Journal of Southeast University (English Edition) Vol. 22, No. 3, pp. 334 – 337 Sept. 2006 ISSN 1003—7985

Novel method for searching ontologies on semantic web

Yu Wei[1] Cao Jiaheng[1] Chen Junpeng[2]

([1]Computer School, Wuhan University, Wuhan 430072, China)

([2]Wuhan Research Institute of Posts and Telecommunications, Wuhan 430072, China)

Abstract: In order to solve the problem of information retrieval on the semantic web, a new semantic information retrieval (SIR) model for searching ontologies on the semantic web is proposed. First, SIR transformed domain ontologies into global ontologies. Then semantic index terms were extracted from these global ontologies. Based on semantic index terms, logical inferences can be performed and the logical views of the concept can be obtained. These logical views represent the expanded meaning of the concept. Using logical views, SIR can perform the information retrieval and inferences based on the semantic relationships in the documents, not only on the syntactic analysis of the documents. SIR can significantly enhance the recall and precision of the information retrieval by the semantic inference. Finally, the practicability of the SIR model is analyzed.

Key words: semantic index term; semantic web; ontology

Information retrieval (IR) on the web is a hot topic and attracts many researchers at present. The purpose of IR is to find useful information from the web documents set. At present, the web search engines such as google and yahoo have always performed IR based on keywords. As such, the IR model may be efficient, but it is not directly suitable for indexing and retrieval on the semantic markup. And retrieval and semantic inference cannot interact and improve together. The most important reason causing the unsatisfactory situation is that the search engine cannot understand the semantics.

In order to improve the situation, a semantic web (SW)[1] is proposed, in which the machine-understandable semantic markups are used to tag the semantics in the documents. With the development of the SW, the need to develop efficient IR, is urgent. Many models have been developed for solving the problem[2–7]. The most studied method for the IR on the SW, such as that adopted by OWLIR, MELISA and LoLaLi is using the domain ontologies and local knowledge base to perform the logic inference and conduct IR by current web search engines. This method can perform semantic inference. Since there are limitations in the domain ontologies and local knowledge base, this method may have difficulties when applied to the entire SW.

In order to improve the situation, the semantic index term for semantic inference and IR in semantic information retrieval (SIR) model are introduced.

1 Framework of SIR Model

To explore the tight integration of information retrieval and semantic inference, we propose a framework of the SIR model designed to meet the following desiderata: First, the framework must support the IR and semantic inference at the same time. Secondly, the SIR model should be easily modified to be applied to the current dominant web search engines, which can save resources and facilitate software reuse. Finally, the SIR model should be friendly to the users as well as being highly efficient.

The SIR model uses the framework shown in Fig. 1 to solve the problems above. The framework can be divided into four parts: First, the domain ontologies on the SW are transformed into global ontologies; secondly, the semantic index terms are extracted

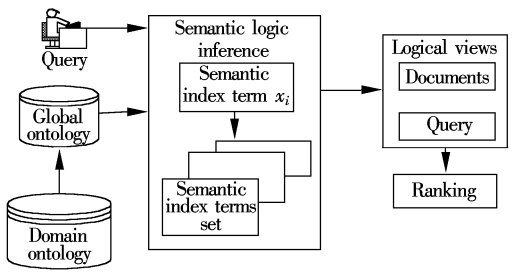

Fig. 1 Framework of the SIR model

Received 2006-05-30.

Foundation items: The National Natural Science Foundation of China (No. 60273072), the National High Technology Research and Development Program of China (863 Program) (No. 2002AA423450).

Biographies: Yu Wei (1979—), female, graduate, yuwei21@whu.edu.cn; Cao Jiaheng (corresponding author), male, professor, jhcao@whu.edu.cn.

from the global ontologies; thirdly, semantic inference is performed based on the semantic index terms; finally, the ranked list is obtained.

2 Design and Implementation of SIR

2.1 Ontology language on semantic web

By markup the metadata on the SW with the semantic terms, ontologies can be used to represent the shared conception and data sets explicitly. The familiar ontology languages are OWL, RDF(S), DAML + OIL and so on.

