

# Promoting Divergent Terms in the Estimation of Relevance Models

Javier Parapar and Álvaro Barreiro

IRLab, Computer Science Department  
University of A Coruña, Spain  
{javierparapar,barreiro}@udc.es

**Abstract.** Traditionally the use of pseudo relevance feedback (PRF) techniques for query expansion has been demonstrated very effective. Particularly the use of Relevance Models (RM) in the context of the Language Modelling framework has been established as a high-performance approach to beat. In this paper we present an alternative estimation for the RM promoting terms that being present in the relevance set are also distant from the language model of the collection. We compared this approach with RM3 and with an adaptation to the Language Modelling framework of the Rocchio’s KLD-based term ranking function. The evaluation showed that this alternative estimation of RM reports consistently better results than RM3, showing in average to be the most stable across collections in terms of robustness.

## 1 Introduction and Motivation

In the history of the Information Retrieval research, efforts to improve retrieval effectiveness have been centred in both developing better retrieval models by including new features or using different theoretical frameworks; and in designing new techniques to be incorporated on top of existing models to improve their performance. Particularly on the later, Query Expansion (QE) has proven to be effective from very early research stages. QE approaches can be classified between global techniques which produce a query rewriting without considering the original rank produced by the query, and local techniques in which the expanded query is generated using the information of the initial retrieval list.

In [19] Salton presented the initial efforts on exploiting the local information to improve the query formulation introducing, among others, Rocchio approach [16] working on the Vector Space Model framework. This family of local techniques is called Relevance Feedback (RF) [17] and it is based on using the relevant documents in the initial retrieval set in order to reformulate the query based on their content. Nevertheless, in a real retrieval scenario it is not realistic to assume that relevance judgements are available. Because of this, Pseudo Relevance Feedback (PRF) algorithms have been investigated [6, 21]. PRF methods are based on assuming relevance of a set of documents retrieved by the original query. The set of documents which are assumed to be relevant and the way in

which their information is exploited to improve the original query varies from one PRF method to another.

Lately, a PRF technique has been presented in the Language Modelling framework and proven very successful to improve retrieval effectiveness. This approach, called Relevance Models (RM) [10], has been established as a high-performance PRF approach showing great improvements over the results obtained with the initial ranking. Since it was originally presented in [10] it has been used in combination with other approaches such as the employment of query variants [5], cluster based retrieval [11], passage retrieval [12] or sentence retrieval [3]. Originally, Lavrenko and Croft presented [10] two different estimations of a relevance model: RM1 and RM2.

Despite the success of the RM, it was only recently when Lv and Zhai [14] tackled the necessity of comparing different estimations for the RMs. In [14] they compared five methods to estimate the query language models: RM3 and RM4 [1]; a divergence minimization model (DMM) and a simple mixture model (SMM) [23]; and a regularized mixture model (RMM) [20]. The main finding of this paper was that, in general, RM3 is the best and most stable method among the others. RM3 and RM4 [1] are extensions of the originally formulated RM1 and RM2 approximations, respectively. These extensions linearly interpolate the original query with the terms selected for expansion using RM1 or RM2.

The contributions of our paper are two PRF techniques that promote divergent terms and their comparison with RM3. Back in 2001 Carpineto et al. [4] presented a discriminational model to score candidate expansion terms in the Rocchio's framework based on the Kullback-Leibler Distance (KLD). This method improved the results of the standard Rocchio method. In this work we adapt this approach to work under the Language Modelling framework, improving also the performance of the original method by interpolating the selected expansion terms with the original query as in RM3. In our second contribution we present a new RM estimation that promotes divergent terms for expansion, i.e., terms that are far from the collection language model. We adopted the evaluation methodology from [14] and the results showed that the new estimated relevance model performs better than RM3 and that its behaviour, in terms of robustness across collections, is more stable than the other methods.

The rest of the paper is as follows. Section 2 presents the background. Section 3 explains the proposed methods for PRF with promotion of divergence. In Section 4 the evaluation and its results are reported. Section 5 describes the related work and, finally, conclusions and future work are reported in Section 6.

## 2 Background

In this section we will introduce the theoretical basis for this work: the retrieval method for the initial ranking, the different formulations for the RM and the KLD based discriminational model presented in [4].

