Is document frequency important for PRF?

Stéphane Clinchant, Éric Gaussier

Is Document Frequency important for PRF?

Stéphane Clinchant[1,2] and Eric Gaussier[2]

[1] Xerox Research Center Europe, Meylan, France
[2] LIG Université de Grenoble, UMR 5217/AMA team
stephane.clinchant@xrce.xerox.com, eric.gaussier@imag.com

Abstract. We introduce in this paper a new heuristic constraint for PRF models, referred to as the *Document Frequency (DF) constraint*, which is validated through a series of experiments with an oracle. We then analyze, from a theoretical point of view, state-of-the-art PRF models according to their relation with this constraint. This analysis reveals that the standard mixture model for PRF in the language modeling family does not satisfy the DF constraint on the contrary to several recently proposed models. Lastly, we perform tests with a simple family of *tf-idf* functions based on a parameter controlling the satisfaction of the constraint. This last series of experiments further validate the DF constraint.

1 Introduction

Pseudo-relevance feedback (PRF) has been studied for several decades, and a lot of different models have been proposed, in all the main families of information retrieval (IR) models. In the language modelling approach to IR, for example, the mixture model for PRF is considered state-of-the-art, and numerous studies use it as a baseline. It has indeed been shown to be one of the most effective models in terms of performance and stability wrt parameter values in [11]. However, several recently proposed PRF models seem to outperform this mixture model, as models based on bagging, models based on a mixture of Dirichlet compound multinomial distributions, geometric relevance models or the log-logistic models of the recent information-based family [4, 14, 2, 13]. This paper *aims at providing an explanation* of such improvements. In a nutshell, many of the recent models tends to favor terms with a high document frequency in the feedback set, a behavior we will capture with the Document Frequency constraint.

The notations we use throughout the paper are summarized in table 1, where w represents a term. We note n the number of pseudo relevant document used, F the feedback set and tc the number of term for pseudo relevance feedback. We call FTF, the feedback set term frequency and FDF, the feedback document frequency. The remainder of the paper is organised as follows. We give in Section 2 some basic statistics on three PRF models, which reveal global trends of PRF models. We then introduce in section 3 the Document Frequency constraint, that PRF models should satisfy, prior to reviewing standard PRF models according to their behavior wrt this constraint in section 4. We then introduce in section 5

Notation	Description
General	
q, d	Original query, document
$RSV(q,d)$	Retrieval status value of d for q
$c(w,d)$	# of occurrences of w in doc d
l_d	Length of doc d
avg_l	Average document length in collection
N	# of docs in collection
N_w	# of documents containing w
$\text{IDF}(w)$	$-\log(N_w/N)$
$tdfr(w,d)$	$c(w,d)\log(1+c\frac{\text{avg}_l}{l_d})$
PRF specific	
n	# of docs retained for PRF
$\mathbf{F}$	Set of documents retained for PRF: $\mathbf{F}=(d_1,\ldots,d_n)$
tc	*TermCount*: # of terms in $\mathbf{F}$ added to query
$FTF(w)$	$=\sum_{d\in F} c(w,d)$
$FDF(w)$	$=\sum_{d\in F} I(c(w,d)>0)$

Table 1: Notations

a simple family of feedback functions which allows us to better understand the relations between the different constraints, prior to discuss some related work in section 6.

2 Some Statistics on PRF

We begin this paper by analyzing the terms chosen and the performance obtained by three different, state-of-the-art, pseudo-relevance feedback (PRF hereafter) methods, namely the mixture model and the divergence minimization method in the language modeling family [15], and the mean log-logistic information model in the information-based family [2]. These models are reviewed later in section 4, and their exact formulation is not necessary here. In order to have an unbiased comparison, we use the same IR engine for the retrieval step. Thus, all PRF algorithms are computed on the *same* set of documents. Once new queries are constructed, we use either the Dirichlet language model (for the new queries obtained with the mixture model and the divergence minimization method) or the log-logistic model (for the new queries obtained with the mean log-logistic information model) for the second retrieval step, thus allowing one to compare the performance obtained by different methods on the same initial set of PRF documents. Two collections are used throughout this study: the ROBUST collection, with 250 queries, and the TREC 1&2 collection, with topics 51 to 200. Only query titles were used and all documents were preprocessed with standard Porter stemming, and all model parameters are optimized through a line search on the whole collection. The results obtained are thus the best possible results

