International Journal of Database Theory and Application
Vol.7, No.4 (2014), pp. 227-236
http://dx.doi.org/10.14257/ijdta.2014.7.4.18

Quality Evaluation Method of Undergraduate Thesis References Based on Hadoop

Jie Wang, Jiwei Liu, Gang Hou, Yanshuo Yu and KuanJiu Zhou*

School of Software Technology
Dalian University of Technology, Dalian, China, 116620
wang_jie@dlut.edu.cn, liujiwei_dut@mail.dlut.edu.cn, hg.dut@163.com

Abstract

References can directly reflect quality of Undergraduate theses. However, it is a thorny problem that how to quantitatively evaluate massive undergraduate thesis references. A quality evaluation method of undergraduate thesis references based on Hadoop is presented which applies pattern matching algorithm to extract metadata and evaluates quality of undergraduate theses by quality evaluation model based on grey relational analysis. Experiment results show that this method can quickly and accurately evaluate quality of undergraduate thesis references on Hadoop.

Keywords: *Metadata extraction; Normative Evaluation; Grey Relation Analysis; Quality Evaluation; Hadoop*

1. Introduction

With development of cloud computing and big data processing, Hadoop has begun to receive a more and more attention as the best solution to the massive data analysis. Many researchers have applied Hadoop to conduct related research. Y. R. Zhao et al. [1] proposed a join query processing algorithm in parallel which can effectively improve the processing efficiency of join queries. Y. Lee et al. [2] developed a hadoop-based packet trace processing tool, which used a new binary input format and an efficient traffic analysis MapReduce job model. With the help of MapReduce and HDFS, they achieve a better performance to process a large set of packet data. To automatic clustering analysis large scale textual datasets, Z. W. Cao et al. [3] paralleled the classical Jarvis-Patrick (JP) algorithm on the Hadoop. All the work they did with the Hadoop has achieve better results.

According to the survey, there are 6.99 million undergraduates in 2013 in our country, an increase of 19 million than in 2012. Assuming that each undergraduate needs to complete a thesis with the size of about 2M, then the size of the all the theses in 2013 are about 12TB. In order to evaluate massive undergraduate theses, a quality evaluation method of undergraduate thesis references based on Hadoop is presented. References contain important information which can help us quickly locate the research directions and evaluate quality of theses.

2. Background

Recently, some related researches on automatic extraction of the reference metadata have been done. The metadata extraction approaches can be divided into two categories: template matching and machine learning based approaches. A template matching approaches take an input citation and match its syntactic pattern against known templates. The template with the best fit to the input is then used to label the citation's tokens as fields. The canonical example

ISSN: 2005-4270 IJDTA

of a template based approach is Para Tool. Disadvantage of Para Tool is that the result it generates is on the basis of available template [4]. M. Y. Dayet *et al.*, [5] applied a hierarchical knowledge representation framework called INFOMAP to automatically extract the reference metadata. The machine learning based approaches apply the learning algorithm to obtain the pattern model of the reference strings from the set of the training documents. At present, the main machine learning algorithms applied to text information extraction are Hidden Markov Models (HMMs) [6] and Conditional Random Fields (CRFs) [7]. J. Laiet *et al.*, [8] proposed a semantic-block-based hidden Markov model, which can effectively increase the recall and precision of information extraction under certain circumstances. M. Romanello *et al.*, [9] developed a parser which recognizes word level n-grams of a text as being canonical references by using a CRF model trained with both positive and negative examples. They demonstrated the suitability of Conditional Random Fields (CRF) for extracting this particular kind of reference from unstructured texts. A. Kovacevic *et al.*, [10] developed a system based on machine learning, which can perform automatic extraction and classification of metadata in eight pre-defined categories. The potential problems of parsing the references include data entry errors, diverse citation formation, common author names and the large-scale citation data.

In addition to automatically extracting the reference metadata, the researchers are also interesting in analyzing the references. N. Vallmitjana *et al.*, [11] analyze 46 doctoral theses presented at the Institut Químic de Sarriá (IQS) from 1995 to 2003 to ascertain the frequencies of the document types in the research process. Some researchers try to use the statistical methods to analyze the references. They analyze the references usage of national excellent doctoral dissertation between 1999 and 2004. These analysis methods are relatively simple, which can only help us to understand the overall situation of references usage. They cannot quantitatively evaluate the quality and the standardization of the theses.

