy entry and processing, such as receipt, voucher payments and transfers in unified format. Accounting summary written to be standardized as far as possible, best use fixed lines.

2. The report generation function is the key to success of the software package.

3. Financial management system should have mandatory control functions.

(2) Principle to regard support decision-making, financial planning forecast as the guiding. A financial plan forecasts subsystem consists of the following functional modules:

1. Financial planning forecasting module function as financial forecasting and preparation of the medium-term funding programs.

2. Cost planning module can achieve its cost estimates and other functions.

3. Project budgeting module. According to the existing standards and prices, calculate the costs of projects and preparation of the project budget.

4. Investment projects feasibility analysis module is applied for feasibility analysis for investment, joint venture projects.

4.2 PRINCIPLE OF DATABASE DESIGN

Hotel management information's main task is to manage the needs of information obtained through large amounts of data, which must store and manage large amounts of data. Therefore, establish a good data structures and databases can make the whole system be quickly, easily, and accurately to manage the data you need is one of the main indicators to measure a hotel management information system development work well or bad. A good data structure and database should fully reflect the evolving situation, fully meet regulatory requirements at all levels of the Organization, should also make subsequent systems development work convenient, fast, overhead is small and ease of maintenance [7]. In order to

better organize your data and designed to meet the actual needs of the database, the system in the design mainly follow the following principles of database design:

(1) Data consistency and integrity guidelines

In order to ensure the consistency and integrity of the database, reduce data redundancy. Associations between tables is a coercive measure, once established, to the parent table and the child table's insert, update, and delete operations are take up overhead, so generally we do not identify the properties in the database design field as the primary key connect with the child table.

(2) Regulatory guidelines

When design and operation and maintenance the database, to ensure that data distributed properly across the tables in your database. Using the correct data structure and corresponding access to the database is not only easy to operate, and can greatly simplify the other contents of the application (queries, forms, reports, code, and so on), the correct table designs means that implement database normalization.

4.3 THE LOGICAL DATABASE STRUCTURE DESIGN

Considering the concepts required building the hotel management system, and does logical structure design for it [8].

Customer (name, card type, ID card number, gender, telephone number, contact address, customer registration, record-keepers, notes).

Levels of rooms (number, room level, housing prices, basic settings).

Room (room number, room level, room status, room size, position).

Reservation form (expected order number, customer ID, customer name, room level, room reservation, booking status, check-in method, an extra bed, extra bed price, deposit, expected arrival time, the expected departure time, recording time, record-keepers, note).

Stay list (check number, customer ID, customer name, room level, number of rooms, occupancy status, check-in methods, extra bed, extra bed price, deposit, arrive time, expected departure time, whether breakfast, whether wake up, whether staying secret, recording time, record-keepers, note).

Stay list's history (number, stay list number, customer ID, customer name, room level, number of rooms, occupancy status, whether an extra bed, extra bed price, deposit, expected arrival time, the expected departure time, recording time, record-keepers, note).

Changes of business list (change order number, change business customer ID, customer name, room level, number of rooms, occupancy status, whether an extra bed, extra bed price, deposit, expected arrival time, the expected departure time, recording time, record-keepers, note).

Consumption list (AutoNumber, check-in list number, customer name, consumption content, spending time, amount, note).

Consumption list (AutoNumber, check-in list number, customer name, consumption content, spending time, amount, note);

Cashier's check-list (check-out number, customer ID, customer name, room level, price, number of rooms, occupancy status, stay methods, whether an extra bed, extra bed price, deposits, accounts receivable, Actual balance, departure time, recording time, record-keepers, note).

4.4 DATABASE ACCESS TECHNOLOGY

This system can consolidate uses Java JDBC technology achieve connection of Java with MySQL. DBC (Java DataBase Connectivity Java database connectivity) is a Java API for executing SQL statements, and can provide unified access to multiple relational databases, which consists of a set of classes and interfaces written in the Java language. JDBC provides a standard API for database developers, can build more advanced tools and interfaces according to it that enables database developers to use pure Java API to write database applications.

JDBC features include a connection to the database, send a statement of the operational database and process the results. Java has a sturdy, safe, easy to use, easy to understand and can be automatically downloaded from the network properties, it is an outstanding language in writing database applications. All you need is Java application and the method of dialogue between different databases and JDBC is a mechanism for this purpose. JDBC extends the functionality of Java [9]. For example, using Java and the JDBC API can publish a Web page containing the aPPlet, the aPPlet using information from remote databases. Enterprises can also use JDBC via hitranet linked all staff to one or more internal database (even though these computers used by staff in a variety of operating systems such as Windows, Macintosh, UNIX, etc.).

T JDBC API supports two-layer model (c/s) for database access, and also supports a three-tier model (B/S). In a two-layer model, the Java aPPlet or application will have a direct dialogue with the database. This will require a JDBC driver to communicate with specific database management system. Users' SQL statement is sent to the database, and the results will be sent back to the user. Databases can be located on a different computer, users connecting over the network to the database. This is called a client/server configuration, in which users' computers is clients, provides database's computer is a server. In a three-tier model, the command is sent to the service's "middle-tier", and then it sends SQL statements to the database. Database process SQL statements and the results are sent back to the middle tier, middle tier and then sent the results back to the user [10].

5 Conclusion

At present, information management technology has been pushed at frontier of time by global information tide, thus the information management system becomes hot issue that all the society concerned about. With the development of science and technology, computer science become matures; they play an increasingly important role in all areas of society. People can give a lot of tedious work to computer systems to process, thus improve work efficiency.

Hotel management system in this paper is to analyse and design on the basis of the current development of the computer technology. Through analysis of the feasibility of the system, and a variety of needs analysis, modularize the system, analyse and give simple introduction on function of every module, then given the logical structure of the database design of the management system, provides a theoretical basis for the realization of the management system. This management system overcome shortcomings of modern operating system that computation are not strict and systematic, such as heavy workload, data transmission is not timely, error-prone management statistics and so on; enhance the competitiveness of the hotels.

References

[1] Ham S W, Kim G, Jeong S 2005 Effect of information technology on performance in upscale hotels *Hospitality Management* (24) 281-94
[2] Jing W 2010 Hotel Management System Design *Government Office* **174** 23-7
[3] Bing G 2010 Hotel Management Information System Based on B/S *Dalian: Maritime Affairs University Of Dalian*
[4] Joseph S V, Joey F G, Jefferey A H 2005 Essentials of System Analysis & Design *Tsinghua University Press* 378-402
[5] Yang Z 2009 Hotel Industry E-commerce System Planning Based on B/S *Modern Economic Information* (8) 208-9

[6] Xiaolei Y 2012 Design and Implementation of Hotel Management System *Jilin: Jilin University*
[7] O' Connor P, Murphy J 2005 Research on Information Technology in the Hospitality Industry Hospitality Management (24) 281-94
[8] Xin L 2011 Design and Implementation of Hotel Management System Base on B/S *Jilin: Jilin University*
[9] Lei X, Zhongwei L 2008 Java Tutorial *Beijing: People's Telecon Publishing House*
[10] Jiajing Z 2008 My SQL and E-commerce Practice *Beijing: People's Telecon Publishing House*

Authors	
	Guirong Guo, born in 1982, Shandong Province of China **Current position, grades:** lecturer **University studies:** Master degree candidate was earned in Shandong University. **Scientific interest:** hotel management, folk tourism
	Yan Lu, born in 1982, Shandong Province of China **Current position, grades:** lecturer **University studies:** Doctor graduate degree was earned in Shandong Normal University. **Scientific interest:** population, resources and environment economics

Operation Research and Decision Making

GIS technology integration design based on university culture resource

Jiahua Zhou*

Suzhou University School of Management Engineering, Suzhou, Anhui, 234000, China

Received 15 June, 2014, www.cmnt.lv

Abstract

3-D GIS technology has become the hot research in current GIS realm. It is applied in various industries. This paper constructed open and interactive 3-D data campus based on Sketch Up applying ArsGIS 3-D space platform and combining traditional information management system and 3-D GIS technology. We integrated various system resource management modules in campus and fulfilled an open campus project, which has strong inquiry and space analysis function. We turned 3-D model data into ArcGIS supportable format and imparted it into digital campus development structure. Browsing system of multi-detail and multi-level scene is also applied. Finally, we constructed a lively 3-D digital campus culture design that has interaction form.

Keywords: GIS technology, open type, interaction type, mutual integration

1 Introduction

Digital campus of GIS technology is generally modelled as real campus [6]. It is constructed with the principle of computer network technology, visualization technology and information integration technology. It is not limited by time and space and can collect and manage various culture resource at any time in LAN for efficient deploy and powerful control. This kind of integration design is good to the improvement of informatization scientific research system of campus, implement and perfect of various measurement of school and planning and decision of campus scientization. Therefore, integration of university culture resource is needed an effective management system to allocate resource.

Foreign university pay high attention on campus informatization development construction. Switzerland allocated 30 million Swiss franc approved by both houses of the Swiss confederation to the state Swiss University to encourage campus informatization construction and construct information technology website of higher education [1]. America is always carrying out policy of rapid development of visual information technology after formulating National Information Infrastructure. In recent years, educational management system of America has witnessed great changes. Many building facilities, office affairs and graphic technology centre in campus are all informationized. Resource is integrated in it, which achieved objective of high efficiency [2]. Compared to abroad, digital campus design of our country still has gap. However, some excellent digital campus platform is constructed with the constant emphasis from Ministry of Education. Chang Hong et al proposed in Goal, Content and Strategy of Digital Campus Construction under the

Macro Digital Background [3] that education informatization pulling the development of modern education has become the strategy of current education growth. They also thought that construction of digital campus construction is the construction goal, content and strategy under macro digital age background. Ma Tian et al [4] realized a set of 3-D visual campus system based on 3DMax and OSG in 3-D Visual Campus System Design Realization of Lintong Campus of Xi'an Science & Technology University. They proposed independent roaming collision detection method to realize automatic positioning by 3-D modelling of the whole campus. Liu Xinyu [5] took GIS as platform on overall design, content construction design, system function design and interfacial design. He studied GIS and map service based on web and realized campus map service prototype in campus GIS map service system design of Kunming Science & Engineering University.

2 ArcGIS introduction

ArcGIS is the most advanced GIS software developed by a series of perfect progress integrating multiaspect computer mainstream technology by American Environment System Research Institute (ESRI). ArcGIs is widely applied in world GIS realm [11].

2.1 INTRODUCTION OF ARCGIS SYSTEM

There is a set of complete GIS application external member in GIS frame covered in ArcGIS. Development group can provide GIS flushbonading application technology. In addition, there are also ArcGIS Server,

---

Corresponding author e-mail: zjhzhoujh@163.com

ArcIMS, ArcGIS Engine and ArcGIS Desktop applied in enterprise.

2.2 ArcGIS ENGINE

Flushbonading GIS can be flexibly used to construct new application procedure in GIS procedure developing process. ArcGIS Engine can also be used to extend specific application procedure of other industries and derived to user. Development tool is shown in Table 1.

TABLE 1 Overview of ArcGIS Engine development tool

ArcGIS Engine development tool		overview
kit	ArcGIS Engine Developer Kit	Provide a series of tools and resources used for comprehensively developing customization application procedure for developer
	ArcGIS Engine Runtime	Provide core component and extended module needed for operating ArcGIS Engine for users
toolbar	Comprehensive and dynamic GIS edit tool	Tools such as translation, rotation, zoom, browsing, analysis and selection are dragged and dropped into customized application to realize interaction function of system.
Object category gallery model	ArcGIS Engine component gallery	Basic service, data access, map expression, development component and extended function
3-D extended controls	SceneControl	Used for 3-D scene and 3-D analysis display with small data size and can realize visualization mutual scene and 3-D surface analysis on GIS data
	GlobeControl	Can realize browsing and query on large space topography, 3-D grid and huge amount of vector data

3 3-D digital campus overall design based on GIS technology

Design philosophy of this paper is based on data model of university campus geographic space. We use GIS technology to develop 3-D digital campus by space management and information analysis ability to realize integration of space information cultural resource of university campus.

3.1 DESIGN THOUGHT AND TECHNOLOGY ROUTE OF DIGITAL CAMPUS

First, we need to install SketchUp ESRI control [10] to add SketchUp toolbar into relative procedure of ArcGIS to realize a real visualization 3-D digital campus model in construction of digital campus model, as shown in Figure.1. We constructed a technology route through data preparation, data processing, 3-D modelling and system development at earlier stage, as shown in Figure 2.
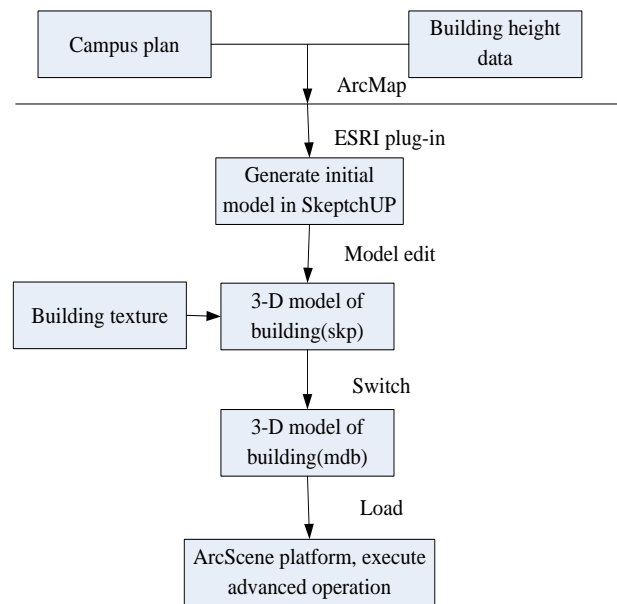
FIGURE 1 Procedure diagram of 3-D campus construction
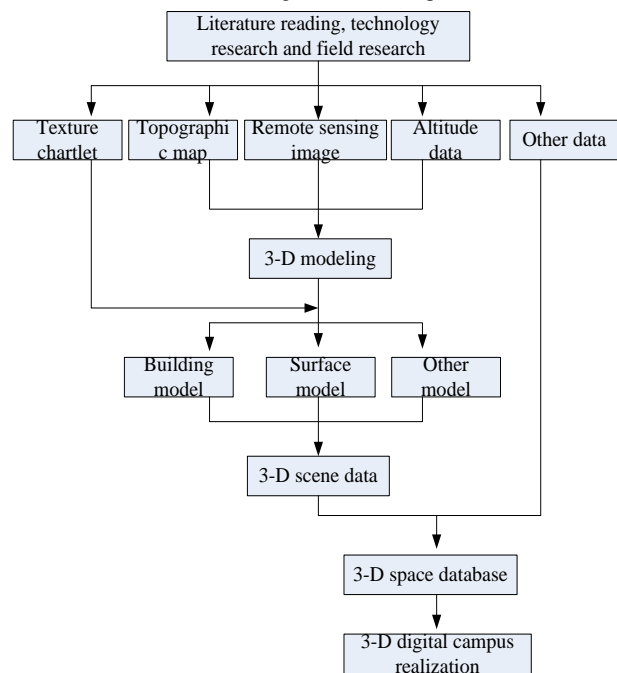
FIGURE 2 Technology route of 3-D digital campus construction

3.2 DIGITAL CAMPUS SYSTEM FUNCTION DESIGN

3-D digital campus design first need 3-D scene modelling including topography with shrink actual ratio, building, playground, road, river, etc. Second, we should construct a 3-D roaming system for realizing roaming browsing and flying browsing in the process of 3-D campus browsing. At last, we should realize information query and information analysis function so that users can do interaction functional operation.

3-D digital campus system can be divided into four system module, that is, scene browsing module,

information query module, space analysis module and measurement statistics module, as shown in Figure 3.
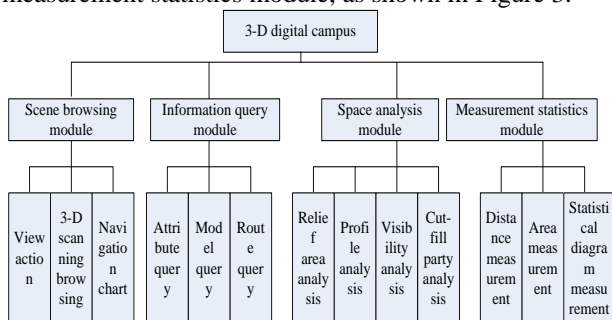
FIGURE 3 System module function diagram

Scene browsing module includes view action. It is used for browsing 3-D scene by perspective switching and can be moved and zoomed for omnibearing observation. 3-D browsing, that is, some function of route browsing, combining with navigation chart has navigation function for the moving direction and position of users. Information query module includes attribute query. It is one of 3-D campus basic function and can be used for effective searching of building, college information, hot event, news, etc. Model query can rapidly found relative building, road, river, etc. Route query can provide the shortest distance and walking time estimation between two places. Module function in space analysis module can make the whole system possess enough advantage on space analysis applying space analysis ability in ArcGIS technology. Users can develop various space analysis functions by tools on control board. Distance, area and statistical figure in measurement statistical module are stored in form of database. If users have need then it can conduct modification order.

3.3 DATABASE STRUCTURE DESIGN

Development of digital campus system constructed by GIS technology need various types of data. Attribute data and space data are two main data types. Attribute data includes information of teacher and student, course notice information and building attribute information. They are stored according to types for flexible processing ability of system, which can combine campus entity object and data of database. We can achieve goal of resource integration by using ArcGIS technology for space data controlling and database software for attribute data controlling [9]. Database structure figure is shown in Figure 4.
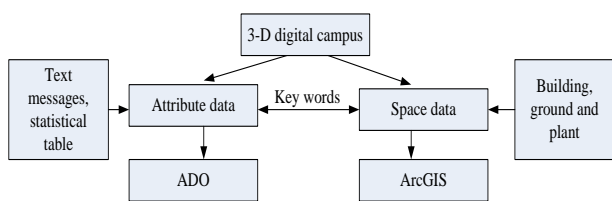
FIGURE 4 Database structure figures

Geodatabase is to integrate and describe data that needed to be processed according to certain model framework and GIS technology based on DBMS [8]. This kind of thought and behaviour can more directly present the real world. Data storage characteristics of GIS technology are hierarchical storage. Geodatabase applies hierarchical mode to use space data and attribute data in specific data storage. Structural representation of geospatial data model on 3-D model file is shown in Figure 5.
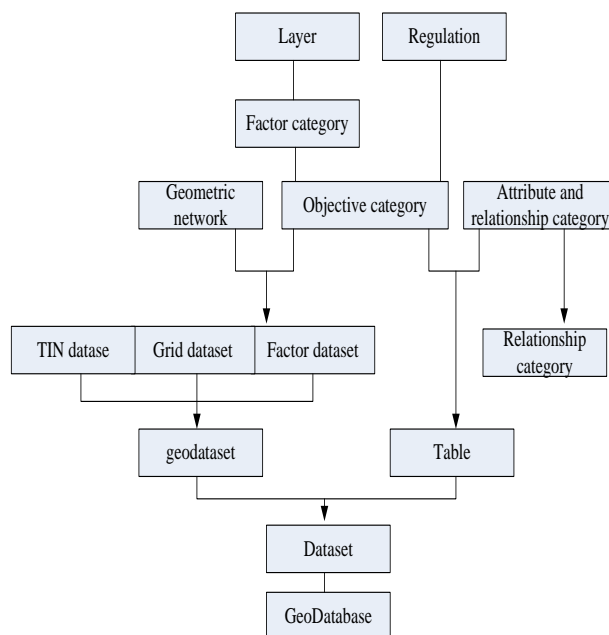
FIGURE 5 Sketch map of Geodatabase organization structure

System database integrates and manages lots of space data information and attribute data information adopting practical multiline relationship distribution database system SQLServere2005 and stable GIS space data engine ArcSDE.

3.4 SPACE AND ATTRIBUTE DATABASE DESIGN

According to the high-resolution remote sensing image [7] and campus cultural environment and geographical environment, space data and attribute data are conducted hierarchical management based on space feature. It includes environment layer such as plant, river, road, land, etc., building layer such as Ball Park, teaching building, dormitory, street, infrastructure, etc., label layer such as name, direction, tendency, distance, etc., and information layer such as college information, teacher information, class information, student information, news notice, etc. We should set up attribute table and add relative data into SQL Server. Table 2 is excerpts of attribute table of database. Of course, we can apply space analysis means of ArcGIS technology to process and distribute relative message segment.

TABLE 2 Part information of attribute database

Classification of information table	Field name	Field type	Field length	Remark
Student	Sno	Int	10	Major key
	Sname	Char	10	
	Ssex	Char	2	
	Sage	Int	20	
Department	Dno	Int	10	Foreign key
Class	Cno	Int	10	Foreign key
Building	Bno	Int	10	Major key
	Bname	Char	10	
	Bnum	Int	50	
	Barea	Int	10000	
	Bdate	date	10	

4 Conclusion

This system developed and designed 3-D digital campus with the help of space data analysis and information resource analysis of GIS technology in purpose of integration of campus human geography resource. It aimed to show real and vivid campus culture, plan out scientific and rapid campus cultural resource framework and promote the stable development of university campus. This paper comprehensively studied overall process of digital campus construction in view of system design thought, development route of technical route, functional module design, system structural database design, space database design and attribute database design. It mainly discussed realization of interaction pattern in 3-D digital campus interaction pattern with SketchUp software technology and ArcGIS software technology and creatively proposed hierarchical management form of applying space database and attribute database, which provides a solid foundation for integration of university cultural information resource.

References

[1] Fabio D F, Giovanni A, Arcangelo D 2009 Configurable Design of Multimodal Non Visual Interfaces for 3D VE's *Springer-Verlag Berlin Heidelberg* 71-80
[2] Kevin S, Eric D K, Winston W et al 2006 Spatial targeting using queries in a 3-D GIS environment with application to mineral exploration. *Computers & Geosciences* (32) 396-418
[3] Changhong Y, Yunwu W 2013 Goal, content and strategy of digital campus construction under the macro digital background *China Electrochemical education* **10** 30-5
[4] Tian M, Guan W 2014 3-D visual campus system design realization of Lintong campus of Xi'an Science & Technology University *J. of Xi'an Science and technology University* **34**(1) 67-70
[5] Xinyu L 2013 Design and realization of campus GIS map service system *Yunnan: Kunming Science & Engineering University*
[6] Hui L, Jianzhen S 2011 Research on digital campus construction based on GIS technology *Science and Technology Education Innovation* (13) 137-42

[7] Zhengxin G, Zulu Z, Lu Z 2008 3-D visualization campus geographical information system design and realization based on GE&GIS technology *Resource Development and Market* **24**(11) 961-4
[8] Tieying L, Yu S, Xiaofan B, Yaolong Q 2006 Research on digital campus model based on component GIS technology *J. of Hebei University* **26**(2) 213-8
[9] Rong S, Fanli M, Wei Z 2012 Discussion on integrating operation strategy of college teaching resource in the environment of digital campus *Modern Education Technology* **22**(5) 46-51
[10] Lin P 2012 Simple analysis on application of SketchUp software in Garden planing and design course of vocational college *Modern Horticulture* **6** 195
[11] Yibo D 2009 University visual campus construction based on ArcGIS *J. of Xuzhou Engineering College (Natural Science Edition)* **24**(3) 46-9

Authors

Zhou Jiahua, born in 1979, Anhui Province of China

Current position, grades: associate professor
University studies: Bachelor degree was earned in major of history, Anhui University in 2000; master degree was earned in major of Historical theory and Historiography, the School of History and Culture, Shandong University in 2008
Scientific interest: University culture and regional economic and social development

Operation Research and Decision Making

An analysis model of urban water supply quality based on extension classification method

Yun Bai[1, 2], Pu Wang[1, 2*], Jingjing Xie[3], Chuan Li[4]

[1]*Key Laboratory of Three Gorges Reservoir Region's Eco-Environment, Ministry of Education, Chongqing University, Chongqing 400045, China*

[2]*National Centre for International Research of Low-carbon and Green Buildings, Chongqing University, Chongqing 400045, China*

[3]*Testing Center for Science and Technology, Chongqing Academy of Science and Technology, Chongqing 401123, China*

[4]*Engineering Laboratory for Detection, Control and Integrated Systems, Chongqing Technology and Business University, Chongqing 400067, China*

Abstract

Directing at the multiple factors and levels in the complex problem of water consumption, this paper studied and analysed the quality of urban water supply and put forward an analysis model of urban water supply quality based on Extension classification method. Based on the Extension Theory, it modelled relative factors of analysis of urban water supply quality by matter-element method. Than classical domain and joint domain of relative factors of analysis of urban water supply quality were established. Via extension correlation function, the mapping relation between the water supply condition of the urban as research subject and the existing grades of urban water supply condition was established. Based on extension correlation, the grade of water supply condition of the present urban could be obtained. Thus, this grade can offer guidance for water supply of the urban and support for water supply strategy. Finally, by a case study, the model and algorithm were tested and proved feasible and operable.

Keywords: urban water supply, water consumption, extension theory, extension classification, model

1 Introduction

Water supply is important for city life, industrial production, city greening, etc., and is a basic factor in maintaining the normal operation of the urban. Thus, in order to maintain the security and reliability, the current water supply condition of the city is needed to be analysed for the study of dynamic regulation in urban water consumption. It will offer support for the prediction of the future development of urban water consumption and the accurate prediction of water consumption can provide data support or the optimal operation and management of urban water supply system [1, 2]. However, urban water consumption is a complex system of multiple factors and levels, and is related to urban population, industrial development, living standard of the people, energy-saving technique improvement, etc. In the system, there are both given information and unknown or unascertained information, so the analysis of the grade of water supply is a decision-making process with uncertain information [3-5]. In recent years, the imbalance between supply and demand in the water supply system in many large and medium-size cities has been becoming prominent. And urban water shortage has become a common phenomenon. Thus, the analysis and prediction of the quality condition of urban water supply is very important. By far, there have been some studies focusing on condition analysis and prediction of urban water supply quality from different aspects and with different analysis method [6-9].

However, there are some limitations in the present condition analysis and the prediction model and method in the processing with uncertain information, and cannot process the positional relation in uncertain information effectively. In this way, the reliability of the analysis and prediction is not high enough. For that, based on the existing studies, this study analysed the grade of quality condition of urban water supply based on extension classification. By applying matter-element model, relative decision-making factors of decision of grade of urban water supply quality conditions were modelled in formal method, and classical domain and joint domain of different grades of quality conditions of urban water supply were established. Then based on extension correlation function, the grade of quality condition of water supply of the current urban was obtained. Finally, the reliability and operability of the model and algorithm were checked via a case study.

2 Basic concept of extenics

2.1 BASIC CONCEPT

Extenics is an intelligent design discipline that focuses on the designation of possibility and feasibility of extension of objects, and studies the general rules and methods of innovative extension designation. It describes design objects with formalized and modelled knowledge modelling. Its logical cell is matter-element. And by

combining qualitative analysis and quantitative analysis, it considers comprehensively the quantitative and qualitative changes of the objects, and establishes the extension model that can solve contradictor issues including incompatible problems and antagonistic problems. By far, the relative theory and framework of extenics has got preliminary development, and is widely applied in engineering application, and has obtained corresponding achievements [10-14]. Thus, an applied technology that based on the Extention Theory has formed, namely the extension engineering.

Matter-element model: in extension model, it is one of the logical cells of extenics and also a formalization tool for formalized knowledge modelling. Model-element applied a sequential triad $R=(O, c, v)$ as the basic element in description object. In the formula, O stands for the name of each object, c stands for the characteristics of the object, v stands for the value of the object O on the characteristic c.

Extension distance: in extenics, stipulates the distance between the point x and the interval $X=[a, b]$ is stipulated. Set x as any point in real axis, $X=[a, b]$ as any interval in real domain, then name:

$$\rho(x, X) = \left| x - \frac{a+b}{2} \right| - \frac{b-a}{2}, \qquad (1)$$

as the extension distance between point x and the interval $X=[a, b]$.

2.2 MATTER-ELEMENT MODELING OF ANALYSIS OF URBAN WATER SUPPLY QUALITY

The analysis of urban water supply is a complex problem with multiple factors and levels. Establishing the matter-element model of analysis of city water supply quality in formalization and modelling method is the basis of the analysis of the grade of quality condition of city water supply. If there are n characters c_1, c_2, …, c_n and corresponding values v_1, v_2, …, v_n in the matter-element model of analysis of city water supply quality, its rectangular array formed by the corresponding matter-element model forms is a matter-element of n dimensions. It can be expressed as:

$$R(O,C,V) = \begin{bmatrix} O & c_1 & v_1 \\ & c_2 & v_2 \\ & \dots & \dots \\ & c_n & v_n \end{bmatrix}, \qquad (2)$$

where $C = (c_1, c_2, ..., c_n)^\mathrm{T}$, $V = (v_1, v_2, ..., v_n)^\mathrm{T}$. $v_i(1 \le i \le n)$ can be certain values or fuzzy interval values. In the process of analyzing the city water supply quality, via conduct element-matter modelling to influence factors in different categories, it provides the computer with formalization tool for intelligent implementation in analysis of city water supply quality.

3 Analysis model and algorithm of urban water supply quality based on extension classification analysis method

3.1 CLASSICAL AND JOINT DOMAINS

Assume that there are m grades of the quality condition of city water supply, and the i-th grade has n condition characteristics. So the classical domain matter-element model $R_i^{dom-cla}$ can be expressed as

$$R_i^{dom-cla} = [O_i, C, V] = \begin{bmatrix} O_i & c_{i1} & v_{i1} \\ & c_{i2} & v_{i2} \\ & \dots & \\ & c_{in} & v_{in} \end{bmatrix} = \begin{bmatrix} O_i & c_{i1} & \langle v_{i1}^a, v_{i1}^b \rangle \\ & c_{i2} & \langle v_{i2}^a, v_{i2}^b \rangle \\ & \dots & \dots \\ & c_{in} & \langle v_{in}^a, v_{in}^b \rangle \end{bmatrix}, \quad (3)$$

where c_{i1}, c_{i2}, …, c_{in} are the quality condition characters of water supply of the city in the i-th grade, v_{i1}, v_{i2}, …, v_{in} are respectively the values of quality condition of city water supply of the city in the i-th grade on the characteristics of n different influence factors. It can be accurate information of point value or fuzzy interval information, namely the classical domain of the i-th grade of the quality condition of the urban water supply satisfies $v_{ij} = \langle v_{ij}^a, v_{ij}^b \rangle, v_{ij}^a \le v_{ij}^b$ ($i=1,2,…, m$; $j=1,2,…, n$).

On the basis of the establishment of classical domain matter-element of grades of quality condition of water supply of m cities, the joint domain matter-element $R_0^{dom-joi}$ of quality condition of water supply can be formed:

$$R_0^{dom-joi} = [O_0, C, V] = \begin{bmatrix} O_0 & c_{01} & v_{01} \\ & c_{02} & v_{02} \\ & \dots & \\ & c_{0n} & v_{0n} \end{bmatrix} = \begin{bmatrix} O_0 & c_{01} & \langle v_{01}^a, v_{01}^b \rangle \\ & c_{02} & \langle v_{02}^a, v_{02}^b \rangle \\ & \dots & \dots \\ & c_{0n} & \langle v_{0n}^a, v_{0n}^b \rangle \end{bmatrix}, \quad (4)$$

where c_{01}, c_{02}, …, c_{0n} are n characters of different influence factors of O_0, v_{01}, v_{02}, …, v_{0n} are the value of O_0 on n characters of different influence factors c_{01}, c_{02}, …, c_{0n}. It can be accurate information of point value or fuzzy interval information, namely the joint domain of O_0. It satisfies $v_{0j} = \langle v_{0j}^a, v_{0j}^b \rangle, v_{0j}^a \le v_{0j}^b$. Obviously, $v_{ij} \subseteq v_{0j}$, and $v_{0j}^a = \min_{1 \le i \le m} v_{ij}^a, v_{0j}^b = \max_{1 \le i \le m} v_{ij}^b$.

3.2 STANDARDIZATION OF INFLUENCE FACATORS

In the process of choosing the characteristics of influence factors in the analysis of the quality condition of city water supply, some characteristics of influence factors can be

described quantitatively and their values are accurate point values. While other characteristics of influence factors are not certain values and are usually fuzzy and uncertain, and their values are described with an interval. On the other hand, in the analysis of quality condition of city water supply, some characteristics of influence factors are positive indexes, while others are negative indexes. Thus, the characteristics of influence factors in the analysis of quality condition of city water supply are complex and are of multiple levels, attributes and types. Thus, it needs to be standardized.

If the characteristic of influence factor of water supply v_{ij} is a positive index and its value $v_{ij} = \left\langle v_{ij}^a, v_{ij}^b \right\rangle$, its corresponding standardized index is:

$$v_{ij}^{\odot} = \left\langle v_{ij}^{\odot a}, v_{ij}^{\odot b} \right\rangle = \left\langle \frac{v_{ij}^a}{v_{0j}^b}, \frac{v_{ij}^b}{v_{0j}^b} \right\rangle, \tag{5}$$

where v_{0j}^b is the maximum value of grade joint domain of quality condition of water supply on the interval of characteristic of influence factors v_{ij}.

If the characteristic of influence factor of urban water supply quality v_{ij} is a negative index, its value $v_{ij} = \left\langle v_{ij}^a, v_{ij}^b \right\rangle$, the corresponding standardized index is:

$$v_{ij}^{\odot} = \left\langle v_{ij}^{\odot a}, v_{ij}^{\odot b} \right\rangle = \left\langle \frac{v_{0j}^a}{v_{ij}^a}, \frac{v_{0j}^a}{v_{ij}^b} \right\rangle, \tag{6}$$

where v_{0j}^a is the minimum value of grade joint domain of quality condition of water supply on the interval of characteristic of influence factors v_{ij}.

Thus, via the above-mentioned standardization of characteristics of influence factors of city water supply, all the characteristics of influence factors are in a unified measure standard and the differentiation among different types of characteristics of influence factors are removed. In this way, the accuracy of the analysis of quality condition of urban water supply will be improved.

3.3 ESTABLISHMENT OF EXTENSION
    CORRELATION FUNCTION

Assume that the analysed quality condition of urban water supply is P. Then the analytic object is described with matter-element model, and its matter-element model is R_P:

$$R_P = \begin{bmatrix} O_P & c_{P1} & v_{P1} \\ & c_{P2} & v_{P2} \\ & \cdots\cdots \\ & c_{Pn} & v_{Pn} \end{bmatrix} = \begin{bmatrix} O_P & c_{P1} & \left\langle v_{P1}^a, v_{P1}^b \right\rangle \\ & c_{P2} & \left\langle v_{P2}^a, v_{P2}^b \right\rangle \\ & \cdots & \cdots \\ & c_{Pn} & \left\langle v_{Pn}^a, v_{Pn}^b \right\rangle \end{bmatrix}, \tag{7}$$

where c_{P1}, c_{P2}, …, c_{Pn} are the n different design characteristics of the matter-element of quality condition of water supply R_P, v_{P1}, v_{P2}, …, v_{Pn} are relatively n values of matter-element of quality condition of water supply R_P on different design characteristics c_{P1}, c_{P2}, …, c_{Pn}. If the characteristic of the matter-element of quality condition of water supply R_P is the characteristic of influence factors that can be described accurately, its corresponding value of matter-element characteristic is an accurate point value. If the characteristic of the matter-element of quality condition of water supply R_P is not a characteristic of influence factors that can be described accurately, its corresponding value of matter-element characteristic is an interval value. Thus, $v_{Pj} = \left\langle v_{Pj}^a, v_{Pj}^b \right\rangle$.

Since the matter-element characteristics of the matter-element of water supply condition R_P can be either accurate point value or interval value, thus the establishments of the correlation functions are different. Thus, these two conditions need to be discussed separately.

a) When the matter-element characteristic of the matter-element of water supply condition R_P is an accurate point value, the extension distance $\rho_{R_i^{dom-cla}}^{R_P}(ij)$ between the matter-element of quality condition of water supply R_P and the classical domain $R_i^{dom-cla}$ of the i-th quality condition of urban water supply on the j-th matter-element characteristics is:

$$\rho_{R_i^{dom-cla}}^{R_P}(ij) = \left| v_{Pj} - \frac{v_{ij}^a + v_{ij}^b}{2} \right| - \frac{v_{ij}^b - v_{ij}^a}{2}. \tag{8}$$

Meanwhile, the extension distance $\rho_{R_0^{dom-joi}}^{R_P}(ij)$ between the matter-element of quality condition of water supply R_P and the joint domain matter-element $R_0^{dom-joi}$ of the i-th quality condition of urban water supply on the j-th matter-element characteristics is:

$$\rho_{R_0^{dom-joi}}^{R_P}(ij) = \left| v_{Pj} - \frac{v_{0j}^a + v_{0j}^b}{2} \right| - \frac{v_{0j}^b - v_{0j}^a}{2}. \tag{9}$$

The value of the extension correlation function $K_{R_i^{dom-cla}}^{R_P}(ij)$ of the matter-element of quality condition of water supply R_P and the classical domain matter-element $R_i^{dom-cla}$ of the i-th quality condition of city water supply on the j-th matter-element characteristics is:

$$K_{R_i^{dom-cla}}^{R_P}(ij) =$$
$$\begin{cases} -\rho_{R_i^{dom-cla}}^{R_P}(ij) / \left| v_{ij} \right|, v_{Pj} \in v_{ij} \\ \rho_{R_i^{dom-joi}}^{R_P}(ij) / (\rho_{R_0^{dom-cla}}^{R_P}(ij) - \rho_{R_i^{dom-cla}}^{R_P}(ij)), v_{Pj} \notin v_{ij} \end{cases}. \tag{10}$$

b) When the matter-element characteristic of the matter-element of water supply condition R_P is an interval

value, the extension distance $\rho_{R_i^{dom-cla}}^{R_P}(ij)$ between the matter-element of quality condition of water supply R_P and the classical domain $R_i^{dom-cla}$ of the i-th quality condition of city water supply on the j-th matter-element characteristics is:

$$\rho_{R_i^{dom-cla}}^{R_P}(ij) = \frac{\left[\left|v_{Pj}^a - \dfrac{v_{ij}^a + v_{ij}^b}{2}\right| + \left|v_{Pj}^b - \dfrac{v_{ij}^a + v_{ij}^b}{2}\right| - (v_{ij}^b - v_{ij}^a)\right]}{2}. \quad (11)$$

Meanwhile, the extension distance $\rho_{R_0^{dom-joi}}^{R_P}(ij)$ between the matter-element of quality condition of water supply R_P and the joint domain matter-element $R_0^{dom-joi}$ of the i-th quality condition of city water supply on the j-th matter-element characteristics is:

$$\rho_{R_0^{dom-joi}}^{R_P}(ij) =$$

$$\frac{\left[\left|v_{Pj}^a - \dfrac{v_{0j}^a + v_{0j}^b}{2}\right| + \left|v_{Pj}^b - \dfrac{v_{0j}^a + v_{0j}^b}{2}\right| - (v_{0j}^b - v_{0j}^a)\right]}{2}. \quad (12)$$

The value of the extension correlation function $K_{R_i^{dom-cla}}^{R_P}(ij)$ of the matter-element of quality condition of water supply R_P and the classical domain matter-element $R_i^{dom-cla}$ of the i-th quality condition of urban water supply on the j-th matter-element characteristics is:

$$K_{R_i^{dom-cla}}^{R_P}(ij) = \begin{cases} -\rho_{R_i^{dom-cla}}^{R_P}(ij)/\left|v_{ij}\right|, \left\langle v_{Pj}^a, v_{Pj}^b \right\rangle \in \left\langle v_{0j}^a, v_{0j}^b \right\rangle \\ \rho_{R_0^{dom-joi}}^{R_P}(ij)/(\rho_{R_0^{dom-cla}}^{R_P}(ij) - \rho_{R_i^{dom-cla}}^{R_P}(ij)). \\ \left\langle v_{Pj}^a, v_{Pj}^b \right\rangle \notin \left\langle v_{0j}^a, v_{0j}^b \right\rangle \end{cases} \quad (13)$$

Directing at the extension correlation functions of different matter-element characteristics obtained in the two methods mentioned above and considering the weight w_j of corresponding matter-element characteristic, comprehensive extension correlation $K(i)$ between the matter-element of quality condition of water supply R_P and the classical domain matter-element $R_i^{dom-cla}$ of the i-th quality condition of the urban water supply is:

$$K(i) = \sum_{j=1}^n (w_j \times K_{R_i^{dom-cla}}^{R_P}(ij)). \quad (14)$$

3.4 REALIZATION OF MODEL AND ALGORITHM OF ANALYSIS OF URBAN WATER SUPPLY BASED ON EXTENSION CLASSIFICATION METHOD

In the above-mentioned analysis of quality condition of city water supply, via constructing the classical domain

matter-element $R_i^{dom-cla}$ and the joint domain element $R_0^{dom-joi}$ of corresponding different grades of urban water supply quality, the extension correlation classification between the matter-element of the quality condition of water supply of the current city and the classical domain matter-element can be conducted. After the comprehensive extension correlation between the matter-element of the quality condition of water supply of the current city R_P and the classical domain matter-element $R_i^{dom-cla}$ is obtained, the extension correlation sequence **K** between the matter-element of quality condition of urban water supply of the current city R_P and the classical domain matter-element $R_i^{dom-cla}$ can be obtained, namely:

$$\mathbf{K} = \max(K(1), K(2), ..., K(n)) = K(t). \quad (15)$$

Thus, the grade of quality condition of city water supply correspondent to the classical domain matter-element $R_i^{dom-cla}$ is the quality condition of water supply of the current city. Thus, the subsequent water supply plans can be adjusted based on the quality condition of city water supply.

In conclusion, the concrete steps of implementation of quality analysis model of city water supply based on extension classification analysis method are as below:

Step 1 Divide the grades of quality condition of city water supply based on design experience and knowledge of this field, and extract the key characteristics of influence factors;

Step 2 Directing at the obtained characteristics of influence factors, conduct matter-element modelling to the classical and joint domain of different grades of quality condition of city water supply based on Equations (3) and (4);

Step 3 Directing at the water supply condition of the current city, extract the corresponding characteristics of influence factors and values of the factors and construct the matter-element model of the water supply condition of the current city based on Equation (7);

Step 4 Standardize different types of characteristics of influence factors based on Equations (5) and (6);

Step 5 If the characteristics of the value is point value information, obtain the extension distance and the extension correlation function of the water supply condition of the current city based on the Equations (8-10);

Step 6 If the characteristics of the value are fuzzy internal information, obtain the extension distance and the extension correlation function of the water supply condition of the current city based on the Equations (11-13);

Step 7 Considering the importance of different characteristics of matter-element, namely the weight, the comprehensive extension correlation between the quality condition of water supply and the classical domain matter-element of the quality condition of water supply can be obtained via Equation (14);

Step 8 Based on the obtained comprehensive extension correlation sequence of quality condition of city water supply, the grade of the quality condition of water supply of the current city can be obtained via Equation (15).

4 Case study

In this paper, the model and algorithm was tested and explained via the analysis of quality condition of water supply of a city. Via diagnoses, analysis and consultation to specialists and relative technician in city water supply engineering and designation, influence factors, including personal average amount of water supply, popularization rate of water supply, water supply peak, average values of water supply, etc, were extracted as characteristics of matter-element of analysis of quality condition of city water supply. Thus, the grade of quality condition of city water supply can be established as "excellent", "good", "fair", and "bad". And the corresponding classical domain matter-element model of grades of quality condition of city water supply are $R_1^{dom-cla}$, $R_2^{dom-cla}$, $R_3^{dom-cla}$ and $R_4^{dom-cla}$, respectively.

$$
R_1^{dom-cla} = \begin{bmatrix} O_1^{dom-cla} & \text{personal average amount of water supply (m}^3) & 45-55 \\ & \text{popularization rate of water supply(\%)} & 95-100 \\ & \text{water supply peak}(10^7 \times m^3) & 5.5\text{-}6.0 \\ & \text{average values of water supply}(10^7 \times m^3) & 5.0-5.8 \\ & \text{water supply price(RMB/m}^3) & 0.9-1.1 \\ & \text{water quality(\%)} & 95-100 \end{bmatrix},
$$

$$
R_2^{dom-cla} = \begin{bmatrix} O_2^{dom-cla} & \text{personal average amount of water supply(m}^3) & 40-45 \\ & \text{popularization rate of water supply(\%)} & 85-95 \\ & \text{water supply peak(m}^3) & 4.5-5.5 \\ & \text{average values of water supply(m}^3) & 4.2-5.0 \\ & \text{water supply price(RMB/m}^3) & 1.1-1.2 \\ & \text{water quality(\%)} & 85-95 \end{bmatrix},
$$

$$
R_3^{dom-cla} = \begin{bmatrix} O_3^{dom-cla} & \text{personal average amount of water supply(m}^3) & 35-40 \\ & \text{popularization rate of water supply(\%)} & 75-85 \\ & \text{water supply peak(m}^3) & 3.8-4.5 \\ & \text{average values of water supply(m}^3) & 3.6-4.2 \\ & \text{water supply price(RMB/m}^3) & 1.2-1.3 \\ & \text{water quality(\%)} & 75-85 \end{bmatrix},
$$

$$
R_4^{dom-cla} = \begin{bmatrix} O_4^{dom-cla} & \text{personal average amount of water supply(m}^3) & 25-35 \\ & \text{popularization rate of water supply(\%)} & 60-75 \\ & \text{water supply peak(m}^3) & 3.4-3.8 \\ & \text{average values of water supply(m}^3) & 3.0-3.6 \\ & \text{water supply price(RMB/m}^3) & 1.3-1.5 \\ & \text{water quality(\%)} & 60-75 \end{bmatrix}.
$$

Transfer the water supply condition of the current city into the matter-element of city water supply condition R_P:

$$
R_P = \begin{bmatrix} O_P & \text{personal average amount of water supply(m}^3) & 40-45 \\ & \text{popularization rate of water supply(\%)} & 86-91 \\ & \text{water supply peak(m}^3) & 5.5 \\ & \text{average values of water supply(m}^3) & 4.8 \\ & \text{water supply price(RMB/m}^3) & 1.2 \\ & \text{water quality(\%)} & 83-86 \end{bmatrix}.
$$

Bai Yun, Wang Pu, Xie Jingjing, Li Chuan

Based on the standardization method of influence factor of quality condition of city water supply in this paper, the extension distance between the water supply condition of the current city R_P and the corresponding classical domain of quality condition of city water supply $R_1^{dom-cla}$, $R_2^{dom-cla}$, $R_3^{dom-cla}$ and $R_4^{dom-cla}$, and that between water supply condition of the current city R_P and the joint domain $R_0^{dom-joi}$ were obtained through relative calculation formula of extension distance. The result is presented in Table 1.

According to the corresponding calculation formula of extension correlation function, the extension correlation function between the water supply condition of the current city R_P and the corresponding classical domain of quality condition of city water supply $R_1^{dom-cla}$, $R_2^{dom-cla}$, $R_3^{dom-cla}$ and $R_4^{dom-cla}$, and that between the water supply condition of the current city R_P and the corresponding joint domain of quality condition of city water supply $R_0^{dom-joi}$. The result is presented in Table 2.

TABLE 1 The extension distance of the water supply condition R_P of the current city

Characteristics	$R_1^{dom-cla}$	$R_2^{dom-cla}$	$R_3^{dom-cla}$	$R_4^{dom-cla}$	$R_0^{dom-joi}$
Personal average amount of water supply	0.046	0	0.046	0.137	-0.228
Popularization rate of water supply	0.065	-0.025	0.035	0.135	-0.115
Water supply peak	0	0	0.167	0.284	0.083
Average values of water supply	0.034	-0.034	0.104	0.206	-0.172
Water supply price	0.068	0	0	0.058	-0.150
Water quality	0.105	0.005	-0.010	0.095	-0.155

TABLE 2 The extension correlation function of the water supply condition of the current city R_P

Characteristics	$R_1^{dom-cla}$	$R_2^{dom-cla}$	$R_3^{dom-cla}$	$R_4^{dom-cla}$
Personal average amount of water supply	-0.168	0	-0.168	-0.375
Popularization rate of water supply	-0.361	0.500	-0.233	-0.540
Water supply peak	0	0	-0.668	-0.774
Average values of water supply	-0.165	0.246	-0.377	-0.545
Water supply price	-0.312	0	0	-0.279
Water quality	-0.404	-0.031	0.333	-0.038

Due to the different weight of different characteristics of influence factors, the weight of each characteristics of influence factors is obtained via the experience of design specialist, namely W=(0.18,0.18,0.13,0.15,0.17,0.19). By using the calculation formula of comprehensive extension correlation, the extension correlation sequence **K** between the water supply condition of the current city R_P and the corresponding classical domain of quality condition of city water supply $R_1^{dom-cla}$, $R_2^{dom-cla}$, $R_3^{dom-cla}$ and $R_4^{dom-cla}$ can be obtained **K**=(-0.235,0.119,-0.186,-0.425). It can be observed that the water supply condition of the current city R_P is the closest to the classical domain of quality condition of city water supply $R_2^{dom-cla}$, namely the city water supply of the current city is in good condition. However, the differentiation between the water supply condition of the current city R_P and quality condition of city water supply $R_3^{dom-cla}$ is smaller than the differentiation between the water supply condition of the current city R_P and quality condition of city water supply $R_1^{dom-cla}$, which indicates that the water supply condition of the current city has a tendency of growing bad. Thus, in order to keep the city water supply in a good condition, corresponding measures should be taken to maintain and adjust the city water supply project.

5 Conclusion

This paper came up with an analysis model and algorithm of city water supply quality based on extension classification analysis method. Via establishing the classical domain matter-element and joint domain matter-element, the extension distances, extension correlation function and the calculation models of extension correlation relatively based on accurate point value and interval value. Based on that, the grade of quality condition of water supply of the current city can be obtained. It offer guide information for the city water supply strategy and can help to improve the reliability of city water supply quality. What's more, it offers a solution to the implementation of intelligent system of city water supply. Finally, in this paper the model and algorithm prove feasible and operable via a case study.

Acknowledgments

Operation Research and Decision Making

References

[1] Wang L, Wang Z, Yue L 2007 Forecast model of urban daily water consumption based on particle swarm optimization *China Water and Wastewater* **23**(7) 89-93 (*in Chinese*)

[2] Li H, Cui J, Zhang X 2004 Optimal research of water consumption prediction model *China Water and Wastewater* **20**(2) 27-8 (*in Chinese*)

[3] Zheng S 1995 Research on forecasting model of urban water supply quantity *Journal of Chengdu University of Science and Technology* **6** 18-26 (*in Chinese*)

[4] Xu H, Yuan Y, Zhao H 2001 Use of gray model GM(1,1) for forecasting annual water demand *Journal of Harbin University of C.E. & Architecture* 2001 **34**(4) 61-4 (*in Chinese*)

[5] Miao Z, Fu Q 2012 Study on prediction model of urban water demand based on grey support vector machine *Computer Simulation* **29**(8) 196-9 (*in Chinese*)

[6] Maruyama Y, Yamamoto H 2006 A study for forecasting water supply demand in the Area-K waterworks service area *The 5th IEEE/ACIS International Conference on Computer and Information Science 2006 and 2006 1st IEEE/ACIS International Workshop on Component-Based Software Engineering, Software Architecture and Reuse ICIS-COMSAR 2006* 136-141

[7] Xie Y, Zheng H 2008 Water supply forecasting based on developed LS-SVM *The 3rd IEEE/ICIEA Conference on Industrial Electronics and Applications 2008 ICIEA 2008* 2228-33

[8] Zou G, Wang H, Wang D, Zhang G 2004 Urban water consumption forecast based on neural network model *Information and Control* **33**(3) 364-8 (*in Chinese*)

[9] Durga Rao KHV 2005 Multi-criteria spatial decision analysis for forecasting urban water requirements: a case study of Dehradun city India. *Landscape and Urban Planning* **71**(2) 163-174

[10] Cai W, Yang C 2013 Basic theory and methodology on extenics *Chinese Science Bulletin* **58**(13) 1190-9 (*in Chinese*)

[11] Wang T, Yang A, Zhong S 2014 Products extension adaptive design based on case reuse *International Journal of Control and Automation* **7**(1) 295-306

[12] Zhao Y W, Zhang G X 2012 A new integrated design method based on fuzzy matter-element optimization *Journal of Materials Processing Technology* **129**(1) 612-8

[13] Gong J, Qiu J, Li G, Li W 2008 Extension approach for product module variant design *Computer Integrated Manufacturing Systems* **14**(7) 1256-67 1312 (*in Chinese*)

[14] Zhao Y, Su N, Zhang F, Chen J 2010 Configuration design method for product family based on extension case reasoning *Chinese Journal of Mechanical Engineering* **46**(15) 146-154 (*in Chinese*)

Authors

Yun Bai, born in January, 1985, Shanxi, China

Current position, grades: PhD Candidate, School of Urban Construction and Environmental Engineering, Chongqing University, China.
University studies: Master Degree in environment engineering, Chongqing Technology and Business University in 2008.
Scientific interests: water supply system optimization and scheduling, computer simulation.

Pu Wang, born in October, 1965, Chongqing, China

Current position, grades: professor, School of Urban Construction and Environmental Engineering, Chongqing University, China.
University studies: PhD degree in municipal engineering, Chongqing University in 2004.
Scientific interests: water supply system optimization, water treatment.

Jingjing Xie, born in September, 1985, Anhui, China

Current position, grades: engineer, Testing Center for Science and Technology, Chongqing Academy of Science and Technology, China.
University studies: Master Degree in environment engineering, Chongqing Technology and Business University in 2008.
Scientific interest: data mining, water treatment.

Chuan Li, born in December, 1975, Chongqing, China

Current position, grades: professor, Chongqing Technology and Business University, China.
University studies: Ph.D. degree in machinery and electronics engineering, Chongqing University in 2007.
Scientific interest: data mining, computer technology, intelligence algorithm, model optimization.

Operation Research and Decision Making

New logistics distribution route dispatching mode based on genetic algorithm-ant colony algorithm

Fasheng Yi[1, 2], Xiaoling Li[1, 2*], Jimin Yuan[3]

[1]*School of Computer Science and Technology, Chengdu University, Chengdu 610106, P. R. China*

[2]*University Key Laboratory of Pattern Recognition and Intelligent information Processing, Chengdu 610106, P. R. China*

[3]*The Department Computer Engineering, Chengdu Aeronautic Vocational and Technical College, Chengdu 610000, China*

Abstract

For multi-target route optimization with constraint conditions, the mathematical model for logistics distribution route optimization is built to accelerate response speed of logistics enterprises to customers, improve service quality, and strengthen the satisfaction of customers, and a new algorithm with the combination of genetic and ant colony algorithms is proposed to solve the selection issues of such logistics route. Initial pheromone is formed with genetic algorithm, based on which the optimal solution is rapidly sought with ant colony algorithm, and complementary advantages are achieved between above two algorithms. Application examples and simulations are available for calculation, and the results show that such algorithm is practical and effective to optimize logistics distribution route.

Keywords: logistics distribution, genetic algorithm, ant colony algorithm, combination, optimization

1 Introduction

Genetic Algorithm (GA) is a bionic optimization algorithm first proposed by John H. Holland, professor of University of Michigan in 1975. Which is based on Darwin's biological evolutionism of "survival of the fittest" and Mendel's genetic variation theory of "biological genetic evolution is mainly dependent on the chromosome on which the genes from parents are orderly arranged", and simulates the biological evolution process.

Ant colony algorithm is a newly-developed bionic optimization algorithm for simulating the intelligent behaviour of ant colony [1] proposed by Italian scholar Dorigo M in 1991, which boasts such advantages as strong robustness, excellently distributed computation mechanism and easy to combine with other methods [2, 3].

The improvement strategy that GA and ant colony algorithm are combined is first proposed by Abbattista F et al. [4]. And good results are obtained from Oliver30TSP and Eilon50TSP simulation experiments; later, the ant colony algorithm is combined with GA to solve various optimization issues in discrete and continuous domains with good application effects achieved. GA and ant colony algorithm are combined by Ding Jian et al. [5] to solve combinatorial explosion and NP issues with excellent effects with respect to optimal performance and time performance. The algorithm in this paper is aimed to conduct clustering analysis, i.e. GA is combined with ant colony algorithm, and the initial clustering centre for data object is established by virtue of rapid and random groupment global searching capacity of GA. The

clustering structure is perfected based on such characteristics as parallelism, positive and negative feedback and high solution efficiency of ant colony algorithm.

2 Basic principle and design concept of combined algorithm

With wide-range rapid global searching capacity, GA will produce massive redundant iteration for solution to a certain extent, and fail to fully use feedback information in the system with low solution efficiency. Due to random distribution of initial data object, insufficient pheromone and slow solution speed, the ant colony algorithm requires a longer searching time, and is easy to produce premature stagnation phenomenon when large optimization issues are solved, as shown in the speed-time curve of GA and ant colony algorithm in Figure 1 [6]. GA boasts higher convergence speed in the early stage ($t_0 \sim t_a$ time bucket), but the efficiency after ta is significantly reduced. Due to the randomness of data and its movement, the searching speed of the ant colony algorithm is slow, but the efficiency is significantly increased after a certain period of time. The basic thought of genetic algorithm-ant colony algorithm is that GA-based rapid global searching capacity and ant colony algorithm-based positive feedback convergence mechanism. GA is adopted in the early stages. The basic thought of which is that GA is available in the process before the algorithm, and the rapidity, randomness and global convergence of GA shall be fully utilized to produce the distribution of initial pheromone for

relevant issues. The ant colony algorithm is adopted in the process after algorithm, and under the condition that certain initial pheromones are distributed, the parallelism, feedback and high solution efficiency of ant colony algorithm shall be fully utilized to achieve exact solution. Such two algorithms complement each other's advantages so as to obtain best time and optimization performance.
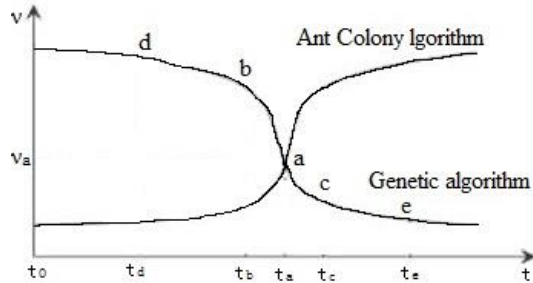
FIGURE1 Speed–time curve of GA and ant colony algorithm

2.1 ESTABLISHMENT OF LOGISTICS DISTRIBUTION MODEL

The goods are delivered from logistics centre to customers scattering in different geographic locations with various vehicles, in which the locations of logistics centre and each customer, customer demand, loading capacity of each distribution vehicle and maximum mileage are established for one distribution of each vehicle. It is required to reasonably arrange the distribution route of vehicle so that two object functions, i.e. total distribution distance and non-full load rate of distribution vehicle, can be optimized.

The following assumptions are proposed in the model Reference [7]:

1) the direction of goods is unidirectional, i.e. delivery only.

2) The loading capacity of each vehicle is limited and shall be greater than aggregate demand of customer in its transportation route;

3) With the demand of each customer known, required goods can only be delivered with one vehicle, and all customers shall be fully served.

4) With distances between distribution centre and each customer and between customers known, the maximum mileage of vehicle must be greater than the length of distribution route;

5) Each route is led from and to distribution centre;

The model building will involve the number of customers, customer demand, loading capacity of vehicle, number of vehicles and maximum mileage, etc. This Paper is based on the single object optimization model proposed in Reference [8], and specifies multi-target optimization model. For the sake of study, the following variable parameters are defined:

K: the number of vehicles in the distribution centre;

L: the number of customer points;

Q_k: loading capacity of vehicle, where ($k =1,2,…,K$);

D_k: maximum mileage of the vehicle;

Q_i: demand of customer points, where ($i = 1,2, –L$);

d_{i-j}: the distance between customer points i and j, wherein ($i, j=1,2,...,L$);

d_{0-j}: the distance between distribution centre and each demand point;

n_k: total number of customers served with vehicle k; when $n_k=0$, meaning k is not involved in distribution;

R_k: a set, which means the set of customer points served with vehicle k;

r_{ki}: an element of R_k in the set, which means that the sequence of such customer point in distribution route by vehicle k is i;

R_{k0}: distribution centre.

The optimization of logistics distribution vehicle route is such that relevant constraint conditions are met so as to minimize the total logistics distribution distance and maximize full load rate of vehicle. The model is as follows:

$$\min z_1 = \sum_{k=1}^{K}\left[\sum_{i=1}^{n_k}d_{r_{k(i-1)}r_{ki}} + d_{r_{knk}r_{k0}}sign(n_k)\right], \quad (1)$$

$$s.t. \sum_{i=1}^{n_k}d_{r_{k(i-1)}r_{ki}} + d_{r_{knk}r_{k0}}sign(n_k) \le D_k, \quad (2)$$

$$\sum_{i=1}^{n_k}q_{r_{ki}} \le Q_k, \quad (3)$$

$$0 \le n_k \le L, \quad (4)$$

$$\sum_{k=1}^{K}n_k = L, \quad (5)$$

$$R_k = \left\{r_{ki}\middle|r_{ki} \in \{1,2,...,L\}, i=1,2,...,n_k\right\}, \quad (6)$$

$$R_{k_1} \bigcap R_{k_2} = \varnothing, \forall k_1 \ne k_2, \quad (7)$$

$$sign(n_k) = \begin{cases}1,(n_k \ge 1)\\0,(else)\end{cases}, \quad (8)$$

where Equation (1) means that the total target distribution distance is shortest; Equation (2) means that the length of each vehicle distribution route shall not exceed maximum mileage for one distribution; Equation (3) means that the sum of demand of each demand point on each vehicle distribution route shall not exceed its maximum loading capacity; Equation (4) means that the number of customer points on each route shall not exceed the number of total demand points; Equation (5) means that that each customer point can be better served; Equation (6) is the combination of customer points on each route; Equation (7) means that that each customer point is served with only one vehicle; Equation (8) is aimed to check whether vehicle k is involved in the distribution: if the number of customers served is ≥ 1, it means that vehicle k is involved, otherwise vehicle k is not involved.

2.2 STRUCTURE OF GA AIMING AT LOGISTICS DISTRIBUTION ROUTE OPTIMIZATION

GA is established to optimize logistics distribution route based on its characteristics.

1) Establishment of coding method: binary coding is adopted: "0" means distribution centre, and "1" means a customer ($L\times$"1" means L different customer(s)). Since the distribution centre is provided with k vehicle(s), there are k distribution routes at most, and each distribution route is led from and to distribution centre. $L\times$"1"or $K-1\times$ "0" is randomly arranged in a chromosome (N+K−1 digit), which is corresponding to a distribution route scheme. For example, if 6 customers are served with 3 vehicles, the 8 characters, i.e. $6\times$"1" and $2\times$"0" (the distribution centre 0 from which the first vehicle is led is default for the reason that all vehicles are led from distribution centre) are randomly arranged to establish a logistics distribution scheme. There are 3 distribution routes in the distribution route scheme represented by chromosome 11011011: route a: (0) − 1 − 1 − 0; route b: 0 − 1 − 1 − 0; route c: 0 − 1 − 1 − 0.

2) Fitness evaluation: for the distribution route scheme corresponding to an individual, the advantages and disadvantages shall be judged: first, check whether constraint conditions for distribution are met; second, the object function value (i.e. the sum of each distribution route length) shall be calculated. According to the coding method determined based on the characteristics of distribution route optimization issues, the constraint conditions that each demand point can be better served with only one vehicle are met, and that the sum of demand of each demand point on each route shall not exceed the loading capacity of the vehicle, and the length of each distribution route shall not exceed the maximum mileage for one distribution are not met. For this reason, Equation (2) is converted into a part of transportation distance as follows Equation (10):

$$Z = \sum_{k=1}^{K}\left[\sum_{i=1}^{n_k} d_{r_{k(i-1)}r_{ki}} + d_{r_{knk}r_{k0}} sign(n_k) \right]$$
$$+M\sum_{k=1}^{K}\max\left[\sum_{i=1}^{n_k} q_{r_{ki}} - Q_k, 0 \right],$$
(9)

where M is a large positive number, which means when the volume of freight of a vehicle exceeds the penalty coefficient of its maximum loading capacity, M is 1000000 in software design. The transportation distance is converted into fitness function: $F_i = \dfrac{\min z}{z_i}$, where F_i is the fitness of the *i-th* chromosome, min z is the target value of optimal chromosome in the current colony, and z_i is the target value of the *i-th* chromosome.

3) Selection operation: N individual(s) in the colony of each generation is arranged from big to small by fitness, and the individual ranking No.1 boasts optimal

performance, and is copied to directly enter into the next generation and ranks No.1. $N–1$ individual(s) in the colony of next generation shall be produced with roulette wheel selection method [4] based on the fitness of N individual(s) in the colony of former generation. Specifically, it shall first calculate the sum (ΣF_j) of fitness of all individuals in the colony of former generation, and then calculate the proportion ($F_j/\Sigma F_j$) of fitness of each individual as the selection probability. With such selection method adopted, the optimal individual will survive to next generation, and the individual with larger fitness may have a greater chance to survive to next generation.

4) Crossover operation: for new colony produced from selection operation, N-1 individual(s), in addition to optimal individual ranking No.1, will be mated, crossed and reorganized based on crossover probability P_c. A crossover method similar to OX method [2] is adopted as follows:

a) randomly select a mating area in the parent individual, e.g. two parent individuals and mating areas are selected as A = 478563921, B = 834691257;

b) A′=4691478563921 and B′=8563834691257 that are obtained by putting B mating area in front of A, and A mating area in front of B;

c) the natural numbers identical to mating area in A′ and B′ shall be deleted in order after mating area so as to obtain two final individuals, i.e. A″=469178532, B″=856349127. Compared with other crossover methods, such method still can produce a certain variation effect under the condition that two parent individuals are identical, which plays a positive role in maintaining the diversification characteristics of such colony.

5) Variation operation: since the optimal sample is maintained in selection mechanism, the variation technology by continuous conversion is adopted to maintain the diversification of individuals in the colony, and make the individual change to larger extent in arrangement of the order. The variation operation is based on probability P_m, and the exchange time J is produced with random method once it occurs, and the genes of individuals required for variation operation are exchanged for J times (the location of exchanged gene is produced randomly).

2.3 ANT COLONY ALGORITHM RULE IN COMBINED ALGORITHM

1) Fitness function: the fitness functions in both ant colony algorithm and GA stages are identical.

2) Route selection strategy:

$$j = \begin{cases} \operatorname*{arg\,max}_{j\in allowedk}([\tau_{ij}(t)]^{\alpha}[\eta_{ij}(t)]^{\beta}), & if\,(q<q_0) \\ s, & else \end{cases},$$
(10)

$\tau_{ij}(t)$ is the amount of information on (i,j) at t time route,

315

wherein $\eta_{ij}(t)$ is $F_i = \dfrac{\min z}{z_i}$, and q_0 is a preset variable in (0,1), which is larger in the early stage of algorithm, but with the development of algorithm, q_0 is gradually reduced. q is a random variable.

$$s = \begin{cases} \dfrac{[\tau_{ij}(t)]^{\alpha}[\eta_{ij}(t)]^{\beta}}{\displaystyle\sum_{s \in allowedk} [\tau_{is}(t)]^{\alpha}[\eta_{is}(t)]^{\beta}}, & j \in allowedk \\ 0, & else \end{cases} \quad . \quad (11)$$

3) Setup rule for initial pheromone value: a set of optimal multicast tree solution is obtained from GA, which has great significance on optimal solution, so the concentration of route pheromone on such multicast tree is strengthened, and initial pheromone value is set as $\tau_0 = \tau_C + \tau_G$, wherein τ_C is a pheromone constant, and τ_G is the pheromone value obtained from GA. There are m sets of routes for the colony produced after iteration, and n set(s) of routes in front of fitness value are selected. For the pheromone on each route, the formula and pheromone constant τ_C are added to obtain the distribution τ_0 of initial pheromone in the ant colony algorithm.

4) Rules for updating pheromone: the pheromone is updated with ant cycle model, i.e. the pheromone of ant with shortest route can only be modified and added in one cycle, and thus, the track renewal equation for all routes is adopted:

$$\tau_{ij}(t+1) = \rho\tau_{ij}(t) + \sum \Delta\tau_{ij}^{k}(t) . \quad (12)$$

3 Combining algorithms for logistics route optimization

Step 1: Initialize the network node and start to encode routing character of the network from source node.

Step 2: Optimize the GA route to generate initial population, and reserve the colony with the highest fitness via selection, crossover and inheritance operations of GA; after the genetic iterations reaches to specified maximum, n group(s) colony with higher fitness generate(s) the initial distribution of pheromone: $\tau_0 = \tau_C + \tau_G$.

Step 3: Initialize the tabu list, tabuk and allow list allowedk. Add the source node into tabuk and all rest nodes into allowedk, and use the source route method to put m ant(s) on the source node.

Step 4: Select the route. Flood ants on links except those between nodes in the tabuk according to ant species. Mark as the k batch ants, and select one route according to Equation (10) and (11).

Step 5: Update the pheromone. Locally update the pheromone of such route through which each ant passes according to the Equation (12). Delete j in the allowedk and add it in the tabuk.

Step 6: For k (k=1, 2, …, M), if there is a target node in the allowedk, turn to Step 4; otherwise, the flooding shall be finished, and the k batch ants are dead.

Step 7: When the algorithm is in a halted state, the current optimal solution shall be saved. It shall judge whether it is lower than the reference value; if so, the result shall be output; otherwise, turn to Step 3.

Step 8: If the algorithm achieves the specified iterations, the result shall be output; otherwise, turn to Step 3.
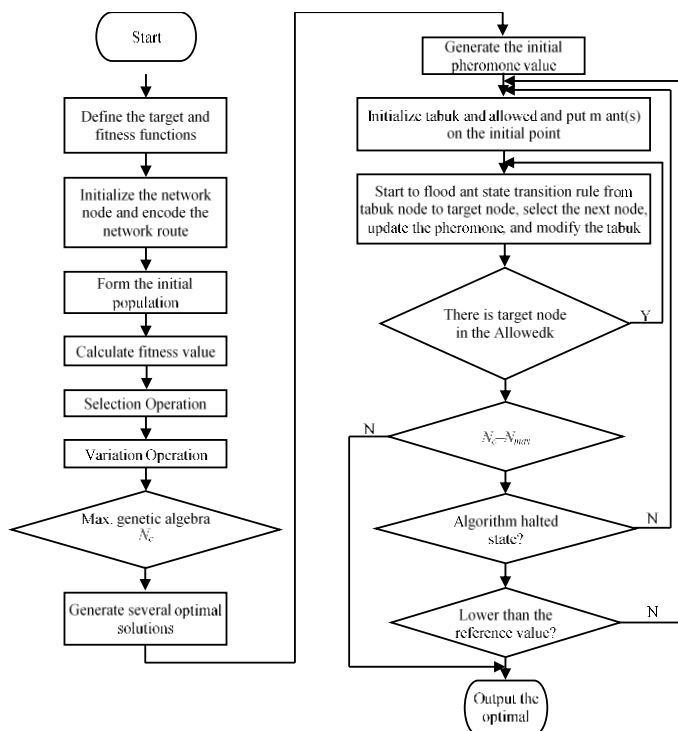
Algorithm process is showed as Figure 2:

FIGURE 2 Algorithm process flow chart

4 Simulation experiment results

The new algorithm, with the combination of GA and ant colony algorithm for logistics distribution route optimization, sets the GA iterations as 50 and initial pheromone values [10] τ_C of all routes in the ant colony algorithm as 60; the pheromone value converted from GA result is programmed with C language by adding 2 via the route and updating the track $\rho = 0.8$ and $Q = 1000$; the calculated example results are analysed and compared.

Example: Some distribution centre is used as an example, and it shall assume that the distribution centre assigns 5 trucks to 30 customers, the maximum loading capacity of each truck is 10t, and the maximum mileage for each distribution is 50km. The coordinates and

demands for 30 customers (to-be distributed locations) are shown in the Table 1, in which the number of 0 represents the distribution centre.

The shortest distance for each iteration process is obtained from 15 iterations of the computer, as shown in the Table 2; wherein the minimum value of all shortest distances is 204.99km, namely that the shortest mileage of this distribution example is 204.99km, and the optimal route corresponding to the distribution scheme is:

Route 1: 0→20→11→4→14→19→22→1→7→23→0
Route 2: 0→16→26→17→5→21→28→30→0;
Route 3: 0→12→6→25→2→15→0;
Route 4: 0→29→8→27→0;
Route 5: 0→9→18→10→24→13→3→0.

TABLE 1 Distance between the Distribution Center and Demand Point/km, and the Demand of each Demand Point Q/t

i\j	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
0	0.00	18.36	18.02	9.43	12.16	9.89	7.28	15.00	14.00	10.63	15.62	9.05	7.21	14.76	14.31	5.09	5.00	10.00	12.04	15.65	5.09	16.28	18.38	3.16	19.10	11.70	5.00	14.31	11.04	2.24	12.16
1	18.36	0.00	29.01	18.44	12.37	17.46	22.13	2.00	9.00	27.02	32.20	15.00	20.61	31.38	8.94	15.52	12.73	18.79	28.42	8.60	13.60	21.21	4.12	13.60	35.61	25.50	11.31	6.00	23.85	15.62	26.40
2	18.02	29.01	0.00	16.55	17.11	27.80	11.31	27.02	30.67	19.70	20.22	14.03	10.82	17.12	21.21	14.87	22.36	27.66	20.25	22.36	16.76	34.06	28.16	20.12	23.02	7.21	24.70	29.83	26.40	20.25	25.55
3	9.43	18.44	16.55	0.00	19.72	5.38	7.07	24.41	22.56	3.16	6.40	15.65	9.22	5.38	22.80	12.20	13.93	14.32	4.00	24.21	13.45	20.25	27.66	12.53	10.00	9.49	14.14	23.41	10.82	10.77	9.22
4	12.16	12.37	17.11	19.72	0.00	19.65	13.45	10.44	16.97	21.93	26.07	4.24	11.31	24.21	4.12	7.61	12.37	20.40	25.63	11.04	11.04	29.68	15.26	15.52	15.52	15.00	23.02	13.00	24.33		
5	9.89	17.46	27.80	5.38	19.65	0.00	16.64	16.76	9.90	14.03	19.23	17.89	17.03	20.00	20.02	14.42	7.28	1.41	5.38	21.00	13.42	6.40	20.88	8.94	21.84	21.09	4..12	12.20	6.71	7.81	10.29
6	7.28	22.13	11.31	7.07	13.45	16.44	0.00	20.25	21.09	10.00	13.00	9.22	2.24	14.32	17.03	6.71	12.16	16.40	11.04	18.44	8.54	22.80	22.82	10.05	16.55	4.47	13.93	21.02	15.26	9.49	14.87
7	15.00	2.00	27.02	24.41	10.44	16.76	20.25	0.00	9.22	25.49	30.61	18.68	29.68	7.21	13.60	11.70	21.02	6.32	3.00	21.95	15.23	3.16	25.61	22.00	34.06	23.54	13.42	6.32	22.82	14.15	25.24
8	14.00	9.00	30.67	22.56	16.97	9.90	21.09	9.22	0.00	22.47	27.86	17.49	20.40	27.89	10.63	15.81	9.00	11.31	23.77	15.65	13.93	12.37	13.04	11.04	30.87	25.32	8.54	3.00	17.03	12.04	20.00
9	10.63	27.02	19.70	3.16	21.93	14.03	10.00	25.49	22.47	0.00	5.38	18.03	12.04	6.08	24.70	14.32	14.42	13.00	1.41	26.08	15.26	18.44	29.00	13.45	8.60	12.65	13.93	23.70	8.54	11.40	6.40
10	15.62	32.20	20.22	6.40	26.07	19.23	13.00	30.61	27.86	5.38	0.00	21.95	15.23	3.16	29.21	18.60	19.72	18.11	4.12	30.61	19.85	23.09	33.97	18.60	3.60	14.03	19.31	29.07	13.04	16.64	10.20
11	9.05	15.00	14.03	15.65	4.24	17.89	9.22	13.00	17.49	18.03	21.95	0.00	7.07	20.00	8.06	4.00	10.81	18.38	19.31	9.43	4.47	24.19	14.56	8.94	25.55	11.18	13.89	16.15	20.10	10.44	21.02
12	7.21	20.61	10.82	9.22	11.31	17.03	2.24	18.68	20.40	12.04	15.23	7.07	0.00	13.04	15.00	5.10	11.70	16.97	13.15	16.40	7.07	23.34	21.02	9.49	18.79	5.00	13.89	20.02	16.55	10.29	19.42
13	14.76	31.38	17.12	5.38	24.21	20.00	14.32	29.68	27.89	6.08	3.16	20.00	13.04	0.00	27.59	16.97	19.31	19.03	5.38	29.00	18.44	24.51	32.80	17.89	6.08	11.18	19.42	28.79	14.56	16.15	12.08
14	14.31	8.94	21.21	22.80	4.12	20.02	17.03	7.21	10.63	24.70	29.21	8.06	15.00	27.59	0.00	10.63	13.04	21.00	26.08	1.41	9.43	25.50	7.00	12.37	25.61	19.23	16.00	12.80	24.51	14.56	26.25
15	5.09	15.52	14.87	12.20	7.61	14.42	6.71	14.42	15.81	14.32	18.60	4.00	5.10	16.97	10.63	0.00	7.81	14.76	15.65	12.04	3.60	20.81	16.12	5.67	22.20	10.05	10.63	15.13	16.12	6.71	17.03
16	5.00	12.73	22.36	13.93	12.37	7.28	12.16	11.40	9.00	14.42	19.72	10.81	11.70	19.31	13.04	7.81	0.00	8.06	15.81	14.14	6.40	13.42	15.26	2.24	23.02	16.49	3.16	9.49	11.70	3.16	13.89
17	10.00	18.79	27.66	14.32	20.40	1.41	16.40	18.03	11.31	13.00	18.11	18.38	16.97	19.03	21.00	14.76	8.06	0.00	14.03	22.02	13.93	6.40	22.13	9.49	20.61	20.81	5.00	13.60	5.83	8.06	8.94
18	12.04	28.42	20.25	4.00	23.26	5.38	11.04	26.90	23.77	1.41	4.12	19.31	13.15	5.38	26.08	15.65	15.81	14.03	0.00	27.46	16.64	19.23	30.41	14.87	7.21	13.34	15.23	25.06	9.22	12.81	6.71
19	15.65	8.60	22.36	24.21	5.38	21.00	18.44	7.07	15.65	26.08	30.61	9.43	16.40	29.00	1.41	12.04	14.14	22.02	27.46	0.00	10.82	26.30	6.08	13.60	34.20	20.59	17.03	13.04	25.71	15.81	27.51
20	5.09	13.60	16.76	13.45	7.07	13.42	8.54	11.70	13.93	15.26	19.85	4.47	7.07	18.44	9.43	2.00	6.40	13.93	16.64	10.82	0.00	19.72	14.42	4.47	23.43	12.04	9.43	13.15	16.00	6.08	17.26
21	16.28	21.21	34.06	20.25	25.63	6.40	22.80	21.02	12.37	18.44	23.09	24.19	23.34	24.51	25.50	20.81	13.42	6.40	19.23	26.30	19.72	0.00	25.08	15.26	25.02	27.20	10.31	9.49	13.60	13.15	9.49
22	18.38	4.12	28.16	27.66	11.04	20.88	22.82	4.12	13.04	29.00	33.97	14.56	21.02	32.80	7.00	16.12	15.26	22.13	30.41	6.08	14.42	25.08	0.00	15.62	37.48	25.63	17.46	10.05	26.83	17.80	29.15
23	3.16	13.60	20.12	12.53	11.04	8.94	10.05	12.04	11.04	13.45	18.60	8.94	9.49	17.89	12.37	5.67	2.24	9.49	14.87	13.60	4.47	15.26	15.62	0.00	22.02	14.32	5.00	11.18	12.16	2.24	13.93
24	19.10	35.61	23.02	10.00	29.68	21.84	16.55	34.06	30.87	8.60	3.60	25.55	18.79	6.08	25.61	22.20	23.02	20.61	7.21	34.20	23.43	25.02	37.48	22.02	0.00	17.26	22.36	32.25	15.13	20.00	12.04
25	11.70	25.50	7.21	9.49	15.26	21.09	4.47	23.54	25.32	12.65	14.03	11.18	5.00	11.18	19.23	10.05	16.49	20.81	13.34	20.59	12.04	27.20	25.63	14.32	17.26	0.00	18.38	25.02	19.21	13.93	18.36
26	5.00	11.31	24.70	14.14	15.52	4..12	13.93	13.42	8.54	13.93	19.31	13.89	13.89	19.42	16.00	10.63	3.16	5.00	15.23	17.03	9.43	10.30	17.46	5.00	22.36	18.38	0.00	10.00	10.00	4.47	12.04
27	14.31	6.00	29.83	23.41	15.00	12.20	21.02	6.32	3.00	23.70	29.07	16.15	20.02	28.79	12.80	15.13	9.49	13.60	25.06	13.04	13.15	9.49	10.05	11.18	32.25	25.02	10.00	0.00	13.60	19.10	21.93
28	11.04	23.85	26.4	10.82	23.02	6.71	15.26	22.82	17.03	8.54	13.04	20.10	16.55	14.56	24.51	16.12	11.70	5.83	9.22	25.71	16.00	10.05	26.83	12.16	15.13	19.21	10.00	13.6	0.00	10.05	3.16
29	2.24	15.62	20.25	10.77	13.00	7.81	9.49	14.14	12.04	11.40	16.64	10.44	10.29	16.15	14.56	6.71	3.16	8.06	12.81	15.81	6.08	14.21	17.80	2.24	20	13.93	4.47	19.10	10.05	0.00	11.70
30	12.16	26.4	25.55	9.22	24.33	10.29	14.87	25.24	20.00	6.40	10.20	21.02	19.42	12.08	26.25	17.03	13.89	8.94	6.71	27.51	17.26	13.00	29.15	13.93	12.04	18.36	12.04	21.93	3.16	11.70	0.00
Q		1.59	0.21	1.78	1.29	1.25	1.69	0.17	0.01	0.71	1.15	0.75	0.59	1.03	1.39	0.98	0.73	0.05	1.58	0.54	1.91	0.62	1.32	0.15	1.33	0.71	1.05	0.15	0.93	0.22	0.97

TABLE 2 Simulation Result

iterations	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Zmin/km	209.719	209.817	209.479	207.935	208.397	207.668	209.848	206.826	207.659	205.742	204.99	207.062	208.967	205.833	208.967

5 Conclusions

The optimal route for logistics distribution is effectively solved by logistics route optimization model built in the Paper, and two optimal targets are considered: shortest total distance of logistics distribution and maximum full load rate of vehicle. The partial optimal solution is avoided by ant colony algorithm effectively, and the searching speed of ant colony algorithm is greatly improved by GA with a strong robustness. The decision maker can select the optimal route combining with the current optimizing target according to optimal route set. The example demonstrates that this algorithm is an effective method for logistics route solution and good in the application prospect.

References

[1] Ling W 2001 Intelligent Optimization Algorithms with Application *Tsinghua University Press:* Beijing 154-9 *(in Chinese)*

[2] Dorigo M, Gambardella L M 1997 *Ant colonies for the traveling salesman problem Biosystems* **43**(2) 73-81

[3] Dorigo M, Gambardella L M 1997 *IEEE Transactions on Evolutionary Computation* **1**(1) 53-66.

[4] Abbattista F, Abbattista N, Caponetti L 1995 *Proceedings of the IEEE International Conference on Evolutionary Computation* **2** 668-71

[5] Ding J-L, Chen Z-Q, Yuan Z-Z 2003 *Journal of Computer Researchand Development* **40**(9) 1351-6

[6] Alshamranl A, Mathur K, Ballou R H 2007 *Computers&Operations Research* **34**(2) 595-619

[7] Lee C-Y, Lee Z-J, Lin S-W, Ying K-C 2010 *Applied Intelligence* 32(1) pp 88-95

[8] Wang Y, Lang M 2008 *Logistics Technology* **9** 70-4 *(in Chinese)*

Operation Research and Decision Making

Authors

Fasheng Yi, born in September, 1968, China

Current position, grades: associate professor in the Department Computer Engineering, Chengdu University and University Key Laboratory of Pattern Recognition and Intelligent information Processing, Sichuan Province, China.
University studies: PhD in Information Sciences from University.
Scientific interest: network technology, signal detection and processing.
Publications: 15.

Xiaoling Li, born in June, 1973, China

Current position, grades: Professor in Chengdu University and University Key Laboratory of Pattern Recognition and Intelligent information Processing, Sichuan Province, China.
University studies: MS degree from University electronic science and technology for circuit and system, China.
Scientific interest: image and video feature extraction, signal detection and processing.
Publications: 40.

Jimin Yuan, born in September, 1969, China

Current position, grades: Professor in the Department Computer Engineering, Chengdu Aeronautic Vocational and Technical College, Sichuan Province, China.
University studies: PhD in Information Sciences from University in 2009.
Scientific interest: signal detection and processing.
Publications: 30.

Operation Research and Decision Making

Research on the management of project cost data based on BIM

Tao Yi[1], Yunfei Zhang[2*], Weichun Shen[3]

[1]*Department of technology economics, North China Electric Power University, Beijing, China*

[2]*Department of Management science and Engineering, North China Electric Power University, Beijing, China*

[3]*Department of technology economics, China Electricity Council, Beijing, China*

Abstract

This article first analyses the differences between the BIM application mode and the traditional mode in the cost data management. With the application of BIM (Building Information Modelling), this article analyses the relationship between the standard framework of BIM and the whole process of cost management, then points out that the BIM application mode has changed the whole process of cost management and refines what the BIM standard framework should reflect about the whole process of cost management.

Keywords: BIM, cost data management, construction stage, participants in project

1 Introduction

The application of BIM in Chinese construction market has already started. Corresponding research on standard framework of BIM has also begun [1]. The standardization of BIM mainly involves two aspects, namely the construction field and IT field. The construction field mainly discusses how to implement the benchmark of BIM. Research based on project integration and full life cycle theory of BIM standards are divided into prophase management, planning management, bidding management, etc. [2].

Among them, the core of investment control management is the whole process of cost management; the most important of whole process cost management is cost data management .Construction projects produce cost data, which has complex types, diverse forms and huge quantities from all participants throughout the life cycle.

2 The conception of construction cost data

Construction cost data is a combination of all information related to construction cost features, conditions and variations. All information could be classified into different categories according to management organization, forms, transfer direction, reflection plane, tense and stability. Construction cost data studied in this article is the representation of this kind of information to meet use requirements. At different construction stages, construction cost managers need to collect construction cost data from all parties involved in the project, and put this data into construction cost management after data processing and arrangement [3].

Data of feasibility study stage and design stage is to meet the requirements of deciding investment estimate, design budget estimate, execution budget estimate and working drawing estimate; moreover, data of construction stage is targeted to determine engineering settlement, completion settlement and complement accounting. With respect to price attribute, the former is virtually estimated, but the latter is settled according to facts; from the point of view of data source, the former comes from quota regulations, price information, contract price and so forth, and the latter is confirmed by all participants in the project. From the perspective of the influence on construction cost, the former barely generates actual cost, however, causes more than 70 % of the construction cost. The latter has larger actual cost, but fewer influence on construction cost. In the view of management requirements, the former is widely used in top management than the latter, which is generally used in basic management. From the collection method, specifically speaking, the former is the data collection and arrangement of the latter. For instance, we can analyse and calculate all kinds of construction cost indexes on the basis of construction cost data. The latter comes from practical engineering field and needs contractors' declaration, the supervisory engineer's audit and the owner's confirmation.

Besides data above, some other functional data about confirming cost is involved in construction cost data arrangement, such as quotas and regulations. In addition to technological process and auxiliary-professional knowledge of construction engineering, some technical figures of special equipment and application of new technique are also included. On top of that, there is relevant information got from the market and network

---

Corresponding author e-mail 1342637910@163.com

system, such as the latest price information of various materials and plant on market, comparisons of construction cost data of various materials and plant from different manufacturers, equipment manufacturers and suppliers.

3 Problems of cost data management in traditional mode

In the process of project construction, all related departments' cooperation is necessary for data's transmission. Firstly, information construction at the early stage of project lacks a long-term planning; only the current requirements are considered for a certain project stage [4]. There is not a unified technical standard for engineering data; the inefficient sharing and connecting cannot make cost data be transmitted accurately and fails to meet user's demand. Moreover, the data cannot be reorganized, thus isolated information islands which have a serious impact on cost data management efficiency are formed.

Secondly, because of lacking scientific data accumulation system, lots of construction cost data cannot be collected, neither can the cost database be updated in time. Competition and game relationships among all participants exist in project construction, and the relationships are more intense in settlement process [5]. As a result, the price offered by contractors in bid, project change and claim are lacking scientific and accurate measurement, and the disputes between the owner and contractors must be solved by negotiation and discussion. Such problems have a severe influence on construction progress.

Thirdly, from the perspective of data accumulation, the cost data in current domestic construction market is short of scientific management; most of the data is only used for participants themselves. With the implementation of project bidding system and the specification of bill of quantities issued, Chinese construction market becomes market-oriented. But the construction cost is still depending on government's quota in the process of investment estimate and design estimate. As a matter of fact, the government quota is going to be given up by the market due to its hysteresis [6].

At last, in China engineering cost is in accordance with regional quota valuation method, so it has obvious regional characteristics. What is more, the loss of cost data and management experience caused by the mobility of construction cost personnel is huge. In construction the building materials not only have great consumption but also have many varieties. According to different types and specifications, the decomposition number is more than 500,000. The existing management process is conducted by the cost information published by various areas, which is inefficient and unscientific. Such problem has caused much inconvenience for cost management.

4 BIM's application in cost data management

4.1 THE CONCEPT OF BIM

BIM is short for Building Information Modelling. It is based on three-dimensional digital technology and integrates all kinds of engineering data models of all related information into construction projects. It is a numerical expression of project facilities and functions [7].

The cost data based on BIM has the following characteristics:

1) Objectivity. BIM is a mature information model [8]. It can combine data, process and resources of different stages in a construction project. It is a full description of an engineering object. With a link between the BIM model and project progress, we can have a dynamic comparison of planned cost and practical cost and real-time supervision on quality and safety, and thus to improve the management level of cost, quality and schedule.

2) Integration. The data in BIM model is produced with engineering projects, and is maintained and updated relying on all participants [9]. In addition, it is synchronously simulated in project construction progress until the project comes to an end; it can be organized rapidly and completely, then form a four-dimensional relationship database. As shown in Figure 1, the problems of isolated information islands exposed among departments' management can be solved.
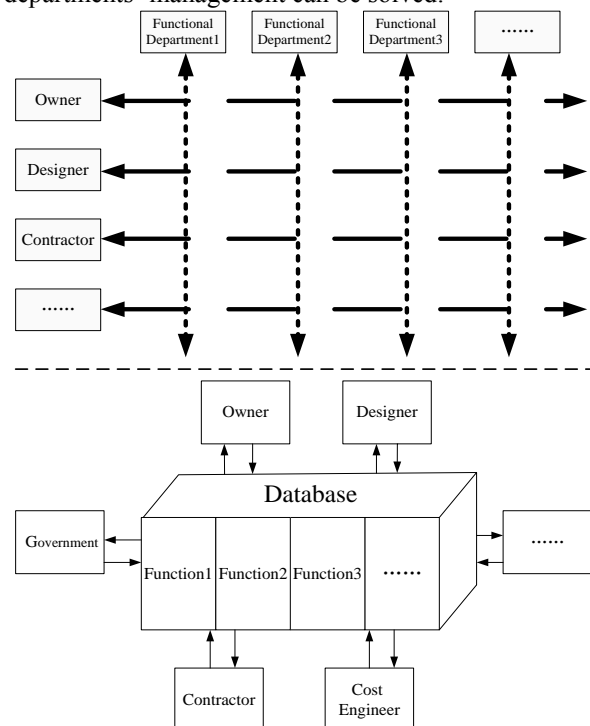

FIGURE 1 Comparison between cost data platform and traditional management mode

3) Accuracy. In BIM database, the data is highly accurate so that it can provide all kinds of data information rapidly for management. It can form the materials and plant list automatically and provide evidence for engineering quantity computation, cost, budget and final

accounts. With the help of BIM technology, the site administrators can compare on-site performance with the planned construction model directly, which can reach a more scientific decision-making level.

4) Real-time. BIM model can reflect construction cost data in a certain time, and provide scientific reference for similar projects' decision-making. At different project stages [10], data can be put in, extracted, updated and shared by different participants, and it has strong function of search, analysis and statistics. This data can also be organized by cost engineers to meet different management requirements.

5) Decision-making. From the perspective of cost management decision, each decision should be based on actual engineering data. BIM model can show the amount of used money, and thus can control money risk and achieve profit target in the short term; it can calculate engineering quantity accurately and rapidly by establishing relevant database, and improve the efficiency and accuracy of construction budget. It can extract project basic information at any time and compare planned consumption with actual consumption, by which we can know project profitability and resource consumption. It can make a dynamic comparison of planned cost and actual cost by combining actual project and BIM model, based on which, we can have a real-time supervision on quality and safety and make adjustments according to the actual situation, then improve the management level of cost, quality and schedule.

These characteristics of cost data based on BIM bring about a change of the cost data management [11]. Compared with the traditional information management model, this kind of management has undergone tremendous changes.

5 Management requirements and construction cost data requirements of all project participants

5.1 GOVERNMENT

The government institution should master and provide information about developing trend of construction cost for society and engineering construction market, and help other participants to make investment decision. The government needs basic data for making construction cost pricing basis and cost management policies and regulations at all stages of project. So the basic data can serve for the government's macroeconomic regulation and control [12].

5.2 OWNER

At feasibility study stage, the owner needs to make project investment estimate, make and examine the project proposal and feasibility study report, make a reasonable determination of project investment to improve benefit, and establish the investment plan. The owner also needs to use the collected construction cost data and the way of

analogy to select out some typical projects, which should be breakdown, conversion and combination. Then the owner should take the factors of plant and materials price changes into account when making investment estimate. At preliminary design stage, the owner makes preliminary design budget estimate. They need to use collected construction cost data to make in-depth and meticulous preliminary design budget estimate when not having the detailed construction drawings. At design accomplishment stage, the owner should make working drawing estimate, then compare it with cost data of similar projects, judge the accuracy, identify causes of deviation and revise construction drawings. Meanwhile, accurate predicting and analysing the unforeseeable factors that will bring cost changes is necessary, as effective control of the working drawing estimate not exceeding the design budget is. At bidding stage, the owner should accurately and reasonably determine contractors' quotation according to accumulative data, and have scientific examination, evaluation and comparison among bids to select out the most suitable contractor. The owner needs to determine the unit price scientifically as well. At construction stage, the owner should get fast, fair and reasonable engineering settlement based on specific work contents. The construction cost data should reflect the actual production level.

5.3 DESIGNER

At preliminary design stage, designers use cost data of typical projects and look for the scientific, economical and rational design scheme to make the engineering technology and economy more effectively integrated and to ensure that the design budget is less than the investment estimate for internal management. Designers should collect and accumulate various types of project cost data, and establish its own database. The required construction cost data is more comprehensive and divided by design characteristics. At design stage, designers should refer to similar projects' design, and reasonably determine project parameters and parameter system design according to the influence of different indexes on cost. At bidding stage, design institute needs make rational design determination. Moreover, design needs to be completed according to the influence of different indexes on cost. At construction stage, design institute should make a reasonable estimate of the cost change in management when design changes occur, as well as scientific economy evaluation.

5.4 CONTRACTOR

Construction contractor should establish its own internal enterprise quota and conduct internal cost accounting for internal management. This kind of construction cost data is professional and divided in detail. At bidding stage, construction contractor should have scientific construction cost data to support scientific quotation. It needs construction cost data from similar projects. At

construction stage, construction contractor needs fast, fair and reasonable engineering settlement, and it should determine specific work contents in strict accordance with the actual situation.

5.5 OPERATION ORGANIZATIONS

Operation organizations should collect engineering data to be convenient for maintenance and management in operation at operational stage. They need various data of complement settlement and engineering information.

5.6 COST ENGINEERS

Cost engineers should organize and take part in measuring work of cost indexes, build their own engineering database system of completed projects, guide industrial development and improve enterprise core competence and technical management. They need to collect various data of different levels, users and purposes, and to accumulate by different use demand.

After the analysis of management requirements and construction cost data requirements of all project participants mentioned above, we need to establish data channel reflected by the standard framework of BIM. Only in this way, can we use the great advantages of the cost data based on BIM

6 Management emphases at different stages for project participants

In order to achieve those goals, the project participants should put the required cost data into BIM model at different stages and get cost data they need, as is showed in Figure 2. Among them, the lines with odd numbers indicate the cost data that a certain participant needs to get from BIM model and the lines with even numbers indicate the cost data that a certain participant should input into BIM model.
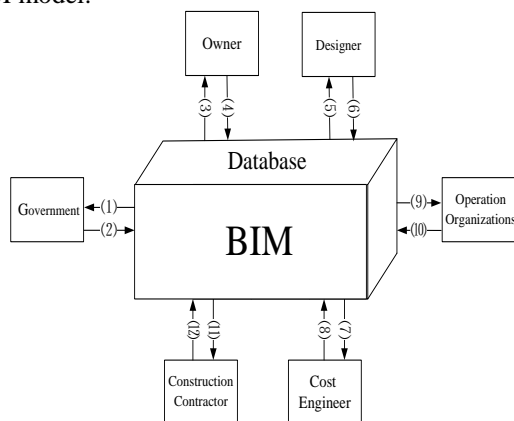
FIGURE 2 Cost data flow of all project participants

1) Construction cost pricing basis and cost developing trend.

2) Policies and regulations, legal provisions.

3) Survey data, data reference of similar projects (feasibility study stage, preliminary design stage, design accomplishment stage); tender documents, data reference of similar projects (bidding stage); project site management documents, settlement supporting information (construction stage).

4) Feasibility study report, approval of budget estimate, working drawing estimate (feasibility study stage, preliminary design stage, design accomplishment stage); tender documents of the successful bidder, contract items (bidding stage); settlement documents (construction stage).

5) Reference of previous construction cost data in database (project preliminary/internal management, design stage, bidding stage); documents of actual project situation (construction stage).

6) Survey report (project preliminary stage/internal management); design budget estimate (design stage); design contract (bidding stage); construction cost data from designers' site management (construction stage).

7) Reference of previous construction cost data in database (project preliminary stage/ internal management, bidding stage); documents of actual project situation (construction stage).

8) Enterprise quota (project preliminary stage/internal management, bidding stage); tender offer, contract price (bidding stage); settlement documents (construction stage).

9) Documents of complement accounting (operational stage).

10) Report of operating situation (operational stage).

11) Construction cost data collected by use demand (internal management).

12) Engineering database system (internal management). The discussion above specifies management emphases for project participants in engineering construction process. What is more, the cost data is distributed among different construction stages according to construction progress. In actual management process, we need to carry out cost data management according to specific stages. For cost data management, timeliness must be guaranteed, thus the effectiveness can be guaranteed.

7 Expectations

At home and abroad, the idea and technology of BIM has been put into practice. However, there are many difficulties and challenges [13]. Besides, BIM itself has some problems occurred during application. When it applies to construction cost data management, some other difficult points include.

Theoretically, construction cost data management eventually need come to the level of industrial management. At the preliminary stage of establishing database, data only could be managed and accumulated as a unit of enterprises or projects. In addition, due to lacking experience and collected data, many troubles may occur.

Some construction cost data, which is involved in the quota level of construction enterprise, is relevant to commercial confidentiality such as internal cost accounting. As a consequence, the push of BIM technology may meet obstacles.

Technically, different participants in the project choose diverse BIM software for different management demand. Also with the heavy workload of data matching, the data formats and levels of data detail are varied. Thus, for one project, it is hard to complete the unified data standard and data collaboration. In addition, project construction duration is long. All parties have different technical levels management levels and application levels of BIM.

As a result, in the project construction, it is by efforts of all participants that the advance of BIM can be expressed in construction cost management process.

References

[1] The BIM research group of Tsinghua University School of software. Research on Chinese Building Information modelling Standard Framework 2009 *Journal of Information Technology in Civil Engineering (in Chinese)*

[2] Feng L, Lu H 2009 Blueprint of construction project management information system based on BIM *Construction Management Modernization* **23**(4) 362-66 *(in Chinese)*

[3] Yi Tao 2002 *Comprehensive knowledge of engineering cost* Beijing China Electric Power Press *(in Chinese)*

[4] Duan Z 2010 *Construction Supply Chain Information Sharing Program Evaluation Research* Xi'an:Xi'an University of Architecture and Technology *(in Chinese)*

[5] Wei J 2004 *Study on the Countermeasures of Asymmetric Information in Project Management* Beijing North China Electric Power University *(in Chinese)*

[6] Hongbin M 2010 *The development direction of architectural engineering quota* www.chinabim.com *(in Chinese)*

[7] Zhang J 2010 Study on Barriers of Implementing BIM in Engineering Design Industry in China *Journal of Engineering Management* **8**(4) 387-92 *(in Chinese)*

[8] He G 2011 *BIM general* Beijing China Architecture & Building Press *(in Chinese)*

[9] Ge W 2013 *The second dimension of BIM—BIM application in different Participants of project* Beijing China Architecture & Building Press *(in Chinese)*

[10] Ge Q 2013 *The first dimension of BIM—BIM application in different stages of project* Beijing China Architecture & Building Press *(in Chinese)*

[11] Zhang S 2012 Research of BIM Application in Project Cost Management *Construction Economy* (2) 20-4 *(in Chinese)*

[12] Shen W, Dong S 2008 Index System and Calculation Model on Engineering Cost *Technology Economics* **27**(10) 62-7 *(in Chinese)*

[13] He Q, Qian L, Duan Y, Li Y 2012 Current Situation and Barriers of BIM Implementation *Journal of Engineering Management* **26**(1) 13-6 *(in Chinese)*

Authors

Tao Yi, born on October 25, 1967, Beijing, China

Current position: associate professor in School of Economics and Management of North China Electric Power University, master instructor of Business Administration of North China Electric Power University.
University studies: Wuhan University.
Scientific interest: technical economics and management, theory and application of project management.

Yunfei Zhang, born on April 18, 1990, Beijing, China

Current position: master in School of Economics and Management of North China Electric Power University.
University studies: Management Science and Engineering in North China Electric Power University.
Scientific interest: management science and engineering, engineering construction management.

Weichun Shen, born on December 5, 1961, Beijing, China

Current position: Vice secretary-general of China Electricity Council, master instructor of Business Administration of North China Electric Power University.
University studies: Engineering project management in Tianjin University.
Scientific interest: engineering project management, engineering economy.

Spatial clustering algorithm with obstacles constraints based on artificial bee colony

Li-ping Sun[1, 2], Yong-long Luo[1, 2*], Xin-tao Ding[2], Fu-long Chen[2]

[1]*College of National Territorial Resources and Tourism, Anhui Normal University, Anhui, 241000, China*

[2]*Engineering Technology Research Center of Network and Information Security, Anhui Normal University, Anhui, 241000, China*

Abstract

Spatial clustering is one of practical data mining technique. In this paper, artificial bee colony (ABC) is used for clustering algorithm, which aims to optimally partition N objects into K clusters in obstacle space. The ABC algorithm used for clustering analysis with obstacles constraints, called The ABC algorithm used for clustering analysis with obstacles constraints ABC-CO, is proposed in the paper. By comparison with the two classic clustering algorithms, k-medoids and COE-CLARANS, demonstrates the rationality and usability of the ABC-CO algorithm.

Keywords: spatial clustering, artificial bee colony, obstructed distance; fitness calculation

1 Introduction

Spatial clustering is the organization of geographical data set into homogenous groups, the aim of which is to group spatial data points into clusters [1-3]. Most spatial clustering algorithms apply Euclidean distance between two sample points to measure the proximity of spatial points. However, physical obstacles (e.g. rivers and highways) often exist in real applications, which can hinder straight reachability among sample points. As a result, the clustering results, which utilize Euclidean distance measure are often unreasonable. Taking the simulated dataset in Figure 1a as an example, where the points represent the location of consumers. The clustering result shown in Figure 1b can be obtained, when the rivers and hill as obstacles are not considered. Obviously, the result is not realistic. If the obstacles are taken into account, the clustering result in Figure 1c can be obtained.

a) Spatial dataset with obstacles   b) Spatial clustering result ignoring obstacles   c) Spatial clustering result considering obstacles
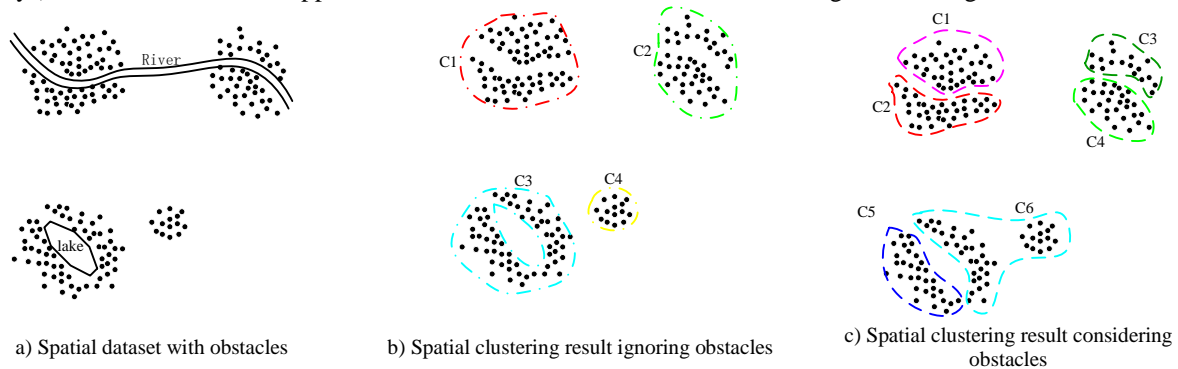
FIGURE 1 Spatial clustering in the presence of obstacles

At present, there are a few algorithms considering obstacles constraints in the spatial clustering process [4-9]. There generally exists the shortcomings, including low robustness and easy to fall into local optimum. Many heuristic clustering algorithms have been introduced to overcome local optima problem, such as evolutionary algorithms [10, 11], swarm intelligence algorithms [12, 13] and so on [14, 15]. Artificial bee colony (ABC) algorithm, which simulates the intelligent foraging behaviour of a honey bee swarm, is a novel category of heuristic algorithms. In this paper, ABC optimization algorithm for optimization problems, which is proposed by Karaboga [16], is applied to spatial clustering analysis. Fathian and et al. proposed the application of honeybee mating optimization in clustering [17]. Zhang and et al. presented a novel artificial bee colony approach for clustering which was compared with other heuristic algorithms such as genetic algorithm, ant colony, simulated annealing and tabu search [18]. Yan and et al. designed a hybrid artificial bee colony (HABC) algorithm for data clustering [19]. Karaboga and Ozturk applied

artificial bee colony (ABC) optimization algorithm to classification benchmark problems [20].

In this paper, ABC algorithm is extended for clustering analysis with obstacles constraints. The performance of the algorithm has been tested on a variety of simulated data sets provided from real-life situations and compared with classic clustering algorithms. The remainder of this paper can be structured as follows. Section 2 gives the related definitions and elaborates the detailed of ABC-CO algorithm. Section 3 shows the experimental results. Section 4 presents the conclusions and main findings.

2 The clustering analysis with obstacles constraints

Let $E = \{E_1, E_2, …, E_i,…, E_n\}$ represents a collection of spatial entities with autocorrelation. Let $(e_{i1}, e_{i2}, ..., e_{ij}, ..., e_{im})$ represents feature vector of a spatial entity E_i. Divide E into k clusters, denoted $E = \{C_1, C_2, ..., C_i, ..., C_k\}$, $C_i = \{ E_1, E_2, ..., E_t\}$. $S(E_{ai}, E_{bj})$ represents the similarity of the i-th entity in the a-th cluster and j-th entity in the b-th cluster. The spatial clustering result satisfies following conditions:

1) $\bigcup_{i=1}^{k} C_i = E$;

2) For $\forall C_a, C_b \subseteq E, a \neq b$, the following conditions need to be satisfied simultaneously:

1) $C_a \cap C_b = \varphi$;

2) $MAX_{\forall E_{ai} \in C_a, \forall E_{bj} \in C_b} (S(E_{ai}, E_{bj})) <$

$MIN_{\forall E_{ai} \in C_a, \forall E_{bj} \in C_b} (S(E_{ai}, E_{bj}))$.

In this paper, we mainly focus on clustering analysis with obstacles constraints. The obstacles have two typical classes: convex obstacle and concave obstacle. Relevant definitions are defined as follows:

Definition 1 (Polygon obstacles). Let O be the set of m non-intersecting obstacles $\{o_1, o_2, …, o_m\}$. Each obstacle o_i is represented by a simple polygon, denoted as $G (V, E)$, where $V = (v_1, v_2, …, v_n)$ is the set of vertices of the polygon, $E= \{(v_k, v_{(k+1) \bmod n}) \mid k = 1, …n\}$, n is the number of the vertices.

Definition 2 (Reachability between two points). For $\forall$ p, q in a two-dimensional space, if segment pq does not intersect with any obstacle, p is called directly reachable from q; Otherwise, p is called indirectly reachable from q.

Definition 3 (Obstructed distance). Given point p and q, $d_o(p,q)$ represents the obstructed distance between two sample points which represent represents the entities. If p is directly reachable from q, $d_o(p,q)$ is Euclidean distance between p and q, denotes as $d(p,q) = \sqrt{(p_x - q_x)^2 + (p_y - q_y)^2}$. If p is indirectly reachable from q, $d_o(p,q)$ is the length of the path which is configured to

bypass the obstacles while p, q respectively is taken as the start and end point.

The degree of similarity between the objects can be identified by some criterions, such as the total within-cluster variance or the total mean-square quantization error (MSE) [21], which is defined as follows:

$$Perf(O,C) = \sum_{i=1}^{n} MIN\{\|o_i - c_j\|^2, j = 1, 2, ..., k\},\qquad (1)$$

where $\|o_i - c_j\|$ denotes the similarity between the j-th centre and the i-th object. The most used similarity metric in clustering procedure is the distance metric. In this paper, the obstructed distance described in Definition 3 is used as a distance metric. The following algorithm is the calculation of the obstructed distance between objects follows:

Algorithm 1: Calculation of obstructed distance
1: **If** (p is directly reachable from q)
2:          $d_o(p,q) = d(p,q)$
3:          return
4: else
5:          shape = polygon_shape($G (V, E)$) /*Judge the shape of the obstacles*/
6:          go 8
7: **End if**
8: **If** (shape = convex)
9:          Construct_visibility_graph(p, $G (V, E)$, q) /*Constuct the visibility graph with obstacles by the method in Literature [22] */
10:          $d_o(p,q)$ = Dijkstra_Calculate_distance (p,q) /*Apply Dijkstra algorithm [23] which is the classic model to calculate the shortest distance of $d_o(p,q)$ based on the visibility graph achieved by the previous operation.*/
11: **else** /*Means the obstacle shape of the obstacles is concave.*/
12:          Eliminate the concave points from V, denoted V' /*If two points is obstructed by concave obstacles, the concave points should not be linked */
13:          Construct_visibility_graph(p, $G (V', E)$, q)
14:          $d_o(p,q)$ = Dijkstra_Calculate_ distance (p,q)
15: **End if**

3 Application of artificial bee colony algorithm in clustering

3.1 HONEY BEE MODELLING

Artificial bee colony (ABC) algorithm was presented by Karaboga [16]. In ABC algorithm, the colony of artificial bees consists of three essential categories of bees: employed bees, onlookers and scouts. Half of the colony consists of the employed bees and the other half contains the onlookers. Corresponding to each food source, there is one employed bee. The food source which is abandoned by the bees is replaced with a new food source by the scout bees. A food source represents a possible solution to the combinatorial optimization problem and the nectar amount of a food source corresponds to the quality (fitness) of associated solution. Pseudo-code of the ABC algorithm can be elaborated as follows:

Algorithm 2: The framework of the classic ABC algorithm.

1: Initialize the population of feasible solutions and evaluate the fitness of the solutions.
2: Generate new solutions and calculate the fitness of the solutions.
3: Evaluate the probabilities of preferable solutions.
4: Choose solutions depending on probabilities and generate new solutions.
5: Replace the abandoned solutions with new solutions.
6: Memorize the best solution so far.

In the initialization phase, the ABC algorithm generates a randomly distributed initial population of N solutions, where N denotes the size of population. Each solution $X_i = (x_{i1}, x_{i2}, \ldots, x_{iD})$ can be generated as follows:

$$x_{ij} = x_j^{\min} + (x_j^{\max} \quad x_j^{\min}) rand(0,1), \tag{2}$$

where $i = 1, 2, \ldots, N, j = 1, 2, \ldots, D$. D is the dimension of the optimization problem. $x_j^{\min}$ and $x_j^{\max}$ are lower and upper bounds of the j-th parameter. The fitness of solutions will be calculated by Equation (3).

$$fit_i = \frac{1}{1+f_i}, \tag{3}$$

where f_i is the objective function value of the solution x_i. In the employed bee' phase, each employed bee is sent to the food source in its memory and finds a neighbouring food source $v_i = (v_{i1}, v_{i2}, \ldots, v_{iD})$, which is generated by Equation (4) as followed:

$$v_{i,j} = x_{i,j} + \varphi_{i,j}(x_{i,j}, x_{k,j}), \tag{4}$$

where k is a randomly selected dimension different from i, j. $\phi_{i,j} \in [-1, 1]$ is a random number.

In the onlooker bees' phase, an onlooker selects a food source depending on the probability value of the food source, which is calculated by the following Equation:

$$p_i = \frac{fit_i}{\sum_{j=1}^{N} fit_j}. \tag{5}$$

3.2 THE ABC ALGORITHM USED FOR CLUSTERING ANALYSIS

In this paper, the ABC-CO algorithm is proposed for clustering analysis with obstacles constraints. In a clustering problem, a set of n objects need to be divided into k clusters. Let $V = \{v_1, v_2, \ldots, v_n\}$ be the set of n objects. Each object has m characters. The character data matrix is denoted $D_{n \times m}$. The i-th row of the profile data matrix $D_{n \times m}$ presents an object v_i. Let $C = \{C_1, C_2, \ldots, C_k\}$ be the set of k clusters, each candidate solution of the population consists of m times k cells c_{ij} ($i \in \{1, \ldots, k\}$, $j \in \{1, \ldots, m\}$). In order to apply the ABC algorithm to solve clustering algorithm, floating point arrays are utilized to solve cluster centres. Each food source presents a set of cluster centre, seen in Equation (6). A food source

can be decoded to a cluster centre following the Equation (7).

$$X_i = \{x_1, x_2, \ldots, x_m, x_{m+1}, \ldots, x_{n \times m}\}, \tag{6}$$

$$C_i = \{x_{(i-1) \times m+1}, x_{(i-1) \times m+2}, \ldots, x_{i \times m}\}, \tag{7}$$
$$(i = 1, 2, \ldots, k)$$

where X_i denotes a food source in ABC-CO algorithm, k is the number of clusters and m is the number of characters of the clustering data.

The total within-cluster variance in Equation (1) is used to evaluate the quality of clusters partition for the ABC-CO algorithm. The pseudo-code of the fitness calculation of ABC-CO algorithm is shown in Algorithm 3.

Algorithm 3: Fitness calculation of a solution.

1: Perform Algorithm 1 to compute the obstructed distance between all objects in V and each cluster centre.
2: Assign objects to the nearest centres.
3: Calculate the total within-cluster variance t_i following Equation (1).
4: fit$_i = t_i$

According to the explanation of classic ABC algorithm and the adjustments for clustering analysis with obstacles constraints, the pseudo-code of the ABC-CO algorithm is shown in Algorithm 4.

Algorithm 4: The ABC algorithm used for clustering analysis with obstacles constraints (ABC-CO).

1: Set the population size N, the control parameter *limit* and the maximum cycle number *MCN*.
2: Randomly generate the initial population of solutions, which can be decoded to the cluster centres following Equation (7).
3: **For** (each food source C_i)
4: Perform Algorithm 3 to calculate the fitness of C_i
5: **End for**
6: Set $trial_i = 0$, where i =1, 2 , …, N and *iteration* = 1
7: **While** (iteration $\leq MCN$)
8: **For** (each food source C_i)
9: Generate a new solution v_i following the Equation(7);
10: Perform Algorithm 3 to calculate the fitness of v_i
11: **If** (the fitness of v_i is better than or equal to the fitness of C_i) then $trial_i = 0$, otherwise $trial_i = trial_i + 1$
12: **End for**
13: Evaluate the probability p_i for each solution following the Equation (5);
14: Set I=1 and the current onlooker bee number $ln = 0$
15: **While** ($ln < N$)
16: **If** (rand() $< p_i$)
17: Generate a new solution v_j following the Equation (7)
18: Perform Algorithm 3 to calculate the fitness of v_j
19: Apply greedy selection process on the original solution and the new solution
20: $ln = ln + 1$
21: **End if**
22: Set $i = i + 1$and if $i > N$ then i =1
23: **End** while
24: **For** (each food source C_i)
25: **If** ($trial_i \geq limit$)
26: Replace the abandoned solution with a new solution which is randomly generated by Equation (7)
27: **End if**
28: End **for**
29: Memorize the best solution so far
30: Set iteration = *iteration* +1
31: **End** while

4 Results and discussion

This paper presents two sets of experiments to prove the effectiveness of the ABC-CO algorithm: The first experiment uses a set of analogue data, which are generated by the simulation of ArcGIS 9.3. Experimental results are compared with k-medoids clustering algorithm [2]. The second experiment carries on spatial dataset, and compares the results with the COE-CLARANS algorithm [8]. All algorithms are implemented in C# language and executed on a Pentium 4.3HZ, 2GB RAM computers. The population size N=40, the maximum cycle number MCN=2000 and the parameter $limit$=100.

The classic k-medoids clustering algorithm has been widely used for its simplicity and feasibility. ABC-CO algorithm uses obstacle distance defined in this paper for clustering analysis, and k-medoids algorithm uses Euclidean distance as similarity measure of samples.

Simulated dataset of the first experiment are shown in Figure 2a. When cluster number k=6, the clustering results of k-medoids clustering algorithm and ABC-CO algorithm are shown in Figures 2b and 2c, respectively. Experimental results show that the clustering results of the ABC-CO algorithm considering obstacles are more efficient than k-medoids algorithm. This paper presents two sets of experiments to prove the effectiveness of the ABC-CO algorithm: The first experiment uses a set of analog data, which are generated by the simulation of ArcGIS 9.3. Experimental results are compared with k-medoids clustering algorithm [2]. The second experiment carries on spatial dataset, and compares the results with the COE-CLARANS algorithm [8]. All algorithms are implemented in C# language and executed on a Pentium 4.3HZ, 2GB RAM computers. the population size N=40, the maximum cycle number MCN=2000 and the parameter $limit$=100.

a) Simulated dataset | b) Clustering results of k-medoids algorithm | c) Clustering results of ABC-CO algorithm

FIGURE 2 Clustering results for analog data
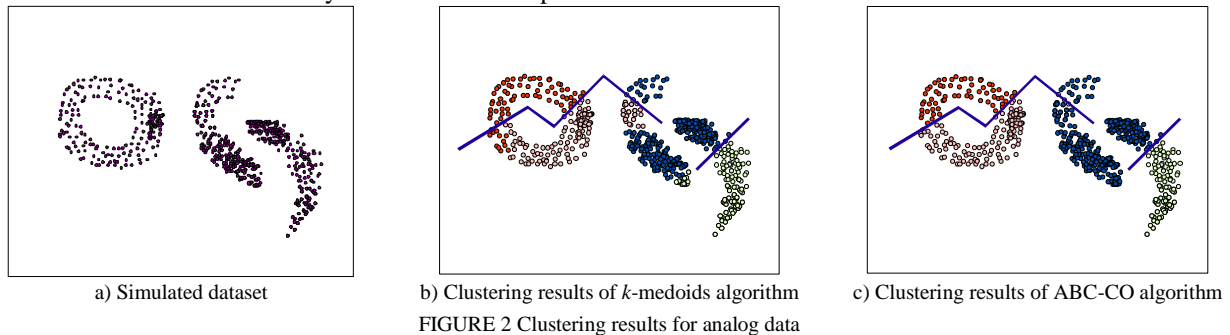
4.1 SIMULATION EXPERIMENT

The classic k-medoids clustering algorithm has been widely used for its simplicity and feasibility. ABC-CO algorithm uses obstacle distance defined in this paper for clustering analysis, and k-medoids algorithm uses Euclidean distance as similarity measure of samples. Simulated dataset of the first experiment are shown in Figure 2a. When cluster number k=6, the clustering results of k-medoids clustering algorithm and ABC-CO algorithm are shown in Figures 2b and 2c, respectively. Experimental results show that the clustering results of the ABC-CO algorithm considering obstacles are more efficient than k-medoids algorithm.

4.2 CLUSTERING ALGORITHM APPLICATION AND CONTRASTIVE ANALYSIS

This paper takes resident communities as cluster points, where the points are represented as (x, y). The hills, rivers and lakes in the territory are spatial obstacles which are described in Definition 1, denoted as G (V, E). Based on above real world datasets, the COE-CLARANS algorithm and the ABC-CO algorithm are compared by simulation experiment. The results are shown in Figure 3:

a) 5 clusters (COE-CLARANS algorithm) | b) 5 clusters (ABC-CO algorithm)

c) 10 clusters (COE-CLARANS algorithm) | d) 10 clusters (ABC-CO algorithm)
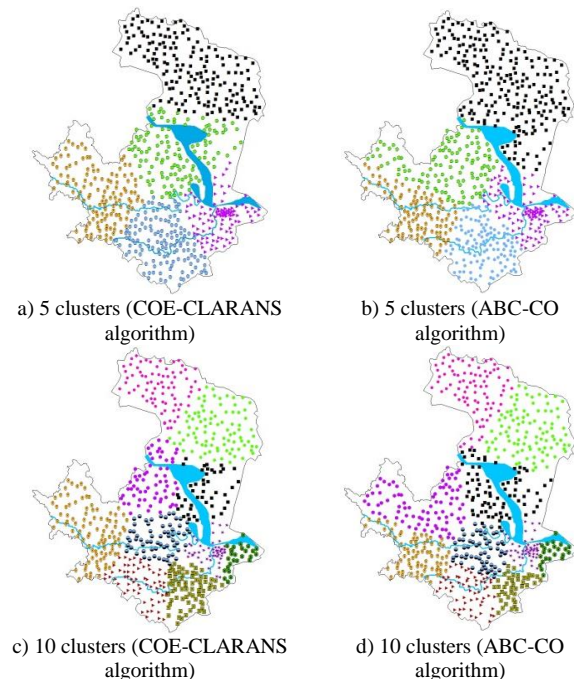
FIGURE 3 Comparison of clustering analysis using COE-CLARANS algorithm and ABC-CO algorithm

In view of the range of radiation area of different type public facility, the paper has carried on the cluster simulation deferred to 5 subclasses and 10 subclasses separately. For Yangtze River is the main obstacle of Wuhu territory, the result of its peripheral region clustering

analysis can demonstrate the validity of the algorithm. As can be seen from the clustering results, the subclasses that the COE-CLARANS algorithm obtains are more centralized. And the community intensive places are divided by obstacles in several subclasses. The actual distance between a cluster sample point and centre, which do not belong to the same side of an obstacle is much farther than the straight line distance between them. The ABC-CO algorithm has solved this problem well. The results of the algorithm demonstrate subclasses of clustering divided by physical obstacles is quite few, which divide at the community sparse place.

5 Conclusions

Spatial clustering analysis is an important method of in the data mining community. Traditional clustering methods use a variety of straight-line distance metrics to measure the degree of similarity between spatial entities. In this paper, an artificial bee colony algorithm is extended to solve clustering problems with obstacles constraints, which is inspired by the bees' forage behaviour. This algorithm was implemented and tested on several real spatial datasets. By comparison with the classic clustering algorithms, it verifies the rationality and usability of the ABC-CO algorithm.

Acknowledgements

References

[1] Oliveira D, Jr J 2011 *Advanced Engineering Informatics* **25** 380-9
[2] Wang Z, Soh Y C, Song Q, Sim K 2009 *Pattern Recognition* **42** 2029-44
[3] Pennerstorfer D, Weiss C 2013 *Regional Science and Urban Economics* **43** 661-75
[4] Tung A, Hou J, Han J 2000 *Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining* 165-8
[5] Ng R, Han J 1994 *Proceedings of the 20th International Conference on Very Large Data Bases* 144-55
[6] Zaïane O R, Lee C H 2002 *Proceedings of the IEEE International Conference on Data Mining* 737-40
[7] Ester M, Kriegel H, Sander J 1996 *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining* 226-31
[8] Wang X, Rostoker C, Hamilton H 2012 *Journal of Intelligent Information Systems* **38** 269-97
[9] Estivill-Castro V, Lee I 2004 *Journal of Photogrammetry and Remote Sensing* **59** 21-34
[10] Jiang B, Wang N, Wang L P 2013 *Communications in Nonlinear Science and Numerical Simulation* **18**(3)134-45
[11] Cura T 2012 *Expert Systems with Applications* **39** 1582-8

[12] Bereta M, Burczyński T 2009 *Information Sciences* **179** 1407-25
[13] Kuo R J, Lin L M 2010 *Decision Support Systems* **49** 45-62
[14] Gou S P, Zhuang X, Li Y Y, Xu C, Jiao L C 2013 *Neurocomputing* **4** 275-89
[15] Liu R C, Zhang X G, Yang N, Lei Q F, Jiao L C 2012 *Applied Soft Computing* **12** 302-12
[16] Karaboga D 2005 Technical Report-TR06 *Erciyes University Engineering Faculty*
[17] Fathin M, Amiri B, Maroosi A 2007 *Applied Mathematics and Computation* **190** 1502-13
[18] Zhang C S, Ouyang D T, Ning J X 2010 *Expert Systems with Applications* **37** 4761-67
[19] Yan X H, Zhu Y L, Zou W P 2012 *Neurocomputing* **97** 241-50
[20] Karaboga D, Ozturk C 2011 *Applied Soft Computing* **11** 652-7
[21] Gűngőr Z, Űnler A 2007 *Applied Mathematics and Computation* **184(2)** 199-209
[22] Lozano-Perez T, Wesley M 1979 *Communications of the ACM* **22** 436-50
[23] Marianov V, Revelle C 1996 *European Journal of Operational Research* **93** 110-20

Authors

Li-ping Sun, born in June, 1980, Anhui, China

Current position: associate professor in Anhui Normal University.
University studies: Computer Software and Theory in Chongqing University.
Scientific interest: computer science, computer modelling and spatial data mining.

Yong-long Luo, born in April, 1972, Anhui, China

Current position: professor in Anhui Normal University.
University studies: Computer Science and Technology in University of Science and Technology of China.
Scientific interest: computer modelling, network security, applied mathematics and spatial data mining.

Xin-tao Ding, born in December, 1979, Anhui, China

Current position: lector in Anhui Normal University.
University studies: Computational Mathematics in East China Normal University.
Scientific interest: computer science, computer modelling and spatial data mining.
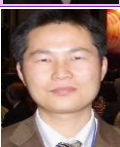
Fu-long Chen, born in May, 1978, Anhui, China

Current position: associate professor in Anhui Normal University.
University studies: Computer Science and Technology in Northwestern Polytechnical University.
Scientific interest: computer science, computer modelling and spatial data mining.

Use of Petri Nets for maximum power point tracking in photovoltaic power generation system

Ning Lu[1], Leilei Yi[2*]

[1]*School of Automation, Wuhan University of Technology, Wuhan, China*

[2]*State Grid Suizhou Power Supply Company, Suizhou, China*

Abstract

Maximum power-point tracking is used to take full use of solar energy. The main object of this paper was to find a simple efficient technique to let photovoltaic power generation system working near the maximum power point of the solar arrays. The way of finding the suitable method of detection of the maximum power point and the strategy, which forces the system to work near this point was discussed. This project proposes a new way to realize the incremental conductance method by using high level Petri nets which is capable of tracking global maximum power point under condition change. This model will be evaluated by using stateflow in MATLAB. The result showed that this model is effective.

Keywords: photovoltaic power generation, maximum power-point tracking, incremental conductance, high level, Petri nets

1 Introduction

In the recent years power generation systems have had great revolutionary changes in terms of the integration of Renewable Energy Sources (RES) as alternatives to reduce the emissions and new intelligent control algorithms, which increase the efficiency in the grid are provide at the same time. The Smart Grid concept and the use of power electronics has been extended to Microgrids (MG), providing reliability, protection and efficiency with the grid interconnection. Control applications in this area have become more relevant due to the necessity to achieve minimum requirements, such as frequency voltage regulation, coordination and protection system [1-3].

In addition, the impact of Distributed Power Systems (DPS) on the Smart Grid requires the use of distributed processes control architectures according to the characteristics of Distributed Generation (DG) like the multiple conversion stages [4-6]. Therefore, several control techniques have been proposed in order to provide some smartness to the grid which is also needed in self-autonomous behaviour [7, 8]. Petri nets are well known to be good at modelling discrete-even systems which given an approach to multi-agent systems and used in distributed control problems [9, 10].

Photovoltaic (PV) Generation System performance depends on solar irradiation and operating temperature. Technique is use to track the maximum possible power from the PV array. Algorithm that included in charge controllers is used for extracting maximum available power from PV module under different conditions.

The voltage, at which PV module can reach maximum power, is called 'maximum power point' (or peak power voltage). The present study aims to evaluate performance and enhance accuracy for maximum power-point tracking. The proposed method is established model, which used a high level Petri nets to realize the incremental conductance method. The performance of proposed models shows that this algorithm tracks the maximum power solar energy effectively.

2 High level Petri nets

2.1 PETRI NETS

The control network of a described system by High Level Petri nets (HLPN) is an extension of Ordinary Petri nets (OPN). OPN are graphic and mathematic formalism. It can represent and describe systems in which concurrence and parallelism concepts are present. An OPN is a tuple. In which, let R={P, T, Pre, Post, m0}, where P is a non-empty set of places, T is a non-empty set of transitions, Pre is predecessor application and Post is a successor application of m0. Places of Petri nets can be marked by tokens; the application of tokens in the places of Petri net indicates the marking of the network. The graph is constructed with two types of nodes: the places and transitions. The rule is that two places cannot be directly linked and two transitions neither. High-level Petri nets (HLPN) are improved in the base of OPN, which are a widely used in modelling and specification language for information system behaviour. Chang et al. used a Behavioural Browsing model based on HLPN, which was made for modelling and generating behavioural pattern of students [11]. Object oriented course modelling based on High level Petri net (HLPN) was proposed by Su et al. [12]. Liu et al. proposed an

---
***Corresponding author** e-mail: susanln@163.com

approach based on Petri net for controlling learning path among learning activities [13].

2.2 COLOURED PETRI NETS

In the following years later extensions of Petri net such as Coloured Petri net and timed Petri net are proposed and both of them are called High level Petri net (HLPN). The following will give the description of Coloured Petri nets (CPN) [14].

In Coloured Petri net a value so called colour is allocated to each token. The colour of input tokens which is used by transitions to determine the colour of output tokens. Relation between input tokens and output tokens is presented by colour consumption function or arc function which is depicted by H. For example in Figure 2, E functions are described as Hf(p1, t1) = a, Hf(t1, p2) = b. If a transition is fired, coloured tokens removed from input places and new colour tokens are produced and placed in output places. In Figure 1a, place p1 is marked with two coloured tokens a, b from a colour set {a, b, c}, denoted m(p1) = {a, b}. t1 is enabled as its input arc only requires one coloured token which is available in place p1. If fired, token a is removed from p1, two new coloured tokens b and c are generated by t1's output arcs and deposited in p2 and p3, respectively, as shown in Figure 1b.

1a Before t1 fires   1b After t1 fires
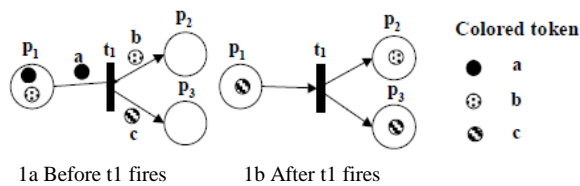
FIGURE 1 Colour Petri nets

In this work it is proposed to use Coloured Petri net approach.

3 Maximum power point tracking

Maximum power point tracking (MPPT) is a technique that grid connected inverters, solar battery chargers and similar devices use to get the maximum possible power from one or more photovoltaic devices, typically solar panels [15]. Solar cells have a complex relationship between solar irradiation, temperature and total resistance that produces a non-linear output efficiency which can be analyzed based on the I-V curve (which is shown in Figure 2) and P-V curve(which is shown in Figure 3).

The purpose of the MPPT system is to sample the output of the Photovoltaic cells and apply the proper resistance (load) to obtain maximum power for any given environmental conditions. MPPT devices are typically integrated into an electric power converter system that provides voltage or current conversion, filtering, and regulation for driving various loads, including power grids, batteries, or motors.
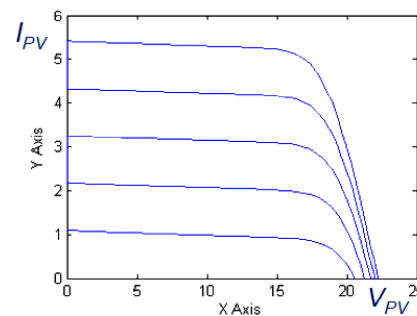
FIGURE 2 I-V Curves
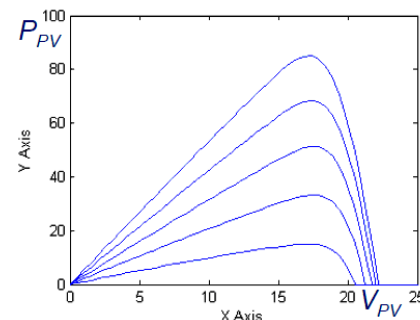
FIGURE 3 P-V Curves

Photovoltaic cells have a complex relationship between their operating environment and the maximum power they can produce. The fill factor (FF), is a parameter, which characterizes the non-linear electrical behaviour of the PV cell. It is defined as the ratio of the maximum power from the solar cell to the product of Open Circuit Voltage Voc and Short-Circuit Current Isc. It is often used to estimate the maximum power that a cell can provide with an optimal load under given conditions. The function to describe the relationship is that P=FF*Voc*Isc. For most purposes, FF, Voc, and Isc are enough information to give a useful approximate model of the electrical behaviour of a photovoltaic cell under typical conditions. For any given set of operational conditions, PV cells have a single operating point where the values of the current (I) and Voltage (V) of the cell result in a maximum power output. These values correspond to a particular load resistance, which is equal to V/I as specified by Ohm's Law. The power P is given by P=V*I. A photovoltaic cell, for the majority of its useful curve, acts as a constant current source [16]. However, at a photovoltaic cell's MPP region, its curve has an approximately inverse exponential relationship between current and voltage. From basic circuit theory, the power delivered from or to a device is optimized where the derivative dI/dV of the I-V curve is equal and opposite the I/V ratio (where dP/dV=0) [17]. This is known as the maximum power point (MPP) and corresponds to the "knee" of the curve.

