Available online www.jocpr.com

Journal of Chemical and Pharmaceutical Research, 2014, 6(6):1958-1966

Research Article

ISSN : 0975-7384
CODEN(USA) : JCPRC5

Application of textual relevance retrieval in patent information service

Li Dan and Yang Ting

School of Information and Computer Engineering, Northeast Forestry University, Harbin, China

---

ABSTRACT

In this paper, the patent search service is based on the patent textual relevance information of the actual application requirements. In depth study of the Lucene full text retrieval tool kit and related technical basis,we extend the Lucene segmentation module,and improve the ordering algorithm Lucene by default, so that the patent text retrieval system designed in this thesis retrieval can support multiple patent document format which is commonly used, and eventually patent information retrieval system used in this study above, can effectively improve the performance of patent related retrieval system.

Key words: Lucene; textual relevance retrieval; patent

INTRODUCTION

The word ‘patent’ comes from Litterae patentes in Latin, which means the open letters or public literature.Patents are issued proof of medieval monarchs used some kind of privilege, and later refers to the exclusive rights of a certificate signed by the King of England himself.A patent is the world's largest source of technical information, according to the empirical statistical analysis, patent contains 90% -95% of the world scientific and technical information. Patent Information widespread and permeate all areas of science and technology, economic and social life, with a variety of information in one set, a huge number of features. A wide range of subjects covered by the patent information, information publishing fast, innovative, highly standardized, as well as reveal the full details inventions content. It has become engaged in scientific research, technology development and other social and legal norms. It has become an important information in scientific research, technological development and legal norms and other essential social and economic activities.At present, the world has about 40 million items over the patent document.In today's rapid development of technology innovation, patent information is at a speed of more than 100 million pieces per annum growth.

Retrieval technologies for the analysis of a number of patents and temporal distribution, status and trends may reflect the development of the technology. With the technology in an increasingly competitive world, countries are increasingly focusing on patent strategy research, and though the processing and combination of patent information in patent specification and Patent Gazette ,and finally make this information into the overall situation and forecasting capabilities with competitive intelligence to provide information to support strategic decision-making enterprise. With the rapid development of computer and network technology, people can retrieve patent databases through the Internet and a variety of patent-related information which has become increasingly diverse. How to take full advantage of such huge information resources so that patents play an important role in many aspects of business research and patents, patent information retrieval systems have come into being.

Full text retrieval is to search the contents of any information stored in a database in the whole book and the paper . It can be obtained according to the need full text of chapter, section, paragraph, sentence, word information, that is similar to the word for the whole book to add a label, which also can be a variety of statistics and analysis.

Full text retrieval technology, is based on data such as text, sound, image and so on as the main content, to retrieve

the document content rather than the appearance. With the development of computer industry, the computer storage devices carriers more and more electronic information, the information can be roughly divided into two categories: structured data and unstructured data. Usually structured data refers to enterprise financial accounts and production data, student score data and so on, and the non structured data is some text, image data sound and other multimedia data and so on. According to statistics, unstructured data occupies more than 80% of the volume of the whole information. Full-text retrieval technology is one of the most common information retrieval applications, and also is a very efficient information retrieval technology. It is a powerful tool for dealing with unstructured data, but also the general search engine information retrieval tools. It greatly improves the data to find specific information from a vast complex data efficiently.

Lucene is a Java full-text search engine. Lucene is not a complete application, but rather a code library and API that can easily be used to add search capabilities to applications. Lucene is a high-performance, scalable full-text information retrieval tool kit, and it is not a full-text search engine, but a full-text search engine architecture, providing a complete query engine and indexing engine, part of the text analysis engine (in English and German, two kinds of Western languages).It provides a simple, but very powerful core API, people can quickly and easily integrate it into the application to increase the indexing and search capabilities. Lucene is designed to provide an easy-to-use software developer kit to facilitate the realization of full-text search function in the target system, or as a basis for establishing a full-text search engine.

