Research in search engine user behaviour based on cloud computing

Wei Liu[1, 2*], Cunchen Tang[1], Fan Kang[1]

[1]*International School of software, Wuhan University, Wuhan, Hubei, 430079, China*

[2]*Department of Information Engineering, Wuhan Business University, Wuhan, Hubei, 430056, China*

**Corresponding author's e-mail: yoxiky2003@tom.com*

Received 20 December 2014, www.cmnt.lv

Abstract

User behavior analysis is important for both Web information retrieval technologies and commercial search engine algorithms. With the expansion of information data, the current search engine is facing some serious problems, such as limited storage space and computing power. The paper discussed the shortcomings and technical bottlenecks of the current traditional search engine. Then, in the understanding of search engine features and technical requirement, it improved the system by means of the cloud computing architecture. With the combination of the static analysis of user behaviour and real-time monitoring, real-time acquisition of Web log and user to access the context information of the page, the paper tested the whole system performance in the laboratory environment, demonstrated the superiority of the system by analysis of experimental data.

Keywords: user behavior analysis, search engine, cloud computing, system performance

1 Introduction

At present, cloud computing is a hot spot in IT industry, almost every IT company is promoting this newly emerging business model and spending huge amounts of money researching cloud computing. With storage system becoming more cheaper, internet bandwidth higher, processing unit faster, the old assumption that moving computing and storage into the cloud is becoming true. Now there is no common interface for cloud computing, nor related standard published. IT companies are focusing on building their own "clouds" in different fields including: e-commerce, internet storage, online office, search engine, online map, etc. Among them, search engine is the forefront of these applications. Cloud Computing will play a very important role in the evolution of search engine. Up to now, most search engine company's backgrounds are not cloudy yet, there is still much work to do. It is a very good practice to research Cloudy Computing from the characteristics of search engine.

User behavior analysis has already drawn the wide attention of scholars at home and abroad. Cen Rongwei, based on 756000000 real network user behavior log on, the user search behavior in the query length, query modification rate, related search hits, the first / last click distribution and query hits distribution information, in order to optimize the search engine algorithm and the improved system cable. Wang Zhenyu, Guo Li and so on, through the HDFS MapReduce distributed file system and parallel computing model to support massive log file analysis, user click behavior, so as to enhance the search engine retrieval algorithm and the retrieval efficiency of service, freeing users from abundant disorderly search results. China Telecom Guangdong Institute's Tao Caixia, Xie Xiaojun proposed an engine analysis solutions of mobile Internet user behavior data based on cloud computing, including the overall system architecture design, data storage and preprocessing module, user behavior data analysis model, the design of the key module. At the same time, the foreign scholars such as Joohee Kim, Chankyou Hwang MapReduce model using cloud computing platform is proposed for IPTV user behavior analysis method of Hadoop, describes the IPTV user regional characteristics.

To sum up, at present, for the analysis of user behavior are mostly concentrate on mining WEB log, the log is the intentions of the user, the actual performance, motivation in action. However, Web log is not enough to describe the scene when the user visits a website, the user must be collected in real time on the client end operation behavior and the context information. We combine the two according to the maximum possible to reproduce the user, the real scene browsing Web pages, to extract the comprehensive of the user behavior trajectory, provide effective data support for the analysis of user behavior.

Analysis of user behavior refers to the site access to basic data, through the study on statistical analysis of the relevant data, found that the laws of the user to access the website, to allow enterprises to more detailed, clear understanding of user behavior, to find out the existing problem of business website, marketing channels, marketing environment, help enterprises to obtain high conversion page, let the enterprise marketing is more accurate, efficient, improve business conversion rate, so as to enhance the enterprise income.

2 User behavior analysis engine based on cloud computing engine architecture

In this study, "user behavior analysis engine" is defined as: according to certain strategy, respectively, to obtain user

dynamic behavior and behavior, and summarize analysis and reasoning, the system user behavior habit and characteristics.

"The scale of user behavior based on cloud computing analysis engine" for user behavior information, the use of cloud technology, storage and analysis of its efficiently find, mining user behavior, its structure as shown in Figure 1. It from the client to obtain real-time dynamic behavior of context information, asynchronous upload to the server to save; trigger server processing module pre-processing, aggregation analysis; access to Web log from the server, filtering, denoising and mining, and according to the point in time restore user history context information; at the same time, the dynamic behavior, treatment history stored in HBase database which: Static behavior analysis: mainly completes the Web log mining, and according to the time point the user to restore the historical context of user behavior, filtering, denoising, fusion operation, save the processed results and user behavior database.

