A Theoretical Analysis of Pseudo-Relevance Feedback Models

Stéphane Clinchant
Xerox Research Center Europe
stephane.clinchant@xrce.xerox.com

Eric Gaussier
LIG, Univ. Grenoble I
eric.gaussier@imag.fr

ABSTRACT

Our goal in this study is to compare several widely used pseudo-relevance feedback (PRF) models and understand what explains their respective behavior. To do so, we first analyze how different PRF models behave through the characteristics of the terms they select and through their performance on two widely used test collections. This analysis reveals that several well-known models surprisingly tend to select very common terms, with low IDF (inverse document frequency). We then introduce several conditions PRF models should satisfy regarding both the terms they select and the way they weigh them, prior to study whether standard PRF models satisfy these conditions or not. This study reveals that most models are deficient with respect to at least one condition, and that this deficiency explains the results of our analysis of the behavior of the models, as well as some of the results reported on the respective performance of PRF models. Based on the PRF conditions, we finally propose possible corrections for the simple mixture model. The PRF models obtained after these corrections outperform their standard version and yield state-of-the-art PRF models which confirms the validity of our theoretical analysis.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

IR Theory, Pseudo Relevance Feedback, Axiomatic Theory

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 modeling 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 with respect to parameter values in [16]. However, several recently proposed PRF models[6, 21, 2] seem to outperform the mixture model. It is nevertheless difficult to compare PRF models and to draw conclusions on the sole basis of studies making use of different collections and different ways of tuning model parameters.

We first analyze in this paper several well-established or recently proposed PRF models and show that they behave differently with respect to the terms they select. We then propose a characterization of PRF models that allow an assessment of their general behavior. In particular, we establish a series of conditions PRF models should satisfy, and review different PRF models according to their behavior with respect to these conditions. This analysis provides explanations on experimental findings reported in different studies of PRF models. From this characterization, we finally introduce variants of the mixture models that both comply with the above-mentioned conditions and outperform their original version.

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 documents used, F the feedback set and tc the number of terms used for pseudo-relevance feedback. An important change of notations concerns TF and DF which are in this paper *related to the feedback set* F.

The remainder of the paper is organized as follows. We give in Section 2 some basic statistics on several PRF models, which reveal significant differences in the way PRF models behave. We then introduce in section 3 general conditions for PRF functions, prior to reviewing standard PRF models according to their behavior with respect to these conditions in section 4. From this analysis, we propose variants of the mixture and geometric relevance models in section 5 that outperform the standard versions of these models. Finally, we discuss some related work in section 6. Throughout this study, we use two standard IR collections: the ROBUST collection, with 250 queries, and the TREC 1&2 collection, with 150 queries corresponding to topics 51 to 200. We only make use of query titles, as this is a common setting when studying PRF [9]. All documents are preprocessed with standard Porter stemming and stopword removal.

ICTIR '13, September 29 - October 02 2013, Copenhagen, Denmark

http://dx.doi.org/10.1145/2499178.2499179

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
$t(w,d)$	normalized # of occurrences (e.g. $\frac{c(w,d)}{l_d}$)
avg_l	Average document length in collection
N	# of docs in collection
N_w	# of documents containing w
$IDF(w)$	IDF of a term (e.g. $-\log(N_w/N)$)
$p(w\|C)$	Corpus language model
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 q
$TF(w)$	$=\sum_{d \in F} c(w,d)$
$DF(w)$	$=\sum_{d \in F} I(c(w,d) > 0)$

Table 1: Notations

2. DO PRF MODELS BEHAVE SIMILARLY?

In order to compare PRF models, our first experiments consists in comparing standard, state-of-the-art PRF models[1], namely (a) the mixture and divergence minimization models presented in [22], (c) the log-logistic feedback model presented in [2], and (d) the Geometric Relevance Model (GRM) presented in [19]. The exact formulation of these models, as well as others, is given in section 4; we just want to illustrate here main differences in the way these standard models behave. For all models, 5-fold cross-validation is used to optimize the different parameters (including the interpolation weight). The parameter ranges are standard and are defined by: $n \in \{10, 20\}$, $tc \in \{10, 20, 50, 75, 100\}$, $\alpha \in \{0.1, ..., 0.9\}$, $\lambda \in \{0.1, ..., 0.9\}$ (parameters of PRF language models), $\beta \in \{0.01, 0.1, 0.25, 0.5, 0.8, 1, 1.2\}$ (parameter for the log-logistic model).

Figure 1 plots the performance of these four different models when the number of feedback terms, tc, varies. As one can see, the log-logistic model only needs 20 terms on both collections to reach its best performance, whereas other models need 100 terms on ROBUST and 150 terms on TREC to attain their best performance. Furthermore, the best performance of the log-logistic model is above the one of the other models, despite the small number of feedback terms it relies on. *What does explain this difference?*

To answer this question, we used the same IR engine for the retrieval step (thus ensuring that all PRF algorithms are computed on the *same set of documents*) and analyzed the terms chosen by the previously mentioned models. We first computed, for each query and for each word, the total number of occurrences of this word in the feedback set (i.e. $TF(w)$), its document frequency in the feedback set (i.e. $DF(w)$) and its inverse document frequency in the collection ($IDF(w)$). We then averaged these quantities over-all feedback terms and queries. For instance, the mean *idf*

[1]By "standard" we mean here models that aim at selecting feedback terms, irrespective of such dimensions as query aspects.

