The research of vertical search engine for recruitment

Linna Li [1,a,*], Zhifeng Li [1,b]

[1] College of Science, Wuhan University of Science and Technology, Wuhan 430081, China
[a]linda020329@163.com, [b] 516750653@qq.com

Keywords: Vertical search engine, recruitment, crawl, index.

Abstract: With the rapid growth of Web information, how to obtain the web page that contains the information of the user needs in an effective and accurate way has become a problem that needs urgent solution, and vertical search engine has become one of the important way in the field of network information retrieval. In this paper, a Web vertical search engine for recruitment with a three-layer architecture is proposed, in which automatic crawling, query processing and interface interaction are contained. The Web vertical search engine that could query mobile information was realized. The three-layer architecture and implementation process provide a theoretical basis and practical guidance for building a complete subject oriented Web vertical search engine.

1. Introduction

With the rapid growth of Web information, it has become a problem that needs urgent solution how to obtain the web page that contains the information that the user needs in an effective and accurate way, and vertical search engine has become one of the important way in the field of network information retrieval. At present, some domestic and foreign research institutions, universities, and companies are in the business of the vertical search engine research, and have developed a series of successful product[1]. Scirus of Eisevier is a kind of search engine which is specially designed to search highly relevant scientific information, and it is the most comprehensive scientific literature portal at present. CiteSee of NEC research Institute is a very famous retrieval system in the field of computer science, and the core is the automatic indexing which can automatically index and classify electronic documents on the Internet. Collection Building Program of the National Science Digital Library aims to create large-scale online digital library for science, mathematics, engineering and technology. FocusedProject of Berkeley take the lead in research and development by S.Charkrabarti who is an India scientist, and it guide the crawler through two procedures, which one is a classifier used to calculate the correlation degree of the theme of the book and download documents, and another is the purifier used to link the page of many related resources. The LIBClient-IRISWeb system, which was jointly developed by the computer science department and the law school of North Carolina University, can retrieve full-text legal information on the Internet using natural language, so the retrieval efficiency is greatly increased[2-4].

Graduate students who need to take part in some campus want to get information about this, so the vertical search engine website providing college recruitment information must be developed. The vertical search engine is a complex information system. However, most researches are concentrated on one detailed problem appearing in an aspect of the search engine, but they lack the correlational research on the complete implementation process of Web vertical search engines. Aiming at this problem, the implementation process of a Web vertical search engine with a three-layer architecture is proposed, in which data preparation, query processing and interface interaction are contained. An actual operation of a certain task describing the implementation process was performed with Java platform and relative open source tools. And by this operation, the Web vertical search engine that could query recruitment information was realized. The three-layer architecture and implementation process provide a theoretical basis and practical guidance for building a complete subject oriented Web vertical search engine.

2. Vertical search engine

Most of the Vertical Search Engines are lying in the phase of scientific research. The portals have appeared that faces some fields after making use of the searched result and after the professional person's processing. The research about the subject-based searching engine is getting hotter overseas. The actual Vertical Search Engine adopts two kinds of technologies list below:

One is based upon the content which is the extension of the traditional information retrieval technologies. Its main way is to establish a word list in the connection with the subject and the Search Engine. The crawler of Search Engine makes index from web according to the word list in the Search Engine. The complexity of establishing the word list is quite differently according to the different systems.

3. Design proposal

3.1 System architecture

The construction drawing describes the relationships between crawler, indexer, searcher, and user interface which are the four core parts of the search engine, and identifies the key points that each core needs to handle. The crawler not only needs to crawl the web page, but also needs to filter the garbage page. The system needs to change the propaganda link into the propaganda entity, and it needs a series of structured processing of the web text information. The indexer adds the preaching entity to the index library. The searcher queries from the index library and sorts the results accordingly. The user interface needs to understand the user's input, record user input, and display the search results to the user.

