Application of Ontology in the Web Information Retrieval

Zimeng Xing[1], Lina Wang[1, *], Wenbo Xing[2], Yongjun Ren[3], Tao Li[4] and Jinyue Xia[5]

Abstract: In this paper, the research advances of ontology and its application are reviewed firstly. With the development of ontology technology, subject-oriented web information retrieval technology combining ontology has been becoming one of the hot scientific issues. The innovative method of the semantic web technology combined with the traditional information retrieval technology is put forward, and the related algorithm based on ontology for judging the relevancy with different topics is also represented, and has proved to be effective in given experiments.

Keywords: Ontology, semantic web, topic relevance.

1 Introduction

Ontology is a subject emerging rapidly in recent years, promoting the development of computer-aided information processing and artificial intelligence. It has been grabbing more and more attention in the field of information science and technology, and its importance has been embodied in many fields, such as knowledge engineering, natural language processing, information retrieval system, intelligent information integration and knowledge management, information interchange and software engineering, database design and integration, etc. At the age of network globalization, a large amount of information on the web has caused "over-loaded information", and how to retrieve information accurately and efficiently has become an urgent problem to be solved [Cheng, Xu, Tang et al. (2018); Yang, Tan and Zhang (2018)]. The application of ontology into the web has resulted in new semantic technology emerging, which can solve the problems of web information sharing, and realize the worldwide knowledge level information integration. It is definitely showing the importance of ontological research [Du, Li and Wang (2006); Zhao, Xu and Luo (2007); Jiang (2007); Maedche and Staab (2002); Gruber (1995)].

[1] School of Electronics and Information Engineering, Nanjing University of Information Science and Technology, Nanjing, 210044, China.

[2] School of Criminal Science and Technology, Nanjing Forest Police College, Nanjing, 210023, China.

[3] School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing, 210044, China.

[4] School of Electronics and Information Engineering, Nanjing University of Information Science and Technology, Nanjing, 210044, China.

[5] International Business Machines Corporation (IBM), New York, USA.

[*] Corresponding Author: Lina Wang. Email: wangllna@163.com.

In fact, ontology is also an important semantic technology, conceptually describing the concept of objective existence and its mutual relationship. As the knowledge base, all kinds of concepts and their relationship are expressed in ontology as the set of concept definition. So, Ontology is regarded as a complex knowledge network [Wang (2003); Lee (2006); Deng, Tang, Zhang et al. (2002); Xiao and Shao (2003)].

In the field of information retrieval and integration, $(Onto)^2$Agent, Ontobroker and SKC are three well-known projects in the world. Each of these three projects has its own emphasis in application. $(Onto)^2$Agent is designed to help users retrieve existed ontology on Internet. It mainly uses reference ontology, i.e., ontologies are built based on existing ontologies on the Internet, to preserve the metadata of different ontologies. Ontobroker is to help users retrieve web pages and collect information which is interested and needed for the specific users. The goal of SKC is to solve the problem of semantic heterogeneity of information systems and realize the interoperability among heterogeneous autonomous systems. This project is to cater for building an algebraic system to realize the interoperability among ontologies as well as the interoperability among heterogeneous systems.

Many scholars also have applied ontology into information retrieval. Wu et al. [Wu, Jiao, Tian et al. (2001)] put forward an information retrieval server based on ontology and multi-agent. The system uses ontology to assist intelligent agents to classify information sources on the network and normalize the pattern of user information retrieval. As it only provides retrieval results in the user-concerned fields, the retrieval accuracy is relatively higher. The content-based information retrieval system is proposed by Wan et al. [Wan and Teng (2003)], using ontology to fill and expand the retrieval conditions, and applying document analyzer to filter the retrieval documents, thus improving the precision. Ontology is regarded as the core part of information retrieval system. By constructing standardized domain ontology, the method proposed is to introduce knowledge representation and its processing into Internet information processing system, which provides a unified semantic model for semi-structured data and relational database on the Internet [Xu, Zhang and Chen (2001)].

2 Methods of constructing ontology

Considering the different purpose, many rules of constructing ontologies have been proposed with their respective process. The most influential set of rules is 5-rule proposed by Gruber [Gruber (1995)], which is represented as follows:

- Rule 1: Explicitness and objectivity: ontology should give explicit and objective semantic definitions of the referred terms in natural language.
- Rule 2: Completeness: the definition given is to completely express the meaning of the term described.
- Rule 3: Conformance: the inference derived from the terms is compatible with the meaning of the term without any contradiction.
- Rule 4: Maximum monotonic scalability: adding generic or specialized terms to ontology requires no modification of the existed content.
- Rule 5: Minimum commitment: minimize constraints on modeling objects.

