Une analyse théorique des modèles de rétro-pertinence

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

1. *Xerox Research Center Europe*
   stephane.clinchant@xrce.xerox.com
2. *UJF-Grenoble 1/CNRS - LIG UMR 5217/AMA team*
   eric.gaussier@imag.fr

RÉSUMÉ. *Nous proposons dans ce papier une analyse théorique des modèles de Pseudo-Relevance Feedback (PRF), expliquant pourquoi certains modèles comme les modèles d'information sont plus performants. Pour ce faire, nous proposons un ensemble de propriétés pour les fonctions de PRF, dont une portant sur la fréquence documentaire (DF) qui est validée expérimentalement. Cette étude théorique révèle que plusieurs modèles de référence ne respectent pas la propriété sur l'effet IDF ou sur l'effet DF. Les modèles satisfaisant toutes les propriétés surpassent ces modèles en partie déficients.*

ABSTRACT. *Our aim in this paper is to provide a theoretical study of Pseudo-Relevance Feedback (PRF) models, explaining why some models, as the ones from the recently introduced information-based family, perform better than others. To do so, we propose general properties for PRF functions, including a Document Frequency (DF) effect, which is experimentally validated. The theoretical study we conduct reveals that several standard PRF models fail to enforce the IDF effect, and thus tend to select terms with a high TF and a low IDF. Such models fail to provide a good PRF weighting function, and are outperformed by other models able to enforce all the properties we review.*

MOTS-CLÉS : *modèles théoriques de RI, boucle de rétro-pertinence, modèles d'information.*

KEYWORDS: *IR theoretical models, axiomatic constraints, pseudo relevance feedback.*

DOI:10.3166/DN.15.1.125-146

1. Introduction

L'utilisation de *Pseudo-Relevance Feedback* [1], dans la suite notée PRF, a fait l'objet d'un grand nombre d'études depuis plusieurs décennies et beaucoup de modèles ont été proposés dans ce cadre. Rappelons rapidement le fonctionnement du PRF :

1. Nous utilisons ici le terme anglais faute d'une traduction adéquate en français.

1. Partant d'une requête, une première recherche est faite à l'aide d'un système de recherche d'information (RI) ; ce système fournit une liste de documents triés suivant un score de pertinence ;

2. On considère les n premiers documents (qui constitueront l'ensemble de *feedback*), desquels on extrait tc termes significatifs ;

3. Ces termes sont ensuite ajoutés à la requête initiale, de façon à produire une nouvelle requête plus complète ;

4. Enfin, les documents retrouvés par cette nouvelle requête (à l'aide du système de RI) sont fournis à l'utilisateur comme résultats de sa recherche.

Le terme *pseudo-relevance* exprime ici le fait que les n documents de l'ensemble de feedback ne sont pas nécessairement pertinents pour la requête ; ce sont juste les documents jugés les meilleurs par le système de RI. L'utilisation de cette stratégie permet en général, mais pas toujours, d'améliorer les rsultats obtenus par la requête originale.

Pour les modèles de langues, le modèle de mélange (Zhai, Lafferty, 2001) est considéré comme l'un des modèles de référence dans de nombreux travaux. En effet, il a été montré que ce modèle est un des plus performants et des plus robustes dans l'étude conduite par (Lv, Zhai, 2009b). Cependant, plusieurs études plus récentes sur les modèles de PRF montrent que ce modèle est surpassé par d'autres modèles comme ceux utilisant du *bagging*, des mélanges de distributions de type *Dirichlet Compound Multinomial* ou par le modèle log-logistique des modèles d'information (Collins-Thompson, Callan, 2007 ; Xu, Akella, 2008 ; Clinchant, Gaussier, 2010). Les performances de ces modèles varient néanmoins d'une étude à l'autre et différentes collections sont utilisées pour les évaluer. Il est donc très difficile de tirer des conclusions sur les caractéristiques des modèles car il n'existe pas de cadre théorique qui permettrait de comparer directement les modèles de PRF, indépendamment des collections utilisées. Le but de cette étude est précisément d'établir un tel cadre théorique, en établissant une liste de propriétés pour les modèles de PRF et en examinant les différents modèles de PRF au regard de ces propriétés. Cette analyse théorique fournit des explications sur les comportements expérimentaux pour plusieurs modèles de PRF et ouvre la voie à une évaluation théorique des modèles de PRF.

Les notations que nous utilisons sont résumées dans le tableau 1, où w représente un mot. Nous notons n le nombre de documents utilisés en PRF, F l'ensemble de documents de feedback et tc le nombre de mots ajoutés à la requête. Un important changement de notations concerne TF et DF qui sont dans cet article *calculés sur l'ensemble de feedback* F.

Le reste de cet article est organisé comme suit. Nous donnons dans la section 2 quelques statistiques sur plusieurs modèles de PRF qui révèlent les tendances générales de ces modèles. Nous introduisons ensuite plusieurs propriétés de PRF dans la section 3, avant de passer en revue les modèles de PRF selon leur capacité à prendre en compte ces propriétés dans la section 4. Finalement, nous discutons de travaux connexes dans la section 5.

Tableau 1. Notations

Notations	Description
Général	
q, d, w	Une requête, un document, un mot (ou terme)
l_d	Longueur de d (tokens)
$RSV(q,d)$	*Retrieval status value*, score de d pour q
$c(w,d)$	Nombre d'occurrences de w dans d
$t(w,d)$	Nombre d'occurrences normalisé (e.g. $\frac{c(w,d)}{l_d}$)
avg_l	Longueur moyenne des documents dans la collection
N	Nombre de documents dans la collection
N_w	Nombre de documents contenant w
$IDF(w)$	IDF (Inverse Document Frequency) de w (e.g. $-\log(N_w/N)$)
$tdfr(w,d)$	$c(w,d)\log(1+c\frac{\text{avg}_l}{l_d})$ (c paramètre)
$p(w\|C)$	Modèle de langue du corpus
Spécifiques au PRF	
n	Nbre de documents retenus pour le PRF
$\mathbf{F}$	Ensemble des documents retenus pour le PRF : $\mathbf{F} = (d_1, \ldots, d_n)$
tc	*TermCount*: Nbre de termes de $\mathbf{F}$ ajoutés à q
$TF(w)$	$=\sum_{d\in F} c(w,d)$
$DF(w)$	$=\sum_{d\in F} I(c(w,d)>0)$

2. Quelques caractéristiques des modèles de PRF

Nous commençons par présenter une comparaison des performances de différents modèles de PRF en fonction du paramètre tc, c'est-à-dire du nombre de termes choisis pour étendre la requête initiale. Cette comparaison porte sur : (a) le modèle de mélange (Zhai, Lafferty, 2001), que nous noterons MIX dans la suite, (b) le modèle de divergence KL [2] (Zhai, Lafferty, 2001), que nous noterons DIV, (c) le modèle log-logistique (Clinchant, Gaussier, 2010), que nous noterons LL, et (d) le modèle de pertinence géométrique (Geometric Relevance Model), que nous noterons GRM. Ces modèles sont présentés en détail dans la section 4 et leur formulation exacte n'est pas nécessaire ici. Pour tous ces modèles, tous les paramètres sont optimisés (avec le paramètre d'interpolation) sur toutes les requêtes d'une collection. La figure 1 montre les performances de ces quatre modèles quand le nombre de mots ajoutés à la requête (tc) varie. Comme nous pouvons le constater, le modèle log-logistique surpasse les autres modèles avec une MAP proche de 0.3 sur ROBUST et TREC 1&2 alors que la majorité des autres modèles ne dépasse pas 28.5 sur TREC et 29.0 sur ROBUST. Ceci étant, le fait le plus remarquable est que le modèle log-logistique n'a besoin que de 20

2. Divergence de Kullback-Leibler.

nouveaux mots pour obtenir d'excellentes performances, alors que les autres modèles obtiennent leur meilleure performance avec 100 mots pour ROBUST et 150 mots pour TREC. C'est typiquement ce genre de comportement que nous voulons comprendre dans cette étude.

