Automatic Indexing for Research Papers Using References

Wei Liu
Institute of Scientific and Technical Information of China
China, 100038
Email: liuw@istic.ac.cn

***Abstract*—An effective way to reveal the contents of research papers is assigning a group of terms against a controlled vocabulary. To the best of our knowledge, a variety of automatic indexing techniques have been studied to enhance the effectiveness and the efficiency. However, the current approaches depended on the content of a research paper, such as title, abstract, etc., which suffering from limitations on the automatic indexing performance. In this paper, we propose a new approach of automatic indexing for single research paper with its references based on Genetic Algorithm (GA). The extensive experiments on four subjects show the effectiveness of the proposed approach.**

***Index Terms*—information retrieval, automatic indexing, research paper, genetic algorithm**

I. INTRODUCTION

Research paper indexing refers to the process of assigning a group of meaningful terms against a controlled vocabulary for one research paper, which the terms can be representative of the research paper. The indexing terms can be regarded as the knowledge summary for one research paper. Indexing terms are facilitated for users to grasp the highlights without going through the whole research paper, and they can also be used as low cost measures of text similarity for research paper search system. Many applications can benefit from it, e.g., research paper retrieval, automatic classification & clustering, content summarization, topic-oriented information visualization. Obviously, manual assignment of high-quality indexing terms is expensive and time-consuming, which is almost an impossible task to handle a large number of papers annually for a library. Therefore, automatic indexing approaches are in great demand, which try to make process of assigning index terms with minimum human intervention.

To the best of our knowledge, a variety of automatic indexing approaches have been proposed for general documents[1,2,4,5,7,14,18] or special types of documents (e.g., news articles, blogs)[3,6,7,8,9,10,11]. Research papers, as a special type of documents, require more precise terms as the highly condensed summary. Most present automatic indexing approaches for research papers only make use of the main content of research papers with statistics, linguistics, machine learning and other techniques. However, they suffer from two limitations in practice. First, many papers are in form of scanned images instead of texts in the database due to some reasons. Thus the content-based approaches could not deal with such papers. Though OCR(Optical Character Recognition) techniques could be used to recognize the texts in scanned images, the accuracy is disturbing for many factors, such as image resolution and scan angle. And usually watermarking techniques[12] are used to prevent illegal copying, which also increases the difficulties of image recognition. Second, word segmentation is one of the prerequisites for analyzing the contents of research papers. It is widely known that this problem has not been well addressed for research papers Chinese). Further, one important fact ignored by the precious approaches is that research papers are the heritage and improvement of academic knowledge, which is manifested as referenced papers. As a result, the indexing terms of the references can be considered as important clues to generate the indexing terms of the paper.

Motivated by this idea, we tries to find a practical approach which can accomplish the automatic indexing task only using the papers' references. This is very important for many applications. For instance, it is always a heavy burden for a library to index a quantity of purchased research papers every year. In this way, the newly published papers can be automatically indexed if all previous papers have been assigned appropriate index terms already. The prophase survey based on the journal paper database, dissertation database and conference paper database of National Science and Technology Digital Library[13] shows the feasibility of our approach: on average, 94.2% index terms of one paper appear in the indexing terms of its references, and 4.3% indexing terms have semantic similar terms(synonym, hypernym or hyponym) in the indexing terms of its references, and only 1.5% indexing terms are completely new for the indexing terms of its references. Without the ambiguity

As a direct way, selecting high-frequency ones or low-frequency ones from the indexing terms of the references is not a good idea because high-frequency ones are too common and low-frequency ones are too rare to represent the topic of the paper. Hence, the key is how to select appropriate ones from the indexing terms of the references? To this end, we regard the indexing task as the knowledge inheritance, and the genetic algorithm is

doi:10.4304/jsw.9.11.2823-2829

adopted help us fulfill the filter process of indexing terms. We believe this is the first work which handles the automatic indexing problem from the knowledge inheritance point of view. Based on the proposed approach, a prototype system AIRPUR(Automatic Indexing for Research Paper Using References), has been implemented and is on trial. The latest experimental results will be reported in Section 5.

