Blind Relevance Feedback with Wikipedia: Enterprise Track

Yefei Peng
Yahoo! Labs
Sunnyvale, CA USA
ypeng@yahoo-inc.com

Ming Mao
SAP Research
Palo Alto, CA USA
ming.mao@sap.com

Abstract: In this year's Enterprise track experiment, we focused on testing Blind Relevance Feedback, especially using online Wikipedia as query expansion collection. We demonstrated that using Wikipedia as query expansion collection returns better infNDCG than not using it.

1. Introduction

This year's TREC Enterprise track is using the same corpus as last year. The document collection is a crawl of the publicly available web pages from the `*.csiro.au` domain, known as the CSIRO Enterprise Research Collection (CERC or .CSIRO). But the topics are different in 2008. They are real information requests from users via telephone and email. The enquiries staffers need to search on CSIRO web to find answers for these requests (Soboroff et al. 2008). The enquiries cover a wide range of topics. The answers could be very specific or very general.

Because of the wide coverage of topics, CSIRO website may not have enough details of each of them. This led us to looking for a source of knowledge which covers a wide range of topics with as much details as possible.

Query expansion based on Blind Relevance Feedback (BRF) has been demonstrated to be an effective technique for improving retrieval results. The documents selected from initial search should contain reasonable number of relevant documents. Normally top k documents will be selected. Most frequent terms from these top documents will be extracted to form query expansion. The expanded query should be able to get more relevant documents if the selected documents share similar genre with target relevant documents.

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 (Evans & Lefferts 1994). 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 (He & Peng 2006). 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.

Wikipedia[1] is an online encyclopedia written collaboratively by its readers. There are several reasons for which we chose Wikipedia as our expansion collection. First, like other encyclopedias,

[1] http://www.wikipedia.org

it covers a wide range of topics. Second, it has a large size. As of Jan 2009, there are 2,679,000 articles in Wikipedia. Third, its content is up to date. People continuously contribute content to it.

In the remaining of this report, we will first talk about our approach in detail, then present the experiment design including the runs that we submitted.

2. Our Approach

CSIRO corpus was index with Indri to form our target collection. A snapshot of Wikipedia at July 2008 was used as query expansion collection. We only focus on English version of Wikipedia. There are totally 6,996,745 articles. All these articles are index with Indri 2.7 [2] to form our query expansion collection.

Indri 2.7 is our retrieval system. We modified Indri's 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.

In total, we submitted 4 different runs. They are:

- **TitDes**: In the Indri query, both the title and the description were used, and each word was treated as a term in the query. No Blind Relevance Feedback is used.
- **TitBrf**: Only the title is used in original query. Target collection (CSIRO) is used as expansion collection. Expansion collection (Wikipedia) is not used. This is traditional BRF method, mentioned earlier as BRF1.
- **TitExp**: Only the title is used in original query. Wikipedia collection is used as expansion collection. This method is different from traditional BRF method, mentioned earlier as BRF2.
- **TitExpBrf57**: Only the title is used in original query. Wikipedia collection is used as expansion collection. The new query is issued in target collection (CSIRO) again with Blind Relevance Feedback. BRF is used twice here.

3. Results

Results of the four runs are shown in Figure 1. As expected, TitDes has lowest performance since no blind relevance feedback is used. TitBrf has best infAP, while TitExp has best infNDCG. It shows that using target collection itself or another collection (Wikipedia) both improves performance. In this case, they give similar results.
As of last run TitExpBrf57, BRF is used twice. The first time, original query is issued on Wikipedia collection, expanded query is combined with original query with weight of 0.5 and 0.5. Then new query is issued on target collection (CSIRO), expanded query is combined with issued query (which is an expanded query) with weight 0.3 and 0.7 respectively. Then final query is issued on target collection. Similar idea could be found in (He & Peng, 2006), but their results

[2] http://www.lemurproject.org/indri/

showed similar twice BRF method returned better results than BRF1 and BRF2. Further experiments and analysis needs to be done to find out the reason.

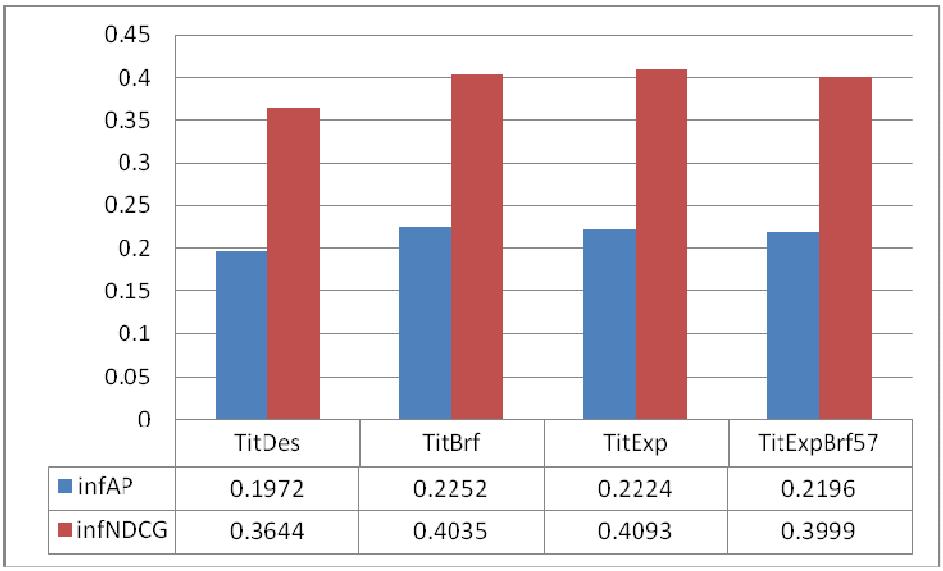

Figure 1. infAP and infNDCG of the runs.

4. Conclusion

In this paper, we are talking about the studies we conducted in the participation to TREC 2008 in the task of document search. The goal of this study is to examine the effects of different Blind Relevance Feedback methods, and using Wikipedia as query expansion collection. Our experiment results demonstrate that using target collection as expansion collection (BRF1) and using Wikipedia as expansion collection both get better results than no BRF.

5. Reference

I. Soboroff, A. de Vries (2008). Overview of the TREC 2008 Enterprise Track, in *Proceedings of TREC 2008*. Gaithersburg MD USA.

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.

Daqing He, Yefei Peng. Comparing Two Blind Relevance Feedback Techniques. In *proceedings of Annual Conference of SIGIR*, Seattle, WA USA 2006