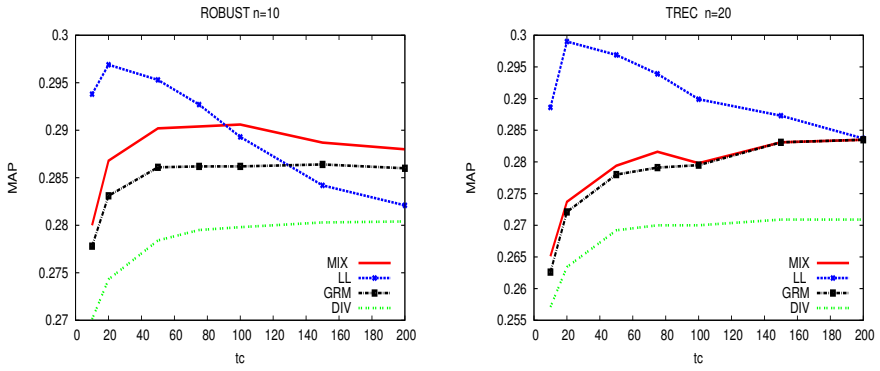

Figure 1. MAP sur toutes les requêtes avec $tc \in \{10, 20, 50, 75, 100, 150, 200\}$. *Les valeurs des paramètres ont été optimisées sur chaque collection, ROBUST* $n = 10$ *à gauche et TREC-1&2* $n = 20$ *à droite*

Pour cela, nous avons analysé les mots choisis par les 4 modèles précédemment mentionnés quand peu de mots sont choisis, à travers deux configurations de PRF : le cas A, avec $n = 10$ et $tc = 10$, et le cas B, avec $n = 20$ et $tc = 20$. Plusieurs études choisissent de comparer les modèles de PRF en fixant ces paramètres. Par exemple, (Hawking *et al.*, 1997) et (Lv, Zhai, 2010) utilisent $n = 20$ et $tc = 30$ dans leurs expériences. (Carpineto *et al.*, 2001) et (Amati *et al.*, 2004) utilisent $n = 10$ et $tc = 40$. Les modèles de pertinence sont eux utilisés avec différentes configurations : (Trevor *et al.*, 2004) choisit $n = 10$ and $tc = 15$ alors que (Cronen-Townsend *et al.*, 2004) prend $n = 50$ et $tc = 1000$. Nous choisissons le cas A et B afin de mieux comprendre les améliorations du modèle log-logistique.

Tout d'abord, nous utilisons le même modèle de RI pour la première étape de PRF de sorte que tous les modèles de PRF soient comparés sur le *même ensemble*. Une fois les nouvelles requêtes construites, nous utilisons soit le modèle de langue avec lissage de Dirichlet pour les modèles de PRF fondés sur les modèles de langues, soit le modèle log-logistique pour la deuxième étape de recherche. Ceci permet donc de comparer les performances des différents modèles de PRF toutes choses étant égales par ailleurs.

Nous comparons d'abord le modèle de mélange et le modèle log-logistique, en calculant l'accord sur les mots choisis par ces méthodes. Cela revient à calculer l'intersection de l'ensemble des mots choisis. Le tableau 2 indique la moyenne, médiane et variance de la taille de cette intersection pour toutes les requêtes d'une collection donnée. Comme on le voit, les méthodes s'accordent sur un peu plus de la moitié

Tableau 2. Statistique sur la taille de l'intersection des termes choisis par les modèles de mélange et log-logistique

Collection	n	tc	Moyenne	Médiane	Variance
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

des mots. On peut obtenir une autre comparaison des mots choisis en regardant le profil d'un mot. Le profil d'un mot est constitué de son nombre d'occurrences dans l'ensemble de feedback $TF(w)$, de sa fréquence documentaire dans l'ensemble de feedback (i.e. $DF(w)$) et de son $IDF(w)$ dans la collection. Pour chaque requête, les statistiques TF, DF et IDF sont moyennées sur tous les mots sélectionnés, puis sur l'ensemble des requêtes :

$$\mu(tf) = \sum_q \frac{1}{|Q|} \sum_{i=1}^{tc} \frac{TF(w_i)}{tc}$$

$$\mu(df) = \sum_q \frac{1}{|Q|} \sum_{i=1}^{tc} \frac{DF(w_i)}{tc}$$

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

Le tableau 3 montre les statistiques précédentes pour les quatre modèles de PRF : modèle de mélange (MIX), log-logistique (LL), modèle de divergence (DIV) et le modèle de pertinence géométrique (GRM). Le modèle de mélange choisit en moyenne des mots qui ont un plus grand TF comparé au modèle log-logistique. Ces mots sont aussi plus fréquents dans la collection (IDF plus petit). En revanche, on observe que les modèles de divergence et de pertinence choisissent des mots très fréquents (faible IDF et grand DF), ce qui est un de leur principal inconvénient.

Enfin, nous évaluons la qualité des mots choisis dans quatre cas différents. Nous calculons d'abord la performance des mots choisis *sans* les mélanger avec la requête initiale, cas que nous appelons *raw*. Dans un deuxième temps, nous ne gardons que les mots qui appartiennent à l'intersection des deux modèles, avec les poids prédits par ces modèles ; nous appelons ce cas *interse*. Le troisième cas, appelé *diff*, mesure la performance des mots n'appartenant pas à l'intersection. Enfin, le dernier cas, *interpo* pour interpolation, est le cas classique du PRF lorsque les mots sont ajoutés à la requête initiale. Le tableau 4 montre les résultats pour le modèle de mélange et log-logistique (les deux autres modèles se comportant comme le modèle de mélange). Ces résultats indiquent que le modèle log-logistique a tendance à choisir des termes plus intéressants pour le feedback que le modèle de mélange (cas *diff*). De plus, pour

Tableau 3. Statistiques sur les mots extraits par diverses méthodes de PRF. Le suffixe -A correspond à $n = 10$ et $tc = 10$ alors que -B correspond à $n = 20$ et $tc = 20$

Collection	Statistique	MIX	LL	DIV	GRM
robust-A	$\mu(tf)$	62.9	46.7	53.9	52.33
	$\mu(df)$	6.4	7.21	8.55	8.4
	$\mu(idf)$	4.33	5.095	2.20	2.40
trec-1&2-A	$\mu(tf)$	114 .0	79.12	92.6	92.27
	$\mu(df)$	7.1	7.8	8.77	8.72
	$\mu(idf)$	3.84	4.82	2.51	2.56
robust-B	$\mu(tf)$	68.6	59.9	65.27	64.57
	$\mu(df)$	9.9	11.9	14.7	14.38
	$\mu(idf)$	4.36	4.37	1.66	1.93
trec-1&2-B	$\mu(tf)$	137.8	100.0	114.9	114.8
	$\mu(df)$	12.0	13.43	15.17	15.23
	$\mu(idf)$	3.82	4.29	2.10	2.25

les termes choisis par les deux modèles (cas *interse*), le modèle log-logistique semble choisir une meilleure pondération que le modèle de mélange.

