Research Journal of Applied Sciences, Engineering and Technology 4(16): 2839-2843, 2012
ISSN: 2040-7467

Submitted: March 31, 2012 Accepted: April 23, 2012 Published: August 15, 2012

Query Expansion Approach Based On Ontology and Local Context Analysis

Jing Wan, WenCong Wang, JunKai Yi, Chong Chu and Kang Song
College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China

Abstract: In information query process, query expansion can effectively resolve the problem related to low query efficiency caused by ambiguous short query terms. This study proposes a query expansion approach based on ontology and local context analysis. Ontology-based query expansion enables the addition of standardized descriptive information for search query, while local context analysis, through document library analysis, offers more relevant information for inquiry. We combine advantages of the two methods to expand the keywords semantically. Experimental analysis indicates that our method can effectively improve the search recall and precision.

Keywords: Local context analysis, ontology, query expansion, relevance calculation

INTRODUCTION

In conventional information query, expression discrepancies of the same concepts in document and user query result in phrase mismatch. Ambiguous keywords and the weight setting of keywords in document are the major reasons affecting query result (Díaz-Galiano *et al*., 2009).

In the field of information query, query expansion has been extensively recognized as one of the techniques which can effectively improve query efficiency (Rahmatollah *et al*., 2008). The basic idea is to use queried keyword-related phrases to correct and complement the query results, so that the recall and precision of information query will be improved. Query expansion approach involves global analysis, local analysis, local context analysis and user logs-based query expansion (Hsi-Ching *et al*., 2006). Global analysis processes the entire archives, with large amount of calculation, thus it is unsuitable for massive data retrieval. Local analysis depends on the initially-retrieved documents, so when the initially-retrieved documents are not highly correlated with original query, the retrieval precision will be much affected (Saeedeh and Katebi, 2010). The results of association rule-based query expansion depend on the quality of inter-phrase association rules. Query expansion based on user logs requires sufficient quantity of user logs.

In order to improve query efficiency, this study proposes a query expansion method which combines ontology-based query expansion and local context analysis.

RELEVANT DEFINITIONS

Definition of ontology: The concept of ontology originated in the field of philosophy. Ontology deals with the scientific connotation of the nature of things. In the field of computer science, "ontology is a formal and explicit specification of the concept of sharing" (Cho-Wei *et al*., 2011).

Definition 1 ontology: Knowledge ontology is a triple:

KO = <KA, Rela, Rule>

KA (Knowledge Atom) represents the set of knowledge atoms, which are the smallest units for knowledge representation, or the smallest component units of knowledge ontology, i.e.,

$KA = \{a_i \mid 1 \le i \le n, a_i \notin \varphi, a_i \in \Omega\}$

KB represents the set of knowledge ontology:

$$KB = \left\{ b_k \mid 1 \le k \le n, b_k = \left\{ \sum a_i \ a_j \vee \prod a_i \ a_j \mid 1 \le i, j \le n, a_i, a_j \notin \varphi, a_i a_j \in KA \right\} \right\}$$

Rela (Relation) represents the set of relations between knowledge atoms and knowledge entities:

$$Rela = \left\{ r_{ij}(a_i, a_j) \vee r_{kl}(b_k, b_l) \mid 1 \le i, j, k, l \le n, r_{ij}, r_{kl} \notin \varphi, a_i, a_j \in KA, b_i, b_j \in KB \right\}$$

Corresponding Author: Jing Wan, College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China

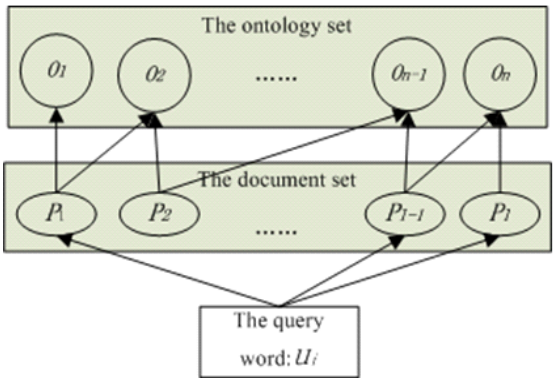

Fig. 1: The relationship among query, document and ontology

Rule represents the set of rules or operations formed by the combination of relations between knowledge atoms or knowledge entities.