Table 2: Statistics of terms extracted by different models, on two collections. Suffix A means $n = 10$ and $tc = 10$ while suffix B means $n = 20$ and $tc = 20$

Settings	Statistics	MIX	LL	DIV	GRM
robust-A	$\mu(tf)$	62.9	46.7	53.9	52.3
	$\mu(df)$	6.4	7.21	8.6	8.4
	μ(idf)	4.3	5.1	2.2	2.4
trec-1&2-A	$\mu(tf)$	114 .0	79.1	92.6	92.3
	$\mu(df)$	7.1	7.8	8.8	8.7
	μ(idf)	3.8	4.8	2.5	2.5
robust-B	$\mu(tf)$	68.6	59.9	65.3	64.6
	$\mu(df)$	9.9	11.9	14.7	14.4
	μ(idf)	4.4	4.4	1.7	1.9
trec-1&2-B	$\mu(tf)$	137.8	100.0	114.9	114.8
	$\mu(df)$	12.0	13.	15.2	15.2
	μ(idf)	3.8	4.3	2.1	2.2

$\mu(idf)$ is computed as

$$\mu(idf) = \frac{1}{|Q|} \sum_{q} \sum_{i=1}^{tc} \frac{IDF(w_i)}{tc}$$

where $|Q|$ represents the number of queries used (the formulas for $\mu(tf), \mu(df)$ are identical).

Many studies choose a fixed parameter strategy either to compare PRF models or when submitting runs to evaluation campaigns [16, 17]. The choice of the settings we use in this study is dictated by the typical behavior of the log-logistic model. As the log-logistic feedback model outperforms the other feedback models with fewer feedback terms, we focus here on two settings with few feedback terms: setting A, with $n = 10$ and $tc = 10$, and setting B, with $n = 20$ and $tc = 20$. Table 2 displays the above statistics for the four feedback models: mixture model (MIX), log-logistic model (LL), divergence minimization model (DIV) and geometric relevance model (GRM). Our experimental results show that: **(a)** The log-logistic model, which is the best performing model in Figure 1, selects feedback terms that have high IDF, small TF and a medium DF; **(b)** The mixture model selects terms with small DF and high TF; **(c)** The GRM and divergence models select terms with small IDF, and relatively high TF and DF. It thus appears that these models focus on terms with different characteristics to enrich the original query. A first explanation of the better behavior of the log-logistic model thus lies in its capacity to focus on words that are not too common (high IDF and small TF) but that still occur in a sufficient number of feedback documents (average DF). We turn now to a theoretical analysis of these aspects.

3. A CHARACTERIZATION OF PRF

We introduce in this section general characterizations of PRF models that will help us understand the behavior of these models from a theoretical point of view. Our approach is reminiscent of the axiomatic approach adopted Fang et al [11] and followed in many studies including [12, 8, 2]. However, whereas the axiomatic approach aims at describing IR functions by constraints they should satisfy, we rather view here these constraints as general properties that can help us understand, from a theoretical point of view, the behavior of PRF functions.

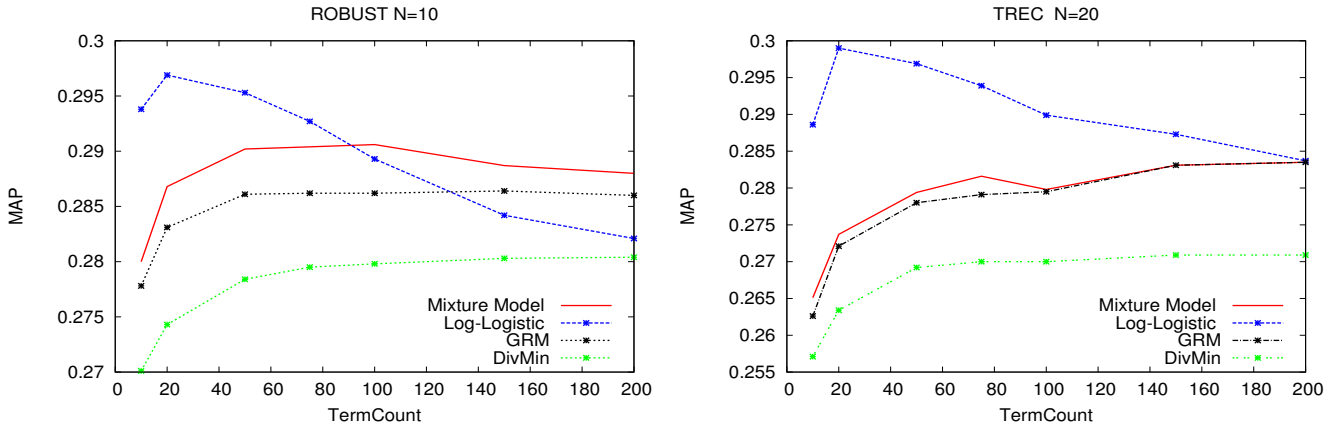

Figure 1: MAP on all queries with $tc \in \{10, 20, 50, 75, 100, 150, 200\}$ with best parameters, ROBUST $n = 10$ left, TREC-1&2 $n = 20$ right

The four main conditions on IR functions considered in the axiomatic theory of IR are: the weighting function should (a) be increasing and (b) concave wrt term frequencies, (c) be increasing wrt IDF and (d) penalize long documents. In the context of PRF, the first two properties relate to the fact that terms frequent in the feedback set are more likely to be effective for feedback (which we refer to as the *TF effect*), but that the difference in frequencies should be less important in high frequency ranges (which we refer to as the *Concavity effect*). The *IDF effect* is also relevant in feedback, as one generally avoids selecting terms with a low IDF. The property regarding 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 (referred as the *Document length effect*).