Incremental conductance is used in this paper for MPPT. In the incremental conductance method, the controller measures incremental changes in array current and voltage to predict the effect of a voltage change. This method can keep up with changing conditions more rapidly than the perturb and observe method (P&O). This method utilizes the incremental conductance (dI/dV) of the

photovoltaic array to compute the sign of the change in power with respect to voltage (dP/dV) [18]. The incremental conductance method computes the maximum power point by comparison of the incremental conductance ($\Delta I/\Delta V$) to the array conductance (I/V). When these two are the same (I/V=$\Delta I/\Delta V$), the output voltage is the MPP voltage and it reach the max power output. The controller maintains this voltage until the irradiation changes and the process is repeated. Figure 4 shows the process of incremental conductance method.
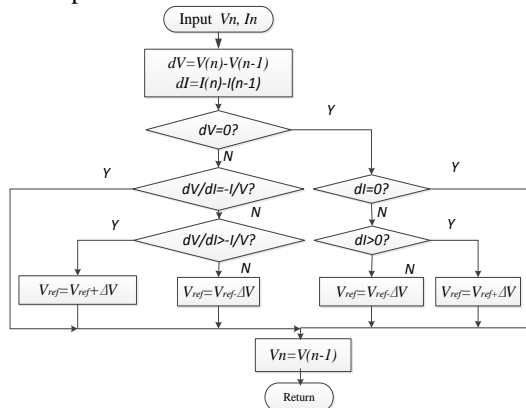
FIGURE 4 Incremental conductance method

4 Case study

Taking a photovoltaic cell as an example, we make test of maximum power point tracking with Coloured Petri net.

Simulation example: 1 PV cell with full sun (1,000 W/m2)

Tables 1 and 2 show the definition of places and transitions of this Coloured Petri net for MPPT.

TABLE 1 Definition of places and transition of the CPN

Places and transition	Description		
	Represent for	Token=Color1	Token=Colour 2	Token=Colour 3
P0	dV	=0	$\neq 0$	NONE
P2	dI/dV	>-I/V	< I/V	= I/V
P11	dI	>0	<0	NONE

TABLE 2 Definition of places and transition of the CPN

Places and transition	Description
	Represent for
P1	V=V(p=pMAX)
T3	dI$\neq$0
T4	dI =0
T5	dV$\neq$0
T6	Vref+$\triangle$V
T7	Vref-$\triangle$V
T8	dI/dV=I/V

According to the task of MPPT and the incremental conductance model, which is showed as Figure 4, Figure 5 described the scheme of MPPT task based on Petri nets.

Four places p0, p2, p11, p1 is used to describe the state of incremental conductance parameters of dV, dI/dV, dI, and the MPP voltage. The colours represent its properties.

T3, T4, T5, T6, T7, T8 is the condition for transition.
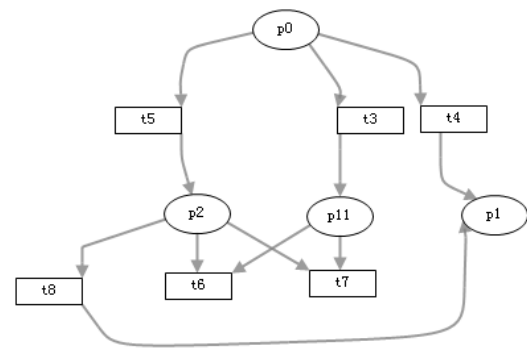
FIGURE 5 Maximum power point tracking scheme based on Petri nets

Figure 5 showed the Maximum power point tracking scheme based on Petri nets.

The simulation used Matlab simulink to establish the photovoltaic power generation system and stateflow to build the Petri net.

This system firstly detected the input voltage and current, and then the output of MPPT module is supplied to the PWM module. It adjusts the duty ratio (FF), and then adjusts the voltage to make the system working in the maximum power point.

The result is shown as Figures 6-9.
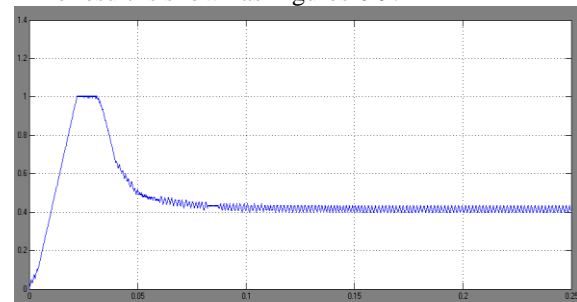
FIGURE 6 Output of MPPT
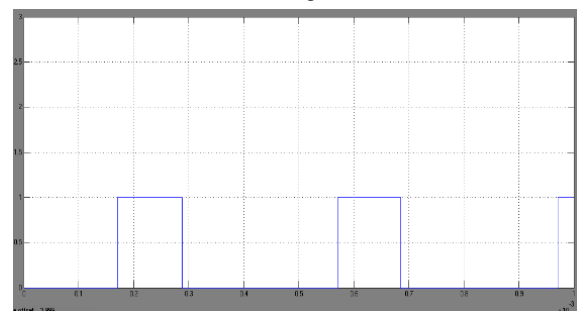
FIGURE 7 Output of PWM
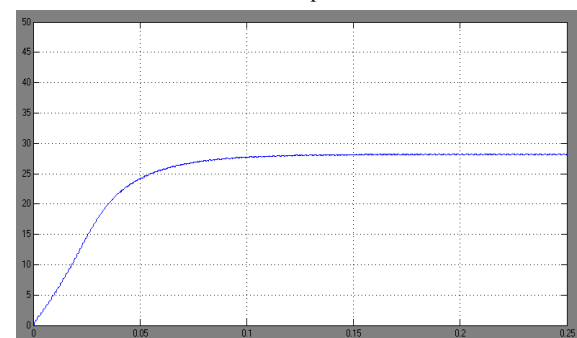
FIGURE 8 Voltage output curve
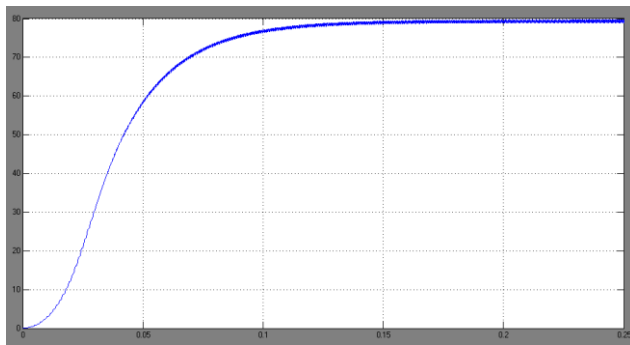
FIGURE 9 Power output curve

The input voltage of approximately stable at 18V, after DC-DC circuit, the output voltage is stable at around 28V as shown in Figure 8. Output power reaches the maximum value at about 0.15s. It is stable at around 80W. The simulation is successful.

6 Conclusions

Controllers with MPPT function increases the utilization efficiency of the solar panel and improves the net system maximum efficiency. On the basis of the analysis of the equal circuit for the solar cell, the paper provide the maximum power point tracking (MPPT) methods, and design a Petri net based on incremental conductance which controls the maximum power point. Trough testing we can track the maximum power point successfully. The experiment proves the improving of stability and reliability of the photovoltaic power generation system.

Acknowledgments

References

[1] Pecas Lopes J A, Moreira C L, Madureira A G 2005 Defining control strategies for analyzing microgrids islanded operation *In Proceedings of Power Tech Russia*

[2] Qin L, Peng F Z, Balaguer I J 2009 Islanding Control of DG in Microgrids *Power Electronics and Motion Control Conference*

[3] Chamorro H R, Jimenez J F 2012 Use of Petri nets for load sharing control in distributed generation applications *In Proceedings of 3rd IEEE International Symposium on Power Electronics for Distributed Generation Systems (PEDG)* 731-6

[4] Luo S, Batarseh I 2006 A review of distributed power systems. Part II. High Frequency AC Distributed Power Systems *Aerospace and Electronic Systems Magazine* **21**(6) 5-14

[5] Simoes M 2006 Intelligent based hierarchical control power electronics for distributed generation systems *Power Engineering Society General Meeting*

[6] Peng FZ, Wei-Li Y, Tolbert LM 2009 Control and protection of power electronics interfaced distributed generation systems in a customer-driven microgrid *Power & Energy Society General Meeting*

[7] Lasseter R H 2002 Micro grids *in IEEE Power Engineering Society Winter Meeting*

[8] Jeon J H, Kim J Y, Cho C H, Nam K Y, Kim J M 2009 Real time digital simulator based test system for micro grid management system *in Transmission & Distribution Conference & Exposition: Asia and Pacific* 1-4

[9] Lukomski R, Wilkosz K 2010 Modeling of multi-agent system for power system topology verification with use of Petri nets

[10] Abouaissa H, Nicolas JC, Benasser A, Czesnalowicz E 2002 Formal specification of multi-agent systems: approach based on meta-models and high-level Petri nets - Case study of a Transportation System *IEEE International Conference on Systems Man and Cybernetics* **5** 6-9

[11] Chang Y C, Huang Y C, Chu C P 2009 B2 Model: A Browsing Behavior Model Based on High-Level Petri nets to Generate Behavioral Patterns for E-Learning *Expert Systems with Applications* **36**(10) 2009 12423-40

[12][12] Su J M, Tseng S S, Chen C Y, Weng J F, Tsai W N 2006 Constructing SCORM compliant course based on high-level Petri nets *Computer Standards & Interfaces* **28**(3) 336-55

[13] Liu X Q, Wu M, Chen J X 2002 Knowledge aggregation and navigation high-level Petri nets based *Learning, International Conference on Machine Learning and cybernetics* **1** 420-5

[14] Omrani1 F, Ali H, Vahid R 2011 An Adaptive Method Based on High-Level Petri nets for E-Learning *Software Engineering & Applications* 559-70

[15] Lowe R A, Landis G A, Jenkins P 1993 The efficiency of photovoltaic cells exposed to pulsed laser light (Report) *NASA Retrieved 3* Nov 2012.

[16] University of Chicago GEOS24705 Solar Photovoltaics EJM May 2011.

[17] Simon M 1981 Physics of Semiconductor Devices (2nd ed) 796

[18] Maximum Power Point Tracking. zone.ni.com Retrieved 2011-06-18

Authors

Ning Lu, China

Current position, grades: an associate professor of Wuhan University of Technology.
University studies Ph.D. in System Analysis and Integration from Huazhong University of Science and Technology in China in 2010.
Scientific interest: artificial intelligence and its applications in electric power system.

Leilei Yi, China

Current position, grades: assistant engineer of State Grid Suizhou Power Supply Company.
Scientific interest: power system analysis.

Operation Research and Decision Making

Research on the knowledge flow evolving model for mechanical product innovation design

Shaohui Su[1], Pengfei Li[1], Zhangming Peng[1*], Fanchao Wu[1], Chang Chen[1], Jiangang Wang[2]

[1]*School of mechanical engineering, Hangzhou Dianzi University, Hangzhou, 310018, China*

[2]*China New Building Materials Design and Research Institute, Hangzhou, 310012, China*

Abstract

The innovation and competitiveness of product require the fulfilment of customers demand. The key is to acquire knowledge of customer needs. In this article, a design knowledge flow cognitive model was built on the traditional FES model, using the acquirement of product demand for a start. It fuses the user's demand, function, effect, structure, constraint and recycling of product for product innovation. According to the model, the level of knowledge classification and description methods was studied and an object-oriented description method of knowledge was proposed. The relation between knowledge and knowledge flow was analysed and so were the vertical relation and horizontal relation between each knowledge point. The evolution process of knowledge was researched based product family Take example for Curve sawing products to validate the proposed model in this article.

Keywords: innovative design, knowledge flow, jig saw

1 Introduction

The decision will be made increasingly based on the data and analysis rather than experience and intuition in the commercial field, economic field and other fields with the coming of "Big Data" era. The ownership and using ability of the knowledge resources has become the source of enterprise's core competitiveness and product innovation in the era of big data. And the effective flow of knowledge is the important premise of innovation and is the main way to improve enterprise competition [1].

The knowledge flow refers to the knowledge's production, dissemination and application between multiple participants according to certain rules or procedures [2]. Design knowledge flow refers to that the design knowledge transmits between all parties of the design activities in the process of product design. Nonaka has proposed the 2D knowledge flow model based on epistemological dimension and the ontological dimension. This model could be used to describe the flow process of implicit and explicit knowledge between different ontology in an organization [3]. Nissen has proposed the 3D model to describe knowledge flow. He made clear that knowledge is dynamic sticky and not distributed within the organization evenly. If there is no relevant triggering mechanism, the static knowledge will always be static, and only the flow of knowledge can create value [4].

Yang Kun etc. have put forward product innovation solving model in which design flow and knowledge flow has interaction of multi-level reciprocating mappings. The different goals of products could be achieved using multi-level reciprocating mapping between design flow and knowledge flow [5]. Qingsong Xing etc. have proposed a kind of product innovation knowledge ontology semantic representation method based on OWL, and established the access mechanism of knowledge sharing and prototype system [6]. Zheng Liu etc., who adopted the method of knowledge flow modelling, has used knowledge node ontology knowledge system to identify knowledge source for the sake of obtaining dynamic creative knowledge [7]. Su Hai has analysed the function of knowledge map in knowledge chain management and presents the knowledge reuse model based on knowledge map [8]. Shengfa Wang has introduced resource space model to describe the knowledge in the knowledge flow. The knowledge is controlled through the integration of knowledge flow and workflow [9]. Haiqiang Liu has put forward a product integration design knowledge model, which could support multidisciplinary design optimization, to acquire product relative design knowledge with an integration of design examples and rules [10]. Qingfeng Bao has put forward a kind of manufacturing enterprise knowledge flow model based on life cycle. The knowledge flow process is divided into four stages: knowledge germination stage, growth stage, mature stage and knowledge recession stage [11].

Currently, the study of design knowledge flow is mainly focused on the aspect of knowledge sharing and

---
Corresponding author e-mail: pengming-620@163.com

reuse. The reuse of enterprise existing knowledge, such as data and information, can be realized by building knowledge management system. A company, whose innovative product is developed based on account of a quick access to relevant knowledge of customers' requirements and design tasks, could fulfil the needs of customers in the tight development cycle to gain advantages in the fierce market competition.

2 Knowledge flow model based on product innovation design process

The essence of product design activities is that the design process is driven by knowledge. The product innovation and its competitiveness of product edge require the fulfilment of customers demand and the key is to acquire knowledge of customer needs. In addition, product recycling and reuse should be taken into consideration in order to meet the requirements of sustainable development and energy saving using green design methods. In this article, a design knowledge flow cognitive model was built on the traditional FES model, using the acquirement of product demand for a start. It fuse the user's demand(R), function(F), effect(E), structure(S), constraint(C) and recycling(R) of product for product innovative design, as shown in Figure 1. The original target of product design is to obtain customer needs and then the function of products could be designed according to customer's demand. Meanwhile partial of the demands will be transformed into constraint during product design process. Following the determination of product function, the function is mapped into generalized physical effect to seek original theory. The product structural design can be conducted on the basis of original. The factors of constraint, recycling and reuse should be taken into account in the process of structural design. Eventually, the overall solution for products will be obtained.

With the proceeding of product development, the design knowledge of product is improved increasingly. During the process, which starts with user's demand and end with a fruit of a complete product solution, product design knowledge is indispensable. At the same time, it flows among every stage of the product design and continuously improves during the process. Meantime, the product's design is iterative with the process of perfection of knowledge and the knowledge flow will get corresponding feedbacks. From the view of natural evolution, knowledge itself in each stage is endowed with a process of evolution. Therefore, in order to support the product innovation design better, different stages of knowledge should be described and relation of knowledge in different stages should be established to construct a knowledge network.
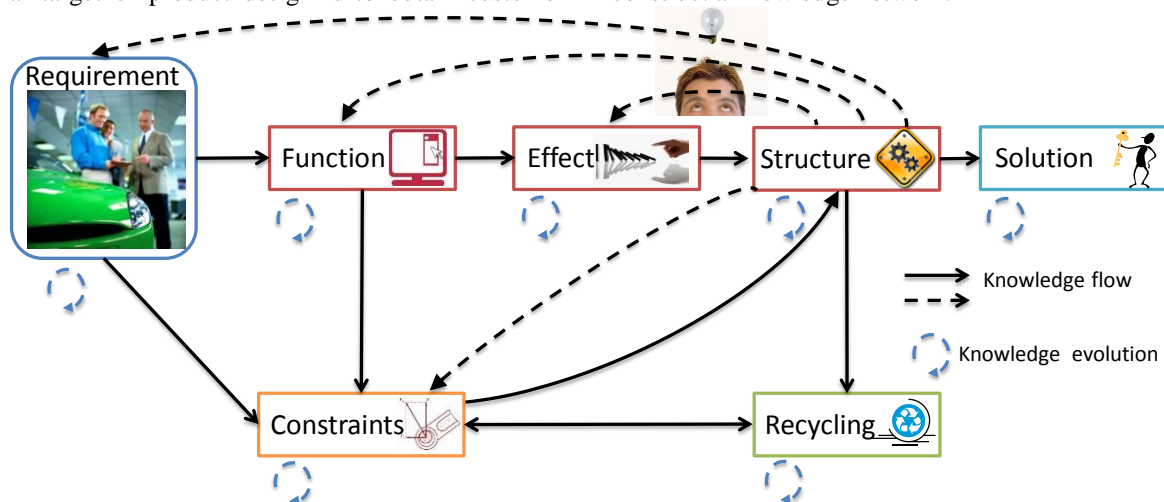
FIGURE 1 Knowledge flow model based on product innovation design process

3 Classification and description of different levels knowledge

According to the knowledge flow model mentioned above, knowledge could be classified from four levels. Here take jig saw products for example to describe four different levels of knowledge, as shown in Figure 2. Jig saw is a kind of normal electric tools, which is widely used in the fields of automobile, shipping and decoration etc. During innovative design of jig saw, customer's demand is reflected mainly in the aspects of material, cutting stroke, sawing thickness. Meanwhile customers also hope it is of low vibration and low noise in the process of using. Its main function is to realize the saw cutting through reciprocating movement. The electrical parts, transmission parts and supporting part should be designed in the developing process of jig saw. Take the transmission part as an example, gear drive is adopted. So the drive ratio of gear drive can be calculated with relevant mechanical principle knowledge. Then the speed, input power and torque of each shaft can be determined. After accomplishment of the designing calculation, corresponding structural parameters can be determined. The drawings and 3D model will be mapped out in the end.

Su Shaohui, Li Pengfei, Peng Zhangming,
Wu Fanchao, Chen Chang, Wang Jiangang

Customer requirements

sawing material, sawing stroke, sawing thickness, weight and noise vibration requirements

Jig Saw

Function

The saw cutting is realized through reciprocating movement.

motor | transmission parts

reciprocating movement Of the saw

Effect and principle

The transmission part of motor is realized by gear transmission

Structure

motor

Gear structure

$$P_1 = P_d \cdot \eta_1 = 1.1 \times 0.98 kw = 1.1 kw$$

$$P_2 = P_1 \cdot \eta_1 \cdot \eta_2 = 1.1 \times 0.98 \times 0.97 = 1.05 kw$$

$$P_3 = P_2 \cdot \eta_1 \cdot \eta_2 = 1.05 \times 0.98 \times 0.97 = 1.00 kw$$

FIGURE 2 Knowledge description of jig saw products in four aspects

The knowledge description should be satisfied the needs of designer and the computer. In other it should be convenient for the usage and exchanging among the designers and also be easily to store in the computer to realize computer-aided designing. Designers communicate with each other via natural language, graphics, formula and model etc. While the computer is apt to realize the management and storage of structured information and knowledge. Therefore an object-oriented method is applied to describe knowledge. The knowledge could be regarded as an object and a management object will be created to describe knowledge. Then the knowledge master record (KMR), which is the structuralized data of knowledge, will be created to be managed easily by computer. As to specific knowledge, general description method can be adopted to describe the specific knowledge for facilitating communication among designers, as shown in Figure 3.

In Figure 3 some attributes of object is defined to organize and manage knowledge. By means of these attributes, the content and features of knowledge can be reflected in different hierarchies. A living example for knowledge is shown below.

<Knowledge>
    <Gear Design notes>
   <Shaft>
   <Knowledge level> Structural hierarchy </ Knowledge level>
   <parts> drive system </ parts>
   <Knowledge category>rule</Knowledge category>
   <Transmission type>Gear transmission </Transmission type>
   <Knowledge code>12345678</Knowledge code>
   </shaft>
   <Founder>Yinsong Lu</Founder>
   <Version>V1.0</Version>
    <Creation date> design group on Flange</Creation date> 2014-2-18>
   <Changer>Yunfeng Zhang</Changer>
   <Storage location>
   http://192.168.112.64:8080/knowledge/message.jsp?message=715
   </Storage location>
   <keyword>Jig Saw, gear, Design Notes.</keyword>
   </gear design notes>
   <Summary of gear design experience>
   </Summary of gear design experience>
   <Living examples of gear design>
   </Living examples of gear design>
</knowledge>

Su Shaohui, Li Pengfei, Peng Zhangming,
Wu Fanchao, Chen Chang, Wang Jiangang
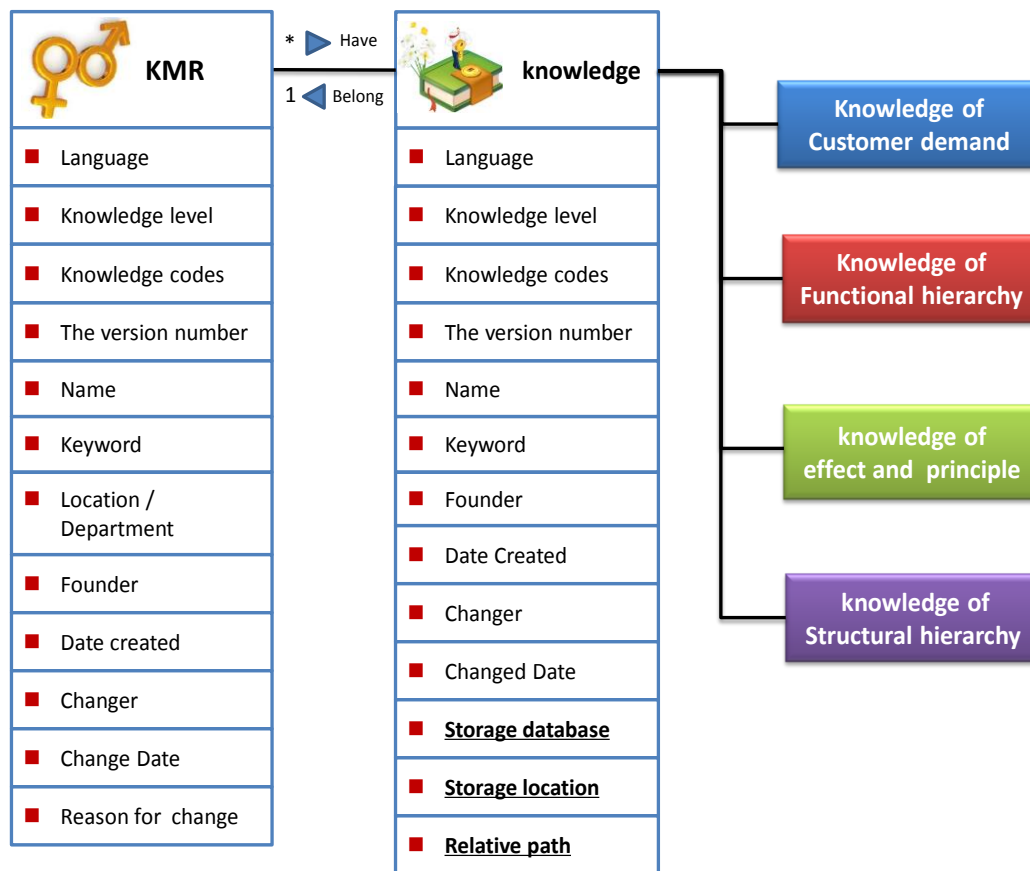
FIGURE 3 Object-oriented knowledge descriptions

Knowledge as an object, a method should also be needed besides the definition of its attributes. Figure 4 describes the "display" method of object via which the relevant properties of the object can be demonstrated.
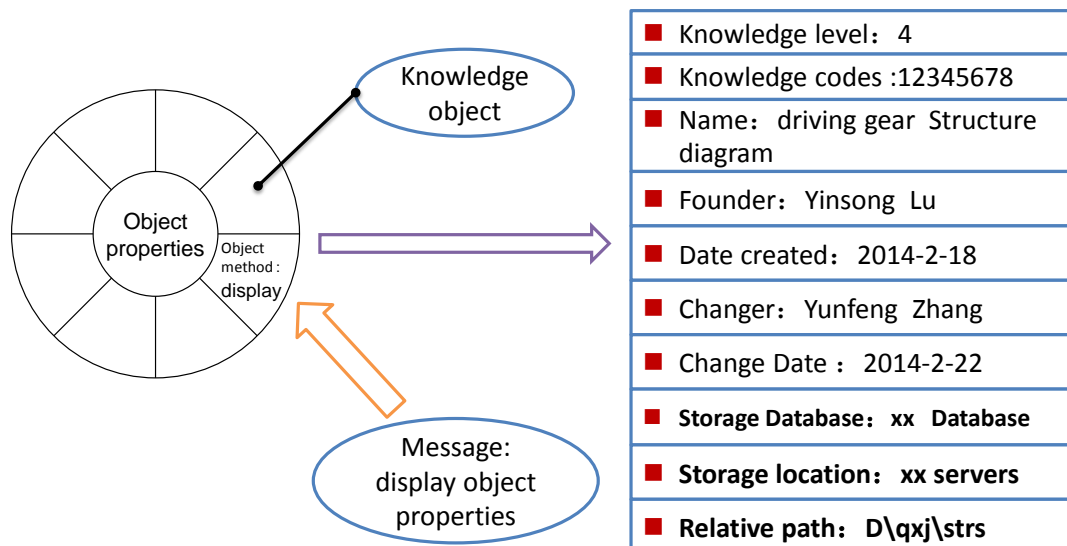
FIGURE 4 Example for knowledge object method

4 Contact and flow between the knowledge of various stages

The process of innovative design for mechanical product can be seen as a process in which the knowledge produces, flows, shares and apply between different designers, as shown in Figure 5. There are many relations between knowledge. Here the relationship between transverse knowledge and longitudinal knowledge is described mainly. The horizontal linkages refer to knowledge flows between the different levels and different stages of knowledge. The vertical linkages

336

Su Shaohui, Li Pengfei, Peng Zhangming,
Wu Fanchao, Chen Chang, Wang Jiangang

refers to the different stages belongs to the innovative design of the same product. It remains relatively integrity of knowledge. The knowledge of different stages is needed to design a product. It is difficult to design a product by only grasping a certain stage of knowledge, which truly meets customer's needs outside and is easy to design and manufacture inside. Horizontal linkages illustrate the flow and timing of knowledge, which can only truly be used for innovative product design and truly reflect the value of knowledge when it flows. Meanwhile the process of knowledge development is a time sequence. Only when the first stage has been completed or partially completed, the next stage could be started. Knowledge of the previous stages may be the input of subsequent stages while the next stage may require the use of knowledge generated previously.

If the enterprises want to win in the fierce market competition, the products must be meet customers' requirements. So knowledge of acquiring customers' needs is the starting point in product innovation design. After a clear understanding of customers' demand, the customers' needs should be mapped to specific product functions. After determination of product function, the following steps are to seek for specific knowledge of effects and scientific principles and then to propose preliminary concept design. The initial solution for concept design should be analysed and compared and the final solution could be concluded. Also the structure design of product will be done via computer analysis. Innovation design process of mechanical product can be regarded as a process in the premise of the given knowledge input where designers are driven by design goals to create knowledge output through specific design behaviour. Throughout the whole process, the design knowledge is constantly improved, the product structure continues to be improved and a complete solution comes into being finally.
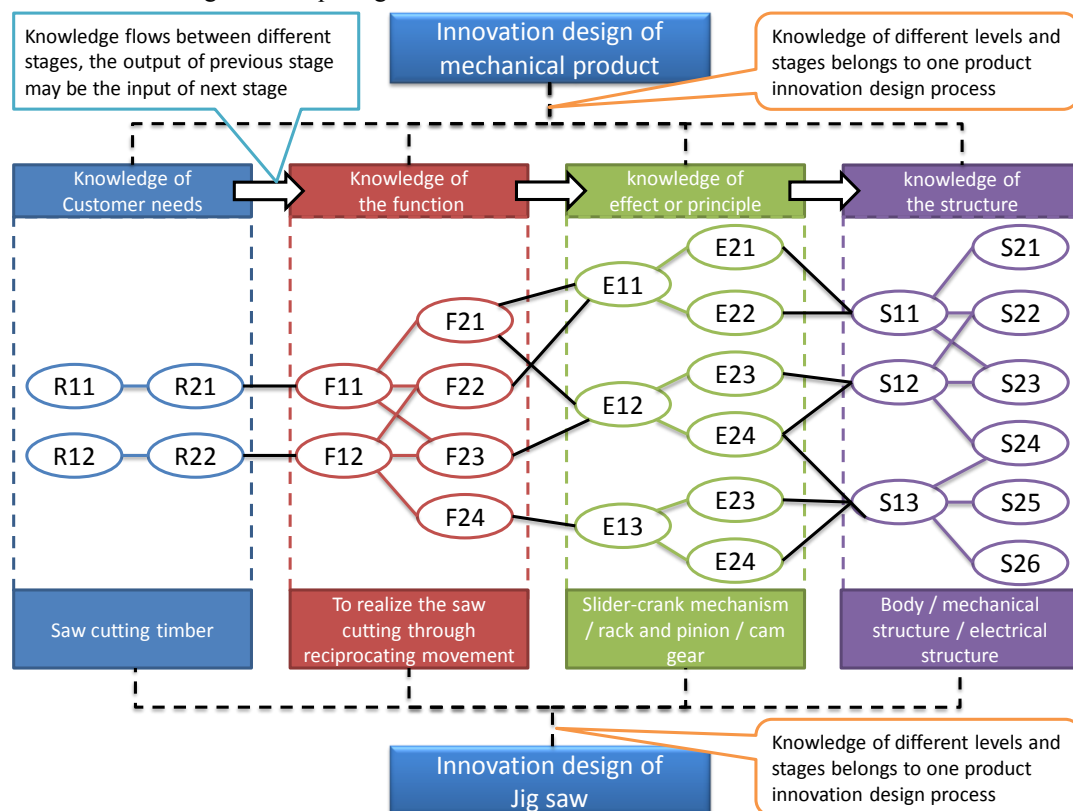
FIGURE 5 Contact and flow between different levels of knowledge

5 Evolution of knowledge itself

Mechanical product innovation design process requires the flow of knowledge. From the perspective of the product family, the product itself is also continually evolving and evolution. Therefore, the knowledge needed to design products is constantly evolving and evolution, as shown in Figure 6. The knowledge flows between different levels and stages in process of product innovation design. The improved design is an improvement on the basis of the original product. And the corresponding knowledge also needs to be evolved based on original knowledge. It is a spiral process. With the improvement of product and the evolution of knowledge, some new changes will be produced for the original product. Consequently, knowledge feedbacks will be inevitably generated.

Su Shaohui, Li Pengfei, Peng Zhangming,
Wu Fanchao, Chen Chang, Wang Jiangang
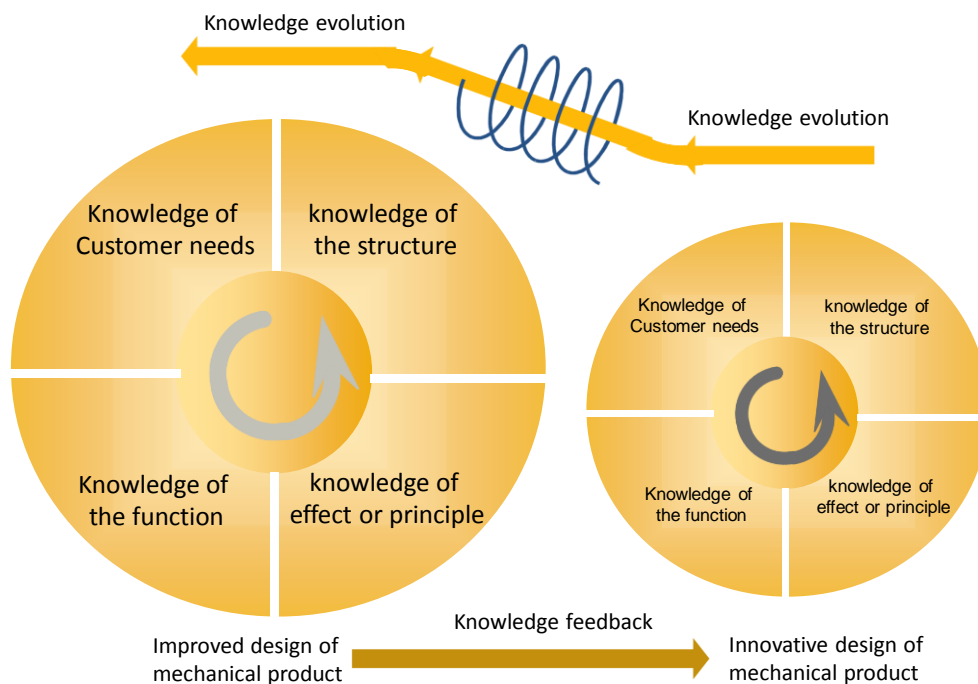
FIGURE 6 Evolution of knowledge itself

5 Conclusions

In order to provide the low-cost, high-quality products that also meet the needs of customers, enterprises' product design is in the direction of rapidity, modularity and knowledgeable. Innovative design of mechanical products is based on knowledge and is aimed at obtaining knowledge of customers' needs as a starting point. Therefore, the evolution model of a knowledge flow's established to describe the flow of knowledge. It will facilitate the flow and sharing of knowledge and serve for the innovative design of mechanical products in a better way. There are many categories of knowledge. How to carry out a unified description of knowledge and realize knowledge integration of the various stages and levels is needed study ulteriorly. And how to build a design knowledge management system is also requires further research. Knowledge can only create value and make better sharing when it flows. But knowledge is sticky, how to improve the flow of knowledge still needs further study.

Acknowledgments

References

[1] Zhang Z 2011 Knowledge flow theory and methods study in process of product design *Shanghai Jiaotong University doctoral dissertation* (*in Chinese*)
[2] Zhang X, Li S 2005 Journal of Software **16**(2) 184-93 (*in Chinese*)
[3] Nonaka I 1994 A dynamic theory of organizational knowledge creation *Organization science* **5**(1) 14-37
[4] Nissen ME 2002 An extended model of knowledge-flow dynamics *Communications of the Association for Information Systems* (8) 251-66
[5] Yang K, Li Y, Xiong Y, etc 2013 *Computer Integrated Manufacturing Systems* **19**(2) 274-83 (*in Chinese*)
[6] Xing Q, Yu Y, Liu A, etc. 2012 *China Mechanical Engineering* **23**(23) 2817-24 (*in Chinese*)
[7] Zheng L, Na L, Sun L 2011 *Computer Integrated Manufacturing Systems* **17**(1) 10-17 (*in Chinese*)
[8] Hai S, Jiang Z, Jun L 2007 *Shanghai Jiaotong University* **41**(7) 1061-5 (*in Chinese*)
[9] Wang S, Gu X, Zhang T, etc. 2008 *Journal of Agricultural Machinery* **39**(11) 122-6 (*in Chinese*)
[10] Liu H, Ji Y, Qi G, etc. 2009 *Journal of Zhejiang University (Engineering Science)* **43**(10) 1841-7 (*in Chinese*)
[11] Bao Q 2008 Research on knowledge management for manufacturing enterprise based on knowledge flow model *Northeastern University doctoral dissertation* (*in Chinese*)

Su Shaohui, Li Pengfei, Peng Zhangming,
Wu Fanchao, Chen Chang, Wang Jiangang

Authors

Shaohui Su, born on September 7, 1978, Luoyang City, Henan Province, China

Current position, grades: associate professor of Mechanical Engineering in Hangzhou Dianzi University.
University studies: PhD in Mechanical Manufacturing and Automation Zhejiang University.
Scientific interest: product digitalize design, manufacturing information engineering.
Publications: 5 papers.

Pengfei Li, born on July 23, 1989, Jiangxi Province, China

Current position, grades: studying for master degree in Hangzhou Dianzi University.
University studies: Bachelor of Hangzhou Dianzi University.
Scientific interest: knowledge management.

Zhangming Peng, born on September 17, 1977, Hubei Province, China

Current position, grades: lecturer of Hangzhou Dianzi University, doctor's degree.
University studies: marine engineering PhD Wuhan University of Technology.
Scientific interest: monitoring and control of diesel engine.
Publications: 2 papers.

Fanchao Wu, born on February 15, 1989, Xinyang City, Henan Province, China

Current position, grades: studying for master degree in Hangzhou Dianzi University.
University studies: Bachelor of Hangzhou Dianzi University.
Scientific interest: integration of TRIZ and other innovative design method.

Chang Chen, born on December 10, 1983, Taizhou City, Zhejiang Province, China

Current position: lecturer of Department of Mechanical Engineering in Hangzhou Dianzi University.
University studies: PhD of Huazhong University of Science and Technology.
Scientific interest: multi-domain unified modeling theory and technology.

Jiangang Wang, born on October 24, 1978, Zhuji City, Zhejiang Province, China

Current position, grades: senior engineer of China New Building Materials Design and Research Institute.
University studies: Bachelor of Hangzhou Dianzi University.
Scientific interest: mechanical design, new materials.
Publications: 2 papers.

Operation Research and Decision Making

HPN simulation model of carrying capacity of combination station for heavy-haul trains

Jinchuan Zhang*, Hao Yang

School of Traffic and Transportation, Beijing Jiaotong University, Shangyuancun 3, Haidian District, Beijing, China

Abstract

Combination station for heavy-haul trains imposed restrictions on the whole heavy-haul railway system. Through analysis of particularities of operation of combination station, the paper established HPN simulation model of carrying capacity of combination station based on Petri net theory, a graphical modelling method. The simulation model took technical operations of arrival, combination and departure of trains as interconnected system, and output parameters related to carrying capacity of the station. Finally, the paper, took Hudong station in Datong-Qinhuangdao railway as an example to verify the validity and practicability of the model.

Keywords: carrying capacity, HPN, combination station, heavy-haul railway

1 Introduction

Heavy-haul railway is transportation corridor for goods and materials among the areas in China. Because of variety of goods flow of sources and whereabouts, such a complicated organizational form of train, combination train, operated in the Chinese heavy-haul railway. Combination train is a kind train with greater tonnage of traction made up by several trains with less tonnage of traction in technical stations, which was called combination station. This kind of transportation organization can make the best use of carrying capacity of heavy-haul railway. The trains reception and departure in combination station included unit trains and combination trains. The work of unit train in the station is simple, consisting of duty shift of locomotives and crew. The work of combination train is more complicated, divided into three stages, train reception, train combination and train departure.

The issue of carrying capacity of combination station for heavy-haul trains is a new study field because the combination station is a completely new thing. So there are few literatures studying the issue according to the characteristics of heavy-haul railway.

So, the paper selected Hybrid Petri Net (HPN) to model the combination station for heavy-haul trains, which is a dynamic system mixed discrete and continuous processes, to describe static attribute of trains stopping in the station and dynamic behaviours in the section 1-3. The discreteness refers to the transition of display status of signal lamps and the stop of the train at the station. The continuance means moving process of the trains in the section or inside the station [4, 5].The continuous behaviour is different when trains running in the section and inside the station, because moving process of the train

limited by the arrival signal, thus forming the interactive status between continuous system and discrete system. The changes among the trains status of moving, stopping at station and signals, are of the characteristics of dynamic, concurrence and synchronization.

2 HPN simulation model of combination station

2.1 BRIEF INTRODUCTION ON HYBRID PETRI NET (HPN)

In 1962, C. A. Petri proposed the Petri network in his thesis for the Doctorate at first. In recent years, the research has already expanded the basic form of Petri network from different side and different angles, and led many kinds of expanded Petri networks of different characteristics and forms. Among them, HPN model can clearly describe systematic organization, the change course of structure and state and time characteristic. It particularly suited to describe the Characteristics, including parallelism, concurrency, synchronization and resource sharing. HPN is a kind of modelling tool with powerful function. It can combine graphical description and mathematical analysis. So HPN concurrently has intuition of graphical methods and generality of logical methods. It can also model the system mixed with discrete and continuous processes.

2.2 HPN SIMULATION MODEL

This paper, taking it as an example that trains moving into and out of the simulation system of combination station, set up HPN system simulation model of combination station for heavy-haul trains.

Combination station system is made up of home section, home throat area, receiving-departure yard,

---

Corresponding author e-mail: jchzh@126.com

starting throat area and starting section. There are combining-receiving-departure tracks and ordinary receiving-departure track in receiving-departure yard. The combining-receiving-departure tracks are tracks for combining trains and ordinary receiving-departure track for general trains.

When the trains moving in the home section, the train motion states are restrained by indicated states of signals. While entering the home throat area, the trains are influenced by switches with limited speed. When in the receiving-departure yard, the trains undertaking combination operation or other technological work as requested. Then, the trains, acceleration from parking state to leave receiving-departure tracks into starting section, restrained by switches with limited speed.

In order to simplify the problem size, the moving distance of trains at station are all established by sum of the length of home throat, receiving-departure tracks and starting throat (In fact, the moving distances of different kinds of trains are of certain differences, especially the train needing combined).

According to the mathematical definition of Petri network, the paper set up the model as multiple groups, $\text{HPN}=\{P,T,h,\text{pre},\text{post},\tau,M_0\}$.

$P = \{P_1, P_2, P_3, P_4, P_5, P_6, R_1, R_2, R_3, R_4, R_5, R_6, S_1, S_2,$
$S_3, S_4, S_5, S_6, S_7, S_8, G, G_1^1, G_1^2, G_2^1, G_2^2, G_3^1, G_3^2\}$.
$T = \{T_1, T_2, T_3, T_{p1}, T_{p2}, T_{p3}, T_G, T_{G1}, T_{G2}, T_{G3}, T_{G4},$
$T_{G5}, T_{G6}, T_{s1} T_{s2}, T_{s3}, T_{s4}, T_{s5}, T_{s6}, T_{s7}, T_{s8}, T_{s9}\}$.

The HPN simulation model of combination station is shown in Figure 1.

The meaning of place and transition are as follows:

P_1: distance of trains from home throat section (assumed to be l_1. In order to simplify the problem, in this paper, it is assumed to be the same that the moving distance of all the trains in home throat section). P_2: distance of trains moving in home section. P_3: length of home throat area (assumed to be l_2, which is also stopping distance, containing part of the length of receiving-departure track). P_4: distance of trains moving in home throat area. P_5: length of starting throat area (assumed to be l_3, which is also departure distance, containing part of the length of receiving-departure track). P_6: distance of trains moving in starting throat area. R_1: generating a train according to distribution discipline. R_2: idle status of beginning of home section. R_3: train number plus one in the simulation network. R_4: occupancy status of beginning of home section. R_5: train number minus one in the simulation network. R_6: trains moving out of the simulation network; S_1: trains in home section. S_2: trains in home throat area. S_3: trains in the beginning of combining-receiving-

departure yard. S_4: trains in ordinary receiving-departure yard. S_5: departure of ordinary trains. S_6: trains moving in starting throat area. S_7: trains moving out of starting throat area. S_8: trains in starting section. G: trains reaching combining-receiving-departure track. G_1^1: trains stopping at first half of combining-receiving-departure track. G_1^2: occupancy status of first half of combining-receiving-departure track. G_2^1: trains stopping at second half of combining-receiving-departure track. G_2^2: occupancy status of second half of combining-receiving-departure track. G_3^1: two trains in the same combining-receiving-departure track combined to a new train(combination train). G_3^2: departure of combination trains.
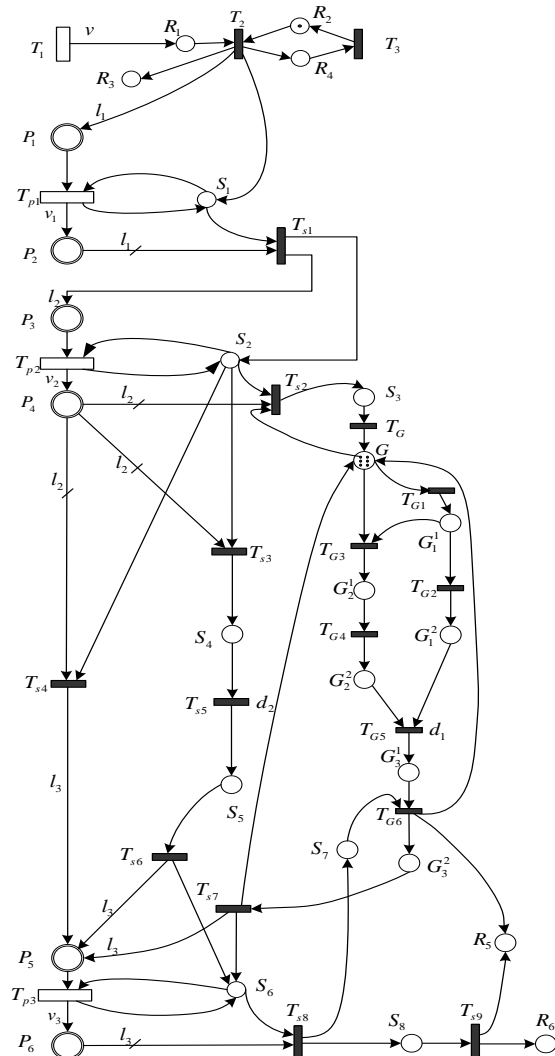

FIGURE 1 HPN simulation model of combination station

T_1: trains randomly generated at the speed of v. T_2: trains moving into simulation network. T_3: occupied to idle status of the beginning of home section. T_{p1}: trains

moving at the speed of $v_1(t)$ in home section. T_{p2}: trains moving at the speed of $v_2(t)$ in home throat area. T_{p3}: trains moving at the speed of $v_3(t)$ in starting throat area. T_G: trains moving into the second half of combining-receiving-departure track, which is in idle status. T_{G1}: idle status of the first half of combining-receiving-departure track which can receive trains. T_{G2}: idle to occupied status of the first half of combining-receiving-departure track. T_{G3}: destination of arriving train same to the train in the first half of combining-receiving-departure track for trains. T_{G4}: idle to occupied status of the second half of combining-receiving-departure track. T_{G5}: combining the two trains in the same combining-receiving-departure track. T_{G6}: occupied to idle status of starting throat area. T_{s1}: trains moving from home section to home throat area. T_{s2}: trains moving from home throat to beginning of combining-receiving-departure track, which is in idle status. T_{s3}: trains moving from home throat to ordinary-receiving-departure track, which is in idle status. T_{s4}: through trains occupying starting throat area. T_{s5}: trains undertaking technology operations in ordinary-receiving-departure yard. T_{s6}: idle to occupied status of starting throat area when trains departure from ordinary-receiving-departure yard. T_{s7}: idle to occupied status of starting throat area when combination trains departure from combining-receiving-departure yard. T_{s8}: idle status of starting section. T_{s9}: trains moving out of simulation network.

On the basis of the above definitions, the inferences can be obtained:

$$h(P_1, P_2, P_3, P_4, P_5, P_6, T_1, T_{p1}, T_{p2}, T_{p3}) \rightarrow C,$$

that is a continuous process.

$$h(R_1, R_2, R_3, R_4, S_1, S_2, S_3, S_4, S_5, S_6, S_7, G, G_1^1,$$
$$G_1^2, G_2^1, G_2^2, G_3^1, G_3^2, T_2, T_3, T_{G1}, T_{G2}, T_{G3}, T_{G4}, T_{G5},$$
$$T_{G6}, T_{s1} T_{s2}, T_{s3}, T_{s4}, T_{s5}, T_{s6}, T_{s7}, T_{s8}, T_{s9}) \rightarrow D,$$

that is a discrete process.

Input correlation mapping associated with the output are shown in Figure 1 directed arc labelling. For example, $pre(P_2, T_{s1}) = l_1$ and $post(P_3, T_{s1}) = l_2$, and the rest are no longer expatiated on. The non-labelled indicates $pre(P_i, T_j) = 1$ and $post(P_i, T_j) = 1$.

As to transition T_1, a nonnegative real number, v represents the speed of train planned to generate. As to transition T_{G5}, d_1 represents the time of combining trains, including technical operation time. As to transition T_{s5}, d_2

represents technical operation time of trains. If home section is in idle status, there is $M_{R_2}^0 = 1$.

Because there is not only one combining-receiving-departure track used in combination station, the value of M_G^0 can be set up according to the size of the actual situation. If there are 6 combining-receiving-departure tracks, thus there is $M_G^0 = 6$.

3 Analysis on operation of simulation model

3.1 ANALYSIS ON OPERATION OF SIMULATION MODEL

According to the given initial state shown in Figure 1, transition T_1 generates the train at the speed of v. Because $M_{R_2}^0 = 1$, so transition T_2 is enabled. The train moves into simulation network, the number of the trains plus one, and the status of beginning of home section become occupied from idle.

Then the train moves into home section, so transition T_3 is enabled, and the status of beginning of home section become idle from occupied so the next train can move into the simulation network.

At the moment, the train is moving at the speed of v_1 in the home section. After the train running the distance of l_1, transition T_{s1} is enabled, so the train moving into home throat area; when the train is the home throat area, the train moves at the speed of v_2.

There are three situations for the train according to the fact whether the train needs combining operation. In the first situation, the train needs combining operation, so transition T_{s2} enabled, and the train running the distance of l_2 into the combining-receiving-departure yard. If the second half of a certain combining-receiving-departure is in idle status, and the destination of the train is same to the train in the first half of the track, transition T_{G3} is enabled, and the train moves into the second half of the track to wait for combining operation. If the first half of a certain combining-receiving-departure is in idle status, transition T_{G1} is enabled, and the train moves into the first half of the track to wait for the successor train. When G_1^2 and G_2^2 exist at once, transition T_{G5} is enabled. The both trains in the same combining-receiving-departure track undertake combining operation. The delay time d_1 is the time of combining operation, which is random Numbers following a certain statistical distribution. If the starting throat is in the idle status, transition T_{G6} is enabled, and the combination train moves into starting throat area. Meanwhile, because the combination of two trains, the number of trains in simulation network minus one.

In the second situation, the train does not need combining operation, so transition T_{s3} enabled, and the

train running the distance of y into the ordinary-receiving-departure yard to wait for technical operations, then, transition T_{s5} is enabled. The delay time d_2 is the time of technical operation, which is random numbers following a certain statistical distribution. At present, the train can start. So transition T_{s6} is enabled, and the combination train moves into starting throat area.

In the third situation, if the train does not need any operations and directly through the station, transition T_{s4} is enabled, and the train moves into the starting throat.

When in the starting throat area, the train is moving at the speed of v_3. After the train runs the distance of l_3, transition T_{s7} is enabled, and the train moves into the starting section. Then, transition T_{s8} is enabled, and the train moves out of simulation network, so the number of trains in simulation network minus one.

At present, there are a lot of simulation tools for Petri net design, for example, widely used Visual Object Net++,

GPNT, and OPMSE. But simulation ability of some software is limited [6-10].

The paper used Matlab as simulation tool. Matlab, as a large commercial software in engineering calculations, has obvious advantages on computing capability, expansibility and openness [11-13].

Stateflow is a tool for modelling, simulation and analysis on complicated system integrated with Simulink in Matlab. It unites the theories together that finite state machine theory, flow diagram and state transition diagrams. It is a creation and simulation tool on complex response system and event-driven system, very suitable for simulation of Petri net.

3.2 BASIC FRAMEWORK OF THE SIMULATION MODEL

Flow chart of simulation of combination station for heavy-haul trains is shown in Figure 2.
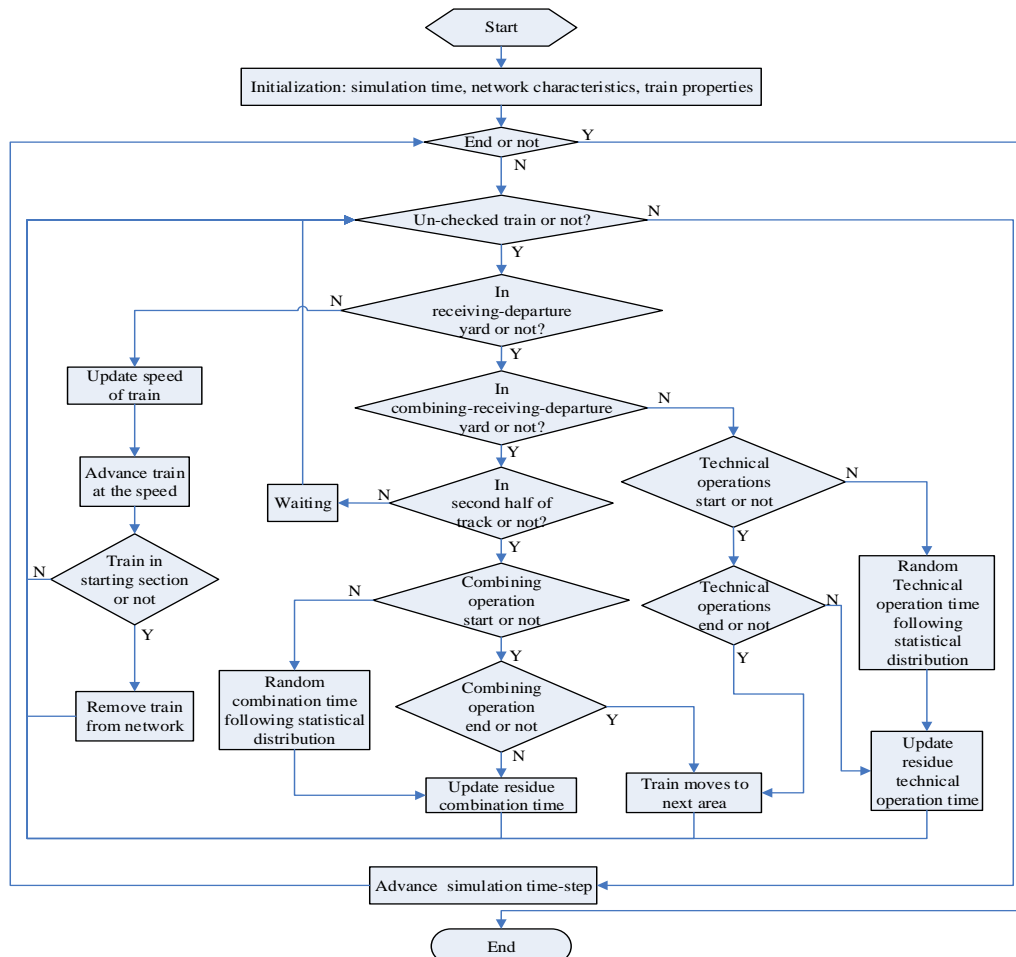
FIGURE 2 Flow chart of simulation of combination station

The simulation system needs traverse detailed information of each train in the network every simulation cycle, computing speed, moving position, and corresponding operations of the trains.

4 Case study on HPN simulation model

The paper, according to the above simulation theory and model, carried on simulation calculation of carrying capacity of Hudong station in Datong-Qinhuangdao

343

railway, based on the yard layout, traffic flow and operation organization method.

The paper mainly aimed at the experiment simulation of the operation that 2 trains weighing 10000 tons are combined to 1 train weighing 10000 tons, and the simulation cycle is for 24 hours.

At start of the simulation, arrival time of trains, which is determined according to the actual statistics data of Hudong station, is randomly generated. When the simulation is operated, the time table of a certain number of train is prior generated, which is read in sequence when simulation is running.

When a train generated reaches a certain combining-receiving-departure track in totally idle status, the train can stop in the first half of the track, and wait for the successor train. While, if just the second half is in idle status, the train can stop in the second half of the track, the combination operation can be undertaken, and the time of combination operation is generated according to the random probability distribution.

According to the realistic situation of train operations in Hudong station, the arrival and departure data and operation time distribution are as follows.

Firstly, train arrival time interval analysis: the average train time interval is 624 seconds, by using statistical analysis software, the train arrival time interval basically obeyed normal distribution.

Secondly, train combination operation time: according to the realistic data, the average time of a train weighing 20000 tons combined by 2 trains weighing 10000 tons is 25.6 minutes, the shortest time is10 minutes, and the longest time is 54 minutes. By using statistical analysis software, the data basically obeyed normal distribution.

Thirdly, according to statistics data of cross interference situation of station operation, the interference time approximately obeyed normal distribution, the minimum time is 2 minutes, the largest time is 40 minutes and the average time is 12 minutes.

The paper introduced random cross interference when simulation of the train operation in the station. In simulation process, the various cross interference possibly occurring in the station are transformed into the impact on the operation time, so, the given operation time were to be added the interference time.

Therefore, the time from the second train reaching to the combination train moving out of the track were as follow: the average time is 161minutes, the minimum time is 68mins and the maximum time is 330 minutes. By using statistical analysis software, the data basically obeyed γ distribution, and the two parameters respectively were $\alpha = 8$ and $\beta = 0.05$.

There are 6 combining-receiving-departure tracks in Hudong station, so, the ability restriction of combination operation should be considered. While the ability restriction of ordinary-receiving-departure yard and carrying capacity of through trains weighing 20000 tons need not be considered.

According to the future tendency, the destinations of combination trains were set to three, including Qinhuangdao Port, Caofeidian Port and other stations, and the ratio is 2.2:2:1.

In the case, the length of trains weighing 10000 tons is 1.2km, the length of trains weighing 20000tons is 2.4km, the limited speed in throat area is 30km/h, the value of l_1, l_2 and l_3 are 1.5km, 4km and 4km, and the value of v is 60km/h.

The speeds were updated according to the functions as follows:

$$v_1(t+1) = v_1(t) - 8.33t , \qquad (1)$$

$$v_2(t+1) = v_2(t) - 1.44t , \qquad (2)$$

$$v_3(t+1) = v_3(t) + 1.17t . \qquad (3)$$

According to the need to promote the simulation process, on above functions, the unit of speed is km/h, and the unit of time is minutes.

In simulation process, the skylight time of 120 minutes were deducted.

In every simulation circle, the given traffic volume was large enough and the simulation time was set to 30 hours. The result is that the traffic volume can be handled by the station within the later 24 hours.

In the 50 times simulation, the daily combination 20000-ton trains are 46 on average, 43at least and 49 at most. The statistical data was shown in Figure 3.
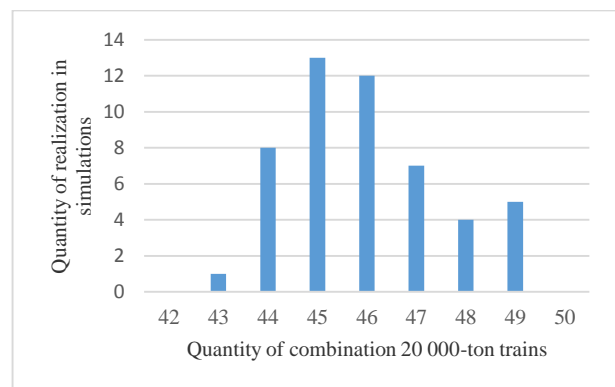
FIGURE 3 Quantity of combination 20,000-ton trains realized in simulations

After 50 times simulations, each simulation can realize the quantity of trains as shown in Table 1.

Attainable probability of quantity of combination 20000-ton trains was shown in Table 2.

TABLE 1 Statistics of simulation results

NO.	Quantity of 10000-tons trains by ordinary technical operation	Quantity of directly through 20000-tons trains	Quantity of combination 20000-tons trains	Total number of trains	Conversion quantity of 10000-tons trains
1	36	11	44	91	146
2	35	11	46	92	149
3	36	12	45	93	150
4	35	12	45	92	149
5	35	13	45	93	151
6	34	13	43	90	146
7	37	11	45	93	149
8	37	10	49	96	155
9	36	11	46	93	150
10	37	12	45	94	151
11	36	12	45	93	150
12	35	12	44	91	147
13	35	11	47	93	151
14	34	12	44	90	146
15	34	12	45	91	148
16	35	10	49	94	153
17	35	13	46	94	153
18	34	13	48	95	156
19	38	12	47	97	156
20	36	11	47	94	152
21	35	13	46	94	153
22	35	12	44	91	147
23	35	13	44	92	149
24	36	12	45	93	150
25	33	13	46	92	151
26	34	12	49	95	156
27	33	12	46	91	149
28	36	11	45	92	148
29	35	13	47	95	155
30	34	11	49	94	154
31	34	13	46	93	152
32	35	11	48	94	153
33	35	12	46	93	151
34	36	11	45	92	148
35	37	12	44	93	149
36	35	12	45	92	149
37	37	12	44	93	149
38	33	12	48	93	153
39	35	14	45	94	153
40	38	11	44	93	148
41	35	12	46	93	151
42	35	12	47	94	153
43	37	13	45	95	153
44	34	13	47	94	154
45	35	13	46	94	153
46	34	10	49	93	152
47	36	10	48	94	152
48	35	12	46	93	151
49	36	12	46	94	152
50	34	12	47	93	152

TABLE 2 Attainable probability of quantity of combined 20,000-ton trains

Quantity of combination 20000-ton trains	≥43	≥44	≥45	≥46	≥47	≥48	≥49	≥50
Attainable probability	1	0.98	0.82	0.56	0.32	0.18	0.1	0

The annual transporting capacity of Datong-Qinhuangdao railway can be calculated through the above realization probability of traffic volume. Suppose the 10000-ton trains all were made up by C80-wagons, so the annual transporting capacity achieved by a single 10000-ton train is $10000 \times 0.8 \times 365 = 2920000$ tons.

According to the carrying capacity of Hudong station, the attainable probability of annual traffic volume of 425 million tons in Daqin railway is 1, the attainable probability of annual traffic volume of 429 million tons is 0.94, the attainable probability of annual traffic volume of 435 million tons is 0.82, the attainable probability of annual traffic volume of 451 million tons is only 0.16.

The analysis shows that when traffic volume from 429 to 435 million tons, attainable probability fell by 0.12 and when from 435 to 451 million tons, attainable probability fell by 0.66, which fully shows that when the transporting capacity was 435 million tons, under the condition of existing equipment, the carrying capacity of Hudong station was saturated and there is no space to increasing the transporting capacity.

5 Conclusion

The paper, on the base of analysis of particularities of operation of combination station, established HPN simulation model of carrying capacity of combination station by using Petri net theory, which was a graphical modelling method. The simulation model took technical operations of arrival, combination and departure of trains as interconnected system, and output parameters related to carrying capacity of the station. Hudong station in Daqin railway was taken as an example to prove the validity and practicability of the HPN simulation model.

Reference

[1] Yan S, Cole C, Spiryagin M, Godber T, Hames S, Rasul M 2013 Conceptual designs of hybrid locomotives for application as heavy haul trains on typical track lines *Proceedings of the Institution of Mechanical Engineers Part F Journal of Rail and Rapid Transit* **227**(5) 439-52

[2] Fenling F, Dan L 2012 Heavy-haul train's operating ratio, speed and intensity relationship for Daqin railway based on cellular automata model *Information Technology Journal* **11**(1) 126-33

[3] Xin L, Makowsky T W, Eadie D T, Oldknow K, Jilian X, Jinzhong J, Guibao L, Xianhong M, Yude X, Yu Z 2012 Friction management on a Chinese heavy haul coal line *Proceedings of the Institution of Mechanical Engineers Part F: Journal of Rail and Rapid Transit* **226**(6) 630-40

[4] Khan S A, Zafar N A, Ahmad F, Islam S 2014 Extending Petri net to reduce control strategies of railway interlocking system *Applied Mathematical Modelling* **38**(2) 413-24

[5] Milinkovic S, Markovic M, Veskovic S, Ivic M, Pavlovic N 2013 A fuzzy Petri net model to estimate train delays *Simulation Modelling Practice and Theory* **33** 144-57

[6] Declerck P, Bonhomme P 2014 *IEEE Transactions on Automation, Science and Engineering* **11**(1) 103-10

[7] Mekki A, Ghazel M, Toguyeni A 2012 *IEEE Transactions on Intelligent Transportation Systems* **13**(2) 714-23

[8] Herajy M, Heiner M 2014 Petri net-based collaborative simulation and steering of biochemical reaction networks *Fundamenta Informaticae* **129**(1) 49-67

[9] Ghomri L, Alla H 2013 Continuous flow systems and control methodology using hybrid petri nets *Journal of Control Engineering and Applied Informatics* **15**(4) 106-16

[10] Fanti M P, Iacobellis G, Mangini A M 2014 *IEEE Transactions on Automation, Science and Engineering* **11**(1) 90-102

[11] Fateh K, Said H, Abdellah E M 2007 A hybrid Petri nets-based simulation model for evaluating the design of railway transit stations *Simulation Modelling Practice and Theory* **15**(8) 935-69

[12] Partha C, Durgesh V 2008 Optimum assignment of trains to platforms under partial schedule compliance *Transportation Research Part B* **42**(2) 169-84

[13] Andrews J 2013 A modelling approach to railway track asset management *Proceedings of the Institution of Mechanical Engineers Part F: Journal of Rail and Rapid Transit* **227**(1) 56-73

Authors

Jinchuan Zhang, born in June, 1981, Henan, China

Current position, grades: assistant professor at the school of traffic and transportation, Beijing Jiaotong University.
University studies: Doctoral degree in traffic and transportation at Beijing Jiaotong University in 2009.
Scientific interests: railway transportation organization.
Publications: more than 15.

Hao Yang, born in December, 1944, Fujian, China

Current position, grades: professor in the school of traffic and transportation, Beijing Jiaotong University.
University studies: Master's Degree in railway transportation in 1978.
Scientific interests: railway transportation organization.
Publications: more than 50.

Operation Research and Decision Making

A new mutton logistics tracking algorithm for Internet of things based on PSO and neural network

Minning Wu[1*], Fei You[1], Feng Zhang[1, 2]

[1]*School of Information Engineering, Yulin University, Yulin 719000, China*

[2]*School of automation, Northwestern Polytechnical University, Xi'an 710072, China*

Received 3 March 2014, www.cmnt.lv

Abstract

In order to improve the particle filtering precision and reduce the required number of particles, to solve the neural network training algorithm has slow convergence speed, easily falling into local optimal solution, proposed a target tracking algorithm based on PSO particle filter, using of Bayesian method to sample the prior information and coupled PSO algorithm. For the existence of intelligent wireless sensor network energy constrained sensor nodes, limited communication features, the PSO optimization is introduced into the distributed particle filter algorithm to solve the existing distributed particle filter network traffic load is heavy and node energy consumption of high disadvantage. Then, we propose a new particle filter algorithm based on PSO and neural integration the algorithm makes full use filter tracking historical information, combined with predictions of particle filter, the detection signal of the sensor nodes were isolated, thus achieving the target tracking. Simulation results show that the target tracking algorithm based on particle filter PSO and neural integration can use a smaller computational cost, multi-target tracking problem solving, and practical system to meet the demand.

Keywords: particle swarm optimization, particle filter, neural network, tracking algorithm, internet of things

1 Introduction

Particle filter (PF) is an estimate based on statistical filtering theory and Bayesian Monte Carlo idea [1]. It is based law of large numbers, using the Monte Carlo method to solve the integral Bayesian estimation problem, namely the sequential Monte Carlo methods. Theoretically applicable to any nonlinear stochastic systems under non-Gaussian background model represents the state space available, its accuracy can approach the optimal estimation. In recent years, it has been successfully applied in targeting [4], the Internet [5] and other fields. Target tracking is one of the typical applications of wireless sensor networks, target tracking can be used for cargo warehouse logistics tracking, traffic planning, monitoring wildlife and intruder monitoring. Since the sensor nodes with small size, low price, easy to deploy, such as communications and computing capabilities, while the network has self-organization, concealment and robust features, compared with the traditional tracking methods, wireless sensor networks for its good characteristics make up the deficiency of traditional tracking methods, making the wireless sensor network more suitable for locating and tracking moving targets. Therefore, research on target tracking based on WSN is of great theoretical and practical significance.

Wireless sensor network target tracking system consists of target detection, target location and target prediction and notification class into [6]. By activation of the target nodes of a dynamic near the tracking area, the node information by voice, vibration information, the image information of the target detection and location [7]. Then sends the position information to the dynamic tracking of the target area of the data processing centre, and the status of the completion of updating the target position in the data processing centre and the prediction.

Particle filter observation system as an effective way to solve the nonlinear, non-Gaussian problems, the target location and tracking system widely used. Since the particle filter complexity and computational requirements of data storage capacity, the particle filter algorithm must consider the impact the computational complexity, computational power nodes and node energy consumption and other factors in wireless sensor network based target tracking environment [8]. The traditional particle filter does not need to consider the energy consumption, computation complexity and data storage capacity and other issues in the operation of Sink node or terminal user centre. When the particle filter run-time dynamic tracking of distributed nodes in the cluster, you need to consider limiting the node energy and computing power, find a new particle filter operating mode, reducing energy consumption under the premise of protecting the target tracking accuracy, balance tracking accuracy and network energy consumption.

In this paper, from the basic theory of target tracking, in-depth analysis of the particle filter principle, on the basis of dynamic clustering, we propose a parallel

---

Corresponding author e-mail: 178578051@qq.com

extended Kalman particle filter algorithm (PEPF) and based on PSO particle filter (PSO-PF) and neural networks particle filter (PSO-PFNN).

2 Targeting particle filter algorithm

Motion model and measurement model is not only the design basis of objective filtering algorithm , and appropriate model can effectively reduce the computational complexity [9], improve tracking accuracy. For manoeuvring target tracking issues, we can use the following motion model Equation (1) and measurement model Equation (2) description:

$$x_t = f(x_{t-1}, v_{t-1}),$$ (1)

$$z_t = h(x_t, w_t),$$ (2)

where x_t represents the system state at time t, for target tracking, it is usually the target position, velocity or acceleration information. z_t represents the measured value of the target, v and w denote the process noise and measurement noise. $f()$ represents manoeuvrings target motion model, function $h()$ represents a measurement model, used to indicate the relationship between the measured value and the target states. General Equation (1) is established in the Cartesian coordinates, while Equation (2) was established in the spherical coordinates, in the solving process usually take the measurement equation is transformed to Cartesian coordinates. So, the target tracking problem is usually considered a highly nonlinear problem.

Measurement model used to describe the relationship between the target measurement data and status variables. It is generally refers to the measured data from the sensor output at all viewing measurement collection. These observations are usually attribute information (type, status, etc.) and a set of goals, motion information (position, velocity, etc.) as well as environmental information (time) correlation measurements. In the target tracking problem, the measured data is the main signal and the target motion state. As for the sensors, typically measured value and the target cannot be obtained directly related to the state, and the measured value contains mostly noise and clutter, so the general model to the measured data using the measured data for filtering, and the original measured value and the target the motion state information associated with it.

2.1 KALMAN FILTER ALGORITHM

Tracking Kalman filter for linear systems, and are assuming a Gaussian posterior probability distribution at any time can be determined by the mean and variance of the distribution. Kalman filter algorithm that is based on the following assumptions:

1) Target transfer noise from the v_{i-1} and measurement noise n_i is Gauss distribution with known parameters.

2) The state transition equation $f_i(x_{i-1}, v_{i-1})$ is a linear function of x_{i-1} and v_{i-1}.

3) The measurement equation $h_i(x_i, n_i)$ is a linear function of x_i and n_i.

Then, the Equations (1) and (2) can be written as:

$$x_t = Fx_{t-1} + v_{t-1},$$ (3)

$$z_t = Hx_t + n_i,$$ (4)

where F and H are matrix definition of linear function, v_{t-1} and n_i are independent of each other, and the mean value is 0, Q and R are variance of a Gaussian distribution. Then it is according to Bayesian estimation theory, we can get a Kalman filter algorithm.

2.1.1 State Prediction

a) Target state and variance forecast:

$$x_{t|t-1} = Fx_{t-1|t-1},$$ (5)

$$P_{t|t-1}^x = Q + FP_{t|t-1}^x F^T.$$ (6)

b) Predict the measured value and variance:

$$z_{t|t-1} = Hx_{t|t-1},$$ (7)

$$P_{t|t-1}^z = R + FP_{t|t-1}^z H_t^T.$$ (8)

c) Correlation matrix is a priori predictions:

$$P_{t|t-1}^{xz} = P_{t|t-1}^x H_t^T.$$ (9)

2.1.2 Measurement update

a) Calculating the Kalman gain:

$$K_t = P_{t|t-1}^{xz}[P_{t|t-1}^z]^{-1} = P_{t|t-1}^x H_t^T [R + HP_{t|t-1}^{xz} P_{t|t-1}^x H_t^T]^{-1}.$$ (10)

b) Using of time t measured values, update the target state prediction and estimation variance:

$$x_{t|t} = x_{t|t-1} + K_t[z_t - z_{t|t-1}],$$ (11)

$$P_{t|t}^x = P_{t|t-1}^x - K_t P_{t|t-1}^z K_t^T = (1 - K_t H_t)P_{t|t-1}^x.$$ (12)

Since the Kalman filter algorithm is simple, small amount of calculation has been widely used in target tracking. Kalman filter algorithm is a Bayesian filter theory of linear, Gaussian distribution analytical results obtained under the assumption, therefore, is the best estimate of the Kalman filter algorithm for linear Gaussian filter under the concept.

2.2 PARTICLE SWARM OPTIMIZATION ALGORITHM

PSO algorithm [10] is an operation based on the fitness of particles, Swarm intelligence generated by the cooperation

and competition between particles to guide the optimization search. The particle as a no weight and volume of the particles. In the *n-dimensional* space, it is keeping a certain speed flight, flight speed dynamically adjusted by the flying experience of individuals and groups of flying experience. PSO algorithm is the use of the substance of the information itself, the next iteration of individual extreme position information and global information on these three kinds of extreme value information to guide particles.

Standard PSO algorithm is to calculate the fitness of each particle *i* in iterative process, by tracking individual extreme p_i and p_g to update their global minimum. Particle *i* is to update its velocity and position values according to the following two equations:

$$V_i(t+1) = \omega V_i(t) + c_1 r_1 [P_i(t) - x_i(t)] + c_2 r_2 [p_g(t) - x_i(t)], \quad (13)$$

$$X_i(t+1) = X_i(t) + V_i(t+1), \quad (14)$$

where $V_i(t)$ is the particle *i* velocity at time *t*, ω is the inertia weight, $x_i(t)$ is a position of the particle at time *t*, r_1, r_2 is a random number between (0, 1), c_1, c_2 is a learning factor, usually $c_1 = c_2 = 2$. It can significantly improve the performance of the PSO algorithm, it is given:

$$\omega = \omega_{\min} + (iter_{\max} - iter) \times (\omega_{\max} - \omega_{\min}) / iter_{\max}, \quad (15)$$

where $\omega_{\min}$, $\omega_{\max}$ respectively the maximum and minimum weighting factor, *iter* is the current iteration number, $iter_{\max}$ is the total number of iterations. $\omega_{\max}$ is the initial inertia weight; $\omega_{\min}$ is the last inertia weight; $t_{\max}$ is the maximum number of iterations. Flight speed is $v_i \in [-V_{\max}, V_{\max}]$, the constraint conditions to prevent particle speed missed optimal solutions, through the improvement of the algorithm further improves the global searching ability of particle swarm.

3 Particle filtering algorithm based on PSO and neural networks

3.1 PSO PARTICLE FILTER ALGORITHM

PSOPF algorithm can be expressed as follows: First, the introduction of the new measured values of the sampling process and the fitness function is defined as:

$$z_k \sim fitness = \exp\left[-\frac{1}{2R_k}(z_k - z_k^i)^2\right]. \quad (16)$$

Particle initialization: PSO $p(x_0)$ generated from the a priori probability density $\{x_0^i\}_{i=1}^N$, it all particle value is $\frac{1}{N}$, calculate the importance weights:

$$w_k^i = w_{k-1}^i \exp\left[-\frac{1}{2R_k}(z_k - z_{k|k-1})^2\right]. \quad (17)$$

Using the PSO algorithm according to the speed and position of each particle to update formula, the particles closer to the true state of constantly, it is given

$$v_{ij}(t+1) = wv_{ij}(t) + c_1 r_{1j}[p_{ij}(t) - x_{ij}(t)] + c_2 r_{2j}[p_{gj}(t) - x_{ij}(t)], \quad (18)$$

$$x_{ij}(t+1) = x_{ij}(t) + v_{ij}(t+1). \quad (19)$$

Flow chart of particle PSO particle filter algorithm as shown in Figure 1.
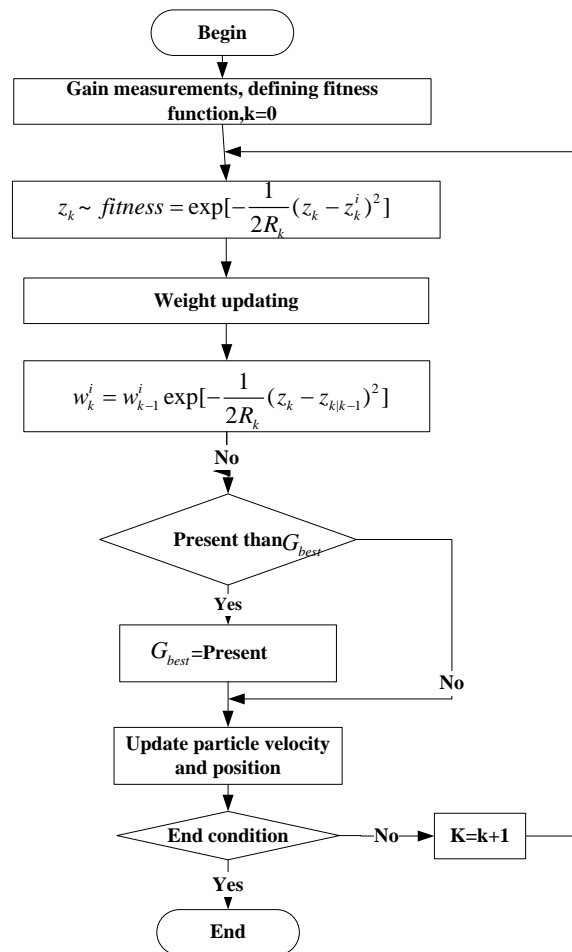
FIGURE 1 Flow chart of PSO particle filter algorithm

3.2 PSO NEURAL NETWORK OPTIMIZATION ALGORITHM

In recent years, with the increasing depth research, many researchers use swarm intelligence optimization algorithm to optimize the training process of neural networks [11]. Because of swarm intelligence optimization algorithm has strong global convergence ability and robustness, and the characteristics of the problem without the help of

information, such as derivatives and other information, therefore, will not only be able to play a combination of both the generalization ability of neural network mapping, but also can improve the convergence speed and learning ability. Among them, PSO algorithm is used as a simple and effective study based on a random search algorithm is shown to have great potential in terms of optimization of neural networks.

PSO neural network optimization (PSO-NN) algorithm is used to replace the traditional parameters of PSO training algorithms such as BP to optimize NN. Each particle is a vector that represents a group of parameters, process the process for the global optimal value is to obtain the optimal parameters.

The training error is used to calculate the fitness value $f(x)$, it is given:

$$f(x) = \frac{1}{1 + \frac{1}{2n}\sum_{k=1}^{n}(y_k - t_k)} , \qquad (20)$$

where, k is the number of samples, y_k is the actual output value, t_k is the output value. When it reach the maximum number of iterations or target error, the program terminates, and get the global optimum value that a group of optimal parameters. Algorithm is detailed as shown in Figure 2.

The initial particle swarm PSO algorithm standard position is randomly generated, causing the initial search space of uncertainty. Set the priori information coupling problem inherent in the object into the PSO, it can narrow the search space. For a large sample of the data set, the posterior probability is calculated first plurality of sets of values of weights in the particulate composition that is under the given conditions. Then initializes the value according to the probability distribution of particle position, thereby reducing the value of the learning process to modify the magnitude of particulates, to improve the learning speed.

Get the priori information is calculated as follows:
set the neural network weights is $W = (w_1, w_2, ..., w_n)$. Sample points are $D = \{(x_1, y_1), (x_2, y_2), ..., (x_m, y_m)\}$, where n and m are respectively the weight and the number of sample points, The position of a particle of W value in the sample points of conditional probability is:

$$P\left(\frac{W}{D}\right) = \frac{P(D/W)P(W)}{P(D)} . \qquad (21)$$
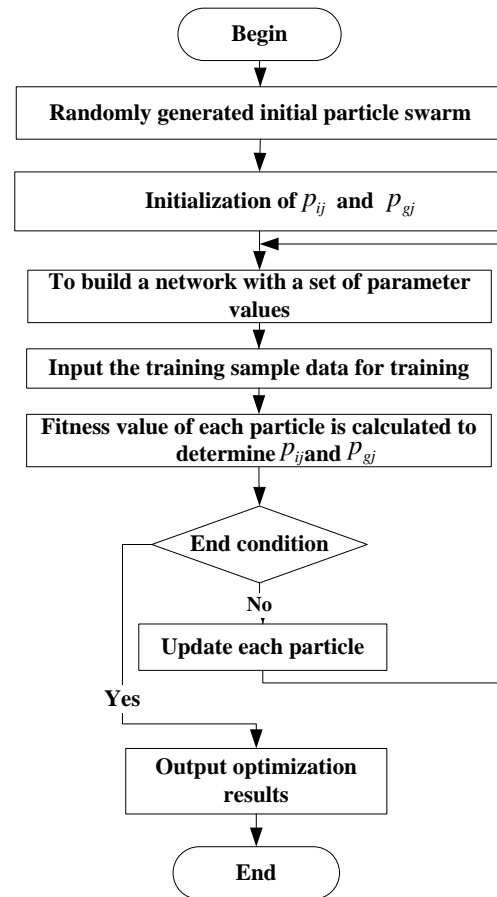
FIGURE 2 PSO neural network optimization algorithm

$P(W)$ is the particle position value of W prior distribution. $P(D/W)$ is the likelihood function, $P(D) = \sum_i P(D/w_i)P(w_i) = 1$ is a normalization factor.

In the absence of a priori information, if the W is too small, then all the incentive function are almost in a linear part, will also reduce the speed of convergence. The general consensus is that W obeys the exponential distribution, the equation is as follows:

$$P(W) = \frac{1}{Zw(\alpha)} \exp(-aEw) . \qquad (22)$$

The $Zw(\alpha)$ is a normalization factor, to ensure $\int P(W)dW = 1$, $Zw = \int \exp(-\alpha Ew)dW$, a is a control parameter of weight distribution form, Ew is a kind of error function. General Ew as shown in the following equation:

$$Ew = \frac{1}{2}\|W\|^2 = \frac{1}{2}\sum_{i=1}^{n} w_i^2 . \qquad (23)$$

Coupled Bayesian prior information using PSO algorithm combines the former BP algorithm to train feed forward neural networks (PSO-BPNN), the algorithm detailed steps can be divided into four stages.

3.2.1 PSO-BPNN initialization step

Step 1: generate training samples

$D = \{(x_1, y_1), (x_2, y_2), ..., (x_m, y_m)\}$;

Step 2: determine the number and network target error each neuron;

Step 3: determine the relevant parameters of PSO: the number of particles, the dimensions of each particle, inertia weight, acceleration factor, the maximum number of steps, such as iteration.

3.2.2 Coupling priori information

Step 1: randomly initialize each particle velocity and position;

Step 2: according to the formula (22) modify each particle's position, that is used to calculate the $P(D/W)$ update particle position value.

3.2.3 Using a priori information coupled PSO training Neural Network Model

Step 1: generate particles using a priori information to adjust the position of the initial value $X_i(0)$;

Step 2: calculate the fitness value $f(X_i)$ of each particle;

Step 3: comparing each individual extreme fitness value and the size of the particles, if the former than the latter, making it a new personal best;

Step 4: If the particle swarm best individual extreme due to the current global optimum value, write down the number of particles, adjust its position, then the optimum value as the new global;

Step 5: Update the velocity and position of each particle;

Step 6: determine whether the PSO stop iteration condition. If yes, stop the search, note the global optimum value, go to the fourth stage; otherwise go to step2.

3.2.4 BP neural network re-training

Step 1: PSO iteration of the global optimal value obtained corresponds to the BP neural network weights and thresholds;

Step 2: Use BP neural network algorithm to continue training until it reaches the target error.

4 Experiment and analysis

In order to prove that the improved classification algorithm results, using three kinds of algorithm PEPF, PSO-PF and PSO-PFNN in MATLAB 13.0 environment classified simulation.

Selected data sets are the breeding process, the production process and sales process for experimental details as shown in Table 1.

TABLE 1 The selected data set

Dataset	Samples number	Attributes number
breeding process	500	12
production process	1240	8
sales process	560	24

Based on the experience and experimental test case, the relevant parameters are set as follows: The number of particles $n = 100$, inertia weight $\omega = 0.92$. Neural network output layer weights in the range [-1, 1], the maximum iteration number is 5000.

Preclude the use of tri-fold cross-validation method to select the training and test samples. Before classification, the use of SNR method attributes data set dimensionality reduction. Relationship dimensions and classification error rate shown in Figure 3. It can be seen that the PSO-PFNN error rate compared to other three algorithms in most dimensions are the lowest.
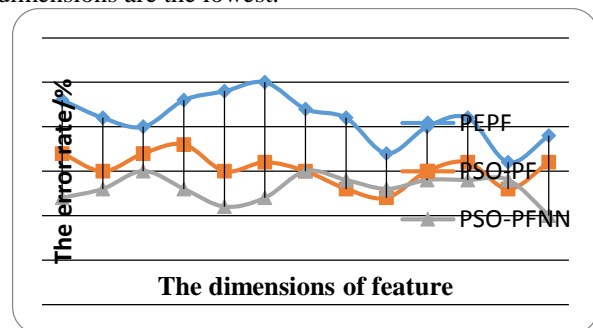
FIGURE 3 The relationship dimensions error rate

Three algorithms average classification accuracy on three data sets shown in Table2.

TABLE 2 Three algorithms average classification accuracy

Dataset	PEPF	PSO-PF	PSO-PFNN
breeding process	93.34%	95.12%	96.52%
production process	81.61%	83.31%	88.51%
sales process	73.12%	75.22%	79.68%

As can be seen from Table 2, both of which are experimental data sets for, PSO-PFNN average classification accuracy rate is the highest in the four algorithms, proved coupling prior information PSO-BPNN algorithm effectively improve the classification accuracy. These results prove that the use of a priori information Bayesian large sample extracted from a centralized data to improve search efficiency PSO algorithm is effective. Because PSO can more accurately search the global optimum value, so that the accuracy of the classification model trained network has been improved.

5 Conclusions

In this paper, the actual demand for the background, expand the research work for the lamb traceability network target tracking problem. Mainly is based on particle filter algorithm as the centre, research on distributed address tracing target tracking. Analysed the characteristics of a large sample and small samples, respectively, combined with Bayesian theory is to obtain the corresponding

representation of a priori information, and then called together into the PSO algorithm. The experimental results show that for any type of sample set, the guiding of prior information, which accelerates the convergence speed and improve the classification accuracy rate of neural network model.

Acknowledgments

References

[1] Jain A K, Duin R P W, Mao J 2000 *IEEE Transactions on Pattern Analysis and Machine Intellience* **22**(1) 4-37

[2] Zhang Y-H, Zhang F 2013 A New Time Synchronization Algorithm for Wireless Sensor Networks Based on Internet of Things *Sensors and Transducers* **151**(4) 95-100

[3] Ji Y, Zhang F 2013 Wireless Sensor Traceability Algorithm Based on Internet of Things in the Area of Agriculture *Sensors and Transducers* **151**(4) 101-6

[4] Sperduti A, Starita A 1993 Speed up learning and network optimization with extended back propagation *Neural Networks* **6** 365-83

[5] Lengelie R, Denceux T 1996 Training MLPs layer-by-layer using an objective function for internal representations *Neural Networks* **9**(1) 83-97

[6] He S, Wu Q H, Wen J. Y, Saunders J R, Paton R C 2004 A particle swarm optimizer with passive congregation *Biosystems* **78**(1-3) 135-47

[7] Monson C K, Seppi K D 2004 The Kalman Swarm: A New Approach to Particle Motion in Swarm Optimization *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2004)* Seattle Washington 140-50

[8] van den Bergh F, Engelbrecht A P 2006 A study of particle swarm optimization particle trajectories *Information Sciences* **176**(8) 937-71

[9] Ratnaweera A, Halgamuge S, Watson H C 2004 *IEEE Transactions on Evolutionary Computation Conference* **8**(3) 240-55

[10] van den Bergh F, Engelbrecht A P 2004 *IEEE Transactions on Evolutionary Computation Conference* **8**(3) 225-39

[11] Zahara E, Hu C-H 2008 Solving constrained optimization problems with hybrid particle swarm optimization *Engineering Optimization* **40**(11) 1031-49

Authors

Minning Wu, born in July, 1984, China

Current position, grades: Lecturer in Yulin University.
University studies: MS degree in Computer Application at Xidian University in 2010.
Scientific interests: computer application in agriculture, mobile internet technology.

Fei You, born in October, 1970, Yulin, Shanxi, China

Current position, grades: Professor at Yulin University.
University studies: PhD degree in Mathematical science from Beijing Normal University in 2005.
Scientific interest: fuzzy mathematics and artificial intelligence.

Feng Zhang, born in June, 1980, Yulin, Shanxi, China

Current position, grades: associate professor in Yulin University.
University studies: MS degree in Computer science from Xidian University in 2009.
Scientific interest: cloud integrated manufacturing technology, the modeling of complex systems, internet of things applications.

Operation Research and Decision Making

A system dynamics-based simulation experiment for aligning two anthrax progression models and their implications

Jiaxiang Hu*

College of Economics & Management, Nanjing Agricultural University, 1 Weigang, Nanjing, Jiangsu, China

Abstract

Two models have been proposed in the literature to describe anthrax progression - the first is referred to as Compartment-B model, which has 22 states, and the other is called Incubation–Prodromal–Fulminant (IPF) model, which has 9 states. How do these two models differ from each other in terms of the indicators considered important by policy or decision makers? Does one always outperform the other based on key performance measures? This paper describes our experience of aligning these two models in the context of anthrax attack. We first develop two simulation models using system dynamics to integrate the key indicators of emergency response, such as treatment rate, detection time, and treatment capacity. We then propose the process of model alignment and examine a large number of numerical examples to see whether the number of deaths, the stabilization time, and the demand for medicine produced by the two models will be reasonably equivalent. This study indicates that it is important for policy makers to understand the differences and similarities between the two models before making decisions. Furthermore, this research provides insights for scholars that rely on simulation tools for investigating bioterrorism attacks and for policy/decision makers that use these tools.

Keywords: Anthrax attack, Model alignment, Compartment model, Simulations, System dynamics

1 Introduction

Recently, the world has grown increasingly concerned about the threats posed by bio-terrorists. In various bioterrorism attacks, anthrax is often chosen for use [1], and was selected by former Soviet Union and the USA as a core microbe for weaponisation [2]. When a large-scale bioterrorist attack (such as anthrax) happens, it is essential to know the diffusion characteristics in order to improve the ability to handle it [3]. Gregory [4] developed a compartment model that has 21 compartments. Chen [5] established a simple Incubation–Prodromal–Fulminant (IPF) model, which has only 9 states. The simple IPF model is a simpler and well-understood model, while the compartment model is complex one, and is difficult to apply in the actual cases. As a result, whether the IPF model can describe the anthrax diffusion rule and whether it can replace the compartment model in coping with an anthrax attack is a research problem and has not been addressed in the literature.

Model alignment [6-7], also referred to as "docking", is the comparison of two computational models to see if they can produce equivalent results. System dynamics (SD) is an analytical modelling approach [8-10], and it deals with the broad behaviour of the system and how it influences its own evolution into the future. We modify the compartment model to include 22 compartments, which are referred to as the "compartment-B model" (B), and we rework the IPF model as "simple IPF model" (IPF). In this paper, our purpose is to demonstrate how to examine the general equivalence between the compartment model and the simple IPF model based on simulated anthrax attacks, and obtain implications from aligning these two models.

The remainder of this paper is organized as follows. Background information about the two models is given in Section 2. The comparison of the two modes and the results are given in Section 3. Finally, conclusions and future research are summarized in Section 4.

2 The Compartment-B and Simple IPF Models based on System Dynamics

2.1 THE COMPARTMENT-B MODEL BASED ON SYSTEM DYNAMICS

When an anthrax attack occurs, the population in that area is divided into mutually exclusive and collectively exhaustive compartments, and there are three disease stages: incubation, prodromal, and fulminant. In the incubation stage, an individual is infected with anthrax but is asymptomatic. The prodromal stage is when the disease is symptomatic with flu like syndrome. The fulminant stage is when the disease is severely symptomatic and is characterized by respiratory distress and followed by death within 24 to 48 hours. There are four categories of awareness and treatment: unaware of

---

* *Corresponding author* e-mail hujiaxiang2000@163.com

exposure, aware of exposure but not receiving prophylaxis or treatment, in prophylaxis, and in treatment. According to the three stages and four categories, any individual can fall into one of the 22 compartments (or states). Note that the first 21 states are established by the study of [4] and the last state – the 22nd state (death) is added in this paper. Individuals move between compartments according to a set of transition rates. We assume that R_{itoj} (i, $j \in (1, 2, 3, 22)$) is the transition rate from compartment i to j, the value of which can be shown in Table 1 [8-9]. In addition, the rate of entry into prophylaxis is assumed to be α_1, and the rate into treatment is assumed as β_1.

TABLE 1 All of the transition rates in the compartment model

Rate	Value	Rate	Value	Rate	Value
R1to4	0.014	R8to10	0.022	R16to17	0.00005
R2to5	0.014	R9to11	0.022	R6to22	1
R3to6	0.021	R13to14	0.024	R10to22	0.968
R19to20	0.021	R7to17	0.002	R11to22	0.968
R1to2	0.012	R8to17	0.0004	R12to22	0.968
R4to5	0.012	R9to17	0.0003	R14to22	0.968
R7to8	0.012	R13to17	0.00005	R15to22	0.968
R2to3	0.028	R14to17	0.00005	R16to22	0.968
R5to6	0.022	R15to17	0.00005		

According to the anthrax progression rule, the simulation model can be shown below:

$$\begin{cases} X_i(t+1) = X_i(t) + \sum_{j=1}^{22} R_{jtoi} X_i(t) - \sum_{i=1}^{22} R_{itoj} X_i(t) \quad i=1,2,3,...,22 \\ DP_1(t) = \alpha_1 * (X_4(t)*DPi + X_5(t)*DPp + X_6(t)*DPf + X_{20}(t)*DPi)*DPP \\ DT_1(t) = \beta_1 * (X_5(t)*DPp + X_6(t)*DPf + X_8(t)*DPp + X_9(t)*DPp + \\ X_{10}(t)*DPf + X_{11}(t)*DPf + X_{12}(t)*DPf)*DTT \end{cases} \quad . (1)$$

In the above equations, $X_i(t)$ is the number of people in stage i at time t. DP_1 (t) are denoted as the demand of medicine during prophylaxis, and DT_1 (t) are during treatment. DPi, DPp, and DPf represent a person's length of treatment in incubation, prodromal, and fulminant stages, respectively, DPP is the daily amount of medicine needed for treating one person during the prophylaxis period, and DTT are during the treatment period.

2.2 THE SIMPLE IPF MODEL BASED ON SYSTEM DYNAMICS

According to [4], the total population exposed to anthrax spores is divided into nine states, which are described as follows: E (exposed but not yet infected), I (incubation), P (prodromal), F (fulminant), IT (incubation with treatment), PT (prodromal with treatment), FT (fulminant with treatment), R (population that recover), and D (population that die). We can calculate the transition rate from the incubation to prodromal stage (Tip), from prodromal to fulminant stage (Tpf), and from fulminant to death stage (Tpd), and these are 0.012, 0.028, and 0.083[9-10]. The transition rate from the incubation, prodromal and fulminant stages to the recovery stage, named as Tir, Tpr, and Tfr, respectively, is set equal to zero [11]. Based on the

results published in [10], the recover rate from incubation ($Titr$), prodromal ($Tptr$), and fulminant stage ($Tftr$) are 1, 0.14, and 0.032. In addition, we define the transmit rate from the prophylaxis, prodromal and fulminant stages to the treatment stage are α_2, β_{2p} and β_{2f}, respectively, and the amount of medicine needed for prophylaxis and treatment are DP_2 (t) and DT_2 (t). Hence, we can establish the equations as follows:

$$\begin{cases} IT(t+1) = IT(t) + Tiit*I(t) - (Titpt + Titr)*IT(t) \\ P(t+1) = P(t) + Tip*I(t) - (Tppt + Tpf + Tpr)*P(t) \\ I(t+1) = I(t) + E*Tei - (Tip + Tir + Tiit)*I(t) \\ PT(t+1) = PT(t) + Tppt*P(t) + Titpt*IT(t) - (Tptft + Tptr)*PT(t) \\ F(t+1) = F(t) + Tpf*P(t) - (Tfpt + Tfd + Tfr)*F(t) \\ FT(t+1) = FT(t) + Tpft*F(t) + Tptft*PT(t) - (Tftd + Tftr)*FT(t) \\ D(t+1) = D(t) + Tfd*F(t) + Tftd*FT(t) \\ R(t+1) = R(t) + Tir*I(t) + Titr*IT(t) + Tpr*P(t) + Tptr*PT(t) + Tfr*F(t) + Tftr*FT(t) \\ DP_2(t) = \alpha_2 * I(t)*DPi*DPP \\ DT_2(t) = (\beta_{2p} * P(t)*DPp + \beta_{2f} * F(t)*DPf)*DTT \end{cases} \quad . (2)$$

2.3 THE PROCESS OF MODEL ALIGNMENT

We align the compartment-B model with the simple IPF model and compare the outputs from both models. To make our comparisons manageable and meaningful, we choose some example values for the following parameters: detection time (t_d) = 0, 48, 120, 240, treatment capacity (V) = 5000, 10000, 20000, α_1, β_1 = 0, 0.4, 0.8, 1, α_2, β_{2p}, β_{2f} = 0, 0.4, 0.8, 1, $S1$ = I = 100000, 200000, 300000, 500000. In addition, $Dpi = DPp = Dpf = 60$, $DPP = 0.2$, $DTT = 0.8$[11]. Thus, we can have 192 cases (4 x 3 x 4 x 4) of numerical problems. We align the components of the two models in the 192 numerical cases, and the model is implemented by a system thinking software, iThink 9.0. We define the following parameter, w, as the percentage of difference between the two models

$$w = (|IPF - B|/B)*100\% , \quad (3)$$

where IPF and B represents the output results, such as the number of death and the medical demand for treatment. According to Sterman [7] and Oliva & Sterman [12], the output results in the two models are the cumulative value over the entire time series.

3 Simulation results

In this section, the final outputs of the simulated attack, including death rate, stabilization time, and medical demand over time are presented and discussed.

3.1 DEATH RATE AGAINST TIME

3.1.1 The impact of the treatment rate

Table 2 shows the average number of deaths in the compartment-B and simple IPF models in a simulation scenario when the number of exposed people and the treatment rate change. In Table 2, E represents the exposed people, D is the average number of death, R is the treatment rate, B represents the compartment-B model, and IPF refers to the simple IPF model. Figure 1(a) shows the number of deaths against time with full prophylaxis and treatment, from which we see that about 95,342 people suggested by the compartment-B model would die, whereas only 834 people in the simple IPF model would die. This large difference is due to a time lag between the unaware exposed state and the aware exposed state, which delays many exposed people's timely treatment. If we suppose that all of the unaware exposed people search for treatment, we can arrive at a
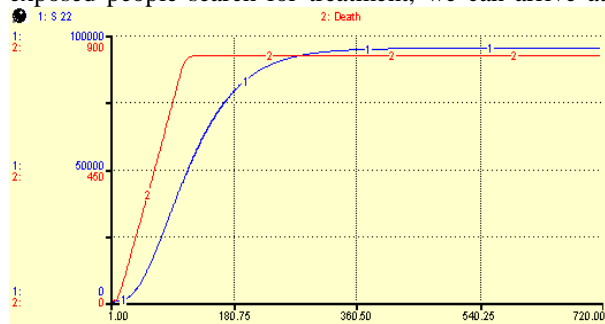
new plot shown in Figure 1(b), from which we see that the number of deaths suggested by both models is very close.

Table 2 reports the quantities of deaths suggested by both models as well as the percentages of difference when E = 100,000; 200,000; 300,000; 500,000, and R = 0, 0.4, 0.8, 1, respectively. As indicated by Table 2, almost all of the exposed people will die when the treatment rate is 0. When the treatment rate is 0.4, about 8.36% of people in the B model will die, whereas 7.28% of those in the IPF model will do. If the treatment rate is 0.8, then 6.86 % and 1.85% of exposed people in the B and IPF model respectively will die. Interestingly, at the full treatment rate (R = 1), 6.55% of people in the B model will die, but only 0.16% of those in the IPF model will die. Consequently, we conclude that the difference between the two models in terms of number of deaths increases with the treatment rate; especially when the treatment rate increase to 0.8 or even 1, the two models differ significantly.

FIGURE 1(a) The Number of Deaths against Time with Full Prophylaxis and Treatment
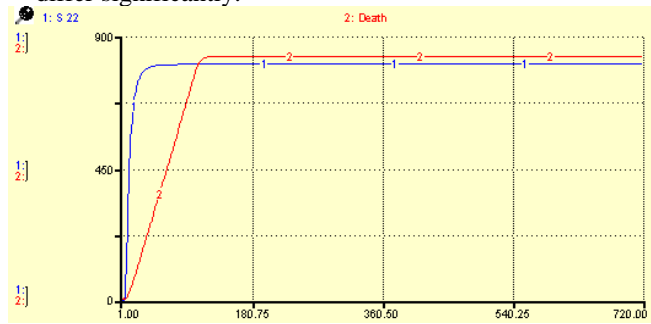
FIGURE 1(b) The Number of Deaths against Time with Full Prophylaxis and Treatment When all of the Unaware Exposed People Search for Treatment

TABLE 2 Average number of death under different treatment rates

	100000			200000			300000			500000		
	B	IPF	w	B	IPF	w	B	IPF	w	B	IPF	w
0	99942	99951	0.01%	199884	199902	0.009%	299827	299853	0.008%	499712	499756	0.009%
0.4	7843	6468	17.53%	16277	13882	14.71%	25304	22190	12.31%	45022	41600	7.6%
0.8	6674	1541	76.91%	13564	3444	76.41%	20660	5688	72.47%	35426	11167	68.47%
1	6550	149	97.93%	13101	317	97.58%	19652	486	97.53%	32753	823	97.49%

3.1.2 The impact of the detection time

We change the detection time to examine its impact on the death rate. As indicated in Table 3, on average 25.54%, 27.78%, 30.83%, 37.56% of the population suggested by the B model, and 25.67%, 26.36%, 27.36%,

29.86% of the same population in the IPF model will die when the anthrax is detected at 0, 12, 24 and 48 hours, respectively. In addition, we observe that the quantities of deaths produced by the two models are very similar if the attack is detected within 12 hours.

TABLE 3 Average number of deaths under different detection times

	100000			200000			300000			500000		
	B	IPF	w	B	IPF	w	B	IPF	w	B	IPF	w
0	25298	25354	0.22%	50894	51097	0.4%	76752	77178	0.56%	129135	130196	0.83%
12	27590	26054	5.57%	55498	52473	5.45%	83440	79229	5.05%	140078	133665	4.58%
24	30671	27085	11.69%	61520	54478	11.45%	92550	82163	11.22%	155114	138568	10.67%
48	37449	29616	20.92%	75004	59496	20.68%	112701	89646	20.46%	188587	150917	19.97%

Under the same assumption, we examine the impact of treatment capacity on the average number of deaths, and the relevant results are reported in Table 4. From this table we can see that the value of w are very close, which

suggests that increasing treatment capacity is not more effective and cost-effective at the margin to reduce the death rate.

TABLE 4 Average number of death under different treatment capacity

	100000			200000			300000			500000		
	B	IPF	w	B	IPF	w	B	IPF	w	B	IPF	w
5000	30326	27156	10.45%	60987	54832	10.09%	91960	83041	9.70%	154815	141170	8.81%
10000	30327	27007	10.95%	60656	54312	10.46%	91255	81860	10.30%	152941	137730	9.95%
20000	30194	26918	10.85%	60477	54015	10.69%	90867	81260	10.57%	151931	136110	10.41%

3.2 STABILIZATION TIME

We define stabilization time as the number of days elapsed when at least 99% of infected population either die or recover, and we obtain these under the different treatment rate, detection time, and treatment capacity, which are summarized in Table 5. We can see that the values of w are very large, implying that the two models are very different in terms of stabilization time when the treatment rate, the detection time and treatment capacity are considered simultaneously. In addition, Table 5 suggests that the stabilization time has a relationship with the treatment rate and the detection time; specifically, if the treatment rate is lower, the stabilization time will be longer. Additionally, if the attack can be detected earlier, the stabilization time will be shorter. However, we cannot find the relationship between the stabilization time and the treatment capacity.

TABLE 5 Stabilization time under different conditions

	100000			200000			300000			500000		
	B	IPF	w	B	IPF	w	B	IPF	w	B	IPF	w
0	384	360	6.25%	384	360	6.25%	384	348	9.38%	384	378	1.56%
0.4	240	118	50.83%	244	122	50%	268	130	51.49%	228	160	29.82%
0.8	242	108	55.37%	240	206	14.17%	252	130	48.51%	242	160	33.88%
1	210	94	55.24%	221	120	45.70%	234	124	47.01%	234	140	40.17%
0	212	152	28.30%	210	162	22.86%	228	164	28.07%	176	162	7.95%
12	254	162	36.22%	265	260	1.89%	274	174	36.50%	272	210	22.79%
24	286	174	39.16%	288	184	36.11%	292	188	35.62%	318	212	33.33%
48	324	192	40.74%	326	202	38.08%	344	206	40.12%	322	244	24.22%
5000	272	171	37.13%	271	255	5.90%	282	197	30.14%	273	225	17.58%
10000	265	170	35.85%	274	178	35.04%	287	179	37.63%	263	206	21.67%
20000	270	170	37.04%	271	172	36.53%	285	174	38.95%	281	191	32.03%

3.3 MEDICAL DEMAND AGAINST TIME

When the parameters such as the detection time, treatment rate and exposed people are changed, different average medical demands can be obtained and summarized in Table 6. It can be seen that when the treatment rate is increased to 1, the medicine for treatment in the IPF model reduces almost to 0. Since there are lots of people who aren't exposed but seek prophylaxis, the medicine demand for prophylaxis in the B model is higher than that in the IPF model. In addition, the medical demand for treatment in the B model is higher than that in its counterpart. All in all, the average quantities of medicine are totally different in the two models in all simulation cases.

TABLE 6 Medical Demand under Different Conditions

# of Exposed People	Treatment Rate	Compartment-B model		Simple IPF model		Comparison
		Prophylaxis	Treatment	Prophylaxis	Treatment	w
100000	0.4	15153881.08	100423.6	280128.8	72214.05	28.09%
100000	0.8	30309525.3	186825	630691.9	65316.81	65.04%
100000	1	37885494.53	220989.8	840000	0	100%
200000	0.4	15181983.34	225798	542529.51	173769.2	23.04%
200000	0.8	30384672.22	392978	1245563.65	157562.7	59.91%
200000	1	37985901.92	441979.5	1680000	0	100%
300000	0.4	15202172.03	372919.62	789662	300201	19.50%
300000	0.8	3086448.41	617196.49	1845563	274992	55.44%
300000	1	38086448.41	662969	2520000	0	100%
500000	0.4	15219257.96	720701.3	1245768.48	612308.99	15.04%
500000	0.8	30587078.94	1115901.04	3004362.63	579089.00	48.11%
500000	1	38287644.53	1104948.81	4200000	0	100%

4 Conclusions

In this paper, we develop a methodology to align two models of simulating disease progression after a biological attack. From this alignment study, we have shown that the average number of deaths and the stabilization time in the simple IPF model are comparable to the compartment-B model to some extent, on the condition that all of the unaware exposed people in the compartment-B model know they are exposed and seek prophylaxis and treatment. But if the detection lag is long and the treatment rate is large, the average number of death people and the stabilization time of the two models will be different. As for the medical demand, the two

models have a different performance because of their different model structures. In addition, the simulation results indicate that the longer detection time delay will result in more deaths, and increasing the treatment capacity is not more effective and cost-effective at the margin for the community.

It is necessary to point out some limitations of this research. Firstly, we can assume that different age group people will use different dose medicine in the two models. Secondly, we can consider that the transition rate from the incubation stage to prodromal stage in the two models followed a lognormal distribution with a mean of 10.95 days and dispersion factor $e^{-0.713}$ [13]. Thirdly, we can validate the two models in the actual case of anthrax attack in order to develop an efficient response and control strategies. All these areas represent our future research directions.

Acknowledgments

References

[1] Franz D R 2009 *Molecular Aspects of Medicine* **30** 503–10
[2] Guillemin J 2005 *Biological Weapons: From the Invention of State-sponsored Programs to Contemporary Bioterrorism* Columbia University Press: NewYork
[3] Hu J X, Zeng A Z, Zhao L D 2009 *Industrial Management & Data Systems* **109** 976–92
[4] Gregory S Z, Dena M B, Holty J C, McDonald K M, Owens D K, Brandeau M L 2008 *Medical Decision Making* **28** 332-50
[5] Chen L C, Kathleen M C, Douglas F, Boris K, Alex Y 2006 *Decision Support System* **41** 654-68
[6] Axtell R, Axelrod R, Epstein J M, Cohon M D 1996 *Computational and Mathematical Organization Theory* **1** 123-42
[7] Sterman J D 2000 *Business Dynamics: Systems Thinking and Modeling for a Complex world* McGraw-Hill Education: Boston
[8] Hu J X, Zhao L D 2011 *International Journal of Mathematics in Operational Research* **3** 490-509
[9] Forrester J W 1975 In *The impact of feedback control concepts on the management sciences*, ed Forrester J W (1975 collection) Wright-Allen Press: Cambridge, MA pp 45-60
[10] Hu J X, Zhao L D 2012 *ICIC Express Letters Part B: Applications* **3** 1209-16
[11] Rich K M 2008 *System Dynamics Review* **24** 67-96
[12] Oliva R, Sterman J D 2001 *Management Science* **47** 894-914
[13] Wein L M, Craft D L, Kaplan E H 2003 *Proceedings of the National Academy of Sciences of the United States of America* **100** 4346-51

Authors

Jiaxiang Hu, born on March 26, 1981, Henan

Current position, grades: Lecturer, College of Economics & Management, Nanjing Agricultural University
University studies: Doctor degree in system engineering from Southeast University in 2010.
Scientific interest: Supply Chain, Emergency management
Publications: 12 papers

Economic benefit evaluation model of E-commerce based on DEA

Xue Wu[1, 2*]

[1]*Business Administration College, Chongqing University of Science and Technology, Chongqing City, China*

[2]*International College, National Institute of Development Administration, Bangkok, Thailand*

Abstract

E-commerce, a new business transaction mode, becomes the key drive of regional economic growth. Without time-space limitations and with different trade mode, higher cost efficiency of commodity circulation and informationization degree of transactions, E-commerce plays an important role in enterprise, industrial and even regional economic development. Based on microscopic and macroscopic perspectives, economic benefit evaluation model of enterprise E-commerce, economic benefit evaluation model of industrial E-commerce and economic benefit evaluation model of regional E-commerce were established using DEA. A deep analysis was conducted to applications of the established models, including economic evaluation before and after E-commerce, economic evaluation of different regions within the same time period, and economic evaluation of same region at different time periods of E-commerce promotion. Meanwhile, three pieces of improvement advices were given to facilitate continuous E-commerce development and increase economic benefit of regional E-commerce.

Keywords: DEA model, E-commerce, economic benefit, evaluation model

1 Introduction

E-commerce is a kind of business communication generated with the rapid network development, is characteristic of high efficiency and economic returns as well as quick fund flow, because it goes over complicated procedures in common business process. E-commerce is now an important component of national economy and social informatization. E-commerce development is a big move to drive industrialization based on informatization, change economic growth mode, improve quality and efficiency of national economic operation, and take a new road to industrialization. It is of important significance to achieve the ambitious goal of building well-off society in an all-around way. Recently, China's E-commerce develops quickly due to IT development and popularity. It has achieved initial success and facilitated development of national economy informatization. However, China's E-commerce has smaller range of application, lower application level and poorer policy environment compared to those in developed countries. E-commerce is the commercial development trend in future and its economic benefit will affect stability and speed of regional economic development directly. Increasing economic benefit of regional E-commerce shall be the key goal of enterprises and the strategic goal of regional economic development. Therefore, it has important practical significance to establish and measure the economic benefit evaluation model of E-commerce. To avoid effect from other macroscopic factors, this paper established the economic benefit evaluation model of enterprise E-commerce firstly, which was expanded to regional E-commerce later.

2 Effect of E-commerce on enterprise and regional economic development

2.1 EFFECT OF E-COMMERCE ON ENTERPRISE AND REGIONAL ECONOMIC DEVELOPMENT

Firstly, E-commerce offers enterprises more opportunities. Different from traditional physical transactions in fixed place and time, E-commerce can make transactions wherever and whenever by taking full use of Internet. Compared to traditional transaction mode, E-commerce not only brings enterprises new vitality, bigger market and more trade opportunities, but also creates a free and fair competitive environment in international market. Enterprises can search business partners around the whole world and larger development space through Internet.

Secondly, E-commerce reduces transaction cost of enterprises, including marketing cost, purchase cost and office cost. Network marketing activity is superior to traditional advertisement. It can achieve 10 times better promotion effect and contribute far higher sales volume at only 1/10 cost of traditional advertisement. Additionally, the online customer support service can reduce telephone counseling significantly, thus saving considerable expenses and personnel investment. E-commerce can lower purchase cost of enterprises. Traditional raw material purchase is a complicated process. With the

---

Corresponding author e-mail: wuxue400020@163.com

Internet, enterprises can search cheapest supplier in the global market. Based on information sharing with suppliers, they can reduce losses caused by intermediate information loss. On the other hand, raw material purchase and product manufacturing can be combined organically into an integrated information transfer and processing system. E-commerce also can reduce office cost of enterprises. Cost for information transfer on network is lower than that through letter, telephone and fax. Moreover, the different transaction mode of E-commerce from the traditional one further reduces information sharing cost, including agency fee, advertisement and printing expenses, file processing fee, shop rent, etc.

Thirdly, E-commerce can improve production quality of enterprises.

1) It reduces product inventory. Long production cycle requires traditional enterprises to keep more inventories, thus resulting in their lagged response to market demand and higher operation cost.

2) It shortens the lead time of enterprises. Product design, development, manufacturing and marketing may involve many affiliated enterprises. Such independent cooperation can be changed into information-sharing collaboration in E-commerce.

3) It reduces intermediate links of product transaction. Business messages can be standardized by Internet and thereby can be delivered around the world and processed by computer automatically. E-commerce overcomes shortcomings of traditional trade mode, including high cost, frequent mistakes and slow processing. It shortens the whole transaction time significantly and changes the mode of whole social economic operation.

2.2 EFFECT OF E-COMMERCE ON REGIONAL ECONOMIC DEVELOPMENT

Online shoppers in China reached 0.24 billion by December, 2012, increasing by 48.07 million (24.8%) than that in 2011. The rate of online shopping increased to 42.9%. Although netizen growth slows down gradually, online shopping still presents strong growth momentum. The absolute user growth in 2012 exceeded that in 2011, 4% up compared to same period last year. Online shopping with mobile phone increased by 6.6% and the online shoppers with mobile phone was 2.36 times that by the end of 2011. The total E-commerce trading value quadrupled and the average growth of online retail sales was 80% in the past five years (Table 1 and Table 2). In Table 1, China's total E-commerce trading value in 2010 was 2.22 billion RMB, which increased to 97.78 billion RMB in 2013, increasing by 43 times. This is expected to reach 250 billion RMB in 2015. The mobile E-commerce trading value in 2010 was 1,800 billion RMB, which increased to 10,600 billion RMB in 2013. This is expected to reach 15,700 billion RMB in 2015. In Table 2, China's internet trades in 2008 was only 128.1 billion RMB, which, however, increased by more than 13 times in 2013, showing an annual average growth of higher than 60%. Such growth is expected to slow down, but still will keep at high growth rate. It is estimated that China's internet trades will reach 3,600 billion RMB in 2016. Correspondingly, the proportion of internet trades in total retail sales of consumer goods increased from 1.1% in 2008 to 7.7% in 2013. It is expected to reach 10.8% in 2016. To sum up, E-commerce is vital to China's regional economic development.

TABLE 1 E-commerce Trades in China

Year	2010	2011	2012	2013	2014e	2015e
E-commerce trading value in China	22.2	114.6	482.7	977.8	1686.5	2574.2
Mobile E-commerce trading value	4.8	6.3	8.4	10.6	13.0	15.7

Data source: iResearch statistics and predicted data, unite: 100 million RMB

TABLE 2 Online shopping development in China

Year	2008	2009	2010	2011	2012	2013	2014e	2015e	2016e
Online transactions (100 million RMB)	1281	2630	4610	7847	13040	18500	24500	30200	36000
Growth rate (%)	138.4	105.2	75.3	70.2	66.2	41.9	32.4	23.3	19.2
Proportion in total retail sales of consumer goods (%)	1.1	2.0	2.9	4.3	6.2	7.7	9.0	10.0	10.8

Data source: iResearch statistics and predicted data.

E-commerce Report of China (2013) reveals that China has become the largest online retail market in the world. As a new business mode, E-commerce propels transformation and upgrading of traditional economy by proving new service, market and economic organization form. It has become one important component of China's strategic emerging industry. The continuous rapid growth of E-commerce is manifested by both numbers and the industrial prosperity. It also increases benefits and efficiency of other industries. In capital market, JUMEI, Jingdong and Alibaba have been listed or prepare to be listed in the United States. They create rich mystery one by one. New industrial inventions are created from time to time. Recently, E-commerce is penetrating into traditional industry, agriculture, service industry to drive the transformation and upgrading of agriculture and traditional service industry. E-commerce industry and E-commerce economy are becoming links and core of modern industrial system. It is even called as the "potential quaternary industry".

3 Economic benefit evaluation model of E-commerce

3.1 DEA MODEL

DEA has been widely used in macroscopic and microscopic benefit model. Enterprise benefit is measured by input and output. Suppose is input, which has m

influencing factors; is output, which includes n products. Under certain conditions, some input will bring some output. Let h represent the economic benefit measurement index and be the weighting function of input and output. Enterprise benefit can be expressed as:

$$h = \frac{v_1 y_1 + v_2 y_2 + ... + v_n y_n}{u_1 x_1 + u_2 x_2 + ... + u_m x_m}.$$

This can be simplified into:

$$h = \frac{\sum\limits_{i=1}^{n} v_i y_i}{\sum\limits_{j=1}^{m} u_j x_j}.$$

(u, v) will change upon big environmental changes. Different combinations will contribute different outputs. Therefore, the enterprise benefit is changed into:

$$h' = \frac{v'_1 y'_1 + v'_2 y'_2 + ... + v'_n y'_n}{u'_1 x'_1 + u'_2 x'_2 + ... + u'_m x'_m}.$$

This can be simplified into:

$$h' = \frac{\sum\limits_{i=1}^{n} v'_i y'_i}{\sum\limits_{j=1}^{m} u'_j x'_j}.$$

Therefore, h of an enterprise or different enterprises in specific period can be known. Suppose there are p time points and q enterprises for comparison. The matrix about h can be gained:

$$W = \begin{vmatrix} h_{11} & h_{12} & ... & h_{1p} \\ h_{21} & h_{22} & ... & h_{2p} \\ ... & ... & ... & ... \\ h_{q1} & h_{q2} & ... & h_{qp} \end{vmatrix}.$$

Changing the environment and the matrix is changed into:

$$W' = \begin{vmatrix} h'_{11} & h'_{12} & ... & h'_{1p} \\ h'_{21} & h'_{22} & ... & h'_{2p} \\ ... & ... & ... & ... \\ h'_{q1} & h'_{q2} & ... & h'_{qp} \end{vmatrix}.$$

The regional economic development can be reflected by h of various local enterprises. The regional input and output can be acquired by calculating h of all local enterprises, thus establishing the evaluation model of regional economic benefit. Similarly, economic benefit of a specific industry in one region can be evaluated by calculating h of local enterprises within the industry. There are a lot of industries in one region and one industry covers many enterprises. The regional economic benefit can be calculated from economic benefits of local industries. Suppose there are k industries in one region and each industry selects g enterprises for economic benefit measurement. Then, the matrix of evaluation model of regional economic benefit is:

$$W* = \begin{vmatrix} h*_{11} & h*_{12} & ... & h*_{1g} \\ h*_{21} & h*_{22} & ... & h*_{2g} \\ ... & ... & ... & ... \\ h*_{k1} & h*_{k2} & ... & h*_{kg} \end{vmatrix}.$$

3.2 MODEL APPLICATION

Economic benefits of enterprise E-commerce and regional E-commerce can be analysed through historical data. Economic indexes of enterprise E-commerce can be acquired from three comparisons.

1) h of an enterprise before and after E-commerce. If h of the enterprise in five years before adopting E-commerce is smaller than 1.2, but h after adopting E-commerce is higher than 1.2, E-commerce brings it higher economic benefits. The difference between these two h can be used to estimate specific economic benefits of E-commerce.

2) h of product involved in E-commerce and h of product uninvolved in E-commerce. If there is no difference between these two h, E-commerce fails to bring the enterprise huge economic benefit. If h of product involved in E-commerce is larger than that of product uninvolved in E-commerce, E-commerce brings the enterprise actual benefit, which is reflected by difference between these two h.

3) Continuous economic benefit of enterprise E-commerce, that is, annual h of the enterprise after adopting E-commerce within the test period. If h of the enterprise increases continuously with the development of its E-commerce, E-commerce contributes great economic benefit. Difference of these annual h reflects the specific economic benefits of enterprise E-commerce. The greater the difference is, the huger the economic benefit will be.

Economic benefits of industrial E-commerce also can be analysed from three comparisons.

1) h of an industry before and after E-commerce. If h of the industry in five years before adopting E-commerce is smaller than 1.5, but h after adopting E-commerce is higher than 1.5, the industry wins economic benefits from E-commerce. The difference between these two h can be used to estimate specific economic benefits of E-commerce.

2) h of industry involved in E-commerce and h of industry uninvolved in E-commerce. If there is no difference between these two h, E-commerce fails to bring the industry huge economic benefit. If h of industry involved in E-commerce is larger than that of industry uninvolved in E-commerce, E-commerce can contribute actual benefit, which is reflected by difference between these two h. Since different enterprises have different economic benefits, enterprises can only be compared and

ranked relatively according to their h.

3) Continuous economic benefit of industrial E-commerce, that is, annual h of the industry after adopting E-commerce within the test period. If h of the industry increases continuously with the development of its E-commerce, the industry benefits a lot from E-commerce. Difference of these annual h reflects the specific economic benefits of industrial E-commerce. The greater the difference is, the huger the economic benefit will be.

Economic benefits of regional E-commerce also can be analysed from three comparisons.

1) h of local industries before and after E-commerce. If h of local industries in five years before adopting E-commerce is smaller than 1.4, but h after adopting E-commerce is higher than 1.4, the region wins economic benefits from E-commerce. The difference between these two h can be used to estimate specific economic benefits of E-commerce.

2) h of local industries involved in E-commerce and h of local industries uninvolved in E-commerce. If there's no difference between these two h, E-commerce fails to bring the region huge economic benefit. If h of local industries involved in E-commerce is larger than that of local industries uninvolved in E-commerce, E-commerce can contribute actual benefit which is reflected by difference between these two h. Since different industries have different economic benefits, industries can only be compared and ranked relatively according to their h.

3) Continuous economic benefit of regional E-commerce, that is, annual h of the region after adopting E-commerce within the test period. If h of the regional economy increases continuously with the development of E-commerce, the region benefits a lot from E-commerce. Difference of these annual h reflects the specific economic benefits of regional E-commerce. The greater the difference is, the huger the economic benefit will be.

4 Suggestions to increase economic benefit of E-commerce

The economic benefit evaluation model of E-commerce is established based on input and output analysis. In fact, (u, v) is another important influencing factor. Different (u, v) shall be supported with different $X(x_1, x_2, ..., x_m)$ and $Y(y_1, y_2, ..., y_n)$ to get the final maximum h. To increase economic benefit of E-commerce, it is suggested to:

1) Choose optimal input. Enterprise, industrial and regional input shall pay attention to proportion of input factors, for example, capital/human resources ratio. According to the production function, human resources can replace capital and capital, in return, can replace human resources. However, different combinations consume different costs even though they contribute same output. Similarly, it is necessary to choose optimal proportion of input factors in order to achieve maximum benefit at lowest cost.

2) Choose optimal output. One enterprise can produce various products. One industry can have enterprises of different types and scales. One region can cover industries of different types. Same input factors will generate different output combinations in different enterprises, industries and regions. For instance, an enterprise can produce product A and B. It has different output combinations (e.g. 10 A and 15 B) under different allocations of same human resources and capital. Since different output combinations have different benefits, output also can be maximized.

3) Perfect E-commerce market. Although E-commerce has achieved rapid development and the E-commerce market is perfecting continuously, it still has many challenges. Firstly, internet security shall be enhanced. Customer information on websites is stolen frequently in recent years, which highlights the security issues of Internet-based E-commerce. Secondly, key attention shall be paid to the credit and management of e-sellers. With the continuous expansion of E-commerce, trade disputes caused by E-commerce increases. How to deal with these disputes becomes a burning question to perfect E-commerce market.

References

[1] Yi X 2012 Study on the performance evaluation model of electronic commerce *Master thesis at Xihua University* 21-7
[2] Qiu L 2004 The Economic Analysis on Information Dissymmetry in E-bussiness *Technoeconomics & Management Research* **02** 30-31
[3] Xiao Y 2004 Discussion on the impact of e-commerce on enterprise management and innovation *Modern Information* **11** 189-91
[4] Chen J 2012 Research on the development of financial accounting based on electronic commerce network *Enterprise Economy* **06** 128-32
[5] Xiong K, Zhang Y, Zhang Z, Wang S, Zhong Z 2014 PA-NEMO: Proxy mobile IPv6-aided network mobility management scheme for 6LoWPAN *Elektronika ir Elektrotechnika* **20**(3) 98-103
[6] Zeng C 2007 Electrical computer business's safety problem and its solution method *Special Zone Economy* **09** 303-4
[7] Qi Y, Tang M, Zhang M 2014 Mass customization in flat organization: The mediating role of supply chain planning and corporation coordination *Journal of Applied Research and Technology* **12**(2) 171-81
[8] Zhang C, Huang L, Zhao Z 2013 Research on combination forecast of port cargo throughput based on time series and causality analysis *Journal of Industrial Engineering and Management* **6**(1) 124-34
[9] Chen S 2008 Analysis of China's electronic commerce mode *HLJ Foreign Economic Relations & Trade* **06** 66-7

Author

Xue Wu, born in January, 1981, Chongqing city, P.R. China

Current position, grades: lecturer at Chongqing University of Science and Technology, Chongqing, China.
University studies: PhD student at National Institute of Development Administration, Thailand, in China.
Scientific interests: the study of international trade and regional economic.
Publications: more than 20 papers published in various journals.
Experience: teaching experience of 6 years, five scientific research projects.

Operation Research and Decision Making

Research on outdoor wind environment of building groups based on computer simulation

Qian Yi[1]*, Shang Tao[1], Qingming Zhan[1, 2], Liming Bo[1], Jie Yin[1, 2]

[1]*School of urban design, Wuhan University, No.8, South Dong Hu Road Wuhan, Hubei, China*

[2]*Research Center for Digital City, Wuhan University, No.8, South Dong Hu Road Wuhan, Hubei, China*

Abstract

Building outdoor wind environment is closely related to indoor air quality and thermal comfort of human, directly affecting the health and quality of life of people living and the building energy consumption at the same time. Enhancing natural ventilation in summer also helps reduce air-conditioning equipment uptime, reducing air conditioning energy consumption and green building should be particularly emphasized natural ventilation. However, in the irrational layout of the buildings or the too high buildings and other factors, outdoor wind will bring no comfort of pedestrians, but also easily lead to loss of energy and increase heating energy consumption, especially in the winter. In this paper, with the careful study of local wind data, wind environment simulation and evaluation and optimized design are conducted about two cases including teaching buildings of Faculty of Engineering in Wuhan University, as well as the modern trade mart layout in Rizhao city of china by using computational fluid dynamics (CFD) technique. Research shows that: 1) Teaching buildings of Faculty of Engineering in Wuhan University have bad ventilation in summer, which cannot meet the green building standards in China. It will bring discomfort to pedestrians on the days of large wind speed in winter and need for windproof design. 2) Through wind environment simulations of modern trade mart plan in Rizhao city, planning of program two is more reasonable than program one in the ventilation by adjusting the building pattern. Using the technique of CFD simulation can guide the existing buildings renovation and architectural planning and design optimization.

Keywords: wind environment, natural ventilation, building environment, computer simulation, CFD

1 Introduction

Wind environment is the status of airflow into and out of buildings and its impact on them [1]. Wind can be a friend of a building because it can naturally ventilate the building, providing a comfortable and healthy indoor environment, as well as saving energy [2]. Natural ventilation can be used for cooling in the spring and autumn for a moderate climate; the spring for a hot and dry climate; the summer for a cold climate; and the spring and summer for a mild climate. Natural ventilation can also be used to cool environments in a hot and humid climate during some of the year [3]. What's more, natural ventilation also has an important significance in the hot summer and cold winter area, especially in the middle of China along the Yangtze River area, take Wuhan city of China as an example. In this area through the summer natural ventilation can take away the heat inside the building group.

On the other hand in the area of hot summer and cold winter wind can be an enemy to a building mostly in winter when it causes discomfort to pedestrians-usually as a result of high wind speed around the building. Table 1 summarizes the effects of wind on people [4]. Beaufort number classifies wind as 0 (calm) to 12 (hurricane). The wind speed is normally referred to as the speed of wind at 10m above an open terrain. The wind

speed at pedestrian level (about 1.5m high) is roughly 70% of the tabulated values. Visser [5] proposed comfort criteria with different activities versus the frequency of wind speed higher than 5 m/s, as shown in Table 2. For example, in an area where the number of day with an averaged wind speed higher than 5 m/s is 200 days per year (or the frequency of wind with a speed higher than 5 m/s is 200 days/365 days x 100 = 55%), people who walk fast would feel unpleasant. Clearly, wind speeds greater than 5 m/s are considered uncomfortable for most pedestrians. Therefore, it is very essential to reduce the wind speed around buildings.

In addition, hot in summer and cold in winter climate, it is very important to keep cold temperatures out during the winter by reducing wind speed and promote natural ventilation enhancing thermal comfort during the summer by increasing wind speed around buildings.

This paper takes the teaching buildings of Faculty of Engineering in Wuhan University and the modern trade mart plan in Rizhao, China as research objects and employs Computational Fluid Dynamics (CFD) technology to perform a digital analysis of the wind environment.

Corresponding author e-mail: qianyiwhu@gmail.com

TABLE 1 Wind effect on people

Beaufort no.	Description	Wind speed(m/s)	Wind effect
2	Light breeze	1.6-3.3	Wind felt on face
3	Gentle breeze	3.4-5.4	Hair disturbed; clothing flaps; newspaper difficult to read
4	Moderate breeze	5.5-7.9	Raises dust and loose paper; hair disarranged
5	Fresh breeze	8.0-10.7	Wind force felt by body; possible stumbling when entering a windy zoon.
6	Strong breeze	10.8-13.8	Umbrellas used with difficulty; hair blown straight; difficult in walking steadily; wind noise on ears unpleasant
7	Near gale	13.9-17.1	Inconvenience felt when walking
8	Gale	17.2-20.7	Generally impedes progress; great difficulty with balance in gusts
9	Strong gale	20.8-24.4	People blown over

TABLE 2 Coal property parameters and experiment ignition temperature

Activities	Acceptable (%)	Unpleasant (%)	Intolerable (%)
Walking fast: car-park, sidewalk, road, cycle-track	35	35-75	75
Strolling : park, shop centre, footpath building entrance, bus station	5	5-35	35
Sitting/standing short: shop centre, square, playground	0.1	0.1-5	5
Sitting/standing long: terrace, swimming pool, open theatre	0	0-0.1	0.1

2 Methods of analysing the wind environment

There are three methods of studying the wind environment around buildings: field measurements, wind tunnel testing, and Computational Fluid Dynamics (CFD) simulation [6].

2.1 FIELD MEASUREMENTS

The Field measurements are a method using spot investigation and first-hand information collection. But it has difficulties in precisely measuring factors of the wind environment and the long-term accumulation of observation in a large range. Yafeng Gao [7] presents a study on the effects of wind induced airflows through urban built form using statistical analysis. The data employed in the analysis are from the year-long simultaneous field measurements conducted at the University of Reading campus in the United Kingdom. Factor analysis of the measured data identified meteorological and architectural layout factors as key factors. The derivation of these factors and their variation with the studied built forms are presented in detail. Figure 1 shows a field weather measurement at the experimental site.

