Query Expansion using Associated Queries

Bodo Billerbeck Falk Scholer Hugh E. Williams Justin Zobel
School of Computer Science and Information Technology
RMIT University, Melbourne, Australia, 3001
{bodob, fscholer, hugh, jz}@cs.rmit.edu.au

Abstract

Hundreds of millions of users each day use web search engines to meet their information needs. Advances in web search effectiveness are therefore perhaps the most significant public outcomes of IR research. Query expansion is one such method for improving the effectiveness of ranked retrieval by adding additional terms to a query. In previous approaches to query expansion, the additional terms are selected from highly ranked documents returned from an initial retrieval run. We propose a new method of obtaining expansion terms, based on selecting terms from past user queries that are associated with documents in the collection. Our scheme is effective for query expansion for web retrieval: our results show relative improvements over unexpanded full text retrieval of 26%–29%, and 18%–20% over an optimised, conventional expansion approach.

1 Introduction

Information retrieval aims to find documents that are relevant to a user's information need. In web retrieval, the need is typically expressed as a query consisting of a small number of words (Spink, Wolfram, Jansen & Saracevic 2001), and answer documents are chosen based on the statistical similarity of the query to the individual documents in the collection. Much research over several decades has led to development of statistical similarity measures that are reasonably effective at finding answers for even the shortest queries (Witten, Moffat & Bell 1999). However, addition of good additional query terms can lead to significant improvements in effectiveness.

A wide range of methods for addition of query terms have been proposed, from manual techniques such as iterative query development and relevance feedback to automatic techniques such as thesaural expansion, anchor-text ranking, and query expansion. In query expansion—which is the focus of this paper—the additional query terms are extracted from highly ranked documents, on the assumption that these are likely to be relevant. It has been shown to be effective on some collections, but results on large collections of web data have been mixed. For example, in our work (Billerbeck & Zobel 2003) using the Okapi approach to ranking, we have found not only that the standard parameters are inappropriate for the web data, but that, even with the best parameters (found by tuning to that data and queries), the performance gains are insignificant.

In this paper we propose an alternative approach to query expansion. The general approach we consider is that the source of expansion terms need not be the collection itself, but could be any document set whose topic coverage is similar to that of the collection, and may thus suggest additional query terms.

Specifically, we investigate whether query associations can be used for query expansion. Given a log containing a large number of queries, it is straightforward to build a surrogate for each document in a collection, consisting of the queries that were a close match to that document. We have shown in earlier work (Scholer & Williams 2002) that these *query associations* can provide a useful document summary; that is, the queries that match a document are a fair description of its content. Here, we investigate whether query associations can play a role in query expansion.

Ranking with query expansion consists of three phases: ranking the original query, against the collection or a document set; extracting additional query terms from the highly ranked items; then ranking the new query against the collection. We show that query associations are a highly effective source of expansion terms. On the TREC-10 data, average precision rises from 0.158 for optimised full text expansion to 0.189 for expansion via association, a dramatic relative improvement of 19.5%.

2 Background

This paper explores refinements to automatic query expansion by considering alternative ways in which candidate expansion terms can be chosen. In this section, we consider related background work in the areas of query expansion, the use of past user queries, and expansion using anchor text from web documents.

Query Expansion

Relevance feedback was proposed over thirty years ago as a method for improving the effectiveness of information retrieval (Salton & McGill 1983). In this approach, a user is presented with a list of answers to a query, which the user would then mark as relevant or irrelevant to the information need. It was observed (van Rijsbergen 1979) that terms closely related to those which successfully discriminate relevant from non-relevant documents are good discriminators themselves. Since it can be assumed that query terms are useful in favouring relevant documents, expansion terms that are related to the original query terms should be useful for ranking. The system could then develop a new query based on this feedback; experiments showed that the new query could be evaluated with significantly greater effectiveness. However, the approach does require that the user takes the time to assess each document.

More recently, query expansion or QE—also known as pseudo-relevance feedback or automatic query expansion—was developed as a variation of relevance feedback (Buckley, Salton, Allan & Singhal 1994). In this approach, an original query is run using conventional information retrieval techniques (Arasu, Cho, Garcia-Molina, Paepcke & Raghavan 2001, Baeza-Yates & Ribeiro-Neto 1999, Witten et al. 1999). Then, related terms are extracted from, for example, the top 10 documents that are returned in response to the original query; the additional terms are selected using statistical heuristics. The related terms are then added to the original query, and the expanded query is run again to return a fresh set of documents, which are returned to the user. Again, experiments showed that the method improves effectiveness.

