An Ontology-based Methodology for Semantic Expansion Search

Guobing Zou
Dept. of Computer Science
Tongji University
Shanghai, 201804, China
guobing278@sina.com

Bofeng Zhang
Shcool of Comp. Eng. and Sci.
Shanghai University
Shanghai, 200072, China
bfzhang@shu.edu.cn

Yanglan Gan
Dept. of Computer Science
Tongji University
Shanghai, 201804, China
ganyl@mail.tongji.edu.cn

Jianwen Zhang
Center of material supply
Yanzhou Coal Mining Co.,Ltd
Zoucheng, 273500, China
jwzhangyzc@sina.com

Abstract

Matching search technology based on query keyword has been widely used by traditional search way. It still belongs to pure keyword matching and can not acquire satisfactory search results. The essential reason is that traditional web search lacks semantic understanding to user's search behaviors. In this study, we propose a novel ontology-based framework for semantic expansion search. Based on constructed domain ontology, semantic annotation algorithm and semantic expansion reasoning algorithm are presented in detail, which are associated with semantic annotation unit and semantic expansion reasoning engine respectively. Then a semantic search prototype system is designed and implemented. The experimental results show that semantic expansion search by proposed methodology can overcome limitations in comparison with traditional keyword search mode, and achieve higher recall ratio and precision ratio.

1. Introduction

As information resources have rapidly increased in recent years, search engine has been widely used and become a prerequisite approach for users to retrieve and acquire information resources on the Internet. At present, two main disadvantages [1] exist in current search engines. Firstly, thousands of irrelevant web pages have been returned from search engine. Secondly, the display order of search results is rather in confusion. Thus, search engine can not deal with search results effectively and efficiently for those returned web pages when it ranks them by all kinds of ranking algorithms. The essential reason of these two issues is that traditional web search lacks semantic understanding to user's search behaviors, which results in low recall ratio and precision ratio. Therefore, it is difficult to make users satisfy their search requirements.

Ontology is a formal, explicit specification of a shared conceptualization [2]. Owning to its function of describing conceptual meaning and relationships among different concepts, domain ontology can form a nice conceptually hiberarchy and provide excellent support for logic reasoning. Consequently, domain ontology has been widely applied to acquire useful knowledge in information retrieval domain. However, two crucial problems related to domain ontology application remain unsolved perfectly. Due to lacking semantic understanding ability to expand user's query keyword, search systems can not provide users with accurate query expansion set for search navigation. Furthermore, it is difficult for current retrieval systems to realize exactly semantic annotation. Therefore, innovative annotation algorithms are necessary for the sake of annotating specific domain document resources and web pages.

In order to solve these existed issues, this paper adopts domain ontology as the way of domain knowledge organization and expression and proposes a semantic expansion search framework and related algorithms.

Our work is distinguished from others for the following reasons.

(1) Rather than using a dictionary or knowledge index, computer science ontology (CSO) has been constructed and used for semantic annotation and semantic expansion reasoning.

(2) Based on CSO domain ontology, we propose a semantic expansion search model named Sem-Exp-M.

(3) In order to annotate domain resources effectively and efficiently, a novel semantic annotation algorithm named DocSemanAnno is designed and implemented.

(4) For the purpose of semantic query expansion, we present a new semantic expansion reasoning algorithm called SemanExpRea, which is responsible for semantic query expansion to user's query keyword.

The rest of paper is organized as follows. In Section 2, we review related works. In Section 3, domain ontology is defined and constructed. In Section 4, we firstly describe the overall framework for semantic expansion search. Then semantic annotation and semantic expansion reasoning algorithms are represented. Experimental results and analysis are shown In Section 5. Finally, Section 6 concludes the paper and our future works.

2. Related work

Domain ontology has been emerged as a mainstream in many application domains. Exploiting a key observation that semantically related items exhibit consistency in

978-0-7695-3305-6/08 $25.00 © 2008 IEEE
DOI 10.1109/FSKD.2008.475

presentation style as well as spatial locality in template-based content-rich HTML documents, Mukherjee S et al. proposed a novel framework for automatically partitioning such documents into semantic structures [3]. The framework tightly couples structural analysis of documents with semantic analysis incorporating domain ontologies and lexical databases. The aim of this literature was to bridge the semantic gap by addressing the fundamental problem of automatically annotating HTML documents with semantic labels. Dill, S et al. [4] developed a system called SemTag for large-scale annotation of web documents. Annotations in SemTag are carried out on the level of concepts in a document using TAP taxonomy [5].