2.2 WIND TUNNEL TESTING

Wind tunnel testing is a method with a high degree of dependability. Gandemer [8] has simulated several aerodynamic effects that may occur around buildings depending on wind tunnel tests. Such effects can be descriptively used to interpret airflow behaviour in urban environments. For example, Ventura effect can be observed when a narrow distance between two buildings exists. In this case, airflow passing in this distance, known as Ventura nick, is restricted by buildings' walls. This increases its velocity until it leaves this narrow area, where airflow slows down. Another effect is the corner effect. Airflow becomes extremely turbulent at building corners, where the zones of positive and negative

pressure interact within a radius equal to building width. Kubota et al. [9] investigated the relationship between the average wind speed at pedestrian-level and building density of actual residential neighbourhoods depending on several wind tunnel tests. These residential neighbourhoods were selected to represent different building coverage ratios and different building heights. The study produced detailed data that emphasize the strong relationship between the coverage ratio and the mean wind velocity ratio, where increasing this ratio decreased wind speed on site. This helped in proposing some planning guidelines regarding the suitability of using detached houses or apartment buildings to achieve acceptable wind environment considering the climatic conditions of Japan. Figure 2 shows a site model placed in a wind tunnel.

2.3 COMPUTATIONAL FLUID DYNAMICS (CFD) SIMULATION

The numerical simulation of fluid flow, discretely analyses air flow by computer, following hydrodynamic equations. The results are expressed visually by computer graphics. Such numerical simulation technology is called Computational Fluid Dynamics (CFD). This technology has been widely used in manufacturing industries since 1974. And in recent years researchers have applied CFD to the simulation of building environments. Cheung JOP [10] explores the effects of building interference on natural ventilation using computational fluid dynamics (CFD) techniques. Building disposition is therefore one of the feasible solutions to improve the natural ventilation performance in our crowded environment. Sun Z [11] presents a CFD-based virtual test method for control and optimization of indoor environment by combining a ventilated room with a ventilation control system. Figure 3 shows CFD simulation figures. In this paper, CFD analysis was conducted by using PHOENICS software. PHOENICS [12] uses the standard k–ε model [13], staggered grid distribution [14], and simple algorithm. It can be used to simulate flow and heat

transfer with PHOENICS calculations. It is the world's first set of computing and computational fluid heat

transfer commercial software. Wind speed and wind pressure around buildings are calculated in this research.

FIGURE 1 Field weather measurement

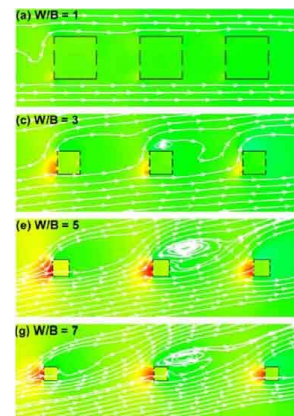FIGURE 2 Building model in a wind tunnel

FIGURE 3 CFD simulation figures

3 Simulation analysis of Outdoor wind environment of building groups

Outdoor wind environment design will be illustrated by two application examples. The first example concerns the teaching buildings of Faculty of Engineering in Wuhan University, and the second is modern trade mart plan in Rizhao city of china.

3.1 TEACHING BUILDINGS OF FACULTY OF ENGINEERING IN WUHAN UNIVERSITY

Wuhan is located in the middle and low latitudes, which has a subtropical monsoon climate, four distinct seasons,

abundant sunshine, hot in summer and cold in winter, frost-free period long. The annual average temperature of Wuhan is 16.2-16.7 degree, mostly northeasterly wind in winter and southeasterly wind in summer. According to Wuhan meteorological data, the annual average wind speed is 2.5-2.9 m/s, with an average wind speed up to 3 m/s or more in thirty percent days in whole year and average maximum wind speed of 14-19 m/s. The main monthly wind direction and frequency of Wuhan with many years' observations is shown in Table 3.

TABLE 3 The main monthly wind direction and frequency of Wuhan

Month	1	2	3	4	5	6	7	8	9	10	11	12	Whole year
Wind direction	NE	NE	NE	NE	NE	SE	WE	NE	NE	NE	NE	NE	NNE
Wind frequency(%)	18	19	16	12	10	9	10	14	18	17	20	20	14

The dominant wind direction in summer and winter can be obtained from the wind rose figure of Wuhan as Figure 4 shows. The typical working conditions of southeast winds in summer and northeast winds in winter were simulated by PHOENICS so that the Teaching buildings of Faculty of Engineering in Wuhan University could be analysed at a technical level. Figure 5 shows the CFD simulation wind direction layout.

Figure 6 shows a model of teaching buildings of Faculty of Engineering in Wuhan University designed by Jingtang He (Chinese Academy of Engineering). Since this campus building has windy surroundings and uncomfortable wind speed affect based on investigation of many students in campus, the outdoor wind comfort in the area is concerned. At one time the main building is too high (about height of 65 meters) which affect the landscape axis of Wuhan University. The University wanted to remove half floors of the main building. Since understanding possibility impact of wind environment on

the area after removing half floors, the University initiated a study of the wind distribution around the teaching buildings of Faculty of Engineering in Wuhan University.
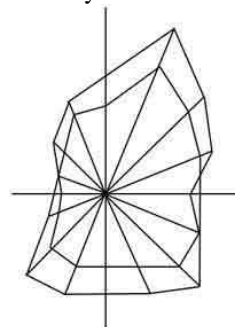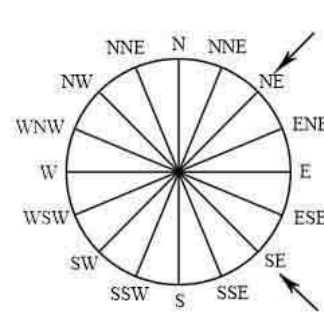

FIGURE 4 Wind rose figure in Wuhan

FIGURE 5 CFD simulation wind direction layout

FIGURE 6 Main teaching building (left) and its surrounding (right)

The models of the teaching buildings of Faculty of Engineering in Wuhan University can be established by AutoCAD. (In Figure 7, there is the original state on the left and the completion of removing half floors of the main teaching building on the right.) Figure 7 is the CFD simplified model of the teaching buildings and the main object is the wind environment in the teaching buildings. The boundary of the model is 1200m × 1200m × 200m; the model scaling is 1:1; the foundation is at the height of 0m.
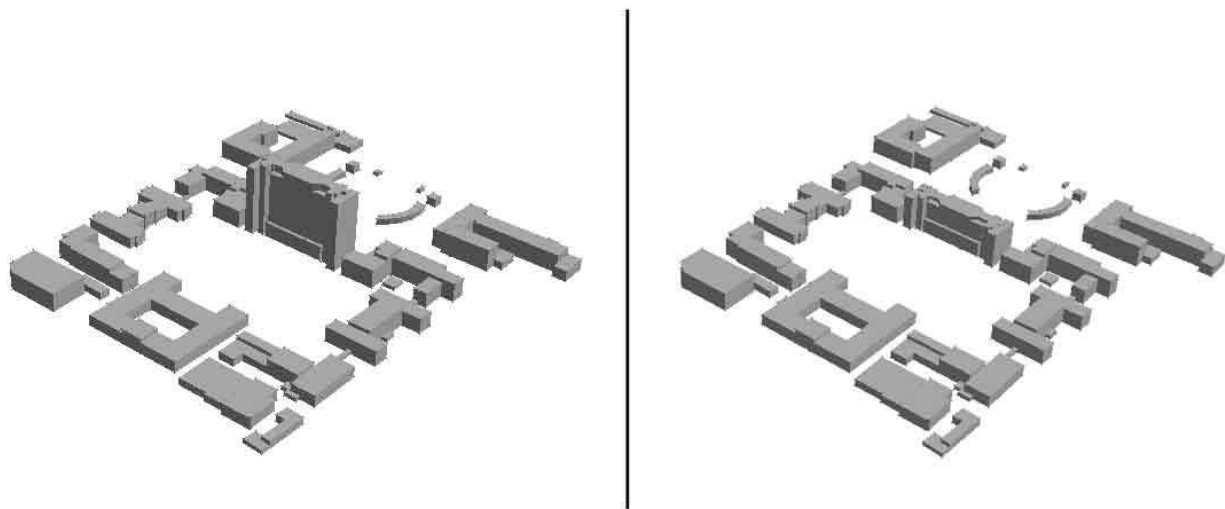

FIGURE 7 The simplified model made by AutoCAD

The reasonable boundary condition must be determined right after the confirmation of mathematical models and the governing equations so that the simulation will be as close to the real situation as possible. To analyse the wind velocity and wind direction in the teaching buildings of Faculty of Engineering in Wuhan University, the wind velocity and wind direction frequency graph mentioned above should be used to confirm the wind velocity and direction which are the input used for simulating the region. The inlet is defined as PHOENICS velocity-inlet boundary condition which is calculated under the typical working conditions in summer (southeast winds, 3m/s) and winter (northeast winds, 2.8m/s). The outlet is defined as the free outflow boundary condition.

Figures 8 to 11 are the wind velocity distribution of teaching buildings of Faculty of Engineering and after removing half floors of the main building in the summer and the winter at the height of 1.5m (pedestrians' height). As can be seen from the comparison between Figure 8 and Figure 10, the main building is too high that lead to bad ventilation in summer in the whole teaching buildings area. Wind speed around teaching buildings behind the main teaching building declines sharply and

wind shaded area becomes larger which is not beneficial to ventilation. The average wind speed of the whole area is larger after removing half floors (about 30 meters). Therefore, removing half floors of the main building is necessary that not only promote the teaching buildings' outdoor natural ventilation to improve comfort, but also can reduce the use of air conditioning in summer and save energy under the hot summer environment in Wuhan. In addition, with the comparison between Figure 9 and Figure 11, wind speed of the road area next to the

main teaching building where students often need go through is very large especially in the days of large wind speed of outdoor atmosphere. Wind speed ratio (Index of wind environment evaluation) around the main teaching building is almost up to two, which means wind in this area is too large causing discomfort of students in winter. Therefore, removing half floors of the main teaching building is beneficial for ventilation in summer and windproof in winter.
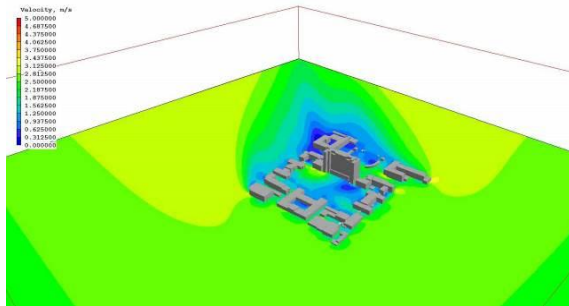

FIGURE 8 Wind velocity distribution of teaching buildings of Faculty of Engineering in summer at the height of 1.5m
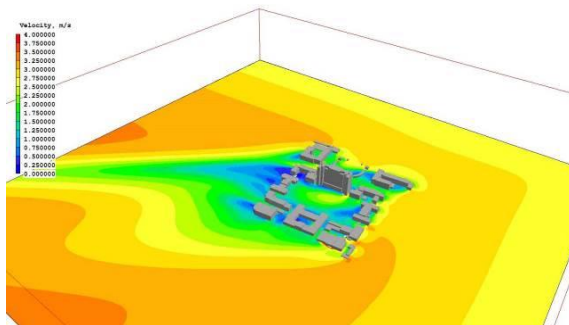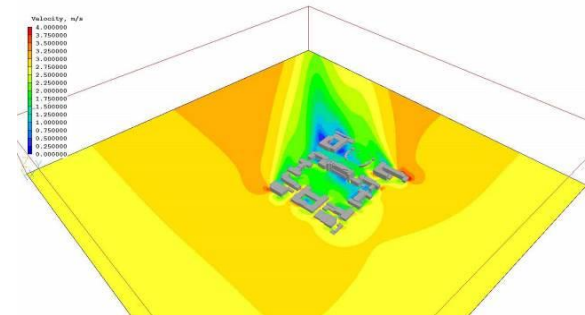

FIGURE 10 Wind velocity distribution of teaching buildings of Faculty of Engineering in summer at the height of 1.5m (after removing half floors)

FIGURE 9 Wind velocity distribution of teaching buildings of Faculty of Engineering in winter at the height of 1.5m
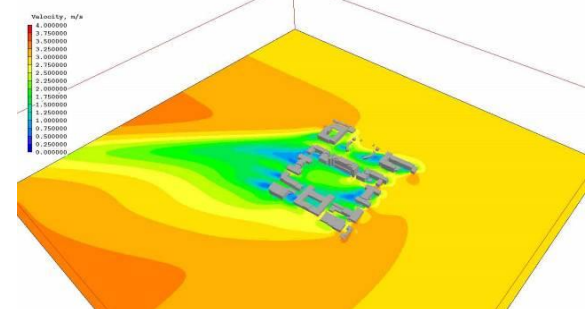
FIGURE 11 Wind velocity distribution of teaching buildings of Faculty of Engineering in winter at the height of 1.5m (after removing half floors)

3.2 MODERN TRADE MART PLAN IN RIZHAO CITY OF CHINA

The project of modern trade mart plan is a building bidding project in Rizhao city of china designed by the author and his design team. The project is a design of commercial complex, including hotel, commercial, office, residential. In order to obtain an excellent design scheme,

two design schemes are created. Since both of the two design schemes can meet the design requirements in building function and form, the author wants to conduct wind environment simulations using CFD to judge which scheme is better. Figure 12 shows the final design rendering of general lay out plan and three-dimensional bird's-eye view.

FIGURE 12 General lay out plan (left) and three-dimensional bird's-eye view (right)

Yi Qian, Tao Shang, Zhan Qingming, Bo Liming, Yin Jie

The models of modern trade mart plan one and two established by AutoCAD are seen in Figure 13. As can be seen from Figure 13, there are some differences between plan one and plan two in building form. Area of (A) of plan two cancels the shops along the street which makes area of (A) more open. Areas of (B, C, and D) of plan two remove the bottom two floors of the buildings which make the walking streets connectivity from south to north. Area of (E) slows down the polyline angle of commercial streets and the smaller angle of the building form is better to natural ventilation.

The inlet is defined as PHOENICS velocity-inlet boundary condition which is calculated under the typical working conditions in summer (southeast winds, 3.7m/s) and winter (north winds, 5m/s) based on wind rose figure of Rizhao city.

FIGURE 13 Models of modern trade mart plan one (left) and plan two (right)

Figures 14 to 17 are the wind velocity distribution of modern trade mart plan one and two in the summer and the winter at the height of 1.5m (pedestrians' height). As can be seen from the comparison between Figure 14 and Figure 16, the average wind speed of plan two is higher than plan one which means plan two has better ventilation than plan one in summer. In addition, the region of low wind speed of plan one is much more than plan two that will cause discomfort for customers or residents. Wind speed of region of (A) in plan two is bigger than plan one. Wind speed of region of (B, C, and D) and streets' wind speed in plan two are significantly increased. Therefore, by changing the buildings' form can promote the building's natural ventilation and improve comfort of people in summer. With the comparison between Figure 15 and Figure 17, wind velocity distribution of plan two in the region of (A, B, C, and D) and other regions is more uniform and better than plan one. Wind speed ratio (Index of wind environment evaluation) around the modern trade mart is less than two, which means wind in this region is reasonable and comfortable for customers or residents in winter. Therefore, changing buildings' form is beneficial for ventilation in summer and winter.

The wind environment distribution, the influence of the main teaching building and changing buildings' form on the nearby wind environment, and the space changing rule of an atmospheric flow field can be visually observed through simulating and analysing two cases in summer and winter. As a result, we can draw three conclusions:

1) Including the use of PHOENICS should become standard in the planning and construction of building groups. Because the wind environment affects energy use and wind comfort, the layout and positioning of buildings in architectural complexes should promote natural ventilation and increased air circulation.

2) There are a diverse range of wind zones in the leeward side of high-rise buildings, and on the windward side, there are high-velocity areas that have serious effects on the surrounding environment.

3) Changing layout and form of buildings' group have beneficial to wind environment that can reduce wind shaded area and make the wind velocity distribution of whole building groups uniform and reasonable.

FIGURE 14 Wind velocity distribution of plan one of modern trade mart in summer at the height of 1.5m
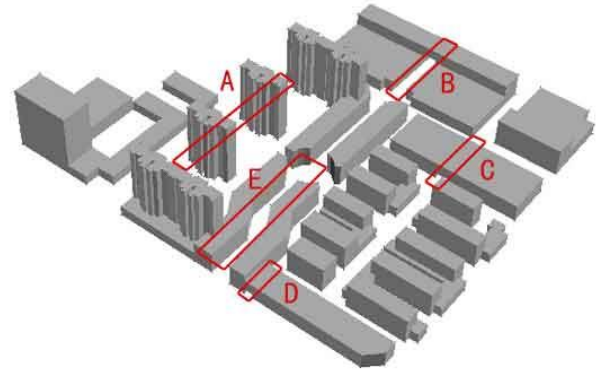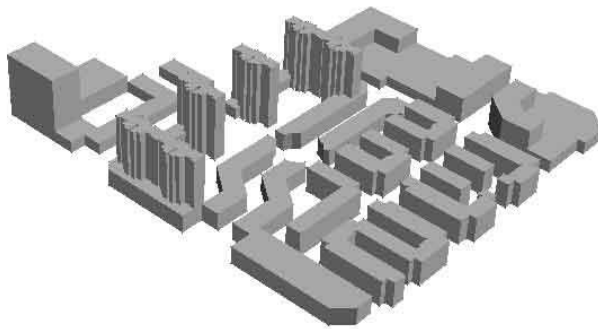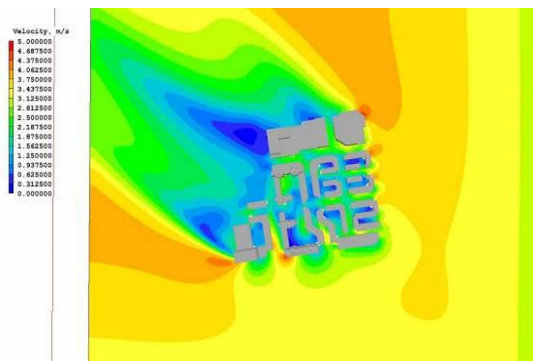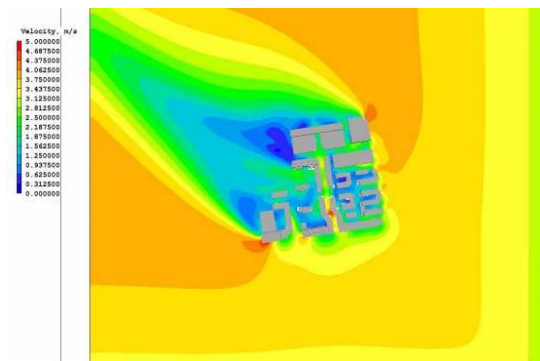

FIGURE 16 Wind velocity distribution of plan two of modern trade mart in summer at the height of 1.5m

FIGURE 15 Wind velocity distribution of plan one of modern trade mart in winter at the height of 1.5m
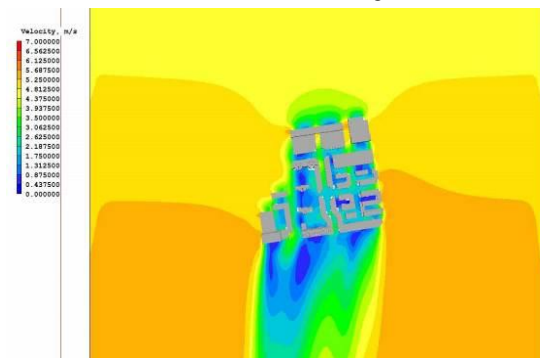

FIGURE 17 Wind velocity distribution of plan two of modern trade mart in winter at the height of 1.5m

4 Conclusions

This paper discusses the available methods for designing a building to take advantage of the wind, such as field measurement, wind tunnel testing, and computational fluid dynamics (CFD). CFD can provide detailed and useful information and is becoming an attractive and popular design tool. Simulation analysis of Outdoor wind environment of building groups illustrate how to collect wind information, develop different strategies for outdoor environment design, and use a CFD software PHOENICS to conduct the design. The outdoor wind design aims at developing pedestrian thermal comfort by varying building form.

In the environment, wind directly affects the lives of people. The wind environment is important not only to climate, but also to architecture's volume, layout, orientation, etc. Architects should factor the wind

environment into their designs. They can do this easily by computer simulations of the wind environment that use CFD software. Furthermore, the wind environment and its contributions to natural heating, cooling, and good air quality should be considered as part of the approval process of residential districts by city planning and city construction departments of local governments. Such considerations can lead to improvements in the quality of life for local residents.

Acknowledgements

References

[1] Boulard T, Haxaire R, Lamrani M A 1999 Characterization and modeling of the fluxes induced by natural ventilation in a greenhouse *Journal of Agricultural Engineering Research* **74** 135-44

[2] Chen Qingyan (Yan) 2004 Using computational tools to factor wind into architectural environment design *Energy and buildings* **36** 1197-209

[3] S Xiong K, Zhang Y, Zhang Z, Wang S, Zhong Z 2014 PA-NEMO: Proxy mobile IPv6-aided network mobility management scheme for 6LoWPAN *Elektronika ir Elektrotechnika* **20**(3) 98-103

[4] Bottema M 1980 Wind climate and urban geometry *Ph.D. Thesis, Eindhoven University of Technology Eindhoven The Netherlands*

[5] Visser GTh 1980 Windhindercriteria: een litertuuronderzoek naar envoorstellen voor het hanteren van unforme TNO-windhindercriteria (in Dutch) *Report 80-02746 IMET-TNO Apeldoorn The Netherlands*

[6] Li Y 2010 Computational fluid dynamics technology and its application in wind environment analysis *Journal of Urban Technology* **17** 67-81

[7] Gao Y, Yao R, Li B, Turkbeyler E, Luo Q, Short A 2012 Field studies on the effect of built forms on urban wind environments *Renewable Energy* **46** 148-54

[8] Gandemer J 1975 Wind environment around buildings: aerodynamic concepts *in: Proceedings of the Fourth International*

Conference on Wind Effects on Buildings and Structures Cambridge University Press Cambridge 423–32

[9] Qi Y, Tang M, Zhang M 2014 Mass customization in flat organization: The mediating role of supply chain planning and corporation coordination *Journal of Applied Research and Technology* **12**(2) 171-81

[10] Cheung JOP Liu C-H 2011 CFD simulations of natural ventilation behaviour in high-rise buildings in regular and staggered arrangements at various spacings *Energy and Buildings* **43** 1149-58

[11] Sun Z, Wang S 2010 A CFD-based test method for control of indoor environment and space ventilation *Building and Environment* **45** 1441-7

[12] Launder B E, Spalding D B 1974 The numerical computation of turbulent flows *Computer Methods in Applied Mechanics and Energy* **3** 269-89

[13] Patankar S V 1980 Numerical heat transfer and fluid flow *Taylor&Francis*

Authors

Yi Qian, born in December, 1986, Jiujiang City, Jiangxi Province, P.R. China

Current position, grades: the PhD candidate of School of urban design, Wuhan University, China.
University studies: his B.Sc. in Industrial Design from Northwest A&F University, M.Sc. from Wuhan University in China.
Scientific interest: architectural design, green building, and digital architecture.
Publications: 5 papers.
Experience: completed 3 scientific research projects.

Tao Shang, born in May, 1956, Xinyang City, Henan Province, P.R. China

Current position, grades: the professor of School of urban design, Wuhan University, China.
University studies: M.Sc. from Jilin University, PhD from Wuhan University in China.
Scientific interest: computer applications; heritage conservation.
Publications: 30 papers.
Experience: teaching experience of 30 years, 5 scientific research projects.

Qingming Zhan, born in September, 1964, Yongan City, Fujian Province, P.R. China

Current position, grades: the professor of School of urban design, Wuhan University, China.
University studies: M.Sc. from Wuhan University in China, PhD from Wageningen University in the Netherlands.
Scientific interest: new technology of urban planning, planning support system, urban environmental analysis and duct ventilation plan.
Publications: 50 papers.
Experience: teaching experience of 25 years, 10 scientific research projects.

Liming Bo, born in October, 1982, Qingdao City, Shandong Province, P.R. China

Current position, grades: the PhD candidate of School of urban design, Wuhan University, China.
University studies: B.Sc. in Urban Planning from Wuhan University, M.Sc. from Wuhan University in China.
Scientific interest: eco-city, green building, and green city planning.
Publications: 5 papers.
Experience: teaching experience of 3 years, 3 scientific research projects.

Jie Yin, born in September, 1987, Songzi City, Hubei Province, P.R. China

Current position, grades: the PhD candidate of School of urban design, Wuhan University, China.
University studies: B.Sc. in Urban Planning from Donghu University of Wuhan, M.Sc. from Wuhan University in China.
Scientific interest: urban sustainability, urban planning.
Publications: 5 papers.
Experience: 4 scientific research projects.

Operation Research and Decision Making

Research on the application of the Z-value analysis method in financial risk management of enterprise

Jinlin Zhou[*]

Department of economy and management, Henan Polytechnic Institute, Henan Province, China, 473000

Received 20 October 2014, www.cmnt.lv

Abstract

As a method of measuring the bankruptcy risk of enterprise invented by the US scholar Altman, the Z-value analysis method is widely applied by people. According to the prediction of the model, an enterprise will go into bankruptcy if the Z-value is less than 1.20; it is the grey area if the value is between 1.20 and 2.90; the enterprise has no bankruptcy risk if it is greater than 2.90. Although the Z-value analysis method has been used for measuring the bankruptcy risk of enterprises for a long time, in fact, it reflects the risks through the financial indicators of the enterprises, so it can be widely used for the financial risk management of enterprise. Through analysing the connotation of the Z-value analysis method, the paper explores the application of the method in the enterprise risk management from the specific calculation perspective and discusses the problems in application. Finally, some policy suggestions are provided for relevant decision-maker.

Keywords: z-value model, enterprise finance, risk management

1 Introduction

Risk management is a way of decision management based on prevention and enterprise risk avoidance. Through the management, it can avoid the enterprise risks in law, finance, social responsibility and investment, reduce the enterprises' loss in the abnormal cases and maintain the stable and sustained business activities. Therefore, for the large companies and those pursuing stability, risk management is particularly important. At the same time risk management is the decision based on facts, which needs to establish the information collection and response mechanism in the enterprise. In addition to the market information the financial information of the enterprise is also the core content. Financial risk is the inherent risk and the key risk of the enterprise. It is an important aspect of enterprise risk management. All along, business risks cannot be fundamentally eliminated but only be prevented and reduced. The Z-value analysis method is to measure the bankruptcy risk of enterprise. It uses the financial data to analyse the enterprise's financial conditions and predict the future possibilities. This paper applies the Z-value analysis method in the specific enterprise risk management and reflects the financial risks of enterprise through the financial data.

2 Relevant theories and research of the financial risks of enterprise

Financial risk of enterprise has been always a research focus which involves various aspects, but no any method can eliminate the risk. Meng Xiangxia combines the management theories of all kinds of enterprise financial risks and finds any theory of financial management has defects [1]. Ling Zhang holds that the financial risks of enterprise are characterized by uncertainty, dispersibility, transfer and decision making; they can be identified; the effective financial risk management can promote the value maximization of the enterprise [2]. Li Sheng argues that the overall financial risk management research is not only the necessary requirement for improving the business management of the enterprise but also the inevitable trend of the financial risk management theory development. With the development of global economic integration, the financing environment of the enterprises has been changing sharply. As a result, the financial risks of the Chinese enterprises are more diversified and complicated; at the same time, the degree of harm of the financial risks to the Chinese enterprises is also increasing, which becomes an important difficulty needing urgent solution for most domestic enterprises. The main purpose of the overall financial risk management is to realize the maximization of the enterprise value with the lowest financial risk. In order to achieve the purpose, it needs to establish a comprehensive and effective financial risk prevention and control system. Compared with the traditional financial risk management, it has the characteristics of systematicness, self-regulation, openness, sustainability and dynamics [3]. Yu Xinhua thinks that in the fierce market competition environment, the role of enterprise risk management is increasingly prominent, but the risk management of the domestic enterprise groups is relatively backward; with the intensification of competitions, the financial risks of those enterprises will be more and more complex and changeable [4]. Huang

[*]*Corresponding author* e-mail: Zhoujinlin678@163.com

Jinliang analyses the production and production process of financial risks [5]. Ren Yongsheng believes that in the market economy, the financial activities of enterprises will encounter all kinds of risks. Financial risks are objective, which cannot be completely eliminated. When pursing the business objectives, the enterprises should also take into account the risks. The full understanding and analysis of financial risks can reduce the negative impact on the enterprises to a minimum. With the access of our country into WTO, the enterprises are facing more fierce market competition. Once the financial risks appear, the enterprises will suffer heavy losses and even go into bankruptcy if handling them, improperly. In order to control the financial risks, the enterprises must start from the source. The first thing is to analyse the causes of financial risks. They should also have measures to control financial risks, so as to reduce the losses to a minimum when the risks occur [6]. Na Pengjie proposes the view that the financial risk management is the core of the risk management based on analysing the causes of financial risks of the enterprise groups and improves it from strategic finance, organization, internal control, early warning, assessment, information and culture [7]. Song Lei holds that as the core of enterprise risk management, the financial risk management of enterprise has caught redoubled attention from enterprises and governments, but the current financial risk management lacks systematicness and practicality; the financial risk management of enterprise is closely related to the diversification of the enterprise financing way, optimization of the enterprise investment strategy selection, maximization of the enterprise fund utilization and rationalization of the enterprise income distribution, but there are also great conflicts between them [8]. This paper makes an in-depth analysis of the application of the Z-value analysis method in enterprise risk management and discusses the application problems from dynamic management, prevention and prediction.

3 Application of the Z-value analysis method in enterprise risk management

3.1 THE CONNOTATION OF THE Z-VALUE ANALYSIS METHOD

As a method of measuring the bankruptcy risk of enterprise invented by the US scholar Altman, the Z-value analysis method is widely applied by people. According to the prediction of the model, an enterprise will go into bankruptcy if the Z-value is less than 1.20; it is the grey area if the value is between 1.20 and 2.90; the enterprise has no bankruptcy risk if it's greater than 2.90. Altman selected 33 bankrupted companies and 33 companies in operation as the samples and found after detection that the Z-value accurately predicted the fate of 63 companies. The research of the UK and Germany in recent years also proves the role of the Z-value in predicting the financial conditions of enterprise. Thus, the

Z-value analysis method has certain predictability of the enterprise risks.

The calculating formula of the Z-value designed by Altman in 1968 was:

$$Z = 0.012 * X1 + 0.014 * X2 + 0.033 * X3 + 0.006 * X4 + 0.999 * X5 \quad ,$$

where $X1$ is working capital/total assets; $X2$ is accumulated retained earnings/total assets; $X3$ is earnings before interest and tax/ total assets; $X4$ is market value of the owners' equity/total liabilities; $X5$ is sales revenue/total assets.

In 2000 Altman made the following modification on the Z discernibility function:

$$Z = 0.717 * X1 + 0.847 * X2 + 3.107 * X3 + 0.420 * X4 + 0.998 * X5 \quad ,$$

where the definitions of $X1$, $X2$, $X3$ and $X5$ remained unchanged; the calculating method of $X2$ changed; $X4$ is the book value of the owners' equity/total liabilities. Here, working capital referred to the narrow one, namely, working capital is current assets-current liability; accumulated retained earnings is surplus reserves plus undistributed profit; earnings before interest and tax is net profit plus interest expense plus income tax; sales revenue is revenue from operation.

3.2 APPLICATION OF THE Z-VALUE ANALYSIS METHOD IN FINACIAL RISK MANAGEMENT OF ENTERPRISE

There should be some changes if applying the Z-value analysis method in the enterprise bankruptcy to the financial risk management of enterprise. This is because the bankruptcy risk is the extreme risk, but for the general enterprises, the risks are reflected only to a certain extent, such as whether the financial risks of the enterprise increase and whether the risks are controllable compared with last year. There is a grey area in the enterprise bankruptcy risk: when the Z-value is between 1.2 and 2.9, the enterprise has no bankruptcy risk, but it does not mean the enterprise has no other risks. In fact, the enterprise has risks in all development stages, but at different levels. In the grey area between 1.2 and 2.9, the enterprise has high risks, but not the bankruptcy risk. When the Z-value is greater than 2.9, the enterprise also has risks, but at a low level. Thus, it can be seen that the Z-value calculated with the financial indicators reflects the financial risk of the enterprise to a certain extent. The greater the Z-value is, the lower the financial risk of the enterprise is; the smaller the Z-value is, the higher the financial risk is. In addition, the specific indicators for calculating the Z-value can show the rough sources of the risks.

$X1$ is working capital/total assets. Working capital is the capital that the enterprise invests in current assets, including accounts receivable, inventory, notes payable

and accrued expenses. It reflects the quota of the enterprise's current assets and current liabilities. Current assets minus current liabilities is working capital. Low $X1$ indicates the low proportion of working capital in total assets, which means that the enterprise has the higher risk of paying off its debts. Conversely, High $X2$ indicates the low risk of paying off the debts. The proportion of working capital in total assets can reflect the enterprise's ability to pay its short-term liabilities and the capital operation condition in the short term. The small proportion indicates that the enterprise has difficulties to pay the short-term liabilities. At this time, it needs to actively raise the short-term borrowings to reduce the financial risk. When $X1$ is negative, it means the enterprise cannot repay the short-term debts with current capital. In other words, it becomes insolvent. Then, the production of the enterprise may be interrupted at any time due to the capital turnover issue.

$X2$ is accumulated retained earnings/total assets. Retained earnings is created by the enterprise in the business process without the profit distributing to the owner of the enterprise. Accumulated retained earnings reflects the enterprise's profitability over the years and the profits drawn out for the enterprise's development. Retained earnings is the important source of capital of the enterprise. The utilization of it does not require the additional costs. It is immune to the market influence and restrictions. If retained earnings ratio is too low and the retained earnings is not enough for meeting the demand of enterprise development, the long-term shortage of funds may cause the enterprise to solve the development needs through getting into debt. Retained earnings is the cumulative net profit of the enterprise. The long-term accumulative quota can reflect the development condition of the enterprise. The higher the quota of retained earnings in total assets is, the faster the enterprise develops, the higher the profit is and the lower the financial risk is. If $X2$ has problems, the enterprise is faced with the long-term slow development. At this time, it should consider adjusting the business strategies.

$X3$ is earnings before interest and tax/ total assets. Earnings before interest and tax refers to the profit not deducting interest and income tax. In other words, it's the profit before paying the income tax under the condition of not considering the interest. It reflects the current actual profitability of the enterprise. The higher the proportion of earnings before interest and tax in total assets is, the higher the current profit ratio of the enterprise is and the lower the financial risk is. Conversely, the lower the proportion is, the lower the current profit ratio is and the higher the financial risk is. If $X3$ is low or negative, it means the enterprise has low profitability or suffers losses. Therefore, if $X3$ has problems, the enterprise needs to consider the balance between earnings and cost, reduce the business cost and improve the profitability.

$X4$ is market value of the owners' equity/total liabilities. The market value of the owners' equity refers to the value of the enterprise reflected in the market,

showing the public recognition of the enterprise. It's not directly related to the financial data, but the public recognition has great influence on the consumption of the enterprise products. The low market value of the owners' equity indicates the low public recognition. In this case, the future sales of the enterprise may have difficulties, which reduces its profitability and increases its financial risks. The high public recognition of the enterprise will promote the product sales and reduce the financial risks. In addition, the public recognition can also affect the enterprise's financing. The enterprises with the low recognition have greater difficulty in financing. If $X4$ is low, the enterprise should consider two factors: the first is the product quality, which is the key that affects the market; the second is the marketing ability and interpersonal skills. If the products of the enterprise have fatal accidents, it must reduce the market value of the enterprise sharply. If the products have the quality problem, the public acceptance must be low.

$X5$ is sales revenue/total assets. Sales revenue reflects the enterprise's market share and indicates its competitiveness and development speed. In general, the greater $X5$ is, the more the revenue of the enterprise obtained by utilizing the limited assets; the smaller $X5$ is, the lower the revenue of the enterprise is. It can be seen that $X5$ reflects the enterprise's operation results and using efficiency of assets. Thus, if $X5$ is too low, the enterprise should consider the problems of insufficient production, selling or product itself.

Given all that, it can be seen that the Z-value can reflect the financial risks of the enterprise and its specific calculating indicators can help find the sources of enterprise risks, like insufficient products, poor interpersonal skills, big gap between current capital and debts, etc.

3.3 THE RELEVANT PROBLEMS OF THE APPLICATION OF THE Z-VALUE ANALYSIS METHOD IN THE FINANCIAL RISK MANAGEMENT OF ENTERPRISE

According to the analysis, the indicators of the Z-values are mainly reflected in two aspects: the first is the ratio of current assets to current liabilities in the enterprise; the second is the short-term and long-term profitability of the enterprise. Specifically, there are mainly the following aspects:

First of all, the Z-value can reflect the enterprise's risks and their directions, but cannot reflect the sources. Both the ratio of assets to liabilities and the profitability are a general direction. In order to know the specific source of the financial risks, it needs a further analysis. Many factors contribute to the financial risks, such as low productivity, market strategy, level of employees and level of management. Therefore, after the Z-value estimates the risks and risk changes, the enterprise should know the reasons and make the analysis according to the actual circumstances.

Secondly, the calculation methods and indicators of the Z-value are simple, and the risks reflected are not complete. In fact, the sources of risks of many enterprises are undisclosed, or the enterprises cannot predict them in advance, such as sudden product accident, sudden jump in material prices, etc. The elements of those changes will cause the financial risks, but the Z-value cannot fully reflect all changing factors.

Thirdly, the Z-value analysis method is a dynamic analysis method, which needs the long-term tracking. The risks of the enterprise are changing, so is the Z-value. If the Z-value is smaller, it means the financial risks of the enterprise reduce, and vice versa. Any enterprise has certain risks, so it should make the dynamic analysis to predict the Z-value regularly and compare the values in different periods to get the changing rule rather than observing it statically. If the Z-value keeps falling, it means the financial risks of the enterprise are reducing; if the Z-value keeps rising, it means the financial risks are increasing. If the value fluctuates, the financial risks of the enterprise fluctuate. If the value fluctuates slightly, the risks are stable; if the value fluctuates widely, the financial risks of the enterprise rise and fall, which highlights the poor control and management ability of the enterprise.

In fact, the defects of application can be made up by changing the Z-value analysis method. The changing path needs the auxiliary conditions. Here, the principal component analysis method is applied to solve the defects. The details are as follows: according to the specific financial indicators of the enterprise, such as production cost, production efficiency, employee cost, employee efficiency, etc., it's assumed that there are n indicators, namely, $X1$, $X2$, $X3$, $X4$,...,Xn.

$$\begin{vmatrix} X11,X12,X13......X1p \\ X21,X22,X23......X2p \\ \ddot{X}n1,Xn2,Xn3......Xnp \end{vmatrix}.$$

Make the normalization processing to get:

$$\begin{vmatrix} Y11,Y12,Y13......Y1p \\ Y21,Y22,Y23......Y2p \\ \ddot{Y}n1,Yn2,Yn3......Ynp \end{vmatrix}.$$

Calculate the relevant coefficient matrix to get:

$$\begin{vmatrix} R11,R12,R13......R1p \\ R21,R22,Y23......R2p \\ \ddot{R}n1,Rn2,Rn3......Rnp \end{vmatrix}.$$

The characteristic value of the matrix is calculated by $|\lambda E - R| = 0$. Then, the proportion of each indicator can be obtained according to Equation (1). The changes of the proportions can show the specific sources of risks.

$$a = \frac{\lambda_j}{\sum\limits_{i=1}^{p} \lambda_i} \tag{1}$$

Therefore, the enterprise can know the degrees and changes of financial risks through the Z-value analysis method and the sources of risks through the principal component analysis method, such as production efficiency, production level, etc.

4 Conclusion and countermeasures

According to the above analysis, the Z-value analysis method can be effectively applied to the financial risk management of enterprise, but there are also many problems. In order to improve its practicality, it is recommended to strengthen the following aspects when measuring and analysing risks with the Z-value analysis method.

First of all, the enterprise should improve its ability of risk identification. The Z-value analysis method just reminds the enterprise of the existence of risks from the general direction, but cannot reflect the sources of risks. Risk identification is a kind of professional management work which should be attached great importance to. If the enterprise is unable to do it, it should hire the exports or intermediary agencies to make the principal component analysis. Otherwise, it cannot solve the risks even when finding them timely. At that time, it will be unable to achieve the goal, and the Z-value risk management will be useless. Thus, the Z-value analysis method must require the enterprise to have the supporting ability for assistance.

Secondly, the enterprise should develop the risk management strategies according to its development. In fact, the financial risks of the enterprise are mainly from technical and management levels of production and operation, sales ability and market competition, The enterprise should develop various risk management strategies according to its environment, like risk avoidance, risk transfer, risk conversion, risk hedging, risk compensation or risk control. It should also make preparations. Otherwise, even if the Z-value can show risks, it cannot take the effective measures timely to make risk management and control.

Thirdly, the top management of the enterprise should fully recognize the advantages and problems of the Z-value analysis method, improve their quality and enhance their ability of risk analysis and treatment. The change of the Z-value and the change and source of enterprise risks largely depend on the risk preference and attitude of the top management as well as their ability of risk treatment. Thus, it is the key measure for realizing the Z-value analysis method to strengthen their understanding and improve their ability; otherwise, it will end in talk.

Fourthly, the enterprise should strengthen the dynamic management and analysis of the Z-value analysis method, pay attention to the changes of the Z-value and coordinate the changes with its ability of risk analysis and control. Although the Z-value analysis method is simple, its application needs to keep consistency to accurately get the change of risks and

analyse the reasons. For example, if X1 remains unchanged in previous years but suddenly drops in recent three years, it indicates that the enterprise has some

problems with its working capital. Then, it should make adjustments in time.

References

[1] Meng X 2007 Thinking about the financial risk management theory research *Research of Financial and Accounting* **01** 40-1

[2] Zhang L, Liu J 2007 On financial risk's characteristics, management and prevention *Special Zone Economy* **04** 291-2

[3] Li S 2005 Research on overall financial risk management *Dissertation of master degree of XiangTan University* 26-9

[4] Yu X 2009 Enterprise financial risk management and control strategies *Friends of Accounting* **20** 291-2

[5] Huang J, Bai F 2004 Study on the basic framework of financial risk management *Research of Financial and Accounti*ng **06** 35-7

[6] Ren Y S 2007 Research on the enterprise financial risk management *Dissertation of master degree of Huazhong University of Science and Technology* 31-4

[7] Na P 2008 Some Thoughts on financial risk management of China's enterprise group *Inquiry into Economic Issues* **12** 142-8

[8] Song L 2011 Enterprise financial risk management shortcomings and Countermeasures *Research of Financial and Accounti*ng **22** 57-9

Author

Zhou Jinlin, born in October, 1971, Nanyang, Henan Province, P.R. China

Current position, grades: the vice-professor of department of economy and management, Henan Polytechnic Institute, China.
Scientific interest: Her research interest fields include accounting, auditing and tax.
Publications: 12 papers.
Experience: teaching experience of 7 years, 6 scientific research projects.

Operation Research and Decision Making

Research on internal control of accounting information in enterprises based on OPM model

Xihui Wu*

Department of economy and management, Henan Polytechnic Institute, Henan Province, China, 473000

Abstract

OMP model is an important computing model in the data field, which is broadly applied in in the field of physics, mathematics, etc. In fact, OMP can be adopted in enterprises as well. In this paper, through in-depth analysis on the calculation method and content of OMP, the application principles, specific application methods of OMP in internal control of enterprise accounting information and existing problems was analysed in accordance with its effect, so as to establish an OPM model software evaluation system from the practical perspective and ultimately propose corresponding countermeasures for the application of OMP model in internal control of accounting information in enterprises.

Keywords: SOFC, OMP, accounting information, internal control, evaluation

1 Introduction

Internal control of enterprises is a basic measure aiming at keeping the security of corporate assets in order to enhance the operational efficiency of materials. There are a lot of methods and measures of internal control, all of which contain the accounting information section. During the construction and improvement of internal control system, all enterprises should first strengthen internal mutual restraint and supervision of accounting organization, improve the quality of accounting, and realize supervision of enterprises through accounting supervision, so as to avoid moral risks in employee behaviours. Although enterprises are constantly strengthening internal control of accounting department, financial information leaking, accounting officers advancing private interests for themselves or other employees by irregularities are often the case, which are difficult to control. To this end, in December 2013, the Ministry of Finance issued <Working Rules for Enterprise Accounting Information>, of which major purpose is to protect the security of corporate accounting information. Accounting information is one of the core secrets of business that will directly affect their competitiveness, therefore is crucial to enterprises. This is also the practical significance of using OPM model to analyse the internal control and evaluation system of enterprise accounting information in this paper.

2 Relevant theories and researches

There are more researches on the corporate internal control, which mainly focus on the problems and countermeasures. Liu Xiaodong believed that internal control plays the role of internal factor during corporate developing process.

Internal factors determine the potential for development of enterprises, which are important criteria for the measurement of enterprise performance. However, imperfect legal person governance structure, inadequate risk assessment, weak awareness for enterprise risks and other issues exist in many enterprises. Zou Hui [1] analysed the problems of internal wage deduction with specific enterprise as an example, and argued that many enterprises in China have vigorous performance in internal control but with low effectiveness, mainly because of a lack of continuity in internal control of domestic enterprise internal control as well as a long-term and effective internal control system. Yang Xiaoling [2] analysed the environmental changes of enterprise internal control in the information environment and believed that enterprises in the information environment should pay more attention to the security of electronic information. Yao Feng [3] analysed the internal problems to be solved in enterprise, first of which is the internal control issue of finance. Li Youhua analysed the effectiveness of internal control of domestic enterprises from the perspective of validity control, and proposed suggestions for the evaluation of internal control. Zhai Yuechun [4] believed that in the process of enterprise informatization, information technology has impact on internal control in control environment, control object, objective setting, risk evaluation, control activities, information and communication, as well as supervision aspects. Zhang Xianzhi [5] believed that the publishing of supporting guidelines for basic norms of internal control of domestic enterprises has laid a foundation for the construction of enterprise, CPA, and relevant regulatory authorities three in one in the internal control system of enterprise. Taking into account the system background in China, limitations

---

and reliability of enterprise internal audit and other factors, on the basis of enterprise internal control disclosure and CPA internal control audit disclosure, indicators for the evaluation result of enterprise internal control reflecting comprehensive status and level of enterprise internal control is established. Wang Yu [7] argued that untrue accounting information quality has serious economic results. The key of ensuring information quality lies in external supervision and internal control, where internal control is the most important affecting factor. Wang Chen [8] believed that internal control system is designed for satisfying the organizational, business demand for management objective in modern enterprises, institutions and other related organizations. Internal control cannot offer certain helps for the management personnel to reach their objectives, but cannot provide absolute assurance [9]. As can be seen from the above, these analysis are mainly conducted from theoretical perspectives, which have defects in the application of model and data.

In order to effectively improve the capacity of internal control for accounting information of enterprises, a new model and method – OPM model and computing method is introduced here with in-depth analysis on its application of internal control for accounting information as well as evaluation on application.

3 Application of OPM model in internal control of accounting information

3.1 CONNOTATION OF OPM MODEL

OPM model and computing method is one of the sparse decomposition methods which decompose the signals. The computing method is as follows:

Suppose the signal to be represented is y, and its length is n. Assume that H refers to space. In space H, dictionary matrix D is composed by a set of vectors $\{X_1, X_2, X_3, ..., X_n\}$, in which each vector can be called a vector with the length same as the length n of signal y to be represented. And these vectors are normalized, namely $\|X_i\| = 1$, which means the length of unit vector is 1.

Basic ideas of MP model: select the best matching atom with signal y from dictionary matrix D (also known as over-complete dictionary of atoms), build a sparse approximation and obtain the signal residual. Then continue to choose the best matching atom with signal residual with reiteration. Linear approximation of signal y can be done by these atoms, which are represented by the final residuals. Obviously, if the residual value is within the range that can be ignored, then signal y is a linear combination of these atoms. Specific steps are as follows:

1) Calculate the inner product of signal y and each column of matrix dictionary, select the atom with maximum absolute value, and it will be the most matching one with signal y in this iteration. Describe with jargon: let the signal $Y \in H$, and select a most matching atom from the dictionary matrix satisfying

$\left|\langle y, X_{r0} \rangle\right| = \sup_{i \in (1,2,......k)} \left|\langle y, X_i \rangle\right|$, r_0 refers to a column index in the matrix dictionary. Thus, signal y is decomposed as the vertical projection and residual two parts of the most matching atom X_{r0}, namely $y = \langle y, X_{r0} \rangle X_{r0} + R_1 f$;

2) Residual $R_1 f$ is decomposed in the same way, then it can obtain in step K: $R_k f = \langle R_k f, X_{r_{k+1}} \rangle X_{r_{k+1}}$, where $X_{r_{k+1}}$ meets $\left|\langle R_k f, X_{r_{k+1}} \rangle\right| = \sup_{i \in (1,2,3,......k)} \left|\langle R_k f, X_i \rangle\right|$. As can be seen, after K steps of decomposition, signal y is decomposed as: $y = \sum_{n=0}^{k} \langle R_n f, X_m \rangle R_k f + R_{k+1} f$, where $y = R_k f$.

OMP is the improvement of OM method which conducts orthogonal processing for all selected atoms in each step of decomposition. Therefore the convergence rate of OMP algorithm is faster.

For example, a k-order model indicating situation of signal f after k steps of decomposition:

$$f = \sum_{n=1}^{k} a_n^k X_n + R_k f, with \langle R_k f, X_n \rangle = 0, n = 1, 2, ... K + 1 . \quad (1)$$

k+1-order model is as follows:

$$f = \sum_{n=1}^{k} a_n^{k+1} X_n + R_{k+1} f, with \langle R_{k+1} f, X_n \rangle = 0, n = 1, 2, ... K + 1, \quad (2)$$

Apply k+1-order model minus k-order and obtain as follows:

$$f = \sum_{n=1}^{k} (a_n^{k+1} - a_n^{+1}) X_n + a_n^{k+1} X_{k+1} + R_{k+1} f - R_k f . \quad (3)$$

Atoms in dictionary matrix D are non-orthogonal. An auxiliary model indicating the independence of X_{k+1} on the former k entries $X_n (1, 2, ...n)$ is introduced, as described below:

$$\sum_{n=1}^{k} b_n^k X_n + \gamma_k, with \langle \gamma_k, X_n \rangle, n = 1, 2, ...k . \quad (4)$$

Conduct orthogonal projection on the one of the span $(X_1, X_2, ..., X_k)$, and the latter entry is residual. This relationship is described by mathematical symbol as:

$$\sum_{n=1}^{k} b_n^k X_n = P_{V_k} X_{k+1}, and \quad \gamma_k = P_{V_k}^{X_{k+1}} .$$

Note that the subscripts a and b here refer to the value of k step.

Substitute Equation (4) into Equation (3) and get:

$$\sum_{n=1}^{k} (a_n^{k+1} - a_n^k + a_n^{k+1} b_n^k) X_n + a_n^{k+1} \gamma_k + R_{k+1} f - R_k f = 0 . \quad (5)$$

If the two equations are true, then Equation (5) must be true.

$$a_n^{k+1} - a_n^k + a_n^{k+1} b_n^k = 0 \,, \tag{6}$$

$$a_n^{k+1} \gamma_k + R_{k+1} f - R_k f = 0 \,. \tag{7}$$

Let $a_n^{k+1} = a_n^k$, then:

$$a_n^k = \frac{\langle R_k f, X_{k+1} \rangle}{\langle \gamma_k, X_{k+1} \rangle} = \frac{\langle R_k f, X_{k+1} \rangle}{\|\gamma_k\|^2} \,. \tag{8}$$

Ultimately, the convergence is obtained by computing matrix *a*. As a result, verification of two convergences exist here, one is basic data *X*, second is calculation of data *a*. If matrix *X* changes, then matrix a will change as well. If the final result is that both are converged, it indicates that the causes of change are internal. If not converged, then it means that the motivation of change comes from external force, namely abnormal effect. OPM is a static model, but can be applied to dynamic data, which is the process of continuous development for enterprises. Therefore issues of short-term convergence and relative convergence exist here. When the enterprise is very slow in short-term growth, a short-term positive relationship exists in various data of its production, which means the original matrix is multiplied or divided by a specific value. The current enterprise is not absolute rest. Even the enterprise does not develop, there are still differences in specific data, and so relative convergence exists. It means that if there are many data in enterprises which are all very large, a slight change in data will not affect its convergence value.

3.2 APPLICATION OF OPM IN THE APPLICATION OF INTERNAL CONTAL

Principles of OPM application: data information of enterprises is the most comprehensive, so the application of equation and model should be more convenient. But this must rely on a certain history to derive certain data law. OPM is to measure the operational conditions of enterprises through historical data and reach to a state of equilibrium. For example, for the production department, certain material costs, wages, and manufacturing costs of workshop form a certain amount of yield. If the production technology has not been significantly improved, then all data formed matrix should be converged without significant increasing trend. Changes in various costs of production can be designed as a set of vectors, and vectors in each time form the matrix. If the finally obtained value is converged, it indicates that there is no fluctuation in production and operation of enterprises. If matrixes *X* and *a* are unskilled, it means that fluctuations exist in corporate accounting information. The reason may come from the following aspects: first, the improvement of technology. For example, for the production department, the improvement of technology may lead to reduction in a part of costs, raw materials, labor costs, etc.; second, improvement of management level, and the outcome of management will eventually be presented as costs; third,

personnel issues - the financial personnel doing account may result in the increase of costs or decrease of efficiency. Therefore, it can be seen that OPM model ultimately conducts analysis on whether financial officer has violations through changes in costs and efficiency. The basis of its implementation it that corporate costs and efficiency will not have drastic changes in a short time.

Analysis on specific application methods of OPM model. Based on the above analysis, OPM calculation process is very complicated, but actually, it can be simplified by using computer even without additional manual operation. Specific steps of application are as follows. First, according to the accounting process and data situation calculation methods and enterprise accounting, OPM calculation software is compiled which should be able to connect with accounting software, in which data can be automatically input to the OPM accounting software. The calculation software should have the alarm function, which can rapidly alerts once there is a problem in calculation result and remind the accounting information managers to pay attention. Second, conduct normal accounting operation. The application of OPM model in accounting model is simple, but its effect is significant. Its calculation process is more complicated, but the operation is simple, as long as the manager conducts accounting in accordance with specific data regularly, the results will be remarkable, while the equation complexity and software confidentiality can effectively prevent violations of law issues within the enterprise. In general, manager of the software should be the executives of enterprises rather than accounting personnel, so that they can supervise from the accounting department. For accounting personnel who do not know the accounting process of software, they are not able to estimate whether entering untrue data will result in changes of software results.

Problems existing in OPM model. Separate OPM application model software development is relatively simple but requiring dock with the accounting software, which is more difficult. So whether OPM can be implemented also needs analysis by software programming experts, it is just assumed that it can obtain here. Second, changes in costs may come from many aspects. Employee turnover rate will change the operating level of corporate employees, thus affecting costs. The increase in prices of raw materials will lead to changes in corporate costs. Failure of strategies will also change the production costs of enterprises. Therefore, if cost increases but the OPM calculation result does not converge, there may be lots of reasons which are difficult to analyse specifically. So OPM can only alert the problem but is not able to discover specific problems. Third, software design is fixed, which may have adaptive issues over time. Fourth, the software should be first on the basis of a lot of past data, therefore only enterprises with longer years of development can use. The scope of application for the software is limited, which can only discover problems by trial for a longer time, therefore cannot have an effect

377

immediately. It is different from system control

4 Application of OPM model in the evaluation system of accounting information

OPM model application in internal accounting information mainly involves three aspects: first, correctness; second, sensitivity; third, time of persistence. Specific analysis is as follows:

First, evaluate the correctness of OPM model reaction through several commissioning. For medium-sized enterprises, corporate accounting division is very clear. There are many types of data, so the evaluation on its correctness must start from the possible risks of each position in the enterprise and debug the effect of all data changing. For example, turn the data of a position up or down while inputting to see whether it will alarm. If it alarms for all data commissioning, then the model calculation is correct. If there are 10 times failing to alarm among 1000 commissioning, then it indicates the presence of the correct rate is 99%; if there are 100 times failing to alarm, the correct rate is only 90%. The percentage of failure in commissioning may reflect the effect of OPM model.

Second, check the sensitivity of the verifier reaction by fine-tuning the data. Fine-tuning means to conduct several adjustments for a data with the same magnitude of each adjustment which is very small, like the number 10000 plus or minus 1 each time to test the sensitivity of the model. If it alarms when increasing by 10, the sensitivity on data is 99%; if it alarms when increasing by 100, its sensitivity on data is 90%. Sensitivity will not be higher with the increase, if it alarms for a little change, then it will alarm when using the model. However, the enterprise is developing, of which data is bound to change. Thus, over high sensitivity cannot play the genuine role of internal control. The sensitivity is appropriate by reacting only when the magnitude of changing is abnormal.

Third, regularly commission to ensure is use persistence. Enterprises are constantly evolving, so OPM may fail to adapt which needs regular commissioning. The regular basis can be one year, two years or three years according to the development of enterprises. Content of regular commissioning should include correctness commissioning and sensitivity commissioning. In addition, when major changes occur in the organizational structure,

market strategy, production technology, and management modes and so on, tests should also be done on the model. Any model has certain adaptability, whether it is able to adapt to all the enterprise development process is yet to be explored. Regular commissioning is an important measure to guarantee the long-term correctness and appropriate sensitivity of the model.

5 Countermeasures for the OPM model application

On the basis of the assumption that the software can be developed, in order to ensure the internal control role on accounting organizations, it should be conducted from the following aspects when enterprise is applying OPM model:

First of all, user of the OPM application software must not be the personnel of accounting agency, but should the executives or supervising department of enterprises. The data of the software comes from accounting, but the parts not belonging to accounting software can be used by non-accounting personnel. In order to ensure the effectiveness of supervision, relevant personnel of accounting agency can be informed that the software is operating, but they cannot be the manager, so as to safeguard the independence of supervision. Executives of corporate shall not be in charge of accounting, whom should be irrelevant with accounting in order to avoid the presence of conflict of interest. If independent supervision department exists in the corporate, then it can be implemented by supervision department, but can only be special personnel, which then accept the supervision of their superiors.

Second, OPM application model software should experience a longer probationary period like a year. This is because seasonal issues usually exist in the operation of enterprises. Only when the software adapts to all time including busy and off seasons and various periods with high or low staff turnover rate can the software be determined as applicable. If other time limits are discovered, then it cannot further be implemented.

Third, evaluation should be conducted regularly for OPM application model software during the trial, including evaluation of correctness and sensitivity. Through evaluation, the adaptability of all production in the enterprise can be fully understood and know whether it can accurately measure abnormal changes in corporate accounting information as well. If it can be satisfied in any time, then the adaptability is good.

References

[1] Liu X 2009 Problems and Countermeasures of enterprise internal control in China *Journal of Insurance professional College* **05** 55-7 *(in Chinese)*
[2] Zou H 2008 Research of application of enterprise internal control *Master degree thesis of Southwestern University of Finance and Economics* 26-7
[3] Yang X, Yang H 2010 The construction of enterprise internal control mechanism in information environment. *The ninth national accounting information annual meeting proceedings (down)* **9** 7 *(in Chinese)*

[4] Yao F 2006 Problem should be solved in enterprise inner control and its countermeasures *Special Zone Economy* **05** 308-9 *(in Chinese)*
[5] Li Y 2006 How to evaluate the enterprise internal control *Chinese financial newspaper* **03** 8 *(in Chinese)*
[6] Zhai Y 2010 Research on internal control in the process of the enterprise *HLJ Foreign Economic Relations & Trade* **06** 101-2 *(in Chinese)*
[7] Zhang X, Dai W 2011 Study on internal control evaluation system for Chinese enterprises *Auditing Research* **01** 69-78 *(in Chinese)*
[8] Wang Y, Hu R 2008 Under the environment of accounting information government research supervision and internal control of

accounting information *The international accounting and business conference set* **7** 287-9 *(in Chinese)*

[9] Wang C 2008 Study on the relationship between internal control system and accounting information quality *Journal of Shandong Youth Administrative Cadres College* **05** 99-103 *(in Chinese)*

Author

Xihui Wu, born in September, 1972, Nanyang, Hennan Province, P.R. China

Current position, grades: the vice-professor of the department of economy and management, Henan Polytechnic Institute, China.
Scientific interests: research and practice of accounting theory.
Publications: more than 16 papers.
Experience: Teaching experience of 7 years, 6 scientific research projects.

Research on product characteristics affecting the transformation of B2B E-commerce

Licheng Ren, Wenhui Pan*

School of Economics and Management, Taiyuan University of Science and Technology, Taiyuan City, Shanxi Province, China, 030024

Abstract

With the explosive growth of B2C and C2C Electronic Commerce (EC) in recent years, B2B EC has been tried in many fields, especially many traditional industries with little permeation into B2B EC, were transformed into B2B EC. However, the key factors influencing B2B EC transformation have not been fully investigated. On the basis of relevant literature analysis, product characteristics affecting B2B e-commerce transformation were divided into seven parts in this paper, namely, the product standardization, differentiation, tangibility, intangibility, time sensitivity, substitutability and complexity. Then, the proposed seven aspects were analysed using the method of the objective index weight, so as to obtain the mathematical model for product characteristics affecting B2B EC transformation. Finally, model test was conducted by taking Shanxi Pingyang Industry Machinery Co., Ltd. as an example, hoping the relevant research conclusion can provide references in decision-making for managers and other researches.

Keywords: B2B electronic commerce transformation, product characteristic, objective index weight

1 Introduction

Electronic commerce (EC) has experienced 20 whole years since 1994 [1], during this period, earth-shaking changes have taken place in various industries, enterprises transform from traditional business way to e-commerce gradually. As one mode of e-commerce, the most one difference for B2B EC to distinguish from other electronic modes is that its application range is between enterprises. In recent years, with the advantages of B2B e-commerce are more and more obvious, large amount of B2B online trade becomes a trend gradually. Under the circumstances, various enterprises, especial, which have little permeation into B2B EC, also begin to seek the way of B2B e-commerce transformation.

Product, as the fundamental source for enterprise profit, is the 'face' facing market. There has few scholars about the topics which kind of product is suitable for enterprise B2B e-commerce transformation, and what affect the transformation from traditional mode to B2B EC are seldom studied systematically at present stage. B2B e-commerce transformation present situation and product characteristics elements were in-depth analysed in the paper, then building decision-making model about product characteristic affecting B2B e-commerce to decomposes and measures various product characteristic, hoping the relevant research conclusion can provide references in decision-making for enterprises.

2 Related work

2.1 RELATED REVIEW ABOUT B2B E-COMMERCE TRANSFORMATION

The existing literature about e-commerce transformation offers a fairly substantial theoretical basis for this paper. Davison (2005) [2] defines enterprise transformation as a way to improve traditional business efficiency and effectiveness through the application of information and communication technologies; Winter (2001) [3] et al. argue that enterprise e-commerce transformation can be caused by the innovation of information technology and the emergence of new business model; Cuixiao Fu, Min Qin, Lihua Huang (2011) [4] et al. combine e-commerce transformation with trilogy of e-commerce model, hold that e-commerce transformation also need to experience the process of 'unfreeze-change-refreeze'.

About the literature about B2B e-commerce transformation, Cuixiao Fu, Lihua Huang and Qingfeng Zeng (2010) [5] suggest it a transform process for enterprises from current operation mode to B2B e-commerce mode; Qingfeng Zeng (2005) [6] also suggests it a process that reshapes enterprise organization behaviour in order to adapt new environment and achieve sustainable advantage, strategic, dynamic, systematic, gradation and innovative; Lei Nie, Cuixiao Fu, Lihua Huang hold enterprise transformation to B2B platform type will be affected by IT ability, B2B electronic service ability and B2B EC transformation ability when considering from enterprise ability perspective. Among

---

Corresponding author e-mail: 975424374@qq.com

Operation Research and Decision Making

these abilities, IT ability including IT infrastructures, IT management and IT match ability with business; B2B EC transformation ability are divided into enterprise and vision transformation ability, organization structure transformation ability, product and market transformation ability, business process transformation and cultural transformation ability.

In conclusion, the research is about the influence on B2B e-commerce transformation in the exploratory stage, the majority of which consider from ability perspective, so as to unable to provide effective theoretical guidance and method support for enterprise B2B e-commerce transformation adequately.

2.2 RELATED REVIEW ABOUT PRODUCT CHARACTERISTICS

Product characteristics point to the external and internal features which form product final quality, every feature affects its performance [8]. Early studies about product mainly focus on three points of view. Firstly, research about supplier selection, T.M.Laseter & K.Ramdas [9] find product characteristics have significant effect on supplier selection; secondly, research about the match between product characteristics and supply chain, Fisher, as the typical representative researcher, put forward classical matching model, mainly explored the most important obstacle for product characteristics and supply chain match smoothly is low enterprise performance level; thirdly, research about how product characteristics affect sales level, R.G.Javalg, I.D.S.White & O.Lee [11] hold product characteristics have big influence on export way and then affect sale capacity.

With the development of e-commerce, some scholar literature of the relationship between product characteristics and e-commerce focus on B2C EC. Qixiang Jin (2007) [12] find the degree of digitization, information content, product characteristics, time-sensibility affect accepting different operating mode under e-commerce sales environment; Zhenhua Wang & Huanchen Wang (2002) argue consumers purchase product mainly considering from its 14 features including descriptive, absolute value et al.

To sum up, product characteristics have important influence between enterprises, however the topic mainly focus on B2C and C2C e-commerce, thereby lacking of in-depth study in B2B e-commerce. The paper analyses the influence on B2B EC transformation from product characteristics perspective, in order to provide theory and method references for enterprises.

3 Product characteristics analysis and indicators construction

3.1 PRODUCT CHARACTERISTICS ANALYSIS

Standardization and differentiation. Product having strict specification and unified standard can be fully showed through information such as standardization and specification, so that standard product have low cognitive uncertainty risk for customers. Product which satisfy specific need belongs to personalized products and have high degree of differentiation in B2B e-commerce transaction, leading to communication between enterprises is still a must under Internet trading mechanism for transaction cost of B2B EC don't reduce effectively based on network mechanism but little profit for enterprises. Now, the research about enterprise information are still at primary stage, standardization and differentiation affect the transformation of B2B e-commerce transformation. In 2003, Sculley & Woods investigated different industries transformation, the result indicated automotive industry, information electron industry, pulp and paper industry are the most successful three industries, the product of which are physical and available for bulk production; on the other side, the most lowest industries are health care etc., which has poor standard, customer different demand becomes the driving force to transform in these enterprises. Therefore, standardization and differentiation are two of the most important characteristics affect B2B e-commerce transformation.

Tangibility and intangibility. Tangible product mainly mean visible physical product like coal, steel etc., while intangible product mean invisible product like information, service, electronic and products etc., the two parts both have own strength in e-commerce field. First, risk aversion type enterprises hold conservative attitude toward network transaction considering from the perspective of transaction, leading to tangible product become the first choice for enterprise transaction; secondly, information flow and cash flow can be conveyed through Internet for tangible product from logistics perspective, but product delivery process need logistics support. Comparing with tangible product, the superiority of intangible product reflects on the channel convenience, product can be conveyed by network directly so that shorten logistics time and reduce cost greatly. Therefore, tangibility and intangibility are two of the most important characteristics affect B2B e-commerce transformation.

Time sensitivity. The influence on B2B e-commerce transformation from time sensitivity perspective mainly embodies on two aspects, namely, the product life cycle and update speed. First of all, short life cycle product are suitable for close deals in the same market in general considering from the perspective of market range, small market scope and relatively stable customer group give priority to shorten logistics time-consuming; secondly, large range of market and customers scope are needed for low update-speed product to sale in short time, so as to finding customers through Internet becomes important, leading to B2B e-commerce become an option in this case for its advantage that gather customers quickly and enhance the effectiveness of transaction. Therefore, time

sensitivity is one of the most important characteristics affect B2B e-commerce transformation.

Substitutability. Product substitutability includes function alternative, performance substitutability, stability substitutability, relevance substitutability and experience substitutability. Now traditional marketing channel can meet enterprises' demand in a vast extent, making B2B e-commerce shelved but reduce interests due to various input of B2B EC. Comparing with low substitutability product have relatively small but stable customer group, high substitutability product have large market demand because of relative low price. B2B e-commerce provides a method to expand customer group for high substitutability product to realize small profits but quick returns. Therefore, substitutability is one of the most important characteristics affect B2B e-commerce transformation.

Complexity. Product complexity embodies in two aspects: firstly, the complexity of product attributes, including system complexity, interaction complexity with outside enterprise, product technology complexity, manufacturing process complexity and project management complexity. Secondly, high product function complexity. In the process of B2B electronic trading, the question whether product can be accepted quickly by customers regardless of the unalterable fact that reflecting different aspects of complicated product fully is hard to address. The higher enterprises product complexity, the

higher barrier into the industry, and the less competitors, in this condition, outside pressure fails to lead to B2B e-commerce transformation for enterprises fully; Conversely, the lower product complexity, the lower barrier into industry, the more competitors for enterprises. Enterprises have to transform to B2B e-commerce due to outside pressure. Therefore, complexity is one of the most important characteristics affect B2B e-commerce transformation.

In short, product characteristics determine industry characteristics and have large influence on enterprises B2B e-commerce transformation, the proposed seven aspects are analysed hoping to get further results of B2B e-commerce transformation.

3.2 INDICATORS CONSTRUCTION

It is hard to measure the seven parts of product characteristics above affecting B2B e-commerce transformation directly these factors accurately, the research indicates higher market share and striving for more customers are the direct and fundamental purpose of B2B e-commerce transformation based on relative reference, therefore, measuring item considering from two perspectives of market and customers are designed in the paper to measure product characteristics, as shown in Table 1.

TABLE 1 Measurement Index System of Product Characteristics Affecting B2B E-commerce Transformation

Product characteristics (level indicator)	Order number	Measuring number
standardization	x_{11}	The internal information(such as function, style, material, etc.) of the same type product has unified standard
	x_{12}	The external information(such as price, color, etc.) of the same type product has unified standard
differentiation	x_{21}	The main product have high concept differentiation
	x_{22}	The main product have high prosperity (such as features, quality, function and style, etc.) differentiation
	x_{23}	The main product have high consumption (such as personnel and consumption process) differentiation
	x_{24}	The main product have high brand (such as personality and users) differentiation
tangibility	x_{31}	The main product is tangible, easy to see and touch
	x_{32}	The main product need to be delivered by the third logistics
intangibility	x_{41}	The main product is intangible, including digital product, information and service, etc.
	x_{42}	The main product need to be delivered through online service
time sensitivity	x_{51}	The main product has long product life cycle
	x_{52}	The main product has faster update speed
substitutability	x_{61}	The main product has high function substitutability
	x_{62}	The main product has high performance substitutability
	x_{63}	The main product has high stable substitutability
	x_{64}	The main product has high associated substitutability
	x_{65}	The main product has high experience substitutability
complexity	x_{71}	The main product has complex components and diversified forms
	x_{72}	The main product has complex function
	x_{73}	Enterprise trades main product via the Internet, the process is complex

4 The decision-making model of product characteristics affecting B2B EC transformation

Based on previous analysis, decision-making model of product characteristics affecting B2B EC

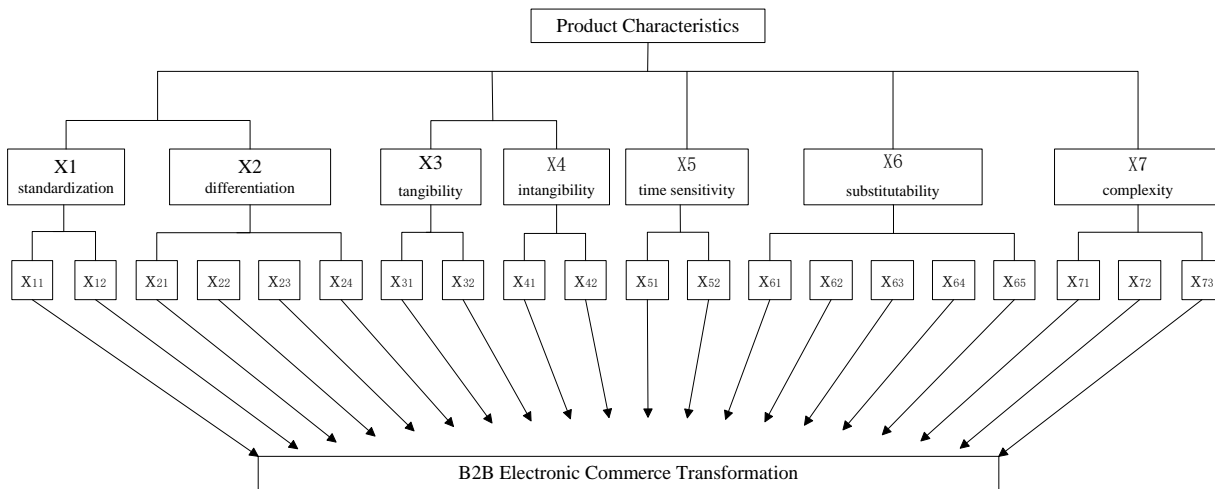transformationare designed combining with Table 1, as shown in Figure 1.

FIGURE 1 Decision-making model chart

The most common method to affirm target weight is Analytic Hierarchy Process (AHP), but in the paper, the weight of product characteristics indicator cannot be confirmed by experience. Therefore, the article adopts the determination method of objective index weight.

4.1 MODEL STRUCTURE

Setting X as product characteristics, $x_1, x_2, \cdots, x_7$, are 7 evaluation objects, meet relationship $X = \{x_1, x_2, \cdots, x_7\}$, $x_i \in X$, $1 \le i \le 7$. In order to quantitative describe the certain degree of x_i, need to measure m indicators, setting I as index space, meet relationship: $I = \{I_1, I_2, \cdots, I_m\}, I_j \in I, 1 \le j \le m$.

Setting x_{ij} as measure values for evaluation object i on indicator j. Each measure values has 5 evaluation degree C_1, C_2, C_3, C_4, C_5. Setting U as evaluation degree space, satisfying relationship:

$$U = \{C_1, C_2, C_3, C_4, C_5\}, C_p \in U, 1 \le p \le 5$$

4.2 JUDGMENT METHOD

Setting w_j as the relative importance of the evaluation index j on the whole index space. Meeting condition $0 \le w_j \le 1$.

1) Setting $u = u_{ijk}$ as the level of x_{ij} belonging to c_k, meet unitary and additivity, assuming each index has the same importance, that is said:

$$w_j = \frac{1}{m}. \tag{1}$$

2) Defining matrix:

$$\left(u_{ijk}\right)_{m \times p} = \left(u_{ijk}\right)_{m \times 5} = \begin{pmatrix} u_{i11} & u_{i12} & u_{i13} & u_{i14} & u_{i15} \\ u_{i21} & u_{i22} & u_{i23} & u_{i24} & u_{i25} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ u_{im1} & u_{im2} & u_{im3} & u_{im4} & u_{im5} \end{pmatrix}, \tag{2}$$

as single index measure evaluation matrix of x_i, defining

$$u_{ik} = \sum_{j=1}^{m} w_j u_{ijk}. \tag{3}$$

Combing Equations (2) and (3), getting Equation (4) and call it multi-index comprehensive evaluation matrix.

$$\left(u_{ik}\right)_{n \times p} = \begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1p} \\ u_{21} & u_{22} & \cdots & u_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ u_{n1} & u_{n2} & \cdots & u_{np} \end{pmatrix}, \tag{4}$$

at the same time, getting comprehensive evaluation vector for each evaluation object

$$A_i = \left(u_{i1}, u_{i2}, u_{i3}, u_{i4}, u_{i5}\right). \tag{5}$$

3) Defining Y_i^+ as the ideal solution for the evaluation result 'evaluation objects i have positive influence on B2B e-commerce transformation', on the contrary, defining Y_i^- as the ideal solution for the evaluation result 'evaluation objects i have negative influence on B2B e-commerce transformation'.

$$Y_i^+ = \max u_{ik}, \ Y_i^- = \min u_{ik}, \ k \in (1,5), \tag{6}$$

At last, defining variables $E_i^+ = 1 - \frac{1}{k} \sum_{k=1}^{5} \left| u_{ik} - Y_i^+ \right|$ and

$$E_i^- = 1 - \frac{1}{k} \sum \left| u_{ik} - Y_i^- \right|. \tag{7}$$

So drawing the paper conclusion 1, if $E_i^+ \succ E_i^-$, it shows that evaluation objects i have positive influence on B2B e-commerce transformation, if $E_i^+ \prec E_i^-$, it shows

that evaluation objects i have negative influence on B2B e-commerce transformation.

4) Because evaluation space U has its order, that shows $C_1 \prec C_2 \prec C_3 \prec C_4 \prec C_5$, making the score for C_k is k, defining:

$$q_{xi} = \sum_{k=1}^{5} k u_{ik} \ .\tag{8}$$

So drawing the paper conclusion 2, the scoring ranking of 7 product characteristics.

5 Case study

Shanxi Pingyang Industry Machinery Co., Ltd. a company which produces various kinds of advanced precision machinery, is one of 156 key construction projects during the period of 'One Five'. The reasons choose Shanxi Pingyang Industry Machinery Co., Ltd. are as follows: first of all, due to the limitation of traditional business model, it takes a long time to trade between upstream suppliers and downstream customers. In today's market, with competition become more fierce, enterprise need to looking for new suppliers and customers constantly. Shanxi Pingyang Industry Machinery Co., Ltd. has established internal website and adopt CAD electronic procurement, OA management

system etc.. In the condition, the company urgent hope to B2B e-commerce transform; Secondly, enterprise product contain both large machinery like hydraulic press machine, hydraulic valve etc. and universal product such as support controller etc., because of the limitation of product characteristics, the way for B2B e-commerce transformation is always rough; Thirdly, Shanxi Pingyang Industry Machinery Co., Ltd. is representative for many large type energy and mechanical state-owned enterprises in China, due to large size and number of employees, leading to heavy enterprise pressure. In the case, B2B e-commerce transformation is a must. Therefore, the paper taking Shanxi Pingyang Industry Machinery Co., Ltd. for example in order to analyse product characteristics affecting B2B E-commerce transformation, at the same time, hoping to provide reference for this type of enterprises.

To sum up, the paper designs questionnaire first according to product characteristics, then distributes the questionnaire to several senior executive E_i. The questionnaire use Likert 5 scale method, with number 1-5 represent very disagree, less agree, general, comparative agree, strongly agree. The number of questionnaire is 14, in last, distributed all. The enterprise comprehensive evaluation matrixes are shown in Table 2.

TABLE 2 Comprehensive evaluation matrix

Measure values / Investigator Measure item	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}	E_{11}	E_{12}	E_{13}	E_{14}
x_{11}	5	5	3	4	5	5	4	4	3	5	4	4	3	3
x_{12}	4	5	2	2	5	5	4	2	4	5	4	5	2	5
x_{21}	5	1	3	1	2	1	5	1	2	1	3	5	3	1
x_{22}	5	1	1	1	3	1	5	1	2	1	3	1	4	1
x_{23}	4	1	2	1	2	5	5	1	2	5	1	2	3	1
x_{24}	5	1	2	1	4	5	5	1	4	1	1	3	3	1
x_{31}	5	1	5	5	5	5	5	4	5	2	1	4	4	1
x_{32}	5	1	5	2	5	5	3	4	1	1	1	3	1	1
x_{41}	3	1	1	2	1	2	2	1	2	3	1	2	3	3
x_{42}	2	1	3	1	2	3	2	2	1	3	1	3	2	1
x_{51}	1	1	1	1	2	1	1	1	1	1	1	3	2	1
x_{52}	1	1	1	1	3	1	1	1	1	1	5	3	1	1
x_{61}	4	3	5	4	4	3	5	3	4	5	3	4	5	5
x_{62}	5	5	4	3	4	2	4	5	4	5	3	3	5	4
x_{63}	5	4	3	2	4	3	4	5	5	4	3	5	3	3
x_{64}	3	3	2	5	4	5	4	5	3	4	5	5	4	3
x_{65}	4	3	5	5	4	3	2	3	5	4	3	4	4	5
x_{71}	5	1	1	1	1	1	5	4	5	1	5	1	4	3
x_{72}	3	1	1	1	1	1	5	4	3	1	5	3	5	3
x_{73}	3	1	1	1	2	2	5	1	1	1	2	1	1	1

Single index measure evaluation matrix are as follows:
Single index measure evaluation matrix of standardization:

$$\left(u_{ijk}\right)_{2\times5} = \begin{pmatrix} u_{111} & u_{112} & u_{113} & u_{114} & u_{115} \\ u_{121} & u_{122} & u_{123} & u_{124} & u_{125} \end{pmatrix} =$$

$$\begin{pmatrix} 0 & 0 & 0.28 & 0.36 & 0.36 \\ 0 & 0.28 & 0 & 0.36 & 0.36 \end{pmatrix}.$$

Single index measure evaluation matrix of differentiation:

$$\left(u_{2jk}\right)_{4\times5} = \begin{pmatrix} u_{211} & u_{212} & u_{213} & u_{214} & u_{215} \\ u_{221} & u_{222} & u_{223} & u_{224} & u_{225} \\ u_{231} & u_{232} & u_{233} & u_{234} & u_{235} \\ u_{241} & u_{242} & u_{243} & u_{244} & u_{245} \end{pmatrix} =$$

$$\begin{pmatrix} 0.43 & 0.14 & 0.21 & 0 & 0.21 \\ 0.57 & 0.07 & 0.14 & 0.07 & 0.14 \\ 0.36 & 0.28 & 0.07 & 0.07 & 0.21 \\ 0.43 & 0.07 & 0.14 & 0.14 & 0.21 \end{pmatrix}.$$

Single index measure evaluation matrix of tangibility:

$$\left(u_{3jk}\right)_{2\times5} = \begin{pmatrix} u_{311} & u_{312} & u_{313} & u_{314} & u_{315} \\ u_{321} & u_{322} & u_{323} & u_{324} & u_{325} \end{pmatrix} =$$

$$\begin{pmatrix} 0.21 & 0.07 & 0 & 0.21 & 0.5 \\ 0.43 & 0.07 & 0.14 & 0.07 & 0.28 \end{pmatrix}.$$

Single index measure evaluation matrix of intangibility:

$$\left(u_{4jk}\right)_{2\times5} = \begin{pmatrix} u_{411} & u_{412} & u_{413} & u_{414} & u_{415} \\ u_{421} & u_{422} & u_{423} & u_{424} & u_{425} \end{pmatrix} =$$

$$\begin{pmatrix} 0.36 & 0.36 & 0.28 & 0 & 0 \\ 0.36 & 0.36 & 0.28 & 0 & 0 \end{pmatrix}.$$

Single index measure evaluation matrix of time sensitivity:

$$\left(u_{5jk}\right)_{2\times5} = \begin{pmatrix} u_{511} & u_{512} & u_{513} & u_{514} & u_{515} \\ u_{521} & u_{522} & u_{523} & u_{524} & u_{525} \end{pmatrix} =$$

$$\begin{pmatrix} 0.79 & 0.14 & 0.07 & 0 & 0 \\ 0.79 & 0 & 0.14 & 0 & 0.07 \end{pmatrix}.$$

Single index measure evaluation matrix of substitutability:

$$\left(u_{6jk}\right)_{5\times5} = \begin{pmatrix} u_{611} & u_{612} & u_{613} & u_{614} & u_{615} \\ u_{621} & u_{622} & u_{623} & u_{624} & u_{625} \\ u_{631} & u_{632} & u_{633} & u_{634} & u_{635} \\ u_{641} & u_{642} & u_{643} & u_{644} & u_{645} \\ u_{651} & u_{652} & u_{653} & u_{654} & u_{655} \end{pmatrix} =$$

$$\begin{pmatrix} 0 & 0 & 0.28 & 0.36 & 0.36 \\ 0 & 0.07 & 0.21 & 0.36 & 0.36 \\ 0 & 0.07 & 0.36 & 0.28 & 0.28 \\ 0 & 0.07 & 0.28 & 0.28 & 0.36 \\ 0 & 0.07 & 0.28 & 0.28 & 0.36 \end{pmatrix}.$$

Single index measure evaluation matrix of complexity:

$$\left(u_{7jk}\right)_{3\times5} = \begin{pmatrix} u_{711} & u_{712} & u_{713} & u_{714} & u_{715} \\ u_{721} & u_{722} & u_{723} & u_{724} & u_{725} \\ u_{731} & u_{732} & u_{733} & u_{734} & u_{735} \end{pmatrix} =$$

$$\begin{pmatrix} 0.5 & 0 & 0.07 & 0.14 & 0.28 \\ 0.43 & 0 & 0.28 & 0.07 & 0.21 \\ 0.64 & 0.21 & 0.07 & 0 & 0.07 \end{pmatrix}.$$

Combining Equations (2) and (3) multi-index comprehensive evaluation matrix is obtained:

$$\left(u_{ik}\right)_{7\times5} = \begin{pmatrix} u11 & u12 & u13 & u14 & u15 \\ u_{21} & u_{22} & u_{23} & u_{24} & u_{25} \\ u_{31} & u_{32} & u_{33} & u_{34} & u_{35} \\ u_{41} & u_{42} & u_{43} & u_{44} & u_{45} \\ u_{51} & u_{52} & u_{53} & u_{54} & u_{55} \\ u_{61} & u_{62} & u_{63} & u_{64} & u_{65} \\ u_{71} & u_{72} & u_{73} & u_{74} & u_{75} \end{pmatrix} =$$

$$\begin{pmatrix} 0 & 0.14 & 0.14 & 0.36 & 0.36 \\ 0.45 & 0.14 & 0.14 & 0.07 & 0.19 \\ 0.32 & 0.07 & 0.07 & 0.14 & 0.39 \\ 0.36 & 0.36 & 0.28 & 0 & 0 \\ 0.79 & 0.07 & 0.11 & 0 & 0.04 \\ 0 & 0.06 & 0.28 & 0.31 & 0.34 \\ 0.52 & 0.07 & 0.14 & 0.07 & 0.19 \end{pmatrix}.$$

At the same time, get comprehensive evaluation vector for each evaluation object:

$$Ai = \left(u_{i1}, u_{i2}, u_{i3}, u_{i4}, u_{i5}\right).$$

If $i=1$, $E_1^+ = 0.84$, $E_1^- = 0.8$, so $E_1^+ \succ E_1^-$, it shows that evaluation objects product standardization have positive influence on B2B e-commerce transformation;

If $i=2$, $E_2^+ = 0.75$, $E_2^- = 0.87$, so $E_2^+ \prec E_2^-$, it shows that evaluation objects product differentiation have negative influence on B2B e-commerce transformation;

If $i=3$, $E_3^+ = 0.81$, $E_3^- = 0.87$, so $E_3^+ \prec E_3^-$, it shows that evaluation objects product tangibility have negative influence on B2B e-commerce transformation;

If $i=4$, $E_4^+ = 0.85$, $E_4^- = 0.79$, so $E_4^+ \succ E_4^-$, it shows that evaluation objects product intangibility have positive influence on B2B e-commerce transformation;

If $i=5$, $E_5^+ = 0.41$, $E_5^- = 0.8$, so $E_5^+ \prec E_5^-$, it shows that evaluation objects product time sensitivity have negative influence on B2B e-commerce transformation;

If $i=6$, $E_6^+ = 0.86$, $E_6^- = 0.8$, so $E_6^+ \succ E_6^-$, it shows that evaluation objects product substitutability have positive influence on B2B e-commerce transformation;

If $i=7$, $E_7^+ = 0.68$, $E_7^- = 0.87$, so $E_7^+ \prec E_7^-$, it shows that evaluation objects product complexity have negative influence on B2B e-commerce transformation;

Combining Equation (8) and relative data, the paper obtains $q_{x1} = 3.94$, $q_{x2} = 2.38$, $q_{x3} = 3.18$, $q_{x4} = 1.92$, $q_{x5} = 1.46$, $q_{x6} = 3.9$, $q_{x7} = 2.31$ and then arrive at the conclusion two as follows:

$$q_{x1} \succ q_{x6} \succ q_{x3} \succ q_{x2} \succ q_{x7} \succ q_{x4} \succ q_{x5}$$

It is said the result arrange in score order is standardization, alternative, tangibility, differentiation, complexity, intangibility, time sensitivity.

From the two results, it has been confirmed the model product characteristics affecting the transformation of B2B E-commerce has high applicability.

6 Conclusions

On the basis of combing with the existing related literature and theory about product characteristics and B2B e-commerce transformation, the paper extracts seven factors affecting B2B e-commerce transformation that product characteristics contained, namely standardization, differentiation, tangibility, intangibility, time sensitivity, substitutability, complexity. Standardization and differentiation as a pair of analysis variables, affect B2B e-commerce transformation by influencing product

transaction cost; tangibility and intangibility are also a pair of analysis variables, affect B2B e-commerce transformation by embodying at two points, namely transaction risk and logistic cost; time sensitivity of product affecting B2B e-commerce mainly reflects on the product life cycle and update speed; product substitutability and complexity affect B2B e-commerce transformation by influencing enterprise earning and bearing pressure intensity outside.

At the same time, the paper put forward the mathematical model of product characteristics affecting B2B EC transformation by combining with the method of the objective index weight, including the way seven aspects of product characteristics affecting B2B e-commerce, as well as the influence degree. In the last, the example of Shanxi Pingyang Industry Machinery Co., Ltd. tests the feasibility of the model.

The theory of B2B e-commerce is infancy, further research can proceed from three aspects as follows:
1) Research from other sides of product characteristics;
2) Research by refering to existing relative cases;
3) Research from external environment, such as suppliers, competitors, partners, customers, etc.

References

[1] Song Wenguan 2000 *The e-commerce Introduction* higher education press: Beijing *(in Chinese)*
[2] Davison R M, Vogel D R, Harris R W 2005 The E-transformation of Western China *Communication of The ACM* **48**(4) 62-6
[3] Winter R 2001 Working for e-business – The business engineering approach *International Journal of Business Studies* **9**(1) 101-17
[4] Fu C, Qin M, Huang L 2011 How to Transform Enterprises into Platform-Based B2B E-Commerce Business Models *Journal of Business Economics* **11**(8) 14-22 *(in Chinese)*
[5] Fu C, Huang L, Zeng Q 2010 Analysis on Enterprise Transformation for B2B Electronic Commerce: Based on Competence View *Science of Science and Management of S&T* **10**(7) 123-9 *(in Chinese)*
[6] Zeng Q 2005 *Research on Enterprise e-Business Transformation: Based on Competence View* Routledge: Shanghai

[7] Nie L, Fu C, Huang L 2012 A Decision-Making Model of Enterprise Transformation to Platform-Based B2B e-commerce Business Model: Enterprise's Competence View *Shanghai Economic Review* **12**(9) 127-46
[8] Lin Z 2005 Product Design and Manufacturing Quality Engineering *Mechanical industry press:* Beijing *(in Chinese)*
[9] Laseter T M, Ramdas K 2002 *An Exploratory Analysis IEEE Transactions on Engineering Management* **49**(2) 107-18
[10] Fisher M L 1997 What is the right supply chain for your product? A simple frame work can help you figure out the answer *Harvard Business Review* **75**(2) 105-16
[11] Javalg R G, White I D S, Lee O 2002 Film Characteristics Influencing Export Propensity: An Empirical Investigation by Industry Type *Journal of Business Research* **34**(47) 217-28
[12] Jin Q 2007 Impact of Product Characteristics on E-Commerce Operation: *Shanghai*

Authors

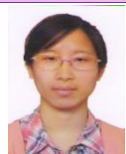**Licheng Ren, born in April, 1968, Datong, Shanxi Province, P. R. China**

Current position, grades: professor at the School of Economics and Management, Taiyuan University of Science and Technology, China.
University studies: Master's degree in management at the University of Science and Technology Beijing.
Scientific interests: electronic commerce, management information system, organization behavior.
Publications: more than 40 papers.
Experience: teaching experience of 30 years, 10 scientific research projects.

Wenhui Pan, born in April, 1989, Jinzhong, Shanxi Province, P. R. China

Current position, grades: the graduate student of School of Economics and Management, Taiyuan University of Science and Technology, China.
University studies: Bachelor's degree in computer sciences at Wuhan Polytechnic University in 2012.
Scientific interests: electronic commerce, management information.
Publications: 1.

Operation Research and Decision Making

Research on supply chain surplus of low carbon supply chain coordination system

Chunjie Yu*

School of Business, Linyi University, Linyi, Shandong Province, 276005, China

Abstract

Supply chain coordination management plays an important role in reducing carbon emission in low carbon supply chain. This study introduces synergetics theory into the research of low carbon supply chain coordination management and reveals the operational mechanism and dynamic mechanism of low carbon supply chain management. Supply chain surplus is an order parameter of low carbon supply chain coordination management, which drives low carbon supply chain system to evolve and develop. It presents the formula of supply chain surplus based on its definition, analyses the components of supply chain surplus, proposes the approaches to realize supply chain surplus, supposes that allocation of supply chain surplus is a Pareto improvement issue, and discusses major basis of supply chain surplus allocation and the program of supply chain surplus allocation.

Keywords: supply chain surplus, low carbon, supply chain management, coordination management

1 Introduction

With the increasing warming of global climate, laws and regulatory policies on reducing the emissions of greenhouse gases are gradually strict worldwide which drive firms find new ways to reduce carbon emissions [1, 2]. Many new technologies and equipment have been applied to reduce the emissions of carbon dioxide, which has also obtained certain effect [3, 4]. Carbon emissions exist in the entire supply chain from the upstream firms to the downstream firms, reducing carbon emission not only need the efforts of a single enterprise, but also requires the joint efforts of all enterprises in supply chain [5]. Recently, there is a growing pressure on supply chain members for reducing the carbon emission of their supply chain. Supply chain coordination management, is the key to effective operation of the entire supply chain system to reduce carbon emission. Numerous enterprises make the target to reduce carbon emissions into the strategy of supply chain management, build low-carbon supply chain to reduce carbon emissions, enhance brand image and competitive advantages.

The concept of supply chain management (SCM) was first proposed in 1980s. It has been used widely in manufacturing industries, and becomes a new business model now. More and more enterprises have been convinced that no longer will enterprises compete against other enterprises, but total supply chains will compete against other supply chains [6]. Next generation of supply chain strategy is coordinated supply chain. Coordination is the key to the competition of supply chain [7].

Most of the earlier models have focused on cost, quality, lead time, etc. issues but not given enough importance to carbon emission for supplier evaluation. Supply chain coordination is that the enterprises in the supply chain coordinate each other to promote the whole supply chain's competition [8]. In the coordinative network, suppliers, manufacturers, distributors and customers can share information dynamically, coordinate smoothly, progress together to achieve their goals [9]. Supply chain coordination management can coordinate and optimize upstream and downstream enterprises' carbon emissions in the supply chain so as to achieve maximize targets for reducing carbon emissions.

Supply chain coordination management plays an extremely important role in carbon emission in low carbon supply chain. Now in academic research of supply chain coordination management, most researches are about method and technology of supply chain coordination, but seldom researches are about the mechanism and theory of supply chain coordination [10-14]. Supply chain surplus is a very important parameter in the study on supply chain coordination mechanism of low carbon supply chain. This paper will discuss the coordination mechanism of low carbon supply chain based on supply chain surplus and realize approaches and allocation program of supply chain surplus in low carbon supply chain.

2 The driven force of low carbon supply chain coordination management is to obtain more supply chain surplus

People have realized that low carbon supply chain is a network, which consists of several independent enterprises. These enterprises can belong to different

---

supply chain. In order to analysis of low carbon supply chain coordination mechanism better, we must introduce new theory to grasp the essence of the low carbon supply chain collaboration mechanism.

2.1 THE OPERRATION MECHANISM OF LOW CARBON SUPPLY CHAIN SYSTEM

Low carbon supply chain management and supply chain management have the same system structure. The supply chain is the connected series of activities which is concerned with planning, coordinating and controlling material, parts and finished goods from supplier to customer. Supply chain consists of all the node enterprises including suppliers, manufacturers, distributors, retailers, etc., which generally has a core enterprise. The node enterprises under the drive of demand information to achieve supply chain added value unceasingly by cash flow, logistics flow and/or service as the media through the division of labour and cooperation of supply chain members.

There is a significant positive correlation between environmental action of enterprises and environmental pressures. Every low carbon supply chain management innovation can be traced to a specific environmental pressure. Environmental stress factors which influence low carbon supply chain management mainly including pressure from consumers, purchasing companies, shareholders and other stakeholders, environmental regulations, environmental groups and the social role of the enterprise and so on. The environmental pressure which plays the leading role in different contexts is different. Laws and regulations are the strongest factor which influences on the environmental innovation of product and process.

Environmental pressure is not direct impetus of low carbon supply chain management, enterprise's fundamental purpose of using part of enterprise's resources to solve environmental problems, is to reduce external risk and ensure maximum economic benefits. In the face of various environmental pressures, the reactions of the supply chain are not always timely and effectively. Only when the environment pressure (external factor) transfer into the supply chain (internal cause), the whole supply chain will bring the factors of environmental protection into the systematic and integrated management process, so as to implement low carbon supply chain management in whole supply chain.

Dynamic mechanism of low carbon supply chain management has two key points. First of all, core enterprises are the medium and bridges of transferring environmental pressure into supply chain pressure. The influence of the core enterprise and the environment pressure transformation ability determines the effectiveness of that transformation. Secondly, economic performance is the "prime movers" of the implementation of low carbon supply chain management. Only when environment pressure affects the core enterprise's

economic performance, and impacts on the rest enterprise's economic performance through the pressure of supply chain, the problems of environmental performance transfer into questions on the economic performance, the whole supply chain has dynamic to response of external environment pressure, and to achieve the optimization of the overall economic performance of the whole supply chain through the low carbon supply chain management practices.

2.2 SUPPLY CHAIN SURPLUS IS THE ORDER PARAMETER OF LOW CARBON SUPPLY CHAIN COORDINATION MANAGEMENT

In essence, low carbon supply chain is a complex system. With the drastic market competition and various customers' demand, low carbon supply chain has been emerged system character of complexity, opening, dynamic and uncertainty. We need introduce new theories to study the mechanism of low carbon supply chain coordination management.

Synergetics is a cross science across nature science and social science, which was founded by German physical scientist H. Haken in 1970s [15]. It is based on the newest systems science theories, adopts the combined method of statistics and dynamics, through analogism it builds a whole set of mathematical model and method to study the evolution rules which how complex system from out-of-order state to order state.

H. Haken indicated that: "If we focus people's economic behaviours on an easiest question, beyond question the answer is profit." Supply chain surplus is borrowed from the conception of economics. Here it means the sum of margin that all enterprises', including core enterprises and node enterprises, profit which earn after supply chain formed subtracts from profit, which earn before supply chain formed. The precondition of forming supply chain cooperate partner relationship is profit increase. The creation of supply chain surplus don't require profit increase and cost drop happen at the same time, if only the extent of profit increase larger than cost increase or cost drop larger than profit drop, supply chain surplus can be created.

In synergetic, supply chain surplus is the order parameter of the whole supply chain system, it works through supply chain system evolution and development, gets the most subsystems' response and drives subsystem work, and so supply chain surplus determines the supply chain system's evolution speed and progress. It roots in coordination of subsystems, and reacts on dominating subsystems' behaviours.

2.3 THE PURPOSE OF LOW CARBON SUPPLY CHAIN COORDINATION MANAGEMENT IS TO GET MORE SUPPLY CHAIN SURPLUS

Enterprise as a rational "economic man", it is subordinate to the principle of self-interest, always pursues the lowest

cost and maximum profit. If participation in the low carbon supply chain is very convenient (assuming the cost tends to zero), assuming an enterprise faces two choices: participates in the low carbon supply chain or does not participate in the low carbon supply chain. If it is assumed that enterprise's profit before participating in the low carbon supply chain is π_{nsc}, enterprise's profit after participating the low carbon supply chain is π_{sc}, enterprise participates in the low carbon supply chain or not depends on the comparison between π_{nsc} and π_{sc}.

$$\pi_{nsc} = R_{nsc} - C_{nsc} , \tag{1}$$

$$\pi_{sc} = R_{sc} - C_{sc} , \tag{2}$$

where:

R_{nsc}= enterprise's revenue before participating the low carbon supply chain.

C_{nsc} = enterprise's cost before participating the low carbon supply chain.

R_{sc} = enterprise's revenue after participating the low carbon supply chain.

C_{sc} = enterprise's cost after participating the low carbon supply chain.

S is defined as one enterprise's supply chain surplus after participating in the low carbon supply chain

$$\begin{aligned} S &= \pi_{sc} - \pi_{nsc} \\ &= \left(R_{sc} - C_{sc} \right) - \left(R_{nsc} - C_{nsc} \right) . \\ &= \left(R_{sc} - R_{nsc} \right) - \left(C_{sc} - C_{nsc} \right) \end{aligned} \tag{3}$$

The enterprise's choice is: it will not participate or quit from low carbon supply chain when $S<0$; it will participate or not participate the low carbon supply chain when $S=0$; it will participate the low carbon supply chain when $S>0$. Accordingly, prerequisite of participating the low carbon supply chain is $S>0$ or $\pi_{sc}>\pi_{nsc}$.

Supply chain surplus refers to the sum total of difference between enterprise's profit before participating in the low carbon supply chain and enterprise's profit after participating in the low carbon supply chain [16]. It assumes that there are several enterprises in low carbon supply chain, and supply chain surplus of the whole low carbon supply chain is defined as S_{sc}.

$$S_{sc} = \sum_{i=1}^{n} S_i = \sum_{i=1}^{n} \left(\pi_{sc} - \pi_{nsc} \right) , \tag{4}$$

$$S_{sc} = \sum_{i=1}^{n} \left[\left(R_{sci} - R_{nsci} \right) - \left(C_{sc} - C_{nsc} \right) \right] , \tag{5}$$

where:

S= one enterprise's supply chain surplus after participating in the low carbon supply chain.

$i{\geq}2$, there are at least two companies in the low carbon supply chain, such as supplier and manufacturer.

Every enterprise in the low carbon supply chain wants more coordination profit, $S_{sc}>0$ is the prerequisite of formation of low carbon supply chain and the prerequisite

of supply chain coordination. The purpose of the existence and coordination of low carbon supply chain is to obtain more supply chain surplus. Enterprises in low carbon supply chain can get more profit than before they participate low carbon supply chain.

3 Implementation approaches of supply chain surplus in low carbon supply chain

In order to facilitate implementation approaches of supply chain surplus in low carbon supply chain, we divide supply chain surplus into two parts: revenue surplus (S_R) and cost surplus (S_C) [16].

$$S_R = \sum_{i=1}^{n} \left(R_{sci} - R_{nsci} \right) , \tag{6}$$

$$S_C = -\sum_{i=1}^{n} \left(C_{sci} - C_{nsci} \right) = \sum_{i=1}^{n} \left(C_{nsci} - C_{sci} \right) . \tag{7}$$

From the two expressions above, it can be seen that supply chain surplus can be achieved by revenue surplus and cost surplus through cooperation of enterprises in the low carbon supply chain.

3.1 IMPLEMENTATION APPROACHES OF REVENUE SURPLUS

Revenue is determined by quantity of sales and prices, then

$$\begin{aligned} S_R &= \sum_{i=1}^{n} \left(R_{sci} - R_{nsci} \right) = \sum_{i=1}^{n} \left(Q_{sci}P_{sci} - Q_{sci}P_{sci} \right) \\ &= \sum_{i=1}^{n} \left[\left(Q_{nsci} + \Delta Q \right)\left(P_{nsci} + \Delta P \right) - Q_{nsci}P_{nsci} \right] \\ &= \sum_{i=1}^{n} \left(P_{nsci}\Delta Q + Q_{nsci}\Delta P + \Delta Q\Delta P \right) . \end{aligned} \tag{8}$$

The above analysis shows that implementation of revenue surplus of supply chain surplus by two ways: increase of sales quantity and price [17]. According to management integration theory, the essence of low carbon supply chain management is an integrated management model based on process. In the formation process of low carbon supply chain, enterprises' integration and cooperation drive the whole function of low carbon supply chain system to be multiplier or emergence, to meet the needs of consumers better, provide value-added service, and create more supply chain surplus.

Revenue surplus of supply chain surplus can be achieved by following ways:

1) Meet more diversified needs of customers. The cooperation of the member enterprises of the low carbon supply chain can understand the needs of the consumer better, and meet the needs of more customers by means of

mass customization flexible manufacturing postponement manufacturing.

2) Present more value-added services and products to consumers. Supply chain is a pull production system based on consumer-centric. The cooperation of the members of the supply chain enterprises give consumers better services and bring higher utility to consumer.

3) Shorten new product development cycles. The cooperation of suppliers, manufacturers, distributors and customers will produce more new product development ideas, speed up the development of new products, and put new products to market more quickly.

4) Improve the quality of products and services. Quality of product and service involving all links of supply chain, such as suppliers, manufacturers, distributors and users, good partnership is good to implement total quality management of supply chain node enterprises.

5) Develop new market. Cooperation of node enterprise in low carbon supply chain will enable the development of new markets, opening up new distribution channels to obtain satisfactory results.

3.2 IMPLEMENTATION APPROACHES OF COST SURPLUS

In consideration of transaction costs, the cost of enterprise should include internal organization costs (C_o), market transaction costs (C_m), and costs of input factors (C_r) [17], then cost surplus can be divide into three parts:

$$S_c = \sum_{i=1}^{n} \left(C_{nsci} - C_{si} \right)$$

$$= \sum_{i=1}^{n} \left[\left(C_{nsci} - C_{sci} \right) + \left(C_{nscoi} - C_{scoi} \right) + \left(C_{nscmi} - C_{scmi} \right) \right].$$

Cost surplus of low carbon supply chain mainly comes from the saving of internal organization costs, market transaction costs, and costs of input factors. Cooperation of enterprises in low carbon supply chain can effectively reduce total costs of low carbon supply chain, then generate cost surplus.

Cost surplus of low carbon supply chain surplus can be achieved by following ways:

1) Reduce inventory cost. Inventory cost savings in the supply chain management are the most credible and most worth pursuing performance [18]. Vender manage inventory (VMI), joint manage inventory (JMI), and other inventory management means play significant roles in reducing inventory costs.

2) Decrease transaction cost. Supply chain member companies makes the transaction object stability, trust, information sharing, which reduce intermediate links, and reduce transaction costs.

3) Reduce organization cost. The internal organization of cost savings can be achieved at different stages of the supply chain more efficient by division of labour and collaboration. The members of the supply chain are independent to each other, cooperate and compete with each other, which helps to avoid the non-productivity of hierarchical organization, and reduces organizational costs.

4) Cut down production cost. Members of the supply chain companies can develop new products through cooperation to amortize the cost of R&D, the stabilization of the cooperation between member companies is conducive to achieve economies of scale, reduce the costs generated by the production of excessive volatility (bullwhip effect). Because enterprises in the supply chain have formed long-term, stable, and cooperative relations, enterprises in low carbon supply chain can invest in special assets, which can improve production efficiency and reduce production costs.

3.3 THE ALLOCATION OF SUPPLY CHAIN SURPLUS

Supply chain surplus is the result of all enterprises cooperation in low carbon supply chain, is the purpose of low carbon supply chain coordination management. Enterprises' purpose of participating in low carbon supply chain is to create "cooperation surplus", how to allocate the "cooperation surplus" and how much right of residue claim can participating in enterprises get from the low carbon supply chain are the focus of attention problems in modern firm theory. Allocation of supply surplus not only influences the profit allocation in low carbon supply chain, but also affects quantity of surplus, even influences the enterprises' participation. How to allocate the residual power between participants in the low carbon supply chain is the core issue of governance of low carbon supply chain.

Allocation of supply chain surplus is a Pareto improvement issue. How to allocate supply chain surplus fairly and reasonably to the enterprises in low carbon supply chain is a critical issue in low carbon supply chain management. The major basis of low carbon supply chain surplus allocation is mainly based on:

1) Degree of information sharing. Demand-side enterprises and supply-side enterprises in supply chain can achieve real-time information by sharing information to reduce the risk of supply-side enterprises' inventory. In the cooperation of Procter & Gamble Company and Wal-Mart, Procter & Gamble Company can call Wal-Mart's sales and inventory data at any time, and develops its production and shipping plans efficiently by sharing information from Wal-Mart. In return, Procter & Gamble Company transfers part of supply chain surplus to Wal-Mart by providing products at preferential prices.

2) The size of the risk. Each link of the low carbon supply chain bears a different risk. Generally, upstream enterprise's risk significantly dropped after the formation of the supply chain partnership. But the risk of end-consumers still cannot be reduced effectively (consumer demand is changing all the time). So the current trend of

allocation of the supply chain surplus is inclined to links directly to downstream consumer.

3) Loss of control. The members of the supply chain enterprises constrained by the core enterprises, and thus lose some control rights. Control rights are parts of the business utility, supply chain must give some income or usufruct as compensation of the loss of control rights, such as preferential pricing, granted exclusive powers, etc.

Correct and suitable process is the safeguards of supply chain surplus allocation. First, it can determine the

need of resource and value which enterprises need put into by the need of low carbon supply chain, sign contracts according to various stakeholders' resources, risk and contribution. Then it will evaluate the profits by effects and actual performance of low carbon supply chain operation, diagnose problems, take supplementary adjustments to the contracts, and make the secondary allocation. The program of supply chain surplus allocation is shown in Figure 1.
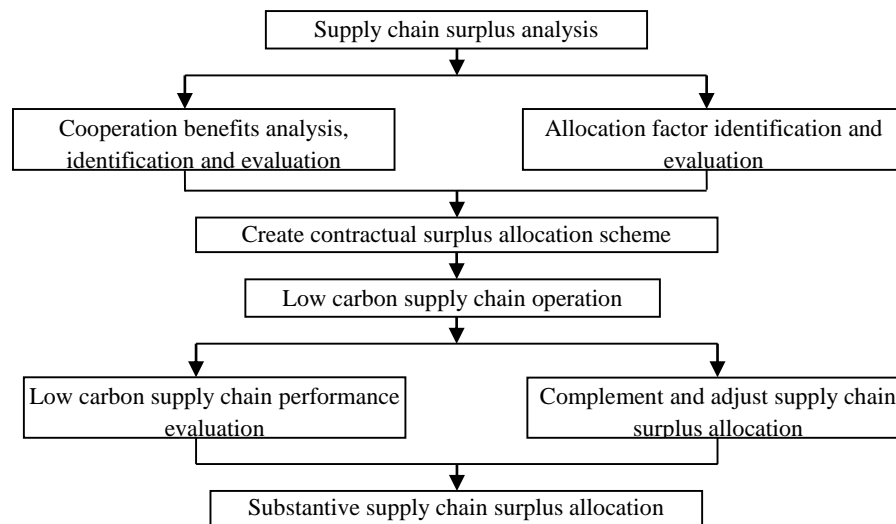
FIGURE 1 The program of low carbon supply chain surplus allocation

Since enterprises in the supply chain collaborative networks have formed a closely strategic alliance, the allocation of supply chain surplus must consider cooperative enterprise's fixed investment, contribution to the whole low carbon supply chain, effort level, and risk

factors to achieve Pareto improvement. Factors affect influenced the distribution of supply chain surplus in low carbon supply chain can be concluded to four factors: specific assets investment, effort level, risk factors and substantial contributions, which are shown in Figure 2.
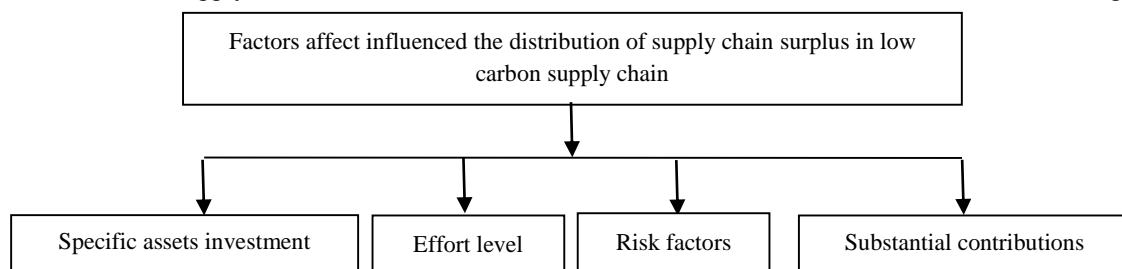
FIGURE 2 Factors affect influenced the allocation of supply chain surplus in low carbon supply chain

The allocation of the supply chain surplus is a game process between member enterprises of the low carbon supply chain, especially enterprises both are strong cooperation. The allocation of the supply chain surplus is a process of repeated game process. Therefore, negotiation mechanism is necessary in relational contract [11]. Supply chain as a network intermediates between market and hierarchy organization, allocation of supply chain surplus can be only realized by coordination and, including compensation of information sharing, risk-sharing, and loss of control right. Generalized supply chain members also include the end users and consumers,

and supply chain surplus also need be allocated to the end-users and consumers to enhance the competitiveness of the whole low carbon supply chain.

4 Conclusions

With the increasing warming of global climate, people pay more attention to reduction carbon dioxide emission. Reducing carbon dioxide emission not only needs the efforts of a single enterprise, but also requires the joint efforts of all enterprises in supply chain. Supply chain coordination management, is the key to effective

operation of the entire supply chain system to reduce carbon emission.

The purpose of existence and coordination of low carbon supply chain is to obtain more supply chain surplus. $S_{sc}>0$ is the prerequisite of formation of low carbon supply chain and the prerequisite of low carbon supply chain coordination. Enterprises in low carbon supply chain can get more profit than before they participate low carbon supply chain. Synergetics theory is good to apply to the study of low carbon supply chain coordination mechanism. In Synergetics, supply chain surplus is the order parameter of the whole low carbon supply chain system, it works through low carbon supply chain system evolution and development, gets the most subsystems' response and drives subsystem work. Supply chain surplus determines the low carbon supply chain systems' evolution speed and progress. The interaction of supply chain surplus with specialization, informatization and knowledge management drives the supply chain system evolving to stably transitional supply chain surplus, and then developing to a new supply chain surplus structure.

Supply chain surplus can be achieved by revenue surplus and cost surplus, revenue surplus of supply chain surplus can be achieved by raising the value or creating new customers, cost surplus of supply chain surplus can be achieved by cutting down inventory cost, transaction cost organization cost, and production cost. How to allocate supply chain surplus fairly, reasonably to the enterprises in low carbon supply chain is a critical issue in low carbon supply chain management. Allocation of supply chain surplus in low carbon supply chain is a Pareto improvement issue. The major basis of supply chain surplus allocation is mainly based on degree of information sharing; the size of the risk; loss of control. Supply chain surplus allocation needs a correct and suitable process.

Acknowledgments

References

[1] Matthews H S, Hendrickson C T, Weber C L 2008 The importance of carbon footprint estimation boundaries *Environmental Science & Technology* **42**(16) 5839-42

[2] Balan S, Robert de S, Mark G, Stephan M W, Sushmera M 2010 Modeling carbon footprints across the supply chain International *Journal of Production Economics* **128**(1) 43-50

[3] Perry S, Klemeš, J, Bulatov I 2008 Integrating waste and renewable energy to reduce the carbon footprint of locally integrated energy sectors *Energy* **33**(10) 1489-97

[4] Hitchcock T 2012 Low carbon and green supply chains: the legal drivers and commercial pressures *Supply Chain Management: An International Journal* **17**(1) 98–101

[5] Shaw K, Shankar R, Yadav S S 2012 Supplier selection using fuzzy AHP and fuzzy multi-objective linear programming for developing low carbon supply chain *Expert System* **39**(9) 8182-92

[6] Barbarosoglu G 2000 An integrated supplier-buyer model for improving supply chain coordination *Production Planning & Control* **11** 732-41

[7] Fugate B, Sahin F, John T M 2006 Supply chain management coordination mechanisms *Journal of Business Logistics* **27**(2) 129-61

[8] Manthou V, Vlachopoulou M, Folinas D 2004 Virtual e-Chain (VeC) model for supply chain collaboration *International Transactions in Operational Research* **87**(3) 241-50

[9] Walker H, Sisto L Di, McBain D 2008 Drivers and barriers to environmental supply chain management practices: lessons from the public and private sectors *Journal of Purchasing and Supply Management* **14**(1) 69-85

[10] Balakrishan A, Geunes J 2004 Collaboration and coordination in supply chain management and E-Commerce *Production and Operations Management* **13**(1) 1-2

[11] Cachon G P, Lariviere M 2005 A supply chain coordination with revenue-sharing contracts: strengths and limitations *Management Science* **51**(1) 30-44

[12] Chen H, Chen J, Chen Y 2006 A coordination mechanism for a supply chain with demand information updating I*nternational Journal of Production Economics* **103**(1) 347-61

[13] Lee CH, Rhee BD 2012 Trade credit for supply chain coordination *European Journal of Operational Research* **214**(1) 136-46

[14] Voigt G, Inderfurth K 2012 Supply chain coordination with information sharing in the presence of trust and trustworthiness *IIE Transactions* **44**(8) 637-54

[15] Zeng J, Zhang Y 2000 *Social synergetics* Science Press: Beijing (*in Chinese*)

[16] Wang Z 2005 Supply chain surplus and mechanism of supply chain cooperation *China's logistics academic frontier report (2005~2006) China Federation of Logistics & Purchasing (Eds.) China Material Press: Beijing* 392-97 (*in Chinese*).

[17] Li N, Yang H 2005 Residual of clusters and the internal coordination of cluster of enterprises *Nankai Besiness Review* **8**(8) 60-4 (*in Chinese*)

[18] Waters D 2008 Inventory control and management *John Wiley & Sons Inc: Chichester*

Author

Chunjie Yu, born in October, 1976, Linyi, Shandong, P.R. China

Current position, grades: associate professor in School of Business, Linyi University, China.
University studies: Ph.D. in School of Humanities & Economic Management from China University of Geosciences (Beijing) in China.
Scientific interest: supply chain management, organization behaviour.
Publications: 10 papers.
Experience: teaching experience of 16 years.

Operation Research and Decision Making

The application of BP neural network optimized by genetic algorithm in logistics forecasts

Huilin Yuan[1, 2], Jia Fu[2], Wei Hong[3*], Jinbo Cao[4], Jing Li[5]

[1]*Beijing University of Aeronautics and Astronautics Mechanical Engineering Post-doctoral Mobile Station, Beijing, China*

[2]*Northeastern University at Qinhuangdao, Hebei, China*

[3]*China petroleum & chemical corp., Hebei oil products co., Hebei, China*

[4]*Yanshan University, Qinhuangdao, Hebei, China*

[5]*TBEA Shenyang Transformer Group Co., high-voltage switch Institute, Shenyang, Liaoning, China*

Abstract

This paper points out disadvantages of traditional forecast methods and elaborates the advantages of the method based on BP neural network. On this basis, the paper puts forward a logistics forecasting model of BP neural network optimized by genetic algorithm. The new method uses historical data to establish and train BP neural network and thus obtain logistics forecasting model. The results implemented by MATLAB show that, neural network possesses memorizing and learning capability, and can forecast logistics development trend perfectly, which is proved by a large amount of actual forecast results. Compared with BP neural network model, the model has the advantages of less number of iterations, convergence speed and strong generalization ability.

Keywords: BP neural network, genetic algorithm, optimized, MATLAB

1 Introduction

In logistics management, if the trend of the market can be grasped through prediction, it would be possible for the management to make decisions on research strategies in advance and take more effective technological strategies to gain even better results. Moreover, enterprises can prevent or minimize the adverse effects on their development and foster more favourable conditions with a more sensible decision on logistics development strategic goals. Since the prediction is determined by many factors, such as market supply and demand, economic situations and transportation, systematic efforts are needed to establish a predicting model [1].

In recent years many innovative and practical predicting methods have been proposed by scholars both home and abroad such as moving average method, exponential smoothing method, neural network method, chaos and nonlinearity investigation, regression analysis, time series, grey theory and Markov analytic approach. However, each of the models or the analysis methods has its own advantages and disadvantages as far as its limitation and applicability are concerned [2]. For example, the moving average forecasting method is based on the time series by item. It forecasts the further phenomenon by calculating exponential smoothing value and cooperating with the time series forecast model. To some extent, this method is more accurate in reflecting the recent changes and the trends [3]. The exponential smoothing method as a time series is based on the moving average method. It calculates the exponential smoothing value, and then forecasts the phenomenon of the future by time series prediction model [4]. Regression analysis method is based on historical logistics data, and the regression mathematical model is established to forecast the future logistics. This kind of logistics forecasting method is effective and accurate, but it must establish the mathematical model, which is very complex. The traditional cost prediction of logistics operation is based on regression analysis. This kind of prediction method is simple and easy to use, but its prediction error is bigger and it cannot meet the requirements of cost control in the process of modern logistics operation [5].

By contrast, the neural network has a good ability of curve fitting, anti-interference ability and learning ability, so the neural network is an effective method to logistics prediction.

The neural network has many good qualities, such as self-organization, self-adaption and is good at making decisions from the approximate, uncertain and even conflicting knowledge environment, and avoids calculating the weights and correlation coefficient artificially [6]. But the traditional neural network has the problems of local minimum and slow convergence, so BP neural network optimized by genetic algorithm is established for the logistics forecasts model. The model combines the self-adaptive and self-organization of BP neural network with rapid global search ability of genetic

---
[*]*Corresponding author* e-mail: 45264824@qq.com

algorithm, solves the problems of slow convergence speed and is easy to fall into local minimum. The practical simulation results prove that the method has stronger practicability.

2 Basic principle of BP neural network optimized by genetic algorithm

Usually the BP algorithm adjusts the weights between neurons through some learning rules. In the learning process, the topological structure and network learning rules are constant. However, a neural network information processing function depends not only on the strength of the connections between neurons, but also on the network topology (neurons connection), characteristics of input and output neurons, and the weights and thresholds of neurons [7].

The traditional BP neural network has shortcomings of long training time, slow convergence and easy to fall into local minimum, but the local searching ability of genetic algorithms can make up for the deficiency of BP neural network. And BP neural network has great dependence on the weights and thresholds, but if the initial values are closer to the true ones, the time of network training will be shorten obviously. This paper uses the basic genetic algorithm to optimize BP neural network weights and thresholds, leads the weights and thresholds close to the real values, rather than a simple random assignment. Then, the paper adopts the trained BP network as the model of logistics forecasts. Simulation results demonstrate that the genetic algorithm has a significant effect to accelerate the convergence speed.

The algorithm that optimizes BP neural network weights and thresholds with basic genetic algorithms is described as follows:

1) Encode BP neural network weights and thresholds with the real, and produce an initial population $W = (W_1, W_2, \cdots W_p)^T$ randomly, which has P individuals. Each individual is a string of real numbers, containing the connection weights of the input layer and hidden layer, hidden layer thresholds, hidden layer and output layer weights and the output layer thresholds.

2) Use the reciprocal of an error function as the fitness function. The error produced by the mean square error function of the desire output and the actual output of the network, determine the fitness of each individual by fitness function. Smaller the error is, greater the fitness is.

3) Use the roulette bet method as the selecting method. In the method, the selected probability of each individual is proportional to the fitness. The selected probability increases with increasing of fitness. Select the part of higher fitness individuals as the parents, and eliminate the lower ones.

4) Handle the parents to produce progeny by crossover and mutation operators. If the offspring's fitness is higher than the parent's, the parent's individual will be eliminated, and the new offspring will become the parent. Keep the number of parent individuals as a constant.

5) Repeat step (2)-(4), a new round of selection, crossover and mutation will be executed to the new group, until the termination condition is satisfied.
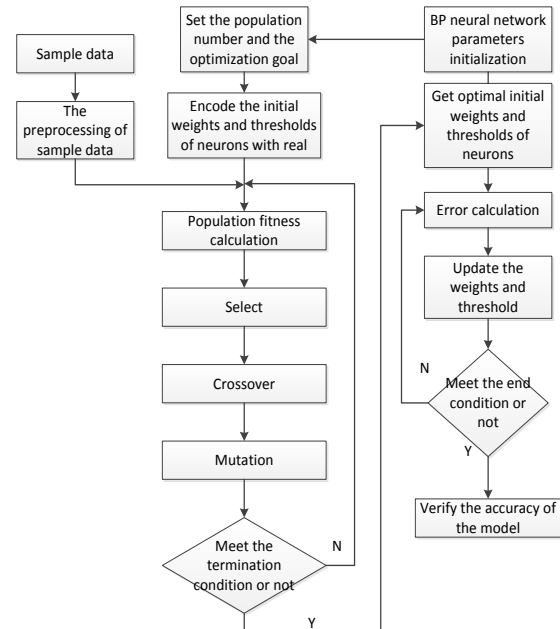

FIGURE 1 Procedure of BP neural network optimized by genetic algorithm

6) Decompose the optimal individuals of current groups into the connection weights and thresholds of BP neural network, and use the optimal weights and thresholds as the initial values of BP neural network.

7) Train BP neural network with preset algorithm parameters.

8) Stop training when the training objective is satisfied.

The procedure of BP neural network optimized by genetic algorithm is shown as Figure 1 the training objective is satisfied.

3 BP neural network model optimized by genetic algorithm

3.1 BP NEURAL NETWORK MODEL

BP neural network is a multilayer feed forward neural network. The network is composed of input layer nodes, hidden layer nodes (hidden layer can be one or more), and output layer nodes.

There is no universal theory guidance for selecting how much hidden layers and hidden layer nodes of each layer, but after a lot of practice, predecessors have accumulated some experience. Theoretical analysis shows that, BP network with a single hidden layer can map the arbitrary continuous nonlinear function. Only when the learning is not a continuous function (such as a saw tooth wave), two hidden layers are needed.

Increasing the hidden layer number can improve the training accuracy and reduce the errors, but the network will become complicated and time-consuming. In fact, the nodes of the hidden layer can be increased to improve the training accuracy, when the increased number of nodes is not significantly to reduce the error, trying to increase the number of hidden layers [8]. So this network uses three layers network, namely the input layer, the hidden layer and the output layer.

1) Design the input and the output layer. The design of input and output layer is according to the specific problems, there are 2 evaluation indexes and one evaluation result, so the input layer is set to n=2, the output m=1.

2) Design the hidden layer. In the BP network, hidden layer nodes are used to extracting and storing the samples inherent law from the samples, so setting the numbers of hidden layer nodes depend on the training sample size and complexity. There is no exact formula for the calculation of the hidden layer nodes, but the summary based on network structure obtained an empirical formula. The hidden layer nodes number $l = \sqrt{n+m} + \alpha$, n and m are the input and output layer nodes numbers, α is a constant between [1,10]. After repeated testing, this paper determines the hidden layer node number is 7.

3) Design the driving function. In the BP neural network, which is established in this paper, hidden layer transferring function uses bipolar S type function.

$$f(x) = \frac{1-e^{-x}}{1+e^{-x}} .$$ (1)

The output layer transferring function uses linear function.

$$f(x) = x .$$ (2)

The structure shown as Figure 2:
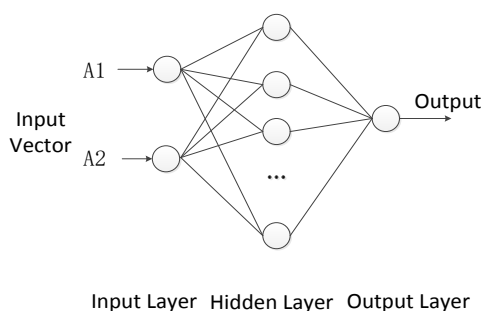
FIGURE 2 The structure of BP neural network model of logistics forecasts

3.2 THE CODING METHOD SELECTION

Genetic algorithm needs to encode the initial weights and thresholds of BP network, and BP network weights values and thresholds values composed of many real values. So it will be the real if uses binary coding. This

will affect the learning accuracy, so this paper adopts real coding method.

Assume each chromosome in the population contains S gene, then

$$S = R \times S_1 + S_1 \times S_2 + S_1 + S_2$$ (3)

where R is the number of input layer nodes; S_1 is the number of hidden layer nodes; S_2 is the number of output layer nodes. So, the chromosome length $S = 2 \times 7 + 7 \times 1 + 7 + 1 = 29$, a total of 29 parameters need to be optimized.

3.3 FITNESS FUNCTION DESIGN

In the genetic algorithm the value of fitness function directly determines the direction of evolution population. The fitness function is designed generally based on the objective function of optimization. In the network designed in the paper, the objective function makes the mean square error of all the sample output to the minimum, namely the minimum of Equation (4).

$$E_{MSE} = \frac{1}{p} \sum_{p=1}^{p} (d_p - y_p)^2 ,$$ (4)

where P is the number of training samples; d_p is the desired output of the p samples; y_p is the actual output of the p sample.

As the individuals with higher adaptive value are selected to evolve in genetic algorithm, the adaptive value function uses the reciprocal of sample output variance.

$$f = \frac{1}{E_{MSE}} .$$ (5)

3.4 DESIGN OF GENETIC OPERATION

Genetic operations include selection, crossover and mutation.

1) Selection. The selection operation is based on fitness proportional selection method, higher the fitness is, greater the probability of selection is. Selection probability of each individual is shown as follows:

$$p_i = f_i \Big/ \sum_{i=1}^{N} f_i .$$ (6)

2) Crossover. The crossover adopts the real crossover method, namely the chromosome marked i and chromosome marked j cross at Position r.

$$a_{ir} = a_{ir}(1-c) + a_{jr} c ,$$ (7)

$$a_{jr} = a_{jr}(1-c) + a_{ir} c ,$$ (8)

where c is a random number between [0, 1].

3) Mutation. Select the j genes of the individual marked i to compile operation.

$$a_{ij} = \begin{cases} a_{ij} + (a_{ij} - a_{max})\, r_2 (1 - g/G_{max})^2 & r_1 \geq 0.5 \\ a_{ij} + (a_{min} - a_{ij})\, r_2 (1 - g/G_{max})^2 & r_1 < 0.5 \end{cases}, \quad (9)$$

where a_{max} and a_{min} are the upper and lower bounds of gene a_{ij}; r_1 is a random number between [0, 1]; r_2 is another random number; g is the current iterations; G_{max} is the maximum number of evolution.

3.5 DESIGN OF TERMINATION CONDITIONS

In the BP neural network algorithm, when the optimal individual fitness reaches a given threshold, or the best individual fitness and population fitness will not rise any more, even or the number of iterations reaches a preset algebra, the algorithm will be terminated.

In this paper the algorithm selects the default algebra as 100 generation, and terminates when the iterations reaches a preset algebra.

3.6 BP NEURAL NETWORK PREDICTION

Take the weights and thresholds optimized by genetic algorithm as the BP neural network weights and thresholds, then use them to predict.

4 The example research

4.1 EXAMPLE

The case in [9] is used to analyse the enterprise material purchasing. The historical data samples of enterprise market demand, market price and the amount of material procurement are shown in Table 1:

TABLE 1 Sample data

Resources /t	2540	2960	2570	…	3110
Price /Yuan	1.60	1.75	1.63	…	1.78
Market purchases /t	235	273	238	…	289

4.2 TRAINING AND RESULTS

200 groups of data as the training sample are used to train BP neural network model, network structure is BP (2,7,1). The BP neural network is trained and optimized by genetic algorithm with MATLAB. The results are shown as Figure 3 and Figure 4.

The model, which uses BP neural network had gained the goal after 235 generations, but the model that uses BP neural network optimized by genetic algorithm had attained the goal just after 76 generations. The results show that convergence speed is greatly improved by genetic algorithm optimization.
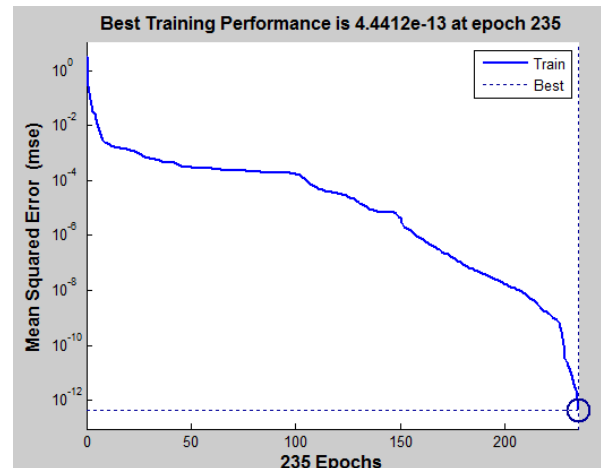
FIGURE 3 The training results of the BP neural network
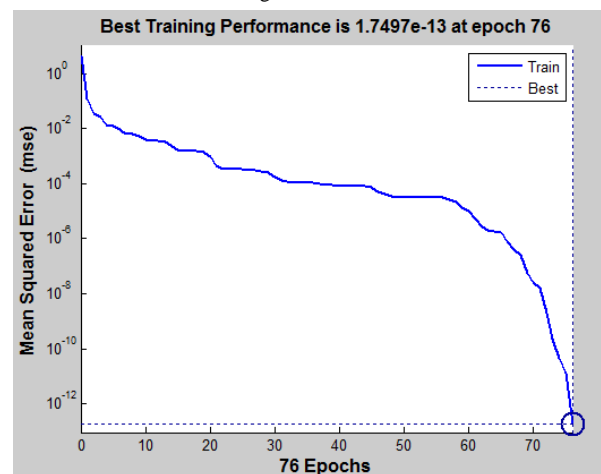
FIGURE 4 The training results of the BP neural network optimized by genetic algorithm

4.3 SIMULATION AND PREDICTION

After training the neural network 11 groups of data, which is not trained are used to test. The results are shown as follows (Figure 5 and 6).
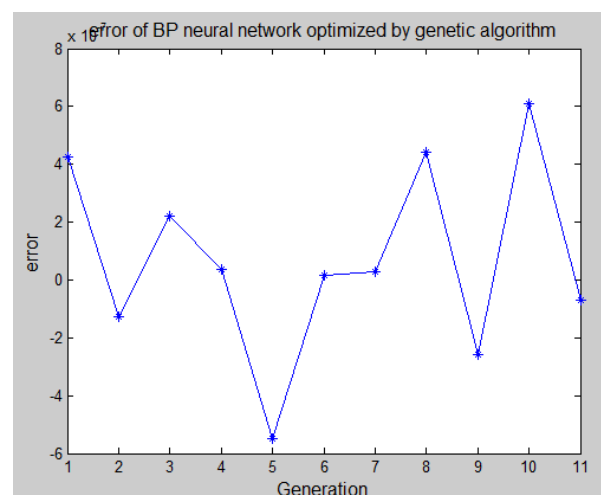

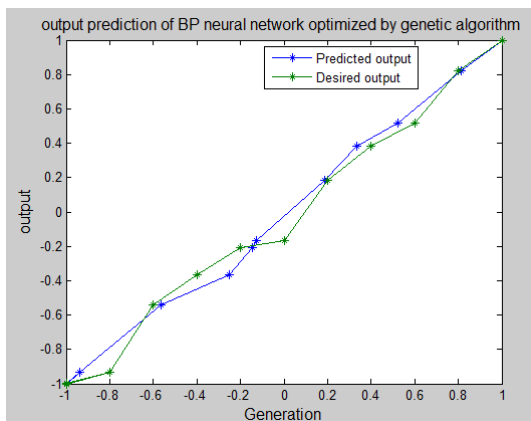FIGURE 5 The error of the BP neural network optimized by genetic algorithm

FIGURE 6 The contrast of the predicted output and desired output

The results show that the BP neural network optimized by genetic algorithm has good performance in logistics forecasting. The predicted output is very similar to the desired one. The errors are small and mostly are around 0 errors. The results proved that the network was trained successful and had better generalization ability.

