Available online www.jocpr.com

Journal of Chemical and Pharmaceutical Research, 2014, 6(7): 2767-2774

Research Article

ISSN : 0975-7384
CODEN(USA) : JCPRC5

Query expansion research based on semantic context

Jun-li Luo

College of Information Engineering, Xuchang University, Xuchang Henan, China

---

ABSTRACT

To solve the problem of lower precision caused by traditional query expansion technology, a new query expansion technique based on semantic context was proposed. The semantic context is constructed by WordNet knowledge base and related feedback documents. Firstly, the query words senses are confirmed by disambiguation with WordNet lexical database. Secondly, the initial expansion words are obtained according to the WordNet semantic hierarchy structure. Finally, the weight of the expansion terms is determined according to the overall correlation of candidate expansion terms and all the query words. These words whose weight is higher than weight threshold will be chosen as the final query expansion word. The experimental results show that the proposed method obviously improves the retrieval precision while preserving higher recall.

Key words: information retrieval; query expansion; WordNet; semantic context; ontology

---

INTRODUCTION

In information retrieval, users tend to use a few keywords to conduct the query, short query cannot accurately express the users' query intention, resulting in some of the relevant documents were not retrieved correctly. In addition, due to the ambiguity of natural language, the search terms those users employed maybe different with the index terms that text dataset used, resulting in the related semantic documents cannot be retrieved correctly. These problems hinder the improvement of information retrieval performance. Therefore, query expansion technology which based on semantics is an important research topic in current related information retrieval fields.

Query expansion technology is a kind of new query method which work through selecting the words related with the original query and adding these words to users' query, consisting new long query that fully express the users' query intention; it contains more information to determine the document correlation, it remedies the defect of users' insufficient information. It is an effective method to improve information retrieval performance. There are 3 kinds of traditional query expansion methods:

(1) Method Based on Semantic Knowledge Dictionary: It uses semantic dictionary or domain knowledge base to obtain the original-query-related concept as expansion words. This method selects expansion words from the perspective of semantics concept, to a certain extent, it overcomes the natural semantic ambiguity problem in information retrieval, yet to get the full field dictionary or knowledge base is not an easy thing.

WordNet is a common semantic knowledge dictionary. Liu et al. 【1】proposed a new method based on the semantic relations between concepts in WordNet. After confirming the corresponding concepts of search terms, selecting synonyms, defining words, hyponyms and compound words as query expansion words were conducted. This method improved the retrieval performance under the circumstance without a network connection. Kim 【2】 got the corresponding original words of search terms from WordNet and automatically selected some relevant words from the document, used the original and related words as query expansion words, this method obtains a better experiment effect in large-scale TREC test sets.

(2) Method Based on Partial Text Set: It selects the related documents from the users' initial query result set, then use statistics to form new query from these related documents set. This method makes full use of the information provided by the users, but it is over-reliance on the first retrieved result sets and users' feedbacks. When the initial result set or user feedback is not accurate enough, new query will reduce the accuracy of query on the contrary.

SMART system 【3】 is a typical application of this method in the information retrieval field, which selects related expansion words from relevant document collection and adjusts the weight of new words based on user feedback. The application shows that this system has indeed improved query precision in the small set aspect. Kelly and Teevan 【4】 used query logs to infer the user query intention in order to avoid the users involving in the result feedback directly, this conduct the expansion query automatically based on the relevant information of the initial retrieval document set. Shen【5】based on partial documents set feedback method, proposed the retrieval algorithm which based on statistical model, it rearrange the document order according to the document summary that users clicks on, to obtains the expansion words of relevant document set.

(3) Method Based on Global Information: It surveys the correlation between each word in all documents set of training corpus. When a user conducts query, it selects expansion words according to the correlation between each word, yet to get the field-related training corpus is not an easy thing. Current common global analysis methods include Latent Semantic Indexing and Statistical Dictionary【6】.