In this paper, the patent search service is based on the patent textual relevance information of the actual application requirements. In depth study of the Lucene full text retrieval tool kit and related technical basis,we extend the Lucene segmentation module,and improve the ordering algorithm Lucene by default, so that the patent text retrieval system designed in this thesis retrieval can support multiple patent document format which is commonly used, and eventually patent information retrieval system used in this study above, can effectively improve the performance of patent related retrieval system.

LUCENE

A. Introduction of Lucene

Lucene is a Java full-text search engine. Lucene is not a complete application, but rather a code library and API that can easily be used to add search capabilities to applications. Lucene is a high-performance, scalable full-text information retrieval tool kit, and it is not a full-text search engine, but a full-text search engine architecture, providing a complete query engine and indexing engine, part of the text analysis engine (in English and German, two kinds of Western languages).It provides a simple, but very powerful core API, people can quickly and easily integrate it into the application to increase the indexing and search capabilities. Lucene is designed to provide an easy-to-use software developer kit to facilitate the realization of full-text search function in the target system, or as a basis for establishing a full-text search engine. The abundant use of Strategy design pattern in Lucene toolkit, make the application interface design flexibility. Then users can take advantage of these interfaces, customized to suit their own needs language parser, query analyzer and crawler.

Doug Cutting originally wrote Lucene in 1999[1]. It was initially available for download from its home at the SourceForge web site. It joined the Apache Software Foundation's Jakarta family of open-source Java products in September 2001 and became its own top-level Apache project in February 2005. Until recently, it included a number of sub-projects, such as Lucene.NET,Mahout, Solr and Nutch. Solr has merged into the Lucene project itself and Mahout, Nutch, and Tika have moved to become independent top-level projects. Apache Lucene is a free/open source information retrieval software library, originally created in Java by Doug Cutting. It is supported by the Apache Software Foundation and is released under the Apache Software License.Lucene has been ported to other programming languages including Delphi, Perl, C#, C++,Python, Ruby, and PHP[2][3].

B. Lucene's structure and function

Lucene full-text retrieval system has two main functions: one is the establishment of index database, which is about to be indexed data source via the parser parse, the content of the cut after word segmentation index storage, the second is the search index library, which meet the conditions of the document to find out from the index database according to the user's query condition, will result in in order to return to the user [4][5] .

The Lucene source code has 7 packages; each package can complete a specific function. There are 3 main core classes: language analysis package, indexing management package and retrieval package. Concrete analysis is as shown in tbale 1:

Table 1. Lucene Function

Name	Function
org.apache.1ucene.analysis	language analysis
org.apache.1ucene.document	document management
org.apache.1ucene.index	index management
org.apache.1ucene.queryParser	query parser
org.apache.lucene.search	search management
org.apache.1ucene.store	data storage
org.apache.1ucene.util	public classes

● **org.apache.1ucene.analysis**

The language analysis package is mainly used for the source file content of word segmentation, this is usually achieved by the extension of Analyzer (classes such as SimpleAnalyzer, StandardAnalyzer etc.), cutting back to a TokenStream,Then use the TokenStream in the next () method to remove a word. According to certain segmentation rules,the method would put an article from A to Z divided into words, and after processing, calculation of each word in the position and frequency of occurrence in the article. The default language parser can provide English, German and Russian analyzer.

● **org.apache.1ucene.index**

Index Management Pack is the core of the whole system, the main provider of the library to read and write interfaces. This package contains IndexWriter and IndexReader within two classes, it can call other classes within the package, in order to complete the creation of index database, add, modify, delete, index and reading index and so on. Initialization and full-text indexing records are loaded through the class to complete.

● **org.apache.lucene.search**

Retrieve the package mainly provides retrieval interface, you can create a search of the bag by calling Searcher class. When you enter a query, retrieval device to retrieve the index file by analyzing the query, get the query result set. The retrieval device using the method of IndexSearcher.search (Query), returned the Hits resultset. In addition, with the query analysis package, you can customize the query rules, support "and", "or", "not", "belonging" among other complex query query.

