The Application of Web Information Retrieval Technology Based on Semantics to Border Entry and Exit Personnel Pre-Declaration Platform

Jing Li, Jinhuan Wang
School of Intelligence Science and Information Engineering
Xi'an Peihua University, Xi'an, China
28328715@qq.com, 360635476@qq.com

Keywords: Information Retrieval; Semantic Web; Pre-Declaration Platform

Abstract. Today, the information is changing rapidly. People's demand for information retrieval is getting higher and higher. The information retrieval technology based on keyword has not fully met the user's requirements in the precision and recall. The information retrieval technology based on semantic is still a hot spot of research. The web information retrieval technology based semantic mainly relies on the semantic web, the smart network, to achieve semantic recognition and semantic extension, so as to get more efficient retrieval results, and make up for the shortcomings of past retrieval technology. This paper studies how to construct a web information retrieval framework based onsemantic.

1 Introduction

The traditional network retrieval technology is based on keyword information retrieval. This retrieval technique ignores the semantic information contained in the keyword. Therefore, the query results have been greatly affected by the recall and precision. The web information retrieval based on semantic can automatically analyze and reasoning the semantics of information resources by some means and methods, retrieve specific information units of specific knowledge meaning, and obtain retrieval results through semantic and semantic analysis.

Information retrieval refers to the process of organizing information in a certain way and finding the relevant information according to the needs of the information users. At present, the more common information retrieval technologies are: data retrieval, text retrieval based on keyword and text retrieval based on semantic.

2 Semantic Web

The semantic web is a kind of idea of the future network. The information in the semantic web is given a clear meaning. Computers can automatically process the available information, analyze and judge it according to the semantics. As an intelligent network that can "understand" the human language, it makes communication between people and computers easier. The semantic web is composed of various parts of the database with very high degree of intelligence and very strong coordination[1]. In the semantic web, the network can not only connect various files, but also recognize the information conveyed in the files, and embody the "intelligence" aspect of the semantic web. In other words, the semantic web can not only "understand" the words and concepts, but also understand the logical relationship between them. It is a smart network.

The Semantic Web involves many technologies, as follows:

HTML, XHTML - a markup language used to represent information;

XML, XSL, XSLT, SMIL - a markup language for the description of the content;

RDF, RDFS, XRDF - semantic description and relationship description;

DAML, OIL, OWL - the ontology language that satisfies the logic and proof requirements.

In the academic circle, Tim Berners-Lee proposed the hierarchical architecture of the semantic web, as shown in Fig. 1.

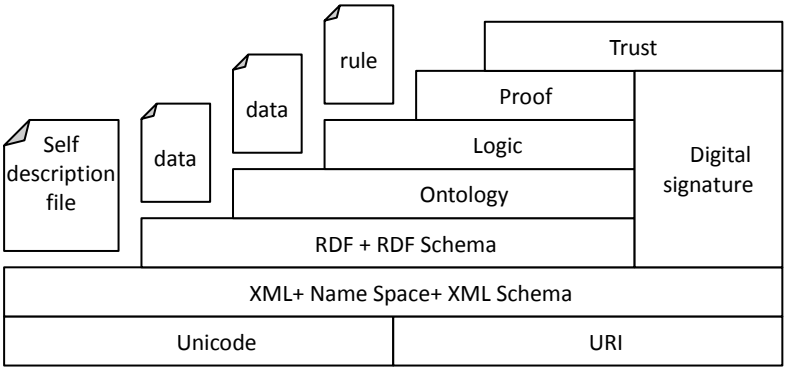

Figure 1. Semantic Network Hierarchy Architecture

3 The Web Information Retrieval Framework Based on Semantic

It is because the semantic web is "smart", so the information retrieval based on the semantic web can get more accurate results. For information retrieval based on Semantic Web, we try to construct a retrieval framework. The main idea of construction is: first of all, according to the content of information retrieval to gather related portals, and construct a web document database with the website information. Read a document from a web document database and preprocess the document. Segmentation is done by segmentation technology, and the words with high frequency in the document database are extracted into the document feature library, and the document feature index library is constructed. Use Protégé tools to create Ontology. Each time information retrieval is carried out, the key words that are input are reasoned and retrieved in the ontology library. The semantics is extended, and the extended keywords are matched and retrieved. Finally, the retrieval results are sorted and displayed according to needs. The retrieval framework structure is shown in Fig. 2.

