# KNN question classification method based on Apriori algorithm

**Caixian Chen[1, 2*], Huijian Han[2], Zheng Liu[1, 2]**

[1]*School of Computer Science and Technology, Shandong University of Finance and Economics, Jinan 250014, Shandong, China*

[2]*Shandong Provincial Key Laboratory of Digital Media Technology, Shandong University of Finance and Economics, Jinan 250014, Shandong, China*

*Received 1 March 2014, www.cmnt.lv*

**Abstract**

KNN (K-Nearest Neighbours) algorithm is a classification algorithm that can apply to question classification. However, its time complexity will increase linearly with the increase of training set size, which constrains the actual application effects of this algorithm. In this paper, based on a discussion of disadvantages of traditional KNN methods, an improved KNN algorithm based on Apriori algorithm was proposed. This method extracts the frequent feature set of training samples of different categories and the associated samples. Next, on the basis of correlation analysis of each category of samples, a proper nearest neighbour number *k* was determined for an unknown category of questions. In the training samples of known categories, *k* nearest neighbours were selected. And then, according to the category of nearest neighbours, the category of unknown question was identified. Compared with the question classification method of traditional KNN, the improved method could efficiently determine the value of *k* and decrease time complexity. Our experimental results demonstrated that the improved KNN question classification method improved the efficiency and accuracy of question classification.

*Keywords:* question classification, KNN, correlation analysis

## 1 Introduction

In the question and answer system, users can give concise, accurate, and user-friendly answers to questions input by natural language. Answers are generally a length of text. In 1999, a special project of the question and answer system evaluation was introduced in TREC conference. Hence, the question and answer system in the open domain has become a key branch and research focus in the field of natural language processing and information retrieval. Generally, the question and answer system is comprised of three modules, namely, question comprehension, information retrieval and answer extraction. For question comprehension, the recognition of question type is a crucial part, that is, question classification. Question classification is a key factor to locate and test answers and formulate answer extraction strategy.

Currently, the often used question classification algorithms include Native Bayes [4], K-Nearest Neighbours (KNN) [5], SVM (Support Vector Machine) [6] etc. In these algorithms, k-nearest neighbour algorithm is most widely used. However, two issues in KNN algorithm need to be solved: firstly, the way to determine the nearest neighbour number *K* of samples to be categorized. Secondly, in the classification, the distance between each sample to be classified and all training samples needs to be calculated. Meanwhile, the size of classification samples is often large. To calculate the similarity between thousands of training samples and the sample to be categorized, the classification performance will be far from satisfactory.

In order to overcome the disadvantages of traditional KNN methods, in this research, an improved KNN algorithm based on Apriori algorithm was proposed. This method extracts the frequent feature set of training samples of different categories and the associated samples. Next, on the basis of correlation analysis of each category of samples, a proper nearest neighbour number *k* was determined for an unknown category of questions. In the training samples of known categories, *k* nearest neighbours were selected. And then, according to the category of nearest neighbours, the category of unknown question was identified. Compared with the question classification method of traditional KNN, the improved method could efficiently determine the value of *k* and decrease time complexity. Our experimental results demonstrated that the improved KNN question classification method improved the efficiency and accuracy of question classification.

## 2 Related work

KNN is an extension of the nearest neighbour algorithm. Based on the thinking of nearest neighbour, *K* nearest neighbours of the test samples are selected, and the type of *K* new nearest samples can be determined. As a non-parameter classification algorithm, KNN has a simple and intuitive principle that is easy to realize. Thus, KNN is widely applied in the field of pattern recognition such as

