ICIC Express Letters
Part B: Applications
Volume **11**, Number **12**, December **2020**

ICIC International ©2020 ISSN 2185-2766
pp. 1187–1193

RETRIEVAL EXPRESSION AUTOMATIC CONSTRUCTION

Tingting Wang 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 April 2020; accepted July 2020

Abstract. *With the development of modern computer technology, electronic information and text consulting showing explosive growth, network has already become the most important way for people to obtain and transmit information. It takes people a lot of time to lock information quickly in the content which they are interested in the mass information. Therefore, effective and simple information retrieving technology is the most needed in the current Internet era. The retrieval automatic construction technology is a powerful tool to guide the searchers, and recall criteria and precision criteria are the judgment standard for evaluating the retrieval results. This paper introduces the retrieval technology and operators, retrieval techniques and methods in detail. It is pointed out that the first step is to determine the retrieving words. Secondly, the classification number is determined. The automatic construction of retrieval process based on combination optimization provides the basic theory and technical support for retrieval automatic construction technology.*

Keywords: Keyword, Operator, Retrieving technology, Retrieval process

1. **Introduction.** Retrieval expression is a query string that can be understood and operated by search engines. It consists of keywords, logical operators and search instructions (search grammar). Key words are the mainstay of retrieval expressions. Logic operators and search instructions define keywords from different perspectives according to specific query requirements. Information retrieval is a process in which the computer retrieval system matches the query words of the searcher with the indexed words of document records. When the question word matches the quotation word, its name is hit. Usually, in order to improve recall and precision, computer retrieval system requires searchers to use various operators prescribed by the retrieval system to connect the logical relationship and location relationship between search terms, and to form a search command expression (abbreviated as retrieval expression) that can be recognized and executed by computer for retrieval. The quality of retrieval writing directly affects the quality of retrieval results [1].

This paper describes in detail the processing of the search query in the retrieval system. After the user inputs the search content, the system performs segmentation words and part-of-speech tagging on the search content, and deletes the useless words, stop words, unregistered words, etc. in the constructed vocabulary according to the scene requirements, and extracts the extracted keywords through the concept synonyms. The vocabulary finds matches and outputs, the keywords are recombined by the method of the rule and the logical operator, and the combined search expression is extracted in the database by the index method, and the search words are presented on the page according to the word frequency, semantic or diversified sorting method. Keywords are key words that describe the content of a search. The keyword in network search is a broad concept.

DOI: 10.24507/icicelb.11.12.1187

It is a non-controlled free word, and any expression with practical significance and its writing form, can be used as search keywords. Words can be divided into subject keywords and feature keywords. Subject keywords refer to the words that must be used to express the main search content. Without the use of subject keywords, specific content cannot be accurately searched. Feature keywords refer to quantifiers, adjectives and nouns that appear at the same time with the subject keywords in the content description and are location close to each other, which further explain and limit the subject keywords. Feature keywords can effectively narrow the search scope and make the required information more in advance of the result ranking. In search practice, we often encounter the situation that the topic keywords are selected accurately and used properly. The query content is still not in the first three pages of the results, if the use of feature keywords is added. Search results will improve significantly.

2. **Retrieving Techniques.** In practical retrieval, there are often more than one subject concept involved in the search query, and the same subject concept often involves multiple synonyms, synonyms, related words, and even some scientific names, popular names and commodity names, different spellings in Britain and America, singular and plural forms, full names and abbreviations, etc. In order to express search questions objectively, comprehensively and correctly, the computer retrieval system requires searchers to first analyze the logical relationship between retrieval terms, and then use different Boolean logical operators to match related search terms. Make some search units with simple concepts form a retrieval expression with complex concepts through logical combination. By specifying the hitting conditions and matching order of the information, the real search demands of the searcher are expressed, and then the searcher inputs the retrieval type into the corresponding retrieval system for retrieval. Where the conditions in the database meet the requirements specified by the logical matching, that is hit information. If you want to improve the probability of information hitting, you must master the retrieval technology and method skillfully. If you master the retrieval technology and method, you can save time, improve efficiency, grasp the overall development of things, and also prevent you from going astray and losing your way.

2.1. **Retrieval technology.** The retrieval technology of database named is analyzed. It can realize the functions of field retrieval, Boolean logic retrieval, phrase retrieval, location retrieval and truncation retrieval.

