KNOWLEDGE MEDIA INSTITUTE

Exploiting Semantic Association To Answer Vague Queries

Technical Report KMI-06-01
February, 2006

Jianhan Zhu, Marc Eisenstadt, Dawei Song, Chris Denham

To appear in Proc. of The Fourth International Conference on Active Media Technology (AMT 2006), June 2006, Brisbane, Australia.

Exploiting Semantic Association To Answer 'Vague Queries'

Jianhan Zhu, Marc Eisenstadt, Dawei Song, Chris Denham
Knowledge Media Institute, The Open University, United Kingdom
{j.zhu, m.eisenstadt, d.song, c.m.denham}@open.ac.uk

Abstract. Although today's web search engines are very powerful, they still fail to provide intuitively relevant results for many types of queries, especially ones that are vaguely-formed in the user's own mind. We argue that *associations* between terms in a search query can reveal the underlying information needs in the users' mind and should be taken into account in search. We propose a multi-faceted approach to detect and exploit such associations. The CORDER method measures the association strength between query terms, and queries consisting of terms having low association strength with each other are seen as 'vague queries'. For a vague query, we use WordNet to find related terms of the query terms to compose extended queries, relying especially on the role of least common subsumers (LCS). We use relation strength between terms calculated by the CORDER method to refine these extended queries. Finally, we use the Hyperspace Analogue to Language (HAL) model and information flow (IF) method to expand these refined queries. Our initial experimental results on a corpus of 500 books from Amazon shows that our approach can find the right books for users given authentic vague queries, even in those cases where Google and Amazon's own book search fail.

Keywords. Query expansion, similarity, association strength, semantic space.

1. Introduction

Search engines are getting increasingly powerful and popular, yet user-frustration with even the best and most popular engines is not an unusual experience. A BBC Radio 4 interview (BBC, December 2004) with booksellers described many common-sense but 'loose' queries that the bookseller could answer, but which made absolutely no sense to today's search engines. For instance, one bookseller described a customer looking for 'that book about a guitar-playing sergeant', which neither Google nor Amazon could locate (answer: (Captain) Corelli's Mandolin).

The root cause of the disparity between such common-sense queries and the keyword approach of today's engines is this: a user's search queries are often an *approximation* and synopsis of his/her information needs, so purely matching against the terms in the search query is a woefully inadequate method for finding the correct or even correlated information. In contrast, human memory can leverage the *associations* between these terms in order to understand what the user really wants to find. These terms may be vague and not even be correct for the specific type of information, especially since the user may only have a rough memory of the target information or be unfamiliar with the mechanism of search engines. In previous work, Anderson and Bower [6] proposed a human associative memory model.

It is clear that in many cases the *combination* of terms in a query may tell a coherent story about the user's information needs. For example, in the 'guitar-playing sergeant' example above the word "sergeant" is related to the word "captain" (they are both a kind of military officer according to WordNet) and the word "guitar" is related to the word "mandolin" (they are both a kind of stringed instrument). For such queries, which we henceforth call *vague queries*, using keyword matching cannot find the correct information. We need to find the related terms of "sergeant" as "captain" and "guitar" as "mandolin" respectively to compose a new query for search.

We propose an approach which exploits the *associations* between terms in a search query consisting of two or more terms in order to help understand the true intention of search users, and thus provide correct search results for even *vague queries*. We use our CORDER method [4] for measuring the association between terms in a query. Queries having terms with weak associations with each other are seen as *vague queries*. We use WordNet (http://wordnet.princeton.edu/) to find extended queries of a *vague query*. For each term in a query, a similarity measure based on the WordNet's network structure is used to find all its related terms. These related terms are used to compose extended queries. Since these extended queries may contain a number of dubious ones, we use the CORDER method to remove queries having terms with weak associations with each other. Thus, we can get good quality extended queries, such as extending "sergeant guitar" to get "captain mandolin".

The context of terms in documents can be represented as vectors in a high dimensional semantic space constructed by the Hyperspace Analogue to Language (HAL) model [1]. In pervious work, information flow (IF) method [2] has been used for query expansion and outperformed benchmark methods in information retrieval (IR). For each good quality extended query, we use the IF method to expand the query, e.g., query "Captain Mandolin" is expanded to "Captain Mandolin Correlli Louis De Bernieres Iannis Arsenios Kyria Pelagia Gandin", where the author "Louis De Bernieres" and main characters in the book in the context of the two original terms are included by the model. The expanded query can be used for search the corpus for improved search results, or be sent back to the searchers for more informative search.