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[2]. For simplicity, we use the notation $FW(w)$, but it is important to bear in mind that this function depends on a feedback set and some parameters. One can formalize the above considerations as follows:

[**TF effect**] *FW should increase with the term frequency $c(w, d)$; in analytical terms, this gives:*

$$\frac{\partial FW(w)}{\partial c(w,d)} > 0$$

[**Concavity effect**] *The above increase should be less marked in high frequency ranges, which can be formulated as:*

$$\forall d \in \mathbf{F},\ \frac{\partial^2 FW(w)}{\partial c(w,d)^2} < 0$$

[**IDF effect**] *Let w_a and w_b two words such that $IDF(w_b) > IDF(w_a)$ and $\forall d \in \mathbf{F},\ t(w_a, d) = t(w_b, d)$. Then*

$$FW(w_b) > FW(w_a).$$

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

We want to study the increase of the feedback weight wrt IDF, *all other things being equal*. This forces the introduction of the condition on the distribution of frequencies over the feedback documents.

[**Document length effect**] *The number of occurrences of feedback terms should be normalized by the length of documents they appear in:*

$$\frac{\partial FW(w)}{\partial l_d} < 0$$

Lastly, an additional characterization, related to the behavior of PRF models with respect to DF and introduced in [3], can be considered. It stipulates that, *all other things being equal*, terms with a higher DF (i.e. terms occurring in more feedback documents) should receive a higher score:

[**DF effect**] *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:

$$\begin{aligned} T(w_a) &= (t_1, t_2, ..., t_j, 0, ..., 0) \\ T(w_b) &= (t_1, t_2, ..., t_j - \epsilon, \epsilon, ..., 0) \end{aligned}$$

with $\forall i,\ t_i > 0$ and $t_j - \epsilon > 0$ (hence, $TF(w_a) = TF(w_b)$ and $DF(w_b) = DF(w_a) + 1$).

Then: $FW(w_a; \mathbf{F}, \mathbf{P}_{w_a}) < FW(w_b; \mathbf{F}, \mathbf{P}_{w_b})$

The following theorem can help establish whether a PRF model enforces the DF effect. This theorem is a slight extension of the one introduced in [3] and its proof is similar.

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

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

with $\mathbf{P'}_w = \mathbf{P}_w \setminus \{c(w,d)\}$ and $f(0; \mathbf{P'}_w) \geq 0$. If the function f is strictly concave in $c(w,d)$ or linear, $f(c(w,d)) = ac(w,d) + b$, with $a \in \mathbb{R}$ and $b > 0$, then FW enforces the DF effect. If the function f is strictly convex in $c(w,d)$ or linear, $f(c(w,d)) = ac(w,d) + b$, with $a \in \mathbb{R}$ and $b \leq 0$, then FW does not enforce the DF effect.

We now assess how different PRF models behave with respect to these conditions.

4. REVIEW OF PRF MODELS

We review in this section different PRF models according to their behavior wrt the characterizations we have defined. We start with language models, then review the recent model introduced in [21] which is related to both generative methods and to the *Probability Ranking Principle* (PRP), prior to review Divergence from Randomness (DFR) and Information models.

4.1 PRF for Language Models

PRF models within the language modeling (LM) approach to information retrieval assume that words in the feedback document set are distributed according to a multinomial distribution, θ_F (the notation θ_F summarizes the set of parameters $P(w|\theta_F)$). Once the parameters have been estimated, PRF models in the LM approach proceed by interpolating the (original) query language model with the feedback query model θ_F:

$$\theta_{q'} \quad = \quad \alpha\theta_q + (1-\alpha)\theta_F$$

In practice, one restricts θ_F to the top tc words, setting all other values to 0. The different feedback models then differ in the way θ_F is estimated. We review the main LM based feedback models below.

