Ontology based Domain Resource Semantic Retrieval Model

Shiliu Wang
Editorial Department of Journal
Jiaying University
Meizhou, China
shliuwang@126.com

Gongjie Zhang
Department of Computer Science and Technology
Jiangsu Normal University
Xuzhou, China
zhanggongjie@126.com

***Abstract*—Retrieval technology has been widely used, but most of the current retrieval models are based on the logic matching of characters without considering user's query requirements and objectives in semantic level, which makes the retrieval results deviate from the retrieval intention of users. Based on the knowledge organization ontology, a semantic retrieval model is proposed. The proposed model abstracts semantic vectors in the form of concept and attributes, and establishes formulas for semantic matching. Based on the proposed model, experiments are performed, and the feasibility and effectiveness are proved by the experimental results.**

Keywords-Ontology; Semantic Retrieval; Retrieval Model

I. INTRODUCTION

Information retrieval plays an increasingly important role in resources acquisition. With the trends of resources specialization, people are more concerned about the quality of retrieval results. But traditional means of retrieval based on the logic matching of characters appears its powerlessness. And users care more about retrieval results in domain and semantic level. For such a bottleneck problem, effective tool or model for resource organization is needed. The introduction of Ontology brings new vigor into information retrieval. It is because that Ontology can not only describe concepts, but also represent complex relationships between concepts. The related studies involve semantic web based semantic representation [1], defining of synonyms [2, 3], concept similarity calculation [4] [5] [6] [7] [8], and so on. But for information retrieval based on ontology, there still lacks of powerful research.

In this paper, with the aid of knowledge organization of ontology, we propose a model for domain resource semantic retrieval. In the proposed model, semantic vector, the combination of concept and attributes, is presented for the representation of domain resource or query requirement, and semantic matching method is established by the calculation of semantic similarity between vectors. The proposed model also provides a normative process for semantic retrieval. Finally, based the proposed model, the experimental study provide further evidence of its feasibility and efficiency.

The rest of this paper is structured as follows: Section 2 introduces traditional model for information retrieval, and the application of ontology in retrieval. In section 3 presents an ontology based information retrieval model. Section 4 designs an algorithm for retrieval. Experimental study is described in section 5. Finally, we draw our conclusions and present possible lines of future work in section 6.

II. ONTOLOGY AND SEMANTIC RETRIEVAL

Information retrieval model can be described as three-tuples [9], $IRM = (D, Q, R)$, where D is the set of documents for retrieval; Q is the requirements from users; and R is the mapping from the Cartesian between Q and D to the real set R, that is similarity measurement. Therefore, for a retrieval model, there are there key problems: (1) the representation of resources; (2) the presentation of the user's retrieval requirements; and (3) the similarity calculation between resource and user's requirement. The traditional retrieval model is based on the logic matching of characters, and the similarity is calculated by the comparison of key words. This leads to much useless and even wrong information in retrieval results, which seriously deviates from use's retrieval intention. These expose the ineffectiveness of the traditional retrieval model. The reason of such phenomena is that the lack of effective representation method for resources makes the ignorance of users' domain and semantic requirements.

Ontology as a model or tool of the knowledge representation, "defines the basic terms and relations comprising the vocabulary of a topic area, as well as the rules for combining terms and relations to define extensions to the vocabulary" [10]. Ontology can express not only concepts, but also complex relations between concepts, and provides clear semantic concepts and relations within a sharing range. Therefore, a uniform cognition can be achieved, and the understanding of domain knowledge can be enhanced. The retrieval model based on ontology will realize semantic matching, instead of the logic matching between characters in traditional model. And such a model will concern more about the semantic process and domain knowledge, and will accurately express semantic information of resources users' query requirements..

