Modèles de RI fondés sur l'information

Stéphane 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. *Dans une premier temps, nous présentons dans cet article une vue analytique des contraintes heuristiques récemment proposées pour les fonctions d'ordonnancement (retrieval function): ces caractérisations permettent ainsi de tester simplement si un modèle de recherche d'information (RI) respecte ces contraintes ou non. De plus, nous examinons un certain nombre de résultats empiriques sur les distributions de fréquences de mots et le rôle central joué par le phénomène de rafale, pour lequel nous proposons une définition formelle. Nous introduisons ensuite une nouvelle famille de modèles probabilistes pour la RI, fondée sur la notion d'information. Lorsque la loi de probabilité sous-jacente est capable de modéliser le phénomène de rafale, alors le modèle devient naturellement valide au sens des contraintes heuristiques. La distribution log-logistique est présentée dans ce contexte et les expériences, menées sur trois collections différentes, illustrent le comportement adéquat de ce modèle; il surpasse Okapi BM25 et les modèles de langues, avec lissage de Jelinek-Mercer ou de Dirichlet, à la fois pour la précision moyenne et la précision en tête de liste, fournit de meilleurs résultats que les modèles DFR (Divergence from Randomness) en précision moyenne et des résultats similaires sur la précision en tête de liste, tout en simplifiant ces modèles.*

ABSTRACT. *We first present in this paper an analytical view of heuristic retrieval constraints which yields simple tests to determine whether a retrieval function satisfies the constraints or not. We then review empirical findings on word frequency distributions and the central role played by burstiness in this context. This leads us to propose a formal definition of burstiness which can be used to characterize probability distributions wrt this phenomenon. We then introduce the family of information-based IR models which naturally captures heuristic retrieval constraints when the underlying probability distribution is bursty and propose a new IR model within this family, based on the log-logistic distribution. The experiments we conduct on three different collections illustrate the good behavior of the log-logistic IR model.*

MOTS-CLÉS : *Modèles théoriques de RI, phénomène de rafale, modèles de langue, modèles DFR*

KEYWORDS: *IR theoretical models, burstiness, language models, Divergence from Randomness*

1. Introduction

Si la recherche d'information (RI) sur le web est dominée par des systèmes apprenant des fonctions d'ordonnancement à partir de log de données, la RI *ad hoc* est largement dominée par des modèles probabilistes avec peu de paramètres à régler, comme Okapi, les modèles de langues et les modèles DFR (*Divergence from Randomness*). Ces derniers sont fondés sur plusieurs distributions de probabilité et hypothèses qui facilitent leur déploiement en pratique. Si ces modèles sont bien fondés d'un point de vue RI (ils satisfont les contraintes heuristiques proposées dans (Fang *et al.*, 2004)), les distributions de probabilités sous-jacentes s'accordent mal avec les données empiriques collectées dans les collections textuelles. Nous explorons dans cet article les liens entre les contraintes heuristiques et les distributions de fréquences de mots afin de proposer un nouveau modèle de RI fondé sur des lois de probabilité qui s'ajustent bien aux données et qui respectent les contraintes heuristiques.

Après avoir présenté une caractérisation analytique des contraintes heuristiques, nous reviendrons sur le phénomène de rafale dans la partie 2. Dans la partie 3, la famille de modèles fondés sur l'information sera introduite, ainsi que le nouveau modèle log-logistique pour la RI. Dans la partie 4, les expériences montreront le bon comportement de ce modèle à travers une série d'expériences qui valideront l'ajustement aux données empiriques et les performances en RI. Nous discuterons alors différents aspects de ce nouveau modèle dans la partie 5 avant de conclure.

Les notations utilisées sont résumées dans le tableau ci-dessous.

Notation	Description
x_w^q	Fréquence du terme w dans la requête q
x_w^d	Fréquence du terme w dans le document d
t_w^d	Version normalisée de x_w^d
N	Nombre de documents dans la collection
M	Nombre de termes d'indexation
F_w	Fréquence de w : $F_w = \sum_d x_w^d$
N_w	Fréquence documentaire de w : $N_w = \sum_d I(x_w^d > 0)$
y_d	Longueur du document d
m	Longueur moyenne des documents d (token)
L	Longueur de la collection (token)
$h(x_w^d, y_d, z_w)$	Fonction d'ordonnancement
z_w	Statistique du mot dans le corpus $z_w = F_w$ or $z_w = N_w$

2. Modèles de RI et distributions de fréquences de mots

Dans cette première partie, nous présentons une caractérisation analytique des contraintes heuristiques des modèles de RI. Nous examinons ensuite le phénomène de rafale et différentes lois de probabilité de fréquences. Cette partie nous permet d'énumérer certains points concernant les modèles de RI et nous aidera à concevoir un nouveau modèle.

2.1. *Contraintes heuristiques*

Nous considérons dans la famille des modèles de RI qui prennent la forme suivante :

$$RSV(q,d) = \sum_{w \in q \cap d} a(x_w^q) h(x_w^d, y_d, z_w, \theta)$$

où θ est un ensemble de paramètres et h une fonction d'ordonnancement qui sera supposée de classe C^2 et définie sur $\mathbb{R}^{+*} \times \mathbb{R}^{+*} \times \mathbb{R}^{+*} \times \Theta$, où Θ représente le domaine des paramètres de θ. a est souvent la fonction identité et x, y et z sont données dans le tableau 1. Les modèles de langues (Zhai *et al.*, 2004), Okapi (Robertson *et al.*, 1994), les modèles DFR (Amati *et al.*, 2002) autant que les modèles vectoriels (Salton *et al.*, 1983) appartiennent à cette famille de modèles. Par exemple, pour le "pivoted normalization retrieval formula" (Singhal *et al.*, 1996), $\theta = (s, m, N)$ et :

$$h(x,y,z,\theta) = \frac{1 + \ln(1 + \ln(x))}{1 - s + s\frac{y}{m}} \ln(\frac{N+1}{z})$$

