

# Information-Based Models for Ad Hoc IR

Stéphane Clinchant  
XRCE & LIG, Univ. Grenoble I  
Grenoble, France  
stephane.clinchant@xrce.xerox.com

Eric Gaussier  
LIG, Univ. Grenoble I  
Grenoble, France  
eric.gaussier@imag.fr

## ABSTRACT

We introduce in this paper the family of information-based models for *ad hoc* information retrieval. These models draw their inspiration from a long-standing hypothesis in IR, namely the fact that the difference in the behaviors of a word at the document and collection levels brings information on the significance of the word for the document. This hypothesis has been exploited in the 2-Poisson mixture models, in the notion of eliteness in BM25, and more recently in DFR models. We show here that, combined with notions related to burstiness, it can lead to simpler and better models.

## Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

## General Terms

Theory, Algorithms, Experimentation

## Keywords

IR Theory, Probabilistic Models, Burstiness

## 1. INTRODUCTION

The purpose of this paper is to introduce the family of information based model for *ad hoc* information retrieval (IR). By information, we refer to Shannon information when observing a statistical event. The informativeness of a word in a document has a rich tradition in information retrieval since the influential indexing methods developed by Harter ([11]). The idea that the respective behaviors of words in documents and in the collection bring information on word type is, *de facto*, not a novel idea in IR. It has inspired the 2-Poisson mixture model, the concept of eliteness in BM25 models and is at the heart of DFR models. In this paper, we come back to this idea in order to present a new family of IR models: *information models*. To do so, we first present,

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.

Copyright 2010 ACM 978-1-60558-896-4/10/07 ...\$10.00.

in section 2, the conditions a retrieval function should satisfy, on the basis of the heuristic retrieval constraints proposed by Fang *et al.* [9]. Section 3 is then devoted to the presentation of information models, and their link with the retrieval conditions and the phenomenon known as *burstiness*. We present two instances of information models, based on two power law distributions, and show how to perform pseudo-relevance feedback for information models. Section 4 provides an experimental validation of our models. Our experiments show that the information models we introduce significantly outperform language models and Okapi BM25. They are on par with DFR models, while being conceptually simpler, when pseudo-relevance feedback is not used. When using pseudo-relevance feedback, they significantly outperform all models, including DFR ones.

## 2. PRELIMINARIES

The notations we use throughout the paper are summarized in table 2 ( $w$  represents a term). They slightly differ from standard notations for convenience reasons, i.e. their easiness of use in the mathematical framework we deploy. We

Notation	Description
$x_w^q$	Number of occurrences of $w$ in query $q$
$x_w^d$	Number of occurrences of $w$ in document $d$
$t_w^d$	Normalized version of $x_w^d$
$y_d$	Length of document $d$
$m$	Average document length
$L$	Length of collection $d$
$N$	Number of documents in the collection
$M$	Number of terms in the collection
$F_w$	Number of occurrences of $w$ in collection: $F_w = \sum_d x_w^d$
$N_w$	Number of documents containing $w$ : $N_w = \sum_d I(x_w^d > 0)$
$z_w$	$z_w = F_w$ or $z_w = N_w$

Table 1: Notations

consider here retrieval functions, denoted  $RSV$ , of the form:

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

where  $\theta$  is a set of parameters and where  $h$ , the form of which depends on the IR model considered, is assumed to be of class<sup>1</sup>  $C^2$  and defined over  $\mathbb{R}^{++} \times \mathbb{R}^{++} \times \mathbb{R}^{++} \times \Theta$ , where

<sup>1</sup>A function of class  $C^2$  is a function for which second derivatives exist and are continuous.

$\Theta$  represents the domain of the parameters in  $\theta$  and  $a$  is often the identity function. Language models [21], Okapi [15] and Divergence from Randomness [3] models as well as vector space models [16] all fit within the above form. For example, for the pivoted normalization retrieval formula [17],  $\theta = (s, m, N)$  and:

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

where  $I$  is an indicator function which equals 1 when its 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. These different considerations can be analytically formalized as a set of simple conditions the retrieval function  $h$  should satisfy:

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

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

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

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

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, already mentioned in this form 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. We will refer to the above conditions as the **form conditions** inasmuch as they define the general shape the function  $h$  should have. They respectively correspond to the heuristic retrieval constraints TFC1, TFC2, LNC1 and TDC<sup>2</sup> introduced by Fang *et al.* [9]. In addition to this form conditions, Fang *et al.* [9] used two additional constraints to regulate the interaction between frequency and document length, i.e. between the derivatives wrt to  $x$  and  $y$ . These conditions, which we will refer to as **adjustment conditions**, allow to adjust the functions  $h$  satisfying the form conditions 1, 2, 3 and 4. They correspond to:

<sup>2</sup>Condition 4 is in fact a special case of TDC, but this is beyond the scope of the current paper.

