Questioning Query Expansion: An Examination of Behaviour and Parameters [*]

Bodo Billerbeck **Justin Zobel**

School of Computer Science and Information Technology
RMIT University, Melbourne, Australia, 3001,
{bodob,jz}@cs.rmit.edu.au

Abstract

In information retrieval, queries can fail to find documents due to mismatch in terminology. Query expansion is a well-known technique addressing this problem, where additional query terms are automatically chosen from highly ranked documents, and it has been shown to be effective at improving query performance. However, current techniques for query expansion use fixed values for key parameters, determined by tuning on test collections. In this paper we show that these parameters may not be generally applicable, and more significantly that the assumption that the same parameter settings can be used for all queries is invalid. Using detailed experiments with two test collections, we demonstrate that new methods for choosing parameters must be found. However, our experiments also demonstrate that there is considerable further scope for improvement to effectiveness through better query expansion.

Keywords: Information retrieval, Search engines, Query expansion, Effectiveness

1 Introduction

Search engines are the principal mechanism used for finding documents on the world wide web (Schwartz 1998). These engines use information retrieval techniques to match queries, expressed as a series of words, to the documents that are judged the most likely to answer the users' needs. When queries are well formulated, typically consisting of topic-specific keywords that together specify the information need with low ambiguity, search engines can return good matches in the top-ranked documents.

However, queries are often not well-formulated. They may be ambiguous, insufficiently precise, or use terminology that is specific to a country – consider for example the US "wrench" versus the UK "spanner". The majority of queries posed to search engines are brief, with about 60% of queries containing only one or two key words, while the average query length in 2001 was 2.6 words (Spink, Wolfram, Jansen & Saracevic 2002). These problems are more acute when the user wishes to find large numbers of relevant documents: all reviews of a particular movie, for example, or all commentary on a particular topic. Many relevant documents may not contain the words used in the query.

A variety of techniques for improving effectiveness are used. Google[1] uses an approach called PageRank (Page, Brin, Motwani & Winograd 1998), which takes note of links embedded in web pages to rank higher pages that are often referenced by other pages. The reasoning behind this concept is that pages with authority or popularity are more likely to satisfy the information needs of users. An alternative methodology based on links is the hub-and-authority approach (Bharat & Henzinger 1998). There have been many proposals for improving effectiveness based on better use of evidence internal to documents, such as locality and HTML structure. However, none of these methods addresses the issue of vocabulary mismatch.

Query expansion has been widely investigated as a method for improving the performance of information retrieval (Buckley, Salton, Allan & Singhal 1994, Carpineto, de Mori, Romano & Bigi 2001, Mandala, Tokunaga & Tanaka 1999, Robertson & Walker 1999, Robertson & Walker 2000, Sakai & Robertson 2001). It is the only successful automatic method for solving the problem of vocabulary mismatch; alternatives, such as thesaurus-based techniques, have not been as successful (Mandala et al. 1999). In query expansion, the original query is used to identify a small set of highly-ranked documents, which are likely to contain other terms that are common in the same context as the original query terms. Topic-specific terms are chosen from the highly-ranked documents and added to the original query, which is then re-evaluated. It has been shown that query expansion can significantly improve effectiveness.

In this paper, we investigate the performance of one successful approach to query expansion, as used in the Okapi system (Sparck-Jones, Walker & Robertson 2000). In common with all query expansion methods, the Okapi approach requires several parameters, in particular the number of documents in the initial ranking and the number of expansion terms. These were determined in experiments on a particular test data set, and in most experiments since then have been used without variation.

Our results show that it is far from clear that these parameter choices are optimal. Using comprehensive experiments on one test collection, we have investi-

[*] Some of the material in this paper is based on a two-page poster "When query expansion fails", B. Billerbeck and J. Zobel", *Proc. International ACM SIGIR Conference on Research and Development in Information Retrieval*, July 2003, pp. 387–388, Toronto, Canada.

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

gated both average effectiveness and per-query effectiveness for a wide range of parameter choices. These results show that other choices of values can give higher effectiveness, but that no fixed choice is robust: entirely different values are preferable for other collections. Worse, the best choices per query vary wildly. Current approaches to query expansion are not well founded.

However, our results also show that the performance of query expansion has significant scope for improvement: individually tuning parameters to queries can give much better performance than use of fixed values. We hope to be able to develop a method for predicting parameter values, and thus obtain greater effectiveness than is available with current methods.

2 Query expansion

Finding information that suits users' needs is the primary concern about information retrieval.

By the most effective search engines today, documents are matched to queries by *ranking*. In this approach, in principle a statistical matrix is used to evaluate document-query similarity for every stored document, then the highest-scored documents are returned to the user as potential matches. In contrast, traditional matching techniques such as Boolean querying have been demonstrated to be ineffective (Salton 1989). Although boolean queries are slowly gaining greater popularity again since the time before ranked queries were used, the overwhelming number of queries are still ranked queries (Spink et al. 2002).