Language Modeling (LM) has been successfully applied for query expansion. Term relationships in LM have been proposed by Jing Bai et al. in literature [6] to expand query model instead of document model, so that query expansion process can be naturally implemented. SC Wang and Yuzuru Tanaka [7] introduced a topic-oriented query expansion model based on the information bottleneck theory that classify terms into distinct topical clusters in order to find out candidate terms for the query expansion. A term-term similarity matrix was defined to improve the term ambiguous problem during the process of query expansion.

These semantic annotation frameworks and query expansion algorithms still have some limitations. Firstly, these frameworks just annotate documents with either lexical databases or taxonomy, which is not enriched and accurate to extract and compute concepts or instances. So the quality of semantic annotation has fallen. Secondly, the lack of efficient semantic expansion reasoning algorithms makes above query expansion models and algorithms impractical for user's query keyword.

Aiming at above problems, we propose a novel ontology-based semantic expansion search method, which is inductive to implement effective documents annotation and semantic query expansion.

3. Domain ontology construction

According to definitions to domain ontology in literature [8], we give our definitions about domain ontology as follows.

Definition 3.1 (Domain ontology). Domain ontology is defined as a five tuple:

$$DO = \{C, R, H^C, I, A\} \quad (1)$$

where C represents concept set; R shows relationship set among concepts or between concept and instance; H^C represents hierarchical structure among concepts and instances. I denotes set of overall instances that belong to concepts. A includes all axioms of domain.

Definition 3.2(Concept representation). One concept in domain ontology is formalized by a four tuple:

$$Concept_i = \{Id_i, Name_i, CRelationSet_i, CSynoSet_i\} \quad (2)$$

where Id_i is the identifier for concept i, which is unique in our defined domain ontology; $Name_i$ represents name of concept i; $CRelationSet_i$ describes relation set where every element exists a kind of relation with concept i; $CSynoSet_i$ includes all synonyms for concept i.

In this paper, three kinds of semantic relationships are considered between two concepts or concept and instance: Part-of, Kind-of and Instance-of. Each relationship is illuminated respectively in definition 3.3.

Definition 3.3 (Concept relationship). Relationship set R defined in definition 3.1 is made up of three kinds of relationships:

$$R = \{Part\text{-}of, Kind\text{-}of, Instance\text{-}of\} \quad (3)$$

Part-of relation depicts relation of part and integrity between two concepts; *Kind-of* relation is represented by characteristic of inheritance relationship of two concepts; *Instance-of* relation describes inclusion relationship between a concept and its subordinate instance.

According to the above basic definitions, by using concept or instance as node and three kinds of relations (Part-of, Kind-of and Instance-of) as joint edge, domain ontology can be represented as a tree structure graph (TSG). A portion of tree structure graph of CSO domain ontology that we have constructed is shown in Figure 1.

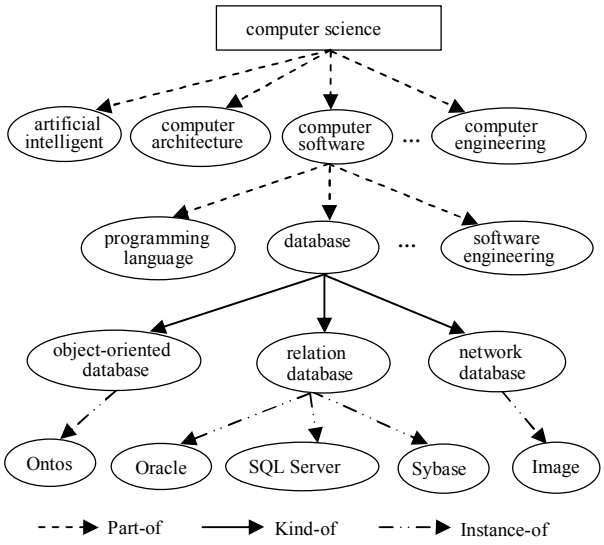

Figure 1. A portion of tree structure graph of CSO

4. Semantic expansion search

4.1 Semantic expansion search model

According to different ways of search realization, search mode can be divided into full text search, semantic search and semantic expansion search. Full text search has been defined in [9]. We give definitions of semantic search and semantic expansion search respectively.

