Axiomatic Analysis for Improving the Log-Logistic Feedback Model

Ali Montazeralghaem[†], Hamed Zamani[‡], and Azadeh Shakery[†]
[†]School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Iran
[‡]Center for Intelligent Information Retrieval, College of Information and Computer Sciences, University of Massachusetts Amherst, MA 01003
{ali.montazer,shakery}@ut.ac.ir zamani@cs.umass.edu

ABSTRACT

Pseudo-relevance feedback (PRF) has been proven to be an effective query expansion strategy to improve retrieval performance. Several PRF methods have so far been proposed for many retrieval models. Recent theoretical studies of PRF methods show that most of the PRF methods do not satisfy all necessary constraints. Among all, the log-logistic model has been shown to be an effective method that satisfies most of the PRF constraints. In this paper, we first introduce two new PRF constraints. We further analyze the log-logistic feedback model and show that it does not satisfy these two constraints as well as the previously proposed "relevance effect" constraint. We then modify the log-logistic formulation to satisfy all these constraints. Experiments on three TREC newswire and web collections demonstrate that the proposed modification significantly outperforms the original log-logistic model, in all collections.

CCS Concepts

•Information systems → Query representation; Query reformulation;

Keywords

Pseudo-relevance feedback; axiomatic analysis; theoretical analysis; query expansion; semantic similarity

1. INTRODUCTION

Pseudo-relevance feedback (PRF) refers to a query expansion strategy to address the vocabulary mismatch problem in information retrieval (IR). PRF assumes that a number of top-retrieved documents are relevant to the initial query. Based on this assumption, it updates the query model using these pseudo-relevant documents to improve the retrieval performance. PRF has been shown to be highly effective in many retrieval models [1, 5, 8, 10].

Several PRF models with different assumptions and formulations have so far been proposed. Clinchant and Gaussier [2] theoretically analyzed a number of effective PRF models. To this end, they proposed five constraints (axioms) for PRF models and showed that the log-logistic feedback model [1] is the only PRF model (among the studied ones) that satisfies all the constraints. They also showed that its performance is superior to the other PRF methods, including the mixture model [10] and the geometric relevance model [9]. Effectiveness of the log-logistic model motivates us, in this paper, to study this state-of-the-art PRF model.

Recently, Pal et al. [8] proposed a sixth constraint for PRF models to improve the PRF performance in the divergence from randomness framework. This constraint, which is called "relevance effect", indicates that the terms in the feedback documents with high relevance scores (i.e., relevance of document to the initial query) should have higher weights in the feedback model compared to those with exactly similar statistics, but appear in the documents with lower relevance scores. Formally writing, if a term w occurs in two documents $d_1, d_2 \in F$ (F denotes the set of feedback documents) such that d_1 is more relevant to the initial query than d_2. Then, we can say that the feedback weight of w given the $F - \{d_1\}$ feedback documents is lower than the weight of the same word in the $F - \{d_2\}$ feedback documents [8]. It can be shown that the log-logistic feedback model does not satisfy the relevance effect constraint.

In this paper, we propose two additional constraints for PRF models. The first constraint considers the semantic similarity of feedback terms to the initial query. Although previous work, such as [4], proposed similar constraints for retrieval models, to the best of our knowledge, it is the first time to study a semantic-related constraint for the PRF task. The second constraint indicates that the weight of each term w in the feedback model not only depends on the distribution of w in the feedback documents, but is also related to the distribution of the other terms in those documents. We further show that the log-logistic model does not satisfy the two proposed constraints. We then propose a modification to the log-logistic feedback model to satisfy the proposed constraints as well as the relevance effect constraint [8].

We evaluate the modified log-logistic model using three standard TREC collections: AP (Associated Press 1988-89), Robust (TREC 2004 Robust track), and WT10g (TREC 9-10 Web track). The experimental results demonstrate that the proposed method significantly outperforms the original log-logistic feedback model in all collections. The proposed method is also shown to be more robust than the original log-logistic model, especially in the web collection.