Tableau 4. Performances (MAP) des différentes méthodes. Le suffixe -B correspond à $n = 20$ et $tc = 20$

Collection	Ensemble de mots	MIX	LL
robust-B	raw	23.7	25.7
	interse	25.3	26.2
	diff	3.0	10.0
	interpo	28.2	28.5
trec-1&2-B	raw	25.1	27.0
	interse	26.1	26.5
	diff	2.1	11.2
	interpo	27.3	29.4

Résumons nos observations :

(a) Le modèle log-logistique (LL) surpasse les autres modèles ;

(b) Les autres modèles choisissent des mots avec un plus fort TF ;

(c) Le modèle de mélange (MIX) choisit des mots avec un plus faible DF ;

(d) Les modèles GRM et DIV choisissent des mots ayant un faible IDF.

3. Une analyse théorique des modèles de PRF

Nous introduisons dans cette section une caractérisation des modèles de PRF qui permet de mieux comprendre le comportement de ces modèles d'un point de vue théorique. Notre approche s'inspire des méthodes axiomatiques (Fang *et al.*, 2004)

et s'inscrit dans la lignée de plusieurs études (Fang, Zhai, 2006 ; Cummins, O'Riordan, 2007 ; Clinchant, Gaussier, 2010). Si l'approche axiomatique vise à décrire les fonctions d'ordonnancement en RI par des contraintes que ces dernières devraient satisfaire, nous considérons plutôt ici des propriétés générales qui permettent d'appréhender théoriquement les modèles de PRF.

Selon (Clinchant, Gaussier, 2010), les propriétés principales caractérisant la fonction de pondération d'un modèle de RI sont le fait que cette fonction doit être croissante et concave suivant le nombre d'occurrences des mots, décroissante suivant l'IDF et pénaliser les documents longs. Nous allons par la suite adapter ces propriétés au cas du PRF.

Soit $FW(w; \mathbf{F}, \mathbf{P}_w)$ la fonction de feedback pour le mot w, avec $\mathbf{P}_w$ un ensemble de paramètres dépendant de w [3]. Nous utilisons aussi la notation abrégée $FW(w)$ mais il est important de garder à l'esprit que cette fonction dépend d'un ensemble de feedback et de paramètres. Nous pouvons alors formaliser les quatre effets principaux des modèles de RI pour le PRF par :

[Effet TF] *FW croît avec $t(w,d)$:*

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

[Effet CONC] *Cette croissance est moins forte lorsque le nombre d'occurrences augmente, ce qui se traduit par le fait que la fonction doit être concave sur $t(w,d)$:*

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

[Effet IDF] *Soit w_a et w_b deux mots tels que $IDF(w_b) > IDF(w_a)$ et $\forall d \in \mathbf{F},\ t(w_a,d) = t(w_b,d)$. Alors*

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

[Effet LD] *Le nombre d'occurrences des mots doit être normalisé par la longueur des documents, ce qui se traduit par :*

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

La condition sur l'effet IDF est guidée par la cadre du PRF dans lequel nous nous plaçons et où nous voulons étudier l'augmentation de la fonction de PRF *ceteris paribus sic stantibus* [4].

3. $\mathbf{P}_w$ dépend du modèle de PRF considéré et contient minimalement $TF(w)$, et souvent d'autres éléments comme $IDF(w)$ par exemple.

4. « Toutes choses étant égales par ailleurs ».

Nous introduisons maintenant une nouvelle condition de PRF, qui se fonde sur les observations de la section précédente. En effet, nous avons vu que les meilleurs résultats de PRF étaient obtenus par le modèle log-logistique qui favorise les mots avec une grande fréquence documentaire ($DF(w)$), ce qui suggère que les mots avec un plus grand DF doivent être favorisés.

[Effet DF] *Soient $\epsilon > 0$, et w_a et w_b deux mots tels que :*
(i) IDF(a) = IDF(b)
(ii) Les distributions des occurrences de w_a et w_b dans l'ensemble F sont telles que :

$$T(w_a) = (t_1, t_2, ..., t_j, 0, ..., 0) \quad T(w_b) = (t_1, t_2, ..., t_j - \epsilon, \epsilon, ..., 0)$$

avec : $\forall i$, $t_i > 0$ et $t_j - \epsilon > 0$. Nous avons donc : $TF(w_a) = TF(w_b)$ et $DF(w_b) = DF(w_a) + 1$).
Alors: $FW(w_a; \mathbf{F}, \mathbf{P}_{w_a}) < FW(w_b; \mathbf{F}, \mathbf{P}_{w_b})$

En d'autres termes, FW doit être *localement* croissante en $DF(w)$.

Le théorème suivant permet d'établir si un modèle de PRF s'écrivant comme une somme de fonctions de pondération sur chacun des documents de feedback satisfait la contrainte DF.

THÉORÈME 1. — *Supposons que FW puisse s'écrire comme :*

$$FW(w; \mathbf{F}, \mathbf{P}_w) \quad = \quad \sum_{d=1}^{n} f(t(w,d); \mathbf{P'}_w) \tag{1}$$

avec : $\mathbf{P'}_w = \mathbf{P}_w \setminus t(w,d)$ et $f(0; \mathbf{P'}_w) \geq 0$. Alors :

1. *Si la fonction f est strictement concave, alors FW satisfait l'effet DF.*
2. *Si la fonction f est strictement convexe, alors FW ne satisfait pas l'effet DF.*

Démonstration : Si f est strictement concave alors, la fonction f est sous-additive ($f(a+b) < f(a) + f(b)$). Soient a et b deux mots vérifiant les conditions de l'effet DF :

$$FW(a) = FW(\underbrace{t_1, ..., t_j}_{DF}, \underbrace{0, ..., 0}_{n-DF}) \tag{2}$$

Nous avons alors :

$$FW(b) - FW(a) = f(t_j - \epsilon) + f(\epsilon) - f(t_j) \tag{3}$$

Cette quantité est strictement positive si f est sous-additive. Si la fonction f est strictement convexe, alors f est super-additive (car $f(0) = 0$), ce qui montre que $FW(b) - FW(a) < 0$. □

On peut remarquer aussi que la somme de fonctions concaves est concave et les fonctions de feedback définies par l'équation 1 vérifient donc l'effet concave en même

temps que l'effet DF. Nous verrons cependant qu'il existe des modèles qui satisfont la condition DF mais pas la condition de concavité.

