

An investigation of query expansion terms

ABSTRACT

This poster describes a framework for investigating the effectiveness of query expansion term sets and reports the results of an investigation on the quality of query expansion terms coming from different sources: pseudo-relevance feedback, web-based expansion, interactive elicitations from the user searchers, and expansion approaches based on query clarity.

The conclusion regarding the experimental framework is that certain different approaches show a substantial level of correlation, and can therefore be used interchangeably according to convenience considerations.

With regard to the actual comparison of different sources of expansion terms, the conclusion is that machines are better than humans at doing statistical calculations and at estimating which query terms are more likely to match documents that are relevant for a given topic. One consequence is a recommendation for research in implicit relevance feedback approaches and novel interaction models based on ostention or mediation, which have shown great potential.

1. INTRODUCTION

The work described here is based in part on our participation in the High Accuracy Retrieval from Documents (HARD) track of the Text Retrieval Conference (TREC), organized by the National Institute for Standards and Technology (NIST) which was introduced with the aim of exploring methods for improving the accuracy of document retrieval systems based on various forms of personalization (Allen 2004, 2005). NIST supported the track by providing assessors to answer “clarification forms”, through which the retrieval systems could get extra information via a brief interaction.

Like most other participants in HARD TREC 2005, we used a range of query expansion approaches, some automatic (based on query clarity, or on mining the web), some interactive (by asking the searcher, in the clarification forms, to provide extra information) and some mixed (we asked the searcher to filter expansion terms coming from automatic methods). We were unpleasantly surprised to realize that, while our sophisticated approaches did better than the baseline (simple search based on the standard topic representation), they were not better than pseudo relevance feedback (PRF). PRF is a simple and automatic procedure implemented as standard functionality on most IR toolkits, which assumes that the top ranking documents following a search are relevant, it extracts the most representative terms and uses them for query expansion, and re-runs the search with the expanded query. At the TREC conference it became rather clear than other participants in the HARD TREC had a similar experience: sophisticated expansion techniques and simulations of interactions with the searcher (which are expensive in terms of time spent and cognitive effort) did not show a significant improvement over the standard PRF.

Those results triggered the work described here. We are systematically investigating sets of expansion terms coming from different sources of evidence and trying to assess their quality and potential to improve retrieval effectiveness. While we are employing the HARD TREC experimental setting (document collection, topics, relevance judgments, clarification forms), the kind of investigation described here was not possible during the TREC experiment, when relevance judgments were not available. Based on these judgments, we can now establish optimal

upper-bounds of performance, and compare them with a number of approaches to query improvement.

2. METHODOLOGY

First, we use relevance judgments to build optimal expansion term sets; details are given in the next section. We then build candidate expansion term sets, based on a variety of sources and with a variety of methods. By evaluating the quality of these sets, we were able to conclude which sources and which methods of query expansion worked better.

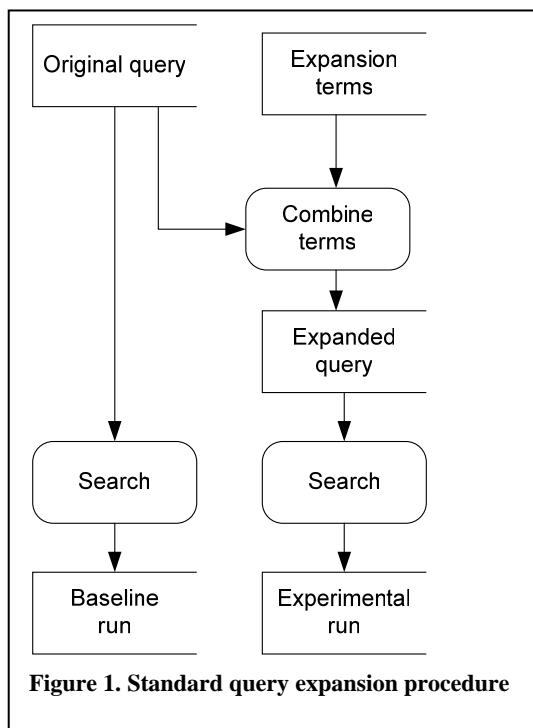
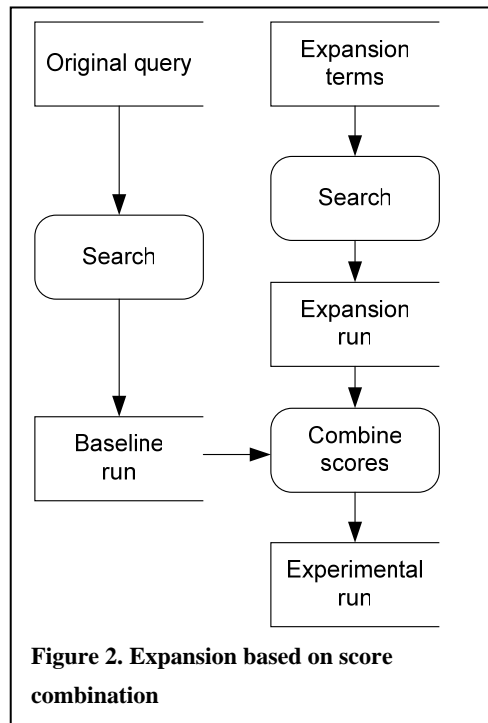