SIGIR '16, July 17-21, 2016, Pisa, Italy

DOI: http://dx.doi.org/10.1145/2911451.2914768

2. METHODOLOGY

In this section, we introduce two constraints that (pseudo) relevance feedback methods should satisfy (in addition to those proposed in [2, 8]). We further analyze the log-logistic model, a state-of-the-art feedback model, and figure out that this model does not satisfy the proposed constraints as well as the "relevance effect" constraint introduced in [8]. Based on these observations, we modify the log-logistic feedback model to satisfy all the constraints.

We first introduce our notation. Let $FW(w; F, P_w, Q)$ be the *feedback weight* function that assigns a real-value weight to each feedback term w for a given query Q. F and P_w respectively denote the set of feedback documents for the query Q and a set of term-dependent parameters. For simplicity, we henceforth use $FW(w)$. In the following equations, TF and IDF denote term frequency and inverse document frequency, respectively. The notation $|\cdot|$ is also used for query/document length or size of a given set.

2.1 Constraints

In this subsection, we introduce two constraints for feedback models.

[Semantic effect] Let Q be a single-term query (i.e., $Q = \{q\}$), w_1 and w_2 be two terms such that $IDF(w_1) = IDF(w_2)$, $\forall D \in F : TF(w_1, D) = TF(w_2, D)$, and

$$sem(q, w_1) < sem(q, w_2)$$

where $sem(\cdot, \cdot)$ denotes the semantic similarity of the given terms. Then, we can say:

$$FW(w_1) < FW(w_2)$$

The intuition behind this constraint is that the feedback terms should be semantically similar to the initial query.

[Distribution effect] Let w_1 and w_2 be two vocabulary terms such that $TF(w_1, D_1) = TF(w_2, D_2)$, $TF(w_1, D_2) = TF(w_2, D_1) = 0$, and $|D_1| = |D_2|$, where D_1 and D_2 are two documents in the feedback set F. Also, assume that w_1 and w_2 do not occur in other feedback documents, and

$$UniqueTerms(D_1) < UniqueTerms(D_2)$$

where $UniqueTerms(\cdot)$ denotes the number of unique terms in the given document. Then, we can say:[1]

$$FW(w_1) < FW(w_2)$$

In other words, this constraint implies that for computing the feedback weight of a term w, the distribution of other terms in the feedback documents should also be considered.

2.2 Modifying the Log-Logistic Model

The feedback weight of each term w in the log-logistic feedback model [1] is computed as follows:

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

where $\lambda_w = \frac{N_w}{N}$ (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), and $t(w, D) = TF(w, D) \log(1 + c\frac{avg_l}{|D|})$ (avg_l denotes the average document length and c is a free hyper-parameter). It is shown that the log-logistic model satisfies all the PRF constraints introduced in [2]. It can be easily proved that this model cannot satisfy the constraints proposed in this paper. In more detail, there is no semantic-related or relevance-related components in the log-logistic formulation and thus it cannot satisfy the proposed "semantic effect" and the "relevance effect" [8] constraints. In addition, the log-logistic formula does not consider the distribution of other terms in computing the weight of each term w, and thus it does not satisfy the "distribution effect" constraint.

To satisfy the "semantic effect" constraint, we modify the log-logistic feedback weight function as follows:

$$FW_{sem}(w) = FW(w) * \beta * \frac{1}{|Q|} \sum_{q \in Q} \frac{s(w, q)}{s(q, q)} \quad (2)$$

where $s(\cdot, \cdot)$ denotes the semantic similarity between the given two terms. The parameter β controls the effect of semantic similarity in the feedback weight function. The semantic weighting component comes from the query-growth function, which was previously proposed by Fang and Zhai [4]. Note that in Equation (2), we can ignore the $1/|Q|$ term and the β parameter, since they are equal for all terms and the feedback weighting function will be normalized. Several methods have so far been proposed to incorporate semantic similarity of terms in various retrieval tasks. In this paper, we consider the mutual information as a basic semantic similarity metric to compute $s(\cdot, \cdot)$. The mutual information (MI) of two terms w and w' is computed as follows:

$$I(X_w, X_{w'}) = \sum_{X_w, X_{w'} \in \{0,1\}} p(X_w, X_{w'}) \log \frac{p(X_w, X_{w'})}{p(X_w)p(X_{w'})}$$

where X_w and $X_{w'}$ are two binary random variables that represent the presence or absence of the terms w and w' in each document. A simple way to compute the mutual information is to consider the whole collection; but, this choice may not be ideal for ambiguous terms. Another way is to compute the mutual information from the pseudo-relevant documents. However, the top-retrieved documents could be a biased corpus for this goal. Therefore, similar to [4], we extract the mutual information from a corpus containing the top m retrieved documents and $r \times m$ documents randomly selected from the collection, where r is a free parameter that controls the generality of mutual information scores.

To satisfy the "distribution effect" constraint, we re-define the function $t(w, D)$ as follows:

$$t^*(w, D) = \frac{t(w, D)}{\log(\frac{|D|}{ut(D)})} \quad (3)$$

where $ut(D)$ denotes the number of unique terms in the document D. A similar approach for modifying the raw TF formula was previously used in [7].

To satisfy the "relevance effect" constraint, we re-define the function $FW(w, D)$ (see Equation (1)) as follows:

$$FW^*(w, D) = FW(w, D) * RS(Q, D) \quad (4)$$

where $RS(Q, D)$ denotes the relevance score of the document D to the query Q. This function can be computed using the relevance score of D in the first ranking phase in PRF. A similar idea was previously proposed by Lavrenko and Croft [5]. They used the query likelihood similarity as a posterior probability in the relevance models. Lv and

[1]The intuition behind this constraint comes from the definition of *information* in information theory literature.

Table 1: Collections statistics.

ID	Collection	Queries (title only)	#docs	doc length	#qrels
AP	Associated Press 88-89	TREC 1-3 Ad Hoc Track, topics 51-200	165k	287	15,838
Robust	TREC Disks 4 & 5 minus Congressional Record	TREC 2004 Robust Track, topics 301-450 & 601-700	528k	254	17,412
WT10g	TREC Web Collection	TREC 9-10 Web Track, topics 451-550	1692k	399	5931

Table 2: Performance of the proposed modifications and the baselines. Superscripts 0/1 denote that the MAP improvements over NoPRF/LL are statistically significant. The highest value in each column is marked in bold.

Method	AP			Robust			WT10g		
	MAP	P@10	RI	MAP	P@10	RI	MAP	P@10	RI
NoPRF	0.2642	0.4260	–	0.2490	0.4237	–	0.2080	0.3030	–
LL	0.3385	0.4622	0.15	0.2829	0.4393	0.33	0.2127	0.3187	0.08
LL+Sem	0.3422^{0}	0.4702	0.18	0.2940^{01}	0.4474	0.31	0.2247	0.3188	0.10
LL+Rel	0.3425^{0}	0.4681	**0.20**	0.2897^{01}	**0.4490**	0.35	0.2289^{01}	**0.3289**	0.17
LL+Dis	0.3386^{0}	0.4671	0.16	0.2831^{0}	0.4401	0.32	0.2194	0.3207	0.13
LL+All	$\mathbf{0.3445^{01}}$	**0.4722**	**0.20**	$\mathbf{0.2979^{01}}$	0.4486	**0.36**	$\mathbf{0.2300^{01}}$	0.3177	**0.19**

Zhai [6] also used a similar technique to improve the divergence minimization feedback model [10].