Structure of Lucene is clear, and each packet has its own mission, such as org.apache.lucene.search for search,org.apache.1ucene.index is responsible for the index,and org.apache.1ucene.analysis is responsible for the segmentation of words. The main action of Lucene is used in the abstract classes, which is facilitate for expansion.

C. Lucene index structure

Lucene uses the inverted index structure as an index of the center of the word to establish the mapping relationship between words==>document. When searching, it is based on the word to search for documents, rather than by looking for documentation to find words.

● **Segment**

Lucene index may consist of multiple sub-index composition, these sub-index called segments. Each section is completely independent of the index that can be searched. The index is: to create a new paragraph, merge paragraphs which already exist for the newly added document. When searching index, more than one segment or multiple indexes may be involved, and each index may be provided by a number of segments.

● **Document**

Lucene uses an integer number of documents to indicate the document. The first index added to the index of the documents is the No. 0, and the order will be added to a document by a number incremented from the previous number. When deleting and inserting, document number will change, so it must be careful to store these numbers within Lucene external storage.

● **Field**

Field is an association of tuples, consisting of a name and a value domain. Field name is a string, and the value of the field is a term, such as "title" to form a field. The title should be used in the search results, so it will be added to the document as a field object. These fields can be indexed, also can not be indexed, while the original data can also choose to save in the index. Field stored in the index of the search results page which is created while searching is useful. Field values can also be tokenized, which means that an input to the analysis procedures will be broken down into the contents of the token that can used in the search engine.

● **Term**

Term is the smallest unit index concept; it directly represents a string and its position in the file, the information such as occurrences.

The Lucene index is composed of several segments, each segment is composed of a number of documents, each document is composed of a number of fields, each field is composed of a number of terms.Term is the smallest unit index concept, it directly represents a string and its position in the file, the information such as occurrences.Field is an association of tuples, consisting of a name and a value domain. Field name is a string, and the value of the field is a term.

D. Application and advantages of Lucene

Lucene API interface design is relatively common, and the input and output structures are like a database table ==> records ==> field.So many traditional application documents, such as the database can be easily mapped to Lucene storage structure interface. Overall, we can use Lucene as a support full-text index database system.Usually followed by a relatively thick books are often attached to keyword index table (for example: Shandong: 12,34 pages, Jiangsu: 3,77 pages.....), it can help readers more quickly find relevant content pages. The database indexes can greatly improve query speed. The database index is not designed for full-text indexing, so in the use of "like"% keyword% "", the database index is ineffective. When you use "like" query, the search process has become ergodic process which is similar to the open books page by page, so the “like” query is harmful for the database service which contains fuzzy query .So the establishment of an efficient retrieval system is the key to establish a similar technology index reverse indexing mechanism, the data source (such as articles)sort order is stored at the same time, there are also a sorted list of key words, used to store the keywords = = > article mapping relationship, the mapping relation between the index of this: [keyword ==> words article number, number of occurrences (including position: start offset, the end offset), frequency of occurrence]. retrieval is a process of fuzzy query into multiple exact query logic combination can use the indexing process. So as to improve efficiency, multi keyword queries so, full text retrieval problem boils down to the end is a sequencing problem. Fuzzy query retrieval process is to become more than one process that is a logical combination of precise queries using the index. Thus greatly improve the efficiency of multi-keyword query. To make a long story short, full text retrieval problem is a scheduling problem.

The core features of Lucene index structure through a special mechanism to achieve the full-text index, and traditional database full-text indexing mechanism is not good, this mechanism provides full-text indexing extension interface to facilitate for different applications from the line customization. The biggest difference is that full-text retrieval and database applications: Let the first 100 results are most relevant to meet the needs of more than 98% of users.

