International Journal of Hybrid Information Technology
Vol. 9, No.9 (2016), pp. 67-76
http://dx.doi.org/10.14257/ijhit.2016.9.9.07

Mobile Internet Search Methods Research Based on Java

Guang Zheng and Yiran Wang

School of Computer Science and Technology, Zhoukou Normal University,
College of Network Engineering, Zhoukou Normal University,
Henan Zhoukou, 466001, China
zhengguang@zknu.edu.cn
286206308@qq.com

Abstract

Due to the rapid development of information technology, especially mobile network development, its information and increase the data exponentially, so in such a huge data through effective retrieval, fast, accurate and convenient access to useful data and information mobile Internet has become a serious problem; for this problem, this paper the characteristics of mobile Internet and data services, using Java to achieve rapid retrieval capabilities, and in accordance with the search function, complete the mobile Internet a variety of online services. According to the data analysis can be obtained to achieve the proposed method can effectively solve the problem of mobile Internet retrieval.

Keywords*: Search engine; Java; The Lucene index； mobile internet*

1. Introduction

With the rapid development and widespread use of technology, the Internet blogs, the Internet search indexes, e-commerce, social networking sites and other technology has brought dramatic growth in data volumes. Widespread use of computer technology in all walks of life also contributed to the generation of data, such as for the measurement and transmission of the relevant position, vibration, humidity and temperature sensor generates massive amounts of data. Astronomy, two years ago, the entire data collection has been accumulated to terabytes of information, but the Sloan Digital Sky Survey in just a few weeks collected more than terabytes of information. In recent years, annual data growth rate, doubling every two years, which indicates that we have entered the era of big data. Big Data era has brought us not only the explosive growth of data volume, complex data structures and diverse, but also the means to deal with these data information becomes complicated. Data is stored in order to obtain useful information from existing data, so data analysis and processing technology as vital. Famous business case "for diapers beer", that is, on the basis of existing data, analyze the results of beer and diapers together can increase the conclusions of both sales[1-3], the two put together after greatly promoted diapers and beer sales, Wal-Mart shocked the world business case. Also according to Google searches, posts and Twitter messages speculate that people's character, mining user habits and preferences, and then get in line with interest and used the services and products in the mass of complex data, and targeted recommendations to users the services and products, thereby increasing sales. This is where the value of the data, the data for promoting the development of various sectors of the effective power. In the era of big data, analyze data not only have an enormous impact on the economic, political and cultural aspects also have a broad and profound impact.

With the development of mobile services, mobile applications diversification, and faster data transfer, so that the mobile Internet has become an indispensable part of

ISSN: 1738-9968 IJHIT

people's lives. Traditional distributed computing, are based on the Internet on. Mobile Internet can not be calculated primarily by two restrictions, namely storage and computing capacity and data transfer speed of the mobile terminal between nodes. However, with the recent development of wireless communication technology, data transmission speed between nodes has been greatly improved; and smart phone to replace the traditional phone, the storage and computing power doubling, which for the cloud computing movement Internet applications provide hardware assurance[4-6].

However, due to the constant introduction of different mobile Internet services, its data access network is very large, and the data type and format varies depending on the business. How under conditions different data formats of different services, mobile data network search. This article is for the different data and different services, we proposed a Java-based indexing and search engine, its search engine can be adapted to different business forms of mobile Internet.

2. Related Works

2.1. Data Mining

At present, because of new technology and database technology combine to make the database in the field of new content, new applications, new technologies, forming a huge database of family. But the application of these databases are based on real-time query processing technology-based, in essence, a database query is a passive use, due to its simple queries only selectively output database contents. And therefore it is expected analysis, forecasting and decision support advanced applications is still a great distance.

