A Document Frequency Constraint for Pseudo-Relevance Feedback Models

Stéphane Clinchant, Éric Gaussier

A Document Frequency Constraint for Pseudo-Relevance Feedback Models

Stephane Clinchant[*†], **Eric Gaussier**[†]

[*] *Xerox Research Center Europe*
[†] *Laboratoire d'Informatique de Grenoble, Université de Grenoble*
stephane.clinchant@xrce.xerox.com, eric.gaussier@imag.fr

RÉSUMÉ. *Nous étudions dans cet article le comportement de plusieurs modèles de rétro-pertinence en mettant en avant leurs principales caractéristiques. Ceci nous conduit à introduire une nouvelle contrainte pour les modèles de rétro-pertinence, contrainte liée à la fréquence documentaire (DF) des mots. Nous analysons ensuite, d'un point de vue théorique, différents modèles de rétro-pertinence par rapport à cette contrainte. Cette analyse montre que le modèle de mélange utilisé en rétro-pertinence pour les modèles de langue ne satisfait pas cette contrainte. Nous réalisons ensuite une série d'expériences qui permettent de valider la contrainte DF. Pour cela, nous utilisons tout d'abord un oracle sur la base de documents pertinents, puis utilisons une famile de fonctons de type tf-idf, mais paramétrée de telle sorte que des individus différents de la famille auront des comportements différents par rapport à la contrainte DF. Ces expériences montrent la validité et l'importance de la contrainte DF.*

ABSTRACT. *We study in this paper the behavior of several PRF models, and display their main characteristics. This will lead us to introduce a new heuristic constraint for PRF models, referred to as the Document Frequency (DF) constraint. 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. We then conduct a series of experiments in order to see whether the DF constraint is valid or not. To do so, we performed tests with an oracle and a simple family of tf-idf functions based on a prameter k controlling the convexity/concavity of the function. Both the oracle and the results obtained with this family of functions validate the DF constraint.*

MOTS-CLÉS : *Modèles de RI, boucle de rétropertinence*

KEYWORDS: *IR theoretical models, pseudo-relevance feedback*

1. Introduction

In the language modelling approach to IR, the mixture model for pseudo-relevance feedback is a state of the art algorithm. Numerous studies use it as a baseline, and it has been shown to be one of the most effective models in terms of performance and stability wrt parameter values in (Lv *et al.,* 2009). However, several recently proposed models outperform this model, as models based on bagging, models based on a mixture of EDCM distributions and information models (Collins-Thompson *et al.,* 2007, Xu *et al.,* 2008, Clinchant *et al.,* 2010). We try here to highlight what these new models have in common. This leads us to formulate a heuristic constraint for pseudo-relevance feedback, which we will refer to as the Document Frequency constraint. Our analysis shows that most proposed models meet this heuristic contraint. Finally, we run experiments to assess the validity of this cnstraint. 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. An important change of notations concerns TF and DF which are in this paper *related to the feedback set* F.

Notation	Description
General	
$c(w,d)$	Number of occurrences of w in document d
l_d	Length of document d
N	Number of documents in the collection
f_w	Number of occurrences of w in the collection
n_w	Number of documents containing w
$IDF(w)$	$-\log(n_w/N)$
PRF specific	
n	Number of (top) documents retained for PRF
$\mathbf{F}$	Set of documents retained for PRF ; $\mathbf{F} = (d_1, \ldots, d_n)$
tc	*TermCount* : number of terms in $\mathbf{F}$ to be added to the query
$TF(w)$	$=\sum_{d \in F} c(w,d)$
$DF(w)$	$=\sum_{d \in F} I(c(w,d) > 0)$

Tableau 1. *Notations*

2. Pseudo-Relevance Feedback Statistics

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 (Zhai *et al.,* 2001), and the mean log-logistic information model in the information-based family (Clinchant *et al.,* 2010). These models are reviewed later in section 4, and their exact formulation is not necessary here. In order to have

Tableau 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

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, which is a common setting when studying PRF (Dillon *et al.*, 2010). All documents were preprocessed with standard Porter stemming.

