Estimation du paramètre de collection des modèles d'information pour la RI

Parantapa Goswami, Eric Gaussier

Université Joseph Fourier Grenoble 1, LIG
Grenoble, France
firstname.lastname@imag.fr

RÉSUMÉ. *Nous explorons dans cet article plusieurs méthodes permettant, a priori, d'estimer le paramètre de collection des modèles d'information. Jusqu'à présent, ce paramètre a était fixé au nombre moyen de documents dans lesquels un mot donné apparaissait. Nous présentons ici plusieurs méthodes d'estimation de ce paramètre et montrons qu'il est possible d'améliorer les performances du système de recherche d'information lorsque ce paramètre est estimé de façon adéquate.*

ABSTRACT. *In this paper we explore various methods to estimate the collection parameter of the information based models for* ad hoc *information retrieval. In previous studies, this parameter was set to the average number of documents where the word under consideration appears. We introduce here a fully formalized estimation method for both the log-logistic and the smoothed power law models that leads to improved versions of these models in IR. Furthermore, we show that the previous setting of the collection parameter of the log-logistic model is a special case of the estimated value proposed here.*

MOTS-CLÉS : *Modèles de RI, modèles fondés sur l'information, estimation des paramètres.*

KEYWORDS: *IR models, information-based models, parameter estimation.*

1. Introduction

Clinchant et Gaussier (Clinchant *et al.,* 2010) ont récemment introduit les modèles d'information pour la tâche de recherche d'information *ad hoc*. Ces modèles sont fondés sur le *contenu informatif* des termes, mesuré sur la base de l'écart entre le nombre d'occurrences observé d'un terme dans un document et celui prédit à partir de la distribution de ce terme sur l'ensemble des documents de la collection ; plus cet écart est important, plus le terme est significatif et porteur d'information dans le document. La fonction de score d'un document est alors calculée comme la moyenne pondérée des contenus informatifs des termes de la requête présents dans le document. La distribution des termes sur les documents de la collection repose, dans les modèles d'information, sur des distributions de probabilité "en rafale" (ou *bursty*) à un seul paramètre, λ_w. Nous dénommons ici ce paramètre *paramètre de collection* dans la mesure où il régule, au sein de la famille de distribution choisie, le comportement du terme w dans la collection. Dans (Clinchant *et al.,* 2010), λ_w est simplement fixé au nombre moyen de documents dans lesquels le terme w apparaît. La principale justification formelle avancée pour ce choix repose sur le fait qu'il garantit que les modèles obtenus satisfont à l'effet IDF ((Fang *et al.,* 2004)). Ce n'est toutefois pas le seul choix qui garantisse cela, et nous explorons dans cette étude différents cadres théoriques qui permettent l'estimation de ce paramètre. Nous nous concentrons ici sur les deux distributions de probabilité couramment utilisées pour les modèles d'information, à savoir la loi log-logistique et la distribution de puissance lissée. L'estimation que nous retenons finalement est motivée d'un point de vue théorique, garantit l'effet IDF, justifie en partie le choix fait précédemment dans (Clinchant *et al.,* 2010) et fournit des résultats meilleurs que les précédents en recherche d'information *ad hoc*.

La description de cette étude est organisée comme suit. Nous présentons tout d'abord en section 2 les modèles d'information et le rôle joué par le paramètre de collection dans ces modèles. La section 3 est dévolue à l'estimation du paramètre de collection λ_w. Enfin, la section suivante, section 4, décrit les expériences réalisées. Ces expériences montrant qu'une bonne estimation du paramètre de collection conduit à une amélioration des systèmes de recherche d'information. Mentionnons également que les notations utilisées dans les descriptions sont résumées dans le tableau 1, où w représente un terme.

2. Modèles d'information et paramètre de collection

Les modèles d'information sont fondés sur la fonction de score suivante :

$$RSV(q,d) \quad = \quad \sum_{w \in q \cap d} -\frac{x_w^q}{l_q} \log P(X_w \geq t_w^d | \lambda_w) \qquad [1]$$

où :

Notations	Descriptions
x_w^d (x_w^q)	nombre d'occurrences de w dans le document d (dans la requête q)
t_w^d	version normalisée de x_w^d, i.e. *normalized term frequency*
l_d (l_q)	longueur du document d (de la requête q)
l_{avg}	longueur moyenne des documents
N	nombre de documents dans la collection
N_w	nombre de documents de la collection contenant w, *i.e.* $N_w = \sum_d I(x_w^d > 0)$

Tableau 1. *Notations*

1) t_w^d est une fonction de normalisation qui dépend du nombre d'occurrences, x_w^d, de w dans d, de la longueur, l_d, de d, et qui vérifie :

$$\frac{\partial t_w^d}{\partial x_w^d} > 0; \;\; \frac{\partial t_w^d}{\partial l_d} < 0; \frac{\partial^2 x_w^d}{\partial (t_w^d)^2} \geq 0$$

Dans cette étude, et suivant en cela (Clinchant *et al.,* 2010), cette fonction de normalisation est définie par : $t_w^d = x_w^d \log(1 + c\frac{l_{avg}}{l_d})$ où c est un paramètre de lissage ;

