Expanding Query Terms in Context

Chris Staff and Robert Muscat

Department of Computer Science and AI,
University of Malta

Abstract. Query expansion is normally performed using a thesaurus that is either generated from a collection of documents, or is otherwise language specific. We present a technique to discover associations between query terms that are synonyms based on past queries and documents common to multiple result sets, to enable query expansion to occur in context.

1 Introduction

Furnas describes the main obstacle to improved recall in Information Retrieval as the Vocabulary Problem [3], which arises from the small probability that authors and searchers of a concept describe that concept using the same terms.

Attempts to improve recall frequently utilise a manually or automatically constructed thesaurus [7, 4, 8]. The query is normally expanded prior to its submission and query expansion involves identifying as many alternative new terms as possible that express the same concept. When a term has many different word senses, choosing the correct synonym set can be challenging, as indiscriminate automatic query expansion may lead to a loss of precision [8].

When precision is low it is frequently because the terms specified by the user in the query do not have sufficient discriminatory power to distinguish between relevant and non-relevant documents containing the terms [6]. On the other hand, recall may be low because there may be several different terms that are used throughout a document collection to describe the concept in which the user is interested. However, the user has expressed fewer terms in the query than there are to describe the concept [1].

We explore a different approach to query expansion that assumes that the presence of the same document in the results sets of different queries indicates that some terms in the different queries may be synonyms. If we discover that they are, using WordNet [5], then we can construct a synonym set that will be used to expand a query when the results set of the original, unexpanded, query contains the document. This method allows us to discriminate between alternative candidate synonym sets when a term is ambiguous.

2 Scenario

Assume that two users searching for information related to the same concept C express the queries Q_1 and Q_2 respectively. Assume that no term is common to both Q_1 and Q_2. Let R_1 and R_2 be the results sets for Q_1 and Q_2 respectively. Let R_{common} be the intersection of R_1 and R_2. Furthermore, let R_{common} be small but non-empty. R_{common} is the set of documents that are relevant to both user queries. Following [2] and [3], we would expect R_{common} to be small if different terms in the original queries potentially describe the same concept but only a few documents contain both

terms. Incidentally, if R_{common} is large, then it means that the majority of documents in the results set of either query contain both terms, in which case users using either term in a query will retrieve approximately the same set of documents. In this case, adding the other term to the query will not improve recall (though it may improve precision). However, if R_{common} is small, and we can demonstrate that a term in Q_1 is a synonym of a term in Q_2, then in future queries that include either term we can successfully expand the query to include the other term to improve recall. Unlike thesaural expansion techniques, we also require that a document in the results set of the future query was previously seen in R_{common} before we expand the terms.

Furnas recommended that an *adaptive index* is constructed by associating terms that users use with the terms that the system knows about [2]. We use terms supplied by users through queries to discover how the same concept may be described using different terms in other documents and user queries. Our assumption is that different users use different terms in their queries to describe the same concept; that there are documents in the collection that can satisfy each independent query; that some of these documents will contain more than one description of the same concept; and that we are able to automatically identify and associate these alternative descriptions, using, for example, WordNet [5].

3 Approach

We assume that a separate vector space-based information retrieval system provides document indexing and retrieval services. When a user query is submitted, the 100 top-ranking documents in the results set retrieved by the information retrieval system are processed to increment the number of times the document has ever been retrieved, and the number of times the document has been retrieved following a query containing each of the terms in the query. The rank at which the document is retrieved is also recorded. We keep a "bag of words" that contains all the terms that have ever been used to retrieve the document. Before adding a new term q_i from the query to the "bag of words" for document d_n we consult WordNet to check if the term is a synonym of any word/s w_j in the "bag of words". For each pair q_i, w_j that are synonyms, we update the synonym sets for w_j, d_n and q_i, d_n, adding q_i and w_j respectively. This gives us synonym sets for w_j and q_i in the context of document d_n. We also record the word category and sense number of the terms in the synonym set (obtained from WordNet).