II. Related Works

The keywords assigned by the authors in most papers are usually are not professional for literature indexing. As a result, the indexing quality could not be stable, and the papers have to be assigned indexing terms manually by pro. Obviously, this is a labour-intensive and error-prone task. To this end, automatic indexing is being a hot topic in the literature retrieval field. Various techniques have been adopted for solutions, such as statistics, machine learning, linguistics, and so on.

The statistics-based solutions are simple for implementation, and no complicated training is necessary. Cohen et al[14] proposed a fast statistics method for generating indexing terms. This method utilizes N-gram to compute term' weight, and only the word split technique is required. Barker and Cornacchia [15] propose a system that takes into account not only the frequency of a "noun phrase" but also the head noun. In addition, other statistics-based solutions, such as words co-occurrence method[16] and PAT tree based method[17]. Yih et al[19] uses a number of features, such as term frequency of each potential indexing term and inverse document frequency to extract new indexing terms from previously unseen pages. The linguistics-based solutions generate indexing terms through lexical analysis, syntactic analysis and discourse analysis. Ercan et al[18] construct lexical chains and extract and qualify effective features from texts. The machine-learning-based solutions obtain the statistical parameters, and then extract indexing terms from samples. Zhang et al[19] use Conditional Random Fields model for training. Thuy Dung Nguyen et al[20] deduce the weights of candidate terms based on the layout structure of scientific and technical papers. [5] presents a term extraction approach which uses the features based on lexical chains in the selection of keywords for a document. Wan et al[2,4] proposes using a small number of nearest neighbor documents to improve document summarization and term extraction for the specified document, under the assumption that the neighbor documents could provide additional knowledge and more clues.

A conclusion is easy to be drawn that all the existing methods only make use of the internal features of literatures, which put forward high demand for natural language analysis techniques. As mentioned above, they heavily limited by the word splitting algorithm and OCR(Optical Character Recognition) techniques. In the left this paper, a new automatic indexing method will be present, which uses the external features to avoid the limitations of previous approaches.

III. Simple Approach Based on Term Frequency of Reference

In this section, a simple approach is introduced, which is simply based on term frequency. Intuitively, one term has more chance to be selected as the indexing term for a research paper if it appeared frequently as the indexing term in the references. So the direct way is to select the most frequent indexing terms of the references as the ones of the paper. Here we assume an indexing term appears only once in one reference. Based on such idea, the frequency of each indexing term is counted first. And then, the indexing terms are ranked by frequency. In order to make the rank more objective, citation frequency is incorporated, and the rank criteria of selecting indexing terms for a paper is as follows:

$$Rank(x) = \sum_{i=1}^{|references|} citation(ref_i) \bullet ref_i(x)$$

where $|references|$ is the number of the references of the paper, $citation(ref_i)$ is the citation number of the i*th* reference, and $ref_i(x)$ is a bool variable, where:

$$ref_i(x) = \begin{cases} 1 & x \text{ in the ith reference} \\ 0 & x \text{ not in the ith reference} \end{cases}$$

However, Simple Approach is not effective and robust enough since frequency is not the most useful factor. In many scenarios, it would bring negative efforts instead. For example, in the fast developing subjects, the indexing terms those represent current progress need to be given more consideration on recent references, but frequent indexing terms are more likely appear past references, which are not representative enough. The experimental results reported in Section 5 will prove this fact.

IV. Automatic Indexing Based on Genetic Algorithm

In this section, a new approach is proposed based on genetic algorithm, which can eliminate the defects of Simple Approach. The idea of our proposed approach is similar with that of Wan et al[2,4], which generates indexing terms from neighbor documents. The main difference is that the neighbor documents of them are obtained by using document similarity search techniques, while ours are references.