2.1.1. *Field retrieval.* Field retrieval, also known as restricted retrieval, improves the hit rate by restricting the location of search terms in different fields of records.

2.1.2. *Boolean logic retrieval.* Boolean logic retrieval is also called Boolean logic search. Strictly speaking, Boolean logic search refers to the method that uses Boolean logic operators to connect search terms, and then the computer performs corresponding logical operations to find the required information. It is the most widely used and the most frequently used. The function of Boolean logic operators is to connect the retrieval words and form a logical search formula.

2.1.3. *Phrase retrieval.* Phrase retrieval is also called string retrieval. Phrase retrieval is to search for exact retrieval terms in all content in order to achieve the purpose of phrase retrieval. It is a method of enclosing a word or phrase with “ ” as an independent operation unit and matching strictly to improve the accuracy of retrieval.

2.1.4. *Location retrieval.* Location retrieval is also called proximity retrieval. The relative order or position of words in document records may express different meanings, while the relative order of words in the same retrieval expression is different, and their retrieval intentions are also different. Boolean logic operators are sometimes difficult to express exact query requirements for certain retrieval topics. Although field-limited retrieval can further satisfy the query requirements to a certain extent, it cannot restrict the relative position of search terms. Location operator retrieval is a technology that uses some special operators (position operators) to express the proximity relationship between search words and search words, and can directly use free words to search without relying on the thesaurus.

2.1.5. *Truncation retrieval.* Truncation retrieval is a commonly used retrieval technique to prevent missed detection and improve recall. Most systems provide the function of truncation retrieval. A truncation word is a truncation at the appropriate position of the search term, and then processed using a truncation character, which saves the number of characters input and achieves a high recall rate. A truncated word search generally refers to a right truncation word and a part supports an intermediate truncation word. Truncation retrieval can help improve the recall rate of searches. The truncation search is mainly used for the singular and plural number of the search term, the ending of the part of speech, the word with the same root, and the case of the same word spelling variation.

The research of domestic and foreign search engine scholars also shows that a large number of users seldom use advanced search services [2]. The complexity of retrieval, such as the use of Boolean logic symbols, has no obvious impact on the retrieval results [3]. Therefore, in the process of system transformation, we should strengthen the function of foolish retrieval, rather than invest a lot of energy in improving the function of advanced retrieval. For example, in order to adapt to the usage habits of some users, a synonym list can be generated for each field, and the user's input "title" can be converted to "TIT", which can improve the retrieval situation immediately; or the function of conceptual retrieval can be provided, that is, the system tries to determine the real content of the retrieval, rather than the surface content of the word [4], which can greatly improve the availability of the system.

2.2. **Retrieval techniques and principles.** Document information retrieval is a process in which people use specific retrieval techniques and methods to quickly locate the target resources from the information set and obtain the information related to the information needs. From the point of view of document information processing, document information retrieval includes two processes as follows in Figure 1: information storage and retrieval. Document information storage process: collecting a large number of scattered document information, indexing them according to their content or appearance characteristics, forming feature representations that represent these document information, and storing them on certain carriers to become retrieval tools with retrieval functions; the process of document information retrieval: users put forward retrieval questions according to their information needs, and then use relevant information. The indexing language standardizes the formulated retrieval questions into retrieval marks for the retrieval process. Among them, indexing refers to the analysis of document content characteristics and external characteristics to form a conceptual identification, which is fully and accurately expressed with the corresponding identification according to certain criteria or rules.

The commonly used method of literature retrieval is to construct the retrieval type manually by domain experts, and then use the retrieval type to match in the scientific and technological literature database to obtain the retrieval results. In this case, the precision of the retrieval type directly determines the quality of the retrieval results. Domain experts are prone to two problems in the process of constructing retrieval style. One is that the keywords used by domain experts in retrieval may not be comprehensive

