Studying Query Expansion Effectiveness

Ben He and Iadh Ounis

Department of Computing Science
University of Glasgow
United Kingdom
{ben,ounis}@dcs.gla.ac.uk

Abstract. Query expansion is an effective technique in improving the retrieval performance for ad-hoc retrieval. However, query expansion can also fail, leading to a degradation of the retrieval performance. In this paper, we aim to provide a better understanding of query expansion by an empirical study on what factors can affect query expansion, and how these factors affect query expansion. We examine how the quality of the query, measured by the first-pass retrieval performance, is related to the effectiveness of query expansion. Our experimental results only show a moderate relation between them, indicating that the first-pass retrieval has only a moderate impact on the effectiveness of query expansion. Our results also show that the feedback documents should not only be relevant, but should also have a dedicated interest in the topic.

1 Introduction

Various approaches have been proposed to improve the query representation by reformulating the queries. Among them, query expansion is arguably one of the most effective approaches. In information retrieval (IR), query expansion is referred to as the techniques, algorithms or methodologies that reformulate the original query by adding new terms into the query, in order to achieve a better retrieval effectiveness. A classical query expansion algorithm is Rocchio's relevance feedback technique, proposed in 1971 [12] for the Smart retrieval system. It takes a set of documents as the feedback document set. Unique terms in this set are ranked in descending order of $tf \cdot idf$ weights. A number of top-ranked terms, including a fixed number of non-original query terms, are then added to the query. Many other query expansion techniques and algorithms were developed in the following decades, mostly derived from Rocchio's relevance feedback algorithm. For example, a popular and successful automatic query expansion algorithm was proposed by Robertson [11] while developing the Okapi system; Amati and Carpineto et al. proposed a query expansion algorithm in his Divergence from Randomness (DFR) framework [1,5].

Despite the marked improvement in the retrieval performance (e.g. [1,11]), query expansion can also fail, leading to a decreased retrieval performance. A typical example is the experiments conducted by various participants in the TREC Robust track, in which query expansion was reported to be unable to

M. Boughanem et al. (Eds.): ECIR 2009, LNCS 5478, pp. 611–619, 2009.

improve retrieval performance for a considerable number of so-called difficult queries [13]. Regarding the effectiveness and robustness of query expansion, there has been a few studies proposed in the literature. Carpineto et al. showed that the size of the feedback document set, and the number of expansion terms can affect the performance of query expansion [5,6]. Amati et al. predicted the effectiveness of query expansion by looking at the following two factors: the divergence of the query term's distribution in the feedback documents from its distribution in the whole collection, and the query term's appearances in the whole collection. A combination of these two factors, called $InfoQ$, is shown to have a moderate while significant correlation with the query expansion effectiveness [2]. Cao et al. use features such as the proximity of expansion terms to the query terms, query terms co-occurrences etc. to predict which expansion terms are useful [4].

In this paper, we aim to investigate the query expansion effectiveness from a perspective that is different from previous work. In particular, we argue that the main reasons for query expansion's failure can be summarised as follows: First, the feedback set contains too many non-relevant documents so that misleading expansion terms are added to the query. Second, documents in the feedback set, although containing relevant information, are sometimes only partially related to the topic, and can therefore yield bad expansion terms. This is also called topic drift in literature [9].

The remainder of this paper is organised as follows. Section 2 introduces the experimental settings of this paper. Section 3 studies how the first-pass retrieval performance affects the effectiveness of query expansion, and Section 4 investigates the connection between the distribution of query terms in the feedback documents, and the effectiveness of query expansion. Finally, Section 5 concludes this work and suggests future research directions.

2 Test Collection and Weighting Models

In this section, we introduce the collections and weighting models that are used in our study on the effectiveness of query expansion. We use the Terrier platform for both indexing and retrieval [10]. We experiment on the disk4&5 (minus the Congressional Record on disk4) of the TREC collections[1]. The test queries used are the 249 queries used in the TREC 2004 Robust track. All the test topics used are ad-hoc ones, which require finding as many relevant documents as possible [13]. We choose the Robust track queries for our study because compared to other TREC tasks, the Robust track has a large set of ad-hoc queries, and has been widely used for studying query expansion (e.g. [2,8]). All documents and queries are stemmed using Porter's stemmer. Standard stopword removal is also applied. We only experiment with the title field of the queries, which are usually very short, containing few keywords.