## 2.1 Language Modelling for the Initial Ranking

The RM for PRF was presented within the Language Modelling (LM) theoretical framework. In Language Modelling the probability of a document given a query,  $P(d|q)$ , is estimated using the Bayes' rule as presented in Eq. 1.

$$P(d|q) = \frac{P(q|d) \cdot P(d)}{P(q)} \stackrel{rank}{=} \log P(q|d) + \log P(d) \quad (1)$$

In practice  $P(q)$  is dropped for document ranking purposes. The prior  $P(d)$  encodes a-priori information on documents and the query likelihood,  $P(q|d)$ , incorporates some form of smoothing. In this paper we consider uniform priors and uni-gram language models with Dirichlet smoothing [24], see Eq. 2.

$$P(q|d) = \prod_{i=1}^n P(q_i|d) = \prod_{i=1}^n \frac{tf(q_i, d) + \mu \cdot P(q_i|C)}{NT_d + \mu} \quad (2)$$

where  $n$  is the number of query terms,  $tf(q_i, d)$  is the raw term frequency of  $q_i$  in  $d$ ,  $NT_d$  is the document length expressed in number of terms, and  $\mu$  is a parameter for adjusting the amount of smoothing applied.  $P(q_i|C)$  is the probability of the term  $q_i$  occurring in the collection  $C$  that is usually obtained with the maximum likelihood estimator computed using the collection of documents.

After obtaining the initial ranking using the original query, the PRF methods assume relevance over a subset of retrieved documents. This set is usually called *relevance set*. The information of those documents is then used to improve the initial retrieval. The most common way of achieving this objective is expanding the original query and producing a second retrieval with the reformulated query. Next, different models to produce expanded queries are analysed.

## 2.2 Relevance Models

The RM approach builds better query models using the information given by the pseudo relevant documents. Two estimations were originally presented in [10]. RM1 assumes that the words in the relevant documents and the query words are sampled identically and independently from the relevance model. The result is an estimation where the query likelihood for every document is used as the weight for the document and the probability of a word is averaged over every document language model. In contrast, RM2 assumes that the query words are independent of each other, but they are dependent of the words of the relevant documents (conditional sampling). The result is that relevant documents containing query words can be used for computing the association of the their words with the query terms. A quite detailed explanation of the RM for PRF is given in the Chapter 7 of the book [7] by Croft et al.

In RM the original query is considered a very short sample of words obtained from the relevance model ( $R$ ). If more words from  $R$  are desired then it is reasonable to choose those words with highest estimated probability when considering the words for the distribution already seen. So the terms in the lexicon of the

collection are sorted according to that estimated probability, which after doing the assumptions using the RM1 method, is estimated as in Eq. 3.

$$P(w|R) \propto \sum_{d \in C} P(d) \cdot P(w|d) \cdot \prod_{i=1}^n P(q_i|d) \quad (3)$$

Usually  $P(d)$  is assumed to be uniform.  $\prod_{i=1}^n P(q_i|d)$  is the query likelihood given the document model, which is traditionally computed using Dirichlet smoothing (see Eq. 2). Then for assigning a probability to the terms in the relevance model we have to estimate  $P(w|d)$ ; in order to do so it is also common to use Dirichlet smoothing. The final retrieval is obtained by four steps:

1. Initially the documents in the collection  $C$  are ranked using their query likelihood. This query likelihood is usually estimated with some kind of smoothing, commonly Dirichlet smoothing as in Eq. 2.
2. A certain top  $r$  documents from the initial retrieval are taken for the estimation instead of the whole collection  $C$ , let us call this pseudo relevance set  $RS$ .
3. The relevance model probabilities  $P(w|R)$  are calculated using the estimate presented in Eq. 3, with  $RS$  instead of  $C$ .
4. To build the expanded query the  $e$  terms with highest estimated  $P(w|R)$  are selected. The expanded query is used to produce a second document ranking using negative cross entropy as in Eq. 4. In this second retrieval Dirichlet smoothing is commonly used.

$$\sum_{i=1}^e P(w_i|R) \cdot \log P(w_i|d) \quad (4)$$

RM3 is a later extension of RM that performs better than RM1 in terms of effectiveness. RM3 interpolates the terms selected by RM1 with the original query as in Eq. 5 instead of using them directly. The final query is used in the same way as in RM1 to produce a second ranking using negative cross entropy.

