Research and Design of Cross-Language Vertical Search Engine for Special Agricultural E-commerce Platform

Yaru Cao¹, Ning Ma², Fucheng Wan³ and Xiangzhen He⁴

^{1,2,3,4}Northwest Minzu University, Lanzhou city, Gansu province, China

Abstract. Based on the research of vertical search engine and cross-language information retrieval, a cross-language vertical search engine design for e-commerce platform is proposed. It aims to solve the problem that it is difficult for Internet users to quickly, efficiently, and comprehensively search for valuable products, especially ethnic minority netizens. Cross-language in this article mainly refers to the conversion of Chinese, English, and Tibetan. Using dictionary-based query translation method to translate query words to achieve cross-language function. Improved Heritrix designed a web crawler information collection method. Using HtmlParser to achieve structured information extraction, and using Lucene to build an index and achieve retrieval.

1 INTRODUCTION

With the rapid development of science and technology, the Internet quickly enters every area of social life and plays a role that cannot be underestimated. According to statistics from CNNIC, as of December 2017, online shopping users reached 533 million, and the search engine penetration rate was 82.8%. But most of the vertical search engines are now based on Chinese. For ethnic minority peoples in ethnic areas, they do not have very perfect Chinese reading ability. When he was faced with a huge Chinese webpage in e-commerce, they couldn't retrieve the Chinese products they wanted. This type of user naturally generates a strong demand for cross-language search information.

2 SYSTEM INTRODUCTION

2.1 Cross-Language Vertical Search Engine Concept

The vertical search engine is a new search model developed to reduce the general search engine's search results in terms of accuracy, accuracy, and professionalism. The vertical search engine is a deep search that serves a specific field, a specific group of people, or a specific demand, and has a distinctive industry color. The cross-language vertical search engine refers to the addition of cross-language functions on the basis of vertical search engines. In this article, when the user belongs to Tibetan language query words, the Chinese description information can be retrieved.

2.2 System Function Requirements Analysis

The cross-language vertical search engine of the e-commerce platform for featured agricultural products is an extension of the vertical search engine. In addition to having "professional, refined, and deep" industry characteristics, it also has a variety of language conversion functions. In addition, it also has the basic functions of a vertical search engine system, including information acquisition, structured information extraction, construction index, information search, user interface. In summary, the system includes six functional modules as shown in Figure 1.

Fig. 1. Functional Module

3 SYSTEM DESIGN

3.1 System Architecture

The system is mainly divided into six modules: information acquisition module, information extraction module, index construction module, search module, query word translation module, and user interface module. The information acquisition module is to extend the specific Web pages of the "Belt and Road Special Product Network" through the extension of Heritrix Extractor and Frontier Scheduler, and download it to a locally created mirrored

Corresponding author: 11195825322@qq.com, 26105112@qq.com, 3306261663@qq.com, 45967148@qq.com

folder. The information extraction module uses HTML Parser to extract the product metadata from the downloaded product page information to form structured data. This system has extracted the picture address, title, content, original URL address and other data in the web page. This system has extracted the picture address, title, content, original URL address and other data in the web page. At the same time, insert this information into the MySQL database. The index module builds an index library by building a full-text search of the metadata information of the commodity by extending Lucene's functionality. The search module uses Lucene to match the query terms entered by the user with the index library, and feeds the search results back to the user. The query word translation module translates into multiple language-written query words through a single-language query word in a multi-language dictionary database, and is collectively handed over to the system for retrieval. The system architecture is shown in Figure 2.

Fig. 2. Cross-Language Vertical Search Engine System Architecture

3.2 Information Acquisition Module Design

This system uses Heritrix as a tool for automatically collecting information. Heritrix does not have a directional capture function. It only captures all web page information and does not meet the requirements of this system. Therefore, in this design, the Extractor is customized according to the characteristics of the URL of the desired web page to replace the originally embedded class. Second, it further expands the Frontier Scheduler implementation to crawl specific types of content. This class is a Post Professor. Its role is to add the link from the previous Extractor analysis to Frontier for further processing. By limiting Heritrix's crawling behavior, download the required web page information.

3.3 Design of Structured Information Extraction Module

The source code of the web pages that Heritrix has crawled contains a lot of useless semi-structured data such as HTML tags. You need to extract the required information from it and save it for use when creating an index. This article chooses the structured information

extraction tool HtmlParser, which extracts different content data according to the meaning and attributes of the web page tag. Although the contents of the HTML files may not be exactly the same, they have the same basic framework:

```
<html> <head> <title> </title> <meta> </meta>
</head> <body> </body> </html>
```

The information to be extracted in this section is the subject part between <body></body>.