New requirements to promote the birth of new technologies, data mining is the soul of deep data analysis methods. Data analysis is the basis of scientific research, many scientific studies are based on data collection and analysis on the basis of, and in the current business activities, data analysis is always highly intelligent behavior, and some special groups linked because not every an ordinary person can predict the future trends from past sales or to make the right decisions. However, as a business enterprise or industry accumulation of data, especially due to the popularity of the database, doing so much to organize and understand the data source already exists efficiency, accuracy and other problems, so investigate automated data analysis technology to provide enterprises It can bring commercial profits decision information become inevitable. In fact data, information and knowledge workers can be seen as different forms of generalized data performance. It is no exaggeration to say that people have a desire for data is greedy, especially the development of computer storage and network technologies to accelerate the people collecting the data and capacity, this greed resulted in the "data rich and information poor" phenomenon[7-9]. The database is one of the most effective way to organize and store data, but the face of the expanding data, database query technology has demonstrated its limitations, information or Intuitively said valid information refers to data to help people, for example, in the real world, if the average reading time in minutes, then a fastest only a day to browse multiple pages of a newspaper, if you subscribe to a newspaper, in fact, you read every day, but only a has been. Faced with huge amounts of data in the computer, it is also in the same embarrassment. Over the last decade, with the development of computer technology, network technology and information technology, people and production capacity greatly improved data collection, large amounts of data in different forms such as databases, documents, *etc*. is collected and used in commercial management, government office, scientific research and engineering development and other fields, and this trend will continue to grow. The rapid growth of huge amounts of data to be collected, stored in a large database and a large number, if not a powerful tool to understand them far beyond people's ability. As a result, the data collected in large databases into a "data grave", and then access them rare, it is difficult to extract

information from them again. No wonder some people a sense of historical data because the data cannot find too little, and today cannot find the data because the data is too much. Thus, a new challenge has been referred to the front of people, in what has been called the knowledge economy society, how to get rid of troubled large amounts of data, and which promptly and efficiently find useful information to help people find the hidden laws support decision making knowledge is a concept, rules, patterns and laws, *etc.*, it does not like that particular data or information, but it is people have been relentless pursuit of the goal. In fact, in our lives, people just data as the source of the formation of knowledge, we are to form and validate knowledge through the front or back of the data or information, but also continued to use the knowledge to gain new information . Thus, with the progress of data expansion and technical environment, people of advanced information processing, decision-making and analysis online more and more urgent, in the strong commercial demand driven, businesses began to notice large amounts of data efficiently solve the utilization of a significant business opportunity, scholars began to think about how to get useful information and knowledge focused on large-capacity data from the method.

2.2. Java Search Technology

JDK (Java Development Kit) called the Java Development Kit or Java development tool, it is a small program written in the Java Applet and applications development environment. JDK Java is the core of the whole, including the Java Runtime Environment (Java Runtime Environments), some Java tools and Java core class libraries[10-12] (Java API). Whatever the substance of Java application servers are built a version of the JDK. Mainstream JDK is Sun's release of JDK, in addition to the Sun, there are many companies and organizations have developed their own JDK, for example, JDK, BEA's Jrocket, there is GNU organization developed by IBM developed JDK.

In addition, the Java API class libraries in Java SE API subset of the Java virtual machine and the two parts referred to as JRE (JAVA Runtime Environment), JRE supports standard Java environment running .

JRE is a runtime environment, JDK is a development environment. Therefore, when the need to write Java programs JDK, and Java programs run when you need JRE. The JDK which already contains the JRE, so as long as installed JDK, you can edit the Java program to be properly run Java programs. However, due to running JDK contains many irrelevant content, the space occupied by the larger, so the ordinary run Java programs without installing JDK, and only need to install the JRE.

Java Web development dynamic Web technology is a combination of Servlet and JSP technology. The Servlet is the use of Java Servlet application programming interface and related classes and methods of a Java program. In addition to the Java Servlet API, the Servlet can use to extend and add API Java class package. The Servlet in enabling Java Web server or application server running on and expand the ability of the server. Run on the Servlet in the Web server and a Web server, and the Applet is put into a Web browser and executed within a Web browser[12-15]. The Java Servlet API defines a Servlet and Java enabled a standard interface between servers, which makes the Servlet has the property of across server platform. Servlet by creating a framework to extend the server capacity, to provide on the Web service request and response. When the client sends a request to the server, the server can request information sent to the Servlet, and to establish the Servlet returned from the server to the client's response. When the Web server or client requests services for the first time, can automatically load the Servlet. After loading, the Servlet to continue running until the other client requests. The JSP technology using the Java programming language class XML tags and Script lets, processing logic to encapsulate generate dynamic web pages. Page also can be accessed through the Tags and Script lets exist in the server-side resource application logic.JSP page logic and Web page design and display of separation, support reusable component-based design, make the development of a web-based application quickly and easily. in fact, at the bottom of the container JSP

page at run time will be compiled into a Servlet for processing again, finally to feedback the information to the user in the browser to see[6-8].

