Exploring Accumulative Query Expansion for Relevance Feedback

Debasis Ganguly, Johannes Leveling, and Gareth J. F. Jones

CNGL, School of Computing, Dublin City University, Dublin 9, Ireland
{dganguly, jleveling, gjones}@computing.dcu.ie

Abstract. For the participation of Dublin City University (DCU) in the Relevance Feedback (RF) track of INEX 2010, we investigated the relation between the length of relevant text passages and the number of RF terms. In our experiments, relevant passages are segmented into non-overlapping windows of fixed length which are sorted by similarity with the query. In each retrieval iteration, we extend the current query with the most frequent terms extracted from these word windows. The number of feedback terms corresponds to a constant number, a number proportional to the length of relevant passages, and a number inversely proportional to the length of relevant passages, respectively. Retrieval experiments show a significant increase in MAP for INEX 2008 training data and improved precisions at early recall levels for the 2010 topics as compared to the baseline Rocchio feedback.

1 Introduction

Query expansion (QE) is a popular technique to improve information retrieval effectiveness by extending the original query. The Relevance Feedback (RF) track at INEX 2010 attempts to simulate user interaction by communicating *true* relevance information between a *Controller module*, with access to the *qrels* file and simulates RF from a user, and a *Feedback module*. This allows re-ranking results by changing the set of retrieved documents in every retrieval iteration. In the RF track, the incremental reporting of relevant text segments from full documents allows the development of a feedback algorithm choosing feedback terms in different ways, compared to standard Blind Relevance Feedback (BRF). The exchange of relevance information between user and system denotes a retrieval iteration and can be repeated multiple times for the same query. In each iteration, the feedback algorithm or its parameters can be adapted to improve retrieval performance.

In this paper, we investigate the relationship between the length of relevant passages and the number of feedback terms. We explore three variants of selecting the number of feedback terms depending on the length of relevant test segments: i) choosing a constant number of feedback terms, ii) choosing a number directly proportional to the lengths of the relevant segments, and iii) choosing a number inversely proportional to the lengths of the relevant segments. All three approaches can be justified in their own way. One might want to choose more

terms from a smaller relevant segment in the hope that it has less or no noisy terms. It might be more effective to choose more terms from larger relevant segments on the assumption that the likelihood of finding useful expansion terms increases with the the length of a relevant section. Finally, the length of a relevant passage may be unrelated to the best number of feedback terms so that a constant number of feedback terms is the best choice.

The rest of this paper is organized as follows: Section 2 describes the motivation of the RF experiments and related work, our RF algorithm is introduced in Section 3, Section 4 reports our results in the RF track and analyzes the results and we conclude the paper with directions for future work in Section 5.

2 Related Work

One of the problems of BRF is that *all* terms which meet the selection criterion for feedback terms are used for QE. This includes terms which are not related to the query, for example semantically unrelated, but highly frequent terms from long (pseudo-)relevant documents or text segments.[1] A number of experiments using small text passages instead of full documents for BRF have been conducted [1–5]. One assumption behind using small passages is that long documents can contain a wider range of discourse and noisy terms would be added to the original query, which can result in a topic shift. A wide range of discourse in long documents means that the relevant portion of such a document may be quite small and feedback terms should be extracted from relevant portions only. Another assumption behind these approaches is that even non-relevant documents can contain passages with useful feedback terms [6].

In contrast, our experiments for the RF track at INEX 2010 aim at investigating if *true* relevant text passages also contain noise so that a segmentation into smaller textual units (in this case: word windows) will improve IR effectiveness. The motivation behind our method is the assumption that even large *true* relevant text passages contain harmful terms for QE. Furthermore, the RF track provides the opportunity to explore how to use true relevant passages for QE.

LCA [7] involves decomposing the feedback documents into fixed length word windows to overcome the problem of choosing terms from unrelated portions of a long document. The word windows are ranked by a score which depends on the co-occurrence of a word with the query term. Similar to LCA, we presume that terms in close proximity to query terms are good candidates for QE. In our RF method, we select feedback terms from word windows which are maximally similar to the query, the similarity being measured by Lucene's default similarity which is a variant of $\texttt{tf} \cdot \texttt{idf}$, thus achieving the same effect of filtering out potentially irrelevant parts of a longer document as in LCA. A major difference with respect to LCA is that we do not compute term co-occurrences explicitly.

[1] We employ the term segment in its most general sense, denoting sentences, paragraphs, and other small text units such as word windows.

3 System Setup

In contrast to other evaluation tracks in IR, submissions to the RF track comprise of an implemented software module (a JAVA .jar file). We submitted 3 RF modules for the RF track at INEX 2010. As a baseline, we use standard Rocchio feedback [8], which was packaged as a default feedback module implementation by the INEX organizers. We use the Lucene API[2] for indexing and retrieval. The RF track simulates a user highlighting relevant passages if any for each document presented to him. The feedback module re-ranks the initial results based on relevance information.

The Term Selection Algorithm We propose the following basic algorithm for RF. Three variations of this algorithm are realized by choosing the terms t_i in different ways (Step 6 of the algorithm).