Semantic search refers to the process that semantic annotation algorithm extracts concepts or instances from domain ontology, annotates document pool for domain resource repository, and generates semantic index repository. According to user's query keyword, Search program carries on search task from semantic index repository and search results with semantic feature are returned to user.

Semantic expansion search refers to the process that semantic expansion reasoning algorithm generates query expansion set for user's query keyword. According to the query expansion set, search program carries on search task from semantic index repository and search results with the characteristic of semantic information are returned to user.

Semantic expansion search model called Sem-Exp-M and relationships are shown in Figure 2.

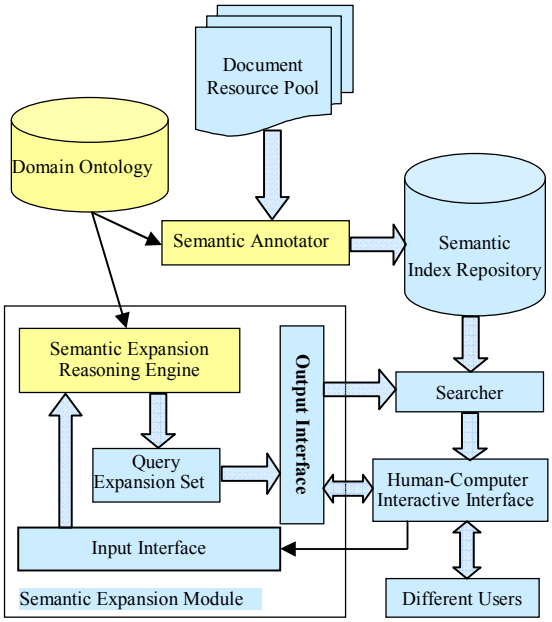

Figure 2. Semantic expansion search model Sem-Exp-M

Human-computer interactive interface belongs to interactive layer of Sem-Exp-M. User submits search task to input interface and interacts with output interface for determining query expansion set, and then achieves search results from searcher. The function of semantic expansion module is to implement semantic expansion for user's query keyword. By the acquisition of search condition from human-computer interactive interface, reasoning engine executes reasoning and generates query expansion set via semantic expansion reasoning algorithm.

Semantic annotator is to make document resource pool possess semantic feature. By use of semantic annotation algorithm, semantic annotator extracts and expands concepts or instances from domain ontology. Then semantic feature domain of document is created and annotated by extracted and expanded concepts or instances. Searcher acquires query expansion set as search condition from output interface and retrieves documents from semantic index repository. Domain ontology is stored in the form of an OWL document.

4.2 Semantic annotation algorithm

Semantic annotation marks the original data (textual or symbolic) in order to give them semantic information and make them understandable for both people and computers. In this paper, a semantic annotation algorithm based on domain ontology is proposed. For further illustration, we define two related concepts as follows.

Definition 4.1 (Equal concept set). In the domain ontology of $DO = \{C, R, H^C, I, A\}$, for $\forall\, c \in C$, equal concept set of c:

$$E(c) = \{syno \mid syno \in CSynSet_c\} \quad (4)$$

Definition 4.2 (Instance set). In the tree structure graph (TSG) of domain ontology, $DO = \{C, R, H^C, I, A\}$, for $\forall\, c \in C$, instance set of concept c:

$$InstanceSet(c)=\{inst \mid H(c, inst) \in H^C.\ c \in C,\ inst \in I\} \quad (5)$$

For example, for domain concept "relation database" shown in Figure 1, InstanceSet("relation database") = {"Oracle", "SQL Server", "Sybase"}.

The description of semantic annotation algorithm called DocSemanAnno is shown in Algorithm 1.

