ICIC Express Letters
Part B: Applications
Volume **11**, Number **9**, September **2020**

ICIC International ©2020 ISSN 2185-2766
pp. **839–845**

KEY TECHNOLOGY RESEARCH ON RETRIEVAL AUTOMATIC CONSTRUCTION FOR SCIENCE AND TECHNOLOGY NOVELTY SEARCH

Yuyan Xing and Yao Liu*

Institute of Scientific and Technical Information of China
No. 15, Fuxing Road, Haidian District, Beijing 100038, P. R. China
*Corresponding author: liuy@istic.ac.cn

Received February 2020; accepted May 2020

Abstract. *The science and technology search is based on the new search points in the new project, with computer retrieval as the main means, obtaining closely related documents as the retrieval target, using comprehensive analysis methods and contrast methods to make a literature evaluation of the novelty of the new projects. The acquisition of the target document is the most important, that is to say, the search has become the most important link. In order to realize the automation of science and technology search, this paper will demonstrate the automatic extraction of search terms, automatic dictionary construction, and automatic construction of search terms. Finally, 30 new cases in 2017 will be selected as experimental data to verify the technology search system, and the effectiveness of this method is verified.*
Keywords: Science and technology search, Automatic retrieval of search terms, Automatic dictionary construction, Automatic construction of search terms

1. **Introduction.** Science and technology search is based on literature search, using contrast and comprehensive analysis methods to provide an objective factual basis for scientific and technological achievements identification and evaluation, evaluation of scientific research projects, technical consultation, patent applications, etc. The consulting service aims to improve the accuracy, seriousness, authority and impartiality of scientific research projects, results appraisal and rewards, and avoid repeating research projects [1]. In the science and technology search work, the most arduous work is the search stage. When the new searcher builds the search formula, he must conduct a test in each database and find the search terms related to the subject words from the literature preliminary work. In order to improve the efficiency of the new search and realize the automation of the science and technology search, this paper conducts the research of the automatic construction of the search type. The searcher can complete the search step through the recommendation of the system, which greatly saves the search time.

2. **Automatic Retrieval of Search Terms.** The interpretation of the search term in the encyclopedia is: it can summarize the relevant vocabulary to retrieve the content. The search term is the basic unit for expressing the information demand and the content of the search subject, and is also the basic unit for matching calculation with the database in the system. The choice of the search term is appropriate or not, which directly affects the search effect [2]. The main task of automatic retrieval of search terms is to extract words from the search points that can express new topics and search for new content. By extracting the search words from the search points, the computer can enable the new searcher to clearly understand the target topics to be searched for in the new case, which greatly improves the efficiency of the search. The automatic extraction of search terms is

DOI: 10.24507/icicelb.11.09.839

similar to the automatic extraction of keywords. The words are extracted from the text to express the words. Therefore, the method is also universal. At this stage, there is no specific introduction to the automatic extraction of search terms in the field of scientific research, so the key is adopted. The method of word extraction performs search word extraction.

The extraction process of the search term first needs to preprocess the target text, including pre-processing such as word segmentation, de-stopping words, and part-of-speech tagging. Word segmentation is the first step in the extraction of Chinese keywords, and it is also an important step. After the word segmentation, each sentence in the text is divided into ordered word segments. There are many existing word segmentation tools, the most typical of which is the ICTCLAS tool of the Chinese Academy of Sciences, which is called in this system. Stop words are redundant data in the process of text analysis. They do not have the ability to express text topics. They often appear in sentences with high frequency, but they are meaningless. For example, the words "though", "but", etc., generally add these stop words to the stop word dictionary, and filter them in the preprocessing step, thereby eliminating the extraction of keywords. Part-of-speech tagging is also an important step. Words include nouns, verbs, adjectives, etc. In general, the keywords required for the text will be nouns, so that some non-used part of speech will be filtered out during the process of keyword extraction, such as, in Sci-tech search, the part of speech such as verbs and adjectives will be filtered out, thus improving the correct rate of the search terms.