2) P est une distribution de probabilité définie pour une variable aléatoire, X_w, associée à chaque mot w et dont les réalisations sont t_w^d. Cette distribution doit être :

- Continue, les réalisations de la variable aléatoire considérée, t_w^d, étant continues ;

- Compatible avec le domaine de t_w^d, *i.e.* si t_{min} est la valeur minimale de t_w^d, alors $P(X_w \geq t_{\text{min}}|\lambda_w) = 1$;

- En rafale, c'est-à-dire qu'elle doit satisfare :
$\forall \epsilon > 0, \;\; g_\epsilon(x) = P(X \geq x + \epsilon | X \geq x)$ est strictement croissante en x ;

3) Et λ_w est le paramètre de collection, fixé dans (Clinchant *et al.,* 2010) à :

$$\lambda_w = \frac{N_w}{N} \qquad [2]$$

Nous cherchons dans cette étude à établir quelle est la meilleure procédure pour estimer ce dernier paramètre.

Les points 1 et 2 ci-dessus assurent que la fonction de score obtenue satisfait les effets TF, de longueur de document et de concavité présentés dans (Clinchant *et al.,* 2010)(Clinchant *et al.,* 2011)(Fang *et al.,* 2004) par exemple. Le quatrième effet, l'effet IDF, ne peut être, quant à lui, garanti que par la valeur prise par λ_w.

Comme on peut le remarquer, l'équation 1 calcule l'information apportée par un document sur chaque mot de la requête ($-\log P(X_w \geq t_w^d|\lambda_w)$), cette information étant pondérée par l'importance du mot dans la requête ($\frac{x_w^q}{l_q}$). Dans ce calcul, comme l'indique le point (2) ci-dessus, il est nécessaire, afin de définir un "bon" modèle de RI,

de choisir une distribution en rafale. Nous nous reposons ici sur les deux distributions en rafale les plus utilisées : la distribution log-logistique, que nous noterons LGD (pour *Log-Logistic Distribution*) et la distribution de puissance lissée, que nous noterons SPL (pour *Smoothed Power Law*). Ces distributions sont définies par :

$$P_{LGD}(X_w \geq t_w^d | \lambda_w) = \frac{\lambda_w}{t_w^d + \lambda_w} \ (\lambda_w > 0)$$

$$P_{SPL}(X_w \geq t_w^d | \lambda_w) = \frac{\lambda_w^{\frac{t_w^d}{t_w^d+1}} - \lambda_w}{1 - \lambda_w} \ (0 < \lambda_w < 1)$$

2.1. *Effet IDF et* λ_w

L'effet IDF stipule que la fonction de score doit croître avec l'inverse de la fréquence documentaire (qui correspond au nombre de documents dans lesquels un terme donné apparaît). La forme des modèles d'information permet de ré-écrire cet effet de la façon suivante :

$$\frac{\partial P}{\partial N_w} = \frac{\partial P}{\partial \lambda_w} . \frac{\partial \lambda_w}{\partial N_w} > 0 \qquad \textbf{[critère IDF]}$$

où P est la distribution de probabilité utilisée dans les modèles d'information.

– Pour la distribution log-logistique, nous avons :

$$\frac{\partial P_{LGD}(X_w \geq t_w^d | \lambda_w)}{\partial \lambda_w} = \frac{t_w^d}{(t_w^d + \lambda_w)^2} > 0$$

comme $t_w^d > 0$. Le critère IDF prend alors la forme : $\frac{\partial \lambda_w}{\partial N_w} > 0$.

– Pour la distribution de puissance lissée :

$$\frac{\partial P_{SPL}}{\partial \lambda_w} = \frac{\overbrace{\big(\frac{t_w^d}{t_w^d+1} \lambda_w^{-\frac{1}{t_w^d+1}} - 1\big)(1-\lambda_w) + (\lambda_w^{\frac{t_w^d}{t_w^d+1}} - \lambda_w)}^{g(\lambda_w)}}{(1-\lambda_w)^2}$$

et :

$$\begin{aligned} \frac{\partial g}{\partial \lambda_w} = g'(\lambda_w) &= -\frac{t_w^d}{(t_w^d+1)^2} \lambda_w^{-\frac{t_w^d+2}{t_w^d+1}} (1-\lambda_w) - \frac{t_w^d}{t_w^d+1} \lambda_w^{-\frac{1}{t_w^d+1}} \\ &\quad + \frac{t_w^d}{t_w^d+1} \lambda_w^{-\frac{1}{t_w^d+1}} \\ &= -\frac{t_w^d}{(t_w^d+1)^2} \lambda_w^{-\frac{t_w^d+2}{t_w^d+1}} (1-\lambda_w) \end{aligned}$$

Or, nous avons : $t_w^d > 0$ et $\frac{t_w^d}{(t_w^d+1)^2} > 0$. De plus, pour $0 < \lambda_w < 1$, $\lambda_w^{-\frac{t_w^d+2}{t_w^d+1}}$ et $1 - \lambda_w > 0$ sont aussi strictement positifs. Donc la fonction $g(\lambda_w)$ est strictement décroissante en λ_w. De plus, $g(0) = \infty$ et $g(1) = 0$. Donc, pour $\lambda_w \in (0, 1)$, $g(\lambda_w)$ est toujours positive. Or $\frac{\partial P_{SPL}}{\partial \lambda_w} = \frac{g(\lambda_w)}{(1-\lambda_w)^2}$, avec, pour $\lambda_w \in (0, 1)$, un dénominateur et un numérateur tous deux positifs. Ainsi, $\frac{\partial P_{SPL}}{\partial \lambda_w} > 0$.