**Condition 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  $RSV(d1, q) \geq RSV(d2, q)$

**Condition 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  $RSV(d1, q) > RSV(d2, q)$ .

We are now ready to proceed to the presentation of information models.

### 3. INFORMATION MODELS

In order to take into account the fact that one is comparing documents of different length, most IR models do not rely directly on the raw number of occurrences of words in documents, but rather on normalized versions of it. 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):

$$t_w^d = x_w^d c \frac{m}{y_d} \text{ or } x_w^d \log\left(1 + c \frac{m}{y_d}\right) \quad (1)$$

The concept of the information brought by a term in a document has been considered in several IR models. Harter [11] observed that 'significant', 'specialty' words of a document do not behave as 'functional' words. Indeed, the more a word deviates in a document from its average behavior in the collection, the more likely it is 'significant' for this particular document. This can be easily captured in terms of information: If a word behaves in the document as expected on the collection, then it has a high probability of occurrence in the document  $p$ , according to the distribution collection, and the information it brings to the document,  $-\log(p)$ , is small. On the contrary, if it has a low probability of occurrence in the document, according to the distribution collection, then the amount of information it conveys is more important. Because of the above consideration, this idea, at the basis of DFR models, has to be applied to the normalized form of the term frequency. This leads to the general and simple retrieval function:

$$RSV(q, d) = \sum_{w \in q} -x_w^q \log \text{Prob}(X_w \geq t_w^d | \lambda_w) \quad (2)$$

where  $t_w^d$  is the normalized form of  $x_w^d$  and  $\lambda_w$  is a parameter for the probability distribution of  $w$  in the collection. We simply consider here that  $\lambda_w$  is set to either the average number of occurrences of  $w$  in the collection, or to the average number of documents in which  $w$  occurs, that is:

$$\lambda = \frac{z_w}{N} = \frac{F_w}{N} \text{ or } \frac{N_w}{N} \quad (3)$$

It is interesting to note that the retrieval function defined by equation 2, which is rank invariant by the change of the logarithmic base, satisfies the heuristic retrieval conditions 1 and 3. Indeed,  $\text{Prob}(X_w \geq t_w^d | \lambda_w)$  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 is the case for all the normalization functions we are aware of, conditions 1 and 3 are satisfied for this family of models.

### 3.1 Burstiness (and condition 2)

Church and Gale [6] were the first to study, to our knowledge, the phenomenon of burstiness in texts. The term “burstiness” describes the behavior of words which tend to appear in bursts, i.e., once they appear in a document, they are much 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, 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* ([4] and [5]) 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 [19] studied the use of a DCM model within the Probability Ranking Principle for modeling the dependency of word repetitive occurrences (a notion directly related to burstiness), and argue that multinomial distributions alone are not appropriate for IR within this principle. More formally, Clinchant and Gaussier [7] introduced the following definition (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  $g_\epsilon$  defined by:

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

is a strictly increasing function of  $x$ . A distribution which verifies this condition is said to be bursty.

which translates the fact that, with a bursty distribution, it is easier to generate higher values of  $X$  once lower values have been observed. We now show that this notion is directly related to the heuristic retrieval condition 2.

In the retrieval function defined by equation 2, the function  $h$  we have considered so far corresponds to:

$$-\log(\text{Prob}(X \geq t_w^d))$$

In this case, condition 2 can be re-expressed as:

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

But:  $\frac{\partial^2 f}{\partial t^2} = \frac{\partial^2 f}{\partial x^2} \left(\frac{\partial x}{\partial t}\right)^2 + \frac{\partial f}{\partial x} \frac{\partial^2 x}{\partial t^2}$ . Furthermore,  $\frac{\partial f}{\partial x}$  is here negative as  $f$  is  $\log(\text{Prob}(X \geq t_w^d))$ . So, as long as  $\frac{\partial^2 x}{\partial t^2} \geq 0$  (which is the case for all the normalization functions we are aware of, in particular the ones provided by equation 1), a sufficient condition for condition 2 is:

$$\frac{\partial^2 \log(\text{Prob}(X \geq t_w^d))}{\partial (t_w^d)^2} > 0$$

The following theorem (the proof of which is given in the appendix) shows that bursty distributions satisfy this condition.

