Research and Implementation of Topic Search Engines System based on Medical Document Classification

Yanmei Hu

Chengdu Medical College, Chengdu, Sichuan, 610083, China

13965808@qq.com

Keywords: Medical documents; Search engines;Topic search; Text categorization

Abstract. This paper studies and analyzes the principle, strategy, structure and working mode of "subject network search engines". A topic search engines system based on medical document classification is designed and implemented in Windows environment using C++. In this paper, the vector space model is used to describe the document, and the document is classified by the simple vector distance method. Through the implementation and testing of the system, it can be concluded that the method used in this paper has a great improvement in the efficiency of the operation and the effect of crawling strategy, which greatly improves the accuracy of the search engines.

Introduction

Search engines has become an important way for people to get the effective information on the Internet. At present, the commonly used search enginess are Goolge, Baidu, SoSo. Search enginess typically use one or more resource collector from the Internet to collect all kinds of valuable data (such as page resources) and indexing the data on the local server and for users provide key information query service. Search enginess for automatic acquisition of data acquisition is also known as the network search (Spider), web crawler or network robot [1] they are an important part of the search engines.

Although the early search system to achieve the information on the Internet automatic collection, but when the subject of medical to document the collection of data is often involved in the field is too wide, too large amount of information [2]. User information may not be accurate or response efficiency is too low, the results are not satisfactory. For this purpose, it is proposed that a search system which can collect some or some special subject resources can be collected on the Internet, which is called the topic search system [3]. Topic specific search system can effectively reduce the number of resource acquisition and improve resource gathering theme regularity, increase the utilization ratio of network bandwidth, at the same time, improve the efficiency of information retrieval is a very practical value for the network information resources acquisition solutions.

Topic search system can automatically search the web subject resource, so as to get rid of the dependence on experts, reduce the manual intervention, improve the speed, efficiency, and quality of theme website resources construction, provides the high quality of information resources and information service for scientific research personnel and associated user [4].In this paper, a prototype of the search engines system based on the automatic classification theory of medical documents and the optimization strategy of hyperlink is realized, and the new direction of the development and research of the subject search engines system is explored.

Working Principle of Search System

The working principle of the search system, mainly based on such a fact: the Internet have contains a large number of the web page URL and the page of the chain to the other page URL, which are respectively a huge picture on the web to (direct graph) graph nodes and edges [1]. By these edges (URL), or from some of the identified nodes (pages), the theory can traverse the entire Internet all nodes (pages), as shown in Fig. 1.

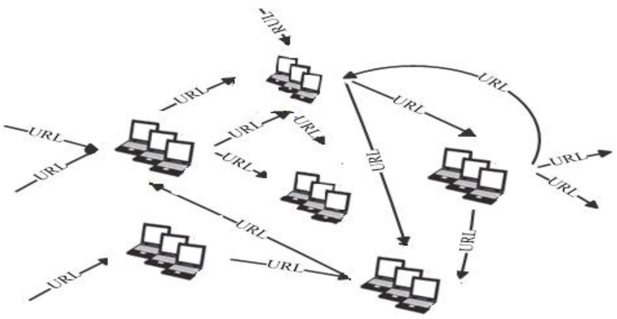

Figure 1. Characteristic of the Internet page on the directed graph

The Principle of Topic Search Engines System for Medical Document Classification

Medical document topic specific search engines system is different from general search engines system, by introducing the URL of the medical subject correlation pruning strategy and page medical subject correlation decision strategy can make the search behavior objective in the Internet [5]. Medical subject search system starting from the initial page, due to the introduction of the medical related theme crawling strategy, crawling with medical subject page has nothing to do with the URL in the would is optional path from the crawling collection cut out, can only crawl or first crawling medical subject related URL [4]. Using these strategies can significantly increase the rate of return on resources and save the limited bandwidth of the network. For the specific area of information query, it can greatly improve the quality and efficiency of information retrieval service.

Structure Model of Topic Search. Topic network search structure is shown in Fig. 2.Theme search engines system adds two important functions, namely, URL value evaluation function and page theme relevance judgment function, respectively. URL evaluation is responsible for assessment if the URL is not relevant to the subject, to determine when or whether the crawl the URL (pruning); page analyzer is responsible for investigation after a URL to download the page content is relevant to the subject. These two modules directly affect the rate of return, coverage and accuracy of the search focused crawling [6].

Thus, through the URL pruning module and subject judgment module of collaborative work and topic search system can anticipate a URL returns the value degree, search system to guide their own on the distribution of a particular subject in a huge library of Internet information resources to make more efficient, more comprehensive more accurate crawling.