The method of building ontology based on formation rules:

Step 1: Determining the purpose and scope of ontology application. It's the first step to set the research scope and the goal and establish the domain ontology or process ontology.

Step 2: Ontology analysis. Defining the meaning of every term of ontology and their relationship requires the participation of domain experts. The deeper the experts understand the scope, the more perfect the ontology will be.

Step 3: The preliminary design of ontology. Refining and optimizing the existed ontology.

Step 4: Ontology representation. In general, ontology is represented by semantic models.

Step 5: Ontology testing.

Step 6: The establishment of ontology. The established ontology should be checked according to the above criteria, and the documents conformed to the rules would be stored, otherwise, transferred to ontology analysis stage.

3 The basic structure of Ontology

The basic design ideas of information retrieval based on ontology can be summarized as follows.

Step 1: Establishing ontology in related fields with the help of domain experts.

Step 2: Collecting data from information sources, and refer to the established ontology, the collected data are stored in the metadata base (relational database, knowledge base) in a prescribed format.

Step 3: After receiving the query information from user retrieval interface, the query converter converts it into a specified format and matches the qualified data set from the metadata based on ontology.

Step 4: The retrieval results are customized and returned to users. If the retrieval system does not need too rigorous reasoning ability, ontology can be expressed and stored in the form of concept maps, and the data can be stored in a general relational database. Graph matching technology is used to complete the information retrieval. The description language (such as Loom, Ontulingua, etc.) will be used to represent ontology with the requirement of rigorous reasoning. The data is stored in the knowledge base, and the function of logical reasoning in description language will realize the retrieval. Because ontology can refine and present concept semantics through relationships of concepts, it can improve retrieval efficiency.

Based on ontology, an intelligent information retrieval prototype system is constructed and consists of the following modules: Ontoserver, user requirements building module, data processing module, document analyzer. Its relation-structure is demonstrated in Fig. 1.

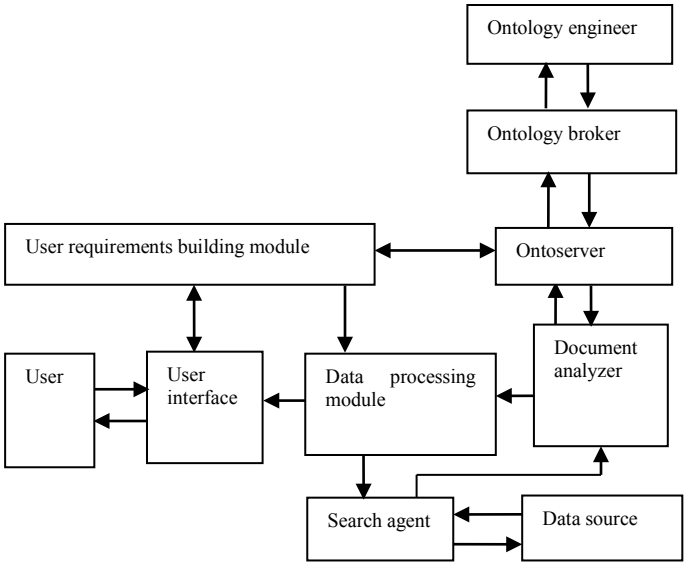

Figure 1: Structure of information retrieval prototype system

3.1 Ontoserver

Ontology, as metadata schemes, provides a controllable dictionary of concepts, in which each concept is definitely defined with understandable semantics. Ontology avails users to concisely communicate with machines supporting semantic exchange. Therefore, establishing a complete and consistent domain ontology library is the fundamental to achieve the above mentioned goals. In the construction process, the following problems should be solved:

a. Developing terminology and relational databases in specific field with expert and linguist. b. There are structural differences in the expression of the same event or process due to the differences of region and habits. Therefore, it is necessary to establish a standard structure to regulate data input and storage. c. Because of the inherent dynamic nature of language, ontology library needs to be constantly developed and updated. Therefore, an ontology broker is set up to expand the ontology database. If the query requests submitted by users with unknown terms, the ontology broker would inform engineer to decide whether new concepts and its relationship are necessary to be added in the ontology database and mapped to the semantic network graph. Ontology and other data are stored on Ontoserver in the form of relational database. The schema information describes the structural pattern of ontology. Ontology, terminology and the relationship of terms can be stored in tables. As concepts and conceptual relationships can be expressed in the form of semantic network graphs, there must be algorithms for deducing nodes (concepts) and arcs (conceptual relationships) on the server, which can provide a method of deductive access to the current data.