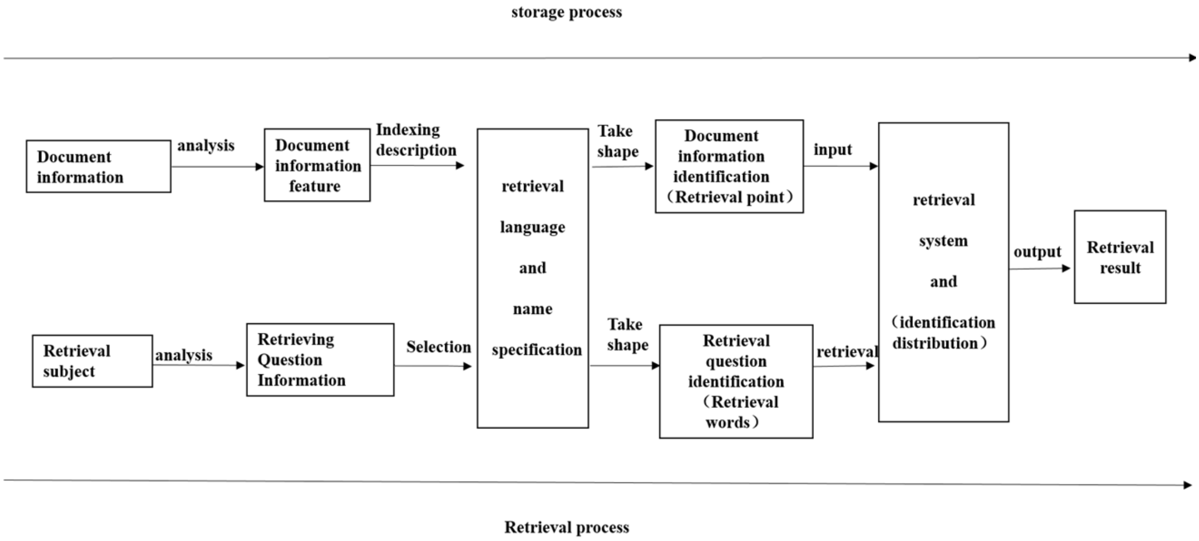
storage process
Document information
analysis
Document information feature
Indexing description
retrieval language and name specification
Take shape
Document information identification（Retrieval point）
input
retrieval system and（identification distribution）
output
Retrieval result
Retrieval subject
analysis
Retrieving Question Information
Selection
Take shape
Retrieval question identification（Retrieval words）
retrieval
Retrieval process

FIGURE 1. Document information retrieval process

enough and some keywords are omitted, which leads to a low recall rate of retrieval results; the other is that the keywords used in retrieval are not only used in the target documents, but also in other non-target documents because of the polysemy of the word. It is possible to introduce unrelated documents into the retrieval results, which results in a low accuracy rate. Therefore, the construction of precise retrieval formula needs to effectively solve these two kinds of problems. Previous work uses such methods as hyponym, synonym, selection of standardized professional terms and so on [5-7], or uses logical operators, location operators and wildcards to improve the accuracy and recall of retrieval results [8,9]. Most of these methods depend on the experience of thesaurus or domain experts that have been constructed, and lack the overall grasp of the whole content of scientific and technological literature database, which is the important reason for the above two types of retrieval problems.

3. **Retrieval Expressions Automatic Construction Process.** Retrieval expression refers to a logical expression that expresses a user's search question in a computer retrieval, and consists of a retrieval expression and various Boolean logic operators, position symbols, and other connection group symbols specified by the retrieval system [10].

Although the keywords used by domain experts are closely related to the target documents, some of them may also appear in other non-target documents. Using such keywords for retrieval may introduce some noise documents to the retrieval results, resulting in low accuracy. For example, the term "Deep Learning" in AI appears not only in the field of AI, but also in the traditional direction of education and teaching research. The retrieval results will contain both AI and related literature of education and teaching, which directly leads to low accuracy. Word embedding generated on the basis of large-scale scientific and technological literature abstracts contains the global semantic information of document content in scientific and technological literature database, so word embedding can be used to solve the above two kinds of retrieval problems to some extent [11].

3.1. **Determination of keywords.** Keyword refers to the key words which can fully reveal and describe the subject content of a document. They often appear in the title, abstract and text of a document. Key words [12] include synonyms, near-synonym, related words and so on. In document retrieval, the main methods to determine keywords are as follows. After the user inputs and submits the search content, the system first performs the word segmentation (IK Analyzer) and part-of-speech tagging processing through the

natural language processing (NLP) technology; then, by combining the scenario requirements, the exception word is filtered by constructing the vocabulary (professional corpus). Useless words, stop words and unregistered words, based on semantic understanding, principal component analysis, select feature keywords (word2vec in Figure 2); then find and embed relevance upper and lower words through conceptual synonym vocabulary, delete and merge the result keywords and logical operator; adjust the retrieval expression; finally, the search expression is indexed by the search engine (Lucene-based Solr or Elastic Search) technology, so that the matching result is presented to the user by word frequency relevance, weight score or diversified sorting.