Lucene is different from most of the search (database) engine, it does not use B tree structure to maintain index. The B tree structure will lead the index updating need a lot of IO operation. Lucene constantly creates new index file in the extended index, then periodically put these new small index file into the index (originally in the update strategy, different batches of size can be adjusted, strategies can be customized). In the premise of not affecting the retrieval efficiency, improve the index efficiency. Lucene as the engine frame a good text retrieval, with open source, cross platform, is not limited to the data source, easy to expand, higher index efficiency advantages, very suitable for constructing full-text retrieval system. Without affecting the efficiency of the premise retrieval, Lucene improve the efficiency of the index. As an excellent framework for full-text search engine, Lucene has open source, cross-platform, and it is not limited to the data source, and easy to expand with index and high efficiency. It is suitable for construction of full-text retrieval system.

CHINESE WORD SEGMENTATION TECHNOLOGY

A. Necessity of Chinese word segmentation

Information retrieval is based on the analysis of the text, and the text analysis is largely a process of language. Currently in natural language processing technology, the Chinese processing technology lags far behind Western

processing techniques.Because of English takes the word as a unit, separated by space, so English segmentation algorithm is easy to implement. And Chinese takes the character as a unit, a word or a plurality of word combination can have their own meaning, no distinction between marked between words, to achieve a certain degree of difficulty, so Chinese word meaning is particularly important.

Chinese word has a great influence on the Chinese search engine. For search engines, the most important thing is not all found, but the most relevant results at the top, which is also known as relevance ranking. Chinese word accurate or not, will directly affect the relevance of search results sorted. For a good search engines, word is essential to a core module, the quality of the analysis will directly determine the search engine results relevance and accuracy.

● Existing segmentation methods

At present the research of Chinese word segmentation method mainly has three aspects: word segmentation method based on string matching, word segmentation method based on the understanding and word segmentation method based on statistics.

● Word segmentation method based on string matching.

Segmentation based on string matching method, also called mechanical segmentation method, which is based on a certain strategy to match the character string and word thesaurus entries .If you find a corresponding entry in the thesaurus, the matching success. Currently practical systems are basically word segmentation method based on string matching, supplemented by a small number of lexical, syntactic and semantic information. The direction of the scan is different, based on string matching segmentation method can be divided into forward and reverse matching methods matching word segmentation method; according to different priorities match lengths can be divided into a maximum matching word segmentation method and the minimum matching method.

In general, there are three basic word segmentation method in mechanical word segmentation method, which is the maximum matching method, the maximum reverse matching method and word by word traversal method. Advantages of lexical dictionary is based on a simple algorithm, easy to implement, maintain updated. But in the case of incomplete dictionary, the algorithm does not recognize text that appears in a large number of unknown words, the lack of global information, resulting in ambiguity segmentation problem. Moreover, because the word itself does not define a standard, and there is no uniform standard dictionary, so different dictionaries have different ambiguity, and poor accuracy of segmentation.

● Word segmentation method based on statistics.

The basic principle of word segmentation method based on Statistics: in view of the form, the word is stable word combinations, so in this context, more adjacent words appear at the same time, the more likely constitutes a word. Therefore the frequency or probability of co-occurrence words adjacent can well reflect the reliability of a word. Through a combination of frequency statistics of each word co-occurrence in adjacent large-scale corpus, to calculate the mutual information of their. Mutual information reflects the relationship between the degree of Chinese characters closely.

When the tightness is higher than a certain threshold, you can think of this word may constitute a word. This method is only on the frequency of word combinations corpus statistics, without segmentation dictionary, which is also called non-lexical or dictionary of probability and statistics methods. Word segmentation method based on statistic learning method is based on the large-scale corpus, this method is more widely used at present, and the effect is better. The statistical model used for: Hidden Markov model, maximum probability model, N model, meta language source channel model.

Advantage of statistical method is: from a large training corpus, it summarizes and analysis correlation information inside the language, and add it to the statistical model. For statistical method, the size of the training corpus, seriously affecting the effect of word segmentation, the credibility of the training set is small model is low, word segmentation effect is poor, while on the one hand, the training corpus large data sparseness problem, greatly reduces the effect of word segmentation;, on the other hand, different areas of the corpus for statistical model plays a decisive role, such as: news corpus and patent literature corpora is a corpus of different professional field, there are great differences in the content, the news corpus training went out of the statistical model segmentation patent corpus, is bound to get good segmentation effect.

