Academic Education Publishing House

Industry Science and Engineering Vol. 1 No. 7, 2024

Research on Concept-based Chinese Literature Retrieval Methods

Peng Xing

Library & Information Management Center, Zhejiang Police Vocational Academy, Hangzhou, Zhejiang, China

Abstract: This study briefly reviews the development of computer-based literature retrieval and summarizes the main issues involved in the research process of Chinese literature retrieval. In response to relevant problems in Chinese literature retrieval, a retrieval model based on "concept groups" was constructed. The model was preliminarily implemented and tested in the study. The test results show that the retrieval model can improve the efficiency of Chinese literature retrieval within a limited scope, reduce the learning and comprehension costs for users, and enhance the user experience and retrieval efficiency in the field of Chinese literature retrieval, providing a concrete direction for its development.

Keywords: Concept Groups; Literature Retrieval; Chinese Literature; Data Processing

1. Research Background and Significance

1.1 Overview of the Development of Computerized Literature Retrieval Methods

In 1954, the U.S. Naval Ordnance Research Center established the first scientific literature retrieval system using the IBM-701 mainframe, enabling searches by logical combinations of unit words and returning document numbers[1]. In 1958, General Electric Company expanded this system to include titles, authors, and abstracts. In 1964, the Chemical Abstracts Service developed an automated system for abstract processing, automating the compilation of abstracts. In the same year, the U.S. National Library of Medicine established the Medical Literature Analysis and Retrieval System (MEDLARS), implementing "OR", "AND", and "NOT" logic operations. From the mid-1960s to the early 1970s, with advances in IT technology, online information retrieval systems were developed, such as the U.S. DIALOG system, ORBIT system, BRS system, and the European ESA-IRS system. From the early 1970s to the present, with further advancements in IT technology, the scale and functionality of online retrieval systems have expanded and improved, with the emergence of commercial international online retrieval services such as Chemical Abstracts, Web of Science, and ACS in the U.S.[2].

In China, research on computerized literature retrieval began in the mid-1970s, with the first introduction of foreign literature databases in 1975. In 1980, the China Construction Technology Development Center established an international online information retrieval terminal, connecting to the U.S. DIALOG and ORBIT systems. In 1981, the Northern Science and Technology Information Institute connected to DIALOG in Beijing, followed by many other departments establishing connections with the DIALOG, ORBIT, and ESA-IRS systems in China[3].

While connecting to foreign systems, China also made progress in developing its own computerized information retrieval systems and databases. In 1978, the Beijing Document Service Information Retrieval System (BDSIRS) was established, containing over 12 million records with 16 Chinese and foreign language databases. In 1989, the Chemical Industry Information Institute's Online System (CHOICE) was established, with 8 Chinese databases and 1 foreign language database. In the same year, the Mechanical and Electrical Industry Information Institute's Online Retrieval System (MEIRS) was established, with 4 Chinese and foreign language databases. Subsequently, many departments set up domestic online retrieval systems. After China officially joined the international Internet in 1994, online retrieval systems and services rapidly developed, especially after the World Bank's 1998 "World Development Report," which introduced the concept of "National Knowledge Infrastructure (NKI)." This led to

the emergence of various commercial knowledge resource services such as CNKI[4].

1.2 Problems in Chinese Literature Retrieval

Based on the development of computerized literature retrieval, the data organization structure, access methods, and retrieval methods for Chinese literature rely on the underlying technology of their native technical environments. The main difference lies in the storage content being in Chinese and using related character sets for encoding. The current primary tools for literature information retrieval are catalog-style searches, search engines, and meta-search engines. These tools mainly use keyword searches, supporting functions like Boolean logic retrieval, phrase search, field search, truncation, proximity search, and subject set search. These functions are based on English, a language with natural space separators, and have solved problems like stemming and lemmatization[5]. In contrast, Chinese lacks natural word separators, and modern Chinese morphology is complex, which limits the effectiveness of keyword searches using native technology environments, significantly impacting precision and recall in Chinese literature searches.