Table 2: Statistics of the size of the Intersection

Collection	n	tc	Mean	Median	Std
robust	10	10	5.58	6.0	1.60
trec-12	10	10	5.29	5.0	1.74
robust	20	20	12	12	3.05
trec-12	20	20	11.8	13	3.14

one can get with these models on the retained collections. We first focus on a direct comparison between the mixture model and the mean log-logistic information model, by comparing the terms common to both feedback methods, i.e. the terms in the intersection of the two selected sets. Table 2 displays the mean, median and standard deviation of the size of the intersection, over all queries, for the collections considered. As one can note, the two methods agree on a little more than half of the terms (ratio mean by tc), showing that the two models select different terms. To have a closer look at the terms selected by both methods, we first compute, for each query, the total frequency of a word in the feedback set (i.e. $FTF(w)$) and the document frequency of this word in the feedback set (i.e. $FDF(w)$). Then, for each query we can compute the mean frequency of the selected terms in the feedback set as well as its mean document frequency, i.e. $q(ftf)$ and $q(fdf)$:

$$q(ftf) = \sum_{i=1}^{tc} \frac{ftf(w_i)}{tc} \text{ and } q(fdf) = \sum_{i=1}^{tc} \frac{fdf(w_i)}{tc}$$

We then compute the mean of the quantities over all queries.

$$\mu(ftf) = \sum_{q} \frac{q(ftf)}{|Q|} \text{ and } \mu(fdf) = \sum_{q} \frac{q(fdf)}{|Q|}$$

An average IDF can be computed in exactly the same way, where IDF is the standard inverse document frequency *in the collection*. Table 3 displays the above statistics for the three feedback methods: mixture model (MIX), mean log-logistic(LL) information model and divergence minimization model (DIV). Regarding the mixture and log-logistic models, on all collections, the mixture model chooses in average words that have a *higher FTF*, and a smaller *FDF*. The mixture model alos chooses words that are *more frequent in the collection* since the mean IDF values are smaller. On the other hand, the statistics of the divergence model shows that this model extracts very common terms, with low IDF and high FDF, which is one of the main drawback of this model. In addition to the term statistics, the performance of each PRF algorithm can also be assessed. To do so, we first examine the performance of the feedback terms *without* mixing them with the original queries, a setting we refer to as *raw*. Then, for each query we keep only terms that belong to the intersection of the mixture and log-logistic models (as the divergence model is a variant of the mixture model,

Table 3: Statistics of terms extracted by. Suffix A means $n = 10$ and $tc = 10$ while suffix B means $n = 20$ and $tc = 20$

Settings	Statistics	MIX	LL	DIV
robust-A	$\mu(ftf)$	62.9	46.7	57.9
	$\mu(fdf)$	6.4	7.21	8.41
	Mean IDF	4.33	5.095	2.36
trec-1&2-A	$\mu(ftf)$	114 .0	79.12	98.76
	$\mu(fdf)$	7.1	7.8	8.49
	Mean IDF	3.84	4.82	2.5
robust-B	$\mu(ftf)$	68.6	59.9	68.2
	$\mu(fdf)$	9.9	11.9	14.4
	Mean IDF	4.36	4.37	1.7
trec-1&2-B	$\mu(ftf)$	137.8	100.0	118.45
	$\mu(fdf)$	12.0	13.43	14.33
	Mean IDF	3.82	4.29	2.0

we do not consider it in itself for this intersection), but keep their weight predicted by each feedback method. We call this setting *interse*. A third setting, *diff*, consists in keeping terms which do not belong to the intersection. Finally, the last setting, *interpo* for interpolation, measures the performance when new terms are mixed with the original query. This corresponds to the standard setting of pseudo-relevance feedback. Table 4 displays the results obtained. As one can note, the log-logistic model performs better than the mixture model, as found in [2]. What our analysis reveals is that it does so because it chooses better feedback terms, as shown by the performance of the *diff* setting. For the terms in the intersection, method *interse*, the weights assigned by the log-logistic model seem more appropriate than the weights assigned by the other feedback models.