$$P(w|q') = (1 - \lambda) \cdot P(w|q) + \lambda \cdot P(w|R) \quad (5)$$

### 2.3 Kullback Leibler Divergence for Pseudo Relevance Feedback

In [4] Carpineto et al. presented a method for term scoring in the context of Rocchio’s framework for PRF. Carpineto et al. tried to maximize the divergence between the probability distributions of the terms estimated in the pseudo relevance set ( $p_{RS}$ ) and the distribution estimated over the whole collection ( $p_C$ ). In order to do so they used the KLD calculated as in Eq 6 because it captures the relative entropy between both distributions. To build the expanded query they selected the terms that mostly contribute to the divergence of both distributions (higher KLD score). In that work they compared the KLD term ranking function with Rocchio’s weights, Robertson’s Selection Value [15], Chi-squared

and Doszko’s variant of Chi-squared [8]. The results showed that the presented KLD term scoring function performed the best.

$$KLD(p_{RS}, p_C) = \sum_{w \in V} p_{RS}(w) \cdot \log \frac{p_{RS}(w)}{p_C(w)} \quad (6)$$

### 3 Promoting the Divergence in Pseudo Relevance Feedback

In this section we describe the two approaches presented under the Language Modelling framework to promote divergence in the PRF context.

#### 3.1 Kullback Leibler Divergence Based Query Expansion in the Language Modelling Framework

Although the KLD method outperformed the other term ranking methods in the Rocchio’s framework, it was not compared with RM in [4]. In our paper we compare the KLD method against the standard RM3 formulation adapting the KLD scoring from the Rocchio’s framework to work under the Language Modelling framework. The KLD scoring function was computed as in Eq. 7

$$kld_{score}(w) = p_{RS}(w) \cdot \log \frac{p_{RS}(w)}{p_C(w)} \approx \frac{tf(tw, RS)}{NT_{RS}} \cdot \log \frac{tf(w, RS) \cdot NT_C}{NT_{RS} \cdot tf(w, C)} \quad (7)$$

where  $tf(w, RS)$  is the term frequency of  $w$  in the pseudo relevance set,  $NT_{RS}$  is the number of terms in the pseudo relevance set  $RS$ ,  $NT_C$  is the total number of terms in the collection and  $tf(w, C)$  is the term frequency of  $w$  in the whole collection.

To obtain a probability for each of the  $e$  terms selected for expansion we re-normalized the scores obtained with Eq. 7 as in Eq. 8

$$KLD(w) = \frac{kld_{score}(w)}{\sum_{i=1}^e kld_{score}(w_i)} \quad (8)$$

In RM3 it was already demonstrated that the interpolation of the original query and the expanded query performs better. So we incorporated this idea in the KLD-based model interpolating the  $e$  terms selected as result of the KLD scoring formula with the original query. Therefore, the second retrieval is processed with an extended query as presented in Eq. in 9:

$$P(w|q') = (1 - \lambda) \cdot P(w|q) + \lambda \cdot KLD(w) \quad (9)$$

#### 3.2 Relevance Models with Promotion of Divergent Terms

The KLD-based introduction of divergence in the Language Modelling framework presented above was made as a plug-in in the Language Modelling framework. According to the analysis presented in [14], the advantage in terms of

stability of RM3 was attributable to the use of the query likelihood scores in the estimation made by RM1, which is not present in the KLD approach. To take advantage of this, we present a new estimation that promotes divergent terms maintaining the benefits from the RM methods, i.e., the use of the query likelihood scores. This new estimation arises naturally when the objective is to select expansion terms that, having high estimated probability in the RS, diverge from the collection distribution, i. e. they are more discriminative terms.

Based on the original RM1 estimation presented in Eq. 3 the most straightforward way of introducing such idea is by replacing the  $P(w|d)$  by  $P(w|d) - P(w|C)$ . In this way those terms whose density is higher in RS than in the collection are promoted, meanwhile those with low density in the RS are demoted. Another important point in order to reinforce the promotion of divergent terms is how  $P(w|d)$  is smoothed. Usually in RM this is done using Dirichlet smoothing choosing as background distribution the collection distribution. In the presented method we decided to apply the smoothing but instead of using the collection distribution as background distribution we chose to use the distribution in the relevance set. Therefore, the objective is to get for expansion the best terms that describe the documents taking into account both the RS and the divergence from the collection distribution. The computation was performed as in Eq. 10.

