Iterative Estimation of Document Relevance Score for Pseudo-Relevance Feedback

Mozhdeh Ariannezhad[1(✉)], Ali Montazeralghaem[1], Hamed Zamani[2], and Azadeh Shakery[1]

[1] School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran
{m.ariannezhad,ali.montazer,shakery}@ut.ac.ir
[2] Center for Intelligent Information Retrieval, College of Information and Computer Sciences, University of Massachusetts Amherst, Amherst, MA, USA
zamani@cs.umass.edu

Abstract. Pseudo-relevance feedback (PRF) is an effective technique for improving the retrieval performance through updating the query model using the top retrieved documents. Previous work shows that estimating the effectiveness of feedback documents can substantially affect the PRF performance. Following the recent studies on theoretical analysis of PRF models, in this paper, we introduce a new constraint which states that the documents containing more informative terms for PRF should have higher relevance scores. Furthermore, we provide a general iterative algorithm that can be applied to any PRF model to ensure the satisfaction of the proposed constraint. In this regard, the algorithm computes the feedback weight of terms and the relevance score of feedback documents, simultaneously. To study the effectiveness of the proposed algorithm, we modify the log-logistic feedback model, a state-of-the-art PRF model, as a case study. Our experiments on three TREC collections demonstrate that the modified log-logistic significantly outperforms competitive baselines, with up to 12% MAP improvement over the original log-logistic model.

Keywords: Pseudo-relevance feedback · Document effectiveness · Axiomatic analysis · Query expansion

1 Introduction

Search queries are usually too short to precisely express the underlying information need, which leads to poor retrieval performance. To address this problem, pseudo-relevance feedback (PRF) technique updates the query model using the top retrieved documents that are assumed to be relevant to the initial query. PRF has been shown to be highly effective in improving the retrieval performance [2,7,8,11,12]. In order to theoretically analyze PRF models, previous work [2,8,9] has proposed various constraints (axioms) that they should satisfy.

J.M. Jose et al. (Eds.): ECIR 2017, LNCS 10193, pp. 676–683, 2017.
DOI: 10.1007/978-3-319-56608-5_65

To satisfy the PRF constraints and thus to improve the accuracy of PRF models, different modifications have been suggested for well-established PRF models, such as mixture model [12] and geometric relevance model [10].

Pal et al. [9] proposed the "relevance effect" constraint as follows: the terms in the feedback documents with higher relevance scores should get higher weights in the feedback model. To satisfy this constraint, they used the initial relevance score of documents as their weight in the feedback model, similar to relevance models [7]. On the other hand, Keikha et al. [5] showed that the initial retrieval score of a document is not a good indicator for its effectiveness in the feedback model. They proposed a supervised algorithm to predict the document effectiveness for this task. In this paper, we argue that the relevance score of feedback documents can be better estimated using the feedback weights of the terms they contain. The intuition is that a document is more useful for PRF if it contains more informative terms for PRF. To this end, we propose the "feedback weight effect" constraint that implies the documents containing terms with higher weights in the feedback model should have higher relevance scores. State-of-the-art PRF models, such as relevance model [7], mixture model [12], matrix factorization-based model [11], and log-logistic feedback model [1], do not satisfy this constraint. In order to satisfy the introduced constraint, we propose a *general* iterative unsupervised algorithm that can be applied to any PRF model. In each iteration, the algorithm alternates between two steps: (1) computing the relevance scores of documents based on the feedback weights of their terms, and (2) computing the feedback weights of the terms with regard to the relevance scores of the documents they appear in.

To study the effectiveness of the proposed algorithm, we modify the log-logistic feedback model [1] to satisfy the feedback weight effect constraint using our iterative algorithm. Log-logistic model is a state-of-the-art PRF model that was previously shown to satisfy many PRF constraints and outperform competitive baselines, including geometric relevance model [10] and mixture model [2]. The experiments on three TREC collections demonstrate that our modification significantly outperforms the baselines.

2 Methodology

In this section, we introduce the "feedback weight effect" constraint and propose an iterative reinforcement algorithm to simultaneously compute the feedback weights of terms and the relevance scores of feedback documents. We use the notation previously used in [2,8]. $FW(w, F)$ and $RS(d, q)$ denote the feedback weight of term w in the feedback set F and the relevance score of document d for a given query q, respectively. $TF(w, d)$ denotes the frequency of term w in document d and $IDF(w)$ represents the inverse document frequency of term w.

2.1 PRF Constraints for Relevance Score of Feedback Documents

Pal et al. [9] introduced the relevance effect constraint for PRF models as follows:

Relevance effect: If a term w occurs in two documents $d_1, d_2 \in F$, and $RS(d_1, q) > RS(d_2, q)$, then: $FW(w, F \setminus \{d_1\}) < FW(w, F \setminus \{d_2\})$.

