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

Research of vertical search engine index module for college entrance examination Forum

Yangyang Guo [1,a], Cao Hui[2,b,*] and Fucheng Wan[3,c]

[1]Northwest University for Nationalities, National languages Information Technology research institute, Lanzhou city, Gansu province, China

[2]Northwest University for Nationalities, National languages Information Technology research institute, Lanzhou city, Gansu province, China

[3]Northwest University for Nationalities, National languages Information Technology research institute, Lanzhou city, Gansu province, China

[a]1597791068@qq.com, [b]147625251qq.com, [c]306261663@qq.com

*Yangyang Guo

Keywords: Vertical search engine, Classification system, Classification index.

Abstract. The vertical search engine index classification method based on ontology, through the introduction of text classification in PubMed forum based on ontology semantic information is added into index, can effectively improve the vertical search engine retrieval recall and precision, improve the relevance of search results ranking results. Firstly, the design of a classification system based on domain ontology, achieve fine-grained multi class text classification; and then design a new classification index structure, the categories of information and information and the formation of effective combination of keywords, classification index; finally the generating algorithm of classification index design, and puts forward the classification index compression, optimization method.

1. Introduction

The index as the core of the search engine, is a bridge capture and retrieval, this paper improved to enhance the vertical search engine retrieval results in the index part. In the index improvement it has been doing a lot of work, such as the introduction of semantic semantic semantic keywords and keyword double index, in the traditional inverted index structure based on the introduction of adjacent bit information, enhances the semantic, improve the recall ratio; introducing the thought of classification, to a certain extent, to speed up the retrieval speed, get better ranking, but they are mainly based on weight classification, category of coarse, not a very good classification system, recall and precision improvement effect is not obvious.

2. Requirement Analysis

This topic is with the characteristics of the vertical search engine design for a university postgraduate forum classification index module, due to the development of modern society, more and more competitive, people demand for education is also gradually attention, then set off a craze for graduate school, ready to graduate students is also through various channels to search for a postgraduate colleges information. A channel which PubMed forum is to obtain information, facing the drawbacks of the existing search engines, the index module is optimized with the advantages of vertical search engines, in order to better index to search more information, they need to graduate students to help university forum information better.

3. Research on system function module

According to the demand analysis, the vertical search engine index module for the university entrance examination forum can be divided into the following 4 parts, as shown in Fig.1. below.

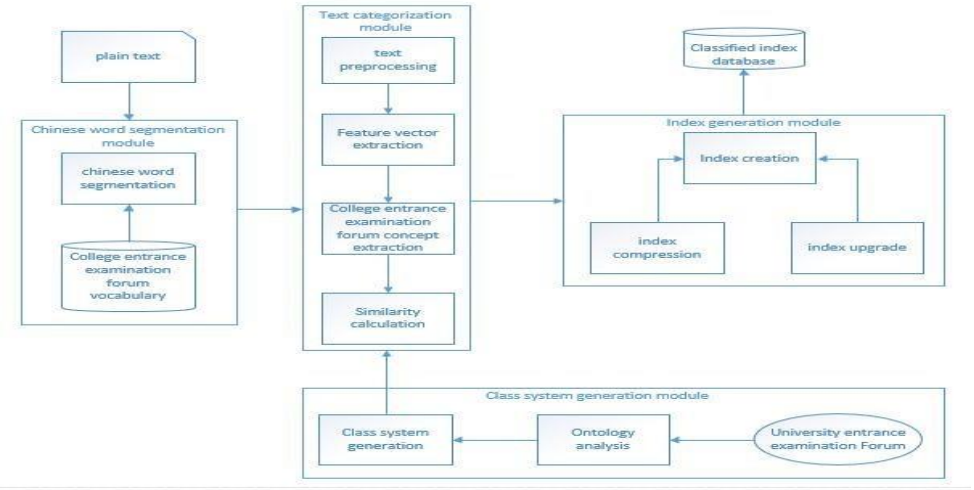

Fig.1. Classification index module diagram

(1)Research on text classification for university entrance examination Forum

The first text data to text preprocessing. Then the feature vector extraction from text information extraction of feature words to quantify the text information, they are from an original unstructured text into structured information processing can be identified by computer.

