Conceptual Query Expansion

Orland Hoeber, Xue-Dong Yang, and Yiyu Yao

University of Regina, Regina, SK S4S 0A2, Canada
Orland.Hoeber@uregina.ca

Abstract. Query expansion has been extensively studied as a technique for increasing information retrieval performance. However, due to the volume of documents available on the web, many of the techniques that have been successful in traditional information retrieval systems do not scale well to web information retrieval. We propose a new technique based on conceptual semantic theories, in contrast to the structuralist semantic theories upon which other techniques are based. The source of the query expansion information is the concept network knowledge base. Query terms are matched to those contained in the concept network, from which concepts are deduced and additional query terms are selected. In this paper, we describe the theoretical basis for this in-progress research, along with some preliminary results.

1 Introduction

Query expansion is a technique for dealing with the word mismatch problem in information retrieval systems. In general, the word mismatch problem is a result of different terms being used in reference to a single concept, both in the documents and in the user queries[2]. Query expansion is the process of adding additional terms to the original query in order to improve retrieval performance [1]. Through query expansion, the effects of the word mismatch problem are reduced, resulting in a higher ratio of relevant documents in the retrieval results (precision) and a higher ratio of the relevant documents from the collection that are retrieved (recall).

Recent studies have shown that users have difficulties choosing good terms to add to their query, even when presented with a list of potentially good terms [8]. Therefore, we focus on automatic query expansion techniques. The general process for automatic query expansion begins with matching the terms from the user query to a knowledge base. From this knowledge base, the best new terms are added to the query automatically, and the expanded query is then used.

A number of different techniques for automatic query expansion have been proposed in recent years. In work by Qiu & Frei [7], a similarity thesaurus was constructed based on the similarity of term pairs across the complete document collection. New terms to be added were based on similarity to all rather than individual query terms. In Voorhees' system [10], the queries were expanded using the term relationships encoded in the WordNet thesaurus [5], a lexical-semantic knowledge base. The relationships used were the synonym, hypernym/hyponym

and meronym/holonym relationships among nouns. Xu & Croft [12] used the top ranked documents returned by the original query as the knowledge base for the query expansion. In this technique, the co-occurrence of terms was calculated using only the passages that contained the query terms, rather than the whole document.

Information retrieval of web documents poses a number of problems for these query expansion techniques. Due to the extremely large volume of documents on the web, analysis of the entire collection (e.g. [7]) is not feasible. In addition, web queries are very short, often consisting of only two or three words [9]. Techniques that have reported success with longer queries (e.g. [12]) may not prove to be very effective with short queries. A third difficulty is that the collection of web documents is very general, necessitating a very general knowledge base. However, a knowledge base which is too general (e.g. [10]) can result in terms being added to the query that are not actually related to the user's query terms in the document collection.

Our approach is to use a concept network to generate a conceptual query expansion for web information retrieval. This is ongoing research being conducted as part of a larger research project consisting of meta-searching, clustering, and visual representations of the search results based on the concepts intended by the user's queries, rather than just the specific terms in the query.

Briefly, a concept network is a bipartite graph consisting of two classes of nodes: concepts and phrases. Such a concept network can be constructed through the statistical analysis of concept-phrase co-occurrence in a concept hierarchy such as the Open Directory Project [6]. Within the concept network, weighted edges between the concepts and the phrases represent the degree to which the phrase has shown to be a good indicator of the concept. Query expansion is performed by selecting additional phrases from those that are connected to the same concepts as the user's query phrases. Weight thresholds ensure that only concepts that are strongly connected to the user's query phrases are selected, and only phrases that are strongly connected to these concepts are included in the expanded query.

By basing the construction of our concept network knowledge base on a human-reviewed subset of the entire collection of web documents (i.e, the Open Directory Project), we avoid the complications introduced by the size of the web, and the generality of the collection. The short queries that are common in web information retrieval remain a problem; there may not be enough evidence in a short query to clearly indicate a single concept. Our larger research project addresses this problem by allowing users to interactively refine their query in a visual manner.

2 Theoretical Basis for the Concept Network

Semantics, the study of the meanings of words in a language, provides many theories that attempt to explain the meaning of a word or phrase. Of interest

for query expansion are the structuralist theories of meaning and the conceptual theories of meaning [4].

The structuralist theories hold that in order to understand the meaning of a word, one has to understand how it functions together with, and in contrast to, other related words. The meaning of a word consists of the relationships it has with other words in the language (i.e., synonyms, homonyms, polysems, antonyms, etc.). In order for two people to communicate, they must have a similar understanding of how the words they use are related to other words in the language [4].

In the conceptual theories, the meaning of a word is the concept in the mind of the person using that word; communication is possible via a shared connection between words and concepts. That is, there is a mapping between the set of words in the vocabulary and the set of concepts in the minds of the users. While there will not be a single unique mapping that all users of a language share, the individual mappings must be similar to one another with respect to the vocabulary used in order for communication to occur [4].

