An Improved Full-text Retrieval for Elementary Education Resource Database System

Cong Liu, Huai Huai
Information Technology Centre, Tsinghua University, Beijing 100084, China
Autolabor Tech Co., LTD, Beijing 100083, China
liuc@tsinghua.edu.cn
huaihuai0922@gmail.com

Abstract. With the rapid development of Internet and information technology, ways of teaching and learning in China’s elementary education have begun to change. In the face of massive resources, how to optimize the searching results and improve users' satisfaction through the search services, has now become a major priority. Based on the in-depth study of full-text retrieval technology, combined with the actual searching needs of the elementary education resources, this paper focuses on Chinese word segmentation technology, and proposes a user-defined professional dictionary based on cell lexicon. Experiments show that using this professional dictionary to segment words improves the professionalism of word segmentation results and the accuracy of search results. Meanwhile, an improved full-text retrieval for elementary education resource database system based on Lucene is designed and implemented, including resource collection, text pre-processing, content indexing and resource retrieval. This system is successfully integrated and applied to the massive knowledge base platform for elementary education, which meets the needs of full-text retrieval of textbooks and supplementary content, and makes a good effect.

1. Introduction
In recent years, with the vigorous development of computer and Internet technology, the channels for people to obtain information and knowledge have gradually changed from paper-based to digitization. The elementary education in China is also in line with the development trend of the times and gradually towards informatization, in order to achieve the goal of innovation and personalization of elementary education resources and education modes by the education department in China. Some education and educational organizations or companies have developed and constructed many education resource database systems for primary and secondary schools. Through the analysis and research, most of the search services provided by the resource database on the market are based on the keywords of the external resource characteristics, such as resource name and type. Using this keyword search method based on Metadata, the range of query content for users is fixed. As the amount of data increases, when facing massive data queries, the database query time increases, and the user's waiting time will become longer. Then in the user-perceivable time, the longer time users wait, the worse user-experience is, the more likely the users give up the query and leave the page, and the lower user-experience that is easier to cause the user loss. It can be seen that the resource search services provided nowadays have been unable to meet the needs of the majority users.
Considering the characteristics of elementary education resources, there are nine basic disciplines including Chinese, mathematics, physics, chemistry, biology, politics, history and geography, spanning 12 grades. Compared with the general information resources in the Internet, such elementary resources are more professional. This paper will discuss and study how to match more accurately in the query, improve the precision on the basis of recall assurance, reduce the minimum resource consumption and improve the search speed for teachers and students. Therefore, the main task of this paper is how to improve the basic search service, optimize the search results and improve the user satisfaction when facing massive resources. Based on Lucene full-text retrieval technology and Java language, with the help of Ansj Chinese word segmentation toolkit and user-defined professional online dictionary, this paper designs and implements the full-text retrieval of the elementary education resources database system.

2. Key Technologies

2.1. Full-text retrieval
Full text retrieval is a branch of modern information retrieval technology, and is a kind of retrieval method that enables any word in the text to be queried. It deals with the entire content of the text rather than the information describing the characteristics of the text[1]. In the information retrieval system, the commonly used evaluation indexes are precision and recall, which were proposed by Kent et al.[2]. For the retrieval system, although the content evaluated by the precision and recall are very important, the two indexes cannot be taken into account at the same time. They influence each other. If one of them is higher, the other will decrease[3].

2.2. Lucene
Lucene is a full-text retrieval engine toolkit developed by Java language based on the foundation of full-text retrieval. It is an open-source project. Since the project was released, Lucene has been widely used in various applications because it can meet the applicability of different platforms[4].

2.3. Chinese word segmentation
Word segmentation technology is one of the key technologies of full-text retrieval. The results of word segmentation directly affect the construction of word-list, which is the source of index data[5]. The research object of this project is the elementary education resources, which includes both Chinese and English, but Chinese takes up a larger proportion. Therefore, considering the actual application situation, Chinese word segmentation will be mainly studied.

Chinese word segmentation is divided into manual segmentation and automatic segmentation[6]. The contents and methods of Chinese word segmentation are all around automatic segmentation. Chinese word segmentation methods can be separated into three categories, including dictionary-based word segmentation, understanding-based word segmentation and statistics-based word segmentation[7]. Through the research and analysis of the above three kinds of methods, we find that the statistical-based word segmentation method is neither limited to the size of the dictionary nor the text category, and can be applied to various scenarios and supports ambiguity discrimination and unknown word recognition. Therefore, compared with other segmentation methods, the statistics-based word segmentation method can be more suitable for the study of this project.

3. Optimization of Chinese Word Segmentation
Word segmentation is a key foundation in full-text retrieval and an important step before content indexing. The precision of word segmentation determines the size of the word-list and the presentation of search results. Chinese word segmentation faces many technical challenges, including overlapping ambiguity, combination ambiguity and unknown word recognition. The Chinese word segmentation tool Ansj selected for this project has a good solution to this[8], but there are still some deficiencies. In the case that proper nouns cannot be well recognized by word segmentation tools, the establishment of
user-defined professional dictionaries can greatly improve the accuracy of segmentation results. The format of user-defined dictionary supported by Ansj Chinese word segmentation toolkit is: [user-defined word] [part-of-speech] [word frequency].

