Available online at www.ijpe-online.com

International Journal of Performability Engineering

vol. 14, no. 7, July 2018, pp. 1421-1430
DOI: 10.23940/ijpe.18.07.p5.14211430

Query Expansion based on Naive Bayes and Semantic Similarity

Zhiyun Zheng, Mengyao Yu, Ning Wang, Xingjin Zhang, Chunyang Ruan, and Dun Li*

School of Information Engineering, Zhengzhou University, Zhengzhou, 450001, China

Abstract

A semantic query expansion method is put forward based on the comprehensive weighted algorithm of semantic similarity. We combine the ontology-based query expansion and corpus-based query expansion. If the query term matches the concept, we calculate the similarity between concepts, construct the connected graph of correlation among the ontology concepts, and expand the semantic query according to the threshold value. Otherwise, we adopt the Naive Bayes algorithm to calculate the co-occurrence probability between the word set and concepts as the relevancy of semantic query expansion. The experimental results show that this method can improve the retrieval performance effectively, with the Pr@30 index being improved by 41.97% compared to the traditional non-extensible query method.

Keywords: semantic query expansion; Naive Bayes; semantic similarity; ontology; corpus

(Submitted on March 28, 2018; Revised on May 5, 2018; Accepted on June 21, 2018)

1. Introduction

The query expansion of Web information retrieval is recognized as the most effective technologies to improve the recall ratio. With the exponential growth of Web data, especially the arrival of the big data age, traditional query expansion methods confront many problems. Among them, prominent problems include keyword-based query expansion bringing some semantic understanding errors, such as synonym problems, ambiguity problems, and variant problems [7]. The development of the semantic web provides a method for resolving these problems, namely semantic query expansion. Semantic query expansion mainly conceptualizes the original query, especially the short-text query, to extract the query semantics of higher accuracy, obtain the expansion concepts and examples from the semantic level through global and partial association rules and query log expansion, and finally achieve better retrieval results through the iterated revision of query keywords [10]. Currently, the semantic query expansion has already become a hot research field, including the vocabulary-based query expansion, corpus-based query expansion, and ontology semantic information-based query expansion [8].

Targeting the query drift in the ontology-based semantic query expansion when the query term does not match the ontology concept, a new semantic query expansion method is proposed by combining the corpus-based semantic query expansion. Meanwhile, in order to improve the low calculating efficiency of the traditional word correlation calculation method in the corpus-based semantic query expansion, the Naive Bayes algorithm is adopted to calculate the world correlation.

2. Related Work

The first information retrieval system emerged in the 1960s, and since then, researchers started to improve the query accuracy and recall rate by extending the vocabulary. Query expansion technology, as an important information retrieval technology, mainly makes use of terms related to the query keywords to modify the query, so that it can describe the information demand of users more precisely [12]. The query expansion technology can compensate the insufficient query information to a certain degree, resolve the mismatching of query terms, and enhance the effect of information retrieval [13]. The semantic similarity calculation is an important method of query expansion, which is established on the conceptual space constructed by the domain ontology. All concepts, corresponding examples, and attributes are organized in a tree-shaped hierarchical structure, to guarantee the accuracy of semantic distance.

* Corresponding author.
E-mail address: ielidun@zzu.edu.cn

Currently, there are two types of word similarity algorithm: one is based on the ontology knowledge, which mainly constructs a dictionary or semantic network of semantic relation for words according to human's understanding of concepts, and measures the word similarity with the relation between concept nodes, the semantic distance, hierarchy depth, density, etc [15]. For instance, Wu and Palmer used the least ancestor concept node to calculate the similarity of two concept nodes on the basis of WordNet [11]; Rada measured the similarity with the shortest path length between two concept nodes [4]; Leacock added the profound impact of concept node on the basis of Rada [2]; Zhang proposed the corresponding similarity algorithm of Chinese words based on HowNet [16]; Ren put forward an improved TF-IDF weight calculation method with the HowNet-based semantic similarity analysis [5]. Another type of the algorithms mainly makes statistics used for the large-scale corpus and measures the word similarity by calculating the co-occurrence probability of two words in the same context. For instance, Ricardo calculated the similarity through the co-occurrence analysis of words [1], and Lin et al. measured similarity with the amount of information shared by two concepts [3]. According to the research analysis, the corpus-based method would conduct correlation processing of the document collection. When the document collection is large, the system overhead may be huge in both time and space. Besides, the corpus requires a process of accumulation, and if it is not updated in a timely manner, the query accuracy would decrease. Compared with the expansion method based on the corpus statistics, the large-scale semantic concept query expansion method does not require the support of corpus or long-term training, and it could realize the expansion from the level of semantic concept. However, many stop words would be added to the concept-tree expansion method. With the increase of the hierarchy of the concept tree, too many query words may lead to useless queries and increase the calculated amount. Besides, if the query term does not match the ontology concept, it would also influence the precision ratio.