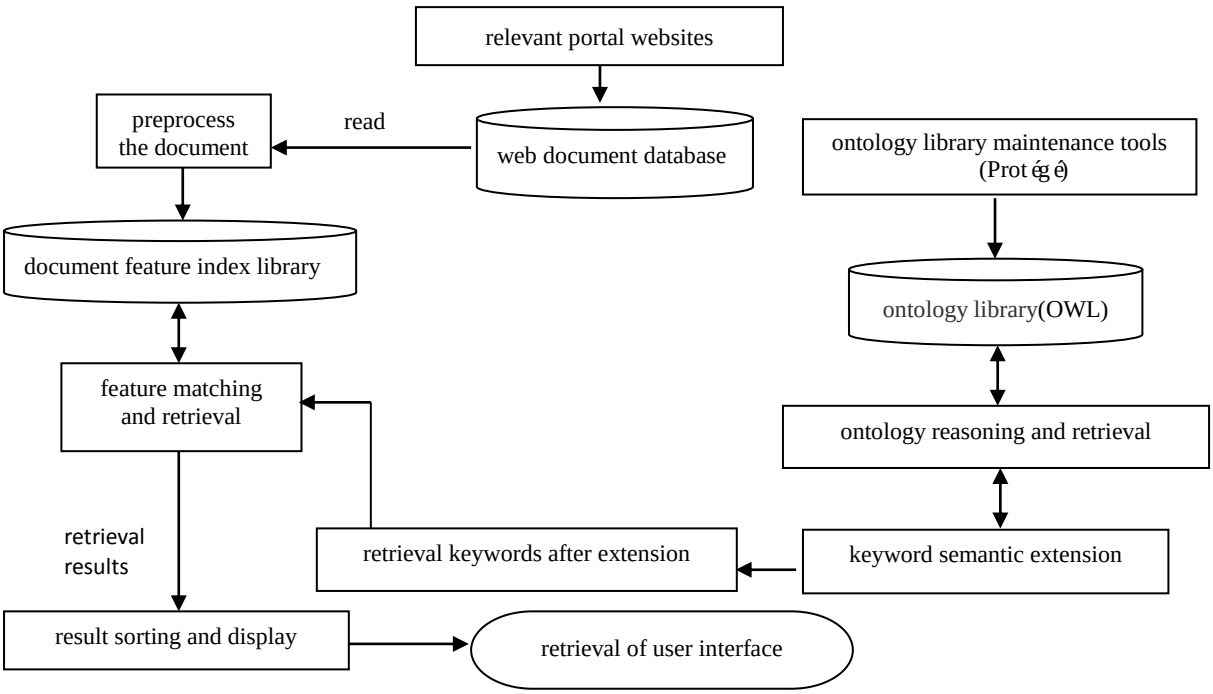

Figure 2. Information Retrieval Framework Based on Semantic Web

4 The Key Technology

4.1 Document Preprocessing

The document preprocessing mainly carries on two aspects of work, which are: extracting text information from HTML and processing Chinese word segmentation. HTML documents are made up of tags and elements. There are a lot of meaningless tags or information stored in the HTML document, which prevents us from reading the document information normally and reduces the use efficiency. Therefore, you need to preprocess HTML documents before using a document, including cleaning HTML documents, removing useless nodes and similar or same basic nodes. Document preprocessing can be implemented by ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System) based on Hidden Markov model. The main functions of ICTCLAS include Chinese word segmentation, word tagging, named entity recognition, new word recognition, and support for user dictionary.

4.2 Feature Vector Extraction

Web document retrieval technology is mainly built on the basis of statistical methods. It is necessary to set up a statistical feature model of Web documents and document similarity comparison method based on this model. The most popular text representation method currently is the vector space model VSM (Vector Space Model). The basic idea of this model is to use vectors to express text: (Wl, W2, W3,... Wn), where Wi is the weight of the ith feature item[2]. In order to be easy to describe, we give the following definition:

Document: generally refers to an article, written as D.

Feature term: refers to the basic language unit that is contained in a document and can represent the nature of the document. It is recorded as t. The document can be expressed as utility itemsets D(t1，t2，…，ti，…，tn), where ti is a ith feature item, each feature represents a dimension, 1<i<n.

Term Weight: Wij indicates the importance of feature item tj to document Di. The weight of the feature item tj in the document Di can be calculated by the TF*IDF[3] formula. The formula is follown as Eq.1.

$$w(t,d)=\frac{tf(t,d)\times\log(N/N_k+a)}{\sqrt{\sum_{t\subset d}\left[tf(t,d)\times\log(N/N_k+a)\right]^2}} \tag{1}$$

Among them, *w(t,d)* is the weight of word *t* in text *d*, *tf(t,d)* is the word frequency of word *t* appearing in text *d*, *N* is the number of documents in document collection, N_k is the number of documents for the appearance of the word *t*, *a* is an offset correction value, it can be adjusted according to the needs.

