# Research and Design of Extraction Technique of Network Education Resources

Zhang Ke[1], Yang Meng[2]

[1]Industrial and Commercial College, Hebei University, Baoding, Hebei, China

[2]North China Electric Power University Science & Technology College, Baoding, Hebei, China

[a]zhangke@163.com

**Keywords:** resource extraction, the JAVA platform, network education resources, extraction technique

**Abstract.** With the development of the Internet and information technology, it also expands constantly in the field of education at present. As for the teachers, students, parents and other education workers in the higher education, accurate extraction of network resources not only helps the users to obtain rapidly the information they need, but also makes the teachers improve their working efficiency, enables the students enrich their knowledge and makes the parents master the latest education information. The paper combines service mode of acquiring Web resources of the higher education based on analyzing the characteristics of the users demanding higher education resources, constructs flexible and high-efficiency distributed and parallel retrieval system to solve the problem of low accuracy for the users to acquire higher education resources with convenient retrieval mode and high-efficiency query process. And the paper starts from search behaviors of the users to study individualized service mode with the basis of user interest and service mode of active push, which can improve time-validity for the users to acquire information.

## Introduction

Searching engine, special website of subjects and integrated website of higher education is the mode which is often used by the users to acquire higher education resources. And searching engine is the most commonly used. Comprehensive application of searching engine is convenient for the users to search online information to some extent, but because it is for the public, emphasizes commonality and search results have a lot of clutter information which has low accuracy, it can't satisfy the needs of higher education users. Most users input one or more keywords for search query, but there are usually two kinds of information after the inquiry by searching multiple keywords, resource category information and subject description information. And search engine can't understand accurately the meaning in certain education field, which result in clutter and excessive information in retrieved results. Therefore, we should dig deep into the resources in this area, filter irrelevant information and make strict and careful classification based on professional needs of higher education users and advantages of topic-specific search engine, which can provide professional information service for higher education users, reduce the time of the users filtering information and make the users invest time and energy into reading and using the information.

## Introduction of system development platform

**JAVA technology.** The most important reason why JAVA technology is selected to develop languages and implement system is that the nature of JAVA is more applicable to the development of application technology for all kinds of enterprises. And JAVA is the preferred technology platform to development languages, especially the server side of JAVA has evident characteristics of developing applications which are embodied in the following aspects.

(1) Independence. It can be applied conveniently in all kinds of enterprises, and there is no need to rewrite the code and recompile for different platforms.

 

(2) Reusability of the code. Java is a language which could be reused and provides a reusable mechanism within the system. JAVA system provides special enterprise-version software which has more evident and better reusable structures for JAVA language.

(3) Modularity. The means of applying modularity in JAVA includes JAVA Servlet, Java Server Pages（JSP）and Enterprise JavaBeans（EJB). And applications are divided into layers with clear functions by applying 3 layers or multilayer structure.

**J2EE system platform.** The reason for choosing J2EE platform is that the platform has some advantages which are embodied in the following aspects. The above-mentioned JAVA language can obtain backwards support from J2EE system computing platform, which can make some standard development and application in J2EE system platform skip their own platform and transplant into other platforms for application. At the same time, the written code are safer and more reliable. It is convenient for the enterprises to use the platform, and the platform provides most services which are needed in calculation for enterprise and it is easy to use. Most ports such as JDBC,JNDI,and JAVA MAIL are made standard definition.

**SQL Server Database.** SQL Server 2000 has introduced some new servers and database, which not only makes the function improve, but also makes it have more practical and convenient features, that is, XML supporting and new data types including BIGINT SQL_VARIANT and TABLE. It is convenient for user-defined functions, indexing function improves greatly, which is embodied in creating index in calculating the list. And the function of full-text retrieval improves a lot, indexed views becomes more concise and beautiful. Distributed query function and the type of trigger improved greatly. Cascading referential integrity constraints. At last, Collation function improves, that is, Collation substitutes for Sort Orders and Code pages.

## Network Structure Design of System

The system uses B/S framework, so the corresponding web server is needed. The connection diagram of hardwares for the whole system is shown in Fig. 1.

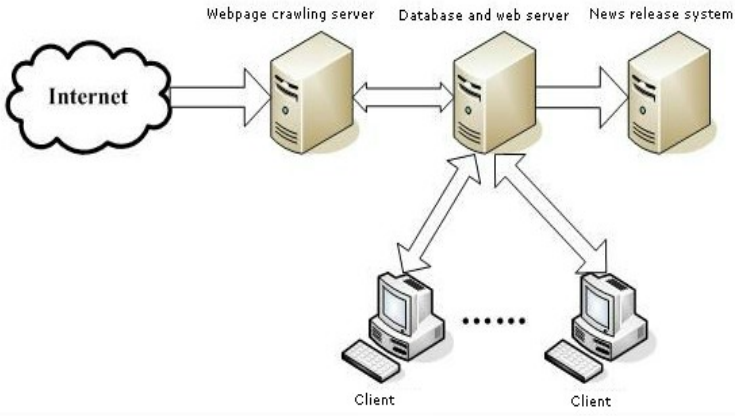


Fig. 1 Connection diagram of hardwares

As indicated above, the relative and the needed web pages are grasped from external internet and are transferred to database and web server. The related operations are carried out in the way of accessing the Web server and are imported into content delivery system. If the server system has better performance, it can combine data acquisition server and database server, and only one computer can be for all operations when te related operations are carried out.