Fang *et al.* (Fang *et al.*, 2004) ont proposé un ensemble de contraintes qui rendent compte des heuristiques que les modèles de RI doivent satisfaire. Tout d'abord, il est important que les documents avec plus d'occurrences des mots de requête obtiennent de plus grand score que des documents avec moins d'occurrences. Cependant, l'augmentation du score doit être plus petite pour de plus grandes fréquences, puisque la différence par exemple entre 110 et 111 n'est pas aussi important que celle entre 1 et 2 (le nombre d'occurrences a doublé dans le deuxième cas, tandis que l'augmentation est relativement marginale dans le premier cas). En outre, un long document, comparé à un document plus court comportant exactement le même nombre d'occurrences d'un mot de la requête, doit être pénalisé car il est susceptible de couvrir des sujets additionnels à ceux traités dans la requête. Enfin, il est important de diminuer l'importance des mots qui apparaissent dans beaucoup de documents, c.-à-d. qui ont une fréquence documentaire importante, car ces mots ont un faible pouvoir de discrimination. Fang *et al.* ont proposé des critères formels qui rendent compte de ces heuristiques. Nous rappelons ces critères et en proposons une version analytique donnant lieu à de nouvelles conditions, plus faciles à valider, sur la fonction h sous-jacente à un modèle de RI donné. Nous utilisons dans la suite les notations suivantes :

$$\begin{aligned}
\forall (y,z,\theta),\ n \in \mathbb{N}^*,\ a_n &= h(n,y,z,\theta) \\
\forall (x,z,\theta),\ n \in \mathbb{N}^*,\ b_n &= h(x,n,z,\theta) \\
\forall (y,z,\theta),\ n \in \mathbb{N}^*,\ c_n &= h(n+1,y,z,\theta) - h(n,y,z,\theta)
\end{aligned}$$

Critère 1 - TFC1 : Soit $q = w$ une requête avec un seul mot w. Supposons que $y_{d1} = y_{d2}$. si $x_w^{d1} > x_w^{d2}$, alors $RSV(d1,q) > RSV(d2,q)$ (Fang *et al.*).

Proposition 1 : TFC1 $\Longleftrightarrow$ a_n croît. Une condition suffisante pour cela est :

$$\forall (y,z,\theta),\ \frac{\partial h(x,y,z,\theta)}{\partial x} > 0 \quad \textbf{(condition 1)}$$

Critère 2 - TFC2 : Soit $q = w$ une requête avec un seul mot w. Supposons que $y_{d1} = y_{d2} = y_{d3}$ et $x_w^{d1} > 0$. si $x_w^{d2} - x_w^{d1} = 1$ et $x_w^{d3} - x_w^{d2} = 1$, alors $RSV(d2, q) - RSV(d1, q) > RSV(d3, q) - RSV(d2, q)$ (Fang *et al.*).

Proposition 2 : TFC2 $\iff c_n$ décroit. Une condition suffisante pour cela est :

$$\forall (y, z, \theta), \frac{\partial^2 h(x, y, z, \theta)}{\partial x^2} < 0 \quad \textbf{(condition 2)}$$

Critère 3 - LNC1 : Soit $q = w$ une requête et $d1$, $d2$ deux documents. Si pour un mot $w' \notin q$, $x_{w'}^{d2} = x_{w'}^{d1} + 1$ mais pour un autre mot de la requête w, $x_w^{d2} = x_w^{d1}$, alors $RSV(d1, q) \geq RSV(d2, q)$ (Fang *et al.*).

Proposition 3 : LNC1 $\iff b_n$ décroît. Une condition suffisante pour cela est :

$$\forall (x, z, \theta), \frac{\partial h(x, y, z, \theta)}{\partial y} < 0 \quad \textbf{(condition 3)}$$

Critère 4 - TDC : Soit q une requête et $w1$, $w2$ deux mots. Supposons que $y_{d1} = y_{d2}$, $x_{w1}^{d1} + x_{w2}^{d1} = x_{w2}^{d1} + x_{w2}^{d2}$. Si ***idf(w1)*** $\geq$ ***idf(w2)*** et $x_{w1}^{d1} \geq x_{w1}^{d2}$, alors $RSV(d1, q) \geq RSV(d2, q)$ (Fang *et al.*).

Un cas particulier de TDC corresponds au cas où $w1$ apparait seulement dans $d1$ et $w2$ seulement dans $d2$. Dans ce cas , la contrainte peut être reformulée :

speTDC : Soit q une requête et $w1$, $w2$ deux mots. Supposons que $y_{d1} = y_{d2}$, $x_{w1}^{d1} = x_{w2}^{d2}$, $x_{w1}^{d2} = x_{w2}^{d1} = 0$. si ***idf(w1)*** $\geq$ ***idf(w2)***, alors $RSV(d1, q) \geq RSV(d2, q)$.

Proposition 4 :

$$\textit{speTDC} \iff \forall (x, y, \theta), \frac{\partial h(x, y, z, \theta)}{\partial z} < 0 \quad \textbf{(condition 4)}$$

Critère 5 - LNC2 : Soit q une requête. $\forall k > 1$, si $d1$ et $d2$ sont deux documents tels que $y_{d1} = k \times y_{d2}$ et pour tous les mots w, $x_w^{d1} = k \times x_w^{d2}$, alors $RSV(d1, q) \geq RSV(d2, q)$ (Fang *et al.*).

Proposition 5 : LNC2 $\iff \forall (z, \theta),\ (m, n) \in \mathbb{N}^*,\ k > 1,\ h(km, kn, z, \theta) \geq h(m, n, z, \theta)$ **(condition 5)**

Critère 6 - TF-LNC : Soit $q = w$ une requête avec un seul mot w w. si $x_w^{d1} > x_w^{d2}$ et $y_{d1} = y_{d2} + x_w^{d1} - x_w^{d2}$, alors $RSV(d1, q) > RSV(d2, q)$ (Fang *et al.*).