5 Conclusions

This article analyses disadvantages of traditional forecasting methods and elaborates the advantages of the method based on BP neural network. Through introducing the genetic algorithm, a new logistics forecasting model based on BP neural network is proposed and applied to a logistical process. The simulation results show that the BP neural network optimized by genetic algorithm improved the convergence speed greatly, and it does not fall easily into local minimum.

Acknowledgements

References

[1] Jilong W, Zhongpeng Z 2005 BP neural network based logistics forecast *Hoisting and Conveying Machinery* (5) 30-2
[2] Xinli W, Kun Z 2010 Study on agricultural products logistics demand forecast based on Neural Network *Journal of Agrotechnical Economics* (2) 62-8
[3] Yingxin Z 2013 Comparison of automobile parts reverse logistics inventory forecast method of high-tech products - BP neural network and the traditional method *The small and medium-sized enterprise management and technology* (31)
[4] Jingming W, Fangyang Z 2010 Research on demand forecasting and development pattern of port logistics for Guangxi Beibu gulf port *Logistics technology* (12) 26-8

[5] Bochao L 2011 Wavelet neural network prediction method of logistics enterprises operating cost *Friend of Accounting* (20) 57-9
[6] Xinhong Z, Pie Z 2002 Neural networks based integrated evaluation method for MIS *Journal of Systems Engineering* **17**(5) 63-8
[7] Xiaofeng L, Jiupin X 2004 The establishment of BP adaptive artificial neural network learning algorithm and its application *Journal of Systems Engineering* **24**(5) 1-8
[8] Liqun H 2006 Artificial neural network tutorial *Beijing University of Posts and Telecommunications press*
[9] Liyan D 2000 Logistics system engineering *Tsinghua University press*

Authors

Yuan Huilin, born in May, 1969, Qinhuangdao City, Hebei Province, P.R. China

Current position, grades: the associate professor of School of Information and Engineering, Northeastern University, China.
University studies: M.Sc. from Yanshan University, D.Sc. from Northeastern University in China.
Scientific interest: system engineering, modeling and optimization.
Publications: 30 papers.
Experience: teaching experience of 15 years, 3 scientific research projects.

Fu Jia, born in March, 1991, Shenyang City, Liaoning Province, P.R. China

Current position, grades: the master of Northeastern University, China.
University studies: B.E. in Automation from Shenyang University of Technology, M.E. from Northeastern University in China.
Scientific interest: students evaluation, logistics forecasts.

Hong Wei, born in October, 1969, Shijiazhuang City, Hebei Province, P.R. China

Current position, grades: the economist of Hebei oil products co. of China petroleum&chemical corp.
University studies: B.Sc. in Management Science and Engineering from Yanshan University of in China.
Scientific interest: management science and engineering, evaluation.

Jinbo Chao, born in December, 1977, Qinhuangdao City, Hebei Province, P.R. China

Current position, grades: the lecturer of School of Information Science and Engineering, Yanshan University, China.
University studies: B.E. in Computer Science and Technology, M.E. in Computer Application Technology from Yanshan University in China.
Scientific interest: web mining and machine learning.

Li Jing, born in October, 1990, Shenyang City, Liaoning Province, P.R. China

Current position, grades: the engineer of TBEA Shenyang Transformer Group Co., high-voltage switch Institute, China.
University studies: B.E. in Automation from Shenyang University of Technology in China.

Product inventory model of iron and steel enterprises

Yuanmin Xie*

The Key Laboratory for Metallurgical Equipment and Control of the Ministry of Education, Wuhan University of Science and Technology, Wuhan, 430081, China

Abstract

As a pillar industry, iron and steel industry has made a significant contribution to China's economic development. Steel and iron inventory is a hot research issue in the supply chain management. How to make use of advanced management methods and mathematical models to obtain reasonable inventory to meet customer service and reduce unnecessary inventory-related costs, and accelerate corporate capital turnover rate, is the goal that the enterprises should pursue in the future. According to the batch production of iron and steel enterprises, and in consideration of inventory-related costs, including fixed costs, storage costs, shortage costs (including deferred delivery costs and lost sales costs), establish the model of finished products inventory cost. When calculate model, we make c + + program and do accurate calculation to the model, optimize the production cycle, production time and production quantities, and define the deferred delivery coefficients. By changing the value of the deferred coefficient, analyse the influence to the production cycle, production time and production quantities and all kinds of costs.

Keywords: Steel and iron industry; supply chain; inventory; finished products inventory; optimization

1 Introduction

In traditional inventory management, inventory was divided into two modes: Independent Demand and Dependent Demand. The problem of inventory demand was always been deal with by using Material Requirement Planning (MRP) [1], while Order Point was always been used in dealing with the independent demand problem. In general, the management of finished goods inventory was regarded as dependent demand management.

For an enterprise, if the demand of product and raw materials could be supplied on time, and the production was able to meet this demand instantly, not keeping the inventory was the most economical state in theory. However, it was always hard to predict the demand accurately, or make sure the supply of raw materials could arrive on time, and there was no way to avoid the fault of machine. All above are the causes of inventory in iron and steel enterprise. Even if the supply and demand of product tends to be exactly equal, the production also need to response quickly, requiring the transport completely reliable and without delivery time. All these are completely impossible at any reasonable range of cost. Therefore, companies should use the inventory to better balance supply and demand and the uncertainty of corresponding demand; adjust the economical production quantities and time to reduce production costs. As for iron and steel enterprises, various processes in production should be continuous, especially in continuous casting production, once out of stock, which will result in broken pouring and severe losses ,so steel companies are not allowed out of stock, and they need stock for more

effective buffer. Therefore, to establish the most reasonable production inventory management model become a pivotal issue [2].

Nowadays, the management of finished products inventory was completely artificial, managers need to deal with a lot of inventory business every day, and fill a number of records, make variety of forms. Those steps make a lot of errors in information transmission, and delay seriously. Therefore, they bring great inconvenience in inventory management. Sometimes when an error, which had been caused by a functional department might contribute to more mistakes in other department, and if the error was not checked in time, might lead to very serious consequences. Since the limitation of inventory capacity, a lot of work being cut down, and the inaccurate and delay of information, the complex task become more difficult. In order to meet customers' requirements and improve the service quality, so only the increase of inventory could cover all managers' mistakes.

As it is can be seen, the control strategy of finished products inventory was very simple, this strategy was obviously not suitable to the modernize of inventory management, and inventory management should keep the balance between improving the service quality to fulfil the customers' requirements and reducing the inventory costs, That is to say, the strategy should achieve low-cost operation on the basis of quickly and efficiently get customers' satisfaction. in order to achieve a reasonable storage purposes. Therefore, the stock of finished products should be made for the goal to improve the production supply, control the inventory reasonably, reduce costs, and improve the customer satisfaction, and

* *Corresponding author* e-mail: wustxie@126.com

Operation Research and Decision Making

to coordinate the healthy development of iron and steel enterprises. In order to meet customers' uncertain requirements, thinking about the mechanical failure, unreliability of transportation, the limitation of inventory capacity, and the delay of information transformation, it was necessary to determine a reasonable production volume and the production time to meet customer demands and ensure continuous production in a production cycle.

An enterprise system, which was composed of interrelated subsystems: supply system, production system, inventory system, sales system, consist the feature to convert the raw materials and semi-finished products to the finished goods. Finished goods inventory system between the production system and sales system plays an important role of regulatory. In turn, production system and marketing system also directly affects the inventory system of finished goods. Therefore, the study the inventory control of finished product should be started with the view of system, the following aspects should be considered.

(1) Inventory levels should be maintained in a stable quantity as far as possible. Unlike the seasonal products, the peaks and valleys in steel product sales were relatively not very obvious. Finished goods inventory control should be not only ensured the needs of sales, so that enterprises could rely on a reasonable opportunity to avoid the loss of chance caused by the shortage of products or delaying the delivery of some products, but also avoided the condition that finished goods inventory funding was so large and velocity of money was so slow caused by product unmarketable happens. Therefore, the number of stocks should be maintained at a reasonable level in the finished product inventory control in order to achieve the lowest total cost of inventory.

(2) Considering the raw materials procurement, some bulk material was expensive and had a long purchasing cycle. Under the premise of maintaining certain stability in the production, these raw materials should be ensured the smooth supply firstly in order to prevent raw material shortages affecting production. In addition, in order to make production plans and production scheduling facilitate, maintain certain stability and management easily for production line, productivity should also be as smooth as possible.

(3) Combined with the development trend of the industry, market's competitive requirements and the need of companies' own strength, cost management level had become the core competitiveness of future success for an enterprise. Therefore, finished goods inventory control must be considered the cost factors, not only reduce inventory storage costs, but also reduce the production cost of the product as far as possible to keep it at a relatively low level.

2 Finished goods inventory model

Inventory management was the core and soul of logistics management [3], because the management process of logistics planning, control, execution, assessment was based on the flow of content to meet the demand of market-oriented customers. The charm of the inventory lies in the future demand forecasts, and forecast uncertainty, it was always necessary to balance the relationship between the demand for services and the stock investments limited in the uncertainty.

Inventory management was an important part of enterprise production and management, which was also an effective way and methods to reduce the cost and improve the economic efficiency of enterprises. Too little inventor would affect the continuity of production, while excess inventory levels would cause the backlog of funds. Therefore, a reasonable inventory management was an important condition to achieve the best combination of both. It was assumed that all the demand was delayed delivery or lost during the empty of the warehouse in general storage model.

2.1 BUILD MODEL

(1) Model assumptions

In order to simplify the research model, making the following assumptions:

1) During the patience period [4], the enterprise had no cost loss; during impatient period, some customers were lost, others would delay delivery, in the meantime, the costs include the cost of lost customer losses and delayed costs of delayed delivery.

2) Demand rate was continuous and uniform, the demand rate was denoted as R, the product productivity was denoted as P, and $P > R$.

3) When you use your own warehouse storage products, storage costs of per unit of product and per unit of time was c_2. When using leased warehouse, storage costs of per unit of product and per unit of time was c_3, and $c_2 \leq c_3$.

4) The maximum capacity of own warehouse was Q_0, Q_{max} was the maximum storage capacity of products, and $Q_0 \leq Q_{max}$.

5) When the product were stored, the product should be stored in own warehouses first, and then deposited the remaining into rented warehouse; when using the product, the product of leased warehouse should be used first, and then use own warehouse products after running out of rented warehouse product.

6) During impatient period, demand rate of backordered was equal to the original demand rate R multiplied by α, $R' = \alpha R$ (R' was demand rate of backordered, $\alpha > 0$, α was backordered coefficient), m was maximum time of patience period.

(2) Establishment of model

Figure 1 showed the variation of the system inventory level with time. In order to establish a more realistic inventory cost model [5], analysis as follows: the section of [0, m] in a period T, said the patient period, customers were willing to wait, all out of stock could delay delivery, there was no case of loss of sales; in the section of [m, t_1], said the impatience period, all out of stock was composed of two parts, namely the backlogging costs and lost sales opportunities lost, then the store capacity was zero, t_1 was the time to begin production; in the section of [t_1, t_2], the products not only met the requirements of use, also needed to fill the shortage which in the section of [0, t_1]; and in the section of [t_2, t_3], in addition to met the production demand, storage capacity was increased to P - R speed, until t_3 moment to fill their warehouse; in the section of [t_3, t], storage capacity was increased to the original growth rate, the increased product was deposited in the leased warehouse; in the section of [t, t_4], using the products which stored in the leased warehouse; in the section of [t_4, t], enterprises used the storage products of its own warehouse at the demand speed of R; at the moment T, the storage capacity was zero. However, due to the iron and steel enterprises patience period was short, and the problem was more complex, the first M segment would not be considered, resulting in the figure 2.
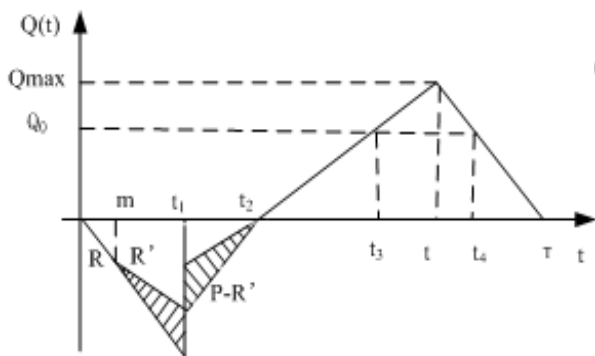
FIGURE 1 Storage capacity change chart
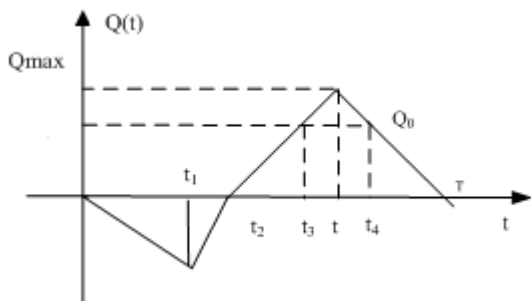
FIGURE 2 Storage capacity change chart

From the figure 1, it could be seen that:
$R't_1 = (R - R')(t_2 - t_1)$, then got:

$$t_1 = \frac{P - R'}{P} t_2,\tag{1}$$

$Q_0 = (t_3 - t_2)(P - R) = (T - t_4)R$, then got:

$$t_3 = \frac{Q_0}{P - R} + t_2,\tag{2}$$

$$t_4 = T - \frac{Q_0}{R},\tag{3}$$

$Q_{max} = (t - t_2)(P - R) = (T - t)R$, then got:

$$t = \frac{t_2(P - R) + TR}{P},\tag{4}$$

$$Q_{max} = \frac{(P - R)R}{P}(T - t_2).\tag{5}$$

1) The shortage cost

When customers placed orders, and the goods could not be delivery by the usual designated warehouse, then it would generate the shortage cost [6]. Shortage cost had two kinds: the lost sales cost and keep the order cost. In each case, assumed that customers would make some response, however, because the customer response could not fathom, the cost was very difficult to accurately measure out.

When out of stock, if customers chose to withdraw his purchase request, it would generate lost sales cost. The cost was the sales profits, which this time should be obtained, and it might also include the negative effects on future sales caused by shortage. If the customer was willing to wait for order fulfilment, then it would not happen lose sales, there would only appear delayed order fulfilment, which produce the retention order cost.

In the section of [0, t_1], one part was tardy delivery quantity, the other part was lost sales for the customer did not want to wait, then backlogging cost was $\frac{1}{2}c_4 R' t_1^2$, as shown in Figure 3, the shaded area represented lost sales opportunities, therefore lost sales cost was $\frac{1}{2}c_5(R - R')t_1^2$.

In the section of [t_1, t_2], there were still some delay delivery and lost sales situations, then backlogging cost was $\frac{1}{2}c_4(P - R')(t_2 - t_1)^2$, as shown in Figure 4, the shaded area represented lost sales opportunities, therefore lost sales cost was:
$\frac{1}{2}c_5[(P - R')(t_2 - t_1)^2 - (P - R)(t_2 - t_1)^2] = \frac{1}{2}c_5(P - R')(t_2 - t_1)^2$

Put the formula (1) and $R' = \alpha R$ into the above equation, it could obtain backlogging cost in the section of [0, t_2] was:

$$\frac{1}{2}c_4[R't_1^2 + (P - R')(t_2 - t_1)^2] = \frac{1}{2}c_4 \frac{(P - \alpha R)\alpha R}{P} t_2^2.\tag{6}$$

And lost sales cost was:
$\frac{1}{2}c_5[(R - R')t_1^2 + (P - R')(t_2 - t_1)^2] = \frac{1}{2}c_5 \frac{(1 - \alpha)R}{P^2}[(P - \alpha R)^2 + (\alpha R)] \cdot t_2^2.$

(In the formula, C_4 was the shortage backlogging cost of unit goods and unit time, C_5 was the loss caused by lost sales opportunity of unit goods.)
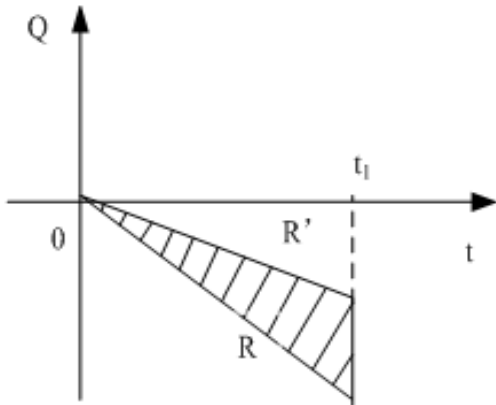

FIGURE 3 The lost sales quantity in the section of $[0, t_1]$
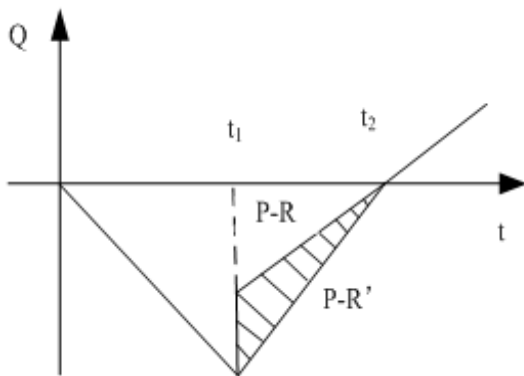

FIGURE 4 The lost sales quantity in the section of $[t_1, t_2]$

2) The storage cost

Total storage cost was composed of two parts, one part was the cost of using their own storage warehouse to store product, the other part was the cost of renting warehouse storing the product. Therefore, the total storage cost was:

$$\frac{1}{2}c_2 Q_0(t_3-t_2)+c_2(t_4-t_3)Q_0+\frac{1}{2}c_2 Q_0(T-t_4)+\frac{1}{2}(Q_{max}-Q_0)c_3(t_4-t_3) .$$

Put the formula (2), (3) and (5) into the above equation, then got:

$$c_2[Q_0(T-t_2)-\frac{Q_0^2}{2A}]+c_3[\frac{1}{2}A(T-t_2)^2-Q_0(T-t_2)+\frac{Q_0^2}{2A}]. \qquad (8)$$

3) The fixed cost

The fixed cost of a cycle of T was C_1 (including warehouse depreciation and workers wages etc.). Therefore, in order to simplify the model, assuming $A=\frac{(P-R)R}{P}$, the average inventory cost of a cycle of T was:

$$C(T,t_2)=\frac{1}{T}[c_1+c_2[Q_0(T-t_2)-\frac{Q_0^2}{2A}]+c_3[\frac{1}{2}A(T-t_2)^2-Q_0(T-t_2)+\frac{Q_0^2}{2A}]+ \qquad (9)$$

$$\frac{1}{2}c_4\frac{(P-\alpha R)\alpha R}{P}t_2^2+\frac{1}{2}c_5\frac{(1-\alpha)R}{P^2}[P-\alpha R^2+(\alpha R)^2]t_2^2]$$

2.2 THE SOLUTION OF THE MODEL

Making that $\frac{\partial C}{\partial T}=0, \frac{\partial C}{\partial t}=0$,

$B=c_4\alpha R(P-\alpha R)P+c_5(1-\alpha)R[(P-\alpha R)^2+(\alpha R)^2]$.

Solved the equations which consist of the above equation, got the optimal production cycle and the beginning of production time were:

$$T^*=\sqrt{\frac{Q_0^2}{A}(1-\frac{c_2}{c_3})(\frac{1}{A}+\frac{c_2 P^2}{B})+2\frac{c_1}{c_3}(\frac{1}{A}+\frac{c_3 P^2}{B})} , \qquad (10)$$

$$t_2=\frac{c_3 AT^*-(c_3-c_2)Q_0}{c_3 A+\frac{1}{P^2}\cdot B} . \qquad (11)$$

The optimal production lot size was:

$$Q^*=R'\times t_2+R(T-t_2)=R[T-(1-\alpha)t_2] \qquad (12)$$

$$=\frac{R(\alpha c_3 AP^2+B)}{c_3 AP^2+B}\cdot\sqrt{\frac{Q_0^2}{A}(1-\frac{c_2}{c_3})(\frac{1}{A}+\frac{c_2 p^2}{B})+2\frac{c_1}{c_3}(\frac{1}{A}+\frac{c_3 P^2}{B})}+\frac{R(1-\alpha)(c_3-c_2)Q_0 P^2}{c_3 AP^2+B}$$

The optimal production time was:

$$t^*=t-t_1=\frac{R}{P}[T-(1-\alpha)t_2] \qquad (13)$$

$$=\frac{R(\alpha c_3 AP^2+B)}{P(c_3 AP^2+B)}\cdot\sqrt{\frac{Q_0^2}{A}(1-\frac{c_2}{c_3})(\frac{1}{A}+\frac{c_2 P^2}{B})+2\frac{c_1}{c_3}(\frac{1}{A}+\frac{c_3 P^2}{B})}+\frac{R(1-\alpha)(c_3-c_2)Q_0 P}{c_3 AP^2+B}$$

3 Cases

3.1 THE ANALYSIS OF THE MODEL SOLUTION

(1) Model solution of different situations
1) When $\alpha=1$

$$T=\sqrt{\frac{Q_0^2}{A^2}(1-\frac{c_2}{c_3})(1+\frac{c_2}{c_4})+\frac{2c_1}{Ac_3}(1+\frac{c_3}{c_4})} ;$$

$$Q=\sqrt{\frac{P^2 Q_0^2}{(P-R)^2}(1-\frac{c_2}{c_3})(1+\frac{c_3}{c_4})+\frac{2c_2}{c_3}\cdot\frac{PR}{P-R}(1+\frac{c_3}{c_4})} ,$$

$$t=\sqrt{\frac{Q_0^2}{(P-R)^2}(1-\frac{c_2}{c_3})(1+\frac{c_2}{c_4})+\frac{2c_1}{c_3}(1+\frac{c_3}{c_4})\frac{R}{(P-R)P}} .$$

Here T, Q, t were the model of the optimal production cycles, the optimal production batch and the optimal production time for the shortage of warehouse capacity was limited to allow delay.

2) When $\alpha=0$

$$T\rightarrow T^*=\sqrt{\frac{PQ_0^2}{(P-R)R^2}(1-\frac{c_2}{c_3})(\frac{P}{P-R}+\frac{c_2}{c_5})+\frac{2c_1}{c_3 R}[\frac{P}{(P-R)}+\frac{c_3}{c_5}]} ,$$

$$Q\rightarrow Q^*$$

$$=\frac{c_5 P}{c_3(P-R)+c_5 P}\cdot\sqrt{\frac{PQ_0^2}{(P-R)R^2}(1-\frac{c_2}{c_3})(\frac{P}{P-R}+\frac{c_2}{c_5})+\frac{2c_1}{c_3 R}[\frac{P}{(P-R)}+\frac{c_3}{c_5}]}$$

$$+\frac{(c_3-c_2)Q_0 P}{c_3(P-R)+c_5 P}$$

$t \to t^*$

$$= \frac{c_5 P}{c_3(P-R)+c_5 P} \cdot \sqrt{\frac{PQ_0^2}{(P-R)R^2}(1-\frac{c_2}{c_3})(\frac{P}{P-R}+\frac{c_2}{c_5})+\frac{2c_1}{c_3 R}[\frac{P}{(P-R)}+\frac{c_3}{c_5}]}$$

$$+\frac{(c_3-c_2)Q_0 P}{c_3(P-R)+c_5 P}$$

Here T^*, Q^*, t were the model of the optimal production cycle; production batch and production time for out of stock all lost sales.

3) When $\alpha = 1$, $c_4 \to \infty, P \to \infty, c_2 = c_3$,

$$T \to T^* = \sqrt{\frac{2c_1}{c_2}\frac{1}{(P-R)}}; Q \to Q^* = \sqrt{\frac{2c_1}{c_2}\frac{R}{(P-R)}}.$$

T^*, Q^* were classic model of the optimal production cycle and the optimal production batch, which were not allowed out of stock storage.

(2) Stability analysis of the model solution

It was necessary to make the stability analysis for model which was not allowed out of stock storage [7], and its application was very wide .It was the foundation of other model when they were deduced.

3.2 THE MODEL OF EXAMPLE VERIFICATION

Through research, data related to inventory system was as follows: P=2000000 tons per year，R=1000000 piece per year, c_1=6000 RMB, c_2= 8.1RMB per ton•year, c_3=13.27 RMB per ton•year,c_4=20 RMB per ton•year,c_5=35 RMB per ton•year, Q_0 =78500 tons. Applying the model to calculate the optimal production cycles, the optimal inventory and a variety of cost under different the coefficient [8] of delaying pay, it was shown in Table 1.

TABLE 1 The influence of the coefficient of delaying pay to the optimal production cycles, the optimal inventory and a variety of cost

α	production cycles (year)	production lot size (ton)	Production time (year)	carrying cost (RMB)	Shortage cost (RMB)	Lost sales cost (RMB)	maximum inventory (ton)
0	0.1134	105010	0.05255	25850	0	1214	52550
0.2	0.1154	106768	0.05338	25704	212	1351	52299
0.4	0.1178	109532	0.05477	25542	605	1350	52015
0.6	0.1204	113598	0.05680	25368	1209	1169	51708
0.8	0.1235	119351	0.05968	25169	2079	788	51350
1	0.1285	128457	0.06423	24875	3602	0	50809

Table 1 showed that, as the delay coefficient alpha, shortage cost increases gradually, the storage cost and the maximum inventory had shrunk, lost sales cost was increased first and then decreases, while increasing production cycle caused by the holding cost, shortage of years lost sales cost increase was greater than the storage cost reduction, the enterprise couldn't extend the production cycle.

4 Conclusions

To summarize finished-parts storage inventory problems, and to describe the factors should be taken into consideration when the finished goods inventory was controlled, the key was to establish model ,in which iron and steel enterprise was considered allowing delayed delivery and lost sales inventory, and it obtained the most optimal production time, the production batch and production cycle within a production cycle, to achieve the minimum total inventory related products, finally, through a set of experimental data and the delay submitted the coefficient alpha effects on various fees.

This paper established model, which allowed the delayed delivery and considered the costs of inventory, and its solution and sensitivity was analysed, the analysis showed as follow:

(1) With the gradual increasing of delay delivery coefficient alpha and shortage cost, the storage cost and the maximum inventory had shrunk, lost sales cost was increased first and then decreased, while increasing production cycle in shortage caused by the holding cost,

lost sales cost increase was greater than the storage cost reduction, the enterprise couldn't extend the production cycle.

(2) When the demand rate was less than 1/2 productivity, classic inventory model had good stability, and closer to the actual inventory management problem.

In a word, using a large number of advanced management was an innovative management to research inventory management in supply chain management thought [9]. Especially, to the state-owned enterprises, they had to admit that a lot of advanced management ideas and management methods, with very low success rate, being carried out in China, one of the important factors was the basic problems of enterprise management. It would be a losing battle to implement a advanced management, in enterprises with weak basic management. As strip factory finished goods inventory management [10] application of supply chain management thinking, at the same time, it should adhere to the principle of grasping both the advanced management and basic management, held "standardization construction" as the breach of the basic management, business process institutionalization, standardization of basic data .Only in this way, management strategies proposed in this paper some could play a role and had practical significance.

References

[1] Wanlei Wang, Jia Xu, Jiyan Wang 2010 Model of Iron & Steel Enterprises Group Raw Material Requirement Planning Based on Consumer Chain *Computer Integrated Manufacturing Systems* **16**(5) 1074-81

[2] Zhiliang Yu, Lixin Wu 2004 Requirements Analysis and Design of Inventory Management System for Mould Enterprise *Mechanical & Electrical Engineering Technology* **33**(8) 39-41

[3] Shaohua Li 2013 Analysis and Countermeasures of Inventory Control Enterprise Logistics Management **35**(228) 70-1

[4] Bing Luo, Na Lu, Shuai Yang 2005 Vendor-managed Inventory Model with Time-limited Free Back-orders *College of Economics and Business Administration* **28**(5) 143-7

[5] Yan Xiao, Shuaihui Tian, Mingquan He, Kangqu Zhou 2007 Inventory Cost Model of Jointly Managed Inventory (JMI) Based on Supply Chain Management. **21**(10) 119-23

[6] Shanghai Wang 1999 Stock Decision Analysis for Random Requirement with Shortage Cost *Journal of Anhui Institute of Mechanical and Electrical Engineering* **14**(3) 19-22

[7] Daijun Wei 2006 Storing Model for Producing and Selling without Short Supply *Journal of Hubei Institute for Nationalities(Natural Science Edition)* **24**(3) 232-5

[8] Yingying Wang, Jinsong Hu 2010 Inventory Model with Controllable Lead Time and Backorder Rate Influenced by Lead Time and Price *Journal of Qingdao University(Natural Science Edition)* **23**(1) 95-9

[9] Fukuan Feng 2006 Discussion About Logistics Supply Chain Management in Iron & Steel Industry in China *Journal of Anhui Vocational College of Metallurgy and Technology* **16**(4) 86-8

[10] Yang Xi, Tieke Li, Cantao Shi 2011 Model and Method for Integrative Optimization of Inventory Matching of the Hot Rolled Coil and Cold Rolled Coil *Metallurgical Industry Automation* **35**(3) 14-8

Authors

Yuanmin Xie, born in November, 1972, born in China

Current position, grades: Master, Lecturer. College of Machinery and Automation, Wuhan University of Science and Technology
Scientific interest: Inventory management, Mechanical design and application
Publications: 10 articles

College ideological instruction teaching method based on multimedia CAI

Zhang He*

Department of Education, Baoji University of Arts and Sciences, Baoji, Shaanxi, China, 721013

Abstract

With the implementation of college ideological instruction teaching method reformation, multimedia and "three dimension" instruction method enjoy more and more popular among colleges. People place high hope on the new instruction pattern and regard it as the effective solution for education in college so far. This paper studies reformation of ideological instruction teaching method to analyse how to make use of multimedia CAI method and "three dimension" teaching method.

Keywords: college ideological instruction, multimedia network, CAI, "three dimension" teaching pattern

1 Introduction

With the progress and development of age, computer network enjoys popularity and develop rapidly. Science and technology is also widely applied in education. People pay more and more attention on spread of ideological instruction knowledge when they emphasize school performance. However, traditional teaching method restricts speed of ideological instruction and input of teaching knowledge. It has a negative effect on the effect of ideological instruction. For many years, ideological instruction always adopts traditional pattern. Teachers repeat what the book says and then ask students to think by themselves or memorize some knowledge and key point mechanically to sit for examination. Under this pattern, teachers are the leading role in the class. They offer few real cases for students to analyse and discuss. Therefore, CAI pattern is born at the right moment. It relieves educators from single education concept.

Computer Aided Instruction, short for CAI, becomes the important research issue of current teaching means in education field. CAI system refers to computer aided instruction system that composed of relative CAI course ware. Its fundamental objective is to realize effective teaching for students, which is consistent with that of traditional class teaching. CAI teaches students with computers. It aims to change traditional ideological instruction pattern and realize combination of leading function of teachers and subject role of students and combination of former lesson and latter lesson. It can fully inspire subjective initiative of learning ability of students and show "student-oriented" teaching concept in the true sense. In the process of lessons preparing, teachers collect and sort materials related to class content, apply network, audio and video technology to make course ware and use multimedia technology to play in class. Multimedia three dimensional teaching methods contain a large amount of multimedia information for students to obtain knowledge by various senses. Students can be more actively, effectively and comprehensively to participate whole process of ideological instruction. And the teaching method creates colourful learning environment of three dimensional ideological instruction teaching knowledge, improves learning efficiency of students and ensures the effect of ideological instruction teaching.

2 Multimedia characteristic

Research on characteristic can optimize the combination of multimedia in course ware and make layout and presentation of media factors have large visual impact and attraction. It can meet the practical needs of current multimedia teaching.

2.1 THREE DIMENSION

According to the current mainstream trend of multimedia, three dimensional images and text content is favoured by learners. Three dimensional multimedia can present plane content more specific and detailed for learners to understand an object more intuitively in different perspectives and aspects. Three-dimensional character presented by media factors break through previous content presentation form. Learners can have different mental feeling when they receive the content presented by multimedia course ware.

2.2 DYNAMIC

If dynamic effect is introduced into multimedia presentation, the presentation means will produce strong

---
* *Corresponding author* e-mail: Zhh.1226@163.com

impact and attraction for learners as well as centralize their attention. Dynamic factors can not only flexible multimedia but also present static things to students by multiaspect and multi-angle controllable means. Primary content is presented carefully in the form of freeze-frame play by the operation of teachers.

2.3 VISUALIZATION

Research shows that combination of different multimedia can play the effect to the largest. Therefore, visualization of content cannot be neglected based on the dynamic and tridimensional design and manufacture of course ware. Critical points of course ware should focus on image and then add words and sound for explanation. It can vivify and visualize content. And learners can be clear at a glance on the key point of the course ware.

3 Multimedia learning theory

American psychologist Richard Meier detailed studied multimedia and proposed multimedia learning theory in 1990's. His theory is based on dual coding theory and constructivism theory of Pevear and cognitive load theory of Sverre. The theory includes dual channels assumption, limited capacity assumption and active processing assumption.

3.1 DUAL CHANNELS ASSUMPTION

External information is transmitted and processed by auditory sense and visual sense channel. Visual information such as text and picture is transmitted and processed by visual channel. External sound information is collected and processed by auditory sense. In the late stage of processing, information such as picture and text will transform into sound after initial processing of vision. At last, it will be transmitted to auditory channel for last processing and storage. It is shown in Figure 1:

FIGURE 1 Dual channels assumption diagram

However, effective inspiration of reception and processing function of two channels is important for effective learning of learners in multimedia teaching because visual sense and auditory sense are mutually independent and the capacity is limited.

3.2 LIMITED CAPACITY ASSUMPTION

Efficiency of multimedia learning is restricted by the limited capacity of working memory. Information in working memory is the noticed content when instant memory is obtaining information. In other words, content in working memory is the knowledge that we need to obtain and other irrelevant information is shielded when it input into instant memory. Working memory connects instant and long term memory. The memory we needed will be screened and put into long term memory after processing to form knowledge base. However, the capacity of working memory is limited, which is consistent with instant memory. If external transmission information exceed the bearing capacity of memory capacity, learning of learners will be affected. External information will be hard to be related to the knowledge in long term memory and cannot be stored into long term memory. Text, image and sound information are included. Therefore, efficiency

improvement of learners is related to the reasonable distribution of text, image and picture in course ware.

3.3 ACTIVE PROCESSING ASSUMPTION

Active processing assumption has three processes. That is, selection, organization and processing. External information enters brain through eyes and ears of learners. It is a selective process. This process is screening the information entering into learners eyes according to the requirement of learners. It screens out important information. Then the information enters system processing through learners' eyes. Meanwhile, the needed external sound information is sent into auditory processing system after screening for learners reorganizing needed content after screening. After that, information becomes more complete and fluent. And image base of visual and auditory sense will be formed according to the ultimate requirement of learning. Last is integration process. Learners build connection of image and sound information and associate these two kinds and previous information of long term memory. It forms new knowledge schema and stored in long term memory after last processing. However, sound and image information is in working memory when it associates with long term memory. Therefore, the size of

working memory capacity directly affects the process of active processing. The process is repeatable. Thus, learners can make up lost information caused by limitation capacity of working memory through many times learning.

4 Multimedia course ware design of college ideological instruction

Broadly speaking, multimedia course ware design is preplanning for a whole set of teaching assisted system, imagination and arrangement for the whole plan and also the detailed regulation and blueprint. Design of multimedia course ware mainly includes visible teaching design, structure design and interface design. Only the comprehensive arrangement and scientific design in teaching course ware design can make the multimedia course ware become scientific and reasonable and possess certain artistry, contemporaneity and technicality.

4.1 DESIGN OF IDEOLOGICAL INSTRUCTION COURSE WARE TEACHING

Teaching design is the basis of multimedia course ware design. It adopts reasonable method to analyse teaching problem systemically, clears teaching process and objective and establishes strategy for teaching problem and process. The purport is to solve teaching objective, content and method. Therefore, teaching design has important function in the process of design process of visual multimedia.

4.1.1 Analysis of teaching content

Teaching content is to realize teaching objective stipulated in outline. It includes various teaching, training and practical activity accepted by students and arranged by system. Analysis of teaching content is the basic working of teaching. It mainly includes depth and breadth of teaching. It reveals the relationship between parts if teaching. This work is usually fulfilled with the help of teachers who have rich teaching experience and a good command of teaching rule. Teaching content has certain design ability and hierarchy. The number of levels of ideological instruction content division is consistent with the levels in communication. And the following procedures should be followed:

a. Select and organize units. Confirm out general objective for the course of ideological instruction. Then the content needed to be learned is divided into several modules and done organizational hierarchy to simplify the content for students to learn step by step.

b. Confirm unit teaching objective and the knowledge that need to be achieved when students is learning the content of ideological instruction unit. Only when the chapter content is confirmed can teachers teach students accordingly. And students can understand which content need to be understood in the class.

c. Evaluate the content and separate the unit content projected in ideological instruction. A big concept is divided into several small specific concepts. Or a complicated knowledge point is divided into several small knowledge points for absorption.

4.1.2 Setting of teaching objective

It is the second procedure of teaching analysis. Its objective is to analyse, which objective and degree can be students achieve through knowledge learning of that part. Teaching objective can be divided into several hierarchies according to learning requirement. For instance, we often require memorizing some concept and evaluating some idea in ideological instruction class. And these objectives should point out degree and standard needs to be achieved for effective inspect and measure learning effect of students.

4.2 DEIGN OF COURSE WARE STRUCTURE

Organization structure of ideological knowledge content is varied. Its basic forms mainly include linear structure, branch structure, net structure and tree structure. Traditional multimedia ideological instruction teaching uses linear structure organized by text, sound and image. It has characteristics of ordering from front to back and taking time as node. However, the system structure of multimedia course ware practically is the organization and presentation means of teaching content. Design of multimedia ideological instruction course ware should make scientific program on course ware system according to the whole teaching content.

4.3 GENERAL PRINCIPLE AND DETAIL ATTENTION

CAI is the common multimedia teaching means in college nowadays. Critical factor that present quality and effect of multimedia teaching is manufacture of course ware. Design of this kind of multimedia course ware has its own specific principle and procedure method.

4.3.1 General principle

This kind of course ware is multimedia application software. Text, sound, diagram, picture, animation and image are integrated together through software design and reasonable composing. The following principles need to be following in design process.

a. Controllable principle. Basic requirement of CAI course ware interface is that initiator or learner always controls course ware operation and the course ware is in the form of interaction. Interface is in the form of facing task for learners to fulfil learning task more rapidly.

b. Easy to learn and use principle. CAI software that can be mastered and used by learners in short time is good multimedia.

c. Consistency principle. It means that real world and course ware content should be consistent. The former can shorten the learning time of learners and improve learning efficiency. The latter provides convenience for teachers to make course ware manufacture more convenient.

d. Guiding principle. Learners fulfil study through direct and visualized means.

e. Economical principle. Interface design should aim at reduce the operation time of learners, such as less words input and replace essay question with choice question.

f. Optimal combination principle of media. Utilization and selection of all media information are all serve for ultimate teaching objective. Excess use of media information should be avoided by all means.

4.3.2 Detail attention

a. Colour collocation.

Teachers should select correct colour keynote including background and colour of words for students to see more nature and beautiful ideological instruction multimedia course ware. The tone of the whole screen is also very important. In the process of making course ware, teachers are required to design course ware with unified and harmonious colour and natural style.

b. Design of cover and title

Cover of course ware includes course name, make time and author. Students can clear about the beginning of course through cover. However, title is the opening words of a new class. It drives students' interest and attention by video or animation. Title design should related to the content of the course. Its length and content should adopt to the whole design and ideological instruction knowledge that imparted in that class.

c. Script design

Script is language description means of software design, which is the bridge between ideological instruction course ware designer and makers and the basis of the whole making. However, course ware designer and maker are assumed by the same person to a large extent in the perspective of current multimedia development. Script designates specific and clear scope on course ware design and making for the maker of course ware. The making of script of ideological instruction generally includes collocation and skip of course ware content arrangement, material organization of page, layout, show time of every page, colour keynote of page, words, music effect and interaction mode of people and computer. Script compiling should design according to the teaching content of ideological instruction teaching plan. Its organization mode, presentation mode and material selection should close to teaching content and plan. Current script design has no specific speculation. Generally, script design includes window presentation size of multimedia material, presentation mode and special effect, collocation and combination of different materials, interaction mode, skip, hot area, hyperlink, control component and a series setting of navigation, return and error control mechanism.
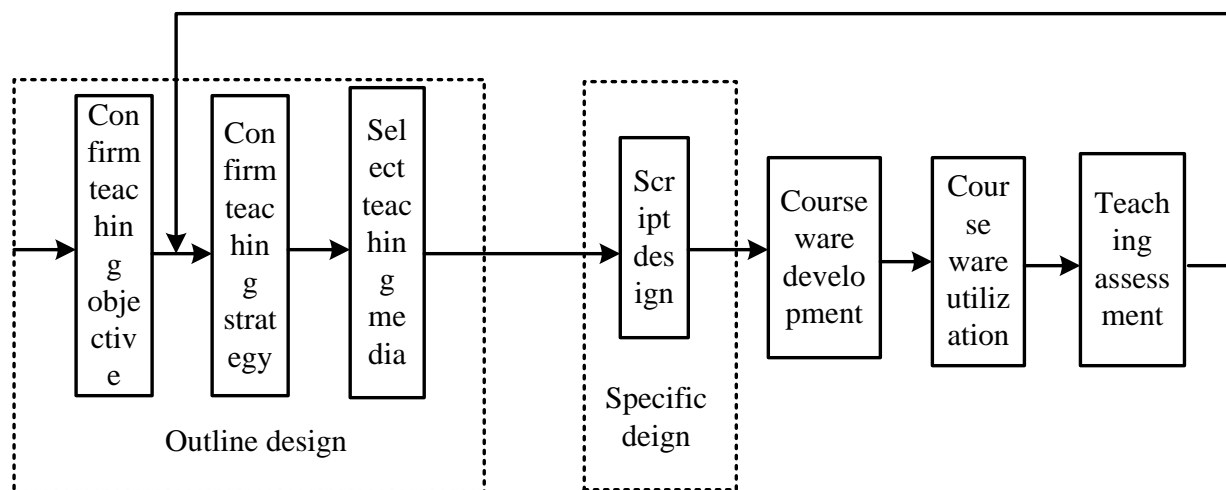

FIGURE 2 Main process of CAI ideological instruction course ware design

Above is the allocation of basic hardware in ideological instruction multimedia teaching. If the equipment is of high quality, then use of multimedia course ware will be handier.

5 Presentation mode of multimedia course ware in college ideological instruction

5.1 DYNAMIC DEMONSTRATION

Multimedia course ware used in ideological instruction class applies PPT as demonstration platform. In teaching process, it is common that teachers switch slide combination manually according to their lecture speed. However, it violates original intention of multimedia

making and utilization. The main function of multimedia course ware is to assist teachers to teach. Only course ware which do not need mutual switch and can operate dynamically can reduce workload for teachers and drew the attention of learners.

Dynamic demonstration means to realize automation demonstration of multimedia teaching content by very few operations. This kind of automation is a demonstration combining teaching content and assisted cognitive material. Some demonstration can present the whole ideological instruction teaching process by one-off animation as showing film. Some require teachers to do separate demonstration according to course content. However, dynamic demonstration is not all dynamic. Importing all

information to learners by one course will cause jam in information reception channel of learners. It is bad for students to understand and receive information and makes learners loss direction and key point. Therefore, teachers should based on outline of ideological instruction teaching in multimedia design and demonstrate chapter. Course ware will stop after the demonstration of one chapter or one knowledge point and pop learning navigation for teachers to explain. And learners can have time to process and memorize knowledge fully. Dynamic demonstration transmit and sort knowledge by visual sense channel of learners for them pay attention to information by sensory stimuli.
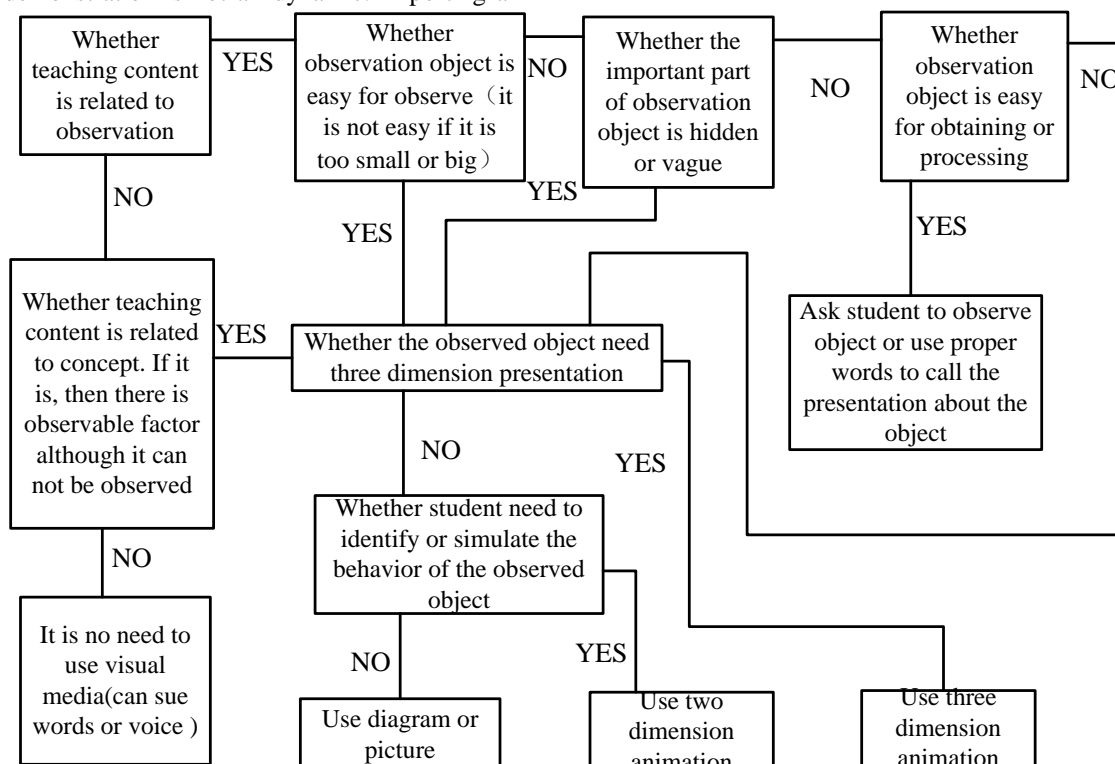
FIGURE 3 Process selection figure of visual media

5.2 ASSOCIATION OF ACTIVITY AND INERTIA

Demonstration of ideological instruction in form of multimedia is various. Association of activity and inertia can achieve teaching assisted task. However, the current educational application pattern is primary in inertia assisted by activity. Generally, it is presented in the screen in form of text. Although picture and video are used to strengthen perceptual understanding, mutual operation of teacher is still needed overall.

Authentic multimedia presentation should be association of activity and inertia. Instant memory of learners is stimulated by means of sound for attention to transmit and sort this kind of information. In the process of demonstration, freeze-frame should be adopted for learners to have enough time to call and memorize demonstration content and do abstract and summary on key

content and theoretical system. Learners can memorize and construct knowledge fully by their means in freeze-frame. In static demonstration, media material such as video, animation and sound should be inserted properly to vivify static text. In addition, addition of sound stimuli can draw the interest and attention of learners and strengthen the understanding and memory for special content structure.

Course ware with association of activity and inertia is primary in visual sense assisted by auditory sense. During ideological instruction teaching, means of auditory sense can make learners to better accept and learn knowledge. Fact proves that rich interest can inspire the effective motivation of learning, mobilize learning enthusiasm and intensify the internal power of learning. Three dimension multimedia teaching can transform perceptual materials in various forms into vivid picture by text, picture, animation and sound. And the abstract and static theoretical

knowledge will be intuitive and vivid. Attractive scene that is created by acousto-optic diagram can act on different sense directly.
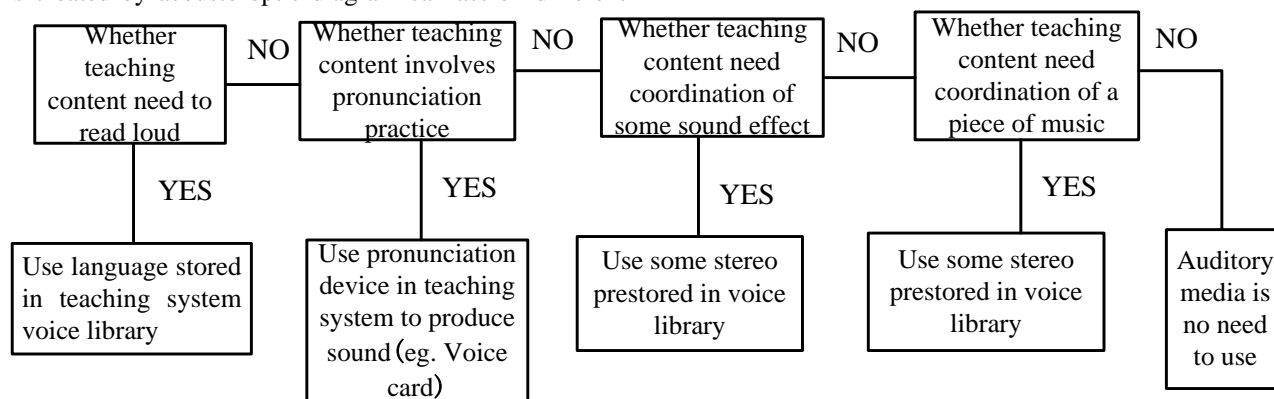

FIGURE 4 Process selection diagram of auditory media

5.3 SUPPORT OF HARDWARE SYSTEM

Design of ideological multimedia CAI course ware can not leave the coordination of hardware system. It is the basic system needed in multimedia course ware development. Hardware system is the basis. That is to say good course design cannot be demonstrated if there is no satisfactory hardware to coordinate.

Multimedia hardware system is the multimedia computer for demonstrating course ware. In course ware design process, collection, sorting, processing and accomplishment of material is a series of work. And multimedia computer with high quality allocation is required. Multimedia computer applies multimedia function in computer. Basic constitution of multimedia computer is as follows:

a. Host computer: PC; b. video and voice input equipment: camera, microphone, sound recorder, etc.; c. video and voice output equipment: projector, projection screen, loudspeaker, stereo headphone, etc.; d. NFFC: video card, sound card, video card, network card, etc.; e.storage device: CD-ROM, mobile HDD and CD witter; f. Intercrossing equipment: keyboard, mouse, etc.; g. Software: operating system, hardware driver and application program.
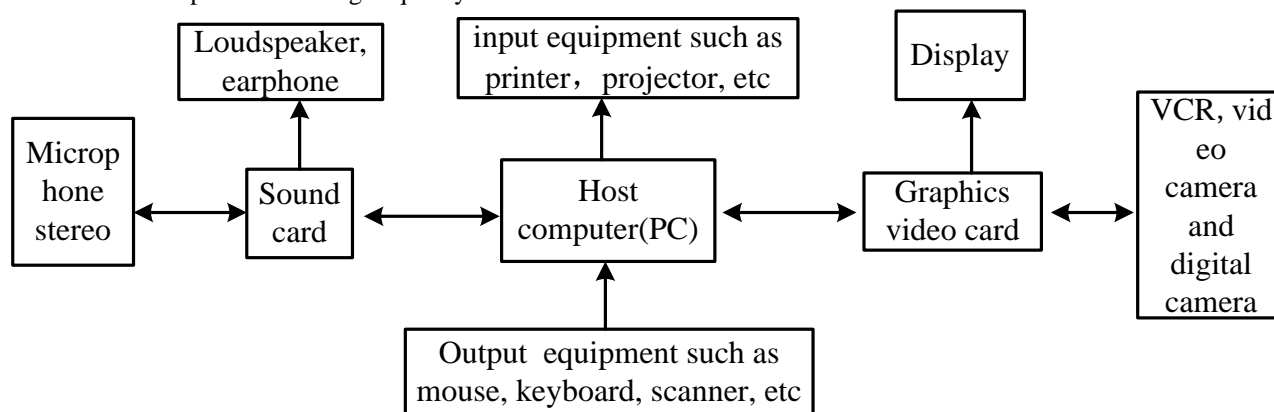
FIGURE 5 Multimedia computer composition figure

6 Conclusion

All in all, traditional method relies on s stick of chalk, a blackboard and a textbook can not keep up with the requirement of age development and meet the need of students' learning for the strong theoretical and synthesis property of ideological instruction teaching and abstraction and complex of ideological instruction teaching content.

Multimedia three dimension teaching is demonstrated to students in form of word, sound, picture, video and music to inspire the learning enthusiasm of students. Meanwhile, it proposes strict requirement for teachers. It requires them to possess prospective educational idea, strong teaching skills and persistence. Introducing multimedia three dimension patterns into the practice of ideological instruction teaching is a general trend of new round course reformation of ideological instruction teaching. It can give full play of guidance, assistance and monitoring and highlight learning leading role of students. It can help students to better understand and master knowledge and improve professional skill and help teachers to teach students in accordance of their aptitude. As a result, students can do some ego early plan and detection effectively in independent study process and achieve

ultimate adjustment to improve lifelong learning ability and innovation ability.

7 Acknowledgements

References

[1] Guan H X 2008 *College ideological and political educational theory and practice* Harbin: Harbin Publisher

[2] *Social Science Department of Educational Ministry edit. Selected literature of ideological and political theory course in ordinary university* Beijing: China Renmin University Press 2008

[3] Luo Y G, Zhang Y D, Mei Y Y 2010 Brief discussion on conversion of textbook system to teaching system of ideological and political classes *Ideological and Theoretical Education Tribune*

[4] Lu S X 2002 Research on the basic structure ofmultimedia technology course content *Journal of Pingdingshan Normal Training College* **2**

[5] Ding Y J 2012 Research on application of multimedia course ware of ideological instruction teaching in higher vocational college *Youth Literator*

[6] Ke S Y 2011 Rustic opinion on integration of ideological and political course teaching and network multimedia *Heilongjiang Sci& Tech Information* **22** 2011

[7] Diao J Q, Huang M X 2012 Application of multimedia in ideological and political course teaching *China Educational Technique & Equipment* **3**

[8] Liao R G 2013 Optimizing political course teaching by skillful use of multimedia *Southern Journal* **2**

[9] Tsiriga Victori, Maria Virvou 2004 A Framework for the Initialization of Student Models in Web-based Intelligent Tutoring Systems *User Modeling and User-Adapted Interaction* **12**

[10] Rossafri Mohamad 2012 The design, development and evaluation of an adaptive multimedia learning environment courseware among history teachers *Technology*

[11] McWilliams E, Taylor P G 2007 Teacher im-material: Challenging the new pedagogies of instructional design *Educational Researcher* **27**

[12] Koper R 2012 PROFIL: a method for the development of multimedia courseware *British Journal of Educational Technology* **94**

Authors

He Zhang, born in 1978, Shannxi Province of China

Current position, grades: lecturer
University studies: Master degree was earned in major of pedagogy, Shannxi Normal University in 2009.
Scientific interest: educational economy and management

Theoretical analysis and experimental verification of hole surface finishing parts

Wenhui Li, Shengqiang Yang*, Xiuhong Li

College of Mechanical Engineering, Taiyuan University of Technology Taiyuan, Shanxi, 030024, P.R. China

Received 1 October 2014, www.cmnt.lv

Abstract

As the main method of improving surface quality, finishing technologies have been developed and applied rapidly in recent years, but they have certain limitations in the hole surface finishing. Based on the principle of centrifugal motion, a kind of hole surface finishing method with self-adaptive ability is putting forward in order to solve the finishing and cleaning problems effectively. The force of the grinding rod is analysed, so the relationship between speed and friction is defined by analysing the minimum speed of the finishing parts theoretically. Finishing parts are designed and experimental study is done. Research results show that hole surface roughness value Ra of seamless steel tube can reduce from 7.0μm to 0.3μm in 50s, which verifies the finishing effect and efficiency.

Keywords: finishing, hole surface, centrifugal principle, finishing effect

1 Introduction

In order to obtain lower surface roughness value and good physical-mechanical performance and improve cleanliness of the hole surface, there have been many kinds of precision machining methods to improve hole surface quality and service life. The traditional finishing methods mainly include honing, grinding, rolling, polishing, etc. Non-traditional finishing methods include magnetic abrasive finishing, electrochemical finishing, abrasive jet, ultrasonic grinding, thermal deburring, two-phase flow hole surface finishing, etc. [1-7]. Because of their different technological characteristics, they are difficult to finishing hole surface [1, 2, 8, 9]. Based on centrifugal principle [11], a kind of hole surface finishing method is putting forward in order to solve hole surface finishing and cleaning problem effectively.

As is shown in Figure 1, the whole finishing parts is placed in the hole of the workpiece, and the whole parts rotate in a certain speed driven by electric motor.

Under the centrifugal force, the floating grinding rods placed on the bracket would exert certain force on the hole surface and realize surface finishing finally under the action of relative rotational motion between grinding rods and the workpiece and the axial motion of the automatic feed device.
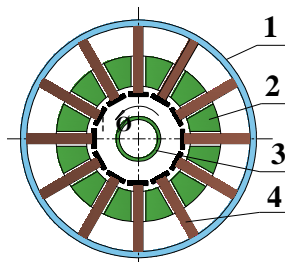
1) Workpiece 2) Bracket 3) Centre frame 4) Grinding rod
FIGURE 1 Schematic of flexible finishing parts

2 Theoretical analysis

2.1 THE MINIMUM SPEED OF THE GRINDING RODS

The force analysis of the grinding rod when it stays on vertical position is shown in Figure 2. The positive pressure F^n is zero because the grinding rods do not contact with the hole surface. When grinding rods accelerate at the starting moment, it would generate inertia force on the bracket, at the same time, the bracket reacts to the grinding rod. In Figure 2, the F^τ represent the tangential inertia force, F_1 and F_2 are the reactive force on the grinding rod, F_1' and F_2' are the friction force between the grinding rod and bracket.

The rods must overcome gravity and friction force in order to achieve the hole surface finishing, which is:

$$m\omega_0{}^2 r > mg + F_1' + F_2', \qquad (1)$$

where, m is the quality of grinding rod, ω_0 is the minimum angular velocity of centrifugal motion of grinding rod, r is the radius between the centroid of grinding rod and the centre of finishing parts.
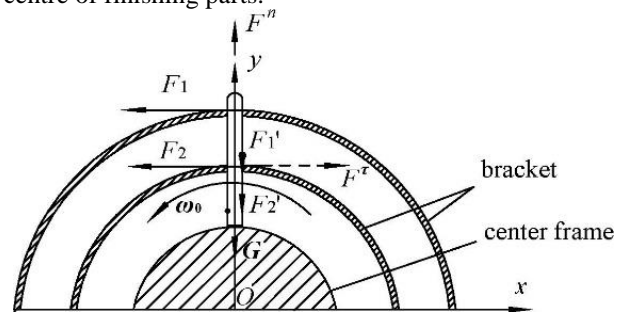

FIGURE 2 Force analysis of grinding rod when finishing parts start rotating

*Corresponding author e-mail:tutysq@263.net.cn

Assuming the angular acceleration is α when the motor started, the tangential inertial force of the grinding rod is:

$$F^{\tau} = mr\alpha . \qquad (2)$$

As $F_1 + F_2 = F_{\tau}$, the friction force between the grinding rod and the bracket is:

$$F_s = F_1' + F_2' = (F_1 + F_2)f = mr\alpha f , \qquad (3)$$

where, f is the static friction coefficient between the grinding rod and bracket, F_s is the static friction force between the bracket and the grinding rod. Leading to:

$$\omega_0 \geq \sqrt{\frac{g + r \cdot \alpha \cdot f}{r}} . \qquad (4)$$

It can be seen from Equation (4) that the minimum angular velocity is related to the distance r, the friction coefficient f and the starting angular acceleration α.

2.2 RELATIONSHIP BETWEEN SPEED AND GRINDING FORCE

As grinding rods do circular motion along the hole surface, grinding rods will bear the maximum resistance at the vertical direction. Figure 3 shows the force of the grinding rod, where, N is the positive force, F_f is the friction force, F_a and F_b are the positive force on the grinding rod from the outer and inner end of the bracket, and ω is the angular velocity.
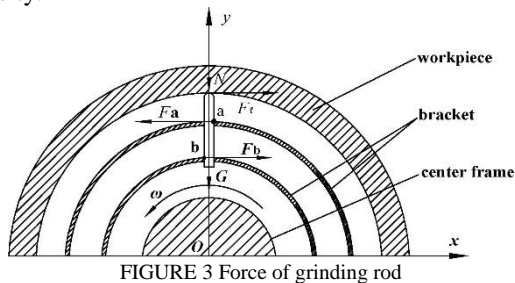
FIGURE 3 Force of grinding rod

Taking the rod as research object, the force balance equations of vertical and horizontal directions are as follows:

$$m\omega^2 r - mg - F_s = 0 , \qquad (5)$$

$$F_a - F_f - F_b = 0 . \qquad (6)$$

The moment balance equation from the section centre of point a is as follows:

$$F_f \cdot l_1 = F_b \cdot l_2 , \qquad (7)$$

where, l_1 is the grinding rod length between the hole surface and the bracket outer end, l_2 is the inner and outer length of the bracket. So:

$$F_a = \left(1 + \frac{l_1}{l_2}\right)F_f , \quad F_b = \frac{l_1}{l_2}F_f , \qquad (8)$$

$$F_s = F_a' + F_b' = (F_a + F_b)f = \left(1 + \frac{2l_1}{l_2}\right)F_f f , \qquad (9)$$

where, F_s is the static friction force between the bracket and the grinding rod. Synthesizing Equations (5)-(9), ω can be expressed as:

$$\omega = \sqrt{\frac{F_f}{f' \cdot m \cdot r} + \frac{g}{r} + \frac{F_s}{mr}} , \qquad (10)$$

thus:

$$\omega = \sqrt{\left(\frac{1}{f' \cdot m \cdot r} + \frac{f}{m \cdot r} + \frac{2l_1 \cdot f}{l_2 \cdot m \cdot r}\right) \cdot F_f + \frac{g}{r}} , \qquad (11)$$

so the relationship between angular velocity ω and friction force F_f is defined according to Equation (11). When the finishing parts is designed, f, f', m, r, l_1, l_2 are certain, and different F_f can be obtained with the change of ω, so finishing efficiency can be altered. Figure 4 is the simulation result of relationship between rotational speed n and friction force F_f.
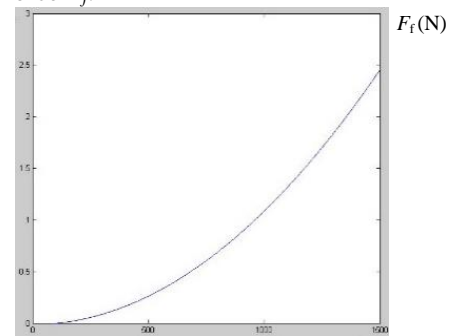

FIGURE 4 Simulation result of relationship between rotational speed and friction force

3 The structure of finishing parts

According to centrifugal principle, the structure design of finishing parts should meet the following requirements:

1) The structure of finishing parts should be reasonable and simple and the size should meet the need of flexible finishing.

2) Lighter quality to reduce the startup load of the motor.

3) The guide hole of the grinding rod should be well-distributed and have good guidance and lubricity.

According to the above requirements, finishing parts are designed and consisted of bracket, centre frame and cover plate. As is shown in Figure 5, bolts connect the bracket and centre frame with cover plate, a certain gap is produced between bracket and mandrel. Placing grinding rod into the bracket and the centre frame, grinding rods are supported into the guide hole. The size of bracket and centre frame can be adjusted to meet the need of finishing. Figure 6 is the photos of finishing parts in different positions.
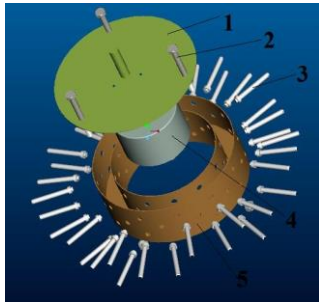
1) Cover plate 2) Bolt 3) Grinding rod 4) Centre frame 5) Bracket
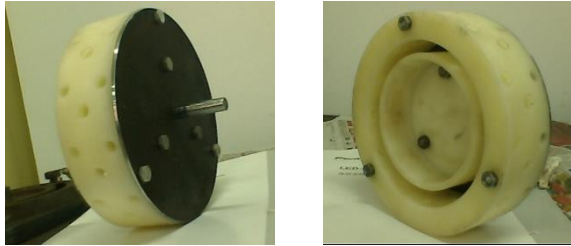FIGURE 5 Structure of flexible finishing head

FIGURE 6 Photos of finishing parts in different positions

4 Experimental analysis

Clamping finishing parts on the spindle of the drilling machines, the workpiece (inner diameter is 180mm seamless steel tube in 45# steel) is fixed on the workbench, Figure 7 shows the experimental device and finishing status.

1) Drill 2) Finishing parts 3)Workpiece
FIGURE 7 Photos of experimental device and finishing status

4.1 EFFECTS OF MOTION PARAMETERS ON FINISHING

The motion parameters of finishing parts mainly include rotational speed and feed speed [9, 12]. Figure 8 shows the change curves of surface roughness value Ra along with finishing time with silicon carbide abrasive of 200 meshes and the spindle rotational speed of 800r/min and the axial feed speed of 2m/s and 4mm/s.
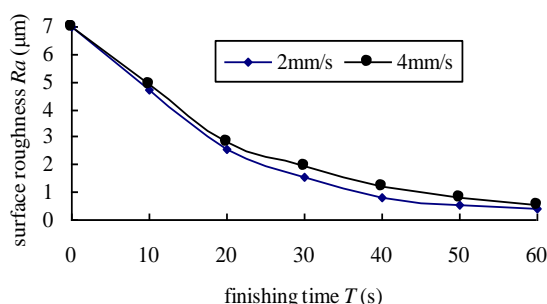
FIGURE 8 Ra with finishing time under different feed speed

The contrast photos before and after finishing is shown in Figure 9. The results show that Ra can be reduced to 0.41μm from 7.0μm within 60s. The time of grinding rods contacting with the certain position of the hole surface depends on the axial feed speed. The lower feed speed is, the longer finishing time on the same position is and the better effect during the same time. Ra decreased significantly when the feed speed varies from 4mm/s to 2mm/s.
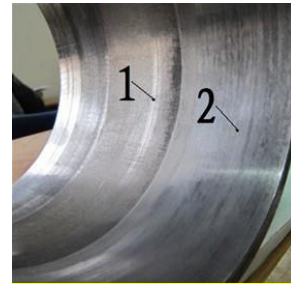
1) before finishing 2) after finishing
FIGURE 9 Contrast photos before and after finishing

With the axial feed speed 2mm/s and rotational speed 1000r/min and 1500r/min under the other same conditions, the change curves of surface roughness value Ra along with finishing time are shown in Figure 10. The results show that Ra can be reduced to 0.3μm from 7.0μm within 50s, which indicates that finishing efficiency will be improved with the increase of rotational speed within certain limits at the same feed speed and the grain size. Experimental results are consistent with theoretical analysis.
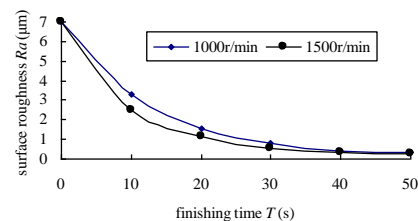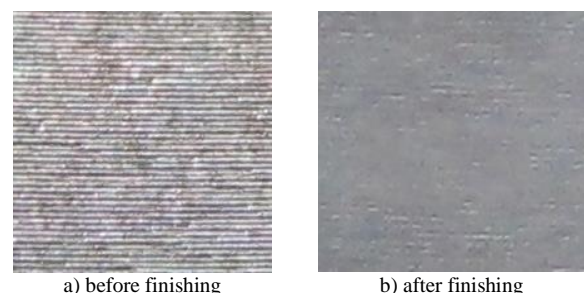

FIGURE 10 Ra with finishing time under different rotational speed

Figure 11 shows the surface texture of the workpiece before and after finishing. Results show that surface texture varied from anisotropy to isotropy, which indicates the surface texture has been significantly improved.

a) before finishing   b) after finishing
FIGURE 11 Surface texture before and after finishing

4.2 EFFECTS OF THE GRAIN SIZE OF GRINDING ROD ON FINISHING

In order to improve processing efficiency, the grain size of grinding rods should be coarse-grained as soon as possible to meet the requirements of the surface quality. Under the

condition that rotational speed is 800r/min, the feed speed is 2mm/s and the grain size of grinding rod is 320 meshes and 600 meshes, the change curves of the surface roughness along with finishing time are shown in Figure 12. The results show that the efficiency of the large grain size is higher than that of the small grain size in a certain range within the same rotational speed, but the surface roughness value is higher under the larger grain size.
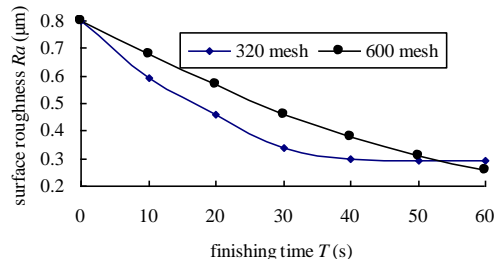

FIGURE 12 Ra with finishing time under different grain size of grinding rod

5 Conclusions

Based on centrifugal principle, a kind of hole surface finishing method is put forward in order to realize hole surface finishing and cleaning effectively. The minimum speed and the relationship of speed and grinding force were derived from theoretical analysis of the grinding rods, which provide theoretical basis for the design of finishing parts and experimental study. Experimental study shows that hole surface roughness value Ra of seamless steel tube has reduced from 7.0μm to 0.3μm in 50s and the surface texture has been significantly improved, which indicates that it has good finishing effect and efficiency.

The structure characteristics of the finishing parts can be described as follows:

1) Good adaptability. The size of finishing parts is flexible and finishing strength can also be controlled. It is suitable for finishing a range of tenuous holes, variable diameter hole, curve hole, elliptical hole and other special hole.

2) Finishing and repairing large parts. For large parts that are not easy to carry, the device can be directly placed on the workpiece to realize finishing and cleaning of the hole surface.

Acknowledgments

References

[1] Yang S C, Li W H, Chen H L 2011 Theory and new technology of the surface integrity *National defense industry press*

[2] Ulutan D, Ozel T 2011 Machining induced surface integrity in titanium and nickel alloys: A review *International Journal of Machine Tools & Manufacture* **51** 250-80

[3] Yuan J L, Zhang F H, Dai Y F, Kang R K, Yang H, Lv B M 2010 Study on the development of the science technology of ultra-precision machining *Journal of mechanical engineering* **46**(15) 161-77

[4] Li Ch H, Hou Y L, Cai P G, Lu N H 2009 The summary of the precision finishing using free abrasive *Precision manufacturing and automation* **1** 4-9

[5] Gavas M, Karacan A, Kaya E 2011 A novel method to improve surface quality in cylindrical grinding *Experimental Techniques* **35**(1) 26-32

[6] Yang S Q, Li W H, YangS C 2010 The surface finishing technology research of inner hole *Modular machine tools and automatic technology* **1** 56-60

[7] Alao A R, Konneh M 2012 Surface finish prediction models for precision grinding of silicon *International Journal of Advanced Manufacturing Technology* **58** 949-67

[8] Li W H, Yang S Q, Yang S C 2011 Technique project and experimental study of gas-particle two-phase compulsive circulation flows finishing *Advanced Materials Research* **305** 438-41

[9] Li W H, Li X H 2012 Technique project and experimental study of viscoelastic magnetic abrasive finishing *Advanced Science Letters* **12** 30-3

[10] Li X H, Yang S Q, Yang S C, Li W H 2009 Simulation research on the swirling airflow compounded with magnetic-field finishing *Key Engineering Materials* **407-408** 658-61

[11] Liu Z F, Yan S J, Feng X G 2012 Dynamic analysis of elastic body with rotation gradient effects in centrifugal field *Journal of Vibration and Shock*, **31**(16) 164-8

[12] Mandal B, Singh R, Das S, Banerjee S 2011 Improving grinding performance by controlling air flow around a grinding wheel, *International Journal of Machine Tools and Manufacture* **51**(9) 670-6

Authors

Wenhui Li, born in May, 1975, Shanxi, China

Current position, grades: Doctor on mechanical manufacture and automation, Associate Professor and master's supervisor in Taiyuan University of Technology, Assistant Dean of College of Mechanical Engineering.
University studies: Mechanical Manufacture in Taiyuan University of Technology.
Scientific interests: precise surface finishing in mechanical engineering.
Publications: 5 patents, 50 scientific papers, 1 monograph.

Shengqiang yang, born in August, 1964, Shanxi, China

Current position, grades: Doctor of mechanical manufacture and automation, Professor and doctor supervisor in Taiyuan University of Technology
University studies: IC Engine Design and Manufacture in Taiyuan University of Technology.
Scientific interests: Precise surface finishing in mechanical engineering.
Publications: 5 Patents, 70 scientific papers, 1 monograph, 5 textbooks

Xiuhong Li, born in June, 1972, Shanxi, China

Current position, grades: Doctor on mechanical manufacture, associate professor and master's supervisor at Taiyuan University of Technology.
University studies: Mechanical Manufacture in Taiyuan University of Technology
Scientific interest: Precise surface finishing in Mechanical Engineering
Publications: 3 Patents, 15 scientific papers, 1 monographs, 2textbooks

Research progress of implantable intra-body communication

Shuang Zhang[1, 2], Yuping Qin[2], Jing Xiao[3], Yihe Liu[1*]

[1]*College of computer science, Neijiang Normal University, Neijiang, 641000, China*

[2]*The engineering & technical college of Chengdu university of technology, Leshan, 614000, China*

[3]*Air force Logistic College, Xuzhou 221000, China*

Abstract

The intra-body communication is an emerging wireless communication technology. According to coupling modes of electrodes, the intra-body communication is classified into two types, the capacitive and the galvanic coupling intra-body communication. The capacitive coupling communication is inappropriate for the medical implant intra-body because this communication mode requires the common grounding, while the galvanic coupling communication can exactly make up for the disadvantage of the former. In existing research overview, prototypes and experiments concerning the two coupling communication modes are thoroughly discussed, and research status of "surface-to-surface", "surface-to-implant", "implant-to-surface" and "implant-to-implant" communication methods is emphasized as per installation positions of electrodes. Furthermore, opportunities and challenges of the communication technology are presented as well as its prospects.

Keywords: intra-body communication, capacitive coupling, galvanic coupling, implantable, sensor

1 Introduction

As an important branch of the wireless sensor technology and an important public application network, the BAN (body area network) technology is significant and in great demand in many fields such as the remote medical treatment, the special population' nursing and health care for the community, which becomes the focus in relevant study and application gradually. With the on-going increase of the aging population all over the world, the medical resource (budgetary outlays, doctors, nurses, sickbeds etc.) is relatively insufficient, so this makes the development of the medical and health care system become the global demand [1]. However, China has a 1.3 billion population, so it is more urgent for China to develop the BAN technology so as to solve the problem of "hard and expensive to see a doctor" in most communities (especially remote mountainous regions) [2]. In addition, most traditional medical treatment methods are used only at the onset of illnesses, and fail to prevent sickness and implement real-time treatment. But new medical technologies represented by BAN can be used to provide early warning for an oncoming illness or give timely warning at the onset of the illness through classified-learning of the existing physiological data and analysis of real-time signals or data, and to save important physiological information at an attack of a disease for subsequent diagnosis and treatment.

BAN is established with some sensors mounted on human body surface or embedded in human body to achieve inter-communication through certain types of connection, and the system is generally called the medical nursing system [3]. In general, these sensors can be connected with the base station through the following connection types: wired connection, wireless connection and intra-body communication.

In the intra-body communication, signals are transmitted by virtue of good conductivity of human body. Therefore, the common short-distance wireless communication, in which air is used as the information transmission medium in the intra-body communication, plays an important role in the wearable medical nursing system due to low power consumption, strong anti-jamming ability against external electromagnetic noise and less radiation energy [3] as well as no complicated wiring for signal transmission.

This paper starts with different coupling modes in the intra-body communication. Then we summarize existing achievements in the capacitive and the galvanic coupling intra-body communication from researchers (or research organizations), coupling modes, carrier frequencies, communication speeds, modeling methods, application fields and other aspects. Moreover, we present opportunities and challenges of the galvanic coupling intra-body communication so as to give comments and suggestions for further research.

2 Research progress of intra-body communication

2.1 CAPACITIVE COUPLING INTRA-BODY COMMUNICATION

The intra-body communication was firstly proposed by the graduate student of MIT, Zimmerman and other

researchers in 1995 [5], they used the transmitting electrode to input a 330 KHz weak signal (30V, 50PA) into human body. The signal was transmitted through human body and was received by the detecting electrode at a transmission rate of 2.4kbit/s, with power consumption of 1.5mW. In the system they designed, a transmitting electrode and a receiving one were mounted near human body to carry out the contactless coupling through the capacitive coupling. However, another pair of electrodes (a transmitting electrode and a receiving one) was utilized to implement the ground coupling, which was called the capacitive coupling intra-body communication.

Since Zimmerman proposed the intra-body communication model, the capacitive coupling intra-body communication technology has been developed further. Gray [6] designed a device, of which the maximum theoretical bandwidth was raised to 2000kbps at the modulation frequency of 100 kHz. He showed that noise sources in the IBC device were mainly from the amplifier in the circuit and external electromagnetic interference. Next, Post [6] achieved the FSK (Frequency Shift Keying) half-duplex communication at a transmission rate of 9600 baud rate in his device, which meant that the intra-body communication had been developed for data and energy transmission through human body. Later on, Partridge [7] from University of Washington designed an improved device at a transmission rate of 38.4kbit/s. In the device, carrier frequencies of 180 kHz and 140 kHz were used in the FSK modulation mode, with the adjustable voltage (the maximum of 22V); on the basis of the above conditions, a large number of experiments were carried out in terms of signal intensity, transmission rate and other communication indexes. Those experiments showed that appearance of the grounding electrode of the transmitter (or the receiver) had no obvious effect on the signal, instead, the distance between signal electrodes produced a remarkable effect upon the signal, and the signal was attenuated more severely with the decrease of the ground contact area.

In 2004 NTT (the Nippon the Telegraph & Telephone) [8] from Japan published the RedTacton developed by themselves, which marked the beginning of practical application of the intra-body communication; in addition, they achieved the maximum half-duplex transmission rate of 10Mb/s. Later, an intra-body communication device named Wearable ID Key was also developed by Sony Corporation and Chiba University [9] of Japan. The transmission rate of the device reached 9600bps in the FSK modulation mode at the frequencies of 10MHz and 14MHz; above all, the first FDTD (Finite Difference Time Domain) simulation model relevant to the intra-body communication was proposed. It revealed the effect of the transmitting electrode's position on the received signal. On the basis of Zimmerman's achievement, the distributed RC model of the whole human body was proposed by Korea Advanced Manufacturing Research Institute [10] from the point of view of electric circuit. Compared with experiments, the model is exactly consistent with human experimentation within the range of 100KHZ to 150MHZ.

It can be seen from the above conclusions that, both the prototype experiments and the finite element models focused on the surface of human body, in which the communication was limited to the "surface-to-surface" communication only, but it was not involved with the internal of tissues. Because the transmitter and the receiver must form a ground return circuit in the process of signal transmission, this method is unsuitable for the implantable intra-body communication.

2.2 GALVANIC COUPLING INTRA-BODY COMMUNICATION

Another kind of intra-body communication was proposed by Handa et al. in 1997 [11], he placed an electrocardiogram monitoring device and a signal receiving device on human chest and human wrist, respectively. The signal transmitting electrode of the electrocardiogram (ECG) monitor sent a weak current signal to human body; the signal was transmitted to the receiving electrode of the receiving device on the wrist through human tissues so as to achieve signal transmission. They inputted about 20uA (effective value) alternating current in human body to realize ECG signal transmission in the body. In the whole process, the transmitting electrode and the receiving electrode must contact with human body which was regarded as a resistor to accomplish resistance coupling. This is termed the galvanic coupling intra-body communication.

According to the installation depth of the transmitter and the receiver in human body, the galvanic coupling intra-body communication may be classified into the following types: "surface-to-surface" communication, "surface-to-implant" communication, "implant-to-surface" communication and "implant-to-implant" communication, which will be analyzed below, respectively.

2.3 SURFACE-TO-SURFACE COMMUNICATION

During installing of communication devices, the signal transmitter and the signal receiver are placed on human skin, this may make devices mounted rapidly and easily and the surgery is not required to complete device installation and achieve communication, therefore it is noninvasive.

Modelling and verification: since Hand et al. [11] proposed the galvanic coupling intra-body communication, Lindsdey et al. [12] from University of California took human body as the transmission medium of current signal, and used a novel biomedical telemetry method to measure the tension of the fore cruciate ligament after surgery. After comparing effects of different carrier frequencies and currents on the signal attenuation, he finally proposed the best communication scheme, in which he used the 3mA current and the carrier wave of 37 kHz.

On this basis, Hachisuka etc. from University of Tokyo presented the waveguide-mode intra-body communication, in which human body is regarded as a waveguide pipe to transmit the high-frequency electromagnetic waves generated at the transmitting terminal. Meanwhile, he built a simplified two-electrode intra-body communication model based on References [13], and presented the equivalent circuit models of the two-electrode model and the four-electrode model. On the basis of the model created by University of Tokyo, Song [14] of Beijing Institute of Technology simplified the four-electrode circuit model and provided the transfer function of the simplified model, finally derived the transfer function of the equivalent circuit through the input-output relationship. Note that a modified value K was firstly given in his paper to correct the difference between the individual simulated result and the measured result.

Oberle [15] of Eidgenoessische Technische Hochschule Zuerich (ETH) designed a set of device capable of producing mA-level alternating current (through coupling) by means of dielectric properties of human tissue. Furthermore, he proposed a simpler engineering channel model as well. According to electromagnetic features of human tissue, Wegmuller et al. [16] subsequently built a preliminary finite element model and a layered tissue model of human forearm, and quantitatively compared effects of electrodes' size and position on signal attenuation. Note that Wegmuller used the carrier wave of 10 kHz to 1MHz as well as 1 mA orthogonal current in his experiment, achieving the maximum transmission rate of 4.8kbit/s. On the basis of the layered tissue model proposed by Wegmuller, Gao and Pun [17] established the galvanic coupling intra-body communication channel model by means of the mathematical modelling for the first time.

They took human arm as a standard cylinder in which all layers of tissue were isotropic, and derived the voltage control equation of the galvanic coupling intra-body communication by means of the Maxwell equation. According to the fact that the biological tissue meets the quasi-static conditions when the frequency is less than 1MHz, they derived the voltage equation.

Although the body-surface communication does not result in injuries to human body, signal distortion and great signal attenuation often arise because the sensors are relatively farther from the signal source. Meanwhile, the environment has a great impact on human body surface because movement is inevitable for a person, so relatively strong noise will be produced. For this reason, an appropriate detection method is required to make the sensors close to the signal source and less affected by the environment in order to reduce signal interference.

2.2 IMPLANT-TO-SURFACE COMMUNICATION

In the implant-surface communication system, the current signal sent from the implantable device is used to implement current coupling with the surface-mounted device through human tissue, so as to achieve the implant-surface communication. In the system, relevant devices can be properly placed at desired positions and be repositioned. Because the signal-source detecting device is installed in human body, the detected source signal is less affected by the environment.

Design and verification of the system: Lindsey [20] implanted a sensor at the intersection of the fore cruciate ligament of a dead body's leg through the surgical operation, so as to measure ligamentous tension. The measured tension was converted into the current signal, which was transmitted to the detecting electrode mounted on human body surface via human tissues to achieve signal transmission. In the system, two electrodes with the diameter of 0.38mm were implanted in the dead body' leg at an interval of 2.5mm, and the electromyography (EMG) detecting electrode was placed on the surface of the leg. A $1 - 3mA$ sinusoidal signal with the carrier wave of 2-160KHZ was sent for test, producing the minimum signal attenuation of 37db. The signal attenuation here is easy to be affected, which is closely related to the current value, the distance between the built-in device and the surface-mounted electrodes and the interval between electrodes. Although the EMG detecting electrode may be repositioned at random, it is difficult to maintain high SNR in the communication due to the large signal attenuation (37-50dB).

A more efficient implant-to-surface communication system was proposed by SUN [21], in which the implanted transmitter was an integrated X-type antenna. They integrated two electrodes of the transmitter (or receiver) on a parabola-shaped surface, and made the X-type antenna on the insulating part generate current within a large range near the antenna by changing the present current, so that the detecting electrode can detect more electric current. They conducted experiments with the saline water and the sedated pig respectively. In the experiments with the saline water, transmission of the X-type antenna signal needed only 1% of the power for the conventional electrode couple. Because the diameter of the X-type antenna was only 9mm, when the transmitter was implanted between the brain surface and the brain diaphragm, the system did not cause great injury to the inside of the brain.

Although the signal source is near and random reposition of the detecting electrodes is available in this kind of communication, the signal must be transmitted through the skin and even more tissues before it is detected by the signal detection electrodes in the process of communication, because the emitting electrodes are installed in the tissues. This will lead to great signal attenuation and reduce the electric conductivity; meanwhile, because the emitting electrodes are installed in the tissues and the signal is relatively steady, it is adverse to transmission of the single-sensor multiple-frequency signal.

2.3 SURFACE-TO-IMPLANT COMMUNICATION

The signal transmitter of the communication system is placed on the surface of the conductor, and the signal receiver is implanted inside a certain tissue. The signal transmitter sends a current signal, and the signal reaches the receiving electrode of the implantable device through the skin and other tissues so as to achieve communication, this communication mode is often used for electric energy transfer in the implantable device.

System design and rationality analysis: Tang [22] proposed to transfer external electric energy to the implantable electronic components through the skin and tissues by virtue of biological tissue's volume conduction characteristics, so as to provide power supply for implantable devices. In order to analyze characteristics of the volume conduction energy delivery system effectively, the X-type equivalent circuit model (with explicit physical significance) of the electrode - skin unit was built, and the computational formula concerning the X-type equivalent-circuit parameters and the open-circuit impedance of electrodes was derived. Meanwhile, through circuit analysis, the input loop impedance ratio (ILIR) was obtained as well as the impedance ratio and the voltage ratio (V1/V2) of the output loop. These parameters were used to determine the efficiency of volume conduction energy delivery. At the operating frequency of 5 kHz, he selected a piece of fresh pig skin to transmit the charging current of 2.8mA to the battery, and the transmission efficiency of the corresponding charging current reached 27%, together with energy transmission efficiency of 11%. It is thus proven that the external electric energy can be efficiently delivered to the implantable electronic devices in the living body through the skin by virtue of biological tissue's volume conduction characteristics.

Although random frequency modulation and electrode layout can be externally realized in this communication mode, there exists signal transmission through the skin, which results in great signal attenuation and poor efficiency.

2.4. IMPLANT-TO-IMPLANT COMMUNICATION

In the implant-to-implant communication system, the device implanted in a tissue sends a signal, which is transmitted to the receiving device in another tissue through in-between tissues to achieve signal transmission. The receiving device can send the signal out by means of the connecting line or radio frequency (RF). This type of communication has less power consumption than the implant-to-surface communication and the surface-to-implant communication, meanwhile, the external environment has less impact on sensors, with less signal attenuation and higher effective power.

Design and verification of the system: design of the implant-to-implant communication system was proposed by M.S.Wegmuller [23]. They simulated human muscular tissue and made two current coupling electrodes in the tissue send an alternating current signal into human body, which were detected by two detection electrodes implanted in the tissue. In order to avoid the common nerve frequency, they selected the current signal below 1uA within the frequency range of 100-500KHZ. To compare effects of different electrodes on signal attenuation, they also designed two kinds of receiving & transmitting electrodes, exposed copper cylindrical electrodes (10mm long, diameter of 4mm) and exposed copper coil (diameter of 4mm). Distance between these transmitting electrodes and receiving electrodes was 50mm. When the cylindrical electrode was chosen, the signal attenuation within 50mm was about 32dB, while the signal attenuation was approximately 47dB when the coil electrode was used. Therefore, it can be seen that different electrodes may produce different effects on signal attenuation and plenty of signal loss is caused by the four-electrode design, because most signals are directly transmitted back to the transmitter and the receiver receives few.

Al-Ashmouny [24] designed a dual-electrode system which was used to experiment with a stupefacient mouse so that signal receiving ability of the receiver was improved. In the system, only two electrodes (included in the transmitter and the receiver respectively) contacted with its tissues. The two electrodes were composed of the platinum - iridium wire with a diameter of 50um and an insulating CMOS chip, and their volume were around 1mm3. They were implanted in the brain of the mouse together with the AC signal electrode. Because the receiving circuit of the transmitter unit is insulating, the communication mode has a larger impedance ratio than the four-electrode circuit, and the signal transmission rate is higher accordingly. Although the system has a high signal transmission rate, it is easy to be affected by the external current receiver. If a conductor with low impedance is used to form a closed circuit to the ground, e.g. human body contacts with the ground directly, signal loss arises easily.

3 Opportunities and challenges

3.1 OPPORTUNITIES

1) In the intra-body communication, human body is considered a conductor capable of transmitting and receiving signals. Therefore, complicated wiring may be avoided.

2) Strong anti-interference capacity. In the galvanic coupling intra-body communication, the current sent from the transmitting electrode flows through human body to achieve current coupling with the receiving electrode. Therefore, the anti-interference capacity of current coupling is stronger and more stable than capacitive coupling, which is beneficial to realization of the high-speed communication. Furthermore, the galvanic coupling electrode directly contacts with human body to form a closed circuit; however, in the capacitive coupling communication, the closed circuit is formed through the

ground coupling. As a result, the capacitive coupling is not suited for communication between implantable medical devices at all.

3) Low carrier frequency, voltage and current, safe communication. Most carrier frequencies for the galvanic coupling intra-body communication are less than 1MHz; the current is several milliampere and the voltage is less than 5V [17, 19]. Compared with the capacitive coupling intra-body communication, the frequency is relatively lower and most of signals are concentrated on tissues, so it is necessary to contact with the living body in order to gather signals, and signal transmission is very safe. Likewise, injuries to the living body are relatively slighter due to low frequency and small voltage and current; and electromagnetic burns will not happen in a short period.

3.2 CHALLENGES

3.2.1 Energy supply

The grand challenge in the medical implant communication is how to provide sufficient energy for devices to gather and transmit signals. Only two methods are adopted to solve the problem, one is to design the biological electric circuit with low power consumption; the other is external energy supply. With the development of the micro electronic technology, the circuit with low power consumption has been well applied to the implantable sensor technology [25]. The commonest energy supplying method for these devices is to use battery, for example, the cardiac pacemaker and deep brain stimulation devices are supplied with power through battery. However, it is difficult to achieve miniaturization in the battery design in order to meet requirements for energy supply. Meanwhile, because the service life of the battery is limited, it must be changed through surgery once the battery is used up, which will injure human tissues greatly. In order to reduce injury to human body, researchers also proposed a method called "external energy supply".

"External energy supply" is to transfer energy to the devices implanted in human body through wireless techniques. The radio-frequency (RF) technique is one important "external energy supply" method. At present, its development remains in the early phase without practical application. The wireless power transmission allows high-efficiency energy transfer through magnetic resonance coupling, but the larger coil is required to accomplish coupling [26], and more space is required for implantation, so it is very difficult to implant devices.

The ultrasonic energy transfer can also be used for implanted devices, but the transmission efficiency is very low (only 0.06% or so) [27], so it is hard to be developed for practical application. A survey on energy harvesting [28] and the optical energy [29] has been conducted, but at present the energy supplied via these means is very little, so it is difficult to meet steady operating requirements for implantable devices. The implantable radio frequency

identification (RFID) method is also used to provide energy. In the overall system, the implantable device is just like an RFID device or an external source produced by conditioning signals. Coupling between the external signal and implantable devices is implemented through human brain tissues. The signal current is generated through the RF signal and the traditional RFID device [30] or by means of volume conduction. Because the battery is not required here, this method enables microminiaturization to the greatest extent. However, the inductance is used in the present research, so it is difficult to realize microminiaturization. The galvanic coupling is also used for power transmission, and Tang has made quite a few achievements [31] in this aspect. Although major improvement has been made compared with previous similar researches, the efficiency is still low on the whole and it is difficult to meet requirements for the working environment of sensors.

3.2.2 Implantation of devices

The way that a device is implanted is closely related to its size. If the implanted device has a large volume, it is usually implanted by means of surgery [16]. If the device is small, it may be implanted through injection. The needle is firstly inserted to the subcutaneous tissue, and then the sensor is pressed into the subcutaneous tissue through the normal saline or air. This technique is often employed to identify dead animals by means of RF identification [31]. For the device implanted in the brain, when the hard needle is inserted in the cerebral cortex or the brain with meninx protection, large force is required to pierce the brain. However, brain tissue has a larger volume than encapsulated tissues, and the pressure applied out of the injector may injure brain tissues [16]. Vacuum tools are used to replace subcutaneous injection, which is similar to the vacuum sensor. After the implantable device is placed in a vacuum tube, once the vacuum tube is inserted until the desired depth, the device is released and the vacuum tube is withdrawn.

The other implantation method is to embed the implantable device by means of magnetic navigation. The magnetic conduit [32] in the brain and the nano-drug delivery technique [33] are firstly developed. In the magnetic navigation, a permanent magnet is embedded in the implantable device, which is made to enter human body through injection or swallowing and is positioned by means of external magnetic guidance. The equipment is capable of 3D fine positioning and reposition. Its shortcoming is to require the implantable device with strong magnetic sensitivity and permanent magnetism. Meanwhile, environmental magnetic force and nuclear magnetic resonance are likely to result in shifting of the device.

In recent years, the fusible thin film has been used to fix the electrode network on the surface of the brain [34], and it is also applied to the implantable miniature device. The fusible thin film will be gradually separated from the

device and fibre as time goes by, and decomposed within several days or weeks. But its safety and stability is required to be investigated and tested.

3.2.3 Communication security

In the actual communication, a man equipped with the implantable device shall not be injured or injury to him shall be minimized. From overseas research conclusions concerning medical implant communication, we learn that, when an implantable device is fixed in human body for a short or long period, complicated reaction will occur on the surface of human body [35]. The reaction may have an effect on functions of the implantable device. In order to reduce damage to human tissues, health organizations are seeking various methods to achieve safe application of implantable devices, including selection of compatible materials and coating [36] as well as delivery of local medicine [37].

It is very important to reduce impacts of the intra-body communication upon human body, including local temperature rise caused by power loss as well as local stimulation. In order to avoid local temperature rise, we choose the intra-body communication. Compared with RF identification, the low-frequency carrier wave is often used in the galvanic coupling intra-body communication, and only the carrier wave less than 1MHz is chosen under ideal conditions [17, 19]. Meanwhile, the carrier frequency should be higher than the frequency for physiotherapy in order to minimize stimulation to human tissues, or at least 100KHZ is chosen as the carrier frequency. Within the range of 100KHZ to 1MHZ, biological tissues have good electrical conductivity [38]. Even if these intermediate frequencies are used, it is necessary to notice specific absorption rate (SAR) of organic energy and select the electric current density below the international standard [38], because the intra-body communication is an emerging communication technology, potential temperature rise of tissues and stimulation to tissues are yet indeterminate. Even if the communication mode has been accepted by the relevant international standard, they must be closely monitored in future experiments.

3.2.4 Physical model of channel

The circuit model is one of the commonest approaches to build the intra-body communication channel model. In the model, electrodes and human tissues are considered to be interconnected ideal circuit elements (resistors and capacitors) which are connected through the ideal wire. The modelling method is quite simple, but it can show effectively the electrical resistance properties of signals and tissues.

The finite element model is a method to solve approximate solutions in modelling the communication channel [40]. In the model, human body is considered to be composed of numerous small interconnected sub-domains called the finite element. Each element is

supposed to have a simple and appropriate approximate solution, and then the general boundary condition (e.g. an equilibrium condition of a structure) of the domain is derived, so as to obtain the solution to the problem. The solution is not the exact solution but the approximate solution, because the practical problem is replaced by the simpler.

Whether the circuit model or the finite element model is used to simplify modelling, both of them cannot present characteristics of the intra-body communication channel well, failing to focus on the general applicability of distribution and transmission mode of the electric signals inputted in human body. Moreover, they cannot be used to interpret essentially internal mechanisms and external factors affecting the quality of the intra-body communication. For this reason, a more accurate modelling method is required to describe these mechanisms.

Pun, Gao et al. [17, 19] abstracted human arm as a homogeneous isotropic multi-layer cylinder; moreover, they derived the control equation of the intra-body communication according to Maxwell's Equation, and established the electromagnetic model of the galvanic coupling intra-body communication (surface communication) in combination with relevant boundary conditions and linking conditions. The classical electromagnetic theory was used to analyse relevant parameters in the model, which demonstrated distribution and transmission of the electric signal when the weak AC was inputted. And then internal and external factors affecting channel characteristics were analysed and causal relation between phenomena and mechanisms was interpreted. However, it was the surface communication channel model built under the ideal condition and in the quasi-static mode [19], leaving out anisotropic property of tissues [41]. In the actual human tissue, the parallel and the transverse characteristics of the tissue are not identical [41], even the parallel and the transverse electrical characteristics of some tissues are greatly different $2.04 < \frac{\delta_L}{\delta_T} < 15.3$ [41]. In most of existing models, only a

certain part of human body is considered. If the whole body is taken into account, how should the geometric model be connected? With the continuous increase of the frequency, the quasi-static condition will be completely true no longer. When the implantable device transmits signals from inside to outside, the existing channel model will not be able to interpret transmission characteristics of signals completely; even the position where the device is mounted will have a direct impact on transmission characteristics of signals. Under various complicated conditions, the intra-body communication channel model will be greatly changed. In order to show signals' transmission characteristics in human body better, these problems will be gradually considered in future channel models.

3.2.5 Communication speed

The galvanic coupling intra-body communication has stable and strong anti-interference performance and is suitable for the implantable communication; instead, its communication speed is very low because of restriction on the communication frequency. The existing research achievements show that, its maximum communication speed is only 9.6Kbps/s. Physiological signals of human body at the speed can be transmitted basically in a real-time way. However, it is very difficult if it is used to transmit consecutive big-data signals such as video and audio signals. The research of the galvanic coupling intra-body communication will focus on how to improve the communication speed within the limited bandwidth.

4 Conclusion

The intra-body communication is an emerging wireless communication technology, in which human tissues are regarded as the communication wire and human body as the communication channel, so as to achieve communication through coupling of electrodes. This type of communication plays a significant role in real-time medical monitoring of human body because of slight wound, no infection and easy positioning of devices.

The galvanic coupling intra-body communication plays an important role in the medical implant communication because of no ground coupling, less effects produced by the external environment and low carrier frequency. But this technology is not proven at present and cannot be applied to practical medical monitoring; especially the mathematical model of the medical implant communication has not been established, and characteristics of the intra-body communication channel cannot be shown, so the technology will be applied practically after a certain period of time.

In this paper, firstly, research status of the capacitive coupling intra-body communication is analysed in chronological order. It is found from the above analysis that the communication mode cannot be applied to the medical implant intra-body communication. Fortunately, the galvanic coupling intra-body communication may make up for the defect. Next, research status of the "surface-to-surface" communication, the "surface-to-implant" communication, the "implant-to-surface" communication and the "implant-to-implant" communication is analysed as per positions where electrodes are mounted. Finally, opportunities and challenges of the technology are presented. In addition, solutions to some problems mentioned in this paper will be shown in a separate paper.

References

[1] Natarajan A, Motani M, de Silva B, Yap K K, Chua K C 2007 Investigating network architectures for body sensor networks *Proceedings of the 1st ACM SIGMOBILE International Workshop on System s and Networking Support for Healthcare and Assisted Living Environments (HealthNet'07)* New York USA ACM 2007 19-24

[2] Yun D, Kang J, Kim J E, Kim D 2007 A body sensor network plat form with two level communications *Proceedings of IEEE International Symposium on Consumer Electronics ICSE* 2007 1-6

[3] Deng Q K 2006 A Novel Model of the Medical Instrumentation – An Overview of Wearable Sensors and Systems" *Chinese Journal of Medical Instrumentation* 30(5) 327-9

[4] Yang G Z 2006 Body Sensor Networks *Springer*

[5] Zimmerman T G 1995 Personal area networks (PAN): near-field intrabody communication *Master thesis Massachusetts Institute of Technology* MA USA

[6] Gray M Physical Limits of Intrabody Signalling *Master thesis Massachusetts Institute of Technology* 1997 1-49

[7] Partridge K, Dahlquist B, Veiseh A, Cain A, Foreman A, Goldberg J, Borrielo G 2001 Empirical Measurements of Intra-body Communication Performance under Varied Physical Configurations Proceedings of the 14th annual ACM symposium on User interface software and technology table of contents 183-90

[8] Shinagawa M, Fukumoto M, Ochiai K, Hakaru K 2004 *IEEE Transactions on Instrumentation and Measurement* 53(6) 1533-8

[9] Fujii K, Takahashi M, Ito K 2007 Electric Field Distributions of Wearable Devices Using the Human Body as a Transmission Channel *IEEE Transactions on Antennas and Propagation* 55(7) 2080-7

[10] Cho N, Yoo J, Song S J 2007 *IEEE Transactions on Microwave Theory and Techniques* 55(5) 1080-6

[11] Handa T, Shoji S, Ike S, Takeda S, Sekiguchi T 1997 A very low-power consumption wireless ECG monitoring system using body as a signal transmission medium *International Solid State Sensor Actuator Conference* 1003-6

[12] Lindsey D P, McKee E L, Hull M L, Howell S M 1998 *IEEE Transactions on Biomedical Engineering* 45(5) 614-9

[13] Hachisuka K, Terauchi Y, Kishi Y, Sasaki K, Hirota T, Hosaka H, Fujii K, Takashi M, Ito K 2006 Simplified circuit modeling and fabrication of intra-body communication devices *Sensors and Actuators* 322-30

[14] Song Y, Hao Q, Zhang K, Wang M, Chu Y, Kang B 2011 *IEEE Transactions on Instrumentation And Measurement* 60(4) 1257-66

[15] Oberle M 2002 Low power system-on-chip for biomedical application PhD Thesis ETH No 14509 IIS/ETH Zurich

[16] Wegmuller M S 2007 Intra-Body Communication (IBC) for Biomedical Sensor Networks *PhD Thesis* Switzerland ETH

[17] Pun S H, Gao Y M, Mak P U, Vai M I, Du M 2011 *IEEE Transactions on Information Technology in Biomedicine* 15(6) 870-6

[18] Plonsey R, Heppner E B 1967 Considerations of quasi-stationarity in electrophysiological systems *Bulletin of mathematical biophysics* 29

657-64

[19] Chen X M, Mak P U, Pun S H, Gao Y M, Lam C T, Vai M I, Du M 2012 Study of Channel Characteristics for Galvanic-Type Intra-Body Communication Based on a Transfer Function from a Quasi-Static Field Model *Sensors* **12** 16433-50

[20] Sun M, Mickle M, Liang W, Liu Q, Sclabassi R J 2003 *IEEE Transactions on Neural System Rehabilitation Engineering* **11**(2) 189-92

[21] Tang Z D A 2007 *Study on energy delivery rationale and method for implantable devices through volume conduction* PhD Thesis China Chongqing University

[22] Wegmueller MS, Huclova S, Froehlich J, Oberle M, Fleber N, Kuster N, Fichtner W 2009 *IEEE Transactions on Instruments and Measurement* **58**(8) 2618-25

[23] Al-Ashmouny K M, Boldt C, Ferguson J E, Erdman A G, Redish A D, Yoon E 2009 IBCOM (intra-brain communication) microsystem: wireless transmission of neural signals within the brain *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society* 2054-7

[24] Sarpeshkar R 2010 Ultra Low Power Bioelectronics: Fundamentals, Biomedical Applications, and Bio-Inspired Systems *Cambridge University Press* Cambridge, UK

[25] Kurs A, Karalis A, Moffatt R, Joannopoulos J D, Fisher P, Soljacic M 2007 Wireless power transfer via strongly coupled magnetic resonances Science **317**(5834) 83-6

[26] Lee K L, Lau C P, Tse H F, Echt D S, Heaven D, Smith W, Hood M 2007 First human demonstration of cardiac stimulation with transcutaneous ultrasound energy delivery: implications for wireless pacing with implantable devices *Journal of American College of Cardiology* **50**(9) 877-83

[27] Justin G A, Zhang Y, Sun M, Sclabassi R 2004 Biofuel cells: a possible power source for implantable electronic devices In: *Proceedings of the 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* **2** 4096-9

[28] Murakawa K, Kobayashi M, Nakamura O, Kawata S 1999 *IEEE Engineering in Medicine and Biology Magazine* **18**(6) 70-2

[29] Riistama J, Aittokallio E, Verho J, Lekkala J 2009 Totally passive wireless biopotential measurement sensor by utilizing inductively coupled resonance circuits Sensors and Actuators A Phys. **157** 313-21

[30] Troyk P R 1999 Injectable electronic identification, monitoring, and stimulation systems Annual Review Biomedical Engineering **1** 177-209

[31] Grady M S, Howard M A, Dacey R G, Blume W, Lawson M, Werp P, Ritter R 2000 Experimental study of the magnetic stereotaxis system for catheter manipulation within the brain *Journal of Neurosurgery* **93**(2) 282-8

[32] Neuberger T, Schöpf B, Hofmann H, Hofmann M, Von Rechenberg B 2005 Superparamagnetic nanoparticles for biomedical applications: Possibilities and limitations of a new drug delivery system *Journal of Magnetism and Magnetic Materials* **293** 483-96

[33] Kim D H, Viventi J, Amsden J J, Xiao J, Vigeland L, Kim Y, Blanco J A, Panilaitis B, Frechette E S, Contreras D, Kaplan D L, Omenetto F G, Huang Y, Hwang K, Zakin M R, Litt B, Rogers J A 2010 Dissolvable films of silk fibroin for ultrathin conformal bio-integrated electronics *Nature Materials* **9**(6) 511-7

[34] Turner J N, Shain W, Szarowski D H, Andersen M, Martins S, Isaacson M, Craighead H 1999 Cerebral astrocyte response to micromachined silicon implants *Experimental Neurology* **156**(1) 33-49

[35] Williams D F 2008 On the mechanisms of biocompatibility *Biomaterials* **29**(20) 2941-53

[36] Onuki Y, Bhardwaj U, Papadimitrakopoulos F, Burgess D J 2008 A review of the biocompatibility of implantable devices: current challenges to overcome foreign body response *Journal of Diabetes Science and Technoloogy* **2**(6) 1003-15

[37] Gabriel C, Gabriel S, Corthout E 1996 The dielectric properties of biological tissues: I. Literature survey Physics in Medicine and Biology **41**(11) 2231-49

[38] International Commission on Non-Ionizing Radiation Protection (ICNIRP) 1998 Guidelines for limiting exposure to time-varying electric, magnetic, and electromagnetic fields (up to 300 GHz) *Health Physics* **74**(4) 494-522

[39] Song Y, Zhang K, Hao Q, Hu L, Wang J, Shang F 2012 A Finite-Element Simulation of Galvanic Coupling Intra-Body Communication Based on the Whole Human Body *Sensors* **12** 13567-582

[40] Gielen F L H, de Jonge W W, Boon K L 1984 Electrical conductivity of skeletal muscle tissue: experimental results from different muscles in vivo *Medical & Biological Engineering and Computing* **22** 569-77

Authors

Shuang Zhang, born in May, 1983, Leshan, China

Current position, grades: lecturer in The Engineering & Technical college of Chengdu University of Technology. PhD student in the Department of Electrical and Electronics Engineering, Faculty of Science and Technology, University of Macau.
University studies: Doctor of Electrical & Electronic Engineering at University of Macau.
Scientific interests: intra-body communication, cryptography.
Publications: 2 patents, 25 papers

Yu Ping Qin, born in March, 1984, Leshan, China

Current position, grades: Lecturer in The Engineering & Technical college of Chengdu University of Technology.
University studies: Master of Basic mathematics at Sichuan Normal University, 2011, China.
Scientific interests: intra-body communication, cryptography.
Publications: 2 patents, 25 papers.

Xiao Jing, born in August, 1976, Xuzhou, China

Current position, grades: Professor in Air force Logistic College
University studies: Master's degree at Chengdu University of Technology in 2007.
Scientific interests: intra-body communication, cryptography, remote sensing and environmental engineering
Publications: 5 papers
Experience: tutor in the Air force Logistic College Since 2005.

Yihe Liu, born in April, 1964, Neijiang, China

Current position, grades: Professor in Neijiang Normal University
University studies: PhD degree in Cryptography in Sichuan University, China, in 2005.
Scientific interests: intra-body communication, cryptography
Publications: 2 patents, 46 papers.

Task-role-based workflow authorization model and its implementation in emergency command system of water traffic

Yuankui Li, Yingjun Zhang∗, Feixiang Zhu, Jiandao Liu

Navigation College, Dalian Maritime University, Linghai Rd. 1, Dalian, China

Received 23 April 2014, www.cmnt.lv

Abstract

Water traffic plays an irreplaceable role in modern traffic as accomplishing heavy transport task in low cost, but the emergency frequency brings serious challenge to ensure safety and rescuing in water. At the same time, the safety supervision departments at all levels lack of modern technical means for prevention and action to emergency, so an emergency command system not only for safety supervision and risk early-warning in normal state, but also for quick response and scientific decision for emergency must be built. In this paper, by the use of the task-role-based workflow authorization model, an emergency command module is realized, which key technology is permission assignment. It accomplishes 7 tasks by 4 roles, besides, by the definition of user, role and task and constrains of relationship among them, and take use of delegation technology and right transmission between roles, emergency command process can executed methodically with high quality and clear target according to emergency plan.

Keywords: water traffic, emergency command, workflow, permission assignment

1 Introduction

Water traffic undertakes heavy transport task, and because of the influence of objective and subjective factors, water emergency occurs frequently, objective factor includes the channel, hydrology, meteorology, aid to navigation, traffic characteristics, hull form, ship(fleet) size, navigation technology; subjective factor includes safety navigation technology, safety management model, quality of employees. Water safety supervision departments at all levels already had related emergency plan and policy, but due to the lack of effective technical support of the emergency command system, some serious problem may often appears in emergency treatment, such as unified scheduling of the manpower and material resources, slow rate of information and data exchange, lack of objective information support for decision-making and commanding, lack of understanding of operators for emergency plan and its operation flow. Therefore, it is urgent to establish a complete set of emergency command system, which not only for safety supervision and risk early-warning in normal state, but also for quick response and scientific decision in emergency.

Emergency command system involves complex business processes, so, workflow technology is introduced and its delegation technology and right transmission between roles is researched, and task-role-based workflow authorization model is constructed for emergency command in this paper according to its actual work logic. The emergency command process is divided into 7 tasks: Record receiving alarm, Department verified, Leader verified, Start order, Disposal action, Final treatment, Task suspension & end, these tasks are dependent and executed in a coordinated manner methodically with high quality and clear target according the emergency plan. This paper is developed from a conference paper of The 5th International Conference on Intelligent Computation Technology and Automation (ICICTA2012), as in [1].

2 Task-role-based workflow authorization model

Workflows typically represent processes involved in manufacturing and office environments, and the definition of the workflows given by WfMC (Workflow Management Coalition) is "The automation of a business process, in whole or part, during which documents, information or tasks are passed from one participant to another for action, according to a set of procedural rules", as in [2, 3].

The task-role-based workflow authorization model associated the role and permission through task, grants different executes right to different users through task state and the roles of users, as in [4]. Besides, a series of constraint mechanism must be used to assign different permission to different users to reduce the mistake and ensure the emergency command execution methodically, as in [5].

1) U (*user*): the set of users, R (*role*): the set of roles, WT (*tasks*): the set of tasks, P (*permission*): the set of permission.

2) $URA \subseteq U \times R$: User Role Assign, represents the many-to-many hibernate from U to R, means the users are fixed to the roles.

∗*Corresponding author* e-mail: ecdis@sina.com

3) $RTA \subseteq R \times WT$: Role Task Assign, represents the many-to-many hibernate from   to *WT*, means the roles have right to do the tasks.

4) $RH \subseteq R \times R$: Role Hierarchy, is partial ordering relation on roles set, represented by "$\succ$". There exists $r_1, r_2 \in R$, if $r_1 \succ r_2$, then it means that superior role r_1 has all rights of lower role r_2, and $WT(r_1) = WT(r_1) \cup WT(r_2)$.

5) (u, wt_i, u'): Delegation relationship, represents that user *u* delegate his task wt_i to u', and u' will take place of *u* to execute task wt_i.

3 Task-role-based workflow authorization model

3.1 TASKS REPRESENT

In emergency command process, it contains 7 tasks, which includes Record receiving alarm, Department verified, Leader verified, Start order, Disposal action, Final treatment, Task suspension & end, as shows in Table 1.

TABLE 1 The task represent table

NO.	Task Name	Task function brief introduction
wt_1	Record received alarm	Record emergency condition after receive alarm
wt_2	Department verified	Related department verify emergency condition
wt_3	Leader verified	Related leader verify emergency condition
wt_4	Start order	Start emergency disposal
wt_5	Disposal action	Dispose the emergency
wt_6	Final treatment	Related task after emergency disposal
wt_7	Suspension & end order	Suspend or end the emergency action

1) Received alarms record: this task is to save emergency alarm record and report them to management department to ask for verifying after alarm received. To record emergency alarm information rapidly, the quantity of users with permission to execute this task is larger than others.

2) Department verified: this task is to verify the newest reported emergency information in alarms record list, it can judge if the emergency is false alarm or not, and ascertain the dangerous situation is excluded or not.

3) Leader verified: this task is to treat verified and reported emergency from department.

4) Start order: this task is to start emergency plan and inquire about the state of the started emergency.

5) Disposal action: this task is to inquire the handling emergency, it only can inquire the emergency in disposal state of "started" in one month in default.

6) Final treatment: this task is to complete final treatment to started or ended emergency.

7) Task suspension & end order: this task is to suspend or end any reported or verified emergency.

3.2 ROLE REPRESENT

In this system, there are 4 roles: Leader, Expert, Commander and Executor as shows in Table 2.

TABLE 2 The role represent table

NO.	Role name	Role function brief introduction
r_1	Leader	The person in over-all command
r_2	Expert	Provide decision opinion and take participate in making disposal scheme and action
r_3	Commander	Organize emergency disposal action
r_4	Executor	Carry out emergency disposal action and provide related information to leaders

4 Workflow Model for Emergency Command

4.1 MODEL FOR EMERGENCY COMMAND WORKFLOW

Refer to the symbol expression in references [6-8] about workflow, in this paper, workflow model for emergency command is represented.

Given: $WT = \{wt_1, wt_2, wt_3, wt_4, wt_5, wt_6, wt_7\}$, they are expressions for all of the tasks of Record received alarm, Department verified, Leader verified, Start order, Disposal action, Final treatment, Task suspension & end order; $U = \{u_1, u_2, u_3, u_4, u_5, u_6, u_7\}$, they are expressions for all of the users of A, B, C, D E, F, G; $R = \{r_1, r_2, r_3, r_4\}$, they are expressions for all of the roles of Leader, Expert, Commander, Executor, *A* is Leader, *B* is Expert, *C* and *D* is Commander, *E*, *F*, *G* is Executor. So, the User Role Assign is:
$URA = \{(u_1, r_1), (u_2, r_2), (u_3, r_3), (u_4, r_3), (u_5, r_4), (u_6, r_4), (u_7, r_4)\}$, and the Role Task Assign is: $RTA = \{(r_1, wt_3), (r_2, wt_5), (r_3, wt_2), (r_3, wt_4), (r_4, wt_1), (r_4, wt_6), (r_4, wt_7)\}$. There also have 3 constraints:

1) The user for task of Start order must be same as the one for Department verified, listed as $C_1 : must\_do(u, wt_4) \leftarrow (u, wt_2) \in C_{wt_4}$, C_{wt_i} is the set of constraints related to wt_i.

2) The user for task of Task suspension & end order must be same as the one for Final treatment, listed as: $C_2 : must\_do(u, wt_7) \leftarrow (u, wt_6) \in C_{wt_7}$.

The user for task of Final treatment can't be same as the one for Record received alarms, listed as: $C_3 : cannot\_do(u, wt_6) \leftarrow (u, wt_1) \in C_{wt_6}$.

4.2 AN EXAMPLE FOR EMERGENCY COMMAND WORKFLOW

The flow graph for emergency command is shown in Figure 1, there are 4 user-sets in 4 roles from left to right,

and the emergency command is executed according to the arrow direction. At first, Record received alarm is executed by u_5 with r_4, this is step 1, then it will be executed according to the arrow direction. The users have fingers in emergency command is $u_5 \rightarrow u_3 \rightarrow u_1 \rightarrow u_3 \rightarrow u_2 \rightarrow u_6 \rightarrow u_6$, they finished 7 tasks methodically. Following is the detail calculative process:

1) Received alarms record:

$$EU\left(wt_1^1\right) = user\left(role\left(J_{WT}\left(wt_1^1\right)\right)\right) = user\left(role\left(wt_1\right)\right) = user\left(r_4\right) = \{u_5, u_6, u_7\},$$

where wt_1 is the task of Record received alarm. wt_1^1 is a task instance of wt_1, EU is users set which can executed wt_1^1. J_{WT} is a task mapping from one task instance to related task, which make each wt_i^k have the equal of $J_{WT}(wt_i^k) = wt_i$, $role\left(wt_i\right) = \left\{r \in R \middle| (r, wt_i) \in RTA\right\}$ is a role mapping function, it maps each wt_i to r_i; $user\left(r_i\right) = \left\{u \in U \middle| (u, r_i) \in URA\right\}$ is a user mapping function, it maps each r_i to u_i.

The task of Record received alarm can be executed by u_5, u_6 or u_7, in this example, it is executed by u_5;

2) Department verified: $EU\left(wt_2^1\right) = \{u_3, u_4\}$, in this example, it is executed by u_3;

3) Leader verified: $EU\left(wt_3^1\right) = \{u_1\}$, there is only one leader in this example, so it must be executed by u_1;

4) Start order: due to the constraint C_1, $EU\left(wt_4^1\right) = U\left(u, wt_2^1\right) = \{u_3\}$, so, in this example, it must be executed by u_3, U is the user who have executed wt_2^1;

5) Disposal action: $EU\left(wt_5^1\right) = \{u_2\}$, there is only one expert in this example, so it must be executed by u_2;

6) Final treatment: $C_3 : cannot\_do\left(u, wt_6\right) \leftarrow \left(u, wt_1\right) \in C_{wt_6}$, this constraint means the user who has done wt_1 is forbidden to executed wt_6, so on the basis of this constraint:

$$EU\left(wt_6^1\right) = user\left(role\left(J_{WT}\left(wt_1^1\right)\right)\right) - U\left(u, wt_1^1\right) =$$ in this $\{u_5, u_6, u_7\} - \{u_5\} = \{u_6, u_7\},$

example, it must be executed by u_6 or u_7, here let u_6 execute it;

7) Task suspension & end order: as C_2, $EU\left(wt_7^1\right) = U\left(u, wt_6^1\right) = \{u_6\}$, in this example, it must be executed by u_6.
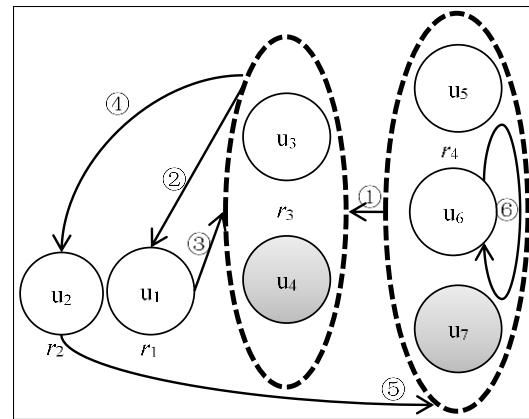
FIGURE 1 The flow graph for emergency command

4.3 DELEGATION IN EMERGENCY COMMAND WORK FLOW

In the above model, any actual task wt_i^k must be accomplished by one certain user u_j with correspond right. But when u_j is out or the tasks are too many for him to accomplish rapidly, wt_i^k will become the bottleneck of the whole wt^k, and because emergency flow is schedule dependent, the whole wt^k cannot be executed due to wt_i^k blocking, which will decrease the emergency command efficiency and acts inimical to safety of lives and property in water traffic. So, the delegation technology must be used in emergency command workflow model.

Theoretically, the delegation is not a model, it is just a very normal problem that needs to be resolved in actual application. If one task wt_i is assigned to user u, but he is too busy to accomplish, by the use of delegation technology, he can delegate wt_i to user u' to ensure the whole workflow can be done rapidly.

Delegation $\left(u, wt_i, u'\right)$ cannot be done successfully unless it can meet the following conditions, as in [9-12]:

1) The agent is not principal own: $u' \neq u$;

2) wt_i has not been delegated yet: $\left(u, wt_i, u'\right) \neq \varnothing$;

The agent and principal must all have operation right to wt_i: $role(wt_i) \in role(u)$ and $role(wt_i) \in role(u')$.

4.4 DELEGATION IN EMERGENCY COMMAND WORK FLOW

The workflow bottleneck is solved only to a certain extent through delegation technology, but because of some matters, the user u neither accomplish wt_i rapidly, nor delegate it to others (called No Answer), which will cause wt_i at a standstill, and it is very harmful to emergency command, and this can be solved by right transmission between users with different roles. According to Role Hierarchy relationship ($r_1 \succ r_2 \succ r_3 \succ r_4$), the system can

automatically send wt_i ($role(wt_i) = r_j$) to $u'' \in user(r_{j-1})$ with higher level role, as in [13, 14]. The No Answer time t is fixed by the system administrator according to need of the actual condition.

As a general rule, there are enough users with r_4 that busy or "No Answer" condition may not come forth; and every emergency must be start the disposal order by leader with r_1, so at least one user with r_1 must be stand by to ensure emergency workflow be accomplished well.

Right transmission mainly be used between r_2 and r_3, when wt_2^k should be executed by one user with r_3, but it is not executed in t, the system will send it to one user with r_2, it will be executed by user with r_2, or when it is not executed in t either, wt_2^k will be executed by user with r_1 to ensure the emergency command be executed rapidly.

For the emergency command workflow in this special condition, the constraints of C_1, C_2 and C_3 can be not followed strictly, because in this condition, all the tasks that users with $r_{i+1}(0 < i < 2)$ should execute are done by user with r_i or even higher level role, which is very useful to ensure the emergency command be executed rapidly.

5 Task-role-based workflow authorization model

To achieve work process management, this system used the method of permission control and process state control to achieve distributed management for each node emergency command. Each module of the system used permission control, the user is unable to operate the module without related distributed permissions, and if the user has got related operation permission, the business trend is specified according to current "disposal status" flag of emergency. And different users with certain operation rights can get different emergency lists, as shown in Table 3.

1) Record received alarm: to record the emergency more rapidly, because there are many users with right to execute this task, all disposal status in one month is inquired without limited in default in this sub-module, at the same time, the received alarm record is saved and reported to management departments to ask for verifying.

TABLE 3 Emergency disposal status table

NO.	Sub-module	Disposal status should be inquired
1	Record received alarm	All status is inquired with no limited
2	Department verified	1.Reported 2.False alarm 3.Excluded
3	Leader verified	1.Reported
4	Start order	1.Reported 2.Started
5	Disposal action	1.Started
6	Final treatment	1.Started 2.Ended
7	Suspension & end order	1.Started 2.Reported 3.Suspended 4.Ended

2) Department verified: this sub-module is used for verifying the newest reported emergency information. In default, the records with status of "Reported", "False alarm" and "Excluded" in one month are inquired, while with others are not. After the alarm is verified by the management department, it can be judged whether the emergency is false alarm or not, and ascertain the dangerous situation is excluded or not.

3) Leader verified: this sub-module is used to inquire and treat verified and reported emergency from department. In default, the records with status of "Reported" in one month are only inquired, to make the leader to treat alarm and give order rapidly.

4) Start order: this sub-module is used to start emergency plan rapidly, and inquire the handling status of the started emergency. In default, the records with status of "Reported" and "Started" in one month are inquired, while other status is not.

5) Disposal action: this sub-module is used to inquire the solving emergency. In default, the records with the status of "Started" in one month are inquired, while with others are not.

6) Final treatment: in this sub-module, the records with status of "Started" and "Ended" in one month are inquired, while with other status is not. Only the emergency with above two mentioned status are final treated in this sub-module.

7) Suspension & end order: in this sub-module, the records with status of "Reported", "Started", "Suspended" and "Ended" in one month are inquired, while with other status are not. Any emergency action can be suspended or ended in this sub-module.

This paper used workflow technology to achieve emergency command subsystem of Three Gorges Reservoir tributaries water transport emergency command system, system interface is shown in Figure 2, in which the alarm record list can be seen, this list contain all records in one month without limited.

FIGURE 2 Emergency command subsystem

6 Conclusion

In this paper, according to the need of emergency command system for water traffic, a task-role-based emergency command workflow model is constructed by the use of workflow technology. Delegation technology and right transmission between roles is researched for properly application in this model, it can ensure smooth completion of emergency command process, improve the efficiency of emergency command, and have important significance in improving water emergency rapid command capability.

References

[1] Zhang Y J, Li Y K, Gu L 2012 Research on Workflow Model for Emergency Disposal of Water Traffic *International Conference on Intelligent Computation Technology and Automation, 12-14 January, Zhangjiajie, Hunan, China* 420-2
[2] Haller A, Marmolowski M, Gaaloul W, Oren E, Sapkota B, Hauswirth M 2009 From Workflow Models to Executable Web Service Interfaces *IEEE International Conference on Web Services July 6-10 2009 Los Angeles CA USA* 131-40
[3] Wang L X 2007 The Analysis of Workflow Reference Model *Microcomputer Application Technology* (4) 30-4 *(in Chinese)*
[4] Sandhu R S, Conyne E J, Feinsteink, H L, Youmank C E 1996 Role-Based Access control Model *IEEE Computer* **29**(2) 38-47
[5] Zhang Q, Zheng H Y, Ding Q L 2011 Task-Role-based Workflow Access Control Model with Multi-constraints *Computer and Modernization* (12) 9-12 *(in Chinese)*
[6] Liu S N 2010 Task-role-based access control model and its implementation *International Conference on Education Technology and Computer, 22-24 June 2010, Shanghai, China* 293-5
[7] Xing G L, Hong F 2005 Workflow Authorization Model Based on RBAC *Mini-Micro system* **26**(3) 76-82 *(in Chinese)*
[8] Oh S, Park S 2003 Task-role-based access control model

Information Systems **28**(6) 533-62
[9] Hsu H J, Wang F J 2011 A delegation framework for task-role based access control in WFMS *Journal of Information Science and Engineering* **27**(3) 1011-28
[10] Wainer J, Kumar A, Barthelmess P 2007 DW-RBAC: A formal security model of delegation and revocation in workflow systems *Information Systems* **32**(3) 365-84
[11] Galoul K, Schaad A, Flegel U, Charoy F 2008 A Secure Task Delegation Model for Workflows *In: International Conference on Emerging Security Information, Systems and Technologies, 25-31 Auguest 2008, Cap Esterel* 10-5
[12] Wei Y H, Shi C J 2011 Delegation Based Workflow Access Control Model *Transactions of Shenyang Ligong University* **28**(3) 27-31 *(in Chinese)*
[13] Zhang S L, Shen Y L 2009 Research and improvement of role inheritance in RBAC model *Application Research of Computers* **26**(6), 2362-4 *(in Chinese)*
[14] Gao D Q, Huang Q, Liu Y L 2011 Role-hierarchy-based task delegation model in workflow *Computer Engineering and Design* **32**(6) 1926-9 *(in Chinese)*

Authors

Yuankui Li, born on February 6, 1987, Gansu, China

Current position, grades: Doctoral candidate in Dalian Maritime University.
University studies: Ms.D in Traffic Information Engineering and Control (2010, Dalian Maritime University).
Scientific interest: intelligent transportation systems, decision theory and method, workflow method, optimization method and algorithm.
Publications: 11 papers.

Yingjun Zhang, born on August 26, 1965, Shandong, China

Current position, grades: Doctor of Technical Sciences, Professor and doctoral supervisor in Dalian Maritime University.
University studies: PhD in Technical Sciences (Admiral Makarov State Maritime Academy, 1995)
Scientific interest:l Large information system development method; decision theory and method; optimization model and algorithm
Publications: 8 patents, 81 papers

Feixiang Zhu, born on July 29, 1978, Hunan, China

Current position, grades: Doctor of Engineering, Associate professor in Dalian Maritime University.
University studies: PhD in Traffic Information Engineering and Control (2008, Dalian Maritime University).
Scientific interest: data mining; decision theory and method; optimization model.
Publications: 15 papers

Jiandao Liu, born on November 4, 1983, Hebei, China

Current position, grades: Master of Engineering, Lecturer in Dalian Maritime University.
University studies: Ms.D in Traffic Information Engineering and Control (2009, Dalian Maritime University).
Scientific interest: data mining; marine communication.
Publications: 5 papers.

Temperature field numerical simulation and experimental study of rapid heat cycle molding in cooling process

Shaofei Jiang*, Zhifei Mao, Jiquan Li, Wei Zheng

Key Laboratory of Special Purpose Equipment and Advanced Processing Technology, Zhejiang University of Technology, Hangzhou 310014, China

Abstract

Rapid heat cycle molding (RHCM) is a new technology aimed at obtaining green and high surface quality of plastic products. In this paper, the finite element model of the mould cavity in cooling process with RHCM was established for the transient heat transfer simulation using ANSYS. Thermal analysis results of temperature field were modified by experimental analysis, the results of which showed a good temperature uniformity and extraordinary efficiency of the cooling rate.

Keywords: rapid heat cycle molding, finite element model, cooling process, temperature field

1 Introduction

Modern injection molding products are developing towards high performance, low-cost and green environmental protection. High stability and performance of moulding process and process equipment are requested due to their use in industry [1-3]. Rapid heat cycle molding technology is the key research direction of injection molding research field at present. This technology is becoming enormously a potential commercial technology, which can get extraordinary surface quality of plastic products without follow-up spraying process and save more material or energy than traditional process method with lower total production cost and a shorter production cycle.

The main research emphases of rapid heat cycle molding technology focus on the research and development of heating methods, which include steam heating, electric heating, high frequency electromagnetic induction heating, infrared radiation heating, flame heating, high temperature gas heating and other heating methods [4-6]. We can improve the fluidity of plastic melt during the filling stage effectively and decrease the quality defect of the plastic to satisfy quality requirement. But production efficiency will be reduced greatly owing to the cooling time increasing. The above disadvantage is not conducive to the application of the actual production. Therefore, research about the effect of heating and cooling stage in rapid heat cycle molding process should be studied to provide a theoretical basis for the development of the technology. Barone and Caulk [7] firstly used the boundary element method for two-dimensional analysis in cooling stage of traditional molding process, and optimized the cooling device settings, size and surface temperature at the same time. Domestic and foreign scholars began to study the three dimensional simulation of steady-state and transient temperature field in traditional molding process with the rapid development of polymer rheology, heat transfer theory, numerical calculation and other related disciplines, as well as a variety of simulation software such as Ansys, Abaqus [8-10]. However related research rarely involves cooling process of rapid heat cycle molding technology. In conclusion the uniformity of the temperature field during the cooling stage and the cooling rate are important influencing factors in the quality as well as the function of the plastic parts. Therefore, the correct simulation and analysis of mold cooling process are conducive to our understanding of the mold temperature field, and help us to adjust the parameters of temperature in the injection molding process effectively. So we can increase the production efficiency and the quality of plastic parts in rapid heat cycle molding technology.

The numerical simulation is an effective way to study the injection molding process [11], and the realization of the mold three-dimensional temperature field simulation provides a scientific basis and analytical tools for the control of the plastic product's quality. It can promote the design and development of the injection mold effectively. In this study, the simulation of temperature field of cooling stage in rapid heat cycle molding with electric heating was proposed helping us knowing the distribution of cavity's temperature field and the factor that influent the temperature variation better, through three dimensional simulation of temperature field in cooling stage and the contrast verification of experimental results. And the results will provide a theoretical basis for the design of electric heating mould and other moulds.

