Term Proximity Constraints for Pseudo-Relevance Feedback

Ali Montazeralghaem
School of ECE, College of Engineering
University of Tehran, Iran
ali.montazer@ut.ac.ir

Hamed Zamani
Center for Intelligent Information Retrieval,
University of Massachusetts Amherst
zamani@cs.umass.edu

Azadeh Shakery
School of ECE, College of Engineering
University of Tehran, Iran
School of Computer Science, Institute for Research in Fundamental Sciences
shakery@ut.ac.ir

ABSTRACT

Pseudo-relevance feedback (PRF) refers to a query expansion strategy based on top-retrieved documents, which has been shown to be highly effective in many retrieval models. Previous work has introduced a set of constraints (axioms) that should be satisfied by any PRF model. In this paper, we propose three additional constraints based on the proximity of feedback terms to the query terms in the feedback documents. As a case study, we consider the log-logistic model, a state-of-the-art PRF model that has been proven to be a successful method in satisfying the existing PRF constraints, and show that it does not satisfy the proposed constraints. We further modify the log-logistic model based on the proposed proximity-based constraints. Experiments on four TREC collections demonstrate the effectiveness of the proposed constraints. Our modification to the log-logistic model leads to significant and substantial (up to 15%) improvements. Furthermore, we show that the proposed proximity-based function outperforms the well-known Gaussian kernel which does not satisfy all the proposed constraints.

KEYWORDS

Term proximity, term position, axiomatic analysis, pseudo-relevance feedback, query expansion

ACM Reference format:
Ali Montazeralghaem, Hamed Zamani, and Azadeh Shakery. 2017. Term Proximity Constraints for Pseudo-Relevance Feedback. In *Proceedings of SIGIR '17, August 07-11, 2017, Shinjuku, Tokyo, Japan, ,* 4 pages.
DOI: 10.1145/3077136.3080728

1 INTRODUCTION

Pseudo-relevance feedback (PRF) is a query expansion strategy to address the vocabulary mismatch problem in information retrieval (IR). In PRF, a small set of top-retrieved documents (i.e., pseudo-relevant documents) are assumed to be relevant to the initial query. These pseudo-relevant documents are further used for updating the query model in order to improve the retrieval performance. PRF has been proven to be highly effective in many retrieval models [2, 10, 13, 14].

Theoretical analysis of PRF models has shown that there are several constraints (axioms) that every PRF model should satisfy. Based on these theoretical studies, several modifications, e.g., [1, 3, 9, 10], have been made to the existing PRF models which lead to significant improvements in the retrieval performance. Although term proximity has been shown to be a strong evidence for improving the retrieval performance [5, 12], especially in the PRF task [6–8], none of the existing constraints for PRF takes term proximity into account.[1]

In this paper, we provide a *theoretical analysis* for the use of term proximity in PRF models. To do so, we introduce three PRF constraints based on the proximity of candidate feedback terms and the query terms in the feedback documents. According to the first constraint ("proximity effect"), the candidate feedback terms that are positionally closer to the query terms in the feedback documents should be given higher weights in the feedback model. The second constraint ("convexity effect") decreases the effect of term proximity when the distance between terms increases. The third constraint indicates that proximity to the less common query terms is more important than proximity to the query terms that are general.

Furthermore, previous work on leveraging term proximity for IR tasks, including the positional relevance model, showed that the Gaussian kernel is an effective way for enhancing IR models with term proximity information [5, 6, 8]. In this paper, we show that the Gaussian kernel does not satisfy all the proposed constraints, and thus it could not be the best way for applying term proximity to PRF models.

The primary contributions of this work can be summarized as follows:

- Introducing three proximity-based constraints for theoretical analysis of PRF models.
- Studying and modifying the log-logistic feedback model [2], a state-of-the-art PRF model that outperforms many existing models, including the mixture model [14] and the geometric relevance model [11] (see [3] for more details).
- Introducing a variant of the Exponential kernel that satisfies all the proposed constraints.
- Evaluating our models using four TREC collections which demonstrates significant improvements over the original log-logistic model as well as the model enriched with the Gaussian kernel, a widely used weighting function for enhancing IR models using term proximity [5, 6, 8].