Proposition 6 : TF-LNC $\iff \forall (z, \theta),\ (m, n, p) \in \mathbb{N}^*,\ h(m + p, n + p, z, \theta) > h(m, n, z, \theta)$ **(condition 6)**

Les conditions 1, 3 et 4 expriment le fait que h doit être croissante avec la fréquence du mot dans le document et décroissante avec la longueur du document et la fréquence documentaire du mot (fréquence du mot dans la collection). Notons que la condition 4 ne représente qu'une condition nécessaire pour la contrainte TDC, puisque nous avons considéré qu'une forme particulière (speTDC) de cette contrainte. La condition 2 montre que h, croissante avec la fréquence du mot dans le document, doit être concave, la concavité garantissant le fait que l'augmentation du score décroît avec la grande fréquence des mots dans le document. Enfin, les conditions 5 et 6 régulent l'interaction entre les fréquences et la longueur des documents, c'est-à-dire entre les dérivées par rapport à x et à y. Elles peuvent être vues comme un complément aux conditions 1, 2, 3 et 4 qui donnent la forme générale de la fonction. Nous appellerons les conditions 1, 2, 3 et 4 **conditions de forme** car ce sont elles qui dictent les contraintes les plus fortes sur la fonction h à choisir. Les conditions 5 et 6, qui affinent la forme de h, seront appelées **conditions d'ajustement**.

Il est à noter que la majorité des conditions obtenues sont des conditions suffisantes. Cela provient du fait que Fang *et al.* ont établi des critères dans un cadre discret (sur des entiers) alors que nous travaillons sur des fonctions continues (la grande majorité des fonctions considérées en RI sont en fait continues). Il est en effet possible de construire des fonctions continues qui se comportent bien au niveau des valeurs entières, c'est-à-dire qui vérifient les critères discrets de Fang *et al.*, et de façon chaotique entre ces entiers (même si de telles fonctions seraient inintéressantes pour la RI).

2.2. *Distributions de fréquences de mots*

Si les modèles de RI doivent remplir les conditions précédentes, la plupart de ces modèles reposent néanmoins sur des probabilités de fréquence. Okapi, par exemple, suppose que les fréquences suivent *un mélange de deux distributions de Poisson* (2-Poisson), dans l'ensemble des documents pertinents et dans l'ensemble des documents non pertinents. Les modèles DFR, proposés par Amati et van Rijsbergen (Amati *et al.*, 2002), utilisent plusieurs types de distributions, parmi lesquelles la distribution géométrique, la distribution binomiale et la loi de Laplace jouent un rôle primordial. Les modèles de langues, quant à eux, utilisent principalement la distribution multinomiale.

Des résultats empiriques sur le comportement des mots dans des collections suggèrent néanmoins qu'aucune de ces distributions n'est appropriée pour expliquer correctement la distribution des fréquences des mots. Church et Gale (Church *et al.*, 1995) ont comparé la distribution binomiale et la distribution de Poisson avec des mélanges de Poisson pour décrire les fréquences de mots. Leurs résultats montrent que la distribution négative binomiale (un mélange infini de distributions de Poisson) s'ajuste aux données bien mieux que les autres distributions considérées. Un phénomène important, mis en avant par Church et Gale (Church, 2000) d'une part et Katz (Katz, 1996) d'autre part, est celui du comportement en rafale, ou crépitement (en anglais *burstiness*), qui décrit le fait que les mots, dans un document, tendent à apparaître

par paquets. En d'autres termes, une fois que l'on a observé une occurrence d'un mot dans un document, il est bien plus probable d'observer de nouvelles occurrences de ce mot. La notion de rafale est identique à celle d'effet de post-échantillonnage, décrite par exemple dans Feller ((Feller, 1968)), et qui se traduit par le fait que plus un mot est observé dans un document, plus on a de chances de l'observer par la suite. Le phénomène de rafale a attiré l'attention de diverses communautés. Madsen (Madsen *et al.,* 2005), par exemple, a proposé la distribution *Dirichlet Compound Multinomial* (DCM) afin de modéliser ce phénomène pour la catégorisation et la classification de textes. Elkan (Elkan, 2006) a ensuite proposé d'approcher la distribution DCM par la distribution EDCM, qui en est une simplification. Le phénomène d'attachement préférentiel (*preferential attachment*, (Barabasi *et al.,* 1999) et (Chakrabarti *et al.,* 2006)) dans les grands réseaux tels que le web ou les réseaux sociaux, s'apparente aussi phénomène de rafale : *plus on a, plus on obtiendra*. En ce qui concerne la RI, Xu et Akella (Xu *et al.,* 2008) ont étudié un modèle DCM dans le cadre du principe d'ordonnancement probabiliste (*Probability Ranking Principle*). Une des conclusions qui ressort de leur étude est que la distribution multinomiale n'est pas entièrement appropriée car elle ne tient pas compte de la dépendance des occurrences des mots lorsqu'ils réapparaissent, *i.e.* du phénomène de rafale.

Peu de descriptions opérationnelles du phénomène de rafale ont été proposées. Une des premières descriptions a été avancée par Church et Gale (Church *et al.,* 1995) : pour une distribution $P(X_w)$, sa capacité à modéliser le phénomène de rafale est mesurée par : $B_P = \frac{E_P[X_w]}{P(X_w \geq 1)}$, où E_P dénote l'espérance par rapport à la distribution P. Si cette mesure fournit bien une méthode pour comparer deux distributions de mots, elle ne permet pas, en revanche, de caractériser directement la distribution P . Pour ce faire, Clinchant et Gaussier (Clinchant *et al.,* 2008) ont introduit la définition suivante :

Définition 1 *[Cas discret] Une distribution discrète P est en « rafale ou crépite » si et seulement si pour tout couple d'entiers $(n', n), n' \geq n$:*

$$P(X \geq n' + 1 | X \geq n') > P(X \geq n + 1 | X \geq n)$$

Nous généralisons ici cette définition au cas continu :

Définition 2 *[Cas général] Une distribution continue P est en « rafale ou crépite » si et seulement si la fonction g_ϵ, $\epsilon > 0$, définie par :*

$$g_\epsilon(x) = P(X \geq x + \epsilon | X \geq x) \text{ est telle que : } \forall \epsilon > 0,\ \frac{\partial g_\epsilon(x)}{\partial x} > 0 \qquad [1]$$

(g_ϵ est une fonction strictement croissante en x pour tout ϵ).