2.1. *Term Statistics*

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. $TF(w)$) and the document frequency of this word in the feedback set (i.e. $DF(w)$). 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. $q(tf)$ and $q(df)$:

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

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

$$\mu(TF) = \sum_{q} \frac{q(TF)}{|Q|} \text{ and } \mu(DF) = \sum_{q} \frac{q(DF)}{|Q|}$$

An average IDF can be computed in exactly the same way. 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

Tableau 3. *Statistics of terms extracted by*

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

the mixture and log-logistic models, on all collections, the mixture model chooses in average words that have a *higher TF*, and a smaller *DF*. The mixture model also chooses words that are *more frequent in the collection* since the mean IDF values are smaller. On the other hand, the satistics of the divergence model shows that this model extracts very common terms, with low IDF and high DF, which, as we will see later, is one of the main drawback of this model.

2.2. *Performance Statistics*

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- we call this setting *raw*. Then, for each query we keep only terms that belong to the intersection of the mixture (respectively the divergence minimization) and log-logistic models , 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 (Clinchant *et al.*, 2010). 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.

Let's summarize our finding here. (a) The log-logistic model performs better thant the mixture and divergence models for PRF. (b) The mixture and divergence models

Tableau 4. *Mean Average Precision for*

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

choose terms with a *higher TF* and a smaller *DF* than the log-logistic one. A first explanation of the better behavior of the log-logistic model can be that the IDF effect is dealt with more efficiently in this model, as shown by the statistics reported in table 3. We also postulate that the log-logistic model tends to favor terms with a *high DF*, while the other models favor terms with a low *DF*. This leads us now to propose a new heuristic constraint for pseudo-relevance feedback.

3. Heuristic Constraints

Axiomatic methods were pioneered by Fang et al (Fang *et al.,* 2004) and followed by many works including (Fang *et al.,* 2006, Cummins *et al.,* 2007, Clinchant *et al.,* 2010). In a nutshell, axiomatic methods describe IR functions by properties. According to (Clinchant *et al.,* 2010), the four main conditions for an IR function to be valid are : the weighting function should (a) be increasing and (b) be concave wrt document term frequencies, (c) have an IDF effect and (d) penalize long documents. In the context of pseudo-relevance feedback, Lv (Lv *et al.,* 2009) mentions a document score heuristic constraint implemented in relevance models (Lavrenko *et al.,* 2001) and in the Rocchio algorithm (Hoashi *et al.,* 2001). The document score heuristic constraint can be formulated as follows :

PRF Heuristic Constraint 1. [Document Score] *Document with higher score should be given more weight in the feedback weight function.*

Another heuristic is related to the term proximity constraint, that is feedback terms should be close to query terms in documents (Lv *et al.*, 2010).

The development made in the previous section however suggests that an additional constraint seems to regulate the good behavior of PRF models. Indeed, as we have seen, the best PRF model we have studied favors feedback terms with a high document frequency in the feedback set, whereas the ohter models we studied fail to do so. This constraint can be formalized as follows :

PRF Heuristic Constraint 2. [Document Frequency] *Let $\epsilon > 0$, and a and b two words such that :*

1) $IDF(a) = IDF(b)$

2) The distribution of the frequencies of a ($c(a,d)$) in the feedback set is given by : $T(a) = \overbrace{(x_1, x_2, ..., x_j, 0, ..., 0)}^{n}$

3) The distribution for b is given by : $T(b) = (x_1, x_2, ..., x_j - \epsilon, \epsilon, ..., 0)$

4) $\forall i$, $x_i > 0$ *and* $x_j - \epsilon > 0$

Hence, $TF(a) = TF(b)$ *and* $DF(b) = DF(a) + 1$. *Then, the feedback weight function* $FW(.)$ *is such that* $FW(a) < FW(b)$

In other words, $FW(.)$ is *locally* growing with $DF(w)$. It is possible to define a constraint based on a globally growing function, but this complicates the matter. Furthermore, the above constraints directly captures the intuition put forward for the document frequency behavior. The following theorem allows one to decide whether a given PRF model agrees or not with the document frequency (DF) constraint for a large class of models (as we will see below) :

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

$$FW(w) \quad = \quad \sum_{d=1}^{n} f(c(w,d)) \qquad [1]$$

with $f(0) = 0$. The we have :

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.