Retrieval can be divided into two categories based on the user's search objectives: knowledge point detection and subject/discipline exploration research. The former focuses on acquiring specific knowledge points, where the search results must accurately return results in the top positions. This is the primary need for most users, who often submit short search terms to search engines without paying attention to query strategy accuracy. To improve the efficiency of such Chinese literature searches, researchers have explored clustering organization of search results, diversified sorting based on topics, and optimization based on multiple document features[6]. For searches aimed at subject/discipline exploration, the primary research directions include retrieval strategy recommendation and optimization[7]. These studies assume that the boundary of Chinese topics proposed by users is unclear, and suggest methods for improvement. Some researchers classify results using methods like knowledge association and central clustering[8], while others evaluate strict retrieval requirements based on saturation/redundancy in specific scenarios[9], using complex retrieval strategies with iterative processes to exhaustively collect evidence[10]. However, none of these studies have focused on improving the native search technology itself to enhance the accuracy of Chinese literature searches.

Moreover, for retrievals aimed at subject/discipline exploration, the ideal scenario is to achieve the search goal through subject-based searches. However, in practice, existing Chinese literature search services compile Chinese subject terms based on the "Chinese Library Classification" standard, using classification numbers to match corresponding category names. Some engines also use segmentation and other methods for complex matching of topics. However, due to variations in author habits, subject terms may not always accurately reflect the content, affecting search precision.

2. Research on Concept-based Chinese Literature Retrieval Methods

2.1 Research Objective

Based on the aforementioned background, traditional Chinese literature retrieval methods mainly use Boolean logic matching based on natural word separators, with later enhancements through dictionary correction, statistical optimization, and multi-parameter optimization to improve accuracy. As a supplement to traditional methods, this study aims to explore and build new Chinese retrieval units and develop a complete retrieval logic structure based on the characteristics of Chinese literature retrieval, with the goal of improving the overall efficiency of Chinese literature searches.

2.2 Main Content

2.2.1 Study on the "word group" concept and processing methods

In English, "Gamma globulin" translates to "bǐng zhǒng qiú dàn bái" in Chinese. However, Chinese authors may use "bǐng qiú dàn bái" or even the abbreviation "bǐng qiú" in actual usage. In practice, medical subject thesauri can be used to address such variations, provided that a standardized subject dictionary exists. However, this

method is not suitable for disciplines with incomplete thesauri. When accuracy is required in disciplinary searches, the searcher must first thoroughly understand the concept being searched and investigate synonyms and abbreviations, requiring significant effort to create complex search strategies using Boolean logic.

This study proposes a model that uses "word groups" as the basic unit of search, minimizing the searcher's workload. A "word group" is defined as a collection of descriptive terms for a specific concept. Word groups may be nested, and users are free to maintain the boundaries and scope of each group. However, to avoid disrupting specific searches with overly broad rules, this study limits word groups and provides the following examples: (as is shown in **Table 1**).

Table 1. Types, Examples, and Processing Methods of Word Groups

Word Group Type	Example	Terminology	Processing Method (PCRE)
Synonyms	shuǐ yín ; gǒng	Mercury	shuǐ yín | gǒng
Abbreviation Variations	bīng zhǒng qiú dàn bái ; bīng qiú dàn bái ; bīng qiú	Gamma globulin	bīng zhǒng ? qiú dàn ? bái ?
Fixed Abbreviation	kē xué jiào yù ; kē jiào	Science education	kē xué ? jiào yù ?
Fixed Replacement	xīn lǐ jiǎo zhèng ; xīn lǐ jiǎo zhì	Psychological correction	xīn lǐ jiǎo [zhèng | zhì]
Word Order Variations	jué chá ; chá jué	Perceived	jué chá | chá jué

Other cases are treated as synonym substitutions. Special cases not covered by the table will be handled using specific expressions in this study, as general user interaction paradigms require extensive feedback, which is beyond the sample size of this study.