Le critère IDF prend donc ici la même forme que celui pour la distribution log-logisitque : $\frac{\partial \lambda_w}{\partial N_w} > 0$.

Nous nous concentrons maintenant sur l'estimation de λ_w.

3. Estimation du paramètre de collection

Nous passons en revue dans cette étude trois méthodes courantes d'estimation de paramètres, à savoir la méthode du *maximum de vraisemblance*, la méthode de *Kaplan-Meier estimation* et la *méthode généralisée des moments*.

3.1. *Estimation par maximum de vraisemblance*

La méthode du maximum de vraisemblance est certainement la méthode la plus utilisée pour estimer les paramètres d'un modèle probabiliste. Toutefois, cette méthode ne fournit pas toujours une estimation des paramètres lorsque la vraisemblance atteint son maximum aux bornes, non incluses dans le domaine de définition, des paramètres. C'est très exactement ce qui se passe pour les distributions log-logistique et de puissance lissée, comme nous allons le voir.

– Pour la distribution log-logistique, la vraisemblance prend la forme :

$$L(\lambda_w, X_w) = (2N_w - N) \log \lambda_w - 2 \sum_{d|t_w^d>0} \log(t_w^d + \lambda_w)$$

Si le nombre de documents contenant w (N_w) est inférieur à la moitié du nombre de documents dans la collection ($\frac{N}{2}$), alors $L(\lambda_w, X_w)$ est maximum quand $\lambda \to 0$. En pratique, $N_w < \frac{N}{2}$ pour la majorité des termes. La méthode du maximum de vraisemblance ne fournit donc pas d'estimation opérationnelle.

– Pour la distribution de puissance lissée, la vraisemblance est définie par :

$$\begin{aligned} L(\lambda_w, X_w) &= -N \log\log \lambda_w + N \log(1 - \lambda_w) \\ &\quad + \sum_d \Big[\frac{t_w^d}{t_w^d + 1} \log \lambda_w - 2 \log(t_w^d + 1) \Big] \end{aligned}$$

La résolution de l'équation aux dérivées partielles (c'est-à-dire ici la résolution de l'équation $\frac{\partial L(\lambda_w, X_w)}{\partial \lambda_w} = 0$) fournit :

$$\frac{\lambda_w}{1-\lambda_w} = \frac{1}{N}\sum_d \frac{t_w^d}{t_w^d+1} - \frac{1}{\log \lambda_w} \quad [3]$$

Soit $g(\lambda_w) = \frac{\lambda_w}{1-\lambda_w}$ et $h(\lambda_w) = -\frac{1}{\log \lambda_w} = -(\log \lambda_w)^{-1}$. L'équation précédente peut être écrite sous la forme :

$$g(\lambda_w) = c + h(\lambda_w)$$

avec $c = \frac{1}{N}\sum_d \frac{t_w^d}{t_w^d+1}$. Comme λ_w appartient à $(0,1)$, $g(\lambda_w)$ et $h(\lambda_w)$ sont toutes deux croissantes de 0 à $+\infty$. Ainsi, $c + h(\lambda_w)$ varie de c à $+\infty$ lorsque λ_w varie de 0 à 1. Pour que l'équation 3 ait une solution, il est nécessaire, pour au moins une valeur de λ_w, d'avoir $g(\lambda_w) > h(\lambda_w)$. Mais $g(\lambda_w) > h(\lambda_w)$ implique que $\log \lambda_w < \frac{\lambda_w - 1}{\lambda_w}$ [1], ce qui est impossible car, pour tout $x > 0$, $\log(1+x) > \frac{x}{x+1}$. Ainsi, ici encore, la méthode du maximum de vraisemblance ne fournit pas une estimation de λ_w.

3.2. *Estimation de Kaplan-Meier*

La méthode d'estimation de Kaplan-Meier à été proposée dans le cadre de l'analyse de la survie, et porte sur les fonctions de survie de forme $P(T > t)$. Dans la mesure où cette forme correspond à celle utilisée dans les modèles d'information, $P(X_w > t_w^d | \lambda_w)$, la méthode d'estimation de Kaplan-Meier peut être utilisée pour estimer le paramètre λ_w. Cette méthode, décrite dans (Kaplan *et al.,* 1958), fournit une estimation des quantités $P(X_w > t_w^d | \lambda_w)$ correspondant aux différentes valeurs observées de t_w^d dans la collection. Soit, pour un terme w, la suite ordonnée par valeurs croissantes des fréquences normalisées : $t_w^{d_1} \leq t_w^{d_2} \leq \ldots \leq t_w^{d_N}$ (cette forme peut toujours être obtenue en re-numérotant les documents de la collection). Pour $i \leq N$, la probabilité $P(X \geq t_w^{d_i})$ est estimée par la méthode de Kaplan-Meier par la quantité :

$$P(X \geq t_w^{d_i}) = \prod_{r=1}^{i} \frac{N-r}{N-r+1}$$

Le biais de cet estimateur croît lorsque $t_w^{d_i}$ croît (Balakrishnan *et al.,* 2004). De façon à se reposer sur un estimateur de faible biais, nous choisissons i de telle sorte que $t_w^{d_i}$ est le premier terme non nul de la suite ordonnée introduite précédemment ; en d'autre termes, $i = N - N_w$. Nous obtenons alors l'estimation suivante :

$$P_{est}(X \geq t_w^{d_i}) = \frac{N-i}{N}$$

Nous allons maintenant voir l'application de ce résultat aux distributions log-logistique et de puissance lissée.