A. Genetic Algorithm

Genetic Algorithms (GA) are adaptive heuristic search algorithms premised on the evolutionary ideas of natural selection and genetic. The basic concept of GA is designed to simulate processes in natural system necessary for evolution, specifically those that follow the principles first laid down by Charles Darwin of survival of the fittest. As such they represent an intelligent exploitation of a random search within a defined search space to solve a problem. First pioneered by Holland (1975), Genetic Algorithms have been widely studied, experimented and applied in many fields[21,22]. GA provides an alternative method to solving problem that consistently outperforms other traditional methods in most of the problems link. Many of the real world

problems involved finding optimal parameters that might prove difficult for traditional methods but ideal for GA. A population of individuals is maintained within search space for a GA, each representing a possible solution to a given problem. Each individual is coded as a finite length vector of components. These individuals are linked to chromosomes and the components are analogous to genes. Thus, a chromosome comprises several genes. A fitness score is assigned to each chromosome representing the abilities of an individual to compete. The individual with the optimal or generally near optimal fitness score is sought. The GA aims to use selective breeding of the solution to produce off springs better than parents do by combining information from the chromosomes. This process is repeated until the strings in the new generation are identical or certain termination conditions are met. GA can be more easily made scalable, concurrent and robust. The procedure of GA is described as follows:

i. Choose the initial population of individuals
ii. Evaluate the fitness of each individual in the population
iii. Population evolves until termination conditions are met (time limit, sufficient fitness achieved, etc.):
iv. Select the best-fit individuals for reproduction
v. Breed new individuals through crossover and mutation operations to give birth to offspring
vi. Evaluate the individual fitness of new individuals
vii. Replace least-fit population with new individuals

From the procedure above, three crucial components have to be well designed, which are coding for individuals, fitness function, and evolutionary process.

B. Coding for Individuals

The initial population is all the references $D=\{d_1,d_2,\ldots\ldots,d_n\}$ of paper d_0. Supposing the indexing term vector is $W=<w_1,w_2,\ldots\ldots,w_m>$ and the indexing term of $d_i(i\neq 0)$ must be in W. Hence can be represented as one-dimensional vector $(x_{i1},x_{i2},\ldots\ldots,x_{im})$, where the value of x_{ij} is:

$$x_{ij}=\begin{cases}0 & w_j \notin d_i \\ 1 & w_j \in d_i\end{cases}$$

C. Fitness Function

GA is actually a process of survival of the fittest. The better genes(indexing terms) should have more possibility to be inherited to the next generation. Hence the individuals with more better genes must be assigned a larger fitness value. In the context of the problem studied in this paper, better genes can be considered as the indexing terms that can be representative of the paper.

To design the fitness function, several factors have to be taken into consideration. The first is the topic relevance. Intuitively, an indexing term would be far more representative if it appears frequently in the references and rare in the whole paper database. The second is the impact of the references. Simply, we use the citation number to express the impact of a reference. The third is the publication dates of the references. We argue that, in most cases, the innovation of a research paper is the improvement on the current progress. Usually one paper is more relevant to the current ones among the references, which is also embodied in indexing terms. Based on the three aspects above, the fitness function is designed below:

$$fitness(d_i)=lg\frac{\sum_{j=1}^{m}|w_j|*C(d_i)}{T(d_i)} \quad (1)$$

The formula includes three parts, where $|w_j|$ is the frequency of indexing term w_j that belongs to reference d_i, $C(d_i)$ is the citation number of reference d_i, and the time distance between the publication date of d_0 and its reference d_i.

D. Evolution Process

In GA, the population keeps evolving until the termination conditions are met. The population halves in each round. And the evolution process stops until there is only one individual left in the population. That is, the individuals of the current generation are paired off, and one of the next generation will be produced by two of the current generation. Hence, two questions should be answered in each evolution from the current generation to the next generation: first, how to make the individuals be paired off? Second, how to produce the individuals of next generation with the paired individuals of the current generation?

For the first question, we believe that the individuals that are relevant on indexing term should be paired off, so that the "better genes" have more chance to be inherited to the next generation. To this goal, an undirected graph model, indexing-term graph(ITG), is proposed to facilitate the pairing process. ITG is constructed as follows: for each distinct reference d_i, there exists a unique vertex $v\in V$. An undirected edge $e_{ij}\in E$ i.f.f. v_i and v_j share common indexing terms. For each e_{ij}, a number is assigned which is computed below

$$weight(e_{ij})=\sum|w_{ij}| \quad (2)$$

where w_{ij} is the shared indexing terms of d_i and d_j. By characterizing the references of one paper using ITG and the fitness, the algorithm of the evaluation for the initial generation P to the final generation P' is described in Fig. 1.