III. THE PROPOSED RETRIEVAL MODEL

Based on traditional retrieval model, a semantic retrieval model is established as shown in Fig. 1. In the proposed model, domain resources (text, document, etc.) and users' query requirements are represented in the form of semantic vector, which is the combination of concept and its attributes. Our model includes resource processing, vector abstraction, ontology analyzing, user's query representation, retrieval operation, etc. In Fig. 1, process is presented by rounded

Published by Atlantis Press, Paris, France.

rectangle, flow of resources process is shown with solid line, retrieval flow is presented by dotted line, and retrieval result is expressed by rectangle.

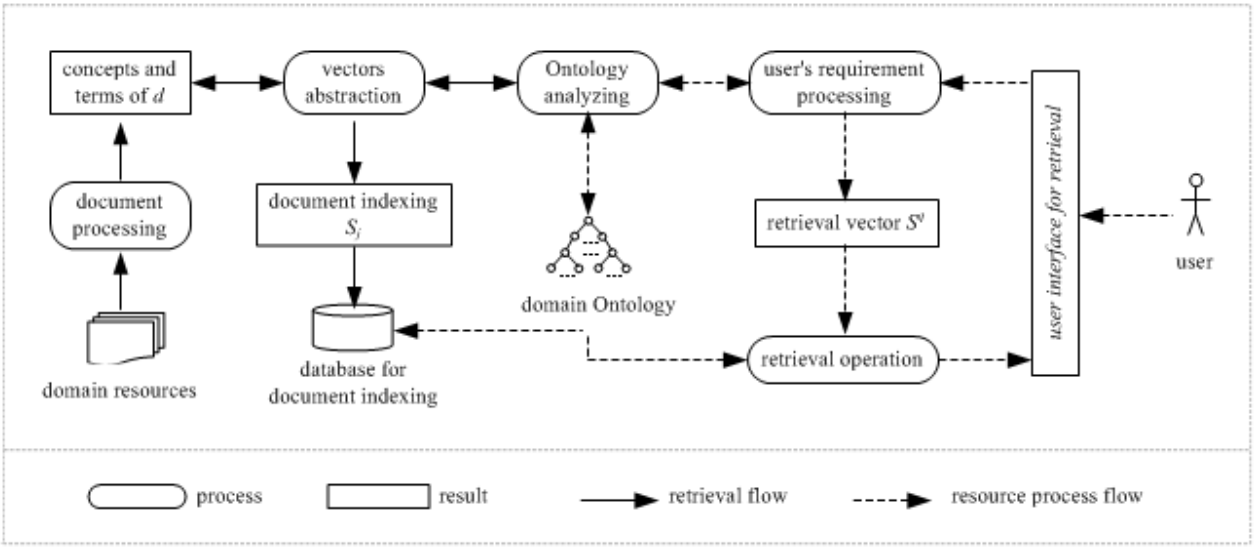


Figure 1. Domain Resources Semantic Retrieval Model.

Suppose that $\boldsymbol{D}$ is the set of domain resources, element d (for example, a document) is an element of $\boldsymbol{D}$, which is $d \in \boldsymbol{D}$, the steps of retrieval are as follows:

1) *Document processing.* Core vocabulary set from a domain resource is abstracted. The mentioned resource can be a document, or document segment, denoted as d. The processes of resource include word segmentation, annotation, purification, etc. If the resource is a web document, the HTML (Hyper Text Markup Language) tags should be removed.

2) *Semantic vector abstraction.* The abstracted core vocabulary and terms from d are organized in the form of semantic vector or vectors. According to the core vocabulary of d, first candidate concepts set $\boldsymbol{C}$ is established by ontology analyzing, the element $c_j \in \boldsymbol{C}$ is a candidate concept of document d by mapping to the concepts in ontology, and attributes with values (p_{i0}, p_{i1}, …, p_{in-1}) can be obtained by the corresponding abstracted concept c_j and terms form d. Therefore, semantic vectors of d are generated as $\boldsymbol{S}_i$=(c_i, p_{i0}, p_{i1},…, p_{in-1}), i=0, 1, 2, …, and stored in the relational database.

