Selecting Discriminative Terms for Relevance Model

Dwaipayan Roy
Indian Statistical Institute, Kolkata
dwaipayan_r@isical.ac.in

Sumit Bhatia
IBM Research, Delhi, India
sumitbhatia@in.ibm.com

Mandar Mitra
Indian Statistical Institute, Kolkata
mandar@isical.ac.in

ABSTRACT

Pseudo-relevance feedback based on the relevance model does not take into account the inverse document frequency of candidate terms when selecting expansion terms. As a result, common terms are often included in the expanded query constructed by this model. We propose three possible extensions of the relevance model that address this drawback. Our proposed extensions are simple to compute and are independent of the base retrieval model. Experiments on several TREC news and web collections show that the proposed modifications yield significantly better MAP, precision, NDCG, and recall values than the original relevance model as well as its two recently proposed state-of-the-art variants.

CCS CONCEPTS

• **Information systems → Query representation**; **Query reformulation**.

ACM Reference Format:
Dwaipayan Roy, Sumit Bhatia, and Mandar Mitra. 2019. Selecting Discriminative Terms for Relevance Model. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19), July 21–25, 2019, Paris, France.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3331184.3331357

1 INTRODUCTION

Query expansion (QE) is a popular technique used for handling the *vocabulary mismatch* problem in information retrieval [6]. In pseudo-relevance feedback (PRF) based query expansion methods, the top-ranked documents retrieved by using the original query are used to select terms for query expansion. As a result, they alleviate the need to use external sources of information for query expansion making PRF based retrieval models [1, 10, 12] as amongst the most widely used query expansion methods in practice.

RM3 [8] is one of the most commonly used method for PRF and experiments reported by Lv and Zhai [11] demonstrated its robustness on different collections when compared with other feedback methods. They also found that RM3 term weighing is prone to select generic words as expansion terms that cannot help identify relevant documents. This occurrence of noise is a well-known shortcoming of PRF methods in general due to presence of non-relevant documents in the pseudo-relevant set (as precision is often less than one) [10]. The pseudo-relevant document set contains considerable noise that can lead to the expanded query *drifting* away from the original query [10]. Cao et al. [2] studied the utility of terms selected for expansion and found that only about a fifth of the expansion terms identified for query expansion contributed positively to an improvement in retrieval performance. Rest of the terms either had no impact on retrieval performance or led to the reduction in final retrieval performance.

Background Work: In order to prevent the inclusion of common terms in expanded queries, different methods have been proposed to compute better term weights for PRF models such as incorporating proximity of expansion terms to query terms [12], classifying expansion terms as good and bad by using features derived from term distributions, co-occurrences, proximity, etc. [2]. Parapar and Barreiro [15] proposed RM3-DT, an alternative estimation of term weighting in RM by subtracting the collection probability of the term from its document probability, thus giving a high weight to terms that have a higher probability of being present in the pseudo-relevant document set rather than the whole collection.

Clinchant and Gaussier [4] described an axiomatic framework and discussed five axioms for characterizing different PRF models. Building upon this framework, various improvements and modifications of the original relevance model have been proposed [1, 13, 14] to select better terms for query expansion and compute better weights by incorporating new constraints and assumptions in the computation of scores of terms present in the set of feedback documents. Recently, Montazeralghaem et al. [14] proposed two additional constraints to consider interdependence among previously established characteristics of pseudo-relevant feedback (PRF) models. The first constraint (TF-IDF constraint) considers the interrelationship of TF and IDF on PRF models and the second constraint (relevance score constraint) focuses on the interdependence of the feedback weight of selected terms and the relevance scores of documents containing the term. They proposed RM3+All, an extension of the RM3 model that accounts for these constraints.

Our Contributions: We propose three alternate heuristics for selecting the terms for query expansion that involve minimal computations on top of relevance model. Our selection process is simple and re-ranks the terms present in feedback documents based on the ratio of likelihoods of these terms being generated by the relevant and non-relevant document sets. We perform a thorough empirical evaluation of the proposed methods on standard TREC collections, ranging from TREC ad hoc collections to ClueWeb09 and compare the performance two state-of-the-art baselines (RM3DT [15] and RM3+All [14]). Despite being simple, the proposed methods significantly outperform RM3, RM3-DT, and RM3+ALL in terms of retrieval performance, are more robust and are computationally more efficient. We make our implementation available as a Lucene plugin[1] to enable further reuse and replication of our experiments.