Most users are familiar with web searches, where a user-formulated query is posed to a search engine that (at least conceptually) compares this query to all web pages and identifies those that are most similar to the query and therefore hopefully satisfy the user need best. While web search engines such as Google, Excite[2] and Lycos[3] can effectively find suitable web pages, there are other applications of searching. Those include searches on confidential collections such as police profiles. In addition, even on subsets of the web, techniques such as the PageRank (Page et al. 1998) are not suitable. For instance if a user is interested only in pages in a specific domain, PageRank information is not available in this context. Another example is corporate data collections, such as archives held by news organisations.

Relevance feedback has been used and extended widely in order to improve upon search effectiveness (Salton & McGill 1983, Rijsbergen 1979, Frakes & Baeza-Yates 1992). Relevance feedback is a mechanism of refining a search process by using knowledge gained by a preliminary search for a final search. Early forms of this process were based on a manual system, were searchers would identify relevant documents returned by a search, extract some features from those documents and use them to augment the original query (Rocchio 1971). A similar approach was to give searchers the option of interactively choosing or ranking additional search terms (Leuski 2000). There are many ways of arriving at additional search terms; an earlier method used thesauruses. Thesauruses can either be derived from the document collection at hand or from other sources (Foskett 1997). However, it has been found that thesaurus-based methods are less less successful (Mandala et al. 1999). In other approaches terms are extracted from the document collection at hand, either by manual selection (as has been referred to above), or – by employment of some heuristics – assuming a group of documents to be relevant which is then used as a base for additional terms. This process is also known as blind relevance feedback (Buckley et al. 1994) and is the most popular strategy employed for query expansion (QE). The purpose of QE is to adapt an original query so that it is better suited to target relevant documents, and it does so by being similar to the documents it is targeting.

As an example, a user might be interested in the rock group Nirvana and therefore use a query *nirvana*. The query might then be expanded to *nirvana cobain kurt band live music*. This new query doesn't only focus on documents that contain the single original term, but will also find documents that are about the rock band, but that don't directly name it.

Note that in the system used, only ranked queries are used and therefore query terms are connected by an implicit "OR". This explains why the final ranking of documents can include different documents than only those that were returned by the unexpanded query.

The QE approach used for the work detailed in this paper is described by Robertson & Walker (2000). It can be split into two tasks, the first is to identify relevant documents from which expansion terms are drawn and the second to rank possible terms by their perceived usefulness.

Finding relevant documents

The first task is relatively straightforward. Documents are originally ranked using conventional retrieval technology (Arampatzis & van der Weide 2001, Baeza-Yates & Ribeiro-Neto 1999, Witten, Moffat & Bell 1999). Using the assumption that documents that are ranked highly are likely to be most relevant to a query, the top-ranked documents are used for extracting terms. Ranking is determined through the Okapi BM25 similarity measure (Robertson & Walker 1999, Robertson, Walker, Hancock-Beaulieu, Gull & Lau 1992) taking into consideration the original query:

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

where t is a term of query q, f_t the number of occurrences of a particular term across the document collection that contains N documents and $f_{d,t}$ is the frequency of a particular term t in document d. K is $k_1((1-b) + b \times L_d/AL)$, k_1 and b are parameters set to 1.2 and 0.75 respectively. L_d is the length of a particular document and AL is the average document length, in a suitable unit, such as the number of terms contained in each document. This formulation is derived from statistical considerations of the nature of text retrieval (Sparck-Jones et al. 2000).

The first term of the above formula is used to dampen the effect of query terms that occur quite frequently in the collection, whereas the second weights