Three increasingly expressive sub-languages are provided by OWL: They are OWL Lite, OWL DL and OWL Full. Since OWL Lite is a suitable tradeoff between expressivity of knowledge and complexity of reasoning problems, we choose it as the standard ontology language in the SIR model. In all the semantic relations in OWL Lite, the most familiar semantic relations, the sub-, super- and equivalent relations can be defined by the semantic markup in OWL Lite as shown in Tab. 1.

Tab. 1 Semantic markup provided by OWL Lite to define the semantic relation

Items	Sub/super relation	Equivalent relation
Class	rdfs: subclassOf	EquivalentClass
Object property	rdfs: subPropertyOf	EquivalentProperty
Datatype property	rdfs: subPropertyOf	EquivalentProperty
Individual		SameAs

2.2 Generation of global ontology

There is different domain information on the web, and hence different domain ontologies and domain ontology languages. In order to ensure that different information can be indexed and searched, we should transform the domain ontologies into global ontologies defined in standard ontology language. Here, we give the definition of the domain ontologies.

Definition 1 Domain means the part of the world which people want to describe by knowledge. Domain conception is the abstraction of the term set, which is extracted from the tasks of the domain. The domain ontology can be regarded as an explicit description of the domain conception and it is usually defined in ontology language. The domain ontology can also be defined as {a conception set of this domain, or a set of the domain knowledge}.

Definition 2 Suppose that the domain A can be divided into n subdomains, then the domain ontology describing A can also be divided into n subdomain ontologies.

Since the domain ontologies on the SW are always defined in different domain ontology languages to satisfy the needs for domain application and description, we should transform these domain ontologies into global ontologies described in standard ontology language, in order to improve efficiency, precision and recall. We also do some integration of domain ontologies when it is needed. The transformation integrates the domain ontologies from the same domain and describe them in OWL Lite. Then the new ontologies are generated and we call them global ontologies. These global ontologies can be regarded as the conception sets in uniform form. The translation mechanism can be achieved with the help of domain experts and tools such as OilEd[8].

2.3 Semantic index term and logic inference

The terms used by the document can be divided into six types according to OWL Lite: class, datatype, object property, datatype property, individual, and data value. After the generation of the global ontologies, the terms in the document are also transformed into semantic markups. So these semantic markups can be used as the semantic index terms. The semantic index terms are identified through URIs.

Using the semantic index terms, we can perform the logic inference and obtain the semantic index terms set. The semantic relations used in the SIR model are sub relation, super relation and equivalent relation. Suppose that x_i is a class, c_i is the sub class of x_i. Then, $c_i \in [C_i]$. In addition, $[F_i]$ and $[E_i]$ represent the super class and equivalent class of x_i. The three semantic relations among the six types of semantic index terms can be described by semantic markups represented in Tab. 1. Logic inference tool OWLJessKB[9] can be used to accomplish key inference in OWL Lite, so we can obtain the sub classes, super classes and equivalent classes of semantic index terms.

2.4 Information retrieval based on semantic index terms

Suppose that there is a documents set D on the SW and $D = \{d_1, d_2, \ldots, d_n\}$, $d_j \in D(1 \leqslant j \leqslant n)$; x_i is a semantic index term, $[C_i]$ $(1 \leqslant i \leqslant n)$ is the set of c_i which is the sub class of x_i. Suppose that x, y and z denote three different strings in document d_j; the semantic index terms representing x, y, and z are c_{i1}, c_{i2} and c_{i3}, respectively. Though x, y and z may be different in d_j, they can all be represented by the semantic index term c_i.

