Automatic Semantic Retrieval and Visualization Model Based on the Integrated Ontology Library

Xiaoyue WANG†, Zewen HU, Ruijiang BAI, Yan MOU

Institute of Scientific & Technical Information, Shandong University of Technology, Zibo 255049, China

Abstract

At present, ontology, by its abundant concept semantic relations, formal share conceptualization model, unified resource description framework, is frequently applied in the expansion of query, semantic indexing of resources, the organization of search results etc. However, by using ontology and ontology-related processing technologies, carrying out an automatic semantic expansion to user query and providing automatic semantic retrieval, conceptual semantic query map of resources etc are still unclear. Therefore, in this paper, we propose an automatic semantic retrieval and visualization model based on the integrated ontology library. This model adopts ontology integration and XML database etc technologies to carry out the ontology integration and storage to the popular ontology libraries, such as WordNet, DBpedia, OpenCyc, SUMO (The Suggested Upper Merged Ontology) etc and forms a normalized integrated ontology library; then under the support of integrated ontology, by XML query language, merging and visualization, query expansion based on conceptual semantic map etc technologies makes an automatic semantic expansion to user query, and provides automatic semantic retrieval, conceptual semantic query map of resources for users. Experiment results show that our method can filter out a large amount of user query-unrelated information, not only greatly improve the precision of information retrieval system, but also satisfy users’ visual and personalized query requirement.

Keywords: Semantic Retrieval; Visualization; Ontology Integration; Ontology Storage; XML Query; Category Theory; Query Expansion; Concept Semantic Map

1. Introduction

With the rapid development of computer technology and internet, information is increasing by explosive speed. Among them, there are a huge amount of valuable information, how to carry out semantic expansion to user query and retrieval the related information that satisfy users’ information needs becomes the facing challenges in present information retrieval systems. At present, an increasing trend in information retrieval domain is the usage of ontology as a semantic means for describing the information content and expanding the user query [1]. Indeed, an ontology, as the explicit specification of domain conceptualization [2] and the conceptual model modeling tool, has the favorable conceptual semantic structure and semantic expression ability, and can carry out reasoning and retrieval according to some rules, is very suitable for the intelligent information retrieval based on semantics, thus becomes a new bright spot in the semantic information retrieval domain [3-4].

Related work. Ref. [5] gives ontology’s definition that is a “formal, explicit specification of a shared conceptualization”, where a “conceptualization” is an abstract model of some phenomenon of the world

*Corresponding author.

Email addresses: wangxy@sdut.edu.cn (Xiaoyue WANG), huzewen915@163.com (Zewen HU).
which identifies the relevant concepts (or entities) and relationships among the concepts of the phenomenon. Again, “shared” means that an ontology captures consensual knowledge, whereas “formal” refers to the fact that an ontology should be machine-understandable; For ontology integration[6][7], there are two comparatively promising methods that are the method based on category theory[8] and the method based on FCA(formal concept analysis)[9]. Category theory and FCA are the theory of disposing structure and the relations among the structures, FCA is only applied to dispose the lightweight ontology. In this paper, we will adopt the ontology integration method based on category theory, mainly because the ontologies that this paper will merge are all heavyweight ontology, and category theory is of very good math theory basis[10], under its guidance, we can implement the automatic ontology integration to a certain extent and reduce the workload of ontology integration; For ontology storage and query, at present, there has occurred many prevalent ontology storage management systems, such as OWLim[11]. OntoEdit[12]. Protégé[13]. Sesame[14]. Jena[15] etc. Unfortunately, these systems are very hard and time consuming to store and query the heavyweight ontology automatically. So we, according to ontology languages’ characteristic of basing on XML syntax, adopt a kind of ontology storage and query method based on XML database Xindice[16] that can not only keep the concept semantic structure of ontology very well, but also greatly improve the automation degree of storing, querying and utilizing ontology, providing the basis for automatic semantic retrieval and visualization based on the integrated ontology library; For query expansion, at present, some query expansion techniques, such as user relevance feedback[17]. query expansion based on thesauri[18] or based on concept[19] etc, can provide some degree of query expansion ability. But these query expansion techniques can’t understand users’ query intention very well and are lack of rich concept semantic relations and reasoning mechanism, such as the user relevance feedback asks users to be familiar with the vocabularies of a document collection etc. So we propose a kind of automatic semantic query expansion method based on concept semantic map. This method can acquire users’ query intention very easy according to the nodes that users click in the conceptual semantic query map. More contents will be discussed detailedly in Sub-section 2.4.