Dumais proposed Latent Semantic Indexing model, mapped the keywords in document and query vector to related concepts, based on concept space to calculate and the similarity between documents and query. This model overcomes the limitations based on keyword matching, but the computational cost and space cost are high. Crouch proposed statistical dictionary by analyzing the document contextual words correlation in the entire training set, related words will be organized into different categories, according to the correlation of words to select expansion words.

Through the above analysis, we found that the traditional query expansion methods did not consider the ambiguity of original search terms, blindly selected the extended words. In this way, it will easily cause "topic offset" because of the ambiguity of original search terms. Moreover, when selecting the expansion words, it often based on a single query term of original query, not fully express the users' query intention. To solve these problems, this paper proposes a semantic context-based query expansion method. Firstly, according to the semantic dictionary WordNet to confirm the meaning of search terms, avoiding the query topic offset. Secondly, while selecting extended words, not only considers the semantic correlation of single original query term in WordNet, but also considers the overall semantic correlation of the expansion words and all query terms in the relevance feedback documents. So that the selected expansion words can get closer to the users' query intention and improve the precision of information retrieval.

RELATED WORK

2.1 WordNet

WordNet is an online English dictionary based on cognitive linguistics【7】. It is designed by Princeton University. The basic unit of WordNet is synonym set. Each synonym set represents a potential concept. These synonym set form a concept network through a variety of semantic relationships. In WordNet word frequency refers to the frequency of synonym set in training set. In addition, the synonyms set that has same part of speech are organized into an up and down hierarchy relationship, among these hierarchy relationships, noun relationships account for the proportion of about 80%.

2.2 PageRank Algorithm

PageRank is an algorithm based on graph theory, it is be used to determine the importance of the pages in Google. The link from page A to page B is considered as A vote to B, so the PageRank value of a page is determined by votes and the importance of the voters. Assuming that G=(V，E) denotes a directed graph, V denotes nodes set, E denotes edges set, the PageRank value of node Vi is defined as follow【8】:

$$\Pr(V_i) = (1-d) + d \times \sum_{j \in In(Vi)} \frac{\Pr(V_j)}{|Out(V_j)|} \tag{1}$$

Where $In(V_i)$ refers to the links source nodes set pointing to the node Vi, $Out(V_j)$ refers to the link target node set pointing out Vj. $|Out(V_j)|$ refers to the number of nodes in $Out(V_j)$, and d is a brake factor, its value is in (0,1).

For the initial PageRank value of graph nodes, through circular computing the PageRank value of nodes, the convergence PageRank values will be as final PageRank value of graph nodes.

SEMANTIC QUERY EXPANSION

The semantic context-based method includes three steps: query word sense disambiguation, query expansion terms selection, and query terms weight measurement.

3.1 QUERY WORD SENSE DISAMBIGUATION

(1) Construct Semantic Relation Graph

Given the fact that the user query is brief, the related documents which are obtained from users first search results by implicit feedback technology are used to disambiguation contexts. The graph nodes are composed with the word senses of query words and all the content words contained in disambiguation contexts, all the word senses in WordNet have a corresponding definition. The graph edges are composed with all kinds of semantic relations defined in the WordNet. Undirected edges are established between sense nodes which have corresponding semantic relationship. The strength of semantic relation is used to the weight of edges. If there are several semantic relations between two nodes, the relationship of maximum intensity will be selected as the weight of edge. Thus, the disambiguation semantic relation graph-G is constructed.

(2) Word Sense Disambiguation based on Improved PageRank Algorithm

The original PageRank algorithm is applied to diagraph. In the diagraph, all the weights of edges are equivalent, however, the semantic relation graph is undirected, and the edge weight varies due to the varied strength of relation. So the original algorithm must be modified to apply to the sense disambiguation.

For the improved PageRank algorithm, the importance value of node not only relates with the vote number and the importance of voting node, but also relate with the edge weight of voter. Suppose G=(V, E) is a digraph that has node set V and edge set E, then the PageRank value of node V_i is defined as:

$$\Pr(V_i) = (1-d) + d \times \sum_{j \in link(V_i)} \Pr(V_j) \times \left(\frac{w_{ji}}{\sum_{k \in link(V_j)} w_{jk}} \right) \tag{2}$$

Where w_{ij} denotes the edge weight from node V_i to node V_j; link (V_i) denotes the node set that have a semantic relation with node V_i. d is a revision factor, the value of which is between (0,1).

