A Semantic Based Query Expansion To Search

Mozhgan Shabanzadeh, University of Isfahan, Iran
Mohammad Ali Nematbakhsh, University of Isfahan, Iran
Naser Nematbakhsh, University of Isfahan, Iran

ABSTRACT

Keyword based information retrieval has difficulties in retrieving relevant information because it is not able to include the semantics of queries. In this paper, we present a method for query expansion based on semantic relations. In our proposed algorithm, related words to the query are extracted from WordNet, and a semantic network of query terms and extracted words is constructed. Spreading activation selects candidate terms for query expansion from the constructed semantic network. Finally, candidate words that don't cause ambiguity and noise in the query are selected. The experimental results based on the TIME data set show that the proposed query expansion method outperforms keyword based information retrieval.

Keywords: semantic search, query expansion, spreading activation algorithm, WordNet, word grouping, semantic relations, semantic network

INTRODUCTION

Currently, the amount of information is increasing dramatically. A sense can be expressed in various ways, for example, all synonym words express a sense. So that, information repositories are becoming more diverse and the language used in documents is more and more dissimilar. Users express their information need in words that might have different means (Fang, & Zhai, 2006).

Query expansion technology appends related words to query and overcomes word mismatch and improves retrieval. So that through using query expansion, documents which are on the same matter can be found even if they do not contain the original query terms (Billerbeck, 2005; Navigli, & Velardi 2003).

There are two main methods for query expansion techniques based on their implementations, in literature: global analysis or corpus specific query expansion and local analysis or query specific query expansion (Aly, 2008).

In global analysis method, new terms are added to an original query before searching. This method needs external resources such as thesaurus, WordNet, and etc. In this method correlation between every pair of terms in the whole document set is pre-calculated and maintained. When users submit a query, the system selects the most related terms according to the pre-calculations and expands the query (Bhogal, Macfarlane, & Smith 2007). Lexical based (Voorhees 1994; Parapar, Barreiro, & Losada 2005; Kwon, Kim, Choi 1994),
statistical based (Bai, Song, Bruza, Nie, & Cao, 2005; Xu & Croft 2000; Qiu, & Frei 1993), hybrid based (Han, Sun, Chen, & Xie 2006; Pekar & Staab 2003) and log based (Cui, Wen, Nie, & Ma 2003) are examples of global analysis method.

Lexical based approaches use lexical thesaurus to select expansion words. Statistical based approaches use statistical measures such as co-occurrence measures or lexical co-occurrence measures in documents to select expansion terms. It should be pointed out that the pre-calculations are very time consuming since requiring the analysis of the entire document collection (Andreou 2005). Hybrid based approaches combine both statistical and lexical approaches. Log based approaches use click behavior of users from system log to select expanded words. In other words, they mine logs to find associations between a query and the documents seen by the user.

In local analysis method, a new query is formulated on the basis of some retrieved documents of search with original query (Qiu, 1995). The correlation between two terms is calculated according to the top N documents in the first retrieval result or obtained by user feedback. Relevance Feedback and Pseudo Relevance Feedback (Rocchio 1971; Croft, Cronen-Townsend, & Lavrenko, 2006) which is developed on the basis of Relevance Feedback are examples of local analysis method. The performance of local analysis method significantly depends on the precision of the top results (Li 2007).

Query expansion has some problems. The most important of them is query drift, which is moving the query away from the user’s intention (Mitra, Singhal, & Buckley 1998; Stenmark 2005). This happens frequently when the query is ambiguous. For example, the query “apple” might be about apple as fruit or apple computers. Selecting words that are different than the user’s intention reduces the precision.

Outweighting is a specific kind of query drift and is well described in (Mahler, 2003). Outweighting is occurred when the augmentation terms are strongly related to individual query terms but not to the whole query.

An important matter in query expansion is the ability of expanded words to distinguish between documents. Suitable words are those that can retrieve more relevant documents and don’t retrieve irrelevant results (Bhogal, Macfarlane, & Smith 2007).

The method in this paper belongs to the global analysis method. However, we have three important improvements in the query expansion. First, it groups query terms based on their semantic similarity, and expands each group on words that show the relationship between group words, so in this method selected words aren’t related to only an individual query term. Second, this method avoids selecting vague and noise words to expand the query. Third, this method assigns appropriate weights to query terms by using spreading activation algorithm.