As an example, a user trying to find information about the Richmond football team might pose a query such as *richmond*. After expansion, the query might be *richmond club football afl tigers*. The reformulated query does not concentrate only on documents that contain the single original query term, but can also retrieve documents that are on the same topic as the original query but for some reason do not name the team.

Variations of QE and relevance feedback involve further interaction (Leuski 2000), and for example require the user to rate any additional query terms proposed by the system. Some of the early web search engines, such as Excite, used a similar approach. In addition, selected techniques such as Rocchio expansion—which we discuss later in this section—are also used in applications such as document categorisation.

In this work, we have used the expansion method described at TREC 8 by Robertson & Walker (2001), which improves effectiveness on average by about 10%—an impressive improvement, given that the effectiveness of the underlying Okapi BM25 similarity measure is already high. In this approach, documents are initially ranked using the Okapi BM25 measure (Robertson & Walker 2000, Robertson, Walker, Hancock-Beaulieu, Gull & Lau 1992) applied to the original query.

The Okapi BM25 measure is as follows:

$$bm25(q,d) = \sum_{t \in q} \log\left(\frac{N - f_t + 0.5}{f_t + 0.5}\right) \times \frac{(k_1+1) f_{d,t}}{K + f_{d,t}}$$

where: q is a query, containing terms t; d is a document; N is the number of documents in the collection; f_t is the number of documents containing term t; K is $k_1((1-b)+b\times L_d/AL)$; k_1 and b are parameters, set to 1.2 and 0.75; $f_{d,t}$ is the number of occurrences of t in d; and L_d and AL are the document length and average document length respectively, measured in some suitable unit.

A detailed explanation of the Okapi BM25 formulation is presented by Sparck-Jones, Walker & Robertson (2000). We have omitted some additional parameters that are not used in this context; for example, we have assumed that query terms are not repeated. The first term in the BM25 measure reduces the impact of query terms that are common in the collection, and the second favours documents that have a high density of query terms.

The top R ranked documents provide a pool from which expansion terms are chosen based on their term selection value:

$$TSV_t = \left(\frac{f_t}{N}\right)^{r_t} \binom{R}{r_t}$$

where r_t is the number of these documents that contain term t. E terms that have the lowest selection value and are not included in the original query are appended to form a new query.

The reformulated query is then used to rank documents, but instead of using the expansion terms' Okapi value as above, the Robertson & Walker (2000) formulation:

$$\frac{1}{3} \times \log\left(\frac{(r_t+0.5)/(R-r_t+0.5)}{(f_t-r_t+0.5)/(N-f_t-R+r_t+0.5)}\right)$$

is used. The division by three helps prevent expansion terms from dominating in the reformulated query. (We tested varying this factor and confirmed that division by three is an appropriate choice.)

In most of the Okapi-related expansion experiments that have been reported, the key parameters are fixed, typically with $R = 10$ and $E = 25$. Sakai & Robertson (2001) have outlined an alternative approach, and Carpineto, de Mori, Romano & Bigi (2001) have investigated the effect of alternative parameter settings.

There has been a great deal of work on the topic of automatic relevance feedback and related areas. One of the earliest and most influential papers is that of Rocchio (1971), who demonstrated the use of training from positive and negative examples of relevance to improve the query formulation. Approaches based on the work of Rocchio continue to be investigated (Carpineto et al. 2001).

However, query expansion is not always effective. Billerbeck & Zobel (2003) and Carpineto et al. (2001) have shown that query expansion using local analysis with fixed parameters is not robust. In some collections, such as the TREC web data, expansion does not appear to work effectively. The experiments reported later show poor results, for example, for an expansion technique that has repeatedly been found to work on the TREC newswire data.

Past queries

Past queries have been shown to be useful for increasing retrieval effectiveness (Fitzpatrick & Dent 1997, Furnas 1985, Raghavan & Sever 1995). Fitzpatrick & Dent (1997) investigated the use of past queries to improve automatic query expansion, by using the results of past queries to form *affinity pools*, from which expansion terms are then selected. The process works as follows: for a query that is to be expanded, up to three past queries that are highly similar to the current query are identified. The top 100 documents that were returned for each of these past queries are then merged, forming the affinity pool. Candidate expansion terms are identified by running the original query against this pool. Individual terms are then selected from the top-ranked documents using a TF-IDF term-scoring algorithm. Fitzpartick and Dent demonstrate that this technique improves relative average precision for the TREC-5 collection by around 15%, from 21.3% to 24.5%. In our work, we propose a different approach to use of past queries. We also choose expansion terms from past queries directly, rather than using them to construct sets of full text documents from which terms are then selected.