Iterative calculating the original PageRank value by the improved PageRank algorithm until the PageRank value of graph node is convergence. In the all query word sense node, the sense whose PageRank value is highest will be treated as exact sense.

$$S_{GT}(w, C) = \{S_i \mid \forall S_k \in w, \Pr(S_i) \geq \Pr(S_k)\} \tag{3}$$

If the node with highest PageRank value more than one, these all senses will be retained as the target word sense. In this case, these senses are hard to distinct due to they are so close.

3.2 SELECT QUERY EXPANSION WORD

After the query word sense is determined, the query words will be expanded by taking advantage of the semantic relationship of WordNet. The specific expansions are described as follows.

(1) Select the Candidate Expansion Word

The root node in the WordNet of each query word sense is considered as the common ancestor node. A query semantic tree is constructed; it includes the common ancestor node, query word sense node, their sub-node and these tree structures in WordNet. The query semantic tree can clearly show candidate concepts of user query in ontology concept hierarchy.

In the query semantic tree, the greater of the distance to query word sense node, the weaker semantic relevance between its concept semantic and query word sense. Assuming the weight of query word sense node is 2, and then its sub-concept node weight is 2-L. L denotes distance between sub-concept nodes and their corresponding query word sense nodes, namely the layers of them relative to query word sense node. The weight of these nodes in path of common ancestor node to query word sense node is L. L has a same meaning as before. The weight of common ancestor node and the other parent node of query word sense equals sum weight of each query word sense node related to it. This weight indicate the relevance degree of common ancestor, the other parent node of query word sense, as well as all the query word sense related to it.

In order to clearly understand the process of query semantic tree generation and the candidate expansion terms weights, give the following examples. Assuming the original query vector is $Q = \{q_1, q_2, q_3, q_4\}$, then query semantic tree and its node weight of Q as shown below.

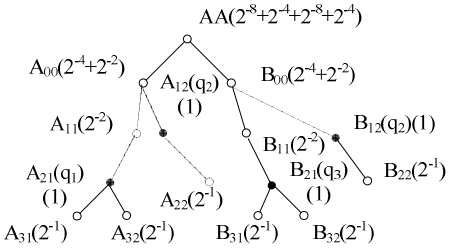

Fig 1 Query Semantic Tree and Node weight

(2) Select the Finally Expansion Words and Calculate the Expansion Word Weight

The goal of semantic query expansion is to find the expansion terms that semantically close to the query Q, not just close to the Q in a separate word. Therefore, it must evaluate semantic related degree of each candidate expansion term and all query items in Q. Related research shows that if the two words often appear together in the same text, then these two words are semantically interrelated, and the higher of the co-occurrence frequency, the stronger of their semantic relation degree. So, in the document sets of first N articles, the co-occurrence frequency of each candidate and all query items in original query is treated as standard which is used to access the overall semantic related degree of candidate expansion words with the original query items.

Rosenfeld has proposed using the average mutual information (AMI) to assess the correlation of words, defined as 【9】:

$$AMI(x,y|S) = p(x,y)\log\frac{p(y|x)}{p(y)} + p(x,\bar{y})\log\frac{p(\bar{y}|x)}{p(\bar{y})} + p(\bar{x},y)\log\frac{p(y|\bar{x})}{p(y)} + p(\bar{x},\bar{y})\log\frac{p(\bar{y}|\bar{x})}{p(\bar{y})} \quad (4)$$

Where $p(x,y)=\frac{c(x,y)}{c(x)}+\frac{c(x,y)}{c(y)}$, $p(x)=\frac{c(x)}{\sum_x c(x)}$, $p(x|y)=\frac{p(x,y)}{p(y)}$. $c(x,y)$ denotes the co-occurrence frequency of word x and y in the same sentence of the training documents set. $c(x)$ denotes frequency of x appears in the training set s. In fact, even in the same window unit, the related degree between words will decrease exponentially with increase the distance between the words. So we add a factor $e^{Space(x,y)}$ to express this nature. The similarity of word x and y is re-defined as follow:

$$SIM(x,y|S) = \frac{AMI(x,y)}{e^{Space(x,y)}} \quad (5)$$

where $Space(x,y)$ denotes the average words number between word x and y in the all window unit.

