The Desgin and Implementationof Vertical Search Engine Based on Nutch

Hui Wang, Jianzhuo Yan, Liying Fang, Xinqing Shi, HairuiLuo
Department of electronic information and control engineering,Beijing University of Technology, Beijing, 100124, China

Abstract. For the convenience that users can query information of traditional Chinese medicine in the network of Marine information, the vertical search engine system based on Nutch is designed. Joined Chinese word segmentation to analyze the web page, added a PageRank algorithm to judge the authority value of the links, achieved a topic relevance algorithm based on VSM to filter the web pages. The results show that the system is feasible.

Keywords: Nutch, Vertical search engine, PageRank, Topic relevance

1. Introduction

Internet has become an extremely handy tool for releasing,disseminating and obtaining information. In the vast network of information, search engine technology has played the role of information navigation. But with the explosive growth of network data, general search engines appear the problems ,such as low precision, Less useful information content and so on.At this time,the vertical search engines that dedicate to searching for a particular discipline or field emerged.It can improve the accuracy and precision of information search. Based on these advantages,vertical search has grown in strength and is running in user-friendly design to seize the general search engine market share[1,2].

Currently, the websites of traditional Chinese medicine (TCM)have sprung up, increasing the amount of information in traditional Chinese medicine. The traditional Chinese medicine is closely related to people's life and it has profound cultural connotation. It tells us that the cure method is varied and in many respects, it can make up for the deficiency of western medicine[]. But the search engine about traditional Chinese medicine is less, the search resource that we can use is very limited.

So this article is going to build a vertical search engine in the field of TCM.

[1]Hui Wang (✉)
Beijing university of technology,Chaoyang district, pingleyuan 100, 100124 Beijing, China
e-mail: wanghuicxx@sina.com

2. The Open SourceProjects of Search Engines

As all kinds of search engine develop popularly, open source project played a role in promoting[3].The main open source project has the following kinds:

1) lucene

Lucene is a full-text retrieval tool kit, providing the index interfaces and query interface. So it is applicable to any full text indexing and search applications. It initially developed by senior experts of full-text retrieval –Dougcutting. In September 2001 it joined the Apache software foundation of open source Java products, and in February 2005 it become the top Apache projects. Lucene system is stable, powerful, classic and widely used.

2) Nutch

Nutch is a newly born complete open source of search engine system. It can be combined with database indexing and can quickly build the required system. Nutch is based on Lucene, Lucene provides Nutch text indexing and search API, so it use Lucene as indexing and retrieval module.

3) Compass

Compass is an open source search engine framework,implemented on Lucene, to provide more concise search engine API. It increase the support of index transaction processing, and can more easily integrate with transaction processing applications such as database integration. Compass update simpler, more efficient, do not need to delete the source document.

4) Larbin

Larbin is an open source web spider, the purpose is to track the URL of the page to extend fetching, finally provides the data source for search engines. Larbin scraping of the page, don't deal with the search engine's other parsing, indexing, retrieval work.

5) Heritrix

Heritrix is a Java open source project and a product on SourceForge. This is different from Nutch open source project. Heritrix is the same as the Nutch from principle and structure, according to the given URL to submit HTTP requests, fetching, timely complete the site content.

3. The Architecture of Nutch

In summary, the Nutch for search engine characteristics is summarized as follows: high transparency, Nutch belongs to the open source program, so any unit or individual can view of the distributed search engine working principle and working process; Scalability, Nutch's application is setted flexible,It can be customized according to user's requirements; High stability, through long time practical application, Nutch's results show that the operation is very stable[4]. So this article has selected Nutch search engine framework.

As a search engine, Nutch includes the basic components of web crawler, indexer and retrieving device. Nutch web crawler is so powerful that it store the download web pages according to certain format,being convenient search index. The function of the Nutch crawler is mainly implemented by the command of crawl. Crawl command is composed by the underlying command Inject, Generate, Fetch and Updatedb. Nutch web crawler is a significant characteristic to distinguish lucene.

