

Design and Implementation of Agricultural Information Resources Vertical Search Engine Based on Nutch

Enjun Ding

Southwest University Library; Chongqing 400715, China
243148793@qq.com

To solve the problems in traditional search engine like large returning number, bad professionalism and low precision, we analyse the working principles of Nutch open source search engine. FMM algorithm is adopted based on word library, and key-word based vector space model is adopted to make topic relevance judgment. Then the key word is extended base on ontology to improve the sorting results. So Nutch-based agricultural vertical search engine is designed, with implementation of agricultural domain ontology on the information acquisition and filtering, retrieval and similar terms recommending of search engine. The experimental results show that our agricultural search engine can improve the precision of user retrieval and satisfies the professional demand of user.

1. Introduction

With the continuous development of China's agricultural information collection, analysis, it has invested large amounts of funds for agricultural knowledge resources in processing and application (Li et al., 2009). The agricultural information resources amount reaches TB level. In the face of such a huge data resources, many people including farmers, agricultural researchers, information service, need quickly and effectively to find suitable for their individual knowledge of agriculture resources and information. Therefore, agricultural professional search engine becomes an effective method to solve these problems.

For the agriculture search engine, sorting according to the contents should be the first factor. For many popular commercial search engines, the query results are strong commercial and professional relationship with the professional and the search result is weak; while other agricultural search engine on the filtering and post sorting related degree is also needed to be improved. Nutch, as an open source project of the Apache, with basic functions of search engine, it adopts the value of the web page itself to sort algorithm (Arafat et al., 2008; Sun et al., 2008). In addition, Nutch is flexible and powerful plugin system. Therefore, we can use Nutch to build agricultural search engine fast and conveniently. This article introduces the design of agricultural search engine system based on Nutch in web environment through agricultural knowledge grid projects which are building the agricultural search engine.

2. Related technology of nutch

Nutch is a Java open source Web search engine provided by Apache software foundation. Its object is to make everyone to spend less time to develop the necessary Web search engine with convenience (Gao et al., 2012). Therefore, it provides all the tools required for the development of search engine, and the developers are just based on secondary development. We can quickly develop its own search engine with Nutch to achieve the minimum cost to provide high quality services. At the same time it is very flexible and it provides a plug-in that can be easily increased with information processing and specific document format, such as the resolution function. Nutch has two core parts: Crawler and Searcher (Tarantino et al., 2013). The core of Crawler is crawling the web pages from the Internet to the establishment of the index database. The searcher core is based on the user's search keywords to search the index database to generate the retrieval results. The structure of Nutch system is shown as figure 1. From Figure 1, we can see the entire workflow from the crawl to the web page to establish the index.

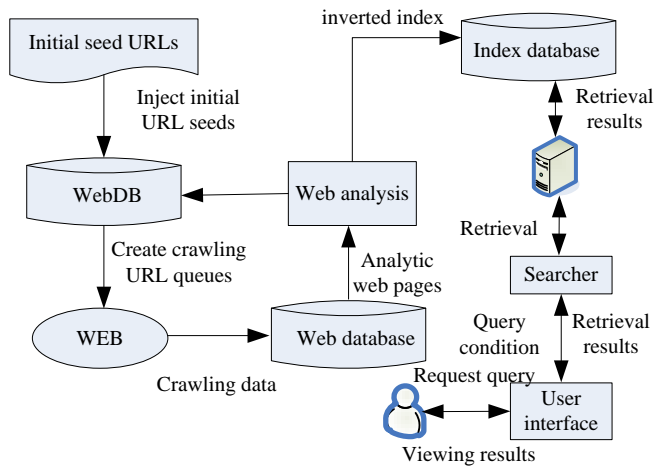


Figure 1: System Structure of Nutch

2.1 System framework

Nutch-based vertical search engine for agriculture serves ordinary Internet agricultural users and professional agricultural users, enabling these service objects and to find their own agricultural information effectively. Our study is based on the research of Nutch Framework of open-source search engine, with reference to Onto seek and Google ranking algorithm (Wang et al., 2013; Zhou et al.,2011) introducing the technologies of ontology, PageRank etc. The realization of information acquisition and filtering, results sorting, information retrieval related functions are also performed. The architecture of agricultural vertical search engine based on Nutch is depicted as figure