To expand some term q_i in a query Q, we first submit Q to obtain the initial ranked results set R. For each document d_k in R, where k is the rank of the document, we retrieve the synonym sets for q_i, d_k along with the document's Inverse Document Relevance (IDR) for term q_i (see next section). The IDR represents the relative frequency with which the document d_k is retrieved in rank k when term q_i occurs in the query Q. The synonym set that will be selected for q_i is based on a re-ranked results set based on each document's Document Relevance Weight.

4 Document Relevance Weight

There may be multiple documents in a query's initial results set R that provide conflicting synonym sets for a term t in the query. If t is ambiguous or has many word senses, then we need to be able to select the most likely synonym set, otherwise we will provide the user with poor results. Let $\mathcal{W}_d$ be the number of times that document d has appeared in any results set, and $\mathcal{W}_{t,d}$ be the number of times that a document d has appeared in the results set of a query containing term t. A document has a rank r in a results set, indicating its relative relevance to the query. Let $\mathcal{W}_{d,r}$ be the number of times that document d has ever been in rank level r in a results set. For simplicity,

we take the significant levels to be top 5, top 10, top 25, top 50, and top 100 (r_5, r_{10}, r_{25}, r_{50}, r_{100}, respectively). Similarly, $\mathcal{W}_{t,d,r}$ is the number of times that document d has occupied rank level r in the results set of a query containing term t.

We calculate $IDR_{t,d}$, a document's inverse document relevance for term t, as $\mathcal{W}_{t,d}$ / $\mathcal{W}_d$. This gives us the relative frequency with which the document appears in a results set because of the appearance of term t in the query. For instance, if $\mathcal{W}_d$, the total number of times d has appeared in a results set, is 1000, and $\mathcal{W}_{t,d}$, the number of times d appeared in a results set of a query containing term t, is 10, then $IDR_{t,d}$ is 0.01.

A Term-Document Relevance score is $TDR_{t,d,r} = IDR_{t,d} \times \mathcal{W}_{t,d,r}/\mathcal{W}_{d,r}$ where r is the rank level of document d in the results list. We then re-rank the documents in R in the order of their term-document relevance scores. The synonym sets of the top 10 newly ranked documents are then merged according to word category and word sense. The synonym set selected to expand the query term is that of the most frequently occurring word category, word sense pair of the synonyms sets of the top 10 ranked documents.

For example, following a query Q with the results set R, we retrieve the synonym sets S for each q_i, d_j, where $q_i \in Q$, and $d_j \in R$, for the top-10 ranked documents in R. A synonym set is a tuple $category, sense, [syn_0, syn_1, ..., syn_n]$. The synonym set belonging to the most frequently occurring category and word sense combination are used to expand query term q_i.

5 Discussion

The ideas presented here are exploratory. Our approach is potentially more discriminating than the typical approach to query expansion, because we associate terms expressed by users with documents, and then we use the documents present in the result set following an initial query to select an appropriate synonym set for each term in the query. In limited trials, the same term occurring in different queries is frequently expanded with a different synonym set in each query. We intend to evaluate the performance using standard performance measures, and experiment with limited sharing of synonym sets within a community of users.

References

1. H. Chen, T. D. Ng, J. Martinez, and B. R. Schatz. A concept space approach to addressing the vocabulary problem in scientific information retrieval: an experiment on the worm community system. *J. Am. Soc. Inf. Sci.*, 48(1):17–31, 1997.
2. G. W. Furnas. Experience with an adaptive indexing scheme. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 131–135. ACM Press, 1985.
3. G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais. The vocabulary problem in human-system communication. *Commun. ACM*, 30(11):964–971, 1987.
4. R. Mandala, T. Tokunaga, and H. Tanaka. Combining multiple evidence from different types of thesaurus for query expansion. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 191–197. ACM Press, 1999.
5. G. A. Miller. Wordnet: a lexical database for english. *Commun. ACM*, 38(11):39–41, 1995.
6. M. Mitra, A. Singhal, and C. Buckley. Improving automatic query expansion. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 206–214. ACM Press, 1998.
7. Y. Qiu and H.-P. Frei. Improving the retrieval effectiveness by a similarity thesaurus. Technical Report 225, Zürich, Switzerland, 1995.
8. E. M. Voorhees. Query expansion using lexical-semantic relations. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 61–69. Springer-Verlag New York, Inc., 1994.