Figure 1. Standard query expansion procedure

We employed two ways to estimate the quality of the expansion sets or, more generally, of queries as representations of topics. The first is simply to compare them with the optimal sets by applying set operations and looking at overlap.



The second approach looks at the actual effectiveness improvement effected by the expansion terms. There are two approaches to use expansion terms. The standard approach, depicted in Figure 1, is to combine the original query terms with the expansion terms either by a simple concatenation. The alternative approach, depicted in Figure 2, uses the expansion terms as a query and it generates an “expansion run”, a list of documents that match the “expansion query”. If a retrieval model with a linear weighting formula is employed, then the same effect as the standard approach can be obtained by combining the scores in the baselines with those in the expansion run, and using the same weighting coefficient.

3. OPTIMAL UPPER-BOUNDS

The ideal query would be one which, for a certain information need, would retrieve all the relevant documents and no non-relevant documents. Although such an ideal query is probably impossible to create, we attempt to use existing relevance judgments to build “optimal queries”, which set upper-bounds of performance. The reasoning is the following: if an IR system with relevance feedback (RF) capability is given all the documents judged relevant for each test topic, then the expansion terms that it derives from them are optimal in terms of query quality, as judged by success in retrieval effectiveness.

Therefore, we employed the RelFBEval functionality found in the open source IR toolkit Lemur¹, which combines queries with relevance feedback, using as input empty queries, and using for relevance feedback the set of all documents judged relevant by the NIST assessors². The direct result of RelFBEval was our “optimal run” that constituted our upper-bound of performance³; the side effect was the weighted list of expansion terms that are representative for the set of relevant document (qrel expansion terms, or the optimal expansion set).

The optimal run does indeed show high level of performance on all three measures of effectiveness used in HARD TREC (mean average precision, precision at 10 and R-precision). Focusing on precision-oriented measures of performance is necessary for interactive retrieval sessions, simulated in our experiment.

4. SOURCES OF QUERY EXPANSION

We considered several source of query expansion:

1. Pseudo relevance feedback (PRF)

We obtained pseudo relevance feedback (PRF) runs with Lemur’s RetEval function. We used a number of topic representations, by combining titles and descriptions, and employed two different sets of parameters, (5, 10) and (10, 20), where the first number, *feedbackDocCount* indicates how many top ranking documents should be assumed relevant and the second, *feedbackTermCount*, indicates how many expansion terms should be used.

2. Web-based expansion

We followed the approach used by Roussinov (2005): we submitted our query to Google and built what we call a *web language model* for it. We employed the top 100 full-text pages returned by Google as “assumed relevant”, retained from them the concepts with the frequency of occurrence larger than their background frequency of occurrence on the

¹ <http://www.lemurproject.org>

² In NIST terminology, these are the “qrels”.

³ Note that the term “optimal” is relative, as it depends on a number of parameters. For consistency we adopted TfIdf as the retrieval model and 10 as the number of expansion terms used in generating relevance feedback runs.

web, and used a logistic regression categorizer (trained on TREC topics from the previous years) to predict concepts representative for the current topics.

3. Clarity-based expansion

Clarity of a query in relation to a document collection (intuitively the opposite of query ambiguity) is defined as the relative entropy between the query language model (LM) and the collection language model (Cronen-Townsend and Croft, 2002). In practice clarity is computed by conducting a search, assuming that the top ranking documents are relevant, and computing the language model of the documents at the top of the list. Apart from computing the overall Kullback-Leibler divergence between the language model of the top documents and that of the entire collection, the contribution of each term can be computed, and the terms with the highest contribution are highlighted as being representative for the query. We used these terms as expansion terms.