Suppose each query word in the original query Q is independent of each other. The semantic correlation of candidate expansion word w and Q is calculated according to logarithm of sum, which obtained by adding the average mutual information value between w and each query item q belongs to Q. The formula is defined as follow.

$$Cohd(w,Q|S)=\log\left(\sum_{q\in Q}(idf(w|C)\times idf(q|C)\times SIM(w,q|S)+1)\right) \quad (6)$$

where $idf(x|C)=\log\frac{N}{df(x|C)+1}$, $df(x|C)$ denotes the documents number where word x appears in corpus C. N denotes documents number of corpus C. joining $idf(x|C)$ is to avoid some unimportant words getting a higher $Cohd$ value. These word tend to have a higher frequency, the worst case is stop word. Adding 1 is to avoid the value equaling 0.

When finally calculate the weight of candidate expansion word, it doesn't only consider the semantic weight of candidate expansion word in the query semantic tree, but also consider its semantic related degree with all the query items of Q. The calculate formula is defined as:

$$Weight(w) = QSTWeight(w) \times Cohd(w, Q \mid S) \tag{7}$$

Where $QSTWeight(w)$ denotes weight of expansion word in the query semantic tree. Setting threshold θ, these word that satisfy $Weight(w) > \theta$ will be treated as final query expansion words.

EXPERIMENTAL SECTION

4.1 Data Set and Evaluation Criteria

Experiments were performed to justify the effectiveness of proposed method. Five different data sets were selected, including NLP, CACM, CISI, Medline and Cranfield【10】. These data sets have different topic contents, document numbers and query numbers. Each data set contains the documents to retrieve, the initial query, the list of document ID associated with each initial query. The number of documents, query subject and number of each data set are shown in Table 1.

Table 1 Statistics on the Data Sets

Data Set	Topic	Number of Documents	Number of Queries	Size(MB)
Medline	Medical	1033	30	1.1
Cranfield	Aerospace Engineering	1400	225	1.6
CISI	Information Science	1460	112	2.2
CACM	Programming Algorithm	3204	64	2.2
NPL	Electronic Engineering	11429	93	3.1

In order to verify whether the proposed method improves the accuracy of information retrieval, the average accuracy (MAP) is introduced to evaluating the experiment sets. The MAP is defined as follows:

$$AP = \frac{1}{R}\sum_{i=1}^{R}\frac{i}{Doc(i)} \qquad MAP = \frac{1}{Q}\sum_{j=1}^{Q} AP_j \tag{8}$$

The R represents the document number related to the query in a single query, represents the ranking of the ith related document, AP represents the average accuracy of a single query, Q represents the total number of query, MAP represents the weighted average of average accuracy of all the query problems. At the same time, the recall is also adopted to evaluate the algorithm performance.

4.2 Experiment process

(1)The first retrieval to get relevant feedback documents

Computing the similarity of each query and the documents to retrieve, and descending order according to the similarity, the first N documents will be selected as a query word disambiguation context. To facilitate the processing, vector space model is adopted to retrieve the relevant feedback document. The similarity of query and document is computed by Angle cosine of vector.

In order to reduce the dimension of vector space, each document of data sets only keeps their body parts. The characteristics dictionary of each data set is constructed through the Participles, removing the stop and extracting stem etc. The weights of query and document are calculated by the widespread method of TF-IDF.