In our approach, we utilize the semantic similarity between words to group query words. The purpose is to extract words that are related to all group words instead of each query term from WordNet (Fellbaum 1998). We construct a semantic network of query words and related words to the query and further employ spreading activation algorithm to determine candidate expand words, at last filter the candidate words by the number of senses that they have to select valuable words for expansion.

This approach has experimentally displayed significant improvements in search precision and recall rate. The remainder of this paper is organized as follows. Section 2 presents the related work.
Section 3 describes our detailed methodology for semantic based query expansion. Section 4 presents experimental results and discussions. Finally the last section presents the conclusions.

RELATED WORK

Query expansion has a long history in information retrieval research, since the correctness of search results is dependent on the query keywords while the user’s queries are often short and vague (Yang, Yang, & Yuan 2007; Kumaran, Allan, 2008). In most cases, users don’t reach their intentions with their first queries and have to change them to close to their goals (Rose, & Levinson 2004). The aim of query expansion is to convert these weak queries to powerful queries. Various methods of how expand a query to improve performance have been discussed in literature. In this section we give a brief review on semantic based query expansion methods which are mostly based on global analysis algorithms and are close to our approach.

(Voorhees 1994) used WordNet to expand queries, and the results showed that individual queries can benefit from query expansion, but query expansion makes little difference in retrieval performance for long queries. (Gong, Wa Cheang, & Hou 2006) used semantic similarity to group query words and for each group, expanded the highest word in WordNet with hypernym and synonym, and the lowest one with hyponym and synonym, and all other words with synonym. They could improve the performance of information retrieval.

(Li 2007) expanded nouns with synonym and hypernym and verbs with words which have closely related sense to it. (Hsu, Tsai, & Chen 2006) used ConceptNet, which is a commonsense knowledgebase, and WordNet to expand the queries. WordNet is more efficient than ConceptNet in query expansion, but using these two resources simultaneously result in better performance.

In (Liu, Y., Li, Ch., Zhang, P., & Xiong, Zh. 2008) related words to noun phrases in the query are extracted from top results of original query and added to the query. In (Lee, Tsai, & Wang, 2007) remove stop words from query and augment the query with synonyms, construct the user intention tree to determine the user’s intention, and expand the query with user’s intention.

PROPOSED QUERY EXPANSION ALGORITHM

This section describes the proposed approach for query expansion. We present a query expansion algorithm based on semantic relations. We use WordNet to extract semantic relations between words. WordNet is a lexical database developed by Princeton University to model the lexical knowledge of a native speaker of English (Fellbaum, 1998), and is one of the most important semantic resources in information retrieval.

Figure 1 shows our algorithm. It at first disambiguates the submitted query, removes stop words, and extracts query keywords. Then groups the query words based on their semantic similarity, and extracts words that are related to all group words instead of each query word. After, constructs a semantic network of query words and related words to the query. In this network words are as its nodes and each node is linked to nodes that are semantically related to it; for example two synonym nodes are linked. Further, runs spreading activation algorithm to determine candidate words for expansion, at last filters the candidate words by the number of senses that they have to remove vague words which cause retrieving more irrelevant documents. In the following, we elucidate the details of proposed algorithm.
**Preprocessing**

Queries are in natural language form and in this phase, the query terms are stemmed, and each query term is disambiguated through assigning a WordNet sense. Finally stop words are removed from query. We use the WordNet-SenseRelate-AllWords\textsuperscript{ii} to disambiguate query words. WordNet-SenseRelate-AllWords package does word sense disambiguation by measuring the semantic similarity between a word and its neighbors.

**Word Grouping**

In case of multi term queries, it is not suitable to add terms that are related to only individual term of original query. Because result in outweighting which is described in introduction. So that in this cases only those terms should be selected that are related to whole query or a subset of query words. For example, in this query: “computer monitor”, as illustrated in figure 2 monitor and computer have the same hypernym: device and are correlated. In this query the intention of user is neither computer nor monitor. If the query augmented by words that are related to only computer or monitor, query will be away from user’s intention and the precision of results will be reduced.