The segmentation algorithm is much more mature and practical method of statistics and dictionary are effectively combined, for example, Chinese Academy of Sciences of the Chinese lexical analysis system ICTCLAS is used in multiple hidden Markov model. They are extended to the original hidden Markoff model, the model is then applied

to the splitting atoms, unknown word recognition based on multiple levels and categories of hidden Markoff the word segmentation, the segmentation algorithm is also insufficient, because hidden Markoff model assumption of independence, the inability to consider the characteristics of context, limited feature selection, so that access to context information of words from the training corpus, ignoring the context information for the segmentation of text. word segmentation method based on the understanding.

Word segmentation method is through the computer simulation based on the understanding of people's understanding of the sentence, to achieve the effect of word recognition. The basic idea is the participle of syntactic and semantic analysis at the same time, the use of syntactic and semantic information to deal with ambiguity. It usually consists of three parts: subsystems of word segmentation, syntactic semantic and total control part. Under the coordination of general control part, word segmentation subsystem can obtain about words, sentences and other syntactic and semantic information to judge the word segmentation ambiguity, which simulates the process of people's understanding of the sentence. This kind of word segmentation method needs to use a lot of knowledge and language Information. Due to the general and the complexity of Chinese language knowledge to the various language information is organized into a machine can directly read the forms, so at present based on the understanding of word segmentation system is still in experimental stage.

In contrast, mechanical segmentation method is simple, more concrete and practical, but also can achieve higher accuracy, but prone to ambiguity segmentation. Statistical segmentation method is not restricted to be working with text fields do not need a machine-readable dictionaries, but it requires a lot of training text to establish parameters of the model, the method of calculation than the larger, while its precision and training text word the selection. Understanding of word segmentation algorithm complexity is high, its effectiveness and feasibility needs to be further verified in practical work.

B. The difficulties of Chinese Automatic Word Segmentation

Automatic word segmentation is the first step in Chinese Natural Language Processing, not only in the Natural Language Processing occupies a very important position, but also restricts the Chinese information processing as a bottle neck in development, many researchers conducted considerable research on it, but there are still many problems not well solved. To sum up, the difficult points in Chinese word segmentation mainly has three aspects as follows: the standard of word segmentation, ambiguity and unknown words recognition.

● Structured word segmentation.

To make Chinese word segmentation, first word should be standardized, so as to allow the machine to correctly distinguish the characters, words and phrases. Although the Chinese word has many years of study course, but so far the country still do not have an open, widely recognized, actionable segmentation specification, nor is there a universal scale evaluation corpus. This makes the findings of many researchers lack a fair comparability, thereby constraining improve Chinese word segmentation technology, real mature public rarely useful segmentation tool used by the general public.

Chinese word segmentation is the first difficulty of the concept of the word is not clear. Written Chinese is the sequence of words, there is no obvious interval mark between word and the word, makes the term defined as the lack of common standards. What word is, namely, the abstract definition of words; What are words, words of specific definition. The definition of "word" in Chinese has been the Chinese language and the focus of the debate, on the one hand, how to distinguish the single words and morphemes; On the other hand, how to distinguish between words and phrases (words).Thus far have failed to come up with a recognized, authoritative dictionary. In addition, for the understanding of the Chinese "word", between the native Chinese speakers tester for testing, the results show that the words in Chinese text only 76% approval ratings, strictly speaking, Chinese word segmentation is a problem that there is no clear definition.