Mixture Model

Zhai and Lafferty [22] propose a generative model for the set $\mathbf{F}$. All documents are i.i.d and each document is generated from a mixture of the feedback query model and the corpus language model:

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

where λ is a "background" noise set to some constant. Finally θ_F is learned by optimizing the data log-likelihood with an Expectation-Maximization (EM) algorithm, leading to the following E and M steps at iteration (i):

$$\begin{aligned} E-step \quad & E(w)^{(i)} = \frac{(1-\lambda)P^{(i}(w|\theta_F)}{(1-\lambda)P^{(i)}(w|\theta_F)+\lambda P^{(i)}(w|C)} \\ M-step \quad & P^{(i+1)}(w|\theta_F) = \frac{\sum_{d\in\mathbf{F}} c(w,d)E(w)^{(i)}}{\sum_w \sum_{d\in\mathbf{F}} c(w,d)E(w)^{(i)}} \end{aligned} \quad (3)$$

where $E(w)^{(i)}$ denotes the expectation of observing w in the feedback set; furthermore, $FW(w) = P(w|\theta_F)$. As one can note, none of the above formulas involve $DF(w)$, neither directly nor indirectly. The mixture model is thus agnostic wrt to DF, and thus does not enforce the DF effect. Regarding the other properties, one can note that the weight of the feedback terms ($P(w|\theta_F)$) increases with $TF(w)$ (which is $\sum_{d\in\mathbf{F}} c(w,d)$), decreases with IDF($w$) (the argument for this is the same as the one developed in [11], a study to which we refer readers). Thus, both TF and IDF effects are enforced. Furthermore, even though counts are normalized by the length (in fact an approximation of it) of the feedback documents, all these documents are merged together, so that the Document length effect is not fully enforced. The situation wrt the Concavity effect is even less clear. In particular, if one approximates the denominator with the total length of the feedback documents (such an approximation being based on the fact that $E(w)^{(i)}$ corresponds to the expectation of w in the feedback set), then the second partial derivative of $P(w|\theta_F)$ wrt to $c(w,d)$ is 0. This suggests that this model does not fully enforce the Concavity effect, and thus that it gives too much weight to high frequency words. This is indeed what we have observed in table 2: the mixture model selects terms with a mean TF which is significantly higher than the mean TF of the other models.

Divergence Minimization

In addition to the mixture model, a divergence minimization model:

$$D(\theta_q|RF) = \frac{1}{|n|}\sum_{i=1}^{n} D(\theta_F \parallel \theta_{d_i}) - \delta D(\theta_F||p(.\parallel C))$$

is also proposed in [22], where θ_{d_i} denotes the empirical distribution of words in document d_i. Minimizing this divergence gives the following solution:

$$\begin{aligned} P(w|\theta_F) \quad \propto \quad & \exp\Big(\frac{1}{(1-\delta)}\frac{1}{n}\sum_{d\in\mathbf{F}}\log(p(w|\theta_d)) \\ & -\frac{\delta}{1-\delta}\log(p(w|C))\Big) \end{aligned}$$

Here again, $FW(w) = P(w|\theta_F)$. As $p(w|\theta_d) = \frac{c(w,d)}{l_d}$, this equation corresponds to the form given in equation 1 with a strictly concave function (log). Thus, by Theorem 1, this model enforces the DF effect[3]. Being based on standard language models, it also enforces the TF, Concavity and Document length effects. Our experiment results, as well as those reported in [16], however show that this model does not perform as well as other ones. Indeed, as shown in table 2, this model selects terms with small IDF and fails to downweight common words. We explain here this phenomenon by examining how this model behaves with respect to the IDF effect.

Let us consider two terms w_a and w_b such that $\forall d \in \mathbf{F}$ $t(w_a,d) = t(w_b,d) = t_d$, and let $p(w_b|C)$ and $p(w_a|C)$ such that $p(w_a|C) < p(w_b|C)$ (i.e. $IDF(w_a) > IDF(w_b)$). The IDF effect stipulates that, in this case, $FW(a)$ should be greater than $FW(b)$. Using Jelinek Mercer smoothing, $\log(FW(w_a)) - \log(FW(w_b))$ is equal to (we skip here the derivation which is purely technical):

$$\sum_{d\in\mathbf{F}}\{\overbrace{\log(\frac{(1-\lambda)t(d)+\lambda p(w_a|C)}{(1-\lambda)t(d)+\lambda p(w_b|C)})}^{<0} - \delta\overbrace{\log(\frac{p(w_a|C)}{p(w_b|C)})}^{<0}\} \quad (4)$$

As $0 < \lambda, \delta \leq 1$, we have:

$$\forall(x,y,z) \in \mathbb{R}^{+*} \; s.t. \; y > x, \log(\frac{z+\lambda x}{z+\lambda y}) > \log(\frac{x}{y}) > \delta\log(\frac{x}{y})$$

Thus $\log(FW(w_a)) - \log(FW(w_b)) > 0$ and $FW(w_a) > FW(w_b)$. The divergence minimization model is thus compliant with the IDF effect. However, this effect is in fact poorly enforced in this model. To see that, let us assume that $P(w_b|C)$ is K times ($K > 1$) larger than $P(w_a|C)$: $P(w_b|C) = KP(w_a|C)$. Then:

$$\log\frac{FW(w_a)}{FW(w_b)} < -\delta\log\frac{1}{K} = \log K^{\delta}$$

[3]This was already noted in [3].

and:

$$FW(w_b) < FW(w_a) < K^{\delta} FW(w_b)$$

The original factor K difference thus amounts, for the PRF weighting of terms, to K^{δ}. In practice, typical values of δ are close to 0.1; in this case, K^{δ} is close to 1 (it is 1.07 for $K = 2$, 1.17 for $K = 5$ and 1.58 for $K = 100$) and there is almost no difference between $FW(w_a)$ and $FW(w_b)$. This explains the small values displayed in table 2 for the IDF statistics.

