Research and Implementation of the Small-scale Search Engine Based on Lucene

Cailan Zhou, Bin Feng, Zhihao Li
College of computer science and technology
WHUT
Wuhan, China
cl_zhou2000@126.com

***Abstract*—Because of the deficiency of general search engines on LAN information retrieve, a solution of establishing a small search engine to help users obtain information through key words on LAN is proposed. After a deep study of the structure design of general search engines, based on Lucene tool kits we designed a small-scale search engine which can be functioned as gathering information, Chinese word segmentation, index and search, information update, and so on.**

Keywords- search engines; Lucene; Chinese word

I. INTRODUCTION

With the rapid development of the network, the information on the intranet of enterprises and campuses has increased significantly. it has become difficult to get the information even on the LAN. Although there have been many outstanding general search engines such as Google and Baidu, they are not good enough to solve the problem. On the one hand, the general search engines do not cover enough information, the information on the LAN is not fully collected; on the other hand, the pages searched by general search engines update quite slow, as a result, the efficiency and accuracy of information can't be guaranteed .In order to improve the efficiency of information retrieval on the LAN, in this paper a small-scale information retrieval system that can be applied on the LAN is designed.

II. SYSTEMATIC STRUCTURE

Through In-depth study of the structure of general search engines [1], we proposed the general plan of the system.

This system is mainly composed of six functional modules, that is, information collectioner, update module, the index module, duplicated-page eliminator, searcher and user interface, and Indexer includes three parts: Parser, Analyzer and Indexer.

The information collectioner uses the Spider to snatch at and download Web information. The update module is mainly responsible for periodically refreshing the downloaded information, so as to assure the consistency between the index information and the web information. Index module indexes the downloaded resources and set up an index database for conveniently storing it in the local disk. The Parser is to extract the text information. Analyzer is responsible for analyzing the texts and segmenting the words, and Indexer is to index words of every document. The searcher is to search the documents which include the querying keywords from the index database. The duplicated-page eliminator functions as eliminating the possibly duplicated pages to improve search efficiency in the index database. The user interface connects the users and the system. It accepts the query keywords which users input and brings the search result back to users.

978-0-7695-3336-0/08 $25.00 © 2008 IEEE
DOI 10.1109/CSSE.2008.1106

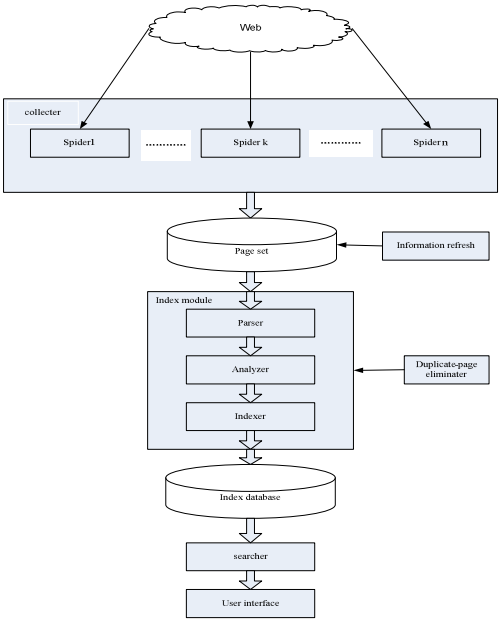

Figure 1. Systematic structure design

III. KEY TECHNOLOGY

A. *Index based on Lucene*

Lucene [2] is a high-performance, scalable Information Retrieval (IR) library. It enables to add indexing and searching capabilities to your applications and hidden the complicated implementation of indexing and searching in a term of easy-to-use API. In this system, Lucene acts as an independent layer, and the application is based on it. Lucene focus on two tasks:

(1) Index, which receives a text-only binary bytes flow, each domain includes three parameters: indexing or not, stored or not and word segment or not.

(2) Searcher, which provides plenty of search API, such as Phrase Query, Fuzzy Query, Prefix-Query, Range Query, Filtered Query, Boolean Query etc.