2.2.2 Study on "concept groups" and processing methods

In this study, a "concept group" is defined as the collection of all word groups covered by the user's target concept. Operations include adding or excluding word groups. For example, in a previous study on "prisoner labor reform", the basic word group was "labor reform," which included abbreviations like "láo gǎi" and "gǎi zào". However, in actual practice, this word group needs to be modified by word groups like "criminal," "penal," "prison," and "police," while excluding terms related to "community correction."

In traditional retrieval methods, effective field selection is determined through iterative testing. In the model proposed by this study, fields are combined, with priority given to bibliographic information. Specifically, the strategy follows the hierarchy "Keywords (subject terms) > Title > Abstract > Table of Contents > Full Text."

2.2.3 Study on user interaction based on concept group retrieval

Users' search behaviors need to find a reasonable balance between convenience and accuracy. Traditional Chinese engines generally address this issue in the following ways: first, by establishing intelligent search functions, where users input simple vocabulary in a single search box, and the actual search content is completed by a word segmentation program and complex thesaurus matching; second, by establishing advanced search functions that require users to first learn specialized Boolean expressions, then combine them with various bibliographic fields to achieve accurate retrieval. The efficiency of the former is determined by factors such as the accuracy of word segmentation, the completeness of the thesaurus, and the accuracy of the matching algorithm. The efficiency of the latter is influenced by the user's mastery of the specialized search expressions.

The user interaction behavior established in this research model differs from traditional engines, focusing on the user's grasp of the hierarchical relationships of target conceptual vocabulary. Taking the retrieval of the concept of "labor reform" as an example, suppose the conceptual vocabulary relationship established by the user is illustrated in the figure below:

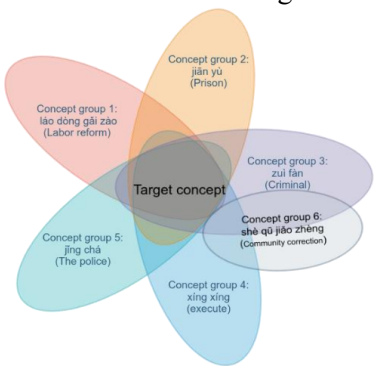

Figure 1. Schematic Diagram of the Formation of the Target Concept Group Case

As shown in **Figure 1**, the user's primary task in this case is to find the constituent words of the retrieval target concept group. Each word group can also be subdivided in this way, tailoring each conceptual group to meet the user's objectives.

2.2.4 Implementation and testing of the concept group-based retrieval model

According to the above model, this research has initially established a concept group-based retrieval engine and testing database. The basic structure of the retrieval part is briefly explained using the pseudocode shown in **Figure 2**.

```
SET CHARACTER ENVIRONMENT $zh_cn;
 // Set Chinese language processing environment
SUB FUNCTION conver_usr_exp( VARCHAR ){...};        //
Function to convert user search expression
SUB FUNCTION data_connect( citation FILE ){...};
 // Function to connect data
SUB FUNCTION match( expression, field list ){...};
// Function to match search expression
SUB FUNCTION degree_strictness( INT,
conver_usr_exp( VARCHAR )){
  // Strictness handling
    ...
    RETURN:1 VARCHAR;
 // Fine-tuned search expression according to strictness
    RETURN:2 search field list;
};
MAIN FUNCTION search(){
// Traverse matched data to obtain results
   $data <- data_connect();
   $exp1 <- conver_usr_exp( GET(user expression) );
   $exp <-  degree_strictness( GET(strictness value), $exp1 )-->1;
   $$list_fields <-  degree_strictness( GET(strictness value), $exp1 )-->2;
   i <- 0;                  // Data item number
   WHILE( $data ) {
     ++i;
     IF( match($exp , $$list_fields ) ){
       RETURN PUSH i INTO $result;
     }
   }
}
```

Figure 2. Pseudocode for the Retrieval Process Implemented in This Research

Based on the retrieval features and user interaction methods constructed in this study, a user retrieval interface based on word groups has been established, with its basic structure and retrieval interaction method illustrated in **Figure 3**. In this retrieval interaction mode, users focus on the descriptive terms contained within the target concept, emphasizing maintaining each level of concepts (word groups) as the focus of their search work. This interaction method avoids additional cognitive overhead and distractions, allowing users to concentrate more on the natural semantic logic of the retrieval target concept.