2 METHODOLOGY

In this section, we first introduce three proximity-based constraints that (pseudo-) relevance feedback methods should satisfy. We further analyze the log-logistic model [2], and show that this model

SIGIR '17, August 07-11, 2017, Shinjuku, Tokyo, Japan
© 2017 ACM. 978-1-4503-5022-8/17/08…$15.00
DOI: 10.1145/3077136.3080728

[1] Tao and Zhai [12] proposed two proximity-based constraints for retrieval models, but not for PRF models.

does not satisfy the proposed constraints. We finally modify the log-logistic feedback model in order to satisfy all the constraints.

We first introduce our notation. Let $FW(w; F, P_w, Q)$ denote the *feedback weight* function that assigns a real-valued weight to each feedback term w for a given query Q based on the feedback document set F. P_w is a set of term-dependent parameters. For simplicity, $FW(w)$ is henceforth used as the feedback weight function. In the following equations, TF and IDF are term frequency and inverse document frequency, respectively. $| \cdot |$ represents the size of the given set.

2.1 Constraints

In this subsection, we introduce three proximity-based constraints for PRF methods.

[Proximity effect] Let $d(w, q, D)$ denote the proximity weight of a candidate feedback term w and a given query term q in a feedback document D. Then, the following constraint should hold:

$$\frac{\partial FW(w)}{\partial d(w, q, D)} < 0$$

According to this constraint, the candidate feedback terms that are closer to the query terms in the feedback documents should have higher weights. Intuitively, the closer terms to the query terms are more likely to be relevant to the query.

[Convexity effect] The feedback weight function should be convex with respect to the distance of candidate feedback terms from the query terms. We can formalize this constraint as follows:

$$\frac{\partial^2 FW(w)}{\partial d(w, q, D)^2} > 0$$

The intuition behind this constraint is that decreasing the effect of the proximity effect should be less marked in high distance ranges.

[Query IDF effect] Let $Q = \{q_1, q_2\}$ be a query with two query terms q_1 and q_2. Let D_1 and D_2 denote two feedback documents with equal length, such that q_1 only appears in D_1, and q2 only appears in D_2. Let w_1 and w_2 be two candidate feedback terms, such that $TF(w_1, D_1) = TF(w_2, D_2)$, and w_1 and w_2 only appear in D_1 and D_2 in the feedback set, respectively. Also assume that $d(w_1, q_1, D_1) = d(w_2, q_2, D_2)$ where d is the function to compute the proximity between two terms in a given document. We can say if $IDF(q_1) > IDF(q_2)$, then we should have:

$$FW(w_1) > FW(w_2)$$

Intuitively, this constraint indicates that proximity to the query terms that are general is less important than proximity to the uncommon query terms. For instance, if a query contains a general term (let say a stopword) then proximity to this term should be less important than the discriminative terms that occur in the query.

2.2 Modifying the Log-Logistic Model

As a case study, we analyze and modify the log-logistic feedback model [2]. The reason is that this method has been shown to outperform many strong baselines, including the mixture model [14] and the geometric relevance model [11]. It has been also shown that this method successfully satisfies all the PRF constraints proposed in [3]. The log-logistic feedback weight function for each term w is computed as:

$$FW(w) = \frac{1}{|F|} \sum_{D \in F} \log(1 + \frac{t(w, D)}{\lambda_w}) \quad (1)$$

where $t(w, d) = TF(w, D) \log(1 + c\frac{avg_l}{|D|})$ (avg_l denotes the average document length and c is a free hyper-parameter that controls the document length effect). The document frequency term λ_w is calculated as:

$$\lambda_w = N_w / N \quad (2)$$

where N_w and N denote the number of documents in the collection that contain w and the total number of documents in the collection, respectively. $FW(w)$ is then interpolated with the original query based on a free parameter (feedback coefficient)[2].