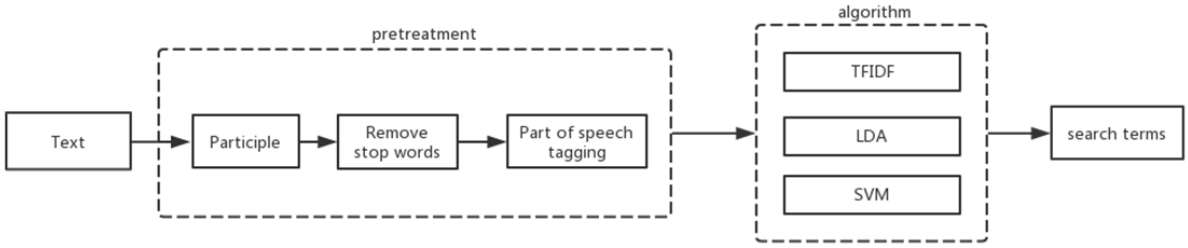

FIGURE 1. Search terms extraction process

At present, the methods of keyword extraction are divided into supervised, semi-supervised and unsupervised, which can be divided into statistical-based methods, topic-based methods and machine learning-based methods. Along the above methods, domestic and foreign scholars have done a lot of research. How and Narayanan [3] may use the CTD (Category Term Descriptor) to improve the TFIDF (Term Frequency Inverse Document Frequency) to improve the skew of the category dataset. In [4], the TFIDF method is applied to the field of image processing, which overcomes the weakness of traditional image processing and considers the color map and shape feature of the image. The information retrieval idea is used to calculate the TFIDF feature weight. Siu et al. [5] found the topic information and keywords by training the HMM speech model, and the system achieved the desired effect on the test corpus. Wei et al. [6] applied the LDA (Latent Dirichlet Allocation) model to the study of Mongolian historical documents, combined with specific languages and fields, and realized automatic keyword pumping, and the results obtained were obvious.

3. **Automatic Dictionary Construction.** At present, the most time-consuming step in the science and technology search work is the retrieval process. In the early stage of the search, the search terms should be determined. In the process of manual search, it is a waste of time to determine the search terms for a large number of documents. The efficiency of the search for new work is greatly reduced, and the automatic construction of the dictionary is to shorten the search time and improve the efficiency of search and new.

The domain dictionary is automatically constructed. On the one hand, it needs to organize and summarize the existing related domain resources. On the other hand, it is necessary to select appropriate algorithms to expand the related fields. The existing related resources include the existing vocabulary in various fields and a large number of new reports. This part is mainly reflected in the manual. It is necessary to search the new recruits in the existing search report based on the existing vocabulary. The words are added to the vocabulary, and the existing dictionary is continuously enriched. Of course, this part is only the initial stage of automatic construction of the dictionary, and manual construction of the dictionary requires a lot of manpower and material resources. The main methods of automatic dictionary construction are automatic vocabulary construction method based on interoperability, automatic vocabulary construction method based on retrieval strategy, automatic vocabulary construction method based on grammar analysis [7].

The automatic vocabulary construction method based on interoperability combines two or more vocabularies into one vocabulary by interoperability technology, and simultaneously guarantees the integrity of the sub-tables during the merging process [8]. The shortcoming of this method is that the existing vocabulary generally reflects the more general subject area information, and requires a large number of additions, deletions and changes, and the method is not applicable to the emerging subject areas.

The vocabulary automatic construction method based on the retrieval strategy means that the user retrieval consciously adopts multiple retrieval strategies to construct the retrieval formula, and the user generates the vocabulary to acquire knowledge from the retrieval strategy as the basis for the recognition of the inter-word relationship. With this method, the construction of vocabulary requires a large number of user interactions, and it is continuously accumulated and improved in the use of vocabulary, so the construction time is long and the workload is large [9].

A vocabulary-based automatic vocabulary construction method uses the external features of the document to identify various semantic relationships implied between words, assuming similarities in similar grammatical contexts, and thus can be grouped into the same concept. This method explores the conceptual relationship from the perspective of linguistics, but the current development of linguistics has not made much breakthrough, so it limits the use and development of this method [10].

An automatic vocabulary construction method based on co-occurrence analysis, which extracts the association between vocabulary by calculating the vocabulary co-occurrence frequency or co-occurrence position, and the generated vocabulary is also called the co-occurrence rate vocabulary [11]. This method is one of the most widely used vocabulary construction techniques, and comprehensively applies theories and methods of natural language processing technology, machine learning, knowledge mining and knowledge discovery. This method uses the literature library covering the subject area as the source of vocabulary construction, using statistical method, knowledge discovery and text mining methods to identify important vocabulary and inter-word relationships in the subject area. For the vocabulary constructed by this method, although the relationship between vocabularies is not like the effect of artificial construction, it can detect the potential knowledge framework in the text library. This method is also a kind of technology search system to adopt in the future method.