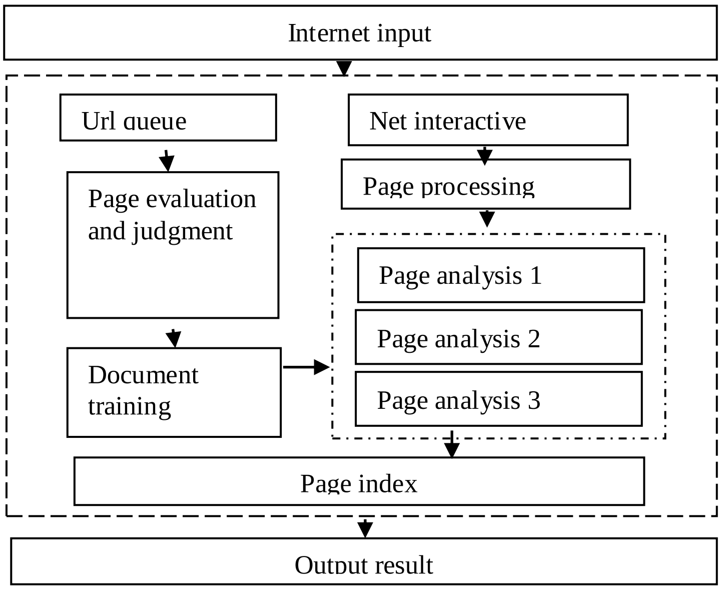

Figure 2. Topic network search structure

Judgment Strategy of Page Topic Relevance . In the implementation process of Web thematic information extraction, the extracted URL has passed the topic relevance judgment, but the extracted page content or may be far from the theme of the set. This phenomenon will affect the accuracy of the topic page information extraction. Therefore, after the extraction of the page, the need to distinguish between the topic relevance of the page to filter out the theme of the page. We use the vector space model based on vector space model to determine the relevance of the topic in the topic search system.

Vector space model (VSM) is a statistical model for document representation [4]. It based on the assumption that the entry appears in order is irrelevant, they for categories of documents are independent of each other, so we can treat a document as a collection of a series of disordered entries. The model is represented by the coordinates of the document as a document, and the document is represented as a point in the multidimensional space. For example, we say a document in a k - dimensional two-valued vector space model:

$$D_i = (d_{i1}, d_{i2}, K, d_{ik}), \qquad d_{ij} = 0 \; or \; 1 \tag{1}$$

There into, d_{ik} is the weight of the feature item in the document. Fig. 3 schematic document vector space model

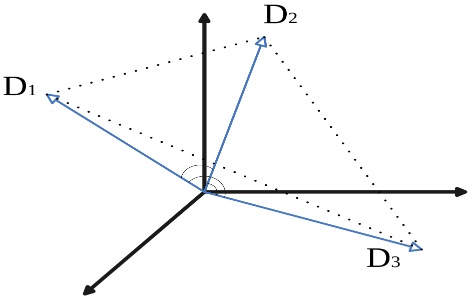

Figure 3. Vector space model of the document

Three thick black arrows in the figure constitute a three-dimensional coordinate space, the thin hollow arrows are 3D vector representation of D1, D2 and D3. It can be seen that if the feature to select the appropriate features of statistical strategies, we can use the quantitative method to achieve the different topics of the document classification.

Prior to the topic of medical documents, we use the inverse document frequency index to establish a generalization model.Inverse document frequency index method (TFIDF) based on a basic assumption: frequent entry (features) with less information, such as "you", "I" and so on, and there are fewer entries with more content information. We set up Tk as a feature entry, the entry in the document Di weights using the TFIDF value, TFIDF value of the calculation using the Eq. 2 [5]:

$$\omega_{ik} = TF_i \times \log({}^{N}/_{DF_i} + c) \tag{2}$$

Among them, TFi is the frequency which appears in the document Di, which is called phase frequency (TF), which is the feature of the Tk; The DFi representation of the corpus statistics collection contains the number of entries in the Tk document, whose reciprocal is IDFi, called the inverse document frequency (IDF); N represents the total number of documents in the collection of statistical data; C is the correction factor, usually 0.1. As a result, any document can be expressed as follows: K dimensional feature vector representation:

$$D_i = (\omega_{i1}, \omega_{i2}, K, \omega_{ik}) \tag{3}$$

Training and Classification of Topic Search Engines System for Medical Document Classification. Common automatic text categorization(ATC) methods, including the method of word matching method, statistical learning methods and knowledge enginesering In this paper, we use simple vector distance method.

