Retrieval Constraints and Word Frequency Distributions - A Log-logistic Model for IR

Stéphane Clinchant, Éric Gaussier

Noname manuscript No.
(will be inserted by the editor)

Retrieval Constraints and Word Frequency Distributions
A Log-logistic Model for IR

Stéphane Clinchant · Eric Gaussier

the date of receipt and acceptance should be inserted later

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 with respect to 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 several collections illustrate the good behavior of the log-logistic IR model: It significantly outperforms the Jelinek-Mercer and Dirichlet prior language models on most collections we have used, with both short and long queries and for both the MAP and the precision at 10 documents. It also compares favorably to BM25 and has similar performance to classical DFR models such as InL2 and PL2.

1 Introduction

If Information Retrieval (IR) on the web is dominated by systems learning ranking functions from log data, standard *ad hoc* IR is largely dominated by probabilistic systems with few parameters to set, as Okapi, the language models or the Divergence from Randomness (DFR) models. These latter models are based on several probabilistic distributions and assumptions which help their deployment in pratical situations. If these models are well founded from an information retrieval point of view (they satisfy the heuristic retrieval constraints given in [9] for example), the probability distributions they rely on yield in general a poor fit to empirical data. Thus, in the "model word frequency distributions to retrieve documents" approach, the first part (model word frequency distributions) is somehow neglected with respect to the second part (retrieve documents) in most models.

Xerox Research Center Europe, 6 chemin de Maupertuis 38240, Meylan France. E-mail: stephane.clinchant@xrce.xerox.com ·
LIG, Univ. Grenoble I, BP 53 - 38041 Grenoble cedex 9, Grenoble France. E-mail: eric.gaussier@imag.fr

We present in this paper a new IR model, based on probability distributions fitting well empirical data, and satisfying heuristic retrieval constraints. To do so, we first explore the links between heuristic retrieval constraints and word frequency distributions. After proposing an analytical view of heuristic retrieval constraints which extends the work presented in [9] and yields simple tests to determine whether a retrieval function satisfies the constraints or not, we review empirical findings on word frequency distributions and the central role played by burstiness in this context. This is the subject of Section 2. In Section 3, we analyze DFR models thanks to the retrieval constraints we reformulated. In Section 4, we introduce the family of information-based IR models and develop, within this family, a new IR model based on the log-logistic distribution. In Section 5, we illustrate the good behavior of our model through a series of experiments which validate the good fit it provides to empirical data and the good performance it yields in IR when compared to language models and Divergence from Randomness models. We then discuss several aspects of our approach in Section 6, prior to conclude.

The notations we use throughout the paper are summarized in table 1.

Table 1 Notations

Notation	Description
x_w^q	Number of occurrences of term w in query q
x_w^d	Number of occurrences of term w in document d
t_w^d	Normalized version of x_w^d
N	Number of documents in the collection
M	Number of terms in the collection
F_w	Number of occurrences of w in the collection: $F_w = \sum_d x_w^d$
N_w	Document frequency of w : $N_w = \sum_d I(x_w^d > 0)$
y_d	Length of document d, in tokens
m	Average document length, in tokens
L	Length of collection d, in tokens
$h(x_w^d, y_d, z_w)$	Base retrieval function with
z_w	$z_w = F_w$ or $z_w = N_w$

2 IR models and word frequency distributions

We first present in this section an analytical version of heuristic retrieval constraints which underlie most IR models. We then review several studies of word frequency distributions, and emphasize the role played by burstiness in these studies. This section thus introduces a few facts concerning IR models deployed over text collections, facts that will help in building a new IR model.

2.1 Heuristic retrieval constraints

Following Fang *et al.* [9], who proposed formal definitions of heuristic retrieval constraints which can be used to assess the validity of an IR model, we introduce here analytical conditions a retrieval function should satisfy to be valid. We consider here

retrieval functions, denoted RSV, of the form:

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

where ω is a set of parameters and where h, the form of which depends on the IR model considered, is assumed to be of class C^2 and defined over $\mathbb{R}^{+*} \times \mathbb{R}^{+*} \times \mathbb{R}^{+*} \times \Omega$, where Ω represents the domain of the parameters in ω and a is often the identity function[1]. Language models [19], Okapi [15] and Divergence from Randomness [2] models as well as vector space models [16] all fit within the above form. For example, for the pivoted normalization retrieval formula [17], $\omega = (s, m, N)$ and:

$$h(x,y,z,\omega) = I(x>0) \frac{1+\ln(1+\ln(x)^{I(x>0)})}{1-s+s\frac{y}{m}} \ln(\frac{N+1}{z})$$

where I is an indicator function which equals 1 when it argument is true and 0 otherwise. A certain number of hypotheses, experimentally validated, sustain the development of IR models. In particular, it is important that documents with more occurrences of query terms get higher scores than documents with less occurrences. However, the increase in the retrieval score should be smaller for larger term frequencies, inasmuch as the difference between say 110 and 111 is not as important as the one between 1 and 2 (the number of occurrences has doubled in the second case, whereas the increase is relatively marginal in the first case). In addition, longer documents, when compared to shorter ones with exactly the same number of occurrences of query terms, should be penalized as they are likely to cover additional topics than the ones present in the query. Lastly, it is important, when evaluating the retrieval score of a document, to weigh down terms occurring in many documents, i.e. which have a high document/collection frequency, as these terms have a lower discrimination power. Fang et al. [9] proposed formal criteria to account for these phenomena. We recall here these criteria and provide an analytical version of them which leads to conditions on h which can be easily tested (the names of the different criteria are directly borrowed from Fang et al. [9]).

Criterion 1 - TFC1: Let $q = w$ a query with only word w. Suppose that $y_{d1} = y_{d2}$. if $x_w^{d1} > x_w^{d2}$, then $RSV(d1,q) > RSV(d2,q)$ (Fang *et al.*).

Proposition 1: TFC1 $\iff \forall (y,z,\omega),\ n \in \mathbb{N}^*,\ h(n,y,z,\omega)$ is increasing. A sufficient condition is:

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

Criterion 2 - TFC2: Let $q = w$ a query with only word w. Suppose that $y_{d1} = y_{d2} = y_{d3}$ et $x_w^{d1} > 0$. If $x_w^{d2} - x_w^{d1} = 1$ et $x_w^{d3} - x_w^{d2} = 1$, then $RSV(d2,q) - RSV(d1,q) > RSV(d3,q) - RSV(d2,q)$ (Fang *et al.*).

Proposition 2: TFC2 $\iff \forall (y,z,\omega),\ n \in \mathbb{N}^*,\ h(n+1,y,z,\omega) - h(n,y,z,\omega)$ is decreasing. A sufficient condition is:

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

[1] A function of class C^2 is a function for which second derivatives exist and are continuous.