Note that our approach is distinct from the traditional query expansion in information retrieval which often does not quantitatively take into account the vagueness of a query by considering the intra-query term associations.

The rest of the paper is organized as follows. In Section 2, we present our novel approach in terms of CORDER based query term association measuring, WordNet based query extension, CORDER based query pruning, and HAL model and IF method based query expansion. In Section 3, we report our initial experimental evaluation on an Amazon Book corpus. Finally, we conclude and propose future work in Section 4.

2. Our Approach for Intelligent Search

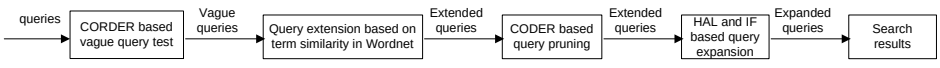

Fig. 1. Process of our intelligent search approach

Our approach for exploiting associations between terms in a query is a combination of previous work on WordNet-based similarity, CORDER, HAL model, and IF method, and the process of our approach is shown in Fig. 1. We can apply our search to a

corpus, e.g., contents and reviews of thousands of books. Our approach can inspire the search on the whole web to help overcome the limitations of current search engines.

This Section is organized as follows. In Section 2.1, we use CORDER to measure the association between terms in a query and queries having terms with weak associations between each other are seen as vague queries. In Section 2.2, we present our method of using term semantic similarity from WordNet for extending vague queries. In Section 2.3, extended queries are ranked and pruned based on CORDER-based association strength between terms in each of these queries. In Section 2.4, we apply HAL model and information flow method to these queries for query expansion towards better IR results and informative search.

2.1. CORDER based Association Measure

We assume that generally users put terms which are semantic related to each other in a search query, such as the query "captain mandolin" is a book's title and the two words "captain" and "mandolin" are strongly related and they will frequently occur with each other in a corpus where the book's title is often mentioned. The probability that the two words co-occur is much higher than by chance. On the contrary, if a query consists of terms that are not strongly associated with each, such as the query "sergeant guitar", we think the query is probably vague and needs to be extended. Various statistical measures can be used, such as mutual information (MI), for association measuring. In our previous work, our CORDER method [4] has outperformed many statistical measures in association measuring. Thus, we use the CORDER method to define the association strength between two terms T_1 and T_2 by taking into account three aspects as follows.

Co-occurrence: Two terms are considered as co-occurring if they appear in the same text fragment, which can be a document or a text window. Generally, if one term is closely related to another term, they tend to co-occur often. To normalize the relatedness between two terms, T_1 and T_2, the relative frequency [5] of co-occurrence is defined as follows.

$$\hat{p}(T_1,T_2) = \frac{Num(T_1,T_2)}{N}$$

where $Num(T_1,T_2)$ is the number of documents in which T_1 and T_2 co-occur, and N is the total number of documents in the corpus for search.

Distance. Two terms which are closely related tend to occur close to each other. The mean distance between T_1 and T_2 in the ith document, $m_i(T_1,T_2)$, is defined as:

$$m_i(T_1,T_2) = \frac{\sum_{j=1}^{f_i(T_1)} \min(T_{1,j},T_2)}{f_i(T_1)}$$

where $f_i(T_1)$ is the number of occurrences of T_1 in the ith document, $min(T_{1,j},T_2)$ is the minimum distance between the jth occurrence of T_1, $T_{1,j}$, and T_2.

Association strength: The overall association strength between T_1 and T_2, defined as follows, takes into account their co-occurrence, mean distance, and frequency in co-occurred documents.

$$R(T_1,T_2) = \hat{p}(T_1,T_2) \times \sum_i \left(\frac{f(Freq_i(T_1)) \times f(Freq_i(T_2))}{m_i(T_1,T_2)} \right)$$

where $f(Freq_i(T_1)) = tfidf_i(T_1)$, $f(Freq_i(T_2)) = tfidf_i(T_2)$, and $Freq_i(T_1)$ and $Freq_i(T_2)$ are the numbers of occurrences of T_1 and T_2 in the ith document, respectively. The term frequency and inverted document frequency measure *tfidf* is defined as $tfidf_i(j) = tf_i(j) * \log_2\left(N / df_j\right)$, where $tf_i(j) = f_i(j) / max(f_j(k))$ is the frequency $f_i(j)$ of term j in the ith document normalized by the maximum frequency of any term in the ith document, N is the number of documents in the corpus, and df_j is the number of documents that contain the term j.

