The Research of Semantic Retrieval Model Based on Idiom Literary Quotation Ontology

RAN Jie
Department of Computer Science
Zhaotong University
Zhaotong, China,(0870)2132754
ztranjie@163.com

Qi Li-juan
Department of Computer Science
Zhaotong University
Zhaotong, China,(0870)2132754
524105981@qq.com

***Abstract*—Aiming at the problem that users often can not retrieve useful information or retrieved information is inaccurate, in this paper we put forward an information retrieval model based on Idiom Literary quotation Ontology. The model and its characteristics were analyzed in detail. Through a systematic in-depth analysis of users' questions, the questions mode and answers mode are proposed, it can improve the efficiency of semantic retrieval, so the precision and recall of semantic retrieval will be better guaranteed.**

Key words-Idiom Literary Quotation Ontology; lexical analysis; semantic retrieval; similarity

I. INTRODUCTION

With the rapidly growth of computer, communication, network and modern information technology, it improved information retrieval technology, at the same time it also improved hardware and software environment and facilitated the development of information retrieval theory and practice. It has brought new challenges for information retrieval. The traditional mechanical matching retrieval ways are based on keywords, it seriously fragmented words and inter-word semantic association, the retrieval processing does not contain any semantic information, leads to the results unsatisfied with users. Therefore, in the retrieval processing we should be import the semantic information of concepts, so that the retrieval processing developed from keywords matching to content matching. In order to overcome the above-mentioned shortcomings, in this intelligent search processing, ontologies play an important role. As an ontology describes intrinsic links between things, by means of ontologies, it can make the retrieve information to get better meet needs of users.

Ontology concept came from philosophy, in which it refers to the subject of existence. It will be introduced ontology concept for knowledge representation and knowledge organization in Artificial Intelligence (AI) areas, the connotation of ontology has changed. Gruber proposed definition of ontology, "An ontology is a specification of a conceptualization."(T R Gruber, 1993) Ontology is composed of a collection of related terms and links among these terms in the specific information field, its main purpose is to describe database. This paper will be based on domain ontology, researched the key technologies of ontology-based semantic retrieval, and on this basis, designed a semantic retrieval model based on Idiom Literary quotation Ontology (ILQO). In order to make clear contents of users' question and users want to know the answer, we put forward questions mode and answers mode, it can improve efficiency of semantic retrieval, it also can improve the precision and recall to a certain extent.

Ⅱ. ILQO SEMANTIC RETRIEVAL MODEL

In this section, we introduced overall design of ILQO semantic retrieval model in detail and explained the function of each module.

According to the above principle, we designed a semantic retrieval system based on ILQO, the system mainly includes several modules: user interface module, query analysis module, retrieval analysis module and data storage modules, as shown in Figure 1.

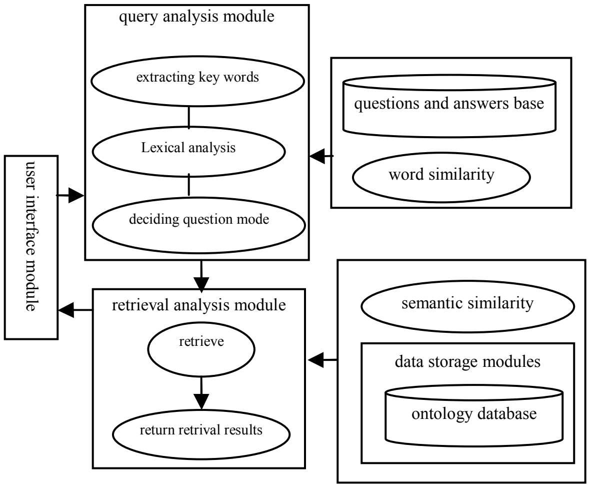

Figure 1. Ontology-based semantic retrieval system

The each module in figure 1 gets together to complete retrieval task. We introduced separately the four modules as follows:

1) User Interface Module. The user interface module provides interface to interact with the system, it can accept users' query requests, then transmits the requests to the query analysis module to deal with it, finally returns the analysis results and displays the results in the user interface.

2) Data storage modules. The module mainly stored ontology's data into relational database. We will introduce ILQO in section Ⅲ. In order to retrieve accessibility and data

structure clear, we designed six tables to store the data of ontology. These table involve concepts table, attribute table, category table, examples tables, relational table among classes and relation table among examples.