Cette définition traduit directement le fait qu'un mot est en rafale s'il est plus facile de le générer une fois qu'il a été généré un certain nombre de fois. Ainsi, les distributions classiques peuvent être classées selon leur capacité à prendre en compte le phénomène de rafale :

– Les distributions binomiales, Poisson et 2-Poisson ne sont pas en rafale. Les modèles de langue et le modèle Okapi ne tiennent donc pas bien compte du comportement des mots dans les collections textuelles.

– Les distributions géométriques et exponentielles sont neutres par rapport au crépitement ($g_\epsilon(.)$ est constante). Le modèle DFR s'appuie donc sur des distributions qui ne sont pas entièrement satisfaisantes.

– La distribution de Pareto est en rafale.

Les distributions négative binomiale et de Weibull peuvent être en rafale ou non, selon la valeur de leurs paramètres (cf. (Clinchant *et al.*, 2008) pour la négative binomiale).

Nous pouvons donc résumer les différents points abordés dans cette partie par les remarques suivantes :

1) Les modèles de RI doivent satisfaire les conditions présentées précédemment ;

2) Les lois de probabilité modélisant les fréquences de mots dans les collections textuelles doivent prendre en compte le phénomène de rafale ;

3) Les modèles de RI actuels utilisent des lois qui ne prennent pas entièrement en compte ce phénomène.

La question naturelle que soulèvent ces remarques est de savoir si l'on peut concevoir un modèle de RI capable à la fois de respecter les contraintes heuristiques (conditions 1 à 6) et de modéliser le phénomène de rafale. La partie suivante est consacrée à la présentation d'un tel modèle.

3. Modèles fondés sur l'information

La plupart des modèles de RI n'utilisent pas simplement le nombre d'occurrences d'un mot dans un document mais plutôt une normalisation de ce nombre. Les modèles de langues, par exemple, utilisent la fréquence relative des mots dans un document et dans la collection. D'autres normalisations incluent la normalisation d'Okapi, et celle de « pivoted length normalization »(Singhal *et al.*, 1996). Récemment, une autre normalisation pour les modèles de langues avec la notion de « verbosité »a été proposée dans (Na *et al.*, 2008). Les modèles DFR, quant à eux, choisissent une des deux normalisations suivantes (où c est une constante multiplicative) :

$$t_w^d \quad = \quad x_w^d c\frac{m}{y_d} \text{ ou } x_w^d \log(1 + c\frac{m}{y_d}) \qquad [2]$$

Plusieurs travaux en RI et en modélisation de collections textuelles se sont attachés à la notion d'information apportée par un terme dans un document. En particulier, plusieurs auteurs, dont l'un des premiers fut Harter (Harter, 1975), ont noté le fait que la distribution des mots significatifs d'un document s'écartait de façon sensible de la distribution des mots non significatifs. De plus, plus le comportement d'un mot au sein d'un document s'écarte du comportement moyen de ce même mot dans la collection, plus ce mot est significatif pour le document. C'est cette idée, qui est une des idées à

la base des modèles DFR, que nous retenons ici pour qualifier les modèles fondés sur l'information. De manière générale, nous considérons la famille de modèles suivante :

$$RSV(q,d) = \sum_{w \in q \cap d} -x_w^q \log Prob(X \geq t_w^d | \lambda_w) \quad [3]$$

où la fonction d'ordonnancement correspond à *l'information moyenne* qu'un document apporte à une requête et est similaire aux modèles Inf_1 pour les modèles DFR. Nous appellerons les modèles de cette famille « modèles d'informations », et montrons maintenant que ces modèles vérifient un certain nombre des conditions de la RI.

$Prob(X \geq t_w^d | \lambda_w)$ est, par définition, une fonction décroissante de t_w^d. Dans la mesure où, pour les normalisations considérées en RI, t_w^d est une fonction croissante de x_w^d et décroissante de la longueur du document y_d, les modèles d"information vérifient les conditions de forme 1 et 3. La condition 2, de concavité, peut se ré-écrire sous la forme :

$$\frac{\partial^2 h(x,y,z,\theta)}{\partial x^2} < 0 \quad \Leftrightarrow \quad -\frac{\partial^2 \log(Prob(X \geq t_w^d))}{\partial (x_w^d)^2} < 0$$

Le théorème suivant (démontré en annexe) montre alors que, si l'on choisit une distribution en rafale pour un modèle d'information, celui-ci vérifie la condition 2.

Théorème 3 *Soit P une distribution de classe C^2. Une condition nécessaire pour que P soit en rafale est :*

$$\frac{\partial^2 \log(P(X \geq x))}{\partial x^2} > 0$$

Ainsi, les modèles d'information, définis par l'équation 3 , dont les distributions sous-jacentes sont en rafale, vérifient les conditions 1, 2 et 3 de la RI, qui correspondent à 3 des 4 conditions de forme. Il ne nous reste donc qu'à choisir, parmi les distributions en rafale, une distribution qui vérifie la dernière condition de forme et les deux conditions d'ajustement. C'est ce que nous faisons dans la suite.