Proof 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. We have :

$$FW(a) = FW(\underbrace{x^1, ..., x^j}_{DF(a)}, \underbrace{0, ..., 0}_{n - DF(a)})$$

and :

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

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 $FW(b) - FW(a) < 0$. □

As we will see in the next section, many recently proposed PRF models follow equation 1, and can be analyzed with the above theorem.

4. Review of PRF Models

4.1. *PRF for Language Models*

Traditional methods, such as Rocchio's algorithm, extract terms from feedback documents and add them to the query. The language modeling (LM) approach to information retrieval follows this approach as it extracts a multinomial probability distribution over words from the feedback document set, parametrized by θ_F. Assuming θ_F has been estimated, the LM approach proceeds by interpolating the query language model with θ_F :

$$\theta_{q'} \quad = \quad \alpha\theta_q + (1-\alpha)\theta_F \qquad [2]$$

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.

4.1.1. *Mixture Model*

Zhai and Lafferty (Zhai *et al.,* 2001) propose a generative model for the set F. All documents are i.i.d and each document comes from a mixture of the relevant topic model and the corpus langague model :

$$P(\mathbf{F}|\theta_F, \beta, \lambda) = \prod_{w=1}^{V} (\lambda\theta_{Fw} + (1-\lambda)P(w|C))^{TF(w)} \qquad [3]$$

where λ is a fixed parameter, which can be understood as a noise parameter for the distribution of terms. Finally θ_F is learned by optimising the data loglikelihood with an Expectation-Maximization (EM) algorithm. It is trivial to show that this mixture model does not meet the DF constraint, since it is DF agnostic.

4.1.2. *Divergence Minimization*

Zhai (Zhai *et al.,* 2001) also propose the divergence minimization model :

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

where θ_{d_i} denotes the empirical distribution of words in document d_i. Miminizing this divergence gives the following solution :

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

This models amounts to the geometric mean of the smoothed document models with a regularization term. Our previous experiments and those of Lv (Lv *et al.*, 2009) show that this model does not perform well. Although it meets the DF constraint (by using a geometric mean leading to a concave function), the IDF effect is not sufficiently enforced, and the model fails to downweight common words, as shown in Table 3. In other words, this model chooses common words which do have a high document frequency, but are not interesting for retrieval.

4.1.3. *Other Models*

A regularized version of the mixture model, known as the regularized mixture model (RMM) and making use of latent topics, is proposed in (Tao *et al.*, 2006) 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 (Lv *et al.*, 2009) show that this model is less effective than the simple mixture model in terms of retrieval performance, for precision and recall. We will thus not study it further here, but want to mention, nevertheless, an interesting re-interpretation of this model in the context of the concave-convex procedure framework (Dillon *et al.*, 2010).

Another PRF model proposed in the framework of the language modeling approach is the so-called relevance model, proposed by Lavrenko *et al.* (Lavrenko *et al.*, 2001), and defined by :

$$FW(w) \quad \propto \quad \sum_{d \in \mathbf{F}} P_{LM}(w|\theta_d) P(d|q) \qquad [4]$$

where P_{LM} denotes the standard language model. Because of its reliance on the language model, the above formulation is compliant with all the classical IR constraints. Furthermore, it 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. As we have mentioned before, it satifies the DS constraint.

The relevance model has recently been refined in the study presented in (Seo *et al.*, 2010) through a geometric variant, referred to as GRM, and defined by :

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

Let us first consider the standard language model with Jelinek-Mercer smothing (Zhai *et al.*, 2004) : $P_{LM}(w|\theta_d) = (1-\lambda)\frac{c(w,d)}{l_d} + \lambda\frac{c(w,C)}{l_C}$, 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 constraint DF, and let us further assume that feedback documents are of the same length l and equiprobable given q. Then $FW(w_a)$ and $FW(w_b)$ respectiveley differ on the two quantities :

$$(i) \qquad (\overbrace{(1-\lambda)\frac{c(w_a,d_j)}{l}+\lambda\frac{c(w_a,C)}{l_C}}^{\alpha})(\overbrace{\lambda\frac{c(w,C)}{l_C}}^{\beta})$$

$$(ii) \quad ((1-\lambda)\tfrac{c(w_a,d_j)-\epsilon}{l}+\lambda\tfrac{c(w_a,C)}{l_C})(\underbrace{(1-\lambda)\frac{\epsilon}{l}}_{\epsilon'}+\lambda\tfrac{c(w,C)}{l_C})$$