(2)Determine the query word sense

The content words in relevant feedback documents are as context of query word sense disambiguation, determining the corresponding WordNet sense node of the query words and the content words in the context. To calculate the PageRank value of each query word sense according to the formula (2), the final query words senses are confirmed according to formula (7).

(3)Calculate the expansion word weight and form the expanded query vector

To determine expansion words according to the expansion word weight formula given in section 3, and then the final query vector is formed according weight formula. Removing the relevant feedback documents of data sets, the remained documents as to retrieve documents, the similarity of expansion query vector and to retrieve documents is calculated.

4.3 Experimental results

The concept tree is constructed according to WordNet.Net interface provided by WordNet.

The number of query expansion word will directly impact the query precision. More expansion word will improve

the recall at the same time reduce the precision. Therefore, the threshold θ of expansion word weight is tested in experiment. Experimental results are shown in figure 2.

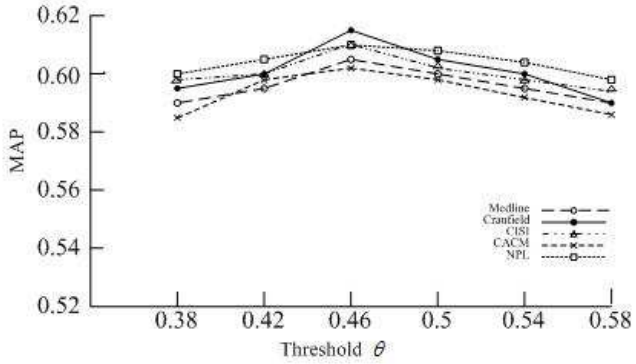

Fig 2 Results for the expansion words threshold θ

The experimental results show that the MAP peaked on five data sets when the expansion words threshold θ is about 0.46. If θ too low, it means the query expansion is not sufficient, and users' query intention is not fully express, so the MAP is lower. Besides, if θ too high, some nose terms will be add to expansion express, so the MAP is also lower.

The finally expansion words are selected by calculating the average mutual information of candidate expansion words with all the original query items. So query performance will be compacted by the number of relevant feedback documents in first retrieve. The experimental result is shown in figure 3. The MAP in figure is average of five data sets.

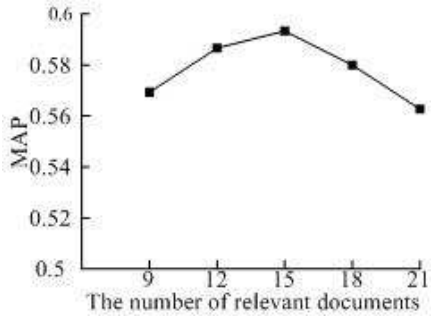

Fig.3. Results for the number of relevant feedback documents

The experimental results showed that the MAP on five data sets obtained best value when the number of relevant feedback documents is about 15. The remained documents removing relevant feedback documents from data sets are used as validation documents of query expansion. Not all the queries are accord with the requirement of verification expansion algorithm. If all the related documents are as relevant feedback documents, there will be no remaining relevant documents to verify the expanded query obviously such a query doesn't conform to the requirements of the experiment. If relevant feedback documents set doesn't have relevant document, it will not be able to provide feedback information, such a query doesn't conform to the requirements of experiment as well. Therefore, if a relevant feedback documents set contains at least 3 related documents, and at least 5 remaining related documents are not retrieved in first retrieve, these queries will be selected to verify query expansion algorithm.

Table.2. The comparisons of MAP value on each data set

Data Set	No Expansion Method	Semantic-based Method	Statistical-based Method	Proposed Method
Medline	0.341	0.561	0.478	0.613
Cranfield	0.362	0.568	0.510	0.620
CISI	0.320	0.515	0.463	0.611
CACM	0.315	0.497	0.449	0.626
NPL	0.353	0.537	0.481	0.642

On the basis of the former two tests, setting expansion words threshold θ as 0.46 and relevant feedback document number as 15, the MAP values of proposed method, semantic-based method, statistical-based method and no

expansion method are compared respectively. The results are shown in Table 2. The recalls of these four methods are also compared respectively. The result is shown in Table 3.