Criterion 3 - LNC1: Let $q = w$ a query and $d1$, $d2$ two documents. If, for a word $w' \notin q$, $x_{w'}^{d2} = x_{w'}^{d1} + 1$ but for another query word w, $x_w^{d2} = x_w^{d1}$, then $RSV(d1, q) \geq RSV(d2, q)$ (Fang *et al.*).

$\forall(x, z, \omega),\ n \in \mathbb{N}^*,$ Let $b_n = h(x, n, z, \omega)$.
Proposition 3: LNC1 $\iff \forall(x, z, \omega),\ n \in \mathbb{N}^*$, $h(x, n, z, \omega)$ is decreasing. A sufficient condition is:

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

Criterion 4 - TDC: Let q a query et $w1$, $w2$ two words. Suppose that $y_{d1} = y_{d2}$, $x_{w1}^{d1} + x_{w2}^{d1} = x_{w2}^{d1} + x_{w2}^{d2}$. If *idf(w1)* $\geq$ *idf(w2)* et $x_{w1}^{d1} \geq x_{w1}^{d2}$, then $RSV(d1, q) \geq RSV(d2, q)$ (Fang *et al.*).

A special case of TDC corresponds to the case where $w1$ occurs only in document $d1$ and $w2$ only in $d2$. In such a case, the constraints can be written as:

speTDC: Let q a query and $w1$, $w2$ two words. Suppose that $y_{d1} = y_{d2}$, $x_{w1}^{d1} = x_{w2}^{d2}$, $x_{w1}^{d2} = x_{w2}^{d1} = 0$. If *idf(w1)* $\geq$ *idf(w2)*, then $RSV(d1, q) \geq RSV(d2, q)$.

Proposition 4:

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

Criterion 5 - LNC2: Let q a query. $\forall k > 1$, if $d1$ and $d2$ are two documents such that $y_{d1} = k \times y_{d2}$ and for all words w, $x_w^{d1} = k \times x_w^{d2}$, then $RSV(d1, q) \geq RSV(d2, q)$ (Fang *et al.*).

Proposition 5: $LNC2 \iff$

$$\forall(z, \omega),\ (m, n) \in \mathbb{N}^*,\ k > 1,\ h(km, kn, z, \omega) \geq h(m, n, z, \omega) \quad \textbf{(condition 5)}$$

Criterion 6 - TF-LNC: Let $q = w$ a query with only word w. if $x_w^{d1} > x_w^{d2}$ et $y_{d1} = y_{d2} + x_w^{d1} - x_w^{d2}$, then $RSV(d1, q) > RSV(d2, q)$ (Fang *et al.*).

Proposition 6: $TF - LNC \iff$

$$\forall(z, \omega),\ (m, n, p) \in \mathbb{N}^*,\ h(m + p, n + p, z, \omega) > h(m, n, z, \omega) \quad \textbf{(condition 6)}$$

Conditions 1, 3 and 4 directly state that h should be increasing with the term frequency, and decreasing with the document length and the document/collection frequency. Conditions 1 and 2 (mentioned by Fang *et al.* [9]) state that h should be an increasing, concave function of the term frequency, the concavity ensuring that the increase in the retrieval score will be smaller for larger term frequencies. Lastly, conditions 5 and 6 regulate the interaction between frequency and document length, i.e. between the derivatives wrt to x and y. They allow to adjust the functions h satisfying conditions 1, 2, 3 and 4. In the remainder, we will refer to conditions 1, 2, 3 and 4 as the **form conditions** and conditions 5 and 6 as the **adjustment conditions**.

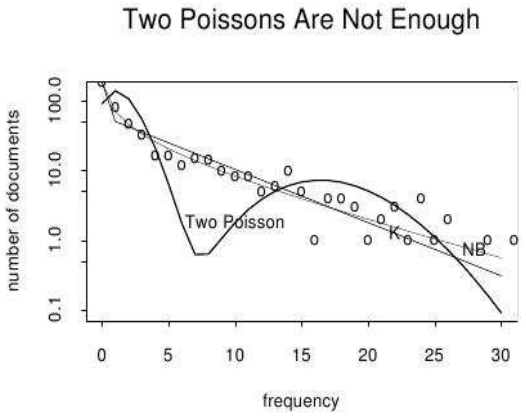

Fig. 1 Typical fit of Two Poisson, Negative Binomial (NB) and Katz mixture (K) for frequencies of a given word

2.2 Word frequency distributions

If IR models have to fulfill the above conditions, the most recent and widely used models also rely on word probability distributions with their own specificities. In Okapi, for example, it is assumed that word frequencies follow *a mixture of two Poisson distributions* (2P), in both the relevant and irrelevant sets. The Divergence from Randomness (DFR) framework proposed by Amati and van Rijsbergen [2] makes use of several distributions, among which the geometric distribution, the binomial distribution and Laplace law of succession play the major role. Language models are, for themselves, built upon the multinomial distribution, which amounts to consider binomial distributions for individual words.

Empirical findings on how words behave in text collections however suggest that none of the above distributions is appropriate for accurately describing word frequencies. Church and Gale ([6]) compared the binomial and Poisson distributions with mixtures of Poisson to model word frequency distributions. Their results indicate that the negative binomial distribution, which is an infinite mixture of Poisson distributions, fits the data better than the other distributions. Figure 1 (from [6]) plots the number of documents (*y-axis*) with exactly x occurrences (*x-axis*) of a given word. As one can observe, the 2-Poisson model yields a poor fit to the data. In a later work, Church also showed experimentally that words tends to reappear in documents [5], a phenomenon referred to as *positive adaptation* or *burstiness*.

The term "burstiness" describes the behavior of words which tend to appear in bursts, i.e., once they appear in a document, they are more likely to appear again. The notion of burstiness is similar to the one of *aftereffect* of future sampling ([10]), which describes the fact that the more we find a word in a document, the higher the expectation to find new occurrences. Burstiness has recently received a lot of attention from different communities. Madsen [13], for example, proposed to use the Dirichlet Compound Multinomial (DCM) distribution in order to model burstiness in the context of text categorization and clustering. Elkan [8] then approximated the DCM distribution

by the EDCM distribution, for which learning time is faster, and showed the good behavior of the model obtained on different text clustering experiments. A related notion is the one of *preferential attachment* ([3] and [4]) often used in large networks, such as the web or social networks. It conveys the same idea: *the more we have, the more we will get*. In the context of IR, Xu and Akella [18] studied the use of a DCM model within the Probability Ranking Principle, and argue that multinomial distributions alone are not appropriate for IR within this principle (quoting):

> Because the multinomial distribution assumes the independence of the word repetitive occurrences, it results in a score function which incorporates undesired linearity in term frequency. To capture the concave property and penalize document length in the score function, a more appropriate distribution should be able to model the dependency of word repetitive occurrences.