Based on TF-IDF schemes introduced in Ref. [10], the weight of the semantic index terms set $[X_i]$

in the document d, $w([X_i], d)$ can be calculated as follows:

$$w([X_i], d) = \log(f([X_i], d) + 1) \log\left(\frac{n}{n_i} + 1\right) \tag{1}$$

where $f([X_i], d)$ represents the frequency of the semantic index terms set $[X_i]$ appearing in document d; n is the total number of documents in the documents set D; and n_i represents the number of documents in which elements from semantic index terms set $[X_i]$ appear. Also, the weight of the semantic index terms set $[X_i]$ in the query q can be obtained from

$$w([X_i], q) = f([X_i], q) \tag{2}$$

where $w([X_i], q)$ represents the weight, and $f([X_i], d)$ is the frequency of the semantic index terms set $[X_i]$ appearing in query q.

After the weights $w([C_i], d)$, $w([F_i], d)$ and $w([E_i], d)$, which represent the weights of $[C_i]$, $[F_i]$ and $[E_i]$ in the document d, have been obtained in Eq. (1), the final weight w_{id} can be calculated by using Eq. (3).

$$w_{id} = w([C_i], d) k_1 + w([F_i], d) k_2 + w([E_i], d) k_3$$
$$k_1 + k_2 + k_3 = 1; \quad k_1, k_2, k_3 \geqslant 0 \tag{3}$$

where w_{id} is the final weight of the semantic index terms set $[X_i]$ in document d; and k_1, k_2 and k_3 are the coefficients of $w([C_i], d)$, $w([F_i], d)$ and $w([E_i], d)$, respectively. The coefficients k_1, k_2 and k_3 can be defined according to the user's information need. For instance, if the user wants to search only the super relationship of the query, then we can define k_1 and k_3 as 0, and k_2 as 1. Hence, only the super relationship of the query will be considered.

By using Eqs. (1) and (3), we can obtain the logic view of d_j, and $d_j = \{w_{1j}, w_{2j}, \ldots, w_{tj}\}$. w_{tj} is the weight of the number t semantic index terms set $[X_t]$ in the number j document d_j in documents set D. This logic view reflects the semantics of semantic index terms. At the same time, in the logic view of $q = \{w_{1q}, w_{2q}, \ldots, w_{tq}\}$, w_{tq} is the weight of the number t semantic index terms set $[X_t]$ in the query q. The ranking function is represented as follows:

$$f = 1 + \frac{\sum_{i=1}^{i} w_{ij} w_{iq}}{\sqrt{\sum_{i=1}^{i} w_{ij}^2 \sum_{i=1}^{i} w_{iq}^2}} \tag{4}$$

The value range of f is $[0, 2]$, and it represents the similarity between q and document d_j. This ranking function defines an ordering among the documents in D with regard to q.

3 Practicability of SIR Model

In this paper, we use precision P and recall R to evaluate the SIR model.

$$P = \frac{N'}{M}, \quad R = \frac{N'}{N} \tag{5}$$

where M is the number of retrieved documents, N is the number of relevant documents, and N' is the number of retrieved relevant documents.

Here, we give a scenario of retrieval, which can reflect the essential differences between the SIR model and the widely-used keyword-based method. That is the SIR model relies on the semantic analysis of the semantic index terms while the keyword-based method relies on the syntax analysis of the keyword. Suppose that there is a documents set D, $D = \{d_1, d_2, d_3, d_4, d_5\}$; string s is in d_1 and d_4, and string t is in d_2 and d_5. There is semantic index term S used to markup s and semantic index term T used to markup t. S and T belong to the same semantic index terms set $[X_i]$. String x is in d_3 and semantic index term R is used to markup x. R is not the element of $[X_i]$. The semantic index terms set $[X_i]$ can represent the user's information need x.

From the statement above, we know that relevant documents to user's information need are d_1, d_4, d_2 and d_5. However, according to the keyword-based information retrieval model, which can only perform the syntax analysis, d_1, d_4 and d_3 are retrieved. Then the precision for this model is 0.67 and recall is 0.50. At the same time, according to the SIR model, semantic analysis can be conducted and the answer set is $\{d_1, d_4, d_2, d_5\}$. Then, the recall and precision for the SIR model are both 1.0. Hence, it shows that the latter model is better in precision and recall.