$$P(w|d) - P(w|C) \propto \frac{tf(w, d) + \frac{\mu \cdot tf(w, RS)}{NT_{RS}}}{NT_d + \mu} - \frac{tf(w, C)}{NT_C} \quad (10)$$

Note that  $P(w|d) - P(w|C)$  could provide negative scores for those terms with less estimated probability in the documents of the relevant set than in the whole collection. To avoid this a re-normalization of such subtraction is done, let us call the re-normalized term  $P_{C-}(w|d)$ . With these considerations the final estimation is computed as in Eq 10.

$$P(w|R) \propto \sum_{d \in RS} P(d) \cdot P_{C-}(w|d) \cdot \prod_{i=1}^n P(q_i|d) \quad (11)$$

After this, the second retrieval was performed as in RM3 (interpolating with the original query) as indicated in Section 2.2.

Another way of introducing the divergence idea would be the use of a document prior to promote documents that are far away from the collections' distribution, acting at document level rather than at term level. Nevertheless no improvements were achieved with our experiments applying that approach.

Now we have to remark an important point that, to the best of our knowledge, was never discussed properly in the context of RM: the different roles of smoothing the parameters in the distinct steps of the process. In RM3 smoothing is applied up to four times (see Section 2.2), and Dirichlet is commonly used in every occasion, so we can distinguish:

1.  $\mu_1$ , the smoothing parameter in the initial retrieval (Eq 2, step 1).
2.  $\mu_2$ , the smoothing parameter in  $P(w|d)$  (Eq 3, step 3).
3.  $\mu_3$ , the smoothing parameter in  $\prod_{i=1}^n P(q_i|d)$  (Eq 3, step 3).

4.  $\mu_4$ , the smoothing parameter in the second retrieval (Eq 4, step 4).

Usually in the literature all the four parameters are considered to be only one and the parameter is even not trained taking default values as for example in [14] ( $\mu = 1000$ ). Although this may produce good values, being a very good property of the method, the roles of the different  $\mu$  parameters are quite different. Meanwhile  $\mu_1$  and  $\mu_3$  parameters are clearly affecting the same query likelihood and should be kept equal, for the other two parameters this is not so clear. The parameters  $\mu_1$  and  $\mu_4$  control the smoothing in the document language model when calculating the query likelihood in order to produce a ranking but the nature of the queries of both retrieval processes is quite different: shorter queries against longer queries. Nevertheless it is demonstrated in [24] that the optimal  $\mu$  values in both scenarios are quite similar, so we can fix  $\mu_1 = \mu_3 = \mu_4$ . On the contrary, the smoothing parameter  $\mu_2$  is used to control the smoothing when estimating the probability of the terms of the relevance model in order to select them to do the expansion. Although it is the language model of the document, here the document is not involved in the computation of a query likelihood, therefore, it can be considered a different parameter. For this reason it does not seem reasonable a-priori to fix the same values for the  $\mu$  parameters used for retrieval as for the  $\mu$  parameter used in the estimation of  $P(w|d)$ . This intuition was confirmed later in the experimentation, being the trained values quite different for both smoothing parameters. In fact, the optimal values trained in the evaluation process of both parameters in RM3 never matched.

## 4 Experiments and Results

This section describes the evaluation methodology and comments the results.

### 4.1 Collections

**Table 1.** Collections and topics for training and test.

<i>Col.</i>	<i># of Docs</i>	<i>Topics</i>	
		Training	Test
AP88-89	164,597	51-100	101-200
WT2G	247,491	401-450	–
TREC-678	528,155	301-350	351-450
WT10G	1,692,096	451-500	501-550

To evaluate the different approaches we chose the same collections used in previous works [14]: a subset of the Associated Press collection corresponding to the 1988 and 1989 years (AP88-89), the Small Web Collection WT2G and the disk 4 and 5 from TREC (TREC-678). Additionally, we decide to use the WT10G

collection, which was not used in [14], to report test values in a web collection. In AP88-89, TREC-678 and WT10G we used training and test evaluation: we performed training for MAP in a set of topics and testing over another set. In WT2G we report well-tuned values over the trained topics, as it was done in [14]. Short queries (title only) were used because they are the most suitable to be expanded. All the collections were preprocessed with standard stop-word removal and Porter stemmer. In Table 1 the evaluation settings are summarized.