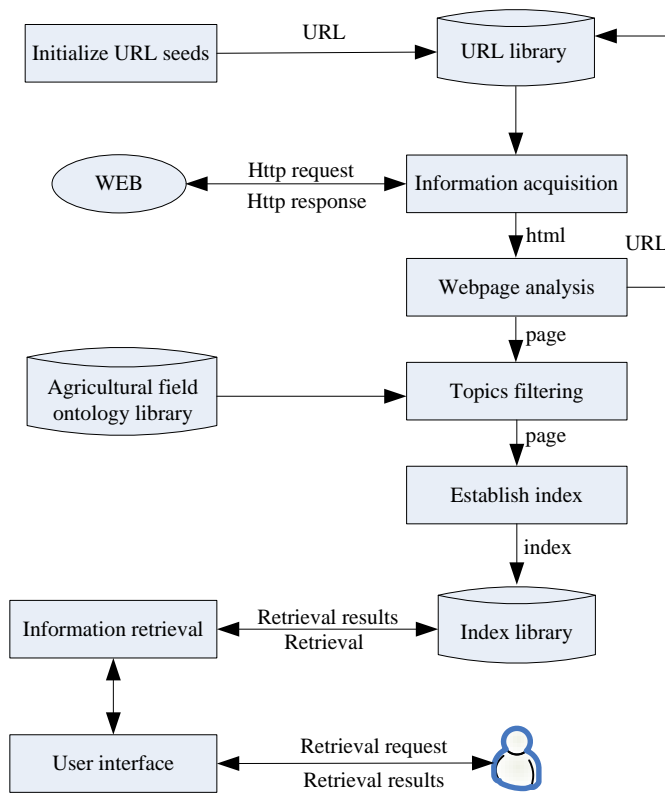


Figure 2: Nutch-based Agricultural Vertical Search Engine Structure

From the logical analysis, the construction of agricultural vertical search engine based on Nutch is divided into 5 stages: the initial URL seed, information collection, information filtering, establishment of index, and information retrieval. The whole process can be described as follows:

The initial URL seed is obtained by the method of artificial sorting and meta search;

According to the initial URL list, use the network spider crawling technology to crawl from the Internet to take the page;

Analyze the pages acquired by crawling, extract links existed in pages to join them to the URL library. and to filter the interference or invalid information with the combination of ontology of the relevance theme of the discriminant method based on vector space model;

Repeat 2 to 3 steps until reach the set number of fetching layers;

Establish the index of the web page which is saved by the filter;

Users make search through the user interface;

Return search results.

2.2 Key Modules Design

2.2.1 Webpages Crawling

Web crawling adopts breadth first algorithm to crawl the web page information (Hou et al., 2014). It uses the number of crawling to limit the number of web pages that can be crawled as much as possible on the web page information. The network crawler crawls on the Internet according to the initial set, and goes on along the access link to the page address, continuing to visit other pages. The Web crawler will crawl the visited web pages to local. The sketch of web page grab is depicted as follows:

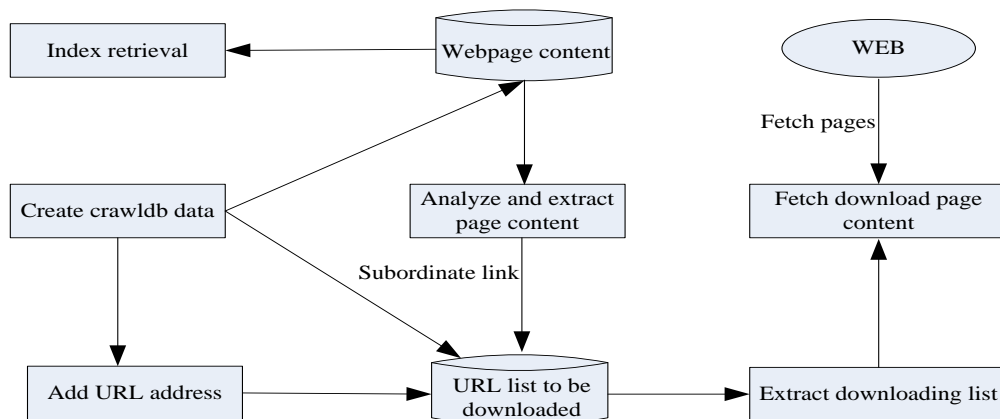


Figure 3: Webpages Crawling Process

The part of the network capture is using the web crawler to crawl the web page information, and saving them in the local database. Establish n initial crawl entry by manual adding URL set of seeds. Through inject operation, the URL set is format and filtered, to eliminate the illegal URL, and the duplication of the URL is merged, to delete the duplication of the URL portal. Then we added the URL to Crawled b database, and crawled database is used to store web crawler to crawl the web page URL and URL state, etc. According to the URLs in Crawled b database, we extract the list to be downloaded through Generate operation and create downloading task. Fetch operation grabs a list of the crawl according to the segment folder to access to web information from the internet. Updated b operation will grab the page information into the Crawled database, update the crawled database, and add new URLs to Crawled database, replacing the old URL Updated operation, which is briefly summarized as the update for the old.

2.2.2 Query Expansion Based on Ontology

In order to effectively overcome the problems that keyword matching-based results excessively depend on the key work. In this paper, we study the keywords for query expansion. The principle idea is: in the original key words we add new keywords associated to the query, in order to solve the problem of lacking information about the user query. In this research, we use the concept hierarchy of ontology in agriculture to carry out logical reasoning, and query the user's retrieval expansion in three modes, the upper, the parallel and the lower level. The detailed query expansion procedures are: Step 1 processes the query word segmentation to get meaningful keywords; Step 2 extends the keyword based on the agricultural ontology library. We use keywords that match the query in the ontology library, to get a set of concepts. If there is not matching results, jumped to step 3; otherwise go to the next step Sequentially, we use the concept of set in every concept of

ontology triples traverse, to get the corresponding upper, lower, and parallel domain concept. These concepts are put into a set of keywords. Step 3 gets the results of the query by search all the key words in the index library.

2.2.3 Result Sorting

To facilitate the user query, the results need to be in accordance with the relevant degree of formation with the sequence from high to low, so as to improve the accuracy and efficiency of query. The traditional web page ranking algorithm is computing the page and keyword matching degree. This paper performs query expansion with the user keyword. Therefore, page and keyword expansion between the relevance of the sort will make effects. The matching degree between the and page keyword can be integrated by using the correlation of web pages and keywords; The expansion degree page and keyword of correlation is between concepts from different areas related degree, denoted by the similarity of domain concepts, to initial keywords and each expansion of keyword similarity weighted average value to represent the correlation degree.

Set $\alpha = \{\alpha_1, \alpha_2, \dots, \alpha_n\}$, $i=1, 2, \dots, n$ to denote the initial keyword group of user query. Its corresponding eigenvectors is $R = (\omega_1, \omega_2, \dots, \omega_n)$. ω_i is the weight of α_i ; The page in returning results is denoted by eigenvector $P = (\bar{\omega}_1, \bar{\omega}_2, \dots, \bar{\omega}_n)$. The weight corresponding to α_1 can be acquired by TF*IDF algorithm. Array T_i is query extended word based on ontology of initial keyword α_i and corresponding correlation degree is $s = (\mu_1, \mu_2, \dots, \mu_n)$. μ_i is the weighted average value of domain similarity of each extended keyword in webpage keyword and T_i . The weight of matching degree of webpage and keyword is r and the weight of the correlation degree between webpage and extended keyword is $1-r$, the correlation degree of querying results can be computed as:

$$Sim(P, R) = \frac{\sum_{k=1}^n \omega_k * \bar{\omega}_k}{\sqrt{\sum_{k=1}^n \omega_k^2 \sum_{k=1}^n \bar{\omega}_k^2}} * r + (1-r) \sum_{k=1}^n \bar{\omega}_k * \mu_k \quad (1)$$

3. System test results analysis

3.1 Function Test

We make functional test and performance test from three aspects: the agricultural domain ontology library management, the search engine index database management and the user search interface. The test results are evaluated. Enter the system management platform on the domain of agriculture concept, and get into the domain of agriculture concept acquisition interface, run sequentially from top to the operation of resources in the domain of agricultural resource data cleaning- agriculture related degree evaluation, to acquire agricultural domain concept, and the running results are shown as figure 4.

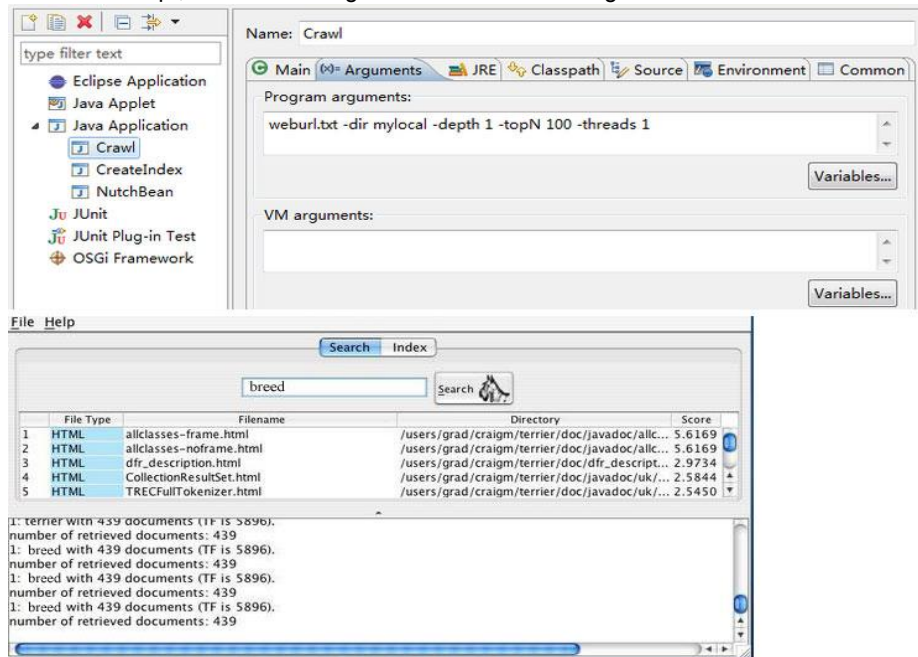



Figure 4: Agricultural Domain Concepts Acquisition

Enter the search engine management interface to make agricultural information retrieval. Under this interface we can carry out under the mode of retrieval, model retrieval and agricultural web site navigation services. The operating results are shown as figure 5.



The screenshot shows search results for the keyword 'Agriculture'. At the top is the 'nutch' logo. Below it is a search bar containing 'Agriculture' and buttons for 'search' and 'help'. The first result is from 'Agriculture News » Topix' with the URL 'www.topix.com/science/agriculture'. The second result is from the 'Department of Agriculture' (HDOA) with the URL 'hdoa.hawaii.gov'. The third result is from Merriam-Webster defining 'agricultural'. The fourth result is from 'Agriculture news, articles and information' with the URL 'www.naturalnews.com/agriculture.html'. The fifth result is from the 'Department of Agriculture, Forestry and...' (DAFF) with the URL 'www.daff.gov.za/daffweb3'.