Aiming at the above-stated problem, this research is mainly based on the method of combining the ontology-based semantic query expansion and corpus-based semantic query expansion. For the established domain ontology corpus, the Naive Bayes algorithm is adopted to calculate the co-occurrence probability of the common concepts and ontology concepts in the corpus and obtain the relevancy between them. In the process of the query expansion, if query terms are covered in the ontology, the hierarchical structure of ontology would be used for expansion. Otherwise, the relevancy would be taken as the weight for expanding all closely-related ontology concepts.

3. Semantic Similarity Calculation based on the Domain Ontology

The semantic similarity calculation is the core of the semantic query expansion. In this section, traditional information-based, distance-based, and attribute-based semantic similarity calculation methods are improved firstly, and then a comprehensive weighted calculation algorithm of concept similarity in the domain ontology hierarchical structure is proposed as the foundation of the construction of a connected graph for concept correlation in section 5.

3.1. Information-Based Calculation Method

The information-based calculation method mainly determines the similarity between concepts through setting the shared information of two concepts in the ontology [14]. The classic formula of calculating the semantic similarity is shown in Equation (1) [9]:

$$\mathrm{Sim}(a,b)_{\mathrm{ic}} = \frac{2IC(Lcan(a,b))}{IC(a)+IC(b)} \tag{1}$$

In which $Lcan(a,b)$ is the least common ancestor node of concept a and b in the ontology tree, and $IC(a)$ and $IC(b)$ are the information content of concept a and b respectively. The traditional information content calculation of the concept node adopts the corpus-based method, but corpora of different categories and scales have distinct concepts, and the occurrence frequency of concepts is also different. Therefore, it is insufficient to calculate the similarity only using the information content of concept nodes.

In the ontology hierarchy, the less the depth of concept, the higher the occurrence frequency, and then we know that the concept is more abstract and the possessed information content is less. In addition, the greater the areal density of concept computed, the more concrete the division of concept in the region and the larger the information content. While the density and path type are same, the path length between concepts will be shorter and the semantic similarity will be greater.

In conclusion, it is considered in this paper that there are three elements impacting the information of concepts and similarity between concepts: (1) depth of compared concepts in the ontology tree; (2) density of compared concepts in the ontology tree; (3) path length of compared concepts in the ontology tree. Therefore, a method of solution is given by making

use of the information content of the ontology concept, as shown in Equation (2):

$$IC(c)=(\frac{mNode(c)}{mNode(T)}+\frac{log(D(c))}{log(mD(T))})\times(1-\frac{log(h(c)+1)}{log(mNode(T))}) \tag{2}$$

In which $mNode(c)$ is the total number of concept nodes in the sub-tree with the root of concept node c as the direct father node, $MNode(T)$ is the concept nodes number in the ontology tree T contained concept c, $h(c)$ is the children nodes number of concept c, $D(c)$ is the depth of concept c, and $mD(T)$ is the maximum depth of the ontology tree T.