Query association (Scholer & Williams 2002) is a technique whereby user queries become associated with a document if they share a high statistical similarity with the document. The association process proceeds as follows: a query is submitted to a search system, and a similarity score is calculated (for example, using the Okapi ranking formula described in Section 2). The query then becomes associated with the top N documents that are returned. For efficiency, an upper bound, M, is imposed on the number of queries that can become associated with a single document. Once a document has a full set of M associations, the least similar associated query can be dynamically replaced with a new, more similar query.

Consider a brief example, where we start with no stored associations, and association parameter settings of $M = 2$ and $N = 5$. A user runs an initial query ($q1$), "stars on crystalline sphere". This query becomes associated with the top 5 answer documents returned by the search system. Suppose that a second query ($q2$), "nicolaus copernicus", retrieves a further 5 documents, one of which was also retrieved by the first query. Then this document now has two associations, while eight other documents in the collection have one. Now consider a final query ($q3$), "geocentric cosmology", which as one of its answers also retrieves the document that already has two associations. If the similarity scores between queries and the common document are ordered such that $q1 < q2 < q3$, then $q1$ will be replaced with the query $q3$ as an association for that document.

Query association was proposed for the creation of document summaries, to aid users in judging the relevance of answers returned by a search system. Scholer & Williams (2002) report that appropriate parameter settings for this task are $M = 5$ and $N = 3$, leading to small summaries composed of high-quality associations. Keeping the summaries small was important for reducing cognitive processing costs for the user. In this work, we use associated queries as a source of terms for query expansion. It is therefore not imperative that the number of associated queries be kept low. We discuss the choice of parameters further in Section 3.

Anchor text

Craswell et al. examined the effectiveness of document surrogates created from *anchor text* for finding entry pages to web sites (Craswell, Hawking & Robertson 2001). In their work, the text content of hypertext links or anchor tags that *inlink* to a document is extracted and compiled into a document surrogate. Experiments show that document surrogates derived from anchor text are significantly more effective than full-text retrieval for a page-finding task (Hawking & Craswell 2001). However, anchor text was not found to be useful for topic-finding tasks. Therefore, for example, anchor text could be expected to aid retrieval for queries such as "richmond football club" but not for queries such as "kicking footballs". We examine the use of anchor text as a source for query expansion terms in Section 3.

3 Generalised Expansion

In this section, we describe our approach to query expansion and, in particular, focus on the novel use of query associations in the expansion process. Our generalised method for query expansion proceeds as follows: first, a query is submitted to our search system, and the top R answer documents are obtained, based on a particular collection. From this initial retrieval run, it is possible to identify a set of candidate expansion terms; these may be based on the top R documents, or surrogates corresponding to these documents. Then the top E expansion terms are selected, using Robertson and Walker's term selection value formula (see Section 2). Finally, selected terms are appended to the original query, which is then run against the target text collection.

Within this general framework, if we use a single collection of documents for all steps, then query expansion is of the standard form, as for example proposed by Robertson & Walker (2001). We call this scheme FULL-FULL, as steps one and two of the expansion process are based on the full text of documents in the collection.

Instead of initially searching or choosing expansion terms from the full text of a document, another possibility is to use surrogate documents constructed from query associations. In this

approach, queries are associated with documents as described earlier, then the set of associated queries for a document is used to represent the document. These can be incorporated into the expansion framework in either the first step (ranking), the second step (term selection), or both. In detail, these three options are as follows.

1. The original query can be run on the full text collection, after which the top E expansion terms are selected from the set of queries that have previously become associated with the top R documents returned from running the original query.

   We call this scheme FULL-ASSOC, as step one of expansion is based on the full text of documents in the collection, step two is based on query associations.

2. Initially rank directly on the surrogates built from associations, then choose expansion terms from the original documents. We call this scheme ASSOC-FULL.

3. Rank on the document surrogates built from associations, then select the E expansion terms from the top R ranked surrogates. We call this scheme ASSOC-ASSOC, as associations are used for both steps 1 and 2 of expansion.

The use of associations for expansion is attractive for several reasons. One is that it means that the additional terms have already been chosen by users as descriptors of topics. Another is that, in contrast to expanding directly from queries, there is more evidence of relevance: a surrogate document constructed from associations has many more terms than an individual query. The fact that the queries have associated with the document means that, in some sense, the terms have topical relationships with each other.

Unlike other methods that make use of the top-ranked documents, such as expansion from document summarisations by Lam-Adesina & Jones (2001), ASSOC-ASSOC does not rely directly on the document collection that is searched. The second and third variations of using associations (schemes ASSOC-FULL and ASSOC-ASSOC) do not rely on ranking the documents in the collection to find relevant associations, but treat the associations themselves as documents that are ranked and used as sources for expansion terms. Thus, in contrast to using thesauruses (Mandala, Tokunaga & Tanaka 1999), the "aboutness" of the individual documents is captured and made use of.