Geared to the needs of different applications, there may be different types of word segmentation system, and the segmentation result is difficult to use uniform segmentation criterion to evaluate, need to reach the effect of word segmentation also have very big difference. So united hard participle specification, one of the important reasons is the application of different systems has different requirements for the participle code. National bureau of standards issued in 1992 as a national standard of "standard of information processing in modern Chinese word segmentation", in this specification, most regulations are through examples and qualitative description to reflect. Many rules require participle unit to "combine closely, the use of stability", this regulation is easily influenced by subjective factors, it is difficult to grasp the operation scale. Thus the "specification" and not let people fundamentally to Chinese words a unified understanding, in this case, establish a fair and open Chinese automatic word segmentation evaluation standard as well. Since July 2003, the first international Chinese word segmentation Bake off evaluation activities

carried out since the Chinese automatic word segmentation technology has made gratifying progress, but due to provide training and test corpus Bake off unit has its own participle standard and vocabulary, is the main purpose of the evaluation is to promote the progress of Chinese word segmentation technology, didn't produce a uniform segmentation criterion. Therefore, for a word segmentation system, formulate a unified code of word segmentation is an important issue.

● **Ambiguity segmentation fields.**

Automatic second term difficulties faced by the ambiguity of the text segmentation field, ambiguity field situation is quite complex, there are two typical situations: the intersection of type and combinational ambiguity. Ambiguous phrases of overlap type. In field AJB, AJ∈W, and JB∈W, says AJB for ambiguous phrases of overlap type, among them A, J, B for string, W for glossary. Combination ambiguity fields. In the field AB, AB∈W, A∈W, B∈W, is called AB combination ambiguity fields. Among them A, B for strings, W for glossary.

● **The unknown word recognition**

Unknown words are defined as those that have not been system dictionary words, its variety, mainly includes the new technical terms and proper nouns, such as names, place names, Chinese translated name, organization name (referring to the organs, organizations and other institutions). Before an unknown word theory is to be expected, to artificial pre added to the vocabulary (but this is only an ideal state, in the real environment is not easy to do); after an unknown word is completely unpredictable, no matter how large vocabulary, also cannot include. In addition, because each specific word segmentation system using the dictionary capacity is not the same, a word is unknown word is relative to the Chinese word segmentation system specific terms. For a certain word segmentation system, it may be unknown words, but in another word segmentation system is not necessarily a nun known word.

Identify unknown words on a variety of Chinese information processing systems not only have direct practical significance, but also a fundamental role. Because a variety of Chinese information processing systems require the use of word frequency and other information, if word on the unknown word recognition system is not correct, statistical information will be a great gap. Such as text proofing system, if the system does not have the ability to identify new words, you can not determine the use of the word in a sentence is reasonable, and therefore could not check the mistake lies. To correct segmentation Chinese text statement requires segmentation system has a certain ability to identify unknown words, thus improving the accuracy of segmentation. The unknown words from universal existence in Chinese text, to identify the unknown words in Chinese automatic word segmentation system plays an important role, the current of the unknown word recognition accuracy evaluation has become an important symbol of word segmentation system is good or bad [7][8] .

FEATURES OF THE SYSTEM

Open your browser and enter "http://localhost:8080/MyTomLucene/index2.jsp", the browser will jump to the "TRIZ theory topics consulting quiz robot" home, in the middle of the pop-up page in the search box to enter your questions to consult, and then click on the search box to the right of the search button, as shown in Fig.1. To enter the question "What is TRIZ theory?" For example, click on the search button, the robot will be given a consistent problem with your input or relatively close to the answer, you can click to see the details related to the answer, the page is shown in Fig.2:

Fig. 1: Features of the System

Fig. 2: Retrieval result of the System

CONCLUSION

In this paper, the patent search service is based on the patent textual relevance information of the actual application requirements. In depth study of the Lucene full text retrieval tool kit and related technical basis,we extend the Lucene segmentation module,and improve the ordering algorithm Lucene by default, so that the patent text retrieval system designed in this thesis retrieval can support multiple patent document format which is commonly used, and eventually patent information retrieval system used in this study above, can effectively improve the performance of patent related retrieval system.