The second quantity amounts to :

$$(\alpha-\epsilon')(\beta+\epsilon')=\alpha\beta+\epsilon'(\alpha-\beta)-(\epsilon')^2$$

But $\alpha-\beta=(1-\lambda)\frac{c(w_a,d_j)}{l}$, a quantity which is strictly greater than $(1-\lambda)\frac{\epsilon}{l}=\epsilon'$ by the assumptions of the DF constraint. Thus the GRM model satisfies the DF constraint when Jelinek-Mercer is used. For the Dirichlet smoothing, setting :

$$\alpha=\frac{c(w,d)+\mu p(w|C)}{l+\mu},\ \beta=\frac{\mu p(w|C)}{l+\mu},\ \text{and}\ \epsilon'=\frac{\epsilon}{l+\mu}$$

leads to exactly the same development as above. The GRM model thus satisfies the DF constraint for both Jelinek-Mercer and Dirichlet smoothing. The use of the exponent $P(d|q)$ also shows that it satisfies the DS constraint.

4.2. *PRF under the Probability Ranking Principle*

Xu and Akella (Xu *et al.*, 2008) propose an instanciation of the Probability Ranking Principle (PRP) when documents are modelled with a Dirichlet Compound distribution. Instead of relying on the PRP to extract new terms, they propose a generative model of documents. In their PRP framework, relevant documents are assumed to come from a Dirichlet Compound Multinomial (DCM) distribution, the parameters of which will be denoted θ_w. In the feedback process, documents arise from a *a mixture of Extended DCM distributions*. Contrary to the mixture model, the mixing parameter for each document is not fixed. Furthermore, several modifications of the EM algorithm to moderate the biais of the generative approach are used. Those modifications are similar to the regularized mixture model studied in (Tao *et al.*, 2006). One can show that maximizing the EDCM likelihood leads to forget TF information. Only DF matters for the EDCM model. Let $s=\sum_{w=1}^{M}\theta_w$, then s verifies the following fixed-point equation :

$$s=\frac{\sum_w DF(w)}{\sum_d \Psi(s+l_d)-n\Psi(s)}$$

Once s is known, the θ_w can be obtained directly by :

$$\theta_w = \frac{DF(w)}{\sum_d \Psi(s + l_d) - n\Psi(s)}$$

It is then easy to show that maximizing the likelihood of an EDCM model entails the DF constraint.

4.3. *PRF in Divergence from Randomness (DFR) and Information Models*

In DFR and information models, the original query is modified, following standard approaches to PRF, to take into account the words appearing in F according to the following scheme :

$$x_w^{q'} = \frac{x_w^q}{\max_w x_w^q} + \beta \frac{\text{Info}_F(w)}{\max_w \text{Info}_F(w)} \quad [5]$$

where β is a parameter controlling the modification brought by F to the original query ($x_w^{q'}$ denotes the updated weight of w in the feedback query, whereas x_w^q corresponds to the weight in the original query).

4.3.1. *Bo2*

The standard PRF model in the DFR family is the Bo2 model (Amati *et al.*, 2003), which is defined by :

$$g_w = (\sum_{d \in F} l_d) p(w|C)$$

$$\text{Info}_{Bo2}(w) = \log_2(1 + g_w) + TF(w) \log_2(\frac{1 + g_w}{g_w})$$

In other words, documents in F are merged together. A Geometric probability model measures the informative content of a word. As this model is DF agnostic, it does not entail the DF constraint.