THEOREM 3. Let  $P$  be a “bursty” probability distribution of class  $C^2$ . Then:

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

We thus see that under certain assumptions, IR models defined by equation 2 satisfy the form conditions 1, 2 and 3. We now summarize these assumptions which characterize information models.

### 3.2 Characterization of Information Models

We characterize information models by the following three elements:

1. **Normalization function** The normalization function  $t_w^d$ , function of  $x_w^d$  and  $y_d$  (respectively the number of occurrences of the word in the document and the length of the document), satisfies:

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

2. **Probability distribution** The probability distribution at the basis of the model has to be:

- Continuous, the random variable under consideration,  $t_w^d$ , being continuous;
- Compatible with the domain of  $t_w^d$ , i.e. if  $t_{\min}$  is the minimum value of  $t_w^d$ , then  $\text{Prob}(X_w \geq t_{\min} | \lambda_w) = 1$  (because of the first inequality above,  $t_{\min}$  is obtained when  $x_w^d = 0$ );
- Bursty according to definition 2 above.

3. **Retrieval function** The retrieval function satisfies equation 2, i.e.:

$$\begin{aligned} RSV(q, d) &= \sum_{w \in q} -x_w^q \log \text{Prob}(X_w \geq t_w^d | \lambda_w) \\ &= \sum_{w \in q \cap d} -x_w^q \log \text{Prob}(X_w \geq t_w^d | \lambda_w) \end{aligned}$$

where the second equality derives from the fact that the probability function verifies  $\text{Prob}(X_w \geq t_{\min} | \lambda_w) = 1$ , with  $t_{\min}$  obtained when  $x_w^d = 0$ . The above 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). Furthermore, the parameter  $\lambda_w$  is set as in equation 3:

$$\lambda = \frac{z_w}{N} = \frac{F_w}{N} \text{ or } \frac{N_w}{N}$$

The general form of the retrieval function and the first two inequalities on the normalization function ensure that the model satisfies conditions 1 and 3. Theorem 3, in conjunction with the last condition on the normalization function, additionally ensures that it satisfies condition 2. Hence, information models satisfy three (out of four) form conditions. The choice of the particular bursty distribution to be used has to be made in such a way that the last form condition and the two adjustment conditions are satisfied.

### 3.3 Two Power-law Instances

We present here two power law distributions which are bursty and lead to information models satisfying all form and adjustment conditions. The use of power law distributions to model burstiness is not entirely novel, as other studies ([4, 5]) have used similar distributions to model preferential attachment, a notion equivalent to burstiness.

#### Log-Logistic Distribution

The log-logistic (LL) distribution is defined by, for  $X \geq 0$ :

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

We consider here a restricted form of the log-logistic distribution where  $\beta = 1$ , so that the the log-logistic information model takes the form:

$$\begin{aligned} RSV(q, d) &= \sum_{w \in q \cap d} -x_w^q \log(P_{LL}(X \geq t_w^d | \lambda_w)) \\ &= \sum_{w \in q \cap d} -x_w^q \log\left(\frac{\lambda_w}{t_w^d + \lambda_w}\right) \end{aligned} \quad (4)$$

The log-logistic motivation resorts to previous work on text modeling. Following Church and Gale [6] and Airoidi [1], Clinchant and Gaussier [7] studied the negative binomial distribution in the context of text modeling. 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, the log-logistic distribution, with its  $\beta$  parameter set to 1, is a continuous counterpart of the BNB distribution since  $P_{LL}(x \leq X < x + 1; r) = P_{BNB}(x)$ .

#### A Smoothed Power-Law (SPL) Distribution

We consider here the distribution, which we will refer to as *SPL*, defined, for  $x > 0$ , by:

$$\begin{aligned} f(x; \lambda) &= \frac{-\log \lambda}{1 - \lambda} \frac{\lambda^{\frac{x}{x+1}}}{(x+1)^2} \quad (0 < \lambda < 1) \\ P(X > x | \lambda) &= \int_x^\infty f(x; \lambda) = \frac{\lambda^{\frac{x}{x+1}} - \lambda}{1 - \lambda} \end{aligned}$$

where  $f$  denotes the probability density function. Based on this distribution, the SPL information model thus takes the form:

$$RSV(q, d) = \sum_{w \in q \cap d} -x_w^q \log\left(\frac{\lambda_w^{\frac{t_w^d}{t_w^d+1}} - \lambda_w}{1 - \lambda_w}\right) \quad (5)$$

From equations 4 or 5, and using the normalization functions defined by equation 1, one can verify (a) that the log-logistic and SPL distributions are bursty, and (b) that their corresponding information models additionally satisfy conditions 4, 5 and 6 (the demonstration is purely technical, and is skipped here). The log-logistic and SPL information models thus satisfy all the form and adjustment conditions.