## 4.2 Methods

We compared four methods:

- **LM**: the baseline Language Modelling retrieval model with Dirichlet smoothing as in Section 2.1
- **RM3**: the standard formulation of RM3, as explained in Section 2.2.
- **KLD3**: the KLD based PRF method adapted as detailed in Section 3.1.
- **RM3DT**: the proposed formulation of RM with estimations promoting divergent terms as described in Section 3.2.

## 4.3 Training and Evaluation

As discussed before, we performed a training and test strategy, more precisely we perform training and test for AP88-89, TREC-678 and WT10G meanwhile well-tuned values are reported for WT2G as in [14].

The parameters tuned were: the smoothing parameter of the initial retrieval  $\mu_1$  ( $\mu_1 \in \{10, 100, 1000, 2000, 3000, 4000, 5000, 6000\}$ ) that was also used for  $\mu_3$  and  $\mu_4$  and which was tuned for LM, KLD3, RM3 and RM3DT. The number of documents in the pseudo relevant set  $r = |RS|$  ( $r \in \{5, 10, 25, 50, 75, 100\}$ ) was tuned for KLD3, RM3 and RM3DT. The number of terms selected for expansion  $e$  ( $e \in \{5, 10, 25, 50, 75, 100\}$ ) was tuned for KLD3, RM3, RM3DT. The interpolation weight  $\lambda$  ( $\lambda \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$ ) was tuned for KLD3, RM3, RM3DT. The smoothing parameter  $\mu_2$  ( $\mu_2 \in \{10, 100, 1000, 2000, 3000, 4000, 5000, 6000\}$ ) was tuned for RM3 and RM3DT.

Finally, test values are reported for Mean Average Precision (MAP) and Robustness Index (RI) over the initial retrieval (LM). The Robustness Index ( $-1 \leq RI(q) \leq 1$ ), also called Reliability of Improvement Index, of a model with respect to a baseline was formulated by Sakai *et al.* in [18] as in Eq 12:

$$RI(q) = \frac{n_+ - n_-}{|q|} \quad (12)$$

where  $q$  is the set of queries over the RI has to be calculated,  $n_+$  is the number of improved queries,  $n_-$  the number of degraded queries and  $|q|$  the total number of queries in  $q$ .

#### 4.4 Results

Analyzing the MAP values for the test topics (see Table 2) it has to be noted that the three PRF methods always outperform the baseline LM as expected. The adaptation of the KLD method to the LM framework using query interpolation performs quite well, obtaining improvements up to the 32% in the AP88-89; this is a very interesting point considering that KLD3 has fewer parameters to tune. Nevertheless the other methods achieve statistically significant improvements over the KLD3 in four occasions.

**Table 2.** Values for Mean Average Precision (MAP) on the test topics. Statistical significant improvements (Wilcoxon  $p < 0.1$ , and Wilcoxon  $p < 0.05$  underlined) with respect to LM, RM3, KLD3, and RM3DT are superscripted with  $l$ ,  $r$ ,  $k$ , and  $d$  respectively. Best values are bolded.

<i>Col.</i>	<b>MAP</b>			
	<i>LM</i>	<i>RM3</i>	<i>KLD3</i>	<i>RM3DT</i>
AP88-89	.2775	.3606 <sup><i>l</i></sup> (+30%)	<b>.3667<sup><i>l</i></sup></b> (+32%)	.3625 <sup><i>l</i></sup> (+31%)
WT2G	.3115	.3445 <sup><i>lk</i></sup> (+10%)	.3352 <sup><i>l</i></sup> (+7%)	<b>.3467<sup><i>lk</i></sup></b> (+11%)
TREC-678	.2190	.2589 <sup><i>l</i></sup> (+18%)	.2586 <sup><i>l</i></sup> (+18%)	<b>.2700<sup><i>lrk</i></sup></b> (+23%)
WT10G	.2182	.2468 <sup><i>l</i></sup> (+13%)	.2238 (+2%)	<b>.2478<sup><i>lrk</i></sup></b> (+13%)

The RM3 method performs also quite well in terms of effectiveness with great improvements over the baseline as expected, as it is the state-of-the art in PRF. RM3 performs better than KLD3 in three collections, achieving in one case statistical significance. In the AP88-89 collection the differences across the three PRF methods are negligible, not being never statistically significant.