With the research background above, this paper proposed a quality evaluation method of undergraduate thesis references based on Hadoop, which applies pattern matching algorithm to extract the metadata and used an evaluation model based on the grey relational analysis to evaluate the qualities of undergraduate theses.

The rest of the paper is organized as follow. Section 3 describes a method of qualitatively evaluating the normative of references. Section 4 illustrates the evaluation model based on the grey relational analysis. Section 5 presents experimental results by applying the proposed method, and is followed by the conclusion and future work in section 6.

3. Reference Normative Evaluation Method

The basic idea of evaluating the reference normative in this paper is applying the pattern matching method to extract the metadata. While extracting, if the reference contain errors, we fix them and reduce some scores which based on the metadata weight we set. And by using the Hadoop parallel framework, the process of extracting is significantly accelerated.

3.1. Select the Metadata and Set the Weight

Reference documents in the undergraduate theses are conformed to the GB/T 7714-2005[12]. Different reference documents have diffident metadata. The main metadata structure mainly contains 4 parts, including authors, titles, types, and publication information. Meanwhile, there is more sub metadata in the types and publication information. Before extracting and evaluating, a weight should be given to all the metadata in the references. To avoid subjectivity, the following algorithm is used to calculate the weight.

Given a set of metadata $T = \{T_1, T_2 \dots T_n\}$, each of the set T_i ($1 \le i \le$ n) indicates a metadata template pattern. The template style library $S = \{S_1, S_2 \dots S_n\}$ is also defined, S_i ($1 \le i \le m$) represents a type of reference template pattern, which is consists of elements in the set T with specific delimiter composition. The set S includes all the standardized templates of the undergraduate thesis references. Each reference is regarded as a set $R= \{w_1, w_2 \dots w_p\}$, w_i ($1 \le i \le p$). The set R represents a string which is separated by some separator character such as comma, dot *etc*.

According to the GB/T 7714-2005, the reference documents should include the reference type and should be put between the symbol "[" and "]". So before parsing, the template pattern S_i can be determined.

When setting the weight, a certain number of theses are selected as the training set. Each entry of the thesis references is saved as a string. Each reference string is parsed by the corresponding template pattern S_i, and then the metadata template T_{ij} ($1 \le i \le n$, $1 \le j \le k$) can be extracted. The weight of the T_{ij} can be calculated by the formula 1:

$$p_{ij} = \ln\left(\left(N_i - N_{ij}\right) + 10\right) \quad (1)$$

Where N_i is the number of the occurrences of reference type i and N_{ij} is the number of occurrences of the metadata j in the reference i.

In order to guarantee that the sum of all metadata weight in each type of references is 100, the formula 2 is applied to calculate final weight value r_{ij}.

$$r_{ij} = \frac{p_{ij} \times 100}{\sum_{j=1}^{k} p_{ij}} \quad (2)$$

Where $\sum_{j=1}^{k} P_{ij}$ is the sum of all the metadata's weight in the reference type i. p_{ij} is the weight of metadata j in the reference i.

3.2. Metadata Extraction Model Based on Pattern Matching

As undergraduate thesis reference is usually not written in the standard format, so if simply using the pattern matching algorithm proposed in section 3.1 to extract the metadata, the result is not very effective and it is also difficult to quantitatively evaluate the normative of the reference. Based on the pattern matching algorithm, an improved algorithm has been proposed in this section. The improved algorithm can fix the errors in the references before extracting the metadata. The algorithm is detailed bellow.

For the errors of punctuations format, dot loss and extra space *etc*. A basic normative score α is assigned to each reference. While fixing the errors, the number of errors is recorded. Then the final score of the normative is calculated by formula 3:

$$\alpha = 10 - a \times h_a - b \times h_b - c \times h_c \quad (3)$$

Where a is the number of error punctuations, h_a is the score reduced by the error of punctuations. b is the number of the extra space, h_b is the score reduced by the error of extra space. c is a flag, if there is a dot in the end, then c=0, Otherwise c=1. h_c is the score reduced by the error of dot loss.