Corresponding author e-mail: zjujsf@163.com

2 Simulation

2.1 ESTABLISHMENT OF THE GEOMETRIC MODEL

Electric heating mould is similar with the traditional injection mould. The difference lies in that the heating pipes used for installation of electrical heating rods are designed in a fixed mould plate. In order to study the thermal response law of mold temperature field in cooling process, and considering the symmetry of the mould plate at the same time, a quarter model of the mould was studied as the analysis object of numerical simulation, model size was set up according to the actual size of the mould. Moreover, the mold cavity surface was simplified as simple plane, ignoring the actual mold flash structure and the gap existed between fixed mould plate and movable mould plate in order to further reduce the complexity of the model. Figure 1 was the picture of geometric model built to analyse the mold temperature field.
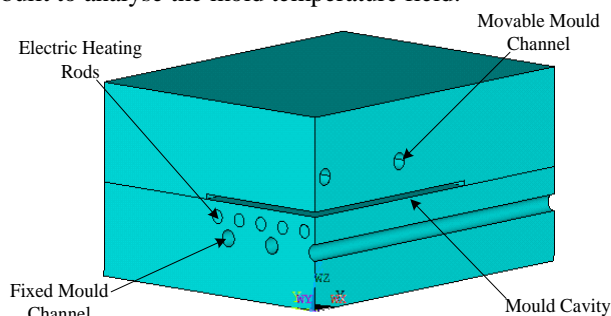
FIGURE 1 Geometric model

2.2 ESTABLISHMENT AND SIMPLIFICATION OF THE MATHEMATICAL MODEL OF COOLING PROCESS

The heat exchange within the injection mold in the cooling stage with electric heating has three main aspects:

1) Plastic melt is contacted directly with the mould cavity, and the heat was transferred to the mould by heat conduction.

2) The cooling pipe inside the mould will transfer most of the heat to the coolant, which is a convection heat transfer process.

3) The mould will release part of the heat directly to surrounding air through thermal convection and thermal radiation. The absolute heat releasing of plastic parts is equal to the sum of the heat taken away from the mould outer surface and the cooling system.

Therefore, the heat transfer process of the cooling stage is composed of the following control equation decision:

$$\rho c \frac{\partial T}{\partial t} - \frac{\partial}{\partial x}(\lambda_x \frac{\partial T}{\partial x}) - \frac{\partial}{\partial y}(\lambda_y \frac{\partial T}{\partial y})$$
$$- \frac{\partial}{\partial z}(\lambda_z \frac{\partial T}{\partial z}) - h(T_m - T_c) - \alpha(t_w - t_\infty) = 0 \quad , \quad (1)$$

where represents the material density of the mold (kg/m³).

Letter c represents the materials specific heat of the mold (Jkg^{-1}K^{-1}), letter T represents the temperature (K), letter t represents the time(s), λx, λy, λz represents the thermal conductivity (Wm^{-1}K^{-1}) along the three main directions of the object (x, y, z), h represents the surface heat transfer coefficient between cooling pipe and the mould (Wm^{-2}K^{-1}), T_m, T_c respectively represent for the mold wall surface temperature and the temperature of the cooling medium (K), α represents the composite surface heat transfer coefficient (Wm^{-2}K^{-1}), t_w, t_∞ respectively represent for the temperature of the air around the mold and the mould surface (K).

The mould exchanges the heat through the cooling channel in the cooling stage. Taking into account the rapid flow of the cooling water, we assume that the initial temperature of the water of mold cooling channel remains consistent.

2.3 PHYSICAL MODELS AND MATERIAL PROPERTIES

Figure 2 describes a part of rapid heat cycle mould with electrical heating, where the size of the fixed mould plate and movable mould plate respectively is 450×350×55.6mm, 450×350×60mm. Compared with the traditional injection molding, the main difference is that a row of heating pipe holes were designed around the mould cavity in rapid heat cycle mould. This can improve the surface quality of plastic though controlling the heating pipe. The fixed mould plate and movable mould plate are made of SP400, which has excellent thermal stability and thermal transmission to ensure the heat generated by the heating rod transfer to the cavity surface rapidly and reduce the heat dissipation of the midway. The material properties of the mould were shown in the following (Table 1).
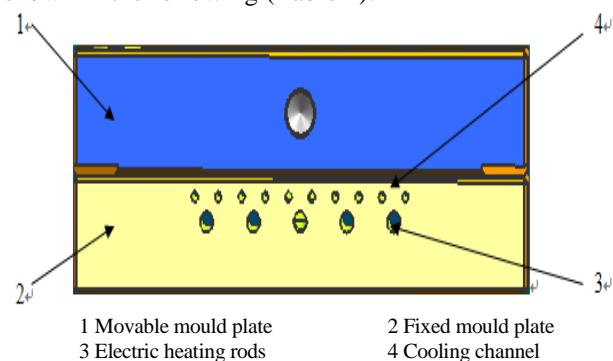

1 Movable mould plate   2 Fixed mould plate
3 Electric heating rods   4 Cooling channel
FIGURE 2 Part of the electric heating mould

TABLE 1 Material properties of the mould plate

Material properties	Parameters
Density	7.8×10³(Kg/m³)
Heat Conductivity	34 (W/mK)
Specific Heat	460 (J/Kg)
Thermal Expansion	11.8×10⁻⁶

2.4 THE BOUNDARY CONDITIONS

Mould and plastic part: The process of heat transfer

between the mould and the plastic parts was simplified as pure heat conduction process owing to the neglect of latent heat, and the physical parameters were considered as a constant. As the temperature value of the cavity surface was obtained by simulating the filling and packing stage, so the process was controlled by the first boundary condition:

$$t_w = f(x, y, z, \tau), \ \tau > 0 \ . \tag{2}$$

Mould and cooling channel: The heat transfer between the mold and the cooling channel was completed by heat convection and heat transfer. The particle of coolant was mixed to reduce the heat. As the temperature difference between the mold wall and the cooling channel, the heat was also transferred by thermal conduction. The basic heat transfer process can be expressed as the expression:

$$q = h(T_m - T_c) \ . \tag{3}$$

We could obtain convective heat transfer coefficient from the third boundary condition. Letter h represents surface coefficient of heat transfer between the mould and cooling channel (Wm^{-2}K^{-1}), T_m, T_c respectively represents the temperature of the mold surface and the cooling medium (K). The speed of the cooling process was determined by the value of h, which is closely related to the flow of the coolant state. This relationship can be reflected by the Reynolds number Re.

Mould and surrounding environment: The outer surface of the mold was exposed in the ways of heat transfer include both heat convection and heat radiation, which belonged to the third kind boundary condition. Expressions of heat transfer boundary condition could be written as follows:

$$q_w = -\lambda \frac{\partial t}{\partial n}\Big|_w = \varepsilon\sigma(T_w^4 - T_{sur}^4) + h(T_w - T_f) = \alpha(t - t_\infty) \ . \tag{4}$$

The significance of each parameter was the same as the heating process.

2.5 INITIAL CONDITIONS

The ambient air temperature was 27°C. The heat transfer coefficient of mold and the surrounding environment was 25Wm^{-2}K^{-1} according to the actual production environment of injection molding cooling stage. The heat transfer coefficient of the mold and cooling pipes was 5707.6 Wm^{-2}K^{-1}.

3 Experimental

3.1 INITIAL CONDITIONS

In order to verify the rapid heat cycle molding technology in improving the quality of plastic product, we used electric heating as an example and designed the structure of plastic product. And then we designed a set of electric heating

mould (as showed in Figure 2) and set up the experimental platform of the electric heating mould process.

The type of the injection molding machine is Haitian HTFX5 series MA3800 in this experiment, maximum clamping force of 3800KN, the biggest melt volume is 1239cm3, the maximum injection pressure can reach 182Mpa. This machine is suitable for demanding precision plastic molding.

The type of the electric heating controller is MTS-32II, which uses a three-phase and five-wire system supplied by 380V power, a single maximum output power of 6.5KW,Temperature control range of 0-399.9, time control accuracy of 10ms, maximum controllable loop digital 24.

Electric heating rods in this experiment are also known as the high-density single end electric heating tube or tube electric tube. The parameters of the electric heating rods respectively are 220V/270W, diameter 4mm, length of 450mm and 220V/210W, diameter of 4mm, length of 134mm, respectively marked as the type I and type II.

3.2 EXPERIMENTAL DESIGN OF RAPID HEAT CYCLE MOLDING WITH ELECTRIC HEATING

Engineering plastics HDPE was used in the experiment. Electric heating controller adjusts the temperature of electric heating rods to heat the mold firstly. Therefore we can know the changes of temperature through the measurement of the thermocouple in a fixed mould plate. We can read and record the temperature of the mould from the screen. The correctness of the simulation process should be verified next by comparing the simulation temperature information at the corresponding position in the thermocouple of the measured temperature information and fully taking into account the impact of the experimental error as well as the simplified model on the results of experiment and simulation. The installation position of the thermocouples was shown in Figure 3, where thermocouple 1. and 2., 3. and 4. respectively on the fixed mould template central symmetric.
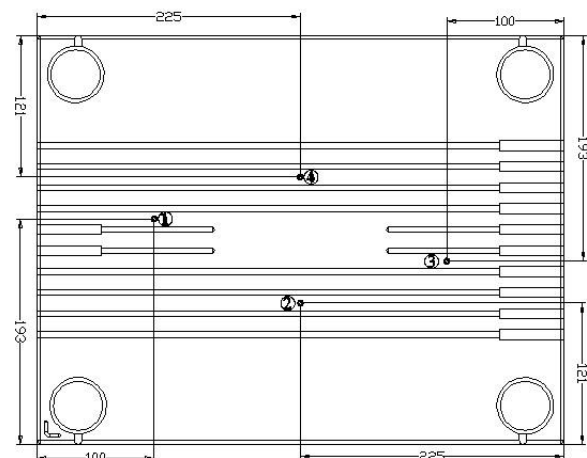

FIGURE 3 The installation position of the thermocouples

Six observation points were selected uniformly in the direction of the mold cavity surface. They were marked by P1, P2, P3, P4, P5 and P6, the distance between each

point is 18mm. Three observation points were selected uniformly in the direction of the vertical cavity surface. And they were marked by P3, P7 and P8. The location of point P1-P8 was shown in Figure 4.
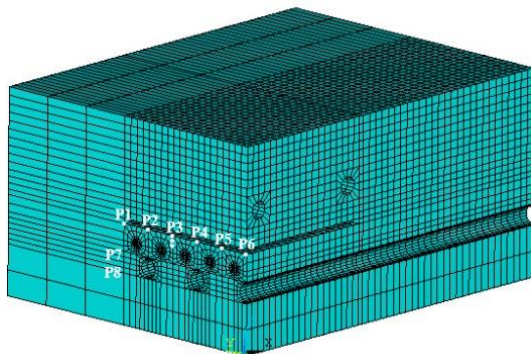

FIGURE 4 The location of observation points

Consequently we recorded the temperature value of the mold cavity corresponding to the position of the thermocouples installed in a fixed mould plate according to the measuring temperature. The location of the thermocouples was shown in Figure 3. We need only observe the thermocouple temperature of position 1. and 4. in this experiment due to the symmetries of the installation location of the thermocouples and the position of electric heating rods. The data of each point were collected every 5 seconds in this experiment. The temperature data of thermocouple 1. and 4. in the entire cooling process (25 seconds) were shown in the following (Table 2).

TABLE 2 Temperature data in cooling process

Time(s)	Thermocouple 1. (℃)	Thermocouple 4. (℃)
0	100.4	95.8
5	95.1	92.0
10	88.0	85.8
15	81.2	78.3
20	75.2	72.7
25	70.6	67.1

3.3 EXPERIMENTAL VERIFICATIONS

The contrast temperature graph of experiment and simulation of thermocouple 1. and thermocouple 4. is shown in Figure 5. From the graph, we can see that the cooling speed of mould is gradually decreased and the temperature of mould drops slightly slows in cooling process. The value of the simulation and experimental maintains the consistent trend of cooling rate. Nonetheless, the mold cooling rate in the simulation was obviously higher than the cooling rate of mold in the experiment at the same time, and the temperature difference gradually expanded along with the cooling process. The case of the tendency is that the release of latent heat of plastic melt has been ignored in model simplification.

The plastic melts generate phase transition by the changing of temperature. The process of phase transition will release plenty of latent heat, which supplies the loss of heat owing to the cooling effect of water. So the mould temperature in experimental is always higher than the temperature in the simulation process at the same time, and the gap can be regarded as the additional value obtained by releasing the latent heat of plastic melt.
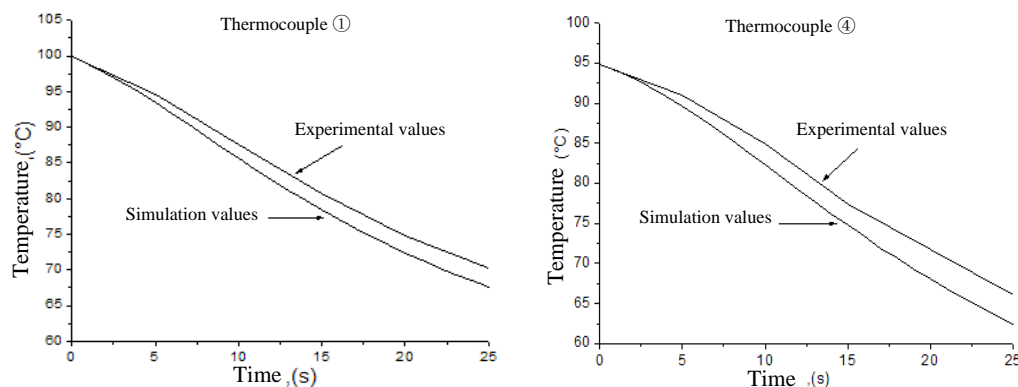
FIGURE 5 The contrast of experiment and simulation temperature in cooling process

We concluded that the simulation process is good to reflect the change of the temperature field in cooling stage with electric heating without considering the latent heat released through the phase transition of plastic melt.

4 Results and discussion

4.1 DISCUSSION ABOUT THE RESULTS OF SIMULATION

It is known that the uniformity and cooling rate of cavity surface temperature in cooling stage has a significant

impact on the quality and production efficiency of plastic products. On one hand, the gradients of cavity surface temperature will result in the inhomogeneity of parts surface and excessive internal residual stress of plastic parts. It may be deformed or even failure in the process of using plastic parts because of the release of stress. On the other hand, cooling rate directly affects the cycle of plastic injection molding process, and the overlong molding cycle will cause a decrease in productivity. So it is necessary to study the temperature field in cooling stage of rapid heat cycle molding with electric heating. We can be faster and more intuitive understanding of the temperature field of

the mold cavity in the cooling stage. It also can provide reasonable structural design of mould for us.

4.2 TEMPERATURE UNIFORMITY

Figure 6 shows the temperature distribution of the mold cavity surface around the fixed mould plate in the cooling time of 25 seconds. The overall temperature of the cavity surface is less than 75℃℃, which is lower than the glass transition temperature of the plastic melt. We can see from the picture that cavity surface temperature is slightly reduced from the cavity centre to the edge of the cavity. The maximum temperature of the cavity surface was 73.7℃℃, the lowest temperature was 40.5℃℃, the average temperature of the cavity surface is 59.1℃℃. The temperature of the red zone in this picture was significantly higher than the cavity edge part. This is caused by the high power of type II electric heating rods in the heating stage, which is caused by local heating too fast. The temperature of the cavity edge is lower than the average temperature of the entire cavity surface. This is because the arrangement of the electric heating rods in the heating stage is not sufficient heating to the cavity edge part. So the temperature is always lower than the centre of the cavity in the cooling stage. And this results in a phenomenon of poor uniformity of the cavity temperature.
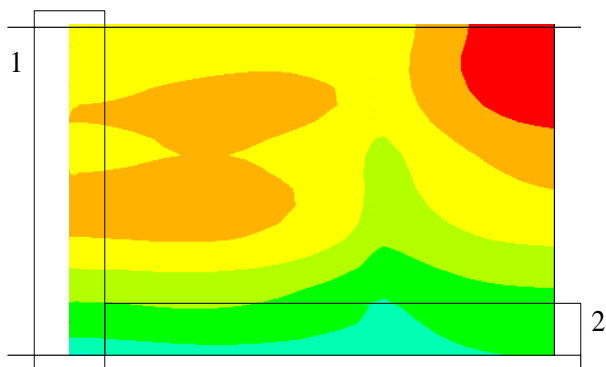
FIGURE 6 The temperature field of the cavity surface

4.3 COOLING RATE

Figure 7 is the temperature change graph of point P1-P6 in the cooling time of 40 seconds. The average cooling rate of P1-P6 respectively are 0.37°C/s, 0.96°C/s, 1.08°C/s, 1.12°C/s, 0.84°C/s and 0.75°C/s. It can be seen through point P1 and point P2 that the temperature of P1 is far less than point P2 at the beginning, but the distance between the cooling pipe the distance between point UP and cooling pipe is smaller than PLY. The heat can be passed rapidly through the mould to cooling water. Moreover, decrease heat of natural convection from the mould surface is much less than the heat, which is loosed from the mould and cooling water. So the application of cooling water can effectively reduce the total cycle of injection molding.
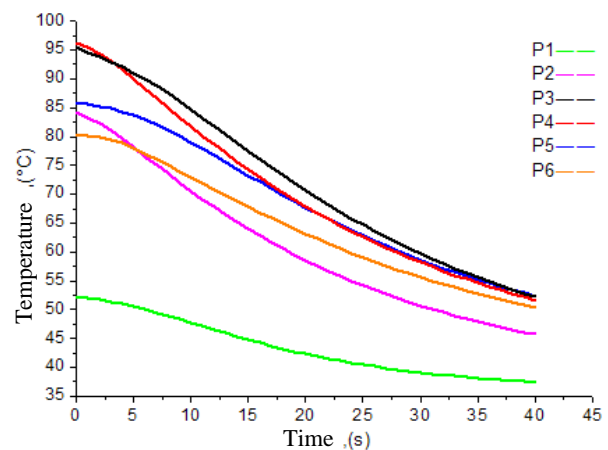
FIGURE 7 Temperature change graph of P1-P6 in cooling stage

It can be seen by comparing the cooling process of point P2 and point P3 that the cooling rate basically consistent in each process. The distance between point P2 and each cooling pipe is very close to point P3. We can see that the cooling rate of point P4 is permanently greater than point P3 in each process of cooling stage, and the distance between point P4 and each cooling pipe is smaller than point P3. In conclusion, we can conclude that the speed of the cooling rate in somewhere was inversely related to the distance between the cooling pipe and the point. Namely, the cooling rate is decreased with the increasing distance from the cooling pipe. It also can be seen from the temperature change trend that the cooling rate decreases with the decrease of temperature. The temperature difference between each point and the cooling water is gradually reduced and the temperature gradient becomes slightly smaller along with the cooling process. In order to shorten the time required for the cooling stage, we should try to reduce the temperature of cooling water and accelerate the flow rate of the cooling water in the cooling pipe under the working conditions. The cooling water which is rose due to heat exchange should be drainage outside the mould quickly.

5 Conclusions

In this study, the main emphasis is placed on the problem of the distribution of cavity's temperature field in cooling stage with RHCM. The geometric model of the temperature field of electric heating mould was establishment. The proposed model is verified through experimental study. The paper showed a good temperature uniformity and extraordinary efficiency of the cooling rate The results was obtained helping us knowing the distribution of cavity's temperature field and the factor which affected the temperature variation better. And the results will provide a theoretical basis for the design of electric heating mould and other moulds.

Acknowledgments

References

[1] Yao D G, Chen S C, Kim B H 2008 *Advances in Polymer Technology* **27**(4) 233-55

[2] Chen S C, Jong W R, Chang J A 2006 *Journal of applied polymer science* **101**(2) 1174-80

[3] Li J Q, Jiang S F, Chen C 2011 *Advanced Materials Research* **189** 2543-6

[4] Jeng M C, Chen S C, Minh P S, Chang J A, Chung C 2010 *International Communications in Heat and Mass Transfer* **37**(9) 1295-04

[5] Jiang S F, Zheng W, Zhang J B, Li J Q 2012 *Applied Mathematics & Information Sciences* **6**(3) 665-71

[6] Reddy B, Kumar J S, Reddy V K, Padmanabhan G 2009 *JESTEC* **2**(1) 56-62

[7] Barone M R, Caulk D A, Panter M R 1986 *Polymer composites* **7**(3) 141-51

[8] Ivanov S V, Perepelica S I, Bikmeyev A T 2012 *Journal of Engineering Science and Technology Review* **5**(3) 1-5

[9] Wang G L, Zhao G Q, Li H P, Guan Y J 2010 *Materials & design* **31**(7) 3426-41

[10] Park H S, Dang X P 2010 *International Journal of Precision Engineering and Manufacturing* **11**(6) 879-90

[11] Singh A K, Baumann G, Henninges J, Gorke U J, Kolditz O 2012 *Environmental Earth Sciences* **67**(2) 497-09

Authors

Shaofei Jiang, born in August, 1975, Datong, China

Current position, grades: the professor of Zhejiang University of Technology, China, a Senior member of Chinese Mechanical Engineering Society.
University studies: Ph.D. degree in mechanical design from Zhejiang University, China, in 2004.
Scientific interest: the theory of product design, the advanced of mold technology, manufacturing industry informatization.
Publications: 35 papers, 17 patents.

Zhifei Mao, born in April, 1990, Quzhou, China

Current position, grades: M.S. degree student in Zhejiang University of Technology, China.
University studies: B.S. degree in mechanical engineering and automation from Lishui University, China, in 2012.
Scientific interest: the theory of product design.

Jiquan Li, born in November, 1977, Jinhua, China

Current position, grades: the associate professor of Zhejiang University of Technology, China.
University studies: Ph.D. degree in mechanical design from Shanghai Jiao Tong University, China, in 2008.
Scientific interest: injection molding process and equipment, polymer material processing and forming principle.
Publications: 20 articles, 10 patents.

Wei Zheng, born in June, 1987, Changsha, China

University studies: M.S. degree in Zhejiang University of Technology, China, in 2014.
Scientific interest: rapid heat cycle injection molding technology and EGAIM.

A novel method to calculate relative permeability of fluids based on fractal theory and core NMR experiment

Xinmin Ge[1, 2*], Yiren Fan[1, 2], Donghui Xing[1, 2], Yongjun Xu[1, 2]

[1]*School of Geosciences in China University of Petroleum. Qingdao 266580, Shandong Province, China*

[2]*CNPC Key Well Logging Laboratory in China University of Petroleum. Qingdao 266580, Shandong Province, China*

Abstract

An analytical relative permeability model based on fractal theory and NMR principle is described. The relationship between NMR transversal relaxation time (NMR T2) and resistivity index is deduced by fractal theory and capillary bundle model. The reciprocity theory of percolation field and electricity field is used to get the relationship between resistivity index and relative permeability. The 'Brooks-Corey/Burdine' equation was adopted to construct the relationship between relative permeability and NMR transversal relaxation time. By introducing the NMR parameters, the relative permeability model is improved since that it quantified the influence of pore structure. The results show that for water (wetting phase in water-gas system), the predicted permeability is exactly consistent with the experiment data, whereas for gas (the non-wetting phase in water-gas system), the fitted results is to some degree deviated from the experiment data and need more extensive research.

Keywords: relative permeability, fractal theory, NMR transversal relaxation time, reciprocity, pore structure

1 Introduction

The relative permeability of multiphase in porous rock is important to fluid recognition, reservoir numerical simulation, productivity evaluation and design of oil/gas field development plan. Purcell (1949) [1] derived an equation between relative permeability and capillary pressure based on a simplified working hypothesis. Burdine (1953) [2] got formulas for calculating relative permeability from pore size distribution data by introducing tortuosity factors and basic laws of fluid flow in porous media, which is a revised version of Purcell's equation since it allowed the impact of irreducible fluid saturation. Corey (1954) [3] investigated the interrelation between gas and oil relative permeabilities then proposed two equations for relative permeability calculation by approximation. Pirson (1958) [4], Brooks et al (1966) [5] also researched the influential factors of relative permeability and got analytical expressions related to capillary pressure and saturation. He, et al (2000) [6], Zhou et al (2007) [7] calculated fractal dimension of pore space from capillary pressure data then established gas/water relative permeability model by fractal theory. O.A.Olafuyi et al (2008) [8] investigated the relationship between capillary pressure and relative permeability of small cores. Li et al (2008, 2010 and 2011) [9-11], conducted systematically research on the interrelationships among relative permeability, capillary pressure and resistivity index then obtained analytical expressions for those parameters. From literatures review we know that most relative permeability models presented ever were deduced from capillary pressure or by simple fitting methods. Many researches noticed that the relative permeability has a great relation of intrinsic pore structure, but the quantitative expressions were not well proposed yet. Nuclear Magnet Resonance (NMR) is a well used technology to study the pore structure of rock-pore system, by which one can obtain pore throat distribution, total porosity, irreducible fluid saturation, absolute permeability etc. In this paper, the author started to investigate the fractal property of porous rock and construct the relationships among fluid saturation, capillary pressure and resistivity index. Then the reciprocity between percolation field and electricity field was discussed and simple capillary bundle model was used to connect the relative permeability and resistivity index. At last, relative permeability model, which related to fluid saturation, NMR transversal relaxation time (NMR T2) based on 'Brooks-Corey/Burdine' equation was established. The model was discussed and analysed by core experiments.

2 Fractal theory and conductivity property rock-pore system

Researches demonstrated that pore space of sediment rock has fractal feature and can be depicted by power law function. (A.J.Katz et al, 1985 [12]). Li et al (2011) [11] got the relationship between water saturation and resistivity by fractal theory, which can be expressed as:

Corresponding author e-mail: gexinmin2002@163.com

$$\frac{1}{R_t} \propto S_w^{\frac{1}{f(3-D_f)}}, \tag{1}$$

where S_w is the water saturation, R_t is the resistivity of rock at a water saturation of S_w, D_f is fractal dimension of pore space and f is the parameter, which affected by the thickness of water film and surface tension.

Accordingly, with the theory of fractal and capillary bundle model, the relationship between water saturation and capillary pressure can be expressed as:

$$S_w \propto P_c^{-(3-D_f)}, \tag{2}$$

where P_c is the capillary pressure at a water saturation of S_w.

Combing Equations (1), (2) and Archie's equation, we can easily obtain:

$$P_c = P_0 I^f, \tag{3}$$

where P_0 is the capillary pressure at a water saturation of 1 and I is the resistivity index.

The NMR transversal relaxation time (NMR T_2) can be expressed as (George Coates, et al, 2000 [13]):

$$\frac{1}{T_2} = \frac{1}{T_{2b}} + \frac{1}{T_{2s}} + \frac{1}{T_{2d}}, \tag{4}$$

where T_2 is the total transversal relaxation time, T_{2b} is the bulk relaxation time, T_{2s} is the surface relaxation time and T_{2d} is the diffusion relaxation time. Practically, the bulk relaxation part and the diffusion relaxation part are often omitted, so the total transversal relaxation time can be summarized as:

$$\frac{1}{T_2} = \frac{1}{T_{2b}} = \rho_2 \left(\frac{S}{V} \right), \tag{5}$$

where ρ_2 is the transversal relaxation ratio, S/V is the ratio of surface to bulk for pore space.

From the capillary theory, for the simplified capillary tube model, one can solve Laplace equation of capillary pressure and get:

$$P_c = \frac{2\sigma \cos \theta}{r}, \tag{6}$$

where σ is surface tension, θ is the contact angle of wetting and non-wetting phase, r is the radius of capillary tube.

For the simple sphere or cylindrical pore structure, the ratio of surface to bulk(S/V) has a linear relationship with pore radius. Combining Equation (5) and (6), the relationship between capillary and NMR T_2 can be expressed as:

$$P_c = C \frac{1}{T_2}, \tag{7}$$

where C is the transforming parameter.

In fact, the pore space is far more complicated than simple sphere or cylindrical-like shape, so the relationship between T_2 and P_c is unknown. But it is definitely that the capillary is a function of NMR T_2:

$$P_c = g \left(\frac{1}{T_2} \right), \tag{8}$$

where g is a function with no specific form.

He et al (2005) [14] got an analytical relationship between capillary pressure and NMR T_2 by large amount of experiments:

$$P_c = m \left(\frac{1}{T_2} \right)^n, \tag{9}$$

where m and n are transforming parameters.

Combing Equation (3) and (9), the relationship between resistivity index and NMR T_2 can be expressed as:

$$I = a \left(\frac{1}{T_2} \right)^b, \tag{10}$$

where a and b are transforming parameters.

Equation (10) is important because it connected the conductivity property with pore structure, which is the bridge for the NMR T_2 and relative permeability discussed below.

3 Reciprocity theory and calculation of relative permeability

The conduction of electricity and the percolation of fluid are similar from the view of mathematical and physical theory. Since the similarities of boundary condition and initial condition, the solving methods of both problems are alike to some extent. Fatt (1956) [15] conducted extensive comparison between the Poiseuille's equation and Ohm's law. King (1989) [16], Li (2000) [17], B.Shimekit, H.Mukhtar et al (2010) [18] also researched the interrelationship between electricity transportation properties and fluid percolation features. With Poiseuille's equation, the flow of capillary bundle-like porous rock (for single phase) can be expressed as:

$$Q = \frac{\pi r^4}{8 \mu l} \Delta P, \tag{11}$$

where Q is flow of single phase, r is the radius of capillary bundle, μ is the viscosity of fluid, l is the length of capillary bundle and ΔP is the pressure difference between inlet and outlet.

By Ohm's law, the electric current of capillary bundle can be expressed as:

$$E_c = G \Delta V, \tag{12}$$

where E_c is the current, G is the conductance and ΔV is the potential difference.

With the reciprocity of percolation field and electricity field, combining Equations (11) and (12), the conductance of the capillary bundle can be expressed as (Fatt, 1956) [15]:

$$G = \frac{\pi r^4}{8\mu l}. \tag{13}$$

The flow of single phase in capillary bundle can be expressed by Darcy law:

$$Q = \frac{AK}{\mu l}\Delta P, \tag{14}$$

where A is the area of cross section of the capillary bundle, K is the absolute permeability of single phase.

Assuming the water and gas in the capillary bundle, combining Equations (11), (13) and (14), the conductance of rock fully saturated with water is:

$$G_o = \frac{AK}{\mu l}. \tag{15}$$

By introducing the relative permeability, the conductance of rock partly saturated with water can be expressed as:

$$G_t = \frac{AKK_{rw}}{\mu l}, \tag{16}$$

where K_{rw} is the relative permeability of water.

From Archie's equation, the resistivity index can be expressed as:

$$I = \frac{R_t}{R_o}, \tag{17}$$

where R_o is the resistivity of core fully saturated with water.

Combining Equations (15), (16) and (17), we obtained:

$$I = \frac{R_t}{R_o} = \frac{G_o}{G_t} = \frac{1}{K_{rw}}. \tag{18}$$

From equations above we know that if rock is fully saturated with water, both the resistivity index and the relative permeability of water are 1, if rock is fully saturated with gas, the relative permeability of water equals to 1 and the resistivity equals to infinite value. But from experiments we know that the relative permeability of water becomes zero when water saturation reduced to irreducible state, so Equation (18) should be revised to suit the actual condition, we called it saturation revision here. After revision, Equation (18) can be expressed as:

$$K_{rw} = \frac{1}{I}\frac{S_w - S_{wir}}{1 - S_{wir}}, \tag{19}$$

where S_w is water saturation and S_{wir} is irreducible water saturation.

Combining Equation (10) and (19), the relative permeability of water can be expressed as:

$$K_{rw} = aT_2^{\ b}\frac{S_w - S_{wir}}{1 - S_{wir}} = ce^{dT_2}\frac{S_w - S_{wir}}{1 - S_{wir}}, \tag{20}$$

where c and d are fitting parameters.

Equation (20) is the new model to calculate the relative permeability of water.

Burdine et al (1953) [2] proposed a new method to calculate relative permeability:

$$K_{rw} = \left(\frac{S_w - S_{wir}}{1 - S_{wir}}\right)^2 \left(\int_0^{S_w} dS_w / P_c^2\right) / \int_0^1 dS_w / P_c^2, \tag{21}$$

$$K_{rnw} = \left(\frac{S_{nw} - S_{wir}}{1 - S_{nwr} - S_{wir}}\right)^2 \left(\int_{S_w}^1 dS_w / P_c^2\right) / \int_0^1 dS_w / P_c^2, \tag{22}$$

$$S_w + S_{nw} = 1, \tag{23}$$

where K_{rw} is the relative permeability of wetting phase, K_{rnw} is the relative permeability of non-wetting phase, S_w is the saturation of wetting phase, S_{wir} is the irreducible saturation of wetting phase, S_{nw} is the saturation of non-wetting phase and S_{nwr} is the irreducible saturation of non-wetting phase. For simplicity, we often assume it is 0.

Olafuyi et al (2008) [8] proposed the relationship between capillary pressure and saturation of wetting phase by fractal theory:

$$P_c = P_o \left(\frac{S_w - S_{wir}}{1 - S_{wir}}\right)^{-1/\lambda}, \tag{24}$$

where P_o is the capillary pressure of core fully saturated with wetting phase, λ is a parameter which relates to fractal feature of pore space.

Substituting Equations (24) to Equations (21) and (22), we can get:

$$K_{rw} = \left(\frac{S_w - S_{wir}}{1 - S_{wir}}\right)^{2/\lambda+3}, \tag{25}$$

$$K_{rnw} = \left(1 - \frac{S_w - S_{wir}}{1 - S_{wir}}\right)^2 \left[1 - \left(\frac{S_w - S_{wir}}{1 - S_{wir}}\right)^{2/\lambda+1}\right]. \tag{26}$$

Equations (25) and (26) are the known 'Brooks-Corey/Burdine' relative permeability equations. In this study, only gas and water are existed in rock, so water is the wetting phase and gas is the non-wetting phase and S_w denotes the water saturation, S_{wir} denotes the irreducible water saturation, S_{nw} denotes the gas saturation, K_{rw}

denotes the relative permeability of water, K_{rnw} denotes the relative permeability of gas from Equation (21) to Equation (26), which has the same meanings from Equation (1) to Equation (20).

Combining Equations (20) and (25) we can obtain:

$$\left(\frac{S_w - S_{wir}}{1 - S_{wir}}\right)^{2/\lambda} = ce^{dT_2}\left(\frac{S_w - S_{wir}}{1 - S_{wir}}\right)^{-2}. \tag{27}$$

Substituting Equation (27) to Equation (26) we can obtain the relative permeability of gas:

$$K_{rnw} = \left(1 - \frac{S_w - S_{wir}}{1 - S_{wir}}\right)^2\left[1 - ce^{dT_2}\left(\frac{S_w - S_{wir}}{1 - S_{wir}}\right)^{-1}\right]. \tag{28}$$

Equation (20) and equation (28) are the new models presented to calculate the relative permeability of water and gas. From the formulas and parameters we know that the new models related not only the fluid saturation, but also the irreducible fluid saturation and NMR T_2, which are more suitable since it reveals that percolation process is affected by pore structure and the relationship between

the relative permeability and pore structure is quantified by NMR T_2.

4 Experiments and discussions

Four shaly sand samples were selected for experiments in order to testify the model proposed. The basic parameters were shown in Table 1. The porosity ranges from 11.49% to 17.53% and the absolute permeability ranges from 1.44mD to 15.99mD.Nacl with a concentration of 8000ppm and a viscosity of 0.7328mPa.s was used for the saturated solution. Nitrogen with a viscosity of 0.0174mPa.s was used as the injecting phase (non-wetting phase). The echo time and waiting time for NMR test were 0.3ms and 6s respectively. The scanning number and echo number were 64 and 2048 respectively. Cores were centrifuged for 30 minutes with a velocity of 8000rpm.SVD decomposition method was used to fitting the NMR T_2 spectrum and 64 points of inversion was adopted. Irreducible water saturation was obtained by relative permeability and centrifuge methods. All the tests were conducted under a temperature of 19°C.

TABLE 1 Basic petrophysical parameters of core samples

Sample ID	Nitrogen porosity (%)	NMR porosity (%)	Nitrogen permeability (mD)	Swi by relative permeability method (%)	Swi by centrifuge (%)
1	11.49	11.13	1.44	68.27	63.94
2	13.72	13.94	3.18	75.02	71.53
3	17.53	17.25	15.99	51.39	53.26
4	16.51	16.88	7.24	39.11	36.75

Figure 1 shows NMR T_2 spectrum of cores with a saturation of 1. Except for No.4 sample, others bear single peak or irregular double peaks in NMR T_2 spectrum. The percentage of immobile part is to some extent larger than that of mobile part, which reveals that the percentage of micro porosity is higher than the macro porosity. Figure 2 depicts the comparison between NMR porosity and nitrogen porosity. Both of methods have similar results, the average absolute error is 0.31% and the average relative error is 2.14%. The irreducible water saturations obtained by gas-water relative permeability method and centrifuge methods were shown in table.1, both methods have exactly the same result, the average absolute error is 3.02% and the average relative error is 5.4%. Discussions above show that the quality of data obtained in our study are fine.

Since the measured number of relative permeability is not equal to that of NMR T_2 spectrum, interpolation was conducted to adjust the number of NMR T_2 and relative permeability to the same for curve fitting. Log-log method was adopted to get 100 points of water saturation and the corresponding NMR T_2 and relative permeability by our experiences. Noticing that the upper limit of water saturation is 1 and the lower limit of water saturation is its irreducible water saturation. Figure 3 shows comparison between interpolated relative permeability and measured data. The interpolated data were in agreement with the experiment data.
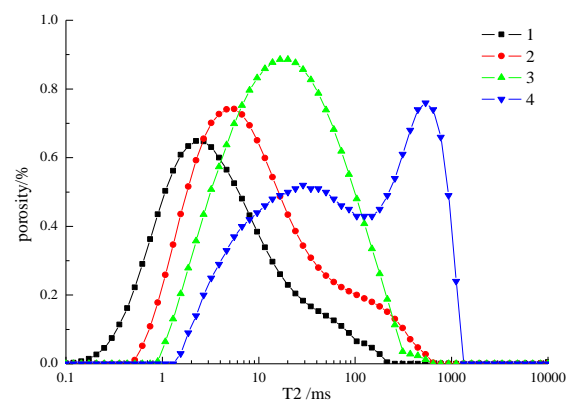

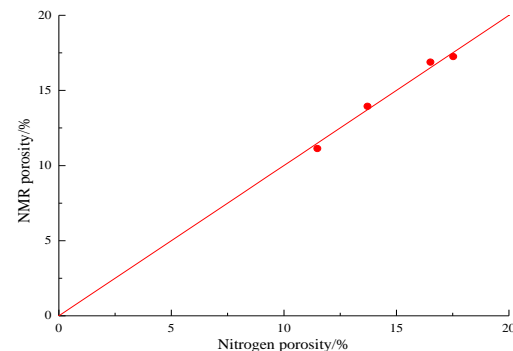FIGURE 1 NMR T2 spectrum of samples fully saturated with salinity

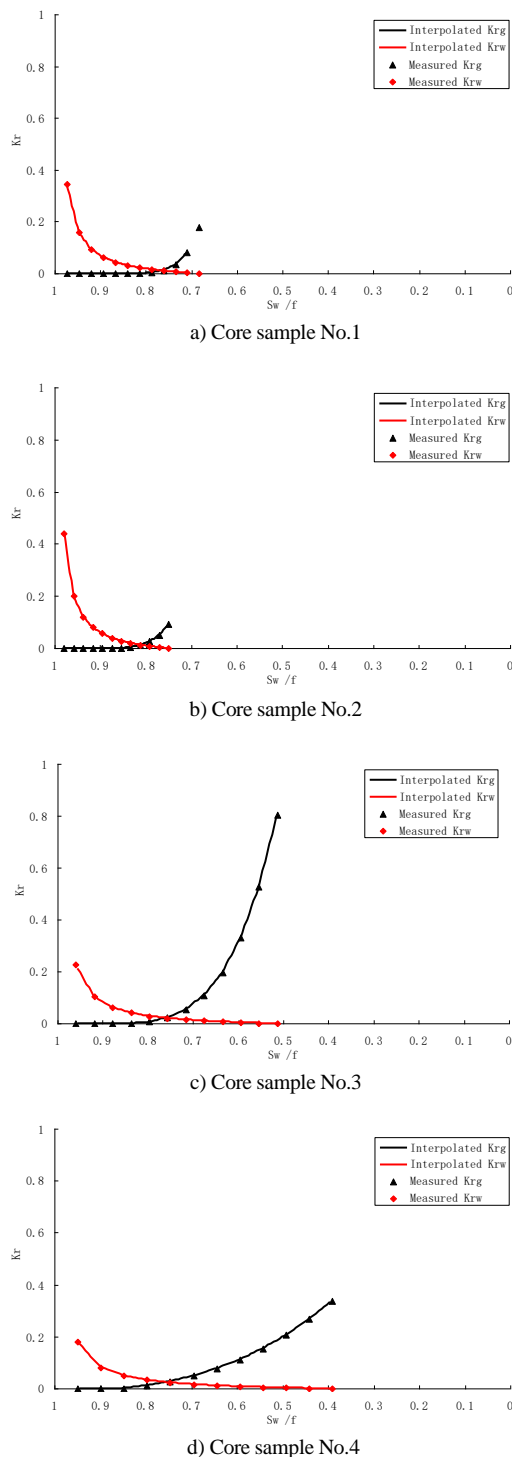FIGURE 2 Comparison between NMR porosity and Helium porosity

a) Core sample No.1

b) Core sample No.2

c) Core sample No.3

d) Core sample No.4

FIGURE 3 Comparison between interpolated water/gas relative permeability and measured data

NMR T_2 and gas/water relative permeability under the same water saturation were obtained after interpolation. The next procedure was the fit of relative permeability. We fitted the relative permeability of water by Equation (20) for every sample firstly, then fitted the relative permeability of gas by Equation (28) with parameters of 'c' and 'd' fitted from Equation (20) for every sample.

Figure 4 shows the relation between $K_{rw}(1-S_{wir})/(S_w-S_{wir})$ and T_2. The black dots are data after interpolation and the black line are fitting curves using equation (20).The fitting results showed that the both has a excellent relationship of power function. The average multiple correlation coefficient is 0.975.

$y = 0.0317e^{0.0423x}$
$R^2 = 0.9684$

a) Core sample No.1

$y = 0.0386e^{0.0112x}$
$R^2 = 0.9853$

b) Core sample No.2

$y = 0.0162e^{0.0205x}$
$R^2 = 0.9946$

c) Core sample No.3

$y = 0.0158e^{0.0031x}$
$R^2 = 0.9892$

d) Core sample No.4
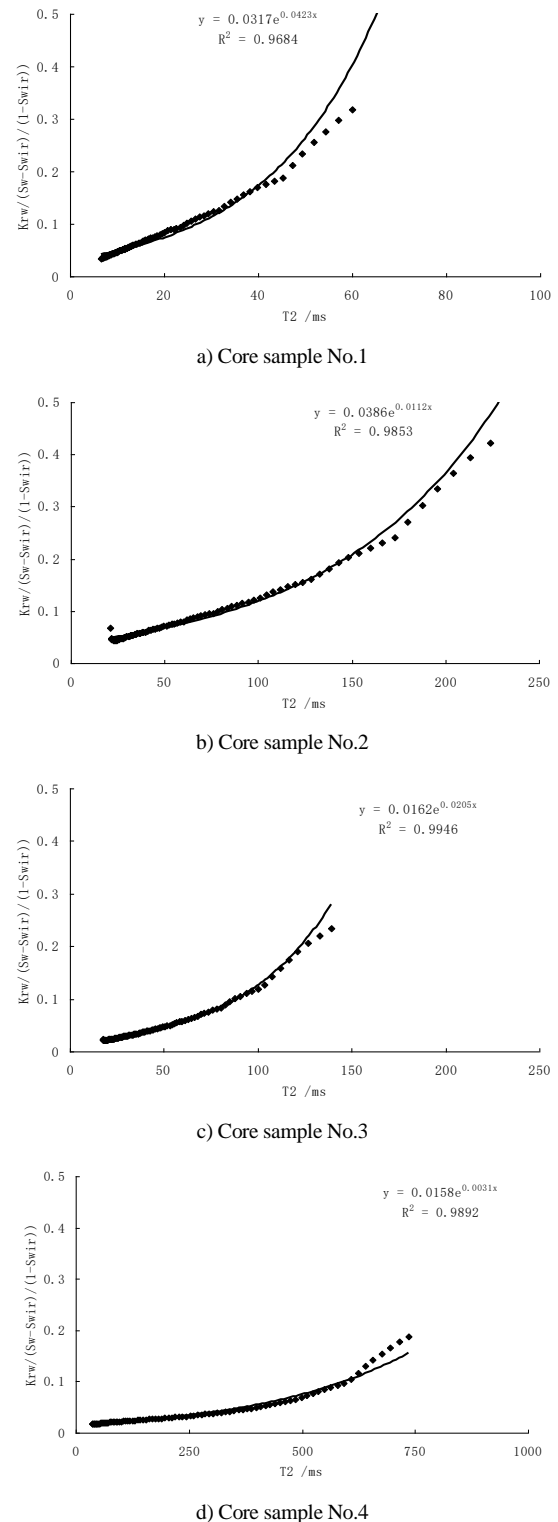
FIGURE 4 Interrelationship between $K_{rw}(1-S_{wir})/(S_w-S_{wir})$ and T2

Figure 5 shows the comparison between fitted water/gas relative permeability and measured water/gas relative permeability. It can be seen from the results that for water (wetting phase), the fitted relative permeability is in accordance with the measured data, the average relative error is 2.45%; but for gas (non-wetting phase), the fitted relative permeability is deviated from the measured data, which may caused by the simplicity of irreducible gas(non-wetting phase) saturation. If we want to get more exact result of the relative permeability of gas (non-wetting phase), the irreducible gas (non-wetting phase) saturation should be considered.

a) Core sample No.1

b) Core sample No.2

c) Core sample No.3

d) Core sample No.4

FIGURE 5 Comparison between fitted relative permeability and measured data
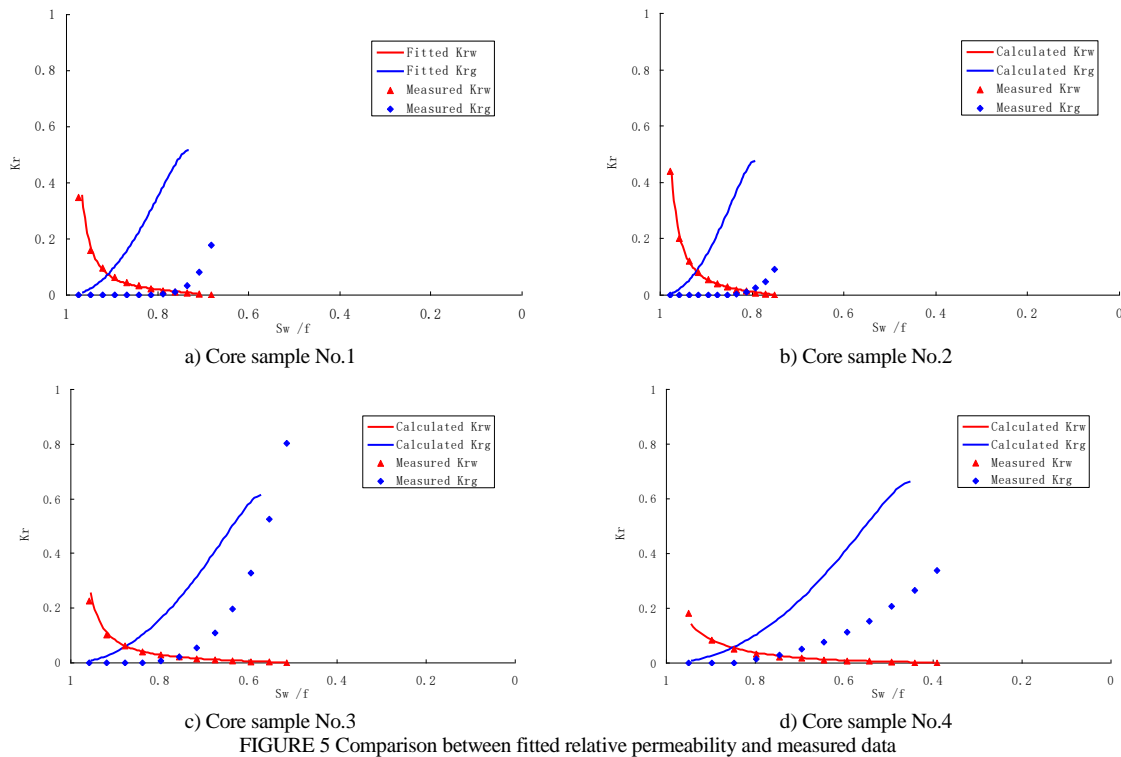
5 Conclusions

With the fractal theory and the reciprocity of percolation field and electricity field, the relationship between relative permeability and NMR T_2 was constructed and testified by experiment; on the basis of this study we make the following conclusions:

1) The power function relationship between resistivity index and NMR T_2 was established by fractal feature of pore space and principles of NMR and capillary pressure, which made the interrelationship of conductivity property and pore structure more quantitative.

2) The boundary conditions, initial conditions and solving methods for percolation field and electricity field are to large extent the same; with the reciprocity of the two fields, the relationship between relative permeability and resistivity index was constructed, which made the conductivity property and percolation property more quantitative.

3) The model proposed for the calculation of relative permeability of water (wetting phase) is fine and exact; but for gas (non-wetting phase), the predicted relative permeability is not coherent with the measured data since the model did not consider the influence of irreducible gas saturation (non-wetting phase). If one need to predict the relative permeability of non-wetting phase exactly, the determination of irreducible saturation is inevitable.

Acknowledgments

References

[1] Purcell WR 1949 Capillary Pressures-Their Measurement Using Mercury and the Calculation of Permeability Therefrom *Journal of Petroleum Technology* **1**(02) 39-48

[2] Burdine N T 1953 Relative Permeability Calculations from Pore Size Distribution Data *Journal of Petroleum Technology* **5**(3) 71-8

[3] Corey A. T 1954 The Interrelation between Gas and Oil Relative Permeabilities *Producers Monthly* **19** 38-41

[4] Pirson S J 1958 Oil Reservoir Engineering *McGraw-Hill New York*

[5] Brooks R H, Corey A T 1966 Properties of Porous Media Affecting Fluid Flow *Journal of the Irrigation and Drainage Division* **92**(2) 6190

[6] He Y 2000 A new method for measuring the relative permeability quantitatively *Petroleum Exploration and Development* **27**(5) 66-8 *(in Chinese)*

[7] Zhou K, Zhang Q, Wang Qun 2007 Calculation of gas and water relative permeability by the fractal model *Natural Gas Industry* **27**(10) 88-9 *(in Chinese)*

[8] Olafuyi O A, Cinar Y, Knackstedt M A, Pinczewski W V 2008 Capillary Pressure and Relative Permeability of Small Cores SPE11386.SPE/DOE *Symposium on Improved Oil Recovery* Tulsa Oklahoma USA 20-3

[9] Li K 2008 A New Method for Calculating Two-Phase Relative Permeability from Resistivity Data in Porous Media *Transport in Porous Media* **74**(1) 21-33

[10] Li K 2010 Analytical derivation of Brooks–Corey type capillary pressure models using fractal geometry and evaluation of rock heterogeneity *Journal of Petroleum Science and Engineering* **73** 20-6

[11] Li K 2011 Interrelationship between Resistivity Index, Capillary Pressure and Relative Permeability *Transport in Porous Media* **88**(3) 385-98

[12] Katz A J, Thompson A H 1985 Fractal Sandstone Pores: Implications for Conductivity and Pore Formation *Physical Review Letters* **54** 1325-8

[13] Coates G, Xiao L, Prammer M 2000 NMR Logging Principles and Applications *Gulf Publishing Company* USA

[14] He Y, Mao Z, Xiao L 2005 A New Method to Obtain Capillary Pressure Curve Using NMR T_2 Distribution *Journal of Jilin University (Earth Science Edition)* **35**(2) 177-81 *(in Chinese)*

[15] Fatt I 1956 The network model of porous media III Dynamic properties of networks with tube radius distribution Petroleum Transactions AIME **207** 164-77

[16] King PR 1989 The use of renormalization for calculating effective permeability Transport in Porous Media **4** 37-58

[17] Li K, Firoozabadi A 2000 Phenomenological modeling of critical condensate saturation and relative permeabilities in gas condensate systems SPEJ 2000 **5**(2) 138-47

[18] Shimekit B, Mukhtar H, Maitra S 2010 Comparison of Predictive Models for Relative Permeability of CO2 in Matrimid-Carbon Molecular Sieve Mixed Matrix Membrane Journal of Applied Sciences **10**(12) 1204-11

Authors

Xinmin Ge, born in September, 1985, Jiangxi, China

Current position, grades: Postdoctoral fellow in China University of Petroleum (East China).
University studies: Master Degree of Geodetection and information technology at China University of Petroleum (East China), 2010. Doctor Degree of Geosciences and Geological engineering at China University of Petroleum (East China), 2013.
Scientific interests: well logging analysis, geophysics and formation evaluation.
Publications: 12 articles.

Yiren Fan, born in October, 1962, Fujian, China

Current position, grades: Professor in China University of Petroleum (East China).
University studies: Master Degree in well logging in University of Petroleum 1988. Doctor Degree in Geosciences at China University of Geosciences (Beijing) 2002.
Scientific interests: well logging information, petrophysics and formation evaluation.
Publications: 50 articles and 12 conferences paper.
Experience: 30 years experiences in well logging principle and application

Donghui Xing, born in May, 1989, Shandong, China

Current position, grades: Applied Physics at China University of Petroleum (East China).
University studies: Mechanical and Electronic Engineering at China University of Petroleum (East China).
Scientific interests: petrophysical experiments, logging principles and comprehensive interpretation.

Yongjun Xu, born in July, 1988, Henan, China

Current position, grades: Master degree student at China University of Petroleum (East China).
University studies: Resources Exploration Engineering at China University of Petroleum (East China).
Scientific interests: petrophysical experiments, logging principles and comprehensive interpretation.

440

Faults diagnosis of railway rolling bearing by using time-frequency feature parameters and genetic algorithm neural network

Dechen Yao[1, 2*], Limin Jia[1], Yong Qin[1], Jianwei Yang[2]

[1]*State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiao tong University, Beijing 100044, China*

[2]*School of Machine-electricity and Automobile Engineering, Beijing University of Civil Engineering Architecture, Beijing, 100044, China*

Abstract

This paper is focused on time-frequency feature parameters and genetic algorithm neural network techniques in fault diagnosis of railway rolling bearings. The time-frequency feature parameters for classification are extracted from vibration signals. However, the weak features of faults in rolling bearing are always immersed in noises of the environment, to solve this problem, Firstly, the wavelet analysis is used to filter and de-nosing and the time domain features are calculated. Secondly, the EMD (Empirical Mode Decomposition) method is used to decompose the signal into a number of intrinsic mode functions (IMFs), and then the IMF energy-torques could be calculated through the de-nosing signal. Finally, the genetic algorithm neural network is used for the classifications of the time-frequency feature parameters. The results of the time-frequency feature parameters and genetic algorithm neural network (GNN) show the effectiveness and the high recognition rate in classifying different faults of railway rolling bearing.

Keywords: time-frequency feature parameters wavelet analysis, EMD, IMF, genetic algorithm neural network, railway rolling bearing, fault diagnosis

1 Introduction

The rolling bearing is one of the important components in the railway sector, whose operation condition directly affects the whole performance of the railway, therefore, bearing fault diagnosis is important, and plays a key role in the reliable operation of the railway. Hence, bearing fault diagnosis and condition-based monitoring are of great significance [1].

There are extensive techniques, such as fast Fourier transform (FFT), envelope analysis, time-domain averaging and other techniques [2-6], have been fully developed and established for processing vibration signals to obtain diagnostic information about rolling bearing. However these methods only provide limited effectiveness for bearing diagnostics. So this paper deals with the application of time-frequency feature parameters and genetic algorithm neural network for fault diagnosis of railway rolling bearing. The vibration signals of bearing can be affected by noise possibly associated with external perturbations and the signal-to-noise ratio (SNR) is so low that feature extraction of signal components is very difficult, so the wavelet analysis is used to filter and de-nosing and then the time domain features are calculated. To obtain additional fault characteristic information, time and frequency domain features are extracted from the rolling bearing vibration signals. Moreover the EMD are applied to decomposing the vibration signals. Then energy features are extracted from the decomposed frequency-band signals of the intrinsic mode functions (IMFs) of

EMD. The time-frequency feature parameters are then put into the classifiers based on genetic algorithm neural network. The results show that this approach can performance effectiveness and high recognition rate in classifying different faults of railway rolling bearing.

The rest of the paper is organized as follows. In section 2, the wavelet analysis and time-domain feature extraction method will be introduced, while EMD algorithm and frequency-domain feature extraction method will be described in section 3. Section 4 describes the principle of GNN. In section 5, the results are provided to demonstrate the effectiveness of the proposed algorithm for faults diagnosis of railway rolling bearing. Finally, the conclusions are drawn in Section 6.

2 Wavelet analysis and time-domain feature extraction

2.1 WAVELET ANALYSIS

The wavelet transform is a linear transform which uses a series of oscillating functions with different frequencies as window functions $\psi_{a,b}(t)$ ("*a*" is the dilation parameter for changing the oscillating frequency and ''*b*'' is the translation parameter) which transforms the signal *x(t)*. At high frequencies, the wavelet reaches to a high time resolution with a low frequency resolution, whereas at low frequencies, a low time resolution with a high frequency resolution is achieved, which make them more suitable for non-stationary signal analysis. The basis function for the

---
Corresponding author e-mail: 290813991@qq.cn

wavelet transform is given in terms of translation parameter b and dilation parameter a with the following mother wavelet [7]:

$$\psi_{a,b} = \frac{1}{\sqrt{a}} \psi\left(\frac{t-b}{a}\right),\ a, b \in R,\ a \neq 0. \tag{1}$$

The wavelet transform, $W(a,b)$ of a time signal $x(t)$ is given by:

$$W(a,b) = \int_{-\infty}^{+\infty} x(t) \frac{1}{\sqrt{a}} \psi^*\left(\frac{t-b}{a}\right) dt, \tag{2}$$

where $\psi^*(t)$ is the complex conjugate of $\psi(t)$.

2.2 TIME-DOMAIN FEATURE EXTRACTION

Three time-domain features, namely, kurtosis factor, margin factor and pulse factor, are calculated. They are defined as follows [8].

Kurtosis factor:

$$K_v = \frac{\sum_{i=1}^{n} x_i^4}{n x_{rms}^4}. \tag{3}$$

Margin factor:

$$CL_f = \frac{x_{peak}}{x_r}. \tag{4}$$

Pulse factor:

$$I_f = \frac{x_{peak}}{|\overline{x}|}, \tag{5}$$

where x_i is i-th sampling point of the signal x; n is the number of points in the signal, x_{rms} is the root mean square of the signal, x_r is the square root of amplitude of the signal, and $|\overline{x}|$ is the absolute average of the signal.

3 EMD algorithm and time-frequency domain feature extraction

3.1 EMD ALGORITHM

EMD is a method of decomposing a non-linear and non-stationary signal into a series of zero-mean amplitude-modulation frequency-modulation (AM-FM) components that represent the characteristic time scale of the observation [9]. This is done by iteratively conducting a sifting process. The zero-mean AM-FM components are called Intrinsic Mode Functions (IMFs); they must satisfy the following requirements:

1) The number of extremes and the number of zero crossings must be equal or only a difference of one is allowed.

2) The mean between the local maxima envelope and the local minima envelope at any point must be equal to zero.

If the two previous conditions are not satisfied, i.e. the resulting signal $x_1(t)$ is not an IMF, then the previous steps are repeated. The procedure becomes iterative and it is called sifting process. Obviously, functions $s_{max}(t)$ and $s_{min}(t)$ are recomputed at each iteration and the newly evaluated $m(t)$ is subtracted from the obtained signal $x_1(t)$.

The sifting process runs until the extracted signal respects the two IMF conditions; then the function obtained represents the first intrinsic mode function $C_1(t)$ and it is subtracted from the initial signal [10]:

$$r_1(t) = x(t) - C_1(t), \tag{6}$$

where $r_1(t)$ is the residual signal. This signal represents the input for the second IMF calculation by means of the sifting process. The EMD algorithm, applied to the original signal $x(t)$, stops when the residual signal $r_N(t)$ is a constant or monotonic function, after the extraction of the N-th intrinsic mode function. The stop criteria can be expressed in terms of a standard deviation and number of extremes:

$$\sigma(r_N(t)) < \sigma_{stop},\ n_{max}, n_{min} = 0. \tag{7}$$

The decomposition stops when both conditions are satisfied.

Since the decomposition objective is the identification of all the original signal structures, the original data can be reconstructed by adding the extracted IMFs to the residual signal.

3.2 TIME-FREQUENCY DOMAIN FEATURE EXTRACTION

Five time-frequency domain feature parameters are extracted from the vibration signal of rolling bearing in this work. The steps of time-frequency domain feature extraction are as follows:

1) The wavelet analysis is adopted to decompose the vibration signal of railway rolling bearing.

2) The de-noised vibration signals are decomposed into some IMFs by using the EMD method, the first n IMFs $c_i(t),\ i = 1, 2, 3, ..., n$, which include the most dominant fault energy are chosen to extract the feature.

3) Calculate the energy-torque of every small time block

The Equation to calculate IMF energy-torque is:

$$E_i = \int_{-\infty}^{+\infty} t |c_i(t)|^2 dt. \tag{8}$$

For discrete signals, the formula to calculate energy-torque is:

$$E_i = \sum_{k=1}^{m} (k \cdot \Delta t) |c_i(k \cdot \Delta t)|^2, \tag{9}$$

where m is the total number of sampling points, k is the sampling points, Δt is the sampling period. Calculating

the energy-torque $E_1, E_2, \cdots$ for each chosen IMF based on the Equation (9).

4) Constructing the feature vector T in the elements of the energy-torque.

$$T = [E_1 \quad E_2 \quad \cdots \quad E_n]. \tag{10}$$

When the energy-torque is a larger numerical, normalizing T and get the normalized feature vector $T^{'}$. Among them:

$$E = \left(\sum_{i=1}^{n} |E_i|^2 \right)^{\frac{1}{2}}. \tag{11}$$

The Equation to calculate IMF energy-torque is [11]:

$$E_i = \int_{-\infty}^{+\infty} |c_i(t)|^2 \, dt. \tag{12}$$

4 Combination of neural network and Genetic algorithm

Genetic algorithms are the algorithms for optimization and learning based freely on several features of biological evolution. Genetic algorithm has been proved to be capable of finding global optima in complex problems by exploring virtually all regions of the state space and exploiting promising areas through mutation, crossover and selection operations applied to individuals in the populations. It applies selection, crossover and mutation operators to construct fitter solutions. A genetic algorithm process the populations of chromosomes by replacing unsuitable candidates according to the fitness function.
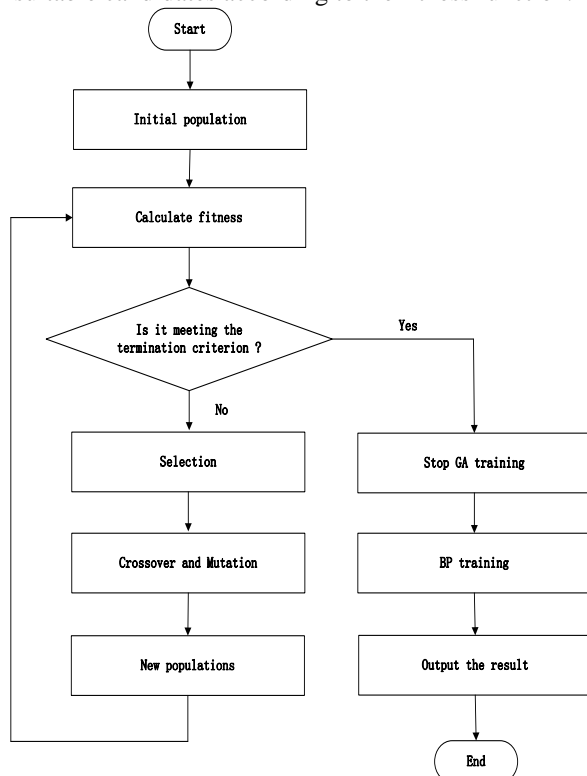
FIGURE 1 Framework of training a neural network using GA

The conventional BP algorithm is prone to falling into the local minimum point. To overcome this problem, this study attempts to combine genetic algorithms, avoiding local minimums.

The steps of combination of neural network and Genetic algorithm are shown in Figure 1 [12].

5 Experimental results

5.1 EXPERIMENTAL SETUP

The effectiveness of the proposed algorithm is further tested using experimental railway rolling bearing vibration signals. The railway rolling bearings are: normal bearing, inner ring fault bearing and rolling fault bearing.

5.2 THE PROCESS OF TIME-FREQUENCY DOMAIN FEATURE EXTRACTION

Here the wavelet analysis is employed to obtain a filtered signal with clear fault-revealing trend lines in its time-frequency representation, the time-domain feature parameters are calculated. The original vibration signals and the de-nosing vibration signals are shown in Figure 2 and Figure 3. The de-nosing vibration signals of railway rolling bearings are decomposed by EMD and the result is shown in Figure 4, and then the time-frequency feature parameters are calculated. All time-frequency feature parameters are divided into two groups, namely, the training group and the testing group. The training group is in Table 1. The testing group is in Table 2. The GNN is used to diagnose the faults; the inputs and outputs of the GNN are all memberships concretely, the inputs are the memberships of the kurtosis factor, margin factor, pulse factor and IMFs energy-torque on each characteristic domain, and the outputs are memberships of each fault.
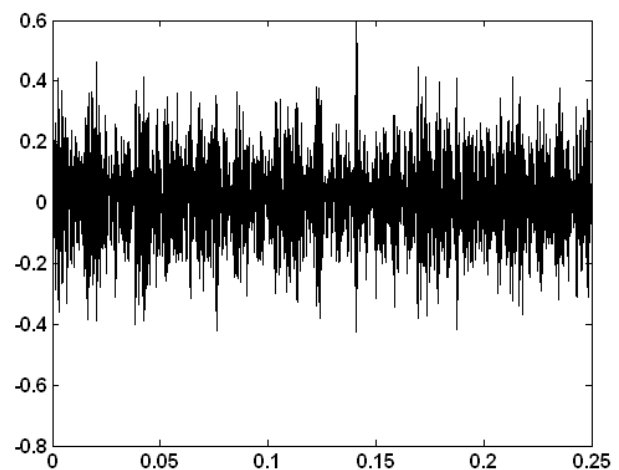
FIGURE 2 The time domain of outer ring fault signal

Yao Dechen, Jia Limin, Qin Yong, Yang Jianwei
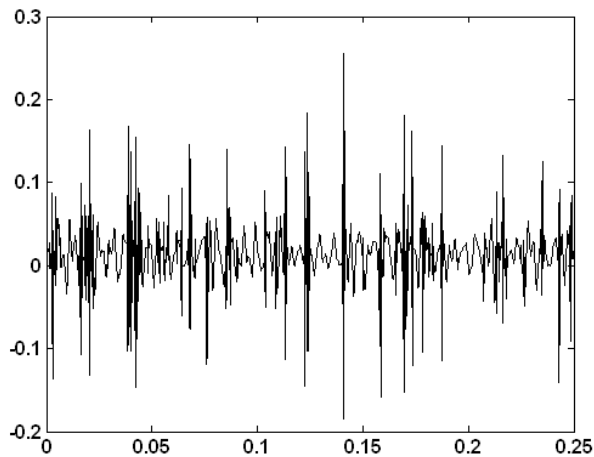
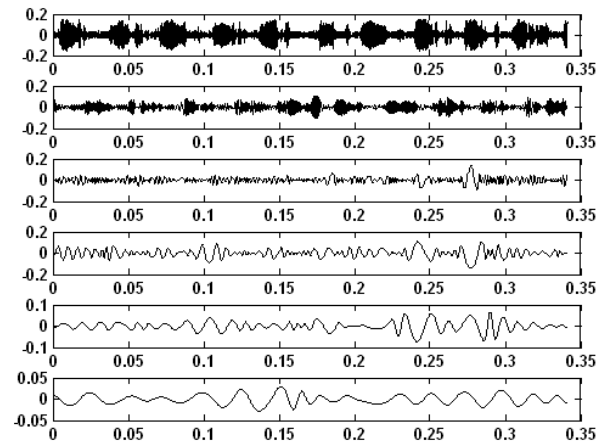FIGURE 3 The de-nosing signal of outer ring fault

FIGURE 4 Decomposition of the outer ring fault signal into intrinsic mode functions

TABLE 1 Sample data of bearing operation

Kurtosis factor	Margin factor	Pulse factor	E_1	E_2	E_3	E_4	E_5	Fault status	Fault vector
3.612	5.979	4.979	0.924	0.147	0.193	0.283	0.075	normal signal	(1 0 0)
3.801	7.919	6.548	0.962	0.162	0.106	0.179	0.059	normal signal	(1 0 0)
3.704	6.015	5.044	0.970	0.120	0.174	0.114	0.032	normal signal	(1 0 0)
3.883	6.687	5.586	0.278	0.668	0.643	0.244	0.047	rolling fault signal	(0 1 0)
3.420	6.091	5.106	0.235	0.734	0.602	0.197	0.047	rolling fault signal	(0 1 0)
3.968	5.673	4.812	0.168	0.650	0.673	0.304	0.037	rolling fault signal	(0 1 0)
23.789	29.190	19.649	0.881	0.424	0.096	0.125	0.135	inner ring fault signal	(0 0 1)
25.857	32.435	21.979	0.871	0.444	0.168	0.096	0.069	inner ring fault signal	(0 0 1)
25.098	34.427	22.848	0.870	0.439	0.128	0.174	0.038	inner ring fault signal	(0 0 1)

TABLE 2 Testing data

Kurtosis factor	Margin factor	Pulse factor	E_1	E_2	E_3	E_4	E_5	Fault status	Fault vector
3.477	7.284	6.088	0.953	0.121	0.144	0.224	0.068	normal signal	(1 0 0)
3.754	6.517	5.504	0.33	0.67	0.587	0.28	0.036	rolling fault signal	(0 1 0)
31.027	33.756	22.602	0.888	0.407	0.107	0.087	0.157	inner ring fault signal	(0 0 1)

5.3 THE PROCESS OF GNN

The genetic algorithm neural network is used for fault diagnosis consist of three layers: an input layer, a hidden layer and an output layer. Before training, those neural networks were optimized, and we obtained the genetic algorithm neural network architecture is 8-6-3.

The testing results of the genetic algorithm neural network are in Table 3. As shown in this table, the classification of genetic algorithm neural network are all close to the corresponding ideal outputs of the examination sample, it is found to be satisfactory and we think that this system can be used in fault diagnosis studies in the future after it is developed.

TABLE 3 Testing results of GNN

Fault status	Ideal outputs	Actual outputs	Testing results
normal signal	(1 0 0)	(1.0387 -0.1913 0.0006)	normal signal
outer ring fault signal	(0 1 0)	(-0.1031 1.4504 0.0011)	outer ring fault signal
inner ring fault signal	(0 0 1)	(0.0215 -0.1257 1.0028)	inner ring fault signal

6 Conclusions

In this paper we develop faults diagnosis of railway rolling bearing by using time-frequency feature parameters and genetic algorithm neural network. The main contributions of our work are as follows:

1) The wavelet analysis is adopted to decompose the vibration signal of railway rolling bearing and the time domain features are calculated.

2) The EMD method is used to decompose the signal into a number of IMFs, and then the IMF energy-torques are calculated.

3) The genetic algorithm neural network is used for the classifications of the time-frequency feature parameters.

The experimental results show that this method can be effectively used in railway rolling bearing faults. So this proposed method gives an effective and feasible approach for the fault diagnosis of the rotary machine.

References

[1] Yang J, Yao D, Cai G 2010 A method of bearing fault feature extraction based on improved wavelet packet and hilbert analysis *International Journal of Digital Content Technology and its Applications* **4**(4) 127-39 (*in Chinese*)

[2] Yin Y, Yang J, Cai G 2010 Fault diagnosis of rolling bearing based on wavelet packet and fourier analysis *Computational Aspects of Social Networks* 703-6 (*in Chinese*)

[3] Shi D F, Wang W J, Qu L S 2004 Defect detection for bearings using envelope spectra of wavelet transform *Journal of Vibration and Acoustics* **4** 567–73

[4] Nikolaou NG, Antoniadis IA 2002 Rolling element bearing fault diagnosis using wavelet packets *NDT & International* 35 197–205

[5] Lei H, Kemao Q, Pan B, Asundi A K 2010 Comparison of Fourier transform, windowed Fourier transform, and wavelet transform methods for phase extraction from a single fringe pattern in fringe projection. –profilometry **48** 141-8 (*in Chinese*)

[6] Prabhakar A S ,Sekhar A R, Mohanty 2001 Detection and monitoring of cracks in a rotor-bearing system using wavelet transform *Mechanical System and Signal Processing* **2** 447–50

[7] Jafarizadeha M A, Hassannejadb R, Ettefagh M M 2008 Asynchronous input gear damage diagnosis using time averaging and wavelet filtering *Mechanical Systems and Signal Processing* **1** 172-201

[8] Lei Y, Zuo M J, He Z 2010 A multidimensional hybrid intelligent method for gear fault diagnosis *Expert Systems with Applications* **37** 1419–30 (*in Chinese*)

[9] Zhang J-w, Zhu N-h, Yang L 2007 A Fault diagnosis approach for broken rotor bars based on emd and envelope analysis *China Univ Mining & Technol* **2** 0205–9 (*in Chinese*)

[10] Ricci R, Pennacchi P 2011 Diagnostics of gear faults based on EMD and automatic selection of intrinsic mode functions *Mechanical Systems and Signal Processing* **25** 821–38

[11] Bin G F, Gao J J, Li X J 2012 Early fault diagnosis of rotating machinery based on wavelet packets-Empirical mode decomposition feature extraction and neural network *Mechanical Systems and Signal Processing* **27** 696-711 (*in Chinese*)

[12] Irani R, Nasimi R 2011 Evolving neural network using real coded genetic algorithm for permeability estimation of the reservoir *Expert Systems with Applications* **38** 9862–6

Authors

Dechen Yao, Beijing, China

Current position: PhD candidate in Beijing Jiao Tong University, China.
University studies: ME degree in Taiyuan university of science and technology, China.
Scientific interest: rotor dynamics, fault diagnosis.

Limin Jia, Beijing, China

Current position: chief professor at Beijing Jiao Tong University, China.
University studies: Beijing Jiao Tong University, in the State Key Laboratory of Rail Traffic Control and Safety, China.
Scientific interest: intelligent automation and rail intelligent transportation system, traffic safety engineering.

Yong Qin, Beijing, China

Current position, grades: professor in Beijing Jiao Tong University, China.
University studies: Beijing Jiao Tong University, in the State Key Laboratory of Rail Traffic Control and Safety, China.
Scientific interest: traffic and transportation information engineering and security.

Jianwei Yang, Beijing, China

Current position, grades: professor in Beijing University of Civil Engineering Architecture, China.
University studies: Beijing University of Civil Engineering Architecture, in the School of Machine-electricity and Automobile Engineering, China.
Scientific interest: fault diagnosis, reliability assessment.

Design and implementation of acoustic target recognition system based on TMS320F2812

Zhigang Ma[1, 2], Wenyi Liu[2*]

[1]*Key Laboratory of Instrumentation Science & Dynamic Measurement(North University of China), Ministry of Education, Science and Technology on Electronic Test & Measurement Laboratory, Taiyuan, 030051, China*

[2]*College of Information Science and Engineering, Shanxi Agricultural University, Taigu, 030801, China*

Abstract

In this paper, TMS320F2812 is adopted as the core component and the hardware circuit system is designed for acoustic target recognition, which includes signal conditioning circuit, A/D acquisition circuit, memory expanding circuit, power management circuit and data communication circuit, etc. The recognition algorithm was transplanted to the acoustic recognition system based on DSP, which enhanced the running rate in DSP. This system can accomplish acoustic signal sampling, pretreatment, feature extraction and recognition. On the basis of the simulation and real tests, it is proved that the acoustic target recognition system proposed in this paper is stable and reliable, which can satisfy the practical requirements.

Keywords: DSP, TMS320F2812, algorithm, acoustic target recognition

1 Introduction

Acoustic target recognition (ATR) is the key technology in battlefield detection system, which attracted people's great concern at home and abroad in recent decades [1-4]. Chinese research institutes and universities have conducted research on acoustic target recognition technology sequentially; the in-depth theoretical research was carried in acoustic sensor technology, acoustic feature analysis of the target signal, signal processing, artificial intelligence and other aspects, and a series of achievements also be made by scientific research. However, due to the lacking research on feature extraction of non-stationary target signal, target recognition method and system synthesis with non-determinacy characteristics, identify systems fusion, etc., there are many theoretical and technical issues need to be addressed in acoustic target recognition technology in china. Especially there are many problems in recognition accuracy, real-time, miniaturization, and other aspects. Therefore, ATR technology is still far from reaching maturity [5, 6].

Generally speaking, the design of target recognition system mainly includes the following three aspects:

1) Self-perception. Perceiving external information by acoustic, magnetic, vibration, infrared sensors;

2) Real-time data processing and transmission. Processing perceived information timely, so that each attack department or combat unit will quickly react and make decision;

3) Microminiaturization. Small size can be distributed more conveniently or concealed easily [4, 7, 8].

Acoustic detection and recognition technology has a long history [9, 10]. Early in World War II, it had been applied in the battlefield for detecting, identifying and localizing the enemy military facilities and armed personnel, so ART played an important role at that time. Compared with other methods, acoustic detection technology has the following advantages:

1) Passive working style is difficult to be detected or located by enemy surveillance equipment;

2) Unrestricted by visual field or obstacles;

3) The acoustic detection equipment has relatively simple structure, low cost, small size, light weight, high mobility, and unrestricted from climate and other natural conditions, so it can work all-weather;

4) Pre-warning network can be composed based on acoustic detection technology, which will improving detection accuracy and reducing false alarm rate [3, 11, 12].

2 Hardware circuit design

2.1 SYSTEM STRUCTURE

The hardware circuit of the acoustic target recognition system proposed in this paper includes: signal conditioning circuit, analogue to digital (A/D) acquisition circuit, memory expansion circuit, power management circuit, data communications circuit and other interface circuit. The overall diagram of system structure is shown in Figure 1.

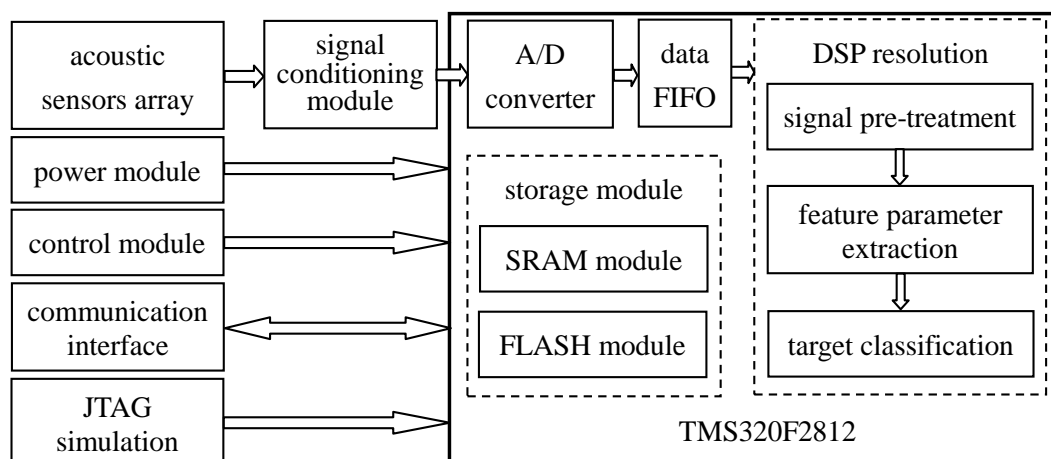[*]*Corresponding author* e-mail: liuwenyi@nuc.edu.cn

FIGURE 1 The overall diagram of system structure

2.2 TMS320F2812 DSP

ATR system requires "high real-time processing" capability, which is: the system must complete all operations and processing for current data before the next sampling time, otherwise it will affect the subsequent processing of sampling data. When data sampling rate is high or the amount of computation is large, the general processor is difficult to achieve "real-time processing" requirement. Digital signal processing (DSP) chip was developed rapidly in recent decades. DSP has a series of advantages, so it can be used in the high-speed real-time processing system [13]. TMS320F2812 is developed by Texas Instruments (U.S.A.), which is a low-power, 32-bit fixed-point DSP. TMS320F2812 has a micro-controller structure, firmware and tool equipment, reduced instruction set (RIS) feature, rich instruction set, multi-pipeline operation, large storage space and faster computing speed [1, 13, 14].

2.3 ACOUSTIC SENSOR

Acoustic sensor is also named as "microphone", and there are several types, such as: piezoelectric type, moving-coil type, capacity type, electrets type, and so on. Among them, because of the poor frequency response and stability, the moving-coil and piezoelectric types had been rarely used. Generally, electrets microphone is most suitable for acoustic target recognition system. According to practical demand, PF0-T6027 is used in this paper, which is a omni-directional electrets microphone.

2.4 SIGNAL ACQUISITION & CONDITIONING CIRCUIT

Acoustic signal acquisition is the first link in target recognition system. An A/D converter (12-bit) in

TMS320F2812 is adopted, which is simple and flexible, and the hardware design can be simplified to some extent. Because of the internal A/D converter module in DSP, it has high anti-interference ability, and the performance was significantly better than other external A/D chip.

During the sampling, quantization, due to the frequency aliasing, the sampling signal may be distorted owing to systematic errors and random errors. In order to ensuring the sampling accuracy, the analogue signal must be properly conditioned before sampling. Generally, the analogue signal is usually filtered and amplified firstly, so as to fully utilize the dynamic range of A/D converter module. Furthermore, the input voltage range of internal A/D is 0~3V, therefore the sampling signal of the microphone must also be limited. Signal conditioning circuit is shown in Figure 2.

Mentioned above, amplitude range of conditioned acoustic signal is 0 ~ 3V, which is connected to the A/D input channel ADCH in TMS320F2812. A/D module is comprised by: front-end analogue multiplexer switches, a sampling and holding circuit, a conversion circuit, a voltage stabilizing circuit, and so on. The digital circuit includes: programmable conversion sequence generation circuit, result registers, analogue power connector, other peripheral interface circuit, and other module interface circuit.

Two 8-channel modules can choose any input from the eight input channels. In the cascade operation mode, the automatic sorting will become a single 16-channel sequencer. In each sequencer, once the conversion is complete, the conversion result will be stored in the corresponding ADCRESULT register. The automatic sorting allows system converts the same channel multiple times in order to achieve over-sampling. Compared with the traditional single-sampling mode, this conversion mode is in favour of improving the data sampling resolution.

FIGURE 2 Signal conditioning circuit

2.5 SERIAL COMMUNICATION CIRCUIT

UART is a standard universal asynchronous receiver/transmitter. The acquisition data can be outputted via UART, and the data communication between DSP and PC will be achieved. Subsequently, the data can be processed by the data analysis software in PC. As a UART interface circuit between DSP and PC, the transmission and receiving of serial data between DSP and PC can be completed only need a few instructions in the program, as well as the rest of the work will be done in UART automatically. All pins' output is TTL level in TMS320F2812, while the PC's serial port follows RS232 standard. As a result, it requires a level conversion interface between PC's serial port and the DSP [1]. In this paper, MAX3160 transceiver (Maxim Company) is selected, which is a programmable multi-protocol transceiver following RS-232, RS-422, RS-485 and other data transmission mode. In addition, +3 ~ +5.5 V power supply will satisfy the RS-232 and RS-485/422 protocol standards. Communication interface circuit between PC and TMS320F2812 is shown in Figure 3.

FIGURE 3 Communication interface circuit between PC and TMS320F2812

3 Embedded system software design

There are some specific requirements when the application program runs in TMS320F2812's internal Flash. Some problem may be concealed when the application program is debugged in RAM by means of CCS simulation software, but the actual use of DSP requires paying more attention to these issues. The following text will discuss how to configure the application software so that it can run properly in Flash [1, 13].

3.1 LINKING PROGRAM BLOCK

TMS320F2812 can be used in embedded systems independently, its on-chip Flash can be saved using an external non-volatile memory, and unnecessary initialization data can be stored in other resources such as SRAM. When the program block is being linked, a user link command file is must be created. In this file, a number of memory cells are defined, and the location area of the block is specified. The format of this file is similar mostly with the ".cmd" file running in RAM, while the main difference between them is where the program block is linked to.

3.2 COPY THE FUNCTION PROGRAM BLOCK FROM FLASH TO RAM

PIE module is used for managing the interrupt requests in TMS320F2812. When system is powered on, all interrupt vectors must have been stored in non-volatile memory Flash in advance. Meanwhile, in the initialization code, the interrupt vectors must be copied to the PIE RAM. PIE RAM is to be located at special RAM area in data space, and the starting address is 0X000D00. Copying PIE vector table needs to be accomplished in a file called as "DSP28_PieVect.c".

3.3 INITIALIZATION THE CONTROL REGISTERS

The control registers in Flash include: FOPT, FPWR, FSTDBYWAIT, FACTIVEWAIT, FBANKWAIT, FOTPWAIT, and so on. Initialization of these registers cannot be completed in Flash, so they must be copied form Flash to RAM and then executed initialization.

Since the Flash control registers are protected by Code Security Module (CSM), if DSP is safe, the Flash register initialization code is must be executed in secured RAM (usually is L0 or L1 RAM). Otherwise the

initialization function InitFlash() will not be able to access the Flash register.

3.4 EXECUTE TARGET RECOGNITION ALGORITHM FUNCTION

In this paper Target recognition algorithm has been designed and simulated in PC, the main functions include: acoustic signal pre-treatment, feature extraction and classification. In this system, the corresponding algorithm needs to be transplanted to the DSP in order to achieve the classification and recognition capabilities.

The clock frequency in TMS320F2812 is 150MHz, and the Flash memory in which can only provide about 120MIPS (Millions of Instruction per Second) of processing speed. As for the on-chip RAM, its performance can reach 150MIPS. So in some cases, some time-sensitive or large computational subprogram should execute in RAM. However, in a separate embedded system, all code must be stored in the non-volatile memory. As a result, the function executed in RAM must be set load and run address. In addition, the copy process from Flash to RAM must be carried out after power on.

3.5 EXECUTE CODE IN FLASH

TMS320F2812 contains a ROM guide loader, which can transfer the code to Flash and execute it after DSP is reset. When the guide mode configures the pin as "Skip to Flash", ROM guide loader will execute the instruction at the specified location. So the user must set up a jump instruction here, which will point to the start address of the user code.

Generally speaking, the jump instruction will jump to the start position of initialization subroutine in "rts2800_ml.lib" supported by C compiler. The entry symbol of this subroutine is "_c_init00", and only after running the routine, the remaining code can be executed.

4 Simulation experiments

4.1 SIMULATION EXPERIMENTS IN CCS

In the simulation experiments, the target recognition system board and DSP hardware emulator is connected together firstly. Secondly, the program is downloaded to the TMS320F2812 chip through CCS development tools. Finally, the pre-treatment, classification and recognition of collected acoustic signals are simulated. The recognition result is displayed on the lower screen of CCS interface, which is shown in Figure 4.

FIGURE 4 Online simulation interface and results in CCS

4.2 SIMULATION EXPERIMENTS RESULTS

In the online simulation in CCS, the nearest neighbour classifying method is applied for target classification and recognition, and the results are shown in Table 1. It is clear that the personnel's acoustic and the vehicles' acoustic can be distinguished effectively in the using of the acoustic target recognition system proposed in this paper, and the average recognition rate is over 90%.

TABLE 1 Simulation experiments results

Category	Personnel		Vehicle	
Sample sum	400		400	
Recognition Results	Correct	Fail	Correct	Fail
	363	37	381	19
Recognition Rate (%)	90.75		95.25	
Average recognition Rate (%)	93			

5 Conclusions

In recent decades DSP technology has been widely applied in various fields, such as: military, industrial inspection and control, smart instrumentation, communications, and medical treatment, etc.

TMS320F2812 chip is specially designed for industrial applications, new-generation 32-bit fixed-point DSP processor, which has high integration, high speed, high precision and low cost advantages, also has a lot of peripherals, interfaces and modules [1, 13, 14]. In this paper, the hardware circuit based on TMS320F2812 is designed; the algorithm of acquisition, processing, feature extraction and recognition is accomplished and simulated; algorithms in each module are transplanted to DSP; part of the algorithms are optimized, and which improved the execution speed in the DSP. A higher recognition rate is achieved in the distinguishing between personnel's acoustic signals and vehicles' acoustic signals. In addition, the system is also be used in numerous practical applications and made several good recognition results.

Acknowledgments

References

[1] Yang C 2008 Research and realization of acoustic target recognition based on DSP *Master's Degree Thesis North University of China* (*in Chinese*)
[2] Wu H, Mel S 1999 *IEEE Transaction on Instrumentation and Measurement* **4**(5) 1005-9 (*in Chinese*)
[3] Lv G, Xu X, Zhao R 2001 Passive sound recognition technology for the targets in battlefield *Journal of Detection & Control* **23**(4) 22-5 (*in Chinese*)
[4] Tang D 2012 A novel multi-target recognition method based on DSP *Telecommunication Engineering* **52**(5) 699-703 (*in Chinese*)
[5] Liu X, Y Wang, Liu Y 2010 Study on a real-time infrared marker

tracking method based on DSP *Chinese Journal of Scientific Instrument* **31**(12) 2654-60 (*in Chinese*)
[6] Zhang D, Zhou Y, Gao P, Huang R 2010 Design of Multi-DSP parallel processing sound detection system *Application of Electronic Technique* **36**(12) 32-5 (*in Chinese*)
[7] Yang H, Jing X, Zeng Z 2013 Study of speech recognition system base on DSP *Computer Measurement & Control* **21**(1) 210-2 (*in Chinese*)
[8] Yang G, Zhao H, Li S, Li L 2010 Time delay estimation of passive acoustic detection based on DSP *Journal of Projectiles, Rockets, Missiles and Guidance* **30**(3) 216-8 (*in Chinese*)

[9] Wang W, Wang G 2006 Study for power of high speed digital signal processing platform based on DSP+ FPGA *Modern Electronic Technique* **23**(6) 44-6 (*in Chinese*)

[10] Mo L, Wang X, Ni W 2011 Design of data acquisition and display system based on DSP *Instrument Technique and Sensor* **51**(7) 41-3 (*in Chinese*)

[11] Li L, Jin H, Chen Z, Yu F 2008 High resolution image data acquisition system based on FPGA and DSP *Journal of Data Acquisition & Processing* **23**(1) 121-6 (*in Chinese*)

[12] Liu X, Chen X, Yu D 2010 Design of woving target recognition and tracking system based on DSP *Video Engineering* **34**(11) 107-10 (*in Chinese*)

[13] Zhang J, Xu F, Yan Z, Zhang L 2009 Design and implementation of data acquisition system base on TMS320F2812 *Petroleum Instruments* **21**(5) 80-2 (*in Chinese*)

[14] Cao J, Chen X, Wu C 2005 Serial communication between TMS320F2812 and PC *Electronic Design Engineering* **5**(8) 38-40 (*in Chinese*)

Authors

Zhigang Ma, born on March 1, 1982, Wenshui, China

Current position, grades: PhD candidate of Instrument Science and technology, lecturer at School of Information Science and Engineering in Shanxi Agricultural University.
University studies: Computer Science and Technology in North China Institute of Technology.
Scientific interest: computer application, electronic information science & technology.
Publications: 11 papers.

Wenyi Liu, born on January 1, 1970, Lanxian, China

Current position, grades: Doctor of Instrument Science and technology, professor and doctoral supervisor at School of Instruments and Electronic in North University of China.
University studies: Measurement and Control Technology & Instrument in North China Institute of Technology.
Scientific interest: wireless sensors network, micro & nano sensor and testing technology.
Publications: 5 patents, 24 papers.

Mesoscopic simulation studies on the aggregation behaviour of glycyrrhizin micelles for drug solubilization

Yuguang Wang[1], Xingxing Dai[2], Xinyuan Shi[2], Yanjiang Qiao[2*]

[1]*Third Affiliated Hospital of Beijing University of Chinese Medicine, Beijing (China)*

[2]*Beijing University of Chinese Medicine, Beijing (China)*

Abstract

Glycyrrhizin is a kind of natural surfactant with the micelle structure for drug solubilization and low toxicity in the human body. In this study, dissipative particle dynamics (DPD) and mesoscopic dynamics (MesoDyn) were carried out to elaborate the aggregation behaviour of glycyrrhizin micelles. Baicalin were selected as the poorly water-soluble drugs. It has been observed from DPD that glycyrrhizin molecules formed core/shell structured spherical, cylindrical and lamellar aggregates with the increase of concentration. Baicalin molecules are solubilized in the hydrophobic core. Glycyrrhizin molecules were easier to form micelles with the addition of drugs. Mesoscopic simulations based on experimental data provided more detailed information for the investigation of solubilization.

Keywords: mesoscopic simulation, dissipative particle dynamics, glycyrrhizin micelle, drug solubilization, aggregation behaviour

1 Introduction

Poor solubility ingredients carry a high risk of failure during the development since insufficient solubility may affect both the bioavailability and clinical effect, and finally affect the develop ability of the compound [1]. Glycyrrhizin, the main constitute of Glycyrrhiza uralensisб which is frequently used in Chinese prescriptions, has been reported to be a kind of natural surfactant with good properties. It has attracted a lot of interest for it can aggregate in water with the core/shell structure for encapsulating the poorly water-soluble drug and has low toxicity to human body [2-5]. It is necessary to investigate the solubilization mechanism of glycyrrhizin for its reasonable development and drug application. However, it is difficult to study the formation process of micelle by common experimental methods, for the solubilization process takes only milliseconds to reach equilibrium. As a result, the underlying mechanism still needs to be further studied, even though a number of experiments have been carried out to investigate the solubilization of saponins.

Mesoscopic simulation provides a powerful tool for the structure study and the performance prediction of surfactants [6-18]. In this work, DPD and MesoDyn mesoscopic simulation methods coupled with experimental studies, were employed to investigate the aggregation behaviour of glycyrrhizin micelles with baicalin, which is the poorly water-soluble drugs, to illustrate the solubilization mechanism of glycyrrhizin.

2 Material and experimental methods

2.1 MATERIALS

Glycyrrhizin (95%), and baicalin (98%) were purchased from Nanjing ZeLang Medical Technology Co., LTD. The drug-loaded glycyrrhizin solution was prepared by dissolting 5mg baicalin in 0.1% glycyrrhizin solution at 25±1°C for 24h, then filtered with 0.45μm filter membrane.

2.2 TEM EXPERIMENTS

The aggregate morphologies of 0.1% glycyrrhizin solution, baicalin solution and drug-loaded glycyrrhizin solution were observed on a transmission electron microscope (TEM, JEOL JEM-1230 microscope) operated at an acceleration voltage of 80 kV. The solution samples were deposited onto copper grids that had been percolated with a thin film of formvar and then coated with a thin carbon film. The liquid was blotted off with filter paper after a few minutes, and the grids were air-dried.

3 Simulation methods

3.1 DPD THEORY

Dissipative particle dynamics (DPD) is a mesoscopic simulation technique, which is suitable to study the collective behaviour of complex fluids. A DPD bead represents a small region of fluid matter and its motion is assumed to be governed by Newton's laws [19]:

---

[*]*Corresponding author* e-mail: yjqiao@263.com

$$\frac{d\mathbf{r}_i}{dt} = \mathbf{v}_i \ , \ m_i \frac{d\mathbf{v}_i}{dt} = \mathbf{f}_i \ , \tag{1}$$

where, $\mathbf{r}_i$, $\mathbf{v}_i$, m_i and $\mathbf{f}_i$ denote the position vector, velocity, mass, and total force acting on particle i, respectively.

The force $\mathbf{f}_i$ between each pair of beads contains three parts: a harmonic conservative interaction force $\left(\mathbf{F}_{ij}^C\right)$, a dissipative force $\left(\mathbf{F}_{ij}^D\right)$ and a random force $\left(\mathbf{F}_{ij}^R\right)$. The expression is given as $\mathbf{f}_i = \sum_{j \neq i}\left(\mathbf{F}_{ij}^C + \mathbf{F}_{ij}^D + \mathbf{F}_{ij}^R\right)$. All forces are short-range with a fixed cutoff radius r_c, which is usually chosen as the reduced unit of length $r_c \equiv 1$.

3.2 MESODYN THEORY

MesoDyn is based on a dynamic variant of mean-field density functional theory, which states that there is a one-to-one mapping between the distribution functions of the system, the densities, and an external potential field. The model used in the MesoDyn project consists of a variety of beads whose interactions are described by harmonic oscillator potentials for the intramolecular interactions. Each bead represents a certain group of atoms.

3.3 MODELS AND INPUT PARAMETERS

The coarse-grained models are shown in Figure 1. The saccharide and aglycone of glycyrrhizin were represented by two types of beads named A (red) and B (green). Baicalinand water were coarse grained by beads BA (yellow) and W (blue), respectively. To show the aggregate clearly, water molecules were not displayed.
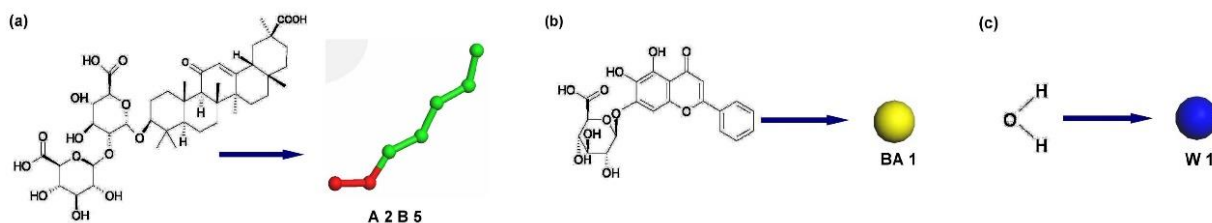
FIGURE 1 Coarse-grained models: (a) glycyrrhizin, (b) baicalin, (c) water

In the DPD simulations, a cubic simulation box with periodic boundary condition is applied in all three directions. A cubic box of 10×10×10 is sufficient to avoid the finite size effects. The grid spacing of 1.0nm is taken and the density is set to be 3.0. The integration time step of 0.05 and the simulation steps of 20000 are used in order to get thermodynamic equilibrium.

In MesoDyn simulations, a cubic simulation box with periodic boundary condition is applied. The dimensions of the simulation lattice are 32×32×32 and the grid spacing of 1.0nm is taken. To ensure isotropy of all grid-restricted operators, the bond length is set to be 1.1543nm. Bead diffusion coefficient is set to be $1.0 \times 10\text{-}7\text{cm}^2 \cdot \text{s}^{-1}$ in order to ensure a stable numerical algorithm. The noise parameter is 75.002, the temperature is 298K and the time step is 50.0ns. Number of steps is 50000 (2.5ms) to ensure a complete process of micelle formation.

All the simulations are carried out using the DPD or MesoDyn module in the commercial software Materials Studio 4.1 from Accelrys Inc.

4 Results and discussion

4.1 AGGREGATION BEHAVIOR OF GLYCYRRHIZIN IN AQUEOUS SOLUTION

The DPD simulation results showed in Figure 2 indicated that glycyrrhizin molecules formed spherical aggregates at low concentration (0.9vol%-19vol%).
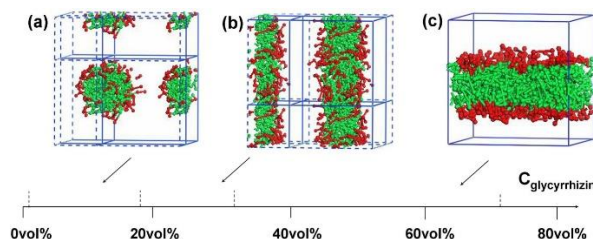
FIGURE 2 Aggregation behaviours of glycyrrhizin at different concentrations: a) sphere, b) cylinder and c) lamella

With the increase of concentration, small aggregates crashed and combined, then spherical aggregates with increased diameter formed. When the concentration of glycyrrhizin was between 20vol% and 32vol%, spherical aggregates disappeared and cylindrical aggregates began to form, whose radial diameter stretched with the increase of concentration. When the concentration of glycyrrhizin was from 33vol% to 72vol%, lamellar aggregates were observed and the lamella became thicker at higher concentration. Moreover, the aggregates did not change as the simulation time prolonged, indicating the above microstructures were the lease in local free energy minima [20].

The spherical glycyrrhizin aggregates observed from DPD (Figure 3a) were further obtained by TEM experiment (Figure 3b). For further investigation the microstructure of glycyrrhizin aggregates, MesoDyn simulation was carried out to analyse the section view of

glycyrrhizin aggregates in concentration of 1.3vol%. Density profiles of beads A and B on the section passing through the centre of the aggregate were shown as Figure 3c. The hydrophilic A formed a peak existed in the position of the shell, while the hydrophobic B was more concentrated in the centre of the aggregate, producing a

core/shell structured micelle. The sections formed by beads A and beads B were given as Figure 3d and e. The hydrophobic core formed for entrapping the poorly water-soluble drugs due to the hydrophobic interactions, while the hydrophilic shell provided a stable interface between the core and water.
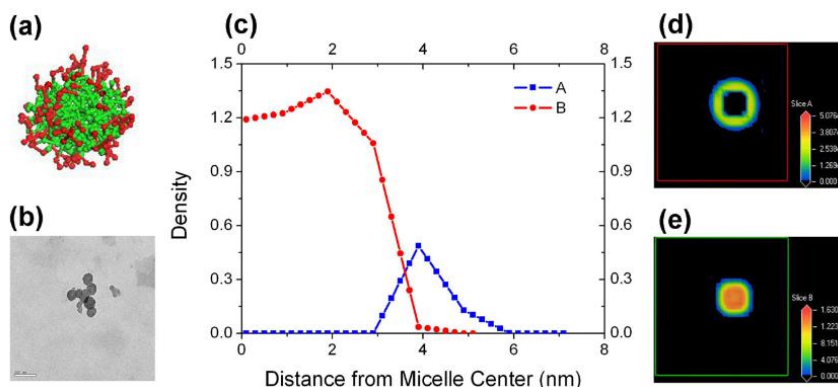
FIGURE 3 Morphological analysis of spherical aggregates of glycyrrhizin:
a) spherical glycyrrhizin aggregate observed from DPD, b) TEM image of spherical aggregates of glycyrrhizin, c) density distributions of beads A and B, and section views of beads A d) and beads B e)

4.2 INTERACTIONS BETWEEN GLYCYRRHIZIN AND POORLY WATER-SOLUBLE DRUGS IN AQUEOUS SOLUTION

Glycyrrhizin molecules can self-aggregate into micelles in aqueous solutions at the concentration above the critical micelle concentration (CMC). The core/shell structure provides a perfect environment for drug solubilization and

eliminates the side effects caused by pharmaceutical excipients. Spherical aggregates of drug-loaded glycyrrhizin micelles were observed with TEM (Figure 4). It could be easily found from the simulation results that baiclin molecules were solubilized in the core of the aggregate formed by the hydrophobic beads B of glycyrrhizin due to the hydrophobic interaction, both in the spherical, cylindrical and lamellar aggregates.
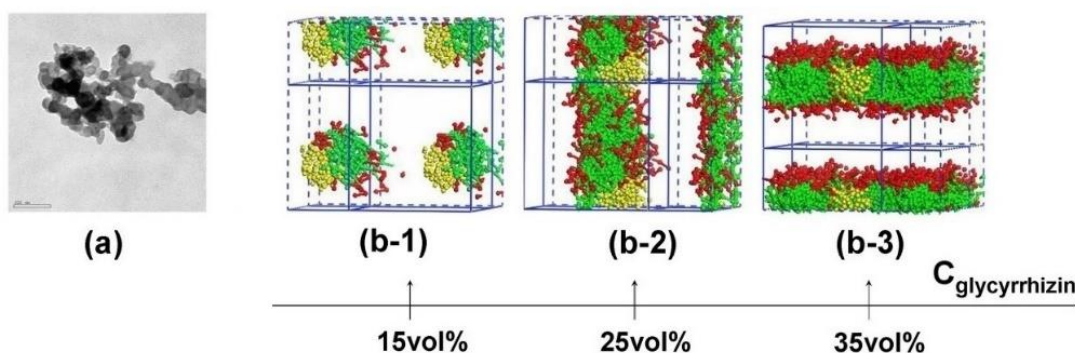
FIGURE 4 Aggregation behavior of baicalin/glycyrrhizin/water system from different methods:
(a) TEM; (b-1), (b-2) and (b-3) mesoscopic simulation

Several snapshots of configurations during the DPD simulation were carried out to study the dynamic formation process of drug-loaded glycyrrhizin micelles. In order to observe the process clearly, the layered aggregate of glycyrrhizin was taken as an example. All components distributed randomly in water in the beginning stage of the simulation. Then some glycyrrhizin molecules aggregated together and small aggregates formed. With the increase of simulation time, larger aggregates formed and the

aggregate changed from cylindrical to lamellar aggregates. In baicalin/glycyrrhizin/water system, baicalin first spread around the surface of glycyrrhizin micelles and then dispersed into the core gradually as the simulation time prolonged. After the simulation of 20000 steps, baicalin molecules dispersed into the core totally and a stable drug-loaded micelle formed (Figure 5). Similar process was observed to puerarin.
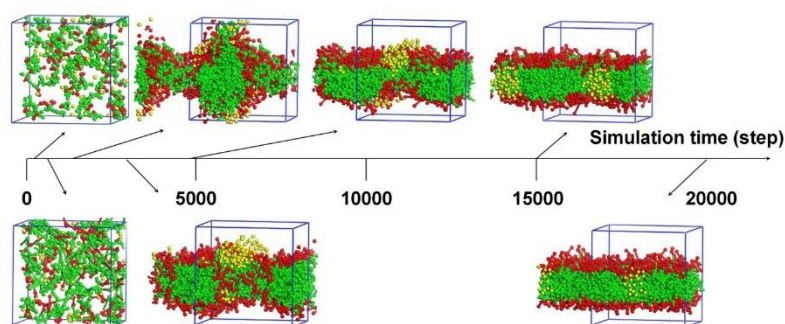
FIGURE 5 The configurations of baicalin/glycyrrhizin/water system at different simulation steps

The micellar property of surfactant is always a research hotspot in the process of preparations. Additionally, the poorly water-soluble drug has effect on the micellar property of surfactant. Figure 6 indicated that the addition of poorly water-soluble drug could facilitate the aggregation of glyrrhizin. As the concentration of the poorly water-soluble drug increased, the time needed for the aggregate formation decreased.
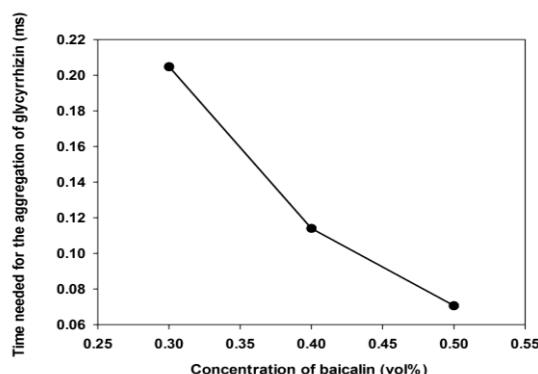
FIGURE 6 Time needed for the aggregate formation of 0.8vol% glycyrrhizin with the addition of poorly water-soluble drugs

5 Conclusion

Glycyrrhizin molecules formed core/shell structured spherical, cylindrical and lamellar aggregates with the increase of concentration. Baicalin, a kind of poorly water-soluble drug, turned to mainly distributed in the hydrophobic core of glycyrrhizin micelle. The solubilization process went through the diffusion from the surface to the core. The addition of poorly water-soluble drug can facilitate the aggregation of glyrrhizin.

Mesoscopic simulations, as the extension of experiments, could provide more details to the aggregation behaviour of glycyrrhizin micelles for drug solubulization, which was a powerful tool for the investigation of solubilization properties and mechanism.

References

[1] Alsenz J, Kansy M 2007 *Advanced Drug Delivery Reviews* **59** 546-67
[2] Güçlü-Üstündağ Ö, Mazza G 2007 *Reviews in Food Science and Nutrition* **47**(3) 231-58
[3] Sasaki Y, Mizutani K, Kasai R, Tanaka O 1988 *World Science and Technology* Beijing School of Traditional Chinese Medicine 1201 *(in Chinese)*
[4] Walthelm U, Dittrich K, Gelbrich G, Schöpke T 2001 *Planta Medica* **67**(1) 49-54
[5] Nokhodchi A, Nazemiyeh H, Ghafourian T, Hassan-Aadeh D, Valizadeh H, Bahary L A S 2003 *IL Farmaco* **57** 883-8
[6] Li Y, Zhang H, Bao M 2012 *Journal of Dispersion Science and Technology* **33**(10) 1437-43

[7] Guo X D, Tan J P K, Kim S H, Zhang L J, Zhang Y, Hedrick J L, Yang Y Y, Qian Y 2009 *Biomaterials* **30**(33) 6556-63
[8] Lin C M, Chang G P, Tsao H K 2011 *J Chem Phys* Taipei National Taiwan University 135 *(in Chinese)*
[9] Jing B, Zhang J, Zhu Y J 2011 *Acta Physico-Chimica Sinica* Beijing State Key Laboratory of Offshore Oil Exploitation 65 *(in Chinese)*
[10] Li Y, Guo Y, Xu G 2013 *Molecular Simulation* 299 *(in Chinese)*
[11] Dai X X, Shi X Y, Yin Q Q 2013 *Journal of colloid and interface science* Beijing School of Traditional Chinese Medicine 165 *(in Chinese)*
[12] Dai X X, Shi X Y, Wang Y G 2012 *Journal of colloid and interface science* Beijing School of Traditional Chinese Medicine 73 *(in Chinese)*

[13] Guo X D, Tan J P K, Zhang L J, Khan M, Liu S Q, Yang Y Y, Qian Y 2009 *Chemical Physics Letters* Guangzhou south china university of technology 336 *(in Chinese)*

[14] Zhao Y R, Chen X, Wang X D 2009 *J Phys Chem B* Shandong Shandong university 2024 *(in Chinese)*

[15] Yang S H, Cheng Y K, Yuan S L 2010 *J Mol Model* Xianggang Baptist University p 1819

[16] Guo X L, Yang S, Lv K, Yuan S L 2011 *Colloids Surf A* 212 *(in Chinese)*

[17] Gong H J, Xu GY, Shi X F, Liu T, Sun Z W 2010 *Colloid Polym Sci* Shandong Beijing 1581 *(in Chinese)*

[18] Dong S J, Yan J T, Xu N, Xu J, Wang H Y 2011 *Surface Science* Jilin Jilin University 868 *(in Chinese)*

[19] Groot R D, Warren P B 1997 *J Chem Phys* Bebington Wirral 4423 *(in Chinese)*

[20] Zheng L S, Yang Y Q, Guo X D, Sun Y, Qian Y, Zhang L J 2011 *J Colloid Interface Sci* Guangzhou south China university of technology 114 *(in Chinese)*

Authors

Yuguang Wang, born in May, 1986, Changchun, China

Current position: pharmacist at the Third Affiliated Hospital of Beijing University of Chinese Medicine.
University studies: M.S. in Chinese Pharmacy, Capital Medical University (2009-2012).
Scientific interest: mesoscale studies on Chinese pharmacy and solubilization mechanism in compound compatibility, clinical rational drug use.
Publications: 4 articles.

Xingxing Dai, born in July, 1987,Fujian, China

Current position: researcher at the University of Chinese Medicine.
University studies: M.S. in Chinese Pharmacy, Capital Medical University (2010-2012).
Scientific interest: computer engineering of medicinal biomaterials, the modernization of Chinese medicine processing.
Publications: 4 articles.

Xinyuan Shi, born in September, 1974, Shanxi, China

Current position: professor at the University of Chinese Medicine.
University studies: Beijing University of Chemical Technology, Northwest University.
Scientific interest: computer simulation study of medicinal biomaterials, the modernization of the traditional fermentation technology.
Publications: 90 articles, 5 patents.
Experience: more than 20 projects.

Yanjiang Qiao, born in November, 1956, Shandong, China

Current position: professor of University of Chinese Medicine.
University studies: Shenyang Pharmaceutical University (1992-1996).
Scientific interest: Information technology for active compounds discovery, information technology for combination rules discovery, and information techniques for quality control of the TCM production process.
Publications: 200 articles, 7 patents.
Experience: more than 20 projects.

Comparison of improved EMD entropy and wavelet entropy in vibration signals of circuit breaker

Dan Xu[1]*, Zhan Zhang[1], Long Yu[2], Yumei Wang[1]

[1]*School of Electrical Engineering and Automation, Henan Polytechnic University, Jiaozuo 454000, China*

[2]*Henan Province Zhoukou Electric Power Bureau, Zhoukou 466000, China*

Abstract

The paper is focusing on extracting the vibration signals of circuit breakers by using Empirical Mode Decomposition (EMD) that improved using least squares to lower the impact on experiment results effectively, which is caused by the EMD inherent end effects. First, work on the EMD and wavelet transform decomposition of both normal and loosening signals, and then to calculate the energy entropy. The results show that the value of improved EMD energy entropy is significantly larger than the wavelet energy entropy. So the improved EMD energy entropy can improve the accuracy of fault diagnosis and provides useful help for the mechanical fault diagnosis based on circuit breakers vibration signals.

Keywords: circuit breakers, vibration signals, energy entropy, EMD, wavelet decomposition

1 Introduction

In traditional vibration, signals analysis theory, some people may totally do signal analysis in time domain, whereas some will do that in frequency domain totally. In order to improve the accuracy of mechanical vibration signals analysis, time- frequency analysis technology is lead up to the fault diagnosis of circuit breaker. For instance, Fourier transform, FFT transform, wavelet transform, wavelet packet transform, and so forth [1]. However, Fourier transform is the fundamental of whether FFT transform or wavelet transform. So it is impossible to get away from the inherence limitations of Fourier transform. That is for the analysis of non-stationary signal, which must be approximated into stationary signal analysis.

Empirical Mode Decomposition (EMD) is a kind of method for the non-stationary signal analysis. According to their own characteristic of time scale to signal decomposition, without setting any base function beforehand. With the base function to automatically generate and self-adaptive multi-resolution feathers, it has the incomparable advantages in the online analysis of non-stationary signal. However, there also exist some self-limitations in EMD, the end effect is the one can't be ignored [2]. In this paper, the EMD was improved, and with the fusion of energy, entropy is applied to fault diagnosis of circuit breaker, to improve the accuracy of diagnosis.

2 End effect

In the EMD, envelope average is a technique to get spline interpolation fitting at the maximum and minimum extreme points of signals and then to calculate the average value of these data. If the endpoints on two ends of the data are not the extreme points, the cubic spline curve which constitutes the upper and lower envelope will produce offset at both ends of the data sequence. Then this will increase the error continuously when we do the spline interpolation and finally generate fitting error in data [3].

During the "screening" process in EMD, due to the uncertainty of extreme value at endpoints, fitting error may be produced in every single spline interpolation and these errors can be accumulated continuously. When we deal with the high-frequency components, end effect will be limited in a very small interval due to a short distance between extreme values and a small time scale. However, when we handle the low-frequency components, situation seems getting worse, because of the large time scale feature existing in low-frequency components, a big fluctuation will occur at the endpoints by the using of cubic spline interpolation. So if we cannot deal with it timely or properly, such fluctuations will "infect", and even cause EMD results in very serious distortion, which led to the failure of EMD [4].

In order to further illustrate the impact on signals of the EMD end effect, here I will build a function to emulate.

$$x(t) = \sin(40\pi t + \pi / 6) + 0.2\sin(20\pi t) , \tag{1}$$

* *Corresponding author* e-mail: xudan@hpu.edu.cn

$$x_1(t) = \sin(40\pi t + \pi / 6)$$

$$x_2(t) = 0.2\sin(20\pi t), \tag{2}$$
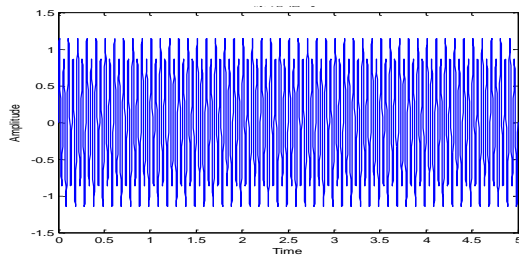
FIGURE 1 Original signal
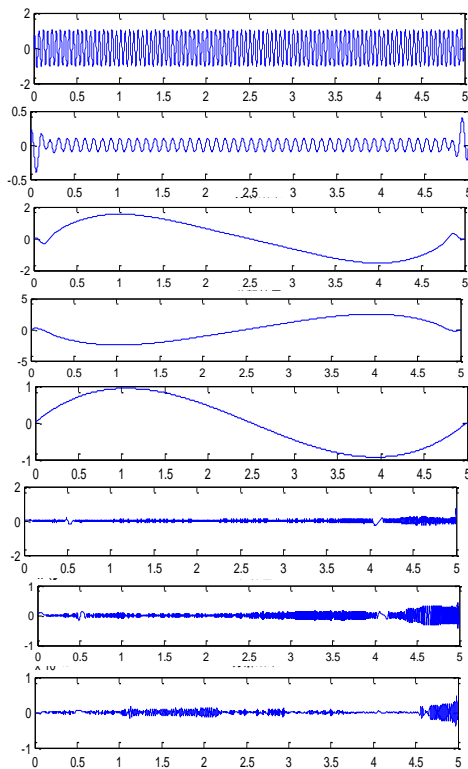
FIGURE 2 The result of Empirical Mode Decomposition

Due to space limitations, only intercept eight IMF components. These eight components are in accordance with the frequency from large to small order. From figure 2, it will be easily noticed that there exist end effect both in IMF2 and IMF3. Waveform aberrance happens both at the left and the right endpoints. Due to the impact of IMF2, the waveform aberrance of IMF3 is getting serious. If we do not deal with these end effects in advance, the Hilbert spectrum will be distorted and the EMD will fail. In this way, we are almost unable to get the correct signal information, which will influence our final judgment.

3 The principle of improved EMD

Now, we are using the endpoint extension to solve the EMD end effect. For the extending method, many scholars have done amounts of research at home and abroad. For instance, Daji Huang [5] adopted the mirror extension method. Qifeng Luo [6] adopted the feature wave method. Yongjun Deng [7] used the RBF neural network method. Yushan Zhang [8] adopted the method of support vector regression machines to extend data. All of them have achieved good effect. In this paper, we will extend boundary by using the least-squares method.

Assume that the sampling time of vibration signals are equally spaced. The sampling data is $\{x(k)\},(k = 1,2,\cdots,n)$ and the sampling interval is $\Delta t = 1$, so the polynomial function is $\hat{x}(k) = a_0 + a_1 k + a_2 k^2 + \cdots + a_m k^m, k = 1,2,\cdots,n$. We will determine the undetermined coefficient $a_j(j = 0,1,2,\cdots,m)$ of $\hat{x}(k)$ to make the quadratic sum of accumulated error be the smallest, which includes $\hat{x}(k)$ and original data $x(k)$.

$$E = \sum_{k=1}^{n}[\hat{x}(k) - x(k)]^2 = \sum_{k=1}^{n}[\sum_{j=1}^{m}a_j k^j - x(k)]^2. \tag{3}$$

The following is an equation to guarantee an average value for E .

$$\frac{\partial E}{\partial x} = 2\sum_{k=1}^{n}k^i[\sum_{j=0}^{m}a_j k^j - x(k)] = 0, i = 1,2,...,m, \tag{4}$$

Let us take the partial derivatives respect to a_j in order to obtain $m + 1$ linear equations. We will get the coefficients a_j respect to every term by solving the equations. The m is the order of the polynomial and its values range from 0 to m. Once we set m equal to 0, we can calculate its constant term.

$$\sum_{k=1}^{n}a_0 k^0 - \sum_{k=1}^{n}x(k)k^0 = 0, \tag{5}$$

The constant term is $a_0 = \frac{1}{n}\sum_{k=1}^{n}x(k)$. And the multinomial coefficient is a_j . Next we will obtain the fitting curve of this polynomial. According to the fitting curve and the data length of instantaneous frequency, we will reconsider the value, get the modified instantaneous frequency.

Figure 4 is the graph of improved EMD. Let us compare figure 4 with figure 2. In figure 4, it can be easily found that the end effects in IMF2 and IMF3 have been improved evidently. This strongly proves that end effect impact on original function can be lowered by the introduction of the EMD, which was improved through least squares. After we process the original signals using improved EMD method, it will be helpful to do the waveform analysis and also help a lot to improve the accuracy of fault diagnosis.

Xu Dan, Zhang Zhan, Yu Long, Wang Yumei
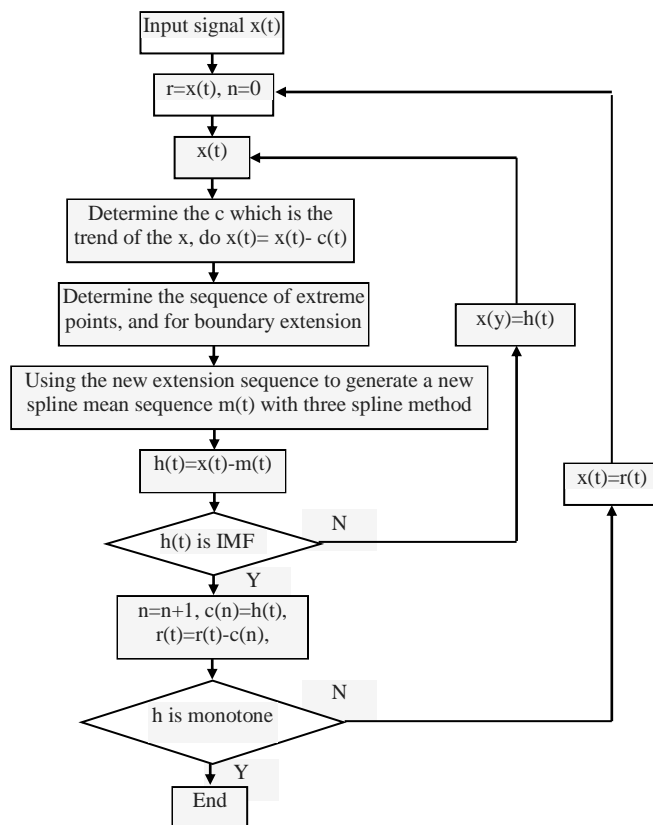
FIGURE 3 The steps of improved EMD method
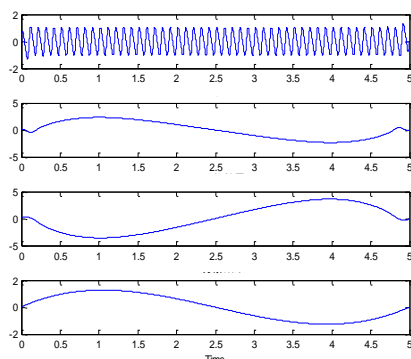
FIGURE 4 The graph of improved EMD

FIGURE 5 original vibration signals
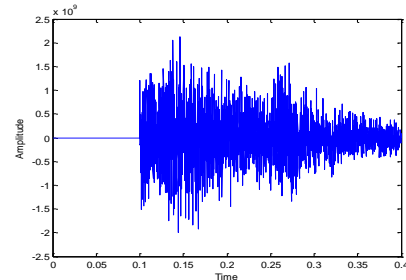
FIGURE 6 Loosening vibration signals
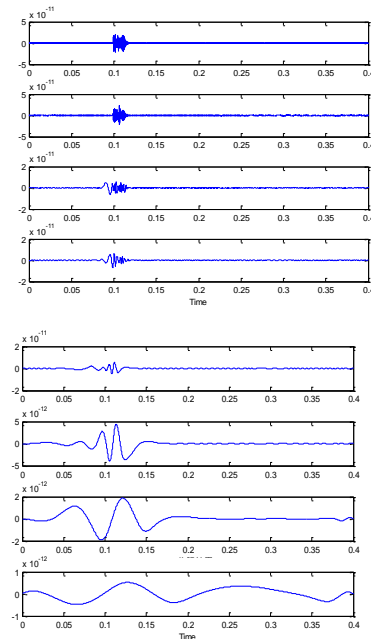
FIGURE 7 Applying the improved EMD to decompose normal signals
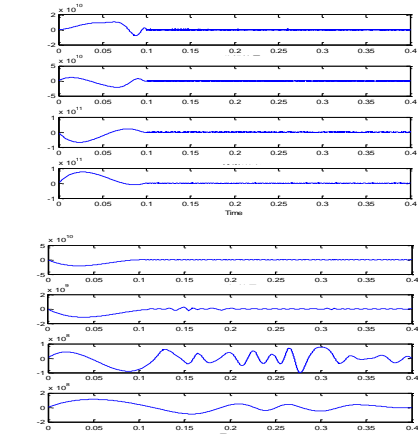
FIGURE 8 Applying the improved EMD to decompose loosening signals

4 The analysis of vibration signals using improved EMD

The original waveform of vibration signals is showing in Figure 5, while the graph of Figure 6 is showing the original waveform of loosening signals. We are able to see that there is some obvious difference in the waveform of the normal and the loosening situation showing on the two time domain graphs. However, we have lots of difficulties in finding out the specific situation and the types of faults just through the comparison in time domain graphs. Therefore, we need to apply the improved EMD method to decompose the normal and the loosening signals separately. The IMF results are showing in both Figure 7 and Figure 8.

After the decomposition, we get 11 IMF components. Here we only list the 8 components in front. In the vibration signal, the key signal mainly concentrated in the high frequency parts, which are the first few components of the IMF.

5 Applying wavelet decomposition to the analysis of vibration signals

The principle of wavelet transform won't be displayed here. We did five wavelet decompositions to obtain the high-frequency components of normal and loosening signals that was shown in Figure 9 and Figure 10 respectively. For an easy observing, we increase the time to 10^4 times of the normal. From Figure 9 we can clearly see that vibration happens at 0.1s, whose signal ends up at 0.12s and the measure of it lasts 1s. However, the vibration continues to 0.5s in Figure 10. The huge denoising function of wavelet transform shows good results in the vibration signals processing.
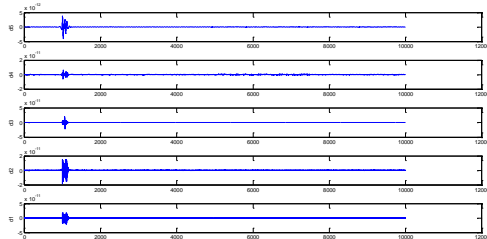
FIGURE 9 Wavelet decomposition of normal waveform
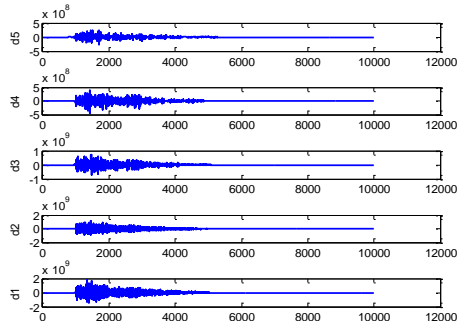
FIGURE 10 Wavelet decomposition of loosening waveform

6 Vibration signals analysis through power entropy and the EMD

Information entropy is described mathematically like this: Assume that $p(p_1, p_2, \cdots, p_n)$ is an uncertain probability distribution and k is an arbitrary constant, so the information entropy of this distribution is

$$s(p) = -k \sum_{i=1}^{n} p_i \ln p_i . \tag{6}$$

After a calculation, we know that if the system information entropy $s = 0$, it will be a certain system. No events will happen except the ones that had happened. The probability of event occurrence is strongly not uniform, so the uncertainty is 0 [9].

The power frequency distributions of vibration signals are different, which are generated in different status of different circuit breakers. Even in the same status, due to the changeable environment, there is difference in this distribution. We need a standard to value the amount of power. So the concept of power entropy is coming up. When we apply the EMD method to decompose a signal, different frequency terms will be included in the n IMF components. The energy is $E_1, E_2, \cdots E_n$, and it's a partition in frequency domain. The definition of EMD power entropy is

$$H_E = -\sum_{i=1}^{n} p_i \ln p_i . \tag{7}$$

The p_i in this equation is the percentage of power in the ith IMF component, that is $p_i = E_i / E$, and $E = \sum_{i=1}^{n} E_i$. According to the definition of information entropy, the more uniform the distribution of p_i is, the larger the H_E value will be, whereas the smaller.
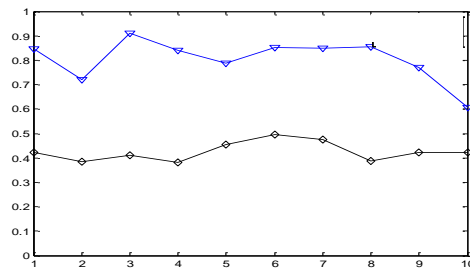
FIGURE 11 The graph of comparison of wavelet power entropy

Let us work on the calculation of the power entropy value after applying the EMD method and wavelet transform decomposition to normal and loosening signals. Power entropy values for ten groups signals by using wavelet transform decomposition are shown in Figure 11. The above line represents loosening signals, while the following one is for normal signals [10].

Power entropy values for normal and loosening signals by using improved EMD method are shown in Figure 12.
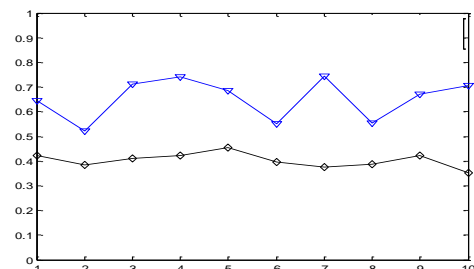
FIGURE 12 Graph of the comparison of improved EMD power entropy

When we compare the Figure 11 with Figure 12, we can find that the switching signals in normal status are smaller than that in fault status and they are relatively centralized. The mechanical fault vibration signals of circuit breakers are often at high frequency, and

resonance is hard to happen. So the power signals are relatively uniform and centralized. Power entropy can be used as the fingerprint for circuit breaker fault.

Now we are going to get the average value of the signals that we measured to obtain the improved EMD and wavelet transform entropy value both under normal and loosening situations. Wavelet entropy under normal state is 0.403 and under loosening situations is 0.656. The improved EMD entropy under normal state is 0.426 and under loosening situations is 0.803.

It tells us that the entropy value of improved EMD under loosening situation is larger than that we get under the normal situation. Meanwhile, the result is much more obvious than that by using wavelet. The bigger entropy value exists in these two statuses, the easier it can be distinguished. From this point, we can realize that the improved EMD method is better than the wavelet transform evidently. This kind of entropy value analysis method gives us a lot of convenience for training once the data permits later.

Using the improved Empirical Mode Decomposition to process the original signal, we can effectively reduce the influence of end effect on the decomposition results and overcome the inherent disadvantages of EMD. Therefore, we combined it with energy entropy to do the analysis of vibration signals and achieved good effect.

Reference

[1] Hoidalen H K, Rundem 2005 Continuous monitoring of circuit breakers using vibration analysis *IEEE Transactions on Power Delivery* **20**(4) 2458-65

[2] Mallat S 1989. A theory for multiresolution signal decomposition the wavelet representation *IEEE Transactions on Pattern Analysis and Machine Intelligence* **11**(7) 674-93

[3] Loh C H, Wu T C, Huang N E, et al. 2001, Application of the empirical mode decomposition-hilbert spectrum method to identify near-fault ground-motion characteristics and structural responses *Bulletin of the Seismological Society of America* **91**(5) 1399-57

[4] Huang N E, Steven R L 1988, The empirical mode decomposition and Hilbert specturm for nonlinear and non-stationary time series analysis *Proc.R.Soc.Lond.A* 903-96

[5] Luo Qi-feng, Shi Chun-xiang 2003. Hilbert-Huang transform and several problems in its calculation method *Journal of TONGJI University (Natural Science)* **31**(6) 637-40

[6] Huang Da-ji, Zhao Jin-ping, Su Ji-lan 2003 Practical implementation of the Hilbert-Huang transform algorithm *Acta Oceanologica Sinica* **25**(1) 1-11

[7] Deng Yong-jun, Wang Wei, et al. 2001 The boundary problem processing for EMD method and Hilbert transform *Chinese Science Bulletin* **46**(3) 363-75

[8] Zhang Yu-shan, Liang Jian-wen 2003 Disposal of boundary problem in EMD method applying autoregressive model *Progress In Natural Science* **13**(10) 1054-9

[9] He Zheng-you, Cai Yu-mei, Qian Qing-quan 2005 A Study of wavelet entropy theory and its application in electric power system fault detection *Proceedings of the CSEE* **25**(5) 38-43

[10] Sun Lai-jun, Hu Xiao-guang, Ji Yan-chao 2007 Fault diagnosis for high voltage circuit breakers with improved characteristic entropy of wavelet packet *Proceedings of the CSEE* **27**(12) 103-8

Authors	
	Dan Xu, born on May 13, 1979, Henan Province Jiaozuo City, China **Current position, grades:** lecturer **University studies:** Henan Polytechnic University **Scientific interest:** electrical safety technology **Publications:** "Modern power supply technology", "University Computer Foundation" **Experience:** She received the BS degree in power system and automation, and MS degree in electric machine and electric equipment from Henan Polytechnic University, China, in 2002 and 2005, respectively. She is a lecturer at Henan Polytechnic University and the national quality curriculum "power technology" speaker. Her research interests include power system analysis and control. Published 11 papers and 3 books, participated in the 4 research.
	Zhan Zhang, born in August 7, 1976, Jiangsu Province Xuzhou City China **Current position, grades:** Associate Professor **University studies:** Henan Polytechnic University **Scientific interest:** transmit electricity, Power Supply and Distribution **Publications:** "Modern power supply technology", "Electrical Manual of diagnosing faults of machine" ect. **Experience:** Now She worked for School of Electrical Engineering and Automation, Henan polytechnic university. She promoted to Associate Professor in 2012 at Henan Polytechnic University. Her major fields of study are Electrical automation, artificial intelligence and Industrial process control. She has edited the "Modern power supply technology", "Electrical Manual of diagnosing faults of machine" ect. About 12 Papers were published.
	Long Yu, born on June 17, 1988, Zhoukou City, Henan Province, China **Current position, grades:** Zhoukou power supply company, assistant engineer: **University studies:** Henan Polytechnic University **Scientific interest:** electric engineering **Experience:** Yu Long was born in Henan, China, in 1988. He received the B.E. and Master of Engineering degrees in 2011 and 2013, respectively, from Henan Polytechnic University, Jiaozuo, China, both in electrical engineering.He is currently with Zhoukou Power Supply Company, where he is an assistant engineer. His areas of interests include design of substation and fault online diagnosis of power system.
	Yumei Wang, born on March 1, 1963, Huixian City, Henan Province, China **Current position, grades:** Henan Polytechnic University, Professor **University studies:** China University of Mining Technology **Scientific interest:** electrical power system, power supply technology, etc. **Publications:** "Modern power supply technology", "Analog Electronic Technology" **Experience:** Yumei Wang was born in Huixian City, China, in 1963. She received the bachelor degree from Xi`an Jiaotong University in Xi'an City in 1986 and master degree from China University of Mining Technology in Beijing in 1996. She works as a teacher in Henan Polytechnic University since graduation. Now she is a professor in electrical engineering field.