Table 4: MAP (%) Performance of different methods. Suffix A means $n = 10$ and $tc = 10$ while suffix B means $n = 20$ and $tc = 20$

	robust-A			trec-1&2			robust-B			trec-1&2-B		
FB Model	MIX	LL	DIV	MIX	LL	DIV	MIX	LL	DIV	MIX	LL	DIV
raw	23.8	26.9	24.3	23.6	25.7	24.1	23.7	25.7	22.8	25.1	27.0	24.9
interse	24.6	25.7	24.	24.2	24.5	23.4	25.3	26.2	22.6	26.1	26.5	24.7
diff	3	11.0	0.9	3	9	0.9	3.0	10.0	0.15	2.1	11.2	0.5
interpo	28.0	29.2	26.3	26.3	28.4	25.4	28.2	28.5	25.9	27.3	29.4	25.7

Let's summarize our finding here. (a) The log-logistic model performs better than the mixture and divergence models for PRF. (b) The mixture and divergence models choose terms with a *higher* FTF. (c) The mixture model selects term with a smaller FDF, whereas (d) the divergence model selects terms with a smaller IDF. A first explanation of the better behavior of the log-logistic model can be that the FDF and IDF effect are dealt with more efficiently in this model, as shown by the statistics reported in table 3.

3 The Document Frequency Constraint

We adopt the axiomatic approach to IR [7] in order to present the Document Frequency constraint. Axiomatic methods were pioneered by Fang et al [7] and followed by many works. In a nutshell, axiomatic methods describe IR functions by constraints they should satisfy. According to [2], the four main constraints for an IR function to be valid are: the weighting function should (a) be increasing and (b) concave wrt term frequencies, (c) have an IDF effect and (d) penalize long documents. We first want to briefly discuss whether these constraints would make sense for PRF models.

In the context of PRF, the first two constraints relate to the fact that terms frequent in the feedback set are more likely to be effective for feedback, but that the difference in frequencies should be less important in high frequency ranges. The IDF effect is also relevant in feedback, as one generally avoids selecting terms with a low IDF, as such terms are scored poorly by IR systems. The constraint on document length is not as clear as the others in the context of PRF, as one (generally) considers sets of documents. What seems important however is the fact that occurrence counts are normalized by the length of the documents they appear in, in order not to privilege terms which occur in long documents.

Let $FW(w; \mathbf{F}, \mathbf{P}_w)$ denote the feedback weight for term w, with $\mathbf{P}_w$ a set of parameters dependent on w.[3]. We now introduce a new PRF constraint which is based on the results reported in the previous section. Indeed, as we have seen, the best PRF results were obtained with models which favor feedback terms with a high *document frequency* ($FDF(w)$) in the feedback set, which suggests that, *all things being equal*, terms with a higher FDF should receive a higher score. This constraint can be formalized as follows:

PRF Constraint 1 [Document Frequency - DF]
Let $\epsilon > 0$, and w_a and w_b two words such that:
(i) IDF(a) = IDF(b)
(ii) The distribution of the frequencies of w_a and w_b in the feedback set are given by:

$$T(w_a) = (x_1, x_2, ..., x_j, 0, ..., 0)$$
$$T(w_b) = (x_1, x_2, ..., x_j - \epsilon, \epsilon, ..., 0)$$

with $\forall i$, $x_i > 0$ and $x_j - \epsilon > 0$ (hence, $FTF(w_a) = FTF(w_b)$ and $FDF(w_b) = FDF(w_a) + 1$).
Then: $FW(w_a; \mathbf{F}, \mathbf{P}_{w_a}) < FW(w_b; \mathbf{F}, \mathbf{P}_{w_b})$