[1] https://github.com/dwaipayanroy/rm3idf

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
SIGIR '19, July 21–25, 2019, Paris, France
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6172-9/19/07…$15.00
https://doi.org/10.1145/3331184.3331357

2 PROPOSED METHOD

2.1 Intuition Behind Proposed Approaches

According to the relevance model (RM) [9], for a given query $Q = \{q_1, \ldots, q_k\}$, the feedback weight of an expansion term w is computed as follows.

$$FW_{rm}(w) = P(w|\mathcal{R}) = \sum_{d \in \mathcal{D}} P(w|d) \prod_{q \in Q} P(q|d) \quad (1)$$

Here, $\mathcal{D}$ is the pseudo-relevant document set.

In RM3 [8], the feedback weight is estimated by linearly interpolating the query likelihood model of w ($P(w|Q)$) with the original relevance model $\mathcal{R}$ to obtain $\mathcal{R}'$ as follows.

$$FW_{rm3}(w) = P(w|\mathcal{R}') = \alpha P(w|Q) + (1-\alpha)P(w|\mathcal{R}) \quad (2)$$

Here, $\alpha \in [0, 1]$ is the interpolation parameter, and $P(w|Q)$ is estimated using maximum likelihood estimation.

Let us consider a term t that is present in a large number of documents in the collection. If t is present in the pseudo-relevant document set, the value of $P(t|\mathcal{R})$ (according to Equation 1) is likely to be quite high, thereby increasing the possibility of selecting t as an expansion term [4, 5, 7]. However, since t is also present in a large number of documents outside the pseudo-relevant document set, this term lacks the discriminating power to distinguish between relevant and non-relevant documents. Furthermore, during retrieval with the expanded query, such terms may adversely affect the computation time by unnecessarily increasing the query length. Additionally, if the term weights in the expanded query are sum-normalized (as done by Lavernko and Croft [9], Jaleel et al. [8]), these common terms may together reduce the weights of important terms in the expanded query. In sum, the inability of the RM3 model to filter out such terms makes it vulnerable to a potential degradation in retrieval performance. Intuitively, terms with a high discriminative power are expected to have different distributions over the set of relevant and non-relevant documents. Mathematically, this difference can be captured by the ratio of the occurrence probabilities of the term t being selected from the relevance ($\mathcal{R}$) and non-relevance ($\mathcal{N}$) classes, similar to the ideas by Robertson and Zaragoza [16]. Formally, for a given query Q, we consider two separate language models, associated respectively with the relevance class ($\mathcal{R}$, estimated following RM [9]) and non-relevance class ($\mathcal{N}$) of the query. To estimate $\mathcal{N}$, we choose the commonly-adopted alternative of using the whole document collection C since, for a typical query Q, most of the documents in the collection are used to be non-relevant. Based on this intuition, we now describe three approaches to compute the feedback term weights to select and weigh terms for query expansion.

2.2 RM3$_1^+$

In our first proposed approach, we compute the weights of candidate expansion terms by using $P(w|\mathcal{R})/P(w|\mathcal{N})$. The probability of selection of a term from the relevance class ($P(w|\mathcal{R})$) can be computed using Equation (1). We adopt the common practice of approximating the non-relevance class is by using the collection model C. Note that the basic intuition behind incorporating the non-relevance model as a factoring component is to identify terms with discriminating features that can also be considered as an estimation of *rareness*. Formally, the feedback weight of term w in the expanded query is computed as follows:

$$FW(w) = \frac{P(w|\mathcal{R})}{P(w|\mathcal{N})} \approx \frac{P(w|\mathcal{R})}{P(w|C)} \approx P(w|\mathcal{R}) * r(w) \quad (3)$$

where $r(w) \approx 1/P(w|C)$ can be interpreted as a measure of *rareness* of w. We may replace $r(w)$ by any other measure of a term's rareness. For our experiments, we use the traditional inverse document frequency factor in place of $r(w)$ (since this yields better retrieval effectiveness). Next, the weights are sum-normalized and the top N terms with the highest $FW(w)$ values are selected as expansion terms. The normalized feedback weights (denoted NFW) are then combined with the query likelihood model Q. Mathematically, the final weight of a term w in the expanded query is given by:

$$FW_{rm3_1^+}(w) = \alpha P(w|Q) + (1-\alpha) * NFW(w) \quad (4)$$

If Dirichlet smoothing is used (as recommended in [18]), then, during retrieval using the expanded query $\mathbf{Q}$, the score of a document d is computed in practice as follows:

$$Score(d, \mathbf{Q}) = \sum_{q \in \mathbf{Q}} NFW_{rm3_1^+}(q) \times \log \frac{\mu . P(q|C) + |d| . P(q|d)}{P(q|C)\Big(\mu + |d|\Big)} \quad (5)$$

where $NFW(q)$ represents the normalized weight of q in the expanded query, μ is the smoothing parameter, and $|d|$ represents the number of indexed words in d.