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

classification and regression [5]. However, two issues in KNN algorithm need to be solved: firstly, the way to determine the nearest neighbour number *K* of samples to be categorized. According to Bayesian Decision Theory, in order to obtain reliable classification, the larger *K* is, the better results will be. However, on the other hand, *K* nearest neighbour samples should be as close to the test samples as possible. Hence, compromises need to be made in reality. The general practice is to determine an initial value first, then keep modifying based on the experimental results and finally reach the optimal value. Many researchers explored into this issue. For example, a comparatively classic reference [7] proposed a *K* nearest neighbour algorithm that can automatically select the optimal *K* value. Reference [8] presented a weighted KNN algorithm, which assigns a comparatively large weight to a comparatively nearer neighbour according to the distance from nearest neighbour samples to the test samples. In this way, even *K* is very large, samples that determine the category of the test sample are nearer samples to it. The weighted method enables the KNN algorithm to be less sensitive to *K* selection and enhance the robustness of the original algorithm. Secondly, in the classification, the distance between each sample to be classified and all training samples needs to be calculated. Meanwhile, the size of classification samples is often large. To calculate the similarity between thousands of training samples and the sample to be categorized, the classification performance will be far from satisfactory. If some samples in the training set can be reduced before the classification and the final classification accuracy can be ensured, this issue will be solved. Based on this goal, researchers have proposed various approaches to reduce the number of training samples, which can be mainly divided into editing and condensing. Editing methods can remove those samples that may generate classification error or samples surrounded by those samples in other categories, such as references [9,10]. The condensing methods are established based on the following views: samples at the decision boundary are crucial to classification accuracy while samples far from decision boundary impose little impact on the classification. Under the premise of not changing decision boundary, this method removes samples far from the boundary and obtains a comparatively small training set. For these reduction algorithms, commonly used methods include condensed nearest neighbour number rule (CNN) algorithm proposed by reference [11] in 1968. This algorithm can effectively reduce the size of training set, but often retain some samples far from classification boundary. In reference [12], a condensing method based on Voronoi diagram was proposed. The condensing set obtained from this algorithm not only accurately classifies training samples but also generates a classification boundary for all training samples. However, since the Voronoi diagram is introduced, the complexity of this algorithm is considerable. Reference [13] proposed a Decremental Reduction Optimization Procedure 1 (DROP1), and a series of improved algorithms on this basis, including DROP2 and DROP5. Subsequently, reference [14] presented the Improved KNN (IKNN). Through repeated iteration, this algorithm reduces most samples in the training set that cannot match the sample to be tested. This algorithm especially applies to the circumstance of high sample feature dimension. In addition, reference [15] proposed a Template Reduction for KNN (TRKNN), which defines a nearest neighbour chain table. Based on the table, the training set can be divided into the condensing set (generally comprised of samples near classification boundary, i.e. the new training set) and reduced set (generally the internal samples). In addition to the above algorithms, references [16, 17] presented a condensing algorithm based on other principles. Reference [18] presented a mixed model algorithm that combines editing and condensing. Based on KNN, our research proposed a *K* nearest neighbour algorithm based on Apriori algorithm.

## 3 Question classification background

### 3.1 QUESTION CLASSIFICATION AND PROCESS

Question classification is an instructional learning process. It identifies the relation model between question features and question category based on a classified training question set. Next, the relation model from the instructional process can determine the category of new question. Let us set a group of conceptual question *C* and a group of training question *Q*. Conceptual questions and questions in the question base may satisfy the hierarchical relation *h* of a concept. There is also a target concept *T*:

$$T : Q \rightarrow C \,. \tag{1}$$

*T* maps a question case to a category. For question *q* in *Q*, *T*(*q*) is known. By instructing the study of the training question set, we can find a model *H* similar to *T*:

$$g_j(x) = \arg_i \max(k_i) \,. \tag{2}$$

For a new question *qn*, *H(qn)* indicates the classification results of *qn*. The establishment of a classification system or the objective of classification study is to identify a *H* most similar to *T*. In other words, when an evaluation function *f* is given, the goal of study should enable *T* and *H* to meet:

$$\min\left(\sum_{i-1}^{|D|} f\left(T(d_i) - H(d_i)\right)\right). \tag{3}$$

Generally, question classification needs to settle 5 problems:

1) To acquire the training question set: whether the training question set is properly selected imposes remarkable impact on question classifier. The training question set should be able to represent question in each category to be processed by the classification system. In

general, the training question set should be widely recognized corpus through manual sorting.

2) To establish question representation model: to select language elements (or question features) and mathematical forms to organize these language elements to represent questions. This is a key technical issue in question classification. At present, most question classification methods and systems adopted characteristics or phrases to represent language elements of question semantics. Representation models mainly include Boolean model and vector space model.

3) To choose question features: language is an open system so that the digitalized question as language should be open. Its size, structure, language elements and the information contained in the question are open, and characteristics of the question are not limited. The question classification system should select as few question features as possible accurately and closely related to the question theme for question classification.

4) To select the classification method: the selection of methods to establish the mapping relation from question features to question category is a core issue of question classification. Commonly used methods include Native Bayes, KNN, class-centre vector, regression model, and Support Vector Machine and so on. In fact, KNN method and Support Vector Machine method are often used, which present efficient classification effects and remarkable stability.

5) Performance evaluation model: ways to evaluate classification methods and system performance. The performance evaluation model that truly reflects the internal characteristics of question classification can be used as the target function to improve the target function of classification system. In the question classification, the selection of valuation parameters is subject to the specific classification question. Single-label classification question (one test question only belongs to one category) and multi-label classification question (one test question can belong to several categories) adopt different evaluation parameters. Currently, commonly used classification performance evaluation indicators include recall ratio and precision ratio, which originate from two terms in information retrieval.