For the errors of punctuations, metadata loss *etc*, a set of reference types is selected, $R=\{R_1,R_2,\dots, R_n\}$, and each R_k represents a type of reference. Each R_k corresponds to a

template pattern library $S^k = \{S_1^k, S_2^k, \cdots, S_m^k\}$, S_i^k represents the *i*-th template pattern of the reference type *k*. In the template pattern library, there is only one completely standardized template pattern. The score of each template is $score_i^k (0 \le score_i^k \le 1)$. Only the completely standardized template has a score 1, and others that contain errors are less than 1.

For the entries of the references, the type of the references is read firstly, and then the right template pattern library is chosen to parse it. The metadata T_j in the template pattern S_i^k has a similarity score, and the formula is as follows:

$$f(S_i^k, T_j) = \begin{cases} r_{ij} & \text{can parse } T_j \\ 0 & \text{cannot parse } T_j \end{cases} \quad (4)$$

When the metadata T_j can be parsed by using the template pattern S_i^k, the similarity score is the value of the weight r_{ij}. Otherwise the similarity score is 0.

In order to determine the template, the similarity of the reference to each template pattern should be calculated. The formula is as follows:

$$sim(S_i^k) = \sum_{j=1}^{n} f(S_i^k, T_j) \quad (5)$$

Then normative score of this reference can be calculated by formula 6.

$$F_{ref} = \alpha + 0.9 \times sim(S_{\max}^k) \times score_{\max}^k \quad (6)$$

Where the $S_{\max}^k$ is the most similar to the template pattern. $score_{\max}^k$ is the score of the $S_{\max}^k$. α is the basic normative score.

The average normative score of all the references can be used to represent the normative of each undergraduate thesis reference. The formula is as follows:

$$F_{nor} = \frac{\sum_{i=1}^{sum} F_{ref}^i}{sum} \quad (7)$$

Where the $\sum_{i=1}^{sum} F_{ref}^i$ is the sum of the normative score of all the undergraduate thesis references and the variable sum is the total number of this undergraduate thesis references.

Finally, according to the reference type, the template with the maximum similarity value is applied to extract the metadata.

4. Quality Evaluation Model Based on Grey Relational Analysis

Grey relational analysis is a branch of the Grey System Theory. Applying the gray relational analysis method [13-14] to comprehensively evaluate the things which are affected by many factors is a very useful approach. Therefore, this paper applies the gray relational analysis method, with five quality evaluation indexes, to evaluate the undergraduate thesis reference.

The five evaluation indexes are the number of references *L*, the journals *J*, the English journals *E*, the average year *Y* and the empty reference *O*. Assuming that there are n undergraduate theses in the system and each undergraduate thesis has *m* evaluation indexes. Then we can get the following attributes Matrix is

$$X' = \begin{bmatrix} x'_{11} & x'_{12} & \cdots & x'_{1n} \\ x'_{21} & x'_{21} & \cdots & x'_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ x'_{m1} & x'_{m2} & \cdots & x'_{mn} \end{bmatrix} = (x'_{ij})_{m \times n} \quad (8)$$

Where x'_{ij} is the *i*-th evaluation index of the *j*-th thesis.

The excellent vector is defined as follows:

$$\vec{G} = (g_1, g_2, \ldots, g_m) \quad (9)$$

Where the g_i is the maximum value of the *i*-th evaluation index in each thesis.

And the poor vector is defined as follows:

$$\vec{B} = (b_1, b_2, \ldots, b_m) \quad (10)$$

Where the g_i is the minimum value of the *i*-th evaluation index in each thesis.

