

Comparing Two Blind Relevance Feedback Techniques

Daqing He, Yefei Peng
School of Information Sciences, University of Pittsburgh
Pittsburgh, PA 15260, USA
daqing@mail.sis.pitt.edu, ypeng@mail.sis.pitt.edu

Categories and Subject Descriptors

H.3.3 [Information Storage and retrieval]: Information Search and Retrieval—*Relevance Feedback*

General Terms

Algorithms, Experimentation, Measurement

Keywords

Blind Relevance Feedback, Automatic Query Expansion, Comparison

1. TWO TYPES OF BLIND RELEVANCE FEEDBACK

Query expansion based on Blind Relevance Feedback (BRF) has been demonstrated to be an effective technique for improving retrieval results. There are two types of BRF-based query expansion. BRF Type 1 (BRFT1) is the original version of BRF, where query expansion is performed on the BRF information extracted from top N documents selected from an initial search on the same collection that the target documents are in [1]. This collection is called “target collection” in this paper. BRF Type 2 (BRFT2) has been explored as an alternative to BRFT1. The query expansion is performed based on the BRF information of the top N documents selected from the initial search on a DIFFERENT collection. Such a collection is called “expansion collection” in this paper. The expanded query is then used to search on the target collection to find the relevant documents.

The effectiveness of BRF depends on two key factors: 1) the documents selected from the initial search for BRF should contain reasonable number of topically relevant documents to the query; and 2) those selected documents should share the similar genre with the target relevant documents so that there is high chance that the important content terms used in these two sets of documents are the same[2]. Both BRFT1 and BRFT2 may encounter situations that at least one of the two conditions cannot be satisfied. For example, there are not enough truly relevant documents in the target collection for many topics in Robust track of TREC evaluation, which makes it difficult to utilize BRFT1 based query expansion techniques to improve the search results.

However, with the amount of electronic resources available, it is often possible that both BRFT1 and BRFT2 can

be performed as means to improve retrieval results. However, there has been no study to examine the relationships between the two BRF approaches, from which we can answer questions like: when both BRF approaches are feasible to perform, should BRFT1 be preferred over BRFT2, or the other way round? Does it make sense to combining both BRF approaches? Will they select similar or different set of terms for expansion? Any relationship among the terms obtained from these two approaches?

In this paper, we will present our initial study about the relationship between BRFT1 and BRFT2 in the context of retrieving news articles on TREC evaluation platform. The research questions that we want to examine are:

1. Is there a performance difference between using BRFT1 and BRFT2 in retrieving news articles?
2. What are the relationships between the query expansion terms obtained from BRFT1 and that of BRFT2?
3. Does it make sense to combine BRFT1 and BRFT2?

2. EXPERIMENT SETTINGS

AQUAINT corpus was used for this study. It contains documents from three news agencies: 239,576 articles from the Associate Press (APW), 314,452 articles from New York Times (NYT), and 479,433 articles from Xinhua News Agency (XIE). For our experiment, we separated AQUAINT into a target collection, including all documents from APW and XIE, and an expansion collection, which contains all articles from NYT. Because there are relevant documents in both collections for almost all topics, and all the documents are news articles, the experiment setting can be seen as favorable for both BRFT1 and BRFT2.

The search topics were 50 HARD 2005 topics. We adopted Indri 2.0 as our retrieval system¹. We modified Indri’s build-in BRF module so that it can perform both BRFT1 and BRFT2 using the same default BRF model. The parameters for BRF were defined as selecting top 20 terms from top 20 documents. The relative weights were set as 0.9 for the original query terms and 0.1 for the expanded terms in the case of BRFT1 (i.e. BRFT1@0.9), and 0.8 for the original query terms and 0.2 for the expanded terms in the case of BRFT2 (BRFT2@0.8). These parameters were obtained by testing on those 50 HARD05 topics. We use the same set of parameters for the run that combines BRFT1 and BRFT2.

¹<http://www.lemurproject.org/indri/>

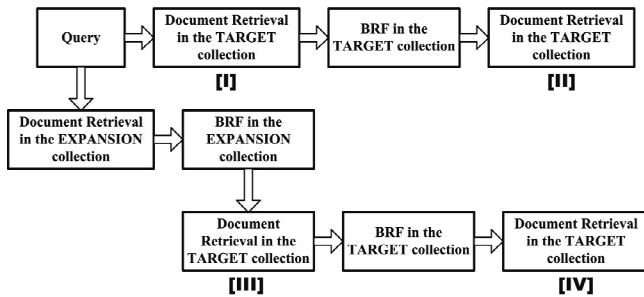


Figure 1: The stages of performing retrievals with BRFT1 and BRFT2 in our studies

3. RESULTS AND DISCUSSION

Four sets of retrieval results were collected in our studies. The first one (marked as [I] in both Figures 1 and 2) is the baseline run that employed no relevance feedback at all. The run corresponding to BRFT1 is marked as [II], and that to BRFT2 is marked as [III]. The one marked with [IV] is the run that contains query expansion based on BRFT2 first, then followed by query expansion based on BRFT1. Run [IV] is an initial exploration to the question whether it makes sense to combine BRFT1 and BRFT2.