1. Nous omettons la dérivation, qui est purement technique.

– Pour la distribution log-logistique, nous avons :

$$P_{est}(X \geq t_w^{d_i}) = \frac{N-i}{N} = \frac{N_w}{N} = \frac{\lambda_w}{t_w^{d_i} + \lambda_w}$$

Ce qui conduit à :

$$\lambda_w = \frac{N_w}{N - N_w} t_w^{d_i}$$

Comme N, N_w et t_w^d sont tous positifs :

$$\frac{\partial \lambda_w}{\partial N_w} = \frac{(N - N_w) - N_w.(-1)}{(N - N_W)^2} t_w^d = \frac{N}{(N - N_W)^2} t_w^d > 0$$

Le critère IDF est donc satisfait, ce qui montre que la méthode de Kaplan-Meier peut être utilisée pour estimer le paramètre de collection du modèle log-logistique. Nous noterons le modèle obtenu avec cette estimation LGD_{KM}.

– Pour la distribution de puissance lissée, nous obtenons :

$$N\lambda_w^{\left(\frac{t_w^{d_i}}{t_w^{d_i}+1}\right)} - N_w = \lambda_w(N - N_w)$$

Cette équation n'a pas toutefois pas toujours de solution dans $]0, 1[$. En effet, considérons un mot pour lequel $t_w^{d_i} = 1$ et $N_w = \frac{N}{2}$ (ce qui correspond à des valeurs observables en pratique). λ_w satisfait alors :

$$N_w^2 \lambda_w^2 - 2N_w^2 \lambda_w + N_w^2 = 0$$

Cette équation a une seule solution, à savoir $\lambda_w = 1$, qui n'appartient pas au domaine de définition de λ_w. La méthode de Kaplan-Meier ne fournit donc pas systématiquement une solution valide pour la distribution de puissance lissée.

Même si la méthode d'estimation de Kaplan-Meier peut être utilisée pour la distribution log-logistique, elle repose sur une observation unique, $t_w^{d_i}$, qui correspond en général à un faible nombre d'occurrences. L'estimateur ainsi obtenu risque donc de ne pas être fiable. Nous allons maintenant présenter une dernière méthode d'estimation qui permet de s'affranchir de ce problème et qui fournit des estimateurs valides pour les distributions log-logistique et de puissance lissée.

3.3. *La méthode généralisée des moments*

La probabilité qu'un mot w soit présent dans un document d correspond à $P(x_w^d \geq 1|\lambda_w)$. De plus, $t_w^d = x_w^d[\log(1 + c\frac{l_{avg}}{l_d})]$, et donc $x_w^d = \frac{t_w^d}{\log(1+c\frac{l_{avg}}{l_d})}$. Nous avons donc :

$$P(x_w^d \geq 1|\lambda_w) = P(t_w^d \geq \log(1 + c\frac{l_{avg}}{l_d}|\lambda_w))$$

L'espérance d'observer le mot w dans la collection s'écrit donc : $\sum_d P(t_w^d \geq \log(1 + c\frac{l_{avg}}{l_d}|\lambda_w)$. La méthode généralisée des moments [2] consiste alors à rechercher la valeur du paramètre λ_w pour laquelle l'espérance ci-dessus est égale à la valeur observée, c'est-à-dire N_w. Ceci conduit à la résolution de l'équation ci-dessous :

$$\begin{aligned} N_w &= \sum_d P(x_w^d \geq 1|\lambda_w) = \sum_d P(t_w^d \geq \log(1 + c\frac{l_{avg}}{l_d}|\lambda_w) \\ &= \sum_d P(t_w^d \geq \alpha_d|\lambda_w) \end{aligned} \quad [4]$$

avec $\alpha_d = \log(1 + c\frac{l_{avg}}{l_d})$.

Le principal avantage de cette méthode est de se reposer sur une quantité, ici N_w, moins sujette à variation que les autres quantités observables. N_w est certainement la statistique la plus robuste que l'on puisse tirer des observations des mots dans la collection de documents. Contrairement à la méthode de Kaplan-Meier, qui repose sur une quantité observée dans un seul document et donc sujette à variations (dues, par exemple, au fait qu'un auteur aurait pu utiliser un mot différent, ou moins d'occurrences du mot donné, pour exprimer la même idée), la méthode généralisée des moments repose, quant à elle, sur une observation plus fiable. Nous allons maintenant voir comment résoudre l'équation ci-dessus pour les distributions log-logistique et de puissance lissée.