2.3 RM3$_2^+$

In RM3$_1^+$, a term w is weighted on the basis of a linear combination between the query language model ($P(w|Q)$) and the ratio between term selection probability from RM (Equation (1)) and from the non-relevance model. As an alternative approach, we consider the RM3 model (instead of RM in RM3$_1^+$) to approximate the relevance class, i.e, a term w is weighted by the probability ratio of how likely w is from the query-relevance model (RM3) and from the non-relevance model. Formally, candidate terms are weighted using a ratio similar to the one used in RM3$_1^+$, but unlike RM3$_1^+$, we use $P(w|\mathcal{R}')/P(w|\mathcal{N})$ to formulate the expansion term weight. Here, $P(w|\mathcal{R}')$ is approximated by Equation 2, and $P(w|\mathcal{N})$ is estimated as in $RM3_1^+$. Thus, the weight of term w in the expanded query is computed as follows.

$$\begin{aligned} FW_{rm3_2^+}(w) &= \frac{P(w|\mathcal{R}')}{P(w|\mathcal{N})} = \frac{P(w|\mathcal{R}')}{P(w|C)} \\ &\approx \alpha P(w|Q) * r(w) + (1-\alpha) P(w|\mathcal{R}) * r(w) \end{aligned} \quad (6)$$

Note that in this approach, the final score for a document incorporates the *rareness* information "twice": first, as a part of the computation of $FW(q)$ (Equation (6)), and again during retrieval because of the factor of $P(q|C)$ in the denominator in Equation (5).

Document Collection	Collection Type	#Docs	Query Fields	Query Set	Query Ids	Dev Set	Test Set
TREC Disk 1, 2	News	741,856	Title	TREC 1 ad-hoc	51-100	✓	
				TREC 2, 3 ad-hoc	101-200		✓
TREC Disks 4, 5 exclude CR	News	528,155	Title	TREC 6 ad-hoc	301-350	✓	
				TREC 7, 8 ad-hoc	351-450		✓
				TREC Robust	601-700		
GOV2	Web	25,205,179	Title	TREC Terabyte 1	701-750	✓	
				TREC Terabyte 2,3	751-850		✓
WT10G	Web	1,692,096	Title	WT10G	451-550	✓	
CW09B-S70		50,220,423		ClueWeb09	1-200		✓

Table 1: Dataset Overview

This overemphasis on the rareness of query terms may promote highly rare (but noisy) terms at the cost of useful, but more frequent query terms leading to a possible drop in retrieval effectiveness.

2.4 RM3$^+_3$

To address the potential issues due to double application of rareness information we adopt an approach similar to Carpineto et al. [3] where candidate expansion terms are ranked using one weighing function, and a different function is used to determine the final weight of the selected expansion terms. Specifically, we first rank the candidate expansion terms using Equation (6), and select the top N terms. Since rareness information is used in Equation (6), we hope that this will avoid the problem of selecting low-IDF words as expansion terms. Next, the final feedback weight for the selected terms is computed using Equation (2) instead of Equation (6). This avoids the use of a "double" IDF factor during final retrieval. Formally, the feedback term weight is computed as follows.

$$FW_{rm3^+_3}(w) = P(w|\mathcal{R}') = \alpha P(w|Q) + (1-\alpha)P(w|\mathcal{R}) \quad (7)$$

3 EXPERIMENTS

3.1 Datasets and Experimental Settings

We evaluate the proposed modifications to RM3 using standard TREC news and web collections. For parameter tuning, we split the topic sets into development and testing sets as summarized in Table 1.All the collections are stemmed using Porter's stemmer and stopwords are removed prior to indexing using Apache Lucene. For the initial retrieval using the original, unexpanded queries, we used the Dirichlet smoothed language model.All the baselines and the proposed methods were implemented using Lucene.

Baselines: For comparing with expansion based methods, we choose three baselines – *(i)* RM3 [8], *(ii)* RM3DT [15], a variant of RM3 that promotes divergent terms, and *(iii)* RM3+All [14], a recently proposed state-of-the-art modification of RM3.