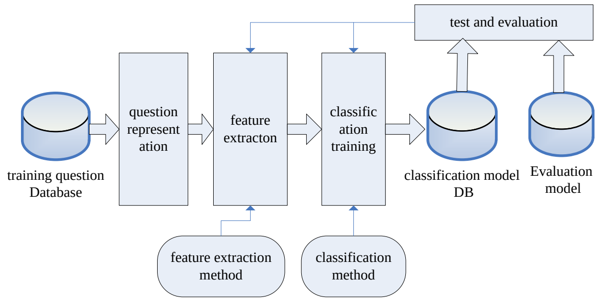

FIGURE 1 Question classification process

As shown in the Figure 1, feature selection, classification training and testing constitute a loop. According to the test results, parameters of feature selection and classification training are adjusted so that the classifier can reach the optimal classification effects.

## 3.2 INTRODUCTION OF TRADITIONAL OF KNN

KNN is a question classification method based on vector space model. Assume *x* is a question, and its vector model is $x=(x_1,x_2,...,x_n)$. Each dimension of question vector *x* corresponds to each word of question representation, also known as attribute. $C_i=(x_1^i,x_2^i,...,x_{ni}^i)$ is a question category with class identifier containing question $x_1^i,x_2^i,...,x_{ni}^i$. The number of questions is *ni*. Let there be *m* question training classification questions $C_1,C_2,...,C_m$, the classification process of KNN is as follows: for a given test question *x*, in the training question set of all categories $C_1,C_2,...,C_m$, the similarity between two questions can be used to find $k(k\geq 1)$ nearest training questions. The number of questions of Category $C_i$ was $k_i(i=1,2,...,m)$ and:

$$\sum_{i=1}^{m} k_i = k\,. \qquad (4)$$

The discrimination function of two commonly used question classification is as follows:

Discrimination function 1: (Let $g_i(x)=k_i$, $i=1,2,...,m$ and $X\in C_j$ ) is decided according to Equation (5):

$$g_j(x)=\arg_i \max(k_i)\,. \qquad (5)$$

In other words, the determination of category of test question adopted the relative majority vote method. In other words, among the category of *K* nearest neighbour question, the category containing the most questions is taken as the category of ultimate test questions.

Reference [19] presented the weighted KNN decision rule as:

$$score(d,C_i)=\sum_{d_j\in KNN(d)} sim(d,d_j)\delta(d_j,C_i), \qquad (6)$$

where $sim(d,d_j)$ is the similarity between *d* and $d_j$ (the samples to be tested), KNN(*d*) is the *k* nearest neighbour set of question d and $\delta(d_j,C_i)$ is used to indicate whether question $d_j$ belongs to $C_i$:

$$\delta(d_j,C_i)=\begin{cases}1, & d_j\in C_i\\ 0, & d_j\notin C_i\end{cases}. \qquad (7)$$

Methods to calculate the question similarity include Euclidean distance, vector inner product and included angle cosine. The included angle cosine is often used to calculate the similarity between questions. The equation is as follows:

$$sim(d_1,d_2)=\frac{\sum_{i=1}^{n}W_{1i}W_{2i}}{\sum_{i=1}^{n}W_{1i}W_{2i}} \tag{8}$$

where $W_{1i}$ and $W_{2i}$ represented the weight of the *i*-th feature word in question vector of question $d_1$ and $d_2$. The greater the cosine is, the more similar the two questions become, and the more likely the question represented by two vectors may belong to the same category, and vice versa.

The question *d* to be tested belongs to the category with the highest $score(d,C_i)$. *X* is used to replace *d* and $X_1^i$ is used to replace the first question of $C_i$ category in *k* nearest neighbours. The actual value of $\delta(d_j,C_i)$ was substituted into it. And the following decision function was obtained:

Discrimination function 2: let

$$f_i(X)=\sum_{l=1}^{k_i}sim(X,X_l^i), i=1,2,...,m, \tag{9}$$

$X \in C_j$ was determined by Equation (10):

$$g_j(x)=\arg_i\max(g_i(X)). \tag{10}$$

## 4 Improved KNN question classification algorithm

In the implementation process of traditional KNN algorithm, the distance between tested question and each training question must be calculated. Besides, the nearest neighbour number *k* cannot be determined. This influences the promotion of KNN algorithm classification. This paper proposed an improved KNN-based question classification method using the correlation analysis.

### 4.1 CONCEPTUAL FRAMEWORK OF ASSOCIATION RULES

Let item set $I=\{i_1,i_2,...,i_m\}$ be the set of *m* different symbols. Each symbol was called an item. *D* is a set comprised by several transactions *T*. *T* is the subset of transaction set *I*. Each item has the unique identifier *TID*. If *X* is a sub-set of *T*, *T* contains item set *X*.

Definition 1, item set: set of items; the set including *k* items is known as *k*-item set.

Definition 2, supportive number of item set: the number of items in *D* is regarded as the supportive number of item set *X*, which can be expressed as:

$$\text{support\_num}(X)=\left|\{T \mid T\in D, X\subseteq T\}\right|.$$

Definition 3, item set support degree indicates the probability of item set occurred in *D*.