3. Based on the Lucene Index and Search

Lucene is a completely open source full-text retrieval tool kit.Lucene is to use Java development in the initial stage.But because of its powerful function, gradually has been translated into many languages.The Java Lucene is a high-performance full-text retrieval toolkit, it USES the index structure is inverted file.Indexing is an important step of search engines work.In this paper with the aid of Lucene kit has two of the most important concepts in the Document and Field domain logic file.They correspond to the Document in the Lucene classes and class Field.The meaning of the Document for the Document, it represents a kind of logical file.Lucene itself cannot be indexed to the physical file, but to recognize and deal with the Document type of Document.In some cases can be corresponding to each Document with a physical file, use a Document as a substitute for a physical file;And more, the Document has nothing to do with physical file, it is as a collection of data sources, to provide the original to Lucene to index text content.Lucene from Document related data source, and corresponding processing according to the configuration properties.For example, when a Document with a physical file corresponding to the up, can extract a variety of data sources, such as file name, file content, file creation time, modified time, *etc*. As shown in Figure 1, can also be extracted from different physical files to the data source, in the same Document.

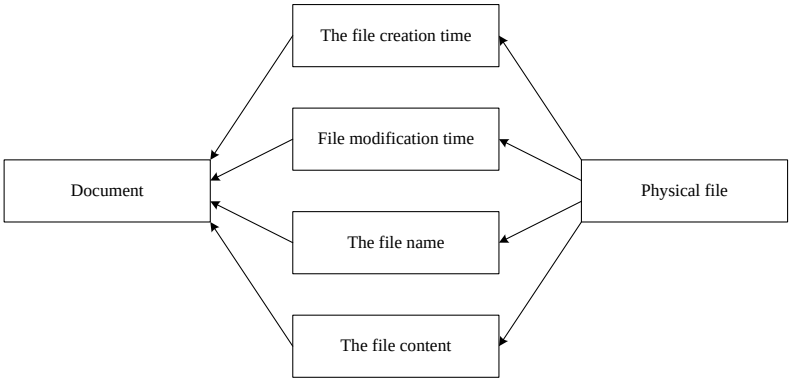

Figure 1. Document and Multiple Data Sources

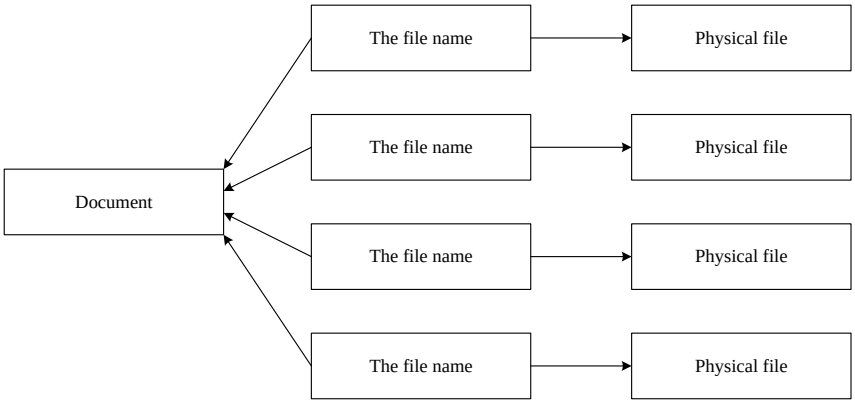

Figure 2. Document and Multiple Files of Data Sources

As shown in Figure 2 will all of the files in a directory file name in a Document, let the

Lucene index for it.Because the Document is only responsible for collecting data and, therefore, can not use physical file to build a Document, a text, Numbers, and even some links are building a Document as a data source.As long as they are added to the Document object, Lucene can build indexes for these data sources, and users can find them.

For a Document, how to represent the collection of data source?In fact, in Lucene, data is represented by the Field type of object.The Field type is mainly used to mark the various attributes of the current data sources, stored data from the data source.Lucene for each Field in the processing, can fully consider the various properties of data source, in order to make a different treatment[9-11].

Here of various properties of data source, actually refers to the following: 1, whether storage: the data source is the data stored in the index to complete.2, whether the index: the data source is the data to be retrieved when the user retrieval.3, whether the participle: whether the data source of data should pass word segmentation.In the actual development application, will meet all kinds of data sources.The importance and function of these data sources may not be the same, this is according to the specific application of developed system needs to determine the role of these sources of data in the index and storage.