3) *Ontology analyzing.* This provides service for semantic abstraction and concepts mapping. The domain Ontology can be modeled in RDF (Resource Description Framework), RDFS(RDF Schema), OWL(Web Ontology Language), or relational database. By analyzing, the structure of domain ontology can be understood by machine (computer); so, the concepts and relations between concepts can be identified, then the attributes of concepts and values of attributes also can be recognized.

4) *User's requirement processing.* User's query requirement is represented in the form of semantic vector. With the aid of ontology analyzing, concept and attributes from user's query phrase are abstracted and combined into query vector $\boldsymbol{S}^q_k$=(c_k, p_{k0}, p_{k1},…, p_{km-1}), k=0,1,2… .

5) *Retrieval operation.* According to user's query vector $\boldsymbol{S}^q_k$, retrieval results by calculation of semantic similarity between query vector $\boldsymbol{S}^q_k$ and the semantic vector $\boldsymbol{S}_i$ of d. The document which has certain similarity with the query vector will be put into the result set. And the retrieval results ranked by the similarity will be presented to user.

User's query requirement is submitted through user interface, and query vector is abstracted by analyzing and mapping. If there is no concept abstracted from the query phrase, certain means of guiding should be presented on the user interface to make user clearly express their query intention. The outputs of the retrieval results are also presented on user interface according to the similarity. The indexed resources by semantic vectors can be stored in format of RDF, RDFS, OWL, and even in the relational database. The Ontology analyzing can be realized by jena[11] component from HP Labs, which can not only analyze RDF, RDFS, and OWL, but also the database.

IV. RETRIEVAL ALGORITHM

We design a simple algorithm for retrieval, as a formalization of the retrieval procedure.

Algorithm 1. Semantic Retrieval

Input: *Vectors abstracted from user's query requirements.*

$\boldsymbol{S^q_k}$=(c_k, p_{k0}, p_{k1}, …, p_{km-1}) // k=0, 1, 2, … m=$\|c_k\|$

Procedure:

```
IR=Ø
for(d ∈ D){
    if (sim(S^q_k, S_i) ⩾ ξ){
      IR=IR ∪ { d }
    }
}
```

Output: Retrieval results ordered by semantic similarity.

Where IR is retrieval result set, and $sim(\boldsymbol{S}^q_k, \boldsymbol{S}_i)$ is a function for similarity calculation.

The calculation of semantic similarity plays very important role in retrieval, for example, the extension the semantic range, the basis of semantic matching, and the basis for the ranking of retrieval results. The semantic vector includes concept and attributes, which play different roles and are with different natures, therefore, in similarity calculation the two parts should be treated respectively. The semantic similarity of query vector and indexing vector can be calculated as:

$$sim(S^q{}_k, S_i) = \delta \times sim_1 + \zeta \times sim_2 \quad (1)$$

In equation (1), sim_1 is the concept similarity, and sim_2 represents the attribute similarity. δ and ζ are weights of concept similarity and attribute similarity respectively, and satisfy the equitation: $\delta + \zeta = 1.00$. In the ideal condition, the attribute values could identify a concept or its instance, and a concept or instance defines the attributes or values of itself, therefore, the value assigned to δ and ζ are respectively 0.50.

Concept similarity sim_1 has very close relationship with the distance between concepts, so it is defined as equation (2), of which, min(dis(c_k, c_i)) represents the nearest distance between c_k and c_i. sim_1 decreases with distance, when c_k and c_i from the same concept, dis(c_k, c_i) is zero, and sim_1 equals 1.

$$sim_1 = \frac{1}{\min(\mathrm{dis}(c_k, c_i)) + 1} \quad (2)$$

Attribute similarity sim_2 is calculated by the ratio of common attribute values in number to the number of attributes in query vector，defined as equation (3).

$$sim_2 = \frac{\{p_{k0}, p_{k1}, \ldots, p_{km-1}\} \cap \{p_{i0}, p_{i1}, \ldots, p_{in-1}\}}{m} \quad (3)$$

