ATLANTIS PRESS
Advances in Engineering Research (AER), volume 132
2017 Global Conference on Mechanics and Civil Engineering (GCMCE 2017)

Research on Cross-language Vertical Search Engine Strategy for E-commerce of Characteristic Agricultural Products

Ning Ma [1,a], Yaru Cao[2,b,*] , Xiangzhen He[3,c], Fucheng Wan[4,d], Jiajia Li[5,e] and Heng Yang[6,f]

[1]Northwest Minzu University, Lanzhou city, Gansu province, China

[2]Northwest Minzu University, Lanzhou city, Gansu province, China

[3]Northwest Minzu University, Lanzhou city, Gansu province, China

[4]Northwest Minzu University, Lanzhou city, Gansu province, China

[5]Northwest Minzu University, Lanzhou city, Gansu province, China

[6]Northwest Minzu University, Lanzhou city, Gansu province, China

[a]6105112@qq.com, [b]1195825322@qq.com, [c]5967148@qq.com, [d]306261663@qq.com, [e]1103089863@qq.com, [f]281830672@qq.com

*Yaru Cao

Keywords: Cross language vertical search engine, E-commerce, Focused web crawler, Cross language information retrieval.

Abstract. Based on the analysis of key technologies such as web crawler, information extraction and cross-language retrieval. This paper searches the related goods using Chinese, Tibetan, Mongolian and Uygur or English from the corresponding index database by using multi-lingual dictionary. In order to achieve a purpose of "one language search multiple languages" vertical search engine.

1. Introduction

At present, in the e-commerce environment, the types and quantities of information resources on e-commerce of characteristic agricultural products are becoming more and more abundant, and the languages used are more and more diversified. With the diversity of e-commerce information resource language and the differences of languages that the network users master, it inevitably brings the language barrier when people use the network to search information, especially when it comes to language barrier of the minority language. This language barrier greatly limits the effective access to information resources, and thus the e-commerce cross-language vertical search engines for agricultural products comes into being.

2. Research on key technologies of cross-language vertical search engine

2.1 Web crawler technology--focused web crawler

Focused web crawler is different from the general crawler. The focused crawler does not seek a large coverage. Its target is to grab web pages related to a specific topic content, and to prepare data resources for the topic oriented user queries. The workflow of focused crawler is more complex. Because it needs to filter the irrelevant links according to a certain webpage analysis algorithm, and retains useful links then put them into the URL queue waiting to be grabbed. Then, it will select the next page URL to be grabbed from the queue according to a certain search strategy, and repeats the process, and it doesn't stop until it reaches a certain condition of the system.

2.2 Web information extraction technology

The so-called web information extraction technology refers to the crawling of the web as a source of information, and it makes the page information more structured, clear and the structure more uniform after a variety of processes. There are five kinds of web information extraction technology: NLP-

based Tools, HTML-aware tools, wrapper induction tools, ontology-based tools, and Web query tools, which have made outstanding contributions to the cross-language vertical search engine. With the continuous development of the future of e-commerce, the web information extraction search method will play a greater application value.

2.3 Cross-language information retrieval technology

Cross-language information retrieval can be divided into the following three processes. First of all, the collection of multi-language information resources and the establishment of multi-lingual information index. Secondly, the unification of the source language and the target language technology realizes via automatic processing of language application. Thirdly, the match of query and index information realizes via using the single language information retrieval technology, and thus gets the retrieval result. Among them, the unification of source language and target language is the key technology to realize the cross-language information retrieval. It can be accomplished mainly in four ways. They are query translation, document translation, triangulated translation and no translation. This design uses query translation. Here we only explain what the question translation is. Query translation is translating the translation of queries into multiple languages supported by the system submitted by a user, and then inquires the different language information. There are two main methods in question translation: dictionary and corpus. The core idea is to translate the query-style by means of the corresponding relation of the same information and different language in the corpus or dictionary, and finally to filter the translation results of the ambiguity. Query translation is to achieve cross-language information retrieval in a more economical way.

3. Cross-language vertical search engine architecture design

Cross-language vertical search engine is an extension of the vertical search engine, with special, refined and deep industry features, and it also has multi-language conversion function. In this paper, the cross-language vertical search engine refers to the transformation of Tibetan, Mongolian, Chinese and English languages. That is to say the user has to submit a familiar language (source language) through the multi-language dictionary translation, and then searches the index library that meets the needs of users of goods from the corresponding language. Finally, the unified presentation of products in multiple languages displays in the same window. The source language is a language, and the target language is a variety of languages. Let’s take Tibetan as an example, cross-language vertical search engine architecture is shown in Fig. 1.