Nutch's indexer is based on Lucene, and implements the interface of the Lucene for further encapsulation.The index is used for index'sestablishment , organization, maintenance and management. Nutch's searching device is also on the basis of building in Lucene, and encapsulate Lucene's query API. Nutch is supported by four main data structures to provide data.They are Web-DB, LinkDB, Segments, and the Index.

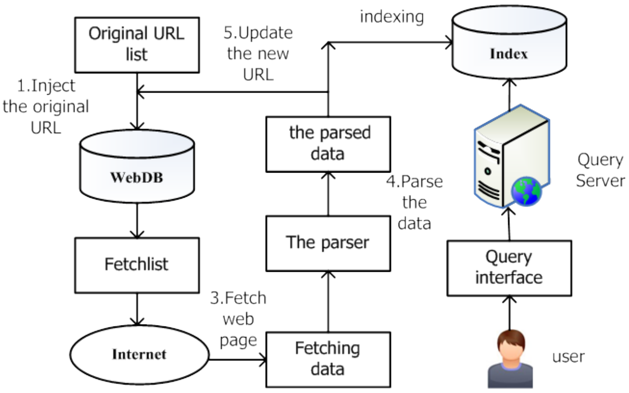

Fig. 1.3 The Workflow of Nutch

The whole working process can be divided into the following steps:

1) Establish the initial URL set.
2) Inject the initial URL set into the database of crawldb, the entire web scraping process will start from these seeds URL, gradually extend to the Internet, or stop until the user specified fetching layer.(inject)
3) Create grab list according to the crawldb database.(generate)
4) Execute crawling and obtain web information.(fetch)
5) Update the database, the web pages which be grabbed down contain a large number of links to other pages, update them to the database. (updatedb)
6) Repeat the steps 3-5, until the preset depth. This cycle process is called “produce/scraping/update” cycle. According to the content of the segments update LinkDB database. (invertlinks)
7) Indexing, generate an index for each Segment. (index)
8) Delete redundant web pages and urls from these indexes.
9) Merged all these small index into one large index, prepare for being searched.
10) User execute query operation through a user interface.
11) Change the user query into Lucene query.
12) Return the result.

4. TheConstruction of AVertical Search Engine

This article build a vertical search engine system based on Nutch. The main research is that we extend some of its function and improve the algorithm on the basis of each functional module of Nutch. As shown in figure 2, the key to improve part is as follows:

1）Nutch is developed based on the English, English is the word for the unit,words and words are separated by spaces. Chinese’s unit is one word, all the words in the sentences together describe a meaning. So the Nutch can only be carried out on the English word segmentation, does not have the function of Chinese word segmentation. Nutch parsing module is used to analyze the content of the pages, so we need to add Chinese word segmentation function in this section.

2）The biggest difference between vertical search engine and general search engine is the topic relevance judgment function. So we need to add the theme filter in nutch. Web page parsing module divide the pages grabbed down into two parts, web links and web content. Pagerank algorithm is adopted in this paper to calculate the priority of web links.The topic correlation discriminant method was carried out on the web page content topic correlation calculation. The priority list of web links are updated to the grab list to be grabbed.The web content whose topic relevance is above a certain threshold is deliveredtonutch indexer to index, for users to search.

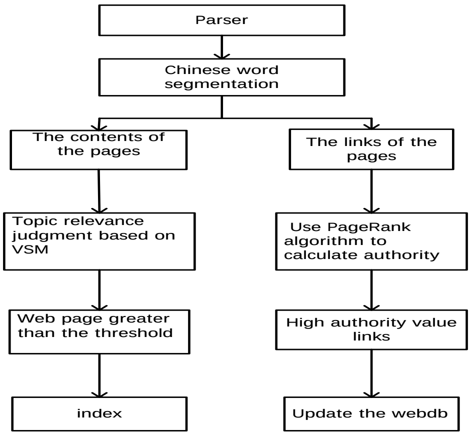
Parser
Chinese word segmentation
The contents of the pages
The links of the pages
Topic relevance judgment based on VSM
Use PageRank algorithm to calculate authority
Web page greater than the threshold
High authority value links
index
Update the webdb