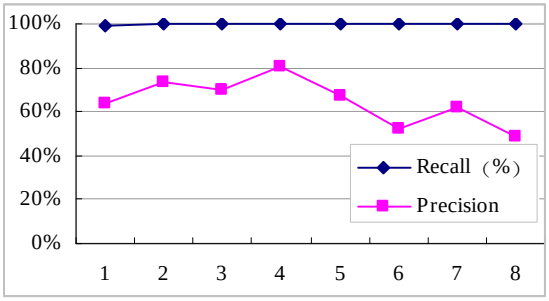
In the above equations, $sim_1 \in [0,1.00]$, $sim_2 \in [0,1.00]$, and $\delta = \zeta = 0.50$, therefore $sim(\boldsymbol{S}^q{}_k, \boldsymbol{S}_i) \in [0,1.00]$.

V. EXPERIMENT

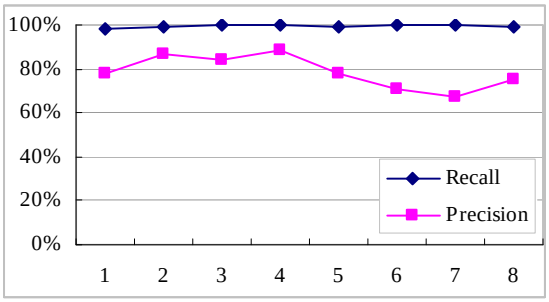
According to Computing Curricula 2001 (CC2001), by collecting and organizing related concepts and terms, we established domain ontology, establish the system structure for Programming Fundamentals (PF), and further detail the Data Structures (PF3). Based on the established ontology, domain resources are gathered from the internet. The gathered resources are structured, semi-structured or unstructured documents, even document fragments, but with clear subjects; and most of the documents are HTML pages after purification. By lexical analysis system ICTCLAS[12], terms and concepts are abstracted, and stored in relational database after indexing. For each document or document fragment, only one semantic vector is abstracted.

Experiment is performed based on the following environments: dual CUP running at 2.79GHz with 2GB ROM, Microsoft Windows XP SP3, Microsoft SQL Server 2000, and ICTCLAS 3 etc.

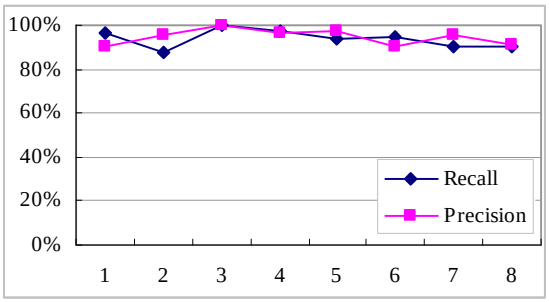
To validate the feasibility and effectiveness of the proposed model, we conducted 4 experiments with different ξ, and carried out 8 times with different query requirements for each experiment. We evaluate the retrieval results by the traditional metric: precision and recall. Precision is the fraction of retrieved instances that are relevant, while recall is the fraction of relevant instances that are retrieved. We can calculate the precision by retrieval relevant documents in the proportion of retrieval documents, and calculates the recall by retrieval relevant documents in the proportion of relevant documents.



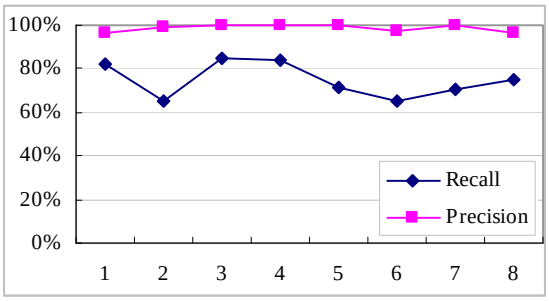
(a) $\xi = 0.40$



(b) $\xi = 0.50$



(c) $\xi = 0.60$



(d) $\xi = 0.70$

Figure 2. The precision and recall with different ξ

Fig. 2 shows the experimental results with different ξ. When ξ is 0.40 as shown in (a), it achieves the highest recall but the lowest precision among the four experiments. In (b)

the recall is slightly improved when ξ equals 0.50. Higher precision and recall are obtained in (c) as 0.60 is assigned to ξ. Comparing with (a), (c) achieves the highest precision but lowest recall, which seems to be a reverse of (a). On the whole, from Fig. 2 it can be clearly seen that the precision increase with the growth of threshold ξ, but on the contrary, the recall are decreased. But more important is that the retrieval results are based on semantic matching instead of logic matching. It avoids such a phenomenon that so many records in the retrieval result are not in accord with the retrieval intention of the user.