4. **Search-Based Automatic Construction.** The search formula is an instruction issued by the searcher to the computer, and is also the language of the human-machine dialogue, and the search expression expresses the search intention of the searcher. The search term usually consists of a search term, a logical operator, etc. [12]. The search terms include synonyms, upper and lower words, and the logical operators include AND, OR, and NOT. Synonym automatic recognition technology and similarity calculation are involved in the automatic construction process of search.

In the field of information retrieval, the meaning of synonyms is not completely equivalent to the synonyms commonly used by people. It mainly refers to words that can be replaced each other when the context is the same or similar in the search [13], regardless of the feelings it has color. At home and abroad, the research on synonym recognition mainly includes the following methods of automatic recognition: the synonym recognition method based on the literal similarity algorithm between vocabularies mainly uses the same characters. In Chinese synonyms, many synonyms contain the same characters. This method can be used to perform rough similarity calculation in synonym recognition. However, there are problems of low efficiency and low accuracy; based on synonyms in the dictionary or lexical classification system, in the synonym dictionary, the semantic vocabulary constitutes a hierarchical concept tree structure. By calculating the distance between the meanings of the two words in the node and performing weighted calculation, the synonym of the two words is finally obtained. The recognition method, the literal similarity algorithm mainly uses the literal similarity of the same characters between vocabularies. In Chinese synonyms, many synonyms contain the same characters. This method can be used to perform rough similarity calculation in synonym recognition, but there is a problem that the accuracy is not high; the automatic recognition method of synonym based on web is mainly realized by the search engine, and the similarity of the vocabulary is calculated by calculating the mutual information of the phrase in the retrieval system, thereby identifying the automatic recognition of the synonyms, using the large corpus. The method of synonym recognition mainly uses the co-occurrence analysis of vocabulary in the corpus and the semantic vector space to calculate the semantic similarity.

All of the above methods are the result of the existing synonym recognition. Although they are also the current methods, most of them are based on traditional synonyms, and the corpus used requires manual maintenance, which is time-consuming and labor-intensive. And in this era of information explosion, new words continue to appear, and synonyms for new words will continue to emerge. Therefore, in order to meet the needs of information growth, it is necessary to improve the algorithm of automatic recognition of synonyms and use new resources to form a complete automatic extraction method of synonyms. The resources used in the automatic identification of synonyms in the science and technology search system are web pages, papers, and authoritative dictionaries. From these resources, automatic extraction and mining of synonyms are implemented, and a synonym dictionary for technology search is generated.

This study will use the method in the synonym automatic identification system to calculate the semantic similarity between vocabularies using hyperlink analysis. Specifically, the following three steps are performed. The first step is to perform a word segmentation operation on the lexical interpretation information, and the relationship between the vocabulary interpretation and the interpreted relationship is represented by a directed graph, and is transformed into a form that is easy to process and understand by the computer-adjacent matrix. The second step uses the sorting algorithm to calculate and rank the semantic similarities between words. Generate a set of candidate synonyms for each word. The third step is to filter the candidate synonym sets and recommend the best synonym [14].

5. **Experimental Analysis.**

5.1. **Experimental data.**

The data of 30 new cases in 2017 were selected for this experiment, which involved the fields of chemistry and chemical engineering, materials science, medicine and health, biology, and agriculture. Using the above methods, the results of the search system automatically generated by the test system were tested. It should be noted that the 30 cases used in the experiment are typical cases encountered by professional