It can be easily shown that the log-logistic feedback model does not satisfy the proximity-based constraints, since its formulation does not contain any proximity-based component. To the best of our knowledge, this is the first attempt to enrich the log-logistic feedback model using term proximity information.

Regarding the query term independence assumption, we propose to modify the log-logistic model as follows to satisfy the proximity-based constraints:

$$FW_{prox}(w) = FW(w) * \sum_{D \in F} \sum_{q \in Q} \delta(w, q, D) \quad (3)$$

where q is a query term and $\delta(w, q, D)$ is a function that computes the proximity of w and q in document D.

To define the function δ, we propose to use the Exponential kernel that satisfies the "proximity effect" and the "convexity effect" constraints. We modify the Exponential kernel by adding an IDF term to satisfy the "Query IDF effect" constraint, as well. The function δ can be computed as follows:

$$\delta(w, q, D) = \exp\left(-\frac{d(w, q, D)}{\alpha}\right) . \log \frac{1}{\lambda_q} \quad (4)$$

where $d(w, q, D)$ denotes the distance function for two given terms w and q in document D. λ_q is the document frequency component for the query term q (see Equation (2)) and α is a free parameter. Several approaches have been proposed to compute $d(w, q, D)$, such as average, minimum, and maximum distances. Tao and Zhai [12] showed that using the *minimum* distance between the term w and the query term q in the feedback document outperforms other distance functions. We also use the minimum distance as follows:

$$d(w, q, D) = \min_{w_i \in \vec{w} \ \& \ q_j \in \vec{q}} |w_i - q_j| \quad (5)$$

where $\vec{w}$ and $\vec{q}$ are two vectors containing the positions of term w and query term q in document D, respectively.

3 EXPERIMENTS

3.1 Experimental Setup

We used four standard TREC collections in our experiments: AP (Associated Press 1988-89), Robust (TREC Robust Track 2004 collection), WT2g (TREC Web Track 2000 collection), and WT10g (TREC Web Track 2001-2002 collection). The first two are newswire collections, while the next two are web collections containing more noisy documents. The statistics of these collections are reported in Table 1. We considered the title of topics as queries. All documents

Table 1: Collections statistics.

Collection	TREC topics	#docs	doc length	#qrels
AP	51-200	165k	287	15838
Robust	301-450 & 601-700	528k	254	17412
WT2g	401-450	247k	645	2279
WT10g	451-550	1692k	399	5931

were stemmed using the Porter stemmer and stopped using the standard INQUERY stopword list. We carried out the experiments using the Lemur toolkit[2].

3.1.1 Parameter Setting. The number of feedback documents, the feedback term count, and the feedback coefficient were set using 2-fold cross validation over the queries of each collection. We swept the number of feedback documents between $\{10, 25, 50, 75, 100\}$, the feedback term count between $\{10, 25, 50, 75, 100\}$, and the feedback coefficient between $\{0, 0.1, \cdots, 1\}$. The parameters c and α were also selected using the same procedure from $\{2, 4, \cdots, 10\}$ and $\{25, 50, \cdots, 500\}$, respectively. The parameter σ in the Gaussian kernel is also set similarly.

3.1.2 Evaluation Metrics. To evaluate retrieval effectiveness, we use mean average precision (MAP) of the top-ranked 1000 documents as the main evaluation metric. In addition, we also report the precision of the top 10 retrieved documents (P@10). Statistically significant differences of average precisions are determined using the two-tailed paired t-test computed at a 95% confidence level. To evaluate robustness of methods, we consider the robustness index (RI) introduced in [4].