The *dependency of word repetitive occurrences* is directly linked to burstiness. More formally, for a word probability distribution $P(X_w)$, [6] measures its burstiness through the quantity:

$$B_P = \frac{E_P[X_w]}{P(X_w \geq 1)}$$

where E_P denotes the expectation with respect to P. The drawback of this measure is that it does not give a clear understanding on whether a given distribution accounts for burstiness or not. Clinchant and Gaussier [7] introduced the following definitions (slightly simplified here for clarity's sake) in order to characterize discrete distributions which can account for burstiness:

Definition 1 [Discrete case] A discrete distribution P is bursty *iff* for all integers $(n', n), n' \geq n$:

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

We generalize this definition to the continuous case as follows:

Definition 2 [General case] A distribution P is bursty *iff* the function g_ϵ defined by:

$$g_\epsilon(x) = P(X \geq x + \epsilon | X \geq x),\ \epsilon > 0$$

is a strictly increasing function of x for all $\epsilon > 0$. A distribution which verifies this condition is said to be *bursty*.

This definition directly translates the fact that, with a bursty distribution, it is easier to generate higher values of X once lower values have been observed. Armed with these definitions, one can characterize standard distributions wrt burstiness:

- The binomial and Poisson distributions are not bursty,
- The geometric and exponential distributions are neutral wrt burstiness, i.e. the function is neither increasing nor decreasing,
- The Pareto distribution is bursty.

2.3 Summary

We can sum up the different points developed in this section as follows:

1. IR models have to fulfill heuristic retrieval constraints stated in conditions 1 to 6,

2. Word frequency distributions should be bursty according to the above definitions,
3. Word frequency distributions used in standard IR models are usually not bursty.

The question which naturally follows from these findings is whether one can build an IR model based on bursty distributions and compliant with the heuristic retrieval constraints. The next section is devoted to the presentation of such a model. Before that, we analyze the Divergence from Randomness framework with respect to the retrieval constraints.

3 The DFR Framework

The Divergence from Randomness (DFR) framework proposed by Amati and van Rijsbergen [2] is currently one of the most successful IR models. It is based on the informative content provided by the occurrences of terms in documents, denoted Inf_1, a quantity which is then corrected by the risk of accepting a term as a descriptor in a document, denoted Inf_2, and associated to the *first normalization principle*. Lastly, raw occurrences are normalized by the length of the document, a normalization which corresponds to the *second normalization principle*. In the remainder, $t(x_w^d, y_d)$ will denote the normalized form of x_w^d. The informative content $Inf_1(t(x_w^d, y_d))$ is based on a first probability distribution and is defined as: $Inf_1(t(x_w^d, y_d)) = -\log Prob_1(t(x_w^d, y_d))$. The risk of accepting a term (first normalization principle) is based on a second probability distribution and is defined as: $Inf_2(t(x_w^d, y_d)) = 1 - Prob_2(t(x_w^d, y_d))$. For example, using the Laplace law of succession for the first normalization principle, one obtains the following retrieval function:

$$RSV(q,d) = \sum_{w \in q \cap d} x_w^q \overbrace{\left(\frac{1}{t(x_w^d, y_d) + 1}\right)}^{\mathrm{Inf}_2(t(x_w^d, y_d))} \mathrm{Inf}_1(t(x_w^d, y_d)) \tag{1}$$

We now review the two normalization principles behind DFR models.

3.1 The Second Normalization Principle

The second normalization principle aims at normalizing the number of occurrences of words in documents by the document length, as a word is more likely to have more occurrences in a long document than in a short one. The different normalizations considered in the literature transform raw number of occurrences into positive real numbers. Language models for example use the relative frequency of words in the document and the collection. Other classical term normalization schemes include the well know Okapi normalization, as well as the pivoted length normalization [17]. More recently, [14] propose another formulation for the language model using the notion of verbosity.

DFR models usually adopt one of the two following term frequency normalizations (c is a multiplying factor):

$$t_w^d = t(x_w^d, y_d) = x_w^d c \frac{m}{y_d} \tag{2}$$

$$t_w^d = t(x_w^d, y_d) = x_w^d \log(1 + c\frac{m}{y_d}) \tag{3}$$

The important point about the second normalization principle is that, to be fully compliant with these definitions, the probability distribution functions at the basis of DFR models should be continuous distributions as the variables considered are continuous[2]. This is not the case for DFR models proposed so far which rely on discrete distributions.

3.2 The First Normalization Principle

The intuition behind Inf_1 is simple. Let $P(t(x_w^d, y_d)|\theta_w)$ represent the probability of $t(x_w^d, y_d)$ (normalized) occurrences of term w in document d according to parameters θ_w which are estimated or set on the basis of a random distribution of w in the collection. If $P(t(x_w^d, y_d)|\theta_w)$ is low, then the distribution of w in d deviates from its distribution in the collection, and w is important to describe the content of d. In this case, Inf_1 will be high. On the contrary, if $P(x_w^d|\theta_w)$ is high, then w behaves in d as expected from the whole collection and, thus, does not provide much information on d (Inf_1 is low). Inf_1 thus captures the importance of a term in a document through its deviation from an average behavior estimated on the whole collection. The question which thus arises is why one should need to normalize it. In other words, what is the role of the first normalization principle?

Amati and van Rijsbergen [2] consider five basic IR models for $Prob_1$: the binomial model, the Bose-Einstein model, which can be approximated by a geometric distribution, the *tf-idf* model (denoted *I(n)*), the tf-itf model (denoted *I(F)*) and the *tf-expected-idf* model (denoted $I(n_e)$). For the last four models, Inf_1 takes the form:

$$Inf_1(t(x_w^d, y_d)) = \begin{cases} t(x_w^d, y_d)\log(1+\frac{N}{z_w}) + \log(1+\frac{z_w}{N}) \\ t(x_w^d, y_d)\log(\frac{N+1}{z_w+0.5}) \end{cases}$$

where the first line corresponds to the geometric distribution, and the second one to *I(n)*, *I(F)* and $I(n_e)$ (z_w being respectively equal to n_w, F_w and $n_{w,e}$, the latter representing the expected number of documents containing term w). We assume in the remainder that $t(x_w^d, y_d)$ is given either by equation 2 or 3. The conclusions we present below are the same in both cases.