For our study, we apply two different weighting models for comparison. The first one is the DPH model [3,7], derived from the DFR framework [1]. Note

[1] Related information of disk4&5 of the TREC collections can be found from the following URL: http://trec.nist.gov/data/docs_eng.html

that DPH is a parameter-free model. All variables in its formula can be directly obtained from the collection statistics. No parameter tuning is required to optimise DPH. We also apply the Okapi's BM25 formula, which is one of the most established weighting models [11]. The parameters are set to $k_1 = 1.2$ and $k_3 = 8$ by default [11]. Moreover, BM25's term frequency normalisation parameter b is set to 0.35 using Simulated Annealing by optimising the mean average precision (MAP) on the queries from the TREC 2004 Robust track.

For query expansion, we measure the Kullback-Leibler (KL) divergence between a term's distribution in the feedback documents and that in the whole collection. In our experiments, the feedback document set contains the *exp_doc* top-ranked documents, from which the *exp_term* most weighted terms by KL are extracted. We scan a wide range of possible values of *exp_doc* and *exp_term*, namely every *exp_doc* value within $2 \leq exp\_doc \leq 10$, and $10 \leq exp\_term \leq 100$ with an interval of 5. We obtain $exp\_doc = 5$ and $exp\_term = 20$, which are used in our experiments in this paper.

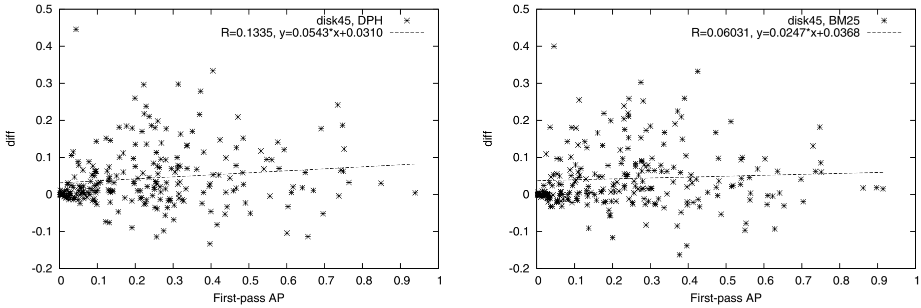
3 First-Pass Retrieval Performance and Query Expansion Effectiveness

In this section, we investigate how the first-pass retrieval performance is related to the effectiveness of query expansion by studying the following question: Does a better first-pass retrieval lead to a better effectiveness of query expansion? In other words, is the retrieval performance improvement brought by query expansion correlated with the first-pass retrieval performance?

We might intuitively consider the first-pass retrieval performance and the query expansion effectiveness to be highly correlated, since the query expansion takes the first-pass retrieval result for feedback, and reformulates the query based on the feedback documents. To test this assumption, we conduct experiments to estimate the correlation between the first-pass retrieval performance, measured by AP, and the improvement brought by query expansion.

In our study, we define the improvement brought by query expansion as the difference in the average precision values between the first-pass (AP) and second-pass (QEAP) retrieval, namely *diff=QEAP-AP*. Using the topics from the TREC 2004 Robust track, we compute the linear correlation between the first-pass retrieval AP and *diff*.

Figure 1 plots the first-pass retrieval performance measured by AP against the improvement in AP brought by query expansion for (a) DPH and (b) BM25. From Figure 1, it is surprising to see that there is almost no correlation between the first-pass AP and the effectiveness of query expansion. The correlation is insignificant for both weighting models used. We argue that this is because the improvement in AP that we expect from query expansion is not linearly related to the first-pass AP. If the first-pass AP is too low, the query expansion mechanism does not have a good pseudo relevance set to extract useful expansion terms. On the other hand, if the first-pass AP is too high, there might be only little room for potential improvement. Therefore, the relation between the first-pass AP (AP) and the improvement in AP brought by query expansion (*diff*) is non-linear.



(a) DPH

(b) BM25

Fig. 1. The average precision obtained before query expansion (AP) and the improvement in AP brought by query expansion (*diff*) using DPH and BM25, respectively. No significant linear correlation R between AP and *diff* is found.