Definition 4, frequent item set: the item set with the minimum support threshold larger or equivalent to that designated by users.

### 4.2 BASIC IDEA OF THE ALGORITHE

Let there be two categories, *P* and *Q*, and 4 feature words, p1, p2, q1 and q2. 4 feature words can produce 15 kinds of non-empty sets {p1}, {p2}, {q1}, {q2}, {p1, p2}… {p1, p2, q1, q2}. Apriori algorithm was used to record question objective that contains the characteristic set under different circumstances. For example, there are 6 objectives that contain the feature word p1, and 2 objectives that contain the feature word p1, p2 and q1. In this way, the frequent item set of feature word was established. When classifying the classification question (p1, p2, q1), we can directly find the question corresponding to the frequent item set of feature word {p1, p2, q1}. As an initial nearest neighbour of classification question, the nearest neighbour number *K* of the questions to be categorized could be determined ultimately according to the number of initial nearest neighbours. Next, the similarity of questions can be calculated. According to the categories of the first *K* nearest neighbours, the category of questions to be classified can be determined. For example, the square category in the Figure 2 made for 2/3. Thus, circular objectives were determined as the square category.

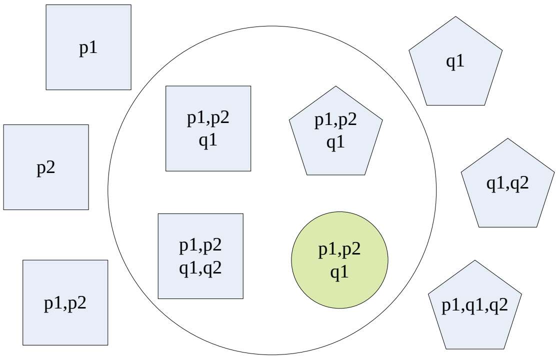

FIGURE 2 Diagram of improved KNN question classification algorithm

It can be seen that the improved classification algorithm could efficient reduce the time complexity of algorithm. However, at the earlier stage, a mapping relation between feature word frequent item set and associated training question objective set can be established, which requires certain time-space consumption.

### 4.3 QUESTION SIMILARITY CALCULATION METHOD

Methods to calculate the question similarity generally include the Euclidean distance, vector inner product and included angle cosine. In order to enhance the accuracy of question similarity calculation, this research applied HowNet to adopt semantics-based question similarity calculation method.

HowNet is a repository with the description objective of concepts represented by Chinese and English terms, which can reveal the relationship between concepts as well as between the basic attributes of concepts. HowNet describes the Chinese terms based on the conceptual framework of “sememe”. Sememe can be regarded as the smallest and fundamental Chinese semantic unit that is not easy to divide. Since Chinese words express different meanings in different contexts, Hownet comprehends Chinese words as a set of several sememes. Each entry of the semantic dictionary of HowNet is comprised of a sememe of a word and its description. In other words, one entry corresponds to one sememe of a word. Meanwhile, each sememe is described by several sememe. HowNet provides a classification tree of sememe, and there is a hyponymic semantic relation between parent-node and child-node. Thus, the classification tree of sememe can be used to calculate the semantic similarity between two words.

### *4.3.1 Calculation of sememe similarity*

$$sim(p_1, p_2) = \frac{2 \times Spd(p_1, p_2)}{Depth(p_1) + Depth(p_2)}, \tag{11}$$

where $p_1$ and $p_2$ represent two sememe sources, and $Spd(p_1, p_2)$ shows the coincidence degree of $p_1$ and $p_2$. $Depth(p)$ is the depth of a sememe in the sememe tree.

### *4.3.2 Similarity calculation of notional words*

In HowNet, the concept of content words (sememes) can be divided into 4 parts:

1) primary fundamental sememe description: the first sememe in DEF item;

2) description of other fundamental sememe: all the other independent sememe or specific words in DEF item;

3) description of relational sememe: in DEF item, “relational sememe =fundamental sememe ” or “relational sememe = (specific words) ” or “ (relational sememe =specific words)” to describe the notional part;

4) description of symbol sememe: in DEF item, “relational symbol fundamental sememe ” or “relational symbol (specific words)” to describe the notional part.