Then, we extract the concept of university entrance examination forum, and then select the domain document set,but pure text is a kind of data that exists in reality, so the main choice of this subject is pure text as data source.Then the document set pretreatment. Then domain term extraction, through artificial selection of seed concept and HowNet set of synonym synonym concept seeds to expand the concept of the concept of learning goal is to all these synonyms for learning, not just a word. The concept of seed as the foundation, from the text in order to find the seed concept as the center of the word term. Automatically contain the seeds noun phrase list in the analysis of text, every noun phrases are considered as a new concept may be, this is only a preliminary material, also need to further clarify the concept. It is not at the same time must meet the terms: low two righteousness and high specificity to become a concept, which is the basis for further processing. Finally the concept set formed in Natural Language Processing, usually put the sentence or part of speech sequence as a string of random. They are then analyzed and studied with statistical information.Finally, the similarity calculation is used, and the semantic similarity algorithm based on vector space is adopted here.

(2)Research on generation module of class system

The traditional classification system are generally constructed, limitations. Ontology as the conceptual system in the field of workers, the relationship between subclassof classification standard very well. There are some relations between the other concepts in the ontology tree, this relationship has certain relations by category number set to save the concept, can be easily retrieved. This thesis adopts the digital classification method, because of its letters and classification and classification of letters and numbers mixed classification method compared with the expansion of flexibility, convenient computer management, and no language barriers, is conducive to communication and promotion. Ontology parsing is to realize the feature words in the field of document vector concept with the concepts in Ontology based on field extraction, concept similarity concept and text classification system for hi layer node calculation. Ontology parsing to OWL ontology is analyzed by using Jena According to the conceptual model of OWL language, we

extract all kinds of information contained in ontology, such as class, attribute, instance and ontology metadata.

(3)Research on index generation module

The basic idea of classification index generation algorithm is if a part of the document is to create an index, will extract them from the document segmentation, and then the content of the document processing, statistical index data needed by compression algorithm to a document belongs to the category, keywords and other related information written memory, when the document number reaches the specified value of theta domain, by merging algorithm layer combined index output will eventually merge to good classification index in the database.

4. Construction of domain web classifier

(1) Text classification overview

Text classification is in accordance with the classification system defined in advance, according to the content of the text will be automatically placed in each text text set corresponding category. At present many text classification algorithm, classification algorithm is shown in Fig.2.

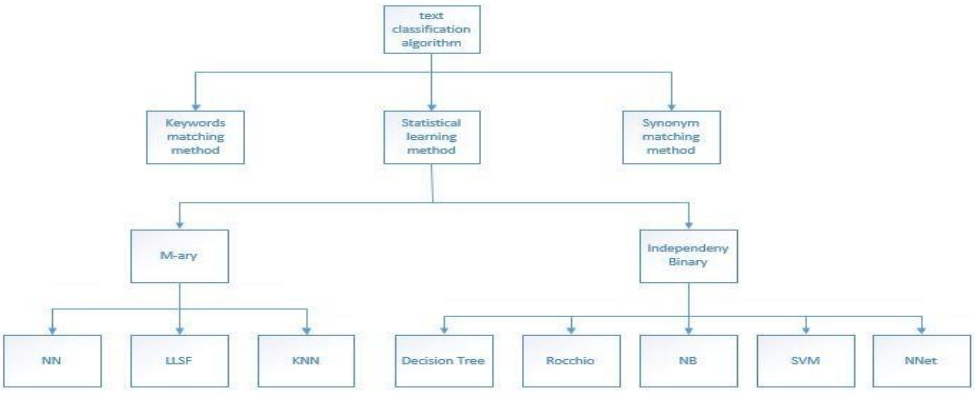

Fig.2. Text categorization algorithm

(2)Class system generation

The traditional classification system are generally made of artificial construction, limitation. Ontology as a conceptual system in the field of recognized, which classification standard subclassof relationship is a good comprehension of the concept of ontology. Subclassof includ es conceptual category between any one field, as long as the constructed ontology, you can through this system ontology ontology tree parsing and generation processing, generate the classification system of field requirements. There is a certain relationship between other co ncepts in the ontology tree, this relationship has certain relations by category number set to save the concepts of the ontology can facilitate retrieval. In this paperthe PubMed forum a s an example to build ontology classification system, such as Fig.3.