3.1. User-defined professional dictionary
The construction of entries in professional discipline dictionaries is inseparable from special dictionaries. The special dictionary is a dictionary that gathers the knowledge of various disciplines, which can meet the needs of the project. Therefore, we propose a professional discipline dictionary based on Sogou cell lexicon[9].

3.2. Sogou cell lexicon
Sogou cell lexicon is a fine differentiated lexicon pioneered by Sogou Pinyin Input Method, which is open sharing and can be upgraded online. Up to now, the number of cell lexicon has reached 27,695, with a total of 48,482,247 entries. It is divided into 12 major categories and 98 sub-categories including natural sciences, social sciences, humanities, and art design. Sogou Pinyin Input Method is the choice of most users. Because Sogou cell lexicon is professional and high-quality, it is the best choice to use it as the source of user-defined entries.

3.3. Design and implementation of improved professional dictionary
The specific steps of the improved professional dictionary are as follows:

Step 1. Analyse the first-level and the second-level categories of Sogou cell lexicon to determine the lexicon category to be crawled. After determining the category, analyse the corresponding link address, as shown in Table 1. Obtain a category ID list, namely cataList = [32, 38, 53, 2, 3, 13, 14, 26, 91]. Obtain the content of each link page according to the rules and get the maximum paging data. The code is as follows:

```python
Table 1 Lexicon links to various disciplines

| disciplines                | Links                                           |
|----------------------------|-------------------------------------------------|
| Humanities | History             | https://pinyin.sogou.com/dict/cate/index/32     |
| Humanities | Literary            | https://pinyin.sogou.com/dict/cate/index/38     |
| Humanities | Language            | https://pinyin.sogou.com/dict/cate/index/53     |
| Natural Science | Mathematics | https://pinyin.sogou.com/dict/cate/index/2     |
| Natural Science | Physics          | https://pinyin.sogou.com/dict/cate/index/3     |
| Natural Science | Chemistry         | https://pinyin.sogou.com/dict/cate/index/13     |
| Natural Science | Biology             | https://pinyin.sogou.com/dict/cate/index/14     |
| Natural Science | Geography and geology | https://pinyin.sogou.com/dict/cate/index/26 |
| Social Science | Politics            | https://pinyin.sogou.com/dict/cate/index/91     |
```

Step 2. According to the paging information and links, get the content of each page list, use regular expressions to get the download link of each lexicon under the category, the code is as follows:
Download it according to its link. The code is as follows:

```python
for i in range(1, int(maxPage) + 1):
    time.sleep(0.3)
    host = url + '/default/*str(1)
    content = urllib.urlopen(host).read()
    list = re.compile(reg, re.MULTILINE).findall(content)
    for j in list:
        print(j)
        urlOutputFile.write(\ + *\n")
```

Step 3. After downloading, convert scel format to text data in txt format.
Step 4. Reorganize the text content in the format of [user-defined word] [part-of-speech [word frequency].

3.4. Experiment and evaluation

After the above steps, the dictionary as a file namely a sougou.txt is obtained, which is 119MB in size and contains 3,999,925 entries. The following experiments are conducted on Baidu-based dictionary, Sogou-based dictionary and default dictionary of Ansj. The experimental data is the content and size of the book named ‘High School Chinese Knowledge List’ as shown in Table 2.

| Files             | Size    |
|-------------------|---------|
| Backcover.html    | 1KB     |
| Chapter_01        | 764KB   |
| Chapter_02        | 435KB   |
| Chapter_03        | 392KB   |
| Chapter_04        | 330KB   |
| Chapter_05        | 201KB   |
| Coverpage.html    | 1KB     |
| Postscript.html   | 42KB    |
| Preface.html      | 3KB     |

Table 3 shows the comparison of the number of the segmentation words based on three different dictionaries. It can be seen that there is not much difference, indicating that it is not appropriate to evaluate only by the number of word segmentation. Therefore, it is necessary to evaluate the results of word segmentation directly. From the segmentation results, it can be seen that among the three word-segmentation methods, the results of introducing user-defined dictionaries are obviously better than those not introducing. If the contents of the two dictionaries are compared separately, the coverage of Baidu-based dictionary is not as high as that based on Sogou-based dictionary.

| Files             | Baidu-based dictionary | Sogou-based dictionary | Default dictionary of Ansj |
|-------------------|------------------------|------------------------|---------------------------|
| Backcover.html    | 2                      | 2                      | 3                         |
| Chapter_01        | 56797                  | 56664                  | 56491                     |
| Chapter_02        | 32184                  | 30049                  | 33031                     |
| Chapter_03        | 25936                  | 24656                  | 26581                     |
| Chapter_04        | 32009                  | 32438                  | 31851                     |
We can see that professional content must be added to the professional dictionary for intervention, otherwise the result of word segmentation will be very fragmented. Some optimizations after adding user-defined dictionaries are also introduced, such as the addition of user-defined stop words to filter out some useless word segmentation items.