Fault diagnosis of nuclear facilities based on hidden Markov model

Fengwei Yuan*, Qian Deng, Jiazhu Zou

College of Mechanical Engineering, University of South China, Hengyang 421001, Hunan, China

Received 1 August 2014, www.cmnt.lv

Abstract

Due to the complex structure of nuclear facilities in a high irradiation environment, people are hard to approach it, In view of these situation a fault diagnosis method based on HMM (Hidden Markov Model) of capturing the audio signal while facilities are operating is proposed. With the strong modelling ability, HMM can be applied to analysing such as audio signal non-stationary time signal. By using this method, the original mechanical structures of nuclear facilities are not destroyed. The proposed sensors were needed as few as possible by the whole diagnosis system and which has a simple structure, low cost structure, the fault diagnosis rate is high and so on. State monitoring and fault diagnosis system of complex nuclear power equipment can timely and effective to provide running status and potential failure information for operating personnel, which has a vital significance for the safe and reliable operation of nuclear power equipment.

Keywords: nuclear facilities, Hidden Markov Model (HMM), fault diagnosis

1 Introduction

In March 2011, the nuclear leakage of the Fukushima nuclear power plant was caused by magnitude 9.0 earthquake, which was a major nuclear accident since the nuclear accident of US Three Island nuclear power station and the Chernobyl accident of the former Soviet Union happened in April 1986. The nuclear accident of the Fukushima as the major nuclear accident in the process of the peaceful use of nuclear power, had caused all countries to re-evaluate nuclear power development plan, what's more, human beings had put forward new reflection on nuclear safety, economy and society effect.

Due to the complex structure of nuclear facilities, strong continuity in production, high security requirements, under high temperature and irradiation environment, so the nuclear accident would cause severe economic losses and serious social consequence no matter which fault caused the nuclear accident [1]. Monitoring running condition and fault diagnosis of complex nuclear power equipment would be usually use traditional threshold method: mainly through preset threshold of monitoring parameters, the system send out alarm signal or take protective action when the monitored amount was more than the prescribed threshold, while the fault may not be able to curb [2, 3]. Therefore, research a fault diagnosis based on HMM of capturing the audio signal and fault diagnosis system had great significance for improving the safety and reliability of the nuclear energy development and utilization.

2 Basic theory and algorithm of the HMM

Baum and Peterie put forward the HMM in 1966, Baker and Jelinek had achieved great success in applying speech recognition. In recent years, more and more attention was paying on the HMM, which had been successfully used in speech recognition, molecular biology, image segmentation, fault diagnosis and many other fields [4, 5].

2.1 BASIC CONCEPT OF THE HMM

The HMM is a double stochastic process. It is using the Markov Chain Model to describe changes of the statistical characteristics of the signal. While these changes are indirectly describe by the observed sequence, it can be divided into two-layer structure - hidden state layer and observed layer, Markov Chain lays in hidden state layer, while the observed layer is the output of the hidden state layer [6, 7].

HMM could be short for $\lambda = (\pi, A, B)$, among them:

1) π stand for initial state probability vector;

2) A is state transition probability matrix, each element a_{ij} of A represents the probability of the HMM transfer from state θ_i to state θ_j.

3) B is confusion matrix, each element b_{jk} of B represents the probability of the HMM appears the observed value V_k as its state is θ_j at t time.

The HMM describe the stochastic process through different distribution of π, A and B. Each probability of state transition matrix and confusion matrix is not related

to the time, so these matrix don't change over time when the system is evaluating.

2.2 TRAINING ALGORITHM OF THE HMM

2.2.1 Forward-Backward algorithm

The main idea of Forward-Backward algorithm is through the Forward and Backward iterative process to calculate two instrumental variable, which is the value of the Forward and Backward variables. In a given observed sequence $O = \{o_1, o_2, ..., o_T\}$ and mode λ. It is effective to use Forward-Backward algorithm to calculate and observe probability $P = (O / \lambda)$ of variable sequence O under given model.

2.2.2 The Viterbi algorithm

The Viterbi algorithm provides an effective method to analyse the observed sequence of the HMM and capture the most likely sequence of hidden state. This algorithm use recursion to reduce amount of calculation and use the whole sequence to do judgment, thus to make good analysis of sequence that contains interference. We need to define the instrumental variables $\psi_t(i)$ and $\delta_t(i)$ for Viterbi algorithm, among them:

$$\delta_t(i) = \max_{q_1, ..., q_{t-1}} P\left(q_1, ..., q_t, q_t = \theta_i, o_1, ..., o_t / \lambda\right). \tag{1}$$

The basic process of algorithm is as follows:
(i) initialization:

$$\delta_1(i) = \pi_i b_i(o_1), 1 \le i \le N, \tag{2}$$

$$\psi_1(i) = 0, 1 \le i \le N. \tag{3}$$

(ii) iterations:

$$\delta_t(i) = \max_{1 \le i \le N}\left[\delta_{t-1}(i) a_{ij}\right] b_j(o_t), 2 \le t \le T, 1 \le j \le N, \tag{4}$$

$$\psi_t(j) = \arg\max_{1 \le i \le N}\left[\delta_{t-1}(i) a_{ij}\right], 1 \le t \le T, 1 \le j \le N. \tag{5}$$

(iii) termination:

$$P^* = \max_{1 \le i \le N}\left[\delta_T(i)\right], \tag{6}$$

$$q_T^* = \arg\max_{1 \le i \le N}\left[\delta_T(i)\right]. \tag{7}$$

(iv) finding the optimal:

$$q_t^* = \psi_{t+1}\left(q_{t+1}^*\right), t = T-1, T-2, ..., 1. \tag{8}$$

2.2.3 Baum-Welch algorithm

Baum-Welch algorithm that use maximum likelihood criterion is currently the main training methods used for the HMM. Its concrete ideas is to estimate one of the most appropriate HMM based on an observed sequence and a set of hidden state, therefore, it is to determine the most appropriate description of a triple (π, A, B) for a known sequence.

To define $\xi_t(i, j)$ for a given training sequence (observation sequence) O and model λ, its probability of Markov Chain in the state θ_i at time t and in the state θ_j at time $t + 1$, namely:

$$\xi_t(i, j) = p(q_t = \theta_t, q_{t+1} = \theta_j | O, \lambda) =$$
$$\{\alpha_t(i) a_{ij} b_j(o_{t+1})\beta_{t+1}(j)\} / p\{O|\lambda\}, \tag{9}$$

where $\alpha_t(i)$ is the forward variable and $\beta_t(i)$ is the backward variable. Then, to define probability of Markov Chain in the state θ_i at time t under the given model $\gamma_t(i)$ is:

$$\gamma_t(i) = p(q_t = \theta_i | O, \lambda) = \sum_{j=1}^{N} \xi_t(i, j) = \alpha_t(i)\beta_t(i) / P(O | \lambda). \tag{10}$$

Then we can get revaluation formula of the HMM is:

$$\bar{\pi}_i = \gamma_1(i), \tag{11}$$

$$\bar{a}_{ij} = \sum_{t=1}^{T-1} \xi_t(i, j) \Big/ \sum_{t=1}^{T-1} \gamma_t(i), \tag{12}$$

$$\bar{b}_{ij} = \sum_{t=1, O_t=v_k}^{T} \gamma_t(j) \Big/ \sum_{t=1}^{T} \gamma_t(i). \tag{13}$$

3 State monitoring and fault diagnosis of complex nuclear power equipment

State monitoring and fault diagnosis system collect related characteristics of monitoring device by sensors, then transmitted to the host computer after appropriate processing, it provided on-line state analysis and fault diagnosis by using computer of high-speed data processing. State monitoring and fault diagnosis system of complex nuclear equipment could timely and effectively provide information of running state and potential failure to operating personnel, which had a vital significance for the safe and reliable operation of complex nuclear equipment.

3.1 RESEARCH OF FAULT DIAGNOSIS TECHNOLOGY OF COMPLEX NUCLEAR POWER EQUIPMENT

At present, all countries have done a lot of research on technology of state monitoring and fault diagnosis of nuclear station, also have obtained certain achievement. The United States, Japan, France, Germany and other developed countries had reached a leading position in the research of state monitoring and fault diagnosis. Japan Atomic Energy institute developed DISKET,

which was a typical expert system based on knowledge base, the system had passed experiment verification on the pressurized water rector. France took six years to develop a SINDBAD system after the accident of US Three Island nuclear power station, the system was mainly used to do some common fault forecast and diagnosis of the plant by using pattern recognition of process and expert system. Norwegian Fantoni Paolo F and Mazzola Alesandro realized to confirm multiple failure signal of boiling water reactor by using autocorrelation neural network. Canadian Lege RP and Garland Wm J used the cumulative total control graph and neural network for fault detection and diagnosis of heat transport system of CANDU reactor. The Peach Bottom plant of America adopt "advanced real-time state monitoring system" for on-line monitoring, which greatly improved the monitoring crack of circulated pump shaft, it was helpful for fault diagnosis through correlation analysis of vibration, pressure, temperature and other parameters [1, 8, 9].

Some domestic Institute also had done a lot of research work on condition monitoring and fault diagnosis of nuclear power plants. The Tsinghua university developed a fault diagnosis expert system of secondary circuit is relative mature, the system had passed the simulator verification. Harbin Institute of Technology carried out a research of intelligent technology on fault diagnosis of nuclear power plant, which was based on neural network; the safety supervision plate system used by Daya Bay nuclear power plant, with fault identification, actuator supervision, selection of accident regulation, condition monitoring of power station and other functions.

3.2 ANTI-RADIATION SENSOR TEST

The fault diagnosis system need to collect vibration, noise, temperature of the system under test by sensors, while there are not mature products to work in high radiation environment for long term operation, sensors need to adopt high resistance strengthening technique, which limited the sort of monitoring information. Therefore, it was completely different between signal monitoring in high radiation environment and regular environment, because the monitoring information of high radiation environment is greatly reduced. Research results of this project group show that the sound detection method could get more operation information of equipment's, and anti-radiation properties of acoustic sensor was very excellent.

CMOS (Complementary Metal-Oxide Semiconductor) image sensor is highly integrated, the pixel array, the timing control circuit, the ADC, the signal processing circuit camera system required for the various functional modules are integrated into a single chip, such as seen in Figure 1 [10].

Common camera with CMOS image sensor, which would completely fail under an environment of ^{60}Co as

radiation source and dose rate is 15GY/h for seven minutes. Because gamma rays, the image would randomly appear a lot of flake when the camera expose to the radiation environment, the number of flake gradually increased over time. The whole image was full of "snow" after 30 seconds' exposure. As gamma rays could cause an irreversible destruction to the CMOS image sensors, the camera was unable to work when the radiation time reaches 7 minutes. The total dose of the whole process was 1.75GY.
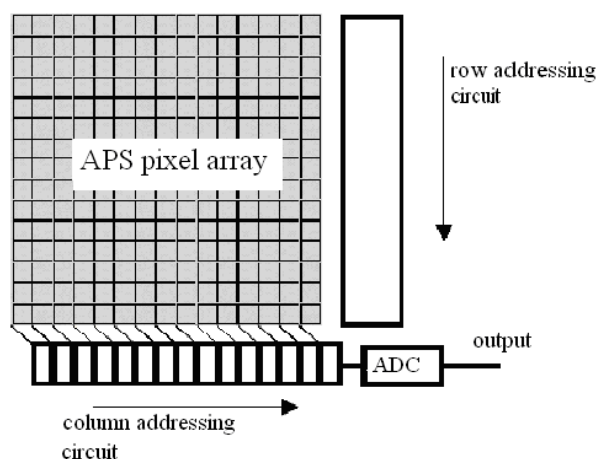
FIGURE 1 CMOS image sensor

FIGURE 2 Not exposed to radiation

Radiation effects lead to performance degradation of CMOS image sensors, mainly due to radiation were trapped charge and interface state generation. The interaction of γ rays and the silicon material to produce secondary electrons by Compton scattering and electron traps and hole traps in the CMOS image sensor. SiO_2 layer of oxide trapped charge (fixed positive charge), the Si-SiO$_2$ interface at some interface traps. These traps to the Si-SiO$_2$ interface potential change, resulting in a MOS structure threshold voltage V_{TH} or flat-band voltage V_{FB} move along the voltage axis. The V_{TH} drift and bias voltage is related to positive bias gate voltage CMOS image sensor, the fixed positive charge and interface trap formation becomes faster, the photodiode of the N+ region to capture the fixed positive charge, and V_{TH} decreases. This had led to the depletion region thickness decreases, the more electrons were able to enter the potential well of the photodiode, the density of

464

the dark current or leakage current of the diode increases, the irradiation induced electron - hole pair concentration with the increase in gamma-ray fluence, leading to the dark current increases with the fluence. The white spots appeared in Figure 3 were caused by the increase in dark current.

FIGURE 3 30 seconds after exposure

Maximum direct output voltage signal peak of dynamic sound sensor can reach 300mV. If signal preamplifier circuit is installed in the irradiation room to enlarge the signal, which is then passed through the cable back to the computer, complexity of experiment will increase. And the more, when the signal changes, there is no way to tell it is the result of sensor failure or preamplifier circuit failure. Taking these factors into account, this experiment chooses to transfer coil sound sensor output signal directly through a 40m cable to the PC outside of the irradiation room, which is read by the sound card and the experimental data is also recorded by the sound card. The experiment used a double-shielded cable, making the space electromagnetic interference to reduce the impact. The experiment was performed at Hunan Academy of Agricultural Sciences Radiation Center, ^{60}Co was used as the radiation source.

Anti-radiation properties of the acoustic sensor were several magnitudes higher than the CMOS image sensor. Selecting silicon micro-acoustic sensor SPM0204 under the environment of the radiation source ^{60}Co and dose rate of 1.5KGY/h, the sensor would fail in 90 minutes of exposure with the total dose rate of 2250GY. Figures 4 and 6 show the time domain waveform, Figures 5 and 7 show the frequency domain wave, the incentive source was common ringer. The failure mechanism of sensors remain to be further study, we preliminary estimated that the amplifying circuit of the silicon micro-acoustic sensor and the quiescent point were damaged by the irradiation.

Dynamic acoustic sensors for mechanical structure. According to the relevant literature, the permanent magnet is able to tolerate the radiation dose of 100000 GY with no significant changes in its magnetic properties. There are no electronic components, which is not resistant to irradiation within the sensor and the sensor can still work properly under high doses of radiation.

If we correctly select anti-radiation sensor it should provide powerful hardware support for fault diagnosis. It is very common to use HMM for voice signal analysis and fault diagnosis.
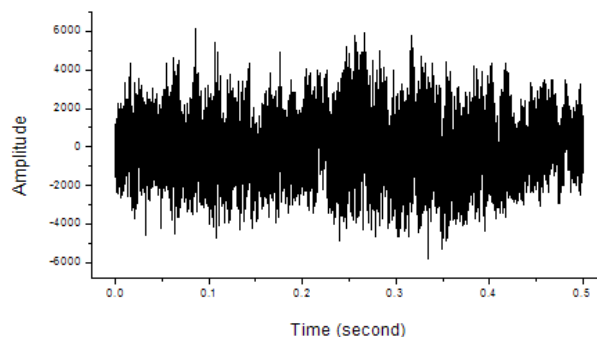
FIGURE 4 time domain waveform (Not exposed to radiation).
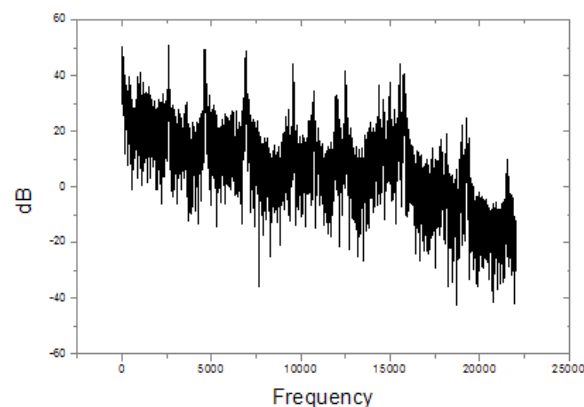
FIGURE 5 frequency domain wave (Not exposed to radiation).
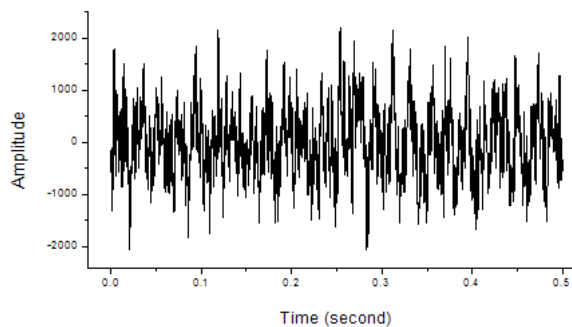
FIGURE 6 Time domain waveform (90 minutes after exposure)
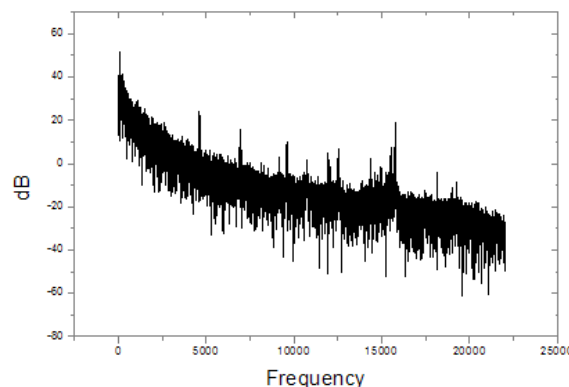
FIGURE 7 frequency domain wave (90 minutes after exposure)

465

Yuan Fengwei, Deng Qian, Zou Jiazhu

3.3 APPLICATION OF THE HMM IN FAULT DIAGNOSIS OF COMPLEX NUCLEAR POWER EQUIPMENT

The HMM is suitable for the dynamic process modelling of time series, which can solve problem of random uncertainty, especially for analysis of non-stationary and poor repeated sequence. At the same time, it can handle sequence of arbitrary length in theory. These characteristics have a strong pertinence for state monitoring and fault diagnosis of complex nuclear power equipment with characteristics of reliability requirements, informative, non-stationary and poor repeated reproducibility.

A fault diagnosis method based on HMM of capturing the audio signal while the nuclear facilities are operating, which just need to choose a place with a small dose rate but can clearly collect voice signal while the device are operating to place sensors around the nuclear power equipment. To extend the working life of sensors by shielding device.

The audio signal will transmit to the host computer for fourier transform after AD transform, to analyse frequency domain and extract the characteristic of audio signals. Extracting various fault of equipment and establishing fault database, the sound signals of operating equipment will match the information in the database, if it can match, then the device may appear the corresponding fault. Fault diagnosis technology is shown in Figure 8.
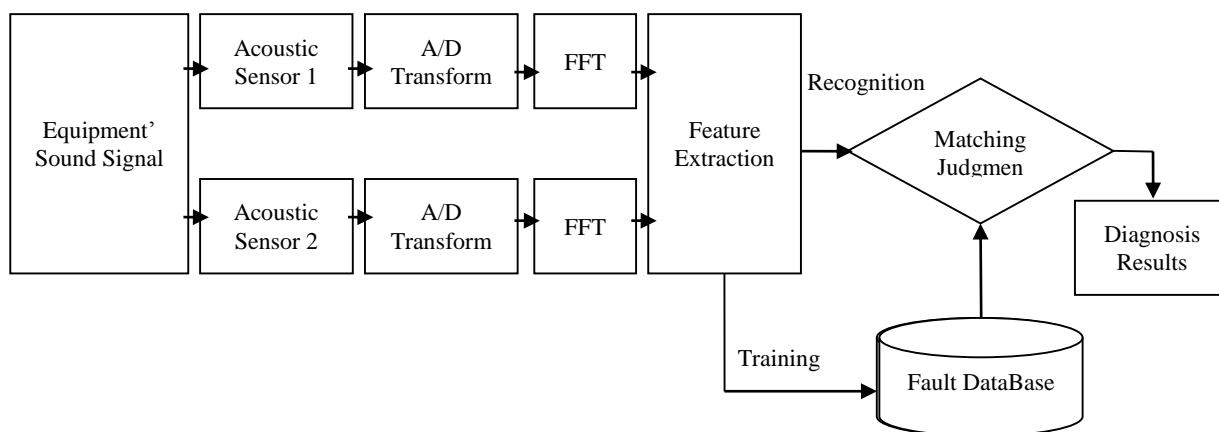
FIGURE 8 System Block Diagram of Fault Diagnosis Technology

4 Conclusion

A fault diagnosis method based on HMM of capturing the audio signal while the nuclear facilities are operating is proposed. Resistance to radiation in the CMOS image sensor in the high radiation environment and moving-coil sound sensor performance study and experimental validation of the mechanical structure of a moving coil sound sensor resistance to radiation performance than the highly integrated semiconductor devices. The advantage of real-time monitoring of voice signal compared with real-time monitoring of video signal lay in anti-radiation of acoustic sensor is obviously stronger than CMOS image sensor. Adopt the method of fault diagnosis of nuclear power equipment with the characteristics of less sensors, high fault diagnosis.

Acknowledgments

References

[1] Feng C J 2002 Theory and method of dynamic pattern recognition based on HMM and application in fault diagnosis of the rotating machinery *Zhejiang University Press:* Zhejiang China *(in Chinese)*
[2] Munteanu D Autran J L 2008 *IEEE Transactions on Nuclear Science* **55**(4) 1854-77
[3] Allen G R, Guertin S M, Scheick L Z, Irom F, Zajac S 2012 Compendium of Test Results of Single Event Effects Conducted by the Jet Propulsion Laboratory *2012 IEEE Radiation Effects Data Workshop (REDW)* 1-10
[4] Bengio S 2004 Multimodal speech processing using asynchronous Hidden Markov Models *Information Fusion* **5**(2) 81-9
[5] Werner S, Rigoll G 2001 Pseudo 2-Dimensional hidden Markov models in speech recognition *IEEE Workshop on Automatic Speech Recognition and Understanding 2001 ASRU'01* 441-4

[6] Liu Y K, Xia H, Xie C L 2008 Research of state monitoring and fault diagnosis system of nuclear power equipment *Atomic energy science and technology* **42**(3) 200-5 *(in Chinese)*
[7] Ma X J, Wang F L, Cai Y, et al 2004 Research of local wave time-frequency analysis in early fault diagnosis of rotor system *Atomic science and technology* **24**(3) 161-4 *(in Chinese)*
[8] Wang N Y 2002 Safety analysis of nuclear power plant *Research and development of Journal of world science and technology* **24**(6) 14-7 *(in Chinese)*
[9] Amusan O A, Casey M C, Bhuva B L, McMorrow D, Gadlage M J, Melinger J S, Massengill L W 2009 *IEEE Transactions on Nuclear Science* **56**(6) 3065-70
[10] Dodd P E, Schwank J R, Felix J A, Paillet P, Ferlet-Cavrois V, Baggio J, Reed R A, Warren K M, Weller R A, Schrimpf R D, Hash G L, Dalton S M, Hirose K, Saito H 2007 *IEEE Transactions on Nuclear Science* **54**(6) 2303-11

Authors	
	Fengwei Yuan, born in June, 1977, Hunan, China **Current position, grades:** associate Professor of Mechanical Engineering, University of South China. **University studies:** Master of engineering on Mechanical Engineering (2001, Central South University of Forestry and Technology, China). **Scientific interest:** virtual reality and computer modelling in mechanical engineering. **Publications:** 11 papers.
	Qian Deng, born in October, 1987, Hunan, China **Current position, grades:** Lecturer of Mechanical Engineering, University of South China. **University studies:** Master of engineering on Mechanical Engineering (2012, University of South China, China). **Scientific interest:** computer modelling in mechanical engineering. **Publications:** 4 papers.
	Jiazhu Zou, born in June, 1977, Hunan, China **Current position, grades:** lecturer of Mechanical Engineering, University of South China. **University studies:** Master of engineering on Mechanical Engineering (2006, University of South China, China). **Scientific interest:** fluid control and computer modelling in mechanical engineering. **Publications:** 7 papers.

Numerical simulation of dynamic responses caused by dynamic compaction on backfilling foundation

Siqiang Wen*, YunPeng Li, Yan Zhou

College of Mechanical and Transportation Engineering, China University of Petroleum, Beijing, 102249, China

Abstract

Since the ground vibration caused by dynamic compaction threatens the structures around the site, a dynamics numerical simulation of the process of dynamic compaction is carried out based on the dynamic compaction experiment in the backfill soil site with the Finite Element Method (FEM) software. The calculated results present the vibration-time curves in radial and vertical directions on the ground in different distances. The characteristics of the vibration-time curves and how the peak velocity and acceleration change with distance are analysed. By comparing the simulated results with field data, the reasons which cause the differences are pointed out. Through comparison it is considered that the near-field dynamic responses in the simulation are more reliable than the far-field ones. According to standards the safe distances of each type of structures are evaluated. The relationship between energy utilization and the vibration energy is discussed, and that raising the aspect ratio of the hammer can reduce vibration is pointed out.

Keywords: dynamic compaction, dynamic response, backfill soil, simulation

1 Introduction

Dynamic compaction is a method of foundation reinforcement that first proposed by L.Menard [1] in 1969. The method for its economy and simplicity are widely used in foundation of miscellaneous fill, loose sand, clay, etc. But in the process of impaction, due to the great stress on the contact surface caused by the powerful collision in a very short time, a strong stress wave is generated in the soil and spreads outward. When the wave transmits to the structures near the tamping point, the energy of wave enters the structures and causes the structures to vibrate. If the vibrational amplitude and frequency reach a certain condition, the structures will be damaged. Therefore, it is necessary to analyse the vibration response of dynamic compaction to avoid the threat to the surrounding structures.

Researches on the dynamic response in the foundation caused by the dynamic compaction are commonly in two ways – the on-site experiments and the theoretical analysis based on a certain mechanical model. Due to the complexity of the dynamic problem, it is difficult to use analytical method to describe the mechanical effects of dynamic compaction. But in this aspect there were some studies, like Scott [2] used lumped equivalent method to establish the motion equation of the hammer-soil system and got the displacement of tamped point; Mayne [3] according to the momentum theorem derived maximum contact stress between hammer and soil, the duration of contact and the maximum stress at a certain distance below the tamping point; Y.K.Chow [4] divided a cylinder which is the same diameter with the hammer from the soil and analysed the propagation characteristics of one-

dimensional stress wave; Kong [5] according to the relationship between the contact stress and displacement, used Laplace-Hankle transform and transfer matrix method to solve the three-dimensional elastic dynamic equations , and got the stress distribution in the layered foundation. However, because of the complexity, the problem is commonly focused on a certain aspect and the model is simplified to meet the feasibility of analysis, like Scott's study only described one-dimensional motion of the hammer; the discussion in reference [5] is based on the assumption of linear elastic medium. These simplifications bring feasibility as well as limitation to the problem solving, so that the numerical simulation is considered as a better method in the research of dynamic compaction. In this aspect, many scholars have studied from different ways, like Niu [6] simulated the finite element dynamic compaction problems and gave the numerical solution of stress field, displacement field and acceleration field; Cai [7] obtained the deformation characteristics, reinforced depth and the stress distribution characteristics of dynamic compaction with large deformation finite element and boundary element method; Jiang[8] introduced the erosion element to eliminate the impact of mesh distortion and on this basis used large deformation finite element to simulate the deformation of tamped pit. Otherwise, there are some researches [9-11] that also have done remarkable works.

This paper in the background of a dynamic compaction project, according to the environmental characteristics of the reinforced area, establishes an axisymmetric FEM numerical model, and by means of FEM software, analyses the dynamic response of the process of dynamic compaction in the backfill foundation with the explicit algorithm which adapt to the large collisional deformation.

---

*Corresponding author e-mail: wsq106@163.com

Then in combination with the real-time monitoring results, the laws of dynamic response of backfill soil consolidation, as well as the influence on surrounding structures are discussed.

2 Analysis model of dynamic compaction

2.1 DYNAMIC EQUILIBRIUM EQUATION

In the process of dynamic compaction, the potential energy is converted into kinetic energy, and then spreads in the foundation under the damping of soil. So it should be simplified as a dynamic system and analysed with dynamic FEM.

For general dynamic FEM question, by Hamilton's principle [12] the dynamic equilibrium equation of system could be received.

$$M\ddot{u} + C\dot{u} + Ku = Q ,\tag{1}$$

where M is the system mass matrix in which are the nodes mass coefficients of soil; C is the system damping matrix that relates to the damping properties of the soil; K is the system stiffness matrix which determined by mechanical parameters of the soil; Q is the equivalent nodal load vector; u is the nodal displacement vector of system.

2.2 NUMERICAL MODEL OF DYNAMIC COMPACTION

The geometric parameters of dynamic compaction model in this paper come from the project. According to the site situation (Figure 1), the model can be divided into hammer and soil in two parts and simplified as a axisymmetric model (Figures 2 and 3). In order to simulate the transmission effect on boundary, the part of soil is divided into finite element zone and infinite element zone.

As shown in Figure 4, the radius of whole model is 120m while the radius of finite element zone is 60m, and outside the zone is the infinite element zone so that the distance from the nodes in the extending direction of infinite element to the pole is twice the distance from the points in the interface to the pole. The function of infinite element is similar with damping boundary condition that when the wave transmits to the interface the reflected wave disappears.

FIGURE 1 The site of dynamic compaction construction
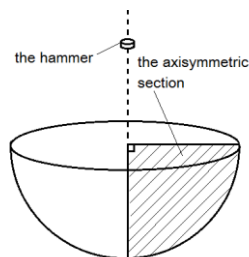
FIGURE 2 The schematic diagram of axisymmetric mechanical model of dynamic compaction
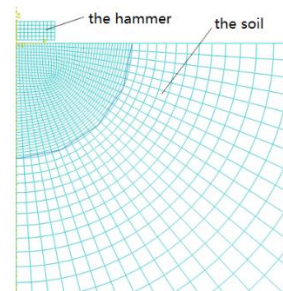
FIGURE 3 The mesh of the FEM axisymmetric mechanical model
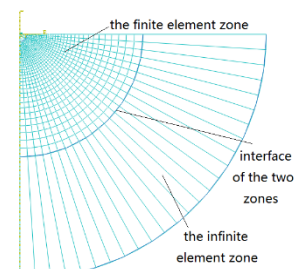
FIGURE 4 The finite element and Infinite element zone in the soil

2.3 CONSTITUTIVE MODEL OF SOIL

In FEM the Drucker-Prager model is often applied to the geo-material in simulation of elastic-plastic dynamic problem. This paper uses the linear Drucker-Prager model expanded from the classical Drucker-Prager model [13]. The yield function of this model is:

$$F = t - p \tan\beta - d = 0 ,\tag{2}$$

where $t = \dfrac{q}{2}\left[\left(1+\dfrac{1}{k}\right) - \left(1-\dfrac{1}{k}\right)\left(\dfrac{J_3}{q}\right)^3\right]$, β is the dip angle of $p \sim t$ stress space that could be conversed by the friction angle; k is the intensity ratio of triaxial extension to triaxial compression; d is the intercept of the yield surface on the $p \sim t$ stress space. J_3 is the third invariant of the deviator stress tensor. The mechanical parameters of the backfill soil are shown in Table 1.

TABLE1 The main input parameters of the soil

Specific weight γ (kN·m^{-3})	Elastic modulus E (Mpa)	Poisson's ratio v	Friction angle φ	Dilation angle θ	Flow stress ratio k
19	12	0.4	25	5	0.778

3 Analysis of the dynamic responses

3.1 TIME HISTORY CURVES OF VELOCITY AND ACCELERATION

When studying the effects on the structures that caused by the ground vibration, a viewpoint [14] thinks of that the vibration velocity should be considered importantly because the dynamic stress of the structure is related to the vibration velocity of the ground, at the same time there is another opinion [15] that by using acceleration the inertial force of structure can directly obtained and then the inertial force of the structure can be calculated, so the acceleration should take the first place. Considering the both aspects, this paper displays the time history curves of vibrational velocity and acceleration respectively. Figure 5 shows the monitoring and simulated time history curves of vibration at different distances. The result of simulation shows the characteristics that the main frequency of the vibration is about 10Hz and the time history curves picked more

closely to the tamping point have an obvious peak value, which is significantly greater than the nearby peaks. At the distance of 6.7m this peak value of velocity and acceleration are about 18mm/s and 1.9m/s2, respectively. While at the further place, due to the dispersion of wave packet in the process of propagation, there is not a distinctive peak, and the vibration attenuates more slowly than that in nearby.

In the test, the sensors laid in a line that crossed by the tamping point at different distances. Now the representative monitoring values collected from the sensors, which are 6.7m, 14.7m and 26.7m away from the tamping point are compared with the simulated values. As Figures 5 shows the simulation and monitoring curves picked from the closer points are more similar than the ones picked from further points. This is because in the simulation the material is isotropic but actually the backfill soil is complex in component and uneven in distribution. In the shorter range the differences are not evident because the wave has transmitted just for a short distance that the differences cannot fully exhibit, but in long range the changes accumulate that the differences are obvious.

The velocity curves in 6.7m and 14.7m to a certain extent are similar, but in 26.7m the similarity reduces. It is because on one hand with the increasing of distance, the vibration energy reduces and the ratio of background vibration increases, and on the other hand, in the process of propagation the wave packet gradually disperses, so the waveform changes. Furthermore, the duration of vibration increases as the distance increases, but vibration has basically subsided in 1s, and in this aspect the simulation values are the same with the monitoring value.
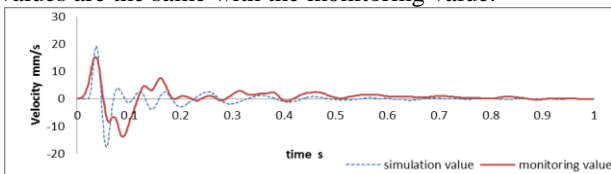
a) Time history curves of radial velocity at distance of 6.7m from the tamping point
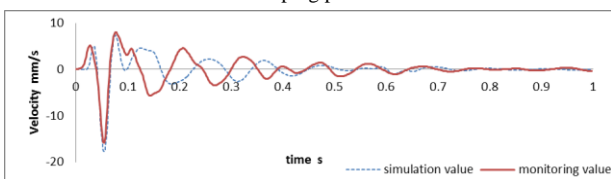
b) Time history curves of vertical velocity at distance of 6.7m from the tamping point
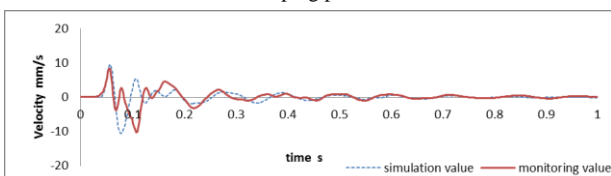
c) Time history curves of radial velocity at distance of 14.7m from the tamping point
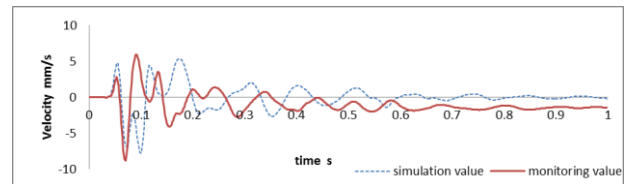
d) Time history curves of vertical velocity at distance of 14.7m from the tamping point
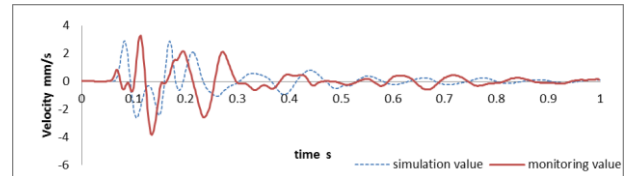
e) Time history curves of radial velocity at distance of 26.7m from the tamping point
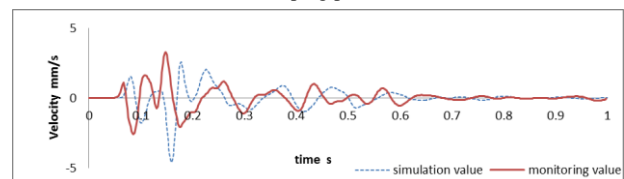
f) Time history curves of vertical velocity at distance of 26.7m from the tamping point
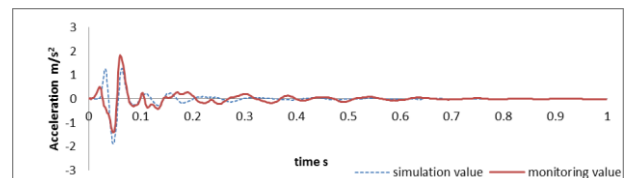
g) Time history curves of radial acceleration at distance of 6.7m from the tamping point
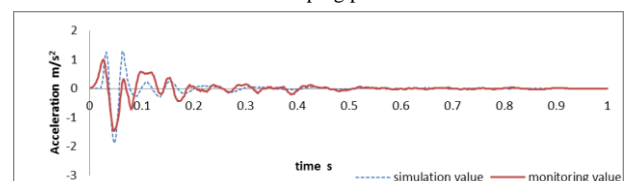
h) Time history curves of vertical acceleration at distance of 6.7m from the tamping point
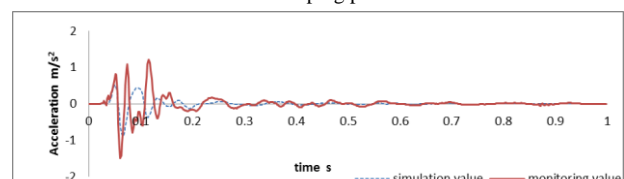
i) Time history curves of radial acceleration at distance of 14.7m from the tamping point
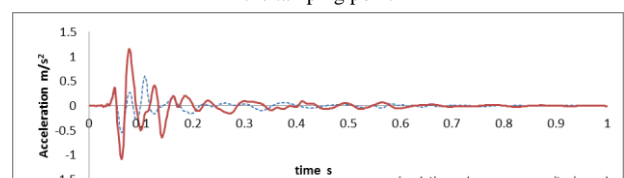
j) Time history curves of vertical acceleration at distance of 14.7m from the tamping point
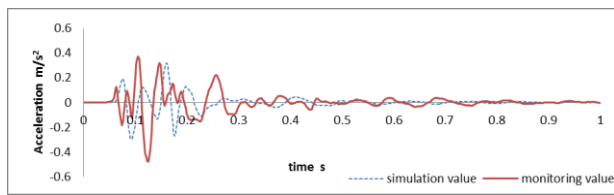
k) Time history curves of radial acceleration at distance of 26.7m from the tamping point
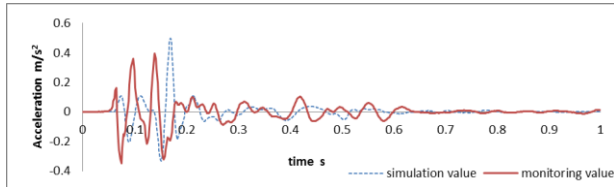
l) Time history curves of vertical acceleration at distance of 26.7m from the tamping point

FIGURE 5 The curves of vibration velocity and acceleration at different distance from the tamping point on the ground

Figure 6 shows the monitoring and simulated vibration frequency spectrum. Through the picture, it can be seen that the two frequency spectrum curves have the similar change tendency, but the high frequency component of the simulated frequency spectrum curves accounts for a larger proportion. Both the two curves have a peak value at about 10Hz. With the increasing of distance, the frequency spectrum concentration decreases and become more abundant.

By the comprehensive analysis above, it can be concluded that the simulated dynamic response, which is in the point that closer to the tamping point is more accurate, and the simulation to the low frequency component is better than the high frequency component.

a) radial velocity frequency spectrum: $r = 6.7$m

c) radial velocity frequency spectrum: $r = 26.7$m

b) vetical velocity frequency spectrum: $r = 6.7$m

d) vertical velocity frequency spectrum: $r = 26.7$m
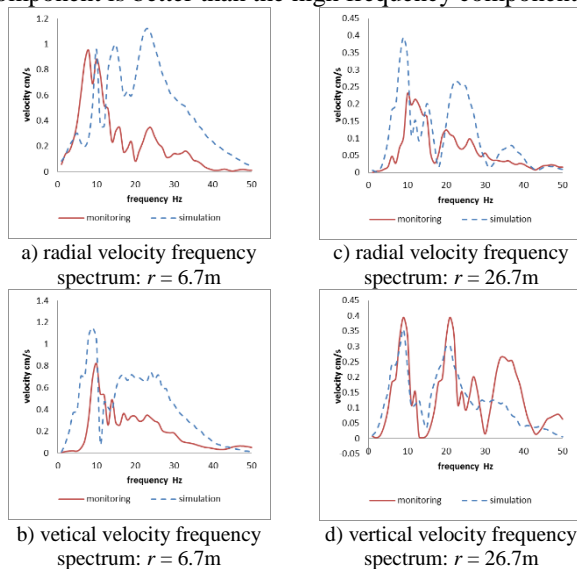
FIGURE 6 Monitoring and simulated vibration frequency spectrum

3.2 RELATION BETWEEN VIBRATIONAL PEAK VALUES AND THE DISTANCE FROM THE TAMPING POINT

In the impaction, the vibrational energy spreads from the tamping point to the environment so the amplitude decreases with the increasing of distance. Studying the relation between distance and the peak vibration could provide reference for determining safe distance in dynamic compaction construction. Figure 6 shows the relation between velocity and distance as well as that between acceleration and distance. The result indicates that the vibration reduces in the form of a negative power function or a negative exponential function and that the radial vibration curves decline faster than the vertical ones. Furthermore, in a same point, the radial peak velocity and acceleration are greater that the vertical ones. The shorter the distance, the more obvious the difference is. By comparing this result with the measured values, it could be found that the law is roughly the same.

Because the ability of ordinary building to resist horizontal vibration is worse than that to resist vertical vibration, in the construction of dynamic compaction, the influence of radial vibration should be considered firstly. The different structures usually have different seismic capacity. The specific values [14] are shown in Tables 2-4.

TABLE 2 Chinese standards for blasting safety

Structure type	Safety vibration velocity (mm/s)
Soil cave, adobe house	10
Non seismic block buildings	20~30
Reinforced concrete frame building	50
Hydraulic Tunnel	100
Traffic tunnel	150

TABLE 3 Swiss standards for building and blasting

Structure type	Frequency range (Hz)	Peak particle Velocity (mm/s)
Steel, reinforced concrete structures	10~60	30
	60~90	30~40
Brick structure	10~60	18
	60~90	18~25
Masonry walls, wooden pavilion	10~60	12
	60~90	12~18
historic and sensitive buildings	10~60	8
	60~90	8~12

TABLE 4 German standards for building and blasting

| Structure type	Frequency range (Hz)	Resultant velocity (mm/s)
Industrial buildings and commercial buildings	<10	20
	10~50	20~40
	50~100	40~50
Residential building	<10	5
	10~50	5~15
	50~100	15~20

By combining Figure 7 with Tables 2-4, on the vibrational frequency of 10Hz, the area around the tamping point can be divided into three parts:

1) The safety zone. The zone is outward of 30m. In the zone peak velocity is under 5mm/s. all the structures in this zone are safe.

2) The slight vibration zone. The zone is 20m~30m away. In this zone, the peak velocity is about 5~8mm/s and the historic and sensitive buildings will be affected.

3) The middle vibration zone. The zone is at about 7m~20m. In this zone, the peak velocity rises to 10~18mm/s. Ordinary brick buildings and wooden pavilions in this zone will be damaged. (4)The strong vibration zone. The zone is within 7m. In the zone peak velocity is up of 20mm/s and ascends quickly with the decreasing of distance. The steel reinforced concrete structure is threatened in this zone.

It needs to point out that all the standards have not considered the vibrational duration. In fact, the vibrational duration is also an important factor. The duration of dynamic compaction is usually within 1s, so the duration has little influence to the vibration effect. Besides, the cumulative damage of structures caused by repeat tamping is also not considered in the standards, so it needs to further research to clarify the relation of structure's seismic resistance and the vibrational parameters under repeat tamping.
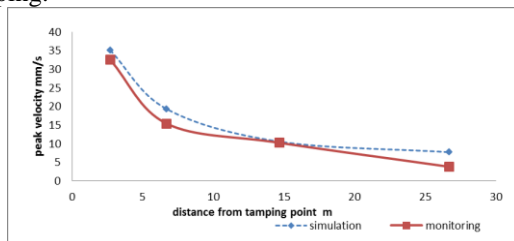
a) The relationship of the peak radial velocity and the distance from the tamping point
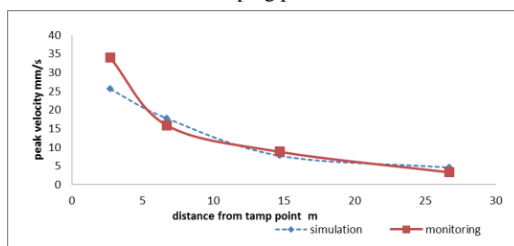
b) The relationship of the peak vertical velocity and the distance from the tamping point
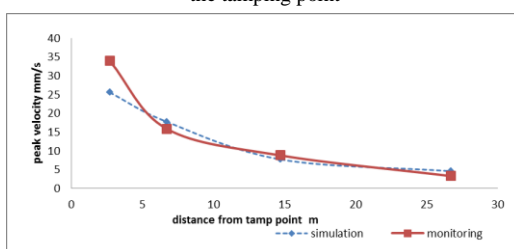
c) The relationship of the peak radial acceleration and the distance from the tamping point
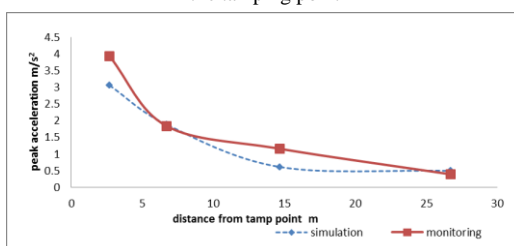
d) The relationship of the peak vertical acceleration and the distance from the tamping point
FIGURE 6 The relation between vibrational peak values and distance from the tamping point

3.3 RELATION BETWEEN VIBRATIONAL PEAK VALUES AND THE DISTANCE FROM THE TAMPING POINT

The total energy in the dynamic compaction is $E = mgh$, in which m is the weight of hammer, h is the drop height. During the impaction a part of the total energy is dissipative in the form of sonic, heat, friction and other way. The other part that consolidates the soil under the hammer in a certain range by doing plastic work called useful energy E_u. Precisely this part of energy makes a plastic zone in the foundation. In the zone, the strength parameters rise so that some scholars call the depth of this zone "satisfactory improvement depth".
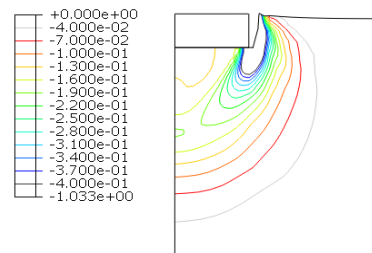
FIGURE 8 The contour map of minimal principal plastic strain
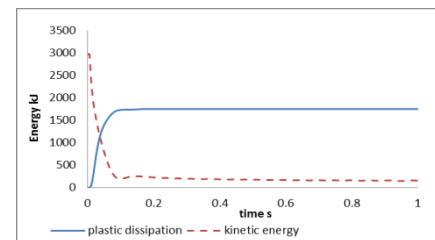
FIGURE 9 The change of kinetic energy and the plastic energy in the simulation

Figure 8 shows the contour map of minimal principal plastic strain in the foundation. In this picture, the plastic zone presents approximate ellipsoid. The energy utilization is $\eta = E_u / E$. Under different conditions include the type of soil, the weight and the drop height of the hammer, η has different values. According to the classical researches, the η is usually about 33%, but some researchers have the results that η may reach 69% in the Clay soil. In this paper, calculated from Figure 9, $\eta = 58.3\%$. It means that 41.7% of the total energy spreads to the environment, including the part, which causes the vibration on the ground. So under the same conditions, as the energy utilization rising, the vibration energy decreases. From paper [16], η is increasing function of m / S, where S is the bottom area of the hammer. It is to say under the same tamping energy level, the vibration can be reduced through reducing the bottom area or increasing the aspect ratio of the hammer.

4 Conclusions

1) In the backfill soil foundation, the duration of the vibration caused by dynamic compaction increases with the increasing of distance, while the amplitude decreases

with the increasing of distance and decays in the form of negative exponential function.

2) The simulated dynamic response, which is in the point that closer to the tamping point is more accurate, and the simulation to the low frequency component is better than the high frequency component.

3) Relative to the peak amplitude of vertical vibration, the radial one is greater, and the shorter the distance, the larger the difference is. In addition, Relative to the vertical vibration, the radial vibration is more threatening to the surrounding structures in the backfill soil in this case, so it is recommended that on the backfill soil foundation, the

radial vibration should be considered more in the general construction of dynamic compaction.

4) According to standards, the safe distance of different type of structures can be evaluated by the peak vibration velocity, but it needs to further research to clarify the relationship between structure's seismic resistance and the vibrational parameters under repeat tamping.

5) The energy that does plastic work to the soil is useful energy, and the soil will form an ellipsoidal plastic zone by this part of energy. The energy utilization is the ratio of the useful energy to the total tamping energy. Raising the aspect ratio of the hammer can raise the energy utilization and then reduce the vibration energy.

References

[1] Menard L F, Broise Y 1975 Theoretical and practical aspects of dynamic consolidation *Journal of Geotechnical Engineering* **25**(1) 3-18

[2] Scott R A, Preace R W 1975 Soil compaction by impact *Geotechnique* **25**(1) 19-30

[3] Mayne P W, Jones S J 1983 Impact stress during dynamic compaction *Journal of Geotechnical Engineering* **109**(10) 1342-6

[4] Chow Y K, Yong D M, Lee S L 1992 Dynamic compaction analysis *Journal of the Geotechnical Engineering* **118**(8) 1141-57

[5] Kong L W and Yuan J X 1999 Stress field distribution characteristics of foundation during dynamic consolidation and its application *Rock and Soil Mechanics* **20**(3) 13-9 *(in Chinese)*

[6] Niu Z R, Yang G T 2006 Dynamic characteristics of soils during and after dynamic consolidation *Engineering Mechanics* **23**(3) 118-25 *(in Chinese)*

[7] Cai Y Q, Chen R W, Xu C J 2005 Numerical analysis of dynamic compaction using large deformation theory *Journal of Zhejiang University(Engineering Science)* (1) 66-70 *(in Chinese)*

[8] Jiang P, Li R Q, Kong D F 2000 Numerical analysis of large deformation impact and collision properties during dynamic compaction *Chinese Journal of Geotechnical Engineering* **22**(1) 222-6 *(in Chinese)*

[9] Song X G, Lu S S, Li W Y 1999 Study on dynamic consolidation by finite dynamic element method *Journal of Hohai University* (5) 22-5 *(in Chinese)*

[10] Wang G Y, Hu Z N, Kuang X L 2008 Large-strain numerical simulation and experimental result research about improving red-sandstone embankment by dynamic compaction *Rock and Soil Mechanics* **29**(9) 2451-6 *(in Chinese)*

[11] Seaman L 1987 Analysis of dynamic in situ backfill property tests Part 2, An improved Lagrangian analysis for stress and particle velocity gage array *Technical report SL-87-11*

[12] Zhang X, Wang T S 2007 *Computational Dynamics* Tsinghua University Press Beijing *(in Chinese)*

[13] Fei K, Zhang J W 2010 *Application of ABAQUS in Geotechnical Engineering* China Water Power Beijing *(in Chinese)*

[14] Yan Z X, Wang Y H, Jiang P, Wang H Y 2003 Study on measurement of blast-induced seism and building safety criteria. *Chinese Journal of Geotechnical Engineering* **22**(11) 1907-11 *(in Chinese)*

[15] Meng J F, Hui H B 1992 The blasting test technology *Metallurgical Industry Press* Beijing *(in Chinese)*

[16] Wang S G, Liu S Y, Fang L 2002 Problems of energy of tamping impaction. *Chinese Journal of Geotechnical Engineering* **24**(3) 290-3 *(in Chinese)*

Authors

Siqiang Wen, born in January, 1987, Guangxi, China

University studies: master student at China University of Petroleum, Beijing.
Scientific interests: rock and soil mechanics and its application in engineering.
Publications: 1 paper.

Yunpeng Li, born in December, 1956, Shanxi, China

Current position, grades: professor at China University of Petroleum, Beijing.
University studies: Solid Mechanics in Xi`an University of Science and Technology.
Scientific interests: mechanical response of engineering structures in complex conditions and security and stability analysis and evaluation.
Publications: 124 papers.

Yan Zhou, born in September, 1987, Henan, China

University studies: Master student at China University of Petroleum, Beijing.
Scientific interests: structural mechanics and its applications in engineering.

Research on planetary bevel gear CVT system based on contact force

Jing Jing Liang[1*], Rui Qin Li[1], Jia Jun Ren[2]

[1]*School of Mechanical and Power Engineering, North University of China, No. 3 school road, Taiyuan, China*

[2]*School of Mechanical Engineering, Taiyuan University of Technology, No.79 West Yingze Street, Taiyuan, China*

Received 1 August 2014, www.cmnt.lv

Abstract

A virtual prototyping model for dynamical characteristic curves based on contact force was established, through the joint modelling (geometric modelling and constraint modelling) of planetary bevel gear CVT system in UG and ADAMS. The virtual prototyping experimental data proved that the system has the feasibility of over-zero variable speed. It is also verified that the model has advantage on continuously variable speed range, compared with planetary cone ring continuously variable transmission system. The main effect factors of the continuously variable speed performance and output torque are obtained. This model could be used to further study on such issues.

Keywords: planetary bevel gear, continuously variable transmission, virtual prototyping experimental, contact force

1 Introduction

Planetary-type continuously variable transmission device with the central roller in planetary motion transmits driving force through the traction force (friction) resulting from rolling. It realizes the transmission by altering either the working radius of sun wheel or planet wheel，which features the advantages in wide range of transmission and good output performance [1]. Nonetheless, some defects are also found like low transmission efficiency and friction loss because the driving force transmission is achieved through the friction force based upon rolling mechanism. Therefore it is of great practical significance and application value to explore a new type of planetary-gear-type continuously variable transmission system with better speed governing performance and lower friction loss.

Scholars both in China and abroad currently focus on the systems with belt-type transmission and V belt mixed with planetary gear transmission and multiple planetary gear grouping [2-6] making less research on the Planetary Bevel Gear continuously variable transmission system. Besides, owing to the complexity of the working conditions of the systems, it is found necessary to carry out the model building and simulation analysis in advance in order to give correct prediction and guidance. The author, in the light of dynamics, applies UG and ADAMS to modelling and simulation analysis so as to obtain a simulation result based upon relevant dynamics, which stands very meaningful in the further research of the subject.

2 Structure and principle of planetary bevel gear continuously variable transmission system

Planetary Bevel Gear Continuously variable transmission System is a bevel gear planetary system in which the structure improvement is made by replacing the friction wheel with bevel gear transmission that passes driving force and the speed governing realization is obtained by changing the contact radius between the speed ring and speed cone. Its structure is shown in Figure 1.
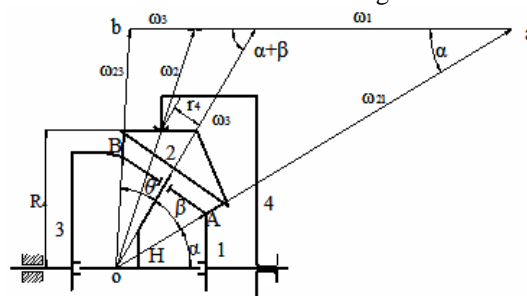

FIGURE 1 Structural diagram of Planetary Bevel Gear step-less speed transmission system

The speed ring keeps static and no turning at work but when the transmission is needed, the speed changing device will allow it to move horizontally along the conical surface of the speed cone. From the output axis different rotating speed and torque can be obtained by changing the contact position between the speed ring and the conical surface as well as the corresponding working radius. This paper aims at the exploration and study on the motion properties by applying virtual prototype technology to obtain the curve of the dynamics characteristics and make a good study of the step-less speed change performance of the system.

---

[*] *Corresponding author* e-mail: s2003126ljj@sina.com

3 Virtual prototype model establishment of main transmission

3.1 GEOMETRIC MODELING

In the previous study, the author has obtained the configuration parameters of the new type planetary step-less speed transmission system, which are optimized is shown in Table 1. Where: m – modulus, z_1 – teeth number of the bevel gear at input terminal, z_2 – teeth number of Planetary Bevel Gear, z_3 – teeth number of bevel gear at output terminal, δ_1 – reference cone angle of bevel gear at input terminal, δ_2 – reference cone angle of Planetary Bevel Gear, δ_3 – reference cone angle of bevel gear at output terminal, value is shown in Table 1.

TABLE 1 Configuration parameters

Parameter	Symbol	Number of Parameters
Modulus	m	$m = 2$
Teeth	z_1, z_2, z_3	$z_1 = 19, z_2 = 21, z_3 = 41$
Sub-degree cone angle	$\delta_1, \delta_2, \delta_3$	$\delta_1 = 28°, \delta_2 = 30°, \delta_3 = 92°$

All the above parameters are used for geometric modelling against various bevel gears, while doing it, each reference cone angle has to be adjusted according to the correct teeth clenching transmission to ensure an equal length of the reference cone generatrix for all teeth clenched bevel gears to satisfy that the teeth of three bevel gears are correctly clenched. In addition, various shared top clearance gear are to be selected for gear tip cones, gear root cones, and gear reference cones when the installation conditions for the bevel gears are taken into account. With a view of what has been stated above, a full bevel gear geometry model is obtained .once we put in the bevel gear geometrical dimensions calculation expression and involutes curve expression in UG, and through Boolean calculation from the curve to the curved surface and to a single tooth. No further description should be repeated here for other components like planetary frame, speed cone, speed ring and axis as well as their modelling based on their assembly relation and dimensions.

Upon the establishment of the model, the initial assembly can be made as per such conditions as the coincident constraint of the output terminal, input terminal bevel gear axes with output terminal, input terminal axis, planetary frame horizontal axes and speed ring axes; the coincident constraint of the planetary frame tilt axis with Planetary Bevel Gear and speed cone axes.

3.2 CONSTRAINT MODELING

Constraint modelling can guarantee the truth and reliability in the simulation analysis. During the process, necessary measures are simplified and the reliability of the result from the simulation analysis can also be ensured [7,

8]. The geometric models established in UG are introduced into ADAMS and based upon the transmission principle and motion relation demonstrated in the Planetary Bevel Gear step-less speed main transmission system, the following constraints are added to the system model.

Joint with fixture pair: input terminal bevel gear and input axis; output terminal bevel gear and output axis; speed cone and Planetary Bevel Gear.

Joint with rotation pair: input terminal bevel gear and planetary frame; connected components of speed cone and Planetary Bevel Gears with the planetary frame; output terminal bevel gear and the planetary frame; output terminal axes and earth.

A model based on Contact Force [9] has to be set up to realize the dynamic simulation for this system to simulate the real body contact. Therefore the constraints based upon Contact Force will be added between input terminal bevel gear and Planetary Bevel Gear; Planetary Bevel Gear and output terminal bevel gear; speed cone and speed ring.

In this paper Contact Force is added by using IMPACT function selected from ADAMS function library, which substantially has been simulated as a nonlinear spring damper. The contact force is composed of two parts: first the spring force which is generated from the mutual cut-in of the two components; the other the damper caused by the relevant velocity [10]. During the analysis of the Contact Force, the Contact Force array is used to define its characteristics. The parameters in the Contact Force array determines the related parameters in the IMPACT function and the main parameters include rigidity k, damped coefficient c, the force nonlinear index e, in-going depth d, static friction coefficient μ_s, dynamic friction coefficient μ_d, static Slip velocity v_s, dynamic Slip velocity v_d [11] and the selection of the parameter value is set by the empirical value is shown in Table 2.

TABLE 2 Empirical value table for contact force parameters

Parameter	Symbol	Number of parameters
Rigidity	$k(N/mm)$	10^5
Damped coefficient	$C(N \cdot S^{-1}/mm)$	50.000
Force nonlinear index	$d(mm)$	0.1
In-going depth	e	1.5
Static friction coefficient	μ_s	0.08
Dynamic friction coefficient	μ_d	0.05
Static slip velocity	$v_s(mm/s)$	0.1
Dynamic slip velocity	$v_d(mm/s)$	1

Given the speed ring is in translational motion, the motion sub-constraint is added between the speed ring and the earth.

Rotating motion stimulation (motion 1) is added to the revolute joint between the input terminal bevel gear and planetary frame. Translational motion stimulation (motion 2) is added to the flat pair between the speed ring and the earth.

Drive add-in: constant rotation speed drive is added to the system input axis and translation drive is added to the speed ring. The constant rotation speed in this paper is set for the system input axis: $1500r/min$, shift to $9000°/s$. Given work radius varying rate of the speed cone is $5mm$, the translation speed of the speed ring in the horizontal direction is converted as: $5/\sin 58° = 5.9mm/s$

Avoid in every way handling step signal in ADAMS resolver. After all, the two drives can be set in the step function form with the exorbitant curve，thus the system drives here are defined as the following: the input axis rotation speed is defined as: $STEP(time, 0, 0, 0.05, 9000d)$; the speed ring translation speed is set: $STEP(time, 0, 0, 0.05, 5.9)$. From the above, we get the virtue prototype simulation model is shown in Figure 2.
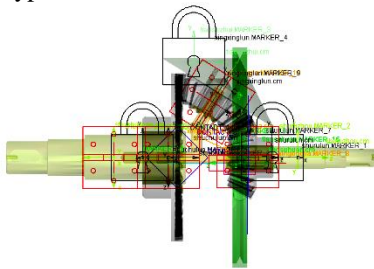
FIGURE 2 Virtue prototype simulation model diagram

4 Performance verification of speed modulation in virtue prototype stimulation model no-load operate

According to the above-set constraint and drive value, we have to ensure the system input axis rotates at a constant angle speed and the speed ring makes parallel move at a constant speed in order to find out the curve patterns in the output angle speed variation during the simulation of the moving process in which the system allows the speed ring to touch initially with the speed cone and then move away. After setting the measuring and simulating environment, we may start simulation and allow the after-treatment module to output the diagrams of the curves is shown in Figure 3.
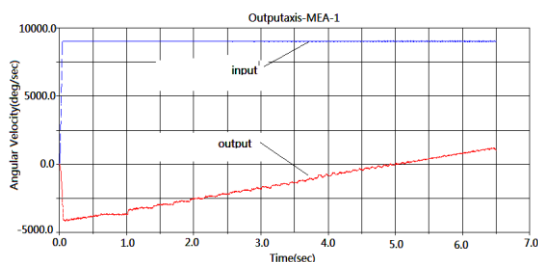
FIGURE 3 under no-load torque input & output rotation speed

4.1 CONSTANT LOAD TORQUE AT WORK

When the input axis reaches the constant speed: $1500r/min$, to the output axis we add constant load torque $6000N·mm$ and with step function definition we add $STEP(time, 0, 0, 0.05, 6000)$, thus the output angle speed curve is shown in Figure 4. From the result of the above

simulation, we perceive the continuous change within a certain range done by the output rotation speed in the Planetary Bevel Gear continuously variable transmission system and also the zero cross speed modulation enablement, which proves the feasibility of the continuously variable transmission, however, in the process of the simulation, the system occasionally shows the phenomenon that the transmission is done at a fixed transmission ratio. Through our study and research, we presume it is caused by the inertia of the Planetary Bevel Gear, for which an improvement of the system model will be made by setting Planetary Bevel Gear and speed cone linkages to a symmetric layout structure is shown in Figure 5.
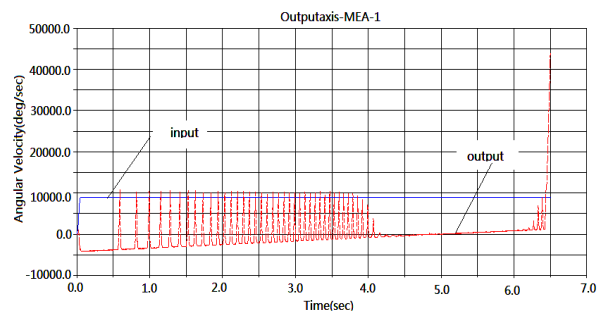

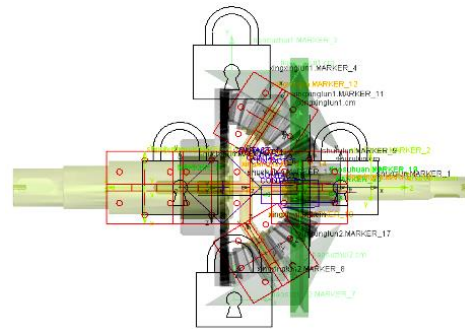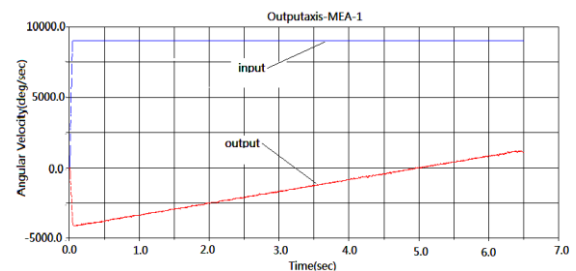FIGURE 4 Under constant load torque input & output rotation

FIGURE 5 Virtue prototype simulation model after modification

4.2 ANALYSIS ON THE MODIFIED MODEL SIMULATION

By giving the simulation analysis to both no-load running of the modified model and constant-load torque, we get the simulation curves is shown in Figures 6 and 7. From the above simulation curve analysis, we find the system model can realize a stable continuously variable transmission upon the improvement, with the transmission rate in the range of $-0.45 \sim 0.14$.

FIGURE 6 Under no-load torque input & output rotation speed diagram of the modified model when the input rotating speed = 9000d/sec
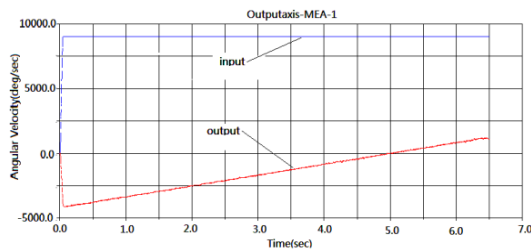
FIGURE 7 Under constant load torque input & output rotation speeddiagram of the modified model when the input rotating

5 Simulation analysis on the system output torque

The output torque presents an important indicator to judge the performance of the continuously variable transmission, of which the parameter simulation is carried out by configuring at the system terminal the input power and constant input rotation speed, meanwhile ensuring a stable running of the whole system with the condition that a certain load torque is given to the output axis, thus the output torque curves are recorded from the output axis. The simulation analysis in this paper is conducted according to the situations provided as following.

5.1 SIMULATION ANALYSIS OF THE OUTPUT TORQUE UNDER THE CONSTANT POWER AND CONSTANT LOAD TORQUE

Setting input power $P = 2.2kW$, load torque $60N \cdot m$, input axis rotation speed $n_1 = 1500r / \min$, $n_1 = 1000r / \min$, $n_1 = 750r / \min$ respectively, we give the simulation analysis to the output torque on the output axis, and the curves can be obtained is shown in Figures 8-10.

From the simulation result analysis, we know that, once the input power and load torque have been rated, the output axis' output torque makes a continual change if the rotation speed of the input axis is changed. The position and scale of the maximum output torque also change: with the continual decrease of the input rotation speed, the maximum output torque continues to increase.
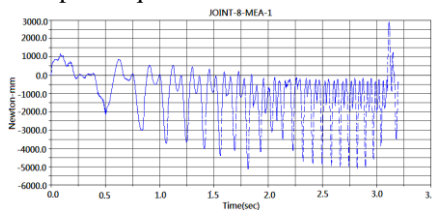
FIGURE 8 Output torque diagram when input rotation speed = 1500r/min
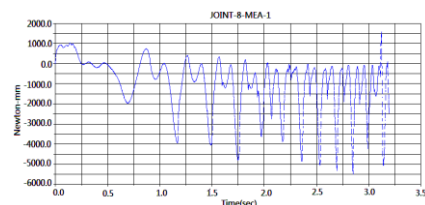
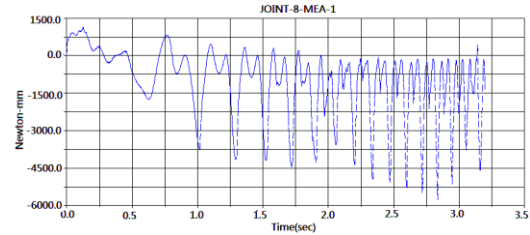FIGURE 9 Output torque diagram when input rotation speed = 1000r/min

FIGURE 10 Output torque diagram when input rotation speed = 750r/min

5.2 OUTPUT TORQUE SIMULATION ANALYSIS UNDER CONSTANT POWER AND CONSTANT INPUT ROTATE SPEED

A simulation analysis is made on the output torque over output axis while keeping the input power and input rotation speed at a fixed rate with variable load torque and system stable running. We set input power $P = 2.2kW$, input rotation speed $n_1 = 1500r / \min$, load torque respectively $0.5N \cdot m$, $5N \cdot m$, $50N \cdot m$, in each case, the output torques over the output axes are recorded and shown in Figures 11-13. From the simulation result analysis, we know that, once the input power and input rotation speed are kept unchanged given any change in load torque, no obvious change is witnessed from the output torque over the output axis, which shows that the load torque does not exercise a clear impact on the system output torque.
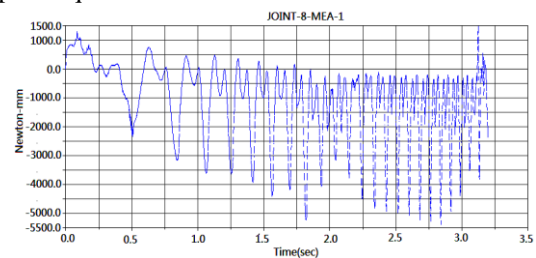

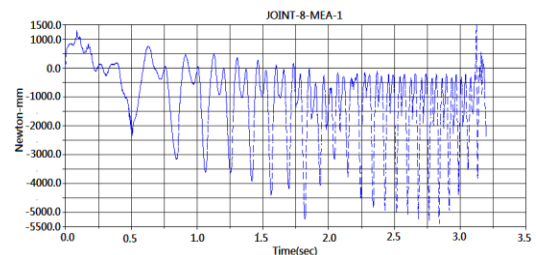FIGURE 11 Output torque diagram when the load torque = 0.5N·m

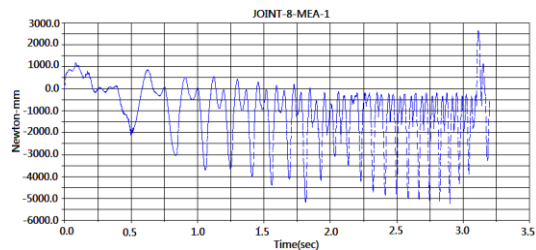FIGURE 12 Output torque diagram when the load torque = 5N·m

FIGURE 13 Output torque diagram when the load torque = 50N·m

6 Conclusions

The output rotation speed under the new type Planetary Bevel Gear continuously variable transmission is able to realize a continual change within a certain range, i.e. continuously variable transmission, and also zero cross speed modulation. This proves the feasibility of the system, and its transmission rate is $-0.45 \sim 0.14$. While the original ring cone transmission system provides a transmission rate. The system model after its improvement conducts a stable and continuously variable transmission, which testifies that there is some influence on the performance of the whole main transmission system's continuously variable transmission imposed by the inertia of both the Planetary Bevel Gears and speed cone linkages,

of which the amount can be changed to improve the performance of the system continuously variable transmission.

The input rotation speed becomes a vital element that affects the output torque. The maximum output torque continues to increase with the continual decrease of the input rotation speed. On the other hand, the influence on the output torque resulted from the load torque is comparatively small.

References

[1] Ruan Z 1999 A Guide to Mechanical Continuously variable transmission Design and Choice *Chemical Industry Press* Beijing Chapter 5
[2] Sun D, Qin D 2008 *Journal of Nanjing Engineering Academy* **6**(4) 9-22 *(in Chinese)*
[3] Duan Q, Yang S 2001 *Journal of Mechanical Design and Research* **17**(3) 40-3 *(in Chinese)*
[4] Zou Z, Gu X, Yang X 2000 *Journal of Agricultural Mechanical Journal* **31**(2) 9-11 *(in Chinese)*
[5] Kim Y S, Park J M , Choi S H 2006 *Journal of Mechanical Science and Technology* **20**(6) 770-81
[6] Bullinger M, Pfeiffer F, Ulbrich H 2005 *Journal of Multibody System*

Dynamics **13**(2) 175-194
[7] Huang H, Zou F 2009 *Journal of Mechanical Design and Research* **6**(3) 114-116 *(in Chinese)*
[8] Feng W, Chen X, Ma Z 2007 *Journal of Liaoning Engineering Technology University* **1**(3) 120-2 *(in Chinese)*
[9] Zhu C, Liu T, Zhu W 2010 *Journal of Hydraumatic Aeromechanics and Seal* **4**(5) 29-31 *(in Chinese)*
[10] Li Z 2006 ADAMS Introduction with Detailed Annotation and Examples *National Defense Industry Press* Beijing 56-96
[11] Shu Z, Shu G 2009 *Journal of Mechanical Transmission* **32**(1) 15-19 *(in Chinese)*

Authors

Jing Jing Liang, born in January, 1975, Taiyuan, China

Current position, grades: lecturer of Mechanical and power Engineering.
University studies: Master's degree at Taiyuan Institute of Technology in 2006, Bachelor's degree at Baotou Iron and Steel Institute in 1997.
Scientific interests: mechanism and robot theory and simulation technology.
Publications: 1 patent, 5 academic papers, 2 textbooks.

Ruiqin Li, born in November, 1964, Taiyuan, China

Current position, grades: professor, PhD supervisor, North University of China.
University studies: PhD degree at Shanghai Jiao Tong University in 2004.
Scientific interests: mechanism and robot theory and simulation technology
Publications: 1 monograph, 7 textbooks, 60 academic papers.

Jiajun Ren, born in May, 1958, Xiaoyi, China

Current position, grades: professor, Vice President of School of Mechanical Engineering of Taiyuan University of Technology.
University studies: Master's degree and Bachelor's degree from Taiyuan University of Technology in 1990/1982.
Scientific interests: optimization of mechanical design and CAD, mechanical transmission, ergonomics and design.
Publications: 2 monograph, 2 textbooks, 30 academic papers.

Vehicle durability test based on user survey

Hudai Fu*, Jingang Gao

School of Mechatronics Engineering, Changchun Institute of Technology, Changchun 130012, JiLin, China

Received 1 March 2014, www.cmnt.lv

Abstract

The vehicle durability test is an important means of assessing and verifying the vehicle reliability. At present, the domestic automobile enterprises generally have some problems such as nonstandard test methods, the test mode insufficiently associating to the users in the vehicle durability test. It proposes a method for establishing the vehicle durability test mode in the paper, based on the theory of fatigue damage. Through the users' survey, the load acquisition and data analysis are carried out from the testing ground and the social roads by using of six-component force tester. It is proved, that the vehicle durability test method can not only shorten the test cycle, but also increase the appearance and countermeasures of adverse conditions. Meanwhile, it can reduce many kinds of adverse problems occur after the new vehicle coming into the market, and improve the quality of vehicle reliability.

Keywords: user survey, durability test, load collection, damage value

1 Introduction

The automobile durability is one of the quality problems which the users are most concerned about. The durability as an important indicator to measure the automobile quality has long been attached great importance by the automobile production enterprises. Because of only those automotive products with good reliability can be durable and popular with users. The vehicle durability test is the most direct and effective evaluation of automotive reliability and durability. The automobile enterprises should carry out the vehicle durability test during the new vehicle design and trial production phase, in order to verify the automobile reliability, and reduce the adverse problem after production [1].

At present, the advanced automobile enterprises in American, in Europe and in other countries have paid more and more attention to the automobile durability. A complete set of durability management system from the design to the service has been established, in order to improve the reliability of the automobile. The automobile durability has become the main factors of its products to win in the market competition. John Hollenbeck took the application of automobile electronic control modules in the vehicle as an example, to discuss the preparation and the implementation method of the durability technology during automobile product development [2]. Early durability tests in China are most to ensure the vehicle component does not break, damage, and can meet certain

quality requirements, during the most extreme conditions. The test mode mainly depends on the experience value. It results in the insufficient association with the users.

In the paper through the investigation on the user's habits and the road environments, combing with the test road conditions in a proving ground, getting the load acquisition and data analysis from the testing ground and the social road by using of six-component force tester, one vehicle durability test mode has been proposed and improved. It can improve the reliability and users' satisfaction, provide the basis for reducing the cost, simplifying the design.

2 Vehicle durability test mode set-up process

The establishment of automobile durability mode derives from the surveys of users' habits, the environment and the real test field conditions. The automobile durability mode will be built on the basis of achieving the endurance test goal, by contrasting the test mode with the social roads intensity. Please see Figure 1.

3 Fatigue damage theory

3.1 S-N CURVE

The cycle of stress or strain before material fatigue failure is called fatigue life. The curve expressing the relationship between the stress (strain) level S and the standard sample fatigue life N is known as the material S-N curve [3].

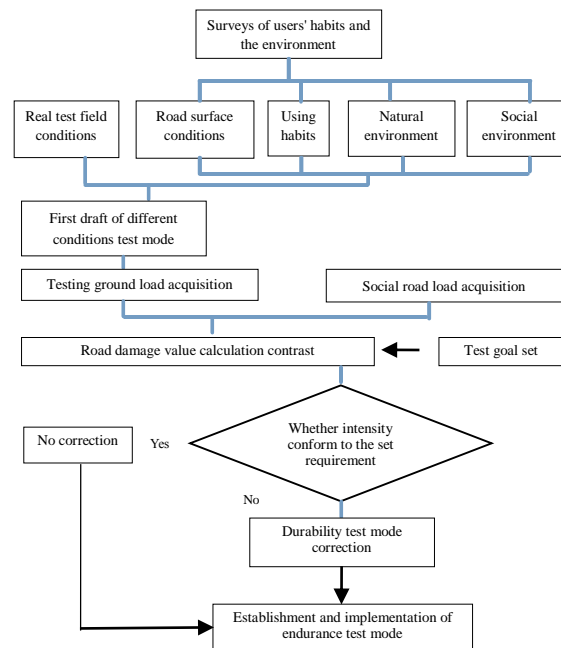---

Corresponding author e-mail: fuhudai@126.com

FIGURE 1 The overall set-up process

3.2 CUMULATIVE FATIGUE DAMAGE THEORY

Fatigue fracture is a process of cumulative damage. At present, there are many kinds of theories and methods for calculating the fatigue life of parts under random load [4]. Miner linear cumulative fatigue damage theory has simple expression, and a considerable degree of agreement between its life estimation and the test results in most cases. So it is currently the most common method for fatigue life prediction. The total damage by Miner cumulative fatigue damage theory is shown in Equation (1).

$$D = \sum_{i=1}^{n} di = \sum_{i=1}^{n} \frac{ni}{Ni} . \qquad (1)$$

In Equation (1), $Ni \rightarrow$ the damage cycles on $\sigma - n$ curve corresponding to the stress level σi ;
$ni \rightarrow$ the actual work cycles under the action of the stress level σi ;
Let N_0 be the limit life corresponding to the fatigue limit σ_{-1} , according to the Basquin relationship, there is:

$$Ni = \left(\frac{\sigma_{-1}}{\sigma_i} \right)^m N_0 . \qquad (2)$$

In Equation (2), $\sigma_{-1} \rightarrow$ fatigue limit; $N_0 \rightarrow$ fatigue life corresponding to the fatigue limit;
Taking Equation (2) into Equation (1) can get calculation formula of the cumulative damage. See Equation (3).

$$D = \sum_{i=1}^{n} \frac{ni}{N_i} = \sum_{i=1}^{n} \frac{ni}{\left(\sigma_{-1}/\sigma i \right)^m N_0} = \sum_{i=1}^{n} \frac{ni\sigma_i^m}{N_0\sigma_{-1}^m} . \qquad (3)$$

According to Miner theory, the fatigue damage will happen when the whole material damage finishing.

4 User habits survey

The use habits and the road environment information include four aspects, which are the road surface conditions, users using vehicle habits, natural environment and social environment. According to the investigation information of these four aspects, the statistics for the users actually driving habits and the road surface features will be obtained. The establishment of durability test mode is based on these features.

1) Road environment, the road conditions and proportion of mountain road, freeway, city road and so on;
2) Use habit, intensity and frequency of braking, use frequency of hand brake, window, CD and so on;
3) Natural environment, maximum and minimum air temperature, sunshine, humidity and other natural factors when using vehicle;
4) Social environment, the quality of fuel, emission standards and other social environmental factors;

5 Road load collection and calculation method of fatigue damage

5.1 ROAD LOAD COLLECTION

The input force to the vehicle from the road is transferred through the tire suspension, shock absorbers and so forth. The sensors for the road load collection are arranged on the

vehicle input force transmission path. According to the principle, tire six-component force tester, strain gauges and acceleration sensors are arranged on the tire, the suspension upper and lower arms, damping spring and the spring seat. Please see Table 1.

TABLE 1 Road load collection illustration

CH	Position		Sensor	Unit
1			six-component force tester	N
2			six-component force tester	N
3	FR	L	six-component force tester	N
4			strain gauges	mm
5			acceleration sensor	G
6			acceleration sensor	G
7			six-component force tester	N
8			six-component force tester	N
9	FR	R	six-component force tester	N
10			strain gauges	mm
11			acceleration sensor	G
12			acceleration sensor	G
13			strain gauges	mm
14	RR	L	acceleration sensor	G
15			acceleration sensor	G
16			strain gauges	mm
17	RR	R	acceleration sensor	G
18			acceleration sensor	G

5.2 ROAD LOAD DATA PROCESSING METHOD

5.2.1 Load data pre-processing

Please see Figure 2. The original data collected from the load are acceleration, force, torque, and the curve when micro strain changes along with the time domain. The disturbed data on the curve should be removed. For example, the unexpected "cusp" on the acceleration curve is generally judged to be disturbed location, and the data corresponding to all channels at this time should be removed.
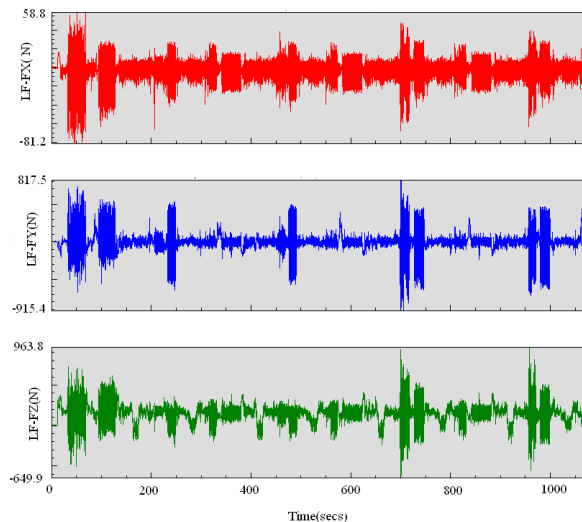

FIGURE 2 The original data collected from the load

The loads collected directly when the car in motion is continuous data changed over time, namely load-time, strain-time data. The load spectrum cannot be used directly for fatigue life estimation, because of the uncertainty of

random load. It must be for statistically processed. The load-time processed is called load spectrum. The load spectrum which is the graph with statistical properties can essentially reflect the load changes of parts. For random load, the statistical analysis method widely used is mainly counting method. The load-time processed into a series of complete cycle or half cycle is called counting method. The counting methods proposed from abroad have more than a dozen kinds. The two-parameter counting method now mostly used is rain-flow counting method. The method considers that the presence of plastic is a necessary condition of fatigue damage, and its plastic property must be the hysteresis loop of stress-strain [5].

The road load data collected are processed by rain-flow counting method. The continuous data changing over time are assigned into 100 intervals, and the output unit is times. Please see Table 2.

TABLE 2 Road load data collected by rain-flow counting method

Interval / Position	1	2	3	4	5	6	7
LF-FX	258026	63524	47322	28346	6035	2382	1160
LF-FY	125407	59336	22858	5172	1546	562	276
LF-FZ	48924	36955	11204	4326	1614	764	463
Interval / Position	8	9	10	11	12	13	14
LF-FX	724	462	289	254	148	92	114
LF-FY	117	73	45	20	21	15	13
LF-FZ	291	129	86	55	52	27	21

5.2.2 Calculation of fatigue damage value

The data processed by rain-flow counting method are fed into the material S-N curve. It will calculate the fatigue damage value Dm, which is the absolute value. Although the vehicle chassis and body are composed of a variety of materials, each material has its own S-N curve. So they cannot be simply fed to calculate the absolute damage value. Introducing the concept of relative damage value is needed here. The ideal S-N curve will be constructed. The input values of reference road are fed into S-N curve, and the Dm value is exactly 1. Then the road input values from the testing ground are fed into the ideal S-N curve. The damage value calculated can reflect the contrast between the pavement strength in testing ground and the reference road [6].

6 Establishment and improvement of durability test mode

6.1 TEST GOAL SET

In order to shorten the test cycle, the road durability test is generally conducted on the testing ground with various features, such as twisted road, stone road and long road etc. The enhancement coefficient that of the various strengthened roads relative to the ordinary roads must be confirmed, in order to make the results comparable. The enhancement coefficient is the ratio of mileage that of the

car running on the social roads and on the strengthened test roads under the same failure mode.

$$K = \frac{L_1}{L_0}.$$ (4)

In Equation (4), $L_1 \rightarrow$ mileages running on the social roads; $L_0 \rightarrow$ mileages running on durability test roads of the testing ground.

6.2 USER HABITS SURVEY

According to the method of user habits survey, many owners use habits, road environments and other data have been collected in a period of time. The user operation class and running class data which provide the basis for establishing the durability test mode have been extracted by analysing.

6.3 PRELIMINARY ESTABLISHMENT OF ENDURANCE RUNNING MODE

Based on user habits survey and one testing ground conditions, the vehicle durability test running mode for highway, city road, mountain road, bad road, traffic jam and other conditions has been preliminarily set up. According to the anchor point on the running track of the testing ground, the operation specification for the velocity, acceleration and braking in each interval is confirmed.

6.4 LOAD DATA ACQUISITION AND ANALYSIS ON TESTING GROUND AND SOCIAL ROAD

6.4.1 Load acquisition on testing ground

According to 5.1 road load data acquisition methods, three loading modes such as no-load, half load, full load are used after the six-component force tester, strain gauges and acceleration sensors are arranged on the experimental car. In accordance with the preliminary testing ground durability test mode driving, 10 groups of the original load data are collected. The data processed by a certain proportion will be added.

6.4.2 Social road load acquisition

Through the surveys of the user habits and the road environment, the users' driving road surface and speed constitution have been gotten. According to this information, the social characteristics roads are selected. The social road load data are collected in accordance with 6.4.1 loading mode. It will be used as the reference road to

generate S-N curve, and to compare with the testing ground durability test mode.

6.4.3 Load intensity contrast between the testing ground and social road

The damage value ratio of the testing ground relative to the social road is calculated, according to the methods described in 5.2. The items in each channel are not the same during collecting the road load in testing. So the calculation of damage value should be separately carried out. Please see Table 3.

TABLE 3 Road load comparison

Channel No.	Data type	Enhancement coefficient	Damage value comparison
1	LF-FX	3	3.26
2	LF-FY	3	4.21
3	LF-FZ	3	2.73

The enhancement coefficient is assumed to 3 in Table 3. By comparing the damage value in the testing ground and in the social road, the values of some channels are below or above the standard value. So we should adjust the proportion of some conditions in the durability mode. For example, the proportion of bad road conditions should be increased when the vertical damage value is relatively low.

6.5 ESTABLISHMENT AND IMPLEMENTATION OF ENDURANCE RUNNING MODE

The established and improved vehicle durability test mode based on the above method has been used into the reliability test of new vehicle developing. It is proved that the mode can effectively find the adverse conditions and propose the countermeasures in advance during the trial stage. It also can reduce the occurrence of the adverse conditions after new vehicle quantity production into the market.

7 Conclusions

With the rapid development of automobile industry in China, how to correctly, quickly and effectively carry out the vehicle reliability test is the research for the domestic automobile enterprises to do. A process and method for establishing the vehicle durability test mode and putting it into practical application have been proposed in the paper, based on the theory of fatigue damage. The results showed that the method cannot only effectively improve the vehicle durability design quality, but also enhance the users' recognition to the automobile products.

References

[1] Bertodo R 1987Reliability and durability of automotive prime movers. *Int J of Vehicle Design* **8**(2) 145-8

[2] Ensor D F 1999 Customer Usage Profiling and Accelerated Development *Journal of the Engineering Integrity Society* **6**(6) 892-8

[3] Bannantine J A, Comer J J, Handrock J L 2004 Fundamentals of metal fatigue analysis *New Jersey Prentice Hall Englewood Cliffs* 168-83

[4] Svensson T 1996 Fatigue damage calculations on block load sequences *Fatigue Fract.Engng Mater Struct* **19**(3) 251-64

[5] Khosrovanch A K, Dowling N E 1990 Fatigue loading history reconstruction based on the rainflow technique *International Journal of Fatigue* **16**(2) 20-4

[6] Lalanne C 2002 Fatigue damage *Mechanical Vibration and Shock* **4**(6) 157-61

Authors

Hudai Fu, born on Novembe 25, 1977, China

Current position: a researcher at Changchun Institute of Technology, China.
University studies: M.E. degree in mechatronics engineering from Changchun University of Technology, in 2004.
Scientific interest: mechatronics technology and intelligent detection.

Jingang Gao, born on September 9, 1976, China

Current position: a researcher at Changchun Institute of Technology, China.
University studies: M.E. degree in mechanical engineering from Huaqiao University, in 2004.
Scientific interest: mechatronics technology and intelligent detection.

Fuzzy control of flue temperature in coke oven heating process

Gongfa Li[1, 2]*, Wentao Xiao[1], Honghai Liu[1, 2], Guozhang Jiang[1], Jia Liu[1]

[1]*College of Machinery and Automation, Wuhan University of Science and Technology, Wuhan 430081, China*

[2]*Intelligent Systems and Biomedical Robotics Group, School of Computing, University of Portsmouth, Portsmouth, PO1 2DJ, United Kingdom*

Abstract

Coke oven production possesses the characteristics of nonlinear, large inertia, large disturbances, and highly-coup ling and so on. Coke oven heating temperature was reflected by flue temperature and adjusted by gas flow. The control method of intermittent heating control is adopted in traditional heating control system of coke oven, and cannot satisfy the command of coke oven heating control. The control principle of combining the intermittent heating control with the heating gas flow adjustment is adopted according to analysing the difficulty and strategy of heating control of the coke oven. On the basis of researching deficiency of the existing control strategy, fuzzy hybrid control is proposed to establish heating intelligent control model of coke oven, which combines feedback control, feed-forward control and fuzzy intelligent control. Carbonization index is utilized in the model to control coking management of coke oven. Then heating fuzzy intelligent control structure of coke oven is built. According to artificial experience and actual conditions, the fuzzy controller is designed. Fuzzy control can deal with fuzzy, inexact or uncertain information and is extraordinarily robust, which can realize intelligent control of heating process of coke oven. Better control result of temperature control is realized by fuzzy intelligent control model. Intelligent control methods were used to adjust stopping heating time and heating gas flow. The practical running results indicate that the system can achieve heating intelligent control of flue temperature, reduce temperature fluctuation, effectively improve quality of coke and decrease energy consumption, and has great practical value.

Keywords: coke oven heating, model, intelligent control, fuzzy control, hybrid control

1 Introduction

These guidelines, written in the style of a submission, discuss how to prepare your paper using Microsoft Word. In addition to the usual guidance on style/formatting, there are notes and links to assist in using some of Word's features such as inserting graphics, formatting equations and so forth.

The most important control in the production of coke oven is temperature control of coke oven [1-3], because temperature of coke oven is a key factor of affecting coke quality, economizing coal gas consumption, reducing smoke and dust pollution during charging coke. Practice in the coke oven has proved that it is of great difficulty in temperature control, which reflects concretely in several following aspects [4].

Coke oven has an uncommonly non-linear characteristic. According to characteristics of coke oven production, relationship between heating gas flow and longitudinal temperature at the top of the regenerator is affected by factors such as coke carbonization cycle, type of coal used in production, longitudinal temperature at the top of the regenerator and production plan of pushing coke and so on. Therefore, the uncommonly non-linear characteristic of coke oven is caused.

Heating process of coke oven possesses characteristics of great inertia characteristics and large time-delay. Because coke oven is bulky, hot capacity is very large, the course of heating-up and heating-down are all

extraordinarily slow. There are some difficulties in measurement of relevant crafts parameters. For example, it is difficult to implement online measurement of longitudinal temperature at the top of the regenerator, calorific value of gas and water content of coal charging of coke oven. Production process of coke oven belongs to the intermittent type, which is operated by single stove according to the operation plan.

A lot of disturbances exist in production process of coke oven. Because there are two kinds of heating gas, namely coke oven gas and blast furnace gas. Calorific value difference between them is great, and flow heating is different. Affected by climate, water content of coal fluctuates greatly. Thus, heating temperature of coke oven is influenced. The Production plan of pushing coke has significant effect on heating temperature of coke oven too. Heating process has the characteristics of time changeable in one cycle, which is affected by such factors as market demand, government policy. Thus, coke carbonization cycle of coke oven often changes, and the kind of heating gas always varies too. These changes cause a lot of interference factors to production process of coke oven.

2 Control methods of heating control system in coke oven

Heating control of coke oven can generally adopt three kinds of control methods, namely feedback control system,

NATURE PHENOMENA AND INNOVATIVE ENGINEERING

feedforward control system and compound control system combined feedback with feed forward.

2.1 TEMPERATURE FEEDBACK REGULATION SYSTEM ADOPTING TEMPERATURE MEASUREMENT IN SUCCESSION

According to deviation between real flue temperature, real finished carbonization time or real coke temperature and target flue temperature, target finished carbonization time or target coke temperature, considering time-delay factor of coke oven temperature [1], settlement value of heating gas flow is adjusted in order to realize the optimal heating control of coke oven. Meanwhile, some control systems provide operation guidance of regulating flue-chamber depending on judgment of carbonization of coke-chamber, wall temperature measurement of coke-chamber or coke button temperature and its distribution. Advantages of feedback control are the influence of various kinds of parameters, which need not be considered. The deviation of controlled parameters is utilized to control. Shortcomings of feedback control are that the disturbance cannot be overcome in time, and the time-delay phenomenon is serious, because feedback control just works after disturbance takes place. Feedback control system is shown as Figure 1.
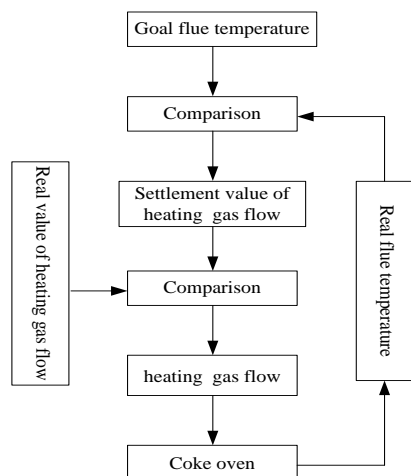
FIGURE 1 Feedback control system of coke oven

2.2 HEATION SUPPLY FEEDFORWARD REGULATION SYSTEM ADOPTION EXPERIENTIAL MODEL

Coking heating amount is calculated in the light of properties of coal and average humidity of coke button, and then coking heating consumption is computed through thermal balance according to amount of coal charging, production task, heat loss of waste gas and heat dissipation. Finally, heating supply is calculated depending on calorific value of gas, air excess coefficient and so on, and then gas flow is calculated. Coking heating consumption is adjusted by real temperature of coke button [2]. Advantages of feedforward control influence the slow course of

temperature when regulation need not be considered. If monitoring point is less, instrument and measure error can be reduced. Shortcomings of feedforward control are to calculate accurate coking heating amount difficultly, and a lot of variables of calculating coking heating amount are difficult to be measured precisely and totally. The feed forward control system is illustrated in Figure 2.
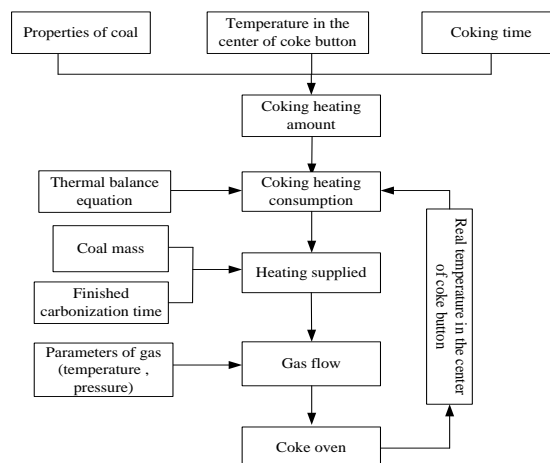

FIGURE 2 Feedback control system of coke oven

2.3 HYBRID CONTROL SYSTEM COMBINED FEEDBACK WITH FEEDFORWARD

When amount of coal charging, coal properties, production task of coke oven and operating time are regarded as input function, heating amount supplied is calculated by the heating supply model, and then settled heating amount supplied is feedback regulation according to deviation between real average temperature of the whole coke oven and settled average temperature of the whole coke oven [3]. The influence of unknown and difficult measure parameters produced is dispelled by feedback control. Influence caused by disturbance is compensated by feed-forward control at the same time in order to improve the time-delay of feedback control in the compound system. The compound control system is shown as Figure 3.
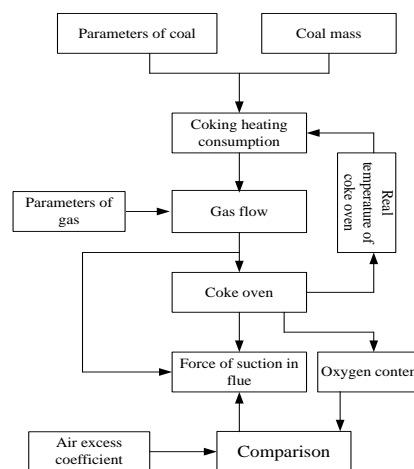

FIGURE 3 Control system of coke oven combined feedback with feedforward

Li Gongfa, Xiao Wentao, Liu Honghai, Jiang Guozhang, Liu Jia

3 Model of heating intelligent control system in coke oven

At present, the iron and steel enterprise usually utilizes "intermittent heating control" control method in the heating control system of coke oven, which can well optimize heating control of coke oven in a situation that the heating energy of coke oven is steady and rich [1-3]. But when pressure in main pipes of blast furnace fluctuates violently and heating coal gas flow is insufficient, the method can't instruct attendants how to operate heat control. The function of control system can only be analysed and judged artificially by the attendants, and during stopping heating time, blast furnace gas and coke oven gas are stopped using at the same time. Blast furnace gas is not fully utilized, either.

A new control principle [4, 5] combining the "intermittent heating control" with the heating gas flow adjustment is adopted in the control system. It analyses and processes data synthetically such as temperature, flow and calorific value of gas, pushing coke, charging coal, coal mass, water content and planned carbonization time, "stopping heating time" of PLC system and the heating blast furnace gas coke oven gas flow of DCS system are calculated and established through the model.

Therefore, heating of coke oven is even and stable and the whole heating level of coke oven is the intelligent control. Heating intelligent control of coke oven is realized. Its control system model is illustrated in Figure 4 [8, 9].
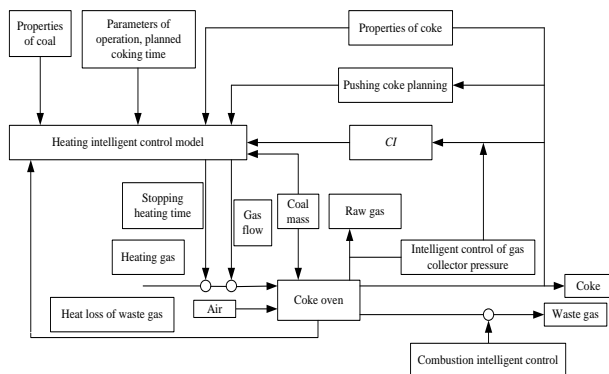
FIGURE 4 Intelligent control system model of coke oven

Carbonization index used in model of the control system is calculated according to the following equation [4]:

$$CI = \frac{t_{coking}}{t_{T\max}} , \qquad (1)$$

where $t_{T\max}$ is the time from coal charging in coke-chamber to the temperature of the waste gas passing the peak, t_{coking} is the time of finished carbonization of each coke-chamber. CI is a carbonization parameter to control coking production management, it fluctuates among $1.2 - 1.24$ suitably. Prediction model of t_{coking} is shown as follows:

$$t_{coking} = A \times t_{T\max} + C , \qquad (2)$$

where $t_{T\max}$ is the time from coal charging in coke-chamber to the temperature of the waste gas passing the peak; A and C are characteristic coefficient of coke oven.

A hybrid control system is proposed to control heating of coke oven, which combines feedback control, feed-forward control and fuzzy intelligent control. Real-time data of production in coke oven are gathered by the system such as pressure, flow, calorific value, temperature of coal gas, water content, and composition of heating gas and dynamic plans and so on. Settlement value of controlled parameters is calculated through the energy prediction model, namely the feed-forward. And then the value is transferred to the basic automated system to be regulated. According to real-time information such as waste gas temperature, coke button temperature, flue temperature and oxygen content offered by the basic automated system at the same time, the energy balance is feedback regulated constantly in the course of heating in accordance with the fuzzy intelligent control model (Figure 5). Then the settlement value is calculated again in order to be kept within the range of control, which satisfy not only necessary temperature required in coking, but also optimal heating control. Its control flow chart is shown in Figure 6.
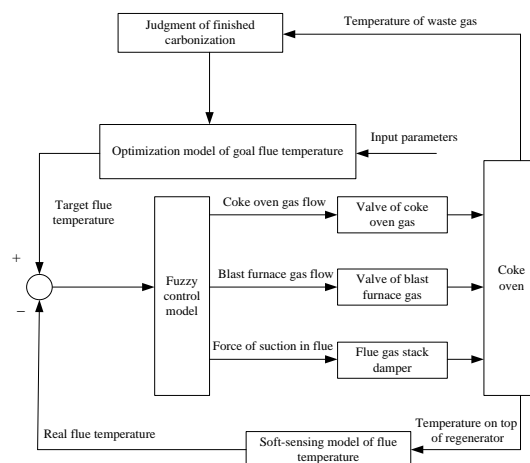

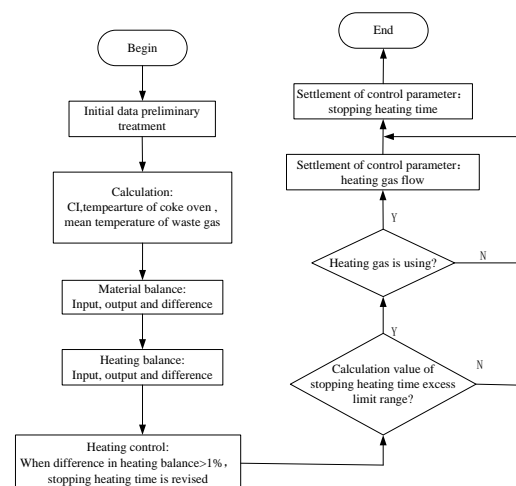FIGURE 5 Heating fuzzy intelligent control structure of coke oven

FIGURE 6 Flow chart of coke oven heating intelligent control model

In addition, gas flow and force of suction in flue (namely air-flue ratio) is intelligently controlled according to real-time processing production data in order to control total burning of heating gas as well as control waste gas content in the rational range. Meanwhile, intelligent control of gas collector pressure is achieved to control it in the suitable range in order to reduce the diffusing of coal gas.

4 Design of fuzzy control mode

4.1 ANALYSIS OF FUZZY CONTROL MODEL

Fuzzy control is the simulation behaviour of artificial intelligence, which utilizes the fuzzy theory to set up some rules of control and controls the production process to reach the satisfactory control result. The basic thinking of fuzzy control model in the system is according to deviation between present gas flow and feedforward gas flow, deviation between standard flue temperature and real flue temperature, and a temperature variation tendency of real flue temperature. Final gas flow is adjusted through certain fuzzy control algorithms [12-14].

ΔV – present gas flow – feed-forward gas flow,   (3)

ΔT – standard flue temperature–real flue temperature,   (4)

ΔC – flue temperature at present – flue temperature at last moment,   (5)

where ΔV, ΔT and ΔC all include three kinds of value, namely greater than zero, equal to zero, less than zero approximately.

The control algorithm of final gas flow adopts the fuzzy control method. ΔV, ΔT and ΔC are defined as the input of the fuzzy controller. The change amount of gas flow is got through fuzzy control algorithm.

4.2 DESIGN OF FUZZY CONTROLLER

Deviation between measured value and settled value will be divided into three grades [6, 7], namely negative, normal and positive respectively. Then every input parameter has three kinds of situation, and therefore these parameters make up as twenty seven kinds of situations. According to the actual conditions, each kind of situation adopts corresponding operation, and opening degree of the gas valve will increase or reduce. Twenty seven rules are shown as Table 1 where NG represents "negative", NM represents "normal", PS represents "positive", IC represents "increasing", DC represents "decreasing", KEEP represents "keeping"

TABLE 1 Control rule of fuzzy intelligent controller

Rule number	ΔV	ΔT	ΔC	ΔU	Rule number	ΔV	ΔT	ΔC	ΔU
1	NG	NG	NG	IC 10%	15	NM	NM	PS	DC 2%
2	NG	NG	NM	IC 10%	16	NM	PS	NG	KEEP
3	NG	NG	PS	IC 5%	17	NM	PS	NM	DC 2%
4	NG	NM	NG	IC 5%	18	NM	PS	PS	DC 5%
5	NG	NM	NM	IC 5%	19	PS	NG	NG	IC 5%
6	NG	NM	PS	KEEP	20	PS	NG	NM	IC 2%
7	NG	PS	NG	KEEP	21	PS	NG	PS	KEEP
8	NG	PS	NM	DC 2%	22	PS	NM	NG	KEEP
9	NG	PS	PS	DC 2%	23	PS	NM	NM	DC 2%
10	NM	NG	NG	IC 5%	24	PS	NM	PS	DC 2%
11	NM	NG	NM	IC 2%	25	PS	PS	NG	DC 2%
12	NM	NG	PS	KEEP	26	PS	PS	NM	DC 5%
13	NM	NM	NG	IC 2%	27	PS	PS	PS	DC 10%
14	NM	NM	NM	KEEP					

Various kinds of influence and variation tendency of gas flow are summed up into twenty seven rules synthetically in the model according to artificial operation experience. The fuzzy control model of gas flow has been set up [10, 11], and thus heating intelligent control of coke oven is realized. Simulation program of fuzzy control is shown as Figure 7.
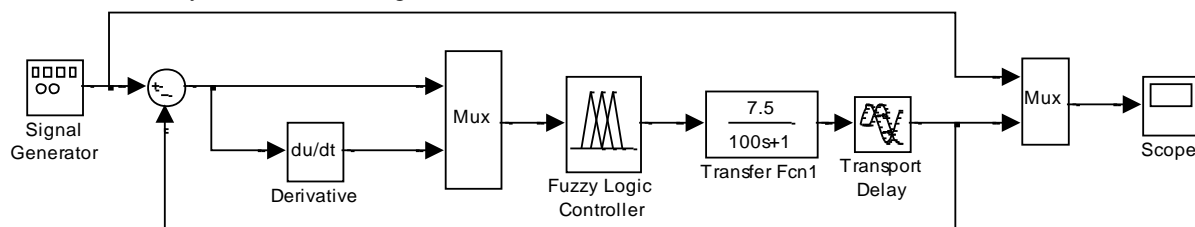

FIGURE 7 Simulink simulation program of fuzzy control

When operation condition of production changes, the change range of flue temperature will often exceed (–6, +6), if the simple control method is still adopted, because of the great inertia of coke oven, big exceeding adjusting amount and too long adjustment time are caused. Aiming at above-mentioned situations, a prediction part in the controlling course has been increased (Figure 8). Furthermore the deviation of temperature is judged firstly,

when the range of deviation does not exceed (−6, +6), fuzzy control is adopted. If it exceeds above-mentioned ranges, Bang-Bang intelligent control is used. So coal gas flow and stopping heating time are adjusted, control precision and fast responses of controlled target are guaranteed [15-17].
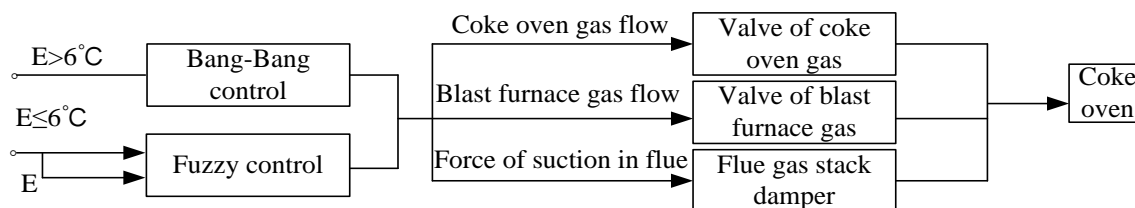

FIGURE 8 Intelligent control

5 Application

This system has already been succeeded in applying intelligent control system of coke oven heating in some iron and steel company. Settled temperatures of machine and coke sides are 1260°C and 1310°C respectively in coke oven production, temperature control of machine side and coke side are shown in Figures 9 and 10, respectively. The statistics show that temperature errors of machine side and coke side are up to 89.5% and 86.7% among ±5°C, meet the real industrial production demand.
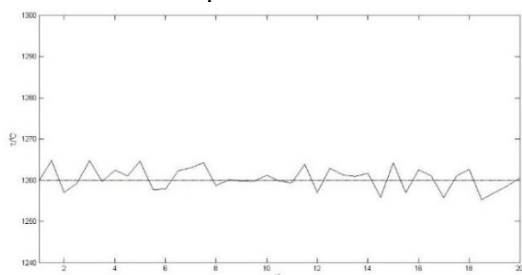
FIGURE 9 Temperature control of machine side
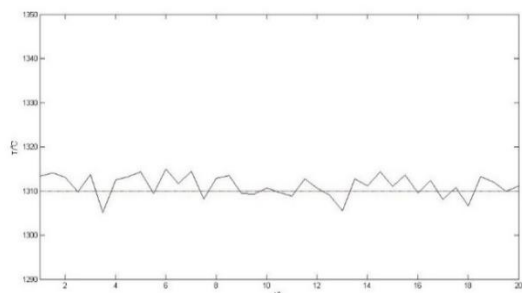
FIGURE 10 Temperature control of coke side

6 Conclusions

According to characteristics of coke oven and actual craft states in coking plants, a fuzzy intelligent hybrid control method is proposed, which combines feedback control, feed-forward control and fuzzy intelligent control. Feed-forward and feedback are utilized to dispel influence of examined disturbance and no examined disturbance respectively. Advantages of the method consider not only change of flue temperature at present, but also demand of heating supplied in the future. Finally fuzzy intelligent control module has been designed to solve temperature fluctuation of coke oven effectively, which reaches the better control result. The system is of great practical value. The system is put into operation in the company for more than three years, the system runs normally, the result is good. It is easy to operate and master for operators, lightens labour intensity, and has improved operation condition. The system not only can regulate the heating supplied amount of coke oven in time, but also have stronger anti-interference ability. The system has reduced fluctuation of whole temperature of coke oven to a great extent, and heating supplied amount needed is also reduced relatively, so consumption of gas has been reduced. At the same time stable coefficient is improved too, therefore, the life of coke oven is lengthened and coke quality is improved. It has reached the anticipated result and very great practical value.

Acknowledgments

References

[1] Swanijung J, Palmu P 1996 Development of Coke-oven Battery Process Management System at Rautaruukki Steelworks *Iron and Steel Engineer* **73**(1) 46-9
[2] Pan C 2007 Temperature control system for thermal field of single crystal growth furnace based on DMC algorithm *Chinese Journal of Mechanical & Electrical Engineering* **24**(8) 53-6 *(in Chinese)*
[3] Shi J, Wu H, Jin X 2011 Automatic control system in regenerative aluminum melting furnace *Chinese Journal of Mechanical & Electrical Engineering* **28**(11) 1342-1344 1390 *(in Chinese)*
[4] Jiang G, Kong J, Li G 2006 Intelligent Control System for Coke Oven Heating *Chinese Journal of Iron and Steel* **41**(11) 73-6 *(in Chinese)*
[5] Li G, Kong J, Jiang G 2008 Research and Application on Compound Intelligent Control System for Coke Oven Heating *Chinese Journal of Iron and Steel* **43**(8) 89-92 *(in Chinese)*
[6] Kim E 2011 A new approach to numerical stability analysis of fuzzy control systems *IEEE Transactions on Systems, Man, and Cybernetics Part C: Applications and Reviews* **31**(6) 107-13
[7] Yan X, Zhang X 2006 Experimental system of temperature controlling based on the algorithm of fuzzy controlling *Chinese Mechanical & Electrical Engineering* **23**(10) 18-20 *(in Chinese)*
[8] Li G, Xiao W, Jiang G 2013 Soft-Sensing Model of Coke Oven Flue Temperature *Sensors & Transducers* **161**(12) 265-70
[9] Wu X, Zhao J, Yang X 2011 Application of intelligent PID algorithm in finance temperature control system *Chinese Journal of Mechanical & Electrical Engineering* **28**(8) 948-950,959 *(in Chinese)*

[10] Wu M, Lei Q, Cao W 2008 Flue temperature fuzzy control for coke oven heating process based on multi-operative modes analysis *Chinese Journal of Centre South University (Science and Technology)* **39**(1) 155-60 *(in Chinese)*

[11] Chen Y, Wu H 2007 Research On Temperature Predictive Model of Cokeoven Flow Based on ANFIS *Computer Measurement & Control* **15**(4) 462-5 *(in Chinese)*

[12] Jang J S R 1993 ANFIS: Adaptive network-based Fuzzy Inference System *IEEE Transactions on Systems Man and Cybernetics* **23**(3) 665-85

[13] Wang X, Shao H 1997 Nonlinear system modelling using the radial basis function neural networks *Control Theory and Applications* **14**(1) 59-66

[14] Peng F 2009 Temperature fuzzy control of thermostat incubator *Chinese Mechanical & Electrical Engineering* **26**(4) 84-6 *(in Chinese)*

[15] Yang H, Yang X, Wang W 2007 Self-learning intersection single control system based on RBF neural network *Chinese Journal of Mechanical & Electrical Engineering* **24**(12) 17-9 *(in Chinese)*

[16] Chaichenko V T, Lavrov K G 2001 Pressure stabilization of direct coke oven gas in the gas collectors of coke batteries *Coke & Cherillstry* **64**(7) 72-6

[17] Yang C, Wu M, Shen D 2001 Hybrid intelligent control of gas collectors of coke ovens *Control Engineering Practice* **9**(7) 725-33

Authors

Gongfa Li, born in October, 1979, Hubei, China

Current position, grades: associate professor at the College of Machinery and Automation, Wuhan University of Science and Technology.
University studies: PhD degree in mechanical design and theory from Wuhan University of Science and Technology in China.
Scientific interests: modelling and optimal control of complex industrial process.
Publications: 100.

Wentao Xiao, born in 1989, Tianmen, China

Current position, grades: M.S. degree student in mechanical design and theory at Wuhan University of Science and Technology.
University studies: B.S. degree in mechanical engineering and automation at City College of Wuhan University of Science and Technology, Wuhan, China, in 2013.
Scientific interests: mechanical CAD/CAE, signal analysis and processing.

Honghai Liu, born in 1973, China

Current position, grades: professor in Intelligent Systems, head of Intelligent Systems and Biomedical Robotics, University of Portsmouth.
University studies: PhD in Intelligent Robotics in 2003 at Kings College, University of London, UK.
Scientific interests: approximate computation, pattern recognition, multi-sensor based information fusion and analytics, human machine systems.
Publications: 300.
Experience: Research at King's College London, University of Aberdeen.

Guozhang Jiang, born in December, 1965, Tianmen, China

Current position, grades: professor of Industrial Engineering, assistant dean at the college of machinery and automation, Wuhan University of Science and Technology.
University studies: PhD degree in mechanical design and theory at Wuhan University of Science and Technology, China, in 2007.
Scientific interests: computer aided engineering, mechanical CAD/CAE and industrial engineering and management system.
Publications: 120.

Jia Liu, born in 1990, Hubei, China

Current position, grades: M.S. degree student in mechanical design and theory at Wuhan University of Science and Technology.
University studies: B.S. degree in mechanical engineering and automation at Wuchang institute of Technology, Wuhan, China, in 2012.
Scientific interests: mechanical CAD/CAE, signal analysis and processing.

Load balancing over redundant wireless sensor networks based on diffluent

Xikui Gao, Yan Bai, Yun Ju*

School of Control and Computer Engineering, North China Electric Power University, 102206, China

Abstract

As the traffic blocking probability of traditional hard load balancing algorithms is generally high over redundant wireless sensor networks, This paper proposes a load balancing algorithm based on dividing the packet flow (LBD) over redundant wireless sensor networks based on the idea of soft load balancing. In the scheme, through numerical analysis, obtain the optimal flow-dividing ratio to determine the volume of traffic delivered to each network, which maintains network load balance. From the theoretical and simulative perspectives, the paper analyses the performance parameters, and the analytical results show that the performance of the scheme is better than other schemes. Simulation results show that the proposed method outperforms traditional hard load balancing techniques in terms of traffic blocking and packet loss probabilities.

Keywords: redundant wireless sensor networks, load balancing, packet flow diversion, access selection, handover

1 Introduction

The future of communication network is a redundant structure system, various wireless access network access to public core network based on all IP through a variety of different access technology [1]. Redundant wireless sensor network resource management technology, especially the load balance technology, is one of the key technologies to realize integration in the redundant wireless network, and is a research hotspot in recent years. Load balancing is the important method to implement effective resource sharing in the redundant wireless network, which can improve the utilization of wireless resources in redundant wireless networks, expand the system capacity, to provide users with better service.

References [2, 3] consider system load affected by the users, the condition of channel, the QoS (quality of service) requirements of business, and other factors, and put forward when the system load is greater than the prescribed threshold method, the high load transfer part of the user in the system to the low load system.

In reference [4] calculate the need of resources consumed by users access each network, the design target of access is. After the user access network, the ratio of consumption of resources and network resources available to is the minimum; reference [5] used multi-objective decision-making method for network selection in the redundant wireless network. Discuss the load balance and other parameters of compromise; in a redundant and distributed grid environment reference [6] designed a kind of immune clone task scheduling algorithm, to achieve the balanced allocation of resource and efficient scheduling tasks. The above references about method of load balancing are all hard load balancing - The user's business at the same time can only access a network, not well QoS guarantee business, business blocking ratio is higher [7]. References [7, 8] propose the concept of soft load balance. The packet of down link can be divided into sub stream, each child flows into the different wireless access networks, this method can more fully use of resources in the redundant wireless network. But references [7] only studied the best diversion ratio in specific network topology circumstances, not universal; and reference [8] only studied the diversion ratio under optimal wireless channel environment, the results deviation with the actual situation.

Aiming at the existing problem of the load balance method, This article propose the load balancing algorithm under the whole cover network environment based on packet flow diversion (load Balancing algorithm –based on dividing packet flow, LBD). Algorithm considered the effect of multipath and path loss, user service in redundant wireless network is obtained by numerical analysis of the optimal split ratio. When users need to access network, or switch, system downstream business will be divided into the subflows according to the optimal split ratio and the subflow will access or switch to the corresponding network, to improve the network capacity, improve the system performance.

2 The system model

2.1 SOFT LOAD BALANCE NETWORK STRUCTURE

In this paper, on the basis of reference [9], improved a layered half centralized network structure as shown in Figure 1. Several location adjacent area mapping into a basic grid, information server-IS, resources allocation-RA, and resources statistics-RS are collectively referred to as resource management unit-RMU, it is responsible for the management of grid resources. RS is set to access node, in the node for statistics and calculation of the resource information; RA is used to collect the RS information, and according to the situation of the node load and basic grid resources for load balancing. IS as RA supervisor server, basic grid structure under the normal state only responsible for edges of resource allocation and the identification of the storage nodes, the information such as position, load condition, but when the management of RA malfunction can't work, IS can quickly take over and perform the function of RA.
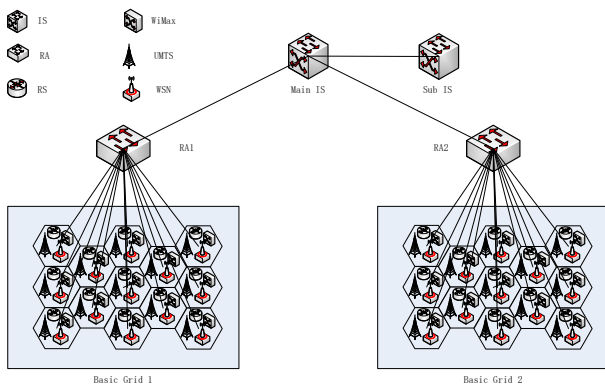
FIGURE 1 Hierarchical semi-centralized architecture in a two-dimensional resource unit can get the biggest transfer rate.

In order to support the soft load balancing. The public link layer are introduced to processing access control signals in RA [7]. In each of the access network, below the public link layer is the media access control layer and physical layer, the MAC layer is located in the RS. Each MAC protocols report it channel state to the public link layer, Public link layer determine the size of the diversion ratio according to the obtained channel condition and system load conditions. The public link layer data received from the MAC layer, according to the data of the serial number on the reorder, and then to transfer the data through the network layer to the transport layer.

2.2 TWO-DIMENSIONAL RESOURCE UNIT

System model have M kinds of overlapping coverage of access system, different access system adopts OFDM (orthogonal frequency division multiplexing) technology. For system m ($m \in 1, 2, \ldots, M$), all subcarrier in sequential mode can be divided into a number of sub-channels. Each channel contains F_m carrier. The frame length T_{fm} is

divided into isometric time slot, a time slot contains S_m OFDM symbols. Such a time slot on the time domain and frequency domain on a child channel will constitute a two-dimensional resource unit, as the basic unit of the resource allocation in OFDM system [4].

In a two-dimensional resource unit can get the biggest transfer rate:

$$b_{i,m} = \frac{F_m S_m C_{i,m}}{T_{f,m}}, \tag{1}$$

where after the user i access system m. Based on adaptive coded modulation mechanism to determine the information bits $C_{i,m}$ carried by each modulation symbols [10].

2.3 CHANNEL MODEL

Considering the influence of multipath and path loss, channel model of this paper by using the Rayleigh fading model, the user receives the base station signal power [4]:

$$P_R = \alpha^2 A d^{-\beta} P_T, \tag{2}$$

where P_T for the access point transmission power. In this paper, redundant wireless sensor networks in each system, assuming that access point transmission power is constant, and evenly distributed in the entire history of the available frequency band; α is the mean of 1 and obey exponential distribution of random variables, used to characterize the fast fading. $A d^{-\beta}$ is used to represent the path loss, constant A decided by the antenna height and carrier frequency, β is attenuation index, d is the distance between the user and the access point.

Assume that the user i in the system m within the coverage of the j access point, the dry ratio $\gamma_{i,m,j}$ at the receiving end can be expressed as:

$$\gamma_{i,m,j} = \frac{\alpha^2 A d_{i,m,j}^{-\beta} P_{T,m}}{\sum_{k=1}^{N_m} \alpha^2 A d_{i,m,j}^{-\beta} P_{T,m} + \omega_{i,m}}, \tag{3}$$

where, $\omega_{i,m}$ is the noise power at the receiving end of user i in the system m, $d_{i,m,j}$ is the distance between user i and the j access point of system m, N_m is the number of other access points which are the disturbance of access point j of system m, $P_{T,m}$ is the transmission power of access points in the system m.

In the AWGN channel with M - QAM modulation the biggest transport bit error rate from the access point to the user i in the system m [4] is:

$$P_{BER,i,m} = \frac{1}{5} e^{-\frac{1.5\gamma_{i,m,j}}{2^{C_{i,m-1}}}}. \tag{4}$$

When the packet length is K bits, in order to obtain the relationship between the packet loss rate and bit error rate, considering the limiting cases. Assume that after through other error correction mechanism, any one bit error means

that the entire packet transmission error, the packet loss rate of the user i in the system m is [11]:

$$P_{PL,i,m} = 1 - \left(1 - P_{BER,i,m}\right)^{K} . \tag{5}$$

Considering automatic retransmission mechanism of MAC layer, an average transfer numbers of receiving a right packet group took [4]:

$$K_{i,m} = \frac{1}{1 - P_{PL,i,m}} . \tag{6}$$

3 The load balancing algorithm based on packet flow diversion

When the user have some group business need network access, or in the process of service because of in the system or between the system switch due to mobile, factors such as the decrease of the quality of the business requirements, System for its first calculated optimal split ratio, then according to the optimal split ratio to access or switch to the appropriate network, the algorithm overall process is shown in Figure 2, algorithm steps as shown in section 2.3.

3.1 TWO-DIMENSIONAL RESOURCE UNIT ALLOCATION MODEL

Assume the rate demand of user i is R_i, user i business in system m transmission rate is $R_{i,m}$, the diversion ratio of users i in the system m is $\omega_{i,m}$, the Equations (7) and (8) are as follows:

$$R_{i,m} = R_i \omega_{i,m} , \tag{7}$$

$$\sum_m \omega_{i,m} = 1 , \tag{8}$$

After user i access system m, each frame required two-dimensional resources unit for an average is:

$$n_{i,m} = \left\lceil k_{i,m} \frac{R_{i,m}}{b_{i,m}} \right\rceil , \tag{9}$$

where, $\lceil x \rceil$ is the minimum integer greater than or equal to x.
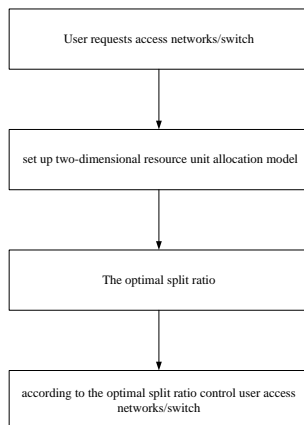
FIGURE 2 Flow chart of LBD

Assume that the user i access system m before the j access point, the access point is covered inside the village which have accessed $I_{m,j,I}$ users, the load in the location can be expressed as:

$$L_{m,j,i} = \sum_{k-1}^{I_{m,j,i}} n_{k,m} . \tag{10}$$

In order to satisfy the QoS requirements of group business, this paper adopts the following strategies:

1) The average service rate of the non real-time business is greater than the minimum service rate, on the basis of using the best scheduling mechanism. So in access to choose, Equation (9) can be said it took an average of network resources.

2) Most of the grouped data real-time business to send in a maximum time delay constraint. This paper adopts the effective bandwidth theory to ensure real-time QoS requirements of the business, If the bandwidth of the network to provide real-time business is greater than the effective bandwidth, its QoS guaranteed [12]. Assume that grouping package (on) the activation state only at fixed intervals, and in the inactive state (off) does not produce. Suppose system m on his stay on and off states respectively obey negative exponential distribution of parameters for the λ_m and μ_m, the probability of packet delay is greater than the given threshold D to ε, the user i equivalent bandwidth of real-time business can be expressed as [13]:

$$W_{i,m}^{'} = \frac{R_{i,m}(\mu_m D - \ln \varepsilon)}{(\lambda_m + \mu_m)D - \ln \varepsilon} . \tag{11}$$

If the network to provide real-time business available data rate is greater than the type of effective bandwidth value in Equation (11), the real-time business QoS can be guaranteed. By Equation (9), real-time business users i access system m 2d resources required for each frame after the average cell number is:

$$n_{i,m}^{'} = \left\lceil k_{i,m} \frac{W_{i,m}^{'}}{b_{i,m}} \right\rceil . \tag{12}$$

3.2 THE BEST DIVERSION RATIO METHOD

This paper takes the real-time business as an example the optimal distribution ratio, real-time business best solution of the diversion ratio is similar.

Set up an access point in the system m available unit for a total of 2d resources $L_{T,m}$, define the user i to the influence degree of the network load after access networks (hereinafter referred to as the load factor) is:

$$\eta_i = \sum_{m=1}^{M} \frac{n_{i,m}}{L_{T,m}} . \tag{13}$$

In order to get the best load balancing effect, user i in the system m the best split ratio $\omega_{i,m}^*$ is obtained by Equation (14):

$$\left[\omega_{i,1}^*,..,\omega_{i,M}^* \right] = \arg\min \eta_i = \arg\min \sum_{m=1}^{M} \frac{n_{i,m}}{L_{T,m}} . \qquad (14)$$

Load factor expressions η_i is received using Equations (4), (5), (7), (9) and (13):

$$\eta_i \sum_{m=1}^{M} \frac{\left| B_{i,m} \omega_{i,m} \right|}{L_{T,m}} , \qquad (15)$$

where:

$$B_{i,m} = \frac{R}{\left(1 - P_{PL,i,m}\right) b_{i,m}} . \qquad (16)$$

When the user i access networks m, R_i, $P_{PL,i,m}$ and $b_{i,m}$ are constant values, so $B_{i,m}$ for the fixed value.

Best diversion ratio of the calculation steps are as follows:

Step 1: when $l = \arg\min_{m \in \{1,...,m\}} \frac{B_{i,m}}{L_{T,m}}$, j access points of system l have enough resources to allow users i all access, when $L_{T,l} - L_{l,j,i} \geq B_{i,l}$, get the objective function for solutions of Equation (14):

$$\omega_{i,m}^* = \begin{cases} 1, & l = \arg\min_{m \in \{1,...,m\}} \frac{B_{i,m}}{L_{T,m}} \\ 0, & other \end{cases} \qquad (17)$$

Step 2 when the j access point of system l does not have enough resources in to allow user i all access, when $L_{T,l} - L_{l,j,i} < B_{i,l}$, make $n_{i,l} = L_{T,l} - L_{l,j,i}$, you can get:

$$\begin{aligned} \omega_{i,m}^* &= \frac{L_{T,l} - L_{l,j,i}}{B_{i,l}} \\ l &= \arg\min_{m \in \{1,...,m\}} \frac{B_{i,m}}{L_{T,m}} \end{aligned} , \qquad (18)$$

make $l' = \arg\min_{m \neq l, m \in \{1,...,m\}} \frac{B_{i,m}}{L_{T,m}}$, if the first j access points of the system l have enough resources to the user i full access, when $L_{T,l'} - L_{l',j,i} \geq \left(1 - \omega_{i,l}^*\right) B_{i,l'}$, then:

$$\omega_{i,m}^* = \begin{cases} \frac{L_{T,l} - L_{l,j,i}}{B_{i,l}}, & l = \arg\min_{m \in \{1,...,m\}} \frac{B_{i,m}}{L_{T,m}} \\ 1 - \omega_{i,l}^*, & l' = \arg\min_{m \neq l, m \in \{1,...,m\}} \frac{B_{i,m}}{L_{T,m}} \\ 0, & other \end{cases} \qquad (19)$$

Step 3 when the sum of all the available resources in system l and l' cannot meet the demand of user business resources, $L_{T,l} - L_{l,j,i} < B_{i,l}$ and $L_{T,l'} - L_{l',j,i} < \left(1 - \omega_{i,l}^*\right) B_{i,l'}$, similar to the step 2 calculation values other $\omega_{i,m}^*$ than $m = 1$.

Step 4: if the sum of all system resources available cannot meet the demand of users i rate R_i, declined to the user's access to the switch or request.

3.3 LOAD BALANCING STEPS

LBD algorithm steps are as follows:

Step 1: when a user i need access to the network or switch, RA or IS (when user i on the verge of basic grid) scanning for the available access points list of users i.

Step 2: RS according to Equation (3) to measure Signal to Interference Ratio $\gamma_{i,m,j}$ from user i to each available access points. According to the Equations (4) and (5) channel bit error rate and packet loss rate are calculated respectively. Combined with the corresponding link adaptive mechanism with Equation (1) to calculate the corresponding each 2d resource unit can obtain the maximum transmission rate of $b_{i,m}$. To transfer the result of the measurement and calculation for RA.

Step 3: RA through using the method of section 2.2 to calculating optimum diversion ratio for the user i $\omega_{i,m}^*$.

Step 4: RA according to the optimal split ratio $\omega_{i,m}^*$, to divide users i business into sub stream, and flow or switch to the corresponding access to the network.

4 Algorithm performance analysis

4.1 THE SIMULATION ANALYSIS

Proposed by reference [4] MLB (maximum load balancing algorithm is hard load balancing method of superior performance of an algorithm, in order to verify the LBD algorithm performance, in this paper, the LBD algorithm and the MLB business traffic congestion, packet loss rate of the algorithm and the network load balancing degree has carried on the simulation analysis. The simulation scenario is shown in Figure 3 [4].
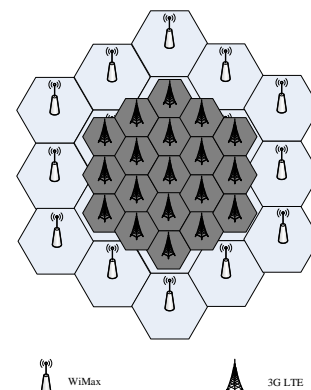

FIGURE 3 Simulation scenario

LTE system area (1.5 km) radius and WiMax system area (radius 3 km) overlapping coverage, two systems centre village access point location collocated. At the beginning of the simulation, the user random distribution in the region of the repeat coverage, random direction, moved to the village edge, handover will happen. When user move at speed 120 km/h, user's location and movement direction is updated once every 100 ms, one direction change ±45° at 0.2 probability at a time. Each user's business source are independent of each other, and with the same probability to generate real-time and non real-time business. Simulation parameter Settings as shown in Table 1.