This is an iterative process. In each round, the next generation is produced according to the selecting operator and the individual fitness. The whole process stops until there is only one individual left in the current generation. The basic idea is described as follows. Firstly, undirected weighted graph G is constructed using the shared indexing terms between references during each iteration. And then, some references are selected to produce the

Population Evolution Algorithm
Input: the initial population P
Output: the final population P' //only one individual in P'
1 if only one individual in P
2 P'=P, and the algorithm ends
3 else perform steps 4-17
4 construct ITG using P
3 if V is not null, perform steps 5-17
4 select vertex v_x with the max fitness from ITG
5 select vertex v_y that has the max weight with v_x
6 initialize an individual d_{xy} as the next generation d_x and d_y;
7 for each indexing term w_j perform steps 7-
8 determine whether w_j should be in d_{xy} with probability $P(w_i)$
9 put d_{xy} into P'
10 remove v_x and v_y from V
11 remove e_{xy} from E
12 remove d_x and d_y from P
13 if P is not null
14 go to step 2
15 else
16 P=P'
17 go to step 1
18 End

Fig. 1. Population Evolution Algorithm

next generation according to the connection relation between references in G. As a result, two crucial problems have to be addressed: the construction of undirected weighted graph G and the selection of the next generation. The two issues will be discussed respectively.Using the shared indexing terms between references, the weighted undirected graph G is constructed as follows. V is the vertex set of G, and each vertex *v* in V corresponds to one conference. E is the edge set of G, and there is an edge two references if they shared indexing terms. The weight of one edge is the total number of the shared indexing between two references if they shared indexing terms. The weight of one edge is the total number of the shared indexing terms. To assure the fitness of the next generation, the higher-fitness individuals in the current generation should be selected to produce the next generation. In the algorithm in Fig. 1, the individual v_x with max fitness is selected from the left ones (see step 3), and then, the one v_y which is most similar to v_x is selected (see step 4). In step 5, the "good" genes have more probability to be passed to the next generation.

V. EXPERIMENTS

To evaluate the proposed approach, a prototype system AIRPUR(**A**utomatic **I**ndexing for **R**esearch **P**aper **U**sing **R**eferences) is developed in C++ by modifying the open source code "Genetic Algorithm Library" downloaded from Website "Code Project"(www.codeproject.com). The key components "individual coding", "fitness function" and "evolution process" of the proposed approach replace the original parts of the open source code. AIRPUR runs on the computer with Intel Core 2 Duo, 2G memory and 500G hard disk.

The comprehensive experiments are conducted over four subjects. We first show the dataset for our experiments, and then present the experiments and discuss for them.