3.2. Distance-Based Calculation Method

The distance-based calculation method mainly calculates the semantic distance of two concept words in the ontology tree according to the geometrical distance between them, as shown in Equation (3):

$$Sim(a,b)=\frac{1}{Dis(a,b)} \tag{3}$$

In which $Dis(a,b)$ is the shortest distance between concepts. Currently, the traditional similarity calculation method based on the semantic distance neglects the impact of edge relation type while different relation types have distinct impacts on the path. In this paper, the edge relation type is added into the calculation process as a weight, and it is shown in Equation (4) considering the three main types of conceptual relation in the ontology tree:

$$Wedge(a,b)=\begin{cases}0.9 & is\ A\\ 0.5 & partof\\ 0.1 & other\end{cases} \tag{4}$$

Combining the WuAndPalmer algorithm [7], an improved calculation model is proposed, as shown in Equation (5) and Equation (6):

$$Sim(a,b)_{dis}=\frac{2\times Path(c,r)+sPath(a,b)}{sPath(a,b)+Path(b,c)+2\times Path(c,r)} \tag{5}$$

$$spath(a_1,a_n)=\sum_{i=1}^{n} wedge_i\times path(a_i,a_{i+1}) \tag{6}$$

In which $wedge$ is the edge weight of adjacent concepts, $path$ is the direct distance between adjacent concepts, $spath(a,b)$ is the shortest weighted path from concept a to concept b, $path(a,c)$ and $path(b,c)$ are the shortest paths from concept a and b to the nearest common node c respectively, and $path(c,r)$ is the shortest path from concept c to root node r.

3.3. Attribute-Based Calculation Method

The attribute-based calculation method mainly calculates the concept similarity using the number of common attributes possessed by concepts. The classic algorithm was proposed by Tversky [6], as shown in Equation (7).

$$\begin{aligned}Sim(a,b)&=\alpha\times Properties(a\cap b)-\beta\times Properties(a-b)\\&-\gamma\times Properties(b-a)\end{aligned} \tag{7}$$

Tversky's calculation method merely measured the similarity with the attributes number shared by concepts. Actually, the structure information of attributes also impacts the calculation of concept similarity; that is, the attributes of the father node must be possessed by the child node, but the attributes of the child node may not be possessed by the father node. In this paper, an improved calculation method is put forward that combines the attribute structural information, as shown in Equation (8):

$$Sim(a,b)_{pro} = \frac{Properties(a \cap b)}{Properties(a \cap b) + \alpha \times Properties(a-b) + \beta \times Properties(b-a)}$$
$$\alpha = \begin{cases} \dfrac{d(a)}{d(a)+d(b)} & d(a) \le b(b) \\ 1 - \dfrac{d(a)}{d(a)+d(b)} & d(a) > d(b) \end{cases}, \alpha + \beta = 1 \quad (8)$$

In which $Properties(a \cap b)$ is the common attribute set of concept a and b, $\Pr operties(a-b)$ is the attribute set of concept a, rather than concept b, and $Properties(b-a)$ is the attribute set of concept b, rather than concept a, and $d(a)$ and $d(b)$ are the depth of concept a and b in the ontology hierarchy respectively.

3.4. Comprehensive Weighted Algorithm of Semantic Similarity

In order to obtain more comprehensive and accurate calculation results, WCA (Comprehensive Weighted Algorithm) is put forward by combining the information-content-based, distance-based, and attribute-based similarity calculation methods with the principal component analysis. This overcomes the subjectivity in the weight setting of the existing comprehensive weighted calculation method.

Suppose the number of concept pairs for comparison is M, let $X_i = (Sim_{ic(i)}, Sim_{dis(i)}, Sim_{pro(i)})$ be a vector in the input sample set of principal components. Among them, each dimensional variable is represented in the result of the semantic similarity calculation in the comprehensive similarity calculation module. Then, the concept semantic similarity matrix is indicated as $x_{sim} = (x_{i1}, x_{i2}, x_{i3})^T (i = 1,2,\cdots,m)$. The principal component analysis is carried out for the constructed semantic similarity matrix x_{sim}. Then, obtain the principal component extracted $Y = (y_{sim1}, y_{sim2}, y_{sim3})$ and the contribution rate of principal components (r_1, r_2, r_3). As a result, the calculation formula of the final concept semantic similarity is shown in Equation (9):

$$Sim_{total} = r_1 \times y_{sim1} + r_2 \times y_{sim2} + r_3 \times y_{sim3} \quad (9)$$

The CWA algorithm is shown as follows:

Algorithm 1: CWA algorithm

Input: node concept a and b, common node concept c, and root node concept r

Output: Sim_{total}

1) For each ontology concept pair do
   Calculate $Sim(a,b)_{ic}$ and $Sim(a,b)_{dis}$
   $Sim(a,b)_{pro}$
   End for
2) Take three calculated similarity elements as three dimensions of the similarity matrix S
3) Conduct S standardized processing to obtain the standard matrix Z and the correlation coefficient matrix R
4) Calculate three characteristic values, and the principal component λ_1, λ_2 and λ_3
5) Solve the set of equations $R \times b = \lambda_i \times b(i = 1,2,3)$, unit feature vector b_i^o
6) The standardized target variable is the principal component $U_i = Z_i^T \times b_i^o (i = 1,2,3)$
7) Weigh the three principal components, sum linearly, and obtain Sim_{total}

4. Naive Bayes' Associated Word Calculation

The context-based correlation analysis relies on a hypothesis: words frequently occurring in documents generally share the same theme, featuring the statistical correlation. Therefore, the context relation obtained through the statistical analysis of document sets can reflect the relationship between these words. While calculating the word relevancy in the corpus, TD-IDF or similar methods are frequently used, but these methods have complicated steps and low efficiency. The Naive Bayes algorithm can capture the uncertainty relation of variables through the posterior conditional probability, characterized robustness, and effectiveness. In this paper, the Naive Bayes algorithm is adopted to calculate the co-occurrence probability and determine the word relevancy. The ideological basis of Naive Bayes is: for the given m terms to be classified $x=\{a_1,a_2,\cdots,a_i,\cdots,a_m\}$, where a_i is a characteristic attribute of x, work out the probability $P(y_j|x)$ of several categories in the set $C=\{y_1,y_2,\ldots,y_k\}$ of k categories in the condition where this term occurs. The term to be classified, whose probability is the greatest, would belong to the category, as shown in Equation (10):

$$If\ P(y_j|x)=max\{P(y_1|x),P(y_2|x),\ldots,P(y_j|x),\ldots,P(y_k|x)\}\ \ then\ x\in y_k \tag{10}$$

In which the conditional probability calculation is crucial, and the formula is shown in Equation (11):

$$p(y_k|x)=\frac{p(x|y_k)p(y_k)}{p(x)} \tag{11}$$

In which, $p(x|y_k)p(y_k)=p(a_1|y_k)p(a_2|y_k)\ldots p(ai|y_k)p(y_k)$, and $p(a_i|y_j)$ is obtained from statistics.

The matching process of concept relevance of the query term and ontology knowledge base can be taken as the classification process of the query term. By referring to the Naive Bayes algorithm, the conditional probability of the occurrence of ontology concept and query term in the corpus is taken as the relevancy of query term and ontology concept. The greater the conditional probability is, the more relevant the two would be. Variables involved in the query term relevance algorithm based on Naive Bayes are shown in Table 1.

Table 1. Variable interpretation

Name of variables	Interpretation of variables
$Count(c_j)$	Occurrences of concept c_j in the documents of the corpus
$Documents$	Number of documents in the corpus
W_f	Feature word set, constituted by w_i $(i=1,2,\cdots)$
N_i	Occurrences of w_i, while concept c_j occurs in the document
$Documents_j$	Number of documents containing the concept c_j
$p(w_f)$	The Probability of words in feature set occurring in the document

The algorithm description is shown as follows:

Algorithm 2: The algorithm of query relative calculation based on the Naive Bayes

Input: candidate word set W, ontology knowledge base C, $Count(c_j)$, $Documents$, N_i, $Documents_j$, $p(w_f)$

Output: $Words\_relation(W_f,c_j)$

1) Calculate $p(c_j)=\frac{Count(c_j)}{|Documents|}$

2) Calculate $p(w_i|c_j)=\frac{N_i+1}{|Documents_j|}$

3) Calculate $Words\_relation(W_f,c_j)=\frac{p(W_f|c_j)p(c_j)}{p(W_f)}$

$$=\frac{p(c_j)\prod_{i=1}^{n}p(w_n|c_j)}{p(W_f)}$$

This method is similar to a global context analysis method, which would analyze each document, count the common words and concepts occurred in the corpus, record the co-occurrence of common words and concepts, and obtain the occurrence rate of common words and each concept in each document. After all documents are processed, the co-occurrence rate of common words and concepts in all documents are combined to get the co-occurrence rate of words and concepts in a global situation. This method calculates the entire query feature set, keeps the semantic association between query terms, and make the calculation much closer to the query intention of users. Meanwhile, the replacement of the common term co-occurrence probability algorithm with the Naive Bayes algorithm has enhanced the efficiency of the batch calculation of feature set.