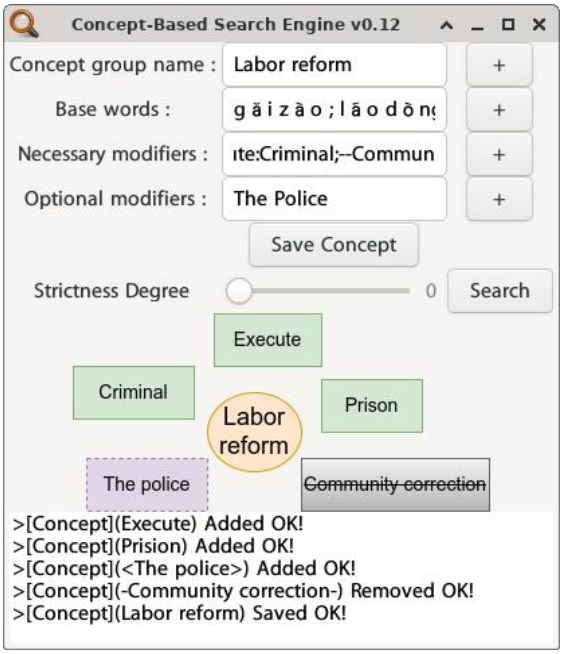

Figure 3. User Interaction Operation Interface for Implementing This Case Retrieval

To test the effectiveness of the retrieval method constructed in this study, a testing database was first established. The method involved obtaining bibliographic information on literature containing the themes "reform" and "labor reform," with each category comprising 3,300 entries from CNKI. After filtering, all literature with the theme "reform" fell into categories like engineering and was unrelated to prison labor reform. The two sets of bibliographic information were mixed as test subjects to detect bibliographic records related to "prison labor reform" for testing the query efficiency of the retrieval scheme.

In the test, traditional Boolean logic retrieval methods were simulated as a control. The testing method of this research involved translating the retrieval expression indicated in the case into a PCRE regular expression. The translated form was like "/(gǎi zào | láo dòng gǎi zào).*(jiān yù | xíng xíng | zuì fàn)/ AND /(jiān yù | xíng xíng | zuì fàn).*(gǎi zào | láo dòng gǎi zào)/" which was further processed according to the testing needs. It is important to note that different retrieval engines have varying implementations of Boolean logic retrieval; this study only implemented it within a controllable scope, and the results represent general conditions defined by this research. The results are shown in **Table 2**.

Since the underlying logic of the two methods in this case is the same, the results are similar,

with differences only in the actual machine search procedure paths, where the latter's path is shorter than the former's.

Table 2. Efficiency Test Results of the Two Retrieval Methods

Testing Method	Recall Rate	Precision Rate	Time Taken (seconds)
Simulated Traditional Boolean Logic	85%	99%	0.074
Concept Group-Based Retrieval Method	85%	99%	0.056

In the user interaction method test, a sample of 160 students from the researcher's college was randomly divided into two groups. Each group received instructional teaching for one class period (45 minutes) on traditional Boolean logic retrieval and the concept group-based retrieval method as indicated in this study, respectively, and then given assignments to conduct retrievals. The evaluation of the retrieval process and results was based on the duration for constructing the retrieval expressions; the retrieval results were scored according to recall and precision rates based on standardized criteria (maximum score of 100). The results are shown in **Table 3.**

Table 3. Experimental Results of the Retrieval Tests in Both Experimental Groups

Group	Average Duration for Expression Construction (minutes)	Average Retrieval Score	Average Retrieval Time (seconds)
Simulated Traditional Boolean Logic Group	33	77	0.389
Concept Group-Based Retrieval Group	16	85	0.148

Due to the limitations of experimental environment and conditions, the sample size and experimental subjects used in this study are relatively small. However, it can illustrate the efficiency of the designed retrieval model within the scope of this research.