3.1. *La distribution log-logistique*

Il existe en fait plusieurs distributions en rafale qui permettent de satisfaire l'ensemble des conditions de forme et d'ajustement. Nous nous concentrons ici sur une des ces distributions, la distribution log-logistique, car elle étend la distribution négative binomiale utilisée dans des travaux antérieurs (comme Church et Gale (Church *et al.*, 1995), Airoldi (Airoldi *et al.*, n.d.) ou Clinchant et Gaussier (Clinchant *et al.*, 2008)). La distribution log-logistique est définie par (nous fixons ici le paramètre de cette distribution à 1) :

$$P_{LL}(X < x; r, \beta = 1) = \frac{x}{x + r}$$

La distribution log-logistique est en rafale :

$$\forall \epsilon > 0,\ g_\epsilon(x) = P_{LL}(X > x + \epsilon | X > x; r) = \frac{r + x}{r + x + \epsilon} \quad [4]$$

est une fonction strictement croissante en x. Cette distribution nous conduit au modèle d'information suivant :

$$RSV(q, d) \quad = \quad \sum_{w \in q \cap d} -x_w^q \log(\frac{r_w}{x + r_w}) \quad [5]$$

Le paramètre r_w peut être fixé selon les principes généraux utilisés dans le modèle DFR. r_w est alors défini au niveau de la collection et prend les valeurs $\frac{F_w}{N}$ ou $\frac{n_w}{N}$. La fonction h du modèle log-logistique vaut donc :

$$h(x, y, z, \theta) = \log(\frac{\frac{z}{N} + t(x, y)}{\frac{z}{N}})$$

où t est une des fonctions de normalisation présentées précédemment. En utilisant les normalisations correspondant à l'équation 2[1], il est aisé de voir que ce modèle vérifie *toutes* les conditions de forme et d'ajustement de la RI, pour toutes valeurs admissibles de x, y and z. Nous obtenons donc un modèle d'information simple (ce modèle ne nécessite pas le premier principe de normalisation des modèles DFR) et fondé sur une loi qui permet de rendre compte du comportement empirique des mots dans les collections de documents (contrairement aux modèles de langue et au modèle Okapi par exemple). Nous allons maintenant illustrer le comportement de ce modèle.

4. Expériences

Les expériences présentées visent deux objectifs. Le premier est de montrer la qualité d'ajustement aux données de la log-logistique et le deuxième est de démontrer ses performances en recherche d'information. Nous avons utilisé des collections standard de RI : TREC (trec.nist.gov) et CLEF (www.clef-campaign.org). Le tableau 1 indique pour chaque collection le nombre de documents (N), le nombre de termes uniques (M), la longueur moyenne des documents ainsi que le nombre de requêtes. Pour la collection ROBUST, nous avons utilisé l'algorithme de Porter pour normaliser les formes de surface. Pour CLEF et GIRT, les textes ont été lemmatisés, avec décomposition des composés nominaux pour GIRT (collection en allemand).

4.1. *Ajustement aux données*

Pour illustrer le bon comportement de la distribution log-logistique par rapport à la distribution de Poisson, nous avons calculé la statistique du χ^2 pour chaque terme, soit

1. D'autres normalisations conduisent au même résultat ; l'étude des normalisations adéquates pour le modèle log-logistique dépasse le cadre du présent article

Tableau 1. *Caractéristiques des collections*

	N	M	Avg DL	# Requêtes
ROBUST	490 779	992 462	289	250
CLEF03	166 754	80 000	247	60
GIRT	151 319	179 283	109	75

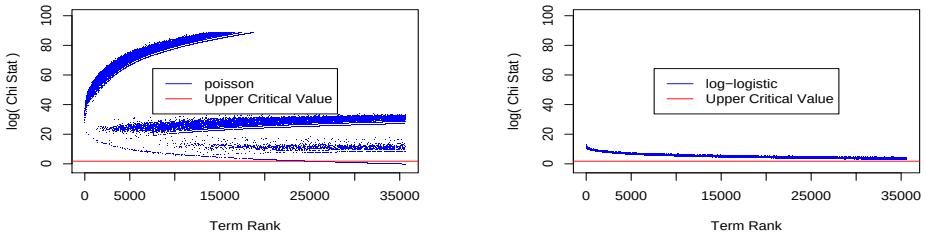

Figure 1. *Statistique du χ^2 pour la Poisson et Log-logistic distributions sur ROBUST*

sous l'hypothèse d'une distribution de Poisson, soit sous celle d'une log-logistique (figure 1). Nous voulons voir les écarts entre la distribution estimée par ces distributions et la distribution empirique. La statistique du χ^2 donne une mesure de cet ajustement. Seuls les termes avec une fréquence documentaire supérieur à 100 ont été étudiés. Les fréquences des termes ont été classés en trois intervalles : $[0, 3[$, $[3, 10[$ et $[10, 100[$, correspondant aux faibles, moyennes et fortes fréquences. On peut montrer mathématiquement que la statistique du χ^2 est la même pour la BNB et la log-logistique sur ces intervalles, grâce à une relation entre la BNB et la log-logistique. Le graphique 1 montre le logarithme de la statistique du χ^2 en fonction du rang du terme dans la collection ROBUST de TREC. Un point sur ce graphique correspond donc à un terme de la collection. La ligne horizontale est la valeur critique supérieure pour un test du χ^2 à 0.05 de confiance. Pour la distribution de Poisson, il y a deux nuages de points principaux. Celui en haut à gauche peut s'expliquer par les termes avec une fréquence dans l'intervalle $[10, 100[$: c'est en effet un événement extrêmement rare avec une loi de Poisson de faible paramètre (ex : 0.05). La deuxième zone, qui ressemble à une large bande, correspond aux termes avec des fréquences dans les deux premières intervalles. Les statistiques des BNB/log-logistique sont dans l'ensemble bien plus faibles. Ces distributions peuvent expliquer le comportement des mots pour toutes les gammes de fréquences, ce qui n'est pas le cas de la Poisson.