5. Semantic Query Expansion

In the ontology-based semantic retrieval, the query term or each word of the query term set may not be able to find the matched concept in the ontology knowledge base. Therefore, different semantic expansion strategies are needed to aim at different query terms. The ontology knowledge base describes the related information of some field, including the concept, concept relational structure, attribute and attribute relational structure, examples, etc. The similarity of different concepts is calculated with the CWA semantic similarity algorithm mentioned in section 2, and the correlation strength of different concepts is obtained. In the ontology hierarchy, concepts stand for nodes, and the dependence relation between concepts is the inter-node path. The similarity can be shown on the path intuitively, and then the relevance connected graph of ontology concepts can be obtained, as shown in Figure 1:

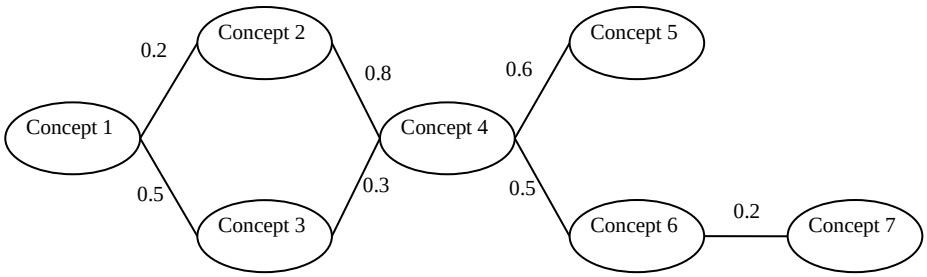

Figure 1. Concept relevance connected graph

Semantic query expansion is carried out according to the ontology connected graph. The expansion method is shown as below: the threshold values are set according to the field; if the query term can match the ontology concept, the concept relevance connected graph should be analyzed, and all concepts whose relevance is greater than the threshold value should be taken out and added to the expanded query term set.

For instance, if the original query term set input by the user contains “concept 4,” it is concluded that four concepts are related to “concept 4” according to the concept connected graph (as shown in Figure 1), and the relevance is 0.8, 0.3, 0.6, and 0.5 respectively. If the relevance threshold value of the expanded query term set in the system is 0.5, then the concept 2, 5, and 6 would be added to the expanded query term set as the target query term set of semantic retrieval.

If the words in the query-term-set fail to find the concept with the same name in the ontology knowledge base, the relevance of words should be calculated according to the associated word based on the Naive Bayes put forward in section 3 to realize the query expansion. In an overall situation, the co-occurrence probability of words and concepts should be calculated with the algorithm, and the relevance of common words and concepts is obtained through the co-occurrence normalization. Then, all ontology concepts that are the most relevant with the query term should be sought according to the relevance. If the relevance is higher than the set threshold value, the corresponding ontology concept should be added to the query term set, and then the query term set can be obtained for the semantic retrieval. The semantic query expansion method is shown as below:

Algorithm 3: Semantic query expansion algorithm

Input: query keyword vector $Q_{key}=\{q_1, q_2,\cdots,q_i,\cdots,q_n\}$, $(i=1,2,\cdots,n)$, in which, n is the vector length of the query keyword, concept c_j

Output: the query term list Q_{exp} after expansion

For (i=1; i<=n; i++)

If q_i matches a certain concept in the ontology knowledge base (indicated by c_j), then

Calculate the semantic similarity between concepts with CWA algorithm, and construct the correlation connected graph

Find the concept set directly related to c_j through the concept correlation connected graph

Select the concept $c_t(t=1, 2,\cdots,m)$ whose similarity is greater than the threshold value θ, and add to the query expanded word set Q_{exp}