Considering the aforementioned modifications, we can rewrite the log-logistic feedback weighting formula as follows:

$$FW^*(w) = \frac{1}{|F|}\sum_{D\in F}\left(\log(\frac{t^*(w,D)+\lambda_w}{\lambda_w}) * RS(Q,D) * \sum_{q\in Q}\frac{s(w,q)}{s(q,q)}\right) \quad (5)$$

3. EXPERIMENTS

3.1 Experimental Setup

We used three standard TREC collections in our experiments: AP (Associated Press 1988-89), Robust (TREC Robust Track 2004 collection), and WT10g (TREC Web Track 2001-2002). The first two collections are newswire collections, and the third collection is a web collection with more noisy documents. The statistics of these datasets are reported in Table 1. We consider the title of topics as queries. All documents are stemmed using the Porter stemmer. Stopwords are removed in all the experiments. We used the standard INQUERY stopword list. All experiments were carried out using the Lemur toolkit[2].

3.1.1 Parameter Setting

The number of feedback documents, the number of feedback terms, the feedback coefficient and the parameter that controls the generally of mutual information scores (parameter r) are set using 2-fold cross validation over each collection. We sweep the number of feedback documents and feedback terms between $\{10, 25, 50, 75, 100\}$, the feedback coefficient between $\{0, 0.1, \cdots, 1\}$, and the parameter r between $\{2, 4, 6, 8, 10\}$.

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 performance are determined using the two-tailed paired t-test computed at a 95% confidence level over average precision per query.

To evaluate the robustness of methods, we consider the robustness index (RI) [3] which is defined as $\frac{N_+ - N_-}{N}$, where N denotes the number of queries and N_+/N_- shows the number of queries improved/decreased by the feedback method.[3] The RI value is always in the $[-1, 1]$ interval and the method with higher value is more robust.

3.2 Results and Discussion

In this subsection, we first evaluate the proposed modifications to the log-logistic model. We further study the sensitivity of the proposed method to the free parameters.

3.2.1 Evaluating the Modified Log-Logistic Model

We consider two baselines: (1) the document retrieval method without feedback (NoPRF), and (2) the original log-logistic feedback model (LL). Although several other PRF methods have already been proposed, since in this paper, we propose a modification of the log-logistic model, we do not compare the proposed method with other existing PRF models.

To study the effect of each constraint in the retrieval performance, we modify the log-logistic model based on each constraint, separately. LL+Sem, LL+Rel, and LL+Dis denote the modified log-logistic model based on the "semantic effect", the "relevance effect", and the "distribution effect" constraints, respectively. We also modify the log-logistic model by considering all of these constraints (called LL+All) as introduced in Equation (5). The results obtained by the baselines and those achieved by the proposed modifications are reported in Table 2. According to this table, LL outperforms the NoPRF baseline in all cases, which shows the effectiveness of the log-logistic model. The improvements on the WT10g collection is lower than those on the AP and the Robust collections. This observation demonstrates that the

[2]http://lemurproject.org/

[3]To avoid the influence of very small average precision changes in the RI value, we only consider the improvements/losses higher than 10% (relatively).

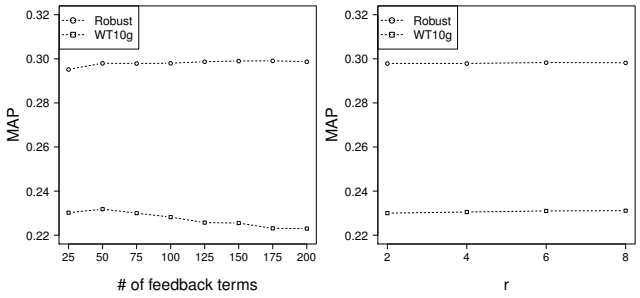

Figure 1: Sensitivity of the proposed method to the number of feedback terms and the parameter r.

log-logistic model is less effective and robust in improving the retrieval performance in the web collection, compared to the newswire collections. LL+Sem and LL+Rel perform better than LL in terms of MAP and P@10, in all collections. The MAP improvements are statistically significant in many cases, especially in the LL+Rel method. Except in one case (i.e., LL+Sem in Robust), both LL+Sem and LL+Rel models are shown to be more robust than the LL baseline. It is worth noting that we use very simple modifications to satisfy these two constraints, and thus using more accurate methods to satisfy these constraints can potentially improve the performance. LL+Dis method in general performs comparable to or sometimes slightly better than the LL baseline. The results achieved on the WT10g collection shows that LL+Dis can be more effective in noisy conditions, such as web collections. Overall, although the theoretical analysis shows that PRF methods should satisfy the "distribution effect" constraint, it does not substantially affect the retrieval performance in the AP and the Robust collections. The reason is that the values of $\frac{|D|}{ut(D)}$ (see Equation (3)) are very close to each other for different documents, especially in newswire collections. Thus, our modification to the log-logistic regarding the "distribution effect" constraint cannot substantially affect the retrieval performance.