Avant de valider cet effet DF, nous voulons mentionner une dernière caractérisation, introduite dans (Lv, Zhai, 2009b) et qu'on appellera effet *Document Score*. Cet effet est pris en compte dans les modèles de pertinence (Lavrenko, Croft, 2001) et dans certains algorithmes inspirés de l'algorithme de Rocchio (Hoashi *et al.*, 2001). Il se formule comme suit :

[Effet Document Score (DS)] *Quand* $FW(w; \mathbf{F}, \mathbf{P}_w)$ *dépend explicitement des documents de* $\mathbf{F}$ *où* w *apparaît, alors les documents avec un plus grand score (obtenu lors de la première recherche) devraient avoir plus d'importance.*

L'intérêt de cet effet est cependant limité et les modèles qui le prennent en compte ne figurent pas parmi les meilleurs modèles de PRF. Par exemple, dans l'étude conduite par (Lv, Zhai, 2009b), le modèle de mélange surpasse des modèles utilisant les scores des documents. Nous discuterons dans la section 5 des stratégies (Collins-Thompson, Callan, 2007) (Lee *et al.*, 2008) qui consistent à échantillonner des documents de feedback et qui pourraient être plus adaptées pour rendre compte de cet effet.

3.1. Validation de l'effet DF

Une manière de valider l'effet DF est de déterminer si les valeurs de DF sont liées aux scores MAP dans la boucle de pertinence. L'effet DF stipule que, toutes choses étant égales par ailleurs, les mots avec un plus grand DF doivent être favorisés. Si cet effet est valide, on doit observer que les mots avec un fort DF contribuent à une augmentation du score MAP. Par conséquent, nous avons calculé l'impact sur la MAP des différents mots sélectionnés à partir des vrais jugements de pertinence et nous avons visualisé cet impact en fonction des variables TF et DF. Nous avons choisi d'utiliser de vrais document pertinents et non des documents de PRF, pour deux raisons : tout d'abord car le cadre PRF vise à approcher la vraie boucle de pertinence, et ensuite car il est plus difficile d'observer une tendance claire en PRF quand la précision à 10 et la MAP ont une large variance. L'utilisation de vrais documents pertinents est donc ici plus propre et facilite l'interprétation.

Nous avons suivi le processus suivant :

1. Faire une première recherche avec un modèle de langue avec lissage de Dirichlet ;

2. Soit R_q l'ensemble des documents pertinents pour la requête q ; retenir les 10 premiers documentes pertinents si possible, sinon garder les $|R_q|(|R_q| < 10)$ documents pertinents ;

3. Extraire sur cet ensemble une nouvelle requête de 50 mots avec le modèle de mélange ;

4. Faire de même avec le modèle log-logistique ;

5. Calculer les TF et DF normalisés, ainsi que $\Delta(\text{MAP})$ (cf ci-dessous).

Pour chacun des mots ainsi obtenus, nous calculons une version normalisée de DF, égale à $DF(w)/|R_q|$, et une version normalisée de TF par la longueur des documents. Nous utilisons dans nos expériences la normalisation $TF(w) = \sum_{d \in R_q} tdfr(w,d)$ sur laquelle nous appliquons la transformation : $\log(1 + TF(w))/|R_q|$. Cette transformation permet d'éviter une trop grande dispersion dans les graphiques. Chaque mot est ajouté indépendamment avec le poids prédit par le modèle de PRF. Pour chaque mot, on mesure alors la MAP de la requête initiale augmentée de ce dernier. La différence de performance avec la requête initiale est ensuite calculée par : $\Delta(\text{MAP}) = \text{MAP}(q \cup w) - \text{MAP}(q)$. Nous obtenons donc, pour chaque mot, les statistiques suivantes :

$$\Delta(\text{MAP}),\ \log(1 + TF(w))/|R_q| \text{ et } DF(w)/|R_q|$$

Les graphes de la figure 2 montrent une vue 3D de ces statistiques pour toutes les requêtes et deux collections: TREC1&2 et ROBUST. Afin d'avoir une meilleure vue des motifs obtenus, nous avons utilisé une grille 30x30 et des noyaux de lissage. L'utilisation d'un noyau de lissage K revient à lisser la valeur en chaque point par la valeur des points voisins comme suit :

$$\Delta(\text{MAP})(x) = \sum_{i=1}^{p} \Delta(\text{MAP})(x_i) K(x, x_i)$$

où $i, 1 \leq i \leq p$ indexe l'ensemble des mots de feedback. Nous avons testé des noyaux exponentiel et gaussien, en obtenant à chaque fois une allure similaire. Nous ne présentons donc ici que les résultats avec un noyau gaussien.

Comme on peut le remarquer, sur tous les graphes de la figure 2, les meilleures régions dans l'espace (TF,DF) correspondent à de grands DF. De plus, pour toutes les valeurs de TF, les mots avec un fort DF pour la boucle de pertinence.

3.2. Validation des différents effets avec une famille TF-IDF

Afin de compléter cette validation des différents effets que nous avons examinés, nous considérons maintenant la famille de fonctions de feedback définie par :

$$FW(w) = \sum_{d \in F} tdfr(w,d)^k \text{IDF}(w) \tag{4}$$

où $tdfr$ est donné dans le tableau 1 et correspond à la normalisation dans les modèles DFR et d'information. Ces modèles encodent simplement une pondération du type *tf-idf*, avec un exposant k qui permet de contrôler la convexité/concavité des modèles de feedback.

Du fait de la forme de $t(w,d)$ et de la manière dont IDF(w) est pris en compte, cette famille de fonctions respecte les effets TF, IDF et LD. Si $k > 1$, alors la fonction est strictement convexe et, d'après le théorème 1, ne satisfait pas l'effet DF. Dans ce

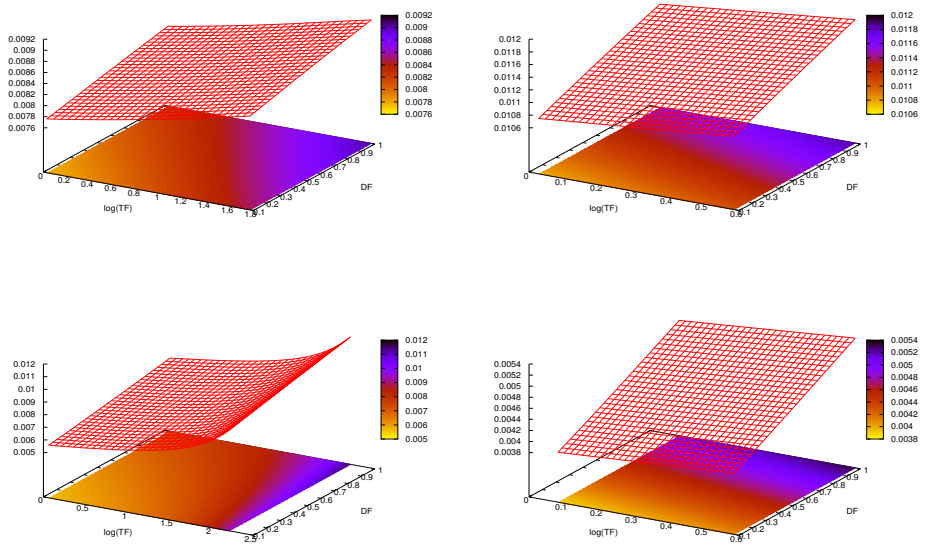

Figure 2. (log(TF),DF) vs Δ MAP. Nous considérons ici de 'vrais' documents pertinents avec $n = 10$ et $tc = 50$. Un noyau gaussien sur une grille (30×30) est utilisé pour lisser les courbes obtenues. En haut: modèle log-logistique, en bas : modèle de mélange. À gauche : collection ROBUST, à droite : collection TREC-12

cas, nous avons également $\forall d \in \mathbf{F}$, $\frac{\partial^2 FW(w)}{\partial t(w,d)^2} > 0$, et donc l'effet CONC n'est pas non plus vérifié. En revanche, si $k < 1$, alors la fonction est strictement concave et satisfait les effets DF et CONC.