4.3 Feature Vector Matching

We can use the method of combining lexical similarity and semantic similarity to improve the accuracy of feature vector matching. The main steps are as follows:

1. Calculate the corresponding lexical similarity and semantic similarity.
2. Select the largest one from all the similarity values, and correspond to the two elements corresponding to the similarity value.
3. Delete the similarity values of the elements that have established the corresponding relationship from all the similarity values.
4. Repeat the above second and third steps until all the similarity values are deleted.
5. There is no correspondence between the element and the empty element.

Similarity: it is used to measure the similarity between a document vector and another document vector, or the similarity between document and user needs, that is to say whether a document is needed by users. The similarity of the vector model of the two documents can be measured by the cosine value of the angle between the vectors[4], as shown in Eq.2.

$$Sim(D_i, D_j) = Cos\theta = \frac{\sum_{k=1}^{n} w(t_k, D_i) \times w(t_k, D_j)}{\sqrt{\left[\sum_{k=1}^{n} w(t_k, D_i)^2\right]\left[\sum_{k=1}^{n} w(t_k, D_j)^2\right]}} \quad (2)$$

Among them, D_i and D_j are the feature vectors of the document, $w(t_k, D_i)$ represents the weight of the kth feature item in the D_i, $w(t_k, D_i)$ and $w(t_k, D_j)$ can be obtained by Eq.1.

After establishing the one-to-one correspondence of the elements according to the above algorithm, we can easily calculate the similarity between the two sets. The similarity of a set is equal to the weighted average of the similarity of its element. Because all elements are equal, so we can take all the weights to the same value, so the similarity of sets is equal to the arithmetic mean of the similarity of their element pairs.

4.4 The Dimension Reduction

Extracting feature vectors from a document in a web document library will get a multidimensional vector representation. The text content is converted into a vector method that is easily processed by mathematics[5]. In order to facilitate the mathematical operation, it is necessary to reduce the dimension of the extracted feature vectors. The dimensionality reduction scheme used in this paper is: Set a word T_i, it appears N_i times in all documents, there are a total of M documents, the average times of the word appearing in the M text documents is $F_i=N_i/M$. Calculate the F_i of all the words, take out the n items with the maximum value, determine the vector of n-dimensional as a feature vector, thus the purpose of reducing dimension is achieved[6]. After reducing the dimension of the feature vector, the extracted feature vectors are added to the document feature library by adding URL, and the records in the document feature index library are formed.

5 Experiment

The main parameters of the retrieval system are the recall and the precision. The calculation of precision is shown as Eq.3. The calculation of recall is shown as Eq.4.

$$P = \frac{a}{a+b} \times 100\% \quad (3)$$

$$R = \frac{a}{a+c} \times 100\% \quad (4)$$

Among them, *a* represents the number of relevant documents that are correctly retrieved, *b* represents the number of unrelated documents retrieved, *c* indicates the number of relevant documents that are not retrieved, that is, the number of missing documents.

In this paper, the system model is applied to the "border entry and exit personnel pre-declaration platform" for testing. In the experiment, we choose two hundred articles on entry and exit laws, regulations, government documents, conference records as the sample files, and selected "notice", "announcement", "exit", "Europe", "France", "entry", "application", "passenger", "hardware" and "piano" as the sample keywords. According to the test results, the comparison of the recall rate is drawn as shown in Fig. 3, and the comparison of the precision rate is shown in Fig. 4.