Algorithm 1 DocSemanAnno(DO, *TSG*, *C*, *I*, *D_S*)
Input: *DO*, *TSG*, concept set *C*, instance set *I*, document set $D\_S = \{d_1, d_2, \ldots, d_N\}$.
Output: semantic document set $SD\_S=\{sd_1, sd_2, \ldots, sd_N\}$.
(1) set $SD\_S = \varnothing$, $D\_S = \{d_1, d_2, \ldots, d_N\}$;
(2) generates set $D\_C = \{c_1, c_2, \ldots, c_N\}$;
c_i corresponds to document d_i, i = 1, 2, …, N;
(3) **For** i=1 to N do begin
acquire c_i from D_C and set $node = c_i$;
If ($node \in C$) then begin
compute $E(c_i)$ in *DO*
compute $InstanceSet(c_i)$ in *TSG*;
annotate d_i with c_i, $E(c_i)$ and $InstanceSet(c_i)$;
generate semantic document sd_i;
End else if ($node \in I$) then **begin**
annotate d_i with instance c_i;
generate semantic document sd_i;
End;
append sd_i to set SD_S;
End;
(4) generate semantic document set SD_S;
(5) output $SD\_S= \{sd_1, sd_2, \ldots, sd_N\}$;

The essential idea of algorithm 1 is to automatically generate semantic documents set from original documents set by annotating semantic feature domain of document.

In the procedure of semantic annotation algorithm, annotation program acquires document sort word c_i (i = 1, 2, 3, …, N) by automatic document classification and then maps it to certain concept or instance in TSG of domain ontology with support of word similarity computation. By the computation of equal concept set and instance set, concepts or instances are effectively used to annotate semantic feature domain of document, which is formed by adopting XML marking language format. As a result of automatic semantic annotation for original documents set D_S, semantic document set SD_S is produced and used to provide semantic index repository for semantic search and semantic expansion search.

4.3 Semantic expansion reasoning algorithm

In virtue of ontology reasoning mechanism, for a concept, descendant concept set and grandfather concept set are defined to describe reasoning algorithm.

Definition 4.3 (Descendant concept set). In the tree structure graph (TSG) of domain ontology, $DO = \{C, R, H^C, I, A\}$, for $\forall\ c \in C$, descendant concept set of c is defined as:

$D(c) = \{node \mid H(c, node) \in H^C \vee$ a route $P(c, d_1, d_2, \ldots, d_n, node)$ existed in $TSG.\ node \in C, d_i \in C, n \geq 1, i = 1, 2, \ldots, n\}$ (6)

In order to make the query expansion set effective, direct descendant concept set has been used in this paper. We define it as $C\_Dir\_Subclass(c) = \{ node \mid H(c, node) \in H^C.\ node \in C\}$.

Definition 4.4 (Grandfather concept set). In the tree structure graph (TSG) of domain ontology, DO = {C, R, H^C, I, A}, for $\forall\ c \in C$, grandfather concept set of c is defined as:

$G(c) = \{node \mid H(node, c) \in H^C \vee$ a route $P(node, g_1, g_2, \ldots, g_n, c)$ existed in $TSG.\ node \in C, g_i \in C, n \geq 1, i = 1, 2, \ldots, n\}$ (7)

Considering the efficiency of semantic expansion results, direct grandfather concept set has been adopted in this paper. It is defined as $C\_Dir\_Parclass(c) = \{node \mid H(node, c) \in H^C.\ node \in C\}$.

To make the algorithm more understandable, a $procedureset = \{PSK, PSE(c), PSI(c), PSS(c), PSP(c)\}$ is used to illuminate the process of semantic expansion reasoning. For a domain concept c:

(1) $PSK = \{Input\_List = Input\_List + Key; Res\_List = Res\_List + Input\_List\}$. *Input_List* is used to save user's query keyword Key, and *Res_List* stores the whole query expansion set.

(2) $PSE(c)=\{Equal\_List= Equal\_List + E(c); Res\_List =Res\_List+Equal\_List\}$. *Equal_List* is applied to save $E(c)$.

(3) $PSI(c) = \{ Instance\_List = Instance\_List + InstanceSet(c); Res\_List=Res\_List+Instance\_List\}$.Instance List is used to keep *InstanceSet(c)*.

(4) $PSS(c) = \{ Child\_List = Child\_List + C\_Dir\_Subclass(c); Res\_List = Res\_List + Child\_List\}$. *Child_List* is applied to conserve *C_Dir_Subclass(c)*.

(5) $PSP(c) = \{ Parent\_List = Parent\_List + C\_Dir\_Parclass(c); Res\_List = Res\_List + Parent\_List\}$. *Parent_List* holds *C_Dir_Parclass(c)*.

