Research the key technologies of the Mongolian full-text retrieval based on Lucene

Ding Guoqiang
Mongolia Normal University, Computer and Information Engineering
IMNU
Hohhot, China
dgqdingding@163.com

Lin Min
Mongolia Normal University, Computer and Information Engineering
IMNU
Hohhot, China
dgqdingding@163.com

***Abstract*—Under the premise of in-depth understanding of Lucene full-text retrieval technology, this paper will apply it to the Mongolian text search. First, several key issues are proposed which are need to be addressed in achieving the Mongolian text search technology, and give the corresponding solutions to achieve the Mongolian full-text retrieval in Lucene. Second, this paper provides a fast, accurate and comprehensive Mongolian information full-text search service, played a key role in promoting the development of the Mongolian search engine.**

Keywords- Lucene;Mongolian;Full-text Search

I. INTRODUCTION

This With the development of computer and network technology, global information resources are shifting the direction to digitization and networking. In recent years, the technology of the Mongolian information processing has made some achievements both in theory and application research. With the help of the strong support of the government and the Mongolia people's hard efforts, the Mongolian cyber source becomes more and more rich . Along with the growing of the Mongolia people's network knowledge, obtaining rich information from the Internet and learning advanced knowledge of the culture have become an urgent need of the majority of Mongolian. The process of accessing to the information from the network does not work without the application of search engine.

On January 2008, the first Mongolian Internet search engine (http://www.qgool.com/) was officially launched, therefore, the Mongolian Internet has had its own search engine; On December 2010, with the trial of the Mongolian search engine of Mong Keli (http://hai.menksoft.com), the technology of the Mongolia Internet search engine was further developed. But so far, the technology of the Mongolian search engine still falls behind the English and Chinese search engine. The Mongolian search engine still exists some deficiencies in the technical aspects:

①Poor universality. It can only search the webpage that are appointed by the specified code's Mongolian search engine, and the others cannot be searched, but the fact is that the quantities of webpage which satisfy the constraints are little;

②Lack of real time. It can not update the content on the web timely, therefore, it can not retrieve the updated content eventually;

③The retrieval result is not optimized. The retrieval results exist the duplicate content, there is not the processing of removing duplicate; the retrieval results are neither according to the importance of content, nor by content time order, there is no regularity, it does not meet the people's habits;

④ The application of retrieval technology based on content was little. It does not play Lucene's great role in search engine.

Lucene as a typical text retrieval engine kit, with its open-source and cross-platform use, gets more and more search engine companies' favored. Many Java projects use Lucene as the full text search engine. Full-text retrieval based on Lucene does not need to consider the data format, it can retrieval all that as long as it can be converted to text type. Furthermore, the realization is simple and the scalability is very strong.

At present, the English and Chinese full-text retrieval technology based on Lucene has been gradually mature, but China minority nationality language and characters are different from English and Chinese, therefore, we must consider the specific language characteristics to study. Yong Cuo[1], Jiang Ming Yuan[6], Li Ying Xing[2]studied the Tibetan information retrieval technology based on Lucene, summarized some key problems that need to be solved in the Tibetan language full-text retrieval technology, and proposed the corresponding solutions. In their papers they referred to the Tibetan segmentation problem, they used the Lucene and a specific tokenizer for the Tibetan segmentation, indexed the segmentation results, and then provided the fast, accurate and comprehensive full-text retrieval service for the Tibetan information, improved the retrieval efficiency of the Tibetan information and promoted the search engine development of the Tibetan language. Therefore, full-text retrieval technology based on Lucene applied to the Mongolian information retrieval will play a vital role in the development of Mongolian search engine.

II. TECHNOLOGY RESEARCH

The Mongolian full-text retrieval based on Lucene need to solve including Mongolian coding is not uniform,

Mongolian word segmentation and Mongolian normal display are difficulties and so on. Therefore, this paper will realize the Mongolian full-text retrieval technology with the steps of pretreatment, create indexing and retrieval.