Index is Lucene, one of the most important process through the IndexWriter addDocument interface, can build a good Document to join index.Shown in Figure 3, the process is as follows:

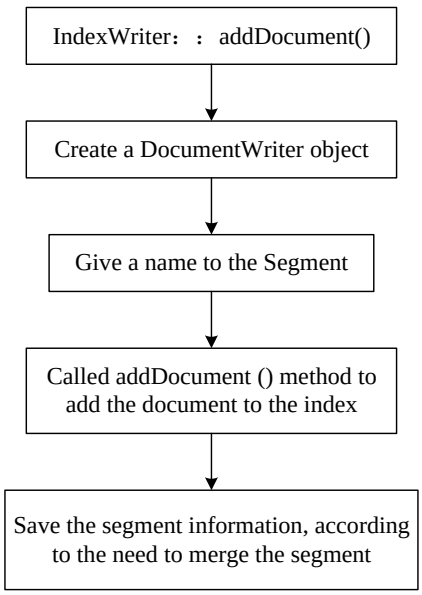

Figure 3. Indexing Process of Lucene

Information retrieval is to deal with the main object of information, and most of the time information is the form of a text.And information retrieval, the first thing is to analyze the text, in order to be able to continue the following processing.The text carries on the analysis of the basic work, is the first word segmentation, split into multiple entry will be a text.At the time of indexing, write index and can be the user retrieval is the entry.Only through the word, can let understand the user's retrieval request information retrieval system, and then search for its related content.Word segmentation tools used in indexing, and on the analysis of user's retrieval request segmentation tools should be used by the same, the reason can be seen from Figure 4.

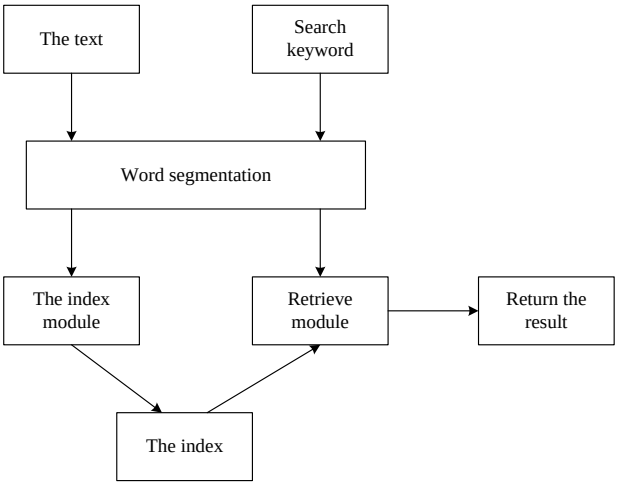

Figure 4. The Function of Analyzer

4. Vertical Search Engine Design Based on Java

With the vigorous development of communication technology, the application of mobile phone has become so common.People's work and study has been inseparable from the mobile communication.However, when people buy mobile phones tend to get to know a brand mobile phone functions and a type of its appearance.This will need to go to a special offer mobile information search in the system, can quickly learn to want to buy this phone is so palatable.Combining with the former technical basis for vertical search engine in this paper, with the help of the existing open source code resources, discusses building a mobile phone product feasibility and implementation approach of vertical search engine.

4.1. The Basic Functional Requirements

In the domestic each big mobile phone portal and the needs of customers on the basis of careful analysis, intends to build a vertical search engine function of mobile phone information query system, providing customers with a search according to all kinds of mobile information platform.Allows customers to easily get what they pay attention to details of one mobile phone.Specifically, the information service system should possess the following two important functions.

(1) automatic acquisition of mobile information

Cell phone information automatic acquisition is one of the core functions of website, main purpose is to use the current vertical search engine technology, selection, access to all kinds of mobile phone information related web pages, and put those pages downloaded to the local image stored.And then analyze the web structure, the use of structured information extraction technology, extract the structural information of cell phone information, such as mobile phone brand, model, price, origin, time to market, mobile phone shape, the size of the home screen, the home screen material, main function, *etc*., deposited in the database, for users to query and retrieve.

(2) information retrieval and display of mobile phone

Vertical search engine system is mobile phone products with the user interface, by providing a friendly query interface and query question type, on the background of mobile database query, and results of the query in the form of a list of pages returned to the user.When the user clicks on a query to the page after go in to see more detailed description, rather than get with information to find by specious, even irrelevant information.