Geometric Relevance Models

(1) A regularized version of the mixture model, known as the regularized mixture model (RMM) and making use of latent topics, is proposed in [20] to correct some of the deficiencies of the simple mixture model. RMM has the advantage of providing a joint estimation of the document relevance weights and the topic conditional word probabilities, yielding a robust setting of the feedback parameters. However, the experiments reported in [16] show that this model is less effective than the simple mixture model in terms of retrieval performance.

(2) Another PRF model proposed in the framework of the language modeling approach is the relevance model, proposed by Lavrenko *et al.* [13], and defined by:

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

where P_{LM} denotes the standard language model. This model is directly based on the standard language models and thus enforces, as its standard counterpart, the TF, Concavity and Document length effect. Furthermore, it corresponds to the form of equation 1 of Theorem 1, with a linear function of the form $f(x) = \alpha x$. This model thus does not enforce the DF effect. Lastly, as for the following model, it fails to enforce the IDF effect, as we will see below (the proof for that is identical to the one for the following model, given below).

The relevance model has recently been refined in the study presented in [19] 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 one can note, the product in the relevance model has now been replaced by a power function. The GRM model enforces the TF, Concavity and Document Length effect (the argument is the same as for the relevance model). We now examine its behavior with respect to the DF and IDF effects.

Let us consider this model with Jelinek-Mercer smoothing [23]: $P_{LM}(w|\theta_d) = (1-\lambda)\frac{c(w,d)}{l_d} + \lambda \frac{c(w,C)}{l_C}$, where $c(w, C)$ denotes the number of occurrences of w in the collection C and l_C the length of the collection. Let w_a and w_b be two words as defined in the DF effect, and let us further assume that feedback documents are of the same length l and equiprobable given q. Let A, B and ϵ' be defined as:

$$A = (1-\lambda)\frac{c(w_a, d_j)}{l} + \lambda \frac{c(w_a, C)}{l_C}, B = \lambda \frac{c(w_b, C)}{l_C}, \epsilon' = (1-\lambda)\frac{\epsilon}{l}$$

where d_j is the document defined in the DF effect. Then:

$$\frac{FW(w_a)}{FW(w_b)} = \frac{AB}{(A-\epsilon')(B+\epsilon')}$$

And: $(A-\epsilon')(B+\epsilon') = AB + \epsilon'[(A-B) - \epsilon']$. But $A - B = (1-\lambda)\frac{c(w_a,d_j)}{l}$, a quantity which is strictly greater than $(1-\lambda)\frac{\epsilon}{l} = \epsilon'$ by the assumptions of the DF effect. Thus, the GRM model enforces the DF effect when Jelinek-Mercer is used. A similar development can be obtained for Dirichlet smoothing. However, this model fails to enforce the IDF effect.

Let w_a and w_b two words such that $p(w_a|C) < p(w_b|C)$ and $\forall d \in \mathbf{F}\ t(w_a, d) = t(w_b, d) = t_d$. Indeed (skipping again the derivation details), $\log(FW(w_a)) - \log(FW(w_b))$ (and hence $FW(w_a) - FW(w_b)$) has the sign of:

$$\sum_d P(d|q) \log \frac{\lambda t_d + (1-\lambda) p(w_a|C)}{\lambda t_d + (1-\lambda) p(w_b|C)}$$

a quantity which is strictly negative as $p(w_b|C) > p(w_a|C)$. This explains the results displayed in table 2, showing that the GRM model selects terms with low IDF.

4.2 PRF under the PRP

Xu and Akella [21] propose an instantiation of the Probability Ranking Principle (PRP) in which relevant documents are assumed to be generated from a Dirichlet Compound Multinomial (DCM) distribution, or an approximation of it, called eDCM and introduced in [10]. The PRF version of this model simply assumes that the feedback documents are relevant. 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 English words occurring frequently in the whole collection, whereas z_{FR} is used to represent terms occurring in the feedback documents and pertinent to the user's information need. The parameters of the two components are estimated through rather time-consuming and complex estimation procedures, typically based on gradient descent or the EM algorithm. [21] furthermore proposes two modifications of the EM algorithm to estimate the parameters of the relevant component, in a way similar to the one followed by [20]. 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, enforces the DF effect. However, it does not rely on the number of occurrences and thus does not enforce the TF effect. That said, the higher the DF of a term, the higher its TF is likely to be in average, so that this model can nevertheless indirectly select high frequency terms by selecting terms with high DF. We conjecture that this is the case with the (e)DCM model, which seems to behave well in practice. Finally, the EM steps also suggest that this model satisfy the IDF condition, as the mixture model does. We however have no theoretical proof for this.

4.3 PRF in DFR and Information Models

In DFR and information models, the original query is modified to take into account the words appearing in $\mathbf{F}$ according to the following scheme:

$$c(w, q') \quad = \quad \frac{c(w,q)}{\max_w c(w,q)} + \beta \frac{\text{Info}(w, \mathbf{F})}{\max_w \text{Info}(w, \mathbf{F})}$$

PRF Model vs Conditions	TF	Concave	IDF	Doc Len	DF
Mixture	yes	not sufficiently	yes	no	no
Div Min	yes	yes	not sufficiently	yes	yes
Geometric Relevance	yes	yes	no	yes	yes
Bo	yes	no	not systematically	no	no
Log-logistic	yes	yes	yes	yes	yes

Table 3: Summary of main PRF models with respect to the conditions of Section 3

where β is a parameter controlling the modification brought by $\mathbf{F}$ to the original query ($c(w,q')$ denotes the updated weight of w in the feedback query, whereas $c(w,q)$ corresponds to the weight in the original query). In this case: $FW(w) = \text{Info}(w, \mathbf{F})$.