3) Query analysis module. The main role of this module is to deal with the user's question. First, it analyzed users' questions and defined content of questions, defined that users want to know the answer. Second, it put forward five kinds of questions and answers mode, the questions and answers mode is one-to-one correspondence. In section Ⅲ of the article we will introduce how to determine the mode of questions in detail. The lexical analysis of the system is ICTCLAS system. According to ILQO semantic search features, we expanded their lexicon base coreDict.dct and added related terms of ILQO.

4) Retrieve analysis module. Retrieval analysis module get keywords from query analysis module, and searched its location in the database based on the keywords and their characteristics, then retrieved ontology library, finally the searching results were returned to the user interface.

Ⅲ.KEY TECHNOLOGIES IN ILQO RETRIEVAL MODEL

In this section, we introduced the characteristics of ILQO and its establishment, and analyzed questions and answers mode, as well as we analyzed the processing of retrieval.

A. ILQO

The development of ontologies construction methods is still stayed in an immature stage, the specific design methods are also different, related mature methods include: Skeletal Methodology, Enterprise Modeling Methodology, Methtontology method and so on. These methods put forward different points of view about ontologies construction from the different angles. (Perez A.G 1999)

In the long history of Chinese culture, Idiom Literary quotation is ancient wisdom and essence of Han language. The system tries to combine Chinese traditional culture with computer technology. For the purpose of idiom semantic retrieval, we built an ILQO ontology-based skeleton method and used OWL language to descript the ontology. The ontology building processing as follows: a) To determine the application of ontology and using scope, as Idiom Literary quotation involved different dynasties, in order to reduce the size of ontology, we limited the scope of ontology between Chu and Han period. We established the corresponding ILQO based on purpose of semantic retrieval ; b) Ontology analysis, we defined the meaning of all terms in ontologies and the relationships among them. The classification is a very critical step in building ontology, we used top-down classification method, and queried various information, finally the Idiom Literary quotations of this period were divided into 11 major categories and 79 small categories. This classification method is also benefit to the future expansion of ontology library; c)Representing and coding the Domain Ontology: ILQO was represented by Protégé 3.2.1. The ontology was stored with OWL files. In short, ontology building is tested through clarity, consistency, integrity and scalability. (Doan A H 2002) We showed the ILQO fragments in Figure 2

.

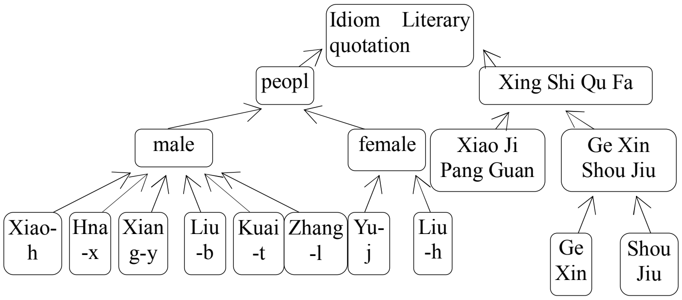

Figure 2 fragment of ILQO

B. questions mode and answers mode in ILQO

When separated words, we can acquire a combination of words. We filtered words that irrelevant with questions, and we can get a group of keywords. Next we can determine the questions mode through analyzed this group of keywords. In order to determine questions mode, we divided concepts in ILQO into three types: main concepts, instances and properties. The main concept refers the classification of ontology, which includes subclasses. The instances are individuals of ontology. On the ontology is concerned, that is, individual is individual phrase. The properties include object properties and datatype properties of ontology. We discussed the five questions modes as following:

Mode 1:There is only a main concept in question, querying result is an instance. This is a common question mode. For example: What idioms about Liu Bang?

Mode 2: There is only an instance in question, querying result is the property value of instance. This is higher frequency question mode. For example: What is the meaning of Duo Duo Yi Shan?

Mode 3:There is only an instance in question, querying result is the class of instance. For example: What are persons involved in Ba Wang Bie Ji?

Mode 4: There are two main concepts in question, querying result is the common instance in two classes. For example: What idioms about Liu Bang and Han Xin?

Mode 5:There are two main concepts in question, querying result is the relationship between two classes. This is lower frequency question mode. For example: What is relationship between Liu Bang and Lv Hou?