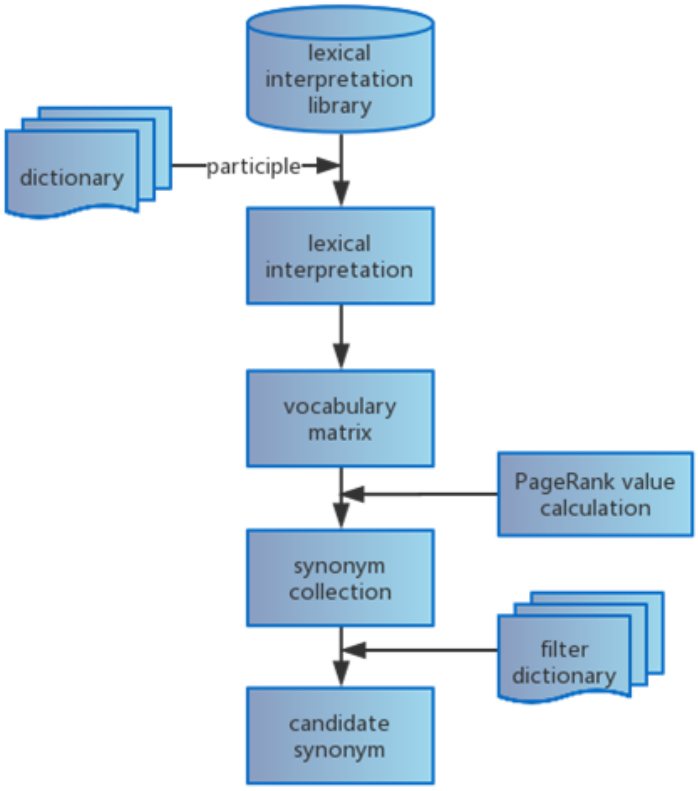

FIGURE 2. Synonym automatic identification process

search personnel in the actual work process. In the manual search process, the difficult cases are summarized.

5.2. **Evaluation criteria.** The evaluation results are evaluated, and the evaluation is completed by using precision, recall, and comprehensive evaluation index (F-Measure). The accuracy is the ratio of the number of documents retrieved to the subject of the new case and the total number of documents returned by the system. It can also be called the recall rate. The recall rate is the same number of documents retrieved as the search report. The ratio of the number of documents listed in the new report can also be called the precision rate. The comprehensive evaluation index (F-Measure) is also called the F factor and the F value (F-score). When precision and the recall rate conflict, at this time, it is necessary to comprehensively consider when the F factor is high. At the time, the method of the test is more effective. The calculation method is as follows:

1) Precision = the number of documents retrieved related to the subject of the new case/the total amount of documents returned by the system.

2) Recall = the number of documents in the hit report/the total number of documents in the new report.

3) F-Measure = $\frac{(\alpha^2+1)P*R}{P+R}$, when $\alpha = 1$, it is the most common F1, F1 = $\frac{2*P*R}{P+R}$.

5.3. **Experimental results and analysis.** Table 1 lists the titles of each case, the number of new points, the number of manual documents (that is, the number of documents listed in the new report), and the number of systematic documents (the types of documents in the system include journal articles, dissertations, patents, conference papers, etc., the number here is the total number of several types of documents), the correct rate, the recall rate and the F value. It can be seen from the table that the correct rate of the system can reach 100%, that is to say, the retrieval method is automatically searched by the system, and the results returned by the system are all related to the subject of the case, thus also proving the system recommendation. Search for words and automatically build the

TABLE 1. Experimental data results

field number precision recall F-measure

领域	编号	查新点(个)	人工（文献数）	系统（文献数）	正确率	召回率	F值
生物领域	1	1	8	6	100.00%	75.00%	0.86
	2	1	6	6	100.00%	100.00%	1.00
	3	2	8	5	100.00%	62.50%	0.77
	4	2	12	9	100.00%	75.00%	0.86
	5	2	16	15	100.00%	93.75%	0.97
	6	1	8	8	100.00%	100.00%	1.00
	7	1	6	4	100.00%	66.67%	0.80
	8	3	16	13	100.00%	81.25%	0.90
	9	2	12	7	100.00%	58.33%	0.74
	10	4	18	15	100.00%	83.33%	0.91
	11	1	6	6	100.00%	100.00%	1.00
化学化工领域	12	1	23	18	100.00%	78.26%	0.88
	13	3	13	11	100.00%	84.62%	0.91
	14	1	16	8	100.00%	50.00%	0.67
	15	1	9	2	100.00%	22.22%	0.36
临床医学领域	16	2	8	6	100.00%	75.00%	0.86
	17	1	15	15	100.00%	100.00%	1.00
	18	1	7	7	100.00%	100.00%	1.00
	19	1	6	5	100.00%	83.33%	0.91
	20	1	6	6	100.00%	100.00%	1.00
	21	1	12	11	100.00%	91.67%	0.96
	22	1	13	7	100.00%	53.84%	0.70
	23	2	17	15	100.00%	88.24%	0.94
医药卫生领域	24	1	4	4	100.00%	100.00%	1.00
	25	1	4	4	100.00%	100.00%	1.00
	26	1	9	4	100.00%	44.44%	0.62
材料科学领域	27	1	14	11	100.00%	78.57%	0.88
	28	1	15	12	100.00%	80.00%	0.89
	29	1	9	8	100.00%	88.88%	0.94
	30	1	3	2	100.00%	66.67%	0.80

validity of the search. The final average recall rate of the system reached 80.22%, and the average value of F reached 0.87, which met the expected standard and proved the effectiveness of the system approach. The above two points are evaluated from the overall effect of the system, and the following is analyzed from the local.