3.2 Building module of user requirement

It's fundamental for realizing intelligent retrieval to fully understand the implication of data. Because of the inherent flexibility of natural language, users have to be guided to express their interests, which are referring to standardize natural language. Usually, a query request from the user is a simple word or phrase. The query request will be semantically processed after being accepted by system.

With the user interface as the intermediary, the implementation steps are described as follows: a. While users input entries, the user requirement building module interacts with the Ontoserver, decomposes and matches the entries with existed ontologies in the ontology library, and deduces them in the semantic network graph according to certain algorithms. Returns all relevant conceptual result sets to the current entry. b. The user manually selects the information from returned set of results, determining the retrieval range, and then submitting the selected query request to the data processing module.

3.3 Data processing module

After receiving the retrieval request, the data processing module matches request with the keywords from the preprocessed keywords set stored in the data source via the search agent. Then the matched documents are transferred into the document analyzer for further analysis and filtering. Finally, the filtered documents are back to the data processing module and stored in the information base. When the same search request encounters again, there is no need to repeat this process, and the results will be available for the user to directly get the satisfied feedback through the user interface.

3.4 Document analyzer

This is the core component of the whole prototype system. It includes free documents and semi-structured documents. Document analyzer adopts different analysis strategies for documents with different structures. For free documents, due to the huge amount of information, complete analysis is difficult to achieve. Therefore, it is necessary to use some methods to extract key words from documents. For semi-structured documents, XML is a language that can better describe structured data. XML can design meaningful tags to facilitate data exchange and information retrieval between heterogeneous systems.

4 Information retrieval for topic relevance based on ontology

Web general search engine cannot provide users with specific topic information very well, and keyword-based technology also makes the search engine lack of semantic support. In order to solve this problem, focused crawlers technology is emerging, which is selectively getting the web pages most relevant to a given topic [Zhou and Lin (2005)]. There are a lot of researches on the description or definition of the crawler technology, the analysis and filtering of the web page or data, and the related algorithm and implementation technology of the URL search control strategy. Topic-focused crawlers require that the objects they search are as consistent as their search objectives in the process of searching resources on the web. Generally speaking, there are two main ways to compute this consistency: One is link-based manner, and the other is similarity-based

method or content-based method. Based on similarity computation, vector space models are generally used. This method is relatively simple. Modeling and computing based on keyword vector space can not accurately reflect the semantic information of web documents. Ontology, which gives explicit formal descriptions of specific domain concepts and terms, not only provides a basis for standardized resource descriptions, but also guarantees a more accurate search for information [Pan and Wu (2003); Tijerino and Sanati (2005); Zhang, Ma and Song (2005); Zhu and Sun (2013)]. This study proposes an ontology-based method for computing topic correlation.

4.1 Keyword weighting

Before calculating the correlation degree, we use ontology to weight the keywords. The weight of the keyword denotes the ability of the keyword to display the topic. In other words, the closer the keyword is relevant to the topic, the higher the weight of the keyword to the topic is, and vice versa. This study defines the weight of keywords according to the position of words in the ontology. Firstly, the core vocabulary set Core within the scope of the topic is given, and then the weights of other keywords are determined according to the semantic distance from the core vocabulary.

Core refers to the collection of core vocabulary of topics. The element in the collection is the core vocabulary of one topic, and the weight value specified by Core is 1.0. The synonyms of core words play an equally important role in expressing information. Therefore, the weight value of synonyms is also set 1.0. It has been proved by the previous studies that the longer the distance between the two words is, the lower their similarity is. The distance refers to the number of edges connecting the shortest path between the two concepts. The correspondence principle is defined as follows:

(1) If the distance between two words is 0, their similarity value is 1. It is deduced that the distance between the word and its synonyms is 0 which is also denoting the distance of the word itself.

(2) If the distance between two words is infinite, the similarity is regarded as 0. In fact, the vocabulary of a finite set is limited, so the maximum distance between two words can be calculated and expressed. The infinity just indicates a trend theoretically. According to the distance between words and core vocabulary sets, the weights of non-core words are defined as follows:

$\mathrm{W}(\overline{\mathrm{Core}})$ denotes the weight of non-core words $\overline{Core}$.

$$W(\overline{Core}) = \frac{\alpha}{\alpha + Dis(Core, \overline{Core})} \tag{1}$$