3.3.1. *Distribution log-logistique*

Pour la distribution log-logistique, l'équation 4 a la forme :

$$\underbrace{\frac{N_w}{\lambda_w}}_{f(\lambda)} = \underbrace{\sum_d \frac{1}{\alpha_d + \lambda_w}}_{g(\lambda_w)}, \ \lambda_w > 0 \quad [5]$$

Comme $f'(\lambda_w) < 0$, $f''(\lambda_w) > 0$, $g'(\lambda_w) < 0$, $g''(\lambda_w) > 0$, f et g sont concaves et telles que $g(0) = \sum_d(\alpha_d)^{(-1)} < f(0) = +\infty$. La figure 1(a)(a) illustre le comportement des fonctions f et g. Soit a un réel strictement positif ; $f(a) = \frac{N_w}{a}$ et $g(a) = \sum_d(\alpha_d + a)^{(-1)}$. L'équation 5 a une solution, pour tous les mots w, dans $(0, a)$ si et seulement si $\exists a \in \mathbf{R}^{+*}$, $g(a) > \frac{N_w}{a}$ pour tous les termes w, *i.e.* si et seulement si :

$$a \sum_d \frac{1}{\log(1 + c\frac{l_{avg}}{l_d}) + a} > N_w^{max} \quad [6]$$

2. Johnson et Kotz (Johnson *et al.,* 1993), par exemple, utilisent une variante de la méthode des moments dans laquelle la variance empirique est remplacée par l'inverse de la fréquence documentaire. Church et Gale (Church *et al.,* 1995) ont introduit, à notre connaissance, le terme "méthode généralisée des moments" pour désigner une méthode dans laquelle le paramètre d'une distribution de probabilité est estimé à partir de contraintes sur des quantités observables (comme l'IDF, relié à N_w).

Comme $\lim_{a\to+\infty} a \sum_d \frac{1}{\log(1+c\frac{l_{avg}}{l_d})+a} = N > N_w^{max}$, une solution existe nécessairement pour a suffisamment grand. Le paramètre libre (c'est-à-dire le paramètre de lissage des modèles de langue, k_1 dans BM25 et c dans les modèles d'information) est en général optimisé sur un sous-ensemble des requêtes pour lesquelles on dispose de jugements de pertinence. Nous adoptons ici la même stratégie, avec une étape additionnelle vérifiant si la condition ci-dessus est satisfaite :

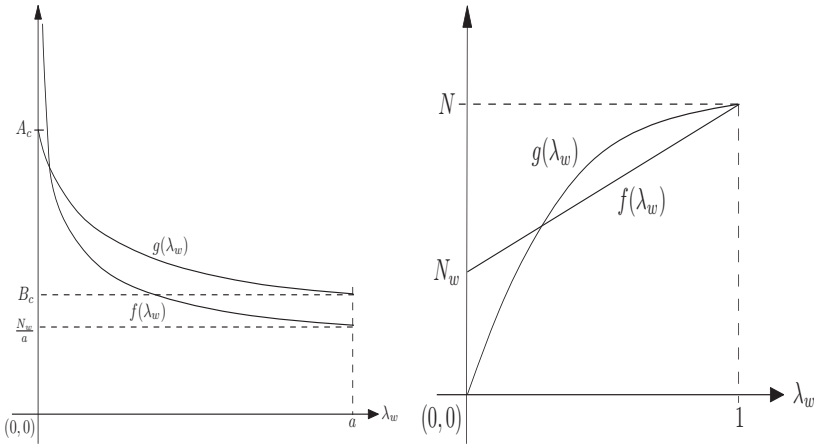
1) Choisir un intervalle pour c ;

2) Choisir a suffisamment grand, par exemple $a = 100$; on cherche λ_w dans $(0, a)\ \forall w$;

3) Pour chaque valeur de c, si la condition 6 est satisfaite, estimer λ_w pour chaque w en résolvant l'équation 5 ;

4) Si la condition 6 n'est pas satisfaite avec la valeur courante de c, augmenter a et retourner à l'étape 3.

Deux remarques doivent être faites ici : (a) l'équation 5 peut-être résolue à l'aide de méthodes standard, comme la méthode de Newton-Raphson ou une dichotomie (en général moins efficace) ; (b) λ_w est estimé pour tous les mots w de la collection une fois pour toute ; cette estimation n'a donc pas d'impact sur le temps de recherche.



(a) LGD (b) SPL

Figure 1. *Forme des fonctions mises en jeu dans la méthode généralisée des moments pour les distributions log-logistique (LGD) et de puissance lissée (SPL).*

Le critère IDF stipule que λ_w doit croître avec N_w. Nous savons que λ_w est solution de l'équation :

$$N_w = \underbrace{\sum_d \frac{\lambda_w}{\alpha_d + \lambda_w}}_{h(\lambda_w)}$$

Comme $h'(\lambda_w) > 0$, h est une fonction croissante en λ_w, ce qui implique que si h croît, λ_w croît aussi. Donc, si N_w croît, λ_w croît aussi. Ceci montre que le critère IDF est satisfait pour l'estimation obtenue avec la méthode généralisée des moments. Nous noterons le modèle obtenu par la procédure ci-dessus LGD_{GMM}.