An alternative way to find candidate terms for query expansion from past user queries is to treat the individual queries as documents. We can then source expansion terms by initially ranking the individual queries, and selecting E terms from the top R past queries returned. We call this scheme QUERY-QUERY. Note that, as the individual queries have no direct relation with any particular full text document in the collection (in contrast to the association case above), it does not make sense to have a FULL-QUERY or QUERY-FULL scheme.

Another source of terms for query expansion that we have experimented with is anchor text. Inlinks (text from anchor tags in other documents that point to a document) have a direct relationship with documents in the collection. We consider one approach using anchor text, where we select the E expansion terms from the top R anchor text surrogates, and then search the surrogates again using the expanded query. We call this LINK-LINK.

Most of the schemes described above have parameters that need to be determined, in particular R and E. Rather than make arbitrary choices of values, in most cases we used the TREC-9 queries and relevance judgements described below to find good parameter settings, using the average precision measure, and then report only these settings on TREC-10. Thus the TREC-9 results are the best possible for that method on that data, with post hoc tuning, while the TREC-10 results are blind runs.

4 Experiments

In this section, we describe our experimental environment, discuss the statistical significance tests used to validate our results, and present the results of our experiments with query expansion techniques for web collections.

Setup

For our experiments, we used the experimental testbed made available by the TREC conferences (Harman 1995) for the evaluation of information retrieval experiments. Our experiments were conducted using the TREC WT10g collection, a 10 gigabyte collection of data crawled from the World Wide Web in 1997 (Bailey, Craswell & Hawking 2001). The collection was constructed to be representative of the web, and consists of 1.69 million documents with a high level of interconnectivity, allowing link-based retrieval methods to be evaluated.

Fifty test queries and corresponding relevance judgements for this collection were developed as part of each of the Web tracks at the TREC-9 and TREC-10 conferences (Voorhees & Harman 2000, Voorhees & Harman 2001). TREC queries consist of four parts: a query number; a title field (for the TREC-9 web track, these were taken from search engine logs (Hawking, Craswell & Thistlewaite 1999)); a description of the user's information need; and a narrative, giving more detail on what kinds of documents should be considered relevant or irrelevant. For our experiments, we only used the title field for the initial query, as this is most representative of a typical web information retrieval task.

We use a variant of the Okapi BM25 ranking measure to obtain our similarity scores (see Section 2 and Robertson & Walker (2000)). The full text retrieval results that we use as a baseline are from runs that use no query expansion.

The approach used for creating associations is that described in Section 2. The query associations were built using two logs from the Excite search engine[1], each taken from a single day in 1997 and 1999 (Spink, Jansen, Wolfram & Saracevic (2002) provide a comprehensive analysis of the properties of these query logs). After filtering the logs to eliminate profanities, and removing duplicates and punctuation, we were left with 917,455 queries to associate with the collection. The average number of associations per document after processing was 5.4, and just under 25% of documents in the collection had zero associations. While we built our associations as a batch job, in a production system associations would be made in real time as each query is submitted to the system; it is unclear what the costs of this process are, and we plan to investigate this in future work.

In separate experiments, we established effective settings for query association (Scholer, Williams & Turpin 2003). We found that association parameters of $M = 19$ and $N = 39$ worked well, that is, each document has a maximum of 19 associated queries and the top 39 documents returned in response to a query are associated with that query. This was determined by creating document surrogates from the associated queries based on different parameter combinations, and testing retrieval effectiveness by evaluating searches on these surrogates. We use these parameter settings for our results reported below.

For our experiments where expansion terms are chosen directly from queries (with no association), we also use the 917,455 filtered entries from the Excite 1997 and 1999 Excite logs.

The anchor text for our experiments was obtained by identifying anchor tags within the WT10g collection, and collating the text of each anchor that points in to a particular document into a surrogate for that document.

Significance

We evaluate the significance of our results using the Wilcoxon signed rank test. This is a non-parametric procedure used to test whether there is sufficient evidence that the median of two probability distributions differ in location. For information retrieval experiments, it can be used to test whether two retrieval runs based, for example, on different query expansion techniques, differ significantly in performance. As it takes into account the magnitude and direction of the difference between paired samples, this test is more powerful than the sign test (Daniel 1990). Being a non-parametric test, it is not necessary to make any assumptions about the underlying probability distribution of the sampled population.