The research purpose of this paper is to apply ontology to information retrieval domain, and propose an automatic semantic retrieval and visualization model based on the integrated ontology library. The model proposed in our paper overcomes the shortcomings of previous model in the following aspects: (1) it greatly improves the automation degree of utilizing ontology to implement semantic and visual retrieval by using XML database to store. query ontology; (2) it can acquire users’ query intention easier according to users’ operations in the visual query interface; (3) it can carry out conceptual similarity computation, semantic mapping and inference between the words, constraint conditions in users’ retrieval conditions and the concepts, relations, constraints of ontology according to ontology’s concept semantic relations and users’ query intention, and form the semantic query vector, making users’ queries express their information needs more clearly.

The rest of the paper is organized as follows. An automatic semantic retrieval and visualization model based on the integrated ontology library is given in Section 2. An experiment is given in Section 3. Section 4 concludes the paper and points out the directions of future work.
2. Semantic Retrieval and Visualization Model Based on the Integrated Ontology Library

According to the above works, we propose a kind of novel automatic semantic retrieval and visualization model based on the integrated ontology library (as shown in Fig.1), which is composed of the following four modules: User Query and Results Feedback Module, Ontology Integration and Storage Module, Merging and Visualization Module, Retrieval Engine Module.

2.1. User Query and Results Feedback Module

The main works of this module are to extract the keywords from users’ query sentences, and organize these keywords, forming the query sentences that accord to XPath query language format, then use this query sentences to retrieval the integrated ontology library and acquire the related ontology documents. By carrying out semantic mapping, merging and visualization between the concepts, constraint conditions in users’ retrieval conditions and the concepts, relations, constraints of ontology, construct the concept semantic map centered by users’ query terms. Then synthetically considering the domain background knowledge and users’ query intention, use query expansion method based on the concept semantic map to make a semantic expansion to users’ query and form the semantic query vector. Use this semantic query vector to retrieval the information resource library and sort the retrieval results. At last, use response, Ajax etc technologies to send the retrieval results to the client.

2.2. Ontology Integration and Storage Module

In this module, we, by using the ontology integration method based on category theory, such as Refs. [8], [20], make an integration to the prevalent ontology libraries at home and abroad, such as WordNet, DBpedia, OpenCyc, SUMO etc. According to most ontology languages’ characteristic of basing on XML syntax, use the ontology edit tools, such as Protégé, OntoEdit etc, to unify the format of ontologies and save them into the XML database Xindice, constructing an unified, comprehensive integrated ontology library.

2.3. Merging and Visualization Module

The main works of this module are as follows: (1) according to users’ query terms, uses XPath query language to retrieval the integrated ontology library automatically, and acquires the corresponding XML ontology documents; (2) uses ontology reasoning engines, such as Jena, Sesame etc to make a parsing to the acquired ontology documents; (3) makes a merging between the concepts, constraints in users’ query and the entities, concepts, relations in users’ query-related ontology documents by using semantic similarity calculation, semantic mapping, reasoning etc technologies; (4) carries out a visualization to the related entities, concepts and relations by using visualization tools such as Protégé, OntoSphere[21], Prefuse[22] etc or visualization program module, thus forming a concept semantic query map centered by users’ query terms; (5) uses response, Ajax etc technologies to send the query map to the client, users can retrieval the resources that are closely related to users’ information needs by clicking the arbitrary nodes in the query map. At the same time, this query map also provides the basis for automatic query expansion based on concept semantic map.
2.4. Retrieval Engine Module

In this module, we mainly use the query expansion method based on concept semantic map to carry out a semantic expansion to users’ query, implementing the automatic semantic retrieval. The concrete processes are as follows: (1) automatically acquires query-related concept semantic map according to users’ query terms; (2) acquires users’ query intention according to users’ operations in the conceptual semantic query map; (3) carries out a semantic association analysis between users’ query terms and the closely related terms to users’ query intention in concept semantic map, then according to the semantic relations between users’ query terms and the related terms, the length of paths among terms, confirms the similarity between users’ query terms and the related terms; (4) according to the similarity among terms, confirms the weight of every term, and sets a threshold “a”, uses the high-weight terms more than “a” to make a semantic expansion to users’ query, forming the semantic query vector. At last, we use this semantic query vector to retrieval the information resource library and sort the retrieval results.

3. Experiment

3.1. Experiment Design

We adopt Google and Yahoo as experiment platform, five original queries and their extended queries as experiment’s query expression (as shown in Table 1), the top 100 results that are retrieved from Google and Yahoo by using every query expression as experiment dataset to evaluate the performance of our proposed model. As we can’t acquire the detailed situation of dataset in Google and Yahoo, we can only use the precision and average precision to evaluate the performance of the proposed model. In this paper, we denote the precision of query results in the top 100 results as Pre@100, Pre@100=P( the relevant document number in the top 100 results | the top 100 results). We adopt AvgPre@100 as the final evaluation index of
experiment results, AvgPre@100 is the average of the precision (Pre@100) of the original queries and their extended queries.