## Establishment of Indexes

**Realization of resources searching technique.** The process of establishing indexes is to extract relevant data from database. And the record of each line automatically generates Document, then the document is written in index file, and the corresponding document is joined with the corresponding method of Lucene.

The following is some source-code:

```
Document doc = new Document();
    doc.add(Field.UnIndexed("url", url));
    doc.add(Field.Keyword("timePublish", time_publish));
    doc.add(Field.Keyword("time_capture", time_capture));
    doc.add(Field.UnStored("content", buf.toString()));
    doc.add(Field.Text("title", title));
    doc.add(Field.Text("fromWhere", fromwhere));
      writer.addDocument(doc);
```

**Searching technology.** After the index is finished, the search is carried out with Lucene. Because the indexed files only store the corresponding master keys of text, the indexed files demand that master keys should be displayed in the foreground. If the users must see detailed content by themselves, these content need to be extract from the databases.

The following is some code:

```
dir = FSDirectory.getDirectory(dirPath, false);
    IndexSearcher searcher = new IndexSearcher(dir);
Query query = QueryParser.parse(input,"content", new MIK_CAnalyzer());
    Hits hits = searcher.search(query);
    for(int i = 0; i < hits.length(); i++){
      Document d = hits.doc(i);
      String url = d.get("url");}
```

**Realization of PDF grasping.** PDF is an important the more format which is common and is used commonly in all documents. The format has better displaying effect,and the format is irrelevant with the platform and has multiple analysis which can be used by third party tools. The paper makes analysis with PDFBox of Lucene, and extracts PDFBOX-0.6.6.jar from PDF.

The following is some grasped node in PDF:

```
COSDocument cosDoc = null;
 try {
   cosDoc = parseDocument (is);
 }catch(IOException e) {
   closeCOSDocument (cosDoc);
 }
String docText = null;
 try {
   PDFTextStripper stripper = new PDFTextStripper ();
   docText = stripper.getText (new PDDocument (cosDoc));
   System.out.println (docText);
 }catch(IOException e){
   closeCOSDocument (cosDoc);
   throw new DocumentHandlerException ("Cannot parse PDF document", e);
 }
```

## System Testing

The paper applies http：//www.edu.cn/ to test. In combination with working background of the paper, three second-level domains of the main web site are tested, that is, http：//English.eol.cn/, http：// Mathematics.eol.cn/ and http：// Macro economics g.eol.cn/. These websites are input in

search background for searching the key words including Higher English, Higher Mathematics, English, test questions, teaching plan and teaching study.

In the experiment, firstly, the models of the users are carried out the matching of category according to the type of resources and subjects. And 2 resources and 3 subjects with the highest matching degree are selected. Then the categories of the lowest matching degree are compared with that of the highest matching degree. And 100 documents in the user models of each category are selected for matching.

The bold section is the related data for the documents with the lowest matching degree. From the data in the table, we can see that accuracy of carrying out the matching of documents is lower when the matching degree of category is lower, which indicates that matching category firstly has no great influence on comprehensiveness of resources. The experiment can prove that the matching algorithms and the effect of strategies are better.

## Conclusion

In theory, the paper observes which information service mode can satisfy professional needs of the users and how to realize educational characteristics. The paper incorporates theoretical research into the development on study and design of retrieval service and individualized notification service. The main wok of the paper are as follows: (1) The principles and the relevant technologies about individualized notification service is studied, and the user modeling method of BERSE in which automatic modeling gives priority is obtained. (2) Combining resources classification system of higher education to design hierarchical matching strategies of user models and document resources, and the comparison of several strategies verifies the superiority of accuracy and efficiency for hierarchical matching strategies. (3) Realizing the prototype of individualized notification service based on user model. At the same time, the paper takes http：//www.edu.cn as the sample data for testing the system. The system developed in the paper meets design requirements of the designers by verification.

## References

[1] Wang Jianhui, Wang Hongwei, Shen Zhan, Hu Yunfa. A Practical and and Efficient Text Classification Algorithm [J]. Computer Research and Development. pp. 85-93, 2005.

[2] Sun Chengjie, Guan Yi. Research on Method of Extracting Text Information from Webpage Based on Statistics [J].Journal of Chinese Information Processing. pp. 17-22, 2004.

[3] Zhao Jian, Wang Xiaolong, Guan Yi, Xu Zhiming. Conditions Random Field Method Based on Trigger Words [J]. High Technology Letters. pp. 795-801, 2006.

[4] Su Jinshu, Zhang Bofeng, Xu Ting. Research Progress in Text Classification Technology Based on Machine Learning [J]. Journal of Software. pp.1848-1859, 2006.

[5] Ding Guoliang, Wang Jiazhen. Design and Realization of Web Searching Information Retrieval System [J]. Journal of Ordnance Engineering College. pp. 58-61, 2000.