Both of these theories infer the meanings of words used by an individual with respect to other objects. That is, the semantics of words are not provided by definitions, reference to instances, or their use, as we are normally accustomed; rather the meaning of words are provided by their relationship to other objects. These theories differ in the type of objects to which words are related (i.e., other words in the structuralist theories, higher-level concepts in the conceptual theories). We note that with the conceptual theories, one can infer a relationship between a pair of words if there is a concept which they have in common. The level of detail of this common concept provides a clue to the degree to which the terms are related: terms linked through a high-level (general) concept may have a weak relationship, whereas terms linked through a low-level (specific) concept may have a strong relationship.

Previous work on query expansion, both those that rely on a general thesaurus as the basis for the query expansion [10], and those that rely on the construction of a specialized thesaurus from the text of the corpus being searched [7, 12], follow the structuralist theories. Queries are expanded to include additional terms that are related to the original terms in the thesaurus. The success of these techniques varies depending on whether the query expansion is manually chosen or automatic, as well as the specificity of the thesaurus used.

Other theoretical approaches for formally specifying concepts in terms of intensions and extensions, such as formal concept analysis [3], have a basis in the conceptual theories. For example, applying formal concept analysis to information retrieval, the intension of a concept is the set of terms that are present in all the documents that represent the extension of the concept. Terms do not have a direct relationship to one another; they have an implied relationship through a common concept. However, such approaches are vulnerable to the word mismatch problem, and the size of the collection of documents in web information retrieval.

In this research, we hypothesize that the mappings between words and concepts proposed by the conceptual theories provides a more effective connection between words (via a common concept) than the thesaurus-based approach of the structuralist theories. Basing the query expansion on such a mapping results in a conceptual query expansion where the basis for the expansion are the concepts intended by the original query.

3 The Concept Network

We define a concept network as a weighted bipartite graph that links concept nodes to phrase nodes. More formally, a concept network $CN = \{C, P, E\}$ consists of a set of concept nodes C, a set of phrase nodes P and a set of edges $E = \{c_i, p_j, w_{ij}\}$, where $c_i \in C$, $p_j \in P$, and c_i and p_j are related with a weight w_{ij}. The weight w_{ij} of an edge in the concept network represents the degree to which the phrase represents the intension of the concept.

While hand-crafting a concept network is possible, it is only feasible for a small set of concepts and phrases. However, given a well defined set of concepts, and a set of documents assigned to these concepts, a concept network can be automatically constructed as shown below. During this process, a bag-of-words approach is used to count the occurrences of noun phrases within each document. The noun phrase frequency is used to calculate the weight values w_{ij}.

For example, consider the set of documents $D_i = \{d_{i1}, \ldots, d_{in}\}$ which are a subset of the extension of the concept $c_i \in C$. For each document d_{ik}, the set of phrases used in this document is $P_{ik} = \{p_{1,ik}, \ldots, p_{m,ik}\}$. We define a function $f(d_{ik}, p_j)$ as the occurrence count of phrase p_j in document d_{ik}. The value for the edge weight between concept c_i and phrase p_j is given by:

$$w_{ij} = \frac{\sum_{k=1}^{n} \frac{f(d_{ik}, p_j)}{\sum_{l=1}^{m} f(d_{ik}, p_{l,ik})}}{n}$$

After all the concepts for which we are given document sets have been analysed, we normalize the edge weights from each phrase in the concept network. For a phrase p_i that is connected to r concepts whose index is given by the relation $f(x)$, $x = 1 \ldots r$, the normalization is performed via the simple calculation:

$$w_{ij} = \frac{w_{ij}}{\sum_{k=1}^{r} w_{if(k)}}$$

Using the normalized average noun phrase frequency rather than a simple total of the noun phrase occurrences reduces the impact of the different document sizes and different numbers of documents that represent a concept in the calculation of the weight values. In particular, without this calculation, a single large document could provide phrase weights that overshadow the weights provided by a number of smaller documents; a similar situation is avoided for concepts that have a large number of documents. Further, without the normalization, common phrases that are included in many documents for many concepts would have a

very high weight value, even though these phrases are of little value in describing the concept. With normalization, the weights for these common terms will be significantly reduced.

We note that the automatic construction of a concept network represents the training phase of this system. In our preliminary research, we used subsets of the Open Directory Project as the training data to construct preliminary concept networks automatically.

4 Conceptual Query Expansion Using the Concept Network

Given a concept network $CN = \{C, P, E\}$, and a query $Q = \{q_1, \ldots, q_n\}$ consisting of query phrases q_i, the process of constructing a query expansion is as follows:

1. Match the query phrases q_i to the phrase set P to obtain a $P' \subseteq P$.
2. Obtain the set of concepts $C' \subseteq C$ which are connected to the phrases in P'. We use two parameters to control this operation: a weight threshold w_e, and an phrase ratio PR. First, all the concepts that are connected to the phrases in P' with a weight greater than w_e are chosen as candidate concepts in C'. Each of these concepts are then evaluated to determine the ratio of the phrases in P' to which they are connected with a weight greater then w_e. If this ratio is less than PR, the candidate concept is dropped from C'.
3. Obtain the set of phrases $P'' \subseteq P$ which are connected to the concepts C'. We use a weight threshold parameter w_d to control this operation. All phrases that are connected to the concepts in C' with a weight greater than w_d are chosen as the phrases in P''.
4. Perform a union of the original query phrases and the new set of phrases to obtain the query expansion: $QE = Q \cup P''$.