1. For the i^{th} request of the next document to return, repeat Steps 2-7.
2. Let R be the accumulated string of relevant passages from the last document returned.
3. Tokenize the string R into words and break it up into fixed length windows of m words after applying stopword removal and stemming.
4. Let $\texttt{tf}(t_i)$ be the term frequency of the i^{th} term in the window and $\texttt{idf}(t_i)$ the inverse document frequency of t_i.
   For each window $\boldsymbol{w} = (w_1, \ldots w_n)$, where $w_i = \texttt{tf}(t_i)^{\frac{1}{2}} \log \texttt{idf}(t_i)$, compute the cosine similarity of $\boldsymbol{w}$ with $\boldsymbol{q} = (q_1, \ldots q_n)$, where $q_i = \texttt{tf}(t_i)$.
5. Rank all windows w by similarity score and choose top p windows.
6. Extract the most frequent T terms from these windows and add them to the query. The three variants for choosing T terms are as follows:
   - RF_{const}: $T = t$, where t is a constant $\forall i$.
   - RF_{invrsl}: $T = \frac{(L_i - r_i)}{L_i} t$, where t is a constant, L_i is the length of the i^{th} document and r_i is the length of the relevant section of the i^{th} document.
   - RF_{rsl}: $T = \frac{r_i}{L_i} t$, with t, L_i and r_i defined as before.
7. Re-retrieve with the expanded query and return the topmost similar document not returned previously.

The first variant (RF_{const}) chooses a constant number of terms regardless of the segment length. For RF_{invrsl} and RF_{rsl} the number of terms added is inversely and directly proportional to the length of the relevant section, which corresponds to choosing a greater number of terms from shorter relevant segments, and choosing a smaller number of terms from shorter segments, respectively.

The default Rocchio feedback implementation serves as a baseline with parameters $(\alpha, \beta, \gamma) = (1, 0.75, 0)$ to weight original terms, positive, and negative feedback terms. The Rocchio feedback uses $T = 20$ terms for query expansion.

Two major differences between the baseline module and our implementation are: a) the baseline method adds expansion terms to the original query at each

[2] `http://www.apache.org/dyn/closer.cgi/lucene/java/`

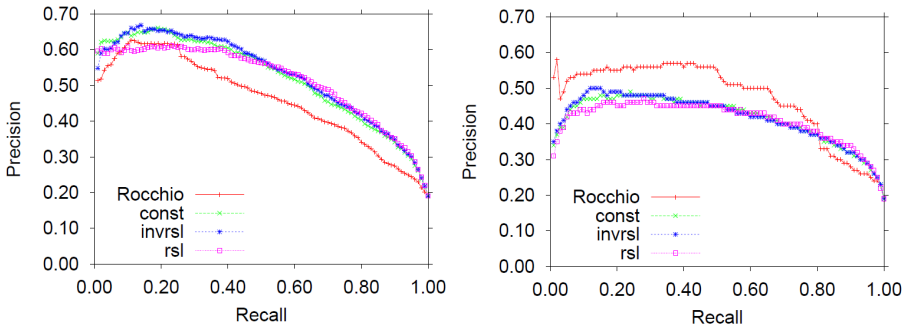

Fig. 1. Interpolated Precision-Recall graphs for INEX 2008 (left) and 2010 data (right).

iteration, whereas we add expansion terms for the i^{th} iteration obtained during the $(i-1)^{th}$ iteration; b) the step-size of the incremental feedback for the baseline method is 5, i.e. it expands the original query after every 5 iterations whereas our method uses a step-size of 1, i.e. we update the query after every iteration. Thus, our query expansion accumulates terms at every retrieval iteration in contrast to the baseline method, which generates a new query in each iteration. This is in contrast to the "save nothing" strategy [9] which was shown to be ineffective for incremental feedback.

Training the System The parameters as outlined in the feedback algorithm are the window length m, the number p of most similar windows to restrict the expansion terms to, and the variable T, which represents the number of feedback terms. After conducting a range of experiments on INEX 2008 topic set we chose the optimal settings of $(m, p, T) = (30, 10, 5)$. The results of these training experiments with the above settings are outlined in Table 1. Wilcoxon tests on the 11 point precision-recall curves reveal that the improvements for the three proposed methods over $\texttt{RF}_{\texttt{Rocchio}}$ are statistically significant.

4 Results and Analysis

For our official submission to the RF track, we used the optimal parameters obtained from INEX 2008 training topics. The graphs of Figure 1 show the document level interpolated precision-recall curves for the four approaches on INEX 2008 and 2010 topics. The graphs of Figure 2 show the interpolated geometric means of per-topic precision values measured at 11-point recall levels on 2008 and 2010 data. The left graph of Figure 1 reveals some interesting characteristics of the two feedback methods $\texttt{RF}_{\texttt{rsl}}$ and $\texttt{RF}_{\texttt{invrsl}}$. While it can be seen that $\texttt{RF}_{\texttt{rsl}}$ yields low precision for lower levels of recall, it outperforms $\texttt{RF}_{\texttt{invrsl}}$ for higher levels of recall which suggests that it might be worth trying a combination of the above two techniques as a part of our future work. The right graph of Figure 1

Table 1. Results for Relevance Feedback on INEX 2008 training topics.