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

[3] http://www.lycos.com

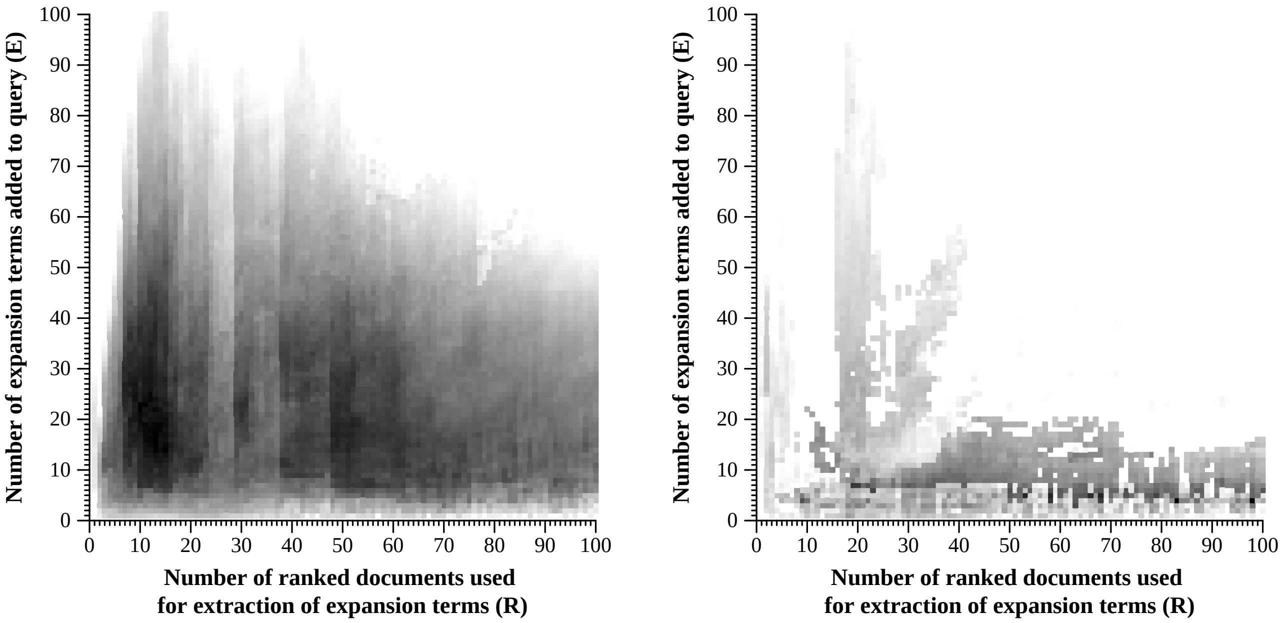

Figure 1: *Cumulative average precision for each number of documents and number of expansion terms. Left-hand side: TREC 8 data, disks 4 and 5. Right-hand side: TREC 9 10-gigabyte web data. Dark areas: high average precision. Light areas: low average precision. White: average precision worse than or equal to no expansion.*

those documents higher that have a high concentration of query terms. A third term of the formula (Sparck-Jones et al. 2000) is neglected here, since it is assumed that query terms occur at most once in each query.

Identifying appropriate expansion terms

Once the top-ranked documents have been found, suitable expansion terms have to be identified. The approach of appending query terms from top-ranked documents corresponds to the observation of Rijsbergen (1979) that terms are good discriminators between relevant and non-relevant documents if they are closely related to terms that are good discriminators.

All terms in the top documents are scored by the following formula and are assigned a term selection value:

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

where R is the number of top-ranked documents examined and r_t is the number of documents that contain a particular term t. Therefore, the more often a term occurs across those R documents and the greater the number of occurrences the smaller is the resultant TSV. R is usually chosen to be 10 (or 5 in similar approaches).

The terms with the smallest TSVs are appended to the original query, but instead of assigning them their Okapi weight, they get assigned the modified Robertson/Spark-Jones weight as follows:

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

This weight is modified by downgrading it to 1/3 of the original value so as to not overpower the original query terms. (Unreported experiments show that this fraction is suitable and maximises effectiveness for a particular test collection.)

A fixed number of 25 terms is appended to the query, and although using thresholds for the TSV have subsequently been used (Robertson & Walker 2000), these were only of marginal effectiveness. Robertson & Walker (1999) showed that the approach as outlined above improves effectiveness by about 10% over an already high Okapi baseline.

The approach of extracting expansion terms from top-ranked documents is also called *local analysis. Global analysis* in contrast relies on information distilled from the collection as a whole, such as the method of Qiu & Frei (1993) that used expansion terms which were found in a similar context as the original query terms. Xu & Croft (1996) use a method where both local and global contexts are taken into consideration.

As discussed later, Sakai & Robertson (2001) have investigated an alternative approach, and Carpineto et al. (2001) have further explored the effect of different parameter settings.

The topic of automatic relevance feedback has been widely researched and one of the earliest and most influential papers is that of Rocchio (1971), where he demonstrates the use of not only terms that enhance a query, but also terms that should be negatively added to a query in order to avoid documents containing these terms.

More recently Kwok (2002) proposed the use of word co-occurrence in small text windows for reranking and query expansion. For his successful method (achieving about 10% improvement in conjunction with re-ranking) he uses a fixed number of 40 terms for every query. In light of the explorations we report, Kwok's approach could potentially benefit from choosing the number of expansion terms individually for each query.

Hoashi, Matsumoto, Inoue & Hashimoto (1999) measures the “word contribution”. In this approach each term occurring in a particular top-ranked document is assigned a score related to how much this