Table 1 Original Queries and Their Extended Queries List

<table>
<thead>
<tr>
<th>Query ID</th>
<th>Original query</th>
<th>Query intention</th>
<th>Extended query</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Mouse</td>
<td>The mouse of a computer</td>
<td>Computer, PC, desktop, notebook, laptop, mouse, wireless, wired, USB, optical</td>
</tr>
<tr>
<td>2</td>
<td>Mouse</td>
<td>A kind of animal</td>
<td>Animal, rodent, mouse, rat, mice, snout, small, rounded ear, long tail</td>
</tr>
<tr>
<td>3</td>
<td>Computer</td>
<td>Computing device</td>
<td>Computer, PC, desktop, notebook, laptop</td>
</tr>
<tr>
<td>4</td>
<td>Memory</td>
<td>The memory of a computer</td>
<td>Computer, PC, desktop, notebook, laptop, hardware, storage device, memory, cache, DDR, RAM</td>
</tr>
<tr>
<td>5</td>
<td>Apple</td>
<td>A kind of fruit</td>
<td>Fruit, apple, red, green, yellow, round</td>
</tr>
</tbody>
</table>

3.2. Experiment Results

Experiment results are shown in Table 2 and Fig.2. Table 2 represents the precision (Pre@100) and average precision (AvgPre@100) of the original queries and their extended queries in Google and Yahoo. Fig.2 represents a comparison of the average precision between the original queries and their extended queries in two different experiment platforms: Google and Yahoo. We can see from Table 2 and Fig.2 that the precision and average precision of the extended queries with our expansion method have much higher improvement than the original queries. For example, we want to search a kind of fruit: apple, generally, we input query keyword: apple, at this time, you can find out that most of retrieval results are the electronic products about apple, the precision of retrieval results are only 2% and 3% in Google and Yahoo, but the precision with its extended query can reach 91% and 94%.

Table 2 The Precision and Average Precision of the Original Queries and Their Extended Queries

<table>
<thead>
<tr>
<th>Query ID</th>
<th>Original query</th>
<th>Extended query</th>
<th>Pre@100 In Google</th>
<th>Pre@100 in Yahoo</th>
<th>Pre@100 in Google</th>
<th>Pre@100 in Yahoo</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Mouse</td>
<td></td>
<td>60%</td>
<td>65%</td>
<td>98%</td>
<td>98%</td>
</tr>
<tr>
<td>2</td>
<td>Mouse</td>
<td></td>
<td>12%</td>
<td>22%</td>
<td>97%</td>
<td>99%</td>
</tr>
<tr>
<td>3</td>
<td>Computer</td>
<td></td>
<td>14%</td>
<td>50%</td>
<td>93%</td>
<td>98%</td>
</tr>
<tr>
<td>4</td>
<td>Memory</td>
<td></td>
<td>30%</td>
<td>44%</td>
<td>94%</td>
<td>96%</td>
</tr>
<tr>
<td>5</td>
<td>Apple</td>
<td></td>
<td>2%</td>
<td>3%</td>
<td>91%</td>
<td>94%</td>
</tr>
<tr>
<td>AvgPre@100</td>
<td>Original query</td>
<td></td>
<td>23.6%</td>
<td>36.8%</td>
<td>94.6%</td>
<td>97%</td>
</tr>
</tbody>
</table>
4. Conclusions and Future Work

In this paper, we apply ontology to information retrieval domain and propose an automatic semantic retrieval and visualization model based on the integrated ontology library. Our approach is to carry out the ontology integration and storage by using category theory and XML database Xindice, constructing the integrated ontology library. By XPath query language automatically retrieval the integrated ontology library and acquire users’ query-related ontology documents, carry out a merging and visualization between the words, constraint conditions in users’ retrieval conditions and the concepts, relations, constraints in users’ query-related ontology documents by the merging and visualization module, thus forming a concept semantic query map centered by users’ query terms. Then by automatic query expansion method based on concept semantic map carry out a semantic expansion to user query and retrieval the information resource library. At last, sort the retrieval results and send the sorted results and the concept semantic query map to the client by using response, Ajax etc technologies.

By this method, we can not only implement the automatic semantic retrieval and provide the visual concept semantic query map for users, but also keep the higher query relevance, improve the precision of user retrieval and satisfy users’ visual, personalized information retrieval demand, but this kind of method can make the recall of user retrieval results have some degree of decrease.

Our future works are to carry out the more detailed research to this model on deeper level and apply it to the concrete retrieval domain, expecting to produce the tremendous economic and social benefits.

Acknowledgement

This work is supported by Natural Science Foundation of Shandong Province (No. 2009ZRB02141). The authors are very grateful for the anonymous reviewers who made constructive comments.

References