The description of semantic expansion reasoning algorithm called SemanExpRea is shown in Algorithm 2.

Algorithm 2 SemanExpRea(*DO*, TSG, *C*, *I*, *Keyword*)
Input: *DO*, *TSG*, concept set *C*, instance set *I*, *Keyword*.
OutPut: query expansion set *Res_List*.
(1) set $Res\_List=Input\_List=Equal\_List=Instance\_List =Child\_List=Parent\_List=\varnothing$;
(2) acquire *Keyword* and turn it into OWL form *c*;
(3) **If** ($c \notin C \vee c \notin I$) then goto step (2);
Else execute procedure PSK;
(4) compute *E(c)* in *DO*;
If ($E(c) \neq \varnothing$) then execute procedure PSE(c);
(5) reason *InstanceSet(c)* in *TSG*;
If ($InstanceSet(c) \neq \varnothing$) then **begin**
execute procedure PSI(c);
goto Step (8);
End;
(6) generate *C_Dir_Subclass(c)* in *TSG*;
If ($C\_Dir\_Subclass(c) \neq \varnothing$) then **begin**
execute procedure PSS(c);
goto Step (8);
End;
(7) produce *C_Dir_Parclass(c)* in *TSG*;
If ($C\_Dir\_Parclass(c) \neq \varnothing$) then
execute procedure PSP(c);
(8) generate query expansion set *Res_List*;
(9) output *Res_List*;

The central idea of algorithm 2 is to expand query keyword. Instance set and direct descendant concept set are closer to semantic relativity for a concept. So they have higher priority than direct grandfather concept set when we execute query expansion. Thus, it will not be expanded if instance set or direct descendant concept set exists.

5. Experimental results and analysis

5.1 Experimental data set and results

In order to validate the efficiency of our proposed semantic expansion search method, we have constructed computer science domain ontology by ontology editor Protégé 2000. The established CSO domain ontology contains approximately 2700 concepts and instances. At the same time, 4558 academic papers have been collected for our experimental data set on the Internet. Based on CSO domain ontology, a semantic search prototype system has been designed, which has implemented full text search, semantic search and semantic expansion search.

In this paper, recall ratio P_{recall} and precision ratio $P_{precision}$ are applied to evaluate efficiency of search results. P_{recall} refers to proportion of retrieved related documents d_{rs} out of all related documents d_{sum} in system. $P_{precision}$ is defined as proportion of retrieved related documents d_{rs} relative to all retrieved documents d_{ss}.

$$P_{recall} = \frac{\text{retrieved related documents } d_{rs}}{\text{all related documents } d_{sum} \text{ in system}} \quad (8)$$

$$P_{precision} = \frac{\text{retrieved related documents } d_{rs}}{\text{retrieved documents } d_{ss}} \quad (9)$$

In order to compare and analyze search efficiency among three kinds of search ways, the experiments used a group of search keywords, such as "computer aided test", "computer produced drawing", "computer aided design", "computer assisted management", "computer aided instruction", "computer aided engineering", "computer aided manufacturing", "computer aided audit", "computer aided programming", "computer aided education". They

are represented respectively by {*K*1, *K*2, *K*3, *K*4, *K*5, *K*6, *K*7, *K*8, *K*9, *K*10}.

Recall ratio and precision ration among three kinds of search ways have been listed in Figure 3 and Figure 4.