The proposed RM3DT estimation achieves statistically significant improvements over the KLD3 method in three occasions and over the RM3 in two, being always better than the later in terms of MAP. Another important point

**Table 3.** Values for Robustness Index (RI) with respect to the LM baseline model for every collection. Best values are bolded.

<i>Col.</i>	<b>RI</b>		
	<i>RM3</i>	<i>KLD3</i>	<i>RM3DT</i>
AP88-89	.38	<b>.56</b>	<b>.56</b>
WT2G	<b>.44</b>	.38	.40
TREC-678	.16	<b>.52</b>	.38
WT10G	.28	-.04	<b>.36</b>

to analyse is the robustness of the methods, and how this is maintained across collections. Considering the values presented in Table 3 we can conclude that the RI numbers of the KLD3 method are quite acceptable and similar across

collections, except in the WT10G collection. RM3 values are still acceptable (always bigger than zero) but are considerable lower than the other methods in the AP88-89 and TREC-678 collections. Contrarily RM3 performs slightly better than the other methods in the WT2G collection. This fact may be explained because the values on the WT2G collection are well-tuned, suggesting that a good parameter setting affects to the robustness of the RM3 method. Comparing both RM methods RM3DT seems to be more stable in terms of RI across collections.

The differences in robustness between RM3 and RM3DT can be analysed observing the queries penalized by RM3 and improved by RM3DT. Let us take as example the query **Parkinsons disease**, for this query LM obtained an average precision of 0.3231, RM3 damaged the query to 0.2927, while RM3DT improved it to 0.5083. Observing the top 25 expansion terms selected in both approaches we can view that many good terms are selected by both methods (for example **patient**, **brain** or **alzheimer**) but the RM3 method introduces terms that are so common that, although being very present in the RS, they introduce a lot of noise in the retrieval such as **page**, **can**, **year**, **will**, **new**, **say**, **may** or **home**, meanwhile those terms are not present in the top 25 RM3DT expansion terms because they were penalized for being so common in the collection.

## 5 Related Work

In [23] the authors explored the divergence idea proposing a Divergence Minimization Model (DMM). The DMM approach tries to minimize the divergence between the query model and the model of the feedback documents. The DMM objective is to build a feedback model that is close to every pseudo relevant document language model and far away from the collection language model, which is assumed as the non-relevance model. This was stated as an optimization problem. The DMM approach was already compared in [14] with Relevance Models showing that DMM performs worse than RM3.

This paper is centred in the Language Modelling framework but it is necessary to say that the idea of using divergence to improve the retrieval performance has been already deeply studied under other retrieval models, to the point of existing whole models based on it. The Divergence From the Randomness (DFR) model [2] is based on a similar idea: the more the terms occurrences in the documents diverge from their expected occurrences considering a random distribution the more information carried by the terms. In the DFR model the QE process is done based on a generalization of the Rocchio's framework [9]. Different weighting schemes, including the aforementioned KLD, were tested being the Bose-Einstein Bo1 model the best in terms of effectiveness, which also select those terms that diverge most from the randomness, using for those estimations the collections' statistics. In another paper [22] the Rocchio's classical feedback method was integrated in the DFR framework for PRF.

In other IR tasks such as adaptive filtering this divergence idea has also been used. In [13] the authors presented different discriminative features for queries and documents to be used in a technique which learns for each query the

interpolation weight of the original query with the expansion terms. Particularly the entropy of the feedback documents and the document clarity are used. With the entropy of the feedback documents basically they capture at term level how heterogeneous is the term distribution in the RS. With the clarity of the feedback documents they try to “explain away” common terms present in the RS.

## 6 Conclusions and Future Work

In this paper we have presented two different methods for PRF based on the idea of promoting the divergent terms in the RS. KLD3 is an adaptation to the LM framework of a KLD based method including the linear interpolation with the original query. RM3DT is a new estimation for the RM that computes the probability of a term given a feedback document by the subtracting to the terms’ probability in the document its probability in the collection and applying the smoothing over the RS. It was also analysed the role of the different smoothing parameters involved in the RM methods, showing the different roles that those smoothing parameters play. We compared the new methods with the LM baseline and the RM3 estimation. Particularly the RM3DT performed, for MAP, better than RM3 in every collection, showing, as the KLD3 method, a very good stability across collections in terms of robustness. We also want to study how the presented ideas may be applied to improve existing techniques for selective query expansion and adaptive relevance feedback.