The $\vec{X}_j = (x_{1j}, x_{2j}, \ldots, x_{mj}), (j = 1, 2, \ldots, n)$ represents the vector composed of each evaluation index in the *j*-th thesis. Then the correlation coefficient on the *i*-th evaluation index between $\vec{X}_j$ and $\vec{G}$ can be calculated by formula 11.

$$\zeta_i(\vec{X}_j, \vec{G}) = \frac{\min\limits_j \min\limits_i \left| x_{ij} - g_i \right| + \rho \max\limits_j \max\limits_i \left| x_{ij} - g_i \right|}{\left| x_{ij} - g_i \right| + \rho \max\limits_j \max\limits_i \left| x_{ij} - g_i \right|} \quad (11)$$

Where the $\zeta_i(\vec{X}_j, \vec{G}) \in [0,1]$ and the ρ is the identification coefficient, $\rho \in (0, 1)$, and Deng[15] stated that the value of 0.5 is normally applied. The $\min\limits_j \min\limits_i \left| x_{ij} - g_i \right|$ is the minimum absolute value of the difference between x_{ij} and g_i . The $\max\limits_j \max\limits_i \left| x_{ij} - g_i \right|$ is the maximum absolute value of the difference between x_{ij} and g_i.

The correlation coefficient on the *i*-th evaluation index between $\vec{X}_j$ and $\vec{B}$ can be calculated by formula 12:

$$\zeta_i(\vec{X}_j, \vec{B}) = \frac{\min\limits_j \min\limits_i \left| x_{ij} - b_i \right| + \rho \max\limits_j \max\limits_i \left| x_{ij} - b_i \right|}{\left| x_{ij} - b_i \right| + \rho \max\limits_j \max\limits_i \left| x_{ij} - b_i \right|} \quad (12)$$

Where the $\min_j \min_i \left| x_{ij} - b_i \right|$ is the minimum absolute value of the difference between x_{ij} and b_i. The $\max_j \max_i \left| x_{ij} - b_i \right|$ is the maximum absolute value of the difference between x_{ij} and b_i.

The relational grade between the evaluation vector $\vec{X}_j$ of j-th thesis and the excellent vector $\vec{B}$ is defined as follows:

$$\gamma(\vec{X}_j, \vec{G}) = \sum_{i=1}^{m} w_i \zeta_i(\vec{X}_j, \vec{G}) \tag{13}$$

Where the w_i is the weight of the i-th evaluation index.

The relational grade between the evaluation vector $\vec{X}_j$ of j-th thesis and the poor vector $\vec{B}$ is defined as follows:

$$\gamma(\vec{X}_j, \vec{B}) = \sum_{i=1}^{m} w_i \zeta_i(\vec{X}_j, \vec{B}) \tag{14}$$

Where the w_i has the same meaning with above.

Assuming the $\vec{X}_j$ is belonging to the $\vec{G}$ with the coefficient u_j, then the $\vec{X}_j$ is belonging to the $\vec{B}$ with the coefficient (1-u_j). In order to establish the comprehensive quality evaluation model of the undergraduate thesis reference, the classical least squares criterion is expanded and the target function is defined as follows:

$$\min \left\{ F(\vec{u}) = \sum_{i=1}^{m} \left\{ \left[(1-u_j)\gamma(\vec{X}_j, \vec{G}) \right]^2 + \left[u_j \gamma(\vec{X}_j, \vec{B}) \right]^2 \right\} \right\} \tag{15}$$

Where the $\vec{u}$ is the optimal solution vector: $\vec{u} = \{u_1, u_2, \ldots, u_n\}$.

In order to minimum the target function $F(\vec{u})$, we can get the relationship as follows:

$$\frac{\partial F(\vec{u})}{\partial u_j} = 0 \tag{16}$$

The bigger the u_j is, the higher the quality of the reference.

Finally the comprehensive evaluation value is calculated as follows:

$$F_j^{qua} = 100 \times u_j = \frac{100}{1 + \left[\frac{\gamma(\vec{X}_j, \vec{B})}{\gamma(\vec{X}_j, \vec{G})} \right]^2} \tag{17}$$

Where F_j^{qua} is the comprehensive evaluation value of the j-th thesis, $F_j^{qua} \subseteq [0,100]$. The bigger the F_j^{qua} is, the higher the quality of this thesis references is.