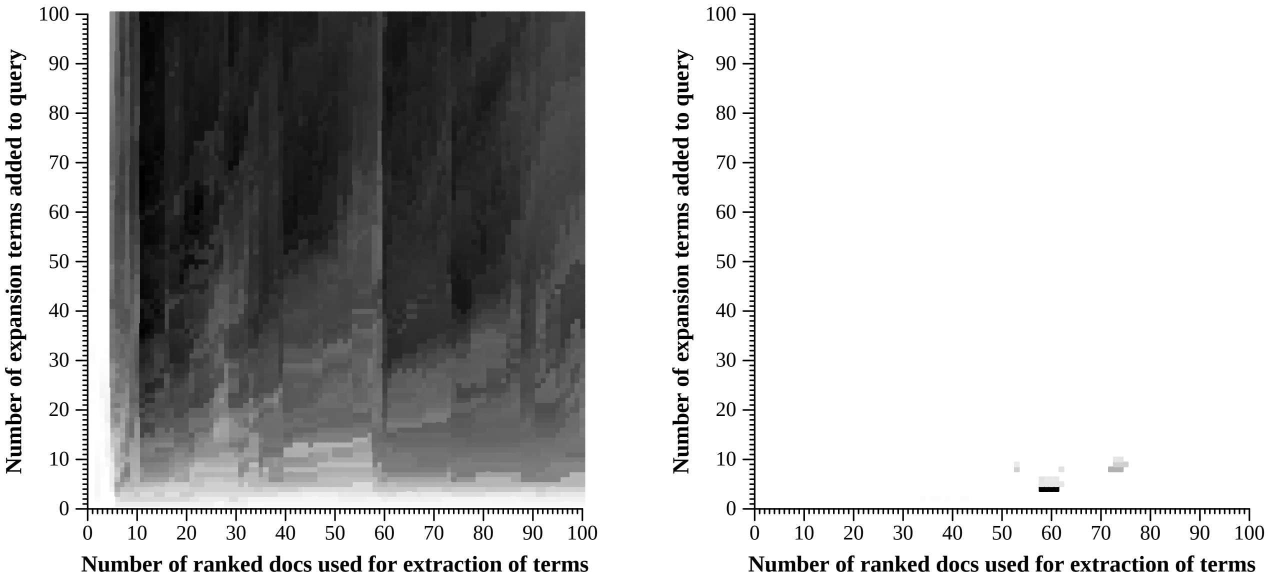

Figure 2: *Parameter pairs for which expansion of TREC 8 query 405 (on the left) and 440 (on the right) achieves higher average precision than non-expansion. Dark spots mark high average precision values, as before.*

term contributes to the similarity of the document and the query. Terms with high scores, usually those that appear in the query and the particular document, indicate that the term contributed heavily to the similarity. Low scores on the other hand mean that a term contributed little to the similarity. Hoashi et al. assume that a term with a low score that is in one of the top-ranked documents would be a good expansion term since it is presumably on the same topic as the query, but is different to the terms the user identified in the original query.

3 Where query expansion fails

In the Okapi work, fixed values were used for key parameters. In some experiments, fixed values were used for R and E, the number of documents in the initial ranking and the number of expansion terms, respectively. These values (10 and 25 respectively) were chosen by experiments on a particular collection, and were observed to give an overall improvement in effectiveness. In other experiments, a fixed value was used for R and a fixed upper bound was imposed for TSV.

Experiments in information retrieval are usually based on a set of test documents, a set of test queries, and manual (human) relevance assessments stating which documents are relevant to which query. Improvements in information retrieval systems are intended to improve average effectiveness according to some metric. The usual metrics are precision (the proportion of answers that are relevant), recall (the proportion of relevant documents that are found), or some combination of these. We focus on average precision and recall at 1000 documents ranked. Different systems tend to do well on different queries; thus an overall improvement may include a decline in effectiveness in specific cases.

We use the TREC data in our experiments[4]. The data includes several multi-gigabyte collections of documents, drawn from sources such as newswires and the web (Voorhees & Harman 1999, Voorhees & Harman 2000). It also includes annotated queries on these documents, and large sets of relevance judgements. This data is the main resource that has been used during the last decade for enhancing information retrieval systems, and is widely used, for example, for benchmarking.

In our experiments, we have primarily used TREC disks 4 and 5 (Harman 1995) and the title field only of queries 401–450, which is the data used in TREC 8 in 1999. In some additional experiments, we have used the TREC 9 10-gigabyte web track with queries 451–500, from TREC 9. The queries were run on our Lucy search engine.[5] Differences in details such as parsers can have a marked impact on effectiveness, making it difficult to exactly reproduce reported experimental results. Some systems stem words aggressively, others do not; some index the content of HTML tags (or of selected tags); others do not; and there are numerous other variations. (Simple experiments suggested that using passages in place of whole documents does not improve effectiveness, however it might be an interesting topic to revisit, as other work has indicated passages can increase effectiveness (Kaszkiel & Zobel 2001).) However, our results with Lucy are highly consistent with those reported by Robertson and Walker, and we are therefore confident our implementation.

As an example, one pair of parameter values we investigated was $R = 13$ and $E = 15$. As discussed later, most queries improve with expansion with these values. For example, query 405 is *cosmic events*. The expansion terms are

> cobe cosmologists asteroids asteroid galaxies astronomers astronomy explorer astronomical particle particles nasa dust earth space

for which average precision increases from 0.0612 to 0.2360, and the recall increases from 13 to 30 documents. Most of the expansion terms are specific

[4] http://trec.nist.gov

[5] Lucy is a prototype search engine being developed at RMIT by the Search Engine Group. The primary aim in developing Lucy is to test techniques for efficient information retrieval. Lucy is available from http://www.seg.rmit.edu.au/lucy/