Methodology	Evaluation Metric		
	MAP	GMAP	MAiP
No feedback	0.3610	0.3087	0.3952
$\texttt{RF}_{\texttt{Rocchio}}$	0.4744	0.4292	0.5011
$\texttt{RF}_{\texttt{const}}$	0.5366	0.4687	0.5519
$\texttt{RF}_{\texttt{invrsl}}$	**0.5442**	**0.4805**	**0.5611**
$\texttt{RF}_{\texttt{rsl}}$	0.5307	0.4596	0.5477

suggests that $\texttt{RF}_{\texttt{Rocchio}}$ starts off with a better precision and outperforms the focused methods until a recall level of 80% is reached. For the focused methods, although the initial retrieval precision (precision at less than 10% recall level) is lower, precision picks-up steadily and does not suffer from a steep down-hill as observed for the Rocchio method. For the RIC metric, we see a different trend for the INEX 2010 topics. The focused methods have a higher precision (thus suggesting that it is more appropriate for precision oriented retrieval tasks such as the focused task) at recall levels of less than 20% after which the Rocchio feedback outperforms each of them.

The fact that the focused RF methods are outperformed by the Rocchio feedback as measured by the standard document level retrieval metric MAP, leads to the question of what changes in the characteristics of the topics and the relevant set, if any, from 2008 to 2010, caused this trend reversal. The corresponding answers should explain the differences in the training results and the official submissions. A possibility is that the average length of relevant passages (average being computed by accumulating the number of relevant characters per document averaged over the number of topics) for the INEX 2010 topic set is higher (453.6 characters) as compared to INEX 2008 (409.9 characters), which means that further reducing the length of relevant passages may be required.

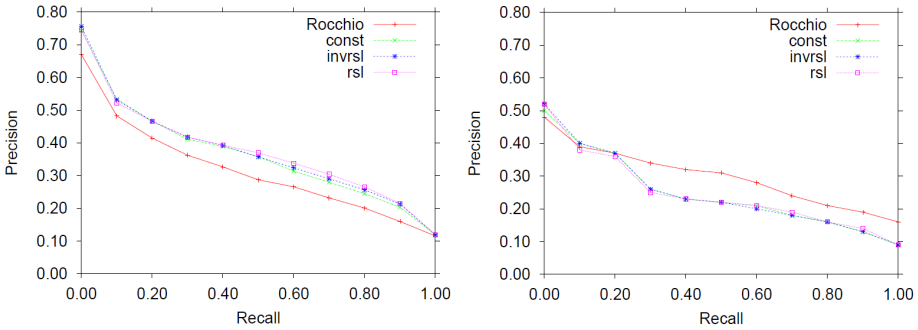

Fig. 2. RIC curves for INEX 2008 (left) and INEX 2010 data (right).

This suggests using smaller values for the number of windows, e.g. decreasing p, could possibly improve results.

5 Conclusions and Future work

For our participation in the RF track at INEX 2010, we implemented a new feedback method which selects feedback terms from maximally similar word windows extracted from reported relevant text passages.

The proposed method significantly outperforms the baseline Rocchio feedback method on the INEX 2008 training data and yields better precision at early recall levels when measured with the RIC metric, but does not show improvement for the INEX 2010 data when evaluated with MAP.

Future work includes optimizing feedback parameters m and p keeping in mind that the average length of relevant segments is higher for INEX 2010 topics. In addition based on the observation from INEX-2010 results that Rocchio gives better precision at early recall levels and our method gives better precision at higher recall levels, we plan to explore a combination of feedback strategies, i.e. selecting or switching the feedback strategies at some retrieval iteration.

Acknowledgments

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (CNGL) project.

References

1. Callan, J.P.: Passage-level evidence in document retrieval. In: SIGIR 1994, ACM/Springer (1994) 302–310
2. Allan, J.: Relevance feedback with too much data. In: SIGIR 1995, ACM Press (1995) 337–343
3. Ganguly, D., Leveling, J., Jones, G.J.F.: Exploring sentence level query expansion in the language model. In: Proceedings of ICON-2010. (2010) 18–27
4. Murdock, V.: Aspects of Sentence Retrieval. PhD thesis, University of Massachusetts - Amherst (2006)
5. Losada, D.E.: Statistical query expansion for sentence retrieval and its effects on weak and strong queries. Inf. Retr. **13** (2010) 485–506
6. Wilkinson, R.: Effective retrieval of structured documents. In: SIGIR '94, NY, USA, Springer-Verlag Inc. (1994) 311–317
7. Xu, J., Croft, W.B.: Query expansion using local and global document analysis. In: SIGIR 1996, ACM (1996) 4–11
8. Rocchio, J.J.: Relevance feedback in information retrieval. In: The SMART retrieval system – Experiments in automatic document processing. Prentice Hall, Englewood Cliffs, NJ, USA (1971) 313–323
9. Allan, J.: Incremental relevance feedback for information filtering. In: SIGIR 1996, NY, USA, ACM (1996) 270–278