As shown in Table 2, the LL+All method, which is our final modification to the log-logistic model, outperforms both baselines in all collections in terms of MAP and P@10. The MAP improvements are always statistically significant. The LL+All method is also shown to be more robust than the LL method, in particular in the WT10g collection.

3.2.2 Parameter Sensitivity

In this set of experiments, we fix one of the parameters r (the generality control parameter for mutual information) and n (the number of feedback terms), and then sweep the other one to show the sensitivity of the method to the input parameters. The results are reported in Figure 1.[4] According to this figure, the method is quite stable w.r.t. the changes in the values of these two parameters, especially for the parameter r. The results also indicate that by increasing the number of feedback terms, performance in the Robust collection generally increases, but in the WT10g collection it is not the case. The reason could be related to the noisy nature of this collection compared to the newswire collections.

[4]For the sake of visualization, we only report the results for the Robust and the WT10g collections. The behaviour of the method in AP is similar to the Robust collection.

4. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed two new constraints for pseudo-relevance feedback models. The first constraint considers semantic similarity of the feedback terms to the initial query. The second constraint focuses on the effect of distribution of all terms in the feedback documents on each term. We further studied the log-logistic model, a state-of-the-art feedback model, and showed that this model does not satisfy the proposed constraints as well as the previously proposed "relevance effect" constraint [8]. We then modified the log-logistic model to satisfy all of these constraints. The proposed modification was evaluated using three TREC newswire and web collections. Experimental results suggest that the modified model significantly outperforms the original log-logistic model, in all collections.

An interesting future direction is to study other feedback methods, such as the language model-based feedback methods, and modify them in order to satisfy the constraints. In this paper, we only consider simple approaches to satisfy the constraints, such as using mutual information for capturing semantic similarities. Future work can focus on more complex and accurate approaches to improve the retrieval performance.

5. ACKNOWLEDGEMENTS

This work was supported in part by the Center for Intelligent Information Retrieval. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

6. REFERENCES

[1] S. Clinchant and E. Gaussier. Information-based Models for Ad Hoc IR. In *SIGIR '10*, pages 234–241, 2010.

[2] S. Clinchant and E. Gaussier. A Theoretical Analysis of Pseudo-Relevance Feedback Models. In *ICTIR '13*, pages 6–13, 2013.

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

[4] H. Fang and C. Zhai. Semantic Term Matching in Axiomatic Approaches to Information Retrieval. In *SIGIR '06*, pages 115–122, 2006.

[5] V. Lavrenko and W. B. Croft. Relevance Based Language Models. In *SIGIR '01*, pages 120–127, 2001.

[6] Y. Lv and C. Zhai. Revisiting the Divergence Minimization Feedback Model. In *CIKM '14*, pages 1863–1866, 2014.

[7] J. H. Paik. A Novel TF-IDF Weighting Scheme for Effective Ranking. In *SIGIR '13*, pages 343–352, 2013.

[8] D. Pal, M. Mitra, and S. Bhattacharya. Improving Pseudo Relevance Feedback in the Divergence from Randomness Model. In *ICTIR '15*, pages 325–328, 2015.

[9] J. Seo and W. B. Croft. Geometric Representations for Multiple Documents. In *SIGIR '10*, pages 251–258, 2010.

[10] C. Zhai and J. Lafferty. Model-based Feedback in the Language Modeling Approach to Information Retrieval. In *CIKM '01*, pages 403–410, 2001.