In other words, FW is *locally* increasing with $FDF(w)$. The above constraint is sometimes difficult to check. The following theorem is useful to establish whether a PRF model, which can be decomposed in the documents of $\mathbf{F}$, satisfies or not the DF constraint:

[3] The definition of $\mathbf{P}_w$ depends on the PRF model considered. It minimally contains $FTF(w)$, but other elements, as IDF(w), are also usually present. We use here this notation for convenience.

Theorem 1. *Suppose FW can be written as:*

$$FW(w; \mathbf{F}, \mathbf{P}_w) = \sum_{d=1}^{n} f(x_w^d; \mathbf{P'}_w) \quad (1)$$

with $\mathbf{P'}_w = \mathbf{P}_w \setminus x_w^d$ *and* $f(0; \mathbf{P'}_w) \geq 0$. *Then:*

1. *If the function f is strictly concave, then FW meets the DF constraint.*
2. *If the function f is strictly convex, then FW does not meet the DF constraint.*

If f is strictly concave, then the function f is subadditive ($f(a+b) < f(a) + f(b)$). Let a and b be two words satisfying the conditions of the DF constraint. Then, we have:

$$FW(b) - FW(a) = f(x^j - \epsilon) + f(\epsilon) - f(x^j)$$

As the function f is subadditive, we have: $FW(b) - FW(a) > 0$. If f is strictly convex, then f is superadditive as $f(0) = 0$, and a comparable reasoning leads to $FW(b) - FW(a) < 0$. In the remainder of the paper, we will simply use the notation $FW(w)$ as a shorthand for $FW(w; \mathbf{F}, \mathbf{P}_w)$.

3.1 Validation of the DF Constraint

The DF constraint states that, all other parameters being equal, terms with higher DF should be preferred. Thus, in average, one should observe that terms with high DF scores yield larger increase in MAP values. To see whether this is the case, we computed the impact on the MAP of different terms selected from true relevance judgements, and plotted this impact against both TF and DF values. Our relying on true relevant documents and not documents obtained from pseudo-relevance feedback is based on (a) the fact that pseudo-relevance feedback aims at approximating relevance feedback, an (b) the fact that it is more difficult to observe clear trends in pseudo-relevance sets where the precision (e.g. P@10) and MAP of each query have large variances. The framework associated with true relevance judgements is thus cleaner and allows easier interpretation. In order to assess the impact of DF scores on the MAP values independently of any IR model, we make use of the following experimental setting:

- Start with a first retrieval with a Dirichlet language model;
- Let R_q denote the set of relevant documents for query q: Select the first 10 relevant documents if possible, else select the top $|R_q|(|R_q| < 10)$ relevant documents;
- Construct a new query (50 words) with the mixture model;
- Construct a new query (50 words) with the log-logistic model;
- Compute statistics for each word in the new queries.

Statistics include a normalized FDF, equal to $FDF(w)/|R_q|$, and a normalized FTF, first using a document length normalization, then using the transformation

$\log(1 + FTF(w))/|R_q|$ to avoid too important a dispersion in plots. Each word w is added independently with weights predicted by the retained PRF model. For each word w, we measure the MAP of the initial query augmented with this word. The difference in performance with the intial query is then computed as: $\Delta(\text{MAP}) = \text{MAP}(q+w) - \text{MAP}(q)$. We thus obtain, for each term, the following statistics: $\Delta(\text{MAP})$, $\log(1 + FTF(w))/|R_q|$, $FDF(w)/|R_q|$.

Figures 1 display a 3D view of these statistics for all queries, based on Gnuplot and two collections: TREC1&2 and ROBUST.

The TF statistics was normalized to account for different lengths and a Gaussian Kernel was used to smooth the data cloud. The shape of the plots obtained remains however consistent without any normalization and a different Kernel.