3.3.2. *Distribution de puissance lissée*

Pour la distribution de puissance lissée, l'équation 4 prend la forme :

$$\underbrace{\lambda_w(N - N_w) + N_W}_{f(\lambda_w)} = \underbrace{\sum_d \lambda_w^{\frac{\alpha_d}{\alpha_d+1}}}_{g(\lambda_w)},\ \lambda_w \in (0,1) \qquad [7]$$

f est une fonction linéaire ; de plus, $g'(\lambda_w) < 0$ et $g''(\lambda_w) > 0$, ce qui montre que g est une fonction concave qui est telle que $g(0) < f(0)$ et $g(1) = f(1)$. L'équation 7 a une solution dans $]0,1[$ si et seulement si f et g se croisent dans $]0,1[$ (voir l'illustration dans la figure 1(b)(b)). Ceci est équivalent à avoir g au-dessus de f dans un voisinage de 1. Soit ϵ un réel positif très petit. Nous avons :

$$g(1-\epsilon) = \sum_d (1-\epsilon)^{\frac{\alpha_d}{\alpha_d+1}} \approx \sum_d 1 - \epsilon\frac{\alpha_d}{\alpha_d+1} = N - \sum_d \epsilon\frac{\alpha_d}{\alpha_d+1}$$

et : $f(1-\epsilon) - g(1-\epsilon) = \epsilon(N_w + \sum_d \frac{\alpha_d}{\alpha_d+1} - N)$. Donc g est au-dessus de f dans un voisinage de 1, pour tous les mots w, si et seulement si :

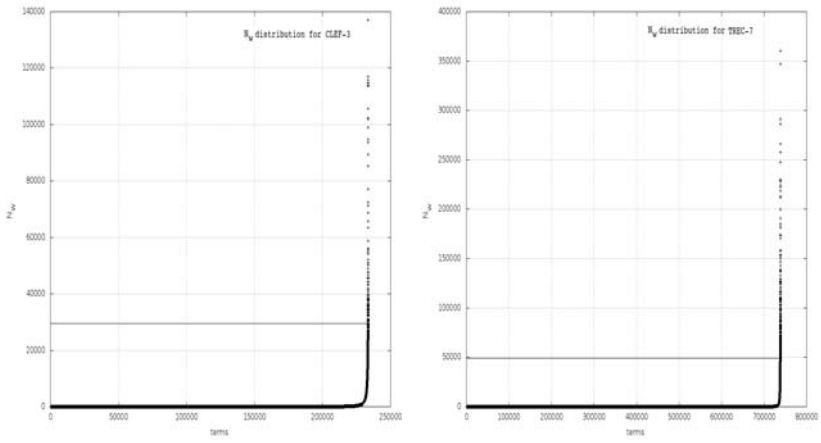
$$\sum_d \frac{\log(1 + c\frac{l_{avg}}{l_d})}{1 + \log(1 + c\frac{l_{avg}}{l_d})} < N - N_w^{max} \qquad [8]$$

ce qui fournit une contrainte sur les valeurs admissibles de c. Pour toutes les valeurs de c compatibles avec cette contrainte, on peut estimer λ_ws en résolvant l'équation 7, ce qui peut se faire avec les méthodes standard.

Ceci étant, la contrainte exprimée par l'équation 8 est forte dès lors qu'il existe des termes dans la collection qui apparaissent dans beaucoup de documents. La présence de tels termes limite l'intervalle dans lequel les valeurs de λ peuvent être trouvées. En pratique, toutefois, très peu de termes apparaissent dans beaucoup de documents. La figure 2 fournit les valeurs de N_w pour tous les termes, et ce dans deux collections différentes qui seront détaillées à la section suivante[3]. Comme on peut le constater,

3. Des courbes similaires peuvent être obtenues pour d'autres collections.

la très grande majorité des termes apparaît dans peu de documents. Il serait alors dommage que seuls quelques termes (très fréquents) limitent l'intervalle d'estimation de l'ensemble de la collection. Nous utilisons donc la stratégie suivante : fixer λ_w à $\frac{N_w}{N}$ (ce qui correspond au choix fait dans (Clinchant *et al.,* 2010)) pour les p termes les plus fréquents, suivant N_w, de la collection, et utiliser la méthode généralisée des moments pour les autres termes, avec N_w^{max} défini sur ce dernier ensemble. Dans nos expériences, p est choisi à 0.05% de N (aucun terme des requêtes considérées n'appartient en fait à cet ensemble).



(a) CLEF3 (b) TREC7

Figure 2. *Distribution de N_w sur 2 collections. La ligne horizontale correspond à $p = 0.05\%$ de N.*

Enfin, l'estimation obtenue avec la méthode ci-dessus satisfait aussi le critère IDF dans la mesure où la fonction $h(\lambda_w) = \sum_d \frac{\frac{\alpha_d}{\alpha_d+1}}{\lambda_w} - \lambda_w}{1-\lambda_w}$ est croissante en λ_w (le raisonnement est le même que celui suivi pour la distribution log-logistique). Nous noterons le modèle ainsi obtenu SPL_{GMM}.