Bo Model

Standard PRF models in the DFR family are the Bo models [1], which are defined by:

$$\text{Info}(w, \mathbf{F}) = \log_2(1 + g_w) + TF(w) \log_2(\frac{1 + g_w}{g_w})$$

where $g_w = \frac{N_w}{N}$ in the $Bo1$ model and $g_w = P(w|C)(\sum_{d \in F} l_d)$ in the $Bo2$ model. In other words, documents in $\mathbf{F}$ are merged together and a geometric probability model (or a different distribution, the choice of the distribution being irrelevant for our argument) is used to measure the informative content of a word. As this model is DF agnostic, it does not enforce the DF effect. Furthermore, when using the geometric distribution, the Concavity effect is not enforced as the second derivative of $FW(w)$ wrt to $TF(w)$ is null. Neither does it enforce the Document length effect, as feedback documents are merged together. Regarding the IDF effect, the derivative of $\text{Info}(w, \mathbf{F})$ with respect to g_w can be positive or negative depending on the values of both g_w and $TF(w)$. There is thus no guarantee that this model is compliant with the IDF condition.

Log-logistic Model

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

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

where $t(w,d)$ is the normalized number of occurrences of w in d (it is set to: $c(w,d) \log(1 + c\frac{\text{avg}_l}{l_d})$), and λ_w a parameter associated to w and set to: $\lambda_w = \frac{N_w}{N}$. Two instantiations of the general information-based family are considered in [2], respectively based on the log-logistic distribution and a smoothed power law (SPL). We focus here on the log-logistic model,which takes a simpler form and performs similarly to the SPL model in PRF. The log-logistic model is defined by:

$$FW(w) = \frac{1}{n} \sum_{d \in F} \log(\frac{t(w,d) + \lambda_w}{\lambda_w})$$

As the logarithm is a concave function, the log-logistic model enforces the DF effect by Theorem 1, as well as the Concavity effect. It is furthermore compliant with the DF and Document length effects as it based on the general information formulation with a bursty distribution (as shown in [2]). Let us take a closer look now at the IDF effect.

Let w_a and w_b, two words such as in the IDF condition (in particular, one has $\lambda_b < \lambda_a$), then:

$$FW(w_a) - FW(w_b) = \frac{1}{n} \sum_{d \in \mathbf{F}} \log \frac{\lambda_a \lambda_b + \lambda_b t_d}{\lambda_a \lambda_b + \lambda_a t_d}$$

which is unconditionally negative. This model thus satisfies the IDF condition.

4.4 Summary

The results we have obtained in this section provides a clear explanation of the experimental results we have reported in Section 2. We provide in Table 3 a summary of the behavior of the main PRF models we have reviewed with respect to the PRF conditions.

We now show that it is possible to exploit them further to improve two well-known models: the mixture and geometric relevance models.

5. IMPROVING MIX AND GRM

We now turn to the problem of improving existing models so that they are compliant with the conditions developed before. We focus here on two models, the mixture and geometric relevance models, as they are the ones, just after the log-logistic model, with the best performance in the experiments reported in Section 2 (see Figure 1).

As noted before, the mixture model is deficient with respect to the DF and Document Length effects, and partly with the Concavity effect. This is due to the fact that the M-step (equation 3, which defines the feedback weight) is a linear function of $c(w,d)$ (as before, one can approximate the denominator of the M-step as the total length of the feedback documents). One way to correct this is to replace $c(w,d)$ by a concave function (this will ensure both the Concavity and DF effects), and to normalize it by the document length (so as to enforce the Document length effect). To do so, we consider a generalization of equation 2 defined by:

$$P^s(\mathbf{F}|\theta_F, \lambda) = \prod_{w=1}^{V} ((1-\lambda)P(w|\theta_F) + \lambda P(w|C))^{A(w)} \quad (5)$$

where $A(w)$ replaces $TF(w)$ and is defined by: $A(w) = \sum_{d \in \mathbf{F}} f(c(w,d), l_d)$ ($TF(w)$ is the special case obtained when $f(c(w,d), l_d) = c(w,d)$).