The relationship among query, document and ontology:

Definition 2 the relationship among query, document and ontology: Attachment degree of the word queried to different ontology categories. u_i, a word to be queried may be included in multiple documents, each of which belongs to one or more ontology categories. The statistical knowledge of ontology categories to which the documents containing the word to be queried belong can be used to obtain the attachment degree of the word queried to different ontology categories. The relationship among query, document and ontology is illustrated in Fig. 1.

The relationship between query and ontology: The larger the number of concepts mapped from the queried word through the document is, the lower the attachment degree of the queried word to a single ontology is. The higher the occurrence frequency of a queried word in the documents attached to ontology, the higher the attachment degree of the queried word to the ontology is. The higher he occurrence frequency of a queried word in documents attached to the ontology is, the higher the attachment degree of the queried word to the ontology is (Bettina *et al*., 2011).

QUERY EXPANSION MODEL

Query expansion model proposed in this study mainly consists of three parts: the construction of ontologically expanded word set, the construction of statistically expanded word set and the relevance calculation of expanded query words. As shown in Fig. 2, the basic idea of this model is as follows:

- User input word set U is expanded through ontology expansion, thus the ontologically expanded word set is obtained. Then the relevance between each ontologically expanded word and user input is calculated.
- Local context analysis is used to obtain statistically expanded word set and the relevance between each statistically expanded word and user input is calculated.
- Union set of the two word sets is taken, these expanded words are sequenced according to the calculated relevance of them, in order to obtain the set of conditions for query expansion.

Construction of ontologically expanded word set: The procedures to obtain ontologically expanded word set are as follows:

Based on the received user input, syntactic and semantic association analysis, semantic disambiguation, keyword extraction and stop word removal are performed, in order to obtain user input word set $U = \{u_i \mid 1 \le i \le n\}$. According to the attachment relationship among query, document and ontology, the input word u_i is mapped into each entity, in order to obtain the set O of ontology.

O is used to expand u_i by domain category, thus the set $Ontv = \{Ontv_{i,j} \mid 1 \le i, j \le n\}$ is obtained.

$g(u_i, Ontv_{i,j})$, the association degree between u_i and $Ontv_{i,j}$, is calculated, in order to obtain

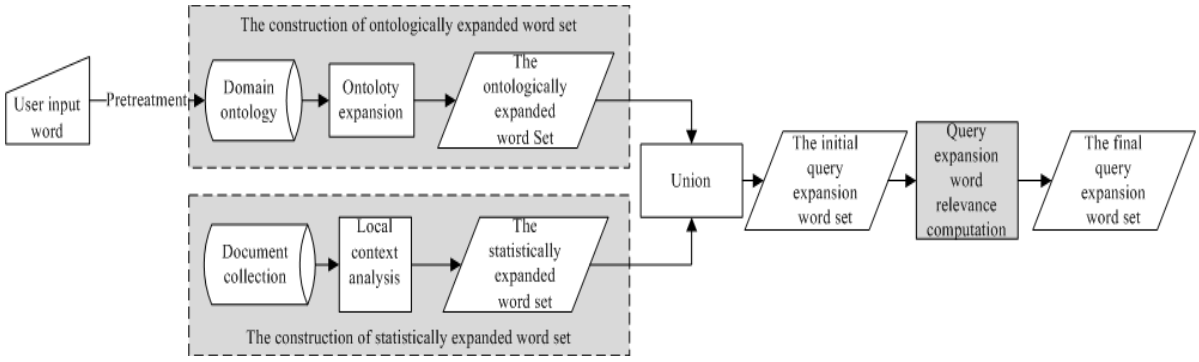

Fig. 2: Query expansion model based on ontology and local context analysis

$Rel_{OntV} = \{Rel(u_i, Ontv_{i,j}) | 1 \le I, j \le n\}$, the set of relevance of ontologically expanded words.
Rel ($u_{i,}$, $Ontv_{i,j}$), is calculated by:

$$g\left(u_i, Ontv_{i,j}\right) = tf\left(u_i, p_l\right) \times \ln\left(\frac{m_i}{M} + 1.0\right) \quad (1)$$

$$Rel\left(u_i, Ontv_{i,j}\right) = g\left(u_i, Ontv_{i,j}\right) / \max\left(g\left(u_i, Ontv_{i,j}\right)\right) \quad (2)$$

where,

If(u_i, p_l) is the occurrence frequency of u_i in document p_l ; m_i is the number of documents which are mapped into the set of ontology according to the attachment relationship between document and ontology; M is the total number of documents attached to entity O_n.

Construction of statistically expanded word set: Local context analysis is used to construct statistically expanded word set in the following procedures. Using data mining techniques such as statistics theory and mutual information method, n articles related to the original query are retrieved, i.e. D = (p_1, p_2, p_3.....p_n) which is used as the source of words for expansion. Then m words which are most relevant to word set U are selected from n articles to form statistically expanded word set StaV, denoted as:

$StaV = \{Stav_{i,j} | 1 \le i, j \le n\}$

The relevance of statistically expanded words is calculated by the following formula (Jinxi and Bruce, 1996):

$$bel\left(u_i, Stav_{i,j}\right) = \left\{\delta + \ln\left(co\left(u_i, Stav_{i,j}\right) + 1\right) \times idf_{Stav_{i,j}} / \ln(n)\right\}^{idf_{u_i}} \quad (3)$$

$$co\left(u_i, Stav_{i,j}\right) = \sum_{l=1}^{n}\left[tf\left(u_i, p_l\right) \times tf\left(Stav_{i,j}, p_l\right)\right] \quad (4)$$

$$idf_{u_i} = \max\left[1.0, \ln\left(N / N_{u_i}\right) / 5.0\right] \quad (5)$$

$$idf_{Stav_{i,j}} = \max\left[1.0, \ln\left(N / N_{Stav_{i,j}}\right) / 5.0\right] \quad (6)$$

where, bel($u_{i,}$, $Stav_{i,j}$), is the relevance between user input keyword u_i and statistically expanded word $Stav_{i,j}$. co($u_{i,}$, $Stav_{i,j}$) is the frequency of co-occurrence of u_i and Stavi, in D; δ is a constant; tf(u_i, p_l) and tf($Stav_{i,j}$, p_l) are respectively the frequency of occurrence of u_i and $Stav_{i,j}$ in document p_l. N is the total number of retrieved documents; N_{u_i} is the total number of documents containing u_i; $N_{Stav_{i,j}}$ is the total number of documents containing $Stav_{i,j}$.

The set of relevance of statistically expanded words is represented as:

$$Rel_{Stav} = \left\{Rel\left(u_i, Stav_{i,j}\right) | 1 \le i, j \le n\right\}$$

where, Rel($u_{i,}$, $Stav_{i,j}$) is the relevance between u_i and $Stav_{i,j}$, , calculated by the following formula:

$$Rel\left(u_i, Stav_{i,j}\right) = bel\left(u_i, Stav_{i,j}\right) / \max\left(bel\left(u_i, Stav_{i,j}\right)\right) \quad (7)$$