4. Validation expérimentale

Toutes nos expériences sont conduites sur la plateforme Terrier (Terrier IR Platform v3.5, terrier.org). Nous avons implanté les nouveaux modèles proposés ici dans cette plateforme, et utilisé les modèles déjà disponibles pour comparaison. Nous avons utilisé quatre collections standard de recherche d'information issues des campagnes TREC (trec.nist.gov) et CLEF (www.clef-campaign.org). Le tableau 4 fournit un résumé (nombre de documents (N), longueur moyenne des documents (l_{avg}) et nombre de requêtes) des différentes collections utilisées : TREC-3, TREC-6, TREC-7, TREC-

8, CLEF-3 de la tâche *ad hoc* en anglais. Pour toutes ces collections, nous avons utilisé le *stemmer* de Porter.

	N	l_{avg}	# Requêtes
TREC-3	741856	261.134	50
TREC-6	528155	295.976	50
TREC-7	528155	295.976	50
TREC-8	528155	295.976	50
CLEF-3	169477	300.789	60

Tableau 2. *Caractéristiques des différentes collections.*

Les modèles LGD$_{KM}$, LGD$_{GMM}$, SPL$_{GMM}$ sont principalement comparés aux modèles log-logistique et de puissance lissée originellement proposés par Clinchant et Gaussier avec $\lambda_w = \frac{N_w}{N}$. Nous notons ces modèles LGD et SPL. Nous avons également comparé ces modèles au modèle de langue avec lissage de Dirichlet (qui est une version du modèle de langue qui fournit de meilleurs résultats que celle fondée sur le lissage de Jelinek-Mercer) (Ponte *et al.,* 1998)(Zhai *et al.,* 2001) et au modèle Okapi BM25(Robertson *et al.,* 2009). Pour chaque modèle, sur chaque collection, nous avons réalisé une validation croisée en 5 parties, ce qui implique que le jeu de requêtes est partitionné en 5 sous-ensembles. Un de ces sous-ensembles est alors utilisé pour évaluer les performances du modèle, les 4 autres servant à l'apprentissage des paramètres libres (c pour les modèles d'information, k_1 pour Okapi et μ pour le modèle de langue avec lissage de Dirichlet). Ce processus est répété 5 fois et les résultats moyennés sur ces l'ensemble de ces répétitions. Nous utilisons la *Mean Average Precision (MAP)* et la précision à 10 documents (P10) pour évaluer et comparer les modèles. Nous nous reposons de plus sur le *paired two-sided t-test* au niveau 0.1 pour établir si les différences obtenues sont statistiquement significatives ou non.

Pour LGD, LGD$_{KM}$, LGD$_{GMM}$, SPL and SPL$_{GMM}$, le paramètre c est optimisé sur l'ensemble des valeurs suivantes : $\{$0.1, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0, 6.0, 7.0, 8.0, 9.0, 10.0, 20.0$\}$. Le paramètre de lissage de Dirichlet est optimisé sur l'ensemble standard : $\{$10, 25, 50, 75, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 1500, 2000, 2500, 3000, 4000, 5000, 10000$\}$. Enfin, pour BM25, les paramètres k_1 et b sont optimisés sur l'ensemble$\{$ 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.25, 1.5, 1.75, 2.0, 2.25, 2.5, 2.75, 3.0$\}$, le troisième paramètre, k_3, étant fixé à la valeur par défaut donnée dans Terrier (7).

Le tableau 3 présente les résultats obtenus avec les modèles d'information : LGD, LGD$_{KM}$ et LGD$_{GMM}$ d'une part, SPL et SPL$_{GMM}$ d'autre part. Les meilleurs résultats sont en gras ; les valeurs de MAP suivies d'un astérisque sont significativement meilleures que les autres valeurs. Comme on peut le constater, les versions des modèles d'information proposées dans cette étude fournisssent de meileurs résultats que les versions standard. Le fait que LGD$_{GMM}$ soit significativement meilleur que LGD

sur une seule collection s'explique par le fait que LGD représente une approximation de LGD$_{GMM}$ (à noter que ceci n'est pas vrai pour SPL et SPL$_{GMM}$). En effet, supposons que $l_d \approx l_{avg}$, alors α_d est une constante ($\alpha_d = \alpha$) et la solution de l'équation 5 est $\lambda_w = \frac{\alpha N_w}{N - N_w}$. Dans la mesure où, pour la plupart des termes (et pour tous les termes des requêtes des collections considérées) $N_w \ll N$, nous avons : $\lambda_w = \frac{\alpha N_w}{N}$, qui est, à un facteur près qui ne change pas l'ordre des documents, la valeur retenue dans (Clinchant *et al.,* 2010).

Collections	TREC-3		TREC-6		TREC-7	
Modèles	MAP	P10	MAP	P10	MAP	P10
LGD	24.56	48.40	24.49	40.40	18.95	44.20
LGD$_{KM}$	24.83	50.00	**24.66**	40.40	**18.96**	**44.40**
LGD$_{GMM}$	**25.60***	**51.80***	24.57	**40.60**	18.92	43.40
SPL	25.17	52.40	25.03	**40.40**	18.44	**45.00**
SPL$_{GMM}$	**26.77***	**54.60***	**25.19**	**40.40**	**19.09**	44.80