[1] http://www.excite.com/

Type	AvP	P@10	P@20	P@30	R-P	R	E
Base	0.1487	0.2714	0.2235	0.2000	0.1710	—	—
ASSOC-ASSOC	0.1893*	0.3429*	0.2888*	0.2503*	0.2204*	06	17
ASSOC-FULL(A)	0.1820*	0.3184**	0.2796*	0.2497*	0.2222*	06	17
ASSOC-FULL(B)	0.1618*	0.3041**	0.2510*	0.2231*	0.1969*	98	04
FULL-FULL(A)	0.1584	0.2796	0.2571*	0.2333*	0.1809	10	25
QUERY-QUERY	0.1567	0.2755	0.2357*	0.2116**	0.1861*	65	02
FULL-FULL(B)	0.1553	0.2857	0.2388	0.2184**	0.1867**	98	04
FULL-ASSOC	0.1549	0.2571	0.2276	0.2068	0.1786	06	17
LINK-LINK	0.1454**	0.2653	0.2153	0.1905	0.1685	37	02

Table 1: Performance of expansion techniques of TREC-10 queries on the TREC WT10g collection, based on average precision (AvP), precision at 10 (P@10), precision at 20 (P@20), precision at 30 (P@30), and R-Precision (R-P). Schemes are ordered by decreasing AvP. Results that show a significant difference from the baseline using the Wilcoxon signed rank test at the 0.05 and 0.10 levels are indicated by * and **, respectively.

Type	Q1	Median	Q3	Variance
ASSOC-ASSOC	-0.0239	0.0040	0.0814	0.7327
ASSOC-FULL(A)	-0.0073	0.0116	0.0738	0.4286
ASSOC-FULL(B)	-0.0119	0.0030	0.0308	0.1201
FULL-FULL(A)	-0.0265	0.0007	0.0302	0.2633
QUERY-QUERY	-0.0150	-0.0000	0.0082	0.1108
FULL-FULL(B)	-0.0211	-0.0004	0.0177	0.1319
FULL-ASSOC	-0.0370	-0.0007	0.0306	0.2077
LINK-LINK	-0.0049	-0.0001	0.0002	0.0410

Table 2: Quartiles and variance of different expansion methods on the TREC WT10g collection (TREC-10). Each number is the effectiveness relative to the baseline of no expansion.

The t-test, an alternative, parametric test for paired difference analysis, assumes that the data is normally distributed. Zobel (1998) analysed the results of retrieval experiments from TREC-5, and concluded that the Wilcoxon signed rank test is a more reliable test for the evaluation of retrieval runs.

It is a mistake to claim a significant change in performance based only on different effectiveness scores (Zobel 1998). While post-hoc analysis of error rates can give valuable information about the properties of a collection—see for example Voorhees & Buckley (2002), who calculate error rates for the TREC-3 to TREC-10 data empirically—such thresholds cannot safely be extended to future runs; as these runs were not themselves part of the calculation, per-query variability would not be taken into account. A statistical significance test, on the other hand, enables conclusions to be drawn about whether a variation in retrieval technique leads to consistent performance gains.

Results

In our experiments, we have compared a baseline full text retrieval run with the expansion variants we have described in Section 3. Our results are presented in Table 1, which shows retrieval performance based on five precision metrics: precision at 10 returned documents (P@10), precision at 20 (P@20), precision at 30 (P@30), average interpolated precision (AvP), and R-precision (R-P).

In these results, the FULL-FULL scheme is the conventional approach to query expansion. We show results with two parameter settings: first FULL-FULL(A), the parameter settings used by Robertson & Walker (2001) of $R = 10$ and $E = 25$; and, second, FULL-FULL(B), the optimal parameters we found in exhaustive tests (we discuss this further below). Perhaps surprisingly—but in agreement with recent observations (Billerbeck & Zobel 2003, Carpineto et al. 2001)—the

Type	AvP	P@10	P@20	P@30	R-P	R	E
Base	0.1895	0.2708	0.2042	0.1806	0.2290	—	—
ASSOC-ASSOC	0.2231*	0.3104**	0.2323**	0.2132*	0.2398	06	17
QUERY-QUERY	0.1996	0.2958**	0.2094	0.1875	0.2402	65	02
ASSOC-FULL(A)	0.1966*	0.2604	0.2115	0.1944	0.2167	06	17
LINK-LINK	0.1939	0.2708	0.2177	0.1799	0.2201	37	02
FULL-FULL(B)	0.1923	0.2813	0.2042	0.1819	0.2287	98	04
ASSOC-FULL(B)	0.1856*	0.2875	0.2229**	0.1924	0.2176	98	04
FULL-ASSOC	0.1804	0.2417**	0.2052	0.1910	0.1954*	06	17
FULL-FULL(A)	0.1607	0.2729**	0.2083	0.1854	0.1806**	10	25

Table 3: Performance of expansion techniques on the training data (TREC-9). These were the runs that we used to determine parameter settings and are included for reference. Schemes are again ordered by decreasing AvP.