The relevance effect constraint indicates that the terms in the feedback documents with higher relevance scores should have higher weights in the feedback model compared to those in the documents with lower relevance scores. An important issue here is how to compute the relevance score? Pal et al. [9] followed the idea behind the relevance models [7] and used the initial retrieval score of feedback documents (e.g., the query likelihood score) as their relevance score. On the other hand, Keikha et al. [5] showed that the initial retrieval score is not an optimal indicator of document effectiveness for query expansion. Based on their observations, we provide a theoretical axiom for estimating the effectiveness of documents for feedback. We argue that the relevance score of feedback documents should depend on the feedback weights of the terms they contain. Since the feedback weight of a term demonstrates the usefulness of the term for PRF, a document that contains more informative terms is more useful for PRF. As a result, such a document should have a higher relevance score. In this regard, we define the feedback weight effect constraint as follows:

Feedback weight effect: If $d \in F$ and w_1 and w_2 are two feedback terms where $TF(w_1, d) = TF(w_2, d) \geq 1$, $IDF(w_1) = IDF(w_2)$ and $FW(w_1, F) > FW(w_2, F)$, then: $RS(d \setminus \{w_1\}, q) < RS(d \setminus \{w_2\}, q)$.

Note that the feedback weight effect is a constraint for the relevance score and can be satisfied regardless of whether the PRF model enforces the relevance effect or not.

2.2 Relevance Score Estimation via an Iterative Reinforcement Model

To satisfy the aforementioned constraints, we provide an iterative approach that simultaneously computes the feedback weight of terms and the relevance score of feedback documents. The relevance effect states that a term should have a high feedback weight if it appears in many feedback documents with high relevance scores, and the feedback weight effect implies that a feedback document should have a high relevance score if it contains many terms with high feedback weights. In other words, the feedback weight of a term is determined by the relevance score of the feedback documents it appears in, and the relevance score of a feedback document is determined by the feedback weights of the terms it contains. For simplicity, we respectively use $FW(w)$ and $RS(d)$ instead of $FW(w, F)$ and $RS(d, q)$, in the equations. The following steps are alternated until convergence, with a uniform initialization for the document and term scores:

1. Computing feedback term weights:

$$\forall w \in V_F : FW(w)^{(n)} = Com(w) \sum_{d \in F} TW(w, d, q) RS(d)^{(n-1)}, \qquad (1)$$

2. Computing document relevance scores:

$$\forall d \in F : RS(d)^{(n)} = \frac{1}{|d|} \sum_{w \in d} TW(w, d, q) FW(w)^{(n-1)}. \quad (2)$$

In the above equations, $TW(w, d, q)$ is a term weighting function that demonstrates the importance of term w in document d with respect to the query q, $|d|$ denotes the length of document d, and V_F represents the set of feedback terms. $FW(w)^{(n)}$ and $RS(d)^{(n)}$ respectively denote the feedback term weight and the document relevance score computed in the n^{th} iteration. In the first equation, $Com(w) = \frac{|F_w|}{|F|}$ ($|F_w|$ denotes the number of feedback documents that contain w) shows how common w is in the feedback documents. $Com(w)$ was previously used in [5] and leads to satisfying the DF effect constraint [2]. Note that in each iteration, the feedback weights and the relevance scores should be normalized subject to $\sum_{w \in V_F} FW(w)^{(n)} = 1$ and $\sum_{d \in F} RS(d)^{(n)} = 1$. The proposed algorithm differs from the one introduced in [4], in that it does not calculate the relevance scores for the feedback documents.

Similar ideas regarding iterative computation of related variables have been used in different tasks, such as in the HITS algorithm [6]. The convergence of our algorithm can be proven, similar to the proof presented in [6].

2.3 Case Study: Log-Logistic Feedback Model

As mentioned above, the proposed constraint and algorithm are general and independent of the feedback model. In this paper, we consider the log-logistic feedback model [1], a state-of-the-art PRF model. Clinchant and Gaussier [2] showed that the log-logistic model satisfies their PRF constraints and outperforms many feedback models, including the geometric relevance model [10] and the mixture model [12]. The log-logistic model calculates the feedback weight of a term w in a document d, as follows:

$$TW_{LL}(w, d, q) = \log\left(\frac{t(w, d) + \lambda_w}{\lambda_w}\right), \quad (3)$$

where $\lambda_w = \frac{N_w}{N}$ (N_w is the number of documents in the collection that contain w and N is the total number of documents in the collection), and $t(w, d)$ is the normalized term frequency component defined as: $t(w, d) = TF(t, d) \log(1 + c\frac{avg_l}{|d|})$, where avg_l denotes the average document length and c is a free hyper-parameter. It is shown that log-logistic satisfies TF, IDF, DF, concavity, and document length constraints [2]. Recently, Montazeralghaem et al. [8] modified the log-logistic model as follows, in order to satisfy the relevance effect constraint:

$$TW_{LLR}(w, d, q) = RS_{init}(q, d) \times TW_{LL}(w, d, q), \quad (4)$$

where $RS_{init}(q, d)$ denotes the initial retrieval score of document d. Both the original log-logistic feedback model and the modified version in [8] use the mean of $TW(w, d, q)$ over all the feedback documents as $FW(w, F)$.

