Research on Hybrid Query Expansion Algorithm

Zhixiao Wang and Qiang Niu

College of Computer Science and Technology, China University of Mining and Technology, 221116 Xuzhou, Jiangsu, China
{zhxwang, niuq}@cumt.edu.cn

Abstract

This paper proposes a hybrid query expansion method named GAO, which derives from the fact that more and more documents have been annotated with one or several ontology concepts based on their semantic. The GAO method employs a combination of global analysis and ontology technology to improve query expansion performance. The global analysis technology is used to obtain term-concept association, and ontology technology is used to carry out semantic reasoning. Experimental results of query expansion on two different corpuses show that, compared with traditional query expansion methods, the GAO method can improve the precision effectively.

Keywords: *Query Expansion, Global Analysis, Ontology, Term-concept Association*

1. Introduction

Query expansion is the process of adding new meaningful terms to the initial query in order to resolve ambiguities and improve results. There are two main approaches for query expansion: probabilistic query expansion and ontological query expansion. The former is usually based on calculating co-occurrences of terms in documents and selecting terms that are most related to query terms. Probabilistic method can not fundamentally eliminate semantic deviation between user query intention and returned results [1]. Ontological method suggests an alternative approach which uses semantic relations drawn from the ontology to select terms. Ontological method generally supposes that query terms are ontology concepts. However, query term may be ordinary words, rather than ontology concepts. In this situation, how to select appropriate ontology concepts associated with ordinary query terms will become very crucial [2].

With the development of semantic web and ontology technology, more and more web documents are added many kinds of semantic information. Annotating web documents with one or several ontology concepts based on their semantic is one of the most common operations [3]. Terms are included in documents, and documents are labeled with concepts. Thus, the relationship between term and concept is established.

This paper proposes a hybrid query expansion method named GAO. The GAO method employs a combination of global analysis and ontology technology to improve query expansion performance. The global analysis technology is traditionally used to compute co-occurrence of terms. This paper uses it to obtain term-concept association. Ontology technology is used to carry out semantic reasoning. Experimental results of query expansion on two different corpus showed that, compared with traditional query expansion methods, the GAO method can improve the precision effectively.

2. Related Works

Typical probabilistic query expansion includes global analysis, local analysis, local context analysis, etc. One of the successful global analysis techniques is term clustering [4]. Term clustering based on the hypothesis that terms related tend to co-occur in the documents corpus. Other well-known global techniques include Latent Semantic Indexing [5], and Phrasefinder [6]. These techniques use different methods to build a similarity matrix of terms and select terms that are most related to the query terms in that matrix. All global techniques extract terms co-occurrence statistics from the whole document collection. Local techniques extract their statistics from the top-n documents returned by an initial query. Local techniques are based on the hypothesis that the top-n documents are relevant to the query. Local Context Analysis [7] combines both local analysis and global analysis. In LCA, expansion terms are selected not based on their frequencies in the top-ranked documents but rather on their co-occurrences with query terms. Hang et al [8] proposed a method for query expansion based on query logs. Kamps [9] explored a feedback technique that re-ranks the set of initially retrieved documents based on the controlled vocabulary terms assigned to the documents. Huang et al [10] proposed a query expansion algorithm of pseudo relevance feedback based on matrix-weighted association rule mining.

Using ontology for query expansion can be dated up to 1994. Elen Voorhees [11] outlined a query expansion method using WordNet. Navigli and Velardi [12] used sense information and ontology for query expansion. The ontology is used to extract the semantic domain of a word and then the query is further expanded using co-occurring words. They concur with the view that query expansion is suitable for short queries. Fu, G. et al [13] presented query expansion techniques based on both a domain and a geographical ontology. Spatial terms such as place names are modeled in the geographical ontology and non-spatial terms such as “near” are encoded in a tourism domain ontology. Nilsson et al. [14] used a domain specific ontology based on Stockholm University Information System (SUiS) to carry out query expansion. SUiS differs from other question answering systems because it does not allow free-form questions. A detailed review of ontology based query expansion can be found in [2].

Lixin Han and Guihai Chen [15] proposed a hybrid query expansion method called HQE. The HQE method employs a combination of ontology-based collaborative filtering and neural networks to improve query expansion.