4.2. The Overall Structure of the System

This mobile phone product vertical search engine USES Eclipse3.1 + MySQL5.0 + Tomcat5.5 development environment.Due to the Eclipse good openness, and supports a variety of import plug-in installation and toolkits.Is a very good choice platform Java language development.When building web applications interface should be installed behind the Tomcat PluginV3 plug-in.In addition, the development of this system also needs to have open source spiders package Heritrix1.12.1 and full-text retrieval kit Lucene2.0 support.

Based on the current development of search engine, and system implementation cost minimization, this paper creatively will combine the two famous open source toolkit, with its good expansibility is not only implements a simple model of vertical search engine system, and due to the expansion of open source code itself a good interface, so the system can also with strong expansibility.For the future continue to improve and improve the system function to lay the foundation.In addition, this system has good portability, as long as to master the principle, can be in very small changes developed in the field of any other vertical search engine system.For example can easily build up to the notebook computer, MP3, digital cameras and other vertical search engine system.Figure 5 is the system running flow chart:

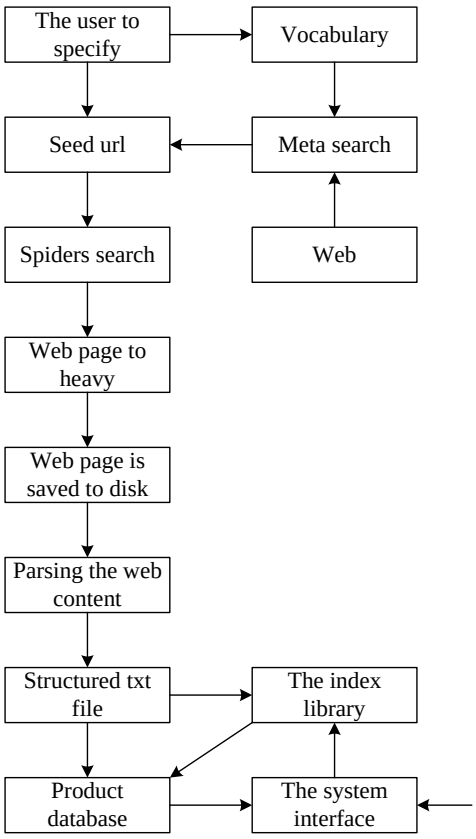

Figure 5. Flow Chart of System Running

For the scraping of the information of the mobile web part, in order to make information has the certain extent and comprehensiveness, in principle by meta search engine to obtain the best seeds site, but this article focuses on the construction of a vertical search engine, in order to easy to switch to the user specified directly seed sites to crawl for Heritrix spider program starting site.In fact, as long as we choose several domestic famous mobile website, the information on the covers almost all common cell

phone information[12-15].

4.3. The Web Information Capture

This system is to collect product information, as much as possible for user query to understand.So our program in crawl the web to choose some traffic is large, rich in content, complete product information of large sites.In that case, we simply by program will this two or three website web crawl down all products stored in the building of the vertical search in the system.In the future, as long as the update product database and index.Of course, in order to make the choice of site representative can also through the baidu, Google and other general search engines for the mainstream professional website, with the result set returned to site as we spider program starting site.To some extent, it can control the theme of fetching information relevance.But also can improve the recall performance of the system.

Because we want to construct a vertical search engine, we have to grab the information about the description of the product information.Even in a professional website and many have nothing to do with the product description page.Such as some of the product evaluation, other products, such as advertising, website help obviously is not something we want to get.If we don't filter the useless information page and eliminate spider crawling burden will increase network, affect system performance.And will be saved to the local after these web pages will also bring to our next page structured extraction work a lot of trouble.Therefore, some seeds to the site page structure and characteristics were analyzed, and the URL of the law applied to the spider program will find can be achieved to some specific grab preservation and filter out useless pages of a web page.In fact, many professional websites web structure and storage products web addresses have certain rules to follow.Of course, this step is not necessary, just in order to improve the efficiency of web scraping.

4.4. The Establishment of the Index and the Database Design

The preparation is complete, this part is relatively simple.Mainly is to write programs to read each product under the specified file path corresponding text files, each corresponding field data into the database, and then returns the actual orderdate Id to Lucene index, thus the database record Id and match them with the Lucene index.Insert data to database and indexed is synchronous.When indexing, should first define the Lucene Document format, the product all the information encapsulated into a logical Document.Index of the content should be less as far as possible, as long as can satisfy user retrieval.This article focuses on a kind of vertical search engines, so strive to specific application data of the simple and clear.System is realized by entering a product brand or product model to retrieve the relevant details of the product function, USES only a data Table to store all the product information.The database table structure of the Product are shown in Table 1 below:

Table 1. Database Table Structure Design

Field name	Data type	Meaning
Id	Int	A primary key
Category	Varchar(128)	Classification
Name	Varchar(128)	Model
Type	Varchar(128)	Detailed parameters
Content	Varchar(5000)	Product Summary
URL	Varchar(1024)	Product information index page
Imageurl	Varchar(1024)	Product pictures stored path

5. Conclusion

The mobile Internet business types and characteristics, the vertical search engine for full-text search, use Java to develop a pan-business class search engine tools, and service data retrieval according to this tool, according to the most widely used Lucene framework and search engine data compared to afford business class pan-developed search engine tool has the advantage of robust business.

Acknowledgements

This work is financially supported by the National Natural Science Fund, China (No 61103143), basic and frontier project of Science and Technology Department of Henan province, China (No 142300410334).

References

[1] C. Y. Wang and L. I. Yu-Fu, "Study of Information Filtering Technology in a Vertical Search Engine", Information Science, **(2014)**.
[2] Wang-Wei, "The Design and Implementation of Computer Vertical Search Engine", Measuring Technology and Mechatronics Automation (ICMTMA), 2015 Seventh International Conference on. IEEE, **(2015)**.
[3] Wang-Wei, "The Design and Implementation of Computer Vertical Search Engine", Measuring Technology and Mechatronics Automation (ICMTMA), 2015 Seventh International Conference on. IEEE, **(2015)**.
[4] Q. Y. Zhang, Y. U. Hui-Hui and Y. Y. Chen, "Construction of Chinese word segmentation dictionary based on agricultural vertical search engine", Guangdong Agricultural Sciences, **(2015)**.
[5] Y.J. Qian and B.X. Cao, "An Improved Ranking Algorithms Based on Vertical Search Engine", Electronic Technology, **(2015)**.
[6] R. Wei and W. U. Zhenqiang, "Design and implementation of vertical search engine for education video resources", 计算机工程与应用, **(2014)**.
[7] Q. Y. Zhou, H. U. Jing-Feng and H. E. Li-Li, "Study and Application of Vertical Search Engine Based on the Depth Mining of Enterprises", Computer Programming Skills & Maintenance, **(2014)**.
[8] H. G. Yue, L. Zhang and F. J. Meng, "Research and Implementation of a Vertical Search Engine in the Financial Domain", International Journal of u- and e- Service, Science and Technology, vol. 7, **(2014)**.
[9] S. Yue, L. I. Wanlong and L. Wang, "Database Full-Text Retrieval Based on Lucene Index", Journal of Jilin University, vol. 52, no. 05, **(2014)**, pp. 995-1000.
[10] A. B. Mathew, P. Pattnaik and S. D. Madhu Kumar, "Efficient information retrieval using Lucene, LIndex and HIndex in Hadoop", Computer Systems and Applications (AICCSA), 2014 IEEE/ACS 11th International Conference on. IEEE, **(2014)**.
[11] L. Diesendruck, R. Kooper and L. Marini, "Using Lucene to index and search the digitized 1940 US Census", Concurrency & Computation Practice & Experience, vol. 26, no. 13, **(2014)**, pp. 2167–2177.
[12] X. Shi and Z. Wang, "An Optimized Full-Text Retrieval System Based on Lucene in Oracle Database", Enterprise Systems Conference (ES), IEEE, vol. 2014, **(2014)**, pp. 61-65.
[13] X. Shi and Z. Wang, "An Optimized Full-Text Retrieval System Based on Lucene in Oracle Database", Enterprise Systems Conference (ES), 2014. IEEE, **(2014)**, pp. 61-65.
[14] X. R. Wang, Q. H. Zheng and H. E. Fa-Mei, "Research and implementation of desktop search engine based on Tika and Lucene", Computer Engineering & Design, **(2014)**.
[15] Y. Guo and Y. Lu, "Research and Application of Full-text Retrieval Technology for Document Based on Lucene", Microcomputer Applications, **(2014)**.

Authors

Guang Zheng, he received B.Eng Degree in Computer Application Technology from Henan University and M.Eng Degree in Computer Application Technology from University of Electronic Science and Technology, China in 1996 and 2009 respectively. He is currently researching on Computer network.

Yiran Wang, he received B.Eng and M.Eng Degree in Computer Science and Technology from ZhengZhou University, China in 1997 and 2005 respectively. He is currently researching on Computer network and Internet of Things.