Calculation of relevance of expanded query words: Through solution of union set of the two word sets, the word set OntV∪StaV = V′ after preliminary query expansion is obtained. The set of relevance of expanded words is:

$$Rel_{V'} = \left\{Rel\left(u_i, v'_{i,j}\right) | 1 \le i, \le n\right\}$$

where, Rel(u_i, $v'_{i,j}$) is calculated by the following formula:

$$Rel\left(u_i, v'_{i,j}\right) = \begin{cases} Rel\left(u_i, Ontv_{i,j}\right) + Rel\left(u_i, Stav_{i,j}\right) v'_{i,j} \in Ontv \bigcap StaV \\ a \times Rel\left(u_i, Ontv_{i,j}\right) + b \times Rel\left(u_i, Stav_{i,j}\right), v'_{i,j} \notin OntV \bigcap StaV \end{cases} \quad (8)$$

where, a, b≥0 and a+b = 1. The members of set $Rel_{V'}$ are arranged in a decreasing order and the first n expanded words with largest Rel(u_i,$v'_{i,j}$) are selected as the most relevant expanded words. Finally, the user input word set *U* is included in the set of expanded words, in order to obtain V = {$v_{i,j}$ | 1≤i, j≤n}, the final set of expanded query words.

EXPERIMENTAL ANALYSIS

The museum knowledge base system was constructed using JAVA language, in order to test the method proposed in this study. The experiment was operated under the following conditions: quad-core 2.30 GHz Intel, 2G memory and Windows 7. The ontology was constructed using Protege 3.1.1; database was SQLServer 2000; the ontology was operated by Jena 2.6.2.

Under the guidance of experts, the museum knowledge ontology was constructed. The model fragment is illustrated in Fig. 3.

The values of parameters in the experimental process are as follows: in Formula (3), δ = 0.5; in Formula (8), a = 0.4, b = 0.6. A total of 50 query experiments were conducted, with the experimental data covering the major 9 aspects in the field of museum: ancient artifacts, ancient literature, cultural relic protection units, museum display, museum management, museum collections, museum

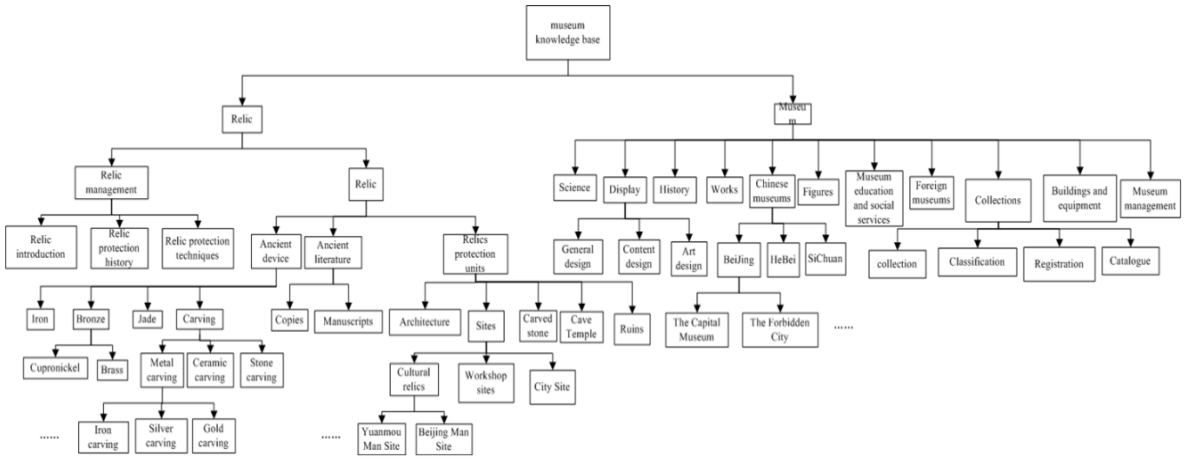

Fig. 3: Fragment of cultural relic data representation model based on ontology

Table 1: Experimental data fragment 1

	User input word	The statistically expanded word set				The ontologically expanded word set	
	Yongle canon	Canon	Ming yongle year	Hong fanzheng kam	Si ku quan shu	Cultural relics publishing	Ancient literature
Yongle canon·	8	6	2	2	0	0	0
Si ku quan shu·	4	2	0	1	12	0	0
Hong fanzheng kam·	0	5	2	4	1	0	0