Figure 1 illustrates the behavior of the log-logistic model, the SPL model and the InL2 DFR model (referred to as *INL* for short). To compare these models, we used a value of 0.005 for  $\lambda$  and computed the term weight obtained for term frequencies varying from 0 to 15. For information models, the weight corresponds to the quantity  $-\log Prob$ , whereas

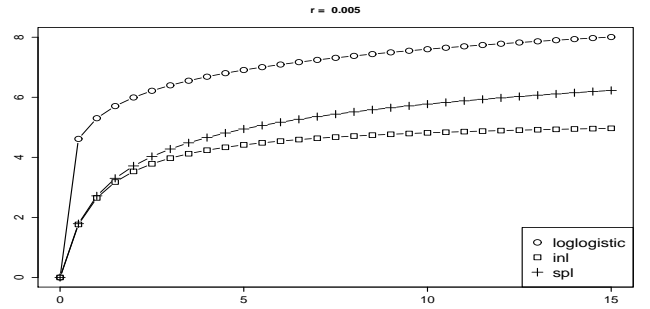


Figure 1: Plot of Retrieval Functions

in the case of DFR models, this quantity is corrected by the  $\text{Inf}_2$  part, leading to, with the underlying distributions retained:

$$\text{weight} = \begin{cases} -\log\left(\frac{\lambda_w}{t_w^d + \lambda_w}\right) & (\text{log-logistic}) \\ -\log\left(\frac{\lambda_w^{\frac{t_w^d}{t_w^d+1}} - \lambda_w}{1 - \lambda_w}\right) & (\text{SPL}) \\ -\frac{t_w^d}{t_w^d+1} \log\left(\frac{N_w+0.5}{N+1}\right) & (\text{InL2}) \end{cases}$$

As one can note, the weight values obtained with the two information models are always above the ones obtained with the DFR model, the log-logistic model having a sharper increase than the other ones for low frequency terms.

### 3.4 PRF in Information Models

Pseudo-relevance feedback (PRF) in information models can be performed following the same approach as the one used in other models: The weight of a term in the original query is updated on the basis of the information brought by the top retrieved documents on the term. Denoting by  $R$  the set of top  $n$  documents retrieved for a given query,  $R = (d_1, \dots, d_n)$ , the average information this set brings on a given term  $w$  can directly be computed as:

$$\text{Info}_R(w) = \frac{1}{n} \sum_{d \in R} -\log(P(X_w > t_w^d | \lambda_w)) \quad (6)$$

where the mean is taken over all the documents in  $R$ . This is a major difference with the approach in [2] where all documents in  $R$  are merged into a single document. Considering the documents in  $R$  as different documents allows one to take into account the differences in document lengths and number of occurrences. The original query is then modified, following standard approaches to PRF, to take into account the words appearing in  $R$  as:

$$x_w^{q2} = \frac{x_w^q}{\max_w x_w^q} + \beta \frac{\text{Info}_R(w)}{\max_w \text{Info}_R(w)} \quad (7)$$

where  $\beta$  is a parameter controlling the modification brought by  $R$  to the original query.  $x_w^{q2}$  denotes the updated weight of  $w$  in the query.

## 4. EXPERIMENTAL VALIDATION

To assess the validity of our models, we used standard IR collections, from two evaluation campaigns: TREC (trec.nist.gov) and CLEF (www.clef-campaign.org). Table 2 gives the number of documents ( $N$ ), number of unique terms ( $M$ ), aver-

age document length and number of test DL queries for the collections we retained: ROBUST (TREC), TREC3, CLEF03 AdHoc Task, GIRT (CLEF Domain Specific Task, from the years 2004 to 2006). For the ROBUST and TREC3 collections, we used standard Porter stemming. For the CLEF03 and GIRT collections, we used lemmatization, and an additional decompounding step for the GIRT collection which is written in German.