Were we to base a retrieval function on the above formulation of Inf_1 only, our model would be defined by:

$$\omega = (x_w^q, m, N)$$

$$h(x, y, z, \omega) = \begin{cases} \left(t(x,y)\log(1+\frac{N}{z}) + \log(1+\frac{z}{N})\right) \\ \left(t(x,y)\log(\frac{N+1}{z+0.5})\right) \end{cases}$$

where the first line still corresponds to the geometric distribution, and the second one to *I(n)*, *I(F)* and $I(n_e)$. It is straightforward to see that models *I(n)*, *I(F)* and $I(n_e)$ meet conditions 1, 3 and 4 and that the model for the geometric distribution verifies conditions 1 and 3, but only partly condition 4, as the derivative can be positive for some values of z, N and t. All models however fail condition 2 as, in all cases, $\frac{\partial^2 h(x,y,z,\omega)}{\partial x^2} = 0$. Hence, Inf_1 alone, for the geometric distribution and the models *I(n)*,

[2] Furthermore, as these variables are positive, the support of the distributions to be considered should be (or included in) $[0; \infty)$.

I(F) and *I(n_e)*, is not sufficient to define a valid IR model[3]. One can thus wonder whether Inf_2 serves to make the model compliant with condition 2. We are going to see that this is indeed the case.

Two quantities are usually used for Inf_2 in DFR models: the normalization L or the normalization B. They both lead to the following form:

$$\mathrm{Inf}_2 = \frac{a}{t(x_w^d, y_d) + 1}$$

where a is independent of $t(x_w^d, y_d)$. Thus integrating Inf_2 in the previous models gives:

$$h(x, y, z, \omega) = \begin{cases} \left(\frac{at(x,y)}{t(x,y)+1} \log(1 + \frac{N}{z}) + \log(1 + \frac{z}{N})\right) \\ \left(\frac{at(x,y)}{t(x,y)+1} \log(\frac{N+1}{z+0.5})\right) \end{cases}$$

As $\frac{\partial^2 h(x,y,z,\omega)}{\partial x^2} = \frac{\partial^2 h(x,y,z,\omega)}{\partial t^2} \left(\frac{\partial t}{\partial x}\right)^2$, for the normalizations considered (equations 2 and 3), and as $\left(\frac{\partial t}{\partial x}\right)^2 > 0$, we have:

$$\mathrm{sgn}\left(\frac{\partial^2 h(x, y, z, \omega)}{\partial x^2}\right) = \mathrm{sgn}\left(\frac{\partial^2 h(x, y, z, \omega)}{\partial t^2}\right)$$

But:

$$\frac{\partial^2 h(x, y, z, \omega)}{\partial t^2} = -\frac{b}{(t(x_w^d, y_d) + 1)^3}$$

with $b > 0$, which shows that the models are now compatible with condition 2. The above development thus explains why the Inf_1 models considered previously need be resized with an Inf_2 model which can take into account burstiness (or, equivalently, the aftereffect of sampling). However, the question remains as whether Inf_1 alone can be used to design an interesting IR model.

4 Information-based IR Models

Several models for IR and textual collections rely on the information brought by a term in a document. In particular, several researchers, Harter [11] being one of the first ones, have observed that the distribution of significant, "specialty" words in a document deviates from the distribution of "functional" words . The more the distribution of a word in a document deviates from its average distribution in the collection, the more likely is this word significant for the document considered. We make use of this notion, underlying DFR models, to define information-based IR models. Indeed, we consider here the family of IR models satisfying the following equation:

$$RSV(q, d) = \sum_{w \in q \cap d} -x_w^q \log Prob(X \geq t_w^d | \theta_w) \tag{4}$$

where θ_w is a set of parameters of the probability distribution considered. This ranking function corresponds to the mean information a document brings to a query (or, equivalently, to the average of the document information brought by each query term)

[3] The same applies to the binomial model, for which $\frac{\partial^2 h(x,y,z,\omega)}{\partial x^2} > 0$. For the sake of clarity, we do not present here this derivation which is purely technical.

and is similar to the Inf_1 part of DFR models. We will refer to models in this family as information-based IR models.

$Prob(X \geq t_w^d|\theta_w)$ is a decreasing function of t_w^d. So, as long as t_w^d is an increasing function of x_w^d and a decreasing function of y_d (which, in practice, is the case for all the normalisation functions used in IR), conditions 1 and 3 are satisfied for this family of models. Furthermore, condition 2 can be re-expressed, in the family of information-based IR models, as:

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

The following theorem (the proof of which is given in the appendix) states that, provided one chooses a bursty distribution, condition 2 is satisfied for information-based IR models , so that IR models defined by equation 4 and based on bursty distributions satisfy conditions 1, 2 and 3 of the previous section.

Theorem 1 *Let P be a probability distribution of class C^2. A necessary condition for P to be bursty is:*

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

Thus, IR models defined by equation 4 and based on bursty distributions satisfy conditions 1, 2 and 3 of the previous section, the concavity property (condition 2) being directly related to the burstiness property of the word frequency distribution used. We now turn to bursty distributions that can be used in such IR models and which will satisfy the last form condition as well as the adjustment conditions (i.e. conditions 4, 5 and 6).

4.1 Log-logistic distribution

Having presented the connection between burstiness and heuristic retrieval constraints for information-based IR models, we now turn to the log-logistic distribution. Following Church and Gale [6] and Airoldi [1], Clinchant and Gaussier [7] studied the negative binomial distribution in the context of text modeling. They showed that this distribution was not appropriate for IR as it relies on two parameters. They then assumed a uniform Beta prior distribution over one of the parameters, leading to a distribution they refer to as the Beta negative binomial distribution, or BNB for short. One problem with the BNB distribution is that it is a discrete distribution and cannot be used for modeling t_w^d. However, there exists a continuous counterpart of the BNB distribution, which is the log-logistic distribution with its β parameter set to 1. The log-logistic distribution is defined by:

$$\begin{cases} X \in [0;\infty) \\ P_{LL}(X < x;\theta,\beta) = \frac{x^\beta}{x^\beta + \theta^\beta} \end{cases}$$