B. *Parser system design*

Search object includes the structured information and half-structured information such as DOC files, HTML documents, PDF files and XML documents. In order to make the system with strong extensibility and strong adaptability, the system applies a standard plug-in design style to implement the objects of different data analysis to be Plug & Play. Its architecture shown in Figure 2:

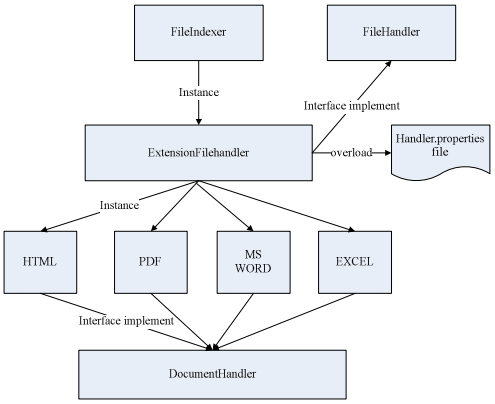

Figure 2. Document parser framework

Through implementing Document Handler class, we accomplish all of parser for corresponding format document, and use Extension File Handler to call the parser. The function of the method get Document (File) is to use a given file to infer this document's specific type and thus call the parser to parser it. The function of Handler properties is to add document resolve class to the document, and map this class to the corresponding file extension.

The process of document parser as follow:

(1) Properties instance map file extension to the implementation class which can resolve this type of file.

(2) Getting the file extension means that the resolve can find the position of the last period in the file. Thereby it can obtain the string which starts from the offset to the file end.

(3) According to the file extension and attributes, instance Document Handler class.

(4) After instanced Document Handler, transfer the file object which was packaged in fileInputStream to DocumentHandler so as to analysis it.

(5) The attributed document of the commend line should be specifically loaded in properties instance.

C. *Chinese Word segmentation*

Chinese word segmentation is an important technique for many areas especial in the Web data and search technology.

Set: C_i symbolizes a char, then this sentence can be expressed to Ci's sequence, namely $C_1C_2...C_i...C_n$. Before cutting words, we should load the cutting- word dictionary to

the memory. To improving the Algorithm's Efficiency, there we select the Singleton model, so that it can assure the Dictionary be loaded only once. Also, we specifically select Tree Map data structure.

Design concept is： Scan the sentence from left to right,when met the longest word in the Dictionary and then cut it out. If there are some char sequences can't be recognized, then cut it into single word [5]. The concrete processing flows is as follows:

(1) To make a variable word to save char sequence in the cutting-process, Init dictionary is loaded only once, set word=Null, i=0;

(2)To scan the sequence, then sentence move forward *C1C2...Ci...Cn* , put out the char *Ci;*

(3) if word's length is 0, then make *Ci* Attach to word,*i*++,return to step (2);

(4) If word's length is more then 0, then make *Ci* and the word connect, and match the char sequence with the dictionary;

(5) If matching is successful, then attach *Ci* to word, and *i*++, return to step (2);

(6) If matching is not successful , it means the char sequence in the word is the longest word that can be matched and cut ,at the same time clear up word (word=0），and keep the *i*, return to step (2)

D. Page refreshing technology

When people search the information, we always hope the search results are latest. While in fact, there are always many pages with outdate information in their search results or some emerged pages can't be searched. This happens because the pages in Web have been deleted or modified, but the page in local database does not change correspondingly in time. Update module aims at guaranteeing the contents of local pages were uncompleted accord with the Web pages. It will call the Spider to crawl the Web again and refresh the pages which were abstracted and modified on the Web.

(1) Batch refresh

For the problem of the content in an extracted page modify, crawling the Web again the setting up a new pages set to replace the old pages set is the best settlement. We named this way the Batch update. In this system, we use Quartz (Job scheduling system) to achieve periodically refreshing the pages set at certain times.

(2) Incremental refresh

For the new page emerging on the Web, we use a method called incremental refresh [3]. It will visit the websites continually when the local page set reach anticipate scope, then it discover the new pages and add them to the pages set. In this system we adopt incremental information retrieve algorithms based on authorize page to find the new pages.

Authorize page is defined as: the authorize page has many links to other pages in this website, the URL of the new pages can be found in this page. Then, we can found the new pages by indexing page. Through analysis, the norm to evaluate the authorize page are:

(1) The page is home page in website.