3.2 Results and Discussion

In this subsection, we first empirically show that satisfying each of the introduced constraints improves the retrieval performance. Our experiments also demonstrate that the Gaussian kernel that has previously been used in the literature [5, 6, 8] is not as effective as the proposed proximity weighting function, since the Gaussian kernel does not satisfy all the constraints.

3.2.1 Analysis of the Proximity-based Constraints. We consider two baselines: (1) the document retrieval method without feedback (NoPRF), and (2) the original log-logistic feedback model (LL). Although there are several effective PRF methods, since in this paper we study the effect of the proposed constraints in the log-logistic model, we do not consider other existing PRF methods.

To study the influence of each of the proposed constraints on the retrieval performance, we consider three different proximity functions: (1) the quadratic function (Quad) that only satisfies the "proximity effect" constraint, (2) the exponential function (Exp) that satisfies both "proximity effect" and "convexity effect" constraints, and (3) a modified version of the exponential function (Exp*) that satisfies all three constraints. More detail is reported Table 2.

The results of the baselines and the aforementioned methods are reported in Table 3. According to this table, modifying the log-logistic method using each of the proximity functions improves the retrieval performance, in all collections. The MAP improvements are statistically significant in nearly all cases. This indicates the necessity of taking term proximity into account for the PRF task.

Table 2: Summary of different proximity functions with respect to the proximity-based constraints ($x = d(w, q, D)$).

Func.	$\delta(w, q, D)$	Proximity effect	Convexity effect	Query IDF effect
Gaus	$\exp[\frac{-x^2}{2\sigma^2}]$	Yes	Partially	No
Quad	$-(\frac{x}{\alpha})^2 + 1$	Yes	No	No
Exp	$\exp[\frac{-x}{\alpha}]$	Yes	Yes	No
Exp*	$\exp[\frac{-x}{\alpha}].\log\frac{N}{N_q}$	Yes	Yes	Yes

The results demonstrate that LL+Exp outperforms LL+Quad and LL+Exp* outperforms LL+Exp, in all collections. The improvements in the web collections are higher than those in the newswire collections. The reason is that these two collections are web crawls and contain more noisy documents compared to the newswire collections. The other reason is that the WT2g and WT10g documents are much longer than the AP and Robust documents on average (see Table 1). The influence of proximity-based constraints can be highlighted in longer documents. Besides that, in terms of RI, LL+Exp* outperforms all the baselines in AP, WT2g and WT10g which shows the importance and robustness of Query IDF effect and these improvements are impressive in WT2g and WT10g which shows that this effect is important in noisy (web) collections.

3.2.2 Analysis of the Gaussian Kernel. In this set of experiments, we study the Gaussian kernel for computing the proximity weight, which has been shown to be the most effective proximity function among the existing ones [5]. As reported in Table 2, employing the Gaussian kernel for PRF satisfies the "proximity effect" constraint. The "convexity effect" constraint is only satisfied when $d(w, q, D) > \sigma$. Therefore, it does not satisfy the "convexity effect" for the candidate feedback terms that are close to the query terms. We evaluate the Gaussian kernel by considering it as a term proximity weight function (LL+Gaus). According to the results reported in Table 3, LL+Exp and LL+Gaus perform comparably in the newswire collections, but LL+Exp outperforms LL+Gaus in the web collections (WT2g and WT10g). LL+Exp* also outperforms LL+Gaus in all collections. The improvements in the web collections are statistically significant.

3.2.3 Parameter Sensitivity. In the last set of experiments, we study the sensitivity of the proposed method to the following hyper-parameters: the number of feedback terms added to the query, (2) the feedback interpolation coefficient, and (3) the parameter α (see Equation (4)). To do so, we sweep one of the parameters and fix the other ones to their default values: 50 for feedback term count, 0.5 for feedback coefficient, and 25 for α. In these experiments, we report the result for LL+Exp* that achieves the best performance in Table 3. The results are plotted in Figure 1, in terms of MAP. In this figure, the first plot shows that the performance of LL+Exp* is stable with respect to the changes in the number of feedback terms, when more than 25 terms are added to the query. In other words, 25 terms are enough for expanding the query in most collections. The second plot in Figure 1 demonstrates that the behaviour of LL+Exp* in newswire collections is similar to each other. LL+Exp* also behaves similarly in the web collections. Interestingly, the feedback model estimated by LL+Exp* does not need to be interpolated with the original query