On a training sample set, we will be the feature vector of all the documents in each class of documents to do arithmetic operations, can be used for document classification evaluation of the feature vector, known as the category of the center vector. Simple vector distance algorithm is to be abstract document classification into feature vectors, center vectors with each category compared calculation between the distance (similarity), will be included in document classification and the centre nearest vector distance represents the category. Usually use the vector included angle as a distance reference mark, calculated by the Eq. 4:

$$\mathrm{SIM}\left(d_i, d_j\right) = \cos\theta = \frac{\sum_{l=1}^{k}(\omega_{il} \times \omega_{il})}{\sqrt{(\sum_{l=1}^{k}\omega_{il}{}^2)(\sum_{l=1}^{k}\omega_{jl}{}^2)}} \qquad (4)$$

There into, d_i is the feature vector of the document to be classified, and d_j is the center vector of the class j document.

Design and Workflow of Topic Search Engines for Medical Document Classification

The topic search engines of medical document classification is divided into the following steps from the start to the system:

(1) the construction of the global data objects of medical documents.

(2) to start the system initialization process, after the completion of the initialization of the system to monitor.

(3) the initialization device creates and starts the working group and sets the control parameters and configure the working thread.

(4) the user controller is created, the system begins to receive the user theme control information.

(5) the end of the life cycle of the working group, the user controller is revoked, the system no longer responds to the interaction between the user and the working group.

(6) global data object analysis, release resources.

Performance Test of Topic Search Engines for Medical Document Classification

The subject search engines of medical document classification is a background subject search program running under the Windows operating system. The system needs to configure the resource file and the control parameters before running.

From the results of Fig. 4, it can be seen that for different operating platforms, the number of concurrent working threads is different from the balance point of the work efficiency. Need according to the specific environment to carry out specific balance adjustment. In addition, the efficiency of the search will be affected by the network bandwidth, HTTP server response efficiency and other important factors.

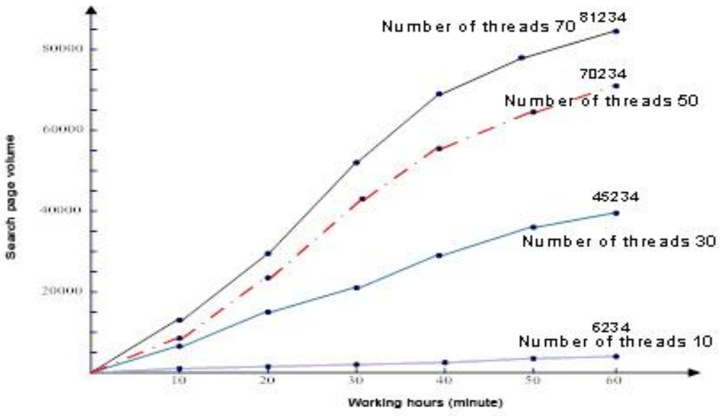

Figure 4. Analysis of test results

Conclusions

Topic search engines is one of the research directions of search engines. Medical topic network search is an important part of medical information collection. The choice of the model of medical subject search engines and the classification strategy of the document directly affect the function of the subject search engines [7]. In this paper, we design and implement a network search system for medical topic. With efficient design strategy, the system has a good system structure, can quickly get medical related topics on the Web page. At the same time, through the implementation of the system and the verification test, the results show that the subject search engines system based on medical document classification has more ideal performance, and can accurately obtain the high quality web pages.

Acknowledgements

This project was supported by Sichuan Education Department Science Project (No. 15ZB0246) and Chengdu Medical college Science Project (No. CYZ12-014).

References

[1] Taylor G. Search Quality and Revenue Cannibalization by Competing Search Engines[J]. Journal of Economics & Management Strategy, 2012, 22(3):445–467.

[2] Lewandowski D. Problems with the use of web search engines to find results in foreign languages[J]. Online Information Review, 2015, volume 32(32):668-672.

[3] White A. Search engines: Left side quality versus right side profits [J]. International Journal of Industrial Organization, 2013, 31(6):690-701.

[4] Cobos C, Mendoza M, León E, et al. TopicSearch - Personalized Web Clustering Engine Using Semantic Query Expansion, Memetic Algorithms and Intelligent Agents[J]. Polibits, 2013, 47:33-46.

[5] Wu M, Hawking D, Turpin A, et al. Using anchor text for homepage and topic distillation search tasks[J]. Journal of the American Society for Information Science & Technology, 2012, 63(6):1235–1255.

[6] Zhang W, Lu G, He H, et al. Exploring large-scale small file storage for search engines[J]. Journal of Supercomputing, 2015:1-13.

[7] Lewandowski D. A Framework for Evaluating the Retrieval Effectiveness of Search Engines[J]. Computer Science, 2015, volume 63(4):354-363.