A Comparative Study of Pseudo Relevance Feedback for Ad-hoc Retrieval

Kai Hui[1], Ben He[1], Tiejian Luo[1], and Bin Wang[2]

[1] Graduate University of Chinese Academy of Sciences, Beijing 100190, China
huikai10@mails.gucas.ac.cn, {benhe,tjluo}@gucas.ac.cn
[2] Institute of Computational Technology, Beijing 100190, China
wangbin@ict.ac.cn

Abstract. This paper presents an initial investigation in the relative effectiveness of different popular pseudo relevance feedback (PRF) methods. The retrieval performance of relevance model, and two KL-divergence-based divergence from randomness (DFR) feedback methods generalized from Rocchio's algorithm, are compared by extensive experiments on standard TREC test collections. Results show that a KL-divergence based DFR method (denoted as *KL1*), combined with the classical Rocchio's algorithm, has the best retrieval effectiveness out of the three methods studied in this paper.

Keywords: Pseudo relevance feedback, Rocchio's algorithm, Divergence from randomness.

1 Introduction

Many PRF algorithms and methods have been proposed in the literature of information retrieval (IR). For example, *RM3* [4] derived from relevance model [3] improves the KL-divergence language model with Dirichlet smoothing (DirKL) [8], and the KL-based DFR feedback (*KL2*) [1,2] improves over the PL2 model [1]. Also based on the KL-divergence, an improved version of Rocchio's algorithm (*KL1*[1]) [7] is applied to enhance retrieval performance over BM25 [5].

Despite the effectiveness of PRF in improving the ad-hoc retrieval effectiveness, there exists a need for further understanding in the relative strength and weakness of different PRF methods [4]. Among the rare previous work, Lv & Zhai compare the effectiveness of relevance model to the model-based feedback [4]. This paper conducts a comparative study on the effectiveness of various popular PRF methods. While the work in [4] focuses on the PRF methods derived based on language model, this work compares the retrieval performance of RM3, KL1 and KL2, which have been previously applied on top of the DirKL, BM25, and PL2 weighting models, respectively. Note that the model-based feedback is not studied in this paper since its performance is comparable to RM3 [4].

[1] The PRF method applied in [7] is denoted as KL1 since it can be seen as a Type I model of the DFR feedback in [1].

G. Amati and F. Crestani (Eds.): ICTIR 2011, LNCS 6931, pp. 318–322, 2011.

2 Related PRF Methods

In this section, we introduce the three PRF methods involved in this study. The algorithms of these PRF methods follow similar steps as described below, where the difference among them is explained:

1. There are two parameters in the PRF methods, namely $|ED|$, the feedback document set size, and $|ET|$, the number of expansion terms. The top-ranked documents in the first-pass retrieval form a feedback document set *ED*.
2. Each candidate term in *ED* is assigned an expansion weight. Different PRF algorithms apply their own weighting methods as follows.

RM3 estimates a feedback model $P(t|ED)$ for a candidate term t as follows:

$$P(t|ED) \propto \sum_{d \in ED} P(t|d)P(d) \prod_{q \in Q} P(q|d) \tag{1}$$

where $\prod_{q \in Q} P(q|d)$ is proportional to the relevance weight $w(Q, d)$, which indicates the relative importance of d in *ED*. $P(t|d)$ is the probability of generating t from the smoothed language model of document d. Moreover, for each d in *ED*, its relevance weight is aggregated by the $w(Q, d)$ of the top-ranked document to normalize the gap in the relevance weights among different feedback documents[2].

KL1 weighs a candidate term t by the KL-divergence of the term's distribution in each feedback document from its distribution in the whole collection:

$$w(t, ED) = \sum_{d \in ED} \frac{P(t|d) \log_2 \frac{P(t|d)}{P(t|C)}}{|ED|} \cdot w(Q, d) \tag{2}$$

where $P(t|d) \log_2 \frac{P(t|d)}{P(t|C)}$ is *0* if t is unseen in d. The relevance weight $w(Q, d)$ works as a quality-biased factor to balance between feedback documents with different importance [7].

KL2, similar to KL1, also uses the KL-divergence measure, but at a larger granularity by considering a candidate term's distribution in the entire feedback document set:

$$w(t, ED) = P(t|ED) \log_2 \frac{P(t|ED)}{P(t|C)} \tag{3}$$

3. Finally, the *ET* most weighted candidate terms, called expansion terms, are added to the original query.

RM3 uses an interpolation of the feedback model with the original query model with a free parameter α:

$$\theta_{Q'} = (1 - \alpha) * \theta_Q + \alpha * \theta_F \tag{4}$$

[2] The interpretation of RM3 follows the implementation in the Lemur toolkit.

Table 1. Information about the test collections used

Coll.	TREC Task	Topics	# Docs
disk1&2	1, 2, 3 ad-hoc	51-200	741,856
disk4&5	Robust 2004	301-450, 601-700	528,155
GOV2	2004-2006 Terabyte Ad-hoc	701-850	25,178,548

where θ_Q, θ_F and $\theta_{Q'}$ are the query model, feedback document model, and the modified query model, respectively.

Using both **KL1** and **KL2**, the vector of query terms weight is modified by taking a linear combination of the initial query term weights with the expansion weight $w(t, ED)$ as follows:

$$Q_1 = \alpha_1 * Q_0 + \beta_1 * \sum_{r \in ED} \frac{r}{R} \tag{5}$$

where Q_0 and Q_1 represent the original and first iteration query vectors, r is the expansion term weight vector, and R is the maximum $w(t, ED)$ of all candidate terms. α_1 and β_1 are tuning constants controlling how much we rely on the original query and the feedback information. In this paper, we fix α_1 to 1 to reduce the number of parameters that require tuning.