Type	Q1	Median	Q3	Variance
ASSOC-ASSOC	-0.0057	0.0040	0.0604	0.5814
QUERY-QUERY	-0.0022	0.0000	0.0049	0.0688
ASSOC-FULL(A)	-0.0044	0.0006	0.0458	0.8446
LINK-LINK	-0.0067	-0.0001	0.0019	0.0938
FULL-FULL(B)	-0.0093	0.0000	0.0075	0.1272
ASSOC-FULL(B)	-0.0028	0.0020	0.0295	0.4789
FULL-ASSOC	-0.0259	-0.0040	0.0197	0.3155
FULL-FULL(A)	-0.0423	-0.0072	0.0086	0.8487

Table 4: Quartiles and variance of different expansion methods on the training data (TREC-9).

FULL-FULL schemes do not offer significantly better results than no expansion, except in the R-P measure for our optimal parameter settings (where the relative improvement is 9%, corresponding to an absolute increase of 0.016).

Our novel association-based schemes are effective for query expansion. In relative terms, the ASSOC-ASSOC scheme is 18%–20% better in all three measures than FULL-FULL(A) expansion (an absolute difference of 0.03–0.05), and 26%–29% better than the baseline no-expansion case (an absolute difference of 0.04–0.07). The ASSOC-FULL schemes are even more effective in the stringent R-P measure; all results are significant at least at the 0.10 level. We conclude that query association is an effective tool in the initial querying stage prior to expansion; this is a particularly useful result since query associations are compact and can be efficiently searched, and we plan to quantify this in future work.

Another perspective on the results is shown in Table 2. All of the methods improve median performance to approximately the same extent—that is, not at all. At the lower quartile, all the methods have degraded performance somewhat; the extent of degradation has little relationship to average effectiveness. At the upper quartile, however, the differences between the methods are clear. However, note that it is often the case that a query improved by one method is not improved by another.

An example of the behaviour of the ASSOC-ASSOC scheme illustrates its utility over the FULL-FULL approach. For the query "earthquakes" (TREC query 513), the AvP for the ASSOC-ASSOC scheme is 0.1706, compared to 0.1341 for no expansion and 0.1162 for the FULL-FULL approach. This is a direct result of the choice of terms for expansion. For the ASSOC-ASSOC scheme, the expanded query is large and appears to contain only useful terms:

> earthquakes earthquake recent nevada seismograph tectonic faults perpetual 1812
> kobe magnitude california volcanic activity plates past motion seismological

In contrast, for the FULL-FULL(B) scheme the query is the more narrow:

earthquakes tectonics earthquake geology geological

These trends in the success of expansion with ASSOC-ASSOC are consistent with our empirical inspections of other queries.

The QUERY-QUERY, FULL-ASSOC, and LINK-LINK schemes offer limited benefit for expansion. Without association to documents, the queries are ineffective: this is probably due to the median length of the queries being two words, that is, the queries have very little content when not grouped together as associations. The FULL-ASSOC scheme is ineffective for similar reasons: query associations are an excellent source of expansion terms, but are less effective as document surrogates in the final step of the retrieval process. The LINK-LINK scheme is significantly worse than no expansion for the AvP measure; this is perhaps unsurprising, since anchor text has been shown to be of utility in home or named page finding tasks, while the queries we use are topic finding tasks.

As discussed earlier, we tuned our parameters prior to our experiments. Specifically, the R and E parameters used in our experiments were identified through an exhaustive search on the same collection, but using TREC-9 queries. The results of this process with the TREC-9 queries are shown for reference in Tables 3 and 4.

We have tried similar experiments on TREC disks 4 and 5, which consist of newswire and similar data. These were unsuccessful. The problem appears to be that the query logs, drawn from the web, are inappropriate for this data: based on a small sample, it seems that many of the queries in the log do not have relevant documents. Thus the process of creating newswire associations from search-engine logs is unlikely to be successful. Query association based expansion is therefore only of utility if queries are available that are appropriate for the collection being searched.

5 Conclusions

Conventional wisdom has held that query expansion is an effective technique for information retrieval. However, recent experiments have contradicted this and shown that parameter settings that work well for one set of queries may be ineffective on another. In this paper, we have investigated alternative techniques for obtaining query expansion terms, with the aim of identifying techniques that are robust for different query sets.

We have identified a successful expansion source for web retrieval. This source is query associations, that is, past queries that have been stored as document surrogates for the documents that are statistically similar to the query. In experiments with almost one million prior query associations, we found that expanding TREC-10 web track topic finding queries using query associations and then searching the full text is 26%–29% more effective than no expansion, and 18%–20% better than an optimised conventional expansion approach. Moreover, our results are significant under statistical tests. We conclude that query associations are a powerful new expansion technique for web retrieval.