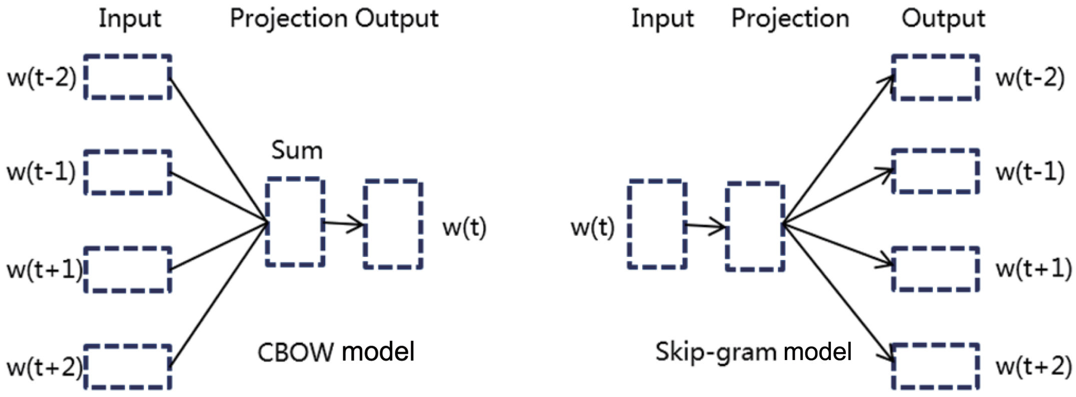

Figure 2. Word2vec model

Automatically construct the retrieval expression model, which inputs the title of the paper into the search box, segment the title of the paper in the background, labels the semantic role and removes the stop words. The commonly used selection methods include as the following. 1) Selecting words through word frequency. Vocabulary can be divided into high-frequency words, medium-frequency words and low-frequency words according to the frequency of occurrence. 2) Selecting words by calculating the value of distinction. The higher the distinctive value of vocabulary, the higher the value of reference words as tags. 3) Words are selected through Poisson distribution. Poisson distribution model can be used to simulate the phenomenon of discrete random distribution. Complete the above steps, and display the processed results in the list box of the page for the retriever to adjust.

3.2. **Organizational retrieval expression.** After determining the keywords and the classification number, the retrieval expression can be organized. Document retrieval based on technical topics can be simply described as the exclusive location or (related location of classification number). On this basis, we can add time constraints to construct a retrieval method for the technical features of each paper topic. Each technical feature needs at least one retrieval method, and different retrieval tools need to construct the retrieval method separately. According to the retrieval strategy, the concept words and the search words under the concept words are selected. All the concept words are “AND” operations, while the search words under the concept words are “OR” operations. Additionally, other operators can be added. According to the norms, it is suggested to construct multiple retrieval formulas from different perspectives according to the content of each novelty retrieval point. Usually, the final retrieval formulas are determined by repeated trial-and-error, word splitting and replacement, composition and field adjustment.

3.3. **Optimizing retrieval expression.** The retrieval module realizes the generation and management of retrieval type. First, it chooses the document database, field (or default), then chooses the concept word and the retrieval term under the concept word, and then clicks on the "Generate the retrieval type" button. In addition, it sets a button to generate the retrieval type of all retrieval tools (field is default) once to improve the efficiency. At the same time, it can edit the retrieval type (such as adding other logic symbols), and next is saved to the list. Different retrieval types of literature databases can be managed separately by lists. It can also jump to the pages of corresponding literature database for retrieval, and the software will automatically copy the search type, and paste on the page. Finally, the available retrieval results are saved in the document management module.

Retrieval expressions and retrieval words are used in the retrieval. After selecting the list box of retrieval words, the retrieval words can be saved by clicking on the "search words" button. The construction of retrieval words is suggested to adopt the logical search forms of "OR" and "AND" symbols. When an existing search formula is selected, the corresponding concept words and the retrieval words in the list box are automatically selected (ticked to add in the list box), and can be added or manually deleted. If there are no subordinate words or related words in the vocabulary, you can also add keywords to the vocabulary in the search page to improve the recall rate.

To a certain extent, the retrieval results will be missed, mistaken or multi-checked. At this time, we should check the upper word, hyponym, related words, conceptual synonyms according to their classification numbers. After viewing the search results, we can do the second search, that is, expand the search. The retrieval can be regenerated under this module. That is to say, narrow or expand the search to achieve the desired results.