Retrieval technologies for the analysis of a number of patents and temporal distribution, status and trends may reflect the development of the technology. With the technology in an increasingly competitive world, countries are increasingly focusing on patent strategy research, and though the processing and combination of patent information in patent specification and Patent Gazette ,and finally make this information into the overall situation and forecasting capabilities with competitive intelligence to provide information to support strategic decision-making enterprise. With the rapid development of computer and network technology, people can retrieve patent databases through the Internet and a variety of patent-related information which has become increasingly diverse. How to take full advantage of such huge information resources so that patents play an important role in many aspects of business research and patents, patent information retrieval systems have come into being.

Full text retrieval is to search the contents of any information stored in a database in the whole book and the paper . It can be obtained according to the need full text of chapter, section, paragraph, sentence, word information, that is similar to the word for the whole book to add a label, which also can be a variety of statistics and analysis.

Full text retrieval technology, is based on data such as text, sound, image and so on as the main content, to retrieve the document content rather than the appearance. With the development of computer industry, the computer storage devices carriers more and more electronic information, the information can be roughly divided into two categories: structured data and unstructured data. Usually structured data refers to enterprise financial accounts and production data, student score data and so on, and the non structured data is some text, image data sound and other multimedia data and so on. According to statistics, unstructured data occupies more than 80% of the volume of the whole information. Full-text retrieval technology is one of the most common information retrieval applications, and also is a very efficient information retrieval technology. It is a powerful tool for dealing with unstructured data, but also the general search engine information retrieval tools. It greatly improves the data to find specific information from a vast complex data efficiently.

Lucene is a Java full-text search engine. Lucene is not a complete application, but rather a code library and API that can easily be used to add search capabilities to applications. Lucene is a high-performance, scalable full-text information retrieval tool kit, and it is not a full-text search engine, but a full-text search engine architecture, providing a complete query engine and indexing engine, part of the text analysis engine (in English and German, two kinds of Western languages).It provides a simple, but very powerful core API, people can quickly and easily integrate it into the application to increase the indexing and search capabilities. Lucene is designed to provide an easy-to-use software developer kit to facilitate the realization of full-text search function in the target system, or as a basis for establishing a full-text search engine.

In this paper, the patent search service is based on the patent textual relevance information of the actual application requirements. In depth study of the Lucene full text retrieval tool kit and related technical basis, we extend the Lucene segmentation module, and improve the ordering algorithm Lucene by default, so that the patent text retrieval system designed in this thesis retrieval can support multiple patent document format which is commonly used, and eventually patent information retrieval system used in this study above, can effectively improve the performance of patent related retrieval system.

Acknowledgments
This work has been supported by the Fundamental Research Funds for the Central Universities Nos. DL12EB01-03 and Heilongjiang Natural science fund in China Nos.F201116.

REFERENCES

[1] S. Brin, L, Page. *The anatomy of a large-scale hypertextual web search engine. Computer Networks*, vol.30, pp, 1~7, **1998**.
[2] A, Fujii., M, Iwayama., N, Kando. *Information Processing Management*, vol.43, n.5, pp, 1149~1153, **2006**. 43(5), 1149–1153. (**2007**)

[3] M, Lux., S. A, Chatzichristofis. *LIRe: Lucene image retrieval: An extensible Java CBIR library. In Proceedings of the 16th ACM international conference on multimedia*, pp. 1085–1088.**2008**
[4] Weng Xianming on the effective use of patent information. Intelligence, **1994**, 13 (6): 430-435.
[5] Lucene (2009) *Apache lucene—overview. Lucene*. http://lucene.apache.org/java/docs/index.html. 2009
[6] S, West whole, M, Ruifang. *Information Theory and Practice*, vol.29, n.1, pp, 29~31, **2006**.
[7] D. G, Lowe. *International Journal of Computer Vision*, vol.60, n.2 , pp, 91~110, **2006**.
[8] J. K, Uhlmann. *Information Processing Letters*, vol.40, n. 4, pp, 175~179, **1991.**