Bridging Language Modeling and Divergence from Randomness Models: A Log-Logistic Model for IR

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

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

Abstract. We are interested in this paper in revisiting the Divergence from Randomness (DFR) approach to Information Retrieval (IR), so as to better understand the different contributions it relies on, and thus be able to simplify it. To do so, we first introduce an analytical characterization of heuristic retrieval constraints and review several DFR models wrt this characterization. This review shows that the first normalization principle of DFR is necessary to make the model compliant with retrieval constraints. We then show that the log-logistic distribution can be used to derive a simplified DFR model. Interestingly, this simplified model contains Language Models (LM) with Jelinek-Mercer smoothing. The relation we propose here is, to our knowledge, the first connection between the DFR and LM approaches. Lastly, we present experimental results obtained on several standard collections which validate the simplification and the model we propose.

1 Introduction

Together with the language modeling approach to IR, Divergence from Randomness (DFR) models, recently introduced by Amati and Van Rijsbergen [2], are among the best performing (and thus most used) IR models in international evaluation campaigns as TREC or CLEF. However, the DFR framework is complex and difficult to comprehend, as it relies on several quantities the role of which is not always clear. We are interested here in trying to better understand this framework so as to simplify it. Interestingly, the simplification we arrive at contains standard language models with Jelinek-Mercer smoothing.

The remainder of the paper is organized as follows: Section 3 introduces an analytical characterization of IR heuristics which will be used throughout the paper; Section 3 describes the DFR framework and lists the problems associated with it; Section 4 describes the simplification we propose on the basis of the log-logistic distribution, and the relation with language models; Section 5 finally presents an experimental validation of our simplification. Throughout the paper, we make use of the following notations: $\mathcal{C}$ is a collection of N documents; for each

L. Azzopardi et al. (Eds.): ICTIR 2009, LNCS 5766, pp. 54–65, 2009.

index term w, x_w^d (resp. x_w^q) will represent the number of occurrences of the term in document d (resp. in query q), n_w the number of documents in which the term occurs, F_w the number of occurrences of the term in the whole collection, and z_w a quantity which, depending on the context, is equal to n_w or F_w, potentially normalized by N (we introduce this quantity to simplify the expression of the different models we are going to consider); y_d will denote the length of document d, and avdl the average length of the documents in the collection.

2 Analytical Characterization of IR Heuristics

Following Fang *et al.* [6], 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:

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

where θ 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 $\mathbb{R}^{+*} \times \mathbb{R}^{+*} \times \mathbb{R}^{+*} \times \Theta$, where Θ represents the domain of the parameters in θ. The above form encompasses many IR models, as the vector space model, language models, or divergence from randomness models. For example, for the pivoted normalization retrieval formula [9], $\theta = (s, \text{avdl}, N, x_w^q)$ and:

$$h(x,y,z,\theta) = \frac{1+\ln(1+\ln(x))}{1-s+s\frac{y}{\text{avdl}}} x_w^q \ln(\frac{N+1}{z})$$

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. We formalize these considerations through the following four conditions (a larger set of conditions as well as their relation to the formal definitions proposed by Fang *et al.* [6] are given in Appendix A):

Condition 1. $\forall(y,z,\theta),\ \frac{\partial h(x,y,z,\theta)}{\partial x} > 0$; **Condition 2** $\forall(y,z,\theta), \frac{\partial^2 h(x,y,z,\theta)}{\partial x^2} < 0$

Condition 3. $\forall(x,z,\theta), \frac{\partial h(x,y,z,\theta)}{\partial y} < 0$; **Condition 4** $\forall(x,y,\theta), \frac{\partial h(x,y,z,\theta)}{\partial z} < 0$

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.* [6]) 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.

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 model. It 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*). 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 first normalization principle is associated with a second information defined from a second probability distribution through: $Inf_2(t(x_w^d, y_d)) = 1 - Prob_2(t(x_w^d, y_d))$. For example, using the Laplace law of succession for the first normalization ($Prob_2$), 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 [9]. More recently, [8] 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):

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

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

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, which is not the case for the distributions usually retained in DFR models.