Parameter Tuning: There are three parameters associated with the RM3 based QE method: number of documents (M), number of terms (N), and smoothing parameter (α) that adjusts the importance of query terms. The parameters M and N were varied in the range $\{10, 15, 20\}$ and $\{30, 40, 50, 60, 70\}$ respectively and α was varied in the range $\{0.1, \ldots, 0.9\}$ in steps of 0.1. The parameters were tuned individually for each of the methods on the development topics, and applied on the corresponding test topics (see Table 1). There are no additional parameters associated with RM3DT, RM3+All, and the proposed methods. The smoothing parameter μ is varied in the range $\{100, 200, 500, 1000, 2000, 5000\}$ and set to 1000 at which the optimal performance is observed on development topic sets. All our experiments were performed on a VM with Intel Xeon 2.80GHz CPU containing 80 cores, and 100GB of RAM.

3.2 Results and discussion

Retrieval Performance: Table 2 presents the results of retrieval experiments with all the baselines over all the test collections. We observe that for all the test collections, all of the three proposed modifications outperform the traditional RM3 [8], as well as the two baselines, for most of the evaluation measures. We also note that the second proposed method (RM3$^+_2$) is seen to be less effective compared to RM3$^+_1$ and RM3$^+_3$. For TREC news topics, the mean average precision (MAP), recall and NDCG at rank 10 are almost always seen to be significantly better for both RM3$^+_1$ and RM3$^+_3$ than the baselines. Improvement is also observed for P@10, however they are not significant for most of the topic sets. Results for TREC Terabyte track 2 and 3 topics exhibit similar improvements over the baselines. For ClueWeb09 topics, RM3$^+_3$ achieves significant improvements in MAP, recall, and P@10 values over RM3 and RM3DT. Compared to RM3+All, RM3$^+_3$ achieves significantly better MAP and recall values. We also report robustness index (RI) [17] to compare the robustness of methods when compared with the original RM3 model. We observe that the performance of both RM3$^+_1$ and RM3$^+_3$ is consistently more robust than the baselines. RM3DT achieves negative RI values for Terabyte 2,3 and ClueWeb collections indicating that more queries suffered in terms of retrieval performance than the queries that gained in retrieval performance. On the other hand, RM3$^+_3$ consistently achieved best RI values across all collections, except for the ClueWeb collection where it is a very close second. Thus, in terms of retrieval performance, RM3$^+_3$ achieves overall best performance among the baselines and the three proposed heuristics, followed by RM3$^+_1$.

Computational Latency: For each test collection, Table 4 reports the total time taken by different methods to execute all the queries (averaged over ten runs). Note that, we only compare the elapsed time for expansion term selection as the actual retrieval times (traversing inverted lists) would be common for all the methods. We observe that the execution time for the three proposed methods are consistently less than RM3+All. Further, RM3$^+_3$ takes the least amount of additional time over RM3 for finding expansion terms for Terabyte 2,3 and ClueWeb collections. Reading results in Tables 4 and 2 together, we conclude that the *RM3*$^+_3$ method, for most cases, achieves best retrieval performance, is robust, and computes the expansion terms faster than other methods studied.

Qualitative Analysis: We compare expansion terms selected by RM3, RM3DT, RM3+All and RM3$^+_3$ (best among the proposed modifications). Table 3 presents the top 15 expansion terms for topic `nuclear proliferation` (from TREC123 collection) for those methods. Observe that as discussed before, RM3 is prone to selecting common terms (low IDF) as exemplified by `year` in the list.On the other hand, RM3+All selects terms with significantly high IDF values. For example, expansion terms `kalayeh, qazvin` and `yongbyon` have collection frequency 1, 4 and 8 respectively in TREC 123 collection. However, such extremely rare terms may often be due to noise rather than topical relevance to the original query. In contrast, RM3$^+_3$ strikes a balance by preventing frequent terms from occupying top weights as well as not prioritizing very rare terms.

4 CONCLUSION

We proposed three modifications to RM3 model to promote selection of discriminative terms for query expansion. A thorough empirical evaluation was performed using TREC news and Web collections and results compared with two state-of-the-art RM3 variants for promoting high IDF terms for query expansion. The