For example, consider the concept network in Figure 1, which was constructed automatically from a subset of the Open Directory Project, as described in the previous section. In this figure, the concepts are represented by the shaded boxes, and the phrases are represented by the ovals. In the interest of clarity, distance is used to represent the edge weights, rather than displaying the weight values; phrases with very low weights (i.e., very common phrases that are used in the documents of many concepts) are excluded from this figure. Suppose we are given a user query $Q =$\{"information", "visualization", "problems", "software", weight thresholds $w_e = 0.05$ and $w_d = 0.1$, and phrase ratio of $PR = 75\%$.

In the first step, this user query is matched to the phrases in the concept network to arrive at $P' =$\{"information", "visualization", "software"\}. In the second step, we follow only the edges that have a weight greater w_e to obtain our set of candidate concepts $C' =$\{"computer graphics", "distributed computing", "artificial intelligence"\}. The candidate concept "computer graphics" is connected to 100% of the phrases with a weight greater than w_e; "distributed

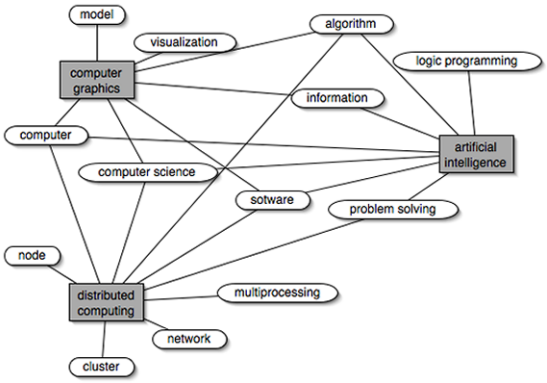

Fig. 1. A sample concept network constructed automatically from a subset of the Open Directory Project.

computing" is connected to 33%; and "artificial intelligence" is connected to 66%. Since these last two concepts have phrase ratios less than our phrase ratio parameter PR, they are dropped from the set, resulting in $C' = \{$"computer graphics"$\}$.

In the third step, we follow the edges that have a weight greater than w_d to obtain our phrases that are related to the concepts in C', resulting in $P'' = \{$"computer", "model", "visualization"$\}$. In the final step, we merge the original query and these concept phrases, giving us $QE = \{$"information", "visualization", "problems", "software", "computer", "model"$\}$.

The end result is a query expansion that has included phrases only from the concepts that the system decided were related to the original query. Such a query will be much more specific to the deduced concepts than the original query. We anticipate that this query will result in a set of documents that are more relevant to the actual information need of the user. Research regarding the effectiveness of this system, along with determining appropriate settings for the parameters w_e, w_d, and PR under various conditions, is ongoing.

5 Conclusion

In this paper, we presented an overview of our ongoing research on the use of a concept network as the knowledge base for inducing a query expansion based on the concepts deduced from the original query terms. We acknowledge that the quality of this conceptual query expansion depends on the quality of the concept network; we are working towards the automatic construction of a large, high quality concept network using the Open Directory Project as the source of concept extensions.

Future work includes determining appropriate initial values for the parameters that control the conceptual query expansion process, measuring the performance of conceptual query expansion using test collections such as the TREC

Web Track collections [11], using conceptual query expansion in a meta-search and clustering system, and using the concept network as a basis for the visualization of search results.

References

1. Efthimis N. Efthimiadis. Query expansion. *Annual Review of Information Systems and Technology (ARIST)*, 31, 1996.
2. G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais. The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11), 1987.
3. Bernhard Ganter and Rudolf Wille. *Formal Concept Analysis: mathematical foundations*. Springer-Verlag, 1999.
4. Cliff Goddard. *Semantic Analysis: A Practical Introduction*. Oxford University Press, 1998.
5. George A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11), 1995.
6. Open Directory Project. A human-edited directory of web pages, http://www.dmoz.org/, 2004.
7. Yonggang Qiu and H. P. Frei. Concept based query expansion. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 1993.
8. Ian Ruthven. Re-examining the potential effectiveness of interactive query expansion. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 2003.
9. Amanda Spink, Dietmar Wolfram, B. J. Jansen, and Tefko Saracevic. Searching the web: the public and their queries. *Journal of the American Society for Information Science and Technology*, 52(3), 2001.
10. Ellen M. Voorhees. Query expansion using lexical-semantic relations. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 1994.
11. Ellen M. Voorhees. Overview of trec 2003. In *The Twelfth Text REtrieval Conference (TREC 2003)*. National Institute of Standards and Technology Special Publication, 2003.
12. Jinxi Xu and W. Bruce Croft. Improving the effectiveness of information retrieval with local context analysis. *ACM Transactions on Information Systems*, 18(1), 2000.