A. Dataset

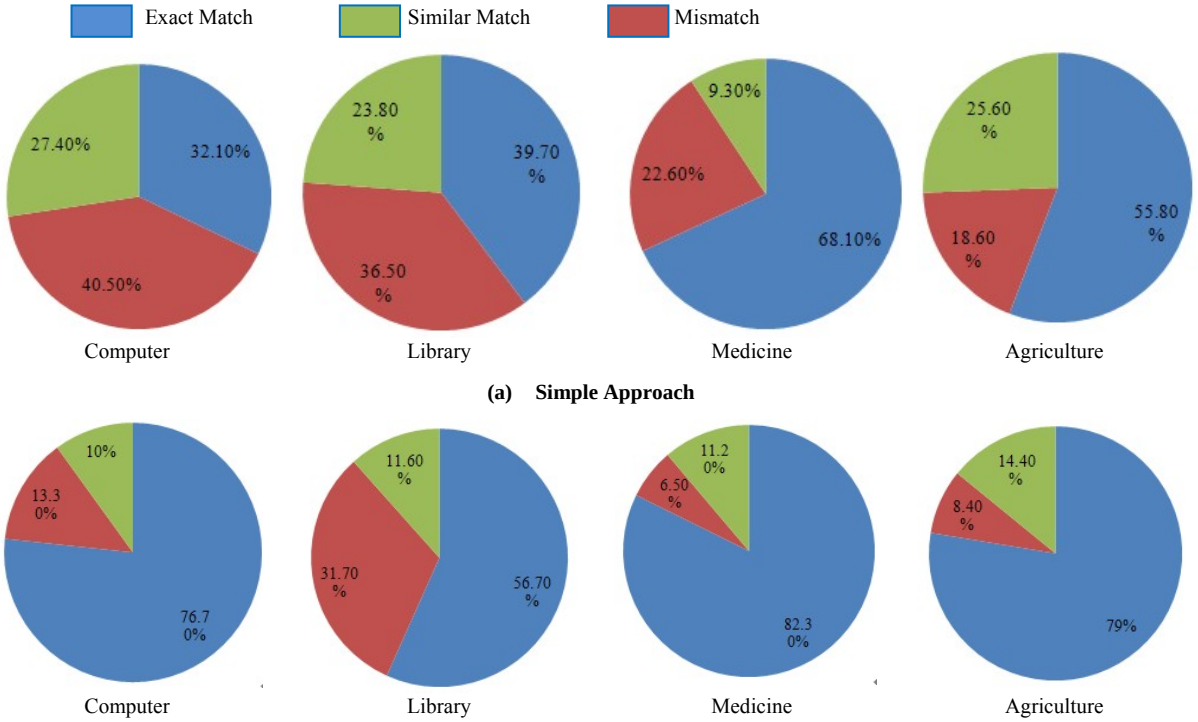
The dataset contains 40 research papers and their references from four subjects which are computer, library, medicine and agriculture. The 40 research papers are averagely selected from the 4 subjects and randomly selected between 2006 and 2010. All the papers and their references have been assigned 6 indexing terms by experts.

B. Evaluation Measures

To qualify the performance, the comparison results are classified into three parts: exact match, similar match and mismatch. In the experiments, we conducted evaluations in terms of exact match, similar match, and not match. Suppose *a* is an automatic indexing term and *b* is a manual indexing term, *a* and *b* being exact match means they are exactly the same, *a* and *b* being similar match means they are synonyms, and *a* and *b* being not match means they are fully different on semantic.

C. Performance Evaluation

We evaluated the performance of the two approaches on the dataset. The experiments are conducted over four subjects respectively, and the results are checked and compared with the indexing terms assigned by experts. By means of the evaluation measures, the experimental results are shown in Fig. 2 and Fig. 3.



