# York University at TREC 2009: Relevance Feedback Track

Zheng Ye[1,2], Xiangji Huang[1], Ben He[1], Hongfei Lin[2]
[1]Information Retrieval and Knowledge Managment Lab, York University, Toronto, Canada
[2]Information Retrieval Lab, Dalian University of Technology, Dalian, China
{yezheng,jhuang,benhe}@yorku.ca, hflin@dlut.edu.cn

**Abstract**

We describe a series of experiments conducted in our participation in the Relevance Feedback Track. We evaluate two traditional weighting models (BM25 and DFR) for the phase 1 task, which are widely used in text retrieval domain. We also evaluate a statistics-based feedback model and our proposed feedback model for the phase 2 task. Currently, we are waiting for the overview paper to facilitate further analyses.

## Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software;

## General Terms

Measurement, Performance, Experimentation

## Keywords

Weighting Model, Relevance Feedback, DFR, BM25, Context

# 1 Introduction

In this paper, we describe the work done by members at York University in Canada and Dalian University of Technology in China for the TREC 2009 Relevance Feedback Track. In particular, we present a series of experiments conducted in Relevance Feedback Track 2009. This is the first year that we participate in this track. Our experiments mainly focus on the following aspects: (1) how the traditional retrieval models perform in identifying useful feedback documents; (2) how different relevance feedback models perform under Rocchio's relevance feedback framework [Roc71].

## 1.1 Relevance Feedback Track

In Information Retrieval (IR), Relevance feedback (RB) is a process that IR systems use the feedback information provided by users to optimize the retrieval results. Relevance Feedback has been one of the most important successes of IR research for the past decades. Feedback information can be from the real users, or from implicit evidence. Relevance feedback has been proven to be effective in both cases [BR08].

However, there has been comparatively few research advances in RF in recent years. There is no general agreement of what the best RF approach is, or what relative benefits and costs of the various approaches are [Buc08]. Relevance Feedback Track is held under this circumstance.

Last year's (2008) TREC Relevance Feedback (RF) Track just concentrated on the RF algorithm itself: Given a topic and a set of judged documents for that topic, how does a system take advantage of the judgments in order to return more documents that will be useful to the user. This year (2009), the track evaluates how well systems can find good documents to be judged, as well as the improvement due to the RF algorithm. In the first phase, each group will identify a small number of documents (e.g. 5 per run) for which they wish relevance judgments. In the second phase, the organizer would like to evaluate how well an algorithm is coupled with documents obtained in different ways, for example, documents ranked by the probability of relevance or docs which represent different aspects of relevance.

### 1.2 Collection

In this year's RF track, a new test collection, ClueWeb09, is used. It contains approximately 1,000,000,000 Web pages. This is the first real attempt to have a test collection be representative of the entire Web. For teams that do not have enough computation power, they can choose B subset of ClueWeb09. Note that the B subset is still quite large - over 3 times the size of the Terabyte GOV2 collection. More detailed information about ClueWeb09 can be found in [1].

The remainder of this paper is organized as follows. In Section 2, we describe the weighting models used in Phase 1. In Section 3, we present two feedback models used in phase 2. In Section 4, we present our official results in TREC 2009 Relevance Feedback Track. In Section 5, we conclude the paper with a look at the future work.

## 2 Weighting Models in Phase 1

There are several choices for identifying documents for the feedback algorithms. For example, we can provide feedback documents according to the following ways:

- 1. the probability of relevance of documents to a query,
- 2. documents likely to draw the line between relevant and non-relevant,
- 3. documents representing different aspects of relevance,
- 4. documents representing different interpretations of a possibly ambiguous topic statement,
- 5. documents which may not be relevant in themselves, but may offer good general background (and thus expansion terms) in the area of the topic [Buc09].