We set up a threshold on the CORDER based association strength. Suppose we have a query, Q, consisting of M terms, $\{T_i\}$. For each term T_i, if its association strengths with all the other terms in Q fall below the threshold, term T_i is a vague term in the query, and the query is a vague query. For example, query "captain guitar" is a vague query and both terms are vague terms. In the next Section, related terms of each vague term in a query are found based on a similarity measure for query extension.

2.2. WordNet-based Semantic Similarity for Query Extension

Given a vague query consisting of a number of vague terms, we extend each of these vague terms to get its related terms in WordNet. The Perl package *WordNet similarity* written by Pedersen et al [3] has eight relatedness or similarity measures. They defined the least common subsumer (LCS) of two concepts A and B as the most specific concept that is an ancestor of both A and B.

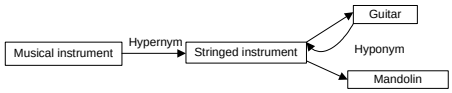

Fig. 2. Part of WordNet taxonomy

In Fig. 2, LCS of "Guitar" and "Mandolin" is "Stringed instrument". They defined the information content of a concept c, $IC(c)$, as the negative log likelihood of the probability of encountering an instance of the concept as follows.

$$IC(c) = -\log\frac{Freq(c)}{N}$$

Where the probability is the frequency of c in the corpus, $Freq(c)$, divided by the total number of concepts in the corpus, N.

Because WordNet allows multiple inheritance, the similarity between two concepts A and B is defined as the *Resnik measure*, i.e., the information content of the least common subsumer of A and B with the highest information content as follows.

$$rel_{res}(c_1, c_2) = \max(IC(c)), \quad where \quad c \in LCS(c_1, c_2)$$

We propose to use the Brown corpus (http://nora.hd.uib.no/corpora.html), which consists of one million words of American English texts printed in 1961, as the basis for the information content calculation. We set up a threshold on the similarity between two terms/concepts. For each vague term in the query as the target, we find its related terms having similarity with the target above the threshold. We get a number of extended queries by the different combinations of the related terms, e.g., for a query consisting of two terms T_1 and T_2, if we get N_1 related terms for T_1 and N_2 for T_2, we get $(N_1+1)*(N_2+1)$ extended queries.

2.3. Association Strength based Extended Query Pruning and Ranking

After query extension, we may get a very large number of queries, such as for query "sergeant guitar", we can get extended queries as "captain mandolin", "lawman guitar", "lawman mandolin", "captain violin", "lawman violin" etc. Similar to the method in Section 2.1, we use the CORDER based association strength to judge whether each of these extended queries is a vague query. Vague queries are seen not to result in good search results, and removed. For example "lawman mandolin" query is vague and removed. Next we rank these extended queries in terms of association strengths among their query terms. Suppose we have a query, Q, consisting of M terms, $\{T_i\}$. For each term T_i, we get its overall association strength, AS_i, as its average association strength with all the other terms in Q. We further average all the M terms' overall association strengths to get the total association strength of the query, AS_Q, as the average of AS_1, AS_2,…, *and* AS_M. We rank these extended queries in terms of their total association strengths, e.g., we can get a ranked list as 1. "captain mandolin", 2. "captain violin" etc.