5. Experimental Results

In this paper, 1200 undergraduate theses were used as the sample data. In the experiment, two operation model of the Hadoop were build: pseudo-distributed mode and real distributed mode. In order to verify the efficiency of the Hadoop platform, a stand-alone program is designed as the comparative experiment.

5.1. Result of the Statistics and Evaluation

Applying the method proposed in Section 3.2 to extract the metadata from the undergraduate thesis references, the rate of the extraction is about 74.42%. Then the normative score of each thesis can be calculated. According to the result, the interval distribution pie chart of normative score is drawn in the Figure 1.

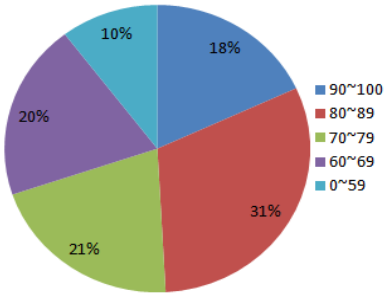


Figure 1. Distribution of the Reference Normative Score

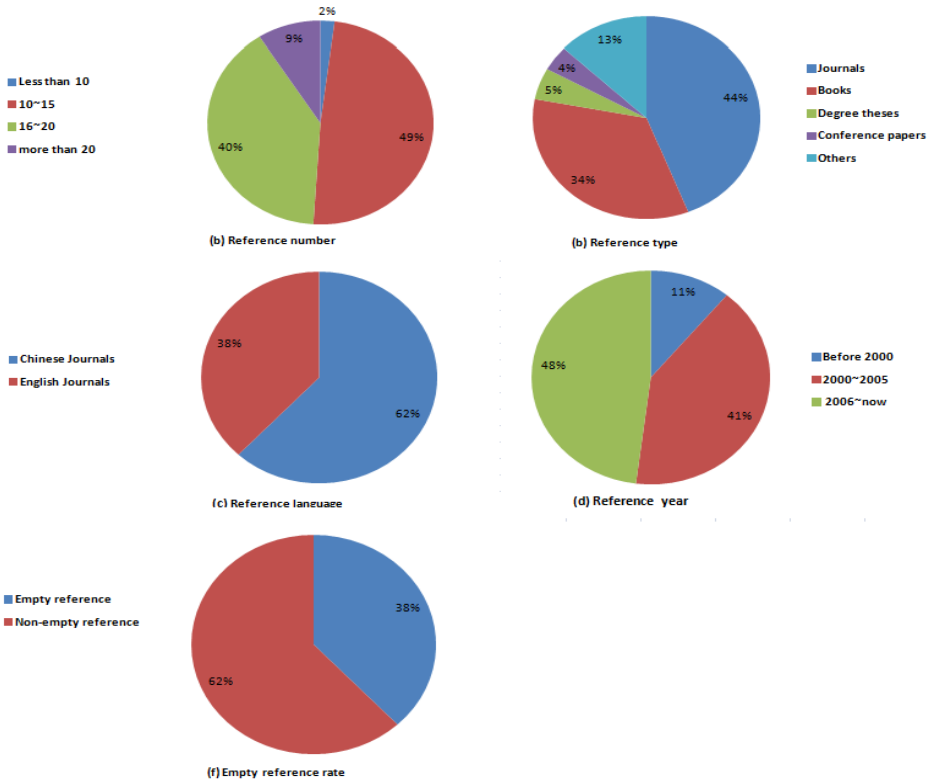


Figure 2. Result of the Statistic

Figure 1 shows that half of the students have a score 80 or more, which indicates that most students pay more attention to the normative of the references. However, the students who get score 70 or lower are about 30%. These students don't know the importance of the reference, and there are lots of errors in their theses.

Statistical information of the five evaluation indexes is shown in Figure 2. The comprehensive evaluation value calculated by applying the evaluation algorithm based on gray relational analysis can reflect the quality merits relationship of various theses. If the value is greater than the lower limit, then the thesis is qualified. Otherwise, it is unqualified. The pass rate is shown in Figure 3(a).

In order to explore the impact factor of the reference quality, a comparative experiment is conducted. The empty reference evaluation index is deleted and the score is recalculated. Then the result is shown in Figure 3(b), which indicates that the number of empty reference is the key impact factor.