Table.3. The comparisons of recall on each data set

Data Set	No Expansion Method	Semantic-based Method	Statistical-based Method	Proposed Method
Medline	0.682	0.710	0.702	0.793
Cranfield	0.679	0.702	0.731	0.821
CISI	0.711	0.736	0.728	0.772
CACM	0.625	0.727	0.758	0.789
NPL	0.691	0.746	0.739	0.806

The comparison results in table show that the proposed method gets a higher precision meanwhile improving the recall, followed by semantic-based method, statistical-based method and no expansion method. For each data set, the increase proportions are not same, relative no expansion method, the MAP and recall of proposed method respectively increased by 84% and 19%, the MAP and recall of semantic-based method respectively increased by 58% and 8%, the MAP and recall of statistical-based method respectively increased by 41% and 9%.

DISCUSSION

Experimental results show that the proposed method has higher precision than traditional query expansion methods. This is because the expansion words are selected without considering the query words senses in traditional methods. When original query words are ambiguous, these methods will easily add too much noise expansion words, so that these traditional technology is unsatisfactory to improve precision. The user query semantics are deduce by query words senses disambiguation using WordNet before query expansion in proposed method, thus avoiding the drift phenomena of the query expansion. In addition, when selecting expansion words, the proposed method not only considers the relevance of expansion words and single query word, but also considers the whole relevance of expansion words and all the query words. Therefore, the expansion words can fully express the user's query intention, and the precision of proposed method is better than the traditional query expansion methods.

CONCLUSION

Based on semantic context, a query expansion algorithm is proposed in this paper. According to query word sense disambiguation, the query expansion words are selected by combining the query semantic tree with local feedback technology. Experiments indicate that the proposed method improve the retrieval precision as ensuring the recall. In order to evaluate the validity and stability of this algorithm, this algorithm will be applied to practical application.

Acknowledgments
The author would like to sincerely thank office colleagues' valuable suggestions. This research was supported by the development plan of science and technology project in the city of Xuchang (1404017).

REFERENCES

[1] Liu S, Yu C, and Meng W, Word Sense Disambiguation in Queries, in CIKM'05: Proceedings of the 14th ACM International Conference on Information and Knowledge Management, pp. 525-532, **2005**.
[2] Kim SB, Seo HC, Rim HC. Information Retrieval using Word Sense: Root Sense Tagging Approach. In: Proc of Conf on SIGIR'04. Sheffield, pp. 258-265, **2004**.
[3] Salton G. The SMART Retrieval System-Experiments in Automatic Document Processing. Englewood Cliffs: Prentice Hal Inc Press, pp.337-354, **1971**.
[4] Kelly D, Teevan J. Implicit feedback for inferring user preference. Journal SIGIR Forum, **2003**, 37(2): 18-28
[5] Shen XH, Tan B, Zhai CX. Context-Sensitive Information Retrieval Us Implicit Feedback. In: Proc of Conf on SIGIR'05. Salvador, pp.43-50, **2005**.
[6] Claudio Carpineto, Giovanni Romano. A survey of automatic query expansion in information retrieval. ACM Comput. Surv., 44(1),pp.1–50, **2012**.
[7] Gadge J R, Sane S S, Kekre H B. Query expansion using WordNet in N-layer vector space model[C]//Engineering (NUiCONE), 2013 Nirma University International Conference on. IEEE, pp.1-5, **2013**.
[8] Haddadene H A, Harik H, Salhi S. On the Pagerank Algorithm for the Articles Ranking[C]//Proceedings of the World Congress on Engineering. pp. 4-6, **2012**.
[9] Kjersten B, Van Durme B. Space efficiencies in discourse modeling via conditional random sampling[C]//Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, pp.513-517,**2012**.

[10] Rahutomo F, Kitasuka T, Aritsugi M. Test collection recycling for semantic text similarity[C]//Proceedings of the 14th International Conference on Information Integration and Web-based Applications & Services. ACM, pp. 286-289. 2012.