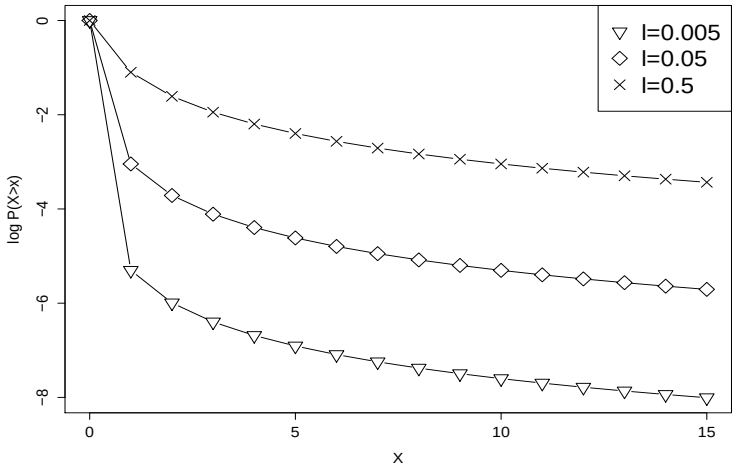

Fig. 2 $\log P(X > x)$ for $\theta \in \{0.5, 0.05, 0.005\}$

Figure 2 shows the density function of the log-logistic distribution for $\beta = 1$ and different values of θ. Setting β to 1, one obtains: $\forall x \in \mathbb{R}^+$

$$\begin{aligned} P_{LL}(x \leq X < x+1; \theta, \beta = 1) &= \frac{x+1}{\theta + x + 1} - \frac{x}{\theta + x} \\ &= \frac{\theta}{(\theta + x + 1)(\theta + x)} \end{aligned} \tag{5}$$

which is exactly the form of the BNB distribution. Furthermore, the following equation shows that the log-logistic is bursty:

$$\forall \epsilon > 0,\ g_\epsilon(x) = P_{LL}(X > x + \epsilon | X > x; \theta, \beta = 1) = \frac{\theta + x}{\theta + x + \epsilon}$$

Finally, using this distribution in the information-based family of IR models leads to the following retrieval function:

$$\begin{aligned} RSV(q,d) &= \sum_{w \in q} -x_w^q \log(P_{LL}(X \geq t_w^d; \theta_w, \beta = 1)) \\ &= \sum_{w \in q \cap d} -x_w^q \log(P_{LL}(X \geq t_w^d; \theta_w, \beta = 1)) \\ &= \sum_{w \in q \cap d} -x_w^q \log(\frac{\theta_w}{t_w^d + \theta_w}) \end{aligned} \tag{6}$$

Following the general idea sustaining the Divergence from Randomness paradigm, θ_w can be defined from the whole collection and can be set to either $\frac{F_w}{N}$ or $\frac{n_w}{N}$. In this way, θ_w represents the probability of observing w in a document, assuming that w is uniformly distributed in the collection. With these settings, it can be shown that the above retrieval function verifies conditions 1, 2, 3 and 4, for all the admissible values

of x, y and z. It can also be shown that it verifies the other conditions associated with IR heuristic constraints.

We are thus now armed with a simplified DFR model, relying solely on Inf_1, which is compatible with the theoretical framework we have developed: our model is based on a continuous distribution that verifies the conditions of retrieval heuristic constraints. We now need to experimentally validate the fact that this model behaves as more complex DFR models on IR collections. We will do that in section 5. Prior to that, we want to show a connection with the language modeling approach to IR.

4.2 Relation to Language Models

Let L be the number of tokens in the collection. Following [19], the scoring formula for a language model using Jelinek-Mercer smoothing can be written as:

$$RSV(q,d) = \sum_{w \in q \cap d} x_w^q \log(1 + s \frac{\frac{x_w^d}{y_d}}{\frac{F_w}{L}}) \tag{7}$$

where λ is the Jelinek-Mercer smoothing parameter and $s = \frac{\lambda}{1-\lambda}$. Using the retrieval formula introduced previously with $\theta_w = \frac{F_w}{N}$ and the length normalization given by equation 2, we have:

$$RSV(q,d) = \sum_{w \in q \cap d} x_w^q \log(1 + c \frac{\frac{x_w^d \times \mathrm{m}}{y_d}}{\frac{F_w}{N}}) \tag{8}$$

Given that $\frac{F_w}{N} = m \times \frac{F_w}{L}$, equation 7 is equivalent to equation 8. The LM model with Jelinek-Mercer smoothing can thus be seen as an information-based model making use of a log-logistic distribution, with a particular length normalization, namely the one given by equation 2, and a particular setting of θ_w.

In the language modeling approach to IR, one starts from term distributions estimated at the document level, and smoothed by the distribution at the collection level. In contrast, the DFR approach uses a distribution the parameters of which are estimated on the whole collection to get a local document weight for each term. Despite the different views sustaining these two approaches, the above development shows that they can be reconciled through appropriate word distributions, in particular the log-logistic one. Lastly, the above connection also indicates that term frequency or length normalizations are related to smoothing. A theory for relating these two elements remains however to be established.

4.3 The LGD model

A choice has to be made for the log-logistic distribution used within the information-based family of IR models, concerning the document length normalization and the value for the parameter θ_w. The previous section provides a possible choice for these elements. We will however not rely on this choice but will rather consider, in the remainder of the paper, the model defined by the following elements:

1. Document length normalization: $t_w^d = x_w^d \log(1 + c\frac{m}{y_d})$

2. t_w^d are distributed according to a log-logistic distribution with $\beta = 1$ and $\theta_w = \frac{N_w}{N}$
3. The retrieval function corresponds to equation 6, which takes the form:

$$RSV(q,d) = \sum_{w \in q \cap d} x_w^q \left[\log(\frac{N_w}{N} + t_w^d) - \log(\frac{N_w}{N}) \right]$$

In other words, we choose the second term frequency normalization of DFR models and the document frequency as the parameter of the word frequency probability distributions. We will refer to this model as *LGD*.

5 Experimental validation

Experiments presented here serve two purposes. The first one is to show the quality of the log-logistic distribution as a word frequency distribution and the second to demonstrate the performance of the IR model. We use the following collections to assess the validity of our model: ROBUST (TREC), CLEF03 AdHoc Task, GIRT (CLEF Domain Specific 2004-2006). Table 2 gives the number of documents (N), number of unique terms (M), average document length and number of test queries for these collections. For the ROBUST collection, we used standard Porter stemming. For the CLEF03, GIRT, we used lemmatization, and an additional decoumpounding step for the GIRT collection which is written in German. In all the following tables, *ROB-t* represents the robust collection with query titles only, *ROB-d* the robust collection with query titles and description fields, *CLEF-t* represent titles for the CLEF collection, *CLEF-d* queries with title and descriptions. The *GIRT* queries are just made up of a single sentence.

5.1 Empirical Fit to Observed Data

We illustrate here the fact that the log-logistic distribution, unlike others like the Poisson distribution, provides a good fit to the data. To do so, we computed the Chi-square statistics for each term under both a Poisson hypothesis and a log-logistic one (figure 3). Our goal here is to see what is the fit between experimental observations and the ones predicted by these distributions: the Chi-square statistics provides us with a measure of this fit.

