RMIT University at TREC 2008: Relevance Feedback Track

Yohannes Tsegay Falk Scholer Simon Puglisi

School of Computer Science and IT
RMIT University, GPO Box 2476V
Melbourne 3001, Australia

1 Introduction

This report outlines TREC-2008 Relevance Feedback Track experiments done at RMIT University.

Relevance feedback in text retrieval systems is a process where a user gives explicit feedback on an initial set of retrieval results returned by a search system. For example, the user might mark some of the items as being relevant, or not relevant, to their current information need. This feedback can be used in different ways; one approach is query expansion, where terms from the relevant documents are added to the original query, with the aim of improving retrieval effectiveness.

This report describes the the query expansion methods that we explored as part of TREC 2008. Our results demonstrate that high weight terms are not always necessarily useful for query expansion.

2 Term Selection

The 2008 Relevance Feedback Track provided a set of relevance judgements for participants to use. These judgements are based on data from previous TREC tracks (2004–2006 Terabyte Tracks, and the 2007 Million Query Track). Different runs for the Relevance Feedback Track made use of varying numbers of relevant and non relevant documents (see Section 4 for details).

Based on the set of available documents with known relevance judgements, a query expansion scheme aims to identify the set of terms that, when added to the original query, is most likely to be able to boost retrieval performance. For all the experiments reported in this paper, queries were expanded using only terms which occur in the documents provided for expansion. No new terms were introduced from external sources.

Let R be the set of documents in the collection that are known to be relevant for the current query (that is, the set of relevant documents provided as part of the Track framework). To expand a query, a set of candidate expansion terms S is first established. In our experiments, we explore two approaches for the construction of the candidate term set, S. In the first, we combine all the terms from the provided relevant documents R into a single *term-pool*. Treating the available expansion documents as a single unit provides a means of selecting expansion terms which are signature to the set of relevant documents as a whole. Furthermore, the question of, how expansion terms from different documents should be combined can be avoided. We call this approach METHOD1.

In the second approach, a set S_d is constructed separately for each relevant document, d. Here, term weights are first calculated for each set S_d independently, using one of the weighting schemes described below. The top ranked terms from each set are then added to the original query by selecting the top terms in an interleaved fashion. Preference is given to terms that occur in over half the expansion documents. This approach ensures that the expansion terms are sourced from a variety of documents; in the first method, it is possible for terms from a small subset of relevant documents to dominate. We call this approach METHOD2.

Once the candidate term sets are constructed, term weighting approaches are used to rank and select the final expansion terms. In our submitted runs, we make use of a TF $\times$ IDF approach. Here, a term's weight is calculated as the product of its occurrence frequency within the set S, and its inverse document

frequency (the reciprocal of the count of documents which contain the term) in the collection. While there are several variations for the calculation of TF and IDF weights, we base our formulation on the logarithmically smoothed approach (Zobel and Moffat, 1998):

$$TF \times IDF = \log(f_{S,t} + 1) \times (\log(N/f_t) + 1)$$

where $f_{S,t}$ is the number of times that term t occurs in the set S, N is the count of documents in the collection, and f_t is the number of documents in the collection in which t occurs. Selected terms are thus those which occur frequently in the set of known relevant documents, but occur rarely across the collection.

An alternative term weighting scheme is based on language modeling, where the Kullback-Leibler divergence (Kullback and Leibler, 1951) — the relative entropy of the model of the set S against the model of the collection C — is used to estimate the weight of terms in S:

$$KL(t, S, C) = P(t|S_{model}) \times \log \frac{P(t|S_{model})}{P(t|C_{model})}$$

where $P(t|M)$ is the probability of observing t in a model M, and S_{model} and C_{model} are the unigram language models of the set S and the collection as a whole, respectively.

Lastly, to minimise the possible confounding effect of query drift, where queries are expanded with terms that lead the query away from relevant answers, the contribution of the expansion terms to the overall similarity score is adjusted relative to the original query terms. Expansion terms are down-weighted by a factor of one-third compared to the original query terms.

3 Document Preprocessing

Prior to term selection, each document was parsed and terms were extracted. A term was defined as a sequence of alphanumeric characters delimited by whitespace Williams and Zobel (2005). Terms were then case-folded and stemmed. We excluded from this set all stopwords[1], URLs, and floating point numbers.

To avoid repetition, expansion terms identical to the original query terms were removed from consideration. Terms were then scored using either of the schemes presented in the previous section. Each query was then expanded using the top 25 unique terms (Billerbeck and Zobel, 2004).

4 Runs

The `gov2` collection was indexed using the ZETTAIR Search engine[2]. Terms were stemmed using the Porter stemmer. The expanded queries were processed using a Dirichlet-smoothed language model.

The Track framework made available a set of training queries (the odd-numbered Million query track topics and corresponding relevance judgments). Based on initial experiments, we found that the TF $\times$ IDF approach performed marginally better than the Kullback-Leibler divergence approach. Due to a limited number of runs that could be officially submitted, we opted to use the former term weighting method only, and to explore the effect of forming the candidate term-pool S with METHOD1 and METHOD2 (these correspond to the official submitted runs named RMIT08.*2 and RMIT08.*1).