4.2. *Performances en RI*

Nous avons comparé le modèle log-logistique aux modèles de langues, avec lissage de Jelinek-Mercer ou de Dirichlet, au modèle Okapi BM25 et aux modèles DFR InL2 et PL2. Pour chaque collection, les requêtes ont été scindées en une partie pour l'apprentissage et une autre pour le test (moitié apprentissage, moitié test). Ce procédé a été répété 10 fois pour chaque collection. Les résultats que nous donnons pour la MAP et la précision à 10 sont les moyennes de ces valeurs sur ces 10 jeux de tests. Les paramètres des modèles sont optimisés sur l'ensemble d'apprentissage et les performances mesurées sur l'ensemble de test. Enfin, un t-test (au niveau 0.05) permet de mesurer la significativité des résultats obtenus. Dans les tableaux suivants, *ROB-t* représente la collection ROBUST avec titres des requêtes, *ROB-d* la collection ROBUST avec titres des requêtes et descriptions (de même pour *CLEF-t* et *CLEF-d*). Les requêtes de la collection GIRT ne sont constituées que d'une seule phrase.

Le modèle log-logistique, dans toutes ces expériences, est tel que $r_w = \frac{n_w}{N}$ et la normalisation de fréquences est celle de l'équation 2 (deuxième membre). Nous notons ce modèle LGD. Comme le paramètre c dans l'équation 2 n'est pas borné, une gamme de valeurs doit être choisie lors de l'optimisation du modèle pendant l'apprentissage. Nous avons choisi des valeurs couramment choisies pour ce type de normalisation dans les modèles DFR : $c \in \{0.25, 0.5, 0.8, 1, 2, 3, 5, 8, 10\}$

Comparaison avec les modèles de langues : Jelinek-Mercer et Dirichlet

Comme le paramètre de Jelinek-Mercer est compris entre 0 et 1, nous avons utilisé une grille régulière de valeurs avec un pas de 0.05 afin de choisir la valeur optimale sur l'ensemble d'apprentissage. Le tableau 2 montre la comparaison de notre modèle (LGD) avec le modèle Jelinek-Mercer (LM). Pour toutes les collections et pour tous types de requêtes, le modèle LGD surpasse significativement le modèle de langue bien que ces deux modèles aient la même complexité.

Tableau 2. *LM-Jelinek-Mercer versus Log-Logistique après 10 divisions ; en gras, les meilleures performances ; * dénote une différence statistiquement significative*

MAP	ROB-d	ROB-t	GIRT	CLEF-d	CLEF-t
LM	26.0	20.7	40.7	49.2	36.5
LGD	**27.2***	**22.5***	**43.1***	**50.0***	**37.5***
P10	ROB-d	ROB-t	GIRT	CLEF-d	CLEF-t
LM	43.8	35.5	67.5	33.0	26.2
LGD	**46.0***	**38.9***	**69.4***	**33.6***	**26.6***

Afin d'observer les comportements de ces deux modèles par rapport à leur paramètre (λ pour Jelinek-Mercer et c pour LGD), la courbe 4.2 dessine la MAP pour différentes valeurs de λ. Pour toutes ces valeurs, c est défini par $c = \frac{\lambda}{1-\lambda}$, ce qui permet de comparer les deux modèles pour n'importe quel λ dans $]0; 1[$. Sauf pour les faibles valeurs de lambda, le modèle LGD domine largement le modèle de langues, ce qui montre que ce modèle est meilleur de façon consistante.

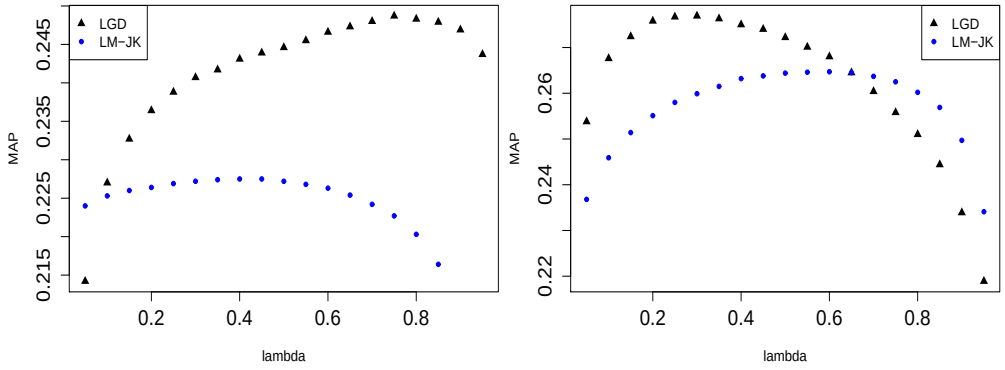

Figure 2. *MAP vs lambda. ROB-t à gauche et ROB-d à droite*

Pour le lissage de Dirichlet, le paramètre a été choisi parmi un ensemble standard de valeurs : $\{10, 50, 100, 200, 500, 800, 1000, 1500, 2000, 5000, 10000\}$.
Le tableau 3 compare le modèle LGD et le modèle de langues avec lissage de Dirichlet (LM). Ces résultats sont similaires aux résultats obtenus avec Jelinek-Mercer, sauf pour ROB-t où Dirichlet surpasse LGD. Sur les autres collections, la MAP et la précision à 10 sont meilleures avec le modèle LGD.

Tableau 3. *LM-Dirichlet vs Log-Logistique après 10 divisions ; en gras, les meilleures performances ;* * *dénote une différence statistiquement significative*

MAP	ROB-d	ROB-t	GIRT	CLEF-d	CLEF-t
LM	27.1	**25.1**	41.1	48.5	36.2
LGD	**27.4***	25.0	**42.1***	**49.7***	**36.8***
P10	ROB-d	ROB-t	GIRT	CLEF-d	CLEF-t
LM	45.6	**44.7***	68.6	33.8	28.4
LGD	**46.2***	44.4	**69.0**	**34.5***	**28.6**

Comparaison avec Okapi BM25

Nous suivons la même méthodologie afin de comparer le modèle log-logistique avec le modèle Okapi BM25. Nous avons seulement optimisé le paramètre k_1 de BM25 parmi l'ensemble des valeurs suivantes : $\{0.3, 0.5, 0.8, 1.0, 1.2, 1.5, 1.8, 2, 2.2, 2.5\}$. Les autres paramètres b et k_3 prennent la valeur par défaut implantée dans Lemur (0.75 et 7). Le tableau 4 montre les résultats de ces deux modèles. Le modèle log-logistique est significativement meilleur sur ROBUST et GIRT alors que sur CLEF les différences de performance ne sont pas significatives.