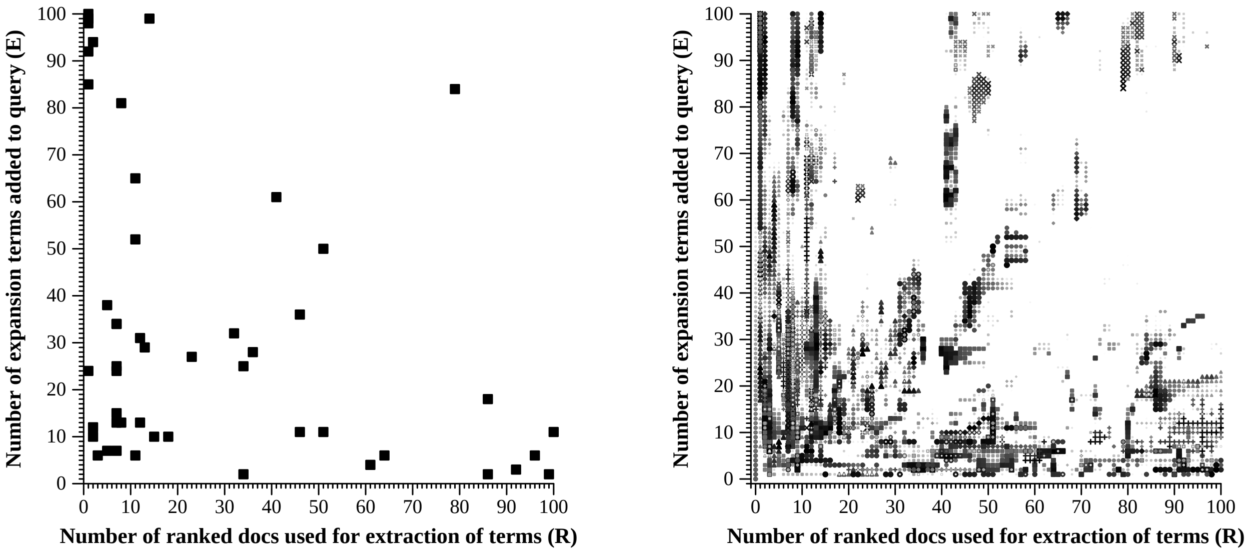

Figure 3: *On the left, parameter pairs for each query where expansion achieves maximum average precision. On the right, the best 100 parameter pairs in descending mark size for each query.*

to that topic and narrow down the search, especially given that the original query is fairly general. Note that we chose not to use stemming, because the query expansion process automatically identifies variant forms of words, and simplistically introducing further variants would be unlikely to be helpful.

In contrast, consider query 441, *lyme disease.* The expansion terms are

> spirochetes syphilis ticks neurological tick arthritis antibiotics deer conn infected infection symptoms diseases patients blood

and average precision falls from 0.6261 to 0.5523. In this case query expansion degrades the query, by introduction of low-relevance and general terms. The decline in average precision is not large, but it is certainly clear that for these parameters query expansion is unhelpful.

4 Exploring query expansion

Our long-term research aim is to find ways of improving query expansion, in particular to make it more robust. As an initial step, we explored the choice of values for parameters R and E, which have been held constant in previous work.

In our first experiment, on the TREC 8 data, we explored all combinations of R and E from 1 to 100, that is, we explored the effect of varying the number of documents used for expansion terms and the number of expansion terms chosen. For each of the 10,001 combinations, we ran all 50 queries and measured average precision.

Results are shown in the left-hand graph in Figure 1. In this graph, the darker the area, the greater the increase in average precision compared to no expansion. Thus, very roughly, the greatest improvement was seen for R between 8 and 16, and for E between 7 and 42. Choosing R of around 50 also gave good results. The vertical stripes correspond to places where average precision for a single query was dramatically improved (or degraded) by retrieval of an additional document with excellent (or awful) terms. On the other hand, sensitivity to the number of expansion terms is low.

The original expansion parameters of $R = 10$ and $E = 25$ are just within the dark "best" area. The average precision at this point is 0.254, up from 0.216 with no expansion. These are not quite the best choices; $R = 13$ and $E = 15$ gave slightly better results overall, of 0.260. However, the original values are impressively close to these settings.

Contrasting results are shown in the right-hand graph in Figure 1, for the TREC 9 10-gigabyte web collection. Expansion has on average been much less successful, with little overall improvement observed. The best effectiveness was at $R = 98$ and $E = 4$ (which is not in the neighbourhood of other successful points), and was only a slight improvement on effectiveness without expansion. Indeed, even at this point most queries were better without expansion.

A possible explanation for why the expansion method used in this paper works with the TREC 8 collection but not with the TREC 9 collection is that TREC 8 consists of newswire articles, whereas TREC 9 is made up of web data. The former consists of carefully reviewed text with a relatively controlled vocabulary on one particular topic, whereas web data is usually interspersed with links and other information that is on a diverse range of topics. The language in web data is often erratic and text commonly serves only to describe or otherwise accompany images, tables, or other items that are infrequent in newswire data.

Although this approach to QE does not lead to higher effectiveness on TREC 9 data, other methods are more successful. Our recent study suggests that using query associations (Scholer & Williams 2002) as a source of expansion terms leads to greater accuracy (Billerbeck, Scholer, Williams & Zobel 2003).