We plan to pursue several directions in our future work. We will investigate the optimal parameters for query association in the context of query expansion; this work uses parameters that were determined through a surrogate retrieval task. We also plan to investigate whether fixed parameters for query association are appropriate, and whether all queries should be associated to documents. In addition, we plan to investigate the efficiency tradeoff between maintaining associations and the likely efficiency improvement of searching associations for expansion terms instead of full text.

Acknowledgements

This research is supported by the Australian Research Council and by the State Government of Victoria.

References

Arasu, A., Cho, J., Garcia-Molina, H., Paepcke, A. & Raghavan, S. (2001), 'Searching the web', *ACM Transactions on Internet Technology (TOIT)* **1**(1), 2–43.

Baeza-Yates, R. & Ribeiro-Neto, B. (1999), *Modern Information Retrieval*, Addison-Wesley Longman.

Bailey, P., Craswell, N. & Hawking, D. (2001), 'Engineering a multi-purpose test collection for web retrieval experiments', *Information Processing and Management* . In revision. Available from www.ted.cmis.csiro.au/∼dave/cwc.ps.gz.

Billerbeck, B. & Zobel, J. (2003), When query expansion fails, *in* C. Clarke, G. Cormack, J. Callan, D. Hawking & A. Smeaton, eds, 'Proceedings of the ACM-SIGIR International Conference on Research and Development in Information Retrieval', Toronto, Canada, pp. 387–388.

Buckley, C., Salton, G., Allan, J. & Singhal, A. (1994), Automatic query expansion using SMART: TREC 3, *in* D. Harman, ed., 'Overview of the Third Text REtrieval Conference (TREC-3)', NIST Special Publication 500-225, pp. 69–80.

Carpineto, C., de Mori, R., Romano, G. & Bigi, B. (2001), 'An information-theoretic approach to automatic query expansion', *ACM Transactions on Information Systems (TOIS)* **19**(1), 1–27.

Craswell, N., Hawking, D. & Robertson, S. (2001), Effective site finding using link anchor information, *in* D. H. Kraft, W. B. Croft, D. J. Harper & J. Zobel, eds, 'Proceedings of the ACM-SIGIR International Conference on Research and Development in Information Retrieval', New Orleans, LA, pp. 250–257.

Daniel, W. (1990), *Applied Nonparametric Statistics*, 2nd edn, PWS-KENT Publishing Company.

Fitzpatrick, L. & Dent, M. (1997), Automatic feedback using past queries: Social searching?, *in* N. J. Belkin, A. D. Narasimhalu, P. Willett, W. Hersh, F. Can & E. Voorhees, eds, 'Proceedings of the ACM-SIGIR International Conference on Research and Development in Information Retrieval', Philadelphia, PA, pp. 306–313.

Furnas, G. W. (1985), Experience with an adaptive indexing scheme, *in* L. Borman & R. Smith, eds, 'Proceedings of the ACM-CHI Conference on Human Factors in Computing Systems', pp. 131–135.

Harman, D. (1995), 'Overview of the second text retrieval conference (TREC-2)', *Information Processing & Management* **31**(3), 271–289.

Hawking, D. & Craswell, N. (2001), Overview of the TREC-2001 web track, *in* E. M. Voorhees & D. K. Harman, eds, 'The Tenth Text REtrieval Conference (TREC 2001)', National Institute of Standards and Technology Special Publication 500-250, Washington, DC, pp. 61–67.

Hawking, D., Craswell, N. & Thistlewaite, P. (1999), Overview of TREC-7 very large collection track, *in* E. M. Voorhees & D. K. Harman, eds, 'The Eighth Text REtrieval Conference (TREC 8)', National Institute of Standards and Technology Special Publication 500-246, Washington, DC, pp. 91–104.

Lam-Adesina, A. M. & Jones, G. J. F. (2001), Applying summarization techniques for term selection in relevance feedback, *in* D. H. Kraft, W. B. Croft, D. J. Harper & J. Zobel, eds, 'Proceedings of the ACM-SIGIR International Conference on Research and Development in Information Retrieval', New Orleans, LA, pp. 1–9.

Leuski, A. (2000), Relevance and reinforcement in interactive browsing, *in* A. Agah, J. Callan, E. Rundensteiner & S. Gauch, eds, 'Proceedings of the ACM-CIKM International Conference on Information and Knowledge Management', McLean, VA, pp. 119–126.

Mandala, R., Tokunaga, T. & Tanaka, H. (1999), Combining multiple evidence from different types of thesaurus for query expansion, *in* F. Gey, M. Hearst & R. Tong, eds, 'Proceedings of the ACM-SIGIR International Conference on Research and Development in Information Retrieval', Berkeley, CA.