4.3.2. *Log-logistic Model*

For information models (Clinchant *et al.*, 2010), the average information this set brings on a given term w is used as a criterion to rank terms, which amounts to :

$$\text{Info}_F(w) = \frac{1}{n} \sum_{d \in F} -\log(P(X_w > t_w^d | \lambda_w))$$

The log-logistic model for pseudo relevance feedback is defined by :

$$t(w,d) = c(w,d) \log(1 + c\frac{avg\_l}{l_d}) \quad [6]$$

$$FW(w) = \sum_{d \in F} [log(\frac{n_w}{N} + t(w,d)) + IDF(w)] \quad [7]$$

As the log is a concave function, the log-logistic model satisfies the DF constraint by theorem 1. Similarly, the SPL model proposed in (Clinchant *et al.*, 2010) satisfies the DF constraint.

Having reviewed several state-of-the-art PRF models wrt to their behavior according to the DF constraint, we now turn to an experimental validation of this constraint.

5. Validation of the DF Constraint

We present here a series of experiments conducted in order to assess whether the DF constraint is a valid constraint in pseudo-relevance feedback. To do so, we first describe the oracle used to escape away from (and thus not being biased by) any given model.

5.1. *Oracle*

Suppose an oracle could tell the performance of each individual word in a pseudo-relevance feedback setting. Then, one could look at the oracle word statistics (mean TF, mean DF) in order to further validate the DF constraint. However, if we use such an oracle on pseudo-relevance feedback sets, there will likely be a significant variation of these statistics, since there is a significant variation in the precision at 10. Indeed, it is difficult to compare the TF statistics for a query with $P@10 = 0.1$ and for a query such that $P@10 = 0.9$. It is difficult to observe a global tendency in such a case. It is however possible to overcome the query variation performance by using true relevance feedback. The experimental setting we follow is thus defined as :

– Start with a first retrieval with a Dirichlet language model ;

– 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 DF, equal to $DF(w)/R_q$, and a normalized TF statistics (the TF is divided by the actual number of document used for relevance feedback, R_q). 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 can be computed as : $\Delta(MAP) = MAP(q+w) - MAP(q)$. We thus obtain, for each term, the following statistics :

– $\Delta(MAP)$
– $log(1 + TF(w))/R_q$
– $DF(w)/R_q$

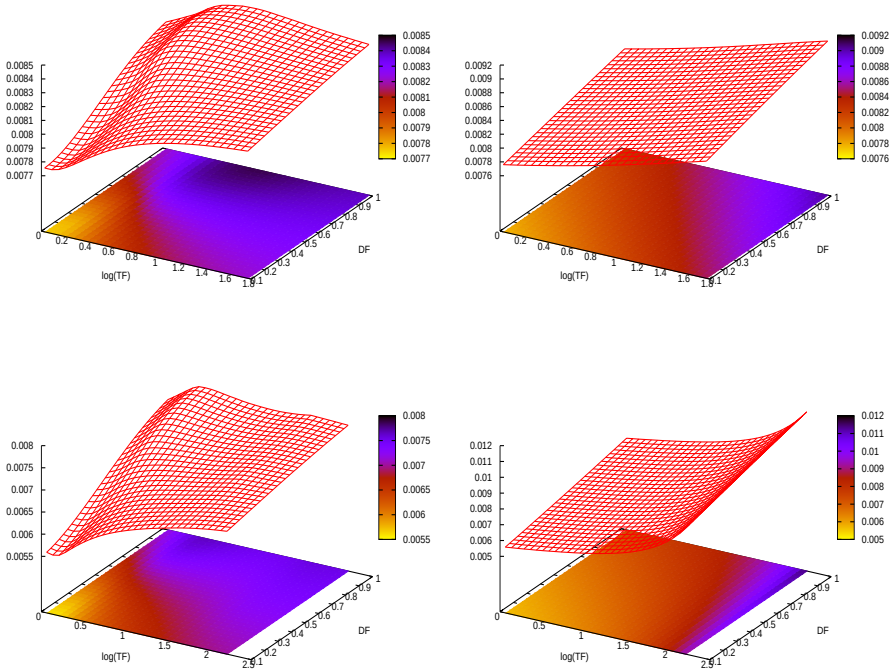

Figure 1. *(log(TF),DF) vs Δ MAP on ROBUST ; true releevant documents are used with $n = 10$, $tc = 50$ and exponential (left) and Gaussian (right) kernel grids (15×15). Top row : log-logistic ; bottom row : language model*

Figures 1 and 2 display a 3D view of these statistics for all queries, using Gnuplot with gaussian and exponential kernel estimators. On all plots, the best performing regions in the (TF,DF) space correspond to large DFs, thus showing the validity of the DF constraint. It has to be noted that the TF statistics was normalized to account for different lengths. In the figures, the DFR normalization was used, but the shape of the plot remains consistent without any normalization or when a language model normalization is used.

