Research on Web Semantic Information Retrieval Technology Based on Ontology*

Hong Zhou and Jun Liu

School of Information Science & Technology, Beijing Wuzi University, 101149, Beijing, China
David_csharp@yahoo.com.cn, liujun@bwu.edu.cn

Abstract. To solve the problems of label clustering, bad theme relevance, etc. of information retrieval especially the semantic retrieval in the context of Chinese environment. The appropriate information retrieval model for the Chinese semantic environment is constructed through introducing ontology and class label mechanism. Based on optimization of the query retrieval submitted by ontology and label to the user and through calculation of the similarity between ontology label and member engine data, the dispatching method for member search engine database is proposed; the method appropriate for extracting this semantic information retrieval model data is proposed through improvement of the traditional STC algorithm. The research on semantic retrieval technology based on ontology and label breaks through the bottleneck of semantic search and information resource management & organization and enhances the scope and quality of the semantic information retrieval.

Keywords: ontology, semantic retrieval, class label, primitive.

1 Introduction

Information retrieval indicates the process of the user’s searching for necessary information from various information sets. It contains information storage, organization, representation, inquiry and access, etc.. Semantic retrieval indicates the retrieval in combination of information retrieval & artificial intelligence technology (AIT) and natural language technology. It analyzes the retrieval request of information object and searcher from the angle of the semantic comprehension, and it is a retrieval match mechanism based on concept and its relevance. Currently, the main technologies in realizing semantic retrieval include three aspects: natural language processing (NLP), method based on conceptual space as well as method based on ontology, etc. NLP indicates the processing of natural language to enable the computer to understand the content of the natural language. No help will be given to the information retrieval effect by adopting the processes, like simply removing stop words and taking wood root, etc. However, the consumption for processing and storage will be increased if

* Supported by Funding Project for Base Construction of Scientific Research of Beijing Municipal Commission of Education (WYJD200902), Beijing-funded project to develop talents(PYZZ090420001451).
the processes, like adopting complicated word sense disambiguation and coreference resolution, etc. The semantic retrieval method based on conceptual space indicates applying the conceptual space into the information retrieval and selecting the content related to the concept based on comprehension of the concept connotation by the conceptual space and the concept conveyed by the key words submitted by the user as the retrieval basis to expand the retrieval scope by dint of this. The traditional information retrieval mechanism can not solve the semantic problem under the context of Chinese effectively. However, ontology, as a conceptual model modeling tool that can describe information system at the layer of semantics and knowledge, focused by numerous research personnel at home and abroad since it is proposed and has been widely applied in many fields of the computer, like knowledge engineering, digital library, software reuse, information retrieval and disposal and semantic net of the heterogeneous information in the internet, etc..

2 Clustering Search Technology Based on Ontology

The main advantages of the semantic information retrieval based on ontology are represented in two points: firstly, the system eliminates the concept ambiguity of the key ambiguous words in semantic concept by dint of domain ontology through utilizing the correlation among the key words in the user’s query method, and ensures the correctness of the return to the document. Secondly, the system can better comprehend the user’s retrieval demand by virtue of domain ontology and perform corresponding reasoning according to the correlation among the key words in the user’s retrieval query method to answer the user’s questions and tap out the user’s real demand.

**Definition 2.1 (Ontology):** \( \mathcal{O} = (D, W, R) \), in which \( D \) indicates a field; \( W \) indicates the clustering of the state of relevant affairs in the field while \( R \) indicates the clustering of the conceptual relation in the field space \(<D \& W>\). The ontology refers to the description of conceptualization in certain language.

However, at present, in information retrieval technology, the retrieval result information interaction method based on “key word query + user’s voluntary browsing” brings about the bottleneck of delivery of user’s information demand. To solve this problem, the label is introduced in this paper. The function of the introduction of label refers to the better orientation between the retrieval result and user interaction.

**Definitions 2.2 Class tag:** Class tag refers to the key terms that indicating the key content of the document but were standardized by ontology techniques, primarily manifested as the search key terms input by users. It possesses three characteristics of subject terms: (1) expressed by letter symbol; (2) reflect the relation model between page data and user search from the perspective of the ontology character; (3) organize structure in the retrieval system according to the ontology model, enabling it to reflect the key content of the document. It composed an assembly concept unit of the retrieved topic. The class tag library, in fact, is a concept table collected with search meaning, which is also a conceptualized key term table for searching and description of various key themes of web page data.

The Information retrieval model based on ontology is shown in Fig 1. The search optimization mainly involves in the optimization of users’ search query. The module of tag