Thus, the similarity corresponding to 4 parts of the 2 notions is respectively marked as $sim_1(C_1, C_2)$, $sim_2(C_1, C_2)$, $sim_3(C_1, C_2)$ and $sim_4(C_1, C_2)$. In this way, the entire similarity of notional words is:

$$\mathrm{sim}(C_1, C_2) = \beta_1 sim_1(C_1, C_2) + \sum_{i=2}^{4} \beta_1 \beta_i sim_i(C_1, C_2), \tag{12}$$

where $\beta_i$ satisfies:

$$\beta_1 + \beta_2 + \beta_3 + \beta_4 = 1, \beta_1 \geq \beta_2 \geq \beta_3 \geq \beta_4 \geq 0.$$

### *4.3.3 Similarity calculation of Chinese words*

For two Chinese words $W_1$ and $W_2$, if $W_1$ has $n$ sememe items: $c_{11}, c_{12}, ..., c_{1n}$ and $W_2$ has m sememe items $c_{21}, c_{22}, ..., c_{2m}$. Let the similarity between $W_1$ and $W_2$ be the maximum similarity of each sememe items. Then we have:

$$sim(W_1, W_2) = \max_{1 \leq i \leq n, 1 \leq j \leq m} sim(C_{1i}, C_{2j}). \tag{13}$$

### *4.3.4 Similarity calculation of Chinese sentences*

For two sentences $S_1$ and $S_2$, $S_1$ has $n$ words: $w_{11}, w_{12}, ..., w_{1n}$. $S_2$ has $m$ words: $w_{21}, w_{22}, ..., w_{2m}$. The calculation method of sentence similarity is: based on the word set of two sentences, one word is selected from one set to calculate the similarity with each word in another set. The word pairs of the maximum similarity are selected. The loop is not stopped until the first set word is empty. Next, the similarity of selected word pairs is added and then divides the word number contained in the first set. Finally, the calculation results based on two sets are averaged to obtain the similarity of two sentences. The calculation equation is as follows:

$$sim(S_1, S_2) = \left. \left( \frac{\sum_{u=1}^{n} \max_{1 \leq v \leq m} sim(w_{1u}, w_{1u})}{n} + \frac{\sum_{v=1}^{m} \max_{1 \leq u \leq n} sim(w_{1u}, w_{1u})}{m} \right) \middle/ 2 \right. . \tag{13}$$

## 4.4 REALIZATION OF ALGORITHM

Our research proposed a KNN question classification method based on Apriori algorithm. With correlation analysis, the nearest neighbour was selected, which avoided defects of traditional KNN. Compared with traditional methods, our method effectively improved time complexity and *K* selection. The improvement methods are mainly divided into two stages:

1) based on correlation analysis, the frequent feature word set and the associated training question were extracted;

2) based on findings of correlation analysis, the initial nearest neighbour of questions to be classified was determined, as well as the final nearest neighbour number K. Next, KNN was used for question classification.

1) Extraction of frequent feature word set and associated training question based on correlation analysis

a) Let the total number of question category be *m*, and the category is $C_1, C_2, ..., C_m$. The number of training samples of each category is noted as $N_1, N_2, ..., N_m$; questions in the training set are processed in advance. With $\chi 2$ statistical approach, a certain amount of questions of different categories in the training set are selected, and noted as the feature word of $N_f$ (for example, 10 characteristics are selected from each category);

b) Scan all training questions and express each question as $m \times N_f$ -dimensional question vector comprised by feature word of all categories. TF-IDF is used to calculate the characteristic weight;

c) Extract the frequent feature set and the associated questions from each category; this step only considers the characteristics of the category of each training question, and the rest will be omitted for now; each category is processed respectively, including the following procedures:

*i)* each question of this category is seen as a single transaction, and the included characteristics of this category are seen as the data item of transaction. Item set is also a feature word set of this category. The minimum supportive degree is set and Apriori algorithm is used to enable the question category to meet all item sets of minimum supportive degree threshold, that is, the frequent item set of this question category is produced;

*ii)* for each frequent item set, the associated training question is preserved, and the training question that contains all features of a frequent item set is the associated training question of this frequent item set;

2) With the findings of correlation analysis, the initial nearest neighbour of the question to be classified is determined as well as the ultimate nearest neighbour number K. The question classification is conducted based on the category of nearest neighbour.

a) For a question to be classified, the pre-treatment is conducted. Next, the extracted feature word of each category can be used to represent this question and to obtain $m \times N_f$ -dimensional question vector. TF-IDF is also applied to calculate the characteristic weight;

b) The feature word weight of each category in question vector of a question to be classified is respectively summed up and arranged in a declining order. Categories in the top 3 are selected and noted as $C_x, C_y, C_z$ and the characteristics;

c) Feature words belonging to the top 3 categories as obtained from *ii* are selected. The maximum frequent item set is found in the corresponding category to acquire the associated training question. These training questions are used as the initial nearest neighbours for a question to be classified. Let the associated training question set be $S_x$, $S_y$ and $S_z$ respectively. The number of questions is $N_x$, $\underline{N_y}$ and $N_z$. Let $k = \min\left(2.5 \times N_x, N_x + N_y + N_z\right)$;

d) The cosine similarity between the question to be classified and each initial nearest neighbour question is calculated;

e) The similarity is arranged in a declining order, and the first *k* training questions are selected. The file number of 3 categories is noted and the similarity is accumulated according to different categories. Thus, the average similarity between the question to be classified and the nearest neighbour question of each category. The category with the maximum mean is determined as the category of the question.