Tableau 5. Statistiques des modèles avec différents exposants (équation 4) sur TREC-12-A

Exposant k	$\mu(tf)$	$\mu(df)$	$\mu(IDF)$
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

On peut donc construire des modèles de PRF à partir de l'équation 4 en variant k et observer si les résultats concordent avec notre analyse théorique. Nous utilisons ensuite l'équation 4 avec l'équation 12 et un modèle log-logistique pour mesurer leur

performance. Le tableau 5 montre les statistiques ($\mu(tf)$,$\mu(df)$, mean IDF) pour différentes valeurs de k. Comme on peut le voir, plus k est petit, plus $\mu(df)$ est grand. En d'autres termes, plus la fonction croît lentement, plus les termes avec un fort DF sont préférés. Le tableau 6 indique la MAP obtenue pour plusieurs valeurs de k. Au moins deux points importants se dégagent de ces résultats. D'abord, les fonctions convexes ont des performances moindres que les fonctions concaves, et moins un modèle respecte les conditions théoriques, moins il est performant. Ceci confirme encore la validité des conditions que nous avons introduites. Deuxièmement, la racine carrée ($k = 0.5$) obtient les meilleures performances sur toutes les collections : elle surpasse même le modèle log-logistique. Quand la fonction croît lentement, ($k = 0.2$), la statistique DF est privilégiée par rapport au TF. La racine carrée offre donc un meilleur compromis entre les informations TF et DF. Ceci est intéressant car cela montre que l'effet TF est important et qu'il ne faut pas trop l'atténuer au profit de l'effet DF.

Maintenant que nous avons introduit les conditions que doivent satisfaire les modèles de PRF, nous allons passer en revue différents modèles existants et regarder comment ils se situent par rapport à ces conditions.

Tableau 6. Score MAP pour différents exposants (équation 4). Le suffixe -A correspond à $n = 10$ et $tc = 10$ et le suffixe -B à $n = 20$ et $tc = 20$. A titre de comparaison, nous donnons aussi les performances du modèle LL

Exposant 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
LL	29.4	28.7	28.5	29.9

4. Revue des modèles de PRF

Nous examinons dans cette section différents modèles de PRF selon leur adéquation aux conditions que nous avons définies. Nous commençons par les modèles de langues avant de passer aux modèles issus du *Probability Ranking Principle* (PRP), puis aux modèles *Divergence from Randomness* (DFR) et aux modèles d'information.

4.1. PRF pour les modèles de langue

Les modèles de PRF pour les modèles de langues (LM) font l'hypothèse que les mots des documents de feedback suivent une distribution multinomiale, θ_F (la nota-

tion θ_F résume l'ensemble des paramètres $P(w|\theta_F)$). Une fois les paramètres estimés, le modèle de langue θ_F est interpolé avec la requête initiale par :

$$\theta_{q'} \quad = \quad \alpha\theta_q + (1-\alpha)\theta_F \tag{5}$$

En pratique, on restreint θ_F aux tc mots les plus importants en annulant la contribution des autres mots. Les méthodes de feedback diffèrent sur leur façon d'estimer θ_F. Nous examinons les principales méthodes dans la suite.

4.1.1. Modèle de mélange

(Zhai, Lafferty, 2001) proposent un modèle génératif pour l'ensemble $\mathbf{F}$. Tous les documents de $\mathbf{F}$ sont supposés être identiquement et indépendamment distribués (ce que l'on note *i.i.d*) et chaque document est considéré comme le résultat du mélange d'une distribution multinomiale propre à la requête et d'un modèle de langue de la collection :

$$P(\mathbf{F}|\theta_F, \beta, \lambda) = \prod_{w=1}^{V} ((1-\lambda)P(w|\theta_F) + \lambda P(w|C))^{TF(w)} \tag{6}$$

où λ est le paramètre de mélange. Enfin θ_F est appris en optimisant la vraisemblance des données avec un algorithme EM (*Expectation-Maximization*), dont les étapes E et M à l'itération (i) sont:

$$Etape\ E: \quad (E[w])^{(i)} = \frac{(1-\lambda)P^{(i}(w|\theta_F)}{(1-\lambda)P^{(i)}(w|\theta_F)+\lambda P^{(i)}(w|C)}$$

$$Etape\ M: \quad P^{(i+1)}(w|\theta_F) = \frac{\sum_{d\in\mathbf{F}} c(w,d)(E[w])^{(i)}}{\sum_w \sum_{d\in\mathbf{F}} c(w,d)(E[w])^{(i)}}$$

où $E[w]$ dénote l'espérance d'observer w dans F. Ici $FW(w) = P(w|\theta_F)$. On peut remarquer qu'aucune de ces formules ne met en jeu $DF(w)$, que ce soit directement ou indirectement. Le modèle de mélange est donc agnostique par rapport à DF et ne satisfait donc pas l'effet DF. Pour les autres propriétés, on peut voir que les poids des mots de feedback ($P(w|\theta_F)$) augmentent avec $TF(w)$ (qui est $\sum_{d\in\mathbf{F}} c(w,d)$) et diminuent avec IDF$(w)$ (l'argument est le même que celui développé dans (Fang *et al.*, 2004)). Ainsi, les effets TF et IDF sont pris en compte par ce modèle. De plus, même si les occurrences sont normalisées par la longueur des documents de feedback (en fait une approximation de celle-ci), tous ces documents sont fusionnés, de sorte que l'effet LD portant sur la normalisation par la longueur des documents n'est pas pleinement pris en compte. Par rapport à la concavité, la situation est encore moins claire. Si l'on approche le dénominateur de l'étape M par la longueur des documents de feedback (en utilisant le fait que $E[w]$ correspond à l'espérance de w dans F), alors la seconde dérivée partielle de $P(w|\theta_F)$ par rapport $c(w,d)$ est nulle. Ceci suggère que le modèle ne prend pas en compte l'effet CONC, et donc qu'il donne trop de poids aux mots avec un fort TF. C'est en effet ce que nous avons observé dans le tableau 3, où le modèle de mélange tend à choisir des mots avec un TF significativement plus grand que les autres modèles.