3.2 The First Normalization Principle

The intuition behind Inf_1 is simple. Let $P(t(x_w^d, y_d)|\lambda_w)$ represent the probability of $t(x_w^d, y_d)$ (normalized) occurrences of term w in document d according to parameters λ_w which are estimated or set on the basis of a random distribution of w in the collection. If $P(t(x_w^d, y_d)|\lambda_w)$ 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 $P(x_w^d|\lambda_w)$ 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:

$$\theta = (x_w^q, \text{avdl}, N)$$

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

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)* verify 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, $\frac{\partial^2 h(x,y,z,\theta)}{\partial x^2} = 0$. Hence, Inf_1 alone, for the geometric distribution and the models *I(n)*, *I(F)* and *I(n_e)*, is not sufficient to define a valid IR model[1]. 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:

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

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

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

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

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

But:

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

with $b > 0$, which shows that the models are now compatible with condition 2.

4 A Simplified DFR Approach to IR

Clinchant and Gaussier [4] propose to use a model relying solely on Inf_1, for which they retain the Beta negative binomial (BNB) distribution. The BNB distribution is based on the negative binomial distribution which has been proposed as an alternative to the standard Poisson or binomial models traditionnaly used in IR ([3,1,5]). The negative binomial is an infinite mixture of Poisson distributions, and thus can be considered as an extension of the Two-Poisson model. In [4], the Beta BNB distribution is introduced by using a uniform Beta prior on one of the negative binomial parameters. As those distributions are conjugate, the BNB results in a simple form:

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

$$P_{BNB}(x_w^d|r_w) = \frac{r_w}{(r_w + x_w^d)(r_w + x_w^d + 1)}$$

where r_w^d is a parameter which can either be learned through maximum likelihood, or directly set from the collection (Clinchant and Gaussier suggest to use $\frac{F_w}{N}$, i.e. z_w with our notation). Using this distribution leads to the following IR model:

$$h(x, y, z, \theta) = -x_w^q \log(z) + x_w^q \log((z + t(x, y))(z + t(x, y) + 1))$$

With both length normalizations we consider here, it is easy to see that the above model verifies conditions 1, 2 and 3. However, condition 4 is only partly verified as, using for example equation 3, one obtains: $\text{sgn}(\frac{\partial^2 h(x,y,z,\theta)}{\partial x^2}) = \text{sgn}(z - x \log(1 + c\frac{\text{avdl}}{y}))$, a quantity which can be negative for words appearing a lot of times in the collection (z high) but not often (x low) in a long document (y high). In addition, the above model still suffers from the fact that the underlying distribution is discrete, and yet applied to positive, real-valued variables. We introduce now a new distribution which corrects this problem.

4.1 A Log-Logistic Model for IR

There exists a distribution which can be seen, under certain conditions, as a continuous approximation of the BNB, namely the log-logistic distribution. The log-logistic distribution is a continuous distribution defined on the set of positive real numbers. Its density and cumulative probability function have the form[2].:

$$f_{LL}(x|\alpha, \beta) = \frac{(\beta/\alpha)(x/\alpha)^{\beta-1}}{(1 + (x/\alpha)^\beta)^2} \quad P_{LL}(X < x|\alpha, \beta) = \frac{x^\beta}{x^\beta + \alpha^\beta}$$

Setting $\alpha = r$ and $\beta = 1$ in the log-logistic distribution, we have:

$$\forall x \in \mathbb{R}^+, \; P_{LL}(x \leq X < x+1|r) = \frac{x+1}{r+x+1} - \frac{x}{r+x} = \frac{r}{(r+x+1)(r+x)} \quad (4)$$

Therefore, for all integer n, $P_{LL}(n \leq X < n+1|r) = P_{BNB}(X = n|r)$ and $P_{LL}(X > n|r) = P_{BNB}(X > n|r)$, which explains why the log-logistic distribution can be considered as a continuous approximation of the BNB distribution.

A simplified DFR model based on Inf_1 only and the log-logistic distribution can thus be defined by:

$$RSV(q, d) = \sum_{w \in q \cap d} -x_w^q \log(P_{LL}(X \geq t(x_w^d, y_d)|r_w))$$

where r_w is defined from the whole collection. In the remainder, we consider that r_w is set to either $\frac{F_w}{N}$ or $\frac{n_w}{N}$, a standard setting for the parameter of DFR models.