We consider here two different forms for the function f:

$$f((c(w,d), l_d) = \quad int(Z\sqrt{c(w,d)})$$
$$f(c(w,d)) = \quad int(Z\sqrt{\frac{c(w,d)}{l_d}})$$

where Z is a large constant, here set to 10000, and where int denotes the integer part. Both forms thus define integers which are proportional to $\sqrt{c(w,d)}$. Both forms are furthermore concave in $c(w,d)$ (the second form additionally contains a normalization by the document legnth). Equation 5

Collection	Settings	MIX	c1MIX	c2MIX
ROBUST	n=10, tc=10	28.0	29.2†	**29.4**†
	n=10, tc=20	28.7	29.8†	**30.0**†
	n=20, tc=20	28.3	29.2†	**29.5**†
	n=20, tc=50	28.5	29.1†	**29.6**†
TREC 1&2	n=10, tc=10	26.3	27.6†	**28.3**†
	n=10, tc=20	27.0	28.1†	**29.0**†
	n=20, tc=20	27.4	28.5†	**29.2**†
	n=20, tc=50	28.0	28.8†	**29.7**†

Table 4: Results (MAP) for the corrected mixture model. A † indicates that the difference with the standard mixture model (MIX) is significant (t-test with p-value set to 0.05). Results in bold correspond to the best results

Coll.	Settings	LL	c2MIX	c2+GRM
ROBUST	n=10,tc=10	**29.4**	**29.4**	28.9
	n=10,tc=20	29.7	**30**	29.6
	n=20,tc=20	28.7	**29.5**	29.0
	n=20,tc=50	28.6†	**29.6**	29.5
TREC 1&2	n=10,tc=10	**28.7**	28.3	27.6
	n=10,tc=20	**29.6**	29	28.2 †
	n=20,tc=20	**29.9**	29.2	28.4†
	n=20,tc=50	**29.7**	**29.7**	28.7†

Table 5: Overall results (MAP) for the three "best" models. A † indicates that the difference with the best model (in bold) is significant (t-test with p-value set to 0.05)

thus corresponds to the likelihood of observations that are no longer the number of occurrences of terms but the results of the application of f on these numbers (hence the notation Ps). The EM algorithm applied to this model leads to the following update rules:

$$E(w)^{(i)} = \frac{(1-\lambda)P^{(i}(w|\theta_F)}{(1-\lambda)P^{(i)}(w|\theta_F) + \lambda P^{(i)}(w|C)}$$
$$P^{(i+1)}(w|\theta_F) \propto \sum_{d \in \mathbf{F}} f(c(w,d), l_d) E(w)^{(i)} \quad (6)$$

As before, $FW(w) = P(w|\theta_F)$. By theorem 1, as $f(c(w,d), l_d)$ is strictly concave in $c(w,d)$, the model satisfies the DF effect. Furthermore, it also satisfies the concavity effect (as the second derivative of FW with respect to $c(w,d)$ has the same sign of the second derivative of f with respect to $c(w,d)$, which is negative as f is strictly concave). In addition, if f integrates a normalization by the document length l_d, then the document length effect is enforced, which is the case for the second function we consider (but not the first one).

Table 4 gives the MAP (mean average precision) results obtained with this new model. The function $f((c(w,d), l_d) = \sqrt{c(w,d)}$ corresponds to the model *c1MIX*, and the function $f(c(w,d)) = \sqrt{\frac{c(w,d)}{l_d}}$ to the model *c2MIX*. As one can note, both corrections significantly improve the performance of the standard mixture model, the best results being provided by the corrected mixture model that enforces all the conditions we have reviewed previously.

The modification of the geometric relevance model we propose makes use of a different approach. We follow here the recommendation developed in [18] stating that the term selection and term weighting functions for PRF should be different. As the GRM model does not satisfy the IDF effect, we make use here, for term selection, of models that strongly enforce this effect and behaves well, namely the corrected mixture model (c2MIX) satisfying all PRF conditions (leading to the model c2+GRM).

Finally, table 5 compares the performance of the three best PRF models considered and developed in this study: the log-logistic model (LL), the corrected mixture model (c2MIX) and the corrected geometric relevance model (c2 +GRM). As one can note, the log-logistic and corrected mixture model behave similarly, the log-logistic model being slightly better on TREC while the corrected mixture model is better on ROBUST. Furthermore, the performance for the corrected geometric relevance model is similar to the other ones on ROBUST, and slightly below the others on TREC. However, the GRM weighting function does not bring improvements compared to the c2MIX model.

6. RELATED WORK

We have studied here the main characteristics of PRF reweighting schemes through several constraints reweighting functions should satisfy. This is to our knowledge the first study to propose general theoretical characterizations for PRF functions. There are however a certain number of additional elements that can be used to improve performance of PRF systems. The study presented in [15], for example, proposes a learning approach to determine the value of the parameter mixing the original query with the feedback terms. Interestingly, such a parameter can be set on a query-dependent manner for improved performance. The study presented in [17] 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. It is not clear yet how one can integrate such an information in the other PRF models we have reviewed, in particular in the LL and SPL models, and this is an aspect one will have to investigate further. Another kind of information that can successfully be exploited in PRF is the one related to query aspects.

The study presented in [7] for example proposes an algorithm to identify query aspects and automatically expand queries in a way such that all aspects are well covered. A similar strategy can be deployed on top of any PRF reweighting function, so as to guarantee a certain aspect coverage in the newly formed query. Another comprehensive, and related, study is the one presented in [5, 9]. In this study, a unified optimization framework is retained for robust PRF. The constraints considered however differ from the constraints we have defined, as they aim at capturing diversity through aspect coverage. The general framework of concave-convex optimization (fully detailed in [4]) is nevertheless interesting and bridges several different models [9]. Lastly, several studies have recently put forward the problem of uncertainty when estimating PRF weights [6, 14]. 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. Interestingly, these approaches can be deployed with any PRF reweighting model and allow a simple and neat integration of the DS constraint in any PRF

model, as the resampling procedure is based on the score of the document obtained in the first retrieval step.