4.1.2. Minimisation de la divergence

Un modèle minimisant la divergence de Kullback-Leibler a été aussi proposé dans (Zhai, Lafferty, 2001) :

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

où θ_{d_i} correspond au modèle de langue du document d_i. La minimisation de cette divergence fournit la solution suivante [5] :

$$P(w|\theta_F) \propto \exp\Big(\frac{1}{(1-\delta)}\frac{1}{n}\sum_{i=1}^{n}\log(p(w|\theta_{d_i})) - \frac{\delta}{1-\delta}\log(p(w|C)\Big) \qquad (7)$$

Ici encore, $FW(w) = P(w|\theta_F)$. Cette équation correspond à la forme donnée dans l'équation 1 avec une fonction strictement concave. Le théorème 1 nous permet donc de conclure que ce modèle respecte l'effet DF. Il prend de plus en compte les effets TF, CONC et LD. Toutefois, nos expériences précédentes ont montré que ce modèle obtient de moins bons résultats que les autres (cf tableau 4 et l'étude dans (Lv, Zhai, 2009b)). Le tableau 3 montre ainsi que l'effet IDF n'est pas assez pris en compte et ce modèle choisit trop souvent des mots fréquents. En effet, considérons deux mots w_a and w_b tels que : (a) $\forall d \in \mathbf{F},\ t(w_a, d) = t(w_b, d) = t_d$ et (b) $p(w_a|C) < p(w_b|C)$. L'effet IDF stipule que dans ce cas $FW(a)$ devrait être plus grand que $FW(b)$. Avec le lissage de Jelinek Mercer (mais le raisonnement est le même pour celui de Dirichlet), $\log(FW(w_a)) - \log(FW(w_b))$ est égal à (nous ne présentons pas la dérivation ici, qui est purement technique) :

$$\sum_{d\in\mathbf{F}}\{\overbrace{\log(\frac{(1-\lambda)t(d)+\lambda p(w_a|C)}{(1-\lambda)t(d)+\lambda p(w_b|C)})}^{<0} - \delta\overbrace{\log(\frac{p(w_a|C)}{p(w_b|C)})}^{<0}\} \qquad (8)$$

Comme $0 < \lambda, \delta \leq 1$, nous avons :

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

Ainsi $\log(FW(w_a)) - \log(FW(w_b)) > 0$ et $FW(w_a) > FW(w_b)$. Le modèle de minimisation de la divergence satisfait donc bien l'effet IDF. Cependant, cet effet n'est en réalité que faiblement pris en compte. En effet, supposons que $P(w_b|C)$ est K fois ($K > 1$) plus grand que $P(w_a|C)$: $P(w_b|C) = KP(w_a|C)$. Dans ce cas :

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

5. Nous renvoyons les auteurs intéressés par le détail de cette minimisation à (Zhai, Lafferty, 2001).

et :

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

Le facteur K originel se traduit finalement, pour la pondération des termes de PRF, en K^{δ}. En pratique, les valeurs obtenues pour δ snt proches de 0.1 ; dans ce cas, K^{δ} est proche de 1 (il est égal à 1.07 pour $K = 2$, 1.17 pour $K = 5$ et 1.58 pour $K = 100$) et il n'y a presque aucune différence entre $FW(w_a)$ et $FW(w_b)$. Ceci explique les faibles statistiques d'IDF dans le tableau 3, ainsi que les moins bons résultats obtenus par ce modèle.

4.1.3. Autres modèles

Une version régularisée du modèle de mélange, connue sous le nom de *regularized mixture model (RMM)*, est proposée dans (Tao, Zhai, 2006) pour corriger certaines déficiences du modèle de mélange. RMM permet d'estimer conjointement les scores de pertinence des documents et des thèmes (distributions sur les mots), ce qui en fait *a priori* un modèle robuste par rapport aux paramètres de feedback. Cependant, les expériences menées dans (Lv, Zhai, 2009b) montrent que ce modèle est moins efficace que le modèle de mélange en terme de performance. Nous ne l'étudierons donc pas plus avant.

Une autre famille intéressante est la famille des *relevance models* (modèles de pertinence), proposée (Lavrenko, Croft, 2001) et définie par :

$$FW(w) \quad \propto \quad \sum_{d \in \mathbf{F}} P_{LM}(w|\theta_d) P(d|q) \tag{9}$$

où P_{LM} correspond au modèle de langue lissé du document. La formule ci-dessus correspond à la forme donnée dans l'équation 1 du théorème 1, avec une fonction linéaire, qui est ni strictement concave ni strictement convexe. Ce modèle est donc neutre par rapport à l'effet DF. De plus, il est direct de montrer que les modèles de pertinence *échouent* à prendre en compte l'effet IDF.

Ces modèles ont été récemment étendus dans (Seo, Croft, 2010) en utilisant l'idée d'une moyenne géométrique, noté GRM et définie par :

$$FW(w) \propto \prod_{d \in \mathbf{F}} P_{LM}(w|\theta_d)^{P(d|q)} \tag{10}$$

Nous montrons ici que ce modèle respecte bien l'effet DF. Considérons en effet ce modèle avec un lissage de Jelinek-Mercer (Zhai, Lafferty, 2004): $P_{LM}(w|\theta_d) = (1-\lambda)\frac{c(w,d)}{l_d} + \lambda\frac{c(w,C)}{l_C}$, où $c(w,C)$ correspond au nombre d'occurrences de w dans la collection C et l_C à la longueur de la collection. Soient w_a et w_b deux mots vérifiant les conditions de l'effet DF. Supposons que tous les documents de feedback ont la

même longueur l et sont équiprobables étant donné q. Soient A, B et ϵ' les quantités suivantes :

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

où d_j est le document défini dans l'efet DF. Nous avons :

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

De plus : $(A-\epsilon')(B+\epsilon') = AB+\epsilon'[(A-B)-\epsilon']$. Or $A-B = (1-\lambda)\frac{c(w_a,d_j)}{l}$ est une quantité strictement supérieure à $(1-\lambda)\frac{\epsilon}{l} = \epsilon'$ d'après les hypothèses de l'effet DF. Le modèle GRM, avec lissage de Jelinek-Mercer, satisfait donc bien l'effet DF. Un développement similaire conduit à la même conclusion pour le lissage de Dirichlet.

Cependant, un des défauts majeurs de ce modèle est qu'il ne satisfait pas l'effet IDF. En effet, soient w_a et w_b deux mots tels que : (a) $\forall d \in \mathbf{F},\ t(w_a,d) = t(w_b,d) = t_d$ et (b) $p(w_a|C) > p(w_b|C)$. Alors, $FW(w_a) - FW(w_b)$ a le signe de :

$$\sum_d P(d|q) \log \frac{\lambda t_d + (1-\lambda)p(w_a|C)}{\lambda t_d + (1-\lambda)p(w_b|C)} \tag{11}$$

qui est strictement positif. Ceci explique les résultats du tableau 3 indiquant que le modèle GRM choisit des mots avec un faible IDF.

4.2. Modèles de PRF pour le PRP

(Xu, Akella, 2008) utilisent le cadre du *Probability Ranking Principle (PRP)* (Robertson, 1977) dans lequel les documents pertinents sont supposés être générés à partir d'une distribution de type *Dirichlet Compound Multinomial (DCM)*, ou l'approximation de celle-ci appelée eDCM et introduite dans (Elkan, 2006). L'extension au PRF de ce modèle suppose simplement que les documents de feedback sont pertinents et que les mots sont générés à partir d'un mélange de deux distributions eDCM. À chacune de ces distributions correspond une variable z_{FR} pour la composante pertinente du mélange censée représenter les documents de feedback, et une variable z_N pour capturer le modèle de langue général du corpus. Les paramètres de ce modèle sont ensuite estimés par des algorithmes plutôt coûteux de type descente de gradient ou algorithme EM (plusieurs modifications de l'algorithme EM, suivant les suggestions de (Tao, Zhai, 2006), sont en fait testées). Au final, en mettant de côté la partie non pertinente du modèle de mélange (z_N), le poids affecté aux mots de de l'ensemble de feedback est donné par l'étape M de l'algorithme EM :

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

Cette formule, fondée sur la présence/absence des termes dans l'ensemble de feedback, est bien compatible avec l'effet DF. En revanche, l'effet TF n'est pas pris en compte ici. Enfin, les étapes de l'algorithme EM suggèrent aussi que le modèle satisfait l'effet IDF, comme le modèle de mélange.