Fig. 1.4 The ImprovementPart of nutch

4.1 Add JE Chinese Word Segmentation into Nutch

Whether the segmentation is accurate or not determines the accuracy of text extraction and the relevance of search results ranking. So Chinese word segmentation technology is a very important part in Chinese vertical search engine.In the Nutch we can use two ways to increase support for Chinese, one is that we directly modify the Nutch system code, modify the default of participle code and make it using a custom in Chinese word segmentation procedure. Another is that we adopt the way of the plug-in. In system design, with the method of the plug-in, implement multi-language support functions on the basis thatsystem code don't be changed. In this paper, we use JE Chinese word segmentation. JE participle is the most widely adopted Chinese word segmentation technology, using positive maximum matching points in the form of word processing. At the same time JE in Chinese word segmentation is Lucene Chinese phrase, the word segmentation result is more conform to people's reading habits and word segmentation method[5].

Nutch's word segmentation is in the package of src.java.org.apache.nutch.analysis. Because Nutch is based on lucene, so its word segmentation is also inherited the lucene. Word segmentation module of Lucene abstract an abstract class--Analyzer.Java. This class contains the abstract methods --TokenStream. Java which is used to analyze the text. NutchAnalyzer inherit the Analyzer class,has implemented Configurable, Plyggable interface. it is the extension point to extend the analysis text in Nutch.

To introduce JE participle,what we have to do is just need to use theJE word segmentation method to replace the Nutch originally word segmentation method and return < JE participle. .tokenStream (fieldName,reader) >. Use tools: javacc, ant, JE participle package[6].

1) Modify the word segmentation tool in index

Make the following changes for tokenStream method:

```
    public TokenStream tokenStream(String fieldName, Reader reader) {
Analyzer analyzer;
analyzer= new MMAnalyzer();
return analyzer.tokenStream(fieldName, reader);
}
```

2) Modify the analysis part in query

For search engine, at the time of information capture and query we should use the same word segmentation method, so that we can achieve the good cutting effect.In order to use the same Chinese plug-in when the user query,we also must change more NutchAnalysis.jj files. Because at the time of producing the query, System uses the parse method of NutchAnalysis to produce query after the call to the Chinese plug-in. But this method produce Chinese vocabulary in default. So we will change < SIGRAM: < CJK >to: < SIGRAM: (< CJK >) + > in the SRC file: \ Java \ org \ apache \ nutch \ analysis \ the NutchAnalysis jj .

3) Use the ant tool to recompile the build. xml project files.

4) Replace the file restored

5) crawland index, restart the tomcat

we use the indexing tool of Luke to view index results . Luke is the graphical tool to query Nutch's index file.It can be intuitively seethat index have stored by phrase as a unit after joining in Chinese word segmentation.

4.2 Add topic Filter to Nutch

Nutch is developed for general search engine and dosen't have the topic relevance judgement in the scraping of the pages. As the vertical search engines is implemented with Nutch, we need to join the topic filter.

4.2.1 The Implementation of the PageRank Algorithm

With the develpopment of network information, there have produced many spam pages. The authority of the pages is low, even these related to the topic is not desirable. So this article will add pagerank algorithm in the theme filter to filter the website whose quality is not high and authority is low. PageRank algorithm is put forward by Lawrence Page and Sergey Brin. It gives each web Page a measure of the importance of the authority of the value. The basic idea is that these referenced by a large number of high quality web page are also a high quality web pages[7].

The prototype of the Pagerangk computation formula is as follows:

$$\mathrm{PR}(\mathrm{u}) = (1-\mathrm{d}) + \mathrm{d}\sum_{i=1}^{n} \frac{PR(Ti)}{C(Ti)} \quad (1)$$

$PR(u)$ isthe PageRank value of the web page u, namely the importance of the web page. $PR(Ti)$ is the PageRank value of the web page who links to u. $C(Ti)$ is the number of chains out for Ti. d is the damping coefficient, 0 < d < 1, usually value of 0.85.