Hence, we assume the following quadratic function for the relation between the first-pass AP and the improvement in AP brought by query expansion (*diff*):

$$diff = f(AP) = -\alpha(AP - \lambda)^2 + \beta \tag{1}$$

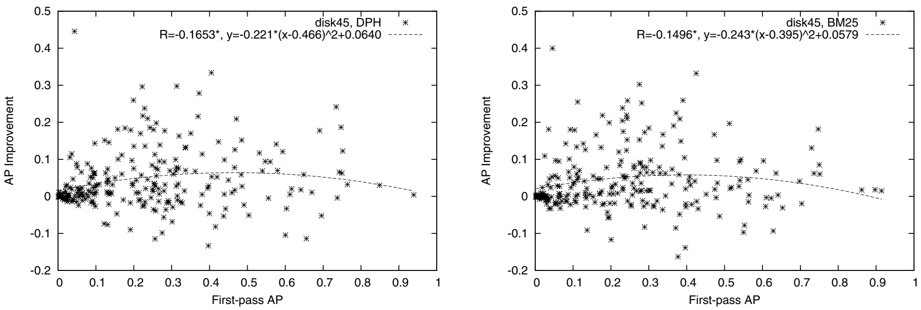
where α, β and λ are parameters of the quadratic function. In particular, when the first-pass AP equals to λ, *diff* is maximised, indicating the maximum potential improvement that query expansion can provide.

Figure 2 plots the results obtained by fitting the above quadratic function. The linear correlation is computed between $|AP - \lambda|$ and *diff*. We found a weak negative correlation between $|AP - \lambda|$ and *diff* that is significant at 0.05 level. The obtained negative correlation can be explained as follows: The further away from λ the first-pass AP is, the less potential improvement query expansion can achieve. Moreover, although the negative correlation is found to be significant, the relative low correlation still indicates a weak association between the first-pass retrieval performance and the effectiveness of query expansion.

In this section, we have studied the relation between first-pass retrieval performance and the effectiveness of query expansion. The results indicate only a weak link between the first-pass retrieval performance and the improvement query expansion provides. Prompted by this, in the next section, we give a closer look at the first-pass retrieval, particularly at the feedback document set.

4 Distribution of Query Terms in the Feedback Documents

In the previous section, we have shown that query expansion can still fail even if 80% of the feedback documents are relevant. We argue that this is due to the second reason that can cause the failure of query expansion, namely topic drift. The query expansion mechanism extracts the most informative terms from the



(a) DPH

(b) BM25

Fig. 2. The average precision obtained before query expansion (AP) and the improvement in AP brought by query expansion (*diff*) using DPH and BM25, respectively. A R value marked with a star indicates a significant linear correlation between $|AP - \lambda|$ and *diff* at the 0.05 confidence level.

feedback documents. In some cases, although a feedback document is relevant, there could be only a subset/paragraph of the feedback document that contains relevant information. Thus, off-topic terms are possibly added to the query, resulting in a decrease in the retrieval performance. Therefore, it is necessary to examine the distribution of query terms in the feedback documents to see to which degree the feedback documents is interested in the topic.

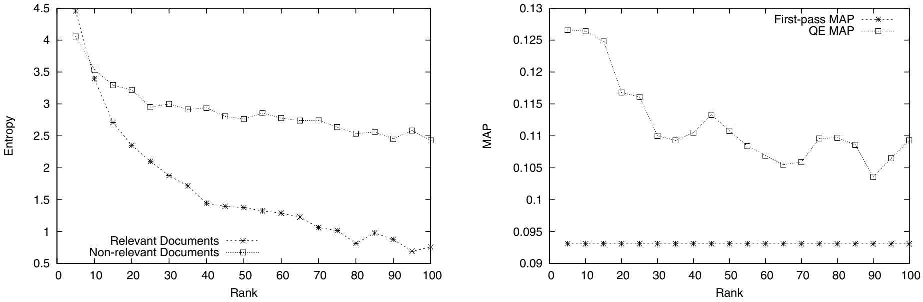
We propose to quantify the degree of interest in the query topic of the feedback document by the Entropy measure, which estimates how the occurrences of a query term spreads over different subsets of a feedback document. The higher Entropy is, the more the feedback document is related to the topic. We define the Entropy measure for a query term t in a document d as follows:

$$Entropy(t, d) = -\sum p_i \cdot \log_2 p_i \tag{2}$$

where p_i is the probability of observing the query term in the ith subset of the document. In order to avoid assigning zero probability to parts where the query term does not appear, we apply Laplace smoothing as follows:

$$p_i = \frac{tf_i + 1}{tf + n} \tag{3}$$

where tf_i is the term frequency in the ith subset of the document, and *tf* is the term frequency in the document. n is the number of subsets that the document is splitted into. In the TREC 2004 Robust track, the average document length of all the judged documents is 1375 words. We arbitrarily assume that each subset of a document has approximately 100 words, and fix n to 14. Note that when the query term is uniformly distributed in the document, i.e. p_i is the same across all subsets of the document, the Entropy measure is maximised. We also define *Entropy(Q, D)*, the Entropy of query Q in a document set D, by the mean of *Entropy(t,d)* of all query terms in all documents in D.



(a) Entropy against the rank of sampled documents

(b) MAP obtained with baseline (First-pass MAP) and query expansion (QE MAP) against the rank of feedback documents

Fig. 3. Entropy and mean average precision (MAP) against the rank of sampled documents. Results in the figure are obtained using DPH. BM25 provides almost identical results.

We study how the query terms are distributed within documents at different levels of ranks in the returned results. We use the 110 submitted runs in the TREC 2004 Robust track for our study. We split the top 100 returned documents into 100/r levels. From the rank 1, which is the top rank, until the rank 100, we randomly sample three relevant and three non-relevant documents for every r ranks. We fix r to 5 in our experiments. For example, if we randomly sample three relevant and three non-relevant documents from all the top-5 ranked documents in the 110 TREC runs, we do the same for all the documents ranked from 6th to 10th in the TREC runs. Moreover, in order to prevent having overlap between samples of different levels of ranks, each sampled document should appear in only the samples at one level of ranks. For example, imagine a document is ranked 4th by run A, and 7th by run B. If this document is in the sample of documents ranked from 1st to 5th, it will not be sampled again to represent documents ranked from 6th to 10th. Moreover, we sample only three relevant and three non-relevant documents at each level of ranks so that there are still enough relevant documents when the sampled rank is around 100. We do not sample further beyond the top-100 documents because in the TREC 2004 Robust track, only the top 100 returned documents in a selected set of submitted runs are judged by assessors. In this case, it is indeed very difficult to find relevant documents that are ranked after the top 100 returned documents.

At each level of ranks, for each query, we compute *Entropy(Q, D)* for the sampled relevant and non-relevant documents, respectively. Figure 3(a) plots the Entropy values against the sampled ranks of the returned relevant and non-relevant documents, respectively. From Figure 3(a), we can see that on one hand, the Entropy measure for relevant documents ranked at top 5 is very high, while it decreases rapidly when the ranking becomes lower. On the other hand, the Entropy measure for non-relevant documents decreases steadily when the ranking decreases, and the curve for the non-relevant documents is nearly flattened at the end. Moreover, we find a significant negative correlation between the Entropy measure and the rank of both the sampled relevant documents and non-relevant documents. The linear correlation values are R=-0.8758 for relevant documents, and R=-0.8845 for non-relevant documents. However, we find no correlation at all when relevant and non-relevant documents are mixed together. The above findings have the following implications:

First, a query can possibly have more than one concept. For example, the query *"radio waves and brain cancer"* has two different concepts: *"radio waves"* and *"brain cancer"*. In a document collection, there could be many documents that cover either of these two concepts, but not both. In this case, these documents are usually non-relevant and have moderate Entropy values.

Second, some relevant documents have a dedicated interest in the topic throughout them, which are top-ranked by retrieval systems and have high Entropy values. Apart from the highly relevant documents, some other documents are not generally about the topic, but contain relevant information in some subsets of them. Therefore, they are also judged relevant. This explains why the top ranked relevant documents have high Entropy values, while other relevant documents' Entropy values are much lower.

As mentioned before, a good feedback document should not only be relevant, but also have a dedicated interest in the topic. Therefore, if we use only the relevant documents for feedback, we expect the feedback documents with higher Entropy values to provide better retrieval performance after query expansion. To test this hypothesis, we run query expansion using the sampled relevant documents at each level of ranks. The sampled relevant documents are removed from both first-pass and second-pass retrieval so that the second-pass document ranking is not biased towards to sampled relevant documents.