## 5 Experiment

### 5.1 DATE SET

For Chinese question classification, there is no uniform standard question test set and training set so far. In our research, the proposed question set is a set of sentences of demarcated question type by a certain question classification system. Since the addition of QA test tasks in TREC-8, TREC conference provides masses of English question sets for QA evaluation in an annual fashion. Thus, the free question set of TREC2013 was used in our experiment, which was adopted into a part of the Chinese question set by translating some questions and transforming some questions. In addition, some questions were extracted from the previously developed recruitment question and answer system of Shandong Economic University. Besides, some questions were collected from the Internet. Both formed the Chinese question set in the context, and a total of 1500 questions were included.

### 5.2 RESULT ANALYSIS

Evaluation of a classifier is a key research topic in question classification. For different goals, researchers have proposed many evaluation approaches for question classification such as: recall ratio, precision ratio, F1 test value, and macro-averaging, micro-averaging and so on. In our research, the classic information retrieval evaluation criteria of recall ratio, precision ratio and F test value are used for our evaluation: recall ratio indicates the ratio between accurately identified sample size by the classifier and the sample size belonging to this category. The accuracy rate indicates the proportion of samples really belonging to this category among samples classified in this category by the classifier. The mathematical equation is:

$$R = \frac{A}{A+C}, \tag{14}$$

$$P = \frac{A}{A+B}. \tag{15}$$

*A* represents the number of questions that belong to the category in the manual sorting criteria and are actually classified by the classifier in this category. *B* represents the number of questions that do not belong to the category in the manual sorting criteria but are actually classified by the classifier in this category. *C* represents the number of questions that belong to the category but are distributed in other categories; *D* represents the number of questions that do not belong to the category and are not classified in the category, as shown in Table 1:

TABLE 1 Meaning of ABCD

| | **Number of questions truly belong to this category** | **Number of questions truly not belong to this category** |
|---|---|---|
| **Number of questions belong to this category** | A | B |
| **Number of questions not belong to this category** | C | D |

For a certain category, recall ratio and precision ratio reflect two aspects of the classification quality. Recall ratio and precision ratio are mutually influenced. Under normal circumstances, precision ratio will decrease with the rise of recall ratio, and it is hard to ensure that both are high. Thus, in order to comprehensively reflect the performance of classification system, recall ratio and precision ratio should be comprehensively considered. The integration of the two indicators will produce a new indicator of evaluation - F1 test value. The mathematical equation is as follow:

$$F1(\mathrm{P},\mathrm{R}) = \frac{2\times P\times R}{P+R}. \tag{16}$$

In order to avoid the influence of a small size of sample on the test results, the cities, dates, specific Figures 3-5, the total number of money, the definitions are selected because they contain 381 questions for the test. The results are shown in Table 2.

TABLE 2 Comparison of experimental results

| **Classification Method** / **Question type** | **Conventional KNN** | | | **Improved KNN** | | |
|---|---|---|---|---|---|---|
| | *precision ratio* | *recall ratio* | *F1 test value* | *precision ratio* | *recall ratio* | *F1 test value* |
| Date | 87.4 | 82.8 | 85.0 | 93.8 | 86.4 | 89.9 |
| City | 84.8 | 84.9 | 84.8 | 83.6 | 89.5 | 86.4 |
| Specific Figure | 87.1 | 84.1 | 85.6 | 90.9 | 87.3 | 89.1 |
| Sum | 55.6 | 90.3 | 69.3 | 66.2 | 93.2 | 77.4 |
| Money Amount | 78.2 | 52.3 | 62.7 | 89.4 | 71.6 | 79.5 |
| Definition | 83.7 | 85.3 | 84.5 | 89.5 | 92.1 | 90.8 |