TABLE 1 Simulation parameters settings

name	UMTS	WIMAX
Carrier frequency/bandwidth	2.0 GHz/10MHz	3.5GHz/10MHz
$T_{f,m}$	10 ms	10 ms
OFDM symbol cycle	0.2 ms	0.2 ms
NFFT	512	1024
$P_{T,m}$	43dBm	46dBm
$W_{i,m}$	-170 dBm/Hz	-174 dBm/Hz
Two dimensional resource unit	24 consecutive subcarrier and 10 symbols of cycle	24 consecutive subcarrier and 10 symbols of cycle
K	512 bit	512 bit
Real-time business rate	Uniform distribution in [50 kb/s, 200 kb/s]	Uniform distribution in [50 kb/s, 200 kb/s]
Non real-time business rate	Uniform distribution in [40 kb/s, 140 kb/s]	Uniform distribution in [40 kb/s, 140 kb/s]
λ_m/μ_m	2.84 s/1.53 s	2.84 s/1.53 s
Business generate time interval	Obey the average poisson distribution for 1 s	Obey the average poisson distribution for 1 s
Business last time interval	Obey the average poisson distribution for 2 s	Obey the average poisson distribution for 2 s
name	UMTS	WIMAX
Carrier frequency/bandwidth	2.0 GHz/10MHz	3.5GHz/10MHz

The statistical simulation results, statistics only two network repeat coverage areas darker areas in (Figure 3). According to the parameters of Table 1, Through the Matlab simulation results as shown in Figure 4 ~ 7, users in the figure is a LTE area within the scope of active users.

Figure 4 for the number of IP when using the algorithm of LBD shunt and the ratio of total access and switching times. When users less than 30, have enough resources to accept user network, so packet flow without shunt. With the increase of number of users, a single network resources began to can't satisfy user needs of the business, user packet of the business flow is divided into two sub flow access two network at the same time.

Figure 5 for the network load is full cause user blocking probability. Due to network users, less than 30 have enough resources to acceptance of users, users under two kinds of algorithm the block probability is 0. With the increase of users, the user's block probability, LBD algorithm can be very high in both the network load will users business packet flow is divided into two sub flow

access two network at the same time, so its block probability MLB was always lower than the algorithm, and the more users, the more obvious advantages in block probability LBD algorithm.

Figure 6 for the user to access the network or when switching the average packet loss rate. It can be seen that with the increase of users, the LBD algorithm proposed in this paper, the packet loss rate than MLB always low packet loss rate of the algorithm.

Figure 7 for two networks under two kinds of algorithm compares the average normalized load, two algorithms is visible in the aspect of load balance ability, has been close to two network load conditions.
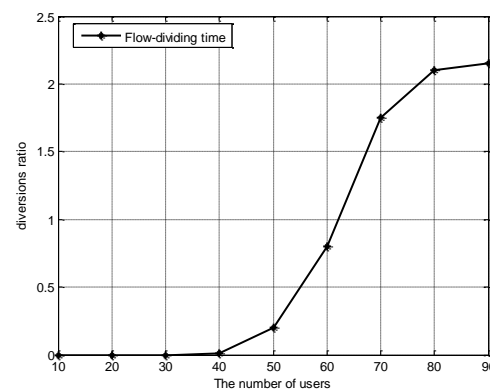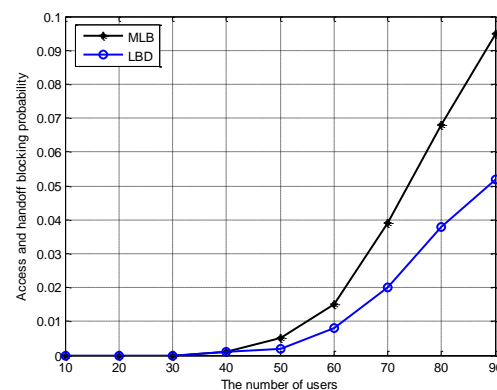

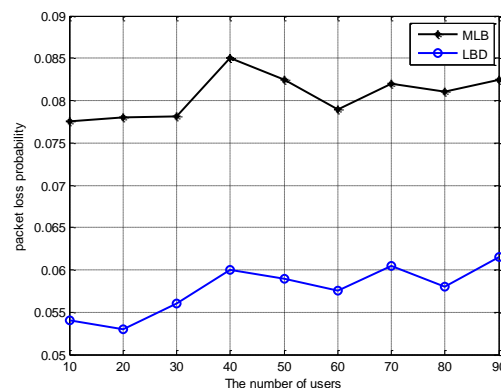FIGURE 4 Flow-dividing time.

FIGURE 5 Blocking probability.

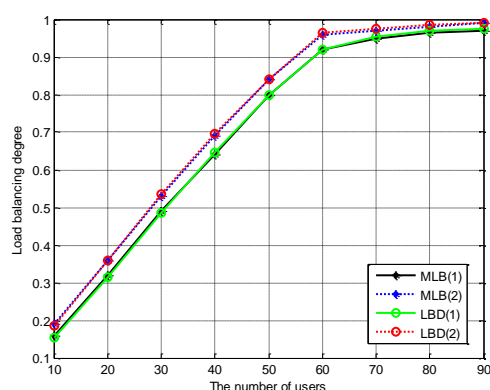FIGURE 6 Packet loss probability.

FIGURE 7 Load balancing degree.

4.2 ALGORITHM APPLICABILITY

Reference [7] study the network structure of two kinds of system coverage crisscross overlap, neighbourhood radius of two systems, the same access point located at the centre of the community, and a system of community centre is located in another system on the edge of the neighbourhood, see Figure 3 in reference [7]; In this paper, we study the network structure as the centre of community access point phase coincidence of the two systems, and neighbourhood radius of two systems can be different, this structure is closer to the actual network. Reference [7] only analyses the two systems centre village access point best

in the middle of the diversion ratio; Algorithm in this paper to consider all the best shunting overlapping coverage ratio, scope of application is more widely. Reference [8] only studied the ideal of the wireless channel environment load balance; In this paper, considering the influence of multipath and path loss. Therefore, compared with the existing soft load balancing algorithm, this algorithm is more universal.

5 Conclusions
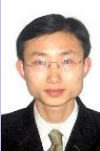
In this paper, based on packet flow diversion of heterogeneous wireless network load balancing algorithm. Algorithm considering the whole network environment, the real-time statistics network status indicators, calculated under different network load balancing target users access to the network or the best split ratio when switching. Choose according to optimal split ratio for access or switch, can achieve less resource consumption, improve the system capacity. The simulation results show that LBD algorithm can effectively balance the network load, and the group business average blocking probability and packet loss spontaneously in a relatively traditional MLB algorithm has obvious improvement.

Acknowledgments

References

[1] Javaid N, Qureshi T N, Khan A H, Iqbal A, Akhtar E, Ishfaq M 2013 EDDEEC: Enhanced Developed Distributed Energy-Efficient Clustering for Heterogeneous Wireless Sensor Networks *Procedia Computer Science* **19** 914-9

[2] Hefeida M S, Canli T, Khokhar A 2013 Cl-mac: A cross-layer mac protocol for heterogeneous wireless sensor networks *Ad Hoc Networks* **11**(1) 213-25

[3] Szczodrak M, Gnawali O, Carloni L P 2013 Dynamic reconfiguration of wireless sensor networks to support heterogeneous applications *IEEE International Conference on Distributed Computing in Sensor Systems (DCOSS) 2013* 52-61

[4] Al-Hamadi H, Chen R 2013 Redundancy management of multipath routing for intrusion tolerance in heterogeneous wireless sensor networks *IEEE Transactions on Network and Service Management* **10**(2) 189-203

[5] Alonso R S, Tapia D I, Bajo J, García O, de Paz J F, Corchado J M 2013 Implementing a hardware-embedded reactive agents platform based on a service-oriented architecture over heterogeneous wireless sensor networks *Ad Hoc Networks* **11**(1) 151-66

[6] Sharma K K 2013 Traffic contention assessment and control for improvement of QoS in heterogeneous wireless sensor networks *In fulfillment of the requirements for the degree of doctor of philosophy in electronics & communication engineering summary of report*

[7] Wang X, Han S, Wu Y, Wang X 2013 Coverage and energy consumption control in mobile heterogeneous wireless sensor networks. IEEE Transactions on Automatic Control **58**(4) 975-988

[8] Parvatkar S U, Gore D V 2014 An Ant Colony Optimization Algorithm for maximizing the lifetime of Heterogeneous Wireless Sensor Network

[9] Kim D Y, Cha S H, Cho K H 2013 Ontology-based methodology for managing heterogeneous wireless sensor networks *International Journal of Distributed Sensor Networks* **2013** 610684

[10] Maia G, Guidoni D L, Viana A C, Squino A L L, Mini R A F, Loureiro A A F, 2013 A distributed data storage protocol for heterogeneous wireless sensor networks with mobile sinks *Ad Hoc Networks* **11**(5) 1588-1602

[11] Kapoor N K, Nandy B, Majumdar S 2013 Dynamic allocation of sensor nodes *in wireless sensor networks hosting multiple applications IEEE International Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS), 2013* 51-7

[12] Senturk I F, Akkaya K, Yilmaz S 2014 Relay placement for restoring connectivity in partitioned wireless sensor networks under limited information *Ad Hoc Networks* **13** 487-503

[13] Younis M, Senturk I F, Akkaya K, Lee S, Senel F 2014 Topology management techniques for tolerating node failures in wireless sensor networks: A survey *Computer Networks* **58** 254-83

[14] Tapia D I, Alonso R S, García Ó, de la Prieta F, Pérez-Lancho B 2013 Cloud-IO: Cloud computing platform for the fast deployment of services over wireless sensor networks *The 7th International Conference on Knowledge Management in Organizations: Service and Cloud Computing* Springer Berlin Heidelberg **172** 493-504

[15] Chen D, Liu Z, Wang L, Dou M, Chen J, Li H 2013 Natural disaster monitoring with wireless sensor networks: a case study of data-intensive applications upon low-cost scalable systems *Mobile Networks and Applications* **18**(5) 651-63

Authors

Xikui Gao, born in October, 1959, China

Current position, grades: senior engineer, professor.
University studies: School of Control and Computer Engineering, North China Electric Power University.
Scientific interests: control theory and control engineering.
Experience: President of harmonious technology co., LTD., Jiangsu, China.

Yan Bai, born in March, 1954, Shenyang, Liaoning province, China

Current position, grades: Doctoral degree, professor at the School of Control and Computer Engineering, North China Electric Power University.
University studies: northeastern university, PhD.
Scientific interests: power plant production process automation, design and teaching.
Publications: more than 50.

Yun Ju, born in August, 1982, Lichuan, Jiangxi province, China

Current position, grades: doctoral degree, lecturer at the School of Control and Computer Engineering, North China Electric Power University.
University studies: PhD at the School of Control and Computer Engineering, North China Electric Power University.
Scientific interest: control theory and control engineering.
Publications: more than 7.

NATURE PHENOMENA AND INNOVATIVE ENGINEERING

A new optimization algorithm for multi-dimensional cloud data centre resources scheduling based on PSO

Dong-Sheng Xu[1]*, Feng Zhang[2]

[1]*Software Institute, Xi'an University, Xi'an 710065, China*

[2]*School of Information Engineering, Yulin University, Yulin 719000,China*

Abstract

In order to solve the problem of multidimensional cloud resources low utilization ratio and the high energy consumption of cloud communication between tasks, proposed a new cloud data centre resource scheduling algorithm, which combined resource fusion principle, particle swarm optimization algorithms and taboo search algorithm and it's with a low-power scheduling computing, storage and bandwidth resources integration scheduling method. Simulation results show that the algorithm has the advantage of cloud resources using stable, high-dimensional cloud resource utilization and low-power cloud data centres.

Keywords: particle swarm optimization, cloud computing, cloud data centre, data scheduling

1 Introduction

With the large-scale development of distributed technology, cloud computing has become a key research area. Cloud computing is a kind of service providers and service users protocol based on the distributed system [1], Cloud users can access the cloud services according to the requirement, there has five main features, that is: high scalability, flexibility, on-demand services and billing, large scale and virtualization. Cloud data centres [2, 3] is composed of hardware and software in cloud computing environment pine that resource sharing framework. The user can according to the needs of the dynamic use of these hardware and software resources, and according to the service quantity to pay service fees. It has five characteristics, that is: resource pooling, intelligent and efficient service and transparent, on-demand service and billed by consumption.

Cloud computing [4] according to the level of abstraction provided resources can be divided into three layers: the cloud data centre layer, application layer and service layer software function operation. Cloud data centre layer through the virtual technology and other infrastructure hardware abstraction and physical resources available to the massive cloud users, cloud data centre layer allows the user to dynamically apply or release the node and on-demand access to resources, according to usage billing [4, 5]. The application ability running layer to the underlying physical resources had higher level of abstraction; it provides a platform to run the cloud user applications, the application ability run layer to provide a platform for software application development as the centre, deployment, operation-related platform resources. Software service layer functions are encapsulated into

services, and through the network to the cloud user. The hierarchy of the cloud data centre layer, application ability running layer and software service layer is as shown in Figure 1.
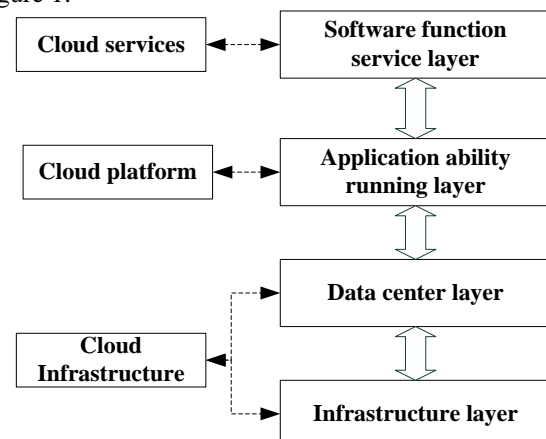

FIGURE 1 The hierarchy of the cloud data center layer

In this paper, cloud data center layer are studied. Because of the high energy consumption of multiple types of cloud communication resources existing cloud data center, and ignore the huge energy, multi resource consumption did not consider the multi-dimensional cloud resource scheduling, energy cloud did not fully consider the communication between tasks, So the presence of an existing cloud data center cloud resource providers low income, multi-dimensional cloud resource utilization and low power consumption high cloud data center issues. This article given a research on the problems of using particle swarm and taboo algorithm based on the above background as the starting point of the research.

Corresponding author e-mail: ylxds@sohu.com

2 Cloud data centre resource scheduling technology

Cloud data center is the user-centric to provide various types of cloud services resources sharing architecture according to need to use the distributed technology [6, 7]. Users can use these on-demand dynamic hardware and software resources, and to pay for services based on service usage. It mainly has five characteristics: resource pooling, intelligent and efficient service and transparent, on-demand service and billed by consumption. As shown in Figure 2:
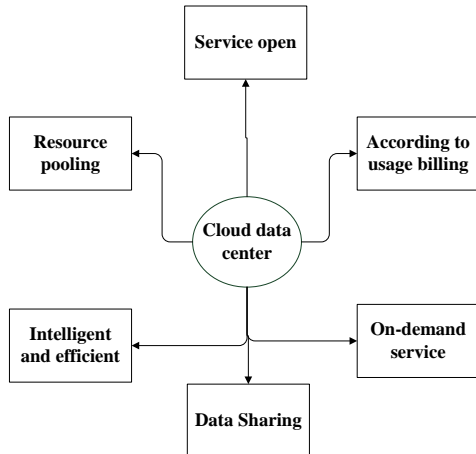
FIGURE 2 On-demand service and billed by consumption.

As the existing cloud data center did not fully consider scheduling a multidimensional cloud resources, did not fully consider the energy consumption of cloud communication between tasks, so there is low cloud resource providers gain, low dimensional cloud resource utilization of existing cloud data center and cloud lack of data center energy consumption is high.

Multidimensional cloud resource efficient scheduling technique is based on three types of resources (compute, storage and bandwidth) resource scheduling technology [8]. Because the demand is not the same of each user task computing, storage, and bandwidth resources, so the cloud data center will have the task of the user can not satisfy a certain category or categories of cloud resources objective situation demands, thereby generating a cloud data center large fragment idle resources [9]. Through using of multi-dimensional cloud resource efficient scheduling technology, more debris resources are fully utilized; more cloud user tasks can be run on the processor, thus effectively improving the multi-dimensional cloud resource utilization and user tasks to meet the more cloud computing, storage and bandwidth requirements.

3 Scheduling algorithm based on particle swarm optimization

3.1 PARTICLE SWARM OPTIMIZATION ALGORITHM

Particle swarm optimization (PSO) algorithm is proposed by Kenney and Eberhart in 1995 population parallel search algorithm based on global optimization [10, 11], through cooperation and competition between groups in the community to achieve optimal particle. Mathematical description of PSO: a population size is n, the i particles in m dimensional search space representation of $X_i = (X_{i1}, X_{i2}, ..., X_{ij}, ..., X_{im})$, flight speed is $V_i = (V_{i1}, V_{i2}, ..., V_{ij}, ..., V_{im})$, the optimal position of individual so far to search is $P_i = (p_{i1}, P_{i2}, ..., P_{ij}, ..., P_{im})$. The particle swarm optimal position is $P_{gbesti} = (p_{gbest1}, P_{gbest2}, ..., P_{gbestm})$. It can update the particle velocity and position according to the Equations (1) and (2):

$$v_i(t) = wv_i(t-1) + \rho_1 \frac{xP_{gbest} - x_i(t)}{\Delta_t} + \rho_2 \frac{xG_{best} - x_i(t)}{\Delta_t}, \quad (1)$$

$$x_i(t) = x_i(t-1) + v_i(t)\Delta_t, \quad (2)$$

$$v_{ij}(t+1) = \omega.v_{ij}(t) + c_1.r_1.(p_{ij} - x_{ij}(t)) + c_2.r_2.(p_{gbestj} - x_{ij}(t)) \quad , \quad (3)$$

$$x_{ij}(t+1) + x_{ij}(t) + v_{ij}(t+1), \quad (4)$$

where, t represents the t iteration, $j = 1,2,...,n$; $j = 1, 2,..., m$; $c_1, c_2 > 0$ are respectively the individual learning factor and social learning factors; t is the current number of iterations, r_1 and r_2 are uniformly distributed random numbers in the range of [0, 1]. ω is the inertia weight coefficient, used to control the effect of history on current speed. In order to balance the global and local search ability, make the ω along with the increase in the number of iterations decreases linearly, can significantly improve the performance of the PSO algorithm. It is given:

$$\omega = \omega_{min} + (iter_{max} - iter) \times (\omega_{max} - \omega_{min}) / iter_{max}, \quad (5)$$

where, ω_{min}, ω_{max} respectively the maximum and minimum weighting factor, $iter$ is the current iteration number, $iter_{max}$ is the total number of iterations. ω_{max} is the initial inertia weight; ω_{min} is the last inertia weight; t_{max} is the maximum number of iterations. Flight speed is $v_i \in [-V_{max}, V_{max}]$, the constraint conditions to prevent particle speed missed optimal solutions, through the improvement of the algorithm further improves the global searching ability of particle swarm.

Particle swarm optimization algorithm works as follows:

1. Random initialization of the position and velocity of a particle.

2. Calculate the fitness value of each particle.

3. If the fitness value is better than the best historical fitness value P_{best}, set the current value of the new P_{best}.

4. Choose the best fitness value of all particles as G_{best}.

5. According to the Equation (1) calculate the particle velocity; according to the Equation (2) to update the position of particle.

6. Stop counting when the termination condition is satisfied, otherwise return to step 2. Generally set as a termination condition is good enough to adapt to reach a preset value or the maximum iteration algebra.

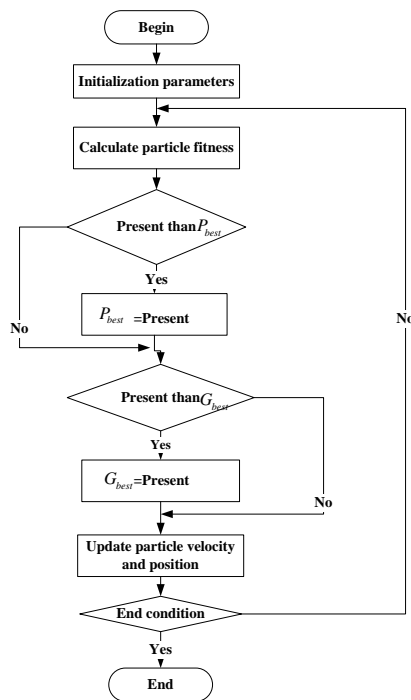Flow chart of particle swarm optimization algorithm as shown in Figure 3.

FIGURE 3 Flow chart of particle swarm optimization algorithm

3.2 MULTIDIMENSIONAL CLOUD RESOURCE EFFICIENT SCHEDULING SYSTEM ARCHITECTURE

There is a multidimensional cloud resources existing cloud data center (computing, storage and bandwidth) [12], Because of not considering multidimensional cloud resource scheduling leads to waste of resources and serious debris, and generally is low cloud resources multidimensional resource utilization, as shown in Figure 4, under the existing cloud data center environments:
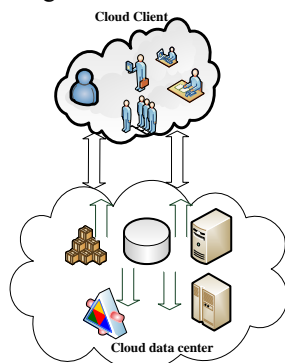
FIGURE 4 cloud data center environments.

1. The advent of cloud user tasks Poisson distribution, resource demand is greater than the actual cloud user tasks demand for resources.

2. Improve resource utilization through effective multidimensional cloud multidimensional resource scheduling methods. Cloud data centers use abstract, key technology virtualization, instantaneous deployment to achieve interconnection cloud data center resources through the network interworking and interoperability. It uses the multi terminal, multi platform, multi network browser access at any time, any place in the distribution according to need, dynamic configuration, flexible expansion, low price, high availability and high reliability computing, storage and bandwidth resources service.

3.3 TABOO SEARCH ALGORITHM

Particle swarm optimization algorithm is a heuristic optimization algorithm proposed by Kennedy and Eberhart, the particle can represent a potential solution to a particular problem, Also is the N dimension optimal solution in $x_i = (x_{i1}, x_{i2}, ..., x_{in})$ vector.

Taboo search is a heuristic method for solving combinatorial optimization problems. It is a global neighbourhood search algorithms, simulation optimization features with human memory function [13]. It is through the local neighbourhood search mechanism and corresponding detour to avoid the taboo search criteria, and to release some of the excellent state of being taboo breaking the ban by the level of diversification and thus guarantee effective exploration to eventually achieve global optimization.

The basic steps taboo search algorithm is as follows.

1. Initialization, create the initial solution; the taboo list is set to null.

2. Determine whether the stop condition is met, if met, the output of the algorithm is stopped; otherwise continue with the following steps.

3. For the best solution candidate solution concentration, whether it meets the aspiration level. If meet the aspiration level, update, update the current solution, turn to step fifth; otherwise, continue to the following steps.

4. Choose the best candidate solutions focused solutions are not taboo as the current solution.

5. Update taboo list.

6. Return to step 2.

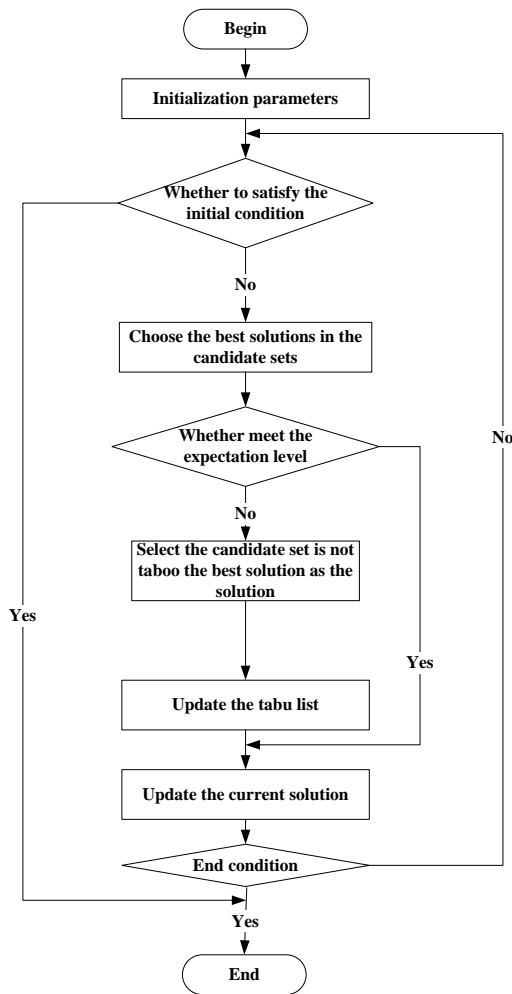Flow chart of taboo search as shown in Figure 5.

FIGURE 5 Flow chart of taboo search

3.4 PARTICLE SWARM AND TABOO UNION ALGORITHM

Based on the particle swarm optimization algorithm is easy to fall into local optimal solution of the shortcomings, combined with a taboo search algorithm "taboo" and "amnesty" thought. Combined with taboos algorithms, this improved particle swarm algorithm from three aspects. That is:

　　1) inertia weight;
　　2) neighbourhood topology;
　　3) learning factor and reproduction.

The proposed algorithm can avoid falling into local optimization solution space, so the three areas for improvement are as follows:

1. The inertia weight is replaced by:

$$\omega = 0.6 + Random / 2 . \tag{6}$$

2. Neighbourhoods topology improved modified for ring topology;

3. Learning factor is modified for the shrinkage factor:

$$v_{id}^{k+1} = k[v_{id}^k + c_1\xi(p_{id}^k - x_{id}^k) + c_2 r(p_{gd}^k - x_{id}^k)], \tag{7}$$

where, K is a function of c_1 and c_2.

$$k = \sigma^2 - \sqrt{\sigma^2 - 2\sigma} / 2 . \tag{8}$$

Reproductive process is as follows: In the each iteration of the PSO algorithm, based on a fixed probability between particles randomized crossover. Then select n optimization particle application of evaluation function in order to maintain the particle population size.

Mapping mechanism is as follows: Cloud data center scheduling matrix corresponds to a user task number line number, column number corresponding to the cloud data center physical processor number. Each element in the matrix represents a user task is running on a physical processor. If the value of the matrix element is "1", then the task on behalf of the user should be able to get the resources and run on the corresponding physical processor. Otherwise, indicating that the user does not get proper task on a physical processor resources and are not operation.

Particle swarm optimization algorithm and taboo process is described as follows:

1. Start, generate the initial population of particles, the number is n, including the position and velocity of each particle. In the particle populations, each particle represents a user task scheduling matrix results.

2. Contraindications cross, the existing population of each particle taboo crossover operation. In the each iteration, the particles in the group are based on certain probability taboo crossover operation.

3. Evaluation, value assessment of each particle in the population is calculated according to the evaluation function of low energy consumption. Reselect the optimization of n particles retain the same population size.

4. Selection. Optimum choice of n particles based on the assessed value of each particle, and the n particles to assess the merits of values in descending order.

5. Movement. According to Equations (9) and (10) is to update the position and velocity of each particle.

$$x_{i+1} = x_i + v_i \cdot t , \tag{9}$$

$$v_{id}^{k+1} = k[v_{id}^k + c_1\xi(p_{id}^k - x_{id}^k) + c_2 r(p_{gd}^k - x_{id}^k)]. \tag{10}$$

6. Update History optimal position and global best position.

P_{best} is the particle's histories that have experienced the best position, and G_{best} is the position of the particles have a global optimum.

In the each iteration, the fitness function of each particle's position and gain new experiences optimized to compare their position and the position of the entire particle swarm optimization withstand algebra. If it is more than the current P_{best} or G_{best}, then update P_{best} or G_{best}, return to the first two steps.

The flow chart of Particle swarm and taboo union algorithm is as shown in Figure 6.
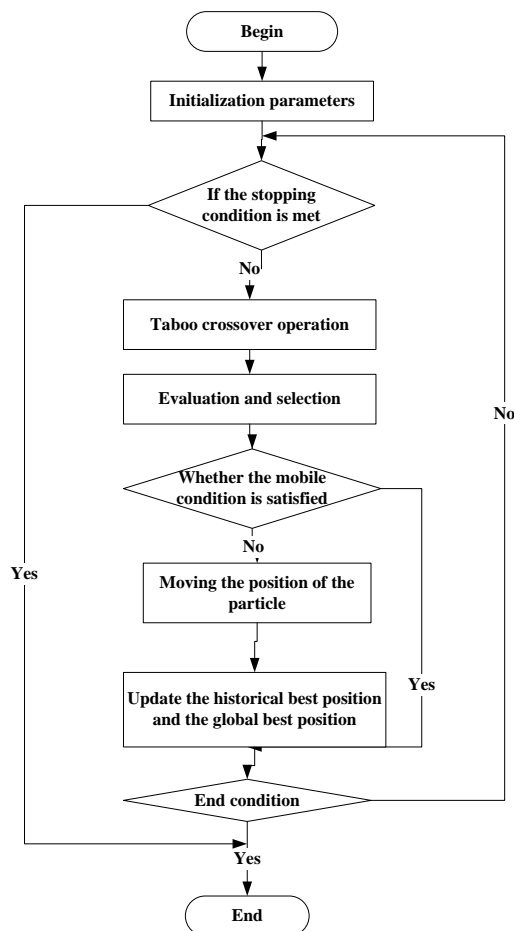
FIGURE 6 Flow chart of Particle swarm and taboo union algorithm

4 Experiment and analysis

In order to verify the stability of particle swarm optimization algorithm and taboo search (PSO-TAO). This paper conducted a series of experiments. Experiments run on the CloudSim architecture. CloudSim combines the flexibility and dynamic characteristics of the economic model and service quality, and these features are not supported other cloud emulator [14].

CloudSim provides four types of entities that is the cloud information service registration, data center; broker represents the interests of users and the auction market manager [15]. CIS cloud information service registry mapping user requests to the cloud data center. Cloud data center combines the distributed hardware, database, storage equipment, application software and operating system to build a pool of resources and according to the needs of users to create virtual machine computing, storage and bandwidth resources. In the cloud data center, users need to consider their offer and reasonable deadlines and restrictions on their tasks offer.

Particle swarm optimization algorithm, tabu search algorithm and random searching algorithm (RS) compare the performance; Use a different set of virtual machines through a lot of experiments. Experimental environment

settings are as follows: Experimental environment setting is 500-3000 different numbers of virtual machines; the processor number is from 300 to 12,000, the interval is 2000. Interval task arrives subject to random distribution of a range of 5-100, and consider the three resources: computing, storage and bandwidth resources. This article is a set of experiments performed, and the resource range is from 100 to 12,000, is 500. In each experiment, automatically generates a virtual machine nodes ranging from 50 to 10,000. Right weight for each virtual machine is a virtual machine needs three resources: computing, storage and bandwidth resources.

First, integration of resources and scheduling mechanisms of particle swarm optimization algorithm proposed taboo compare and random search algorithm. Secondly, the resource scheduling mechanism fusion experiments with the proposed scale can reduce the energy consumption of a cloud data center.
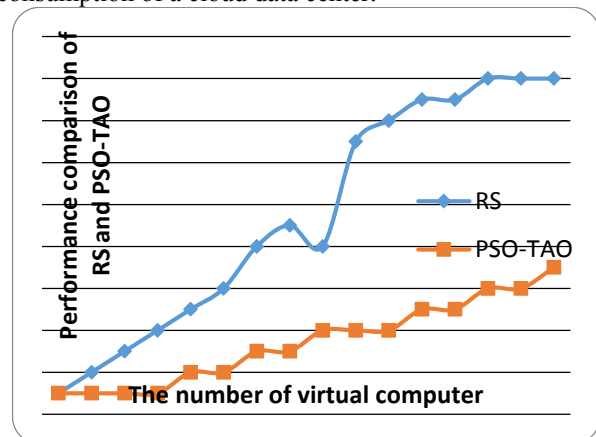

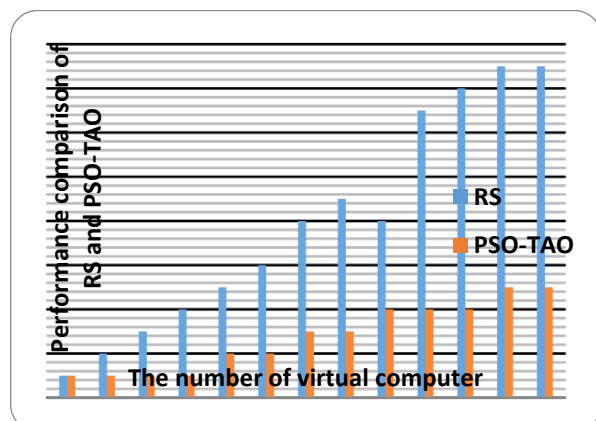FIGURE 7 Curve graph for the performance comparisons of RS and PSO-TAO

FIGURE 8 Bar chart for performance comparisons of RS and PSO-TAO

Figures 7 and 8 show the particle swarm and taboos through histograms and graphs random search algorithm to optimize the energy consumption of the algorithm comparison. The results from the 10 groups on particle swarm optimization algorithm and taboo search algorithm randomly display: First, particle swarm optimization algorithm and taboo energy consumption is lower than the random search algorithm from 50.31% to 76.30%, it average value is 63.6%. Secondly, the proposed particle

501

swarm optimization algorithm and taboos less than the effect of random search algorithm became clear, as the number of virtual machines increases. Finally, when the number of virtual machines is increasing, energy particle swarm optimization algorithm is proposed and taboos consumption stabilized. In summary, the proposed particle swarm optimization algorithm and taboos energy consumption was significantly lower than the random search algorithm, and with the increasing number of virtual machines, the energy saving effect became clear and stable. Particle swarm optimization algorithm and taboo search is proposed heuristic optimization. So the particle swarm optimization algorithm and taboo energy consumption is lower than the random search algorithm stability. With the increasing number of virtual machines, idle processor also increases the number of idle processors will consume more energy. Thus, the particle swarm optimization algorithm and taboo energy consumption along with the increasing number of virtual machines is less than the effect of random search algorithm increasingly apparent. All in all, as heuristic optimization algorithms and resources to integrate new ideas, particle swarm optimization algorithm and taboos raised more than random search algorithm suitable for large-scale cloud data centers. These experiments show that the energy consumption according to the characteristics and needs of the proposed resource scheduling mechanism integration cloud users can fully meet the quality service and effectively reduces the computing, storage and bandwidth consumed energy.

5 Conclusions

Current cloud data center resource scheduling technology exists residual computing and storage resources are still consuming energy in an idle state, inter task communication bandwidth resource energy consumption has been neglected. To solve these problems, we propose a scheduling mechanism based on PSO optimized integration resources designed to minimize the task completion time under the premise of reducing the energy consumed by data centers in the cloud. Based on the experimental architecture CloudSim comprehensive experiments and randomized algorithms by comparing the energy consumption, this paper verifies the minimum completion time under the premise of the proposed particle swarm optimization algorithm and taboos can save energy consumption by up to 63.6%. With the increase in the number of virtual machines, particle swarm optimization algorithm and taboo energy consumption will be lower than the random search algorithm and stabilized. This fully verifies the advantages of the integration of resources scheduling mechanisms include: large-scale, stability and integration of resources.

Acknowledgments

References

[1] Ostermann S, Iosup A, Yigitbasi N, Prodan, R, Fahringer T, Epema D 2010 A Performance Analysis of EC2 Cloud Computing Services for Scientific Computing *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, Cloud Computing* **34**(4) 115-31

[2] KurkovskyS, Harihar K 2006 Using ubiquitous computing in interactive mobile marketing *Personal and Ubiquitous Computing* **10**(4) 227-40

[3] Hadji M, Louati W, Zeghlache D 2011 Constrained Pricing for Cloud Resource Allocation *The 10th IEEE International Symposium on Network Computing and Applications (NCA)* 359-65

[4] Mihailescu M, Teo Y M 2010 Strategy-Proof Dynamic Resource Pricing of Multiple Resource Types on Federated Clouds *Algorithms and Architectures for Parallel Processing* **6081** 337-50

[5] Shang S, Jiang J, Wu Y, Huang Z, Yang G, Zheng W 2010 DABGPM: A Double Auction Bayesian Game-Based Pricing Model in Cloud Market *Network and Parallel Computing* **6289** 155-64

[6] Feng Y, He Z 2003 Optimization of agitation, aeration, and temperature conditions for maximum *β*-mannanase production *Enzyme and Microbial* **32**(2) 282-9

[7] Zhang Y-H, Zhang F 2013 A New Time Synchronization Algorithm for Wireless Sensor Networks Based on Internet of Things *Sensors and Transducers* **151**(4) 95-100

[8] Cao X, Chen J, Sun Y 2009 An Interface Designed for Networked Monitoring and Control in Wireless Sensor Networks *Computer Standards and Interfaces* **31**(3) 579-85

[9] Younge A J, von Laszewski G, Wang L, Lopez-Alarcon S, Carithers W 2010 Efficient resource management for Cloud computing environments *International Green Computing Conference* 357-64

[10] Menache I, Ozdaglar A, Shimkin N 2011 Socially Optimal Pricing of Cloud Computing Resources *Proceedings of the 5th International ICST Conference on Performance Evaluation Methodologies and Tools* 16-25

[11] Peoples C, Parr G, McClean S 2011 Energy-aware data centre management *National Conference on Communications (NCC)* 1-5

[12] DhimanG, Marchetti G, Rosing T 2010 Green: A System for Energy-Efficient Management of Virtual Machines *Transaction on Design Automation of Electronic System* **16**(1) 1-27

[13] Liao X, Hu L, Jin H 2010 Energy optimization schemes in cluster with virtual machines *Cluster Computing* **2**(13) 113-26

[14] Ma S 2012 A Review on Cloud Computing Development *Journal of Networks* **7**(2) 305-10

[15] Rodero I, Jaramillo J, Quiroz A, Parashar M 2010 Energy-efficient application-aware online provisioning for virtualized clouds and data centers. International Green Computing Conference 31-45

Authors

Dong-sheng Xu, born in November, 1970, Yulin, Shanxi, China

Current position, grades: professor in Xi'an University.
University studies: MS degree in Computer Application from Northwestern Polytechnical University in 2008.
Scientific interest: computer application in agriculture, mobile internet technology.

Feng Zhang, born in June, 1980, Yulin, Shanxi, China

Current position, grades: associate professor in Yulin University
University studies: MS degree in Computer science from Xidian University in 2009.
Scientific interest: cloud integrated manufacturing technology, the modeling of complex systems, internet of things applications.

Temperature field and thermal stress field of continuous casting roller bearing

Ze Liu[1], Gongfa Li[1, 2], Honghai Liu[2*], Guozhang Jiang[1], Jia Liu[1]

[1]*College of Machinery and Automation, Wuhan University of Science and Technology, Wuhan 430081, China*

[2]*Intelligent Systems & Robotics Group, School of Computing, University of Portsmouth, Portsmouth, PO1 3HE, United Kingdom*

Received 1 August 2014, www.cmnt.lv

Abstract

The work condition of continuous casting roller bearing was high temperature, large load and complex. If the heat, which the bearing accepted could not be distributed effectively and cooling was not in place, the high temperature would cause bearing thermal deformation and rupture under high loads, resulting in bearing premature failure, affecting the entire casting steel production. The cooling way for the continuous casting roller bearing seat was by the internal cooling water holes which in the bearing seat. Based on the ANSYS, the temperature field of continuous casting roller bearing seat was analysed. The three-dimensional geometric model and the CAE model of the bearing seat was established, and the synthesize convection heat transfer coefficient of bearing seat would be obtained through analysed the principle of heat transfer. Took it as the boundary condition, and load the temperature in the real-time condition. Analysed temperature field of the continuous casting roller bearing seat in the stable work condition, got the overall distribution of bearing seat. Meanwhile, it could draw the conclusion that thermal load what bearing carried was the cause of bearing damage through the thermal stress analysis of the continuous casting roller bearing. It provided advice and guidance on appropriate optimization of the bearing to improve bearing life. And just the above analysis provided a theoretical guidance for the design and installation of continuous casting roller bearing

Keywords: continuous casting roller, bearing seat, temperature field, CAD three-dimensional geometric model, CAE model, stress field, thermal coupling field

1 Introduction

Continuous casting roller bearing was an important part in continuous casting machine, which provided important support for continuous casting roller, and the work environment was complex. The billet which come out from the continuous casting machine crystallizer was forced cooling through spraying water in the secondary cooling zone(the continuous casting machine's line segment, curved segments and sector segments called the secondary cooling zone), in order to make the billet shell quickly cooled to prevent breakout. Nevertheless, when the billet come out from the crystallizer and reached the casting machine straightening section, it was still at a temperature above 900 ℃.And the billet contacted with the continuous casting roller directly, at this moment, the highest point of the bearing seat at the two ends of the continuous casting roller was very close with the billet. If the heat that bearing accepted could not be effectively dissipated, and the cooling inadequate, then the high temperature would cause bearing thermal deformation and rupture under large load, resulting in premature bearing failure, thereby affecting the whole casting steel production. Under the current experimental conditions, most researchers used a thermal sensor to measure the temperature of the research object. However, due to the structure of the bearing, it could not be installed a temperature sensor inside the bearing,

resulting the temperature of the bearing inner ring raceway, outer ring raceway and rolling element could not be measured. So that the data of theoretical analysis could not be verified, this was a serious impediment to the development of theoretical bearing thermal analysis studies. The bearing was analysed by ANSYS finite element analysis software in this paper, the ANSYS finite element analysis software had a great function of thermal analysis and processing, and it could easily solve the coupled thermo mechanical problem. Through the simulation analysis on the continuous casting roller bearing, the distribution of temperature field, stress field and comprehensive stress of continuous casting roller bearing could be obtained. Investigated the mechanism of bearing damage, and analysed the impact of thermal stress on the bearing damage and bearing life, it had the vital significance to the structure optimization of bearing under high temperature. Therefore, the interior of the bearing seat must be cooling to ensure the bearings run under the specified temperature range.

*****Corresponding author** e-mail: honghai.liu@port.ac.uk

2 The CAD and CAE model of continuous casting roller bearing

2.1 THE CAD THREE-DIMENSIONAL GEOMETRIC MODEL OF CONTINUOUS CASTING ROLLER BEARING

According to the bearing model and drawings that provided by a steel plant, using the current mainstream three-dimensional CAD design software to build a three-dimensional CAD integral assembly model of the casting roller bearing, as shown in Figure 1.

In the assembly model of continuous casting roller bearing, the bolt hole on the end cover has been simplified without affecting the casting roller bearing thermal stress, temperature field and stress field analysis.
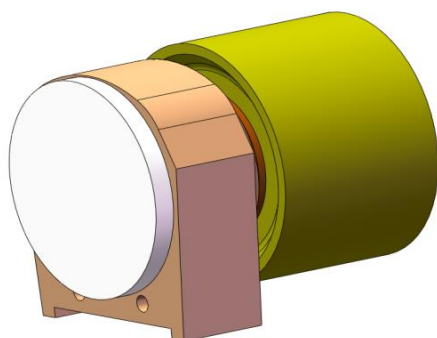
FIGURE 1 The three-dimensional CAD assembly model of continuous casting roller bearing

2.2 THE CAE MODEL OF CONTINUOUS CASTING ROLLER BEARING

In strict accordance with the size and structure of 3D CAD model, using CAE pre-processing simulation software to build the grid model for CAE analysis.

In the process of solving thermal analysis，using ANSYS for its calculation and analysis. In the ANSYS finite element analysis software, according to the requirements of the ANSYS unit type, while taking into account the entire bearing assembly models were three-dimensional solid model, the author chose the common SOLID70 unit. Mesh generation was performed in the bearing assembly model, as shown in Figure 2. And meshing assembly model was the most difficult task, because of its internal cooling water pipe, the internal gridding must be distinguished, in the shape of extremely irregular, we used tetrahedral mesh and pentahedral mesh. However, in this part, the mesh quality was not better than hexahedral mesh, but the number of tetrahedral meshes and pentahedral mesh was few, it had less impact on the temperature field of the entire model.
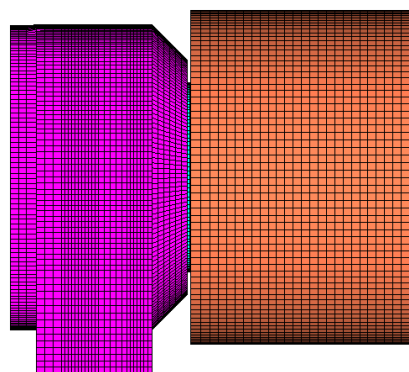

FIGURE 2 The three-dimensional mesh model of continuous casting roller bearing

Mechanical properties of materials of the main components of continuous casting roller bearing assembly, such as shown in Table 1 [1].

TABLE 1 Mechanical properties of the materials of the main components

Main components	Materials	Strength limit σ_b (MPa)	Yield limit σ_s (MPa)
bearing seat, end cover	Q345-D	470-630	345

Continuous casting roller assembly was comprised of a bearing seat, a bearing cover, bearing ring and continuous casting roller. Physical parameters of model materials was shown in Table 2 [2-4], bearing seat material was Q345-D.

TABLE 2 Physical parameters of the materials

Materials	Density (kg/m³)	Specific heat (J/kg·°C)	Thermal conductivity (W/m·K)	Linear expansion coefficient (10^{-6}/°C)	Elastic modulus (Gpa)	Poisson ratio
Q345-D	7.85E3	460	44	13	206	0.28

3 The mathematical model description and boundary conditions determination of continuous casting roller bearing thermal analysis

Thermal simulation analysis was used to calculate the temperature distribution of a system or components, resulting in obtaining the corresponding thermal physical parameters (such as the system gained or loss calorific value, thermal gradient, heat flux etc.). Because it was calculated by temperature as the basic parameters, the thermal analysis was also known as the temperature field

calculation. Due to temperature rise caused by the thermal radiation of continuous casting roll bearing in the steel-making process did not change with time, so the heat transfer process of continuous casting roll bearings could be approximately regarded as the steady-state heat transfer process, met the following heat conduction equation [5],

$$\frac{\partial T}{\partial t} = \frac{k_2}{\rho C_p} \left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} + \frac{\partial^2 T}{\partial z^2} \right), \tag{1}$$

In the equation, x, y, z was the coordinates of system; k_2 was the thermal conductivity coefficient of the material, $W/(m \cdot K)$; ρ was the density of the material, Kg/m^3; C_p was the specific heat of the material, $J/Kg \cdot °C$; T was the temperature function, $°C$; t was the time, s.

Throughout the heat transfer system, also included thermal conduction and thermal radiation of high temperature billet for continuous casting roller, and thermal radiation of high temperature billet for bearing seat. Taking into account a lot of radiation heat transfer plane, and it was highly nonlinear, we have to take a long time to calculate. Therefore, the author simplified the radiation to take the form of convective heat transfer instead of the form of radiation heat transfer, which was to select an equivalent convective heat transfer coefficient instead of radiation heat transfer.

In general, the natural convection coefficient was 5 to 10, the equation for the calculation was [6],

$$h = 1.826\left(\frac{T_s}{T_s - T_a}\right)^{1/3}, \qquad (2)$$

where h was the convective heat transfer coefficient of casting roller bearing and the ambient air; T_s was the temperature of continuous casting roller bearing, $°C$; T_a was the temperature of the air around the continuous casting roller bearing, $°C$.

According to the theory of heat transfer [7], using the Equation (3) to calculate the convection heat transfer coefficient of continuous casting roller bearing, bearing seats, continuous casting roller and air.

$$h_c = N_u \lambda / L, \qquad (3)$$

where h_c was the surface heat transfer coefficient (i.e., heat exchange coefficient), $W/(m^2 \cdot K)$; N_u was the Nusselt number; λ was the thermal conductivity, $W/(m \cdot K)$; L was the characteristic length, m.

The author translated the radiation heat transfer of billet and surrounding environment into the convective heat transfer, their equivalent convective heat transfer coefficient expression [8] was:

$$h_r = \varepsilon B(T_s^2 + T_a^2)(T_s + T_a), \qquad (4)$$

where h_r was the equivalent convective heat transfer coefficient of the continuous casting roller and the ambient air, $W/(m^2 \cdot K)$; B was the Boltzman number, 5.67e-8 $W/(m^2 \cdot K^4)$; ε was the blackness, the radiation heat transfer coefficient of the billet and surrounding environment was 0.8.

Therefore, the composite heat transfer coefficient was,

$$h = h_c + h_r. \qquad (5)$$

If the temperature of ambient air around the continuous casting roller was 300°C, and the convective coefficient of the continuous casting roller and ambient air was $15W/(m^2 \cdot K)$, it could use the Equation (4) to calculate the convection coefficient after conversion, so, the convection coefficient of the continuous casting roller and surrounding environment was 33.4 $W/(m^2 \cdot K)$.

4 The thermal simulation analysis of continuous casting roller bearing

Because the heat source of the temperature field was the high temperature billet which was out of the crystallizer, the working environment of continuous casting roller bearing was complex, and the bearing was wrapped in the inside of the bearing seat. According to the bearing seat temperature, roller temperature, ambient temperature and cooling water temperature to calculate the bearing temperature, and used ANSYS to simulate.

The distribution of temperature field of the continuous casting roller bearing assembly has been calculated through the thermal analysis, as shown in Figure 3. The simulation results indicated the temperature of the entire assembly reduced from the continuous casting roller to bearing seat, showed stratified distribution. And the temperature field was essentially a symmetrical distribution.
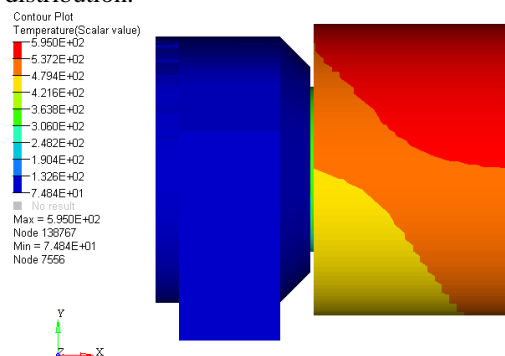

FIGURE 3 The temperature field of continuous casting roller bearing assembly

The analysis result of temperature field of continuous casting roller bearing seat as shown in Figure 4. Due to the bearing seat was provided with cooling water, it made the temperature inside the bearing seat cavity remain at normal working temperature range, so the temperature inside the bearing seat was not very high [9]. The highest temperature place was the part, which close to the casting roller end. Because of the continuous casting roller temperature was very high, this part was subjected to the thermal radiation from the millet and continuous casting roller, then the temperature of this part increased.
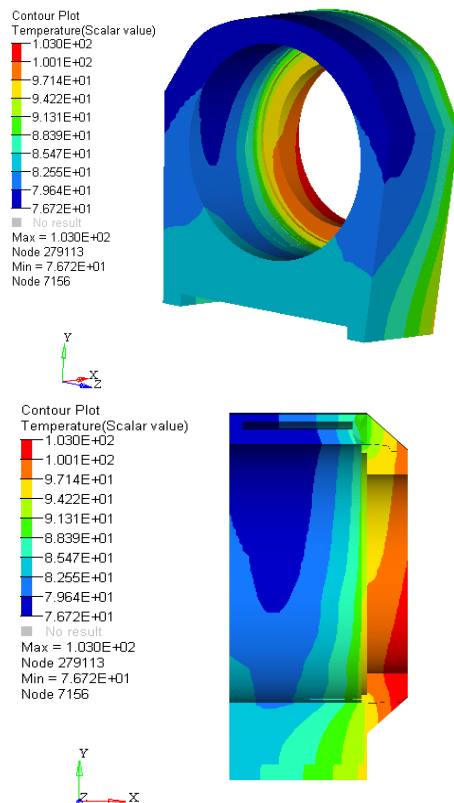
FIGURE 4 The temperature field of continuous casting roller bearing seat

Figures 5 and 6 were the stress field and thermal-stress coupled field simulation results of the continuous casting roller bearing when the bearing seat was provided with cooling water. It could be seen from the figure that the distribution of continuous casting roller bearing stress field was that the part close to the bottom of the bearing seat had a maximum stress and the stress of the part at the top of the bearing seat was 13.77MPa, which was similar to the actual situation. Thermal-stress coupled field distribution of casting roller bearing were the stress increased in the direction of X axis and decreasing in the direction of Y axis, which was consistent with the distribution of temperature field, also consistent with the actual situation.
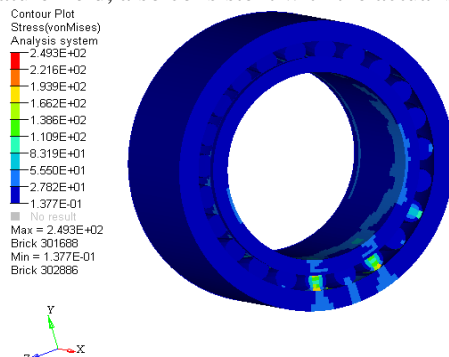
FIGURE 5 The stress field of the continuous casting roller bearing when the bearing seat was provided with cooling water
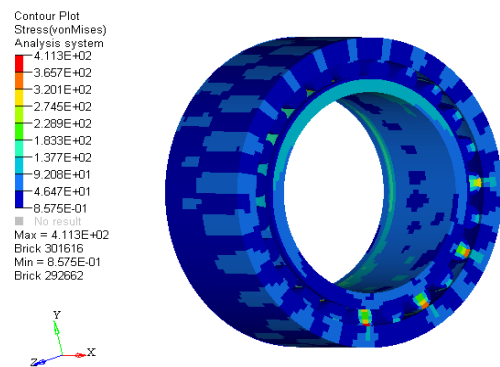

FIGURE 6 The coupled thermo-mechanical field of the continuous casting roller bearing when the bearing seat was provided with cooling water

Extracted the results and compared with the measured temperature of the temperature sensor (provided by the steel mill), then we could know that the simulation temperature of the outer surface of the bearing seat and continuous casting roller was essentially same to the measured temperature, and the maximum error between them was less than 9%. It indicated that the finite element simulation analysis results of the continuous casting roller bearing temperature field were correct and effective (Table 3).

TABLE 3 The simulation temperature of the main parts of the continuous casting roller bearing seat contrasted with the measured temperature

Position	Simulation temperature (°C)	Measured temperature (°C)	Error
Top of bearing seat	83.8	90	6.9%
Inclined plane of bearing seat	98.9	110	8.4%
Side of bearing seat	81.2	75	8.3%

5 The thermal stress analysis of continuous casting roller bearing

In order to easily observe the impact of thermal stress caused by temperature on bearing, here listed the stress field [10] and the thermal-mechanical coupling field sketch map of the bearing under the condition of cooling water. By comparison, we could clearly see that the bearing was subjected to a great stress caused by temperature.

Figures 7 and 8 were the stress field and coupled thermo-mechanical field simulation results of the continuous casting roller bearing outer ring when the bearing seat was provided with cooling water. It could be seen from the figure that the outer ring was obviously subjected to the thermal stress caused by the temperature under the action of coupled thermo-mechanical.
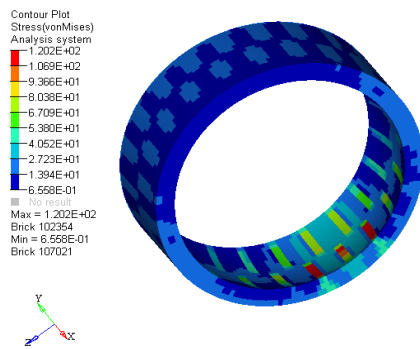
FIGURE 7 The stress field of the continuous casting roller bearing outer ring when the bearing seat was provided with cooling water
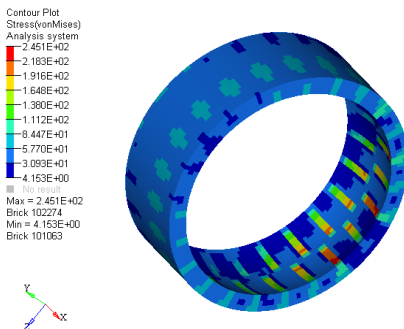

FIGURE 8 The coupled thermo-mechanical field of the continuous casting roller bearing outer ring when the bearing seat was provided with cooling water

Figures 9 and 10 were the stress field and coupled thermo-mechanical field simulation results of the continuous casting roller bearing inner ring when the bearing seat was provided with cooling water. It could be seen from the figure that the stress obviously concentrated in the contact part of the rolling element and inner ring.

FIGURE 9 The stress field of the continuous casting roller bearing inner ring when the bearing seat was provided with cooling water
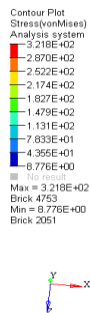

FIGURE 10 The coupled thermo-mechanical field of the continuous casting roller bearing inner ring when the bearing seat was provided with cooling water

Figures 11 and 12 were the stress field and coupled thermo-mechanical field [10] simulation results of the continuous casting roller bearing rolling element when the bearing seat was provided with cooling water. It could be seen from the figure that the part which the bearing was subjected to the maximum stress was the rolling element, and the minimum stress was the outer ring. Meanwhile, each of the rolling element was subjected to stress differently and the gap was obvious. In this paper, the stress analysis was the steady-state analysis, it could be speculated that when the bearing during operation, the stress which the rolling element was subjected was variable, and had a great span, but it would cause the fatigue failure of the bearing.
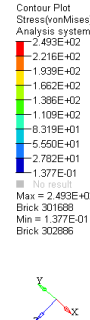

FIGURE 11 The stress field of the continuous casting roller bearing rolling element when the earing seat was provided with cooling water
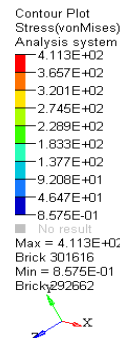

FIGURE 12 The coupled thermo-mechanical field of the continuous casting roller bearing rolling element when the bearing seat was provided with cooling water

As shown in Table 4, statistical analysis of the maximum stress and thermal stress simulation results of the casting roller bearing showed that the stress of continuous casting roller bearing almost doubled, the influence of temperature on the bearing load was also very obvious, and less than the allowable stress only 20.4MPa.Therefore, when the continuous casting roller bearing worked in high strength conditions, it was rather liable to be invalidated. And this situation indicated the continuous casting roller bearing failure was significantly related to temperature. Although the inner ring temperature was the highest, the maximum stress which it suffered was still smaller than the rolling element, thus, the rolling element was the part which suffered the maximum stress.

TABLE 4 The maximum stress and thermal stress of each part of the continuous casting roller bearing result comparison

	Maximum stress (MPa)	Maximum thermal stress (MPa)	Allowable stress of bearing material (MPa)
Continuous casting roller bearing	249.3	411.3	
Outer ring	120.2	245.1	431.7
Inner ring	176.0	321.8	
Rolling element	249.3	411.3	

6 Conclusion

When the bearing seat in the steady state operation, using the finite element analysis software to analysis the distribution of the temperature field, then, according to the analysis result, it could be seen that the inner ring of the bearing had a highest temperature and the outer ring lowest when the bearing seat was provided with cooling water, and the temperature transfer from the inner to the outer ring, but in the direction of X axis (i.e. from the continuous casting roller to the end cap),a large temperature gradient would generate a high thermal expansion stress. Under the extreme conditions, the bearing temperatures up to 103 ℃, higher than the normal operating temperature of high-temperature grease, which made the lubrication system failure, resulting in the failure of the bearing. The analysis provided theoretical guidance for the design and installation of the continuous casting roller bearing seat.

Meanwhile, by the result of the bearing thermal stress analysis under the steady-state operation of the bearing seat, it could be seen that the stress was concentrated in the rolling element, and the maximum stress was close to the allowable stress of the bearing material. Therefore, in order to improve the life of the bearing by reduced the maximum stress of bearing, we could consider to change the structure of the rolling element without changing the overall structure of the bearing. In summary, it provided theoretical guidance for the design and installation of continuous casting roller bearing.

Acknowledgments

References

[1] Zhang R, Zhao J, Wang H 2014 Structural Strength Analysis of a Single-Cylinder Diesel Engine Block and Main Bering Considering Bolt Preload *Small Internal Combustion Engine and Motorcycle* **43**(2) 17-21 *(in Chinese)*

[2] Yang S, Tao W 2006 *Heat Transfer* (4th edition) *Higher Education Press:* Beijing Chapter 16 *(in Chinese)*

[3] Xu H, Cai C, Yan J 2003 *Mechanical Design Handbook* (2nd edition) China Machine Press Beijing *(in Chinese)*

[4] Datong Qin, Liyang Xie 2013 *Modern Mechanical Design Manual (bearing) Chemical Industry Press* Beijing *(in Chinese)*

[5] Tian J, Hong J, Zhu Y, Li X, Guo J 2012 Thermo-Mechanical Coupling Model and Dynamical Characteristics of Machining Spindle-Bearing System *Journal of Xi'An JiaoTong University* **46**(7) 63-8 *(in Chinese)*

[6] Ding C 2010 Research on The Heat-Transfer Mechanism of High Speed Precise Angular Contact Ball Bearing *Mechanical Research & Application* 33-5 *(in Chinese)*

[7] Liu Y, Fan C 2014 Development and Prospect of Heat Treatment Technology for Rolling Bearing Parts Made of High Carbon Chromium Bearing Steel *Heat Treatment of Metals* **39**(1) 25-9 *(in Chinese)*

[8] Ren F, Wang Y 2013 Influence of Transient Effect on Thermal Elastohydrodynamic Lubrication of Roll Bearing *Forging & Stamping Technology* **38**(3) 133-6 *(in Chinese)*

[9] Chen Y, Lu J, Ni X 2007 Temperature and Thermal Stress Fields during the Laser Irradiating a Metal Plate *Journal of Huazhong University of Science and Technology (Natural Science Edition)* **35** 129-32 *(in Chinese)*

[10] Chen S, Li S, Li X 2014 Thermal-Stress Coupled-Field Analysis of Auto Gearbox Test Platform Spindle System *Machine Tool & Hydraulics* **42**(7) 107-11 *(in Chinese)*

Authors

Ze Liu, born in 1989, Hubei, China

Current position, grades: M.S. degree student in mechanical design and theory at Wuhan University of Science and Technology.
University studies: B.S. degree in mechanical engineering and automation from Wuhan Institute of Bioengineering, Wuhan, China, 2013.
Scientific interests: mechanical CAD/CAE, signal analysis and processing.

Gongfa Li, born in October, 1979, Hubei, China

Current position, grades: associate professor at the College of Machinery and Automation, Wuhan University of Science and Technology.
University studies: PhD degree in mechanical design and theory from Wuhan University of Science and Technology in China.
Scientific interests: modelling and optimal control of complex industrial process.
Publications: 100.

Honghai Liu, born in 1973, China

Current position, grades: professor in Intelligent Systems, head of Intelligent Systems and Biomedical Robotics, University of Portsmouth.
University studies: PhD in Intelligent Robotics in 2003 at Kings College, University of London, UK.
Scientific interests: approximate computation, pattern recognition, multi-sensor based information fusion and analytics, human machine systems.
Publications: 300.
Experience: Research at King's College London, University of Aberdeen.

Guozhang Jiang, born in December, 1965, Tianmen, China

Current position, grades: professor of Industrial Engineering, assistant dean at the college of machinery and automation, Wuhan University of Science and Technology.
University studies: PhD degree in mechanical design and theory at Wuhan University of Science and Technology, China, in 2007.
Scientific interests: computer aided engineering, mechanical CAD/CAE and industrial engineering and management system.
Publications: 120.

Jia Liu, born in 1990, Hubei, China

Current position, grades: M.S. degree student in mechanical design and theory at Wuhan University of Science and Technology.
University studies: B.S. degree in mechanical engineering and automation at Wuchang institute of Technology, Wuhan, China, in 2012.
Scientific interests: mechanical CAD/CAE, signal analysis and processing.

Review of the development of ocean data assimilation

Zhenchang Zhang*, Changying Wang

College of Computer and Information Science, Fujian Agriculture and Forestry University, Fuzhou 350002, China

Received 1 September 2014, www.cmnt.lv

Abstract

Data assimilation compensates for the deficiency of a numerical model and minimizes the short-term forecasting error by combining observation data and numerical results. Data assimilation has become a popular research topic all over the world in recent years. The development of ocean data assimilation is introduced in this paper. 4D variational and Kalman filter methods are considered the best means of data assimilation. Thus, these two methods are described in detail. Several novel research methods of assimilation, including assimilation with a constraint condition and dimensionality reduction, are discussed.

Keywords: data assimilation, 4D variational, Kalman filter

1 Introduction

Data assimilation originated from early meteorological methods. Assimilation methods integrate different space–time observations into a numerical model through physical and temporal constraints. During the development of ocean observation technology, the observations spread in a larger space–time range and promote assimilation methods. Data assimilation methods have been applied in many ocean operational numerical models. Using observations from various sources, assimilation methods improve the initial background accuracy and forecast capability. In this study, we review the development process of data assimilation, introduce the characteristics of these methods, and make a simple comparison of these methods. Furthermore, 4D variational and Kalman filter methods are discussed in detail.

2 Development of data assimilation

2.1 INTERPOLATION METHOD

Early methods are simple. They maximize the interpolation (such as linear or polynomial interpolation), and the information at the observation position is interpolated in the background grid. Interpolation methods ignore the error between observations and numerical results and do not consider the relationship among multi variables. Therefore, interpolation methods are less theoretical. Nudging, an interpolation method, was proposed by Gilchrist [1] in 1954. The nudging factor is introduced into the model equation. The observations affect the numerical values in the grid points within the radius of influence, and the nudging factor is inversely proportional to the distance from the observation position. The difference between model simulation and analysis of observations is reduced by revising the numerical values

during the model process. An appropriate nudging factor should be determined. If a large factor is selected, model simulation will converge to the observations very fast, and the time steps become insufficient for dynamic adjustment. By contrast, a small factor increases the model error before the nudging adjustment is implemented.

2.2 OPTIMAL INTERPOLATION METHOD

Statistical theory was incorporated into assimilation methods until Gandin proposed the optimal interpolation (OI) method in 1963 [2]. OI considers the observation and model errors and determines the maximum joint probability in the law of maximum likelihood estimation. Several assumptions exist in OI method. Examples of such assumptions include the background and observation errors are unbiased, error distributions have a Gauss function, and the observation operator is linear. The amount of computation is small, and implementing the method in a time-invariant model is easy. Since 1970, OI method has been utilized widely in many operational numerical forecast systems. The analysis field is derived from observation increment multiplied by the optimal weight matrix and added to the background field. The optimal weight matrix is equal to the background field error covariance matrix in observation space multiplied by the inverse of the total error covariance matrix (background error covariance matrix plus observation error covariance matrix) in model space. White [3] assimilated Geosat altimetry sea level observations into a realistic wind-driven numerical synoptic ocean model of the California current in 1990. Mellor and Ezer [4] and Ezer and Mellor [5] proposed a continuous OI assimilation scheme with a primitive equation and multilayer numerical model and projected the surface observation information into a deep ocean. Tacker et al. [6] implemented the OI method and assimilated expendable bathythermographic

---

*Corresponding author e-mail: stdin@fafu.edu.cn

(XBT) data for 1972 to 1991 into a hybrid coordinate ocean model (HYCOM) for the Atlantic Ocean. Bluelink [7] is Australia's contribution to the Global Ocean Data Assimilation Experiment (GODAE). BODAS is an ensemble optimal interpolation system that estimates background error covariance.

2.3 3D VARIATIONAL METHOD

Variational methods are utilized to determine the maximum value of the objective function to measure the distance between the model and observation fields through the Lagrange function. 3D variational method is essentially equivalent to OI method. Owing to the difference in the solutions of the two methods, the analysis fields are not exactly equal. Fu [8] discussed the similarities and differences between 3DVAR and Ensemble Optimal Interpolation. Dobricic et al. [9] described the development and evaluation of an oceanographic 3D variational (3D-VAR) data assimilation scheme based on a novel specification of the background error covariance. The new 3D-VAR scheme allows for regional variability of the background error covariance matrix, complex coastal boundary conditions, and variable bottom topography.

2.4 PHYSICAL SPACE STATISTICAL ANALYSIS

Another 3D assimilation method called physical space statistical analysis system (PSAS) was proposed by Cohn in 1998 [10]. The amount of computation in PSAS is less than that in 3D variational or OI method. The reduction in computation is due to two aspects. First, the objective function is based on observation space rather than model space. Generally, the dimensions of observation space are far less than those of model space. Second, solving the inverse background covariance matrix is avoided in the resolution process of PSAS [11].

2.5 4D VARIATIONAL METHOD

The methods described above are time invariant. 3D variational method is expanded to 4D method when the time dimension is involved in the objective function. Among variational approaches, 4D variational method provides the best estimate of the initial condition, which leads to an accurate fitting forecast during the assimilation of time windows. Powell et al. [12] applied the 4D variational method in the Intra-Americas Sea. Smith and Ngodock [13] also applied this method to the base of the Navy Coastal Model.

2.6 KALMAN FILTER

Kalman filter is another 4D method. The amount of computation in the extended Kalman filter is considerably large, and the method cannot be applied in operational forecasting systems. Thus, many simplified versions of the

Kalman filter have been proposed; Ensemble Kalman filter [14] is one of them. Ensemble Kalman filter (EnKF) method with an ensemble size of K allows for K number of model integrations (such as OI or 3D variational) computation. However, the computation cost is less than that in the extended Kalman filter method. Simplified methods include reduced-order extended Kalman filter (ROEK) [15, 16], singular evolutive extended Kalman filter (SEEK) [17], ensemble Kalman smoother (EnKS), error subspace statistical estimation (ESSE) [18], singular evolutive interpolated Kalman filter (SEIK) [19], and reduced-order information filter (FOIF) [20].

The final analysis field of 4D variational method is similar to that of extended Kalman filter. The covariance matrix is implicit in the 4D variational computation process, and the implicit covariance matrix in the final state is inaccessible. However, Kalman method can explicitly generate the error covariance matrix in model space.

2.7 SUMMARY OF THE DEVELOPMENT PROCESS

Data assimilation methods developed from simple interpolation methods, such as nudging, to advanced assimilation methods with mathematical and physical theories. 4D variational and Kalman methods are popular topics in international research on data assimilation methods. For most countries and regions with forecast centers, the computation involved in the two methods is overwhelming because of the large amount of calculation. OI method is widely utilized in most operational forecast systems. The operational 4D variational system was applied in the European Centre for Medium-Range Weather Forecasts in 1997 and in France in 2000. However, Kalman method is rarely employed in operational forecasting.

3 Introduction of 4D variational method

4D variational method minimizes the objective function. The deviation between model results and observation data is minimized via adjustment of the control variable. In 4D variational method, observations at different times and locations can be employed in the same manner to obtain an accurate estimate of the initial condition. Satellite and radar observations are difficult to use in OI but easy to use in 4D variational method. 4D method is an extension of 3D method. Thus, the processing of observations, background error, and optimal algorithm are similar in the two methods. Nonlinear, tangent, and adjoint models are introduced in the assimilation method because the observations and the model field at different times are considered in the 4D method. In mathematics, the objective function of 4D method can be expressed as:

$$J(X_0) = \frac{1}{2}[X_0 - X_0^b]^T B_0^{-1}[X_0 - X_0^b] +$$

$$\frac{1}{2}\sum_{i=0}^{p}\left[H(X_i) - y_i^o\right]^T R^{-1}\left[H(X_i) - y_i^o\right], \tag{1}$$

where X_0 is the initial state of the forecast model, X_0^b is the background field at initial time, y_i^o is the i observation, H is the observation operator, and X_i is the model result with the same time as y_i^o and originates from initial state X_0 through the nonlinear model, $x_i = M_{t_0,t_i}(x_0)$. The formula presents the process of the model from time t_0 to t_i.

This objective function is related to initial state X_0, and the model results during the assimilation window originate from initial state X_0. Thus, the objective function is composed of two factors, namely, J_b and J_o. J_b is the deviation between the background and analysis fields in the initial state. J_o is the deviation between the observations and analysis field during the assimilation of the window.

4D method requires an iterative solution of nonlinear, tangent, and adjoint model. Hence, the computation is extensive and related to the resolution of X_0. Ideally, the resolution of X_0 is similar to the model's. The cost of minimizing the objective function in high resolution is tremendous. Courtier's [21] research indicates that minimizing the incremental analysis of X_0 instead of X_0 itself significantly reduces the computational cost. The objective function is transformed into an incremental form, and the incremental analysis is minimized in low resolution. The low resolution incremental analysis result is transformed back to high resolution, and the final analysis result comprises the high resolution incremental analysis and initial state X_0. Thus, the computational cost is reduced with a low resolution.

Furthermore, the preconditioning process is necessary because it enhances the iterative efficiency. Preconditioning transforms the coordinator of initial state X_0 [22], decreases the Hessian matrix (the second derivative of the objective function) condition number, and accelerates the convergence rate of the iterative algorithm. Andrew 2011 [23] utilized preconditioning technology to reduce the computation time.

3.1 INCREMENTAL FORM OF 4D METHOD

Standard variational methods, such as steepest descent and conjugate gradient methods, result in large amounts of computation. Incremental method generates a Taylor series expansion at approximate solutions and approaches the more accurate approximation solution iteratively.

In the beginning of incremental method, the background field is the iterative initial value, $x_0^0 - x_b$. For the n-th iterator, the analysis value is $x_0^n = x_0^{n-1} + \delta x_0^n$. Thus, the incremental objective function is:

$$J\left(\delta x_0^n\right) + \frac{1}{2}\left[\delta x_0^n + \left(x_0^{n-1} - x_b\right)\right]^T B^{-1}\left[\delta x_0^n + \left(x_0^{n-1} - x_b\right)\right] +$$

$$\frac{1}{2}\sum_{i=0}^{p}\left\{H_i\left[M_{t_0,t_i}\left(x_0^{n-1} - \delta x_0^n\right)\right] - y_i^o\right\}^T \tag{2}$$

$$R^{-1}\left\{H_i\left[M_{t_0,t_i}\left(x_0^{n-1} - \delta x_0^n\right)\right] - y_i^o\right\}$$

For the nonlinear model M_{t_0,t_i}, the Taylor series at x_0^{n-1} that ignores the second-order term is:

$$M_{t_0,t_i}\left(x_0^{n-1} - \delta x_0^n\right) = M_{t_0,t_i}\left(x_0^{n-1}\right) + L_{t_0,t_i}\left(\delta x_0^n\right).$$

Based on the hypothesis of the linear observation operator, H_i is:

$$H_i\left[M_{t_0,t_i}\left(x_0^{n-1} - \delta x_0^n\right)\right] \approx H_i\left[M_{t_0,t_i}\left(x_0^{n-1}\right)\right] + H_i L_{t_0,t_i}\left(\delta x_0^n\right). \tag{3}$$

Thus, the final incremental form is:

$$J\left(\delta x_0^n\right) + \frac{1}{2}\left[\delta x_0^n + \left(x_0^{n-1} - x_b\right)\right]^T B^{-1}\left[\delta x_0^n + \left(x_0^{n-1} - x_b\right)\right] +$$

$$\frac{1}{2}\sum_{i=0}^{p}\left\{H_i\left[L_{t_0,t_i}\left(\delta x_0^n\right)\right] - d_i^{n-1}\right\}^T R^{-1}\left\{H_i\left[L_{t_0,t_i}\left(\delta x_0^n\right)\right] - d_i^{n-1}\right\}, \tag{4}$$

d_i^{n-1} is the deviation between the model result and observation y_i^o, is expressed as:

$$d_i^{n-1} H_i\left[M_{t_0,t_i}\left(x_0^{n-1}\right)\right] - y_i^0 = H_i\left[x_i^{n-1}\right] - y_i^0, \tag{5}$$

L_{t_0,t_i} is the tangent model to calculate disturbance quantity δx_i^n at time t_i.

The partial derivative of objective function is:

$$\frac{\partial J}{\partial \delta x_0^n} = B^{-1}[\delta x_0^n - (\delta x_0^{n-1} - x_b)] +$$

$$\sum_{i=0}^{p} L_{t_i,t_0}{}^T H_i^T R^{-1}\left(H_i\left[L_{t_0,t_i}\left(\delta x_0^n\right)\right] - d_i^{n-1}\right), \tag{6}$$

where $L_{t_i,t_0}{}^T$ is the adjoint model for resolving the initial disturbance from the final state reversely.

3.2 PRECONDITIONING

In the iteration process, preconditioning technology was adopted to accelerate the convergence. The inverse matrix of background error covariance matrix B^{-1} was utilized to determine the pre-conditioner as follows [24, 25]:

$$\delta x^n = U v^n. \tag{7}$$

Transformational matrix U is the Cholesky decomposition of background error covariance matrix B.

Thus, $B = UU^T$. The new variable v^n is introduced as a control variable, so the new objective function is:

$$J(v^n) + \frac{1}{2}\left[\sum_{i=1}^{n} v^i\right]^T \sum_{i=1}^{n} v^i +$$
$$\frac{1}{2}\sum_{i=0}^{p}\left\{H_i L_{t_0,t_i} U v^n + d_i^{n-1}\right\}^T R^{-1}\left\{H_i L_{t_0,t_i} U v^n + d_i^{n-1}\right\}. \quad (8)$$

After preconditioning, the condition number of the Hessian matrix is reduced and thus accelerates the convergence of the algorithm.

4 Kalman filter

The process of the Kalman filter is composed of forecast and analysis stages.

4.1 EXTENDED KALMAN FILTER

The extended Kalman filter can be expressed as follows:
1) Forecast stage:

$$x_i^b = M x_{i-1}^a, \quad (9)$$

$$P_i^b = L_{i-1} P_{i-1}^a L_{i-1}^T + Q. \quad (10)$$

2) Analysis stage:

$$K_i = P_i^b H^T [H P_i^b H^T + R]^{-1}, \quad (11)$$

$$x_i^a = x_i^b + K_i\left(y_i^o - H x_i^f\right), \quad (12)$$

$$P_i^a = [I - K_i H] P_i^b = \left[\left(P_i^b\right)^{-1} + H^T R^{-1} H\right]^{-1}, \quad (13)$$

M is the linear model, P is the error covariance matrix, Q is dynamic noise, K is the gain matrix, y is observation, and H is the observation operator. In the forecast stage, the forecast result and error covariance matrix were obtained. In the analysis stage, the gain matrix was obtained from the error covariance matrix. The analysis result and error covariance matrix are thus renewed.

Resolution of the error covariance matrix is computationally intensive in the extended Kalman filter. The dimension of P and L matrix is $n \times n$. In a typical operational forecast system, the number of is 10^7. Thus, the ensemble Kalman filter is utilized to reduce the computation.

4.2 EnKF

Extended Kalman filter is inapplicable because the operational ocean forecast system is a high-dimension nonlinear model, and the computation of the background error covariance matrix is extensive. EnKF is proposed to estimate the background error covariance matrix with an ensemble scheme. Monte Carlo short-term ensemble forecast is utilized to estimate the background error

covariance matrix. At the initial moment, K initial model fields are generated from random disturbance. The optimal estimation model predictions are averaged by corresponding forecasts and expressed by the equation:

$$P_i^b \approx \frac{1}{K-1}\sum_{k=1}^{K}\left(x_k^f - \overline{x^f}\right)\left(x_k^f - \overline{x^f}\right)^T = \frac{1}{K-1}X^b X^{bT}. \quad (14)$$

Since Evensen [14] introduced EnKF into the ocean model in 1994, an increasing number of models adopted this method [26]. Miyazawa et al. [27] adopted the local ensemble transformation Kalman filter algorithm based on 20 members' ensemble simulations of the parallelized Princeton Ocean Model (Stony Brook Parallel Ocean Model) with a horizontal resolution of 1/36°. Deng [28] assumed that the statistical properties of the background errors do not change significantly at neighbouring analysis steps within a short time window and thus allow the ensembles generated in the previous steps to be used in the current steps. As such, a joint ensemble matrix that combines the ensembles of previous and present steps can be constructed to form a larger ensemble for estimating the background error covariance.

5 Frontier of ocean data assimilation

Current data assimilation research mainly focuses on two aspects: data assimilation with several constraints and reducing the computational complexity of assimilation.

5.1 DATA ASSIMILATION WITH CONSTRAINTS

In many cases, the value range of the control variable is limited by certain constraints. For example, salinity is distributed in a certain range and sea surface temperature is higher than the freezing point. These constraints contain useful information to improve the calculation precision. However, linear estimation methods (such as Kalman filter) cannot take advantage of this information on constraints. Recently, several methods were proposed to introduce constraints into data assimilation.

1) Adjustment operator method. Multivariate satellite observations [29] are assimilated into an isopycnic coordinate ocean model (Miami Isopycnic Coordinate Ocean Model). If a cold core ring of the Gulf Stream is absent from a model forecast and has to be introduced by the analysis, several layers must be corrected to outcrop at the bottom of the mixed layer. In this case, linear analysis would certainly introduce a number of negative layer thickness values that would need to be reset to zero. Simon [30] and Thacker [31] adopted similar methods.

2) Introduction of non-second-order terms to the objective function. The variational methods described above are based on the hypothesis of a quadratic objective function. The new algorithm should be proposed to minimize the non-quadratic objective function. Fujii [32] adopted two types of constraints in 3D oceanic variational analysis for the equatorial Pacific. One is the constraint for the variational quality control procedure, and the other is

employed to avoid density and temperature inversions.

3) Non-linear transformation for constrained variables. Bertino [33] reported that several positive variables are incorrectly described by the Gaussian model and can be addressed through the assimilation of their log transform. However, this approach may result in an asymmetric probability distribution.

4) Non-Gaussian error probability distribution function. Generally, assimilation methods with an error distribution assume that error probability distributions are Gaussian. The non-Gaussian distribution approach adopts an assimilation method with certain implicit constraints by modifying the probability distribution function. Lauvernet [34] showed that an optimal filter dealing with inequality constraints can be formulated under the assumption that the probability distributions are truncated Gaussian distributions. The statistical tools required to implement this truncated Gaussian filter were described. This method was then applied to a 3D hybrid coordinate ocean model (HYCOM) of the Bay of Biscay (at 1/15° resolution). The results revealed that the algorithm can deal with the hydrostatic stability condition in isopycnic and z coordinates.

A comparison of these four methods with constraints indicates that adjustment operator method lacks theoretical support because the process does not introduce the constraints and adjusts the model result forcibly to satisfy the constraints. A non-second-order objective function introduces the constraints via non-second-order terms, but the minimizing algorithm should be modified correspondingly. Non-linear transformation method may result in an asymmetric probability distribution. The non-Gaussian method has sufficient theoretical basis and allows for assimilation with certain implicit constraints by modifying the probability distribution function.

5.2 DIMENSIONALITY REDUCTION METHODS

To reduce the computation cost of assimilation, the objective function can be minimized in a subspace. By reducing the dimensions, the required memory and CPU can be reduced considerably. Generally, the control variables are projected onto a set of feature vectors, and most of the energy in the original space is reversed. Cai [35] employed bred vectors as feature vectors and found that minimizing the projection of the bred vectors on the observation-minus-analysis field may be a beneficial factor to achieving an operational forecast system.

Many dimensionality reduction methods for the Kalman filter have been developed. Cane [36] presented an approach to the Kalman filter that employs reduced state space representation for the required error covariance matrices. Kaplan [37-39] and Canizares [40] conducted similar dimensionality reduction research for meteorological and oceanographic historical datasets.

Blayo [41] and Durbiano [42] employed a low-dimensional space based on the first few EOFs or empirical orthogonal functions, which can be computed from a sampling of the model trajectory. Hoteit [43] and Robert [44] suggested reduced-order reduction method to improve the convergence rate of optimization by projecting the control vector onto a limited number and reducing the size of the control vector.

6 Summary

The development of data assimilation techniques was outlined in this paper. Two assimilation schemes, namely, 4D variational and Kalman filter, were introduced. In addition, several frontier ocean data assimilation methods, such as assimilation with a constraint condition and dimensionality reduction, were discussed.

Acknowledgements

References

[1] Gilchrist B, Cressman G P 1954 An experiment in objective analysis *Tellus* **6**(4) 309-18

[2] Guan Y H, Zhou G Q, Lu W S, Chen J P 2007 Theory Development and Application of Data Assimilation Methods *Meteorology and Disaster Reduction Research* **30**(4) 1-8 *(in Chinese)*

[3] White W B, Tai C K, Holland W R 1990 Continuous Assimilation of Geosat Altimetric Sea Level Observations into a numerical Synoptic Ocean Model of the California Current *Journal of Geophysical Research* **95**(C3) 3127-48

[4] Mellor G L, Ezer T 1991 A Gulf Stream Model and an Altimetry Assimilation Scheme *Journal of Geophysical Research* **96** 8779-95

[5] Ezer T, Mellor GL 1994 Continuous assimilation of Geosat altimeter data into a three-dimensional primitive equation Gulf Stream model *Journal of Physical Oceanography* **24** 832-47

[6] Thacker W C, Lee S K, Halliwell G R 2004 Assimilating 20 years of Atlantic XBT data into HYCOM: a first look *Ocean Modelling* **7** 183-210

[7] Oke P R, Brassington G B, Griffin D A, Schiller A 2008 The Bluelink ocean data assimilation system (BODAS) *Ocean Modelling* **21** 46-70

[8] Fu W W, Jiang Z, Yan C X 2009 A comparison between 3DVAR and EnOI techniques for satellite altimetry data assimilation *Ocean Modelling* **26** 206-16

[9] Dobricic S, Pinardi N 2008 An oceanographic three-dimensional variational data assimilation scheme *Ocean Modelling* **22** 89-105

[10] Cohn S E, Silva A D, Guo J, Sienkiewicz M, Lamich D 1998 Assessing the effects of data selection with the DAO Physical-space Statistical Analysis System *Monthly Weather Review* **126**(11) 2913-9

[11] Wang X D, Xu D F, Xu X H 2007 The different variational solutions review in variation data assimilation *Journal of Marine Science* **25**(3) 103-11 *(in Chinese)*

[12] Powell B S, Arango H G, Moore A M, Di Lorenzo E, Miliff R F, Foley D 2008 4DVAR data assimilation in the Intra-Americas Sea with the Regional Ocean Modeling System (ROMS) *Ocean Modelling* **25**(3-4) 173-88

[13] Smith S R, Ngodock H E 2008 Cycling the Representer Method for 4D-variational data assimilation with the Navy Coastal Ocean Model *Ocean Modelling* **24**(3-4) 92-107

[14] Evensen G 1994 Sequential data assimilation with nonlinear quasi-geostrophic model suing Monte Carlo methods to forecast error statistics *Journal of Geophysical Research* **99**(C5) 10143-62

[15] Carmillet V, Brankart J M, Brasseur P, Drange H, Evensen G, Verron J 2001 A singular evolutive extended Kalman filter to assimilate ocean color data in a coupled physical-biochemical model of the North Atlantic ocean *Ocean Modelling* **3**(3-4) 167-92

[16] Barth A, Alvera-Azcárate A, Beckers J M, Rixen M 2006 Coupling a two-way nested primitive equation model and a statistical SST predictor of the Ligurian Sea via data assimilation *Ocean Modelling* **13**(3-4) 255-70

[17] Pham D T, Verron J, Roubaud M C 1997 A singular evolutive Kalman filter for data assimilation in oceanography *Journal of Marine System* **16**(3-4) 323-40

[18] Lermusiaux P F J, Robinson A R 1999 Data assimilation via error subspace statistical estimation. Part I. Theory and schemes *Monthly Weather Review* **127**(7) 1385-407

[19] Pham D T, Verron J, Gourdeau L 1998 Singular evolutive Kalman filter for data assimilation in oceanography *Canadian Research Academic Science Paris* **326** 103-14

[20] Chin T M, Haza A C, Mariano A J 2002 A reduced-order information filter for multilayer shallow-water models: profiling and assimilation of sea surface height *Journal of atmospheric and oceanic technology* **19** 517-33

[21] Courtier P, Derber J, Errico R, Louis J F, Vukićević T 1993 Important literature on the use of adjoint, variational methods and Kalman filter in meteorology *Tellus A* **45**(5) 342-57

[22] Davidon W C 1959 Variable metric for minimization *Report ANL 5990 Argonne National Laboratory Argonne Illinois*

[23] Moore A M, Arango H G, Broquet G, Powell B S, Weaver A T, Zavala-Garay J 2011 The Regional Ocean Modeling System (ROMS) 4-dimensional variational data assimilation systems: Part I – System overview and formulation *Progress in Oceanography* **91**(1) 34-49

[24] Lorenc A C 1992 Iterative analysis using covariance functions and filters *Quarterly Journal of the Royal Meteorological Society* **118**(505) 569-91

[25] Lorenc A C 1997 Development of an operational variational assimilation scheme *Journal of the Meteorological Society of Japan* **75** 339-46

[26] Evensen G 2003 The ensemble Kalman filter: theoretical formulation and practical implementation *Ocean Dynamics* **53**(4) 343-67

[27] Miyazawa Y, Miyama T, Varlamov S M, Guo X, Waseda T 2012 Open and coastal seas interactions south of Japan represented by an ensemble Kalman filter *Ocean Dynamics* **62**(4) 645-59

[28] Deng Z, Tang Y, Chen D, Wang G 2012 A Time-Averaged Covariance Method in the EnKF for Argo Data Assimilation *Atmosphere Ocean* **50**(1) 129-45

[29] Brankart J M, Testut C E, Brasseur P, Verron J 2003 Implementation of a multivariate data assimilation scheme for isopycnic coordinate ocean models: application to a 1993–1996 hindcast of the North Atlantic Ocean circulation *Journal of Geophysical Research* **108**(C3) 1-20

[30] Simon D, Simon D L 2005 Aircraft turbofan engine health estimation using constrained Kalman filtering *ASME Journal of Engineering for Gas Turbines and Power* **127**(2) 323-8

[31] Thacker W C 1998 Data assimilation with inequality constraints *Ocean Modelling* **16**(3-4) 264-76

[32] Fujii Y, Ishizaki S, Kamachi M 2005 Application of nonlinear constraints in a three-dimensional variational ocean analysis *Journal of Oceanography* **61**(4) 655-62

[33] Bertino L, Evensen G, Wackernagel H 2002 Combining geostatistics and Kalman filtering for data assimilation in an estuarine system *Inverse Problems* **18**(1) 1-23

[34] Lauvernet C, Brankart J M, Castruccio F, Broquet G, Brasseur P, Veron J 2009 A truncated Gaussian filter for data assimilation with inequality constraints: Application to the hydrostatic stability condition in ocean models *Ocean Modelling* **27**(1-2) 1-17

[35] Cai M, Kalnay E, Toth Z 2003 Bred Vectors of the Zebiak-Cane Model and Their Application to ENSO Predictions *Journal of Climate* **16** 40-56

[36] Cane M A, Kaplan A, Miller R N, Tang B, Hackert E C, Busalacchi A J 1996 Mapping tropical Pacific sea level: Data assimilation via a reduced state space Kalman filter *Journal of Geophysical Research* **101**(C10) 22599-617

[37] Kaplan A, Kushnir Y, Cane M A, Blumenthal B 1997 Reduced space optimal analysis for historical data sets: 136 years of Atlantic sea surface temperatures *Journal of Geophysical Research Oceans* **102**(C13) 27835-60

[38] Kaplan A, Cane M A, Kushnir Y 2003 Reduced space approach to the optimal analysis interpolation of historical marine observations: Accomplishments, difficulties, and prospects *Advances in the Applications of Marine Climatology: The Dynamic Part of the WMO Guide to the Applications of Marine Climatology WMO/TD1081* World Meteorological Organization Geneva Switzerland 199-216

[39] Kaplan A, Kushnir Y, Cane M A 2000 Reduced space optimal interpolation of historical marine sea level pressure: 1854-1992 *Journal of Climate* **13**(16) 2987-3002

[40] Canizares R, Kaplan A, Cane M A, Chen D, Zebiak S E 2001 Use of data assimilation via linear low-order models for initialization of El Nino Southern Oscillation predictions *Journal of Geophysical Research Oceans* **106**(C12) 30947-59

[41] Blayo E, Blum J, Verro J 1998 Assimilation Variationnelle de Donnees en Ocenaographie et reduction de la dimension de l'espace de controle *In Equations aux Derivees partielles et Applications* 199-218

[42] Durbiano S 2001 Vecteurs caracteristiques de modeles oceaniques pour la reduction d'ordre en assimilation de donnees *PhD thesis Universite Joseph Fourier Laboratorie de Modelisation et calcul Grenoble France*

[43] Hoteit I, Kohl A, Stammer D 2006 Efficiency of reduced-order, time-dependent adjoint data assimilation approaches *Ocean Modeling* **62**(4) 539-50

[44] Robert C, Durbiano S, Blayo E, Verron J, Blum H, Dimet F X 2005 A reduced-order strategy for 4D-Var data assimilation *Journal of Marine Systems* **57**(1-2) 70-82

Authors
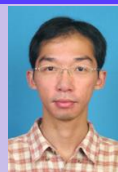
Zhenchang Zhang, born in March, 1980, Fuzhou, P.R. China

Current position, grades: Lecturer of Computer and Information Science College, Fujian Agriculture and Forestry University.
University studies: PhD at Xiamen University in China.
Scientific interests: ocean data assimilation.
Publications: 3 papers.
Experience: more than 3 projects including the National Natural Science Foundation.

Changying Wang, born in January, 1963, Fuzhou, P.R. China

Current position, grades: Department Chair & Associate Professor, Department of Computer Science and Technology, Computer and Information Science College, Fujian Agriculture and Forestry University.
University studies: PhD from Northwest University in China.
Scientific interest: pattern recognition, machine learning.
Publications: more than 10 papers.
Experience: more than 5 projects including the National Natural Science Foundation.

Simulations of standard Brownian motion

Anatolii Pashko[*]

Taras Shevchenko National University of Kyiv, Prospect Hlushkov, 4D, 03187 Kyiv, Ukraine

Abstract

This paper investigates algorithms for simulation of the trajectories of a Brownian motion (Wiener process) with given accuracy and reliability. Spectral representation of Wiener process as random series examines as a model. Estimates of the accuracy and reliability investigated in various function spaces - spaces of measurable integrated functions, Orlicz spaces and spaces of continuous functions. Given the accuracy of the numbers and simulation algorithms error of Gaussian random variables in the model are used strictly sub-Gaussian random variables. Examples of simulation are represented below.

Keywords: Wiener process, simulation, sub-Gaussian model, accuracy and reliability

1 Introduction

Statistical models of Wiener processes used to solve many practical problems, particularly, in the calculation of integrals over Wiener processes or in the numerical solution of stochastic differential equations [1, 2].

As a model, we consider the spectral decomposition processes as a random series or integrals.

Moment of difference of model and process is an estimation of model accuracy in the problems of statistical simulation [1, 2]. In papers [3, 4] investigated accuracy and reliability estimation of the simulation of random processes in various functional spaces.

In this paper, we investigate the accuracy and reliability of the Wiener process simulation in various functional spaces - L_2, L_p, C, Orlicz spaces.

2 Basic definitions

Let (T, Ξ, μ) - be some measurable space and $\mu(T) = 1$.

Definition 1. Generalized Wiener process with parameter $\alpha, \alpha \in (0,1]$ will be called a Gaussian random process with zero mean and correlation function

$$R_\alpha(t,s) = EW_\alpha(t)W_\alpha(s) = \frac{1}{2}\left(|t|^{2\alpha} + |s|^{2\alpha} - |t-s|^{2\alpha}\right).$$

When $\alpha = \frac{1}{2}$ we have a classical Wiener process. This paper considers the classical Wiener process. As a model of the Wiener process we consider random series

$$S_M(t) = \sum_{i=1}^{M} f_i(t)\eta_i,$$

where $\{\eta_i\}$ - Gaussian random variables from $N(0,1)$ (in the general case is optionally dependent).

Let all $S_M(t)$ and $W_\alpha(t)$ belong to some function space $A(T)$.

Definition 2. Model $S_M(t)$ approximates the process $W_\alpha(t)$ with specified accuracy $\delta > 0$ and reliability $0 < \varepsilon < 1$ in the norm of a function space $A(T)$, in case

$$P\left\{\|W_\alpha(t) - S_M(t)\|_A > \delta\right\} \leq 1 - \varepsilon.$$

When simulating a sequence of Gaussian random variables due to the accuracy of computational tools, simulation algorithms obtain sub-Gaussian random variables.

Definition 3. Random variable ζ called sub-Gaussian, if for $\forall \lambda$ $\exists a$ takes place

$$E \exp(\lambda\zeta) \leq \exp\left(\frac{a^2\lambda^2}{2}\right).$$

Space of sub-Gaussian random variables is a Banach space with norm

$$\tau(\zeta) = \inf\left\{a \geq 0 : E\exp\{\lambda\zeta\} \leq \exp\left\{\frac{a^2\lambda^2}{2}\right\}, \lambda \in R^1\right\}.$$

When $E\zeta^2 = a^2$ we have a strictly sub-Gaussian random variables.

Definition 4. C is a continuous, steam, convex function $U(x)$, such as $U(0) = 0$, $U(x) > 0$ when $x \neq 0$.

Definition 5. Orlicz space generated by the C- function $U(x)$ is a family of functions $\{f(t), \ t \in T\}$ such as, for every $f(t)$ exists a constant r such as

$$\int_T U\left(\frac{f(t)}{r}\right)d\mu(t) < \infty.$$

Orlicz space is a Banach space under the norm

---

[*]**Corresponding author** e-mail: pashkoua@mail.ru

$$\|f\|_{L_U} = \inf\left\{ r>0 : \int_T U\left(\frac{f(t)}{r}\right) d\mu(t) \le 1 \right\}.$$

Will consider C-function $U(x)$ for which function $G_U(t) = \exp\left\{\left(U^{(-1)}(t-1)\right)^2\right\}$ is convex and when $t \ge 1$, where $U^{(-1)}(x)$ is inverse function of $U(x)$.

This condition is fulfilled for C-function

$$U(x) = \exp\left\{|x|^a\right\} - 1, \quad 1 \le a \le 2,$$

then $G_U(t) = \exp\left\{(\ln t)^{\frac{2}{a}}\right\}$.

Theorem [5]. For any real $T \ne 0$ and any real t_0 each of random functions $\sqrt{T}W\left(\frac{t}{T}\right)$, $W(t+t_0) - W(t_0)$ and $|t|W\left(\frac{1}{t}\right)$ is similar to random function $W(t)$.

Thus, without loss of generality, we can consider the Wiener process on the interval $T = [0,1]$ and consider different views of the Wiener process in the form of series.

Decomposition of the Wiener process in the Eigen functions of the correlation operator of the Brownian bridge $t \in [0,1]$ has the form [6]

$$\xi_1(t) = t\eta_0 + \sqrt{2}\sum_{i=1}^{\infty} \frac{\sin(i\pi t)}{i\pi}\eta_i,$$

where $\{\eta_0, \eta_1, \eta_2, ...\}$ - independent standard Gaussian random variables.

$\lambda_i = i\pi$ - the eigenvalues of the correlation operator. Fourier series expansion on $t \in [0,1]$ [5]

$$\xi_2(t) = t\eta_0 + \sqrt{2}\sum_{i=1}^{\infty}\left(\eta_{1i}\frac{\sin(2\pi it)}{2\pi i} + \eta_{2i}\frac{1-\cos(2\pi it)}{2\pi i}\right),$$

where $\{\eta_{1i}, \eta_{2i}\}$ - independent standard Gaussian random variables.

3 Estimation of Karhunen – Loeve model

As a model for the expansion of the Wiener process in the eigenfunctions of the correlation operator consider

$$S_1(t,M) = t\eta_0 + \sqrt{2}\sum_{i=1}^{M} \frac{\sin(i\pi t)}{i\pi}\eta_i.$$

Based on the results [3, 4] it is easy to obtain the following assertion.

Assertion 1. Model $S_1(t,M)$ approximates the process $\xi_1(t)$ with accuracy $\delta > 0$ and reliability $1-\varepsilon$, $0<\varepsilon<1$:

a) in $L_2([0,1])$ if inequalities hold $\delta^2 > J1_{M+1}$

$$\exp\left\{\frac{1}{2}\right\}\frac{\delta}{\sqrt{J1_{M+1}}}\exp\left\{-\frac{\delta^2}{2J1_{M+1}}\right\} \le \varepsilon$$

or $\delta^2 > J1_{M+1}$.

$$\left(\frac{\delta^2 - J1_{M+1}}{J2_{M+1}}+1\right)^{\frac{1}{2}}\exp\left\{-\frac{\delta^2 - J1_{M+1}}{2J2_{M+1}}\right\} \le \varepsilon,$$

where

$$J1_{M+1} = \sum_{i=M+1}^{\infty}\lambda_i^{-2} \text{ and } J2_{M+1} = \left(\sum_{i=M+1}^{\infty}\lambda_i^{-4}\right)^{\frac{1}{2}}.$$

b) in $L_p([0,1])$, $p>1$, $p \ne 2$ or inequalities hold

$$\exp\left\{\frac{1}{2}\right\}\frac{\delta}{\sigma_{M+1}}\exp\left\{-\frac{\delta^2}{2\sigma_{M+1}^2}\right\} \le \varepsilon,$$

where $\sigma_{M+1}^2 = \sup_{t\in[0,1]}\left(\sum_{i=M+1}^{\infty}\frac{\sin^2(i\pi t)}{(i\pi)^2}\right)$

and inequalities $\delta^2 > \sigma_{M+1}^2$ when $1 \le p < 2$ or $\delta^2 > (p+1)\sigma_{M+1}^2$ when $p > 2$.

c) in $L_U([0,1])$ if inequalities hold

$$\exp\left\{\frac{1}{2}\right\}\frac{\delta U^{(-1)}(1)}{\sigma_{M+1}}\exp\left\{-\frac{\delta^2\left(U^{(-1)}(1)\right)^2}{2\sigma_{M+1}^2}\right\} \le \varepsilon,$$

$$\delta^2 > \left(2+\left(U^{(-1)}(1)\right)^{-2}\right)\sigma_{M+1}^2.$$

d) in $C([0,1])$ if for $\beta \in \left(0,\frac{1}{2}\right]$ inequalities hold.

$$2\exp\left\{-\frac{1}{2}\left(\frac{\delta}{G_{M+1}}\right)^2+1+\sqrt{2}\left(\frac{\delta}{G_{M+1}}\right)^{\frac{4\beta+1}{2\beta+1}}\left(F_\beta+\frac{\pi}{2}\right)\right\}\times\exp\left\{2\left(\frac{\delta}{G_{M+1}}\right)^{\frac{4\beta}{2\beta+1}}\left(F_\beta q_\beta\left(\frac{\delta}{G_{M+1}}\right)+\frac{\pi^2}{8}\left(\frac{\delta}{G_{M+1}}\right)^{\frac{1-2\beta}{1+2\beta}}\right)\right\}\le\varepsilon$$

and $\delta>2G_{M+1}$, where $G_{M+1}=\left(\sum_{i=M+1}^{\infty}\lambda_i^{-2}\right)^{\frac{1}{2}}$,

$$F_\beta=\sum_{i=M+1}^{\infty}\left|\ln\left(w_R^{(-1)}\left(\frac{1-\left(1-2G_{M+1}^2\delta^{-2}\right)^{\frac{1}{2}}}{\pi i}\right)\right)\right|^{\frac{1}{2}}\frac{\pi^2 i^2}{G_i^{2(1-\beta)}},\quad w_R(h)=\sup_{|u-v|\le h}\left(\int_0^1\left(R(u,x)-R(v,x)\right)^2 dx\right)^{\frac{1}{2}},$$

$$q_\beta=\begin{cases}1,&\beta\in\left[\frac{1}{6},\frac{1}{2}\right],\\\left(\frac{\delta}{G_{M+1}}\right)^{\frac{1-6\beta}{2(2\beta+1)}},&\beta\in\left(0,\frac{1}{6}\right).\end{cases}$$

Since $\sup_{t\in[0,1]}\left|\sin\left(i\pi t\right)\right|\le1$, then $\sigma_{M+1}^2=\sum_{i=M+1}^{\infty}\lambda_i^{-2}$.

For implementations simulation and given δ and ε let's find M. The Table 1 shows the estimates for M in the various functional spaces. Calculations for $L_U\left([0,1]\right)$

is not represented. Depending on the function $U(x)$ results are between $L_2\left([0,1]\right)$ and $C\left([0,1]\right)$.

Figure 1 shows the implementation of a Wiener process to represent $\xi_1(t)$.

TABLE 1 Values of M for different functional spaces

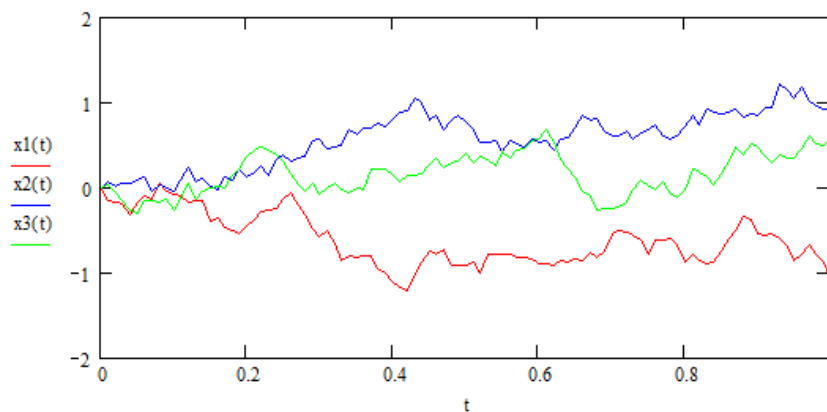| δ	ε	a) $L_2\left([0,1]\right)$	b) $L_2\left([0,1]\right)$	$C\left([0,1]\right)$
0.1	0.05	110	36	10000
0.05	0.05	750	120	100000
0.01	0.05	18700	2250	>1000000
0.1	0.01	260	45	
0.05	0.01	1050	130	
0.01	0.01	26000	2350	

FIGURE 1 Implementation of the Weiner process

4 Estimation of Fourier model

As a model for the expansion of the Wiener process in the Fourier series consider

$$S_2(t,M)=t\eta_0+\sqrt{2}\sum_{i=1}^{M}\left(\eta_{1i}\frac{\sin(2\pi it)}{2\pi i}+\eta_{2i}\frac{1-\cos(2\pi it)}{2\pi i}\right).$$

Based on the results [3, 4] it is easy to obtain the following assertion.

Assertion 2. Model $S_2(t,M)$ approximates process $\xi_2(t)$ with accuracy $\delta>0$ and reliability $1-\varepsilon$, $0<\varepsilon<1$:

a) in $L_2\left([0,1]\right)$ if inequalities hold $\delta^2>J1_{M+1}$

$$\exp\left\{\frac{1}{2}\right\}\frac{\delta}{\sqrt{J1_{M+1}}}\exp\left\{-\frac{\delta^2}{2J1_{M+1}}\right\}\leq\varepsilon$$

or $\delta^2 > J1_{M+1}$

$$\left(\frac{\delta^2-J1_{M+1}}{J2_{M+1}}+1\right)^{\frac{1}{2}}\exp\left\{-\frac{\delta^2-J1_{M+1}}{2J2_{M+1}}\right\}\leq\varepsilon$$

where

$$J1_{M+1}=2\sum_{i=M+1}^{\infty}(\pi i)^{-2}\quad\text{and}\quad J2_{M+1}=2\left(\sum_{i=M+1}^{\infty}(\pi i)^{-4}\right)^{\frac{1}{2}}.$$

b) in $L_p([0,1])$, $p>1$, $p\neq 2$ if inequalities hold

$$\exp\left\{\frac{1}{2}\right\}\frac{\delta}{\sigma_{M+1}}\exp\left\{-\frac{\delta^2}{2\sigma_{M+1}^2}\right\}\leq\varepsilon$$

where

$$\sigma_{M+1}^2=2\sup_{t\in[0,1]}\left(\sum_{i=M+1}^{\infty}\left(\left(\frac{\sin(2\pi it)}{2\pi i}\right)^2+\left(\frac{1-\cos(2\pi it)}{2\pi i}\right)^2\right)\right)$$

and inequalities $\delta^2 > \sigma_{M+1}^2$ when $1\leq p<2$

or $\delta^2 > (p+1)\sigma_{M+1}^2$ when $p>2$.

c) in $L_U([0,1])$ if inequalities hold

$$\exp\left\{\frac{1}{2}\right\}\frac{\delta U^{(-1)}(1)}{\sigma_{M+1}}\exp\left\{-\frac{\delta^2\left(U^{(-1)}(1)\right)^2}{2\sigma_{M+1}^2}\right\}\leq\varepsilon,$$

$$\delta^2 > \left(2+\left(U^{(-1)}(1)\right)^{-2}\right)\sigma_{M+1}^2.$$

d) in $C([0,1])$ in case of $\beta\in\left(0,\frac{1}{2}\right]$ inequalities hold

$$2\exp\left\{-\frac{1}{2}\left(\frac{\delta}{2G_{M+1}}\right)^2+1+\sqrt{2}\left(\frac{\delta}{2G_{M+1}}\right)^{\frac{4\beta+1}{2\beta+1}}\left(F_\beta+\frac{\pi}{2}\right)\right\}\times\exp\left\{2\left(\frac{\delta}{2G_{M+1}}\right)^{\frac{4\beta}{2\beta+1}}\left(F_\beta q_\beta\left(\frac{\delta}{2G_{M+1}}\right)+\frac{\pi^2}{8}\left(\frac{\delta}{2G_{M+1}}\right)^{\frac{1-2\beta}{1+2\beta}}\right)\right\}\leq\varepsilon$$

and $\delta > 4G_{M+1}$, where $G_{M+1}=\left(\sum_{i=M+1}^{\infty}(2\pi i)^{-2}\right)^{\frac{1}{2}}$

$$F_\beta=4\pi^2\sum_{i=M+1}^{\infty}\left|\ln\left(w_R^{(-1)}\left(\frac{1-\left(1-8G_{M+1}^2\delta^{-2}\right)^{\frac{1}{2}}}{\pi i}\right)\right)\right|^{\frac{1}{2}}\frac{i^2}{G_i^{2(1-\beta)}},$$

$$q_\beta=\begin{cases}1,&\beta\in\left[\frac{1}{6},\frac{1}{2}\right],\\\left(\frac{\delta}{2G_{M+1}}\right)^{\frac{1-6\beta}{2(2\beta+1)}},&\beta\in\left(0,\frac{1}{6}\right).\end{cases}$$

The Table 2 shows the estimates for the M in various functional spaces for presentation $\xi_2(t)$. Figure 2 shows the implementation of the Wiener process for submission $\xi_2(t)$.

TABLE 2 Values of M for different functional spaces

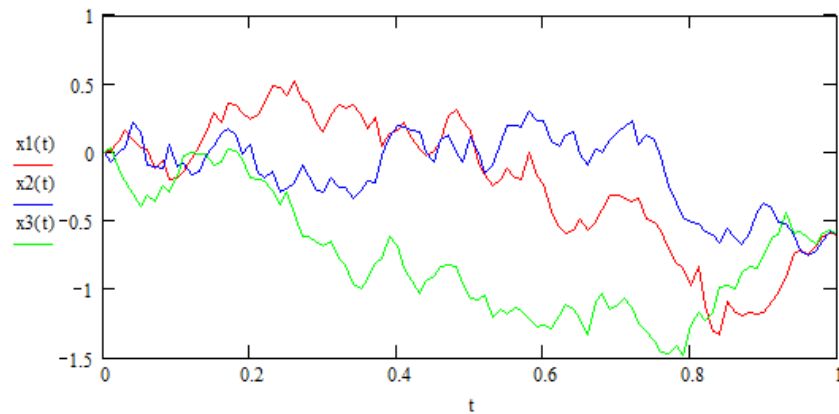| δ	ε	a) $L_2([0,1])$	b) $L_2([0,1])$	$C([0,1])$
0.1	0.05	380	60	300000
0.05	0.05	1500	220	1000000
0.01	0.05	38000	4400	>1000000
0.1	0.01	400	70	
0.05	0.01	2100	240	
0.01	0.01	58000	4500	

FIGURE 2 Implementation of Wiener process

When constructing a model we use strictly sub-Gaussian random variables obtained by the Equation $\eta_i = \sum_{j=1}^{12} \gamma_j - 6$, where $\{\gamma_j\}$ - uniformly distributed on $[0,1]$ random numbers. When building multiple implementations - the algorithm has a natural parallelization. Figure 3 shows the implementation of $\xi_1(t)$ and $\xi_2(t)$ retrieved from one sequence of strictly sub-Gaussian random variables. Figure 4 shown the implementation of presentation $\xi_1(t)$ at various M ($M=110$ and $M=36$).
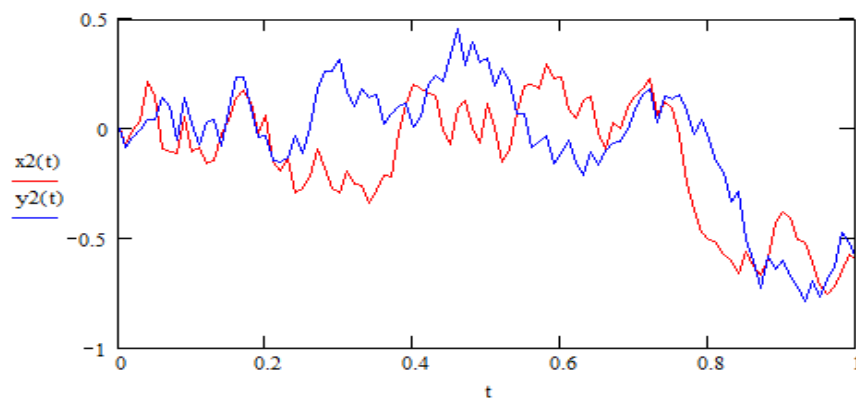
FIGURE 3 Implementation of the Wiener process for the views $\xi_1(t)$ and $\xi_2(t)$
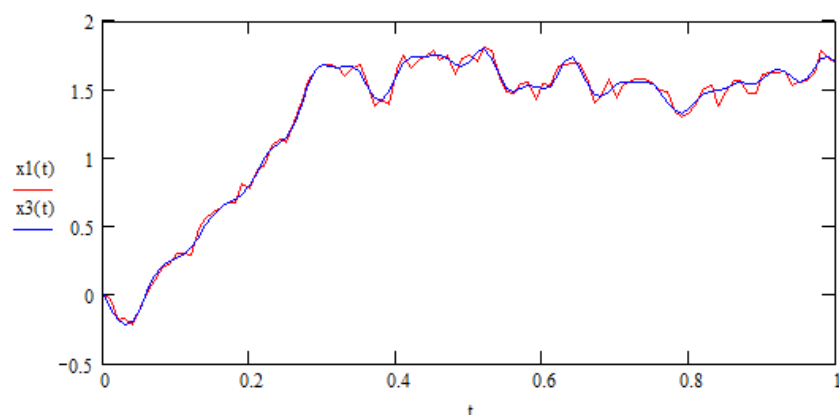
FIGURE 4 An example: Implementation of the Wiener process for the view $\xi_1(t)$ when *M=110* and *M=36*

5 Conclusions

We obtain estimates for the construction of strictly sub-Gaussian model of Wiener process. The model is constructed with the specified accuracy and reliability for various functional spaces. We found the implementation of the Wiener process for different views. Has interest for obtaining similar estimates for the generalized Wiener process.

References

[1] Prigarin S, Belov S 1998 An application of the Wiener process expansions in series *Preprin*t 1107

[2] Ermakov S, Mikhailov G 1982 Statistical simulations *Nauka Moscow* 280

[3] Kozachenko Y V, Pashko A A 1998 The accuracy of the simulation of random processes in the Orlicz spaces I *Probability Theory and Math. Statistics* 58 75-90

[4] Kozachenko Y.V., Pashko A A 1998 The accuracy of the simulation of random processes in the Orlicz spaces II *Probability Theory and Math. Statistics* 59 45-60

[5] Kahane J-P *1973* Some random series of functions *Mir Moscow* 302

[6] Gikhman I, Skorokhod A 1977 Introduction to the theory of stochastic processes *Nauka Moscow* 568

Author

Anatolii Pashko, born on January 28, 1962, Kuznetsovsk, Ukraine

Current position, grades: assistant professor.
University studies: Taras Shevchenko National University of Kyiv.
Scientific interest: stochastic processes theory, simulation of stochastic processes and fields.
Publications: 100 papers.

Order analysis method based on instantaneous phase

Rong Zeng[1*], Zhengfeng Jiang[1], He Ling[1], Wei Hu[1], Xing Wan[2]

[1]*School of Mechanical and Electronic Engineering, Wuhan University of Technology, Wuhan 430070, China*

[2]*Dongfeng Motor Group Co., LTD Technology Centre, Wuhan 430058, China*

Received 6 October 2014, www.cmnt.lv

Abstract

Order analysis is an effective method to analyse non-stationary signal of rotating machinery. The key of this method is to acquire the time sequence under even angle re-sampling. This paper proposed an order analysis method based on instantaneous phase using Hilbert-Huang Transform (HHT), and achieved the order spectrums of torsional vibration signals of rotating machinery from simulation and experiment. Being different from the order analysis method based on instantaneous frequency, this method directly uses the instantaneous phases obtained by HHT to get rotating angle over time. Thus, it is faster and more convenient. Meanwhile, this method is less affected by the 'boundary effect'. Hence, it can achieve higher precision. The simulating and experimental analysis both verified the feasibility and accuracy of this method.

Keywords: order analysis, even angle sampling, HHT, instantaneous phase

1 Introduction

Order analysis method is a highly effective approach to analyse the non-stationary vibration signals of rotating machinery [1]. To achieve this method, the key is to acquire the time sequence of order sampling according to the base speed or frequency of the rotating shaft. The traditional way is to install a phase pulse generator on the shaft system [2-3], which increases the complexity of the shafting and also generates signal synchronization problems. Moreover, sometimes it is not convenient to install the other sensors on the rotating shafts. As a result, scholars at home and abroad have put forward many order analysis methods without tachometers. Shao. H et al put forward an order tracking algorithm based on instantaneous frequency estimation using Gabor transform [4]; Guo Yu et al fulfilled an instantaneous frequency estimation order tracking method based on Short-time Fourier transform spectrum [5]; Xu Minqiang et al achieved an instantaneous frequency estimation method of rotating machinery vibration signals in high speed start process for Haar wavelet analysis [6]; Jia Jide et al realized an order analysis method based on instantaneous frequency estimation using Hilbert-Huang (HHT) transform [7]. Since HHT is not influenced by Heisenberg uncertainty principle, it is more effective to achieve can achieve multi-component signals analysis than the others. Therefore, it becomes a new approach in order analysis with its unique advantages.

The implementation steps of the order analysis method based on instantaneous frequency estimation using HHT are as follows [7]:

Firstly, complex signal was divided into a finite number of Intrinsic Mode Functions (IMFs) by Empirical Mode Decomposition (EMD). Secondly, for main components of the signal, the Hilbert Transform was used, obtaining the instantaneous frequency. Thirdly, the rotating speed was estimated by the instantaneous frequency, and then the rotating angle over time was decided by the speed. Finally, adopting even angle re-sampling on original signal, the angle domain signal was gotten, after conducting spectral analysis, the order spectrum of the original signal was acquired. The instantaneous frequency computed by HHT was influenced by 'boundary effect' and 'Gibbs effect', which introduced serious errors to the order analysis.

Aiming at the error issue brought by the instantaneous frequency acquired by HHT, this paper proposed an order analysis method based on instantaneous phase. Using the instantaneous phase information of the main component acquired by HHT directly, the corresponding relation between the rotating angle and time was plotted. Then according to the rotating angle, the time sequence of even angle re-sampling was obtained, and finally the order analysis was achieved. The simulating and experimental analysis both verify the feasibility of this method. It is simple and reliable, which provides a new approach to analyse the non-stationary signal of rotating machinery.

2 Theories of order analysis method based on instantaneous phase

2.1 EMPIRICAL MODE DECOMPOSITION

HHT is based on the study of instantaneous frequency of non-stationary signal. Usually, we define a signal's instantaneous frequency as the differential of its instantaneous phase. With regard to a multi-component

signal, there will be several instantaneous frequencies at a certain time. In order to make the definition of the instantaneous frequency owing a physical meaning, Norden E Huang proposed to decompose the signal into mono-component, which was IMF. Each IMF should meet the following two demands:

1) in the whole data set, the number of zero crossings and the number of extrema must either equal or differ at most by one;

2) at any point, the mean value of the envelopes decided by the local maxima and local minimum point is zero, namely, the signal is axisymmetric by time axis [8]. For signal $x(t)$, the EMD algorithm is as follows [9]:

Step 1 Determine all extrema of the signal (maximas and minimas). Then interpolate them and fit the points by cubic spline curve, obtaining the upper envelopes $e_u(t)$ and (lower envelopes) $e_l(t)$;

Step 2 Calculate the mean value of the upper and lower envelope, getting $m_1(t) = (e_u(t) + e_l(t))/2$, and then make the original signal sequence minus the mean value, acquiring $h_1(t) = x(t) - m(t)$;

Step 3 Repeat the above steps, until $h_k(t)$ meets the IMF requirements: $c_1(t) = h_k(t)$;

Step 4 Calculate the difference between $x(t)$ and $c_1(t)$, yeilding the residual $r_1(t) = x(t) - c_1(t)$, and iterate all above steps until the final residual $r_n(t)$ becomes a monotonic function.

After EMD, the original signal $x(t)$ is composed of n IMFs and the final residual component, as shown in Equation (1). The residual component is a mean trend or a constant:

$$x(t) = \sum_{i=1}^{n} c_i(t) + r_n(t) . \tag{1}$$

2.2 HILBERT TRANSFORM

The Hilbert Transform is applied to each IMF component obtained from EMD:

$$H(c_i(t)) = \frac{1}{\pi} P \int_{-\infty}^{\infty} \frac{c_i(t')}{t - t'} dt' , \tag{2}$$

where P is the Cauchy principal value of this integral. $c_i(t)$ and $H(c_i(t))$ form a complex conjugate pair, then an analytic signal $z_i(t)$ can be gotten:

$$z_i(t) = c_i(t) + iH(c_i(t)) = a_i(t)e^{i\varphi_i(t)} . \tag{3}$$

Then, the instantaneous amplitude is shown as:

$$a_i(t) = [c_i(t) + H(c_i(t))]^{1/2} . \tag{4}$$

The instantaneous phase can be obtained from Equation (5):

$$\varphi_i(t) = \arctan \frac{H(c_i(t))}{c_i(t)} . \tag{5}$$

The instantaneous frequency is shown as Equation (6):

$$f_i(t) = \frac{1}{2\pi} \frac{d\varphi_i(t)}{dt} . \tag{6}$$

2.3 INSTANTANEOUS PHASE PROCESSING

With regard to the rotating shafts of a rotating machinery, the phase of the base frequency component is correspond to the rotating angle of the shafts, and the phases of the multiple frequency components are integer multiples of rotating angle of the shafts. The instantaneous phases acquired by HHT always keeps $\varphi_i(t) \in (-\pi, \pi)$, which are not continuous. In order to obtain continuous phase information, which is the real rotating angle, the instantaneous phases should be processed.

The steps of the instantaneous phase processing are as follows:

1) Make the first point to be zero, which means that the first point of the curve should be processed to be the component's original phase.

2) Deal with the breakpoint. When the curve appears mutation from π to $-\pi$, connect each breakpoint and add 2π to the connection point.

3) After acquiring the continuous phase, use the quadratic polynomial to fit it.

Then, the instantaneous phase will be changed as follows:

$$\varphi_i'(t) = b_0 + b_1 t + b_2 t^2 . \tag{7}$$

2.4 EVEN ANGLE RE-SAMPLING

The phase after curve fitting reflects the relation between the rotating angle and time. In order to obtain the angle-domain signal, each time increment related to the even angles should be calculated. Assuming the even angle increment is $\Delta\theta$, the number of sampling points in angular domain (FFT points) is N, the corresponding time sequence is $\{T_j\}$, $j = 1, 2, \cdots N$. Therefore, the followed equation can be obtained:

$$\begin{cases} b_0 + b_1 T_1 + b_2 T_1^2 = 0 \\ b_0 + b_1 T_2 + b_2 T_2^2 = \Delta\theta \\ \dots \\ b_0 + b_1 T_N + b_2 T_N^2 = (N-1) \cdot \Delta\theta \end{cases} , \tag{8}$$

where $\Delta\theta$ is determined by the maximum ratio range order:

$$\Delta\theta = 2\pi / O_{max} , \tag{9}$$

where O_{max} is the maximum ratio range order. Thus,

Zeng Rong, Jiang Zhengfeng, Ling He, Hu Wei, Wan Xing

according to Equation (8), the time sequence known as $\{T_j\}$ can be calculated.

The time sequence of the original time-domain signal $x(t)$ in time domain is $\{t_m\}$, where $m = 1,2,...,m,...,M$ and M is the sampling points.

Appling the linear interpolation to the original signal $x(t)$ in the light of the time sequence $\{T_j\}$, the angle domain signal can be obtained, shown in Equation (10).

$$x(T_j) = x(t_m) + \frac{x(t_{m+1}) - x(t_m)}{t_{m+1} - t_m} \cdot (T_j - t_m), \qquad (10)$$

where $t_m \leq T_j \leq t_{m+1}$.

3 Simulations

This paper took the torsional vibration of the rotating machinery as an example, which usually was a multi-components signal with a base frequency corresponded to the rotating speed, described in Equation (11):

$$x(t) = \sum_{i=1}^{n} A_i \cos(2\pi i f_0 t + \varphi_i), \qquad (11)$$

where f_0 is the frequency corresponded to the rotating speed. A compound signal being composed of three chirp signals was established, shown in Equation (12). The signal in time-domain was plotted, illustrated in Figure 1. It was assumed that the signal didn't include noise.

$$\begin{aligned} S = S_1 + S_2 + S_3 = \\ 0.5\sin(2\pi f_0 t) + \sin(2\pi 2 f_0 t) + 1.5\sin(2\pi 4 f_0 t) \end{aligned} \qquad (12)$$
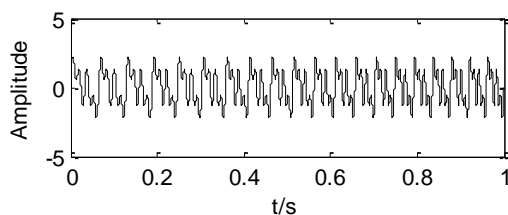
FIGURE 1 Simulating signal

If the rotating speed accelerated from 900r/min to 1200r/min within one second, the base frequency increased from 15Hz to 20Hz, we have:

$$f_0 = 15 + 5t. \qquad (13)$$

Applying EMD, the original signal was decomposed into 6 IMFs and one residual component, shown in Figure 2. IMF1, IMF2 and IMF3 were corresponding to the real signal. IMF1 was equal to the quadruple frequency component; IMF2 and IMF3 were corresponded to the double and base frequency components, respectively. However, IMF4, IMF5 and IMF6 were not equal to the real signal because of over decomposition, which was caused by the use of cubic spline curve to form envelopes and the small threshold value of the 'screening'.
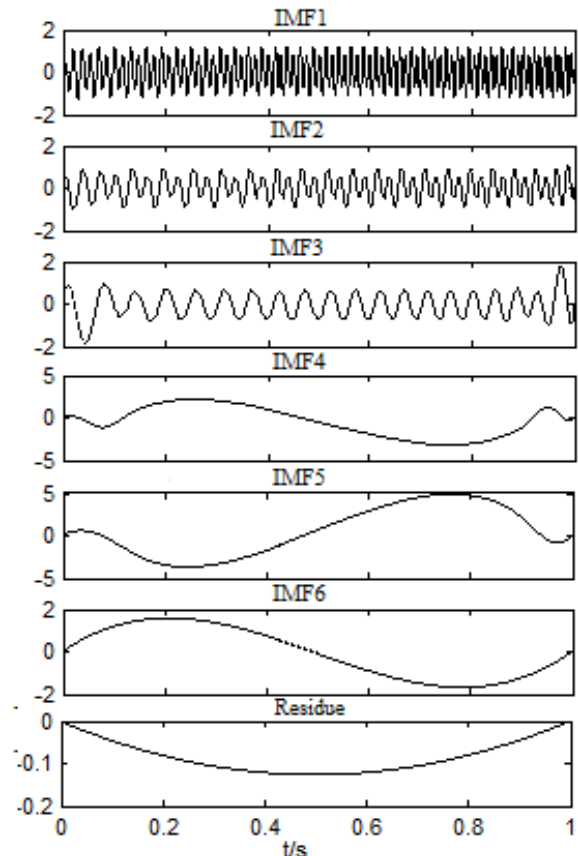

FIGURE 2 EMD of original signal

Appling Hilbert transform to IMF1, IMF2 and IMF3, the instantaneous frequency and instantaneous phase curves were acquired, described in Figure 3 and Figure 4.

The Hilbert spectrum clearly reflected the trend of the instantaneous frequencies of IMF1, IMF2 and IMF3. As shown in Figure 3, however, 'boundary effects' occurred obviously at the beginning and the end of the curve. What's more, the instantaneous frequencies fluctuated visibly because of 'Gibbs effect'. 'Boundary effects' and 'Gibbs effect' will increase errors of the following analysis [7].
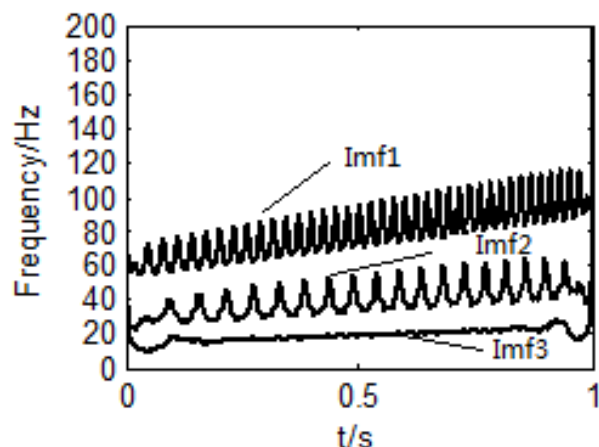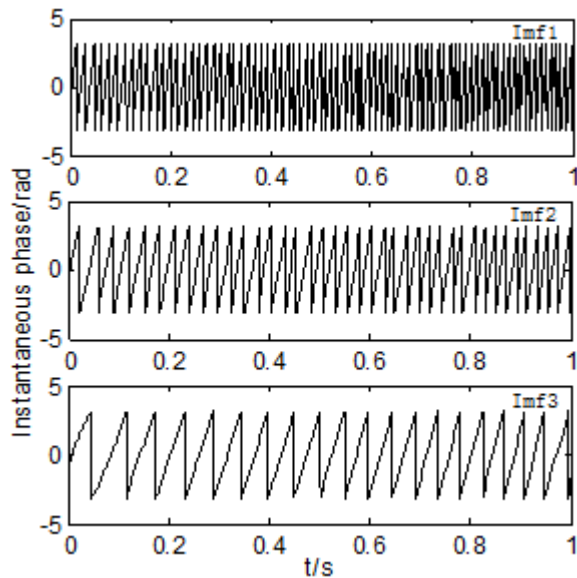

FIGURE 3 Hilbert spectrum

FIGURE 4 Instantaneous phase in time

After processing the instantaneous phases, the continuous phases of IMF1, IMF2 and IMF3 were obtained firstly, shown in Figure 5.
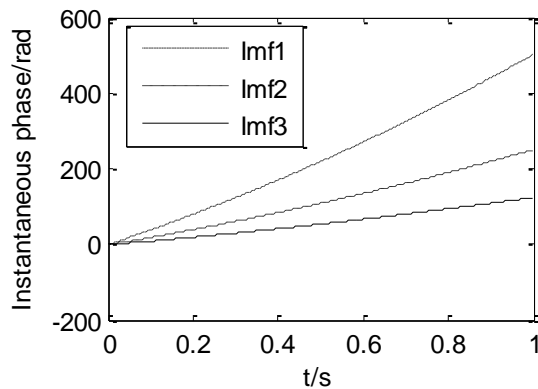

FIGURE 5 Continuous phase in time

$\phi_i(t)$ indicated the continuous phases, where $i = 1,2,3$, therefore:

$$\phi_1(t) = 2\phi_2(t) = 4\phi_1(t). \tag{14}$$

Extracting the instantaneous phases of the base frequency component or the multiple frequency components, the rotating angle changed by time can be obtained. Figure 6 showed the continuous phases of IMF1, IMF2 and IMF3 after quadratic polynomial curve fitting and the actual phase curve of the simulating signal, where the actual phase is shown in Equation (15).

$$\theta(t) = 2\pi(10 + 5t)t, \tag{15}$$
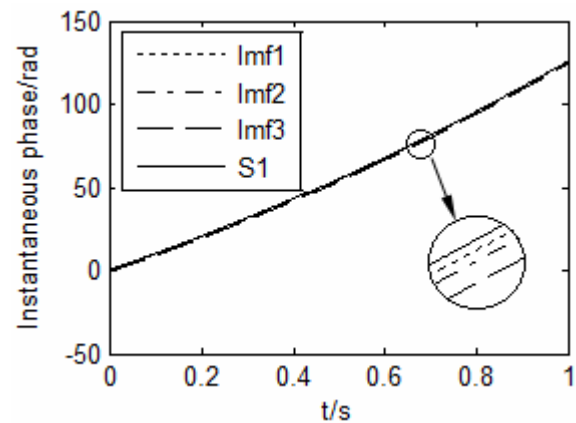
FIGURE 6 Fitting continuous phase in time

Figure 6 showed that the phases obtained by this method were close to the real phase of the original signal, and the error is small. According to the continuous phases, the angle domain signals were acquired through the re-sampling and interpolation of the original signal, as shown in Figure 7.
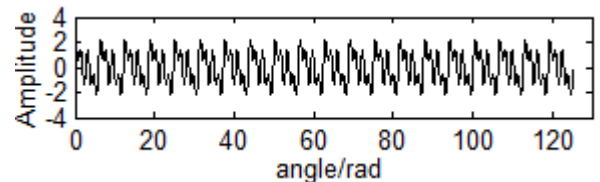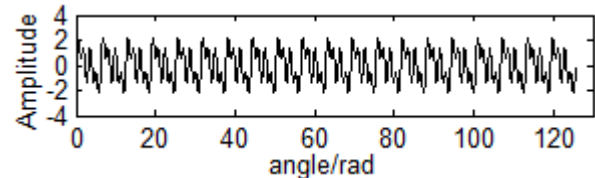

a) Resample signal in angle domain phase 1
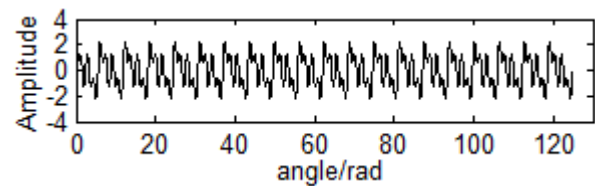
b) Resample signal in angle domain using phase 2
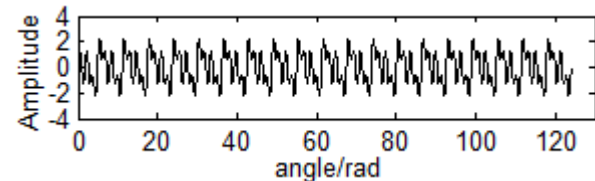
c) Resample signal in angle domain using phase 3

d) Resample signal in angle domain using theta

FIGURE 7 Resample signal in angle domain

Appling FFT operation to the angle domain signals, the order spectrums of original signal were obtained, illustrated in Figure 8. It was shown that the order spectrums and the one analysed by actual phase coincided. The results show that the order spectrum can be yielded through the phase information of each component of the original signal.