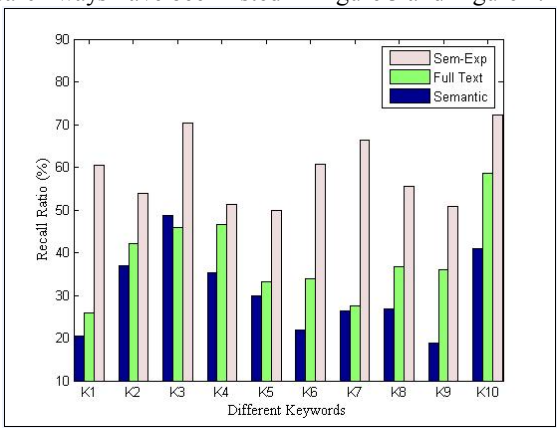

Figure 3. Comparison of recall ratio among three search ways

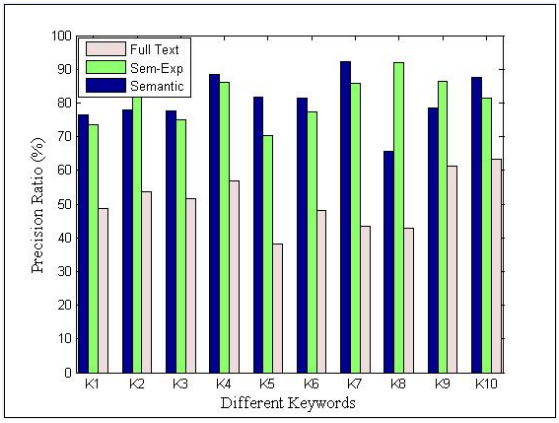

Figure 4. Comparison of precision ratio of three search ways

From above comparison of recall ratio and precision ration among three search ways, we can calculate mean search efficiency shown in the following Table 1.

Table 1. Comparison of mean efficiency of three search ways

	Full text	Semantic	Seman-Exp
Mean recall	38.63%	30.61%	59.18%
Mean precision	50.81%	80.73%	81.08%

5.2 Experimental analysis

Although full text search returns more retrieved documents to user than that of semantic search, only the retrieved documents, which are simultaneously included in search results by semantic search, are the ones user really needs. Thus, even if recall ratio of full text search is higher than that of semantic search, it is no practical value for the higher part. However, precision ratio of full text search is far below semantic search.

By semantic query expansion of user's keyword, retrieved documents of semantic expansion search are markedly more than semantic search. Therefore, recall ratio has obviously improved in comparison with semantic search. Meanwhile, precision ration does not differ significantly from each other between semantic search and semantic expansion search.

In aspect of recall ratio, semantic expansion search and full text search are superior to semantic search. However, semantic expansion search and semantic search are far higher than full text search in respect of precision ratio. So we can conclude that semantic expansion search has higher search efficiency not only in recall ratio but also in precision ratio.

6. Conclusion

Based on domain ontology, this paper proposes a new method for semantic expansion search. Especially, we present annotation algorithm and expansion reasoning algorithm in detail. In comparison with full text search and semantic search, semantic expansion search proposed in this paper can overcome limitations and achieves higher recall ratio and precision ratio. In order to improve semantic expansion search performance, several key issues will be further addressed. Our future work will deal with the integrity of domain ontology construction, the creation of efficiently semantic annotation algorithm and the optimization of query expansion set.

References

[1] XG Zhang, MS Li. Research on Search Engine Technologies. Journal of Computer Engineering and Applications, 24:67-70, 2001.

[2] Studer R, Benjamins V R, Fensel D. Knowledge Engineering, Principles and Methods. Data and Knowledge Engineering, 25(12):161-197, 1998.

[3] Mukherjee S, Yang GZ, Ramakrishnan IV. Automatic Annotation of Content-Rich HTML Documents: Structural and Semantic Analysis. Lecture Notes in Computer Science, 2870: 533-549, 2003.

[4] Dill, S., Eiron, N., Gibson, D., Gruhl, D., Guha, R., Jhingran, A., Kanungo, T., Rajagopalan, S., Tomkins, A., Tomlin, J., and Zien, J. Semtag and seeker: bootstrapping the semantic web via automated semantic annotation. In Proceedings of the 12th International World Wide Web Conference (Budapest, Hungary, May 2003), 178-186, 2003.

[5] Guha, R., and McCool, R. Tap: Towards a web of data, http://tap.stanford.edu/.

[6] J. Bai, DW Song, P. Bruza, JY Nie and GH Cao. Query Expansion Using Term Relationships in Language Models for Information Retrieval. In Proceedings of the 14th ACM International Conference on Information and Knowledge Management, pages 688-695, Bremen, 2005.

[7] SC Wang, Y. Tanaka. Topic-Oriented Query Expansion for Web Search. In Proceedings of the 15th international conference on World Wide Web, pages 1029-1030, 2006.

[8] SY Lao, L. Bai, YL Hu, et al. New Video Retrieval Using Domain Ontology. Journal of Chinese Computer Systems, 28(8): 1470-1476, 2007.

[9] H. Xiao. SCORM Resource Semantic Research to Knowledge Organization and Retrieval. Beijing: Peking University, 2006.