We restricted our study to terms appearing at least in 100 documents of the 'robust' TREC collection. For each selected term, we want to compare two candidate distributions modeling the term frequencies in the documents, namely the Poisson and Log-Logistic distributions. Furthermore, we assume that the parameters of these distributions are set according to:

- Poisson: $\hat{\theta_w} = \frac{F_w}{N}$
- Log-Logistic: $\hat{\theta_w} = \frac{F_w}{N}$

We then used a standard Chi-Square test. For each selected word w and document d, x_w^d is binned into one of the following intervals: $[0,3)$, $[3,10)$ and $[10,100)$. These intervals corresponds roughly to low, medium and high frequency. The number of observations falling into each interval constitutes statistics that the Chi-Square compares to an expected number predicted by the assumed distribution. For each selected term, we

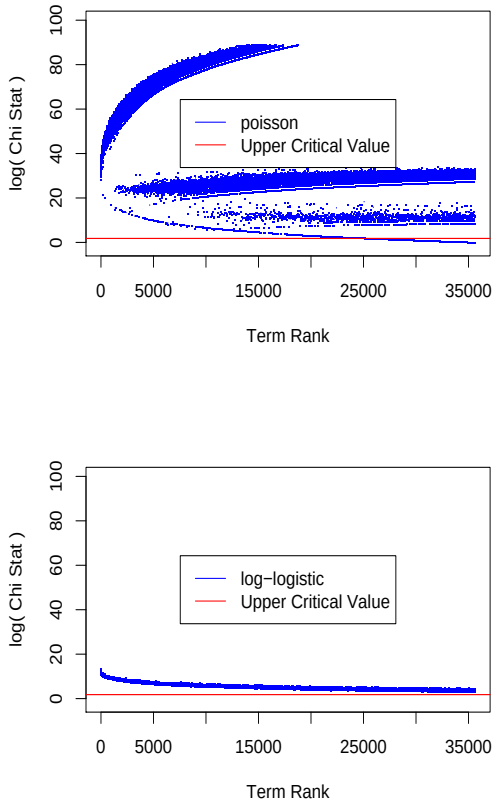

Fig. 3 Distribution of the Chi-square statistics for the Poisson and the BNB/log-logistic distributions on the ROBUST collection

then computed the Chi-square statistics under a Poisson hypothesis and a Log-Logistic hypothesis[4]. To display the results, we first ranked the selected terms by their frequency in the collection in order to get their term rank, as is done in Zipf's Law. We then plotted the term rank against the log of the Chi-Square statistics for both the Poisson and Log-Logistic distributions. Figure 3 shows the log of the Chi-square statistics against the term rank for the 'robust' TREC collection. One dot with coordinate (x, y) on the graph corresponds to a given word in the collection, where x is the term rank and y is the log of the Chi-square statistics for the distribution considered. The horizontal line is the upper critical value for Chi-square test at the 0.05 confidence level.

Concerning the Poisson plot, there are 2 main clouds of points. The upper left area can be explained by words from the interval $[10, 100)$: this is an extremely unlikely event under a Poisson distribution with a very small mean (ex: 0.05). The second area,

[4] Due to relation 5, the Chi-square statistics is the same for the BNB and the log-logistic distributions on the given intervals.

Table 2 Characteristics of the different collections

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

which looks like a thick band, corresponds to words from the first two intervals only. As one can note, the fit provided by the BNB/log-logistic distribution is good inasmuch as the values obtained by the Chi-square statistics are small. These distributions can thus well explain the behavior of words in all the frequency ranges. The same does not hold for the Poisson, for which large values are observed over all the frequency ranges, many words getting a value above the upper critical value.

5.2 Comparison with Language Models

We evaluated the LGD model against the LM model, with both Jelinek-Mercer and Dirichlet Prior smoothing. For each dataset, we randomly split queries in train and test (half of the queries are used for training, the other half for testing). We performed 10 such splits on each collection. The results we provide for the Mean Average Precision (MAP) and the precision at 10 documents are averaged over the 10 splits. The parameters of the different models are optimized (respectively for the MAP and the precision at 10) on the training set. The performance is then measured on the test set. To compare the different methods, a two-sided t-test (at the 0.05 level) is performed to assess the significance of the difference measured between the methods.

For the LGD model, as the parameter c in equation 3 is not bounded, we have to define a set of possible values from which to select the best value on the training set. We make use of the typical range proposed in works on DFR models, which also rely on equation 3 for document length normalization. The set of values we retained is:

$$c \in \{0.25, 0.5, 0.8, 1, 2, 3, 5, 8, 10\}$$

As the smoothing parameter of the Jelinek-Mercer language model is comprised between 0 and 1, we use a regular grid on $[0, 1]$ with a step size of 0.05 in order to select, on the training set, the best value for this parameter. Table 3 shows the comparison of the LGD model (LGD) with the Jelinek-Mercer language model (LM). On all collections, on both short and long queries, the LGD model significantly outperforms the Jelinek-Mercer language model. This is an interesting finding as the complexity of the two models is the same (in a way, they are both conceptually simple). As the results displayed are averaged over 10 different splits, this shows that the LGD model consistently outperforms the Jelinek-Mercer language model and thus yields a more robust approach to IR.

In order to assess the relative behaviors of the log-logistic and Jelinek-Mercer models wrt to their parameter (λ for the Jelinek-Mercer model and c for the log-logistic one), we display in Figure 4 the MAP scores obtained with different values of these parameters, c being set to $c = \frac{\lambda}{1-\lambda}$, which allows one to compare the two models for any λ in $[0, 1]$. As one can note, with the exception of small values of λ, the log-logistic model dominates the Jelinek-Mercer model, which again shows that the log-logistic model is consistently better than the Jelinek-Mercer one.