	Metrics	LM	RM3	RM3DT	RM3+All	$RM3^+_1$	$RM3^+_2$	$RM3^+_3$
TREC23	MAP	0.2325	0.2936^{0}	0.2989^{0}	$0.2996^{0,1}$	$0.3054^{0,1,2,3}$	$0.3030^{0,1}$	$\mathbf{0.3075}^{0,1,2,3}$
	P@10	0.4990	0.5440^{0}	0.5438^{0}	0.5410^{0}	$\mathbf{0.5540}^{0}$	0.5490^{0}	0.5410^{0}
	Recall	0.6142	0.6734^{0}	0.6783^{0}	0.6796^{0}	$\mathbf{0.6906}^{0,1,2,3}$	$0.6859^{0,1,2,3}$	$0.6830^{0,1,2,3}$
	NDCG@10	0.5065	0.5449^{0}	0.5440^{0}	$0.5511^{0,1,2}$	$\mathbf{0.5538}^{0,1,2,3}$	$0.5510^{0,1}$	0.5399^{0}
	RI	-	-	0.11	0.16	0.12	0.11	**0.30**
TREC78Rb	MAP	0.2550	0.2906^{0}	$0.2931^{0,1}$	$0.2946^{0,1,2}$	$0.2998^{0,1}$	$0.2952^{0,1}$	$\mathbf{0.3036}^{0,1,2,3}$
	P@10	0.4372	0.4513^{0}	0.4591^{0}	0.4617^{0}	$0.4704^{0,1,2}$	$\mathbf{0.4714}^{0,1,2,3}$	$0.4678^{0,1,2}$
	Recall	0.7172	0.7828^{0}	0.7881^{0}	0.7891^{0}	0.7963^{0}	0.7923^{0}	$\mathbf{0.7976}^{0,1,2,3}$
	NDCG@10	0.4406	0.4478	$0.4503^{0,1}$	$0.4549^{0,1}$	$0.4551^{0,1}$	0.4560^{0}	$\mathbf{0.4577}^{0,1,2}$
	RI	-	-	0.12	0.19	0.11	0.06	**0.27**
Terabyte 2,3	MAP	0.2918	0.3201^{0}	0.3168^{0}	$0.3212^{0,2}$	$\mathbf{0.3372}^{0,1,2,3}$	$0.3237^{0,2}$	$0.3323^{0,1,2,3}$
	P@10	0.5680	0.6010^{0}	0.5995^{0}	0.6004^{0}	0.6130^{0}	0.6120^{0}	$\mathbf{0.6170}^{0,1,2,3}$
	Recall	0.7076	0.7364^{0}	0.7385^{0}	0.7391^{0}	$\mathbf{0.7503}^{0,1,2,3}$	0.7339^{0}	$0.7446^{0,1,2}$
	NDCG@10	0.4943	0.5155^{0}	0.5134^{0}	$0.5231^{0,1,2}$	$0.5234^{0,1,2}$	$0.5202^{0,2}$	$\mathbf{0.5252}^{0,1,2}$
	RI	-	-	-0.05	0.14	0.24	0.02	**0.32**
CW09B	MAP	0.1065	0.1081	0.1078	0.1081	$0.1137^{0,1,2,3}$	0.1080	$\mathbf{0.1148}^{0,1,2,3}$
	P@10	0.2258	0.2301^{0}	0.2297^{0}	0.2321^{0}	0.2318^{0}	0.2328	$\mathbf{0.2374}^{0,1,2}$
	Recall	0.4494	0.4539^{0}	0.4514^{0}	0.4591^{0}	$\mathbf{0.4702}^{0,1,2,3}$	0.4487	$0.4687^{0,1,2,3}$
	NDCG@10	0.1693	0.1699	0.1695	0.1702	0.1692	0.1664	**0.1737**
	RI	-	-	-0.01	0.10	**0.23**	0.00	0.22

Table 2: Performance of the proposed methods and different baselines. Statistically significant improvements (measured by paired t-Test with 95% confidence) over LM, RM3, RM3DT and RM3+ALL are indicated by superscript 0, 1, 2 and 3, respectively, with maximum improvement in bold-faced.

nuclear proliferation	**RM3**	nuclear. prolifer, weapon, 10, nation, year, soviet, 1, state, spread, date, regim, call, goal, limit
	RM3DT	nuclear, prolifer, intellig, cia, gate, mr, countri, weapon, soviet, iraq, chemic, effort, commun, agenc
	RM3-All	nuclear, prolifer, treati, weapon, farwick, pakistan, ntg, spector, qazvin, signatori, southasia, argentina, kalayeh, yongbyon, mcgoldrick
	$RM3^+_3$	nuclear, prolifer, treati, weapon, soviet, intern, signatori, nation, armamaent, assess, review, regim, union, spread, peac