Dynamic behavior analysis: based on the Markoff model, the dynamic behavior of reasoning, to collect, analysis of user behavior and characteristics.

The dynamic behavior of acquisition and preprocessing: access to information users real-time operation page from the client's behavior, and pre-processing, storing the results in HBase database. Including data cleaning, transformation, reduction, delete the useless content, check the information completeness and consistency.

User behavior information storage based on HBase: storage of user behavior information from the client and the server, the dynamic and static user behavior data, results and analysis.

The polymerization behavior of dynamic user: dynamic user behavior data filtering, integration, excluding those correct but invalid information user behavior but invalid.

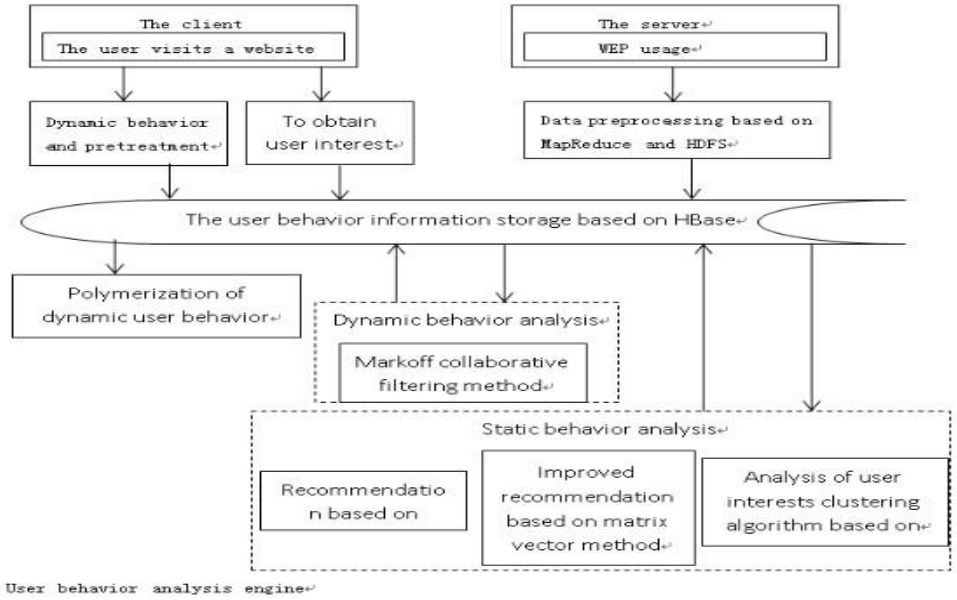

FIGURE 1 User behavior analysis engine based on cloud computing

3 The dynamic behavior of the user acquisition and preprocessing

The dynamic behavior of user refers to the user (including login and not logged in two cases. The user login, registered account user identification by acquiring ID; for users who are not logged in, record visit their website SessionID logo) occurred in accessing the page a moment of behavior, the behavior includes the occurrence time, the page (contains the page title and page URL), related to the operation and behavior subject, real-time capture them and carry on the effective analysis, has an important significance for understanding user behavior characteristics.

From the client gets information including the user dynamic behavior and context information, the context, including: behavior occurrence time; the current user ID or SessionID; the current page title; the address URL of the current page; the current user search conditions; access to the same page number; page retention time; do you want to save the page; printed the page whether or not; whether add to favorites; copy or cut the page content and so on; the environmental context information includes: the client machine configuration, current network condition, the server working condition etc.

Because the user behavior data, acquired it, by using MapReduce model in cloud environment, including filtering, eliminate duplication, delete the useless content, check the information completeness and consistency. As the following methods used:

1) Data cleaning: removal of the incomplete data, delete duplicate data, delete access to pictures, delete pages of animation, the user behavior analysis of useless data [8].

2) Data conversion: the pages print, collection, presser-vation, downloaded operation, in the acquisition, will be converted into the corresponding data format in the database.

3) Data reduction: the user behavior data in large quantity, to standardize the data quantity, reduce the very necessary, but must maintain the integrity of the data.