As one can note, on all plots of Figures 1, the best performing regions in the (TF,DF) space correspond to large DFs. Furthermore, for all TF values, the increase in MAP parallels the increase in DF (or, in other words, $\Delta(\text{MAP})$ increases with DF for fixed TF). This validates the DF constraint and shows the importance of retaining terms with high DF in relevance feedback. Interestingly, the reverse is not true for TF. This implies that if terms with large TF are interesting, they should not be given too much weight. The results displayed in Table 3 suggest that the mixture model [15] suffers from this problem.

4 Review of PRF Models

We review in this section different PRF models according to their behavior wrt the DF constraint we have defined. We start with language models, then review the recent model introduced in [14] which borrows from both generative approaches *à la* language model and approaches related to the *Probability Ranking Principle* (PRP), prior to review Divergence from Randomness (DFR) and Information-based models.

Mixture Model: Zhai and Lafferty [15] propose a generative model for the set **F**. All documents are i.i.d and each document is generated from a mixture of the feedback query model and the corpus language model:

$$P(\mathbf{F}|\theta_F, \beta, \lambda) = \prod_{w=1}^{V} ((1-\lambda)P(w|\theta_F) + \lambda P(w|C))^{FTF(w)} \quad (2)$$

where λ is a "background" noise set to some constant. For this model, $FW(w) = P(w|\theta_F)$ and θ_F is learned by optimising the data log-likelihood with an Expectation-Maximization (EM) algorithm. The above formula shows that the mixture multinomial model behaves as if all documents were merged together. As a result, the mixture model is agnostic wrt to DF, and thus does not satisfy the DF constraint.

Divergence Minimization: For language models, a divergence minimization model was also proposed in [15] and leads to the following feedback model:

$$FW(w) = P(w|\theta_F) \propto \exp\Big(\frac{1}{(1-\lambda)}\frac{1}{n}\sum_{i=1}^{n}\log(p(w|\theta_{d_i})) - \frac{\lambda}{1-\lambda}\log(p(w|C)\Big)$$

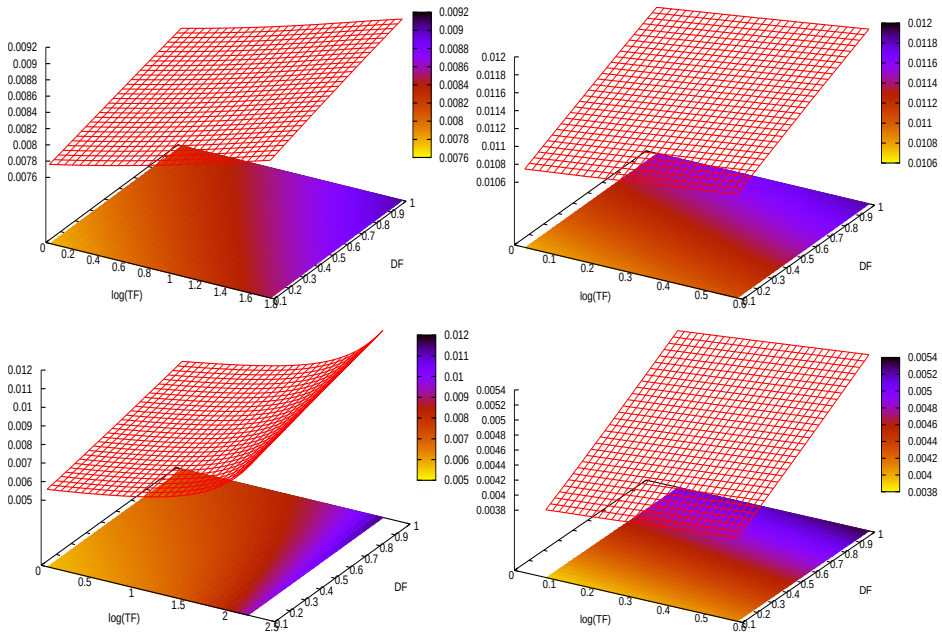

Fig. 1: (log(FTF),FDF) vs Δ MAP; true relevant documents are used with $n = 10$, $tc = 50$ and Gaussian kernel grids (30×30). Top row: log-logistic model; bottom row: mixture (language) model, left column:ROBUST Collection and right column: TREC-12 collection