On the other hand, this method is not applicable in an environment like that of TREC 8, for which there is no suitable collection of queries from which query associations could be constructed. This seems to suggest that no one particular method of query expansion can be used to expand queries on all possible collections.

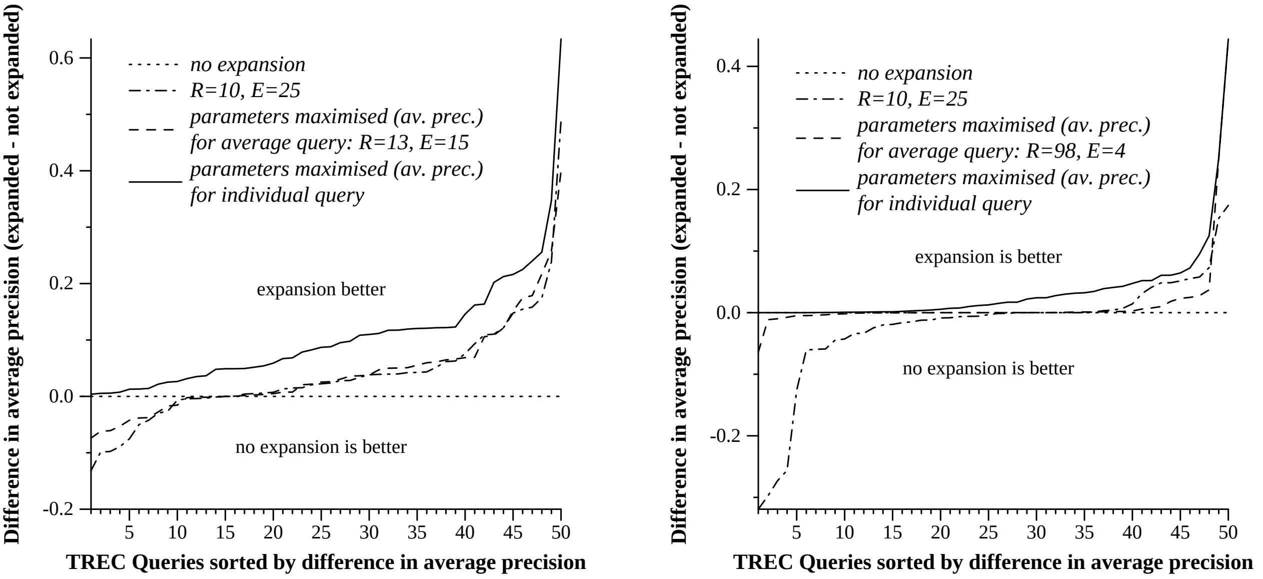

Figure 4: *Average precision at 1000 documents ranked with different parameter settings shown. TREC 8 is shown on the left and TREC 9 is on the right.*

Further issues with these results can be seen in Figure 2, which shows the parameters at which expansion is beneficial for two queries. For query 405, it can be seen that, in terms of average precision, expansion outperforms the unexpanded query for almost any parameter pair (in the region of the graph). For query 440, however, expansion is only beneficial at very limited numbers of parameter pairs; for almost all other combinations of those parameters, the query gets degraded.

However, most of the difficult-to-expand queries do have sets of parameters at which expansion is beneficial. Although one of the adhoc queries from TREC 9 could not be expanded at all, for all other queries at least some parameter pairs were found at which average precision was improved.

Interestingly, neither the default ($R = 10$, $E = 25$) nor the optimal ($R = 13$, $E = 15$) parameter settings are particularly effective for either of these two queries. It can also be seen that parameter pairs that work well for the one query don't work anywhere near as well for the other. Query 405 is best with (11,65), with average precision 0.3170, while query 440 is best with (61,4), with average precision 0.1240. Using (61,4) on query 405 gives average precision of 0.1193, worse by almost a factor of three; using (11,65) for query 440 gives 0.0149, worse by a factor of 9. This phenomenon can be observed for most of the pairs of queries.

More generally, the best expansion parameters vary wildly between queries, as illustrated in Figure 3, which on the left shows the optimal parameter pair for each of the 50 queries. If QE was applied with the optimal parameters for each query, the average precision would increase to 0.330 from 0.260, the best result observed when the same parameters are used for all queries.

The graph on the right of Figure 3 shows 100 coordinates with the highest average precision with descending mark sizes for each of the 50 queries. Most top coordinates for each query are clustered around the top spot (shown on the left hand side). Coordinates associated with a individual query are clustered in one of three possible ways:

- They are located along the x-axis, which means that a particularly important term is found (often in a specific document and therefore the corresponding streak starts at a certain x position);
- Similarly to the above, they are located along the y-axis, which means a document that has particularly good expansion terms has been added to the list of documents used for expansion; or
- A combination of both where coordinates are more closely centred on a particular coordinate.

The coordinates of a query where average precision has greatly changed (either improved or decreased) are likely to be arranged in one of the first two ways.