Table 3 LM-Jelinek-Mercer versus Log-Logistic after 10 splits; bold indicates best performance, * significant difference

MAP	ROB-d	ROB-t	GIRT	CLEF-d	CLEF-t
LM-JM	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-JM	43.8	35.5	67.5	33.0	26.2
LGD	**46.0***	**38.9***	**69.4***	**33.6***	**26.6***

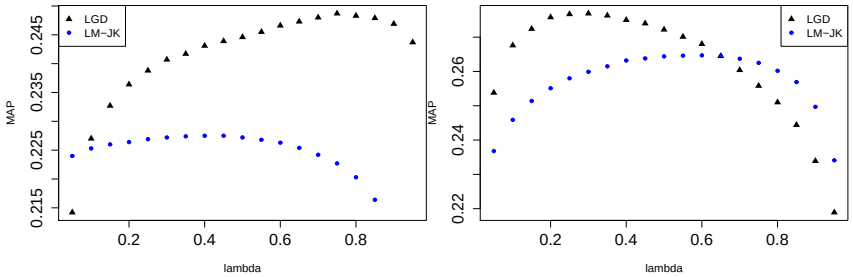

Fig. 4 MAP against lambda. ROB-t are plot on the left side and ROB-d on the right side

For the Dirichlet prior language model, we optimized the smoothing parameter from a set of typical values, defined by:

$$\{10, 50, 100, 200, 500, 800, 1000, 1500, 2000, 5000, 10000\}$$

Table 4 shows the results of the comparison between LGD and the Dirichlet prior language model (LM). These results parallel the ones obtained with the Jelinek-Mercer language model, except for the ROB collection with short queries where the Dirichlet prior language model outperforms the LGD model (the difference being significant for the precision at 10 only). On the other collections, with both short and long queries and on both the MAP and the precision at 10, the LGD model outperforms the Dirichlet prior language model, the difference being significant in most cases. Again, this shows that the LGD model consistently outperforms the Dirichlet prior language model.

Table 4 LM-Dirichlet versus Log-Logistic after 10 splits; bold indicates best performance, * significant difference

MAP	ROB-d	ROB-t	GIRT	CLEF-d	CLEF-t
LM-DIR	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-DIR	45.6	**44.7***	68.6	33.8	28.4
LGD	**46.2***	44.4	**69.0**	**34.5***	**28.6**

5.3 Comparison with BM25

We adopt the same methodology to compare information models with BM25. We choose only to optimize the k_1 parameter of BM25 among the following values: {0.3, 0.5, 0.8, 1.0, 1.2, 1.5, 1.8, 2, 2.2, 2.5}. The others parameters b and k_3 take their default values implemented in Lemur (0.75 and 7). Table 5 shows the comparison of the LGD model with Okapi BM25. The LGD model is either better (3 collections out of 5 for MAP, 2 collections out of 5 for P10) or on par with Okapi BM25. Overall, the LGD model outperforms Okapi BM25.

Table 5 LGD versus BM25 after 10 splits; bold indicates best performance, * significant difference

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**

5.4 Comparison with DFR Models

To compare the LGD model with DFR ones, we chose, in this latter family, the InL2 model, based on the geometric distribution and Laplace law of succession. This model has been used with success in different works ([2,7] for example). As it also relies on equation 3 for document length normalization, we make use here of the same set of possible values for c as the one used for the LGD model, namely:

$$c \in \{0.25, 0.5, 0.8, 1, 2, 3, 5, 8, 10\}$$

It is however interesting to note that InL2 makes use of discrete distributions (geometric and Laplace) over continuous variables (t_w^d) and is thus theoretically flawed. This is not the case of the LGD model which makes use of a continuous distribution, the log-logistic one.

Table 6 provides the results of the comparison between the LGD and the InL2 models. This time, the results are more contrasted than with the language model. In particular, for the precision at 10, both models perform similarly (LGD being significantly better on GIRT whereas InL2 is significantly better on ROB with long queries, the models being on a par in the other cases). For the MAP, the LGD model outperforms the InL2 model as it is significantly better on ROB (for both sort and long queries) and GIRT, and on a par on CLEF. These results are all the more so interesting that the log-logistic model is simpler than InL2: it directly relies on an information measure (see equation 4) without the re-normalization (Inf_2 part) used in DFR models. Lastly, we give in Appendix B a comparison, provided by one reviewer, between the LGD model and Terrier's parameter-free model DFRee[5]. As one can note, the results

[5] http://terrier.org/

obtained with this last model are on par with the models LGD and PL2 (another DFR model). If DFRee is still more complex than LGD, it does not rely on any parameter, which is definitely an advantage. Parameter-free versions of LGD need be determined, maybe along the line used to derive DFRee.

Table 6 INL versus Log-Logistic after 10 splits; bold indicates best performance, * significant difference

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

6 Discussion

The log-logistic model we have introduced is compliant with the heuristic retrieval constraints reviewed in section 2 and is based on a word frequency distribution which can account for burstiness. As we have noted before, this model bears strong similarities with DFR ones. The Divergence from Randomness (DFR) framework proposed by Amati and van Rijsbergen [2] is based on the informative content provided by the occurrences of terms in documents, a quantity which is then corrected by the risk of accepting a term as a descriptor in a document (*first normalization principle*) and by normalizing the raw occurrences by the length of a document (*second normalization principle*). The informative content $Inf_1(t_w^d)$ is based on a first probability distribution and is defined as: $Inf_1(t_w^d) = -\log Prob_1(t_w^d)$. The first normalization principle is associated with a second information defined from a second probability distribution through: $Inf_2(t_w^d) = 1 - Prob_2(t_w^d)$. The overall IR model is then defined as a combination of Inf_1 and Inf_2:

$$\begin{aligned} RSV(q,d) &= \sum_{w \in q \cap d} x_w^q \mathrm{Inf}_2(t_w^d) \mathrm{Inf}_1(t_w^d) \\ &= \sum_{w \in q \cap d} -x_w^q \mathrm{Inf}_2(t_w^d) \log Prob_1(t_w^d) \end{aligned} \tag{9}$$

This latter form shows that DFR models can be seen as information models, as defined by equation 4, with a correction brought by the Inf_2 term, and with the inappropriate use of discrete distributions for modeling continuous variables. With this in mind, we can see the log-logistic model as a simplified DFR model, without the correction through the first normalization principle advocated by Amati and van Rijsbergen (this principle aims at justifying the use of Inf_2). It is thus interesting to see that the LGD model, while being simpler, performs similarly to the InL2 DFR model in our experiments. The use of an appropriate distribution, able to model burstiness, is thus fully justified for this class of models.

Moreover, as we showed in section 4.2, the Jelinek-Mercer model can also be derived from a log-logistic model. However, the Jelinek-Mercer language model and the LGD model differ on the following points:

1. The term frequency normalization;
2. The parameter θ_w;
3. The theoretical framework they fit in.

We want to stress an important point: it is *because* we adopted a new theoretical framework, the information-based family, that we could easily use other term frequency normalizations or settings of θ_w. In fact, a language model with the same term frequency normalization as LGD is clearly not straightforward to obtain in the language modeling approach to IR when using multinomial distributions to model documents. We know of no way so far to do so

As we mentioned previously, other works have tried to model burstiness to come up with more accurate probabilistic models of text collections. We have proposed here a formal definition of burstiness, which allows one to characterize probability distribution wrt this phenomenon, and hence choose appropriate distributions in a more informed manner. We have also shown that burstiness implied the satisfaction of the concavity constraint (condition 2 of section 2) for the family of information models. Indeed, because of its form, heuristic retrieval constraints are naturally captured by models of this family relying on bursty distributions. The LGD model we finally arrive at is thus well founded theoretically. As we have seen, it also outperforms the Jelinek-Mercer and Dirichlet prior language models on most of the collections we have used in our experiments.