Furthermore, this equation corresponds to the form given in equation 1 with a strictly concave function (log). Thus, by Theorem 1, this model satisfies the DF constraint. Despite this good theoretical behavior, our previous experiments, reported in Table 4, as well as those reported in [11], show that this model does not perform as well as other ones. Indeed, as shown in Table 3, the IDF effect is not sufficiently enforced, and the model fails to downweight common words.

Relevance Model: Another PRF model proposed in the framework of the language modeling approach is the so-called relevance model, proposed by Lavrenko *et al.* [8], and defined by:

$$FW(w) \propto \sum_{d \in \mathbf{F}} P_{LM}(w|\theta_d) P(d|q) \tag{3}$$

where P_{LM} denotes the standard language model. The above formulation corresponds to the form of equation 1 of Theorem 1, with a linear function, which is neither strictly concave nor strictly convex. This model is neutral wrt the DF constraint. The relevance model has recently been refined in the study presented in [13] through a geometric variant, referred to as GRM,

and defined by:

$$FW(w) \propto \prod_{d \in \mathbf{F}} P_{LM}(w|\theta_d)^{P(d|q)}$$

As the log is on a concave function, the GRM model satisfies the DF constraint according to Theorem 1.

EDCM: Xu and Akella [14] propose a mixture of eDCM distributions to model the pseudo relevance feedback set. Terms are then generated according to two latent generative models based on the (e)DCM distribution and associated with two variables, relevant z_{FR} and non-relevant z_N. The variable z_N is intended to capture general words occurring in the whole collection, whereas z_{FR} is used to represent relevant terms occurring in the feedback documents. Disregarding the non-relevant component for the moment, the weight assigned to feedback terms by the relevant component is given by (M-step of the EM algorithm):

$$P(w|z_{FR}) \propto \sum_{d \in \mathbf{F}} I(c(w,d) > 0) P(z_{FR}|d,w) + \lambda c(w,q)$$

This formula, being based on the presence/absence of terms in the feedback documents, is thus compatible with the DF constraint.

DFR Bo: Standard PRF models in the DFR family are Bo models [1]:

$$FW(w) = \text{Info}(w, \mathbf{F}) = \log_2(1 + g_w) + FTF(w) \log_2(\frac{1 + g_w}{g_w}) \qquad (4)$$

where $g_w = \frac{N_w}{N}$ in $Bo1$ model and $g_w = P(w|C)(\sum_{d \in F} l_d)$ in $Bo2$ model. In other words, documents in $\mathbf{F}$ are merged together and a geometric probability model is used to measure the informative content of a word. As this model is DF agnostic, it does not satisfy the DF constraint.

Log-logistic: In information-based models [2], the average information brought by the feedback documents on given term w is used as a criterion to rank terms, which amounts to:

$$FW(w) = \text{Info}(w, \mathbf{F}) = \frac{1}{n} \sum_{d \in \mathbf{F}} -\log P(X_w > tdfr(w,d)|\lambda_w)$$

where $tdfr(w,d)$ is given in table 1, and λ_w a parameter associated to w and set to: $\lambda_w = \frac{N_w}{N}$. Two instanciations of the general information-based family are considered in [2], respectively based on the log-logistic distribution and a smoothed power law (SPL). The log-logistic model for pseudo relevance feedback is thus defined by:

$$FW(w) = \frac{1}{n} \sum_{d \in F} [\log(\frac{N_w}{N} + tdfr(w,d)) + \text{IDF}(w)] \qquad (5)$$

It is straightforward to show that both the log-logistic and the SPL models lead to concave functions. So, according to Theorem 1, these models satisfies the DF constraint.