Collections	TREC-8		CLEF-3	
Modèles	MAP	P10	MAP	P10
LGD	25.82	**45.60**	38.55	31.36
LGD$_{KM}$	25.77	45.00	38.47	**31.92**
LGD$_{GMM}$	**25.87**	**45.60**	**39.51**	31.53
SPL	25.56	46.20	38.90	32.82
SPL$_{GMM}$	**26.28***	**47.00**	**40.42***	**33.94***

Tableau 3. *LGD$_{KM}$, LGD$_{GMM}$ vs LGD et SPL$_{GMM}$ vs SPL (après validation croisée).*

La comparaison de LGD$_{KM}$, LGD$_{GMM}$ et SPL$_{GMM}$ avec Okapi BM25 et le modèle de langue est présentée dans le tableau 4. Comme précédemment, les meilleurs résultats sont en gras. Si la différence entre le modèle M et le meilleur modèle est significative, les résultats de M sont en italique. Comme on peut le constater, le modèle SPL$_{GMM}$ fournit les meilleurs résultats sur 4 des 5 collections retenues pour la MAP, et sur 3 des5 collections pour la précision à 10 documents. De plus, lorsque ce modèle est le meilleur, la différence avec les autres modèles est en général significative. Quand ce n'est pas le meilleur modèle, la différence avec le meilleur modèle n'est pas significative. Ceci montre que les nouvelles versions des modèles d'information introduites dans cette étude ne sont pas seulement meilleures que les versions standard. Elles fournissent également de meilleurs résultats que les autres modèles de recherche d'information sur les collections retenues ici.

Collections	TREC-3		TREC-6		TREC-7	
Modèles	MAP	P10	MAP	P10	MAP	P10
BM25	**27.32**	**56.20**	*23.76*	39.40	*19.08*	43.00
LM_{DIR}	*26.85*	*55.80*	*24.27*	*39.00*	18.88	*42.40*
LGD_{KM}	*24.83*	*50.00*	*24.66*	40.40	18.96	44.40
LGD_{GMM}	*25.60*	*51.80*	*24.57*	**40.60**	18.92	*43.40*
SPL_{GMM}	26.77	54.60	**25.19**	40.40	**19.09**	**44.80**

Collections	TREC-8		CLEF-3	
Modèles	MAP	P10	MAP	P10
BM25	*25.89*	46.40	39.88	33.38
LM_{DIR}	*25.43*	*45.20*	*39.38*	*31.58*
LGD_{KM}	25.77	*45.00*	*38.47*	*31.92*
LGD_{GMM}	*25.87*	*45.60*	*39.51*	*31.53*
SPL_{GMM}	**26.28**	**47.00**	**40.42**	**33.94**

Tableau 4. *LGD_{KM}, LGD_{GMM} et SPL_{GMM} vs BM25 et LM_{DIR} (après validation croisée).*

5. Conclusion

Le paramètre de collection des modèles d'information est lié à la nature du comportement d'un terme dans la collection. Dans les études précédentes, ce paramètre était fixé au nombre moyen de documents dans lesquels le terme apparaît, sans réelle justification pour ce choix. Dans cette étude, nous avons exploré plusieurs méthodes d'estimation de paramètres afin de se reposer sur une valeur qui soit motivée. Nous avons pu montrer que, parmi les méthodes retenues, la méthode généralisée des moments fournissait des estimations valides pour les deux distributions standard des modèles d'information, à savoir la distribution log-logistique et la distribution de puissance lissée. Ces estimations garantissent de plus l'effet IDF. Les expériences conduites sur 5 collections de recherche d'information ont de plus montré que les nouvelles versions des modèles d'information obtenues se comportent mieux que les versions standard. De plus, la nouvelle version du modèle de puissance lissée fournit des résultats meilleurs que tous les autres modèles (y compris les modèles de langue et le modèle BM25) dans la majorité des cas.

6. Bibliographie

Balakrishnan N., Rao C., *Advances in Survival Analysis*, vol. 23 of *Handbook of Statistics*, first edn, North Holland, chapter 5, p. 96, February, 2004.

Church K. W., Gale W. A., « Poisson Mixtures », *Natural Language Engineering*, vol. 1, p. 163-190, 1995.

Clinchant S., Gaussier E., « Information-based models for ad hoc IR », *SIGIR 2010 : Proceedings of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 234-241, 2010.

Clinchant S., Gaussier E., « Retrieval constraints and word frequency distributions a log-logistic model for IR », *Information Retrieval*, vol. 14, n° 1, p. 5-25, 2011.

Fang H., Tao T., Zhai C., « A formal study of information retrieval heuristics », *SIGIR 2004 : Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 49-56, 2004.

Johnson N., Kemp A., Kotz S., *Univariate Discrete Distributions*, John Wiley & Sons, Inc., 1993.

Kaplan E. L., Meier P., « Nonparametric estimation from incomplete observations », *Journal of the American Statistical Association*, vol. 53, n° 282, p. 457-481, June, 1958.

Ponte J. M., Croft W. B., « A Language Modeling Approach to Information Retrieval », *SIGIR 1998 : Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 275-281, 1998.

Robertson S. E., Zaragoza H., « The Probabilistic Relevance Framework : BM25 and Beyond », *Foundations and Trends in Information Retrieval*, vol. 3, n° 4, p. 333-389, 2009.

Zhai C., Lafferty J. D., « A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval », *SIGIR 2001 : Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 334-342, 2001.