4. Interactive elicitations

We were inspired by Belkin's ASK model (1980) and UNC work in HARD TREC 2004 (Kelly et al, 2005) to use a clarification form to ask the searcher three specific questions additional terms: (i) "Describe what you already know about this topic"; (ii) "What sort of information would you like to have as a result of this search?"; and (iii) "Please input any additional keywords that describe your topic". We used their answers to derive expansion terms.

For each of these source of expansion terms, we compared three ways to estimate quality:

- term overlap with the optimal set
- effectiveness in a score combination expansion (Figure 1)
- effectiveness in a score combination expansion (Figure 2)

For the last two, we employed a matched pairs Wilcoxon test to compare mean average precision, precision at 10, and R-precision.

Apart from looking at the correlation between these measures of performance, we actually compared the different sets of expansion terms. Moreover, we compared them with the original TREC topic representations, which consisted in various combinations of title and description, terms from each of them being considered zero times, once, or more than once.

5. DISCUSSIONS AND CONCLUSIONS

This poster addresses core questions of information retrieval: how to estimate the quality of query expansion terms, and what are good sources of query terms. The comparisons between the evaluation approaches indicate high correlation, which gives a researcher flexibility in using one approach or another, based on circumstances.

The effectiveness results obtained here corroborate with our HARD TREC results and explain them: the automated interaction with a human information seeker is less likely to produce good query terms, and therefore less likely to achieve retrieval effectiveness superior to that obtained via fully automatic methods. This could be attributed to the human searcher's inability to grasp the statistics of a document collection and to generate terms that are representative for the relevant documents, and also distinguish them from non-relevant documents.

Algorithms are obviously better than humans at doing statistical calculations. One conclusion could be that the power of the algorithms, and especially of machine-learning procedures, should be harnessed even for highly interactive retrieval systems. This could be done by employing implicit relevance feedback, where the system "observes" behavioral cues that indicate interest in the document being examined, builds mathematical models of the topics of interest to the searcher, and retrieves more documents that match the topic model and the user profile, with the searcher's query just one source of evidence about what the user is interested in finding.

Such approaches dictate a re-evaluation of current interactive models, with more attention given to system based on ostention (Campbell, 1996) or on mediated retrieval (Muresan and Harper, 2004), which have shown substantial potential.

6. REFERENCES

- [1] Allen, J. (2004) HARD Track Overview in TREC 2004 – High Accuracy Retrieval from Documents, *Proceedings of TREC 2004*, Gaithersburg, November 2004.
- [2] Allen, J. (2005) HARD Track Overview in TREC 2004 – High Accuracy Retrieval from Documents, *Proceedings of TREC 2005*, Gaithersburg, November 2005.

- [3] Belkin, N.J. (1980) Anomalous states of knowledge as a basis for information retrieval. *Canadian Journal of Information Science*, v. 5: 133-143.
- [4] Campbell, I. (1996), The Ostensive Model of Developing Information Needs, Proceedings of COLIS-96, 2nd International Conference on Conceptions of Library Science.
- [5] Cronen-Townsend, S. and Croft, W.B. (2002) Quantifying Query Ambiguity, *Proceedings of HLT 2002*, San Diego, CA, March 24-27, 2002, pp. 94-98
- [6] Kelly, D., Dollu, V. D., & Fu, X. (2005, August 15-19). *The loquacious user: A document-independent source of terms for query expansion*. Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '05), Salvador, Brazil.
- [7] Muresan, G. and Harper, D.J. (2004) Topic Modelling for Mediated Access to Very Large Document Collections, in *JASIST*, 55 (10): 892 – 910, August 2004.
- [8] Roussinov, D., Zhao, L., and Fan, W. Mining Context Specific Similarity Relationships Using The World Wide Web. *In Proceedings of 2005 Conference on Human Language Technologies*.
- [9] Wilkinson, R. (1997) Using combination of evidence for term expansion, in *Information Retrieval Research – Proceedings of the 19th Annual BCS-IRSG Colloquium on IR Research*, Aberdeen, Scotland, April 1997.
- [10] ***Authors*** (2004) Our HARD TREC paper, Proceedings of TREC 2004, Gaithersburg, November 2004.
- [11] ***Authors*** (2005) Our HARD TREC paper, Proceedings of TREC 2005, Gaithersburg, November 2005.