4 The historical behavior mining

Analyzing the historical behavior mainly comes from on the WEB log mining, due to a surge in the number of Internet, the amount of data the WEB log and exponential growth, which makes the analysis platform of based on the single node has been unable to meet the needs of massive data analysis. Therefore, the analysis engine user behavior based on cloud platform, the cloud storage technology will be the mass of WEB log is stored in the HDFS distributed file system, and the calculation model by using MapReduce, the log file cleared, denoising, protocol, finally, on a variety of data mining algorithms are parallelized modification, in the form of services to the users with mining analysis functions, including: analysis, research, user behavior improved matrix vector association rules method based on clustering algorithm based on user interest degree of user behavior query vector space model based on cloud environment, mining results, stored in the HBase user behavior information base, in order to combine with the dynamic behavior, reasoning user interest. In order to enhance the efficiency analysis of log file.

5 Study on the vector space model of retrieval based on user behaviour

Vector space model (VSM:Vector Space Model) proposed by Salton and others in twentieth Century 70 years, it is the basic idea of each text and query contains some features independent properties reveal its content, and each feature attributes can be regarded as a dimension vector space, then the text can be expressed as a collection of these attributes, ignoring the complex relationship between paragraphs, sentences and words in the text structure. At the same time, given the feature weight vocabulary certain (weight), anti should vocabulary in the importance and the value of the contents of the file identification, this value is called the indexing vocabulary "significant value (Term Significances)" or "weight", by the lexical statistics calculate the document and to, such as: the feature words appear frequency (Term frequency, TF). Vector of each file is in fact all the document feature through a combination of computing, called "the document feature item vocabulary matrix". And then all of the document vector based on specific computing methods of similarity measure between each other.

Vector space retrieval model can be described as $I=(D,T,Q,F,R)$. Among them: $D=\{d_1,d_2,...d_n\}$ as a collection of text, *n* text collection number; $T=\{t_1,t_2,...t_n\}$ set as a feature, *m* feature of all. A text *m* feature indexing can be represented as a vector space $d_i=\{w_{i1},w_{i2},...w_{im}\},i=1,2,...,n$, w_{ij} is characteristic t_j for the text d_i of the weight, if the weight value w_{ij} is 0, indicating t_j that it is not appeared in d_i, $Q=\{q_1,q_2,...q_m\}$ for the query set, a query q_r can be represented by vectors $q_r=\{q_{r1},q_{r2},...q_{rm}\}$, q_{rj} is a characteristic to t_j the query q_r weights, if the weight value q_r is 0, indicating that t_j is not appeared in q_r.

Further definition: frequency tf_{ij}: t_j is the feature for text d_i appear in the frequency.

Inverse document frequency word idf_i (inverse document frequency): the word in the quantitative distribution of document collection, the calculation $\log(N/n_k+0.5)$ is usually, where N is the total number of document centralized, *n* represents a number of documents containing *K*, called the document frequency of the term.

The normalization factor: in order to reduce the inhibitory effect of high frequency characteristics of individual word on other low-frequency feature words, the standardization of components.

Based on the above three factors to term weighting Equation (1):

$$w_{ik}=\frac{tf_{ik}\log(N/n_k+0.5)}{\sqrt{\sum_{k=1}^{n}(tf_{ik})^2\times[\log(N/n_k+0.5)]^2}} \tag{1}$$

The similarity between the text and the query can be used to measure the distance between two vectors. There are many kinds of calculating method of similarity, commonly used methods of inner product, Dice coefficient, Jaccard coefficient and cosine coefficient, usually uses the cosine coefficient method, namely the cosine of the angle between two vectors to represent the similarity between the text and the query $Sim(d_i,q_j)$, see Equation (2). Cosine similarity calculation method is a normalization, the angle between the two vectors of the smaller, the greater the degree of correlation between documents, correspondence cos is higher. Two vector included angle cosine is equivalent to their standard vector inner product unit length, it reflects the similarity term component two vector of relative distribution.

$$Sim(d_i,q_j)=\cos\theta=\frac{\sum_{k=1}^{n}w_{ik}\times w_{jk}}{\sqrt{(\sum_{k=1}^{n}{w_{ik}}^2)\times(\sum_{k=1}^{n}{w_{jk}}^2)}}. \tag{2}$$

6 Algorithm design

Input: user access to key words each time the user query log in.

Output: the similarity with the query keywords vector existing values in the database is not paper and similarity of 0, and according to the similarity value from big to small order.