In our experiments, we provide the feedback documents according to the probability of relevance. In particular, we explore two traditional weighting models, BM25 [HBGH+96] and DFR [Ama03], which perform well on a large number of IR collection. The corresponding weighting functions are as follows:

- BM25

$$\omega = \frac{(k_1+1)*tf}{k_1*((1-b)+b*dl/avdl)+tf} * \log\frac{N-n+0.5}{n+0.5} * \frac{(k_3+1)*qtf}{k_3+qtf} \quad (1)$$

- DFR

[1] http://boston.lti.cs.cmu.edu/Data/clueweb09/

$$\begin{aligned} \omega &= TF * qtf * NORM * log_e(\frac{N+1}{n\_exp}) \\ TF &= tf * log_2(1 + c * avdl/dl) \\ NORM &= (tf + 1)/(df * (TF + 1)) \\ n\_exp &= idf * (1 - e^{-f}) \\ f &= qtf/df \end{aligned} \tag{2}$$

where $w$ is the weight of a query term, $N$ is the number of indexed documents in the collection, $n$ is the number of documents containing the term, $tf$ is within-document term frequency, $qtf$ is within-query term frequency, $dl$ is the length of the document, $avdl$ is the average document length, $nq$ is the number of query terms, the $k_i$s are tuning constants (which depend on the database and possibly on the nature of the queries and are empirically determined).

In our experiments, the values of $k_1$, $k_3$ and $b$ in the BM25 function are empirically set to be 1.2, 8 and 0.35 respectively, which has proven to perform well on a large number of collections. For the DFR weighting, its parameter $c$ is default to 7.

# 3 Our Methods for Phase 2

In this section, we first present Rocchio's Query Expansion method and a DFR-based weighting model. Then, we describe the proposed term weighting model for query expansion under Rocchio's framework.

## 3.1 Rocchio Query expansion

Rocchio's classical algorithm [Roc71] provides a general framework for implementing relevance feedback. It models a way of incorporating relevance feedback information into the vector space model. In particular, it takes a set of documents for feedback. Candidate terms in this set of documents are ranked according to the following formula:

$$Q_1 = \alpha * Q_0 + \beta * \sum_{rel} \frac{D_i}{|D_i|} - \gamma * \sum_{nonrel} \frac{D_i}{|D_i|} \tag{3}$$

where $Q_0$ and $Q_1$ represent the initial and first iteration query vectors, $D_i$ represents document weight vectors, $|D_i|$ is the corresponding Euclidian vector length, and $\alpha$, $\beta$, $\gamma$ are tuning constants.

Many other relevance feedback techniques and algorithms have been developed, mostly derived under Rocchios framework. For example, a popular and successful relevance feedback algorithm was proposed by Robertson [Rob90] while developing the Okapi system. Okapis relevance feedback algorithm is similar to Rocchios, while using a different term weighting strategy called the Robertson Selection Value (RSV) weights [Rob90]. More recently, Amati proposed a relevance feedback algorithm in his Divergence from Randomness (DFR) framework [Ama03], which similarly follows Rocchios algorithm. However, in Amatis method, term weights are assigned by a DFR term weighting model, such as the Kullback-Leibler divergence (KLD) [CdMRB01].

In our experiments, we explore two weighting schemes under Rocchio's framework, and the parameters $\alpha$, $\beta$, $\gamma$ are empirically set to be 1, 0.4 and 0.15 respectively. In addition, the number of expansion terms, $exp\_term$, is empirically set to be 35. In the following subsection, we describe the algorithms in detail.

## 3.2 Bose-Einstein distribution Weighting Scheme

The first term weighting model used in our experiments is DFR-based weighting model described in [Ama03]. The basic idea of these term weighting models for query expansion is to measure the divergence of a term's distribution in a pseudo relevance set from its distribution in

the whole collection. The higher this divergence is, the more likely the term is related to the query topic.

We use the Kullback-Leibler (KL) divergence model in this set of experiments. Using the KL model, the weight of a term $t$ in the *exp_doc* top-ranked documents is given by:

$$w(t) = P(t|D) \log_2 \frac{P(t|D)}{P(t|C)} \tag{4}$$

where $P(t|D) = \frac{c(t,D)}{c(D)}$ is the generation probability of term $t$ from $D$, the set of feedback documents. $c(t, D)$ is the frequency of $t$ in $D$, and $c(D)$ is the count of words in $D$. $P(t|C) = \frac{c(t,C)}{c(C)}$ is the collection model. $c(t, C)$ is the frequency of $t$ in collection $C$, and $c(C)$ is the count of words in the whole collection $C$. *exp_doc* usually ranges from 3 to 10 [Ama03]. Another parameter involved in the query expansion mechanism is *exp_term*, the number of terms extracted from the *exp_doc* top-ranked documents. *exp_term* is usually larger than *exp_doc* [Ama03].