2.4. HAL Model and Information Flow Method based Query Expansion

Users typically capture the meaning of a term by the contexts in which the term appears. The meaning of the term in a corpus can be captured by examining its co-occurrence patterns with other terms in the corpus. Burgess and Lund [1] showed that the HAL vector constructed from the contexts of a term can be used to simulate the semantic meanings of the term. We use the HAL model to construct vector representations of terms in a query. A text window, size W, moves over the corpus. Terms within the window are seen as co-occurring with a relation strength inversely proportional to the number of words between them in the window. After traversing the corpus, an accumulated co-occurrence matrix for all terms is produced. For a query Q consisting of M terms $\{T_i\}$, their corresponding HAL vectors are $\{V_i\}$. We use the Information Flow (IF) method developed by Song and Bruza [2] for query expansion. We use a heuristic method to determine the degree of dominance of a term in a query, as the term frequency in the query (*qtf*) divided by the document frequency of the term (*df*). We rank query terms by *qtf/df* to get a new ordered list $\{T_i\}$. We use the concept combination heuristic presented in [2] to get the combined concept $T_1 \oplus T_2$, where T_1 dominates T_2. There are four parameters for the heuristic. We use β_1 and β_2 to enhance the weights of intersecting dimensions of two terms, and α_1 and α_2 to separately re-scale the weights of dimensions of two term vectors, respectively. We have followed Song and Bruza [2]'s parameter settings. The combination process goes recursively, i.e., $((\ldots(T_1 \oplus T_2) \oplus T_3) \oplus \ldots) \oplus T_M)$. We get a single vector to represent the query Q. For example, the HAL vector for a query "Captain Mandolin" is <Captain:0.22, Mandolin:0.22, Correlli:0.17, Louis:0.09, De:0.09, Bernieres:0.08, …>. We follow Song and Bruza's settings to select up to 100 dimensions with highest weight in the query vector for query expansion. A cosine similarity between the query vector and the vector of each document in the corpus is computed. Cosine similarities of documents are used to rank them and up to 20 documents with the highest cosine similarities are returned as the search results.

3. Initial Experimental Evaluation

We have carried out an experiment with 500 books on Amazon website. Book descriptions and reviews are retrieved using Amazon web services. Each book is associated with its description and its reviews. We have asked one user to compose 20 queries in which 14 are vague ones, and used the CORDER method to identify vague ones. CORDER identified 14 vague queries in which 12 are correct. We then applied our WordNet based query extension method and CORDER based query pruning method to these 14 queries. The user has specified their corresponding accurate queries, e.g., "cup of flame" → "goblet of fire", and "dark mansion" → "bleak house". Among the 14 extended queries ranked as No.1 for the 14 vague queries, respectively, by CORDER, 10 are correct, and the other 2 correct ones are ranked as No.2 and No. 5, respectively. We then use the 14 No.1 ranked extended queries to search the corpus. Out of 14 correct books, 10 books appear as the first on the list of search results, and 2 books appear as the second search result. For example, the query "cup of flame" returns the book "Harry Potter and the Goblet of Fire", and "panda spell" returns "Eats, Shoots & Leaves: The Zero Tolerance Approach to Punctuation" etc. We applied the HAL model and IF method for query expansion. The extended queries are expanded and presented to the user for more informative search. The user had been able to better understand the context of these search queries and select the right query for search. The HAL vector had also been seen by the user as a summary of the search target.

4. Conclusions and Future Work

In this paper, we propose a term association based method for extending and expanding vague search queries to salient ones, which are used to find the correct information in the user's mind. The method can complement mostly keyword based approach in current search engines for more intelligent search. Our initial experiment on an Amazon book corpus has shown that our approach is effective in detecting vague queries, extending them for the salient ones, and finding the correct books. We are planning to experiment with our approach on the TREC collections in order to formally compare with some well-regarded query expansion models.

5. Reference:

1. Lund, K. and Burgess, C. (1996) Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28(2), 203-208.
2. Song, D. and Bruza, P.D. (2001) Discovering information flow using a high dimensional conceptual space. In *Proc. of the 24th Annual International Conference on Research and Development in Information Retrieval (SIGIR'01)*, pp. 327-333.
3. Pedersen, T., Patwardhan, S., and Michelizzi, J. (2004) Wordnet::similarity –measuring the relatedness of concepts. In *Proc. of Fifth Annual Meeting of the North American Chapter of the ACL (NA ACL-04)*, Boston, MA.
4. Zhu, J., Gonçalves, A., Uren, V., Motta, E., and Pacheco, R. (2005). Mining Web Data for Competency Management. In *Proc. of Web Intelligence (WI 2005)*, France, pp. 94-100, IEEE.
5. Resnik, P. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research (JAIR)*, vol. 11, 1999, 95-130.
6. Anderson, J. R., Bower, G. H. (1973). Human associative memory. Washington: Winston and Sons, 1973.