5 Well-founded, Simple PRF Reweighting

Let us introduce the family of feedback functions defined by:

$$FW(w) = \sum_{d \in F} tdfr(w,d)^k \text{IDF}(w) \tag{6}$$

with $tdfr$ is given in table 1 and corresponds to the normalization used e.g. in DFR and information-based models. This equation amounts to a standard *tf-idf* weighting, with an exponent k which allows one to control the convexity/concavity of the feedback model.

If $k > 1$ then the function is strictly convex and, according to Theorem 1, does not satisfy the DF constraint. On the contrary, if $k < 1$, then the function is strictly concave and satisfies the DF constraint. The linear case, being both concave and convex, is *in-between*.

One can then build PRF models from equation 6 with varying k, and see whether the results agree with the theoretical findings implied by Theorem 1. We used the reweighting scheme of equation 6 and a log-logistic model to assess their performance. The new query $q_w^{'}$ was updated as in DFR and information-based models:

$$q_w^{'} = \frac{q_w}{\max_w q_w} + \beta \frac{FW(w)}{\max_w FW(w)} \tag{7}$$

Figure 2 a) displays the term statistics ($\mu(ftf)$, $\mu(fdf)$, mean IDF) for different values of k. As one can note, the smaller k, the bigger $\mu(fdf)$ is. In other words, the slowlier the function grows, the more terms with large DF are preferred. Figure 2 b) displays the MAP for different values of k. At least two important points arise from the results obtained. First, convex functions ($k > 1$) have lower performance than concave functions for all datasets, and the more a model violates the constraints, the worse it is. This confirms the validity of the DF constraint. Second, the square root function ($k = 0.5$) has the best performance on all collections: it also outperforms the standard log-logistic model. When the function grows slowly (k equals to 0.2), the DF statistics is somehow preferred compared to TF. The square root function achieves a different and better trade-off between the TF and DF information. This is an interesting finding as it shows that the TF information is still useful and should not be too downweighted wrt the DF one.

6 Related Work

There are a certain number of additional elements that can be used in PRF settings. The document score hypothesis states that documents with a higher score (defined by $RSV(q,d)$) should be given more weight in the feedback function as in relevance models [8]. Moreover, the study presented in [10], for example, proposes a learning approach to determine the value of the parameter mixing the original query with the feedback terms. In addition, the study presented

Power k	$\mu(ftf)$	$\mu(fdf)$	Mean IDF
0.2	70.46	7.4	5.21
0.5	85.70	7.1	5.09
0.8	88.56	6.82	5.14
1	89.7	6.6	5.1
1.2	91.0	6.35	5.1
1.5	90.3	6.1	5.0
2	89.2	5.8	4.9

(a) Statistics on TREC-12-A

Power k	robust-A	trec-12-A	robust-B	trec-12-B
0.2	29.3	28.7	28.7	30.0
0.5	**30.1**	**29.5**	**29.4**	**30.5**
0.8	29.6	29.3	29.4	30.3
1	29.2	28.9	29.1	29.9
1.2	28.9	28.6	28.6	29.6
1.5	28.6	28.1	28.3	28.9
2	28.1	27.2	27.4	28.0
log-logistic	29.4	28.7	28.5	29.9

(b) MAP (%) for different power function

Fig. 2: (a) Statistics on TREC-12-A. (b) MAP (%) for different power function. Suffix A means $n = 10$ and $tc = 10$ while suffix B means $n = 20$ and $tc = 20$

in [12] focuses on the use of positional and proximity information in the relevance model for PRF, where position and proximity are relative to query terms. Again, this information leads to improved performance. Furthermore, the study presented in [5] for example proposes an algorithm to identify query aspects and automatically expand queries in a way such that all aspects are well covered.

Another comprehensive, and related, study is the one presented in [3, 6]. In this study, a unified optimization framework is retained for robust PRF. Lastly, several studies have recently put forward the problem of uncertainty when estimating PRF weights [4, 9]. These studies show that resampling feedback documents is beneficial as it allows a better estimate of the weights of the terms to be considered for feedback.