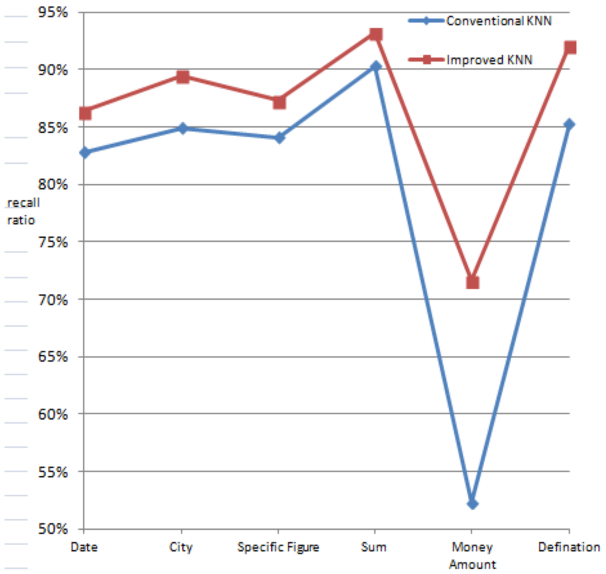

FIGURE 3 Recall ratio contrast

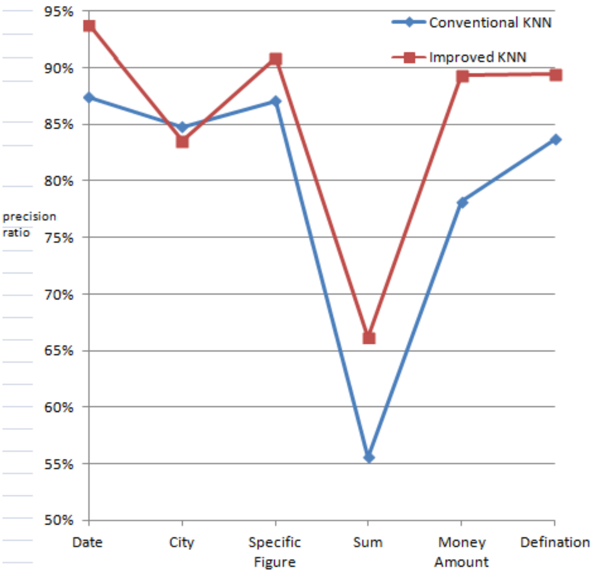

FIGURE 4 Precision ratio contrast

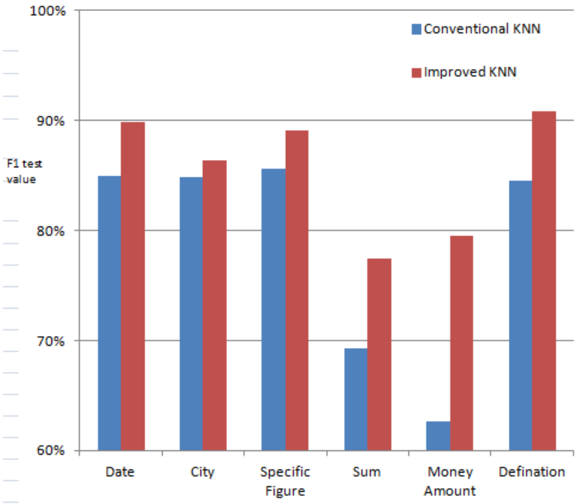

FIGURE 5 F1 test value contrast

It can be observed from the experimental results of traditional KNN algorithm that the recall ratio of money amount is comparatively low, and the precision ratio of the total number is low. It should be noted that the training question number of the category of total number is larger than other categories, and the money amount features share similarity with that of the total number. These led to the mis-categorization of money amount questions in the category of the total number. Meanwhile, the improved KNN algorithm conducted correlation analysis for question vectors in each category and applied the average similarity between the question to be classified and nearest neighbour samples of each category in category determination. This improved algorithm effectively reduced the mis-categorization of text category.

The above experiments presented that feature word from correlation analysis and related information between training questions can be feasibly used to select the nearest neighbour of nearest question. Compared with the traditional KNN method, the calculation load of selecting nearest neighbour can be substantially reduced to about 1/3 of time complexity of traditional methods; meanwhile, nearest neighbours in training samples of different categories can be found via the frequent feature word set. The nearest neighbour number in different categories is of great reference values for determining the nearest neighbour number *k* of test question. The experimental results presented that the proposed nearest neighbour selection approach effectively reduced the involvement of training questions with limited similarity in the category judgment of test question. Meanwhile, as frequent item set was produced by Apriori algorithm and the minimum supportive degree was set, the proposed method is not very sensitive to the local features of samples so that it can improve the accuracy of the classification to some degree.

## 6 Conclusions

Our research proposed an improved KNN question classification method based on Apriori algorithm, which modified the determination method of nearest neighbour number *k* and reduced the time complexity of classification. The experiment demonstrated that compared with traditional methods, time complexity of the improved KNN question classification method was relatively smaller and the classification accuracy was high. Of course, this method presents certain disadvantages: firstly, correlation analysis using Apriori algorithm of various categories of questions causes remarkable consumption temporally and spatially. Secondly,

For a test question, the retrieval of frequent item set that meets the minimum supportive degree cannot accurately find all the neighbour question vectors. In the experiment, about 80% nearest neighbours of test question can be found. In addition, if the characteristic of a category is not remarkably obvious, this will result in the failure of or greatly limited extraction of associated information that abides by the minimum supportive degree. The solution of the question needs further improvement of Apriori algorithm to extract the associated information, thus improving the recall ratio of test question nearest neighbour.