In this phase we use stemmed query keywords, and do try to determine correlated words of query to avoid outweighting problem. We use the proposed approach in (Gong, Wa Cheang, & Hou 2006) to assign
query terms to different groups. In (Gong, Wa Cheang, & Hou 2006) words are grouped based on their semantic similarities. We use WordNet-Similarity package to calculate the semantic similarity between each two query words. If two words be semantically similar, then they will be in a group, other than they will be in different groups.

![Diagram of WordNet](image)

*Figure 2 Relationship between words in WordNet*

**Network Construction**

Following previous phase, a set of groups produces. This phase then constructs a network of query terms and their related words. For each group which contains more than two words, the words that show the relationship between the words are extracted from WordNet, and for other groups which contain only one word, if the word is a noun, then its synonyms, hyponyms and hypernyms are extracted from WordNet. This algorithm doesn’t expand adverb, verb and adjective in groups that contain only a single term to avoid producing a large number of expanded words. In the last example; “computer monitor” that is shown in figure 2; these two words are in a group and “machine, device, electronic, display” senses that show the relationship between them are added to semantic network.

Query terms and all extracted words are added to network as its nodes. At the beginning the weight of query terms is 1, and other nodes’ weight is 0. Two nodes that are semantically related are linked, so that the structure of this network is like to WordNet.

**Performing Spreading Activation**

Spreading activation was proposed by (Quillian 1968; Collins, & Loftus 1975) as a method to search associative networks, neural networks, or semantic networks. The search process is initiated by activation a set of source nodes and then their activation iteratively spread out to other nodes linked to the source nodes until some termination specification is met. Activation spreading process is only iterated for nodes that their activation value is greater than a threshold, called firing threshold.

Here is an illustration of how the spreading activation algorithm works. In the network that is shown in figure 3, the activation value of node A is 1, so this node is active and the algorithm starts. The activation value of this node spreads in the network. When the algorithm terminates every node in this network have a weight. The weight of each node calculated by this formula:

\[ W_B = W_B + W_A * W_{AB} * D \]

*Equation 1*

Where \( W_B \) is the weight of B, \( W_A \) is the weight of A, \( W_{AB} \) is the weight of link between nodes A and B, and D is the decay factor to reduce the weights of network nodes.
In figure 3 the weight of each link is 0.9, decay factor is 0.6, and firing threshold is 0.1. The output of the algorithm is shown in figure 4.

After network construction and activating query term nodes through labeling them with weight 1 in previous phase, spreading activation algorithm is run on this network. At the end, nodes that their weight is greater than firing threshold regarded as candidate words for expansion. We experimentally set the firing threshold in our implementation equals to 0.15.

Filtering the Candidate words

If a candidate word which is produced in the previous phase for query expansion has many senses, adding it to query results in reducing performance. Because such that word is vague and adding it to the query, makes the query noisy. Adding a vague word into the query causes in retrieving irrelevant documents. So that this phase removes words that have many senses, and only adds valuable words to the query. We obtain the threshold for number of senses through experiments.

EXPERIMENTAL RESULTS AND DISCUSSION

In order to test our approach, we have first used WordNet-SenseRelate-AllWords package to disambiguate queries and used the disambiguated queries in the rest, then expand the queries, and retrieve results, and analyze retrieved documents to calculate precision and recall.

It should be mentioned that we use vector space model (VSM) (Salton, Wong, & Yang, 1975) to retrieve documents. VSM represents documents and queries as vectors of terms. The similarity between documents and queries measured by inner product between corresponding term vectors.

The Dataset

We used one test data set of information retrieval from the SMART archive at the Computer Science Department of Cornell University, namely TIME. It is a
collection of 1963 Time Magazine news articles which contains 425 short articles and 80 queries. These queries are natural language sentences. This experimental data set was originally designed for the examination of automatic indexing and document retrieval methods.