Tableau 4. *BM25 versus Log-Logistique après 10 divisions ; en gras, les meilleures performances ; * dénote une différence statistiquement significative*

MAP	ROB-d	ROB-t	GIRT	CLEF-t	CLEF-d
BM25	26.8	22.4	39.8	**34.9**	46.8
LGD	**28.2***	**23.5***	**41.4***	34.8	**48.0**
P10	ROB-d	ROB-t	GIRT	CLEF-t	CLEF-d
BM25	45.9	42.6	62.6	28.5	33.7
LGD	**46.5**	**44.3***	**66.6***	**28.7**	**34.4**

Comparaison avec les modèle DFR InL2 et PL2

Nous avons choisi les modèles InL2 et PL2 pour comparer notre modèle aux modèles DFR, car ces modèles ont été validés dans plusieurs travaux précédents. InL2 suppose une distribution géométrique des termes dans la collection, distribution normalisée avec une loi de Laplace, alors que PL2 repose sur une loi de Poisson et une loi de Laplace. On peut cependant remarquer que ces deux modèles utilisent des distributions discrètes avec des valeurs continues. Il sont donc déficients d'un point de vue théorique, ce qui n'est pas le cas du modèle LGD. Comme ces modèles utilisent la même normalisation de fréquences que LGD (équation 2), nous utilisons le même ensemble pour optimiser c : $c \in \{0.25, 0.5, 0.8, 1, 2, 3, 5, 8, 10\}$. Le tableau 5 compare le modèle LGD avec le modèle InL2 et PL2. Cette fois, les résultats sont plus contrastés qu'avec le modèle de langues. Plus précisément, les deux modèles obtiennent des résultats comparables pour la précision à 10, l'un meilleur sur GIRT, l'autre sur ROBUST. Pour la MAP, LGD est significativement meilleur sur ROBUST et GIRT, les deux modèles ayant un comportement similaire sur CLEF. Ces résultats sont d'autant plus intéressants que le modèle log-logistique est plus simple que le modèle InL2. PL2 et LGD obtiennent aussi des performances similaires avec un avantage pour LGD sur la MAP et un léger avantage pour PL2 sur la P10. Nous revenons sur ce point dans la discussion suivante.

5. Discussion

Le modèle log-logistique satisfait les contraintes de la RI décrites dans la partie 2 et utilise une loi de probabilité modélisant le phénomène de rafale. Ce modèle a de nombreuses ressemblances avec les modèles DFR. La famille de modèles DFR proposée par Amati et van Rijsbergen (Amati *et al.*, 2002) repose sur le contenu informatif des mots dans un document, une quantité qui est ensuite corrigée par le risque d'accepter chaque terme comme un descripteur d'un document (*first normalization principle*). La renormalisation des fréquences par rapport à la longueur d'un document est le deuxième principe de normalisation utilisé (*second normalization principle*). Il est en fait commun à tous les modèles de RI. Dans DFR, le contenu informatif $\mathit{Inf}_1(t_w^d)$ d'un terme est tout d'abord calculé à partir d'une première distribution de probabilité : $\mathit{Inf}_1(t_w^d) = -\log Prob_1(t_w^d)$. Le premier principe de normalisation introduit une

Tableau 5. *INL2 et PL2 versus Log-Logistique après 10 divisions ; en gras, les meilleures performances ; * dénote une différence statistiquement significative*

MAP	ROB-d	ROB-t	GIRT	CLEF-d	CLEF-t
INL	27.7	24.8	42.5	47.7	**37.5**
LGD	**28.5***	**25.0***	**43.1***	**48.0**	37.4
P10	ROB-d	ROB-t	GIRT	CLEF-d	CLEF-t
INL	**47.7***	43.3	67.0	**33.4**	**27.3**
LGD	47.0	**43.5**	**69.4***	33.3	27.2

MAP	ROB-d	ROB-t	GIRT	CLEF-t	CLEF-d
LGD	**27.3***	24.7	40.5	**36.2**	**47.5**
PL2	26.2	**24.8**	**40.6**	36.0	47.2
P10	ROB-d	ROB-t	GIRT	CLEF-t	CLEF-d
LGD	**46.6**	43.2	66.7	28.5	**33.7**
PL2	46.4	**44.1***	**68.2***	**28.7**	33.1

deuxième distribution de probabilité, conduisant à un modèle complet qui combine ces deux types d'information. Les modèles DFR peuvent donc être vus comme des modèles d'information, définis par l'équation 3, mais corrigés par Inf_2. Ils sont donc plus complexes de ce point de vue. De plus, les modèles DFR choisissent des distributions discrètes pour $Prob_1$ et $Prob_2$, avant de les appliquer à des données continues (issues du deuxième principe de normalisation). Ils sont donc déficients d'un point de vue théorique. Au contraire, les modèles d'information et leur réalisation au travers du modèle log-logistique sont plus simples, bien fondés théoriquement et conduisent à des résultats aussi bons, voire meilleurs, sur les collections considérées ici. Ils conduisent aussi à des formules plus simple (comparer PL2 à LGD). Le cadre théorique que nous avons développé montre de plus que le choix de distributions en rafale dans cette famille de modèles conduit naturellement à la satisfaction d'une des conditions de forme de la RI.