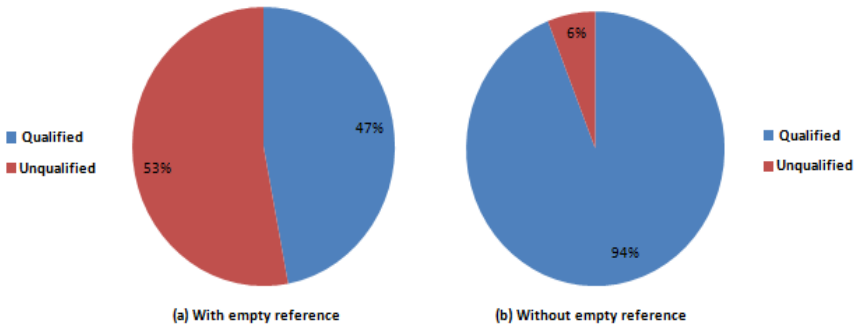


Figure 3. Pass Rate of Reference Quality

5.1. Result of the Statistics and Evaluation

The proposed method is applied to three experimental environments. In the experiments, the running time with different data size is recorded and the relationship between run-time environment and the data size is shown in Figure 4.

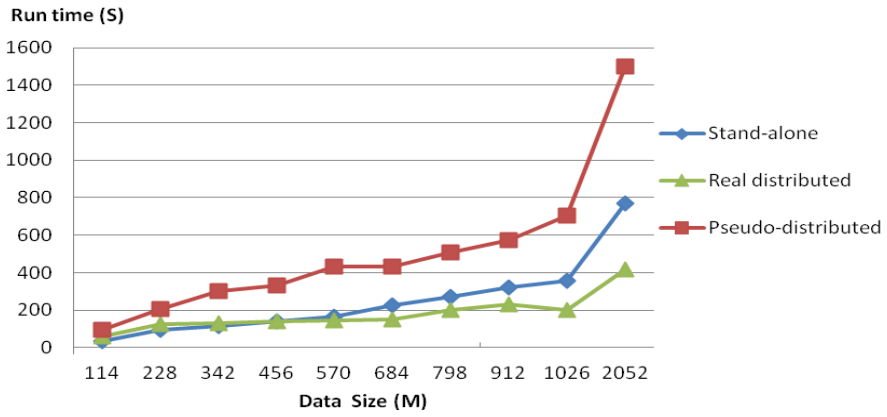


Figure 4. Comparison of Three Run-time Environment

Figure 4 shows that when the data size is small, the stand-alone program is faster, without data transmission and communication overhead. However, when the amount of data

increasing, the MapReduce in the real distributed environment can evenly distributed computing tasks to diffident computers, which can greatly improve the performance of the system. In our experiment, the largest amount data is GB. When the amount of data is upon to TB, then the system will has a better performance, taking advantage of its efficient parallel computing capabilities and storage capacity of massive data.

6. Conclusions

This paper proposes a quality evaluation method of undergraduate thesis reference based on Hadoop, which applies the improved pattern matching algorithm to extract the metadata and uses quality evaluation model based on grey relational analysis to quantitatively evaluate the undergraduate thesis reference. Meanwhile, the algorithm is realized on Hadoop platform, which significantly accelerates the process. According to the experimental result, our approach can quickly analyze massive thesis references and can give a relatively accurate evaluation.

Future work includes improving the evaluation algorithm by identifying other possible indexes of the theses and increasing the statistical measure of teachers, professional, college, university *etc* in the statistical analysis, as an indicator of teaching quality evaluation.

Acknowledgements

This work was supported by National Nature Science Foundation of China with fund code of 61272174 and 61202443. We also would like to acknowledge helpful suggestions and comments from Professor He Guo.