## Acknowledgments

This work was supported in part by a grant from The National Nature Science Foundation of China (61272431, 61303090, 61303089), Nature Science Foundation of Shandong Province (ZR2011FL020, ZR2011FL029, ZR2012FM002), Independent Innovation plan of Universities in Jinan (201303020), Humanities and Social Sciences Project of the Ministry of Education in China (13YJC860023) and the Foundation for Outstanding Young Scientist in Shandong Province (BS2012DX028).

## References

[1] Cai L, Zhou G, Liu K, Zhao J 2011 Learning the Latent Topics for Question Retrieval in Community QA *Proceedings of 5th international joint conference on Natural Language Processing* Chiang Mai Thailand 273-281

[2] Zhou T C, Lin C Y, King I, Lyu M R 2011 Learning to Suggest Questions in Online I:orums *Proceedings of the 25th MI Conference on Artificial Intelligence* San Francisco CA USA 1298-303

[3] Ji Z, Xu F, Wang B 2012 Question-answer Topic Model for Question Retrieval in Community Question Answering. *Proceedings of the 21st ACM international conference on Information and knowledge management* Maui HI USA 2471-4

[4] Gong X, Sun J, Shi Z 2002 The Classifier of Initiative Bayesian Network *Journal of Computer Research and Development* **39**(5) 74-9

[5] Zhang N, Jia Z, Shi Z 2005 Text Classification based on KNN Algorithm *Computer Engineering*, **31**(8) 171-3

[6] Joachims T 1998 Text categorization with support vector machines: learning with many relevant features *Proceeding of ECML – 98 10th European Conference on Machine Learning* Berlin: Springer-Verlag 137-42

[7] Gora G and Wojna A 2002 A classifier combining rule induction and k-NN method with automated selection of optimal neighbourhood

*Proceedings of the Thirteenth European Conference on Machine Learning* Heidelberg Springer Berlin **2430** 111-23

[8] Dudai S A 1976 The distance-weighted k-nearest neighbour rule. *IEEE Transactions on Systems, Man and Cybernetics* **6**(4) 325-7

[9] Ferri F and Vidal E 1992 Colour image segmentation and labelling through multiedit-condensing *Pattern Recognition Letters* **13**(8) 561-8

[10] Segata N, Blanzieri E, Delany S J, Cunningham P 2010 Noise reduction for instance-based learning with a local maximal margin approach *Journal of Intelligent Information Systems* **35**(2) 301-31

[11] Hart P E 1968 *IEEE Transactions on Information Theory* IT **14**(3) 515-6

[12] Toussaint G T, Bhattacharya B K, Poulsen R S 1985 The Application of Voronoi Diagrams to Nonparametric Decision Rules *Proceedings Computer Science and Statistics: 16th Symposium on the Interface* North-Holland Amsterdam 97-108

[13] Wilson D R, Martinez T R 2000 Reduction techniques for instance-based learning algorithms *Machine Learning* **38**(3) 257-86

[14] Wu Y Q, Ianakiev K, Govindaraju V 2002 Improved k-nearest neighbour classication *Pattern Recognition* **35**(10) 2311-8

[15] Fayed H A, Atiya A F 2009 *IEEE Transactions on Neural Networks* **20**(5) 890-6

[16] Huang D, Chow T W S 2006 Enhancing density-based data reduction using entropy *Neural Computation* **18**(2) 470-95

[17] Paredes R, Vidal E 2006 Learning prototypes and distances: a prototype reduction technique based on nearest neighbour error minimization *Pattern Recognition* **39**(2) 171-9

[18] Brighton H, Mellish C 2002 Advances in instance selection for instance-based learning algorithms *Data Mining and Knowledge Discovery* **6**(2) 153-72

[19] Tan S 2005) Neighbor-weighted k-nearest neighbour for unbalanced text corpus. *Expert Systems with Applications* **28**(4) 667-71

## Authors

**Caixian Chen, born in September, 1979, China**

**Current position, grades:** researcher at Shandong University of Finance and Economics and Shandong Provincial Key Laboratory of Digital Media Technology, China.
**University studies:** Master's degree in computer science and technology at Wuhan University of Technology, China in 2005.
**Scientific interests:** natural language processing and information retrieval.

**Huijian Han, born in December, 1971, China**

**Current position, grades:** professor at Shandong University of Finance and Economics. Researcher at Shandong Provincial Key Laboratory of Digital Media Technology, China
**University studies:** PhD degree in computer science and technology from Shandong University, China in 2010.
**Scientific interests:** CG&CAGD and animation.

**Zheng Liu, born in April, 1980, China**

**Current position, grades:** associate professor at Shandong University of Finance and Economics. Researcher at Shandong Provincial Key Laboratory of Digital Media Technology, China.
**University studies:** PhD degree in computer science and technology from Shandong University, China in 2010.
**Scientific interests:** information retrieval and data mining.