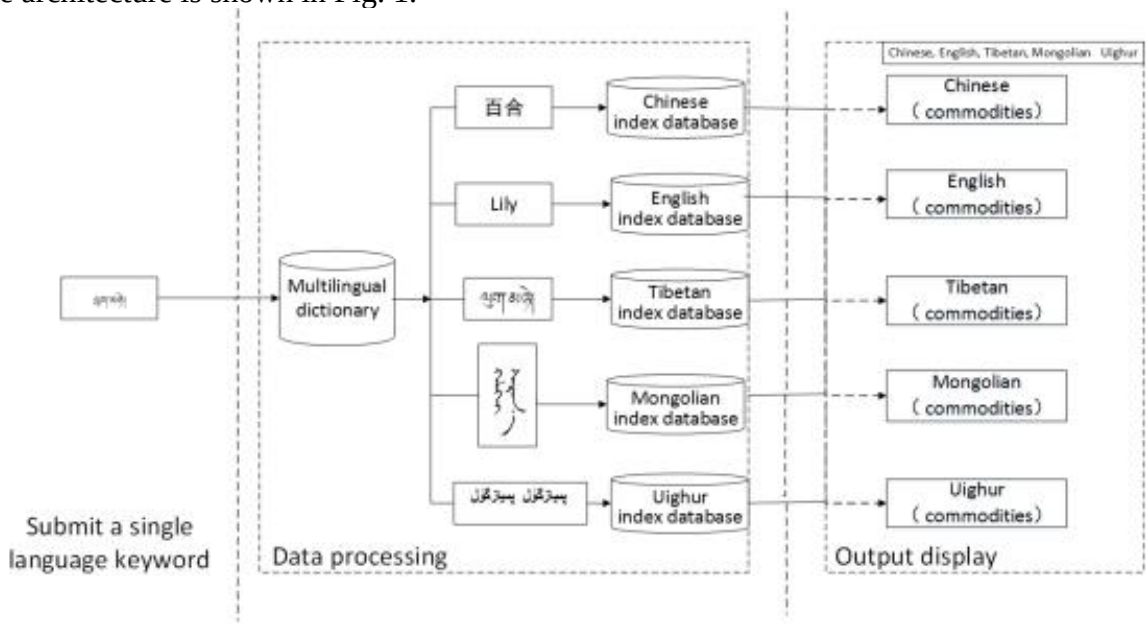

Fig. 1. Cross-language vertical search engine architecture design.

4. Cross-language vertical search engine work flow

According to the design architecture. I sort out the research idea, and draw their work flow shown in Fig. 2. The vertical search engine design has five modules: information acquisition, information extraction, create index, information retrieval, and user interface. Each module cooperates to complete its own functions. There is a certain work among them, and the former module is to prepare for the back of the module.

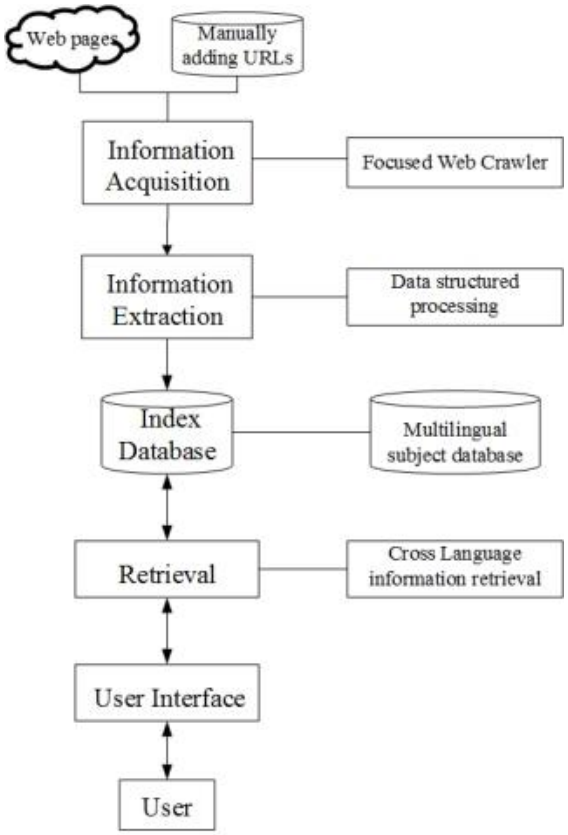

Fig. 2. Cross-language vertical search engine work flow.

4.1 Information acquisition module

The function of this module is to collect the information of HTML pages, and thus collect and save the users' needs of the web page information. Firstly, we should set the acquisition rule of all users' needs to collect the site, and then we can collect the information. As the different pages of the page structure are inconsistent, the same sites among the different pages structure are roughly the same. So we should set customization rules for the characteristics agricultural products trading site, and then use the web crawler software to web crawl. The web crawler software used in this article is open. Extensible Heritrix crawler framework, and its mainly work is to download the required data from the Internet.

4.2 Information extraction module

The function of this module is to filter out HTML tags that are not related to the web page. It only needs to extract useful information and store the contents of the web pages according to its format. The useful information includes: text content, title, link, source and so on. This system uses the extraction tool Html Parser, which mainly contains two aspects: conversion and extraction.