3 Experiments

We experiment on title-only ad-hoc topics over 3 standard TREC test collections as shown in Table 1. Porter's English stemmer, and standard English stopword removal are applied. On each collection, each of the baseline models and the corresponding PRF method are evaluated by a two-fold cross-validation, where the test topics associated to each collection are split into two equal-sized subsets by parity. Each pair of baseline model and PRF method has several free parameters that require tuning on the training topics. In our experiments, we first tune the length normalization/smoothing parameter using Simulated Annealing, and then, scan a wide range of values for the parameters $|ED|$, the feedback set size, and $|ET|$, the number of expansion terms, namely $2 < |ED| < 50$ and $10 < |ET| < 100$. Finally, the linear combination parameter that merges the expansion terms with the original query is tuned by Simulated Annealing.

The experimental results are summarized in Table 2. To examine the effectiveness of PRF with different evaluation purposes, the results are reported in three evaluation measures respectively: mean average precision (MAP), precision at 10 (Pre@10), and normalized discounted cumulative gain (nDCG). The best results obtained by the baseline models and the PRF methods are in bold. The improvement over the corresponding baseline model in percentage is also given in the table. Moreover, a $*$ or † indicates a statistically significant difference over DirKL+RM3 or PL2+KL according to the Wilcoxon matched-pairs signed-ranks test at 0.05 level.

Table 2. Experimental Results

Coll.	DirKL	PL2	BM25	DirKL+RM3	PL2+KL2	BM25+KL1
Results in MAP						
disk1&2	0.2351	0.2336	**0.2404**	0.2744, 16.72%	0.2814, 20.46%	**0.3036∗†, 26.29%**
disk4&5	0.2565	**0.2570**	0.2535	0.2832, 10.41%	0.2886, 12.30%	**0.2950∗, 16.37%**
GOV2	0.3028	**0.3042**	0.2997	0.3352, 10.70%	0.3227, 6.08%	**0.3434†, 14.58%**
Results in Pre@10						
disk1&2	0.4967	0.4986	**0.5106**	0.5266, 6.02%	0.5373, 7.78%	**0.5626∗†, 10.18%**
disk4&5	0.4400	**0.4420**	0.4405	0.4404, ≈ 0	0.4477, 1.29%	**0.4557, 3.45%**
GOV2	0.5617	0.5657	**0.5810**	0.5980, 6.46%	0.5758, 1.78%	**0.6053, 4.18%**
Results in nDCG						
disk1&2	0.4990	0.4978	**0.5018**	0.5390, 8.02%	0.5434, 9.16%	**0.5688, 13.35%**
disk4&5	0.5297	**0.5320**	0.5303	0.5592, 5.57%	0.5668, 6.54%	**0.5776, 8.92%**
GOV2	0.5924	**0.5960**	0.5876	**0.6110, 3.14%**	0.6036, 1.28%	0.6076, 3.40%

According to the results, the baseline models have in general comparable retrieval performance on all three test collections. As for the effectiveness of PRF, apart from on GOV2 in nDCG, BM25+KL1 provides the best retrieval performance on all three test collections used. The three PRF methods have shown comparable retrieval performance, although BM25+KL1 can lead to statistically significant better effectiveness on disk1&2 and disk4&5. Overall, out of the three PRF methods used, KL1, a DFR feedback method derived from Rocchio's algorithm, provides the best effectiveness on the datasets used. A possible explanation is that KL1 evaluates the importance of the candidate expansion terms in individual feedback documents separately. In this case, the effectiveness of KL1 has a less chance of being affected by poor feedback documents than KL2, while the latter could risk contaminating the feedback documents by considering the high-quality and poor feedback documents as a single sample from the collection.

4 Conclusions and Future Work

This paper has conducted a large-scale comparative study on the effectiveness of three popular PRF methods on standard TREC test collections. As shown by the experiments, KL1, a variant of the DFR feedback derived from the classical Rocchio's algorithm, has the best retrieval effectiveness on the datasets used. In the future, we plan to extend this study by including more recently proposed PRF methods, and by experimenting on larger test collections such as the ClueWeb dataset.

Acknowledgements. This work is supported in part by the President Fund of GUCAS.

References

1. Amati, G.: Probabilistic models for information retrieval based on divergence from randomness. PhD thesis, DCS, Univ. of Glasgow (2003)
2. Carpineto, C., de Mori, R., Romano, G., Bigi, B.: An information-theoretic approach to automatic query expansion. ACM Trans. Inf. Syst. 19(1), 1–27 (2001)
3. Lavrenko, V., Croft, W.B.: Relevance-Based Language Models. In: SIGIR, pp. 120–127 (2001)
4. Lv, Y., Zhai, C.: A comparative study of methods for estimating query language models with pseudo feedback. In: CIKM, pp. 1895–1898 (2009)
5. Robertson, S.E., Walker, S., Hancock-Beaulieu, M., Gatford, M., Payne, A.: Okapi at TREC-4. In: TREC (1995)
6. Rocchio, J.: Relevance feedback in information retrieval. In: The SMART Retrieval System. Prentice-Hall, Englewood Cliffs (1971)
7. Ye, Z., He, B., Huang, X., Lin, H.: Revisiting Rocchio's Relevance Feedback Algorithm for Probabilistic Models. In: Cheng, P.-J., Kan, M.-Y., Lam, W., Nakov, P. (eds.) AIRS 2010. LNCS, vol. 6458, pp. 151–161. Springer, Heidelberg (2010)
8. Zhai, C., Lafferty, J.D.: Model-based feedback in the language modeling approach to information retrieval. In: CIKM, pp. 403–410 (2001)