Intuition suggests that queries that are effective prior to expansion should be good candidates for QE, since many relevant documents with well-suited expansion terms are used as sources in the QE process. This intuition is strengthened by the observation that QE based only on relevant documents in the top R is superior to QE based on all documents, as we have seen in our experiments and as reported for example by Mano & Ogawa (2001). However, we found that there is no relationship between the average precision that the original query achieves and by how much QE improves average precision. Using the Pearson product moment correlation, no correlation was found between the improvement of average precision through query expansion and the average precision of the original query. This is illustrated in Figure 5, for which the Pearson product moment correlation rejects the null hypothesis of correlation between the two axes of the graph.

There is also an argument to be made for the converse: queries that did not perform well to begin with have greater scope for improvement through expansion. This also is not supported by the Pearson product moment correlation.

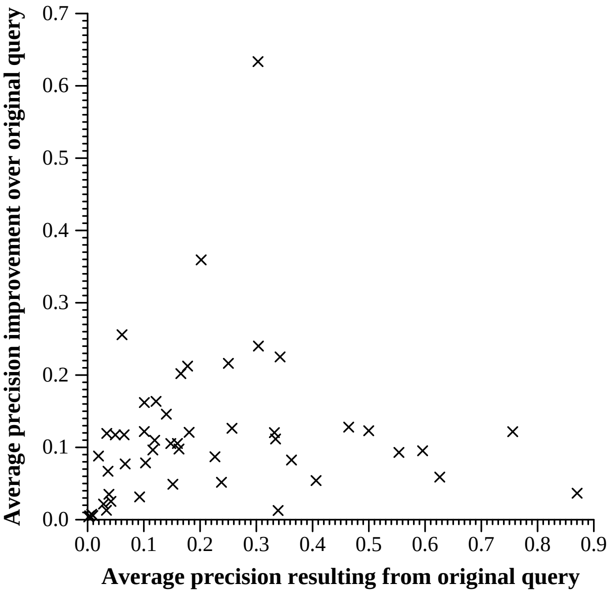

Figure 5: *Improvement through QE versus average precision of original query.*

An interesting question is then whether some property of the original query can be used to predict whether expansion will be effective. We explored a range of query metrics, but without clear success. These included the similarity score of the documents fetched in the original ranking; a measure of how distinct these documents were from the rest of the collection; specificity of the query terms; and an approximation to query clarity (Cronen-Townsend, Zhou & Croft 2002). None of these was effective.

Classification of queries – such as that by Broder (2002) into navigational, informational, and transactional – and using the hypothesis that only informational queries are expandable might lead to improvements. Employing an automatic query classification scheme (Kang & Kim 2003) makes selective expansion feasible and warrants further investigation in the future.

Per-query improvements in average precision due to QE are shown in Figure 4. The baseline (at 0.0) is the performance without expansion achieved by Lucy. The next line is the change in effectiveness from Lucy using the standard Robertson-Walker parameters, and the next is the change using the best parameters for that collection. The solid line is the performance with optimal parameters per query.

For the TREC 8 data, the standard parameters and best parameters are little different. However, choosing optimal parameters per query gives much greater effectiveness. For TREC 9, the standard parameters are very poor, with around half of the queries degraded and less than a third improved. The best parameters give much better performance; a small number of queries are degraded, but only slightly. With choice of the best parameters per query, all but 10 queries improve to some degree.

Sakai & Robertson (2001) have suggested varying parameters per query by classifying queries into one of 10 bins according to measures such as the similarity score of the highly-ranked documents. As we did not observe any correlation between such scores and improvements due to expansion, we are not convinced that such a strategy is likely to be successful.

Carpineto et al. (2001) experimented with varying R for some fixed E, and with varying E for some fixed R, considering the impact on average effectiveness for two data sets. They conclude that some limit on the number of expansion terms is warranted, but did not observe that the various settings had different impact on different queries. Our work generalises their results.

5 Conclusions

Query expansion is a successful method for improving the effectiveness of an information retrieval system, particularly for cases where there is a vocabulary mismatch between query and relevant document. Expansion is most successful when the documents that match the original query include topic-specific terms that can be automatically identified and then used to fetch further documents. For some queries, however, automatic expansion can introduce irrelevant terms that degrade effectiveness.

Furthermore, the precision of an expanded query of a certain proportion of queries is greatly increased or decreased, to the point where differences are approaching 50% in absolute terms. Surprisingly, the success or failure is often determined by a single expansion term, while most expansion terms have almost no effect on the query at all.

We have quantified the performance of a successful query expansion technique, by exploring behaviour as parameters are varied. This exploration has identified an upper bound on the improvement available via the Okapi approach to query expansion on two test collections, and showed that use of fixed parameters for all queries can be significantly improved upon.

We have identified that query expansion is much less reliable than previously suggested in the relevant literature. Despite the positive results reported in many previous papers, in our experiments query expansion failed on many queries and behaviour was highly inconsistent from collection to collection.

What is not clear is how the parameters should be chosen. We have preliminarily explored a range of options, but have not identified a metric that provides a method for guiding expansion. We are exploring how to use the results in this paper to develop new techniques for robust query expansion, as well as techniques for predicting whether expansion will be of value. Nonetheless, with appropriate parameter choices QE is a successful way of enhancing the effectiveness of queries, particularly on collections with consistently-written documents.