3.4 Design of Index Module

The role of the index is similar to the role of the book catalog. According to the page number in the catalog, the required detailed information can be quickly found and the retrieval speed can be greatly improved. The index is to index the structured data extracted from the previous link and stored in the index database. Such as product name, product brand, product price and other data, the schematic diagram of the establishment of the index shown in Figure 3.

Fig. 3. Index establishment

3.5 Query Word Translation Module Design

Cross-language information retrieval is essentially a match between single-language query words and multi-lingual retrieved resources. The system uses a dictionary-based query translation method, which is based on a single language information retrieval system and adds a language converter on top of it. After the user enters the query word, the query word written in other languages supported by the system is translated by the language converter, and then the single language information search can be performed. Obviously, this language converter is the core of the technology that is the most important for the translation of query terms. The multilingual dictionary in this design is this language converter. The multilingual dictionary used by this system is the research team of the National Institute of Information Technology of Northwest China University for Nationalities through the excavation and organization of ethnic agricultural resources. A handmade multi lingual national dictionary library of agricultural products (abbreviated as a multi-language dictionary) Based on the analysis and design of the translation of query terms, the model is set up in the three languages of English, Chinese, and English as an example. As shown in Figure 4.

Fig. 4. query word translation model

This module is based on a multilingual dictionary of human translation. It guarantees the correct correspondence between words and words. Compared to machine translation, it solves the problem of "word polysemy" and greatly increases the accuracy rate.

3.6 Search Module Design

The search module searches for all document records according to a certain search strategy in the index file to be generated, and feeds the result set back to the user in a certain order. The system uses Lucene to achieve retrieval. The process is actually connected to the user while the other side is connected to the index library and the search module client. The input query type matches a source language to other languages through a query word translation module, and then is delivered to the user terminal of the search module. The end of the index database is to take the user's query words to the index database to match the index entries. Finding the document record through the pointer of the index item generates a plurality of inverted document lists. Finally, all the document records are merged and classified and returned to the user in a certain order. As shown in Figure 5.

Fig. 5. Search Module Workflow

3.7 User Interface Design

The user interface module adds two pages to the system. One is the search engine homepage index .html, and the other is the search results page search .jsp. Design related pages based on the principles of natural harmony. As shown in Figure 6.

Fig. 6. User Interface

4 Conclusions

This paper first analyzes the universality of search engines through data analysis and the difficulties of ethnic minority netizens in searching for valuable information. This leads to the importance and urgency of the research and design of the cross-language vertical search engine for the e-commerce platform for featured agricultural products. Secondly, the six major functional modules of the system are identified by analyzing the needs of cross-language vertical search engines. Finally, analyze and design the overall framework of the system and each functional module.

At present, China's study of Tibetan, Mongolian and Uyghur language processing is not mature enough. Combining cross-language and vertical search engines in an e-commerce environment is rare. Mature use is a goal in the future. In the future research work, we need to improve and optimize cross-language.

Acknowledgement

This research was supported by the National Science-technology Support Plan Projects (Grant NO. 2015BAD29B01) and the Fundamental Research Funds for the Central Universities (31920170155).

References

1. China Internet Network Information Center, The 31th China Internet Development Statistics Report, <http://www.cnnic.cn/>, 2017.12.
2. Pu Hong, 2016. Research and Implementation of Lucene-Based Search Engine. Dalian University of Technology.
3. Mingqing Wang, 2010. Research and Implementation of Web Vertical Search Engine Technology Based on JAVA+Lucene+Heritrix. Hebei University of Technology,
4. Zhou, Yachao Li, TaoJiang, Hongzhi Yu, Yangji Jia, 2011. Dictionary-based Cross-Language Retrieval System in Tibetan and ChineseJournal of Northwest University for Nationalities(Natural Science), vol. 32, pp. 40-43. (in Chinese)
5. Heng Wang, Shaoshan Wang, 2013. Research and Implementation of Subject-Oriented Intra-domain Vertical Search Engine System Journal of Ningxia University (Natural Science Edition), vol. 34, pp. 54-57 (in Chinese).
6. Story R.E, 1996. An Explanation of the Effectiveness of Latent Semantic Index by Means of a Bayesian Regression Model. Information Processing and Management, vol. 32, pp. 329-344.

7. Chuanming Jin, 2011. Research and Implementation of Vertical Search Engine. Yunnan University (in Chinese).
8. Zhe Qiu, Taotao Fu. Lucene, 2017. 2.0+Heriteix develops its own search engine. Posts & Telecom Press (in Chinese).
9. Fucheng Wan, Dongchen Li, Xiangzhen He, 2014. Tao Xu. Research on Tibetan Text Indexing Strategy for Information Retrieval. Computer Engineering and Applications, vol. 50, pp 208-211 (in Chinese).
10. YunQiang Liu, 2010. Research and Design of Vertical Search Engine. Computer Applications and Software, vol.27, pp 130-133 (in China).