(a) Simple Approach

(b) GA Appraoch

Fig. 2. The experimental results over four subjects

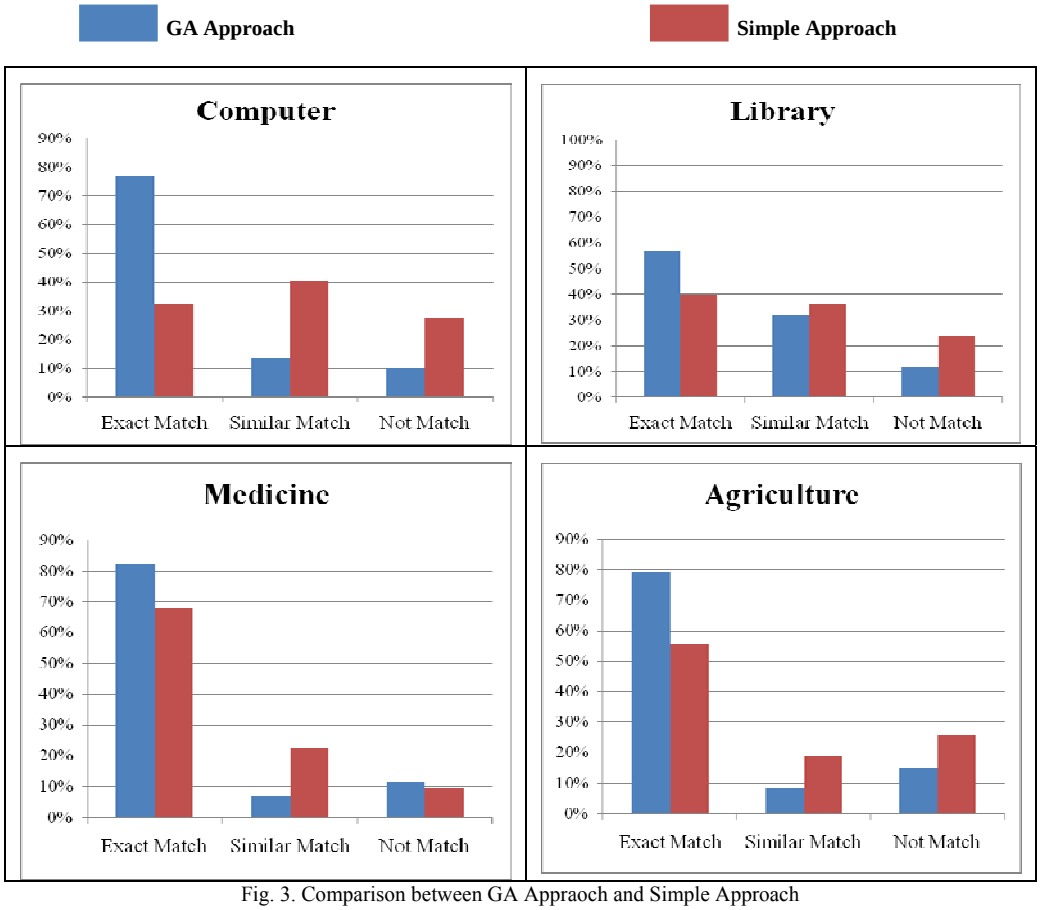


Fig. 3. Comparison between GA Appraoch and Simple Approach

Through comparing the experimental results in Fig. 2 and Fig. 3, we can reveal several interesting phenomena. First, the performance on "extract match" of the simple approach is very poor over most subjects except

Medicine. We also find that the performance on "extract match" of Simple Approach is good enough, and analytical results indicated that the indexing terms in Medicine are more formal and stricter than those in the other three subjects. Second, the performance of GA Approach is much better than that of Simple Approach, especially on "extract match". This indicates simple statistics on frequency is not effective and robust enough; while using the idea of genetic selection, the performance can be greatly improved. Third, for GA approach, the "exact match" instances make up the majority(74% of average), which is much more than the "not match" instances(12% of average) and "similar match"(15% of average). This means GA algorithm performs better on overall performance. Forth, for GA approach, the experimental result on "exact match" in Medicine is much better than those over other three subjects. Fifth, the experimental result on "not match" in Library is much worse than those over other three subjects. From the comparisons among the four subjects, we think that the terms in science are more rigorous and clear than those in arts, which lead to fewer synonyms under each concept.

VI. Conclusion

Auto indexing is always a crucial issue in the field of digital library. However, most previous approaches suffer from the serious limitations by extracting indexing terms from title, abstract, main body, and so on, which. In this paper, a new auto indexing approach is proposed for research papers by selecting indexing terms from those of the references. In the approach, the genetic algorithm is improved to select appropriate indexing terms as the ones of the research paper. The experimental results over 4 subjects show the effectiveness of the proposed approach, and the performance difference among the 4 subjects is also discussed.

Acknowledgment

This work is partially supported by National Key Technology R&D Program Grant #2011BAH10B01 and the Advanced Research Program of ISTIC, Grant #1746.

References

[1] X. Jiang, Y. Hu, H. Li. A ranking approach to keyphrase extraction[C]. In Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. ACM, 2009: 756-757.

[2] X. Wan, J. Xiao. Exploiting neighborhood knowledge for single document summarization and keyphrase extraction[J]. ACM Transactions on Information Systems (TOIS), 2010, 28(2): 8.