Based on the above, making retrieval expression needs to go through the following steps. At first segmentation is the smallest segmentation of words contained in the topic. Secondly, deletion is to delete words that are too broad or too specific, boundary words and function words that have no practical meaning, such as "of", "and". Thirdly, it is substitution. Substitute words that are unclear or prone to retrieval errors. Fourthly, to supplement or adjunction, this section is to add synonyms and related words to the selected words. The addition of these words can prevent omission and affect recall. Finally, for combination the search words can be combined with logical symbolic links to form retrieval words.

4. **Conclusions.** Faced with the boundless Internet, it is like looking for a needle in a haystack to quickly, accurately, reliably and appropriately retrieve the required information. In the actual retrieval, we may get some results, but the results are often insufficient. For the principles and methods of the retrieval strategy, there is still a very broad space to explore. In the process of information retrieval, literature retrieval and analysis become an important means to grasp the status quo of technology development, observe the trend of competitive enterprises, and analyze and predict technology trends. With the development of knowledge economy, the construction of national innovation system and the improvement of innovation level have been put on an important agenda. It is a complex problem to construct a precise retrieval mode and improve the recall and precision of retrieval results. On the basis of word embedding, on the one hand, the recall rate is improved by extending the scientific semantics of the retrieval keywords of domain experts; on the other hand, the author keywords of retrieval results are identified by semantic outliers, so as to improve the accuracy rate. Although it can play a certain role, there are still some problems. Therefore, in future experiments, BERT (which is better than word2vec) can be used to associate keywords in the similarity calculation of keywords in order to better optimize keywords and accurately query. We think BERT is a milestone in NLP, and there is no doubt that there will be a long-term impact on the research and industrial applications of the latter NLP.

Acknowledgment. In the process of writing this paper, my tutor, Mr. Liu Yao, has devoted a lot of effort to pointing out the specific problems in each draft from topic selection to time and again. I would like to express my heartfelt thanks to my elder sister and classmates who have given me great concern and support during my study. While writing my thesis, I learned a lot of useful knowledge, especially the guidance of my thoughts and methods. These useful things have always played an important role in guiding my study and life. I believe that these things will accompany me on the path of a complete life.

REFERENCES

[1] Y. Teng, Information retrieval expression, *Qingdao Medical Journal*, vol.41, no.1, pp.76-78, 2009.

[2] W. Hua, Struts 1.x application development based on MyEclipse 6, *Journal of Anhui Electronic and Information Vocational and Technical College*, vol.9, no.1, pp.4-5, 2010.

[3] C. M. Eastman and B. J. Jansen, Coverage, relevance, and ranking: The impact of query operators on web search engine results, *ACM Trans. Information Systems*, vol.21, no.4, 2003.

[4] C. Hao, Analysis of user retrieval expressions in NSTL network service system, *Library and Information Work*, vol.51, no.6, pp.120-122,146, 2007.

[5] H. Wang and H. Cao, Retrieve learning: Meaning, way and development, *China Distance Education (Comprehensive Edition)*, vol.51, no.4, pp.39-43, 2009.

[6] H. Ma and J. Feng, An analysis of user retrieval characteristics of chinese search engine, *Journal of Information Science*, vol.24 no.6, pp.718-722, 2005.

[7] J. Chen, Reflections on the code for writing the novelty search report of the science and technology novelty search workstation of the ministry of education, *Library and Information Work*, no.S1, pp.237-240, 2013.

[8] Y. Li, Discussion on the methods of improving recall and precision in document retrieval, *Library Science Research*, no.11, pp.92-93, 2002.

[9] K. Zhu, The influence of synonym acquisition on recall and precision of medical science and technology novelty retrieval, *Chinese Journal of Medical Library and Information*, vol.21, no.3, pp.78-80, 2012.

[10] J. Sun and T. Chen, An effective way to improve the recall and accuracy of literature – The flexible operation of logical operators, position operators and wildcards, *Modern Information*, vol.26, no.10, pp.167-169, 2006.

[11] L. Li, B. Jiang and H. Sun, How to improve the recall rate and precision rate in document information retrieval, *Science and Technology Literature Information Management*, vol.24, no.1, pp.23-25, 2010.

[12] T. He, G. Wang and M. Yang, Accurate retrieval construction method based on word embedding semantics, *Modern Information*, vol.38, no.11, pp.55-58, 2018.