7 Conclusion

We have in this paper first introduced an analytical characterization of heuristic retrieval constraints and reviewed several DFR models wrt this characterization. This review showed that the first normalization principle of DFR is necessary to make the model compliant with retrieval constraints. We have then introduced a new model based on the log-logistic distribution to derive a simplified DFR model, and have shown that this simplified model contained, as a special case, the standard language model with Jelinek-Mercer smoothing. This relation is, to our knowledge, the first connection between the DFR and language modeling approaches to IR.

We have then reviewed empirical findings on word frequency distributions and the central role played by burstiness in this context. This has led us to propose a formal definition of burstiness which can be used to characterize probability distributions wrt this phenomenon. We have then introduced the family of information-based IR models which naturally captures heuristic retrieval constraints when the underlying probability distribution is bursty. In particular, theorem 1 guarantees that the concavity constraint is satisfied for bursty distributions, whereas the form of the family guarantees the other constraints when the length normalization function is increasing in x_w^d and decreasing in y_d, which is the case for all the normalization functions we know of. We have then proposed a new IR model within this family, based on the log-logistic distribution.

The experiments we have conducted on three different collections illustrate the good behavior of the LGD model: this model significantly outperforms the Jelinek-Mercer and Dirichlet prior language models on most collections, with both short and

long queries and for both the MAP and the precision at 10 documents. The LGD also yields results similar to DFR ones, while being simple. Future work will investigate an extension of information models for pseudo-relevance feedback and the use of other bursty distributions in the framework we have developed.

Acknowledgements This research was partly supported by the Pascal-2 Network of Excellence ICT-216886-NOE and the French project Fragrances ANR-08-CORD-008. We thank the anonymous reviewers for their comments on the first version of this paper.

References

1. E. M. Airoldi, W. W. Cohen, and S. E. Fienberg. Bayesian methods for frequent terms in text: Models of contagion and the δ^2 statistic.
2. G. Amati and C. J. V. Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst.*, 20(4):357–389, 2002.
3. A. L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, October 1999.
4. D. Chakrabarti and C. Faloutsos. Graph mining: Laws, generators, and algorithms. *ACM Comput. Surv.*, 38(1):2, 2006.
5. K. W. Church. Empirical estimates of adaptation: the chance of two noriegas is closer to p/2 than p2. In *Proceedings of the 18th conference on Computational linguistics*, pages 180–186, Morristown, NJ, USA, 2000. Association for Computational Linguistics.
6. K. W. Church and W. A. Gale. Poisson mixtures. *Natural Language Engineering*, 1:163–190, 1995.
7. S. Clinchant and É. Gaussier. The bnb distribution for text modeling. In Macdonald et al. [12], pages 150–161.
8. C. Elkan. Clustering documents with an exponential-family approximation of the dirichlet compound multinomial distribution. In W. W. Cohen and A. Moore, editors, *ICML*, volume 148 of *ACM International Conference Proceeding Series*, pages 289–296. ACM, 2006.
9. H. Fang, T. Tao, and C. Zhai. 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*, 2004.
10. W. Feller. *An Introduction to Probability Theory and Its Applications, Vol. I.* Wiley, New York, 1968.
11. S. Harter. A probabilistic approach to automatic keyword indexing, part 1: On the distribution of speciality words in a technical literature, part 2: An algorithm for probabilistic indexing. *Journal of the American Society for Information Science*, (26):197–206, 1975.
12. C. Macdonald, I. Ounis, V. Plachouras, I. Ruthven, and R. W. White, editors. *Advances in Information Retrieval , 30th European Conference on IR Research, ECIR 2008, Glasgow, UK, March 30-April 3, 2008. Proceedings*, volume 4956 of *Lecture Notes in Computer Science*. Springer, 2008.
13. R. E. Madsen, D. Kauchak, and C. Elkan. Modeling word burstiness using the dirichlet distribution. In L. D. Raedt and S. Wrobel, editors, *ICML*, volume 119 of *ACM International Conference Proceeding Series*, pages 545–552. ACM, 2005.
14. S.-H. Na, I.-S. Kang, and J.-H. Lee. Improving term frequency normalization for multi-topical documents and application to language modeling approaches. In Macdonald et al. [12], pages 382–393.
15. S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 232–241, New York, NY, USA, 1994. Springer-Verlag New York, Inc.
16. G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA, 1983.
17. A. Singhal, C. Buckley, and M. Mitra. Pivoted document length normalization. In *SIGIR '96: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 21–29, New York, NY, USA, 1996. ACM.

18. Z. Xu and R. Akella. 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*, pages 427–434, New York, NY, USA, 2008. ACM.
19. C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214, 2004.

A Proof of Theorem 1

We recall here theorem 1:
Let P be a probability distribution of class C^2. A necessary condition for P to be bursty is:

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

Proof: Let P be a continuous probability distribution of class C^2. $\forall y > 0$, the function g_y defined by:

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

is increasing in x (by definition of a bursty distribution).

Let F be the cumulative function of P. Then: $g_y(x) = \frac{F(x+y)-1}{F(x)-1}$. For y sufficiently small, using a Taylor expansion of $F(x + y)$, we have:

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

where F' denotes $\frac{\partial F}{\partial x}$. Then, taking the derivative of g wrt x and considering only the sign of g', we get:

$$sgn[g'] = sgn[F''F - F'' - F'^2] = sgn[(\frac{F'}{F-1})']$$
$$= sgn[(\log(1-F))''] = sgn[(\log P(X \geq x))]$$

As g_y is increasing in x, so is g, and thus $\frac{\partial^2 \log(P(X \geq x))}{\partial x^2} > 0$, which establishes the property.

B Comparison with DFRee and PL2

We display here results provided by one reviewer, whom we gratefully thank, on a comparison between the LGD model and the Terrier's parameter-free DFRee and PL2 models. As one can note, all these models perform similarly.

Model	parameter c	MAP
DFRee		0.2030
LGD	1	0.1964
LGD	2	0.2001
LGD	3	0.2017
LGD	5	0.2025
LGD	6	0.2030
LGD	8	0.2031
LGD	10	0.2001

Model	parameter c	MAP
DFRee		0.2030
PL2	1	0.1767
PL2	2	0.1926
PL2	3	0.2017
PL2	5	0.2061
PL2	6	0.2080
PL2	8	0.2090
PL2	10	0.2095