1) Extracted from each Webpage keywords as the feature word, and these feature words and keywords query every time the user binding, rearrangement and according to a lexicographic order, combined together to form a standard feature set of words;

For example, there are Webpage document set (NDoc1, NDoc2, NDoc3, NDoc4), all the feature words together for (W1, W2, W3, W4, W5, W6), and when the query words (WQ1, WQ2), where wq1= W4, the standard set of words

as features (W1, W2, W3, W4, W5, W6, WQ2), Webpage document feature item vocabulary matrix in Table 1.

TABLE 1 Webpage document vector space model

	W1	W2	W3	W4	W5	W6	WQ2
Q1	0	0	0	0	0	0	1
NDoc1	14	21	33	0	0	0	0
NDoc2	0	11	15	0	0	22	0
NDoc3	8	0	0	14	15	17	0
NDoc4	0	8	9	12	0	15	0

2) t_j is calculated for each term tf_{ij} appears in the Webpage text d_i frequency; calculation of $\log(N/n_k+0.5)$ words formula of inverse document frequency idf_i. Then the formula with weight (3.1) to calculate the weight of each feature words each Webpage of document vectors, forming the Webpage document vector in a vector space.

3) To calculate the similarity of each document vector and the query vector Webpage between the cosine coefficient methods, see Equation (2). The interception of similarity values greater than 0.2000 articles, and from high to low return results.

Association rule mining is used to find correlation between the attributes of databases. Association rules is the initial motive of shopping basket analysis problem, the goal is to find the different commodities of association rules mining in transaction database, the relationship between the useful knowledge description data item value. These knowledge characterizes customer buying behavior and mode, use these rules, can effectively guide the scientific arrangement and design business purchase goods shelves. The form of association rule is a rule is, "to buy milk and bread customers, 90% of people bought butter", namely "(milk, bread) → butter" issue.

Let be the $I=\{i_1,i_2,...,i_m\}$ set of items. A related task data D is a collection of database transactions, where each transaction is a set of T $T\subseteq I$, so. Each transaction is an identifier, called TID. Let A be a set of transaction, $A\subseteq T$ T contains A if and only if. Association rules are shaped implication $A\Rightarrow B$, such as one $A\subset I, B\subset I$, and $A\cap B=\varnothing$. Rule $A\Rightarrow B$ D in the transaction set, with the support of S, where s is the D transaction contains the percentage of $A \cup B$, namely $P(A,B)$. Rule $A\Rightarrow B$ D C has confidence in the transaction set, where C is contained in the D A transaction also includes a percentage of B, namely $P(B|A)$.

$$Support(A\Rightarrow B)=P(A\cup B) \tag{3}$$

$$Confidence(A\Rightarrow B)=P(B\mid A) \tag{4}$$

The support and confidence are two important concept description of association rules, the former for statistical measure of the importance of association rules in the data, said the rules which is used to measure frequency; credible degree of association rules, said the strength of the rules. In general, only the support and confidence of association rules are high may be only the interesting rules, useful. Association rules mining is mainly realized by the two steps:

Step 1, according to the minimum support degree to find the database in D all the frequent item sets.

Step 2, according to the frequent item sets and minimum confidence generated Association rules.

Task one step is to quickly and efficiently find all frequent item sets in D, is the central problem of the association rule mining algorithm of association rules mining, is a measure of the standard; step two, relatively easy to achieve, so now all association rules mining algorithm is designed for the first step forward.

7 The matrix vector association rules algorithm

Definition 1: Boolean matrix: transaction set on the database and project sets a Boolean matrix representation. The specific method for each transaction set in a row, item sets are arranged according to a column. The transaction said row vector, the project said column vector respectively, if the first I project in the j transaction, then matrix line j, the I column value is 1, or 0, Boolean matrix called the matrix for the database.

Definition 2: Vector inner product: for any two n dimensional vector $\alpha=\{x_1,x_2,...x_n\}$, $\beta=\{y_1,y_2,...y_n\}$, $<\alpha,\beta>=\sum_{i=1}^{n}x_iy_i$ is defined as the inner product of α and β.

Lemma 1: Any k of item in the set C_{k-1} is a superset of C_k.

Lemma 2: Frequent item sets corresponding to the n dimensional row vector 2 transaction database in D lemma and Boolean matrix R in each row vector inner product results do not exceed the number of frequent item sets containing project.