6. Conclusion

Nous avons présenté dans cet article une caractérisation analytique des contraintes heuristiques des modèles de RI, contraintes initialement proposées dans un cadre discret par (Fang *et al.*, 2004). Cette caractérisation conduit à des conditions simples permettant de déterminer si une fonction d'ordonnancement est correcte ou non. Ensuite, nous avons examiné différents résultats expérimentaux sur la modélisation probabiliste des fréquences de mots et avons mis en avant le rôle important du phénomène de rafale. Ceci nous a amené à proposer une définition formelle qui permet de caractériser les lois de probabilités capables de modéliser ce phénomène. Nous avons ensuite introduit les modèles d'information, pour lesquels la condition de concavité (condition 2) est naturellement respectée si l'on choisit une distribution en rafale (théorème 3). De plus, deux autres conditions de forme (conditions 1 et 3) sont satisfaites si l'on

choisit des renormalisations de fréquences classiques. Enfin, le modèle log-logistique a été introduit dans ce cadre. Outre les conditions précédentes, ce modèle satisfait la dernière condition de forme (condition 4) et les deux conditions d'ajustement (conditions 5 et 6). Les expériences, menées sur trois collections différentes, illustrent le bon comportement de ce modèle. Il surpasse notamment les modèles de langues avec lissage de Jelinek-Mercer et de Dirichlet (à la fois sur la MAP et la P@10), ainsi que le modèle Okapi BM25. Il surpasse le modèle DFR InL2 en terme de MAP. Il a des performance comparables à ce dernier pour la P@10 et obtient des résultats comparables à PL2. Dans le futur, nous présenterons une nouvelle distribution qui peut être utilisée dans le cadre des modèles d'information, ainsi qu'une extension naturelle des modèles d'information pour la boucle de rétro-pertinence (*pseudo-relevance or blind feebdack*).

Remerciements Nous remercions les relecteurs pour leurs remarques utiles.

7. Bibliographie

Airoldi E. M., Cohen W. W., Fienberg S. E., « Bayesian Methods for Frequent Terms in Text : Models of Contagion and the &Delta ;&sup2 ; Statistic », n.d.

Amati G., Rijsbergen C. J. V., « Probabilistic models of information retrieval based on measuring the divergence from randomness », *ACM Trans. Inf. Syst.*, vol. 20, n° 4, p. 357-389, 2002.

Barabasi A. L., Albert R., « Emergence of scaling in random networks », *Science*, vol. 286, n° 5439, p. 509-512, October, 1999.

Chakrabarti D., Faloutsos C., « Graph mining : Laws, generators, and algorithms », *ACM Comput. Surv.*, vol. 38, n° 1, p. 2, 2006.

Church K. W., « Empirical estimates of adaptation : the chance of two noriegas is closer to p/2 than p2 », *Proceedings of the 18th conference on Computational linguistics*, Association for Computational Linguistics, Morristown, NJ, USA, p. 180-186, 2000.

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

Clinchant S., Gaussier É., « The BNB Distribution for Text Modeling », *ECIR*, p. 150-161, 2008.

Elkan C., « Clustering documents with an exponential-family approximation of the Dirichlet compound multinomial distribution », *in* , W. W. Cohen, , A. Moore (eds), *ICML*, vol. 148 of *ACM International Conference Proceeding Series*, ACM, p. 289-296, 2006.

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.

Feller W., *An Introduction to Probability Theory and Its Applications, Vol. I*, Wiley, New York, 1968.

Harter S. P., « A probabilistic approach to automatic keyword indexing. Part I : On the distribution of specialty words in a technical literature », *ASIS*, 1975.

Katz S. M., « Distribution of content words and phrases in text and language modelling », *Nat. Lang. Eng.*, vol. 2, n° 1, p. 15-59, 1996.

Madsen R. E., Kauchak D., Elkan C., « Modeling word burstiness using the Dirichlet distribution », *in* , L. D. Raedt, , S. Wrobel (eds), *ICML*, vol. 119 of *ACM International Conference Proceeding Series*, ACM, p. 545-552, 2005.

Na S.-H., Kang I.-S., Lee J.-H., « Improving Term Frequency Normalization for Multi-topical Documents and Application to Language Modeling Approaches », *in* , C. Macdonald, , I. Ounis, , V. Plachouras, , I. Ruthven, , R. W. White (eds), *ECIR*, vol. 4956 of *Lecture Notes in Computer Science*, Springer, p. 382-393, 2008.

Robertson S. E., Walker S., « Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval », *SIGIR '94 : Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, Springer-Verlag New York, Inc., New York, NY, USA, p. 232-241, 1994.

Salton G., McGill M. J., *Introduction to Modern Information Retrieval*, McGraw-Hill, Inc., New York, NY, USA, 1983.

Singhal A., Buckley C., Mitra M., « Pivoted document length normalization », *SIGIR '96 : Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, New York, NY, USA, p. 21-29, 1996.

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

Preuve du Théorème 3 Rappelons la propriété 3 : *Soit P une distribution de probabilité de classe C^2. Une condition nécessaire pour que P soit en rafale est* : $\frac{\partial^2 \log(P(X \geq x))}{\partial x^2} > 0$

Preuve Soit P une loi de probabilité continue de classe C^2 en rafale. $\forall y > 0$, la fonction g_y definie par :

$$\forall y > 0,\ g_y(x) = P(X \geq x + y | X \geq x) = \frac{P(X \geq x + y)}{P(X \geq x)}$$

est croissante en x par définition. Soit F la fonction cumulative de probabilité. Alors, $g_y(x) = \frac{F(x+y)-1}{F(x)-1}$. Pour y suffisamment petit, une expansion de Taylor de $F(x + y)$ donne :

$$g_y(x) \simeq \frac{F(x) + yF'(x) - 1}{F(x) - 1} = g(x)$$

où F' signifie $\frac{\partial F}{\partial x}$. En dérivant g par rapport à x et en considérant le signe de g', on obtient :

$$sg[g'] = sg[F''F - F'' - F'^2] = sg[(\frac{F'}{F-1})']$$

$$= sg[(\log(1-F))''] = sg[(\log P(X \geq x))]$$

Comme g_y est croissante en x, g l'est aussi et donc $\frac{\partial^2 \log(P(X \geq x))}{\partial x^2} > 0$.