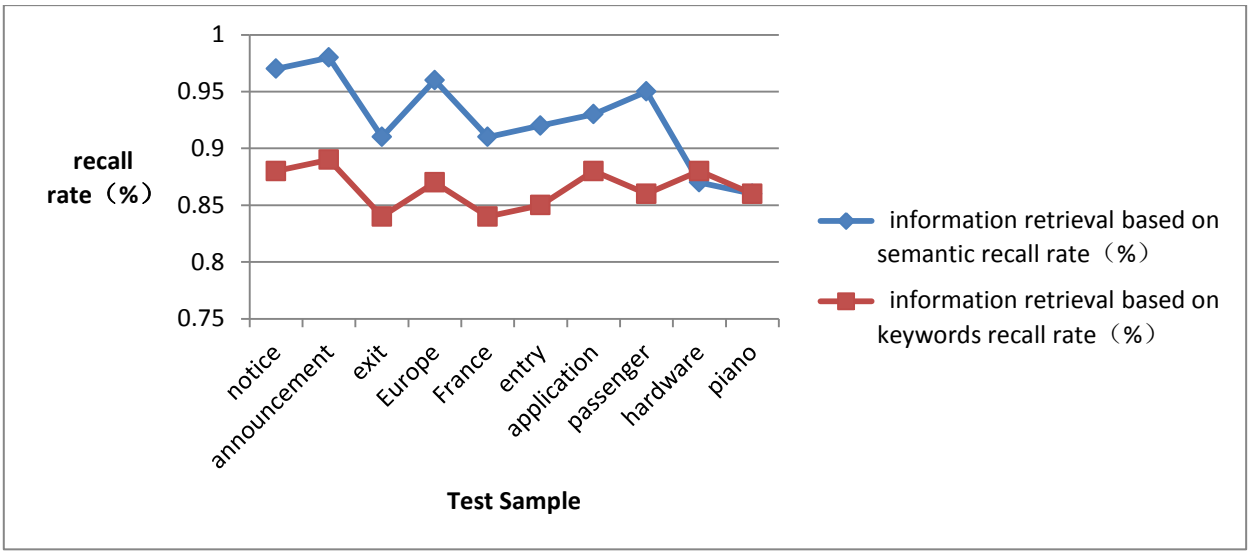

Figure 3. The Comparison of the Recall Rate

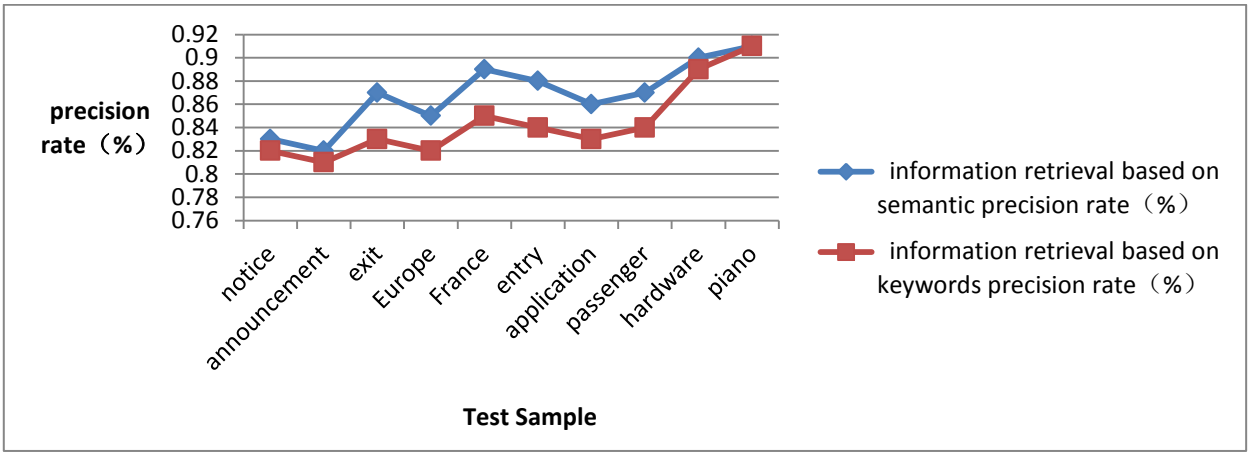

Figure 4. The Comparison of the Precision Rate

6 Conclusions

The web information retrieval technology based on semantic overcomes the limitations of traditional network retrieval technology. It supports knowledge retrieval. It can understand natural language well by using ontology and semantic based on reasoning mechanism. Compared with the traditional retrieval technology, it has good superiority. It can improve the accuracy and coverage of the retrieval, and improve the precision and recall of the retrieval results.

References

[1] Haiying Jiang. 2017. Research on the correlation of network information retrieval based on Semantic Web, Ability and Wisdom, (17): 228.
[2] Weikang Rui.2017.4. Research on the technique of text vector based on semantic, University of Science and Technology of China.
[3] Tong Wei. 2017.4. Terminology recommendation and visualization based on semantic similarity calculation, Liaocheng University.
[4] Zhanbiao Sun,Hongjun Zhang. 2017. Research on the method of word analysis based on semantic similarity, Journal of Library Science, (01):74-79.
[5] Liang Zhang. 2016. A Web service matching method based on semantic similarit, Information Science, (2):21-23.
[6] Pengfei Ji, Yuangang Li, Shengqi Lu,Kaiyu Dai. 2016. Tourism route personalized customization system based on semantic Web, Computer Engineering, (10):308-817.