The track contained five runs, A–E described below:

- **Run A:** Baseline retrieval, no relevance information was provided and queries are run with no expansion.
- **Run B:** For each query, a single relevant document was provided for expansion.
- **Run C:** Three relevant documents and three not relevant documents were provided.
- **Run D:** Ten judged document were provided where at least 3 were relevant and 3 were not.

[1] The list of stopwords used is available at `http://www.csse.unimelb.edu.au/~jz/resources/stopping.zip`

[2] http://www.seg.rmit.edu.au/zettair/

Run	P@10	P@20	P@100	MAP	BPREF
A	0.0742	0.0645	0.0410	0.0378	0.0803
B	0.0742	0.0645	0.0410	0.0378	0.0803
C	0.0323	0.0242	0.0116	0.0114	0.0271
D	0.0355	0.0258	0.0119	0.0106	0.0261
E	0.0355	0.0242	0.0100	0.0087	0.0248

Table 1: METHOD1 runs, where the expansion terms were selected from a single pool of terms from all relevant document, ordered by decreasing TF $\times$ IDF weight.

Run	P@10	P@20	P@100	MAP	BPREF
A	0.0742	0.0645	0.0410	0.0378	0.0803
B	0.0742	0.0645	0.0410	0.0378	0.0803
C	0.0226	0.0242	0.0174	0.0077	0.0395
D	0.0387	0.0323	0.0135	0.0137	0.0345
E	0.0935	0.0774	0.0284	0.0219	0.0597

Table 2: METHOD2 runs, where a terms in each document are ranked separately by decreasing TF $\times$ IDF order, and the top ranked terms are intersected.

- **Run E:** A larger number of judged documents (40–800) documents was provided.

For our runs, the non-relevant documents provided in the above sets were disregarded, and only information in relevant documents was utilised.

A single variant, METHOD1, was submitted for runs A and B. For runs C – E, where several expansion documents were available, METHOD1 and METHOD2 runs were submitted.

5 Results

The effectiveness of the query expansion strategies is shown in Tables 1 and 2. Adding expansion terms negatively impacts on performance as measured by P@10, P@20, P@100, MAP and BPREF. Run C yields the lowest effectiveness for both methods, leading to a relative decrease in P@20 of 62%. The losses are even greater for MAP and BPREF. Failure analysis shows that while expansion did help for some queries, a larger number were harmed by the expansion process (of the 31 judged queries, 5 have shown an increase in performance, while the performance of 15 queries was decreased). The two queries which have shown the best improvements due to expansion and the two queries whose performance decreased most after expansion are shown in Figure 1 and 2 respectively.

Run D demonstrates slight improvements from run C and these are more pronounced in METHOD2. This can be attributed to the fact that precedence is given to terms which occur in the largest number of expansion documents. In METHOD1 a term's weight is computed based on its frequency in the term-pool S, and therefore it does not make any distinction between a high weight term sourced from a single expansion document and another that occurs in a large number of the expansion documents.

The above observation are consistent with run E, where for METHOD2 it is the only run to have performed better than the baseline in P@10 and P@20. Although more effective than runs C and D, the MAP and BPREF scores for run E are still worse than the baseline.

6 Conclusions

As part of the 2008 Relevance Feedback Track, we experimented with two TF $\times$ IDF based query expansion approaches. While some queries benefited form both approaches, the techniques failed to lead to consistent

Figure 1: Terabyte track topics 782 and 772 which have yielded the best improvements. The $*$ next to a term indicates that it is an expansion term.

```
Topic:  782
orange varieties seasons fruit* valencia* season* october*
midseason* florida* lemons* navel* nfc* concentrated* oranges*
economic* growers* exporter* juice* crop* grapefruit* fresh*
citrus* conditions* september* tangerines*


Topic:  772
rules flag display patriotic* manner* army* speakers* anthem*
dipped* tribute* honor* free* flown* july* fastened* military*
inclement* lapel* individuals* executive* pledge* horizontally*
international* shoulder* federal* position* flagstaffs* blue*
```

Figure 2: Terabyte track topics 814 and 850 which have shown the most decrease in performance. The $*$ next to a term indicates that it is an expansion term.

```
Topic:  814
flood johnstown general* prolific* press* pittsburgh* frick* clara*
strayer* club* dam* cambria* oconnor* say* gertrude* mellon*
hildegarde* time* clay* barton* reverend* conemaugh* quinn* colonel*
city* henry* twovolume*


Topic:  850
flood mississippi river army* elevation* moines* levees* louis*
engineers* federal* late* missouri* basin* general* upstream*
drainage* upper* downstream* series* corps* nearrecord* feet*
business* application* ohio* magnitude* waterresources* skunk*
```

improvements. In the case of run E, even though improvements are observed in early precision scores, only 42% of the queries had shown improvements in MAP. We plan to conduct further failure analysis to try and identify the properties of queries for which the approach harmed performance.

References

Bodo Billerbeck and Justin Zobel. Questioning query expansion: an examination of behaviour and parameters. In *Proceedings of the 15th Australasian database conference*, pages 69–76, Darlinghurst, Australia, 2004. ACS, Inc.

S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951. URL `http://www.jstor.org/stable/2236703`.

H.E. Williams and J. Zobel. Searchable words on the web. *International Journal of Digital Libraries*, 5(2):99–105, 2005.

Justin Zobel and Alistair Moffat. Exploring the similarity space. *SIGIR Forum*, 32(1):18–34, 1998.