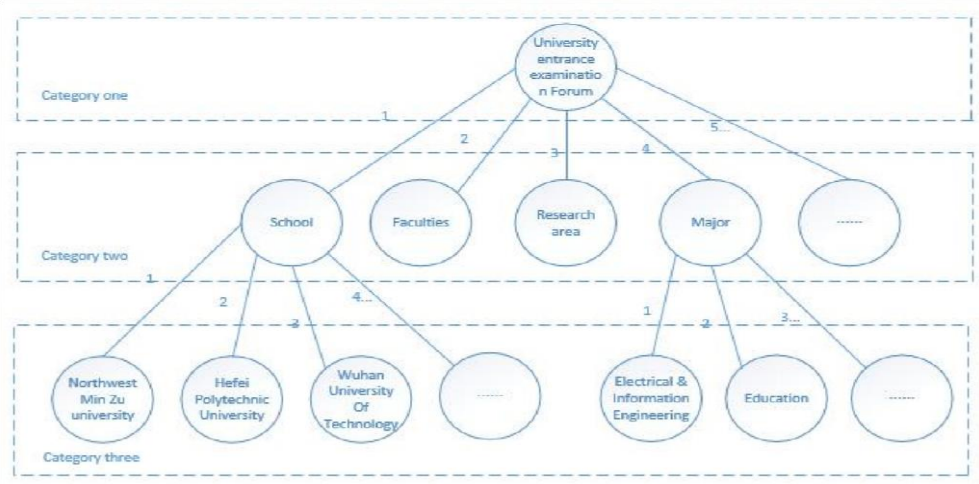

Fig.3. Example of university entrance examination forum classification system

5. Summary

This project can realize the information indexing University Forum Based on the completion of the examination, examination information analysis, classification, index and so on. The vertical search engine index module of this project is fully automated, without artificial part icipation, only need to automatically generate index input text to be indexed can. And in th e process of text automatic classification index. The index is based on keywords and doubl e index categories, compared to other simple keywords or semantic indexing based on, can improve the recall and precision, and can better improve the sorting effect. The vertical sea rch engine classification index system has two main functional modules, text classification a nd text classification accuracy index creation. Directly affects the accuracy of retrieval index return results, so this paper based high accuracy in the classification of the verification of the Department Experimental evaluation.

References

[1]Zhijian Fang,Ruilin Zhang, Xiaosu Tong. *a comprehensive analysis of search engine*[J]. *computer engineering and design*,2007,28 (16).4038-4041.

[2]R.Studer,V.R.Benjamins,D.Fensel.*Knowledge Engineering,Principles and Methods*[J].*Data and Knowledge Engineering*,1998,25(1-2):161-197.

[3]Ruiling Zhang,Wenbin Wang,Xiufeng Wang,Qiushuang Chen.*Case driven adaptive ontolo gy learning* [J]. *computer engineering and applications*, 2009,45(28).31-34.

[4]Zhanting Yuan,Aimin Zhang,Qiuyu Zhang. *Concept based web information retrieval*, [J]. *computer engineering and applications*,2003,39(36):173-181.

[5]Jianguo Liu. *Summary of search engines* [J].*Peking University computer andscience and technology*,1999.10.20.

[6]Jianhua Xu.*Principle of network search engine, characteristic analysis and future devel opment* [J]. *library and information work*,2000(8).34-38.

[7]Zhanting Yuan,Aimin Zhang, Qiuyu Zhang. *Concept based web information retrieval*[J], *computer engineering and applications*, 2003,39 (36): 173-181.