A. Pretreatment

The Mongolian joined the international standard encoding later, and the traditional Mongolian exists many non uniform encoding, such as "Sai Yin", "Mong Keli", "Oyuta (intelligent)", "founder", "Burigude" , "Ming Antu " and so on, but these encoding is not compatible with each other. Therefore, the webpage which under the different encoding are different too[3], and a lot of sources are not compatible and share with each other. So we must make the encoding unified if we really want to establish standardized language data resources. So far, there is not a mature tool for Mongolian code conversion. Inner Mongolia University and Inner Mongolia Normal University are doing the appropriate research in this field now. The method of Inner Mongolia University's Latin Transliteration is more effective. But this method is not only trouble, but also very easy to confuse with Latin text, and for part of the rules of complex Mongolian cannot identify.

In addition, due to the slow development of Mongolian input method and not uniform, there is still not a real authority and general Mongolian input method until now. The result is that many Mongolian editors are not standardized in the input corpus.

In order to solve a series of questions which led by the Mongolian full-text retrieval in different encoding, this paper will set the test corpus data into the international standard code -- Unicode encoding, all of its preservation for the "UTF-8 ". At the same time, we also set the Java project file to the "UTF-8" format that modifies the configuration file attribute value character-encoding="UTF-8", which avoids the encode problems in the process of implementing procedures.

In addition, in order to better retrieval by Lucene, we still need to pretreatment the data following two steps:①Cut the test corpus data file into many small files with the same size (such as 10KB). So we can be more rapidly positioning to find words in the text file when we need to output of the retrieval results, which can further improve the retrieval efficiency. ②Unified the special characters (such as punctuation, digital symbol) in the test corpus data files into alphanumeric format, which can solve the issues of confusion that caused by the full-width and half-width format in the Mongolia characters, and make Lucene can distinguish the delimiter better, which provide the basis for Mongolian segmentation and index.

B. Create Index

There are two steps for the full-text search based on Lucene: index and query. They are necessary to use the word segmentation. The process of creating index by Lucene is to transform the inverted list by Lucene into the unique format (such as .fdt, .fdx, .fnm, .frq, .nrm, .prx, .tii, .tis) of the index file and then to turn these files into a special format file(.cfs). This can reduce the time of accessing to the small files, and improves the speed of retrieval greatly.

The Lucene source code contains seven packages, each package is responsible for the completion of a specific function (see Table 1). This package of "org.apache.lucene.analysis" is used to word segmentation [2].

TABLE I. THE TABLE OF LUCENE'S PACKAGES

Package name	Function
org.apache.lucene.analysis	Language parser
org.apache.lucene. document	Storage structure of the index management
org.apache.lucene.index	Index management
org.apache.lucene. queryParser	Query Analyzer
org.apache.lucene.search	Retrieval management
org.apache.lucene.store	The bottom of the I/O data storage
org.apache.lucene.util	Some public class

For English, because it has spaces between words as natural delimiters, so the technology of word segmentation is simple. We can use the spaces and punctuation marks as the separator for segmentation. The segmentation such as StandardAnalyzer, SimpleAnalyzer etc. that Lucene own can achieve it. But for Asian character especially for Chinese, since there are no natural delimiters as separate between words, the word segmentation technology is relatively more complex. For the Mongolian, it is a kind of typical agglutinative language, which is different from the English and Chinese, although it also has a space between words as a separator, but it has the control characters (Mongolian control characters and functions are shown in table 2[4]) in its word internal, which can bring the confusion delimiter. So it will lead to the correct rate of decline in the word segmentation. The tokenizer of Lucene own and the Chinese word segmentations are not as one wishes. Therefore, according to the characteristics of Mongolian to design our own tokenizer is a problem that needed to be solved urgently.

TABLE II. MONGOLIAN CONTROL CHARACTER AND FUNCTION

Control character	Code	name	function
	U+ 180B	Mongolian free variant selection 1	For the difference in the same under conditions of the same character variants
	U+ 180C	Mongolian free variant selection 2	
	U+ 180D	Mongolian free variant selection 3	
			Used between

	U+ 180E	Vowel delimiter	the character A/E and the preceding consonant.
	U+ 202F	Narrow width no-break space	Used to write additional ingredients before
	U+ 200C	Zero width Joiner	Used to express a single character variant or word is disconnected, and the preservation of its original shape
	U+ 200D	Zero width joiner ban	For mandatory to disconnect the word normal writing

In order to let Lucene more accurately segment in indexing, we must construct our own tokenizer according to the characteristics of Mongolian. First of all, we need to expand the Lucene language lexical analysis interface to achieve the support of Mongolian, which is introduced into the project files, that is "import cn.edu.imnu.lucene.mong.analysis.MongAnalyzer". Here the "MongAnalyzer" is the Mongolian analyzer which is this paper designed. It is a successor to the Lucene Analyzer, called the Mongolian word segmentation (MongTokenizer), Mongolian stop word filter (MongStopFilter) and Mongolian stems and affixes filter (MongStemFilter), eventually returned to a TokenStream, provided the basis for index.