This paper proposes a hybrid query expansion method named GAO. GAO method derives from the fact that more and more documents have been annotated with one or several ontology concepts based on their semantic, and the relationship between term and concept is established. Different from HQE, the GAO method employs a combination of global analysis and ontology technology to improve query expansion performance. The global analysis technology is used to obtain term-concept association. The ontology technology is used to carry out semantic reasoning.

3. Hybrid Query Expansion Algorithm

3.1. User Query Mode

We divide user query into three typical modes, namely concept mode (C-mode for short), ordinary term mode (O-mode for short) and hybrid mode (H-mode for short).

(1) C-mode

The C-mode query consists of ontology concepts only. The relationships among ontology concepts, such as synonymous relationship, can be used to carry out query expansion and accurately express the semantic of user query. At present, we use three kinds of expansion techniques for C-mode query. These techniques include synonymous substitution, concept generalization and concept refinement. Synonymous substitution refers to the expansion through synonymous relationship among concepts. Concept generalization refers to the expansion through combination of concept itself and its direct-super-concepts. Concept refinement refers to the expansion through combination of concept itself and its direct-sub-concepts.

(2) O-mode

The O-mode query consists of ordinary terms, rather than ontology concepts. One ordinary term may strongly be correlated with some ontology concept. For example, if the ordinary term “bandwidth” frequently appears in one document, the document usually annotated with ontology concept “Network Measure”. We can use global analysis technology to find these association relationships between ordinary terms and ontology concepts [16], and establish term-concept association thesaurus. For O-mode query, we carry out expansion based on the thesaurus and map ordinary term to related ontology concept.

(3) H-mode

H-mode query consists of both C-mode concepts and O-mode ordinary terms. For C-mode terms, we carry out expansion through synonymous substitution, concept generalization or concept refinement mentioned above. For O-mode terms, we carry out expansion based on term-concept association thesaurus.

3.2. Term-concept Association Computation

With the development of semantic web and ontology technology, more and more web documents are annotated with one or several ontology concepts based on their semantic. Terms are included in documents, and documents are labeled with concepts. Thus, the relationship between term and concept is established. Taking Figure 1 as an example, term t_2 is included in document d_1, and d_1 is labeled by c_1. Thus, term t_2 establishes internal relationship with concept c_1 through document d_1. We can use global analysis technology to compute co-occurrence of term and concept in document corpus, and obtain term-concept association.

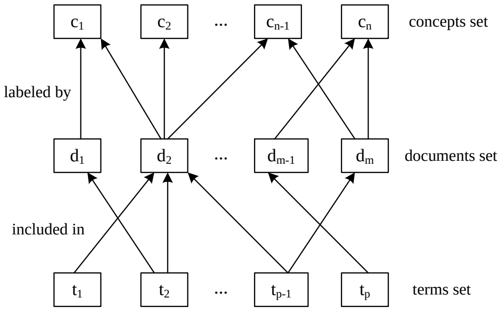

Figure 1. Term, Document and Concept Relationship

This paper puts forward a term-concept association computation formula as follow.

Suppose document set is D, $d_j\,(j=1,...,m)$ is j-th document, m is document number in D; Ontology concept set is C, $c_i\,(i=1,...,n)$ is i-th concept, n is concept number in C; Term set is T, $t_k\,(i=1,...,p)$ is k-th term, p is term number in T. The association degree between term t_k and concept c_i is defined as:

$$association(t_k, c_i) = \log(\frac{n}{n_k} + 1.0) \cdot tf_{k,i} \cdot \log(\frac{num_{k,i}}{num_i} + 1.0) \qquad (1)$$

where n_k is concept number associated with term t_k; num_i is document number labeled by concept c_i; $num_{k,i}$ is document number including the term t_k and simultaneous labeled by concept c_i.

$$tf_{k,i} = \sum_{d_j \in D_i} \frac{count(t_k, d_j)}{len(d_j)} \qquad (2)$$

where D_i is document set labeled by concept c_i, i.e. $D_i = \{d_j \mid d_j \in D \wedge\ d_j$ is labeled by concept c_i $\}$; $count(t_k, d_j)$ is frequency of term t_k appearing in document d_j; $len(d_j)$ is document length of d_j.