The article consider that the same site design a lot of internal links due to the site navigation purposes. The links have subjective service color of this site and can't reflect the value of the contents of this page objectively. And website design many meaningless links in order to improve search rankings. In view of this, the internal links and external links will give different weights to calculate respectively put forward in this paper. the PageRank improved formula is as follows,

$$PR(u) = (1-d) + d[a\sum_{i\in V1} \frac{PR(i)}{C(\mathrm{i})} + (1-a)\sum_{j\in V2} \frac{PR(j)}{C(j)}] \quad (2)$$

Among them, V_1 is the web page who chain in u in the deffernet site with u. $C(i)$

is the number of chain out web pages i. V_2 is the web page who chain in u in the same site with u. Obviously, The right of the links outside the station is bigger than the links in the station.a is the proportion factor controlingexternal links and internal links, and 0 < a < 1 .a has an value greater than 0.5 usually.Therefore, in this article a=0.7.

We will give a larger value of $PR(u)$ to the update command of Nutch crawler to update grab list in the webdb.

4.2.2 The Implementation of Topic Relevance Algorithm

The Internet is an intricate information net, Nutch crawler spread outside from the initial list of urls to crawl a certain depth, It could reach some pages deviating from the theme. So after scraping of the pages and before indexing we need to analyze the content of the page to judge whether or not pertinent to the topic areas. Currently, according to the text content correlation discriminant classic methods mainly include the following: full text scan, Boolean model, probability model, vector space model. we choose the space vector model here.The model has good effect and broad application. The unit of vector space model is web page.It choose several subject keywords that can express the theme of the web page as the feature. The document is expressed in the feature vector, $D = D(t_1, t_2, t_3 ... t_n)$. t_n is a Keywords in web page, each t_n said a dimension, a web page can be represented as n dimension. In a web page, the important of each feature is discriminating. Here we introduced the weight of each feature w_{ik} . Generally the keyword frequency algorithm TF - IDF is used for calculating weight . It's main idea is that feature weighting is decided by the feature frequency in the web and the page number who contain the feature[8] . But since the web has the structural characteristics, a feature in the same page but defferent position has different weights . This article concluded the calculation formula of the feature item frequency in a web page is as follows:

$$\mathrm{Tf}=\mathrm{mTf}_{title}+nTf_{meta}+pTf_{anchor}+qTf_{nomal}+kTf_{EM} \quad (3)$$

The Tf said the frequency of feature keywords. Tf_{title} 、 Tf_{meta} 、 Tf_{anchor} 、 Tf_{nomal} 、

Tf_{EM} in turn, said the title of the page, page META information, web page hyperlinked text, normal text and keywords frequency in the emphasis part of page. m, n, p , q and k respectively said their weighting coefficient. Feature weighting can be expressed as:

$$w_{ik}=tf_{ik}\log(N/n_i+0.01) \quad (4)$$

Tf is the frequency of keywords calculated by theabove formula . N is total number of web pages. n_i is the number of pages containing the keywords .

Topic sample pages said by t,Pageswho will be evaluated said by A, the correlation between them can be expressed in the inner product of both keywords weighting .

$$sim(A,T)=\frac{T\times A}{|T||A|}=\frac{\sum_{i=1}^{n}w_{iT}w_{iA}}{\sqrt{(\sum_{i=1}^{n}w_{iT}{}^2)(\sum_{i=1}^{n}w_{iA}{}^2)}} \quad (5)$$

At this point, the bigger inner product ,the higher correlation. We set a threshold value, and when it is greater than the threshold ,we will pass it to nutch index to index. when less than the threshold,it will be abandoned . Usually threshold is set as 0.85.