4. System Design and Implementation

The technological framework of the improved full-text retrieval system is shown in Fig.1. The full-text retrieval system has four modules, which are the resource collection module, the text pre-processing module, the content index module and the resource retrieval module.

The resource collection module is the source of full-text retrieval data content. Its main working process is to obtain the stored and analysed path information of the resource when it is put into the database. According to the information, obtain the resource storage location, find the directory of the text content of the resource, and get the HTML file to prepare for the subsequent text pre-processing.

The text pre-processing module is mainly responsible for the pre-processing of structured text extraction, word segmentation, and removal of stop words on the collected information, and constructing the word-list of index. The main working process is to first make it a format that the system can handle, that is, extract the content of the collected HTML files into structured plain text after a series of actions such as tag filtering, irrelevant information removal and compression; second use the Ansj Chinese word segmentation toolkit and combine with a user-defined professional dictionary to achieve word segmentation; then introduce user-defined stop word dictionary to remove irrelevant interference; finally obtain the processed text lexicon.

The content index module is the core of retrieval, which is to achieve the index of text lexicon. Through research and analysis, the index work in this paper is based on the index technology of Lucene, and the API provided by Lucene is used to create and maintain the index.

The resource retrieval module includes two submodules, namely the foreground and the background. The foreground is responsible for the search interaction process with the user in the Web interface, in order to provide search services for the user. Users can query in the search interface and view the returned search results on the foreground interface. The background is responsible for providing search interface and user interface data display interface. According to the search keywords, submit them to Lucene's search module, call the retrieval API to search, merge, score and finally return the sorting results. The foreground and background interact with each other to provide users with a complete search service.
5. **Application Effect**

The improved full-text retrieval system has been accomplished and integrated into the resource database which is successfully applied to the massive knowledge base platform for elementary education. It can run well after testing.

Up to now, there have been more than 1,000 textbooks and tutorial books in the platform. It cannot meet the users' needs for resource content retrieval and knowledge acquisition only using the title and classification search. The improved full-text retrieval system has the following functions:

1. Provide full-text retrieval function of textbooks and tutorial books. Combine and display related search results in the same resource file, and display the context summary of the search keywords. Users can preview the results directly in the search list without having to perform the second query in the search results, which saves the query time and improves the query efficiency.

2. Click the summary of search results to locate the keywords directly. The system will highlight the keywords to facilitate users to view the content.

3. Provide basic information about the resources, including the name, applied discipline, version, applicable grade, published version, cover and ISBN of the book. Click the book name to view the detailed content of the book.

4. Provide users to load search results on demand. When the user scrolls to the bottom of the screen, the user requests data once until there is no more data to display. Returning to the top is provided to facilitate the modification of query conditions for re-query.

6. **Conclusion**

Whether or not education resources can be used effectively by teachers and students is a key issue in the development of education informatization. Providing fast and efficient resource search services for teachers and students is a function that the current elementary education resource database system needs to have. The retrieval method based on full-text retrieval can provide users with fast and accurate retrieval services.

This paper establishes an improved full-text retrieval system based on the elementary education resource database system to provide users with full-text retrieval services of resources, which has achieved the expected goals. The paper focused on the following works:

1. Study the Chinese word segmentation methods and the common Chinese word segmentation toolkits for comparison and evaluation purposes.

2. Establish a professional dictionary for subject resources, optimize the results of the segmentation and reduce the irrelevance of the searching results set, as well as reduce mistakes and omissions in choosing the results by combining the word segmentation tools.

3. Design the workflow of the full-text retrieval system framework and various sub-modules.

4. Achieve full-text retrieval of elementary education resources database system based on Lucene technology.

In the future, we will do further research and discussion on relevance recommendation, Chinese word segmentation improvement, search term subject judgment and search intention understanding.

**Acknowledgments**

We acknowledge financial support from the National 863 Project of China No. 2015AA015401.

**References**

[1] Sun Xiquan, Ma Ruifang, et al. Research and application of information retrieval based on Lucene[J]. Information Studies: Theory & Application, 2006, 29(1):125-128(in Chinese).

[2] Wikimedia Foundation, Inc. Precision and recall[EB/OL]. https://en.wikipedia.org/wiki/Apache_Lucene.

[3] Liu Xueqin, Wu Hongxia, et al. Research on full-text retrieval technology for Chinese character[J]. Journal of Information, 2007, 26(1):52-53(in Chinese).

[4] Wikimedia Foundation, Inc. Apache lucene[EB/OL]. https://en.wikipedia.org/wiki/Apache_Lucene.
[5] Lin Shen. Research on word segmentation algorithm and improvement of national characters [J]. Computer knowledge and technology, 2017,13(31):199-200.
[6] Zhang Aimin. Research on natural language processing and intelligent search engine model[D]. Lanzhou University of Technology, 2003.
[7] Gaga Mao. Study on Chinese word segmentation[J]. Advances in Higher Education, 2019,3(3):1-4.
[8] Sun Jian. Ansjseg[EB/OL]. https://github.com/NLPchina/ansj_seg.
[9] Sogou. Sogou Cell Lexicon[EB/OL]. http://pinyin.sogou.com/dict.