3. Discussion

The retrieval model constructed in this study aims to optimize the literature retrieval methods for users in a Chinese language environment, transitioning from the traditional application of Western Boolean logic retrieval models to a retrieval model based on Chinese conceptual word groups. The field and direction involved in this research work are relatively broad. This research, limited by the research environment and conditions, has completed a limited scope of work, but the author believes it still has certain enlightening effects.

Areas for improvement in this study include: (1) Although the basic logic of the retrieval method exists, practical applications are lacking, and there is a deficiency in improving this model. It is hoped that it can be applied to actual retrieval environments in the future and updated based on practical feedback; (2) The testing environment for retrieval in this study was limited, resulting in a smaller sample size, and it is hoped that future large-scale retrieval environments and sample size tests can enhance this work; (3) This research points to users' proactive retrieval, but with the development of artificial intelligence technologies, literature retrieval methods may undergo significant changes. Future research may utilize AI technologies for large-scale conceptual group data processing, sharing, and network collaboration, but these designs still require the implementation of actual retrieval environments.

Through the above content, it is hoped that the retrieval method constructed in this study can contribute to the development of Chinese literature retrieval in future research advancements.

4. Conclusion

This study delves into literature retrieval methods within the Chinese context and presents a model based on concept groups. By reflecting on traditional Boolean logic retrieval, the paper emphasizes the importance of users understanding the hierarchical relationships of target vocabulary in their searches. Experimental results indicate that the new model outperforms traditional methods in retrieval efficiency, accuracy, and user interaction experience, demonstrating its potential for practical application.

Despite the limitations of this study, such as small sample size and insufficient practical application, it offers new insights and directions for the development of Chinese literature retrieval technology. Future research should aim to expand the sample size, optimize retrieval algorithms, and explore the application of artificial intelligence in literature retrieval to promote continuous progress in this field. By continuously improving and applying

this model, we can effectively enhance user experiences in literature retrieval and meet the growing information demands of the digital age.

Acknowledgments

This paper is supported by Library and Information Committee of Zhejiang Universities and Colleges, China. (No. 2022TKT012)

References

[1] Snyder, S. Samuel. Computer Advances Pioneered by Cryptologic Organizations. Annals of the History of Computing, 1980, 2(1): 60-70.

[2] J. M. Griffiths, D. W. King. US information retrieval system evolution and evaluation (1945-1975). IEEE Annals of the History of Computing, 2002, 24(3): 35-55.

[3] C. Z. Wang. The development of China's scholarly publications in library and information science, 1979-2009: An analysis of ISI literature. Library Management, 2011, 32: 435-443.

[4] C. Cao, Q. Feng, Y. Gao, et al. Progress in the development of national knowledge infrastructure. Journal of Computer Science and Technology, 2002, 17: 523-534.

[5] X. P. Qiu, Z. Qi, X. J. Huang. FudanNLP: A Toolkit for Chinese Natural Language Processing. ACL 2013, 2013: 49.

[6] X. Liu, S. Wang, S. Lu, et al. Adapting feature selection algorithms for the classification of Chinese texts. Systems, 2023, 11(9): 483.

[7] J. He, E. Meij, M. Rijke. Result diversification based on query-specific cluster ranking. Journal of the American Society for Information Science and Technology, 2011, 62(3): 550-571.

[8] A. K. Nikhath, K. Subrahmanyam. Feature selection, optimization and clustering strategies of text documents. International Journal of Electrical and Computer Engineering (IJECE), 2019, 9(2): 1313-1320.

[9] R. J. Whittaker, F. Rigal, P. A. V. Borges, et al. Functional biogeography of oceanic islands and the scaling of functional diversity in the Azores. Proceedings of the National Academy of Sciences, 2014, 111(38): 13709-13714.

[10] Y. Zhang. A Study of Automated Deep Classification of Literature Based on Chinese Library Classification. Libraly Journal, 2024, 43(395): 61.