Else

Use the Naive Bayes algorithm to calculate the co-occurrence probability of the feature set W_f constituted by the mismatching query term and the ontology concept c_j $(j=1,2,\cdots)$, as the relativity coefficient $Words\_relation(W_f, c_j)$

Select the concept C_t $(t=1,2,\cdots,r)$ whose relevance is greater than the threshold value θ, and add to the query expanded word set Q_{exp}

End if

End for

6. Experiments

6.1. Experiment Data Preparation

The purpose of this experiment is to find the optimal scale of the expanded word set by setting different similarity threshold values and verifying the precision of the comprehensive semantic query expansion method proposed in this paper. An ontology knowledge base is established on the basis of the semantic retrieval prototype system of ontology and corpus in the tourism field. By combining the related data of tourism in the LOD cloud associated data and tourism data resource of Henan Province, the tourism ontology is generated through the protégé. Moreover, 7050 webpage documents on Henan Tourist Information website are taken as the corpus for testing, and Precision@n (Pr@n) is used to evaluate the query precision of this method. The calculation formula of Precision@n is shown in Equation (12):

$$\Pr ecision@n = \frac{r}{n} \tag{12}$$

In which *n* stands for the first *n* result document and *r* stands for the number of related documents in *n* documents. The formula indicates the query precision of the first *n* result documents. Considering that users usually pay attention to the first 30 retrieval results, we let *n*=30 here, and information of the test set is listed in Table 2.

Table 2. Test set

dataset	value
Content	tourism data of Henan Province
Document number	7050
Average document length	326
Word number	41316279
Average unique word number in every document	201

6.2. Experimental Results and Analysis

The experiments mainly apply two different strategies to calculate the relevance for the semantic query expansion according to whether the query term shares the same name with the concept in the ontology knowledge base. Three examples of the initial query term and expanded term are listed in Table 3. The semantic query expansion aims to optimize the query, and the scale of the expanded word set can be controlled through the setting of the similarity threshold value θ. The experiments set different correlation thresholds θ with the comprehensive semantic query expansion method for several tests, and the results show that when the expanded terms are less than 31, the Pr@30 value retrieved presents a rising trend; that is, the precision ratio keeps rising. Otherwise, since noises are introduced into the initial query constantly, the Pr@40 value keeps dropping, which leads to the phenomenon of "query drift." Figure 2 reflects the impact of the number of expanded word sets on the query precision.

Table 3. Examples of query expansion

	Similarity calculation method	The expanded words and its similarity sorted by similarity
Shaoli temple	Semantic similarity calculation	Tallinn（0.760）
		Songshan Mountain（0.740）
		Kung Fu performance（0.610）
May	the Naive Bayes algorithm	Luoyang Peony（0.757）
		Shaolin Temple（0.746）
		Yuntai Mountain（0.743）
Luoyang	Semantic similarity calculation	Longmen Grottoes（0.762）
		Luoyang Peony（0.759）
		White Horse Temple（0.742）

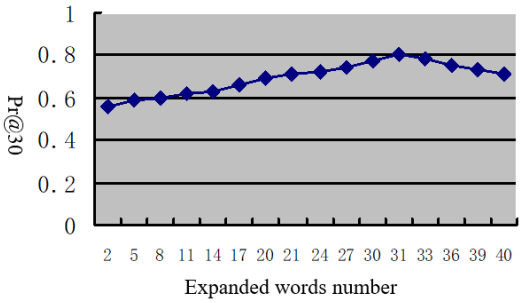

Figure 2. The influence of the extended words number on the query precision

It can be seen by analyzing Figure 2 that when the number of expanded words increases from 27 to 36, the query precision is relatively high. Therefore, the effective threshold θ is set as 0.74 in the experiment, to control the number of expanded words within this range.

To verify the effectiveness of this method, comparison tests are carried out for the ontology-based semantic query expansion (OBSQE) and local context analysis (LCA), and the retrieval performance of the non-extended traditional query is taken as the standard. In this paper, the indexes frequently used in the query performance evaluation of information retrieval are adopted, such as precision, recall, F-measure, and Pr@30. These evaluation indexes are applied to test the query quality of the comprehensive semantic query expansion method proposed in this paper, and the test results are shown in Table 4.