**Table 2: Characteristics of the different collections**

	N	M	Avg DL	# Queries
ROBUST	490 779	992 462	289	250
TREC-3	741 856	668 648	438	50
CLEF03	166 754	80 000	247	60
GIRT	151 319	179 283	109	75

We evaluated the log-logistic and the SPL model against language models, with both Jelinek-Mercer and Dirichlet Prior smoothing, as well as against the standard DFR models and Okapi BM25. 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 (P10) is the average of the values obtained 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. All our experiments were carried out thanks to the Lemur Toolkit ([www.lemurproject.org](http://www.lemurproject.org)). 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, *CL-t* represent titles for the CLEF collection, *CL-d* queries with title and descriptions and T3-t query titles for TREC-3 collection. The GIRT queries are just made up of a single sentence.

The version of the log-logistic model used in all our experiments is based on  $\lambda_w = \frac{n_w}{N}$  and the second length normalization in equation 1 (called L2 in DFR). We refer to this model as the LGD model. The same settings are chosen for the SPL model. As the parameter  $c$  in equation 1 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 1 for document length normalization. The set of values we retained is: {0.5, 0.75, 1, 2, 3, 4, 5, 6, 7, 8, 9}.

#### Comparison with Jelinek-Mercer and Dirichlet language models

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 our models, LGD and SPL, 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). Further-

more, 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. Lastly, the SPL model is better than the Jelinek-Mercer model for most collections for MAP and P10.

**Table 3: LGD and SPL versus LM-Jelinek-Mercer after 10 splits; bold indicates significant difference**

MAP	ROB-d	ROB-t	GIR	T3-t	CL-d	CL-t
JM	26.0	20.7	40.7	22.5	49.2	36.5
LGD	<b>27.2</b>	<b>22.5</b>	<b>43.1</b>	<b>25.9</b>	<b>50.0</b>	<b>37.5</b>
P10	ROB-d	ROB-t	GIR	T3-t	CL-d	CL-t
JM	43.8	35.5	67.5	40.7	33.0	26.2
LGD	<b>46.0</b>	<b>38.9</b>	<b>69.4</b>	<b>52.4</b>	<b>33.6</b>	<b>26.6</b>

MAP	ROB-d	ROB-t	GIR	T3-t	CL-d	CL-t
JM	26.6	23.1	39.2	22.3	<b>47.2</b>	37.2
SPL	26.7	<b>25.2</b>	<b>41.7</b>	<b>26.6</b>	44.1	37.7
P10	ROB-d	ROB-t	GIR	T3-t	CL-d	CL-t
JM	44.4	39.8	66.0	43.9	34.0	25.6
SPL	<b>47.6</b>	<b>45.3</b>	<b>69.8</b>	<b>56.0</b>	34.0	25.6

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 our models and the Dirichlet prior language model (DIR). These results parallel the ones obtained with the Jelinek-Mercer language model on most collections, even though the difference is less marked. For the ROB collection with short queries, the Dirichlet prior language model outperforms in average the log-logistic 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 log-logistic model outperforms in average the Dirichlet prior language model, the difference being significant in most cases. The Dirichlet model has a slight advantage in MAP over the SPL model, but SPL is better for precision. Overall, the information-based models outperform in average language models.

**Table 4: LGD and SPL versus LM-Dirichlet after 10 splits; bold indicates significant difference**

MAP	ROB-d	ROB-t	GIR	T3-t	CL-t	CL-d
DIR	27.1	25.1	41.1	<b>25.6</b>	36.2	48.5
LGD	<b>27.4</b>	25.0	<b>42.1</b>	24.8	<b>36.8</b>	<b>49.7</b>
P10	ROB-d	ROB-t	GIR	T3-t	CL-t	CLF-d
DIR	45.6	43.3	68.6	54.0	28.4	33.8
LGD	<b>46.2</b>	43.5	<b>69.0</b>	<b>54.3</b>	<b>28.6</b>	<b>34.5</b>

MAP	ROB-d	ROB-t	GIR	T3-t	CL-t	CL-d
DIR	<b>26.7</b>	25.0	40.9	<b>27.1</b>	36.2	<b>50.2</b>
SPL	25.6	24.9	<b>42.1</b>	26.8	36.4	46.9
P10	ROB-d	ROB-t	GIR	T3-t	CL-t	CL-d
DIR	45.2	43.8	68.2	52.8	27.3	32.8
SPL	<b>46.6</b>	<b>44.7</b>	<b>70.8</b>	<b>55.3</b>	27.1	32.9

#### Comparison with BM25

We adopt the same methodology to compare information models with BM25. We choose only to optimize the  $k_1$  pa-

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 log-logistic and SPL models with Okapi BM25. The log-logistic is either better (4 collections out of 6 for mean average precision, 3 collections out of 6 for P10) or on par with Okapi BM25. The same thing holds for the SPL model, which is 3 times better and 3 times on par for the MAP, and 4 times better, 1 time worse and 1 time on a par for the precision at 10 documents. Overall, information models outperform in average Okapi BM25.