4. Experiments and Results

Simulation program implemented with java and relative tools in the Linux environment. Two ontology were used in simulation experiment, one is widely used general ontology called WordNet, the other is domain specific ontology in computer science called ACMCCS98 (ACM Computer Classification 98). In view of above ontologies, we selected two document corpus, one is Reuters Corpus Volume 1（http://trec.nist.gov/data/reuters/reuters.html）, which contains 804414 English language news stories. We call this corpus dataset 1. The other corpus is annotated metadata downed from ACM digital library (http://porta.acm.org/portal.cfm), the data scale is 29030. We call this corpus dataset 2.

In actual practice, many users may only pay attention to first n returned results. Therefore, this paper use $\Pr ecision@n$ to evaluate algorithm performance, and n is set 20.

$$\Pr ecision@n = \frac{\text{relevant documents in top-n retrieved}}{n} \qquad (3)$$

Firstly, we compared three different expansion algorithms: (1) GAO (the hybrid query expansion algorithm proposed by this paper); (2) HQE[15](a typical hybrid method for query expansion); (3) LCA[7](a typical probabilistic query expansion method). We used dataset 1 and corresponding ontology ACMCCS98 as data source. Each method operated 20 queries, including 4 C-mode queries, 4 O-mode queries and 12 H-mode queries.

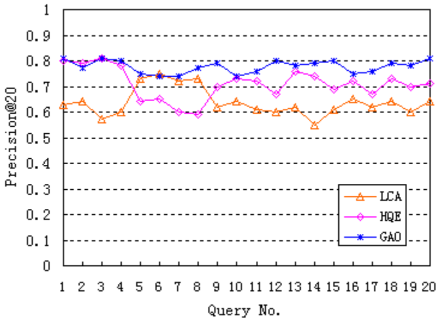

Figure 2. Precision@20 of Three Expansion Algorithms

Figure 2 shows Precision@20 of three expansion algorithms. Query No. 1 to 4 are C-mode queries, and these query terms are ontology concepts. GAO and HQE methods can carry out expansion utilizing ontology knowledge, therefore they have higher Precision@20 than that of LCA. Query No. 5 to 8 are O-mode queries, and these query terms are all ordinary words. HQE method can not utilize ontology knowledge anymore, and performance decreased. GAO method maps ordinary terms to related ontology concepts based on term-concept association thesaurus, and then process user query at concept level. Therefore, the performance of GAO didn't declined under this situation. Query No. 9 to 20 are H-mode queries, and these queries contained both ordinary terms and ontology concepts. GAO method could efficiently process H-mode query, and remained good performance. In actual practice, most user query belongs to H-mode.

Secondly, we carried out experiment to analysis the influence of different ontologies on GAO method. There are two different schemas in the experiment. One schema took WordNet and dataset 1 as data source; the other schema took ACMCCS98 and dataset 2 as data source. For each schema, GAO method operated 20 H-mode queries.

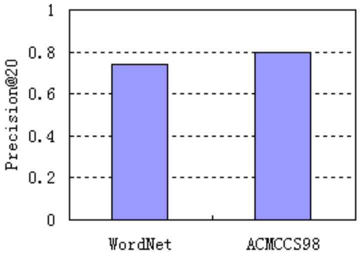

Figure 3. Influence of Ontology

Figure 3 shows the corresponding average Precision@20 of two schemas. There are two factors may affect the Precision@20. One is dataset scale, the other is ontology characteristic. Fang H et al found that dataset scale has little effect on document retrieval performance [17]. Thus, Precision@20 difference comes from ontology factor. ACMCCS98 is domain specific ontology, WordNet is general ontology. Though with simple hierarchical structure, the terminology of ACMCCS98 is less ambiguous and can be expanded with a higher chance of accuracy. Therefore, the performance of ACMCCS98 was better than that of WordNet, as Figure 3 shows. This result is consistent with other research findings [18].

5. Conclusion

This paper proposes a hybrid query expansion method named GAO. The GAO method derives from the fact that more and more documents have been annotated with one or several ontology concepts based on their semantic, and the relationship between term and concept are established. The GAO method employs a combination of global analysis and ontology technology to improve query expansion performance. The global analysis technology is used to obtain term-concept association, and ontology technology is used to carry out semantic reasoning. Experimental results of query expansion on two different corpuses show that, compared with traditional query expansion methods, the GAO method can improve the precision effectively.