**Evaluation Measures**

There are two important measures to consider for evaluating information retrieval systems: Precision and recall. As defined in Equations 2 and 3 precision is the proportion of the number of the relevant retrieved documents to the number of the retrieved documents, and recall is the proportion of the number of the relevant retrieved documents to the number of the relevant documents.

\[
\text{Precision} = \frac{|\text{relevant retrieved documents}|}{|\text{retrieved documents}|} \quad \text{Equation 2}
\]

\[
\text{Recall} = \frac{|\text{relevant retrieved documents}|}{|\text{relevant documents}|} \quad \text{Equation 3}
\]

**Results and Discussion**

In our proposed algorithm, there are some parameters that should be determined, such as the number of senses that each word has as filtering threshold and the spreading activation algorithm parameters include: the weight of links between nodes, decay factor and firing threshold. To determine the best values for these parameters, we select average precision as objective function. In other words, the optimal values for these parameters maximize average precision.

In our proposed algorithm only nodes that their weights are greater than firing threshold will be select as candidate words for query expansion. Based on equation 1 the weights of links and decay factor have a straight effect on the weights of nodes, so we do some experiments to determine the optimal values for them.

There are synonym, hypernym and hyponym links in the constructed semantic network. Table 1 is the performance of average precision via the weights of links. From this table it is obvious that retrieval performance reaches its maximum when the weight of synonym relations is 0.7 and the weight of hypernym/hyponym relations is 0.3.

**Table 1 Average precision versus weights of links**

<table>
<thead>
<tr>
<th>Weight of Synonym relations</th>
<th>Weight of Hypernym/Hyponym Relations</th>
<th>Average Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3</td>
<td>0.7</td>
<td>0.230</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>0.275</td>
</tr>
<tr>
<td>0.7</td>
<td>0.3</td>
<td>0.290</td>
</tr>
<tr>
<td>0.8</td>
<td>0.2</td>
<td>0.284</td>
</tr>
<tr>
<td>0.9</td>
<td>0.1</td>
<td>0.215</td>
</tr>
</tbody>
</table>

Decay factor should be set properly because if its value be low, the weight of nodes will be reduced and therefore a lot of nodes will be discarded in query expansion and if its value be high a lot of nodes will be regarded as candidate word for expansion and may make the query noisy.

**Table 2** is the performance of average precision via the decay factor. From this table it is obvious that retrieval performance reaches its maximum when the decay factor is 0.7.
### Table 2 Average precision versus decay factor

<table>
<thead>
<tr>
<th>Decay Factor</th>
<th>Average Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8</td>
<td>0.273</td>
</tr>
<tr>
<td>0.7</td>
<td>0.290</td>
</tr>
<tr>
<td>0.6</td>
<td>0.261</td>
</tr>
<tr>
<td>0.5</td>
<td>0.186</td>
</tr>
</tbody>
</table>

Firing threshold determines active nodes that iterate the spreading activation algorithm, and is the threshold for selecting candidate words from semantic network. Therefore, regarding it as a low value results in selection of unsuitable words, and regarding it as a high value results in elimination of valuable words.

Table 2 is the performance of average precision via the firing threshold. From this table it is obvious that retrieval performance reaches its maximum when the decay factor is 0.15.

### Table 3 Average precision versus firing threshold

<table>
<thead>
<tr>
<th>Firing threshold</th>
<th>Average Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.45</td>
<td>0.260</td>
</tr>
<tr>
<td>0.21</td>
<td>0.281</td>
</tr>
<tr>
<td>0.15</td>
<td>0.290</td>
</tr>
<tr>
<td>0.10</td>
<td>0.278</td>
</tr>
</tbody>
</table>

As in our discussion of filtering the candidate words section, number of each candidate word senses for augmenting the query is important. In this section, we are going to determine the optimal number of senses threshold for selecting words. This threshold should be determine exactly, because a large number results in selecting vague words and retrieving more irrelevant documents, and a little number results in losing valuable words and missing relevant documents.

Figure 5 is the performance of average precision via number of senses threshold. From this figure it is obvious that retrieval performance reaches its maximum when threshold is at 6, so that only words that have less than 6 senses regarded as expansion words.

![Figure 5- Average precision versus Sense counts](image)

As in figure 6, we have used expansion method and no expansion in our experiments in order to compare their effectiveness. The descriptions of these methods are in below.

No Query Expansion: use original queries without any expansion for test.

Weighted Query Expansion: use candidate words obtained from proposed algorithm with regarding weights. In this method the words in original query (query keywords) have the highest weights and words which are more related to original query have more weights than less related words.