5.2. *Experimental Validation*

Theorem 1 can help us further validate the DF constraints. Indeed, let us use the family of feedback functions defined by :

$$FW = \sum_{d \in F} t(w,d)^k IDF(w) \qquad [8]$$

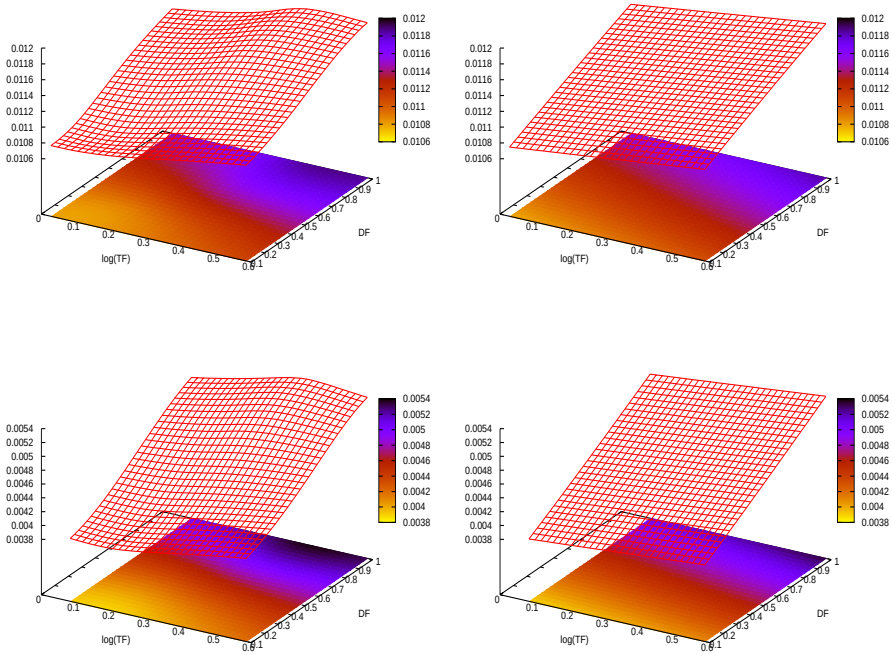

Figure 2. *(log(TF),DF) vs Δ MAP on TREC-12 ; true releevant documents are used with $n = 10$, $tc = 50$ and exponential (left) and Gaussian (right) kernel grids (15×15). Top row : log-logistic ; bottom row : language model*

with $t(w,d) = c(w,d)\log(1 + c\frac{avg\_l}{l_d})$, which corresponds to the second DFR normalization. Equation 8 amounts to a standard d *tf-idf* weighting, with an exponent k which allows one to control the convexity/concavity of the feedback model. According to Theore 1, if $k > 1$ then the function is strictly convex and does not satisfy the DF constraint. If $k < 1$, then the function is strictly concave and satisfies the DF constraint, while the linear case, being both concave and convex, is *in-between*. We can then build PRF models from equation 8 with varying k, and see whether the results agree with the theoretical findings implied by Theorem 1. We used these PRF models with equation 5 and a log-logistic model to assess their performance (as the log-logistic model was the best performing model in our preliminary experiments). Table 5 displays the term statistics ($\mu(tf)$,$\mu(df)$, mean IDF) for different values of k. As one can note, the smaller k, the bigger $\mu(df)$ is. In other words, the slowlier the function grows, the more terms with large DF are preferred. Table 6 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

Tableau 5. *Statistics on TREC-12-A*

Power k	$\mu(tf)$	$\mu(df)$	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