Figure 5: User Retrieval Results

3.2 Performance Test

Since there are huge amounts of data available on the Internet, the recall of measuring the quality of a search engine becomes no significance and the precision for search engine is very valuable. In addition, the indices like dead link rate, repetition rate, response time evaluation of search engine are also very important. Therefore, in this paper, we use five key words search combined with the above the search engine evaluation indices in Baidu, Sounong search which are more representative of the search engine to be analysed and compared. The precision and rate of dead catenary, repetition rate, the responding time are the average precision, and the analysis result are shown in table1.

From table 1 it can be found, the precision of general search engine Baidu is the lowest one. The reason is due to a lot of businesses with a particular keyword registered trademarks of non-agricultural commodity; The recall and precision rate of Sounong is relatively inferior; The scheme proposed in this paper based on Ontology of vertical search engine of agricultural information search get the highest precision

Table 1: Comparison of different search engines

Search engine	Precision	Dead links rate	Repletion rate
Baidu	2/10, 4/20, 8/30/, 8.40, 16/50	0/10, 0/20, 0/30, 0/40, 0/50	0/10, 0/20, 0/30, 0/40, 0/50
Sounong	7/10, 17/20, 25/30, 34/40, 43/50	0/10, 0/20, 1/30, 2/40, 1/50	0/10, 0/20, 0/30, 0/40, 0/50
Nutch-based search engine	9/10, 20/ 20, 25/30, 38/40, 46/50	0/10, 0/20, 1/30,1/40, 1/50	0/10, 0/20, 0/30, 0/40, 0/50

4. Conclusion

There are problems such as precision, low rate of utilization in existed search engines. This paper studies these problems in the open source search engine on the basis of development, to realize the topic relevance discrimination and improved page scoring algorithm, to improve the accuracy of the search engine, and the utilization of extended user retrieval interface to improve search engine. Then we construct a Nutch-based vertical search engine for agriculture, to provide professional, fast and direct service for agricultural users, satisfying the user query demand of agricultural information. Therefore, the test results show the research for agricultural vertical search engine has great practical value.

Acknowledgment

This work is supported by the Fundamental Research Funds for the Central Universities (XDJK2013C071).

References

- Arafat A.H., El Desouky A.I., Saleh A.I., 2008, A new approach for building a scalable and adaptive vertical search engine, *International Journal of Intelligent Information Technologies*, 4, 1, 52-79.
- Gao B.L., Sun J.Y., Yu X.L., 2012, OKN: A presentation method for web resource based on ontologies, *Advances in Intelligent and Soft Computing*, 148, 1, 317-322.
- Hou G.Y., Zhang L., Ha S.H., 2014, Research and implementation of a vertical search engine in the financial domain, *International Journal of u- and e- Service, Science and Technology*, 7, 5, 117-128.
- Li W.Y., Zhao Y., Liu B., 2009, The research of vertical search engine for agriculture, *Advances in Information and Communication Technology*, 294, 799-803.
- Sun J.Y., Yu X.L., Li X.H., 2008, Research on recommender system based on expert profiles, *Journal of Computational Information Systems*, 4, 6, 2709-2714.
- Tarantino E., 2013, A simple model of vertical search engines foreclosure, *Telecommunications Policy*, 37, 1, 1-12. doi:10.1016/j.telpol.2012.06.002.
- Wang S., Liu S.H., 2013, Research on the key technology of agricultural production and market information matching system under the cloud computing background, *Applied Mechanics and Materials*, 41, 1, 2141-2147.
- Zhou J.Q., Wu Y., Jiang J.H., 2011, Object cache optimization strategy for real-time vertical search engine, *Journal of Zhejiang University (Engineering Science)*, 45, 1, 14-19.