Table 4. The results comparison of three query expansions

query expansion methods	Precision	Recall	F-measure	Pr@30
NETQ（standard）	0.513	0.485	0.454	0.467
OBSQE	0.579	0.486	0.529	0.538
LAC	0.509	0.631	0.487	0.518
Our method	0.639	0.629	0.634	0.663

It can be seen from the experimental results that the comprehensive semantic query expansion method is improved in performance compared with the ontology-based semantic query expansion method and local context analysis method in terms of the precision and recall rate. Specifically, the precision is improved by 24.56%, while the recall, F-measure, and Pr@30 are improved by 29.69%, 39.64%, and 41.97% respectively. A bar chart is constructed for the recall, precision, F-measure, and Pr@30 in Table 3, as shown in Figure 3.

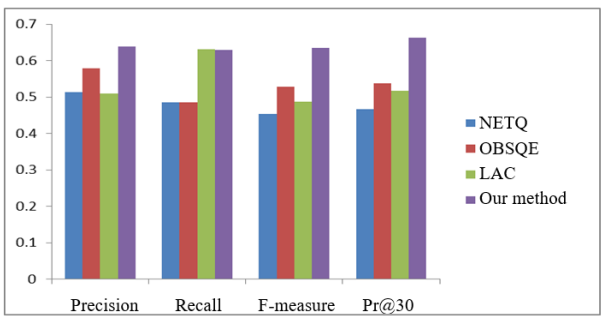

Figure 3. The performance of query expansion methods

It is clear in Figure 3 that the ontology-based semantic query expansion improves the query precision; however, its recall is lower than the standard, since it mainly inquires through the concept expansion in the ontology knowledge base, and it may not be able to find related documents if the word in the query term set fails to match the concept. The query expansion of local context analysis improves the retrieval performance and recall rate, but the big dataset includes some irrelevant expanded words and decreases the precision. The comprehensive semantic expansion method proposed in this paper resolves the defects of the above stated methods, and it not only enhances the query precision but also guarantees the recall rate. Additionally, the top 30 query results also meet the user demand completely. It was shown that the method proposed in this paper has good application value in actual information retrieval. The reasons are analyzed as below: (1) The query term expansion conducted on the ontology knowledge base concept is closer to the user intention, which avoids of noises; (2) While making use of the ontology knowledge base, if it fails to match the ontology concept, it would conduct the relevance calculation by combining the corpus to make the expansion results much more comprehensive; (3) While conducting the corpus-based relevance calculation, the Naive Bayes algorithm is used to replace the traditional method of calculating the co-occurrence rate through the word frequency (e.g. TD-IDF) for the relevance calculation to improve the efficiency of the algorithm.

7. Conclusions

In this paper, a comprehensive semantic query expansion method is put forward according to different queries of users in the semantic retrieval. Firstly, the co-occurrence probability of words and concepts are calculated with the corpus-based Naive Bayes algorithm to get the relevance as the weight. Secondly, the similarity of query terms sharing the same concepts in the ontology knowledge base and other concepts is calculated. Finally, if the query term input is similar to the concept, the expansion would be conducted with the concept relevance connected graph, or the correlation value is used for the expansion. The recall, precision, F-measure and Pr@n of the traditional non-extended retrieval method, the context-based query expansion method, and ontology-based query expansion method are compared in the experiments. According to the retrieval results, our semantic query expansion method performs better than the local context analysis and ontology-based analysis method. This paper resolves the problem that query keywords input by users in the actual semantic retrieval cannot be mapped as the ontology concept and enhances the performance of the retrieval system by improving the relevance algorithm. Reducing the number of irrelevant words in the expanded word set, dealing with the temporal words in the query, and making use of the Top-K ranking are problems to be resolved in the next step of research.

Acknowledgments

The authors are grateful to the editors and reviewers for their suggestions and comments. This work was supported by National Social Science Foundation project (17BXW065), Science and Technology Research project of Henan province (162102310616).