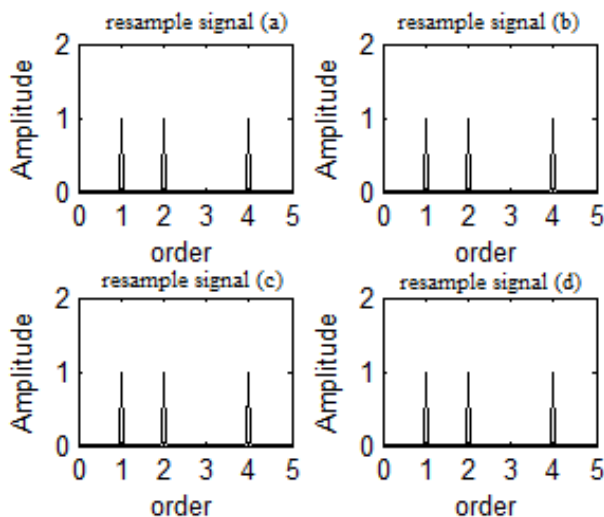Zeng Rong, Jiang Zhengfeng, Ling He, Hu Wei, Wan Xing

FIGURE 8 Order spectrum

The above results were obtained when the original signal was without noise. Since EMD is not capable to eliminate noise automatically, Hilbert spectrum will occurs large deviation at noisy conditions. Figure 9 showed the order spectrum of the original signal with noise by the illustrated method, the SNR of which is 5. It was shown that apparent frequency confusion and amplitude error existed. Consequently, signal denoising is significant before EMD.
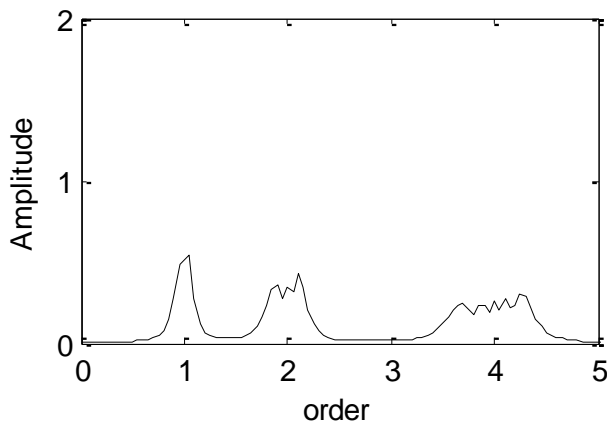

FIGURE 9 Order spectrum (SNR=5)

4 Implementation of the method

In summary, the implementation process of the order analysis based on instantaneous phase is as follows:
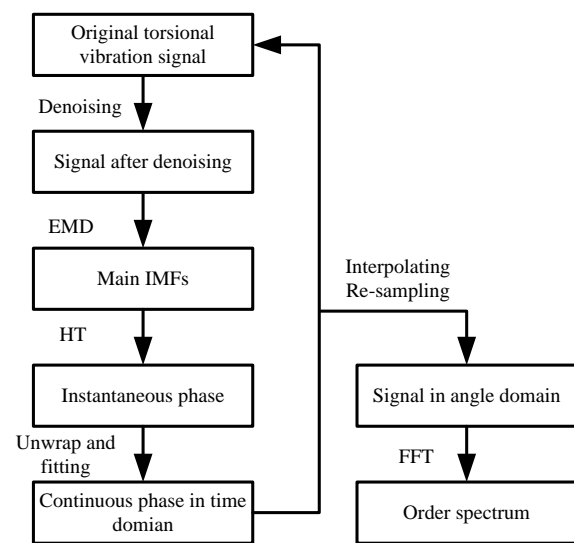

FIGURE 10 Implementation process of the method

5 Experiments and verification

Torsional vibration test was implemented on an dynamic characteristics test bench of a vehicle torsional damper. Figure 11 showed the test bench. The dedicated frequency inverter was used to control the motor, which generated torsional vibration. The eddy current dynamometer simulated load, and the encoders were used to detect the rotating speed of the shaft system. The test was carried out under a constant load condition. The motor speed controlled by the inverter increased from 900r/min to 1200r/min within 5s, where the angular acceleration was π rad/s2. Since the encoder generated 1440 pulses when the shaft rotated one circle, the sampling frequency was set to be 100 KHz. The data within 0.8s was intercepted to be processed by pulse interval method [11], yielding the torsional vibration data, as shown in Figure 12a. After wavelet denoising [12-13], the torsional vibration data was shown in Figure 12b.
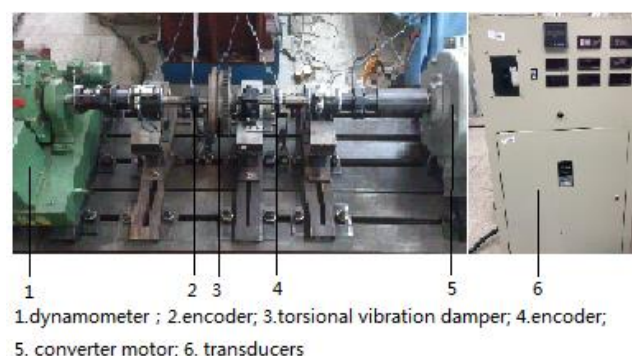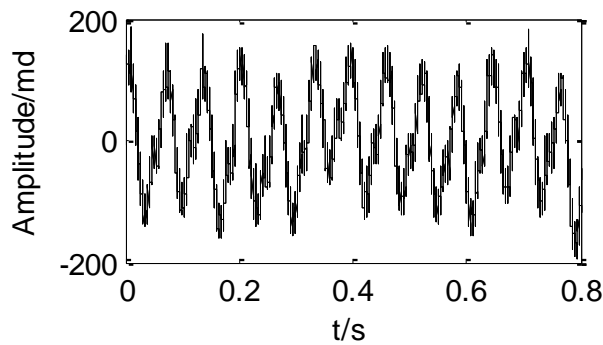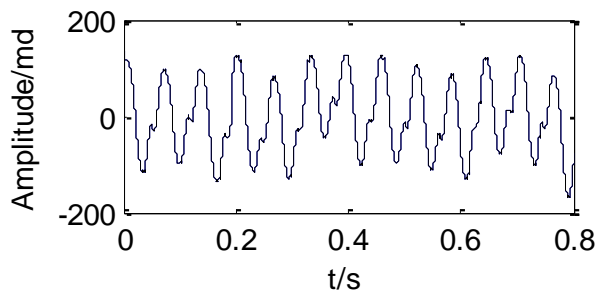

1.dynamometer ; 2.encoder; 3.torsional vibration damper; 4.encoder;
5. converter motor; 6. transducers

FIGURE 11 Torsional vibration test bench

Zeng Rong, Jiang Zhengfeng, Ling He, Hu Wei, Wan Xing

a) Original torsional vibration signal

b) torsional vibration signal after denoising

FIGURE 12 Torsional vibration 3signal

Using EMD, the torsional vibration signal was decomposed, obtaining the IMFs, where IMF1 and IMF2 corresponded to the original signal. Then by Hilbert transform of IMF1and IMF2, the Hilbert spectrum of the signal can be obtained, illustrated in Figure 13. It showed that IMF2 was corresponding to the base frequency component of the torsional vibration signal. Thus, order analysis of the original signal was based on the instantaneous phase of IMF2. Figure 14 described the instantaneous phase of IMF2, and Figure 15 showed the continuous phase curve over time. According to the continuous phase curve, the interpolation re-sampling was implemented, getting the signal in angular domain, shown in Figure 16. Finally, the angular domain signal was calculated by FFT, which obtained the order spectrum of the original signal, as shown in Figure 17.
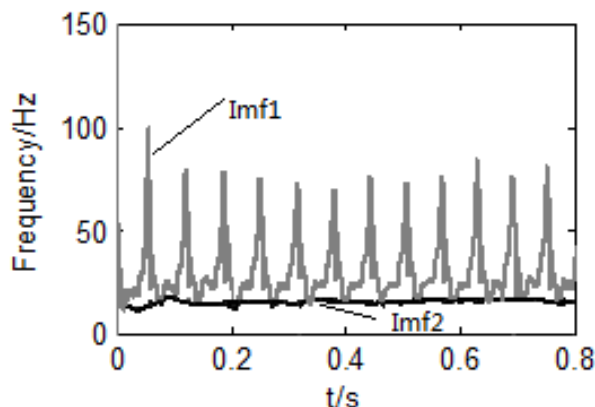

FIGURE 13 Hilbert Spectrum
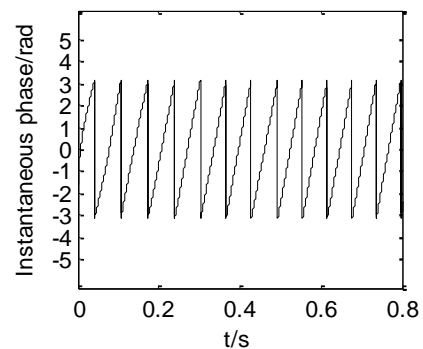
FIGURE 14 Phase in time of IMF2
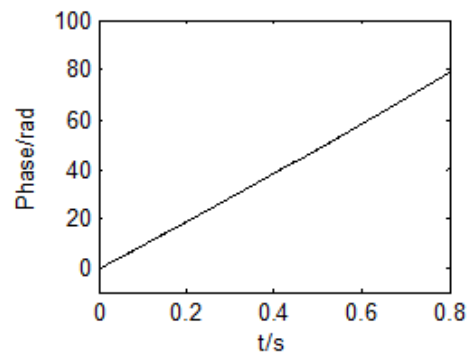
FIGURE 15 Continuous phase in time of IMF2
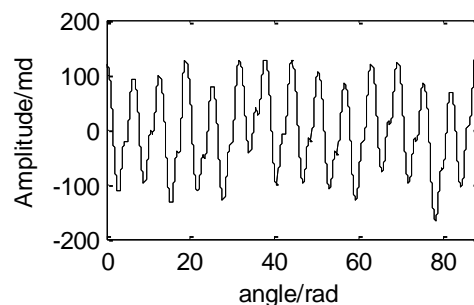
FIGURE 16 Torsional vibration signal in angle domain
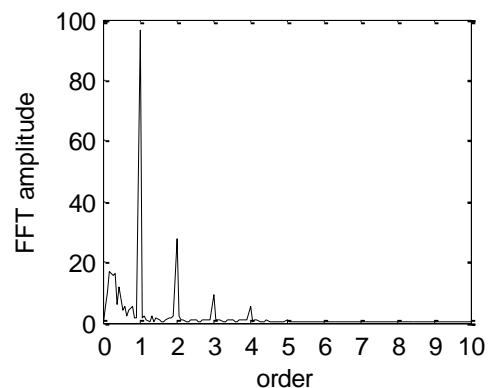
FIGURE 17 Order spectrum

The Signal order spectrum showed that the original torsinal vibration signal was mainly composed four harmonic components, and the amplitudes of each harmonic component were 96.1867md, 27.5165md, 9.005md and 5.5002md respectively. The continuous phase was processed by quadratic polynomial fitting, acquiring the phase function of time as shown in Equation (13). The angular acceleration obtained by the Equation

(13) was 6.2084/2=3.1043 rad/s^2, which was coincided with the actual angular acceleration. The result validates the reliability and accuracy of the order analysis method based on instantaneous phase.

6 Conclusions

The vibration signals of rotating machinery are usually compound signals relevant to the rotating frequency. Using HHT method, the instantaneous phases of the rotating frequency component and multiple frequency components can be extracted, which can be computed to the rotating angle, thus the time sequence of the even angle resample can be gained.

The non-stationary torsional vibration signals of rotating machinery were theoretically and experimentally analysed, using order analysis method based on instantaneous method. The order spectrums of the torsional vibration signals were achieved. The simulating and experimental results both verified the feasibility and accuracy of the proposed method. The proposed order analysis method takes advantage of the instantaneous phase obtained by HHT. Compared with the method using instantaneous frequency, this method is more convenient and faster, and it is less effected by the 'boundary effect', thus can obtain higher accuracy.

Acknowledgments

References

[1] Jiongming Y, Shuren Q, Zhong J 2005 Research on the Methods to Implement Order Sampling in Rotating Machinery Order Analysis Technique *China Mechanical Engineering* **16**(3) 249-53 *(in Chinese)*

[2] Fyfe K.R, Munck E D S 1997 Analysis of Computed OrderTracking *Mechanical Systems and Signal Processing* **11**(2) 187-205

[3] Xiong K, Zhang Y, Zhang Z, Wang S, Zhong Z 2014 PA-NEMO: Proxy mobile IPv6-aided network mobility management scheme for 6LoWPAN *Elektronika ir Elektrotechnika* **20**(3) 98-103

[4] Hui S, Wei J, Qian S 2003 *IEEE Transactions on Instrumentation and Measurement* **52**(3) 754-61

[5] Guo Y, Qin S, Tang B, Ji Y 2003 Order Tracking of Rotating Machinery Based on Instantaneous Frequency Estimation *Chinese Journal of Mechanical Engineering* **39**(3) 32-6 *(in Chinese)*

[6] Xu M, Huang W, Zhang J 2000 Application of Haar Wavelet on Analysis of Vibration Signal of Rotating Machinery in Fast Run-up State *Journal of Vibration Engineering* **13**(2) 216-21 *(in Chinese)*

[7] Jia J, Kong F, Wang J 2005 Order Analysis of Internal Combustion Engine Signal Based on Instantaneous Frequency Estimation *Chinese Internal Combustion Engine Engineering* **26**(3) 15-21 *(in Chinese)*

[8] Qi Y, Tang M, Zhang M 2014 Mass customization in flat organization: The mediating role of supply chain planning and corporation coordination *Journal of Applied Research and Technology* **12**(2) 171-81

[9] Duffy D G 2004 The Application of Hilbert–Huang Transforms to Meteorological Datasets *Journal of Atmospheric and Oceanic Technology* **21**, 599-611

[10] Li Q, Wang H 2013 A Research Review of Hilbert-Huang Transform Used for Rolling Bearing Fault Diagnosis *Applied Mechanics and Materials* **397-400** 2152-5

[11] Zeng R, Jiang Z, Zhou B 2010 Pulse Interval Method for Engine's Torsional Vibration Measurement Based on LabVIEW *2010 International Conference on Computational Intelligence and Software Engineering (CiSE)* 1-4

[12] Zhang C, Huang L, Zhao Z 2013 Research on combination forecast of port cargo throughput based on time series and causality analysis *Journal of Industrial Engineering and Management* **6**(1) 124-34

[13] Sardy S, Tseng P, Bruce A 2001 *IEEE Transactions on Signal Processing* **49**(6) 1146-52

Authors

Rong Zeng, born in February, 1988, Wuhan City, Hubei Province, P.R. China

Current position, grades: PhD, School of Mechanical and Electronic Engineering, Wuhan University of Technology, since 2010.
University studies: BSc in Mechanical Engineering and Automation at Wuhan University of Technology in 2008. MSc in Mechanical Manufacturing and its Automation at Wuhan University of Technology in 2011.
Scientific interests: dynamics and vibrations, parameter identifications and vibration signal processing.

Zhengfeng Jiang, born in January, 1949, Wuhan City, Hubei Province, P.R. China

Current position, grades: professor, School of Mechanical and Electronic Engineering, Wuhan University of Technology.
University studies: BSc in Mechanical Engineering and Automation from Wuhan University of Technology in 1976. MSc in Mechanical Manufacturing and its Automation at Wuhan University of Technology in 1982.
Scientific interests: dynamics and vibrations, vibration control.

He Ling, born in February, 1982, Wuhan City, Hubei Province, P.R. China

Current position, grades: lecturer, School of Mechanical and Electronic Engineering, Wuhan University of Technology.
University studies: BSc in Mechanical Design Manufacturing and Automation at Huazhong University of Agriculture in 2004. MSc in Mechatronics Engineering at Wuhan University of Technology in 2007. PhD in Mechanical Manufacturing and its Automation at Wuhan University of Technology in 2011.
Scientific interests: advanced manufacturing technology and equipment.

Wei Hu, born in November, 1988, Wuhan City, Hubei Province, P.R. China

Current position, grades: master, School of Mechanical and Electronic Engineering, Wuhan University of Technology.
University studies: BSc in Mechanical Engineering and Its Automation at Tianjian Polytechnic University in 2011. MSc in Mechanical Manufacturing and its Automation at Wuhan University of Technology in 2014.
Scientific interests: dynamics and vibrations, signal processing.

Xing Wan, born in March, 1986, Wuhan City, Hubei Province, P.R. China

Current position, grades: Intermediate Engineer, Dongfeng Motor Group Co., LTD Technology Centre
University studies: BSc in Mechanical Engineering and Automation at Wuhan University of Technology in 2004. MSc in Mechanical Manufacturing and its Automation at Wuhan University of Technology in 2011.
Scientific interests: NVH, vibration testing.

The study of communication fibre amplifier based on doped nano-scale semiconductor materials

Yanru Xue[*], Yinghua Yao, Min Liu, Feng Wang

Hebei Normal University of Science & Technology, Hebei, 066004, China

Abstract

This paper firstly describes the development of communication fibre with InP-doped nano-scale semiconductor materials in detail, and then discusses its important dispersion characteristics, starting from the definition of fibre materials dispersion to explore the affected factors of dispersion, different dispersions on fibre as well as the dispersion features of single mode fibre. From the perspectives of experimental and theoretical calculations, it analyses the dispersion characteristics of drawing and doped nano-scale fibre. Thus, it will have much broader prospects for sano-scale semiconductor materials as doping fibre amplifier in communication.

Keywords: nano materials, communication optical fibre, fibre amplifier

1 Introduction

Composite material is now becoming application program which is accepted in much major structure, especially in space and aerospace, marine, civil engineering and automotive industry. In modern optical network, signal remote transmission relies on optical fibre transmission, and current backbone network in communication network also mainly depends on fibre-optic transmission. Fibber optic transmission system as long distance, high-speed and high-capacity backbone network, continues to be enlarged in transmission after the light signal attenuation is unavoidable [1]. The emergence of optical fibre amplifier has better settled down the enlargement of relaying, avoiding over-input of light-electricity-light transformation in time and cost, which provides transmission of high speed and long distance. Current existing amplifier with light mainly can be divided into three categories, including semiconductor laser amplifier, non-linear optical amplifier and doped optical amplifier [2]. Semiconductor laser amplifier is used for semiconductor laser unit working below the threshold, input from one end by amplified optical signal and enlarged in active terminal, output from another end. Non-linear optical fibre amplifiers mainly refer to fibre Raman amplifier using simulated Raman scattering effects of silica fibre in hard light with reasonable wavelength, to amplify the light signal. Doped fibre amplifier refers to silica fibre dope with rare-earth elements to form a multilevel system, under the continuous pumping of pump light realizing population inversion, and then use signal light to guide induction, the particles simulated for radiation transition, thus amplified the signal light.

Scale-up experiment of rare earth doped fibre firstly was held in 1964. With the development of GaAs semiconductor diode laser amplifier, the first diode pumping fibre laser realized the application in optical communication system in 1974. The reduction of optical fibre transmission loss and application development of InGaAs and InGaAsP diode laser in the 70s leads to the development of modern optical fibre communication system with 1.3urn and 1.55um, which promotes the rare-earth doped device development [3]. Semiconductor optical amplifier is verified experimentally, which can increase the receiver sensitivity (as the preamplifier) or transmission length (as a repeater), also can realize multiple amplification in a communication system. However, the nonlinear effects of signal-mode have been widely concerned in the meantime (e.g. Raman scattering, Brillouin is scattering and four-wave mixing); these nonlinear effects can have a potential to realize signal amplification, to help the standard quartz matrix fibre become amplification medium [4].

For different optical fibre substrates, the ways of doping are various, including two aspects: quartz matrix doped fibre and glass matrix doped fibre. As for the former, Won-Taekhan et al from South Korea Gwangju Institute of Science and Technology use MVCD method, combining solution immersion with high-temperature processing, and applying traditional physical adsorption doping technique for preparing optical fibre, doping PbTe nano-scale semiconductor compounds in fibre, its amplified spontaneous emission can be achieved in waveband of 1537nm. Recent researched about optical fibre doping by using MVCD method, including doped InP nano-particle in silica fibre cladding, radiation wavelength of 1080-1350nm and Lipu light of 532mn. For the latter, glass matrix doping, as the glass matrix of transmission optical signals, generally can be divided into four types, namely oxide glass, halide glass, oxide glass and

chalcogenide glass. According to these amplified fibres with different matrixes doping nano-scale semiconductor materials, many research groups have achieved good results. Therefore, some related studies can judge that, in order to meet the demand of bandwidth and optical amplifier gains, the doped sources are advanced from rare-earth elements to nano-scale semiconductor materials, and doped matrixes also have silica fibre, glass fibre, photonic crystal fibre and so on. It also can judge the research for further amplification fibre, mainly focusing on amplified fibre using nano semiconductor materials as doped sources.

2 The preparation of optical fibre doped with InP nano-scale materials

InP nano-scale materials-doped optical fibre in this paper is mainly realized by MVCD method [5].

2.1 PREFORM FABRICATION

Here, we use MVCD method, successfully fabricating two new structures of optical fibre, namely InP nano-particle-doped fibre and fibre of InP nano inner cladding. The preform fabrication of these two fibres is presented in Table 1.

TABLE 1 Optical fibre preform fabrication method

InP nano-particle-doped fibre	Fibre of InP nano inner cladding
1. The first step is to make the core rod: filling the fixed reaction tube with halide $SiCl_4$, C_2F_6 steam and mixed gases of InP materials without heating vaporization by using N_2. As previous said that mixed gases moves towards the direction of principle axis, using oxy-hydrogen flame to form dust through chemical vapour reaction, and the dust (the generated oxides, e.g. SiO_2) is deposited in tube walls, finally sintering shrinkage quartz tube to form solid core rod.	1. The first step is to make the core rod: filling the fixed quartz tube with halide $SiCl_4$ and C_2F_6 steam for chemical vapour deposition at a high temperature, and sintering the quartz tube with SiO_2 tectorium under the reaction of $SiCl_4(G)+O_2(G)->$ and $SiO_2(S)+2I_2(G)$ at high temperature to be solid core rod;
2. The second step is cladding frication: similarly as before, placing the quartz tube on the lathe, putting $SiCl_4$ into the quartz tube by oxygen with maintained high-temperature conditions for depositing outsourcing layer of SiO_2, and then adjusting the flow rate of C_2F_2 to control cladding index, the principle of which has been discusses in the previous section;	2. The second step is that with the gasification of heat light, it fills with the same reaction gas as the step 1, using same method to deposit loose outer layer of SiO_2 in quartz tube;
3. The third step is to insert the finished solid core pf step 1 to the deposited quartz tube of step 2, fixed in the lathe; it can fill inert gas in this process to prevent oxidation of the deposited particles;	3. The third step is to change the conveying gas of the second step to inert gas for transporting InP steam, depositing InP film sediments in the loose layer of SiO_2;
4. Finally, heating the mixed quartz tube of step 3, to make the quartz tube collapse into an optical fibre perform with doped InP, preparing the next step of wiredrawing.	4. The fourth step is to insert the core rod of step 1 to the deposited quartz tube of step 3 with the maintained high temperature and inert gas in the process;
	5. Finally, heating to a suitable temperature, to make the quartz tube of step 4 collapse into a preform for the next step of wire drawing.

2.2 THE WIREDRAWING OF INP NANO-PARTICLE-DOPED FIBRE

These two kinds of optical fibres with the same drawing convey the preform to the heating furnace with a certain speed and uniform, controllable manner by using feeder, and drawing general fibres need to be heated to 2000℃. The degree of rod is decreased, relying on its own weight gradually tapering into fibres, and using accurate mechanical control to draw the dissolved preform to fibre. The main components of optical fibre drawing machine include conveying preform, heating devices, controlling assembly of wiredrawing diameters, the parts of coating fibre, curing fibre components, traction wheel, winding and other control units. Laser diffraction light of fibre diameters ejected from the fibre can change the distraction pattern by optical controlling diameter variations, successively changing the photodiode. This immediate change will be in a form of signals, passing to servo control machinery for adjusting the speed of winding fibre. Fibre diameter variations can rely on this technique to stability control within 0.1%. As a control factor in drawing, it needs to consider the tension of all components in doping fibre, and keep drawing balanced with controllability; while ensuring all parts of doped InP materials without breakage and volatilization, it leads to the affected part undoped InP nano-materials, and drawing should choose a relatively temperature of 1600-170 (with an appropriate

speed of less than 5m/min [6]. The typical value of wrapped fibre diameter is 250 um, and it can be increased to 900 um with multiplayer package.

2.3 INTEGRATION SCHEME WITH ALD TECHNOLOGY

Through the above analysis, it has been successfully drawing the required doped fibres, to further improve the characteristics of optical fibre and advance the drawing, which can consider the combination of MVCD and atomic layer deposition (ALD) technologies, to be improved, as shown in Figure 1.
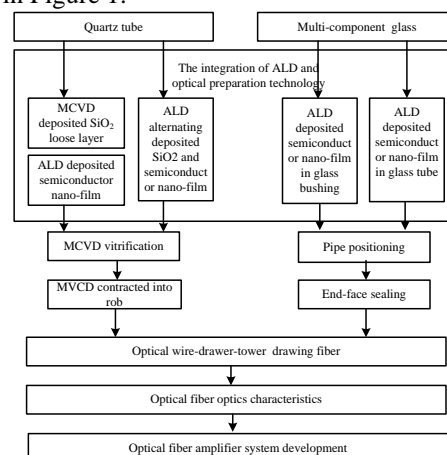

FIGURE 1 the integrated study of MVCD and ALD

For the purpose of the study, we mainly adopt the scheme that after depositing SiO₂ with general MVCD, it uses ALD to wrap the doped semiconductor materials in the surface of loose body with the form of atomic layer, and then uses MVCD to vitrify the loose body; in this process with high temperature, the doped source in atomic layer diffuses and splits material structures of SiO₂, to realize the nano-scale semiconductor doping, but due to the large surface of loose body, thus this scheme has high efficiency of doping.

2.4 NANO-SCALE SEMICONDUCTOR DOPED WITH SILICA FIBRE PREFORMS

In MVCD, by controlling the parameters including tube pressure, flow rate, glass transition temperature and head movement speed, it can acquire different semiconductor doped fibre matrix materials, using a variety of testing methods, such as XRD, SEM, TEM, absorption spectrum, Raman spectrum and so on, to analyse the obtained verification doped layer and then adjust and optimize technological parameters. After verification of doped loose body, collapsing into rod through base tube, it uses preform to analyse refractive index distributions of system testing preform, and acquire the material samples in doped area of preform by grinding and particle beam reduction, then analysing its microstructure characteristics, in order to achieve phase transition properties with high temperature of doped fibre waveguide matrix. According to the highly volatile of nano-scale semiconductor at high temperature, it can be settled down through the following techniques:

1. selecting different precursors of Lead, increasing the solubility of doped materials in SiO2 structures to improve the concentration;

2. using ALD to alternating deposit semiconductor doped sources and other oxides, to form a protective layer in the surface of semiconductor atomic layer, so as to improve the doping concentration and avoid volatile;

3. using ALD technology first to dope the nano-scale semiconductor into the loose layers, and then using MVCD to deposit SiO2 as a protective layer;

4. tubes with protective gas during the vitrification, for instance, avoiding sulfur volatilization can access sulfur-containing to restrain its diffusion volatility.

While in drawing, preform diameter is greatly reduced, and the film thickness of semiconductor in substrate interface is sharply decreased. Which is because semiconductor film is pulled by axial drawing is a molten state to form flow diffusion, resulting in the ruptures of film, surface shrinkage and semiconductor particles changing from condensate to nano-cluster, so as to form nano-scale semiconductor particles between the core and cladding of the fibre. Through the adjustment of wire temperature and drawing speed, it can control the scale of nano-scale semiconductor with its distribution, improving amplifier efficiency and reducing the loss.

3 The analysis of doped fibre characteristics

3.1 MATERIALS DISPERSION

The emergence of material dispersion is due to the different material for producing optical fibre, and refractive index changes with the light impulse frequency. According to the root causes, when materials absorb electromagnetic radiation, raw materials dispersion is relative to intrinsic resonant frequency. However, it has nothing to do with the materials resonance, the refractive index expressed as Sellmeier Equation [7]:

$$n^2(\omega) = 1 + \sum_{j=1}^{M} \frac{B_j \omega_j^2}{\omega_j^2 - \omega^2}, \tag{1}$$

where, ω_j represents the resonant frequency of material, β_j oscillation intensity. n stands for n_1 or n_2, depending on the dispersion characteristics belonging to core or cladding [1]. The calculations and results contain all materials of resonance characteristics, relying on frequency. As for optical fibres, analysing filtration by characteristic curve, the parameters β_j and ω_j can get the empirical value $M = 3$.

Simply assuming the impulse as plane wave, only considering the effects of materials dispersion, and based on this can ignore the influences of waveguide dispersion, it can use Equation (2) to represent the group delay τ:

$$\tau(\omega) = \frac{d\beta}{d\omega}, \tag{2}$$

where, β presents the transmission constant of plane wave, ω angular frequency of the light, and $\beta = \omega n(\omega)/c$, $n(\omega)$ for refractive index and C for the speed of light in vacuum. It can obtain that:

$$\tau(\omega) = \frac{1}{c}\left[n(\omega) + \omega \frac{dn(\omega)}{d\omega} \right]. \tag{3}$$

Substituting $\lambda = 2\pi c/\omega$, then:

$$\tau(\omega) = \frac{1}{c}\left[n(\lambda) - \lambda \frac{dn(\lambda)}{d\lambda} \right]. \tag{4}$$

The dispersion parameter is:

$$D = \frac{d\tau}{d\lambda} = -\frac{\lambda}{c}\frac{d^2 n}{d\lambda^2} \text{ (Ps/km/nm)}. \tag{5}$$

3.2 WAVEGUIDE DISPERSION

Any lights transmission in optical fibre, its group velocity varies with optical wavelength, called as waveguide dispersion [8]. To determine the structure parameters by fibre refractive index distribution is the main cause of waveguide dispersion. At the meantime, different light modes transmitted in optical fibre will lead to the various waveguide dispersions. The sum of different patterns of

waveguide dispersions in optical fibre shall be the waveguide dispersion of multimode fibre. To analyse waveguide dispersion can start from the group velocity. Considering the length of single mode fibre (SMF) as L, when light frequency is ω, and fibre transmission reaches the end, with the transmission time as $T = L / v_g$, while v_g represents group velocity, it can be defied as:

$$v_g = (d\beta / d\omega)^{-1}, \tag{6}$$

where propagation constant $\beta = \bar{n}k_0 = \bar{n}\omega / c$, the group velocity can be expressed as $v_g = c / \bar{n}_g$, while $\bar{n}_g$ stands for group refractive index:

$$\bar{n}_g = \bar{n} + \omega(d\bar{n} / d\omega). \tag{7}$$

Supposing $\Delta\omega$ as impulse frequency bandwidth, the impulse bandwidth when optical length is L should be:

$$\Delta T = \frac{dT}{d\omega}\Delta\omega = \frac{d}{d\omega}\left(\frac{L}{v_g}\right)\Delta\omega = L\frac{d^2\beta}{d\omega^2}\Delta\omega = L\beta_2\Delta\omega, \tag{8}$$

where $\beta_2 = d^2\beta / d\omega^2$ expresses group velocity dispersion parameters, which indicate the broadening degree of optical impulse in transmission. In some optical communication systems, the frequency bandwidth $\Delta\omega$ is determined by the width of emission wavelength $\Delta\lambda$. Therefore, it generally uses $\Delta\lambda$ instead of $\Delta\omega$, introducing $\omega = 2\pi c / \lambda$ and $\Delta\omega = (-2\pi c / \lambda^2)\Delta\lambda$ to the Equation (8) can obtain:

$$\Delta T = \frac{d}{d\lambda}\left(\frac{L}{v_g}\right)\Delta\lambda = DL\Delta\lambda, \tag{9}$$

where:

$$D = \frac{d}{d\lambda}\left(\frac{L}{v_g}\right) = -\frac{2\pi c}{\lambda^2}\beta_2. \tag{10}$$

For the convenient analysis below, the dispersion parameter D is expressed as the above equation, with the unit ps/km/nm. Thus, it can be concluded that the dispersion parameter D relying on wavelength variations is composed by refractive index $\bar{n}$ of the mode depending on frequency. For the above, it can get:

$$D = -\frac{2\pi c}{\lambda^2}\frac{d}{d\omega}\left(\frac{1}{v_g}\right) = -\frac{2\pi c}{\lambda^2}\left[2\frac{d\bar{n}}{d\omega} + \omega\frac{d^2\bar{n}}{d\omega^2}\right]. \tag{11}$$

Thus, it can express D as the sum of two items:

$$D = D_M + D_W, \tag{12}$$

where D_m stands for the detail expression of material dispersion, D_w for waveguide dispersion, which can obtain respectively:

$$D_M = -\frac{2\pi}{\lambda^2}\frac{dn_{2g}}{d\omega} = \frac{1}{c}\frac{dn_{2g}}{d\lambda}, \tag{13}$$

$$D_W = -\frac{2\pi\Delta}{\lambda^2}\left[\frac{n_{2g}^2}{n_2\omega}\frac{Vd^2(Vb)}{dV^2} + \frac{dn_{2g}}{d\omega}\frac{d(Vb)}{dV}\right]. \tag{14}$$

While, n_{2k} presents group refractive index of the cladding:

$$V = k_0 a\left(n_1^2 - n_2^2\right)^{\frac{1}{2}} \approx (2\pi / \lambda)an_1\sqrt{2\Delta},$$

$$b = \frac{p / k_0 - n_2}{n_1 - n_2} = \frac{n - n_2}{n_1 - n_2},$$

V for normalized frequency, k_0 for free space-wavenumber $k_0 = \omega / c = 2\pi / \lambda a$ while a stands for optical radius, n_1 and n_2 for fibre core and cladding refractive index respectively, and $\Delta = (n_1 - n_2) / n_2$ for relative refractive index unrelated to frequency, b for normalized transmission constant [9].

3.3 THE ANALYSIS OF SINGLE-MODE OPTICAL DISPERSION

In order to analyse the simplification of the model, it can take the effect of removing mode distortion, thus setting single mode fibre as the ideal single mode fibre with circular symmetry, where the optical dispersion can be simplified as the derivative of group delay to wavelength, while dispersion is a major cause of impulse broadening. Supposing refractive index as:

$$n^2(r) = n_2^2\left[1 + 2\Delta f\left(\frac{r}{a}\right)\right], \tag{15}$$

where, $f\left(\dfrac{r}{a}\right)$ represents distribution function of refractive index and $0 \le V\left(\dfrac{r}{a}\right) \le 1$.

The wave equation of wave light transmitting in optical fibre is shown as below:

$$\frac{1}{R}\frac{d}{dR}\left(\frac{Rd\psi}{dR}\right) + \left[V^2 f\left(\frac{r}{a}\right) - V^2 B - \frac{m^2}{R^2}\right]\psi = 0. \tag{16}$$

While, $\psi\left(\dfrac{r}{a}\right)$ stands for radial variable Eigen function of the wave light, and $R = \dfrac{r}{a}$, m for radical mode number, $V = ka\left(n_1^2 - n_2^2\right)^{V_2}$, and B for normalized transmission constant, which can be defined as:

$$B = \left(\beta^2 / k - n_2^2\right) / \left(n_1^2 - n_2^2\right). \tag{17}$$

While, N_1, N_2 are expressed as the group refractive index of n_1 and n_2, namely:

$$N_1 = n_1 - \lambda(dn_1/d\lambda), N_2 = n_2 - \lambda(dn_2/d\lambda), k = 2\pi/\lambda.$$

Equation (17) can be simplified based on the weak-guide as:

$$B = \frac{\beta - kn_2}{kn_2\Delta}. \tag{18}$$

Using Equation (17) for the derivation of k:

$$\frac{d\beta}{dk} = N_2 + (N_1 - N_2)\frac{d(VB)}{dV}. \tag{19}$$

The group delay of unit length is:

$$\tau = \frac{d\beta}{d\omega} = \frac{1}{c}\frac{d\beta}{dk}. \tag{20}$$

Thus:

$$\tau = \frac{N_2}{c}\left[1 + \frac{N_1 - N_2}{N_2}\frac{d(VB)}{dV}\right] \approx \frac{N_2}{c}\left[1 + \Delta\frac{d(VB)}{dV}\right].$$

This can reach the dispersion parameter D:

$$D = \frac{d\tau}{d\lambda} = M_2\left[1 + \Delta\frac{d(VB)}{dV}\right] - \frac{N\Delta}{\lambda c}\left[V\frac{d^2(VB)}{dV^2} - P\frac{d(VB)}{dV}\right], \tag{21}$$

where, $M_2 = \frac{1}{c}\frac{dN_2}{d\lambda}$ expresses material dispersion, which can obtain $M_2 = \frac{1}{c}\frac{d^2n_2}{d\lambda^2}$, $P = \frac{\lambda}{\Delta}\frac{d\Delta}{d\lambda}$ as profile dispersion parameter.

It can be seen from the equation of dispersion parameter D in Equation (21) that the first term of the equation at the right is material dispersion. The second item is waveguide dispersion, and it can been seen from the parameters of the expression that two items both contain optical structure parameter V which represents the characteristic parameter of material dispersion, relative to the material dispersion as a whole. From the above equations, it can use the numerical calculation of wave equation in optical fibre can calculate the relation curve between normalized transmission constant B and optical structure parameter V under the different modes of light transmissions, further introducing material dispersion M_2 can get the corresponding values of dispersion parameter D in different transmission modes. According to the previous analysis of simplified calculations, as for ideal single mode fibre with circular symmetry $m = 0$, it can simplify the previous wave equation of optical transmission light for calculation.

When optical transmission bit rate is R, the effect of its dispersion can estimate by the principle of $R\Delta T < 1$, while ΔT is the maximum multipath time delay of light transmission with transmission length L, and $\Delta T = DL\Delta\lambda$, $\Delta\lambda$ corresponding to $\Delta\omega$ and $\Delta\omega$ for the bandwidth of light wave frequency. Based on the above analysis, the effect of dispersion on optical fibre transmission can be clearly expressed as criteria $RL|D|\Delta\lambda < 1$. Aiming at single mode fibre with size of RL, it can estimate an order of magnitude of dispersion parameter by the criteria, to evaluate its dispersion characteristics. And for a standard silica fibre, the dispersion parameter D around the range of wavelength with 1.3 μm is relatively small (D – lps/km/nm). For semiconductor lasers, even operating on several longitudinal modes, the spectrum width $\Delta\lambda$ is only 2-4 nm. The product of bit rate and optical length in this light-wave system RL can exceed 100(Gb/s) km. And actually, the typical bit rate of communication system of length 1.3 μm is 2Gb/s, regenerated length of 40-50 km. While applying single-mode semiconductor lasers, the product of single-mode fibre bit rate and optical length can exceed l (Tb/s) km, and now the spectrum width $\Delta\lambda$ will be smaller than 1 nm.

3.4 DISPERSION ANALYSIS OF DOPED INP NANO-PARTICLE OPTICAL FIBRE

The application of drawing amplifier fibre in this paper in real optical devices needs to analyse its dispersion characteristics to identify its impact on the system, this dispersion analysis fixed in the wavelength range of 1200-1600 nm.

The expression of dispersion parameter D is presented as below:

$$D = -\frac{2\pi c}{\lambda^2}\beta = -\frac{\lambda}{c}\frac{d^2n}{d\lambda^2}. \tag{22}$$

From the right side of the equation, it can judge that D associates with the refractive index [10], therein β_2 stands for group velocity dispersion, c for light speed in vacuum. The relation curve of dispersion parameter and wavelength is shown as Figure 2.

Due to the dispersion effects, various frequency components of optical pulse in optical fibre transmission have differences. When the parts of low frequency is faster than high frequency, then the light pulse lies in the normal dispersion region of the optical fibre, namely $\beta_2 > 0$; when high frequency is faster than the low frequency, then the light pulse locates in the abnormal dispersion region, namely $\beta_2 < 0$, which is achieved by the basic definition of normal dispersion and abnormal dispersion. Looking back to the Figure 2, the horizontal line in the diagram represents zero dispersion, and dash curve stands for the relations between the dispersion parameter and wavelength of undoped common single mode fibre, while solid curve expresses the relations between dispersion

parameter and wavelength of single mode fibre with doped InP nano-scale semiconductor materials. It can be seen clearly from the Figure that, as zero dispersion of undoped common single mode fibre lies in $\lambda_D = 1312$ nm. When the wavelength $\lambda < \lambda_D$ namely it is less than 1312 nm. On the left side of intersection between dashed lines and black lines which shows the undoped common single mode fibre lies in the normal dispersion region with low frequency faster than high frequency. On the right side after the higher than zero dispersion, which high frequency of fibre is faster than the low frequency of abnormal dispersion range. As for the single mode fibre with doped InP nano-scale semiconductor materials, from the wavelength range of 1200-1600 nm as shown in solid curve (below), dispersion parameter is less than zero all the time and

positively correlated with wavelength changes, besides, it also shows a normal dispersion of low frequency faster than high frequency transmission. The single mode fibre after doping InP nano-scale semiconductor materials, the performance of its dispersion parameter varies from dashes lines from solid lines, and nano-doped materials make the dispersion parameter curve change down, the main reason of which is the special properties of nano-materials, generating fundamental changes in the dispersion of single mode fibre materials [11]. This fundamental change will be discusses as below, with its main factor derives from quantum effect of nano materials, affecting the optical fibre itself. Compared nano-InP materials with block InP materials, its absorption peak has a blue shift, owing to the doped drawing fibre to affect the absorption peak.
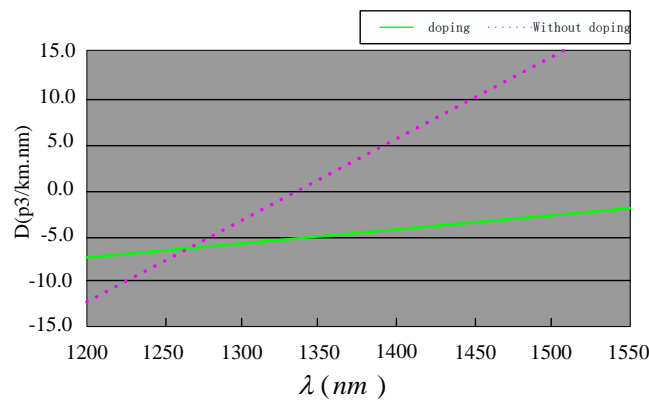

FIGURE 2 Curve of dispersion parameter D changing with wavelength

3.5 OPTICAL AMPLIFICATION CHARACTERISTICS

Transmission length is often limited by fibre loss for any optical fibre communication system [12]. And for long distance transmission system, confinement loss is usually overcome by photoelectric conversion, firstly converting optical signals into electric signals, and then regenerated into optical signals by transmitter [13]. This regenerator will be very complex and expensive to WDM optical system. Thus, another optical amplifier is an important way to overcome this loss, but such a method is to amplify optical signal itself without transferring optical signal into electrical signal [14]. Some specific discussions about fibre amplifiers have been held before, which can be judged that the amplification properties of doped InP particles fibre in this study will become important references for producing optical fibre amplifiers [15].

    Most of optical amplifiers enlarge the incident light by stimulated radiation, which is similar with the structures and principles of lasers. Actually, a fibre amplifier is only a laser without feedback. While amplifiers realizing population inversion by pumping (light or electricity), it realizes the important function of its optical gain. In general, the optical gain not only depends on the frequency of incident light (or wavelength), but also associates with local intensity insider the amplifier. The detail analysis of optical gain frequency dependent and intensity dependent

is related to the amplification medium inside the optical amplifiers, and generally, considering gain medium as model for homogeneous broadening. This gain coefficient of the medium is:

$$g(\omega) = \frac{g_0}{1 + (\omega - \omega_0)T_2^2 + P/P_s}, \qquad (23)$$

where, g_0 represents peak gain, ω incident signal light frequency, ω_0 for atomic jump frequency and p for amplified signal light power. The saturation power p relies on gain medium parameters including burst time and transition process, etc. Parameter T_2 stands for dipole relaxation time with its typical value lower than 1 ps. Burst time T_1 is also known as population relaxation time, varying in 100ps-10ns and relying on gain medium. It can use the above equations to discuss some important features of optical amplifiers, such as gain bandwidth, amplification coefficient, output saturated power and so on. Considering the optical amplifiers under unsaturated conditions, leaving out P/P_s, the gain coefficient can be expressed as:

$$g(\omega) = \frac{g_0}{1 + (\omega - \omega_0)T_2^2}. \qquad (24)$$

The equation indicates that when incident frequency ω is consistent with atomic jump frequency ω_0, the gain coefficient can reach the maximum. Gain bandwidth can be defined as half-high bandwidth of gain spectrum $g(\omega)$. As for Lorentz line, the gain bandwidth is presented by $\Delta\omega_g = 2/T_2$, or:

$$\Delta v_g = \frac{\Delta\omega_g}{2\pi} = \frac{1}{\pi T_2}, \tag{25}$$

amplifies with an appropriate size of bandwidth is quite important for optical communication system. This is because even for a multi-channel signal, the gain is continuous over the entire bandwidth seems very important. The concept of amplifier bandwidth often alternates the notion of gain bandwidth. While considering the amplifier gain G, the difference will be more distinct. The measurement of amplifier gain that widely used is presented as below, thus the amplification coefficient can be defined as:

$$G = P_{out}/P_{in}, \tag{26}$$

where, P_{in} and P_{out} represent the input and output power of amplified continuous wave signal, respectively. And then, the expression of G can be obtained:

$$\frac{dP}{dz} = gP, \tag{27}$$

while, $P(1)$ expresses the optical power from input terminal to the distance Z. Using original state of $P(0) = P_{in}$ to directly integrate to achieve the expression of signal power changing with index:

$$P(z) = P_{in}\exp(gz). \tag{28}$$

According to $P(L) = P_{out}$ it can reach the relationship between amplification coefficient and distance L as:

$$G(\omega) = \exp\left[g(\omega)L\right]. \tag{29}$$

It uses the ratio of input optical power and output optical power to the following tests, so as to measure the amplification characteristics, realizing the optical amplifier bandwidth of the fibre to be measured.

The following steps are to determine the optical amplification features of fibres with drawing InP nano-scale materialsThe working system consists of several parts, including fibre amplifier inside InP nano-film cladding, two wavelength division multiplexers (WDM), pump light source, signal source and output terminals, and between the two WDM is the fibre with InP nano materials to be measure for determining amplification characteristics, pump light entering into the fibre for its optical pumping by coupling of WDM. From the above tests, it shows in Figures 4 and 5. that there have gains in the wave band of 906-1044 nm, 1080-1491 nm and 1524-1596 nm.

3.6 TESTS ON THE GENERATION OF SUPER-CONTINUUM

Super-continuum system as shown in Figure 3, it uses a 120fs passive mode-locking laser with 1550 nm of central wavelength and repetitive rate of 50MHZ as light pulse to be amplified by EDFA, and then inputted into the fibre with InP nano-film cladding. At the output terminal of the fibre, the light beam is divided into two parts by beam splitter with ratio of 1:9, while 90% of light power is inputted into the power meter, and 10% of them are inputted into the spectrum analyser (YOKOGAWA-AQ6370), so as to measure the output spectra. In order to facilitate connecting with conventional fibre and be conductive for optical power to couple into the fibre, the two terminals of InP nano-film cladding will be connected with two transition fibres with length of 1m.
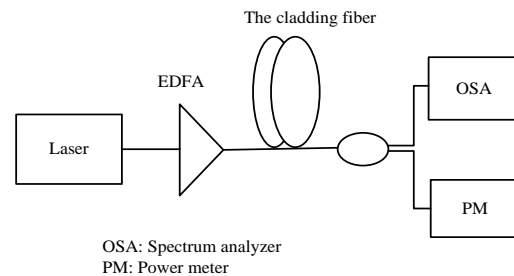

OSA: Spectrum analyzer
PM: Power meter

FIGURE 3 Super-continuum system

The input power is 21.3dBm, the cladding fibre doped nano-scale semiconductor will generate super-continuum spectrum and the fibre length is 30cm and 90cm respectively, as shown in Figures 4 and 5.

FIGURE 4 Super-continuum spectrum of fibre length 30cm
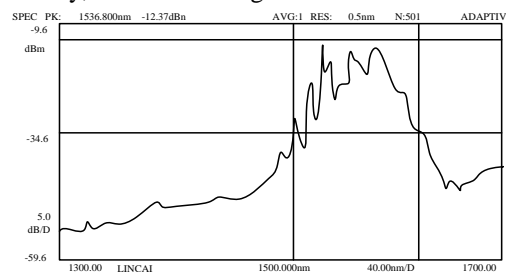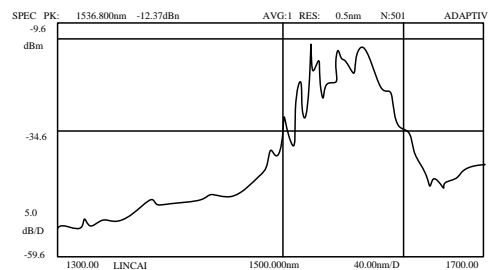

FIGURE 5 Super-continuum spectrum of fibre length 90cm

Figure 4 shows that the fibre length of 30cm acquires 20dB pulse bandwidth of 140 nm. Under the same condition, the measured fibre length of 90cm acquires 20dB pulse bandwidth of 175 nm. The results indicate that the fibre length is longer, and the pulse bandwidth is greater, optical non-linear effects more obvious.

4 Conclusion

This paper mainly studies for the amplification fibre of doped nano-scale semiconductor materials. Firstly, it successfully produces two new structures of optical fibre based on MCVD, and analyse the optical characteristics of amplifies fibre. The appearance of optical fibre amplifier well solves the relay amplification of optical signal in optical transmission system with long distance, high speed and large-capacity backbone network, and amplifies optical fibre as the core part of fibre amplifier will be a focus, while the study on fibre amplifier of nano-scale semiconductor material as doping source will be the main direction and focus of communication optical fibre in the future.

CFX module can ensure accuracy and stability of calculation result as well as basic conservation characteristics and accuracy of numerical value. Pre-processing of 3-D geometric model mainly involved functional operation in DesignModeler on model that have been imported to realize the modelling process of rotating field and external flow field. Through pre-processing of windmill 3-D geometric model by ANSYS Workbench, we used Mesh module to divide the mesh of windmill 3-D geometric model to generate the needed calculation mesh. After division of windmill 3-D geometric model mesh, we entered into fluid analysis setting including designing boundary condition and properties of solver. We only need to do simple solver setting after calculation of fluid and then solution result could be achieved. At last, the numerical simulation of aerodynamic characteristics of windmill was achieved.

Acknowledgement

References

[1] Severin I, El Abdi R, Corvec G, Caramihai M 2014 Optical fibre embedded in epoxy glass unidirectional fibre composite system *Meterials* **7**(1) 44-57

[2] Morakinyo M K, Rananavare S B 2011 Positional control over nanoparticle deposition into nanoholes *The 11th IEEE Conference on Nanotechnology IEEE-NANO 2011* 1677-82

[3] Webb A S, Boyland A J, Standish R J, Lin D, Alam S, Sahu J K 2010 In-situ solution doping technique for novel geometry rare-earth doped fibre fabrication, Conference on Lasers and Electro-Optics (CLEO) and Quantum Electronics and Laser Science Conference (QELS) 1-2

[4] Xia Z, Lunz M, Gerard V A, Gun'ko Y K, Lesnyak V, Gaponik N, Bradley A L 2011 Enhanced quantum efficiency in mixed donor-acceptor nanocrystal quantum dot monolayers *2011 13th International Conference on Transparent Optical Networks (ICTON)* 1-4.

[5] Coleman J J 2011 Quantum dot device *2011 37th European Conference and Exhibition Optical Communication (ECOC)* 1-24

[6] Jackson B R, Sazio P J A, Badding J V 2008 Single-Crystal Semiconductor Wires Integrated into Microstructured Optical Fibres *Advanced Materials* **20**(6) 1135-40

[7] Panniello, Ingrosso C, Coupillaud P, Tamborra M, Binetti E, Curri M L, Agostiano A, Taton D, Striccoli M 2014 Nanocomposites based on luminescent colloidal nanocrystals and polymeric ionic liquids towards optoelectronic applications *Materials* **7**(1) 591-610

[8] Strobel O, Lubkoll J 2010 Fibre-optic communication – An overview *2010 20th International Crimean Conference Microwave and Telecommunication Technology (CriMiCo)* 16-20

[9] Rahayu R, Damia Y 2011 Quantum size effect simulation and Ge composition on SiGe quantum dot for intermediate band solar cell applications *2011 The 2nd International Conference Instrumentation, Communications, Information Technology and Biomedical Engineering (ICICI-BME)* 321-5

[10] Urquhart P, Lopez O G, Boyen G, Bruckmann A 2007 Optical amplifiers for telecommunications, *IEEE International Symposium on Intelligent Signal Processing WISP* 2007 1-6.

[11] Menashe D, Shlifer A, Ghera U 2006 Optical amplifiers for modem networks *2006 International Conference on Transparent Optical Networks* **1** 115-8

Authors

Yanru Xue, Hebei Province, China

Current position, grades: lecturer.
University studies: Master degree at Ocean University of China in 2006.
Scientific interests: optical fibre communications.

Yinghua Yao, Hebei Province, China

Current position, grades: lecturer.
University studies: Master degree in Software Engineering at Beijing University of Technology in 2010.
Scientific interests: computer technologies.

Min Liu, Hebei Province, China

Current position, grades: lecturer.
University studies: Master degree in Software Engineering at Beijing University of Technology.
Scientific interest: computer technologies.

Feng Wang, Hebei Province, China

Current position, grades: lecturer.
University studies: Graduated from the Hebei University of Technology in 2006.
Scientific interests: optical fibre communications.

Cumulative Index

Mathematical and Computer Modelling

Jiansheng Xia, Shasha Dou The study on steel elliptical cup drawing based on finite element analysis
Computer Modelling & New Technologies 2014 18(10) 7-13

The sheet metal elliptical cup drawing is a complex process, Under the assumption of Prandtl-Reuss flow rule and von Mises yield criterion, the incremental Elasto-plastic large deformation finite element model was established based on the Updated Lagrangian Formulation (ULF).The Elasto-plastic conversions of boundary and deformation are reduced with r_{min} rule. The friction phenomenon of slippage and viscosity at the boundary interface is revised with increment of revision Coulomb rule. The increment rules are led into the whole stiffness matrix, and derived out the stiffness equation. The studies show that the influence on steel elliptical cup drawing deformation is influenced by die structure and parameter. The dates show that finite element simulation and experimental result have a good consistency.

Keywords: elasto-plastic, FEM simulation, elliptical cup drawing

Hongjuan Zhang, Long Quan, Yan Gao Dynamic modelling and small signal analysis of push-pull bidirectional DC-DC converter
Computer Modelling & New Technologies 2014 18(10) 14-18

Bidirectional DC-DC converter can not only act as a contact bridge between two different voltage levels systems, but also can achieve the energy flow in both directions. However, due to the operating characteristics of the active switch and diode in the converter, thereby the inverter becomes a strong nonlinear circuit. In light of this, the push-pull bidirectional DC-DC converter is designed and the method of state space small signal analysis is proposed. First of all, the state-space equations of the converter are established. Then each variable of inverter circuit is averaged in a switching cycle to eliminate the influence of the switching ripple. The respective average variable expressions are decomposed into the DC component and AC small signal component. After the DC component of the signal is eliminated, the expression of the AC small signal component is obtained, to achieve the purpose of separating the small signal. Finally, the expression of small signal component is linearized, thereby the nonlinear system is approximated the DC operating point. This study will provide the theoretical foundation ready for the analysis and design of converter controller.

Keywords: dynamic modelling, state space averaging method, small signal analysis, linearization

Feijiang Huang, Jun Yang, Xiaochun Lu, Qingxiao Shan, Yongbin Zhou, Jianyun Chen, Zhuli Hu An inter-satellite dynamic ranging algorithm based on two-way time synchronization
Computer Modelling & New Technologies 2014 18(10) 19-24

Inter-satellite precise ranging is the foundation for all aerospace application systems in realizing autonomous navigation. To acquire a high-accuracy inter-satellite range, this study investigates an inter-satellite dynamic ranging algorithm. Referring to the simulation of inter-satellite range variation rules in constellation, this study analyzes the negative impact of satellite motion on inter-satellite ranging and proposes corresponding improved methods to eliminate the major error caused by satellite motion. This algorithm solves the minimal error in inter-satellite range using a combination of inter-satellite range fitting polynomial and inter-satellite clock-offset fitting polynomial, both of which are generated by two-way time synchronization data. Simulation calculation results show that the accuracies of inter-satellite ranging can be controlled within 3m provided that simulation error is considered. The algorithm can be used to improve the accuracy of inter-satellite dynamic ranging of various aerospace application systems.

Keywords: aerospace application systems, autonomous navigation; inter-satellite communication, two-way time synchronization, inter-satellite dynamic ranging

Shuo Zhang, Yingzi Li Multi-agent simulation of partner selection behaviour based on matching degree in collaborative product development process
Computer Modelling & New Technologies 2014 18(10) 25-30

Collaborative Product Development (CPD) process is characterized by autonomous task control, dynamic task sequence, and frequent team collaboration, which endow the process with high flexibility and uncertainty. To make

the process predictable and improve process efficiency, it is essential to model, simulate, and analyse the process by considering all these characteristics. Our work focuses on studying the human working behaviours in CPD process by agent-based simulation, which we think is the main source of process uncertainty and flexibility. In this paper, the partner selection behaviours are studied under the frame of agent-based simulation. In the simulation, the design agent selects his partner according to matching degree including ability and character. The simulation results indicate that the proposed utility strategy can effectively shorten the project total time of the case.

Keywords: collaborative product development, partner selection, multi-agent simulation

Meng Kong, Zhong-xiang Zhang, Xiao-jing Kuang, Ming-sheng Chen, Xian-liang Wu Solving electrically-large objects RCS based on 3-D vector parabolic equation method
*Computer Modelling & New Technologies 2014 **18**(10) 31-34*

The vector parabolic equation (VPE) method is introduced to calculate bistatic RCS of three-dimensional (3-D) electrically-large objects and polarization effects are fully taken into account. According to an approximate form of the vector wave equation and divergence-free condition the VPE was derived in this paper. The numerical results conducted on the scattering from perfectly conducting cube show the VPE agree with the exact method, and computation time is acceptable compared with the traditional full wave method.

Keywords: vector parabolic equation, electrically-large objects, radar cross section

Yanhong Guo, Wei Liu Empirical research on the diffusion of home appliances to the rural areas in china based on Bayesian estimated bass model
*Computer Modelling & New Technologies 2014 **18**(10) 35-39*

Quantitative research on the diffusion of Home Appliances to the Rural Areas in China is a very important topic both in theory and in application. Most of the research on diffusion focuses on the discussion and qualitative analysis of the policy. This paper expands the Bass model applications by leading an empirical research based on the history sales data of Home Appliances to the Rural Areas. The model has been estimated by both linear and nonlinear least squares method and Bayesian method separately, and predicts the maximum potential market in the end. This empirical research shows that Bayesian method performs best in predicting the diffusion of Home Appliances to the Rural Areas. Finally, the main factors influencing the policy implementation was analysed. It predicts the periodical sales and gives some suggestions to the government and enterprises accordingly.

Keywords: home appliances to the rural areas, diffusion of innovation, Bass model, Bayesian method

Xuhui Zhang Dynamic multi-species coevolution large-scale optimize based on the fuzzy clustering and trust region
*Computer Modelling & New Technologies 2014 **18**(10) 40-47*

Based on the above-mentioned studies, this article put the modified Fuzzy Clustering method into the particle swarm optimization, which solved the curse of dimensionality existing in the conventional algorithms. The large-scale parameter optimization method applied the modified Fuzzy C-Means method to clustering the large-scale dimension under the coevolution frame and achieved the valid dimension grouping. Later on, the dynamic neighbourhood topology multi-species particle algorithms divided the whole species into packets and constructed the subspecies sharing neighbourhood information, which improve the searching efficiency. The bring-in trust region could have the self-adapt adjustment for the particle optimization range, accelerate the optimizing speed, and decrease the iterations in the dead space. We use 20 standard large-scale testing functions for simulation. Compared with the top-ranked tournament algorithm, the mentioned algorithm achieved a better optimizing result in most functions, which surely laded the foundation of the large-scale neural network parameter optimization and the application in the control system.

Keywords: multi-species coevolution, particle swarm optimization, guzzy clustering

Jian Li, Xiyong Wu, Long Hou Analytical solution on the sensitivity of matric suction profile in soil layer which is under the condition of one-dimensional steady flow
*Computer Modelling & New Technologies 2014 **18**(10) 48-51*

To simulate the suction profile within unsaturated soil layer, which is under the condition of one-dimensional infiltration, the point that water potential energy is mainly composed of gravitational potential energy and suction potential energy was used. And this paper also takes the standpoint that the relationship between these two variables is reciprocal. Combined with the mass conservation law, *Darcy*'s law and *Gardner* empirical equation, an analytical solution was got. This solution could be used to describe the suction profile within unsaturated soil layer when the seepage field of this soil layer reaches steady stage and the rainfall infiltrates to the soil layer along with the vertical direction. Several conditions with different rainfall intensities q, different inverse values of air-entry pressure α and different ratio values of q/k_s were examined by using this analytical solution, respectively. Compared with other factors, the rainfall intensity will take more influence on the suction profile, namely, the suction profile within unsaturated soil layer is more sensitive to the change of rainfall. The results obtained in this paper could be used as one useful reference for the research work about rainfall-induce landslide and the corresponding computer simulation.

Keywords: analytical solution, unsaturated soil, matric suction, sensitivity, steady seepage

Pu Han, Li Meng, Biao Wang, Dongfeng Wang Improved particle swarm optimization algorithm with unidimensional search
*Computer Modelling & New Technologies 2014 **18**(10) 52-57*

In this paper, a strategy of unidimensional search is introduced to particle swarm optimization (PSO). The global exploration capability of PSO is used to identify a promising region in search space. With the region as the starting point, a unidimensional local search is applied to search a more accuracy solution. The local search does not rely on the population information, which makes it can jump out of a local optimum when the population stagnates. With combination of global exploration and local exploitation, the algorithm can discover more favourable search area effectively and obtain a better solution. The improved PSO method is tested on eight benchmark functions. Experimental results show that the method can not only improve the accuracy of solution, but also reduce the influence of initial population distribution upon the algorithm performance. Finally, the influence of parameter variation on algorithm is analysed.

Keywords: particle swarm optimization, unidimensional local search, population initialization, population distribution

Chengfang Tan, Caiyin Wang, Lin Cui SMS text similarity calculation based on topic model
*Computer Modelling & New Technologies 2014 **18**(10) 58-62*

The traditional text similarity calculation is mainly based on the statistical method and the semantic method, it exists data sparse and high-dimensional problems and so on. In order to improve the ability of SMS text similarity calculation, this paper puts forward a kind of similarity calculation method based on topic model. By using LDA (Latent Dirichlet Allocation) to model SMS document set and inference parameter via Gibbs sampling algorithm. The topic-word probability distribution and document - topic probability distribution of the SMS document set are generated. Then use JS (Jensen-Shannon) distance formula to calculate SMS text similarity, finally perform the text clustering experiments on the similarity matrix by single-pass incremental clustering algorithms. Compared with traditional text similarity calculation method, experimental results show that this proposed method can obtain better F-measure, which proves the effectiveness and superiority of the proposed text similarity calculation method.

Keywords: SMS text, similarity calculation, topic model, text clustering, latent Dirichlet allocation

Qinglin Huang, Lixin Zhang, Chaoyang Sun, Xiang Zhang An anti-collision algorithm for adaptive search matrix of cotton seed traceability system based on RFID
*Computer Modelling & New Technologies 2014 **18**(10) 63-68*

Focusing on the tags collision problem of traceability management system for the cotton seed quality and safety based on RFID, we proposed an adaptive search matrix for anti-collision algorithm based on ABS in this paper. Meanwhile, the concept of collision stack is put forward to effectively lower the time of request and transmission of redundant information in the proposed algorithm. The theoretical analysis indicates that this new algorithm performs superiority to the ABS.

Keywords: anti-collision algorithm, adaptive search matrix, RFID, traceability

Min Zhang Default assumption reasoning based on fuzzy description logics
Computer Modelling & New Technologies 2014 18(10) 69-75

Fuzzy description logics (DLs for short) provide a convenient tool for dealing with inconsistency and uncertainty. People can infer with uncertain and incomplete information. According to the characteristics and requirement of the knowledge representation, fuzzy DLs can play an important role in the commonsense reasoning. Default rules express concise pieces of knowledge having implicit exceptions, which is appropriate for reasoning under incomplete information. Default assumption reasoning based on fuzzy DLs is proposed. Possibility theory is used for representing both uncertainty and defeasibility. Inference service is considered in the logic and algorithms are provided for it.

Keywords: fuzzy description logics, default assumption reasoning, fuzzy reasoning

Wei Zhang, Qiu-li Wu, Yu-rong Deng, Ze-cheng Lv Case-based reasoning adaptive optimization algorithm for power transformer fault diagnosis
Computer Modelling & New Technologies 2014 18(10) 76-81

The adaptive learning rate for the introduction of case-based reasoning transformer fault type identification. The adaptive learning rate theory, through improved data normalization, typicality and best filtering diversity to extract the original example and optimal neural network. In the sample processing and analysis process to be solved according to the type of fault feature automatically adjusts the data processing methods, processes, boundary conditions and constraints to adapt statistical distribution, the probability characteristics. Examples show that this method can overcome the DGA data ambiguity and dispersion problems in the recognition accuracy and convergence speed advantage.

Keywords: Adaptive, Case-Based reasoning, neural network, fault type, normalization, data filtering

Huaping Zhou, Xutong Zhang The model construction and implementation of discrete physical system in industrial CPS
Computer Modelling & New Technologies 2014 18(10) 82-89

Cyber physical system referred as CPS, or information physical system, achieves the integrated coordination of the cyber world and the physical world. The collaborative system is more reliable and efficient. By analyzing the cyber physical system of integrated sense and control, the constructing approach of the discrete physical model is proposed and applied to mine surface production systems for CPS of the combination of discrete-time input and output and control functions. By modelling, analyzing and verifying the system, the cyber system and the physical system achieve deep fusion. On this basis, trusted software design, formal description methods and reasoning theory are established in the cyber physical system. It provides the authentication methods for the trusted software analysis and modelling of discrete physical system in the cyber physical system. The coal surface production system model is established by the model construction method of discrete physical systems in CPS proposed in this paper. The technology of computer, computer network and embedded systems is adopted to build the information world and achieve the integrated design and collaborative control of the subsystem in CPS. It is applied in a coal mine in the Huaibei Coal Mine Shares Limited Company and achieves good results. To speed up the wide applications to industries and enterprises of the CPS, a useful exploration is carried out in the paper.

Keywords: information world, physical world, discrete physical system, information fusion, credible design

Zhiyong Zhang, Hongli Guo, Lvwen Huang, Xiaoting Zhang Research on adaptive H-Infinity tracking for inhibition fluttering of picking robot arm
Computer Modelling & New Technologies 2014 18(10) 90-98

This paper aim at solve issue that conventional frequency domain theory is unsuitable for MIMO system and LOG theory is unsuitable for model perturbation, we commit to the research on H-infinity stability and focus on discuss design principles of H-Infinity stability system and methods on Riccati equation or inequality solution. As for the uncertainty and external interference, we discuss two-output equation or an algebraic Riccati state feedback equation using character on Riccati equation. According to the state space theory, we derive the controller to make structural equation to meet the requirement of state feedback and observer, and then draw suboptimal solutions in the form of

engineering management experience. To reduce the impact on interference to control stability by selecting the appropriate interference attenuation coefficient γ, so that its stability could be casted to meet harvest scene.

Keywords: H-infinity, Riccati equation, Interference attenuation coefficient, Domain theory

Yuliang Cong, Shuyang Zhang, Limin Xu, Lili Sun Spectrum allocation algorithm based on user requirements under the circumstance of advanced user existence
*Computer Modelling & New Technologies 2014 **18**(10) 99-103*

Since wireless spectrum is a non-renewable resource, how to improve the spectrum utilization is always the problem to be resolved by wireless communication technology. With the development of wireless communication technology, the contradiction between supply and demand for spectrum resource has been more and more intense. In this case, cognitive radio technology emerged. The traditional list-colouring algorithm aims at maximizing the number of allocated bandwidth; CSGC algorithm (Colour Sensitive Graph Colouring algorithm) is to achieve the maximum benefits of bandwidth for the cognitive users; local bargaining algorithm is an improved algorithm on time complexity based on CSGC. However, the three algorithms do not take the bandwidth demand of cognitive users into consideration. Even with the proportion allocation of CSGC algorithm, the problem of ill-considered for the bandwidth demands of cognitive users also exists, which results in irrational allocation of spectrum resources. To solve this problem, this article proposes a priority order with the consideration of cognitive users in spectrum allocation based on advanced users.

Keywords: Cognitive Radio, Spectrum Allocation, Graph Colouring, User Needs

Zhi Liu Approximate completed trace equivalence of real-time linear algebraic Hybrid Automata
*Computer Modelling & New Technologies 2014 **18**(10) 104-108*

In allusion to design simpler software system, the paper proposes approximate completed trace equivalence of real-time linear algebraic Hybrid Automata. Firstly, it pulls real-time linear algebraic program into Hybrid Automaton and establishes real-time linear algebraic Hybrid Automaton. Next, it uses matrix Frobenius norm to analyse approximation of real-time linear algebraic Hybrid Automata. Afterwards, it gets approximate completed trace equivalence of real-time linear algebraic Hybrid Automata. Finally, the Email virus spreading automata example shows that approximate completed trace equivalence of real-time algebraic Hybrid Automata can simplify automaton.

Keywords: Hybrid Automata, approximate, completed trace equivalence, algebraic program

Lei Huang, Zhi Liu Completed trace equivalence of inhomogeneous linear algebraic Hybrid Automata
*Computer Modelling & New Technologies 2014 **18**(10) 109-113*

In order to reduce states of inhomogeneous linear algebraic Hybrid Automaton, the paper proposes completed trace equivalence of inhomogeneous linear algebraic Hybrid Automata. Firstly, it introduces inhomogeneous linear algebraic programs into Hybrid Automata and establishes inhomogeneous linear algebraic Hybrid Automata. And then, it uses mathematical computation and completed trace equivalence to get completed trace equivalence of inhomogeneous linear algebraic Hybrid Automata. Finally, the travel queue automata example shows that completed trace equivalence of inhomogeneous algebraic Hybrid Automata can reduce states.

Keywords: Hybrid Automata, completed trace equivalence, algebraic program

Xiangyang Chen Local reconstruction and local fisher discriminant based semi-supervised dimensionality reduction algorithm
*Computer Modelling & New Technologies 2014 **18**(10) 114-120*

Local reconstruction and global preserving based semi-supervised dimensionality reduction (LRGPSSDR) algorithm gives no consideration to data locality when processing intra-class relationship and class relationship. Enhanced semi-supervised local fisher discriminant analysis algorithm (ESELF) also neglects locality of data manifold structure when maintaining data manifold structure. To address these problems, the local reconstruction and local fisher discriminant based semi-supervised dimensionality reduction (LRLFSDR) algorithm was proposed in this paper. It depicts significance of sample distance with an improved thermonuclear weight. In this way, intra-class relationship and class relationship of the same cluster attracts more attentions, thus enabling to shorten or widen intra-class distance or class distance firstly. Moreover, it uses idea of LLE algorithm to make neighbourhood linear reconstruction relationship of

each point in low-dimensional space to be similar with that in high-dimensional space, which takes locality of data manifold structure into account. Test result confirmed that the proposed LRLFSDR algorithm is superior to other semi-supervised dimensionality reduction algorithms in classifying standard libraries like COIL20, Extended YaleB and CMU PIE.

Keywords: local fisher discriminant, local reconstruction, semi-supervised learning, dimensionality reduction

Changhong Yan, Qin Dong, Hong Wang DDoS attacks defence strategies based on nonparametric CUSUM algorithm
*Computer Modelling & New Technologies 2014 **18**(10) 121-125*

In the Internet network attacks, distributed denial of service (DDoS) has aroused world attention because of its destructive power. It seems particularly difficult to defend against DDoS attacks for they have characteristics such as abrupt attacks, attacking host computer in a very wide distribution, and so on. To guard against network security and defend distributed denial of service attacks (DDoS), research should begin from the detection of DDos attacks. On the basis of deep research of DDoS attacks, the thesis summarizes and analyses the mechanism and principles of intrusion detection firstly. This paper starts with the analysis of the principle of DDoS attacks. Followed by inquiry and analysis of data packet of DDoS attacks detection, the thesis gives out the computation method for detecting DDos attacks based on Flow Connection density and presents a defending model against DDos attacks based on the temporal series of Flow Connection Condensity (Density). With the defending module based on the temporal series of Flow Connection Condensity (Density), data packet can be effectively filtered so that DDos attacks can be effectively defended and prevented. Finally, experiments prove that the module can effectively filter data packet from network.

Keywords: network security, distributed denial of service, flow connection density, time series, defence strategies

Kai Zhang, Tingsong Du, Tianbo Wang, Wenqing Liu Optimization model of power system unit commitment allocation problem considering the value-point effect and its simulation analysis
*Computer Modelling & New Technologies **18**(10) 2014 126-131*

Based on the studies of the allocation problem of large-scale unit commitment in the power system, a mathematical optimization model is established involving the valve point effect of unit commitment. The optimal solution obtained from the method that the standard artificial fish swarm algorithm (AFSA) is applied to the commitment allocation problem of three-units improves the result recently reported in literature. Considering the visual selection of AFSA affects foraging, huddling and other activities and convergence performance much when increasing the unit size, the proposed improved artificial fish swarm algorithm(IAFSA) in this paper uses the linear decreasing vision function instead of the fixed vision. It can speed up the convergence, jump out of local convergence effectively, and obtain a global optimal solution. Finally, the simulation comparing experiment is conducted for commitment allocation problem of ten-units. The simulation result shows that the IAFSA not only improves the convergence but also enhances the global search capability.

Keywords: unit commitment, valve point effect, linear decreasing

Information and Computer Technologies

Liang Li Software development for water quality's monitoring centre of wireless sensor network
*Computer Modelling & New Technologies 2014 **18**(10) 132-136*

Water quality's monitoring centre software of wireless sensor network is designed through applying C#.net and SQL database technology. The monitoring and querying of sensor node data is realized through adopting C/S (Client/Server) pattern. Standard structured query language (SQL) and ADO.NET database access technology are adopted to realize the rapid operation and efficient management on database. Graphical interfaces could display the topology and node status of sensor network, as well as the real-time and history parameters collected by each sensor node and so on. The practice has proved that this software could satisfy the data collection, as well as monitoring and management requirements of wireless sensor network monitoring system.

Keywords: wireless sensor network, water quality parameters, database, c#.net

Zhicheng Chen Automation control and design application of factory sewage disposal system
*Computer Modelling & New Technologies 2014 **18**(10) 137-140*

Intellectualization and automatization become the key technology of sewage disposal for effective realization of sewage disposal technology, continuous and stable operation of system and water quality standard. This paper took a full consideration on the function of online instrumentation in autonomous system and confirmed monitoring system scheme of three-in-one network (data, video and voice) and PLC control scheme adopting means of centralized management and decentralized control combining with sewage disposal technology. Factory adopts Ethernet looped network of optical fibre industry as backbone network for communication and located monitoring system at the central control room of comprehensive office building in sewage disposal factory. It realizes collection and monitoring of procedure parameter, analysis and processing of data, remote monitoring of core equipment. Coarse screen system subprogram, fine screen system screen, immersible pump subprogram, return sludge system program and return sludge system subprogram were detailed described according to technology requirement and design principle of PLC substation. It also discussed test quality guarantee of autonomous system.

Keywords: Sewage disposal, autonomous system, data communication, PLC system

Chenxiang Zhang Research on computer information integration based on some wireless sensor network model
*Computer Modelling & New Technologies 2014 **18**(10) 141-146*

Network model in wireless sensor integrates wireless communication technology, sensor technology and embedded computer technology. This new computer pattern is the joint elements in current network hot technology. Self-organizing feature Map is also termed as SOM network. It can solve conception problem, which can only be fulfilled by human brain nervous tissue. We integrated computer information applying SOM and wireless sensor technology to realize information integration model. We made a conclusion of the characteristics of wireless sensor network model and studied its protocol architecture. We induced information integration system into software engineering and developed it into application in all industries and areas in the perspective of definition, structure, classification and calculation of information integration. This paper integrated information's, reduce redundancy of information, decrease energy consumption and lengthen the service time of network by SOM wireless sensor network model. It also made a analytical research of case on information integration.

Keywords: wireless sensor, SOM network, information integration, network model, evaluation

Lingqiang Ran, Xiangxu Meng Example-based geometric texture synthesis: a survey
*Computer Modelling & New Technologies 2014 **18**(10) 147-150*

3D object modelling is a key step in computer animation industry. How to generate models rich in high quality geometric details is still a challenging task and affects the final visual effects of animation. Example-based geometric texture synthesis (EGTS) is a powerful tool to automatically build models with rich geometric details. Given a small patch of geometric texture as an example and an arbitrary model as a target model, using EGTS, we can change the surface of the target model to the new geometric texture style. EGTS has been studied a lot by the literature and many papers have been published. In this paper, we will summarize the influential papers and present our thoughts on this topic.

Keywords: example-based, geometric texture synthesis, survey

Xiaoyan Gao Two duality problems for a class of multi-objective fractional programming
*Computer Modelling & New Technologies 2014 **18**(10) 151-157*

In this article, we investigate the duality results for a class of non-differentiable multi-objective fractional programming problems. The parametric dual models and Wolfe dual models are formulated for this fractional programming. Weak, strong and strict converse duality theorems are established and proved based on the generalized invexity assumptions. Some previous duality results for differentiable multi-objective programming problems turn out to be special cases for the results described in the paper.

Keywords: invexity, multi-objective fractional programming, parametric dual, Wolfe dual

Zhuo Shi, Yinghui Li, Ke Yu, Yuanquan Cheng, Changshao Zhou Research of hand gesture using Kinect based on finger recognition
*Computer Modelling & New Technologies 2014 **18**(10) 158-162*

Using the depth image from Kinect, we present a novel method to archive the Hand Gesture Recognition (HGR). Firstly we overview the major technical components of the complete method. Then we elaborate several key challenges such as palm recognition, contour analysis and fingertip detection. We present the Biggest concave point detect schemes to contour analysis which is the key to achieving the HGR. Finally, we demonstrate the feasibility and effectiveness of the method by running the system.

Keywords: human-computer interaction, hand gesture recognition, finger recognition, Kinect

Zhenyu Zhang, Yong Qin Test method of railway video surveillance system
*Computer Modelling & New Technologies 2014 **18**(10) 163-167*

Test method of railway video surveillance system was mainly explored in this paper. In order to evaluate accurately railway video surveillance system, based on railway video surveillance system test platform, an objective, direct and effective test system was established for testing equipment requirements, system functionality and performance. In addition, it provides standard of railway video surveillance system and makes field of video surveillance standardization.

Keywords: railway, video surveillance, test method

Wang Li An Improved model of product design case reuse based on extension theory
*Computer Modelling & New Technologies 2014 **18**(10) 168-174*

This article studies the question of product design case reuse based on extension theory for complex product with various information, categories and attributes. It also proposes to use case space elementary system and case subject index elementary model based on elementary model of product design case research to describe complex design case. Besides, it also builds design case research and reuse model based on extension correlation function to acquire the most similar design result. The application of this method can acquire the design case closest to the design objective rapidly, thus improving the efficiency of product configuration, and shortening design period of complex product. Finally, it testifies the effectiveness of the model with application cases.

Keywords: Product Design, CBR, Design Reuse, Extension Theory, Artificial Intelligence

Shunliang Huang Definability of concept in incomplete information systems
*Computer Modelling & New Technologies 2014 **18**(10) 175-179*

An incomplete information table (a set) can be expressed as a family of complete information tables (sets). The family of complete information sets maybe constructs an interval set. A concept in incomplete information situation, called partially known concept, is said to be definable if its extension can be expressed as an interval set. The new definition of definability proposed in this paper, named interval definable, is different from its usual meaning in the rough sets theory where a concept is definable means that its extension is a definable set, which is the union of some equivalence classes. The new definition of definability not only provides a new interpretation of interval sets, but also endows more general meaning and deeper understanding of definability.

Keywords: incomplete information, definability, concept representation, interval sets, uncertainty

Wenqing Huang Research on multi-virtual queue based on flow estimation
*Computer Modelling & New Technologies 2014 **18**(10) 180-187*

To solve the equity issues in the network bandwidth competition of TCP/UDP mixed flow, we have raised an AQM algorithm framework based on "granularity", which is aimed to realize the divide-and-rule goal of the conservative TCP and the greedy UDP so as to alleviate the equity issues of the network resources caused by different protocols by designing a common framework. Furthermore, under the AQM framework of "granularity", we have brought forth Multi-Virtual Queue based on Flow Estimation (VFQ) and we have also given the detailed implementation schemes for different modules of VFQ algorithm. In order to further verify the feasibility and validity of VFQ algorithm, we have designed 5 computer network congestion scenarios in AQM network simulation platform and made simulation

comparisons with the classical ARED, PI and Blue algorithms. Numerous simulation results demonstrate that in contrast with other AQM algorithms, VFQ can better adapt to various sudden network congestion scenarios and it has better response speed and queue management performance.

Keywords: multi-virtual queue, flow estimation, TCP/UDP mixed flow

---

Yang Shi, Huipeng Li, Xianmu Li, Lu Wang Application of federated particle filter to SINS-GPS/BDS integrated navigation system
*Computer Modelling & New Technologies 2014 **18**(10) 188-191*

When integrated navigation information is filtered, there may be non-linear sub-filter. The paper proposes federated particle filter. It is based on the framework of federated Kalman filter and uses the method of particle filter to process non-linear sub-filter, which enhances adaptability of federated Kalman filtering model. The paper applies federated particle filter to SINS-GPS/BDS integrated navigation system to establish filtering model. The simulation is made to verify the effectiveness of federated particle filter.

Keywords: federated Kalman filter, particle filter, integrated navigation

---

Liming Wu, Wei Han, Songbin Zhou, Xin Luo, Yaohua Deng A compressed-domain audio fingerprint algorithm for resisting linear speed change
*Computer Modelling & New Technologies 2014 **18**(10) 192-196*

Existing compressed-domain audio fingerprint algorithms have been able to be used to recognize audio information effectively according to hearing content, and are robust to common time-frequency domain distortion, including echo, noise, band-pass filtering, 32Kbps@MP3 and so on. However, they are poor in resisting linear speed change, which is a very common method for audio processing. In this paper, we propose a novel compressed-domain audio fingerprint algorithm. It is robust to large linear speed change via using auto-correlation function to reduce unaligned degree of MDCT spectrum sub-bands' energy. Besides, it is similar with existing compression-domain audio fingerprint algorithms on the other aspects.

Keywords: compressed-domain audio recognition, audio fingerprint, linear speed change, robustness

---

Shoushan Liu, Maoyong Cao, Yan Chen System model of reconfigurable embedded motion control system based on IEC61499
*Computer Modelling & New Technologies 2014 **18**(10) 197-202*

Based on the function block (FB), and distributed system model of IEC61499 standard, the application and system model of embedded and reconfigurable motion control system are modelled. According the motion control system architecture built by embedded microprocessor and reconfigurable logic devices, the function block of IEC61499 can modified to describe the function unit which being used as the basic element to construct the application. Based on the modified function units, in combining with the distribute resources of reconfigurable logic devices, the distributed application model of the motion control system can be configured. To testify the application model, the interpolation process based on the digital differential analyser (DDA) of computer numerical control (CNC) system is carried out, and the results show that the application model is well performed to describe the embedded reconfigurable motion control system.

Keywords: IEC61499, function block, embedded reconfigurable system, motion control, function unit

---

Jie Zhang, Fuli Song, Jiali Tang Identification of crop weed based on image texture features
*Computer Modelling & New Technologies 2014 **18**(10) 203-206*

By using computer image processing technology, texture features of weed in the corn seedling field are analysed, and then we present an algorithm combining Support Vector Machine (SVM) to form a classifier and promote an improved method of RBF network. The experimental results show that the proposed method is effective.

Keywords: crop, weed identification, image texture features, support vector machine (SVM)

Xiang Zhang, Wei Zhang, Xiaoling Xiao Rapid detection of bedding boundaries based on borehole images
*Computer Modelling & New Technologies 2014 **18**(10) 207-211*

The bedding is an important sedimentary structure phenomenon. The rock bedding structure, the direction of

sedimentary transportation and the ancient sedimentary environment analysis can be studied by extracting the bedding boundaries and dips. Electric imaging logging can provide rich information of a borehole wall and circumference, which reflects formation resistivity variations. The bedding boundaries are detected by using the electrical imaging logging data based on an image recognition method in this paper. On an oriented, unwrapped image of a cylindrical borehole, the trace of a planar-bedding boundary appears as a sine wave. The bedding boundaries are detected by the recognition of the sine curves in borehole image. The influence problems of bedding boundary detection caused by fractures and other geological events are solved by statistical analysis technology. Through the techniques of the slope fitting, the speed and accuracy problems of bedding boundary detection are solved, which has good anti-interference performance. The processed results of the theoretical models and the measured borehole images at the varied dip segment indicate that the detected bedding boundaries reflect the real situation, which are identical to those derived by the Autodip.

Keywords: imaging logging, bedding boundary, object detection, image recognition

Xiaowei Hu, Jiexin Wang, Xuemiao Xu, Biao Zhou Moving vehicle detection algorithm based on motion edge extractor
*Computer Modelling & New Technologies 2014 **18**(10) 212-218*

In this letter, we propose a moving vehicle detection method based on motion edge extractor (MEE). In the course of the vehicle detection, the motion and contour information are prominent, so we combine them together to extract the moving objects with complete contours. We first modify the Gaussian Mixture Model (GMM) to estimate the background more precisely. Then an object extraction method in static image is proposed. The original image and the background image go through the object extraction method and a series of logical operations to get the moving areas. At last we apply a simple filling method to refine the result and accurately extract the vehicle areas. The experiment result shows that our algorithm is not sensitive to illumination and can detect the vehicles with similar colour to the road robustly.

Keywords: vehicle detection, motion edge extractor, video monitoring

Lei Zheng, Defa Hu A resource schedule method for cloud computing based on chaos particle swarm optimization algorithm
*Computer Modelling & New Technologies 2014 **18**(10) 219-223*

In order to improve the cloud computing resource scheduling efficiency, this paper proposed a method for cloud computing resource schedule based on chaos particle swarm optimization algorithm. Firstly, the resource scheduling options were taken as the position of the particle, and resource load balancing was taken as the objective function. Then the optimal resource scheduling solution was obtained by sharing information and the exchange of particles, while chaos mechanism was introduced to guarantee the diversity of particle swarm to prevent appearing premature convergence and local optimal solution. The simulation test was carried out in CloudSim platform, and the results show that the proposed method can quickly find the optimal scheduling solution for cloud computing resources and improve the efficiency of resource, and the method has better practicability and feasibility.

Keywords: cloud computing, resource scheduling, chaotic particle swarm optimization algorithm

Yulong Huang, Benyue Su, Jianqing Xi CUBPT: Lock-free bulk insertions to B+ tree on GPU architecture
*Computer Modelling & New Technologies 2014 **18**(10) 224-231*

B+-tree is one of the most widely-used index structures. To improve insertion process, several batch algorithms are proposed, which all use one thread to complete one node insertion and cannot make full use of GPU's parallel throughput. So, a batch building and insertion method on GPU named CUBPT is proposed in this paper. During the process of bulk building and insertion, CUBPT use one thread to insert one key, which can maximize the performance by GPU. The experimental results show that when build a 10M tree, the overall performance of CUBPT improved 25.03 times compare with four threads PBI. When insert 10M uniform keys into a 10M tree, the overall performance of CUBPT improved 13.38 times compare with four threads PALM; when insert 10M highly skewed keys into tree with same size, the overall performance of CUBPT improved 15.23 times compare with four threads PALM.

Keywords: in-memory B+-tree, bulk build, Lock-free batch insertion, GPGPU

Ping Yan, Teng Lv, Weimin He Probabilistic XML functional dependencies based on possible world model
Computer Modelling & New Technologies 2014 **18**(10) 232-238

With the increase of uncertain data in many new applications, such as sensor network, data integration, web extraction, etc., uncertainty both in relational databases and XML datasets has attracted more and more research interests in recent years. As functional dependencies (FDs) are critical and necessary to schema design and data rectification in relational databases and XML datasets, it is also significant to study FDs in uncertain XML datasets. This paper first proposed XML functional dependencies (XFDs) of deterministic XML dataset based on tree tuple models. Then two new kinds of functional dependencies based on possible worlds model for probabilistic XML dataset are introduced: probabilistic XML functional dependencies (pXFDs) and probabilistic approximate XML functional dependencies (pAXFDs). pXFDs extend the concept of XFDs of deterministic XML dataset by considering the probability of each possible world of probabilistic XML dataset, and pAXFDs extend the concept of probabilistic XML functional dependencies of probabilistic XML dataset by considering the degree of truth of tree tuples in each possible world of probabilistic XML dataset.

Keywords: uncertain XML, functional dependency, inference rule, closed set

Yu Bengong, Wang Liu, Guo Fengyi Security evaluation model for the enterprise cloud services based on grey fuzzy AHP
Computer Modelling & New Technologies 2014 **18**(10) 239-244

This paper analyses the application status of cloud services to identify four factors that affect the security of enterprise cloud services (ECSs), including platform facilities, operational safety, operations management, and legal factors. Based on the four factors, the grey fuzzy analytic hierarchy process (GFAHP) is used to construct an evaluation model for the security of ECSs. An example is investigated to demonstrate the proposed model.

Keywords: enterprise cloud services, cloud safety, grey fuzzy AHP, evaluation model

Liangong Song, Ke Han A novel distributed network database mapping scheme analysis
Computer Modelling & New Technologies 2014 **18**(10) 245-251

This research is based on distributed client-server model, and intends to add lacking fields in the target database by mapping in order to complete the database for subsequent analysis and tab the maximum economic value of a single database. This research takes China Industry, Commerce and Service Census Database (TTICSCDB) as the target database, and China Technology Innovation Survey Database (TTTISDB) as the assistant database. Then this research will employ techniques to mapping common fields from the assistant database to the target database and evaluate the accuracy of mapping. We will prove from the statistical perspective the simulation and accuracy of extension field of the network database proposed in this research. This network database will present its advantages in many ways.

Keywords: database, mapping, statistics, expanding

Jingbin Hao, Liang Fang, Haifeng Yang An improved boundary extraction method of STL model based on edge curvature estimation
Computer Modelling & New Technologies 2014 **18**(10) 252-258

To efficiently extract feature boundaries of the STL model, an improved method is proposed based on edge curvature estimation. Three curvature parameters (dihedral angle, perimeter ration and convexity) are used to estimate the surface curvature information of the STL model. Genetic Algorithm (GA) is used to determinate the threshold of feature edges. The extracted feature edges are grouped and filtered using the best-fit plane (BFP), which is calculated by Least Square Method (LSM). The Dijkstra's algorithm is used to close the incomplete feature boundaries. Several experimental results demonstrate that the amount of feature edges is significantly reduced, and useful feature edges are reserved to construct feature boundaries. The improved boundary extraction method has important significance in decomposing large complex STL models meaningfully.

Keywords: boundary extraction, curvature estimation, STL model, genetic algorithm, least square method

Tao Guo, Zhengqi Liu Application of TV image compression technology based on neural network

Aiming at the disadvantages of digital TV, including a large amount of information and redundant information, a method of TV image compression technology based on neural network combining neural network with image compression technology is proposed in the work. Firstly, the TV image is divided into blocks as the input of neural network to build the network; secondly, the blocks are rebuilt to realize image compression recovery. The simulations show that the neural network algorithm can achieve the TV image compression effectively and the number of neurons of the hidden layer based on the neural network algorithm has great influence on the building and training of the network by contrast. When the number of neurons of the hidden layer is less, the image compression ratio will be higher and the image compression quality will be lower.

Keywords: digital TV technology, image compression, neural network, hidden layer, compression ratio

Zhaozhun Zhong, Pengjie Qi, Miao Guan, Yuedong Xia, Yuanhui Fu Fabric defect detection system based on digital image processing

A fabric defect detection system based on digital image processing for textile fabric is proposed in this paper. The approach for the classification and identification of three commonly encountered classes of fabric defects (holes sto, missing end and mispick) is studied. The developments of both the hardware and software structures are presented. Firstly, median filter preprocessing and image segmentation based on Otsu threshold are applied to localize the fabric defects. Secondly, the features based on grey-level histogram and geometry are extracted. Thirdly, the classification and identification are accomplished by the method of artificial neural network based on the extracted features. Finally, a variety of textile images with different defects are tested to evaluate the performance of the proposed defect detection system. The experiment results indicate that the proposed system works efficiently with high accuracy, which can meet the requirements of the textile industry.

Keywords: fabric defect detection system, digital image processing, image segmentation, defects classification

Shuanglin Huang, Jianjun Tan Joint resource allocation based on Nash bargaining game for wireless cooperative networks

This paper considers the problem of resource sharing among selfish nodes in wireless cooperative networks. In the system, each node can be acted as a source as well as a potential relay, and both nodes are willing to achieve an extra rate increase by jointly adjusting their channel bandwidth and power levels for cooperative relaying. Nash bargaining solution (NBS) is applied to formulate the JBPA problem to guarantee fairness. Simulation results indicate the NBS resource sharing is fair and the fairness of resource allocation only depends on how much contribution its partner can make to its rate increase.

Keywords: Nash bargaining solution, bandwidth allocation, power control

Xiaoyu Zhang, Ning Liu Adhesive image segmentation based on watershed algorithm

Since adhesive image has a deficiency of over-segmentation, the work applied watershed algorithm, whose merits include quick computing speed, closed outline and accurate location, to adhesive image segmentation. In the first few sessions, the basic principle and arithmetic steps of watershed algorithm were illustrated in detail. On this basis, following sessions correspondingly introduced the MATLAB simulation analysis and verification. Through simulation and contrast of the segmentation results by different algorithms, it could be concluded that watershed algorithm, the most effective method, could help split adhesive objects into single ones, while greatly reducing or even eliminating the over-segmentation phenomenon.

Keywords: Image Segmentation, Edge Detection, Threshold Segmentation, Watershed Algorithm, Image Processing, MATLAB Software

Ning Liu, Xiaoyu Zhang An improved sift image matching detection

Aiming at the problems - edge response in the traditional SIFT descriptor and the insufficient correct matching feature points, the work proposed a kind of improved SIFT Image Matching Detection Algorithm. The candidate key point

was firstly detected by the SIFT algorithm; Canny edge detection algorithm was used to detect image edge points; it was judged whether the candidate key point needed to be eradicated by comparing whether the candidate key point equals to the coordinates of edge points; K-means clustering pattern, which is combined by the vector space cosine similarity and vector Euclidean distance similarity, was adopted to perform global image similarity matching. Finally, RANSAC algorithm was used to further get rid of the wrong matching. The experimental result indicates the improved method greatly enhances the stability of SIFT algorithmic and the accuracy rate of matching.

Keywords: SIFT operator, Canny operator, K-means cluster, accuracy rate, key point, edge detection

Changyou Guo, Xuefeng Zheng, Jianjun Liu Uncertain random fault tree analysis based on cloud security protection framework
*Computer Modelling & New Technologies 2014 **18**(10) 288-295*

Based on uncertainty theory and chance theory, this paper proposes a method that constructs and analyses fault tree. The fault tree is constructed based on logical relations between bottom events. Fault rate of bottom event would be characterized as random variable if it is obtained from historical data, it would be characterized as uncertain variable if it has no statistical data but is obtained from expert's subjective judgment. The chance that top event occurs is an uncertain random variable. The minimal cut set of fault tree is obtained by Boolean algebra method and at same time, the simplest standard disjunction expression of top event is obtained. This paper also constructs hybrid simulation algorithm to calculate the chance that top event occurs. At last validity of this method is confirmed by taking cloud security protection framework risk fault tree as example.

Keywords: Fault tree, Chance theory, Uncertain random variable, Cloud Computing, Cloud security

Operation research and decision making

Guirong Guo, Yan Lu Design and analysis of hotel management system based on information technology
*Computer Modelling & New Technologies 2014 **18**(10) 296-301*

With the rapid development of China's economic and tourism, competition in the hotel industry has become intense. Only by improving the means and methods of their management and improve their level of service constantly, can they receive adequate and healthy development. Therefore, the hotel operation computer management has become a priority task. These articles give a brief introduction to hotel management system design, in accordance with means of software engineering to do feasibility analysis, requirements analysis and design. The whole system divided into modules separately and introduce the function, while gives a set of criteria and the logical structure of the database of database management. These analyses and the guidelines basically meet the design requirements of hotel management system based on information technology, so as to improve the service quality and efficiency of the management of the hotel.

Keywords: hotel management, feasibility analysis, requirements analysis, database, logical structure

Jiahua Zhou GIS technology integration design based on university culture resource
*Computer Modelling & New Technologies 2014 **18**(10) 302-305*

3-D GIS technology has become the hot research in current GIS realm. It is applied in various industries. This paper constructed open and interactive 3-D data campus based on Sketch Up applying ArsGIS 3-D space platform and combining traditional information management system and 3-D GIS technology. We integrated various system resource management modules in campus and fulfilled an open campus project, which has strong inquiry and space analysis function. We turned 3-D model data into ArcGIS supportable format and imparted it into digital campus development structure. Browsing system of multi-detail and multi-level scene is also applied. Finally, we constructed a lively 3-D digital campus culture design that has interaction form.

Keywords: GIS technology, open type, interaction type, mutual integration

Yun Bai, Pu Wang, Jingjing Xie, Chuan Li An analysis model of urban water supply quality based on extension classification method
*Computer Modelling & New Technologies 2014 **18**(10) 306-312*

Directing at the multiple factors and levels in the complex problem of water consumption, this paper studied and analyzed the quality of urban water supply and put forward an analysis model of urban water supply quality based on

Extension classification method. Based on the Extension Theory, it modelled relative factors of analysis of urban water supply quality by matter-element method. Than classical domain and joint domain of relative factors of analysis of urban water supply quality were established. Via extension correlation function, the mapping relation between the water supply condition of the urban as research subject and the existing grades of urban water supply condition was established. Based on extension correlation, the grade of water supply condition of the present urban could be obtained. Thus, this grade can offer guidance for water supply of the urban and support for water supply strategy. Finally, by a case study, the model and algorithm were tested and proved feasible and operable.

Keywords: urban water supply, water consumption, extension theory, extension classification, model

Fasheng Yi, Xiaoling Li, Jimin Yuan New logistics distribution route dispatching mode based on genetic algorithm-ant colony algorithm
*Computer Modelling & New Technologies 2014 **18**(10) 313-318*

For multi-target route optimization with constraint conditions, the mathematical model for logistics distribution route optimization is built to accelerate response speed of logistics enterprises to customers, improve service quality, and strengthen the satisfaction of customers, and a new algorithm with the combination of genetic and ant colony algorithms is proposed to solve the selection issues of such logistics route. Initial pheromone is formed with genetic algorithm, based on which the optimal solution is rapidly sought with ant colony algorithm, and complementary advantages are achieved between above two algorithms. Application examples and simulations are available for calculation, and the results show that such algorithm is practical and effective to optimize logistics distribution route.

Keywords: logistics distribution, genetic algorithm, ant colony algorithm, combination, optimization

Tao Yi, Yunfei Zhang, Weichun Shen Research on the management of project cost data based on BIM
*Computer Modelling & New Technologies 2014 **18**(10) 319-323*

This article first analyses the differences between the BIM application mode and the traditional mode in the cost data management. With the application of BIM (Building Information Modelling), this article analyses the relationship between the standard framework of BIM and the whole process of cost management, then points out that the BIM application mode has changed the whole process of cost management and refines what the BIM standard framework should reflect about the whole process of cost management.

Keywords: BIM, cost data management, construction stage, participants in project

Li-ping Sun, Yong-long Luo, Xin-tao Ding, Fu-long Chen Spatial clustering algorithm with obstacles constraints based on artificial bee colony
*Computer Modelling & New Technologies 2014 **18**(10) 324-328*

Spatial clustering is one of practical data mining technique. In this paper, artificial bee colony (ABC) is used for clustering algorithm, which aims to optimally partition N objects into K clusters in obstacle space. The ABC algorithm used for clustering analysis with obstacles constraints, called The ABC algorithm used for clustering analysis with obstacles constraints ABC-CO, is proposed in the paper. By comparison with the two classic clustering algorithms, k-medoids and COE-CLARANS, demonstrates the rationality and usability of the ABC-CO algorithm.

Keywords: spatial clustering, artificial bee colony, obstructed distance; fitness calculation

Ning Lu, Leilei Yi Use of Petri Nets for maximum power point tracking in photovoltaic power generation system
*Computer Modelling & New Technologies 2014 **18**(10) 329-332*

Maximum power-point tracking is used to take full use of solar energy. The main object of this paper was to find a simple efficient technique to let photovoltaic power generation system working near the maximum power point of the solar arrays. The way of finding the suitable method of detection of the maximum power point and the strategy, which forces the system to work near this point was discussed. This project proposes a new way to realize the incremental conductance method by using high level Petri nets which is capable of tracking global maximum power point under condition change. This model will be evaluated by using stateflow in MATLAB. The result showed that this model is

effective.

Keywords: photovoltaic power generation, maximum power-point tracking, incremental conductance, high level, Petri nets

---

Shaohui Su, Pengfei Li, Zhangming Peng, Fanchao Wu, Chang Chen, Jiangang Wang Research on the knowledge flow evolving model for mechanical product innovation design
*Computer Modelling & New Technologies 2014 **18**(10) 333-339*

The innovation and competitiveness of product require the fulfilment of customers demand. The key is to acquire knowledge of customer needs. In this article, a design knowledge flow cognitive model was built on the traditional FES model, using the acquirement of product demand for a start. It fuses the user's demand, function, effect, structure, constraint and recycling of product for product innovation. According to the model, the level of knowledge classification and description methods was studied and an object-oriented description method of knowledge was proposed. The relation between knowledge and knowledge flow was analysed and so were the vertical relation and horizontal relation between each knowledge point. The evolution process of knowledge was researched based product family Take example for Curve sawing products to validate the proposed model in this article.

Keywords: innovative design, knowledge flow, jig saw

Jinchuan Zhang, Hao Yang HPN simulation model of carrying capacity of combination station for heavy-haul trains
*Computer Modelling & New Technologies 2014 **18**(10) 340-346*

Combination station for heavy-haul trains imposed restrictions on the whole heavy-haul railway system. Through analysis of particularities of operation of combination station, the paper established HPN simulation model of carrying capacity of combination station based on Petri net theory, a graphical modelling method. The simulation model took technical operations of arrival, combination and departure of trains as interconnected system, and output parameters related to carrying capacity of the station. Finally, the paper, took Hudong station in Datong-Qinhuangdao railway as an example to verify the validity and practicability of the model.

Keywords: carrying capacity, HPN, combination station, heavy-haul railway

---

Minning Wu, Fei You, Feng Zhang A new mutton logistics tracking algorithm for Internet of things based on PSO and neural network
*Computer Modelling & New Technologies 2014 **18**(10) 347-352*

In order to improve the particle filtering precision and reduce the required number of particles, to solve the neural network training algorithm has slow convergence speed, easily falling into local optimal solution, proposed a target tracking algorithm based on PSO particle filter, using of Bayesian method to sample the prior information and coupled PSO algorithm. For the existence of intelligent wireless sensor network energy constrained sensor nodes, limited communication features, the PSO optimization is introduced into the distributed particle filter algorithm to solve the existing distributed particle filter network traffic load is heavy and node energy consumption of high disadvantage. Then, we propose a new particle filter algorithm based on PSO and neural integration the algorithm makes full use filter tracking historical information, combined with predictions of particle filter, the detection signal of the sensor nodes were isolated, thus achieving the target tracking. Simulation results show that the target tracking algorithm based on particle filter PSO and neural integration can use a smaller computational cost, multi-target tracking problem solving, and practical system to meet the demand.

Keywords: particle swarm optimization, particle filter, neural network, tracking algorithm, internet of things

---

Jiaxiang Hu A system dynamics-based simulation experiment for aligning two anthrax progression models and their implications
*Computer Modelling & New Technologies 2014 **18**(10) 353-357*

Two models have been proposed in the literature to describe anthrax progression - the first is referred to as Compartment-B model, which has 22 states, and the other is called Incubation–Prodromal–Fulminant (IPF) model, which has 9 states. How do these two models differ from each other in terms of the indicators considered important by policy or decision makers? Does one always outperform the other based on key performance measures? This paper describes our experience of aligning these two models in the context of anthrax attack. We first develop two simulation

models using system dynamics to integrate the key indicators of emergency response, such as treatment rate, detection time, and treatment capacity. We then propose the process of model alignment and examine a large number of numerical examples to see whether the number of deaths, the stabilization time, and the demand for medicine produced by the two models will be reasonably equivalent. This study indicates that it is important for policy makers to understand the differences and similarities between the two models before making decisions. Furthermore, this research provides insights for scholars that rely on simulation tools for investigating bioterrorism attacks and for policy/decision makers that use these tools.

Keywords: Anthrax attack, Model alignment, Compartment model, Simulations, System dynamics

Xue Wu Economic benefit evaluation model of E-commerce based on DEA
*Computer Modelling & New Technologies 2014 **18**(10) 358-361*

E-commerce, a new business transaction mode, becomes the key drive of regional economic growth. Without time-space limitations and with different trade mode, higher cost efficiency of commodity circulation and informationization degree of transactions, E-commerce plays an important role in enterprise, industrial and even regional economic development. Based on microscopic and macroscopic perspectives, economic benefit evaluation model of enterprise E-commerce, economic benefit evaluation model of industrial E-commerce and economic benefit evaluation model of regional E-commerce were established using DEA. A deep analysis was conducted to applications of the established models, including economic evaluation before and after E-commerce, economic evaluation of different regions within the same time period, and economic evaluation of same region at different time periods of E-commerce promotion. Meanwhile, three pieces of improvement advices were given to facilitate continuous E-commerce development and increase economic benefit of regional E-commerce.

Keywords: DEA model, E-commerce, economic benefit, evaluation model

Qian Yi, Shang Tao, Qingming Zhan, Liming Bo, Jie Yin Research on outdoor wind environment of building groups based on computer simulation
*Computer Modelling & New Technologies 2014 **18**(10) 362-369*

Building outdoor wind environment is closely related to indoor air quality and thermal comfort of human, directly affecting the health and quality of life of people living and the building energy consumption at the same time. Enhancing natural ventilation in summer also helps reduce air-conditioning equipment uptime, reducing air conditioning energy consumption and green building should be particularly emphasized natural ventilation. However, in the irrational layout of the buildings or the too high buildings and other factors, outdoor wind will bring no comfort of pedestrians, but also easily lead to loss of energy and increase heating energy consumption, especially in the winter. In this paper, with the careful study of local wind data, wind environment simulation and evaluation and optimized design are conducted about two cases including teaching buildings of Faculty of Engineering in Wuhan University, as well as the modern trade mart layout in Rizhao city of china by using computational fluid dynamics (CFD) technique. Research shows that: 1) Teaching buildings of Faculty of Engineering in Wuhan University have bad ventilation in summer, which cannot meet the green building standards in China. It will bring discomfort to pedestrians on the days of large wind speed in winter and need for windproof design. 2) Through wind environment simulations of modern trade mart plan in Rizhao city, planning of program two is more reasonable than program one in the ventilation by adjusting the building pattern. Using the technique of CFD simulation can guide the existing buildings renovation and architectural planning and design optimization.

Keywords: wind environment, natural ventilation, building environment, computer simulation, CFD

Zhou Jinlin Research on the application of the Z-value analysis method in financial risk management of enterprise
*Computer Modelling & New Technologies 2014 **18**(10) 370-374*

As a method of measuring the bankruptcy risk of enterprise invented by the US scholar Altman, the Z-value analysis method is widely applied by people. According to the prediction of the model, an enterprise will go into bankruptcy if the Z-value is less than 1.20; it is the grey area if the value is between 1.20 and 2.90; the enterprise has no bankruptcy risk if it is greater than 2.90. Although the Z-value analysis method has been used for measuring the bankruptcy risk of enterprises for a long time, in fact, it reflects the risks through the financial indicators of the enterprises, so it can be

widely used for the financial risk management of enterprise. Through analysing the connotation of the Z-value analysis method, the paper explores the application of the method in the enterprise risk management from the specific calculation perspective and discusses the problems in application. Finally, some policy suggestions are provided for relevant decision-maker.

Keywords: z-value model, enterprise finance, risk management

Xihui Wu Research on internal control of accounting information in enterprises based on OPM model
Computer Modelling & New Technologies 2014 **18**(10) 375-379

OMP model is an important computing model in the data field, which is broadly applied in in the field of physics, mathematics, etc. In fact, OMP can be adopted in enterprises as well. In this paper, through in-depth analysis on the calculation method and content of OMP, the application principles, specific application methods of OMP in internal control of enterprise accounting information and existing problems was analysed in accordance with its effect, so as to establish an OPM model software evaluation system from the practical perspective and ultimately propose corresponding countermeasures for the application of OMP model in internal control of accounting information in enterprises.

Keywords: SOFC, OMP, accounting information, internal control, evaluation

Licheng Ren, Wenhui Pan Research on product characteristics affecting the transformation of B2B E-commerce
Computer Modelling & New Technologies 2014 **18**(10) 380-386

With the explosive growth of B2C and C2C Electronic Commerce (EC) in recent years, B2B EC has been tried in many fields, especially many traditional industries with little permeation into B2B EC, were transformed into B2B EC. However, the key factors influencing B2B EC transformation have not been fully investigated. On the basis of relevant literature analysis, product characteristics affecting B2B e-commerce transformation were divided into seven parts in this paper, namely, the product standardization, differentiation, tangibility, intangibility, time sensitivity, substitutability and complexity. Then, the proposed seven aspects were analysed using the method of the objective index weight, so as to obtain the mathematical model for product characteristics affecting B2B EC transformation. Finally, model test was conducted by taking Shanxi Pingyang Industry Machinery Co., Ltd. as an example, hoping the relevant research conclusion can provide references in decision-making for managers and other researches.

Keywords: B2B electronic commerce transformation, product characteristic, objective index weight

Chunjie Yu Research on supply chain surplus of low carbon supply chain coordination system
Computer Modelling & New Technologies 2014 **18**(10) 387-392

Supply chain coordination management plays an important role in reducing carbon emission in low carbon supply chain. This study introduces synergetics theory into the research of low carbon supply chain coordination management and reveals the operational mechanism and dynamic mechanism of low carbon supply chain management. Supply chain surplus is an order parameter of low carbon supply chain coordination management, which drives low carbon supply chain system to evolve and develop. It presents the formula of supply chain surplus based on its definition, analyses the components of supply chain surplus, proposes the approaches to realize supply chain surplus, supposes that allocation of supply chain surplus is a Pareto improvement issue, and discusses major basis of supply chain surplus allocation and the program of supply chain surplus allocation.

Keywords: supply chain surplus, low carbon, supply chain management, coordination management

Huilin Yuan, Jia Fu, Wei Hong, Jinbo Cao, Jing Li The application of BP neural network optimized by genetic algorithm in logistics forecasts
Computer Modelling & New Technologies 2014 **18**(10) 393-397

This paper points out disadvantages of traditional forecast methods and elaborates the advantages of the method based on BP neural network. On this basis, the paper puts forward a logistics forecasting model of BP neural network optimized by genetic algorithm. The new method uses historical data to establish and train BP neural network and thus obtain logistics forecasting model. The results implemented by MATLAB show that, neural network possesses

memorizing and learning capability, and can forecast logistics development trend perfectly, which is proved by a large amount of actual forecast results. Compared with BP neural network model, the model has the advantages of less number of iterations, convergence speed and strong generalization ability.

Keywords: BP neural network, genetic algorithm, optimized, MATLAB

Yuanmin Xie Product inventory model of iron and steel enterprises
*Computer Modelling & New Technologies 2014 **18**(10) 398-403*

As a pillar industry, iron and steel industry has made a significant contribution to China's economic development. Steel and iron inventory is a hot research issue in the supply chain management. How to make use of advanced management methods and mathematical models to obtain reasonable inventory to meet customer service and reduce unnecessary inventory-related costs, and accelerate corporate capital turnover rate, is the goal that the enterprises should pursue in the future. According to the batch production of iron and steel enterprises, and in consideration of inventory-related costs, including fixed costs, storage costs, shortage costs (including deferred delivery costs and lost sales costs), establish the model of finished products inventory cost. When calculate model, we make c + + program and do accurate calculation to the model, optimize the production cycle, production time and production quantities, and define the deferred delivery coefficients. By changing the value of the deferred coefficient, analyse the influence to the production cycle, production time and production quantities and all kinds of costs.

Keywords: Steel and iron industry; supply chain; inventory; finished products inventory; optimization

Zhang He College ideological instruction teaching method based on multimedia CAI
*Computer Modelling & New Technologies 2014 **18**(10) 404-410*

With the implementation of college ideological instruction teaching method reformation, multimedia and "three dimension" instruction method enjoy more and more popular among colleges. People place high hope on the new instruction pattern and regard it as the effective solution for education in college so far. This paper studies reformation of ideological instruction teaching method to analyse how to make use of multimedia CAI method and "three dimension" teaching method.

Keywords: college ideological instruction, multimedia network, CAI, "three dimension" teaching pattern

NATURE PHENOMENA AND INNOVATIVE ENGINEERING

Wenhui Li, Shengqiang Yang, Xiuhong Li Theoretical analysis and experimental verification of hole surface finishing parts
*Computer Modelling & New Technologies 2014 **18**(10) 411-414*

As the main method of improving surface quality, finishing technologies have been developed and applied rapidly in recent years, but they have certain limitations in the hole surface finishing. Based on the principle of centrifugal motion, a kind of hole surface finishing method with self-adaptive ability is putting forward in order to solve the finishing and cleaning problems effectively. The force of the grinding rod is analysed, so the relationship between speed and friction is defined by analysing the minimum speed of the finishing parts theoretically. Finishing parts are designed and experimental study is done. Research results show that hole surface roughness value Ra of seamless steel tube can reduce from $7.0\mu m$ to $0.3\mu m$ in 50s, which verifies the finishing effect and efficiency.

Keywords: finishing, hole surface, centrifugal principle, finishing effect

Shuang Zhang, Yuping Qin, Jing Xiao, Yihe Liu Research progress of implantable intra-body communication
*Computer Modelling & New Technologies 2014 **18**(10) 415-422*

The intra-body communication is an emerging wireless communication technology. According to coupling modes of electrodes, the intra-body communication is classified into two types, the capacitive and the galvanic coupling intra-body communication. The capacitive coupling communication is inappropriate for the medical implant intra-body communication because this communication mode requires the common grounding, while the galvanic coupling communication can exactly make up for the disadvantage of the former. In existing research overview, prototypes and experiments concerning the two coupling communication modes are thoroughly discussed, and research status of "surface-to-surface", "surface-to-implant", "implant-to-surface" and "implant-to-implant" communication methods is emphasized as per installation positions of electrodes. Furthermore, opportunities and challenges of the communication technology are presented as well as its prospects.

Keywords: intra-body communication, capacitive coupling, galvanic coupling, implantable, sensor

Yuankui Li, Yingjun Zhang, Feixiang Zhu, Jiandao Liu Task-role-based workflow authorization model and its implementation in emergency command system of water traffic
*Computer Modelling & New Technologies 2014 **18**(10) 423-427*

Water traffic plays an irreplaceable role in modern traffic as accomplishing heavy transport task in low cost, but the emergency frequency brings serious challenge to ensure safety and rescuing in water. At the same time, the safety supervision departments at all levels lack of modern technical means for prevention and action to emergency, so an emergency command system not only for safety supervision and risk early-warning in normal state, but also for quick response and scientific decision for emergency must be built. In this paper, by the use of the task-role-based workflow authorization model, an emergency command module is realized, which key technology is permission assignment. It accomplishes 7 tasks by 4 roles, besides, by the definition of user, role and task and constrains of relationship among them, and take use of delegation technology and right transmission between roles, emergency command process can executed methodically with high quality and clear target according to emergency plan.

Keywords: water traffic, emergency command, workflow, permission assignment

Shaofei Jiang, Zhifei Mao, Jiquan Li, Wei Zheng Temperature field numerical simulation and experimental study of rapid heat cycle molding in cooling process
*Computer Modelling & New Technologies 2014 **18**(10) 428-433*

Rapid heat cycle molding (RHCM) is a new technology aimed at obtaining green and high surface quality of plastic products. In this paper, the finite element model of the mould cavity in cooling process with RHCM was established for the transient heat transfer simulation using ANSYS. Thermal analysis results of temperature field were modified by experimental analysis, the results of which showed a good temperature uniformity and extraordinary efficiency of the cooling rate.

Keywords: rapid heat cycle molding, finite element model, cooling process, temperature field

Xinmin Ge, Yiren Fan, Donghui Xing, Yongjun Xu A novel method to calculate relative permeability of fluids based on fractal theory and core NMR experiment
*Computer Modelling & New Technologies 2014 **18**(10) 434-440*

An analytical relative permeability model based on fractal theory and NMR principle is described. The relationship between NMR transversal relaxation time (NMR T2) and resistivity index is deduced by fractal theory and capillary bundle model. The reciprocity theory of percolation field and electricity field is used to get the relationship between resistivity index and relative permeability. The 'Brooks-Corey/Burdine' equation was adopted to construct the relationship between relative permeability and NMR transversal relaxation time. By introducing the NMR parameters, the relative permeability model is improved since that it quantified the influence of pore structure. The results show that for water (wetting phase in water-gas system), the predicted permeability is exactly consistent with the experiment data, whereas for gas (the non-wetting phase in water-gas system), the fitted results is to some degree deviated from the experiment data and need more extensive research.

Keywords: relative permeability, fractal theory, NMR transversal relaxation time, reciprocity, pore structure

Dechen Yao, Limin Jia, Yong Qin, Jianwei Yang Faults diagnosis of railway rolling bearing by using time-frequency feature parameters and genetic algorithm neural network
*Computer Modelling & New Technologies 2014 **18**(10) 441-445*

This paper is focused on time-frequency feature parameters and genetic algorithm neural network techniques in fault diagnosis of railway rolling bearings. The time-frequency feature parameters for classification are extracted from vibration signals. However, the weak features of faults in rolling bearing are always immersed in noises of the environment, to solve this problem, Firstly, the wavelet analysis is used to filter and de-nosing and the time domain features are calculated. Secondly, the EMD (Empirical Mode Decomposition) method is used to decompose the signal into a number of intrinsic mode functions (IMFs), and then the IMF energy-torques could be calculated through the de-nosing signal. Finally, the genetic algorithm neural network is used for the classifications of the time-frequency feature parameters. The results of the time-frequency feature parameters and genetic algorithm neural network (GNN) show the

effectiveness and the high recognition rate in classifying different faults of railway rolling bearing.

Keywords: time-frequency feature parameters wavelet analysis, EMD, IMF, genetic algorithm neural network, railway rolling bearing, fault diagnosis

Zhigang Ma, Wenyi Liu Design and implementation of acoustic target recognition system based on TMS320F2812
*Computer Modelling & New Technologies 2014 **18**(10) 446-451*

In this paper, TMS320F2812 is adopted as the core component and the hardware circuit system is designed for acoustic target recognition, which includes signal conditioning circuit, A/D acquisition circuit, memory expanding circuit, power management circuit and data communication circuit, etc. The recognition algorithm was transplanted to the acoustic recognition system based on DSP, which enhanced the running rate in DSP. This system can accomplish acoustic signal sampling, pretreatment, feature extraction and recognition. On the basis of the simulation and real tests, it is proved that the acoustic target recognition system proposed in this paper is stable and reliable, which can satisfy the practical requirements.

Keywords: DSP, TMS320F2812, algorithm, acoustic target recognition

Yuguang Wang, Xingxing Dai, Xinyuan Shi, Yanjiang Qiao Mesoscopic simulation studies on the aggregation behaviour of glycyrrhizin micelles for drug solubilisation
*Computer Modelling & New Technologies 2014 **18**(10) 452-456*

Glycyrrhizin is a kind of natural surfactant with the micelle structure for drug solubilization and low toxicity in the human body. In this study, dissipative particle dynamics (DPD) and mesoscopic dynamics (MesoDyn) were carried out to elaborate the aggregation behaviour of glycyrrhizin micelles. Baicalin were selected as the poorly water-soluble drugs. It has been observed from DPD that glycyrrhizin molecules formed core/shell structured spherical, cylindrical and lamellar aggregates with the increase of concentration. Baicalin molecules are solubilized in the hydrophobic core. Glycyrrhizin molecules were easier to form micelles with the addition of drugs. Mesoscopic simulations based on experimental data provided more detailed information for the investigation of solubilization.

Keywords: mesoscopic simulation, dissipative particle dynamics, glycyrrhizin micelle, drug solubilization, aggregation behaviour

Dan Xu, Zhan Zhang, Long Yu, Yumei Wang Comparison of improved EMD entropy and wavelet entropy in vibration signals of circuit breaker
*Computer Modelling & New Technologies 2014 **18**(10) 457-461*

The paper is focusing on extracting the vibration signals of circuit breakers by using Empirical Mode Decomposition (EMD) that improved using least squares to lower the impact on experiment results effectively, which is caused by the EMD inherent end effects. First, work on the EMD and wavelet transform decomposition of both normal and loosening signals, and then to calculate the energy entropy. The results show that the value of improved EMD energy entropy is significantly larger than the wavelet energy entropy. So the improved EMD energy entropy can improve the accuracy of fault diagnosis and provides useful help for the mechanical fault diagnosis based on circuit breakers vibration signals.

Keywords: circuit breakers, vibration signals, energy entropy, EMD, wavelet decomposition

Fengwei Yuan, Qian Deng, Jiazhu Zou Fault diagnosis of nuclear facilities based on hidden Markov model
*Computer Modelling & New Technologies 2014 **18**(10) 462-467*

Due to the complex structure of nuclear facilities in a high irradiation environment, people are hard to approach it, In view of these situation a fault diagnosis method based on HMM (Hidden Markov Model) of capturing the audio signal while facilities are operating is proposed. With the strong modelling ability, HMM can be applied to analysing such as audio signal non-stationary time signal. By using this method, the original mechanical structures of nuclear facilities are not destroyed. The proposed sensors were needed as few as possible by the whole diagnosis system and which has a simple structure, low cost structure, the fault diagnosis rate is high and so on. State monitoring and fault diagnosis system of complex nuclear power equipment can timely and effective to provide running status and potential failure information for operating personnel, which has a vital significance for the safe and reliable operation of nuclear power equipment.

Keywords: nuclear facilities, Hidden Markov Model (HMM), fault diagnosis

Siqiang Wen, YunPeng Li, Yan Zhou Numerical simulation of dynamic responses caused by dynamic

compaction on backfilling foundation
*Computer Modelling & New Technologies 2014 **18**(10) 468-473*

Since the ground vibration caused by dynamic compaction threatens the structures around the site, a dynamics numerical simulation of the process of dynamic compaction is carried out based on the dynamic compaction experiment in the backfill soil site with the Finite Element Method (FEM) software. The calculated results present the vibration-time curves in radial and vertical directions on the ground in different distances. The characteristics of the vibration-time curves and how the peak velocity and acceleration change with distance are analysed. By comparing the simulated results with field data, the reasons which cause the differences are pointed out. Through comparison it is considered that the near-field dynamic responses in the simulation are more reliable than the far-field ones. According to standards the safe distances of each type of structures are evaluated. The relationship between energy utilization and the vibration energy is discussed, and that raising the aspect ratio of the hammer can reduce vibration is pointed out.

Keywords: dynamic compaction, dynamic response, backfill soil, simulation

Jing Jing Liang, Rui Qin Li, Jia Jun Ren Research on planetary bevel gear CVT system based on contact force
*Computer Modelling & New Technologies 2014 **18**(10) 474-478*

A virtual prototyping model for dynamical characteristic curves based on contact force was established, through the joint modelling (geometric modelling and constraint modelling) of planetary bevel gear CVT system in UG and ADAMS. The virtual prototyping experimental data proved that the system has the feasibility of over-zero variable speed. It is also verified that the model has advantage on continuously variable speed range, compared with planetary cone ring continuously variable transmission system. The main effect factors of the continuously variable speed performance and output torque are obtained. This model could be used to further study on such issues.

Keywords: planetary bevel gear, continuously variable transmission, virtual prototyping experimental, contact force

Hudai Fu, Jingang Gao Vehicle durability test based on user survey
*Computer Modelling & New Technologies 2014 **18**(10) 479-483*

The vehicle durability test is an important means of assessing and verifying the vehicle reliability. At present, the domestic automobile enterprises generally have some problems such as nonstandard test methods, the test mode insufficiently associating to the users in the vehicle durability test. It proposes a method for establishing the vehicle durability test mode in the paper, based on the theory of fatigue damage. Through the users' survey, the load acquisition and data analysis are carried out from the testing ground and the social roads by using of six-component force tester. It is proved, that the vehicle durability test method can not only shorten the test cycle, but also increase the appearance and countermeasures of adverse conditions. Meanwhile, it can reduce many kinds of adverse problems occur after the new vehicle coming into the market, and improve the quality of vehicle reliability.

Keywords: user survey, durability test, load collection, damage value

Gongfa Li, Wentao Xiao, Honghai Liu, Guozhang Jiang, Jia Liu Fuzzy control of flue temperature in coke oven heating process
*Computer Modelling & New Technologies 2014 **18**(10) 484-489*

Coke oven production possesses the characteristics of nonlinear, large inertia, large disturbances, and highly-coup ling and so on. Coke oven heating temperature was reflected by flue temperature and adjusted by gas flow. The control method of intermittent heating control is adopted in traditional heating control system of coke oven, and cannot satisfy the command of coke oven heating control. The control principle of combining the intermittent heating control with the heating gas flow adjustment is adopted according to analysing the difficulty and strategy of heating control of the coke oven. On the basis of researching deficiency of the existing control strategy, fuzzy hybrid control is proposed to establish heating intelligent control model of coke oven, which combines feedback control, feed-forward control and fuzzy intelligent control. Carbonization index is utilized in the model to control coking management of coke oven. Then heating fuzzy intelligent control structure of coke oven is built. According to artificial experience and actual conditions, the fuzzy controller is designed. Fuzzy control can deal with fuzzy, inexact or uncertain information and is extraordinarily robust, which can realize intelligent control of heating process of coke oven. Better control result of temperature control is realized by fuzzy intelligent control model. Intelligent control methods were used to adjust stopping heating time and heating gas flow. The practical running results indicate that the system can achieve heating

intelligent control of flue temperature, reduce temperature fluctuation, effectively improve quality of coke and decrease energy consumption, and has great practical value.

Keywords: coke oven heating, model, intelligent control, fuzzy control, hybrid control

Xikui Gao, Yan Bai, Yun Ju Load balancing over redundant wireless sensor networks based on diffluent
Computer Modelling & New Technologies 2014 18(10) 490-496

As the traffic blocking probability of traditional hard load balancing algorithms is generally high over redundant wireless sensor networks, This paper proposes a load balancing algorithm based on dividing the packet flow (LBD) over redundant wireless sensor networks based on the idea of soft load balancing. In the scheme, through numerical analysis, obtain the optimal flow-dividing ratio to determine the volume of traffic delivered to each network, which maintains network load balance. From the theoretical and simulative perspectives, the paper analyses the performance parameters, and the analytical results show that the performance of the scheme is better than other schemes. Simulation results show that the proposed method outperforms traditional hard load balancing techniques in terms of traffic blocking and packet loss probabilities.

Keywords: redundant wireless sensor networks, load balancing, packet flow diversion, access selection, handover

Dong-Sheng Xu, Feng Zhang A new optimization algorithm for multi-dimensional cloud data centre resources scheduling based on PSO
Computer Modelling & New Technologies 2014 18(10) 497-502

In order to solve the problem of multidimensional cloud resources low utilization ratio and the high energy consumption of cloud communication between tasks, proposed a new cloud data centre resource scheduling algorithm, which combined resource fusion principle, particle swarm optimization algorithms and taboo search algorithm and it's with a low-power scheduling computing, storage and bandwidth resources integration scheduling method. Simulation results show that the algorithm has the advantage of cloud resources using stable, high-dimensional cloud resource utilization and low-power cloud data centres.

Keywords: particle swarm optimization, cloud computing, cloud data centre, data scheduling

Ze Liu, Gongfa Li, Honghai Liu, Guozhang Jiang, Jia Liu Temperature field and thermal stress field of continuous casting roller bearing
Computer Modelling & New Technologies 2014 18(10) 503-509

The work condition of continuous casting roller bearing was high temperature, large load and complex. If the heat, which the bearing accepted could not be distributed effectively and cooling was not in place, the high temperature would cause bearing thermal deformation and rupture under high loads, resulting in bearing premature failure, affecting the entire casting steel production. The cooling way for the continuous casting roller bearing seat was by the internal cooling water holes which in the bearing seat. Based on the ANSYS, the temperature field of continuous casting roller bearing seat was analysed. The three-dimensional geometric model and the CAE model of the bearing seat was established, and the synthesize convection heat transfer coefficient of bearing seat would be obtained through analysed the principle of heat transfer. Took it as the boundary condition, and load the temperature in the real-time condition. Analysed temperature field of the continuous casting roller bearing seat in the stable work condition, got the overall distribution of bearing seat. Meanwhile, it could draw the conclusion that thermal load what bearing carried was the cause of bearing damage through the thermal stress analysis of the continuous casting roller bearing. It provided advice and guidance on appropriate optimization of the bearing to improve bearing life. And just the above analysis provided a theoretical guidance for the design and installation of continuous casting roller bearing

Keywords: continuous casting roller, bearing seat, temperature field, CAD three-dimensional geometric model, CAE model, stress field, thermal coupling field

Zhenchang Zhang, Changying Wang Review of the development of ocean data assimilation
Computer Modelling & New Technologies 2014 18(10) 510-515

Data assimilation compensates for the deficiency of a numerical model and minimizes the short-term forecasting error by combining observation data and numerical results. Data assimilation has become a popular research topic all over the world in recent years. The development of ocean data assimilation is introduced in this paper. 4D variational and Kalman filter methods are considered the best means of data assimilation. Thus, these two methods are described in

detail. Several novel research methods of assimilation, including assimilation with a constraint condition and dimensionality reduction, are discussed.

Keywords: data assimilation, 4D variational, Kalman filter

Anatolii Pashko Simulations of standard Brownian motion
*Computer Modelling & New Technologies 2014 **18**(10) 516-521*

This paper investigates algorithms for simulation of the trajectories of a Brownian motion (Wiener process) with given accuracy and reliability. Spectral representation of Wiener process as random series examines as a model. Estimates of the accuracy and reliability investigated in various function spaces - spaces of measurable integrated functions, Orlicz spaces and spaces of continuous functions. Given the accuracy of the numbers and simulation algorithms error of Gaussian random variables in the model are used strictly sub-Gaussian random variables. Examples of simulation are represented below.

Keywords: Wiener process, simulation, sub-Gaussian model, accuracy and reliability

Rong Zeng, Zhengfeng Jiang, He Ling, Wei Hu, Xing Wan Order analysis method based on instantaneous phase
*Computer Modelling & New Technologies 2014 **18**(10) 522-528*

Order analysis is an effective method to analyse non-stationary signal of rotating machinery. The key of this method is to acquire the time sequence under even angle re-sampling. This paper proposed an order analysis method based on instantaneous phase using Hilbert-Huang Transform (HHT), and achieved the order spectrums of torsional vibration signals of rotating machinery from simulation and experiment. Being different from the order analysis method based on instantaneous frequency, this method directly uses the instantaneous phases obtained by HHT to get rotating angle over time. Thus, it is faster and more convenient. Meanwhile, this method is less affected by the 'boundary effect'. Hence, it can achieve higher precision. The simulating and experimental analysis both verified the feasibility and accuracy of this method.

Keywords: order analysis, even angle sampling, HHT, instantaneous phase

Yanru Xue, Yinghua Yao, Min Liu, Feng Wang The study of communication fibre amplifier based on doped nano-scale semiconductor materials
*Computer Modelling & New Technologies 2014 **18**(10) 529-536*

This paper firstly describes the development of communication fibre with InP-doped nano-scale semiconductor materials in detail, and then discusses its important dispersion characteristics, starting from the definition of fibre materials dispersion to explore the affected factors of dispersion, different dispersions on fibre as well as the dispersion features of single mode fibre. From the perspectives of experimental and theoretical calculations, it analyses the dispersion characteristics of drawing and doped nano-scale fibre. Thus, it will have much broader prospects for sano-scale semiconductor materials as doping fibre amplifier in communication.

Keywords: nano materials, communication optical fibre, fibre amplifier