References

[1] M. Mitra, A. Singhal and C. Buckley, "Improving automatic query expansion", Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, **(1998)** August 24-28; Melbourne, Australia.

[2] J. Bhogal, A. Macfarlane and P. Smith, “A review of ontology based query expansion”, Information Processing and Management, 43(4):866-886 **(2007)**.

[3] A. F. Zazo, C. G. Figuerola, J. L. A. Berrocal, “Reformulation of queries using similarity thesauri”, Information Processing and Management, 41(5):1163-1173 **(2005)**.

[4] K. S. Jones, “Automatic keyword classification for information retrieval. Butterworths”, London, UK **(1971)**.

[5] S. C. Deerwester, S. T. Dumais and T. K. Landauer, “Indexing by latent semantic analysis”, Journal of the American Society for Information Science, 41(6):391-407 **(1990)**.

[6] Y. Jing and W. Croft, “An association thesaurus for information retrieval. Proceedings of RIAO, **(1994)** October 11-13; New York, USA.

[7] Y. F. Xu and W. B. Croft, “Improving the effectiveness of information retrieval with local context analysis”, ACM Transaction on Information Systems, 18(1):79-112 **(2000)**.

[8] C. Hang, W. Ji-Rong and N. Jian-Yun, “Probabilistic query expansion using query logs”, Proceedings of the eleventh international conference on World Wide Web, **(2002)** May 7-11; Honolulu, Hawaii , USA.

[9] J. Kamps, “Improving retrieval effectiveness by re-ranking documents based on controlled vocabulary”, Proceedings of 26th European Conference on Information Retrieval, **(2004)** April 5-7; Sunderland, UK.

[10] M. Huang, X. Yan and S. Zhang, “Query expansion of pseudo relevance feedback based on matrix-weighted association rules mining”, Journal of Software, 20(7): 1854-1865 **(2009)**.

[11] E. M. Voorhees, “Query expansion using lexical-semantic relations”, Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, **(1994)** July 3-6; Dublin, Ireland.

[12] R. Navigli and P. Velardi, “An analysis of ontology-based query expansion strategies”, Workshop on Adaptive Text Extraction and Mining, **(2003)** Sept 23-25; Cavtat Dubrovnik, Croatia.

[13] Lin Fu, Dion Hoe-Lian Goh and Schubert Shou-Boon Foo, “Evaluating the effectiveness of a collaborative querying environment”, Proceedings of the 8th international conference on Asian digital libraries, **(2005)** Dec; Bangkokj, Thailand.

[14] K. Nilsson, H. Hjelm and H. Oxhammar, “SuiS – cross-language ontology-driven information retrieval in a restricted domain”, Proceedings of the 15th NODALIDA conference, **(2005)** May 20-21; Joensuu.

[15] L. Han and G. Chen, “HQE: A hybrid method for query expansion”, Expert Systems with Applications, 36(4), 7985-7991**(2009)**.

[16] X. Tian, X. Du and H. Li, “Computing term-concept association in semantic-based query expansion”, Journal of Software, 19(8):2043-2053 **(2008)**.

[17] H. Fang, T. Tao and C. X. Zhai, “A formal study of information retrieval heuristics”, Proceedings of the 27th Annual International ACM SIGIR Conference on Research and development in information retrieval, **(2004)** July 25-29; Sheffield, UK.

[18] J. Lin and D. Demner Fushman, “The role of knowledge in conceptual retrieval: a study in the domain of clinical medicine”, Proceedings of the 29th Annual International ACM SIGIR Conference on Research and development in information retrieval, **(2006)** August 06-10; Seattle, WA, USA.

Authors

Zhixiao Wang

He is currently working towards the PhD degree in Tongji University. His research interests include information retrieval and data mining. He has published more than 10 peer-reviewed research papers in journals and international conferences.

Qiang Niu

He received PhD degree from China University of Mining and Technology in 2010. His research interest is data mining. He has published more than 20 peer-reviewed research papers in journals and international conferences.