4.3. PRF pour les modèles DFR et les modèles d'information

Pour les modèles DFR et d'information, la requête initiale est modifiée pour prendre en compte les mots de $\mathbf{F}$ comme suit :

$$x_w^{q'} = \frac{x_w^q}{\max_w x_w^q} + \beta \frac{\text{Info}(w, \mathbf{F})}{\max_w \text{Info}(w, \mathbf{F})} \quad (12)$$

où β est un paramètre contrôlant la modification apportée par $\mathbf{F}$ à la requête initiale ($x_w^{q'}$ correspond au nouveau poids de w dans la nouvelle requête, et x_w^q au poids de w dans la première requête). Dans ce cas : $FW(w) = \text{Info}(w, \mathbf{F})$.

4.3.1. Modèles DFR : modèle Bo

Les modèles standard dans la famille DFR sont les modèles Bo (Amati *et al.*, 2003), définis par :

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

où $g_w = \frac{N_w}{N}$ pour le modèle $Bo1$ et $g_w = P(w|C)(\sum_{d \in F} l_d)$ pour le modèle $Bo2$. Autrement dit, les documents de $\mathbf{F}$ sont concaténés et une distribution géométrique [6] est utilisée pour mesurer le contenu informatif des mots. De par l'absence d'information de type DF (absence résultant ici de la concaténation des documents de feedback), ces modèles ne satisfont pas l'effet DF. De plus, le choix de la distribution géométrique implique que l'effet CONC n'est pas pris en compte car la deuxième dérivée de $FW(w)$ en $TF(w)$ est nulle. Enfin, ces modèles ne prennent pas en compte les longueurs des documents de feedback puisque ces documents sont fusionnés.

4.3.2. Modèles d'information

Pour les modèles fondés sur l'information (Clinchant, Gaussier, 2010), l'information moyenne d'un mot w dans l'ensemble de feedback sert de critère pour sélectionner les mots :

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

où $t(w, d)$ est le nombre d'occurrences normalisé de w dans d, et λ_w un paramètre associé à w et égal à $\lambda_w = \frac{N_w}{N}$. Deux exemples de modèles d'information sont donnés

6. Le choix de cette distribution n'est pas important pour le développement sur l'effet DF. Il l'est en revanche pour l'effet CONC.

dans (Clinchant, Gaussier, 2010), utilisant soit une loi log-logistique soit une loi de puissance lissée. Nous nous concentrons ici sur le modèle log-logistique qui fournit les meilleurs résultats dans les expériences décrites dans (Clinchant, Gaussier, 2010). Le modèle log-logistique (LL) est donné par :

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

où $tdfr(w,d)$ est donné dans le tableau 1. Comme le logarithme est une fonction concave, le modèle log-logistique prend en compte l'effet DF d'après le théorème 1. Il respecte de plus les effets TF, CONC et LD par construction (cf. (Clinchant, Gaussier, 2010)). Pour l'effet IDF, considérons deux mots w_a et w_b vérifiant les conditions de cet effet (en particulier, nous avons $\lambda_b < \lambda_a$). Alors :

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

qui est toujours strictement négatif. L'effet IDF est donc bien pris en compte par ce modèle, qui vérifie toutes les propriétés théoriques développées dans la section 3.

4.4. Résumé

Notre étude théorique a révélé les éléments suivants sur les modèles de PRF, éléments qui sont résumés dans le tableau 7 :

1. Pour les modèles fondés sur le modèle de langue, le modèle de mélange (MIX) ne satisfait ni l'effet DF ni l'effet LD. Le modèle de divergence (DIV) ne garantit pas suffisamment l'effet IDF. Étonnamment, les modèles de pertinence RM et GRM ne satisfont pas l'effet IDF. Le modèle RM ne satisfait pas non plus l'effet DF.

2. Pour les modèles DFR, les modèles *Bo* ne respectent ni l'effet DF, ni les effets CONC et LD. Au contraire, le modèle d'information LL satisfait toutes les conditions de PRF.

Cette analyse théorique fournit une bonne explication des statistiques collectées sur deux grandes collections et résumées dans le tableau 3. Elle permet aussi d'expliquer le bon comportement du modèle LL développé dans le cadre des modèles d'information, modèle qui surpasse les autres modèles dans le cadre du PRF.

Tableau 7. Prise en compte des différents effets dans les modèles de PRF

Modèles de PRF vs effets	TF	CONC	IDF	LD	DF
MIX	oui	partiellement	oui	non	non
DIV	oui	oui	partiellement	oui	oui
GRM	oui	oui	non	oui	oui
Bo	oui	non	partiellement	non	non
LL	oui	oui	oui	oui	oui

5. Travaux reliés

Nous avons étudié ici les principales caractéristiques des fonctions de PRF à travers un certain nombres d'effets. C'est à notre connaissance la première étude qui propose un cadre théorique général pour les modèles de PRF. Il y a un certain nombres d'éléments, orthogonaux à ce que nous avons vu, qui permettent d'améliorer les performances d'un modèle de PRF. L'étude présentée dans (Lv, Zhai, 2009a), par exemple, propose une approche à base d'apprentissage pour estimer le mélange des termes originaux et des termes issus de l'ensemble de feedback. Ce paramètre peut être ajusté de façon optimale pour chaque requête, ce qui conduit à un système certes plus lourd, mais aussi plus performant. L'étude présentée dans (Lv, Zhai, 2010) se concentre elle sur l'utilisation d'information de position et de proximité dans les modèles de pertinence. Ici encore, la prise en compte d'une telle information permet d'améliorer les résultats en PRF. La prise en compte de ces informations dans d'autres modèles de PRF, comme ceux issus des modèles d'information, n'a pas encore été étudiée à notre connaissance.

Un autre type d'information qui peut se révéler utile en PRF est celui lié à la prise en compte des différents aspects d'une requête. L'étude présentée dans (Crabtree *et al.*, 2007), par exemple, propose un algorithme permettant d'identifier les différents aspects d'une requête et de l'étendre de façon à bien couvrir tous ces aspects. Une telle stratégie peut *a priori* être déployée sur tous les modèles de PRF. Une autre étude intéressante est celle présentée dans (Collins-Thompson, 2009) (Dillon, Collins-Thompson, 2010) où un cadre général d'optimisation est retenu pour le PRF. Les contraintes considérées diffèrent toutefois des effets que nous avons mis en avant dans la mesure où le but est ici encore de rendre compte de tous les aspects couverts par une requête. Le cadre général d'optimisation concave-convexe, pleinement détaillé dans (Collins-Thompson, 2008), est néanmoins intéressant car il établit des ponts entre plusieurs modèles.

Enfin, plusieurs études (Collins-Thompson, Callan, 2007) ou (Lee *et al.*, 2008) se sont récemment intéressées au problème de la modélisation de l'incertitude dans l'estimation des poids de PRF. Ces études montrent que le ré-échantillonnage des documents de feedback permet une meilleure estimation des poids de PRF. Un des intérêts de ces approches est qu'elles peuvent être facilement déployées sur tous les modèles de PRF. Elles permettent de plus une prise en compte simple et propre de l'effet DS que nous avons mentionné en section 3.