In practice, we found that nouns and verbs play very important roles in sentence. The nouns bear more information than verbs according to the characteristics of this system, and verbs belong to requesting words such as "ask", "please," "please answer" and so on. These verbs almost have nothing with main contents of question, so focus of questions were nouns, we only extracted nouns as keywords. In order to more accurately determine the question modes, we designed questions mode table and answers mode table. The questions mode table includes fields as follows: question_id, key1, key2(the value of this field may be null) and questionexplain. The answers mode table includes fields as follows: answer_id, question_id (this field is same as the corresponding field in questions mode table), answer_key and answerexplain. There are five records in the two tables, that is, the five modes were shown as above. The two tables get together to determine the questions mode. We extracted the first noun, the second noun and the last noun.

The steps to determine question mode as follows:

Step 1: Lexical analysis, to extract nouns in the sentence such as: noun1 + noun2 + ... + nounn;

Step 2: To determine the noun1 is which one type of concept in ontology, and match it with key1 field in the questions mode table;

Step 3: To determine the noun2 is which one type of concept in ontology, and match it with key2 field in the question mode table. If the extracted nouns are only two nouns, the noun2 is null, then skip this step, turn to step 4;

Step 4: To determine the nounn is which one type of concept in ontology, and match it with answer_key field in the answer mode table.

Step5. Integrate step 2,3,4 as above, a) If the matching is successful, then we can determine questions mode; b) If the matching is unsuccessful, then compute word similarity and match it again, this time if the matching is successful, we can determine questions mode, otherwise, the question does not belong to questions mode in the system, then the system prompts user to ask question again according to the modes in the system.

$$Sim\ (C1, C2) = 1 - \sqrt{\frac{\alpha - 1}{\alpha} \times Dist\ \ (c1, c2)}$$

At present concepts are used to measure the correlation among links of concept in the field of information retrieval, we mainly consider semantic similarity and semantic relevancy two factors from the angle of natural language. Semantic similarity is concept consistent in the sense. We compute semantic similarity through semantic distance of the concept in semantic tree. Semantic distance of concept is inversely proportional to the semantic similarity. In this paper, we compute semantic similarity by semantic distance. With the concept of Cl and C2, the semantic similarity computing method as follows:

Where Dist (C1, C2) is semantic distance between C1 and C2, α is an adjustable parameter.

Retrieve Analysis Module. The type of retrieval analysis can be divided into two types: one is precise searching, it is a particular search to a concept, such as the first three modes in questions mode; another is about the retrieval of semantic relations, this retrieval is usually two or more keywords and keywords which exists semantic relations, such as the latter two modes in questions mode.

To sum up, we give the steps of ILQO retrieval model:

Step 1: We cut the words for user's query, and obtain keywords of the query. Next we determine questions mode, and then submit them to query analysis module.

Step 2: Query analysis module analyzes the submitted question, and question is divided into two kinds of situations to deal with:

Case 1: It is a precise searching, in other words, it is a simple query. It can be directly search through ontology database;

Case 2: It is about the retrieval of semantic relations. To the retrieval, first we compute semantic similarity between the main concepts, definite the description of its semantics and understand the user's searching intention, and then transmit the semantic description to the retrieval module, finally we get searching results through ontology database.

Step 3: Sum up the above two steps, we return the searching results to user.

Ⅳ. Conclusion

In this paper, we put forward an information retrieval system based on ILQO, it can make up for defects as following: you can not find the information, or the information which found is not accurate. In this system, we proposed questions mode and answers mode, and combined similarity algorithms with them, thereby improved the efficient of query. However, natural language is flexible, so it will still affect the accuracy and integrity of understanding questions, such as how to determine the specific questions, this will be our next research work. In addition, the improvement about similarity computing is a problem what we need to be resovled later.

References

[1] (Perez A.G 1999)Perez A.G., Benjamins V. R.. Overview of Knowledge Sharing and Reuse Components: Ontologies and Problem Solving Methods[A].Stockholm V R, Benjamins B,Chandrasekaran A, eds. Proceedings of the IJCAI 99 workshop on Ontologies and Problem Solving Methods (KRR5)[C]. 1999:1-15.

[2] (T R Gruber, 1993)T R Gruber.A translation approach to portable ontology specifications. Stanford University,Tech Rep:Logic-92-1,1993.

[3] (Doan A H 2002)Doan A H, Madhavan J, Domingos P, et a1 . Learning to Map between Ontologies on the Semantic web[C]//Proceedings of the1lth International Conference on World Wide Web. New York, USA, 2002 : 662-673.