### 3.3 A Context Sensitive Weighting Scheme

In traditional QE weighting models, the expansion terms are selected only by their statistics in the top $k$ documents and the whole collection. In the process of the selection of expansion terms, the context informations are always ignored, for example, the domain of users' interest, knowledge about the query's subject. Zhai et al. [BNCB07] studied using query-specific contexts to boost IR performance. It showed that context factors can bring significant performance improvements in terms of MAP. In this paper, we propose a context sensitive weighting to select the expansion terms. In particular, the candidate terms are ranked according to the following formula:

$$\begin{aligned} P(t|C) &\propto P(t)P(C|t) = P(t)P(c_1, \ldots, c_m|t) \\ &= P(t) \prod_{i=1}^{m} P(c_i|t) \end{aligned} \tag{5}$$

where $C$ represents the context for a query and it consists of a number of feature contexts. A feature context $c_i$ represents a certain kind of context, such as click information and users' background.

The probability $P(t)$ can be interpreted as the prior probability. It means that how likely it is that candidate term $t$ can be selected as an expansion term without taking into account any context information. The probability $P(c|t)$ can be interpreted as: given the expansion candidate term $t$, how likely it is that the feature context $c_i$ will be observed. This probability is estimated according to the type of context. In this paper, we define a co-occurrence feature context, which means the probability that a candidate expansion co-occurs with the query. We only explore the co-occurrence feature context in this RF track.

#### Co-occurrence with the query terms

J. Xu et al. [XC00] proposed an PRF approach, called "local context analysis", in which it is suggested that useful expansion terms tend to co-occur with the original query. In this paper, we define a co-occurrence context, and the corresponding probability $P(c_i|t)$ can be interpreted as, given a candidate term $t$, how likely it is that $t$ will co-occur with the original query. In this paper, we propose to use the term weighting function in [XC00] to estimate the probability $P(c_i|t)$, which is shown as follows:

$$P(c|t_i) \propto g(t_i, Q) = \prod_{w_i in Q} (\sigma + co\_degress(t_i, w_i)) \tag{6}$$

$$co\_degress(t_i, w_i) = \log_{10}( \sum_{d \ in \ S} tf(t_i, d) tf(w_i, d)) idf(t_i) / log_{10}(n) \tag{7}$$

where $S$ is the set of documents for PRF, $Q$ is the original query. $\sigma$ is a smoothing parameter, and we empirically set it to be 0.001 in our experiments.

# 4 Experiments

We preprocess the collection by removing all the HTML tags. Words in the collection are segmented by spaces and punctuations. Porter stemming and stopword removal are conducted in both indexing and searching processes. Beside these simple procedures, no further technologies have been used. In the following, we present our official experimental results.

Table 1 shows our official runs for phase 1 task. The values in parentheses are the counts of worse or better when the run is used as input for RF for each evaluation measure. The final score is the ratio of $better/(better + worse)$. For phase 1 runs, we did not lowercase the words for indexing, which is a kind of mistake. So the results do not reflect the real performance of BM25 and DFR.

Table 1: Phase 1 results

| Run | emap | mapA | P10A | stAP | score |
|---|---|---|---|---|---|
| BM25 (YUIR.1) | (4, 9) | (14, 0) | (14, 0) | (5, 8) | 0.3148 |
| DFR (YUIR.2) | (18, 13) | (0, 0) | (22, 9) | (22, 9) | 0.3548 |

Table 2 shows our official runs for phase 2 task. Three runs marked by superscript "c" are obtained by using our proposed weighting model described in Section 3.3. Since these three runs are obtained based on the un-lowercase index, they also do not reflect the real effectiveness of the proposed context-based feedback approach. In Table 2, we provide the corrected results in terms of "stapMAP". For the other runs, the feedback weighting function used is the Bose-Einstein distribution weighting scheme under Rocchio's framework. For the base run "YUIR.base", we use the top 5 documents and top 35 terms to conduct PRF.