The study we have conducted here differs from the above ones as it aims at explaining, through a specific constraint, why some PRF systems work and others do not. Our experimental validation has revealed that the DF constraint is an essential ingredient to be used while designing PRF models, and our theoretical development has shed light on those models which or which do not comply to this constraint.

7 Conclusion

The main contributions of this paper are the formulation of the Document Frequency constraint and its validation.

The performance of PRF models varies from one study to another, as different collections and different ways of tuning model parameters are often used. It is thus very diffcult to draw conclusions on the characteristics of such or such models. What is lacking to do so is a theoretical framework which would allow one to directly compare PRF models, independently of any collection. The theoretical analysis we conduct provides explanations on several experimental findings reported for different PRF models, and thus paves the way towards a theoretical assessment of PRF models.

First, two widely used models in the language modeling family, the simple mixture and the divergence minimization models, are deficient as one does not satisfy the DF constraint while the other does not sufficiently enforce the IDF effect. Second, the mixture of eDCM distributions [14], the geometric relevance model [13], the log-logistic and the smoothed power law models [2] were shown to satisfy the DF constraint. Hence, we argue that the DF constraint do capture the behavior of these recent models and yield an explanation to the obtained improvements.

Finally, we have introduced a simple family of reweighting functions which allow to further compare the different ingredients of PRF models. The experiments conducted with this family bring additional confirmation of the well-foundedness of the DF constraint.

References

1. G. Amati, C. Carpineto, G. Romano, and F. U. Bordoni. Fondazione Ugo Bordoni at TREC 2003: robust and web track, 2003.
2. S. Clinchant and E. Gaussier. Information-based models for *ad hoc* IR. In *SIGIR'10, conference on Research and development in information retrieval*, 2010.
3. K. Collins-Thompson. Reducing the risk of query expansion via robust constrained optimization. In *CIKM'09 conference on Information and knowledge management*.
4. K. Collins-Thompson and J. Callan. Estimation and use of uncertainty in pseudo-relevance feedback. In *SIGIR'07*.
5. D. W. Crabtree, P. Andreae, and X. Gao. Exploiting underrepresented query aspects for automatic query expansion. In *SIGKDD international conference on Knowledge discovery and data mining*, KDD '07, 2007.
6. J. V. Dillon and K. Collins-Thompson. A unified optimization framework for robust pseudo-relevance feedback algorithms. In *CIKM*, pages 1069–1078, 2010.
7. H. Fang, T. Tao, and C. Zhai. A formal study of information retrieval heuristics. In *SIGIR '04: conference on Research and development in information retrieval*.
8. V. Lavrenko and W. B. Croft. Relevance based language models. In *SIGIR '01:conference on Research and development in information retrieval*.
9. K. S. Lee, W. B. Croft, and J. Allan. A cluster-based resampling method for pseudo-relevance feedback. In *SIGIR '08,conference on Research and development in information retrieval*, 2008.
10. Y. Lv and C. Zhai. Adaptive relevance feedback in information retrieval. In *conference on Information and knowledge management*, CIKM '09, 2009.
11. Y. Lv and C. Zhai. A comparative study of methods for estimating query language models with pseudo feedback. In *CIKM '09: conference on Information and knowledge management*, pages 1895–1898, 2009.
12. Y. Lv and C. Zhai. Positional relevance model for pseudo-relevance feedback. In *SIGIR '10,conference on Research and development in information retrieval*, 2010.
13. J. Seo and W. B. Croft. Geometric representations for multiple documents. In *SIGIR '10: conference on Research and development in information retrieval*, 2010.
14. Z. Xu and R. Akella. A new probabilistic retrieval model based on the dirichlet compound multinomial distribution. In *SIGIR '08: conference on Research and development in information retrieval*, 2008.
15. C. Zhai and J. Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *CIKM '01*, 2001.