As we expect, the performance of Weighted Query Expansion is more than No Query expansion. Because the Weighted Query Expansion to some extent solves the
mismatch problem in keyword based information retrieval. In addition, it avoids query outweighting problem in query expansion through assigning appropriate weights to words in expanded query. As shown in figure 6, Weighted Query Expansion results are more relevant than No Query Expansion results in top retrieved results.

![Figure 6- Precision-Recall diagram](image)

**CONCLUSIONS AND FUTURE WORKS**

In this paper, we propose a method for semantic query expansion. We construct a network from query terms and related words to the query, and use spreading activation algorithm to select candidate words and assigning appropriate weight to query terms. In order to avoid noise words in query expansion, we remove those words that have so many senses in WordNet and are vague. This method can improve the retrieval performance.

This research contributes to improvement of information retrieval in several ways. First, it demonstrates that semantic relations can improve query expansion. Second, this research illustrates that vague words reduce the performance of information retrieval. Third, this research emphasizes that it is important to assigning appropriate weights to query terms.

As future works, we will determine the phrases in the query to select better words for query expansion, determine queries that don’t benefit from query expansion and find a solution for them, and test our approach on another dataset. Furthermore, taking into account user’s interests and personalizing semantic query expansion is a good avenue for future work into making query expansion more powerful.

**REFERENCES**


personalization and recommender systems
digital libraries, 49–54.
Query expansion by mining user logs. IEEE
Transaction on Knowledge and Data
Engineering, vol. 15, no. 4, 829-839.

Fang, H., & Zhai, Ch. (2006). Semantic
Term Matching in Axiomatic Approaches to
Information Retrieval. Proceedings of the
29th Annual International ACM SIGIR
Conference on Research and Development
in Information Retrieval (SIGIR'06), 115-122.

Electronic Lexical Database, The MIT Press

Gong, Z., Wa Cheang, C. & Hou U., L.

Han, L., Sun, L., Chen, G., & Xie, L.
semantic similarity. Advances in egnineering

Hsu, M.H., Tsai, M. F., & Chen, H.H.
(2006). Query Expansion with ConceptNet
and WordNet: An Intrinsic Comparison,

information retrieval systems to user
queries. In: Information Processing and

Kwon, O.W., Kim, M.C., & Choi, K.S.
(1994). Query Expansion Using Domain
Adapted Thesaurus in an Extended Boolean

A Practical Ontology Query Expansion
Algorithm for Semantic-Aware Learning
Objects retrieval. Computers &amp;
Education, 1240–1257.

Based On Semantic Element. Eighth ACIS
International Conference on Software
Engineering, Artificial Intelligence,
Networking, and Parallel/Distributed Computing, IEEE.

Liu, Y., Li, Ch., Zhang, P., & Xiong, Zh.
(2008). A Query Expansion Algorithm
based on Phrases Semantic Similarity.
International Symposiums on Information
Processing, IEEE.

Mahler, D. (2003). New Directions in
Question Answering, Chapter 24 – Holistic
Query Expansion using graphical models.

Mitra, M., Singhal, A., & Buckley, C.
(1998). Improving automatic query
expansion. In Proceedings of the 21st
annual international ACM SIGIR conference
on research and development in information
retrieval, 206–214.

Navigli, R., & Velardi, P (2003). An
analysis of ontology-based query expansion
strategies. In Proceedings of the 14th
European Conference on Machine Learning,
Workshop on Adaptive Text Extraction and
Mining.

Parapar, D., Barreiro, A. & Losada, D.
(2005). Query expansion using Wordnet
with a logical model of information
retrieval, IADIS International Conference
(Algarve (Portugal), Feb 2005), IADIS'2005.

classification based on combined measures
of distributional and semantic similarity. In
Proceedings of the Research Note Sessions
of the 10th Conference of the European
Chapter of the Association for
Computational Linguistics, 147–150.

based on a similarity thesaurus. ETH Zurich.
Query Expansion. In Proceedings of ACM
SIGIR, 160-169.


---

i [http://wordnet.princeton.edu](http://wordnet.princeton.edu)

ii [http://www.d.umn.edu/~tpederse/senserelate.html](http://www.d.umn.edu/~tpederse/senserelate.html)

iii [http://www.d.umn.edu/~tpederse/similarity.html](http://www.d.umn.edu/~tpederse/similarity.html)