Compared to [3] and [16], the study presented here goes beyond these previous studies by considering a larger set of properties (related to the classical IR conditions), and by analyzing a large set of PRF models according to their behavior wrt to all the properties considered. Only through this complete theoretical study were we able to spot the IDF deficiencies of both the Relevance Model (including its generalized version) and the Divergence Minimization model. The development presented here provides a clear explanation on why the log-logistic model outperforms the other models in PRF settings. Lastly, the framework we have developed here is consistent with different empirical observations.

7. CONCLUSION

We have analyzed in this paper how different PRF models behave, through the characteristics of the terms they select and through their performance on two widely used test collections. We have then introduced conditions PRF models should satisfy, namely the TF, Concavity, IDF, Document length and DF effects. We have then studied whether standard PRF models satisfy these conditions or not. In addition to explaining the experimental analysis we have conducted, the results of this study revealed that:

(a) the mixture model, from the language modeling family, as well as the Bo model, from the DFR family, are deficient with respect to the Concavity, Document length and DF effects;

(b) the divergence minimization and geometric relevance models, also from the language modeling family, do not sufficiently enforce the IDF effect (the geometric relevance model simply fails to enforce this effect);

(c) the log-logistic model, from the information family, satisfies all the PRF conditions.

In addition to explain the results reported here, these findings also explain the results reported in other studies.

Lastly, we have proposed possible corrections of the mixture model (certainly the most widely used PRF model) and the geometric relevance model. The correction of the mixture model, based on a generalization of the equation at he basis of the model, satisfies all the PRF conditions and significantly outperform the standard formulation. It yields a model on par with the best PRF models we have reviewed.

8. 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*, SIGIR '10, pages 234–241, New York, NY, USA, 2010. ACM.

[3] S. Clinchant and E. Gaussier. Is document frequency important for prf? In *ICTIR*, pages 89–100, 2011.

[4] K. Collins-Thompson. Estimating robust query models with convex optimization. In *NIPS*, pages 329–336, 2008.

[5] K. Collins-Thompson. Reducing the risk of query expansion via robust constrained optimization. In *CIKM'09*, CIKM '09, pages 837–846, 2009.

[6] K. Collins-Thompson and J. Callan. Estimation and use of uncertainty in pseudo-relevance feedback. In *SIGIR'07*, SIGIR '07, pages 303–310, 2007.

[7] D. W. Crabtree, P. Andreae, and X. Gao. Exploiting underrepresented query aspects for automatic query expansion. In *KDD'07*, KDD '07, pages 191–200, New York, NY, USA, 2007. ACM.

[8] R. Cummins and C. O'Riordan. An axiomatic comparison of learned term-weighting schemes in information retrieval: clarifications and extensions. *Artif. Intell. Rev.*, 28:51–68, June 2007.

[9] J. V. Dillon and K. Collins-Thompson. A unified optimization framework for robust pseudo-relevance feedback algorithms. In *CIKM*, pages 1069–1078, 2010.

[10] C. Elkan. Clustering documents with an exponential-family approximation of the Dirichlet compound multinomial distribution. In W. W. Cohen and A. Moore, editors, *ICML*, volume 148, pages 289–296. ACM, 2006.

[11] H. Fang, T. Tao, and C. Zhai. A formal study of information retrieval heuristics. In *SIGIR '04*, 2004.

[12] H. Fang and C. Zhai. Semantic term matching in axiomatic approaches to information retrieval. In *SIGIR'06*, SIGIR '06, pages 115–122, 2006.

[13] V. Lavrenko and W. B. Croft. Relevance based language models. In *SIGIR '01*, pages 120–127, New York, NY, USA, 2001. ACM.

[14] K. S. Lee, W. B. Croft, and J. Allan. A cluster-based resampling method for pseudo-relevance feedback. In *SIGIR'08*, 2008.

[15] Y. Lv and C. Zhai. Adaptive relevance feedback in information retrieval. In *CIKM'09*, CIKM '09, pages 255–264, New York, NY, USA, 2009. ACM.

[16] Y. Lv and C. Zhai. A comparative study of methods for estimating query language models with pseudo feedback. In *CIKM '09*, 2009.

[17] Y. Lv and C. Zhai. Positional relevance model for pseudo-relevance feedback. In *SIGIR'10*, 2010.

[18] S. Robertson. On term selection for query expansion. *Journal of Documentation*, 46, 1990.

[19] J. Seo and W. B. Croft. Geometric representations for multiple documents. In *SIGIR '10*, pages 251–258, New York, NY, USA, 2010. ACM.

[20] T. Tao and C. Zhai. Regularized estimation of mixture models for robust pseudo-relevance feedback. In *SIGIR'06*, SIGIR '06, pages 162–169, 2006.

[21] Z. Xu and R. Akella. A new probabilistic retrieval model based on the dirichlet compound multinomial distribution. In *SIGIR '08*, pages 427–434, 2008.

[22] C. Zhai and J. Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *CIKM '01*, pages 403–410, 2001.

[23] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214, 2004.