From Table 2, in general, the performance of feedback based on the judged documents is significantly better than that based on pseudo relevance documents. Although feedback from users requires additional efforts, it brings great benefits for improving the retrieval performance. For the relevance feedback in phase 2, the performance is not determined by the results in phase 1 in our experiments. From Table 2, we do not see any correlation between the performance in phase 2 and the performance in phase 1, which is different from that in PRF. This indicates that the judged irrelevant documents (top ranked in phase 1) are beneficial to feedback. Actually, we also conduct experiments of relevance feedback based solely on the relevant documents, the performance of which is not as good as that based on all the judged documents.

Table 2: Phase 2 results

| Run | MAP | emap | stAP (corrected) |
|---|---|---|---|
| YUIR.base | | | 0.2113 |
| YUIR.CMIC.1$^c$ | 0.0780 | 0.0367 | 0.1546 (0.2585) |
| YUIR.UCSC.2$^c$ | 0.0650 | 0.0392 | 0.1586 (0.2540) |
| YUIR.YUIR.2$^c$ | 0.0301 | 0.0322 | 0.1386 (0.2103) |
| YUIR.FDU.1 | 0.0258 | 0.0511 | 0.2471 |
| YUIR.ugTr.1 | 0.0481 | 0.0523 | 0.2460 |
| YUIR.UMas.2 | 0.0426 | 0.0536 | 0.2638 |
| YUIR.YUIR.1 | 0.0320 | 0.0474 | 0.2403 |

# 5 Conclusions

In this paper, we present our participation in Relevance Feedback Track 2009. First, we evaluate two traditional weighting models (BM25 and DFR) for phase 1 task, which are widely

used in text retrieval domain. Second, we evaluate a statistical-based weighting model and our proposed weighting model for phase 2 task.

In future work, we will work on the following two directions. First, we plan to explore different strategies for identifying documents for relevance feedback. Second, we plan to incorporate more feature contexts into our proposed weighting model.

# 6 Acknowledgements

This research is jointly supported by NSERC of Canada, the Early Researcher/Premier's Research Excellence Award, Natural Science Foundation of China (No. 60373095, 60673039 and 60973068), the National High Tech Research and Development Plan of China (2006AA01Z151) and Doctoral Fund of Ministry of Education of China (No.20090041110002).

# References

[Ama03] G. Amati. Probabilistic models for information retrieval based on divergence from randomness. *PhD thesis, Department of Computing Science, University of Glasgow*, 2003.

[BNCB07] Jing Bai, Jian-Yun Nie, Guihong Cao, and Hugues Bouchard. Using query contexts in information retrieval. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 15–22, New York, NY, USA, 2007. ACM.

[BR08] Chris Buckley and Stephen Robertson. Relevance feedback track overview: Trec 2008. In *Proceedings of the 17th Text Retrieval Conference TREC 2008*, 2008.

[Buc08] Chris Buckley. Proposal for a trec 2008 relevance feedback track. In *Proceedings of the 17th Text Retrieval Conference TREC 2008*, 2008.

[Buc09] Chris Buckley. Trec 2009 relevance feedback guidelines. In *Proceedings of the 18th Text Retrieval Conference TREC 2009*, 2009.

[CdMRB01] Claudio Carpineto, Renato de Mori, Giovanni Romano, and Brigitte Bigi. An information-theoretic approach to automatic query expansion. *ACM Trans. Inf. Syst.*, 19(1):1–27, 2001.

[HBGH+96] Micheline Hancock-Beaulieu, Mike Gatford, Xiangji Huang, Stephen E. Robertson, Steve Walker, and P. W. Williams. Okapi at trec-5. In *Text REtrieval Conference (TREC) TREC-5 Proceedings*, 1996.

[Rob90] S. E. Robertson. On term selection for query expansion. *J. Doc.*, 46(4):359–364, 1990.

[Roc71] J. Rocchio. *Relevance Feedback in Information Retrieval*, pages 313–323. 1971.

[XC00] Jinxi Xu and W. Bruce Croft. Improving the effectiveness of information retrieval with local context analysis. *ACM Trans. Inf. Syst.*, 18(1):79–112, 2000.