The Mongolian word segmentation (MongTokenizer) inherited from the class of CharTokenizer in Lucene, and it rewrote the method of TokenChar (char c) which in the class of CharTokenizer. Because the original method of “isTokenChar (char c)” can not identify the Mongolian control characters, the Mongolian control characters will be mistaken for other "delimiter", which led to the word segmentation is not accurate, and appeared to the phenomenon of "excessive segmentation". Therefore, this paper in accordance with the international standard code of the Mongolian spaces, punctuation and end-of-line to distinguish delimiter, which can preclude the confusions be brought by control characters.

Mongolian document will also appear the Mongolian stop word similar to the English and Chinese, which appeared high frequency and meaningless, and if these words are effectively filtered, the retrieval efficiency will be greatly improved. Therefore, this paper according to the Mongolian characteristics design a contains 99 Mongolian words Mongolian words dictionary, and then design a Mongolian stop word filter (MongStopFilter), which is a successor to the class of TokenFilter in Lucene. Its main method is the MongStopFilter (Boolean enablePositionIncrements, TokenStream input, File stopwordsFile) that filtered the Mongolian stop word.

Mongolian is a typical agglutinative language. Its formation and the configuration contain root, stem and different affixes. Each word formation and its grammatical are dependent on different affixes, so the correct segmentation of root, stem and affix to reveal its word’s attribute and grammatical relations is a great significance. Therefore, this paper according to the relations of the Mongolian root, stem and affix design the Mongolian affixes filter(MongStemFilter) based on the dictionary of Mongolian stems and affixes. It is also a successor to the class of TokenFilter in Lucene. It output the token formed in previous step with string type, and then segment the Mongolian word by the method of reverse maximum matching algorithm based on the affix lexicon, the final return a result of string type, which covers the original vocabulary unit and be used for index.

At the same time, we need to modify the procedures in corresponding place. For example, find the procedures entrance statement applied the segmentation "IndexWriter writer = new IndexWriter (INDEX_STORE_PATH, new StandardAnalyzer (),true" in the class (IndexProcesser)of Lucene indexed, where Lucene make the standard tokenizer (StandardAnalyzer) as the default segmentation. We should change it to this: "IndexWriter writer = new IndexWriter (INDEX_STORE_PATH, new MongAnalyzer (),true ".This used the segmentation that our designed. It effectively solved the problem that the other word segmentation are not very good for Mongolian word indexing, which improved the accuracy of the results.

C. Retrieval

When the Lucene indexed for the test corpus data file, the next is retrieval in the index. The specific steps are as follows:

Firstly, determine the index file’s storage location, according to the index file’s storage location to build a query object, i.e. "IndexSearcher searcher = new IndexSearcher (INDEX_STORE_PATH)";

Secondly, establish the search unit, "Term t = new Term (searchType, searchKey)". The searchType represents the filed of need to search, and the searchKey represents the key;

Then, access to the object of <document, frequency>, "TermDocs termDocs = searcher.getIndexReader ().TermDocs (t) ";

Finally, search the keywords position and frequency of occurrence through retrieved circulating retrieval in the index file.

III. EXPERIMENTAL RESULTS AND ANALYSIS

Based on the above experimental procedure, this paper selects Inner Mongolia archives data as the test corpus, whose size is 33.4MB, retrievals the part of the Mongolian

words, counts the total time of retrieval, and compares to the method of used of string matching search. The experimental results of partial screenshot below, which Fig.1 is the results that use of Mongolian word segmentation, the selected keywords contains Mongolian control character (U+180D). Fig.2 and Fig.3 are the results that using the method of Lucene and string comparison for the full-text retrieval.