TABLE I. TABLE 1 AVERAGE RECALL AND PRECISION BY DIFFERENT THRESHOLD VALUES

R&S / ξ	ξ=0.40	ξ=0.50	ξ=0.60	ξ=0.70
Recall（%）	99.84	99.37	93.89	74.81
Precision（%）	64.82	78.65	94.59	98.68

Table 1 shows the average recall and precision achieved by different threshold values, which illustrates the obvious trend of the improvement of precision and the decreasing of recall. The experimental results demonstrate that it is easy to perform the retrieval only by setting some parameters, and it is efficient to get the ideal retrieval results with simple adjustment of ξ. A higher value of ξ will enhance semantic expression of resources and query requirements, which make the retrieval results more accurate in semantic level. But a lower ξ will make the similar but insufficient resources in semantics fall into retrieval results. So, threshold ξ plays a role of tradeoff between higher precision and higher recall.

VI. CONCLUSION

The introduction of Ontology brings new vitality and vigor to the information retrieval, and which is considered an efficient solution to current retrieval techniques. By the establishment of the semantic retrieval model, we find a way for semantic matching between query requirements and domain resources. Based on the proposed model, experiments are performed, the statistical data of retrieval shows that it is easy to achieve higher precision and recall. The experimental results provide further evidence that the proposed model is feasible and efficient.

Recent work has tended to focus on Ontology automatic modeling instead of the manual or Semi automatic mode. We have an interesting in the semantic enhancement, to make the semantic vector much closer to the domain resources and query requirements, and the proposed model should also be proved for practical application. We also consider the opinion mining based on the corrected Ontology. We hope in the near future semantic retrieval will be put into practice.

ACKNOWLEDGMENT

This is the science and technology project of Guangdong Province in China in 2012. Project number:2012B060200013

REFERENCES

[1] J. Köhler, S. Philippi, et al.. Ontology based text indexing and querying for the semantic web. Knowledge-Based Systems, 2006, 19(8), pp. 744-754.

[2] S. Yang, Y. Jia. Building method on domain ontology in not growthful knowledge system. Computer Engineering and Applications ,2008, 44(24), pp. 153-155.

[3] Y. Chen, J. He. Ontology-driven extracting of semi-structure web biological data. Computer Engineering, 2006, 32(5), pp. 192-194.

[4] B. Shi, J. Yan, P. Wang, et al. Ontology-based measure of semantic similarity between concepts. Computer Engineering, 2009, 35(19), pp. 83-85.

[5] S. Chen, J. Wu. Ontology-based concept similarity computation and its application. Microelectronics and Computer, 2008, 25(12), pp. 96-99.

[6] D. Xu, C. Zheng, et. al.. Concept semantic similarity research based on SUMO. Journal of Computer Applications, 2006, 26(1), pp. 180-183.

[7] L. Zheng, G. Li, et. al.. The computation of conceptual similarity in ontology. Computer Engineering and Applications, 2006 (30), pp. 25-27,61.

[8] A. Formica. Ontology-based concept similarity in Formal Concept analysis[J]. Information Sciences, 2006(176), pp. 2624-2641.

[9] X. Wang, Y. Guan et. al.. Natural Language Processing. BeiJing: Tsinghua University Publishion, 2005.

[10] R. Neches, R. E. Fikes, et. al.. Enabling Technology for Knowledge Sharing. AI Magazine, 1991, 12(3), pp. 36-56.

[11] http://jena.sourceforge.net/.

[12] http://ictclas.org/.