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

MAP	ROB-d	ROB-t	GIR	T3-t	CL-t	CL-d
BM25	26.8	22.4	39.8	25.4	34.9	46.8
LGD	<b>28.2</b>	<b>23.5</b>	<b>41.4</b>	<b>26.1</b>	34.8	48.0
P10	ROB-d	ROB-t	GIR	T3-t	CL-t	CL-d
BM25	45.9	42.6	62.6	50.6	28.5	33.7
LGD	46.5	<b>44.3</b>	<b>66.6</b>	<b>53.8</b>	28.7	34.4

MAP	ROB-d	ROB-t	GIR	T3-t	CL-t	CL-d
BM25	26.9	24.2	38.5	25.3	35.1	47.3
SPL	27.1	<b>25.4</b>	<b>40.5</b>	<b>26.8</b>	34.5	47.0
P10	ROB-d	ROB-t	GIR	T3-t	CL-t	CL-d
BM25	45.7	41.4	62.8	51.0	<b>28.5</b>	36.1
SPL	<b>47.6</b>	<b>44.1</b>	<b>67.9</b>	<b>57.0</b>	28.0	35.4

### Comparison with DFR models

To compare our model with DFR ones, we chose, in this latter family, the InL2 model, based on the Geometric distribution and Laplace law of succession, and the PL2 model based on the Poisson distribution and Laplace law. These models have been used with success in different works ([3, 7, 18] for example). All the models considered here make use of the same set of possible values for  $c$ , namely: {0.5, 0.75, 1, 2, 3, 4, 5, 6, 7, 8, 9}. It is however interesting to note that both PL2 and InL2 make use of discrete distributions (Geometric and Poisson) over continuous variables ( $t_w^d$ ) and are thus theoretically flawed. This is not the case of the information models which rely on a continuous distribution.

The results obtained, presented in tables 6 and 7 are more contrasted than the ones obtained with language models and Okapi BM25. In particular, for the precision at 10, LGD and InL2 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. SPL is better than InL2 for precision but on a par for MAP. Moreover, LGD and PL2 are on a par for MAP, while PL2 is better for P10. Lastly, PL2 is better than SPL for MAP but not for the precision at 10 documents. Overall, DFR models and information models yield similar results. This is all the more so interesting that information models are simpler than DFR ones: They rely on a single information measure (see equation 2) without the re-normalization ( $Inf_2$  part) used in DFR models.

**Table 6: LGD and SPL versus INL after 10 splits; bold indicates significant difference**

MAP	ROB-d	ROB-t	GIR	T3-t	CL-t	CL-d
INL2	27.7	24.8	42.5	27.3	37.5	47.7
LGD	<b>28.5</b>	<b>25.0</b>	<b>43.1</b>	27.3	37.4	48.0
P10	ROB-d	ROB-t	GIR	T3-t	CL-t	CL-d
INL2	<b>47.7</b>	43.3	67.0	52.4	27.3	33.4
LGD	47.0	43.5	<b>69.4</b>	53.2	27.2	33.3

MAP	ROB-d	ROB-t	GIR	T3-t	CL-t	CL-d
INL	<b>26.9</b>	24.3	40.4	24.8	<b>35.5</b>	49.4
SPL	26.6	<b>24.6</b>	40.7	<b>25.4</b>	34.6	48.1
P10	ROB-d	ROB-t	GIR	T3-t	CL-t	CL-d
INL	47.6	42.8	63.4	52.5	28.8	33.8
SPL	47.8	<b>44.1</b>	<b>68.0</b>	<b>53.9</b>	28.7	33.6

**Table 7: LGD and SPL versus PL2 after 10 splits; bold indicates significant difference**

MAP	ROB-d	ROB-t	GIR	T3-t	CL-t	CL-d
PL2	26.2	24.8	40.6	<b>24.9</b>	36.0	47.2
LGD	<b>27.3</b>	24.7	40.5	24.0	36.2	47.5
P10	ROB-d	ROB-t	GIR	T3-t	CL-t	CL-d
PL2	46.4	<b>44.1</b>	<b>68.2</b>	55.0	28.7	33.1
LGD	46.6	43.2	66.7	53.9	28.5	33.7