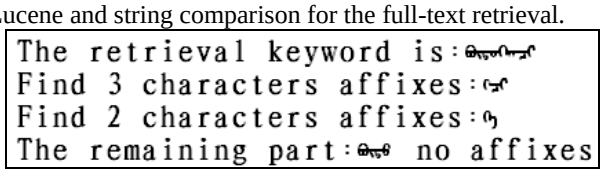

Figure 1. The results that use of Mongolian word segmentation

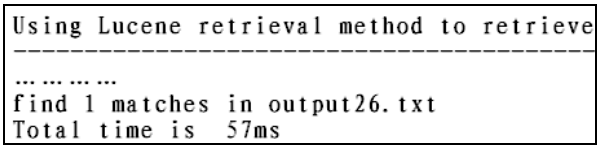

Figure 2. Using a Lucene retrieval method to retrieve

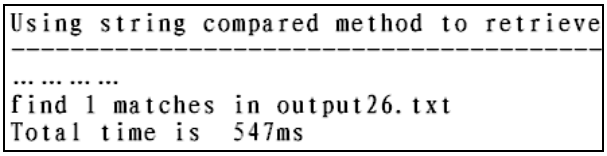

Figure 3. Using a string compared method to retrieve

It can be seen from the experimental results that the use of Lucene and the Mongolian word segmentation to Mongolian full-text retrieval capable of excellent identification of Mongolian control characters, and can correctly according to Mongolian affixes dictionary to affix thematic segmentation, can be effective for document indexing and retrieval of the Mongolian words, and retrieval of total time consuming than the string matching method to improve by more than ten times, greatly improves the efficiency of retrieval.

IV. CONCLUSIONS

This paper studied the Mongolian full-text retrieval technology based on Lucene, summarized several key problems in the Mongolian full-text retrieval, and put forward the corresponding solutions, realized the Mongolian full-text retrieval. This technology not only can provide services to Mongolian search engine, and provide support for small and medium enterprises website search, but also can build individual user desktop search engine and a specific document retrieval database, thus realized the target document convenient retrieval management and improved the efficiency of retrieval.

However, for more complex language keywords, the retrieval accuracy is not very high, the reason is that the Mongolian word segmentation is not mature enough, for more complex Mongolian words, it can not well segmentation, so the indexing is not accurate, eventually make the retrieval results accuracy rate is not high, therefore, the next research work focus is complete the tokenizer. In addition, this paper is only experimental to Mongolian keywords retrieval, and none of the document highlighted or in the browser output, but not optimization and scheduling, did not fully play out the powerful function of Lucene, so this is the next step of the research content too. Due to the Mongolian in the document is from top to bottom and from left to right in written norms shows, this is different from English in from left to right horizontal display rules, At present, real research on this question are few, but the IE browser version 8 or above can solve this problem very well, therefore, the next step is that show search results with the help of IE browser version 8 or above.

ACKNOWLEDGMENT

This paper was supported by the Management Software Project Foundation of the Inner Mongolia Archives(nation archives 10-X-2007) and the Master Graduate Student Research and Innovation Project Foundation of Inner Mongolia Normal University in 2011(CXJJS11071).

REFERENCES

[1] Yong Cuo, “Based on the Lucene Tibetan text retrieval study [J]”, Journal of University of Tibet (NATURAL SCIENCE EDITION) .2009,24 (1): 58-60

[2] Li Yingxing, Fu Ting and Li Yong, “Lucene based Tibetan information retrieval research and application of the national language information technology”, the Eleventh National Minority Language Information Symposium [C] .2007

[3] Wang Rui, Mongolian webpage crawling and identification of coding conversion of Inner Mongolia University. In May.2008

[4] Meirong Bao, Sriguleng Wang, GaRidi and Min Lin, “Amendments to the Rules of Traditional Mongolian OpenType Font”, 2012 in computer science and information engineering, security international conference in June.2012

[5] Jiang Hua, “Based on Lucene topic oriented search engine research and design [D]”, East China Normal University.2007

[6] Jiang Mingyuan and Kong Lingde, “Based on the Lucene Tibetan information acquisition and retrieval system”. North Central University, electronics and computer science and technology, 50 (2011) 02-00 34-04 .1003-58

[7] Wu Haiming, Lucene based search engine technology research and improvement of Jinan University.2006 [D].

[8] Che Dong, “Lucene: Java based full-text search engine”, 2007 year in April.

[9] GospodneticO and HatcherE, Lucene in Action, [s.l.]:Manning Publicatings Co.2009