Definition 3: Column vector counting: counting is the column vector sum of a column element. Similarly, a row vector count is the sum of a line element.

Lemma 3: If the Boolean matrix a column vector count is less than the minimum support count, then delete this column. (Transaction compression).

Lemma 4: In a k-frequent item sets, if the Boolean matrix a row vector count less than k, then delete (Project compression).

Definition 4: Calculation γ_i method: R in a tuple vector α_i (i=1,2,..., M, and the line scanning $a\neq$-1), $<\alpha_i,\alpha_j>=<\alpha_i,\alpha_i>$ the number of calculated γ_i, j=1,2,..., *In M* and $j\neq I$.

8 Algorithm design

Input: All the dynamic behavior of the user to query the URL user input information.

Output: Recommended URL possible next set.

1) For cleaning and pretreatment of all the dynamic behavior of the user context information, see Table 2, picking out each record in the User ID, Danymic Behavior

Time, Search URL, Page-Stay Time, Save-Page, Print-Page, Favorites field.

2) To establish the matrix of Markov model is selected to identify each user: User-ID. According to the users to search the record time, sorting out the user's Search URL front to back Mark off sequence, with the users to search the record all Search URL matrix row and column, statistics of the users in each of the current Search URL to the other Search URL jump number, ratio of the number of the users and the total number of jumps as state the transfer matrix in the position value. In order to establish the Mark off state of each user transfer matrix. Among them, each gear position matrix is set to User-ID_Search URL, each column head position is set to Search URL, matrix stored in the Hbase table in user Shift Matrix.

3) Mark-off state transition matrix weighted: select Page-Stay Time (page retention time, in seconds) as the state transfer of additional weight matrix an element value calculation of one of the conditions, if Page-Stay Time (0,30), in the corresponding matrix elements *1, if Page-Stay Time (30,60), in the matrix corresponding to the element value * (1+1/20), if Page-Stay Time (60,), in the matrix corresponding to the element value * (1+2/20); one additional value selection of Save-Page, Print-Page, Favorites as the state transition matrix elements corresponding value calculation, as long as one parameter value is 1, the corresponding matrix element values * (1+2/20).

4) The similarity calculation method based on the cosine factor: interest model parameters in Table 3, the interest in the registered information in the U keyword user selected as vector u, the similarity between the user employs cosine score vector measurement, cosine value is greater, the higher the similarity between users. Set (U,V) that similarity of $sim(u,v) = \cos(u,\mathrm{v}) = \frac{u \cdot v}{|u||v|}$ user u and user v.

5) Results: recommended for querying URL user input, transfer URL possible matrix find next in the current user owned by the state, is greater than a certain threshold (e.g. 0.1000) probability values that are consistent with the conditions recommended values. In addition, choose the maximal similarity of N users in the input current query URL in the transition matrix of the user's requirements to meet the threshold of a probability value as the recommended URL reasonable set.

TABLE 2 Parameters of dynamic behavior

Table name	**Danymic Behavior_T**			**Dynamic behavior description table dynamic behavior description table**	
Explain	This table to record the dynamic behavior description, representing the dynamic behavior by the following parameters, behavior analysis				
Primary key	Danymic Behavior Time				
The field name	Data type	Whether can be empty	The field description	Default value	Remarks
Danymic Behavior Time	Date time	No	The time of their occurrence		PK
User ID	Var char(10)	No	The current user ID		
Page Title	Text	No	The current page title		
Search URL	Text		Search URL for the current user The current user search conditions corresponding to the URL	Null	
Page-Stay Time	Int		Page retention time	0	Represents a user from entering into the page to the current trigger time duration of stay
Save Page	Char(1)		Do you want to save the page	0	0 not preserved, 1 saves
Print Page	Char(1)		Whether the print page	0	0 means no print, 1 said the print
Favorites	Char(1)		Whether to add to Favorites	0	0 do not add, add 1 representation
Copy Or CutContent	Text		Copy or cut the page content		

TABLE 3 Interest model parameters

Table name	**Favorites Model_T**			**Interest degree model table**	
Explain	This table records the user interest model description (each for a period of time, six months or a year, must put the backup data)				
Primary key	User ID				
The field name	Data type	Whether can be empty	The field description	Default value	Remarks
User ID	Var char(10)	Whether or not	Online user ID		Matching with the online user ID type
Favorite Field	Text		The user's domain of interest		
UserField	Text		User field		
Deduction	Text		User current interest keywords		

9 Application examples

Using the above research results, relying on the research project, the analysis engine system user behavior based on cloud computing is a design and development.