4 Conclusion

The SW has gradually changed the current web by using semantic markup to describe the web content to facilitate users access to the information. In order to integrate the logic inference and IR, we propose an IR model SIR for the SW. In the SIR model, the global ontologies are generated and the semantic index terms are extracted from these ontologies. Then, logic inference is performed based on the semantic index terms and the ranked list is obtained. Since the SIR model relies on semantic index terms, not the syntax of keywords, IR can be performed based on the semantic analysis and, hence, there are higher recall and precision.

References

[1] Berners-Lee T, Hendler J, Lassila O. The semantic web [J]. *Scientific American*, 2001, **284**(5): 34 - 44.

[2] Shah U, Finin T, Joshi A. Information retrieval on the semantic web [A]. In: *Proc of the Eleventh International Conference on Information and Knowledge Management* [C]. Mclean, Virginia, USA, 2002. 461 - 468.

[3] Yang Lingpeng, Ji Donghong, Li Tang. Chinese information retrieval based on terms and ontology [A]. In: Noriko Kando, Masao Takaku, eds. *Proceedings of the Fifth NTCIR Workshop Meeting on Evaluation of Information Access Technologies*: *Information Retrieval*, *Question Answering and Cross-Lingual Information Access* [C]. Tokyo: National Institute of Informatics, 2005. 136 - 142.

[4] Saias Jose, Quaresma Paulo. A methodology to create ontology-based information retrieval systems[A]. In: *Lecture Notes in Artificial Intelligence*[C]. Springer-Verlag, 2003, **2902**: 424 - 434.

[5] van Hage Willem Robert, de Rijke Maarten, Marx Maarten. Information retrieval support for ontology construction and use [A]. In: McIlraith Sheila A, Plexousakis Dimitris, van Harmelen Frank, eds. *Proceedings of the Third International Semantic Web Conference, LNCS* [C]. Hiroshima, Japan, 2004, **3298**: 518 - 533.

[6] Jose M Abasolo, Mario Gomez. MELISA: an ontology-based agent for information retrieval in medicine[A]. In: *Proceedings of the First International Workshop on the Semantic Web*[C]. Lisbon, Portugal, 2000, **3**: 73 - 82.

[7] Jacques Guyot, Radhouani S, Gilles Falquet. Ontology-based multilingual information retrieval[EB/OL]. (2005) [2005-11-25]. http://www. clef-campaign. org/2005/working_notes/CLEF2005WN-Contents. html.

[8] Bechhofer Sean, Horrocks Ian, Goble Carole, et al. OilEd: A reason-able ontology editor for the semantic web[A]. In: Baader F, Brewka G, Eiter Th, eds. *Proceedings of Joint Austrian/German Conference on Artificial Intelligence, LNAI* [C]. Vienna, Austria, 2001, **2174**: 396 - 408.

[9] Kopena J. OWLJessKB: Library for OWL inference using Jess [EB/OL]. (2005-01-21) [2005-10-25]. http://edge. cs. drexel. edu/assemblies/.

[10] Baeza-Yates Ricardo, Ribeiro-Neto Berthier. *Modern information retrieval* [M]. Switzerland: Addison Wesley, 1999.

一种对语义网上本体查询和检索的新方法

虞 为[1] 曹加恒[1] 陈俊鹏[2]

([1] 武汉大学计算机学院,武汉 430072)

([2] 武汉邮电科学研究院,武汉 430072)

摘要:针对语义网信息检索中存在的问题,提出了一个基于语义索引词的语义网信息检索模型 SIR (semantic information retrieval). 其核心思想是将领域本体转换成全局本体,并从全局本体中提取语义索引词. 通过语义索引词进行语义推理,可得概念的逻辑视图. SIR 通过语义索引词间的语义关系对网络资源进行检索,解决了在传统的基于关键字的信息检索中只能从句法上对关键字进行分析,无法根据信息资源中的语义关系进行检索的问题. 最后分析了 SIR 的可用性,证明了 SIR 可极大地提高语义网上信息检索的查全率和查准率.

关键词:语义索引词;语义网;本体

中图分类号:TP393