Tableau 6. *MAP for different power function. Suffix A means* $n = 10$ *and tc* $tc = 10$ *while suffix B means* $n = 20$ *and tc* $tc = 20$

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

datasets, as predicted by the DF cosntraint and Theorem 1. As convex functions do not entail the DF constraint, this suggests that the DF constraint is valid and leads to better performance. Second, the square root function ($k = 0.5$) has the best performances on all collections : it also outperforms the stadard 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. Conclusion

We have studied in this paper the behavior of several PRF models, and have displayed their main characteristics through a first series of experiments. This led us (a) to show that the divergence minimization PRF model was deficient wrt the IDF effect (i.e. this model selects terms with large IDF), and (b) to introduce a new heuristic constraint for PRF models, referred to as the *Document Frequency (DF) constraint*. We have then analyzed, from a theoretical point of view, state-of-the-art PRF models according

to their relation with this constraint. This analysis revealed that the standard mixture model for PRF in the language modeling family does not satisfy the DF constraint.

We have then conducted a series of experiments in order to see whether the DF constraint is valid or not. To do so, we performed tests with an oracle and a simple family of *tf-idf* functions based on a prameter k controlling the convexity/concavity of the function. Both the oracle and the results obtained with this family of functions validate the DF constraint. Furthermore, our experiments suggest that the square root function should be preferred over the mean log-logistic information model introduced in (Clinchant *et al.*, 2010) for pseudo-relevance feedback, as the square root funciton achieves a better tradeoff between the DF and TF statistics.

7. Bibliographie

Amati G., Carpineto C., Romano G., Bordoni F. U., « Fondazione Ugo Bordoni at TREC 2003 : robust and web track », 2003.

Clinchant S., Gaussier E., « Information-based models for ad hoc IR », *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, ACM, New York, NY, USA, p. 234-241, 2010.

Collins-Thompson K., Callan J., « Estimation and use of uncertainty in pseudo-relevance feedback », *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, ACM, New York, NY, USA, p. 303-310, 2007.

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

Dillon J. V., Collins-Thompson K., « A unified optimization framework for robust pseudo-relevance feedback algorithms », *CIKM*, p. 1069-1078, 2010.

Fang H., Tao T., Zhai C., « A Formal Study of Information Retrieval Heuristics », *SIGIR '04 : Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, 2004.

Fang H., Zhai C., « Semantic term matching in axiomatic approaches to information retrieval », *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, ACM, New York, NY, USA, p. 115-122, 2006.

Hoashi K., Matsumoto K., Inoue N., Hashimoto K., « Query expansion based on predictive algorithms for collaborative filtering », *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, ACM, New York, NY, USA, p. 414-415, 2001.

Lavrenko V., Croft W. B., « Relevance based language models », *SIGIR '01 : Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, New York, NY, USA, p. 120-127, 2001.

Lv Y., Zhai C., « A comparative study of methods for estimating query language models with pseudo feedback », *CIKM '09 : Proceeding of the 18th ACM conference on Information and knowledge management*, ACM, New York, NY, USA, p. 1895-1898, 2009.

Lv Y., Zhai C., « Positional relevance model for pseudo-relevance feedback », *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, ACM, New York, NY, USA, p. 579-586, 2010.

Seo J., Croft W. B., « Geometric representations for multiple documents », *SIGIR '10 : Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, ACM, New York, NY, USA, p. 251-258, 2010.

Tao T., Zhai C., « Regularized estimation of mixture models for robust pseudo-relevance feedback », *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, ACM, New York, NY, USA, p. 162-169, 2006.

Xu Z., Akella R., « A new probabilistic retrieval model based on the dirichlet compound multinomial distribution », *SIGIR '08 : Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, New York, NY, USA, p. 427-434, 2008.

Zhai C., Lafferty J., « Model-based feedback in the language modeling approach to information retrieval », *CIKM '01 : Proceedings of the tenth international conference on Information and knowledge management*, ACM, New York, NY, USA, p. 403-410, 2001.

Zhai C., Lafferty J., « A study of smoothing methods for language models applied to information retrieval », *ACM Trans. Inf. Syst.*, vol. 22, n° 2, p. 179-214, 2004.