(2) The page URL is first or second or third rank catalogue in site.

(3) The page URL end with index, default, English subject word and html or dynamic page

(4) The page is XML document and conform RSS or Atom specification.

The authorize page is the page accord with one of the term above. The incremental information retrieves algorithms process as follows:

(1) Search the index database, abstract the authorize pages according to the authorize page norm. Then set up a authorize page list (page_list) storing the authorize pages.

(2) Obtain a page from page_list, analyze it and abstract the hyperlink in it. If we find a new hyperlink, crawl the page and add it to local page set. Then we analyze next authorize page in page_list. If not find a new hyperlink, we go to next page straightly. Repeat this procedure above until all the pages have been abstracted in page_list.

E. The design of duplicated-pages eliminate

When using the search engines, people will find that the

results of search on the inside pages are always repetitive. The results show that the approximate numbers of pages mirror of the total number of pages on the ratio are as high as 29% of all pages [4]. How to find the similar contents pages on the Web search engine quickly and accurately, it has become one of the key technologies to improve the quality of search service. In this system, we have adopted fingerprint information based on MD5 algorithm. Information fingerprint algorithm to re-process the following three major steps:

(1) Feature extraction. The search is completed and formed a search results list. According to the list from the content of each page select key words from the content of each page, and in accordance with the frequency of keywords order keywords and choose the first 10 keywords as the signature of the pages.

(2) Encoding. Use MD5 to encode Signature, and form a 32 or 64 characters FP. This is because the string is more convenient to compare, especially on comparing large-scale data pages. And fingerprint information is like the fingerprint of human being, as long as the content is slightly different, so does the same fingerprint information. *Pi* can be established that the first *i* page, the page maximum weight of the *N* key words constitute a collection *Ti = (t1, t2, ... , tn),* its corresponding weight is *Wi = (w1, w2, ... , wi), N* key words of this sorts is a string with Sort (*Ti*). Then the *FPi=MD5* (Sort (*Ti*));

(3) Similarity calculation method. For a large number of search results, it is unfeasible to directly make judgments on two arbitrary documents' information. Our approach is to firstly sort the encoded document by the value of FP and then calculate the document or the adjacent document the difference between the fingerprint on the document through the code of computing and digital logic, if the two after the operation and the location of many At the same time as one, then two documents are considered to be copied page. Logic and Computing is directly operated on bits, so the speed of operating is faster.

After adopting the algorithm above to eliminate duplicated-pages, we take three experiments in different pages date, the experimental result data as follow table.

TABLE 1. DUPLICATED-PAGES ELIMINATE EXPERIMENT DATE TABLE

Initial pages number	Process time(s)	The pages after processing	The rate of eliminating Duplicated-pages
479	7.2	417	13.6%
5246	75.4	4532	13.6%
30532	784.2	23688	22.4%

IV. CONCLUSION

The paper fully illustrates the structure of a small search engine and its development process, which focus on explaining the structure of the parser, the design for Chinese word segmentation algorithm, and updating pages and pages to re-design idea. After testing, the search engine can meet the requirements.

The design of the search engine was over, but still there are some inadequacies need to be further improved and strengthen .For instance, the problems of search engines cutting the word, how to excellently solve the identification of problems of the unknown word and unambiguous terms. And in pages de-emphasis, keywords are taken as the characteristic strings, so it causes many pages deleted because of having the same keywords, which is also needed to be further study.

V. REFERENCES

[1] Brin.s, Page L. The anatomy of a large_scale hypertextual Web search engine.Computer networks, 1998, 30 (1-7):107-117.

[2] CuttingD.The Lucene Seareh Engine:Powerful,Flexible and Free:JavaWorld. NewYork: John Wiley Sons Inc, 2000-09.

[3] J.Cho,H.Garcia-Molina.Estimating Frequency of Change.ACM Transactions on Internet Technology, 2003,3(3):265-290.

[4] J.C.Mogul, Y.M.Chan, and T.Kelly.Design, Implemention, and evaluation of duplicate transfer detection in HTTP. In NSDI, pages 43-56.

[5] Qi Wenqing, "An improved Maximum Matching Method for Chinese Word Segmentation", journal of Huangshi Institute of Technology. 2007.