α is an adjustable parameter. As the distance between a word and the words in the core vocabulary set may not be sole. Therefore, $\forall x, x \in \overline{Core}, y \in Core$, the weights of the word x in the collection of non-core vocabulary are weighted as follows:

$$W_X = \frac{\alpha}{\alpha + Min(Dis(y, x))} \tag{2}$$

4.2 Relevance calculation

Relevance calculation refers to the data captured from the web after pre-processing will be transmitted to the calculation module to measure whether the captured content is related to the defined content by the topic and how much the correlation is. Studies have shown that the content in the web page plays a different role in describing the specific theme. For example, the content of the *title* usually describes the main content of the current page, and bold and italics in the page are likely to indicate what the author wants to highlight. In order to highlight this feature and not to excessively enlarge the weights range, the page content is divided into two parts--the title and the body, and we give higher weights to the content in the title when processing. The similarity between pages and themes is

$$sim(D, p) = \lambda_T sim(D,T) + \lambda_B sim(D, B) \tag{3}$$

In formula (3), D defines the theme content, *p* represents a page. B, T, λ represent the body of the page, page title and weight respectively.

$\lambda_B + \lambda_T = 1.0$, and $0 < \lambda_B, \lambda_T < 1.0$.

Further decompose the upper form into this mode:

$$sim(D, p) = \lambda_T \sum_{t \in T} sim(D,t) + \lambda_B \sum_{b \in B} sim(D,b) = \lambda_T \sum_{t \in T} W_t + \lambda_B \sum_{b \in B} W_b \tag{4}$$

The relevance between the vocabulary *t* in the title and the theme is directly represented by the weight of *t* in the ontology. Similarly, the relevance between the body word *b* and the topic is expressed directly by the weight of *b* in the ontology, and the similarity after normalization process is presented as follows:

$$sim(D, p) = \frac{\lambda_T \sum_{t \in T} W_t + \lambda_B \sum_{b \in B} W_b}{\lambda_T N_T + \lambda_B N_B} \tag{5}$$

N_T and N_B represent number of keywords in the title and the body after word segmentation in the page. Using the formula (5), the best case is that the weight of every keyword in the page is 1.0, and $sim(D, p) = 1.0$.

4.3 Topic relevance estimation algorithm

According to the formulas in Section 4.2, topic relevance estimation algorithm is as follows:

Algorithm 1 Topic Relevance Estimation Algorithm

Input: Domain ontology, web pages after preprocessing and adjust parameter λ (usually 0.7 to 0.9 based on the experience).

Output: Conclusion of whether the page is relevant or not.

1: weight of initialize page W, weight of title W_t and weight of text W_b;

2: while (there are still unprocessed words in page) // Vocabulary has been preprocessed.

if (title *t* is existed)

for(each word *y* in title *t*)
if(*y* exists in ontology)
$W_t += W_y$;// Modify title weight, W_y is the weight of vocabulary *y*
End for
if (text *b* is existed)
for(each word *z* in text *b*)
if(*z* exists in ontology)
$W_b += W_z$; // Modify weight of text, W_z is the weight of vocabulary *Z*.
End for
End while

3: calculate the topic relevance of the page W based on formula (5);

4: if ($W >= W_{td}$)

save the page; // W_{td} indicates whether the page is related to a threshold.

5: else discarding pages;

6: $\widetilde{W} = \sum W_x / N$; // Calculate the average weights of the first *N* pages, N represents the page that has been processed.

7: $W_{td} = \tilde{W}$; // Threshold feedback is an automatic adjustment of threshold for non manual intervention.

The method can effectively prevent the loss of useful information by extending the user's retrieval requirements semantically by interacting with ontology. Also, the retrieved original documents are filtered through document analyzer, and the retrieval requirements can be better matched.

4.4 Experiment

Referring to the classification standard of Chinese Chart Classification about computer software, the preliminary ontology of computer software subject is established. After discussion, the core vocabulary of a series of computer software topics including software, software packages, software engineering and programming has been determined. The weights of the title λ_T and the weight of the text λ_B are set at 0.6 and 0.4 respectively. The adjustable parameter α is 0.9, and some resources are selected based on the network as the test set to calculate.

The algorithm evaluation criteria can adopt the evaluation index of web information retrieval, including the precision *P* (precision), recall rate *R* (recall) and *F* value. Then,

precision(*P*)= The actual number of documents affiliated to this topic(*D*)/The number of documents collected(*A*); (6)

recall rate(*R*)= The actual number of documents affiliated to this topic(*D*)/The number of documents that actually affiliated to the topic in the test set(*T*); (7)

$F = 2 * (R*P) / (R + P)$ (8)

Based on the proposed ontology-based topic correlation algorithm (marked as method 1 in the table), the test set is subject judged and compared with the results retrieved by the keyword-based matching method (marked as method 2 in the table), as shown in Tab. 1.