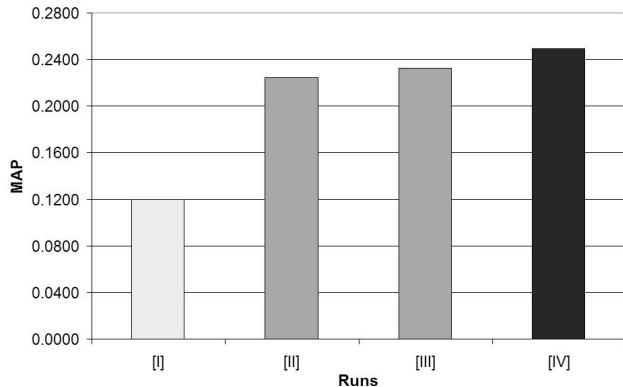


Figure 2: The Mean Average Precision (MAP) of the runs. The color difference on bars indicates significant difference between the two results.

As shown in Figure 2, both BRF approaches ([II] and [III]) achieved significant improvement over non-expansion baseline ([I]) (using the paired t Test, and both $p = 0$). However, there was no significant difference between the two BRF results. In fact, further analysis of their performance at individual topics demonstrated that these two BRF approaches achieved comparable results from individual topics down to individual returned relevant documents:

- *The two BRF runs made noticeable impacts on similar sets of individual search topics.* Among 50 search topics, the two BRF runs had made noticeable changes (increase or decrease) in total 19 topics. Here a noticeable change means that the difference between the MAP value of the baseline [I] and that of the runs [II] or [III] was greater than or equal to 0.1. 17 of the 19 topics were the same topics.

Table 1: The results of the four retrieval runs. Ret-Rel-Num means the total number of returned relevant documents across 50 topics.

Runs	runs id	MAP	Ret-Rel-Num
Baseline	[I]	0.1200	1643
BRFT1@0.9	[II]	0.2244	3284
BRFT2@0.8	[III]	0.2325	3311
Comb	[IV]	0.2493	3376

- *The two BRF runs failed to make impacts on similar sets of low performance topics.* Among the 31 topics that both BRF runs failed to make noticeable impact, 16 topics had MAP values lower than the baseline average MAP (0.12), 23 topics had MAP values lower than the average MAP value of BRF2, which is the better one among the two BRF runs.
- *Although the terms obtained for query expansion using BRFT1 ([II]) and that using BRFT2 ([III]) have only a few overlap, the two runs shared about 95% of returned relevant documents across 50 topics.* Being Consistent with Harman’s observation that different query expansion mechanisms select different expansion terms [2], the two runs averagely only had 6 terms (after stemming) in common among 20 possible terms for each topic. However, the two runs in total shared 3213 out of 3382 retrieved relevant documents across 50 topics.

Our results show some potential for combining BRFT1 and BRFT2. The run [IV], which applied BRFT1 after performing BRFT2 on the expanded collection, achieved significant improvement over the two runs using either BRFT1 or BRFT2 alone (paired t-test and both $p = 0.01$). The improvements is also reflected by the higher number of returned relevant documents in [IV] (see Table 1). However, no definite claim can be made before further experiments on 1) the difference between multiple iteration of BRFT1 or BRFT2 and the combination of these two, and 2) different combination settings.

4. CONCLUSION

By examining query expansion based on BRF using the target collection only (BRFT1) or that using the expansion collection first then the target collection (BRFT2), we demonstrated that the two BRF approaches are similar in improving retrieval effectiveness. They have succeeded and failed on the similar topics, and have retrieved similar set of relevant documents even though the terms that they picked up are different in majority. Future work includes studies on the effectiveness of combining these two approaches.

5. ACKNOWLEDGEMENT

This work has been supported in part by DARPA contract HR0011-06-2-0001.

6. REFERENCES

- [1] D. Evans and R. Lefferts. Design and evaluation of the clarit-trec-2 system. In *Proceedings of the Second Text REtrieval Conference (TREC-2)*, 1994.
- [2] D.K. Harman. Relevance feedback revisited. In *Proceedings of ACM-SIGIR 92*, 1992.