Raghavan, V. V. & Sever, H. (1995), On the reuse of past optimal queries, *in* E. A. Fox, P. Ingwersen & R. Fidel, eds, 'Proceedings of the ACM-SIGIR International Conference on Research and Development in Information Retrieval', Seattle, WA, pp. 344–350.

Robertson, S. E. & Walker, S. (2000), Okapi/Keenbow at TREC-8, *in* E. M. Voorhees & D. K. Harman, eds, 'The Eighth Text REtrieval Conference (TREC-8)', NIST Special Publication 500-264, Gaithersburg, MD, pp. 151–161.

Robertson, S. E. & Walker, S. (2001), Microsoft cambridge at trec-9: Filtering track, *in* E. M. Voorhees & D. K. Harman, eds, 'The Ninth Text REtrieval Conference (TREC-9)', NIST Special Publication 500-249, Gaithersburg, MD, pp. 361–368.

Robertson, S. E., Walker, S., Hancock-Beaulieu, M., Gull, A. & Lau, M. (1992), Okapi at TREC, *in* D. K. Harman, ed., 'The First Text REtrieval Conference (TREC-1)', NIST Special Publication 500-207, Gaithersburg, MD, pp. 21–30.

Rocchio, J. J. (1971), Relevance feedback in information retrieval, *in* E. Ide & G. Salton, eds, 'The Smart Retrieval System — Experiments in Automatic Document Processing', Prentice-Hall, Englewood, Cliffs, New Jersey, pp. 313–323.

Sakai, T. & Robertson, S. E. (2001), Flexible pseudo-relevance feedback using optimization tables, *in* D. H. Kraft, W. B. Croft, D. J. Harper & J. Zobel, eds, 'Proceedings of the ACM-SIGIR International Conference on Research and Development in Information Retrieval', New Orleans, LA, pp. 396–397.

Salton, G. & McGill, M. (1983), *Introduction to Modern Information Retrieval*, McGraw-Hill, New York.

Scholer, F. & Williams, H. E. (2002), Query association for effective retrieval, *in* C. Nicholas, D. Grossman, K. Kalpakis, S. Qureshi, H. van Dissel & L. Seligman, eds, 'Proceedings of the ACM-CIKM International Conference on Information and Knowledge Management', McLean, VA, pp. 324–331.

Scholer, F., Williams, H. & Turpin, A. (2003), Document surrogates for web search. (Manuscript in submission).

Sparck-Jones, K., Walker, S. & Robertson, S. E. (2000), 'A probabilistic model of information retrieval: development and comparative experiments. Parts 1&2', *Information Processing and Management* **36**(6), 779–840.

Spink, A., Jansen, M. B. J., Wolfram, D. & Saracevic, T. (2002), 'From e-sex to e-commerce: Web search changes', *IEEE Computer* **35**(3), 107–109.

Spink, A., Wolfram, D., Jansen, M. B. J. & Saracevic, T. (2001), 'Searching the web: the public and their queries', *Journal of the American Society for Information Science and Technology* **52**(3), 226–234.

van Rijsbergen, C. (1979), *Information Retrieval*, second edn, Butterworths.

Voorhees, E. M. & Buckley, C. (2002), The effect of topic set size on retrieval experiment error, *in* K. Jrvelin, M. Beaulieu, R. Baeza-Yates & S. H. Myaeng, eds, 'Proceedings of the ACM-SIGIR International Conference on Research and Development in Information Retrieval', Tampere, Finland, pp. 316–323.

Voorhees, E. M. & Harman, D. K. (2000), Overview of the Ninth Text REtrieval Conference (TREC-9), *in* E. M. Voorhees & D. K. Harman, eds, 'The Ninth Text REtrieval Conference (TREC 9)', National Institute of Standards and Technology Special Publication 500-249, Gaithersburg, MD, pp. 1–14.

Voorhees, E. M. & Harman, D. K. (2001), Overview of TREC 2001, *in* E. M. Voorhees & D. K. Harman, eds, 'The Tenth Text REtrieval Conference (TREC 2001)', National Institute of Standards and Technology Special Publication 500-250, Gaithersburg, MD, pp. 1–15.

Witten, I. H., Moffat, A. & Bell, T. C. (1999), *Managing Gigabytes: Compressing and Indexing Documents and Images.*, 2nd edn, Morgan Kaufman Publishing, San Francisco.

Zobel, J. (1998), How reliable are the results of large-scale information retrieval experiments?, *in* W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson & J. Zobel, eds, 'Proceedings of the ACM-SIGIR International Conference on Research and Development in Information Retrieval', Melbourne, Australia, pp. 307–314.