MAP	ROB-d	ROB-t	GIR	T3-t	CL-t	CL-d
PL2	26.3	25.2	42.8	<b>25.8</b>	37.3	<b>45.7</b>
SPL	26.3	25.2	42.7	25.3	37.4	44.1
P10	ROB-d	ROB-t	GIR	T3-t	CL-t	CL-d
PL2	46.0	45.2	69.3	54.8	26.2	32.7
SPL	<b>47.0</b>	45.2	69.8	55.4	25.9	32.9

### Pseudo-relevance feedback

There are many parameters for pseudo-relevance feedback algorithms: The number of document to consider ( $N$ ), the number of terms to add the query ( $TC$ ) and the weight to give to those new query terms (parameter  $\beta$  in equation 7). Optimizing all these parameters and smoothing ones at the same time would be very costly. We thus modify here our methodology. For each collection, we choose the optimal smoothing parameters for each model ( $c, \mu, k_1$ ) on all queries. The results obtained in this case are given in table 8, where LM+MIX corresponds here to the Dirichlet language model. They show, for example, that on the ROBUST collection there is no difference between the baseline systems we will use for pseudo-relevance feedback in terms of MAP. Overall, the precision at 10 is very similar for the different systems, so that there is no bias, with the setting chosen, towards a particular system. We compare here the results obtained with the information models to two state-of-the-art pseudo-relevance feedback models: Bo2, associated with DFR models ([2]), and the mixture model associated with language models ([20]). For each collection, we average the results obtained over 10 random splits, the variation of  $N$  and  $TC$  being made on each split so as to be able to compare the results of the different settings. For each setting, we optimize the weight to give to new terms:  $\beta$  (within {0.1, 0.25, 0.5, 0.75, 1, 1.5, 2}) in information and Bo2 models,  $\alpha$  (within {0.1, 0.2, ..., 0.9}) in the mixture-model for

feedback in language models. In this latter case, we set the feedback mixture noise to its default value (0.5). As before, we used Lemur to carry our experiments and optimize here only the mean average precision. Table 9 displays the results for the different models (as before, a two-sided t-test at the 0.05 level is used to assess whether the difference is statistically significant, which is indicated by a \*). As one can note, the information models significantly outperform the pseudo-relevance feedback versions of both language models and DFR models. The SPL model is the best one for  $N = 5$  and  $TC = 5$ , while the LGD model yields the best performance in most other cases. Although DFR and information models perform similarly when no feedback is used, their pseudo-relevance feedback versions do present differences, information models outperforming significantly both language and DFR models in this latter case.

**Table 8: Performances of baseline setting for PRF ( $N = 0, TC = 0$ ): bold indicates significant difference**

MAP	ROB-t	GIRT	T3-t	CLEF-t
LM+MIX	25.4	41.1	<b>28.3</b>	37.0
LGD	25.4	<b>42.4</b>	27.1	<b>37.5</b>
P10	ROB-t	GIRT	T3-t	CLEF-t
LM+MIX	44.6	68.3	<b>56.3</b>	<b>27.5</b>
LGD	44.1	68.7	55.3	27.2

**Table 9: Mean average precision of PRF experiments; bold indicates best performance, \* significant difference over LM and Bo2 models**

Model	N	TC	ROB-t	GIR	T3-t	CL-t
LM+MIX	5	5	27.5	44.4	30.7	36.6
LGD	5	5	<b>28.3*</b>	44.3	<b>32.9*</b>	37.6
INL+Bo2	5	5	26.5	42.0	30.6	37.6
SPL	5	5	<b>28.9*</b>	<b>45.6*</b>	<b>32.9*</b>	<b>39.0*</b>
LM+MIX	5	10	28.3	45.7*	33.6	37.4
LGD	5	10	29.4*	44.9	<b>35.0*</b>	<b>40.2*</b>
INL+Bo2	5	10	27.5	42.7	32.6	37.5
SPL	5	10	<b>29.6*</b>	<b>47.0*</b>	34.6*	39.5*
LM+MIX	10	10	28.4	45.5	31.8	37.6
LGD	10	10	<b>30.0*</b>	46.8*	<b>35.5*</b>	38.9
INL+Bo2	10	10	27.2	43.0	32.3	37.4
SPL	10	10	<b>30.0*</b>	<b>48.9*</b>	33.8*	<b>39.1*</b>
LM+MIX	10	20	29.0	46.2	33.7	38.2
LGD	10	20	<b>30.3*</b>	47.6*	<b>37.4*</b>	38.6
INL+Bo2	10	20	27.7	43.5	33.8	37.7
SPL	10	20	29.9*	<b>50.2*</b>	34.3	<b>39.7*</b>
LM+MIX	20	20	28.6	47.9	32.9	37.8
LGD	20	20	<b>29.5*</b>	48.9*	<b>37.2*</b>	<b>41.0*</b>
INL+Bo2	20	20	27.4	44.3	33.5	36.8
SPL	20	20	28.8	<b>50.3*</b>	33.9	39.0*

