Research and Design of Key Technology of Vertical Search Engine for Educational Resources

XIAOZHENG WANG

ABSTRACT

This paper explores several key technologies involved in the vertical search engine of educational resources, including the selection of URL seeds, the crawl of hidden web, the decision of relevance between topics and the sorting algorithm of search results. On this basis, we design a result ranking algorithm for educational vertical search engine based on neural network. The algorithm collects the user's usage and evaluation information, and then adjusts the weights, coefficients and thresholds of the neurons in the neural network according to the information. Finally, the search results are sorted, and the situation is approaching the optimal situation.

KEYWORDS
Vertical search engine, Educational resources, Neural network.

INTRODUCTION

Vertical search engines for educational resources have distinctive industry characteristics and require vertical crawlers to crawl relevant pages only. Therefore, it is necessary to improve the crawler and increase the topic correlation discriminant function. In addition, the web search engine crawlers are shielded in recent years, which result in the number of dark web pages is 100 times greater than that of the open web [1]. In the field of educational resources, the importance of "the dark network [2] is often better than Ming network, such as K12 database [3], is almost the common focus of all educational personnel. But the crawler is hard to crawl to the dark mesh by conventional means, so the system based on the general crawler must increase the crawl capability of the hidden crawler. In addition, the ranking of web pages should reflect the importance of the industry, so it is necessary to introduce web relevancy into the ranking of crawler pages.

INTRODUCTION OF RELATED TECHNOLOGIES

Vertical search profiles

Vertical search is to provide valuable information and related services for a specific domain, a particular group of people or a specific demand. The core technologies of vertical search engine include vertical network spider technology, information extraction technology, indexing technology, word segmentation technology and search technology. In this paper, an ontology based framework for educational search engine is proposed, and the automatic construction of ontology

Xiaozheng Wang, xz_wang@163.com, School of information engineering, Nanjing Xiaozhuang University, Jiangning, Nanjing, China
and search technology based on ontology are introduced.

**Introduction of Nutch**

Nutch is an open source search engine based on Lucene, and it is a perfect application, which implements the integration of crawl, index and retrieval. Since commercial search engines allow competitive ranking, this results in index results that are not entirely relevant to the site content. The Nutch search results can give a fair ranking result, which makes Nutch [4] a good choice for vertical search, academic search and search for government class sites.

**KEY TECHNOLOGIES OF VERTICAL SEARCH ENGINES**

**The establishment of the URL seed**

The quality of URL seeds has a direct impact on the accuracy and coverage of search engines. The screening of the initial URL mainly considers the relevance and authority of the new education topic, and also needs to consider the rationality of the website organization and layout, the quantity, quality and uniqueness of the web page information. Due to the limited number of seeds, the education discussion group with authority of educational institutions has been gathered from previous organization, and (news group) extraction and education and training company website, and then identified by the educational staff and educational informatics personnel.

**Dark net crawling**

The mechanism of this system is to grab the dark net [5] simulate the behavior of the program, fill in the corresponding query field site to the site, the query submitted the generated form, the final results obtained in the query results page temporary table, and regular injection system every night.

There are 5 key problems in the crawl of the hidden network: the discovery of the entrance of the dark net and the entrance of the dark net, the web site of the important database; Generation of the template with information rich, the template is a combination of database query fields on the page, which is rich in information refers to the template has the following characteristics: fill in the template query field in the template with the corresponding keywords, to generate a series of forms, delivery website retrieval results returned after significant differences between pages; The establishment of the initial keyword table. The keyword here is to fill in the numeric value and text of the retrieval field; The follow-up maintenance of table; New additions to the dark portal and their initial keywords

Because previous work has accumulated a large number of educational resources information collection experience and information, have been familiar with educational resources with high correlation with authoritative dark web site, and the dark net entrance a list of URLs and the initial keyword table of related websites are directly identified. After the system is put into operation, the new dark network entrance and the initial keyword list of the corresponding website are input manually through the function menu, and the new important dark net is added in half a year.
In the above 5 key questions, the information rich template generation and maintenance of the website keyword table's following are realized automatically by the program. The program flow is shown in Fig. 1:

**Topic relevance determination**

The topic relevance judgment mainly includes two parts: topic dictionary and decision algorithm. A method for determining the initial theme dictionary is as follows: firstly, using crawler to crawl URL seed web pages, and then call the parser to parse the web page to produce a set of words, and finally, the educators and educational informatics staff collaboratively screen and assign weights. Here, the weight data is divided into 1, 2 and 3 levels, which represent general professional vocabulary, two commonly used professional vocabulary and one class of commonly used professional vocabulary. After the system is put into operation, the maintenance of the subject dictionary is realized by the function menu, and the period is in March. When a topic filter filters a web page, a list of words generated by the parser is added to the candidate vocabulary for pages that are greater than the threshold. After a full cycle, these words are removed duplication and added to the system's subject lexicon, which is still assigned by professionals to distribute weights, and then empties the candidate thesaurus for backup. The main basis of the topic relevance judgment is the position, frequency and weight of the key words in the web page, and the judgment method is implemented by the most commonly used space vector method.
Search results sorting algorithm

The Beijing Baidu Netcom Science Technology Co., Ltd invented a method of sorting search results, and the system is an algorithm for sorting search results based on user satisfaction results.

The algorithm flow is shown in Fig. 2:

The Baidu search results sorting algorithm with its emphasis on the user to search and different interest, different demand results from the single to the development of web search results closer to the user. Many special search results are developed, such as search results can include in the last few days of weather information and so on. As for the type of search results, the ranking of search results should be adjusted accordingly. This algorithm is applied to this kind of demand.

This paper presents an algorithm for ranking results of educational vertical search engines based on neural networks. The algorithm is composed of 5 parts: query condition, education key word filter, satisfaction calculation, feature extraction, neural network training, sorting and querying. The algorithm flow is shown in Fig. 3.
The working procedure of the algorithm is as follows:

1) Users input query terms in educational vertical search engines.
2) According to the key word library of education, the collection of educational key words is extracted from the query condition.
3) Start the search engine according to the education key collection and get the result URL collection.
4) The extracted keyword sets are sorted according to their order of arrangement (or syntax order) in query conditions, which form a set of input vectors.
5) According to the results of URL concentration in Link DB URL satisfaction to calculate the value of (sum or average). If the satisfaction value is greater than the threshold value, the URL set is sorted directly according to the satisfaction value and output to the user for evaluation, and then tenth steps are carried out; otherwise, the sixth step is executed.
6) The input vectors are used as input to the training neural network which obtained from the fourth step correspond to the input layer neurons of the sorting neural network.
7) According to the URL set obtained from the query results, the weights, coefficients and threshold fields stored in the Link DB in URL are removed sequentially, and the intermediate layer neurons are built to form the middle layer of the neural network.
8) The output of each intermediate neuron is calculated by the neural network, which corresponds to the user satisfaction value of URL, and writes the satisfaction value to the corresponding URL field of Link DB (used for presorting).
9) The URLs in the URL set are sorted according to the user satisfaction vector of the neural network output; thus, the sorting result is obtained and output to the user for evaluating use.
10) Users use and evaluate the results of this sort.
11) Extracting characteristic indicators from user usage results.
13) By comparing the characteristic index and ideal feature index extracted from the user's results, the ideal sorting result is introduced.

14) Compare the results of each URL and the ideal sort of the actual sorting result to the URL. If the location is the same, it will not be adjusted. If different, then adjust the weights, coefficients and threshold of the corresponding URL neuron. The specific implementation is as follows: If the actual URL is ranked relatively well in the ideal sorting result, and then increases the weight coefficient of the corresponding URL neurons and decreases the threshold in the ideal ranking results; else if, and then decreases the weight coefficient of the corresponding URL neurons and increases the threshold in the ideal ranking results.

15) The weights, coefficients, and thresholds of each URL in the sorting result are adjusted sequentially until all the adjustments are completed.

16) Wait for the user to re-enter the query condition and repeat the 1~13 step.

17) The network training will be stopped when the number of URL adjusted is less than $k$ and the absolute value of the adjusted value of the neuron adjustment and the threshold value is less than $\xi$ by comparing the results after enter queries two times.

CONCLUSION

This paper explores several key technologies involved in the vertical search engine of educational resources, and then designs a result ranking algorithm for educational vertical search engine based on neural network. The algorithm collects the user's usage and evaluation information, and then adjusts the weights, coefficients and thresholds of the neurons in the neural network according to the information. Finally, the search results are sorted, and the situation is approaching the optimal situation.

ACKNOWLEDGEMENTS

This work was financially supported by Key Laboratory of Trusted Cloud Computing and Big Data Analysis (Nanjing Xiao Zhuang University), Jiangsu Engineering Research Center for Networking of Elementary Education Resources(BM2013123).

REFERENCES