Nous avons introduit l'effet DF dans (Clinchant, Gaussier, 2011). L'étude présentée ici va plus loin car elle considère un nombre plus important de propriétés (reliées aux propriétés de la recherche d'information *ad hoc*) et de modèles. C'est grâce à cet élargissement que nous avons pu mettre en évidence les déficiences, notamment sur l'effet IDF, de nombreux modèles utilisés en PRF et que nous avons pu expliquer le bon comportement des modèles d'information.

6. Conclusion

Nous avons introduit dans cet article différents effets que les modèles de PRF doivent satisfaire. Ces effets peuvent être mis en parallèle avec les contraintes standard de la recherche d'information, avec toutefois un effet supplémentaire lié à la considération d'un ensemble de documents de feedback (effet DF). Nous avons ensuite analysé un certain nombre de modèles de PRF et observé leur comportement par rapport à ces effets. Cette analyse a révélé que les modèles directement issus du modèle de langue (à savoir le modèle de mélange et de minimisation de la divergence) sont déficients car le premier ne satisfait pas l'effet DF alors que le second ne garantit pas suffisamment l'effet IDF. Les modèles de pertinence, à la fois dans leur version standard et sous leur variante géométrique, sont également déficients car ils ne satisfont pas l'effet IDF. Pour les modèles issus du paradigme DRF, leur déficience est liée à la non satisfaction de l'effet DF. Seul le modèle LL, issu des modèles d'information, satisfait toutes les conditions.

Remerciements

Nous tenons à remercier les re-lecteurs pour leurs suggestions et leur relecture attentive de la première version de cet article.

Bibliographie

Amati G., Carpineto C., Romano G. (2004). Query difficulty, robustness, and selective application of query expansion. In *Ecir*, p. 127-137.

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

Carpineto C., Mori R. de, Romano G., Bigi B. (2001, January). An information-theoretic approach to automatic query expansion. *ACM Trans. Inf. Syst.*, vol. 19, p. 1–27. http://doi.acm.org/10.1145/366836.366860

Clinchant S., Gaussier E. (2010). Information-based models for *ad hoc* IR. In *Proceeding of the 33rd international acm sigir conference on research and development in information retrieval*, p. 234–241. New York, NY, USA, ACM.

Clinchant S., Gaussier É. (2011). Is document frequency important for prf? In *Ictir*, p. 89-100.

Collins-Thompson K. (2008). Estimating robust query models with convex optimization. In *Nips*, p. 329-336.

Collins-Thompson K. (2009). Reducing the risk of query expansion via robust constrained optimization. In *Proceeding of the 18th acm conference on information and knowledge management*, p. 837–846. New York, NY, USA, ACM.

Collins-Thompson K., Callan J. (2007). Estimation and use of uncertainty in pseudo-relevance feedback. In *Proceedings of the 30th annual international acm sigir conference on research and development in information retrieval*, p. 303–310. New York, NY, USA, ACM.

Crabtree D. W., Andreae P., Gao X. (2007). Exploiting underrepresented query aspects for automatic query expansion. In *Proceedings of the 13th acm sigkdd international conference on knowledge discovery and data mining*, p. 191–200. New York, NY, USA, ACM.

Cronen-Townsend S., Zhou Y., Croft W. B. (2004). A framework for selective query expansion [IR]. In *Proceedings of cikm '04*, p. 236-237.

Cummins R., O'Riordan C. (2007, June). An axiomatic comparison of learned term-weighting schemes in information retrieval: clarifications and extensions. *Artif. Intell. Rev.*, vol. 28, p. 51–68. http://portal.acm.org/citation.cfm?id=1485044.1485049

Dillon J. V., Collins-Thompson K. (2010). A unified optimization framework for robust pseudo-relevance feedback algorithms. In *Cikm*, p. 1069-1078.

Elkan C. (2006). Clustering documents with an exponential-family approximation of the Dirichlet compound multinomial distribution. In W. W. Cohen, A. Moore (Eds.), *Icml*, vol. 148, p. 289-296. ACM.

Fang H., Tao T., Zhai C. (2004). A formal study of information retrieval heuristics. In *Sigir '04: Proceedings of the 27th annual international acm sigir conference on research and development in information retrieval.*

Fang H., Zhai C. (2006). Semantic term matching in axiomatic approaches to information retrieval. In *Proceedings of the 29th annual international acm sigir conference on research and development in information retrieval*, p. 115–122. New York, NY, USA, ACM. http://doi.acm.org/10.1145/1148170.1148193

Hawking D., Thistlewaite P., Craswell N. (1997). Anu/acsys trec-6 experiments. In, p. 275–290.

Hoashi K., Matsumoto K., Inoue N., Hashimoto K. (2001). Query expansion based on predictive algorithms for collaborative filtering. In *Proceedings of the 24th annual international acm sigir conference on research and development in information retrieval*, p. 414–415. New York, NY, USA, ACM. http://doi.acm.org/10.1145/383952.384063

Lavrenko V., Croft W. B. (2001). Relevance based language models. In *Sigir '01: Proceedings of the 24th annual international acm sigir conference on research and development in information retrieval*, p. 120–127. New York, NY, USA, ACM.

Lee K. S., Croft W. B., Allan J. (2008). A cluster-based resampling method for pseudo-relevance feedback. In *Proceedings of the 31st annual international acm sigir conference on research and development in information retrieval*, p. 235–242. New York, NY, USA, ACM.

Lv Y., Zhai C. (2009a). Adaptive relevance feedback in information retrieval. In *Proceeding of the 18th acm conference on information and knowledge management*, p. 255–264. New York, NY, USA, ACM.

Lv Y., Zhai C. (2009b). A comparative study of methods for estimating query language models with pseudo feedback. In *Cikm '09: Proceeding of the 18th acm conference on information and knowledge management*, p. 1895–1898. New York, NY, USA, ACM.

Lv Y., Zhai C. (2010). Positional relevance model for pseudo-relevance feedback. In *Proceeding of the 33rd international acm sigir conference on research and development in information retrieval*, p. 579–586. New York, NY, USA, ACM. http://doi.acm.org/10.1145/1835449.1835546

Robertson S. (1977). The probability ranking principle in IR. *Journal of Documentation*, vol. 33.

Seo J., Croft W. B. (2010). Geometric representations for multiple documents. In *Sigir '10: Proceeding of the 33rd international acm sigir conference on research and development in information retrieval*, p. 251–258. New York, NY, USA, ACM.

Tao T., Zhai C. (2006). Regularized estimation of mixture models for robust pseudo-relevance feedback. In *Proceedings of the 29th annual international acm sigir conference on research and development in information retrieval*, p. 162–169. New York, NY, USA, ACM. http://doi.acm.org/10.1145/1148170.1148201

Trevor D. M., Strohman T., Turtle H., Croft W. B. (2004). *Indri at trec 2004: Terabyte track.*

Xu Z., Akella R. (2008). A new probabilistic retrieval model based on the dirichlet compound multinomial distribution. In *Sigir '08: Proceedings of the 31st annual international acm sigir conference on research and development in information retrieval*, p. 427–434. New York, NY, USA, ACM.

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

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