**Acknowledgments:** This work was funded by *Ministerio de Ciencia e Innovación* under project TIN2008-06566-C04-04.

## References

1. N. Abdul-Jaleel, J. Allan, W. B. Croft, O. Diaz, L. Larkey, X. Li, M. D. Smucker, and C. Wade. UMass at trec 2004: Novelty and hard. In *In Proceedings of TREC-13*, 2004.
2. G. Amati and C. J. Van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst.*, 20:357–389, October 2002.
3. N. Balasubramanian, J. Allan, and W. B. Croft. A comparison of sentence retrieval techniques. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR ’07*, pages 813–814, New York, NY, USA, 2007. ACM.
4. C. Carpineto, R. de Mori, G. Romano, and B. Bigi. An information-theoretic approach to automatic query expansion. *ACM Trans. Inf. Syst.*, 19(1):1–27, 2001.
5. K. Collins-Thompson and J. Callan. Estimation and use of uncertainty in pseudo-relevance feedback. In *SIGIR ’07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 303–310, New York, NY, USA, 2007. ACM.
6. W. B. Croft and D. Harper. Using probabilistic models of document retrieval without relevance information. *Journal of Documentation*, 35:285–295, 1979.
7. W. B. Croft, D. Metzler, and T. Strohman. *Search Engines: Information Retrieval in Practice*. Addison-Wesley Publishing Company, USA, 1st edition, 2009.

8. T. Doszkocs. Id, an associative interactive dictionary for online searching. *Online Review*, 2:163–173, 1978.
9. B. He and I. Ounis. Combining fields for query expansion and adaptive query expansion. *Inf. Process. Manage.*, 43:1294–1307, September 2007.
10. V. Lavrenko and W. B. Croft. Relevance based language models. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 120–127, New York, NY, USA, 2001. ACM.
11. K. S. Lee, W. B. Croft, and J. Allan. A cluster-based resampling method for pseudo-relevance feedback. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 235–242, New York, NY, USA, 2008. ACM.
12. X. Li and Z. Zhu. Enhancing relevance models with adaptive passage retrieval. In *Proceedings of the IR research, 30th European conference on Advances in information retrieval*, ECIR'08, pages 463–471, Berlin, Heidelberg, 2008. Springer-Verlag.
13. Y. Lv and C. Zhai. Adaptive relevance feedback in information retrieval. In *Proceeding of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 255–264, New York, NY, USA, 2009. ACM.
14. Y. Lv and C. Zhai. A comparative study of methods for estimating query language models with pseudo feedback. In *Proceeding of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 1895–1898, New York, NY, USA, 2009. ACM.
15. S. E. Robertson. On term selection for query expansion. *J. Doc.*, 46:359–364, January 1991.
16. J. Rocchio. Relevance feedback in information retrieval. In *The SMART Retrieval System*, pages 313–323, 1971.
17. I. Ruthven and M. Lalmas. A survey on the use of relevance feedback for information access systems. *Knowl. Eng. Rev.*, 18(2):95–145, 2003.
18. T. Sakai, T. Manabe, and M. Koyama. Flexible pseudo-relevance feedback via selective sampling. *ACM Transactions on Asian Language Information Processing (TALIP)*, 4(2):111–135, 2005.
19. G. Salton. *The SMART Retrieval System: Experiments in Automatic Document Processing*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1971.
20. T. Tao and C. Zhai. Regularized estimation of mixture models for robust pseudo-relevance feedback. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 162–169, New York, NY, USA, 2006. ACM.
21. J. Xu and W. B. Croft. Query expansion using local and global document analysis. In *SIGIR '96: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 4–11, New York, NY, USA, 1996. ACM.
22. Z. Ye, B. He, X. Huang, and H. Lin. Revisiting rocchios relevance feedback algorithm for probabilistic models. In *Information Retrieval Technology*, volume 6458 of *Lecture Notes in Computer Science*, pages 151–161. 2010.
23. C. Zhai and J. Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the tenth international conference on Information and knowledge management*, CIKM '01, pages 403–410, New York, NY, USA, 2001. ACM.
24. C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214, 2004.