We use the term weight definition provided in Eq. (4) in our iterative algorithm proposed in Sect. 2.2. This method is referred to as *LLIR*. In order to prove that LLIR satisfies the proposed axiom, we consider two terms w_1 and w_2 and a feedback document d. When $TF(w_1, d) = TF(w_2, d) \geq 1$ and $IDF(w_1) = IDF(w_2)$, it can be shown that $TW_{LLR}(w_1, d, q) = TW_{LLR}(w_2, d, q)$. Considering the case where $FW(w_1, F) > FW(w_2, F)$, it is obvious that $TW_{LLR}(w_1, d, q)FW(w_1, F) > TW_{LLR}(w_2, d, q)FW(w_2, F)$, which implies $RS(d \setminus \{w_1\}, q) < RS(d \setminus \{w_2\}, q)$, if we use Eq. 2 to compute the score of feedback documents.

3 Experiments

Collections. In our experiments, we used three standard TREC collections whose statistics are provided in Table 1. AP and Robust are newswire collections, whereas WT10g is a Web collection containing more noisy documents.

Experimental Setup. We used the titles of TREC topics as queries. All indexes and topics were stopped using the standard INQUERY stopword list and stemmed using the Porter stemmer. All experiments were carried out using the Lemur toolkit[1]. Initial retrieval results were obtained using the query likelihood model with Dirichlet prior smoothing ($\mu = 1000$).

Parameter Setting. The number of feedback documents, the number of feedback terms, the feedback coefficient, and the parameter c are set using 2-fold cross validation over the queries of each collection, for all methods. We sweeped the number of feedback documents between $\{10, 25, 50, 75, 100\}$ and the number of feedback terms between $\{10, 50, 100, 150, 200\}$. We changed the feedback coefficient from 0 to 1 in the increment of 0.1, and the parameter c from 1 to 10 in the increment of 1.

Evaluation Metrics. We use three metrics to measure the retrieval quality: (1) mean average precision (MAP) of the top-ranked 1000 documents, (2) the precision of the top 10 retrieved documents (P@10), and (3) the robustness index (RI) [3]. Statistically significant differences of performance are determined using the two-tailed paired t-test at a 95% confidence level.

3.1 Results and Discussion

We consider three baselines: (1) the document retrieval model without pseudo-relevance feedback (NoPRF), (2) the original log-logistic feedback model (LL) [1], and (3) the enhanced log-logistic model (LLR) which satisfies the relevance effect, proposed in [8]. Furthermore, we also report the results achieved by the proposed method both after one iteration (LLIR-1-iteration) and after convergence (LLIR-converged).

[1] http://lemurproject.org/.

Table 1. Summary of TREC collections and topics.

ID	Collection	Queries	#docs
AP	TREC 1-3 Ad-hoc track, Associated press 88–89	Topics 51–200	165k
Robust	TREC 2004 Robust track collection	Topics 301–450 & 601–700	528k
WT10g	TREC 9-10 Web track collection	Topics 451–550	1692k

Table 2. Retrieval effectiveness of the iterative model compared to the baselines. Superscripts 0/1/2/3 indicate that the improvements over NoPRF/LL/LLR/LLIR-1-iteration are significant.

Method	AP			Robust			WT10g		
	MAP	P@10	RI	MAP	P@10	RI	MAP	P@10	RI
NoPRF	0.2663	0.4309	–	0.2490	0.4237	–	0.2080	0.3030	–
LL	0.3300^{0}	0.4691	0.44	0.2798^{0}	0.4394	0.29	0.2089	0.3071	0.08
LLR	0.3381^{01}	0.4624	**0.47**	0.2822^{0}	**0.4450**	0.29	0.2230^{01}	0.3101	0.17
LLIR-1-iteration	0.3406^{01}	0.4698	0.42	0.2876^{012}	0.4365	0.28	0.2219^{01}	0.3101	0.17
LLIR-converged	$\mathbf{0.3507}^{0123}$	**0.4765**	0.45	$\mathbf{0.2926}^{0123}$	0.4442	**0.31**	$\mathbf{0.2344}^{0123}$	**0.3121**	**0.21**