[2] http://lemurproject.org/

Table 3: Performance of different proximity functions applied to the log-logistic model. Superscripts 0/1/2 denote that the MAP improvements over NoPRF/LL/LL+Gaus are statistically significant. The highest value in each column is marked in bold.

Method	AP			Robust			WT2g			WT10g		
	MAP	P@10	RI	MAP	P@10	RI	MAP	P@10	RI	MAP	P@10	RI
NoPRF	0.2642	0.4260	–	0.2490	0.4237	–	0.3033	0.4480	–	0.2080	0.3030	–
LL	0.3385	0.4622	0.15	0.2829	0.4393	**0.33**	0.3276	0.4820	0.36	0.2127	0.3187	0.08
LL+Gaus	0.3471^{01}	0.4695	0.19	0.2926^{01}	0.4454	0.27	0.3351^{0}	**0.4920**	0.38	0.2393^{01}	0.3157	0.16
LL+Quad	0.3441^{01}	0.4682	0.18	0.2920^{01}	**0.4530**	0.30	0.3309^{0}	**0.4920**	0.34	0.2195^{0}	0.3075	0.02
LL+Exp	0.3468^{01}	0.4688	0.19	0.2936^{01}	0.4442	0.28	0.3418^{012}	0.4840	0.38	0.2435^{01}	0.3247	0.19
LL+Exp*	**0.3475^{01}**	**0.4702**	**0.21**	**0.2950^{01}**	0.4430	0.30	**0.3449^{012}**	0.4820	**0.47**	**0.2461^{012}**	**0.3278**	**0.24**

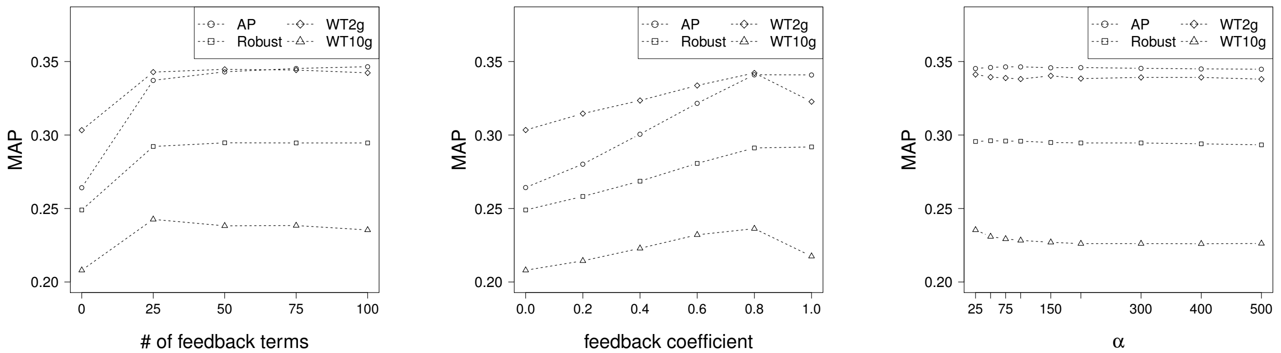

Figure 1: Sensitivity of LL+Exp* to the number of feedback terms, the feedback coefficient, and the parameter α.

model in the newswire collections. In the web collections (WT2g and WT10g), giving a small weight (i.e., 0.2) to the original query model can help to improve the retrieval performance. The reason could be related to the noisy nature of the web collections compared to the newswire collections. The last plot in Figure 1 shows that the proposed method is not highly sensitive to the parameter α when it is higher than 100. The results indicate that 25 and 50 would be proper values for this parameter. The results on the web collections are more sensitive to this parameter. The reason is that the documents in the web collections are much longer than those in the newswire collections (see Table 1).