Table 1: Comparison of experimental results

Parameter	Method 1	Method 2	
		software	computer software
D	19	20	1
A	20	53	7
T – *D*	1	0	20
R %	95	100	5
P %	95	37.7	14.3
F %	95	54.6	7.4

From the data in Tab. 1, it is revealed that the recall rate of ontology-based information acquisition method is similar to that of keyword-based matching method in searching "software", but the precision and *F* value are much higher than the latter. Similarly, the method of this study is far better than the result of search keyword "computer software".

5 Conclusion

The ontology-based information retrieval system model proposed in this paper can effectively prevent the loss of useful information by extending the user's retrieval requirements semantically by interacting with ontology. In addition, the retrieved original documents are filtered through a document analyzer, well matching with user's retrieval requirements with the higher precision. Ontology is adopted to carry out topic analysis and semantic analysis in this experiment, improving the accuracy of topic discovery process through the human-computer interaction. Next, we will further study the descriptive features of vocabulary. Then we can further improve the rationality of lexical semantic relevance calculation.

Acknowledgement: This research was supported by the National Natural Science Foundation of China (NSFC) (Grant No. 61772280).

References

Cheng, J. R.; Xu, R. M.; Tang, X. Y.; Sheng, V. S.; Cai, C. T. (2018): An abnormal network flow sequence prediction approach for DDoS attacks detection in big data environment. *Computers, Materials & Continua*, vol. 55, no. 1, pp. 95-119.

Deng, Z. H.; Tang, S. W.; Zhang, M.; Yang, D. Q.; Chen, J. (2002): Overview of ontology. *Acta Scientiarum Naturalum Universitis Pekinensis*, vol. 38, no. 5, pp. 730-738.

Du, X. Y.; Li, M.; Wang, S. (2006): A survey on ontology learning research. *Journal of Software*, vol. 17, no. 9, pp. 1837-1847.

Gruber, T. R. (1995): Towards principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies*, vol. 43, pp. 907-928.

Jiang, Z. G. (2007): Information retrieval in web based on ontology. *Journal of Information*, vol. 2, no. 6, pp. 103-105.

Lee, T. B. (2006): Semantic web architecture. http://www.w3.org/2000/Talks/1206-xml2k-tbl/slide10-0.html.

Maedche, A.; Staab, S. (2002): *Ontology Learning for the Semantic Web*. Kluwer Academic Publishers, Norwell.

Pan, C. H.; Wu, G. S. (2003): Design and implementation of a topic-focused web information-gathering system. *Mini-micro Systems*, vol. 24, no. 12, pp. 2150-2154.

Tijerino, Y.; Sanati, R. (2005): Onto TEMAS: an ontology-based teaching materials search engine. *Journal of Computing Sciences in Colleges*, vol. 20, no. 4, pp. 177-182.

Wan, J.; Teng, Z. Y. (2003): Application of ontology in content-based information retrieval. *Computer Engineering*, vol. 29, no. 4, pp. 122-123,152.

Wang, J. (2003): VISION: integrated classification, thesaurus and semantic metadata concept network. *Journal of Intelligence*, no. 4, pp. 412-418.

Wu, C. G.; Jiao, W. P.; Tian, Q. J.; Shi, Z. Z. (2001): An information retrieval server based on ontology and multi-agent. *Journal of Computer Research & Development*, vol. 38, no. 6, pp. 641-647.

Xiao, Y. H; Shao, S. H. (2003): An agent-realized model of personalized internet information retrieval system based on ontology. *Microcomputer Information*, vol. 19, no. 6, pp. 77-78.

Xu, Z. N.; Zhang, W. M.; Chen, W. W. (2001): Intelligent information retrieval based on ontology. *Computer Science*, vol. 28, no. 6, pp. 21-26, 44.

Yang, K. H.; Tan, T.; Zhang, W. (2018): An evidence combination method based on DBSCAN clustering. *Computers, Materials & Continua*, vol. 57, no. 2, pp. 269-281.

Zhang, M.; Ma, S. P.; Song, R. H. (2005): DF or IDF? on the use of primary feature model for web information retrieval. *Journal of Software*, vol. 16, no. 5, pp. 1012-1020.

Zhao, J. G.; Xu, D. Z.; Luo, Q. Y. (2007): Ontology and its applications. *Journal of Sichuan University of Science & Engineering (Natural Science Edition)*, vol. 20, no. 6, pp. 102-106.

Zhou, L. Z.; Lin, L. (2005): Survey on the research of focused crawling technique. *Computer Application*, vol. 25, no. 9, pp. 1965-1969.

Zhu, Z. Y.; Sun, J. H. (2013): Improved vocabulary semantic similarity calculation based on HowNet. *Journal of Computer Applications*, vol. 33, no. 8, pp. 2276-2279, 2288.