5. Test and Analysis

This article has builded a vertical search engine in the field ofTCM.Through artificial selection, we select the following websites of traditional Chinese medicine as the initial web crawling seeds.

http://www.zhzyw.org/
http://www.teseyao.com/
http://www.cn939.com/
http://cm.39.net/

http://www.99.com.cn/

We create a text files named urls. txt under the Nutch folder , and put those sites in the text. After all preparations is appropriate, let's run the command Crawl of Nutch to scrap of the page.The parameters are crawl urls -dir crawled -depth 5 -topN 100 -threads 5. After operation is completed, the index has been established.We can retrieve in the retrieval interfaces.

Recall ration and precision ration are put forward by American scholar j. w. Perry in 1955. The recall ratio and precision ratio is an important index to evaluate algorithm and system performance. Recall is the ratio of relevant web page number that have been retrieved and all relevant page number. Precision is the ratio of related web pages number and the total pages number that have been retrieved.

As a result of the recall is difficult to calculate, here we use precision to evaluate the performance of the vertical search engine[9].

We let the system collect 5000 pages and we randomly select 1000 pages.The 1000 pages contains higher authority page and advertising garbage, but also contains topics pages and the pages that have nothing to do with the theme. Separately we do the artificial evaluation about authority value and topic relevance on the result of system evaluation. If we have higher requirements on authority values and topic relevance degree, we can adjust the threshold to get more accurate results.

Table 1.5.1 TheResult of PageRank

System evaluation results		Artificial evaluation results		Authoritative evaluation accuracy
evaluation results	The Number of the pages	Authority Pages	Advertising garbage	
Authority Pages	891	803	88	90.12%
Advertising garbage	119	25	94	78.99%

Table 1.5.2 The Result of Topic Relevance

System evaluation results		Artificial evaluation results		Authoritative evaluation accuracy
Evaluation results	The Number of the pages	Topical Relevance	Unrelated topic	
Topical	867	728	139	83.97%

Relevance				
Unrelated topic	133	41	92	69.17%

Query item "zhenjiu" in Chinese search engines and Google search engine . Respectively do the artificial statistics for 200 results before .Analysis the number of pages about higher topic correlation and higher authority in the query result shown in the following table.

Table 1.5.3 Compared with Google

Search engine	Higher topic correlation and higher authority	Precision ration
The system	165	82.5%
Google	90	45.0%

From the table , we can see the TCM search engine on the recall ratio and precision ratio had the very big enhancement.

6. Conclusion

This article obtain the high precision of vertical search engines in TCM through joining JE Chinese word segmentation, the improved PageRank algorithm and topic relevance judgment algorithm based on VSM in the search engine framework Nutch. It makes the information positioning and search more accurate, reduces the interference of the irrelevant information, and improves the system's power of processing the complicated Internet environment.

Refenrence

1. Chen, X.: Analysis of vertical search engine. Modern Information 9, 133–134 (2004)
2. XiaoYanXu: Research on the Development of Vertical Search Engines.Advances in Intelligent and Soft Computing, 579-584 (2012)
3. Doug Cutting.Nutch:Open-Source Web Search Software. November .26th 2004 University of Pisa
4. Tao Hu, HongyingLu:The study of search engine based on Nutch[J]. The computerage,(01): 57-59（2005）
5. FeiGao, Yun Liu.: Achieve Chinese word segmentation for Nutch[J]. Network Security Technology and Application (09): 71-72（2008）
6. http://wiki.apache.org/nutch/

7.Page L,BrinS,MotwaniR,et al.: The PageRank citation ranking: Bringing order to the Web [EB/OL]. http://www-db. stanford. edu/一backrub/pageranksub.

8.Hao,Hong-Wei:An improved topic relevance algorithm for focused crawling. Conference Proceedings- IEEE International Conference on Systems.Man and Cybernetics,p 850-855 （2011）

9.Si, M.: Construction and Research of school website based on vertical search engine. Technology Online 21(01), 108–109 (2011)