[3] S. N. Kim, O. Medelyan, M. Y. Kan, et al. Automatic keyphrase extraction from scientific articles[J]. Language resources and evaluation, 2013, 47(3): 723-742.

[4] X. Wan, J. Xiao. Single Document Keyphrase Extraction Using Neighborhood Knowledge[C]. In Proceedings of AAAI. 2008, 8: 855-860.

[5] G. Ercan, I. Cicekli. Using lexical chains for keyword extraction[J]. Information Processing & Management, 2007, 43(6): 1705-1714.

[6] Z. Liu, X. Chen and M. Sun. Mining the interests of Chinese microbloggers via keyword extraction[J]. Frontiers of Computer Science in China, 2012, 6(1): 76-87

[7] S. Marinai, B. Miotti and G. Soda. Digital Libraries and Document Image Retrieval Techniques: A Survey. Studies in Computational Intelligence, 2011, 375: 181-204

[8] W. Zhao, J. Jiang, J. He, et al. Topical keyphrase extraction from twitter[C]. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 2011: 379-388.

[9] F. Ferrara, C. Tasso. Extracting Keyphrases from Web Pages[C]. In Proceedings of Digital Libraries and Archives. Springer Berlin Heidelberg, 2013: 93-104

[10] W. You, D. Fontaine, J. P. Barthès. An automatic keyphrase extraction system for scientific documents[J]. Knowledge and information systems, 2013, 34(3): 691-724.

[11] W. Yih, J. Goodman, V. R. Carvalho. Finding advertising keywords on web pages[C]. In Proceedings of the 15th international conference on World Wide Web. ACM, 2006: 213-222.

[12] C. Chen, Z. Jiun, C. Hsu. An Adaptive Reversible Image Watermarking Scheme Based on Integer Wavelet Coefficients[J]. Journal of Computers, 2013, 8(7).

[13] http://www.nstl.gov.cn/

[14] J. D. Cohen. Highlights: Language and Domain-independent Auto Indexing Terms for Abstracting[J]. Journal of American Society for Information Science. 1995, 46(3): 162-174

[15] K. Barker and N. Cornacchia. Using Noun Phrase Heads to Extract Document Keyphrases[C], In Proceedings of the 13th Biennial Conference of the Canadian Society on Computational Studies of Intelligence, 2000: 40–52.

[16] Y. Matsuo, M. Ishizuka. Keyword extraction from a single document using word co-occurrence statistical information[J]. Journal of Artificial Intelligence Tools, 2004, 3(1): 157-169.

[17] L. Chien. PAT-tree-based keyword extraction for Chinese information retrieval[C]. SIGIR 1997: 50-59.

[18] Ercan G, Cicekli I. Using lexical chains for keyword extraction[J]. Information Processing & Management, 2007, 43(6): 1705-1714.

[19] C. Zhang, H. Wang, Y. Liu. Automatic Keyword Extraction from Documents Using Conditional Random Fields. Journal of Computational Information Systems. 2008, 4(3):

[20] T. D. Nguyen, M. Kan. Keyphrase Extraction in Scientific Publications[C]. In Proceedings of ICADL 2007: 317-326

[21] J. Zhao W. Li. Intrusion Detection Based on Improved SOM with Optimized GA[J]. Journal of Computers, 2013, 8(6).

[22] D. Liu, X. Chen, J. Du. A Hybrid Genetic Algorithm for Constrained Optimization Problems[J]. Journal of Computers, 2013, 8(2).

Wei Liu was born in Shandong Province of China. He received the M.S. degree in Computer Science from School of Computer Science and Technology at ShanDong University in 2004, and the Ph.D. degree in Computer Science from School of Information at Renmin University of China in 2008. He did his Postdoc research with Prof. Jianguo Xiao in Institute of Computer Science & Technology at Peking University. His current research interests include Web data extraction, Deep Web data integration, and Science and technology information retrieval.

He is currently an associate research fellow at the information resource centre, Institute of Scientific and Technical

Information of China. He has published over 20 papers in some reputational international journals, such as IEEE Transactions on Knowledge and Data Engineering, Journal of Intelligent Information Systems, etc.