Table 3: Top 15 expansion terms selected by RM3, RM3+All and $RM3^+_3$

Topics	Latency					
	RM3	**RM3DT**	**RM3+ALL**	**$RM3^+_1$**	**$RM3^+_2$**	**$RM3^+_3$**
TREC23	314	**320** (1%)	390 (24%)	325 (3%)	353 (12%)	349 (11%)
TREC78Rb	756	773 (2%)	847 (12%)	**769** (1%)	818 (8%)	778 (2%)
Tb 2,3	997	1005 (0.7%)	1046 (4%)	1004 (0.7%)	1009 (1%)	**1002** (0.5%)
CW09B	1547	1789 (15%)	1897 (22%)	1825 (17%)	1806 (16%)	**1743** (12%)

Table 4: Average computational latency in milliseconds for different methods on test queries. Percentage increase over RM3 (in parentheses); lowest increase in bold.

results suggested that the proposed heuristics (especially $RM3^+_3$) yield significant improvements in performance when compared with the baselines. Further, the improvements are more robust and the proposed heuristics are computationally more efficient than the baselines.

REFERENCES

[1] M. Ariannezhad et al. 2017. Iterative Estimation of Document Relevance Score for Pseudo-Relevance Feedback. In *Proc. of ECIR*. 676–683.

[2] G. Cao, J. Nie, J. Gao, and S. Robertson. 2008. Selecting Good Expansion Terms for Pseudo-relevance Feedback. In *Proc. of 31st ACM SIGIR*. ACM, 243–250.

[3] C. Carpineto, R. de Mori, G. Romano, and B. Bigi. 2001. An information-theoretic approach to automatic query expansion. *ACM Trans. Inf. Syst.* 19, 1 (2001), 1–27.

[4] Stéphane Clinchant and Eric Gaussier. 2013. A Theoretical Analysis of Pseudo-Relevance Feedback Models. In *Proc. of ICTIR 2013*. Article 6, 8 pages.

[5] Ronan Cummins. 2017. Improved Query-Topic Models Using Pseudo-Relevant PóLya Document Models. In *Proc. of the ACM ICTIR 2017*. 101–108.

[6] G. Furnas, T. Landauer, L. Gomez, and S. Dumais. 1987. The Vocabulary Problem in Human-system Communication. *Commun. ACM* 30, 11 (1987), 964–971.

[7] H. Hazimeh and C. Zhai. 2015. Axiomatic Analysis of Smoothing Methods in Language Models for Pseudo-Relevance Feedback. In *Proc. of ICTIR 2015*. 141–150.

[8] N. A. Jaleel, J. Allan, W. B. Croft, F. Diaz, L. S. Larkey, X. Li, M. D. Smucker, and C. Wade. 2004. UMass at TREC 2004: Novelty and HARD. In *Proc. TREC '04*.

[9] Victor Lavrenko and W. Bruce Croft. 2001. Relevance Based Language Models. In *Proc. of 24th SIGIR (SIGIR '01)*. ACM, New York, NY, USA, 120–127.

[10] Kyung Soon Lee, W B. Croft, and J. Allan. 2008. A cluster-based resampling method for pseudo-relevance feedback. In *SIGIR*. 235–242.

[11] Y. Lv and C. Zhai. 2009. A Comparative Study of Methods for Estimating Query Language Models with Pseudo Feedback. In *Proc. of 18th ACM CIKM*. 1895–1898.

[12] Y. Lv and C. Zhai. 2010. Positional relevance model for pseudo-relevance feedback. In *Proc. of 33rd ACM SIGIR*. ACM, 579–586.

[13] A. Montazeralghaem, H. Zamani, and A. Shakery. 2016. Axiomatic Analysis for Improving the Log-Logistic Feedback Model. In *Proc. of 39th ACM SIGIR*. 765–768.

[14] A. Montazeralghaem, H. Zamani, and A. Shakery. 2018. Theoretical Analysis of Interdependent Constraints in Pseudo-Relevance Feedback. In *SIGIR*. 1249–1252.

[15] Javier Parapar and Álvaro Barreiro. 2011. Promoting Divergent Terms in the Estimation of Relevance Models. In *Proc. of Third ICTIR'11*. 77–88.

[16] Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.* 3, 4 (April 2009), 333–389.

[17] T. Sakai, T. Manabe, and M. Koyama. 2005. Flexible pseudo-relevance feedback via selective sampling. *ACM TALIP* 4, 2 (2005), 111–135.

[18] C. Zhai and J. Lafferty. 2004. A Study of Smoothing Methods for Language Models Applied to Information Retrieval. *ACM TOIS*. 22, 2 (2004), 179–214.