References

1. Y. Baeza, A. Ricardo, and N. Ribeiro. “Modern Information Retrieval,” vol.43, no.1, pp.26–28, 1999.
2. C. Leacock, and M. Chodorow. “Combining Local Context and WordNet Similarity for Word Sense Identification,” *An Electronic Lexical Database*. pp.265-283, 1998.
3. D. Lin. “An Information-Theoretic Definition of Similarity,” *Fifteenth International Conference on Machine Learning*, pp.296-304, Morgan Kaufmann Publishers, 1998.
4. R. Rada, H. Mili, and E. Bicknell. “Development and Application of a Metric on Semantic Nets,” *IEEE Transactions on Systems Man & Cybernetics*, vol.19, no.1, pp.17-30, 1989.
5. Y. P. Ren, L. C. Chen, Y. J. Zhang, and Y. Yuan. “Research of Term Weighting Algorithm Combining Semantics,” Computer Engineering and Design, vol.31, no.10, pp.2381-2383, 2010.
6. A. Tversky. “Features of Similarity,” Readings in Cognitive Science, vol. 84, no. 4, pp. 290-302, 1988.
7. X. Tian, X. Y. Du, and H. H. Li. “Computing Term-Concept Association in Semantic-Based Query Expansion,” Journal of Software, vol. 19 no. 8, pp.2043-2053, August 2008.
8. J. D. Wang, Y. Zhang, and N. Li. “Research and Implementation of Semantic Retrieval Technology based on Ontology,” *Computer Technology and Development*, vol.19, no.10, pp. 134-137, October 2009.
9. T. WANG, L. Wang, J. Y. Wu, and H. Xu. “Semantic Similarity Calculation Method of Comprehensive Concept in WordNet,” *Journal of Beijing University of Posts and Telecommunications*, vol.36, no.2, pp.98-101, 2013.
10. Y. Z. Wang, Y. T. Jia, D. W Liu, X. L Jin, and X. Q. Cheng. “Open Web Knowledge Aided Information Search and Data Mining,” *Journal of Computer Research and Development*, vol.52, no.2, pp.456-474, 2015.
11. Z. Wu, and M. Palmer. “Verb Semantics and Lexical Selection,”. *In Proceedings of Annual Meeting on Association for Computational Linguistics*, pp.133-138, New Mexico, USA, June 1995.
12. Q. L. Yang, T. S. Li, and J. Nong. “Semantic Query Expansion based on Domain Ontology Knowledge Base,” *Computer Engineering and Design*, vol.32, no.11, pp.3853-3856, 2011.

13. Y. H. Yang, J. P. Du, and Y. Ping. "Ontology-based Intelligent Information Retrieval System," *Journal of Software*, vol.26, no.7, pp.1675−1687, 2015.
14. C. Zhang, Y. Yang, and X. Guo. "the Improved Algorithm of Semantic Similarity based on the Multi-dictionary," *Journal of Software*, vol.9, no.2, pp.324-328, 2014.
15. H. Y. Zhang, C. Y. Wen, D. B. Liu, and G. Ye. "Improved Ontology-based Semantic Similarity Computation Algorithm," *Computer Engineering and Design*, Vol. 36, no.8, pp. 2206-2210, August 2015.
16. L. Zhang, C. Y. Yin, and J. J. Chen. "Chinese Word Similarity Computing based on Semantic Tree," Journal of Chinese Information Processing, vol. 24, no. 6, pp.23-31, November, 2010.

Zhiyun Zheng received a Ph.D. degree from the Beijing Institute of Technology in 2005. She is currently a professor of Computing Science and the dean of the Department of Software Engineering in the School of Information Engineering at Zhengzhou University. Her research focuses on cloud computing, semantic network, and large-scale linked data management.

Ning Wang is a graduate student at Zhengzhou University. Her research focuses on large-scale data linked management, cloud computing, and semantic network.

Xingjin Zhang focuses on cloud computing, semantic network, and large-scale linked data management.

Chunyang Ruan is a graduate student at Zhengzhou University. His research focuses on large-scale data linked management, cloud computing, and semantic network

Dun Li received a Ph.D, degree from the Beijing Institute of Technology. Her research focuses on cloud computing, semantic network, and large-scale linked data management.