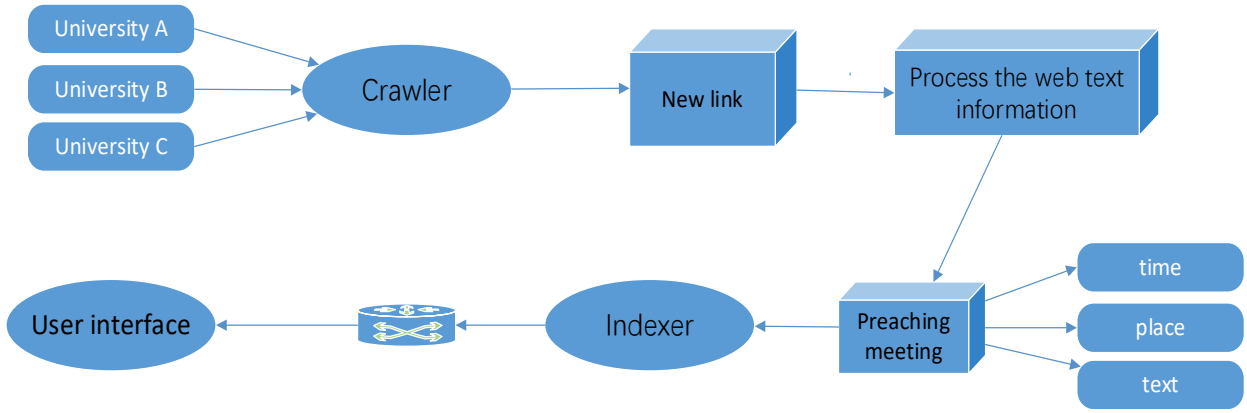

Fig. 1. The construction drawing of the vertical search engine facing recruitment

3.2 System function module

The search engine system of campus conference theme is composed of four functional modules which are crawler, text information processing, retrieval device and user interface. The following describes the main functions of the four modules. The crawler regularly accesses the employment information network of each university and links the newly released recruitment information to the database. The main problem is how to traverse the site, discover new links, and determine if the link is about recruitment. The text information processing is to extract the key nodes of the new propaganda, and then to add the thesaurus to the index library. The searcher is responsible for querying the specified information from the index library and database and sorting the results of hits. The user interface is responsible for rendering the data and providing the search box for the user.

3.3 Software development environment

The development environment is Win7 and IIS, and the software and tools used in the development process include VS2012, Mysql, HTMLParser and Lucene.net. The Internet Information Service (referred to as IIS) is the basic service of Internet provided by Microsoft Corp based on Microsoft Windows, and it is a web service component, which makes the release of information become a very easy thing on the internet. Visual Studio is very easy to use and has the strong IDE function, and the user can use it to easily build a console application, Windows services and web applications. MySQL is a very popular small relational database management system, which

is open source, free, stable and has other features, so many small and medium-sized sites choose Mysql as a web site database.

4. Key technology

4.1 Crawl web information

All The main function of the crawler is constantly finding links to the latest releases. In order to meet the different demand of the information, the general search engine must periodically scan an IP server to maintain its' database can be updated. The vertical search engine is relatively simple, because it only needs to collect a specific website. So only a number of links need to provide to the seed gripper, then new links can be captured. The following is a specific implementation of the crawler in this search engine system.

School = new School (schoolID, schoolName, schoolURL, FilterString);

The first parameter is the ID of the school, the second parameter is the name of the school, the third parameter is the link of the school announcement page, and the fourth parameter is the filter condition of the useful link. In the Handle function, we can get the content of the page the corresponding to schoolURL by calling HttpResponseMgr, and then extract all the links from the page using regular expressions and filter the uncorrelated links using FilterString. If the link is not in the system database, the grabber successfully captured a new tutorial information. The crawler will record the new link and the time when it is take.

4.2 Text information processing

The results obtained by the HttpResponseMgr is only composed of a HTML tag and a text string, and the system needs further mining of text content, so using Htmlparser we must convert the text data into point structure tree which can be traversed by computer. Briefly, Htmlparser provides a convenient and simple method for processing the HTML file. It will resolve each HTML tag into a node in a tree structure, and a type of node corresponds to a class. Then we can easily access any tag content through access interface.

5. Conclusions

The primary focus about search engine has varies from how to find more information to how to find the exact and useful information. The precision ratio becomes the first priority for many of the search engine. The Vertical Search Engine that faces recruitment seeks the thematic information appropriately to make the user search what they want effectively. It can affect the development of search engine deeply.

5. Acknowledgments

This work was financially supported by the innovation and entrepreneurship training program of Wuhan University of Science and Technology fund (Grant: 201410488041).

References

[1] Li W., Zhao Y., and Li Q. The research of vertical search engine for agriculture, Computer and computing technologies in agriculture,2009,pp.799-803.

[2] Yuangui Lei, Victoria Uren, Enrico Motta.SemSearch: A Search Engine for the Semantic Web.Berlin:Springer Berlin Heidelberg, 2006,pp.208-245.

[3] Michael Kohlhase, Ioan Sucan.A Search Engine for Mathematical Formulae. Berlin:Springer Berlin Heidelberg, 2006,pp.241-253.

[4] Albert Bifet, Carlos Castillo, Paul-alexandru Chirita An Analysis of Factors Used in Search Engine Ranking. Adversarial Information Retrieval on the Web - AIRWEB , 2005, pp.48-57.