The main contents that Html Parser extract are as follows: text extraction, extraction of link, link extraction and site inspection. Html Parser conversion mainly includes: URL rewrite, clear ads. The transformation of HTML pages into XML pages and clear of XML pages.

4.3 Create index

This module refers to making use of the indexer to extract the structured data from previous extraction to generate a similar number of index files, its purpose is to achieve a rapid retrieval. This module also chooses to use Solr to build indexes. Solr not only encapsulates and extends index acquiescently, but also provides an easy operation user management interface. The Solr only builds English index acquiescently. So before the establishment of index, it should add the Chinese word segmentation, Tibetan word segmentation, Mongolian word segmentation, and Mongolian Tibetan Uyghur word into Solr index. Its purpose is to achieve the function of cross language index. After using Solr to build index data.You can call the solr directly to search.

4.4 Information retrieval

The main task of this module is to retrieve in the index library rapidly according to the key words input by the user. The search system can find all the relevant pages from the web index database. Then, it estimates the web text content and keywords related degree via a specific algorithm, and ranks the results in accordance with the relevant degree. With a higher relation, it ranks near the front. Finally, it returns to the user in the form of hyperlink address and the main contents of the page by using the specific sorting algorithm to generate the results.

4.5 User interface

User interface is an important component of vertical search engine. Its function is to provide query interface for users and display query results. The main purpose is to provide users with search engines to get effective search results. User interface design and implementation use human-computer interaction theory and methods, so as to be fully adapt to human thinking habits. In brief, the user interface module provides a bridge for the user to query the interface and to build the index library.

5. Conclusions

With the development of the market diversification, specialization and multilingual, cross-language vertical search will be more adaptive to the development trend of the future market. At the same time, with the further study of the rapid development of network and the key technology. I believe that more and more industries are realizing the value of cross language search engine. Research on characteristics of agricultural products e-commerce cross-language vertical search engine not only eliminates language barriers for the characteristics of e-commerce of agricultural products in the process of vertical search in an effective way, but also breaks down language barriers if the technology is applied to the related field of vertical search engine, so as to promote the comprehensive development of the field of computer. We should attach great importance to the research of cross language search engine vertical search and to analyze research from multi angle and multi aspects, so as to research a much better, more conducive cross-language search engine.

Acknowledgement

This research was supported by the National Science-technology Support Plan Projects (Grant NO. 2015BAD29B01).

References

[1] Razieh Rahimi, Azadeh Shakery, and Irwin King, Extracting translations from comparable corpora for cross-language information retrieval using the language modeling framework, *Information Processing and Management*, vol. 52, pp. 299-318, 2016.

[2] Tao Peng, F. He, and Changli Zhang, Adaptive Topical Web Crawling for Domain-Specific, *Lecture Notes in Computer Science*, vol. 4293, pp. 963-973, 2006.

[3] I. Vulic, W. De. Smet, and M.F. Moens, Cross-language information retrieval models based on latent topic models trained with document-aligned comparable corpus, *Information Retrieval*, vol. 16, pp. 331-368, 2013.

[4] D.Zhou, M.Truran, and T.Brailsford, etal, Translation techniques in cross-language information retrieval, *Int. J. ACM Computing Surveys*, vol. 45, pp. 1-44, 2012.

[5] Pilsung Choe, Mark R. Lehto, Jan P. Allebach, Query translation-based cross-language print defect diagnosis based on the fuzzy Bayesian model, *Journal of Intelligent Manufacturing*, vol. 22, pp. 43-55, 2011.

[6] Heng Wang, Shaoshan Wang, and Yuzhuo Gao, Research and implementation of topic- oriented vertical search engine system in domain, *Journal of Ningxia University (Natural Science Edition)*, vol. 34, pp. 54-57, 2013(in Chinese).

[7] Chunan Wang and Yufu Li, Research on information filtering technology in vertical search engine, *Int. J. Information Science*, vol. 32, pp 93-97, 2014(in Chinese).

[8] Min Zhu and Shengxian Luo, Research on focused topic oriented crawler based on Heritrix, *Computer Technology and Development*, vol. 22, pp 65-68, 2012(in Chinese).

[9] Qiyu Zhang, Huihui Yu, and Yingyi Chen, etal, Research on construction of Chinese word segmentation dictionary based on agricultural vertical search engine, *Guangdong Agricultural Sciences*, vol. 42, pp 165-169, 2015(in Chinese).

[10] Congjun Long, Analysis of several key issues in Tibetan text information processing, *Computer CD Software and Applications*, vol. 3, pp 51-58, 2012(in Chinese).

[11] Jianxia Chen, Ri Huang, and Zhongbao Ma, Optimization and implementation of Lucene sorting algorithm based on Pagerank, *Computer Engineering and Science*, vol. 34, pp 123-127, 2012(in Chinese).