Table 2 summarizes the results achieved by the proposed method and the baselines. As shown in the table, LL performs significantly better than NoPRF on all the collections, which shows the effectiveness of the log-logistic model. LLIR-converged outperforms NoPRF and LL on all collections, indicating the importance of the proposed constraint for PRF. The significant improvements achieved by LLIR (after convergence) over LLR show that the document scores estimated by our model are more effective than the initial retrieval scores which are used by LLR. According to Table 2, the LLIR results after convergence are significantly higher than those obtained after the first iteration. It is worth mentioning that LLIR converges after 8 to 10 iterations, indicating its efficiency and low computational cost. The performance of LLIR-converged in terms of P@10 and RI is also superior to the baselines, except in two cases (RI on AP and P@10 on Robust) where the results are comparable to the highest values. In general, the results show that our method have impressive overall ranking performance, e.g., the MAP value achieved by LLIR-converged is up to 12% higher than those obtained by the original log-logistic model (LL).

Figure 1 plots the sensitivity of the proposed method with respect to the number of feedback terms added to the query. According to this figure, after 50 terms, the performance becomes stable in the newswire collections (AP and Robust), while by increasing the number of terms, we lose the performance in the WT10g collection. To have an insight into the term weights computed by the proposed method, Table 3 reports the top 10 terms for the query "gulf war syndrom" in the Robust collection, computed by the LLR and the proposed method. As shown in the table, the order of the terms have changed and also

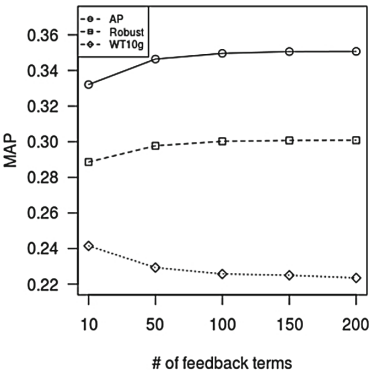

Fig. 1. Sensitivity of the LLIR method to the number of feedback terms.

Table 3. The top terms added to the query "gulf war syndrom" (topic 630) by LLR and LLIR methods.

LLR		LLIR	
Syndrom	0.1862	Syndrom	0.2452
Gulf	0.1507	Gulf	0.2015
War	0.1126	War	0.1317
Veteran	0.0932	Veteran	0.0754
Vietnam	0.0867	Defenc	0.0730
Desert	0.0804	Militari	0.0694
Defenc	0.0757	Desert	0.0570
Soldier	0.0735	Serv	0.0501
Militari	0.0734	Time	0.0494
Diarrhoea	0.0677	American	0.0473

the term "vietnam" which is irrelevant to the initial query does not appear in the list of top terms estimated by LLIR, while a relevant term ("american") is added to the list. For this query, the average precision achieved by LLIR is 0.6034 which is much higher than the one obtained by LLR (i.e., 0.2867).

4 Conclusions

In this paper, we introduced a new constraint concerning the relevance score of the feedback documents. The constraint states that the documents containing more informative terms for PRF should have higher relevance scores. We further proposed a general iterative algorithm that can be applied to any PRF model in order to guarantee the satisfaction of the proposed constraint. We applied our algorithm to the log-logistic feedback model as a case study. Our experiments on three TREC collections showed that the proposed modification significantly improves the results.

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.

References

1. Clinchant, S., Gaussier, E.: Information-based models for ad hoc IR. In: SIGIR (2010)
2. Clinchant, S., Gaussier, E.: A theoretical analysis of pseudo-relevance feedback models. In: ICTIR (2013)

3. Collins-Thompson, K.: Reducing the risk of query expansion via robust constrained optimization. In: CIKM (2009)
4. Dehghani, M., Azarbonyad, H., Kamps, J., Hiemstra, D., Marx, M.: Luhn revisited: significant words language models. In: CIKM (2016)
5. Keikha, M., Seo, J., Croft, W.B., Crestani, F.: Predicting document effectiveness in pseudo relevance feedback. In: CIKM (2011)
6. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. J. ACM **46**(5), 604–632 (1999)
7. Lavrenko, V., Croft, W.B.: Relevance based language models. In: SIGIR (2001)
8. Montazeralghaem, A., Zamani, H., Shakery, A.: Axiomatic analysis for improving the log-logistic feedback model. In: SIGIR (2016)
9. Pal, D., Mitra, M., Bhattacharya, S.: Improving pseudo relevance feedback in the divergence from randomness model. In: ICTIR (2015)
10. Seo, J., Croft, W.B.: Geometric representations for multiple documents. In: SIGIR (2010)
11. Zamani, H., Dadashkarimi, J., Shakery, A., Croft, W.B.: Pseudo-relevance feedback based on matrix factorization. In: CIKM (2016)
12. Zhai, C., Lafferty, J.: Model-based feedback in the language modeling approach to information retrieval. In: CIKM (2001)