Table 2: Experimental data fragment 2

	The statistically expanded word set				The ontologically expanded word set	
	Canon	Ming yongle year	Hong fanzheng kam	Si ku quan shu	Cultural relics publishing	Ancient literature
$co(u_i, Stav_{i,j})$	0.267309	0.133409	0.051048	0.059321	None	None
$bel(u_i, Stav_{i,j})$	0.508266	0.714340	0.700954	0.674097	None	None
$g(u_i, Ontv_{i,j})$,	None	None	None	None	0.297665	0.774528
$Rel(u_i v'_{i,j})$	1.000000	0.937093	0.794983	0.850975	0.517657	1.000000

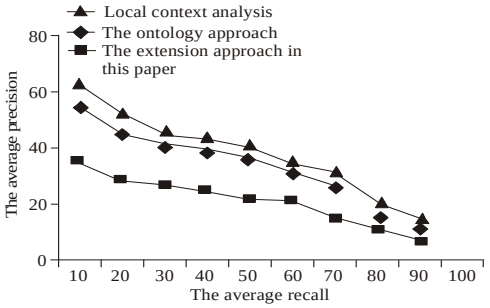

Fig. 4: Comparison among our method, local context analysis and ontology-based query expansion

construction and facilities, museum works and figures and foreign and domestic museums. Data fragments derived in one experiment are shown in Table 1 and 2. In this experiment, the input query term was Yongle Encyclopedia and 10 documents were retrieved. The figures in Table 1 represent the frequencies of occurrence of each expanded word in the most relevant 3 documents; Table 2 lists the calculated data concerning each expanded word.

In terms of recall rate and precision, local context analysis, ontology-based query expansion and the method proposed in this study were compared, thus the results as shown in Fig. 4 were obtained. The experimental results indicated that the application of our method in museum knowledge base had the highest recall rate and precision among all the methods.

CONCLUSION

This study proposes the query expansion method based on ontology and local context analysis. Ontology-based query expansion adds the standardized query words for the retrieval, while local context analysis plays an important role in the improvement of retrieval completeness. Relevance calculation ensures the accuracy of expanded word set. The experimental results indicated

that our method yielded the retrieval results which matched up to the users' demand better.

REFERENCES

Bettina, F., G. Giorgio and G. Georg and L.C. Thomas, 2011. Semantic Web search based on ontological conjunctive queries. Web Semantics, 12: 453-473.

Cho-Wei, S., C. Ming-Yen and C. Hui-Chuan, 2011. Enhancement of domain ontology construction using a crystallizing approach. Exp. Syst. Appli., 38: 7544-7557.

Díaz-Galiano, M.C., M.T. Martín-Valdivia and L.A. Ureña-López, 2009. Query expansion with a medical ontology to improve a multimodal information retrieval system. Comput. Biol. Med., 5: 396-403.

Hsi-Ching, L., W. Li-Hui and C. Shyi-Ming, 2006. Query expansion for document retrieval based on fuzzy rules and user relevance feedback techniques. Exp. Syst. Appli., 31: 397-405

Jinxi, X. and W. Bruce Croft, 1996. Query expansion using local and global document analysis. Proceedings of the 19th annual international ACM, 21: 3414-3418.

Rahmatollah, F., S.W. Concepcion and C. Fletcher, 2008. An alternative approach to natural language query expansion in search engines: Text analysis of non-topical terms in Web documents. Inform. Proc. Manage., 44: 1503-1516.

Saeedeh, S. and S.D. Katebi, 2010. Modeling and evaluation of trust with an extension in semantic web[J]. Web Semantics, Science, 8: 26-36.