6 Acknowledgements

This research is supported by the Australian Research Council and by the State Government of Victoria. We thank the members of the Search Engine Group at RMIT University for assistance and feedback.

References

Arampatzis, A. & van der Weide, T. (2001), "Document filtering as an adaptive and temporally-dependent process".

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

Bharat, K. & Henzinger, M. R. (1998), Improved algorithms for topic distillation in a hyperlinked environment, *in* "Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval", Melbourne, AU, pp. 104–111.

Billerbeck, B., Scholer, F., Williams, H. E. & Zobel, J. (2003), Query Expansion using Associated Queries, *in* "Conference on Information and Knowledge Management", to appear.

Broder, A. (2002), "A taxonomy of web search", *ACM SIGIR Forum* **36**(2), 3–10.

Buckley, C., Salton, G., Allan, J. & Singhal, A. (1994), Automatic query expansion using SMART: TREC 3, *in* "Text REtrieval Conference".

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.

Cronen-Townsend, S., Zhou, Y. & Croft, W. B. (2002), Predicting query performance, *in* "Proceedings of the 25th Annual International Conference on Research and Development in Information Retrieval", SIGIR Forum, ACM Press, New Orleans, Louisianna, USA.

Foskett, D. J. (1997), Readings in information retrieval, *in* K. S. Jones & P. Willet, eds, "Thesaurus", Morgan Kaufman, San Francisco, California, USA, pp. 111–134.

Frakes, W. B. & Baeza-Yates, R., eds (1992), *Information Retrieval: Data Structures and Algorithms*, Prentice-Hall, Englewood Cliffs, New Jersey.

Harman, D. (1995), "Overview of the second Text REtrieval Conference (TREC-2)", *Information Processing & Management* **31**(3), 271–289.

Hoashi, K., Matsumoto, K., Inoue, N. & Hashimoto, K. (1999), Query expansion method based on word contribution (poster abstract), *in* "Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval", ACM Press, pp. 303–304.

Kang, I.-H. & Kim, G. (2003), Query type classification for web document retrieval, *in* "Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval", ACM Press, pp. 64–71.

Kaszkiel, M. & Zobel, J. (2001), "Effective ranking with arbitrary passages", *Journal of the American Society of Information Science* **52**(4), 344–364.

Kwok, K. L. (2002), Higher precision for two-word queries, *in* "Proceedings of the twenty-fifth annual international conference on Research and development in information retrieval", ACM Press, pp. 395–396.

Leuski, A. (2000), Relevance and reinforcement in interactive browsing, *in* "Conference on Information and Knowledge Management", pp. 119–126.

Mandala, R., Tokunaga, T. & Tanaka, H. (1999), Combining multiple evidence from different types of thesaurus for query expansion, *in* "Proceedings of the 22nd Annual International Conference on Research and Development in Information Retrieval", ACM Press, Berkeley, California.

Mano, H. & Ogawa, Y. (2001), Selecting expansion terms in automatic query expansion, *in* "Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval", ACM Press, pp. 390–391.

Page, L., Brin, S., Motwani, R. & Winograd, T. (1998), The pagerank citation ranking: Bringing order to the web, Technical report, Stanford Digital Library Technologies Project.

Qiu, Y. & Frei, H.-P. (1993), Concept based query expansion, *in* "Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval", ACM Press, pp. 160–169.

Rijsbergen, C. J. V. (1979), *Information Retrieval, 2nd edition*, Dept. of Computer Science, University of Glasgow.

Robertson, S. E. & Walker, S. (1999), Okapi/Keenbow at TREC-8, *in* "The Eighth Text REtrieval Conference (TREC-8)", NIST Special Publication 500-264, Gaithersburg, MD, pp. 151–161.

Robertson, S. E. & Walker, S. (2000), Microsoft Cambridge at TREC-9: Filtering Track, *in* "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* "Text REtrieval Conference", 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* "Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval", ACM Press, pp. 396–397.

Salton, G. (1989), *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*, Addison-Wesley, Reading, MA.

Salton, G. & McGill, M. J. (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, "Conference on Information and Knowledge Management", McLean, VA, pp. 324–331.

Schwartz, C. (1998), "Web search engines", *Journal of the American Society for Information Science* **49**(11), 973–982.

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., Wolfram, D., Jansen, M. B. J. & Saracevic, T. (2002), "From e-sex to e-commerce: Web search changes", *IEEE Computer* **35**(3), 107–109.

Voorhees, E. M. & Harman, D. K. (1999), Overview of the Eighth Text REtrieval Conference (TREC-8), *in* E. M. Voorhees & D. K. Harman, eds, "The Eighth Text REtrieval Conference (TREC 8)", National Institute of Standards and Technology Special Publication 500-249, Gaithersburg, MD, pp. 1–23.

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.

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

Xu, J. & Croft, W. B. (1996), Query expansion using local and global document analysis, *in* "Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval", ACM Press, pp. 4–11.