4 CONCLUSIONS AND FUTURE WORK

In this paper, we proposed three constraints for the pseudo-relevance feedback models, that focus on the proximity of the candidate feedback terms and the query terms in the feedback documents. To show the effectiveness of the proposed constraints, we considered the log-logistic model, a state-of-the-art feedback model, as a case study. We first showed that the log-logistic model does not satisfy the proximity-based constraints. We further modified it based on the proposed constraints. Our experiments on four standard TREC newswire and web collections demonstrated the effectiveness of the proposed constraints for the PRF task. The modified log-logistic model significantly outperforms the original log-logistic model, in all collections. We also showed that the Gaussian kernel that has been used in previous proximity-based methods does not satisfy all the constraints. We show that the performance of the proposed variant of the exponential kernel is superior to those obtained by employing the Gaussian kernel. As a future direction, the other existing PRF models could be analyzed and modified based on the introduced constraints.

Acknowledgements. This work was supported in part by the Center for Intelligent Information Retrieval and in part by a grant from the Institute for Research in Fundamental Sciences (No. CS1396-4-51). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsors.

REFERENCES

[1] Mozhdeh Ariannezhad, Ali Montazeralghaem, Hamed Zamani, and Azadeh Shakery. 2017. Iterative Estimation of Document Relevance Score for Pseudo-Relevance Feedback. In *ECIR '17*. 676–683.

[2] Stéphane Clinchant and Eric Gaussier. 2010. Information-based Models for Ad Hoc IR. In *SIGIR '10*. 234–241.

[3] Stéphane Clinchant and Eric Gaussier. 2013. A Theoretical Analysis of Pseudo-Relevance Feedback Models. In *ICTIR '13*. 6–13.

[4] Kevyn Collins-Thompson. 2009. Reducing the Risk of Query Expansion via Robust Constrained Optimization. In *CIKM '09*. 837–846.

[5] Yuanhua Lv and ChengXiang Zhai. 2009. Positional Language Models for Information Retrieval. In *SIGIR '09*. 299–306.

[6] Yuanhua Lv and ChengXiang Zhai. 2010. Positional Relevance Model for Pseudo-relevance Feedback. In *SIGIR '10*. 579–586.

[7] Yuanhua Lv, ChengXiang Zhai, and Wan Chen. 2011. A Boosting Approach to Improving Pseudo-relevance Feedback. In *SIGIR '11*. 165–174.

[8] Jun Miao, Jimmy Xiangji Huang, and Zheng Ye. 2012. Proximity-based Rocchio's Model for Pseudo Relevance. In *SIGIR '12*. 535–544.

[9] Ali Montazeralghaem, Hamed Zamani, and Azadeh Shakery. 2016. Axiomatic Analysis for Improving the Log-Logistic Feedback Model. In *SIGIR '16*. 765–768.

[10] Dipasree Pal, Mandar Mitra, and Samar Bhattacharya. 2015. Improving Pseudo Relevance Feedback in the Divergence from Randomness Model. In *ICTIR '15*. 325–328.

[11] Jangwon Seo and W. Bruce Croft. 2010. Geometric Representations for Multiple Documents. In *SIGIR '10*. 251–258.

[12] Tao Tao and ChengXiang Zhai. 2007. An Exploration of Proximity Measures in Information Retrieval. In *SIGIR '07*. 295–302.

[13] Hamed Zamani, Javid Dadashkarimi, Azadeh Shakery, and W. Bruce Croft. 2016. Pseudo-Relevance Feedback Based on Matrix Factorization. In *CIKM '16*. 1483–1492.

[14] Chengxiang Zhai and John Lafferty. 2001. Model-based Feedback in the Language Modeling Approach to Information Retrieval. In *CIKM '01*. 403–410.