First, there are generally no more than three new search points in the new case. In order to verify the effectiveness of the method in the system, four cases with new search points were selected, "Key technology research on edible fungi after harvesting". From the 18 articles, 15 hits, the recall rate reached 83.33%, reaching the standard, which means that the system method is also suitable for special situations. Second, it can be seen from the experimental results that the recall rate of the new environmentally friendly vanadium-free FCC regenerative flue gas catalyst was only 22.22%, and the system recalled two articles. There are two reasons. First, this search point is more compliacted than other search points, and the search point is also merged by the new recruits. Secondly, there are three kinds of composite oxides to be searched for in the new point. There are many other oxides similar to other oxides. When the founder proves that the retrieved documents are not the oxides, several different documents are selected to prove that, this leads to many human factors. In the experimental data, most of the cases with low recalls are for this reason. This does not prove that the system is not effective, but it is an uncontrollable factor. Third, from the experimental results, when the system handles the case of chemical industry, the recall rate is lower than other fields, because most chemical cases involve many chemical formulas and English abbreviations. In this regard, the system is not treated separately, which will cause the chemical formula to be dispersed and cannot be

identified. This requires special treatment in the later stage, and it is also the aspect that the system needs to improve afterwards.

6. **Conclusion.** This paper is still a starting stage for the exploration and practice of the automatic retrieval technology. The next step still needs to be constantly improved and revised. The experimental data in this paper belongs to the small test set. In the later work, as the case in the system increases, it will be verified again with the large-scale case data. In the next research work, it is necessary to expand the target range of search terms, from the previous search points to case titles, search points, project backgrounds, and search terms in the order, and join the relevant algorithms of network learning to study how to calculate the upper and lower position and synonym relationship between vocabulary, which determines the level of automation of science novelty retrieval, and it is an urgent need in the field of science novelty retrival, and is also an important issue that we need to solve urgently. Exploring retrieval automatic construction is the demand of science and technology and the development of the times, and it is an irreversible technical direction.

REFERENCES

[1] J. Ren, C. Jiang and Y. Ji, Construction of a new technology file management system based on NoteExpress software, *Science and Technology Information*, no.19, pp.241-242, 2011.

[2] *https://baike.baidu.com/item/%E6%A3%80%E7%B4%A2%E8%AF%8D/6196289?fr=aladdin.*

[3] B. C. How and K. Narayanan, An empirical study of feature selection for text categorization based on term weightage, *Proc. of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence*, PA, pp.599-602, 2004.

[4] Y. Suzuki, M. Mitsukawa and K. Kawagoe, A image retrieval method using TFIDF based weighting scheme, *Proc. of the 19th International Conference on Database and Expert Systems Application*, France, pp.112-116, 2008.

[5] M. Siu, H. Gish et al., Unsupervised training of an HMM-based self-organizing unit recognizer with applications to topic classification and keyword discovery, *Computer Speech & Language*, vol.28, no.1, pp.210-223, 2014.

[6] H. Wei, G. Gao and X. Su, LDA-based word image representation for keyword spotting on historical Mongolian documents, *International Conference on Neural Information Processing*, pp.432-441, 2016.

[7] J. Wang and L. Wang, Research on multi-feature keyword extraction algorithm, *Computer System Application*, no.7, pp.162-166, 2018.

[8] T. Petersen, The AAT: A model for the restructuring of LCSH, *Journal of Academic Librarianship*, vol.9, no.4, 1983.

[9] V. Guntzer, Automatic thesaurus construction by machine learning from retrieval sessions, *Information Processing & Management*, vol.25, no.3, 1989.

[10] H. Tsurumaru, T. Hitaka and S. Yoshida, An attempt to automatic thesaurus construction from an ordinary Japanese language dictionary, *Proc. of the 11th International Conference on Computational Linguistics*, pp.445-447, 1986.

[11] Y. H. Tseng, Automatic thesaurus generation for Chinese documents, *Journal of the American Society for Information Science and Technology*, vol.53, no.13, 2002.

[12] L. Zhang, L. Cong, N. Zeng, X. Tan and P. Cui, *Engineering Master's Professional Information Retrieval Reference Guide*, Tianjin University Press, 2011.

[13] J. Mei et al., *Synonyms and Words*, Shanghai Dictionary Press, Shanghai, 1983.

[14] Y. Lu, *Chinese Synonym Recognition for Information Retrieval*, Master Thesis, Nanjing Agricultural University, Nanjing, 2016.