Figure 3(b) plots the mean average precision obtained by query expansion against the level of ranks at which the relevant documents are sampled. From Figure 3(b), we can see that although the feedback documents are all relevant, the effectiveness of query expansion decreases when the ranks of the feedback documents decrease. This can be explained by our previous experiments using the Entropy measure: if the relevant feedback documents are not highly ranked, they are likely to be only partially related to the topic. In this case, it is unlikely that all the extracted query terms are useful, and hence, the improvement brought by query expansion decreases.

In summary, when the feedback documents are all relevant, the effectiveness of query expansion is still affected by the degree of interest the feedback documents show in the topic. The highly ranked relevant documents are very closely related to the topic, and therefore have high Entropy values. However, relevant documents that are not ranked highly are less likely to have a strong, dedicated interest in the topic, due to the fact that most relevant documents are only partially related to the topic. Consequently, the effectiveness of query expansion also decreases together with the ranking of the relevant feedback documents.

5 Conclusions and Future Work

In this paper, we have conducted an empirical study on the effectiveness of query expansion. On the TREC 2004 Robust track topics, we investigate the two possible reasons for the failure of query expansion, namely the low query quality and topic drift. Our experimental results show that the quality of the query, measured by the first-pass retrieval performance, has a moderate association with the effectiveness of query expansion. Moreover, in case of the real relevance feedback, where the feedback documents are known to be relevant, the feedback documents should contain a dedicated interest in the topic.

This paper is a step towards a better understanding of query expansion. Our findings suggest various future research directions. For example, we may be able to utilise the Entropy measure to select good feedback documents for query expansion, in which a strong interest in the topic exists. We may also find good expansion terms by looking at the co-occurrences of candidate expansion terms and the query terms in paragraphs where the query terms are particularly frequent. This is the objective of our future research.

Acknowledgements

We would like to thank Craig Macdonald for his helpful comments on the work. This work is funded by the SIMAP: Simulation modelling of the MAP kinase pathway. EC project 2006-2009.

References

1. Amati, G.: Probabilistic models for information retrieval based on divergence from randomness. PhD thesis, Department of Computing Science, University of Glasgow (2003)
2. Amati, G., Carpineto, C., Romano, G.: Query Difficulty, Robustness, and Selective Application of Query Expansion. In: McDonald, S., Tait, J.I. (eds.) ECIR 2004. LNCS, vol. 2997, pp. 127–137. Springer, Heidelberg (2004)
3. Amati, G., Ambrosi, E., Bianchi, M., Gaibisso, C., Gambosi, G.: FUB, IASI-CNR and University of Tor Vergata at TREC 2007 Blog Track. In: Proceedings of TREC 2007 (2007)
4. Cao, G., Nie, J., Gao, J., Robertson, S.: Selecting good expansion terms for pseudo-relevance feedback. In: Proceedings of SIGIR 2008 (2008)
5. Carpineto, C., de Mori, R., Romano, G., Bigi, B.: An information-theoretic approach to automatic query expansion. ACM Transactions on Information Systems 19(1) (2001)
6. Carpineto, C., Romano, G., Gianini, V.: Improving retrieval feedback with multiple term-ranking function combination. ACM Transactions on Information Systems 20(3) (2002)
7. He, B., Macdonald, C., Ounis, I., Peng, J., Santos, R.L.T.: University of Glasgow at TREC 2008: Experiments in Blog, Enterprise, and Relevance Feedback Tracks with Terrier. In: Proceedings of TREC 2008 (2008)

8. Kwok, K., Grunfeld, L., Sun, H., Deng, P.: TREC 2004 Robust Track Experiments Using PIRCS. In: Proceedings of TREC 2004 (2004)
9. Macdonald, C., Ounis, I.: Expertise drift and query expansion in expert search. In: Proceedings of CIKM 2007 (2007)
10. Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., Lioma, C.: Terrier: A high performance and scalable judgements retrieval platform. In: Proceedings of the OSIR Workshop (2006)
11. Robertson, S.E., Walker, S., Beaulieu, M.M., Gatford, M., Payne, A.: Okapi at TREC-4. In: Proceedings of TREC 4 (1995)
12. Rocchio, J.: Relevance feedback in information retrieval, pp. 313–323. Prentice-Hall, Englewood Cliffs (1971)
13. Voorhees, E.: TREC: Experiment and Evaluation in Information Retrieval. MIT Press, Cambridge (2005)