## 5. DISCUSSION

The Divergence from Randomness (DFR) framework proposed by Amati and van Rijsbergen [3] 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  $Inf_2$ :

$$\begin{aligned}
 RSV(q, d) &= \sum_{w \in q \cap d} x_w^q Inf_2(t_w^d) Inf_1(t_w^d) \\
 &= \sum_{w \in q \cap d} -x_w^q Inf_2(t_w^d) \log Prob_1(t_w^d)
 \end{aligned}$$

The above form shows that DFR models can be seen as information models, as defined by equation 2, with a correction brought by the  $Inf_2$  term. If  $Inf_2(t_w^d)$  was not used in DFR models, the models with Poisson, Geometric, Binomial distributions would not respect condition 2, i.e would not be concave. In contrast, the use of bursty distributions in information models, together with the conditions on the normalization functions, ensure that condition 2 is satisfied. Another important difference between the two models is that DFR models make use of discrete distributions for real-valued variables, a conceptual flaw that information models do not have. Lastly, if the log-logistic, SPL and INL models have very simple forms (see for example the formulas given above for the weight they generate), the PL2 DFR model, one of the top performing DFR models, has a much more complex form ([18]). Information models are thus not only conceptually simpler, they also lead to simpler formulas.

## 6. CONCLUSION

We have presented in this paper the family of information models. These models draw their inspiration from a long standing idea in information retrieval, namely the one that a word in a document may not behave statistically as expected on the collection. Shannon information can be used to capture whenever a word deviates from its average behavior, and we showed how to design IR models based on this information. In particular, we showed that the choice of the distribution to be used in such models was crucial for obtaining good retrieval models, the notion of good retrieval models being formalized here on the basis of the heuristic retrieval constraints developed in [9]. Our theoretical development also emphasized the notion of "burstiness", which has been central to several studies. We showed how this notion relates to heuristic retrieval constraints, and how it can be captured through, e.g., power-law distributions. From these two distributions, we have proposed two effective IR models. The experiments we have conducted on four different collections illustrate the good behavior of these models. They outperform in average the Jelinek-Mercer and Dirichlet prior language models as well as the Okapi BM25 model. They yield results similar to state-of-the-art DFR models (INL2 and PL2) when no pseudo-relevance feedback is used. When using pseudo-relevance feedback, however, the information models we have considered significantly outperform all the other models.

## Acknowledgements

This research was partly supported by the Pascal-2 Network of Excellence ICT-216886-NOE and the French project Fra-grances ANR-08-CORD-008.

## 7. REFERENCES

- [1] E. M. Airoldi, W. W. Cohen, and S. E. Fienberg. Bayesian methods for frequent terms in text: Models of contagion and the  $\delta^2$  statistic.
- [2] G. Amati, C. Carpineto, G. Romano, and F. U. Bordoni. Fondazione Ugo Bordoni at TREC 2003: robust and web track, 2003.
- [3] 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.
- [4] A. L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, October 1999.
- [5] D. Chakrabarti and C. Faloutsos. Graph mining: Laws, generators, and algorithms. *ACM Comput. Surv.*, 38(1):2, 2006.
- [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. P. Harter. A probabilistic approach to automatic keyword indexing. *Journal of the American Society for Information Science*, 26, 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] I. O. V. Plachouras, B. He. University of Glasgow at TREC 2004: Experiments in web, robust and terabyte tracks with terrier, 2004.
- [19] 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.
- [20] C. Zhai and J. Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *CIKM '01: Proceedings of the tenth international conference on Information and knowledge management*, pages 403–410, New York, NY, USA, 2001. ACM.
- [21] 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.

## APPENDIX

### A. PROOF OF THEOREM 3

Let us recall what property 3 states: *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, derivating  $g$  wrt  $x$  and considering only the sign of  $g'$ , we get:

$$\begin{aligned} sg[g'] &= sg[F''F - F'' - F'^2] = sg\left[\left(\frac{F'}{F-1}\right)'\right] \\ &= sg[(\log(1-F))''] = sg[(\log P(X \geq x))'] \end{aligned}$$

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.