The analysis engine platform running on the Ubunt12.10 user behavior, the software mainly includes: jdk-1.7.0_11, Jena-2.6.4, Myeclipse-8.0, Hive-0.10.0, HBase-0.94.4, Hadoop-1.0.4, Tomcat-6.0, Jquery-1.6, Spring-3.0, Struts2-2.2.1, the browser is above IE8.0. The context aware behavior analysis as an example, system operation is as follows.

The system calculates the Mark off matrix intermediate important, as shown in Figure 2.

Select a user's current URL, system using Markov model and collaborative filtering, given URL probably next, as shown in Figure 3, Figure 4.

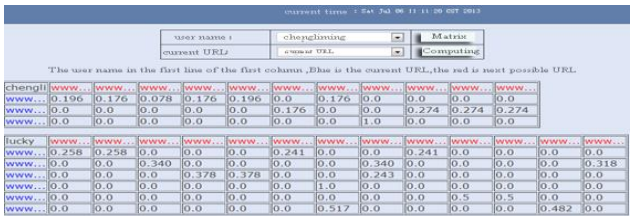

FIGURE 2 Markoff matrix

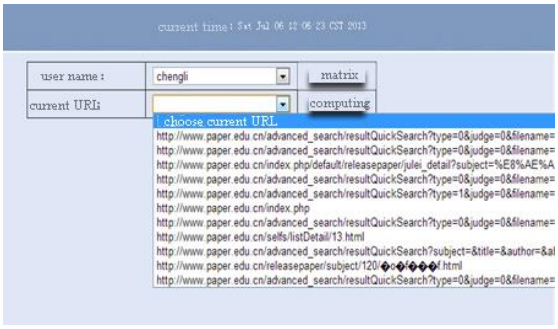

FIGURE 3 To select the current URL

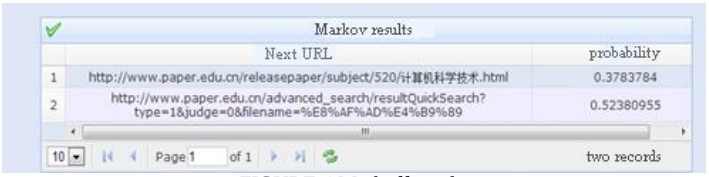

FIGURE 4 Markoff results

10 Research and prospect

In order to improve the retrieval system query processing ability, we have proposed user behavior analysis engine based on the use of cloud environment in which MapReduce parallel computing model, HBase cloud storage capacity, and uses the relevant data mining algorithms, static analysis, dynamic monitoring of user behavior characteristics and synthesis reasoning are included.

That can efficiently push their interested information and provide the basis for the site structure adjustment for the user.

To sum up, this work is currently only integrating three cloud platf 7gvvborm of the data mining model, looking for more scene data mining model, integrated with the cloud platform. Making the system more universal, is the next step of work to do.

Acknowledgments

The National Natural Science Foundation of China (U1204609), the Education Department of Henan Province Science and Technology Key Project (14A510011), the Youth Science Foundation of Henan Normal University (2012QK21).

References

[1] Voas J, Zhang J 2009 Cloud Computing: New Wine or Just a New Bottle **11**(2) 15-7
[2] Youseff L, Butrieo M, Da Silva D 2008 Toward a Unified Ontology of Cloud Computing Grid Computing Environments Workshop.GCE'08 12-16 Nov 1-10
[3] Grossman R L 2009 The Case for Cloud Computing IT Professional **11**(2) 23-7
[4] Mika P, Tununarello G 2008 Web Semantics in the Clouds Intelligent Systems IEEE **23** 82-7
[5] Hewitt C 2008 ORGs for Sealable, Robust, Privacy-Friendly Client Cloud Computing Internet Computing IEEE **12** 96-9
[6] Jepson T C 2004 The basics of reliable distributed storage networks IT Professional **6**(3) 18-24
[7] Beny M W, Do T, O Brien GW, Krishna V, Varadhan S 1993 SVDPACKC (Version 1.0) User, 5 Guide University of Tennessee
[8] Foster I, Yong Zhao, Raicu I, Lu S 2008 Cloud Computing and Grid Computing 360-Degree Compared Grid Computing Environments Workshop GCE 1-10