References

[1] Y. R. Zhao, W. P. Wang, D. Meng, S. B. Zhang and J. Li. Efficient Join query processing algorithm CHMJ based on Hadoop. Journal of Software, vol. 23, no. 8, **(2012)**, pp. 2032-2041.
[2] Y. Lee, W. Kang and Y. Lee. A hadoop-based packet trace processing tool. Traffic Monitoring and Analysis, Springer Publishers, Berlin, vol. 6613, (2011), pp. 51-63.
[3] Z. W. Cao and Y. Zhou. Parallel Text Clustering Based on MapReduce. Cloud and Green Computing (CGC), 2012 Second International Conference on, Xiantan, China, November, **(2012),** pp**.** 226-229
[4] C. C. Chen, K. H. Yang, C. L. Chen and J. M. Ho. BibPro: A citation parser based on sequence alignment. Knowledge and Data Engineering, IEEE Transactions on, vol. 24, no. 2, **(2012)**, pp. 236-250.
[5] M. Y. Day, R. T. H. Tsai, C. L. Sung, C. C. Hsieh *et al*., Reference metadata extraction using a hierarchical knowledge representation framework. Decision Support Systems, vol. 43, no. 1, **(2007)**, pp. 152-167.
[6] E. Hetzner. A simple method for citation metadata extraction using hidden markov models. Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries, New York,USA, June **(2008)**, pp. 280-284.
[7] F. Peng and A. McCallum. Information extraction from research papers using conditional random fields. Information Processing & Management, vol. 42, no. 4, **(2006)**, pp. 963-979.
[8] J. Lai, Q. Liu and Y. Liu. Web information extraction based on hidden Markov model. Computer Supported Cooperative Work in Design (CSCWD), 2010 14th International Conference on, Shanghai, China, April **(2010),** pp. 234-238.
[9] M. Romanello, F. Boschetti and G. Crane. Citations in the digital library of classics: extracting canonical references by using conditional random fields. In Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries, PA, USA, August **(2009)**, 80-87.
[10] A. Kovacevic, D. Ivanovic, B. Milosavljevic, Z. Konjovic., Automatic extraction of metadata from scientific publications for CRIS systems. Program: electronic library and information systems, vol. 45, no. 4, **(2011)**, pp. 376-396.
[11] N. Vallmitjana, L. G. Sabaté. Citation analysis of Ph. D. dissertation references as a tool for collection management in an academic chemistry library. College & Research Libraries, vol. 69, no. 1, **(2008)**, pp. 72-82.
[12] GB/T 7714－2005, Rules for content, form and structure of bibliographic references **(2005)**.

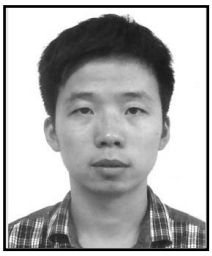
[13] M. L. Tseng. Using linguistic preferences and grey relational analysis to evaluate the environmental knowledge management capacity. Expert systems with applications, vol. 37, no. 1, **(2010)**, pp. 70-81.
[14] J. W. K. Chan and T. K. L. Tong. Multi-criteria material selections and end-of-life product strategy: Grey relational analysis approach. Materials & Design, vol. 28, no. 5, **(2007)**, pp. 1539-1546.
[15] J. L. Deng. The Fundamentals of grey theory. Huazhong University of Science and Technology Press, Wuhan **(2002)**.

Authors

Jie Wang, he was born in Dalian of China in 1979. He obtained PhD of computer architecture at Harbin Institute of Technology, China in 2009. He entered School of Software Technology, Dalian University of Technology as a lecture until now. He has published more than 20 papers. He interests parallel computing, network security and trusted software. Dr. Wang is a committee member of CCF YOCSEF (Dalian).

Jiwei Liu, he was born in Jiangsu of China in 1989. He is a graduate student in School of Software Technology, Dalian University of Technology. He interests in parallel computing and FPGA program technology.

Gang Hou, he was born in Liaoning of China in 1982. He is a Ph.D. candidate in School of Software Technology, Dalian University of Technology. He interests in distributed computing and Trusted Software.



Yanshuo Yu, he was born in Dandong of China in 1989. He is a graduate student in School of Software Technology, Dalian University of Technology. He interests in parallel computing.

Kuanjiu Zhou, he was born in Liaoning of China in 1966. He is a professor in School of Software Technology, Dalian University of Technology. He interests in Trusted Software and data mining.