[2] For more information on the Log Logistic distribution refer to `http://en.wikipedia.org/wiki/Log-logistic_distribution`

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). Using the notation of previous sections, the IR model thus defined corresponds to:

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

This time, this model 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 and given in Property 3 of Appendix A.

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 between our model and the language modeling approach to IR.

4.2 Relation to Language Models

Let L be the number of tokens in the collection. Following [10], the scoring formula for a language model using Jelinek-Mercer smoothing can be written as:

$$RSV(q, d) = \sum_{w \in q \cap d} x_w^q \log(1 + s \frac{\frac{x_w^d}{y_d}}{\frac{F_w}{L}}) \tag{5}$$

Using the log-logistic model introduced previously with $r_w = \frac{F}{N}$ and the length normalization given by equation 2, we have:

$$RSV(q, d) = \sum_{w \in q \cap d} x_w^q \log(1 + c \frac{\frac{x_w^d \times \text{avdl}}{y_d}}{\frac{F_w}{N}}) \tag{6}$$

Given that $\frac{F_w}{N} = \text{avdl} \times \frac{F_w}{L}$, equation 5 is equivalent to equation 6. The LM model with Jelinek-Mercer smoothing can thus be seen as a log-logistic model with a particular length normalization, namely the one given by equation 2.

In the language modeling approach to IR, one starts from term distributions estimated as 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. The DFR framework, and its simplification introduced here, is in a sense more general than the language modeling approach to IR since several length normalizations and several distributions can

be considered, leading to a class of model encompassing the language modeling one, as shown above. 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.

5 Experimental Validation

We use the following collections to assess the validity of our model: TREC-3, ROBUST (TREC), CLEF03 AdHoc Task, GIRT (CLEF Domain Specific2004-2006), TEL British Library (CLEF'08 AdHoc). Table 1 gives the number of documents (N), number of unique terms (M), average document length and number of test queries for these collections. For the TREC-3, ROBUST and NTCIR collections, we used standard Porter stemming. For the CLEF03, GIRT and TEL collections, we used lemmatization, and an additional decoumpounding step for the GIRT collection which is written in German.

5.1 IR Results

As our model is a simplification of the DFR framework making use of a single distribution, the log-logistic one, we want to show experimentally that this simplification does not degrade IR results. The methodology we follow for that is straightforward: compute the mean average precision (MAP) for DFR models and the log-logistic one on several IR collections, and assess whether the difference in the MAP is significant or not.

We used three variants of the log-logistic model: a discrete variant based on the BNB distribution and two direct models, one with r_w set to the mean frequency of the word in the collection, $r_w = \frac{F_w}{N}$, model LG, and one with r_w set to the document frequency $\frac{n_w}{N}$, model LGD. These models were first tested against the standard PL2 and InL2 DFR models. In all cases, we used the length normalization corresponding to equation 3, and chose 4 different settings for c: $c = (0.5, 1, 5, 10)$, which corresponds to the typical range recommended for DFR models. Table 2 shows the MAP and precision at 10 for all the DFR and log-logistic models. TREC3-t refers to the TREC-3 collection with query title only, whereas TREC3-d uses both title and description fields (and similarly for the ROBUST collection with ROB-T and ROB-d). CLEF03 and GIRT queries are long queries with descriptive fields, whereas TEL-BL queries are typically short

Table 1. Characteristics of the different collections

	N	M	Avg DL	# Queries
TREC-3	741 856	668 482	262	50
ROBUST	490 779	992 462	289	250
CLEF03	166 754	80 000	247	60
GIRT	151 319	179 283	109	75
TEL-BL	870 246	259 569	9	50

Table 2. Mean average precision (MAP) and Precision at 10, for the different models and datasets. Bold indicates best performance per line (c value); $*$ indicates a significative difference with the best result by line at the 0.05 level for both T-test and Wilcoxon, whereas † indicates a significant difference with one test only.

		MAP					P10				
	c	BNB	LG	LGD	PL2	INL2	BNB	LG	LGD	PL2	INL2
TREC3-t	0.5	24.7	25.0*	**25.3**	21.6*	23.8*	**50.0**	49.2	**50.0**	46.8	48.6
	1	25.6	25.7	**25.8**	24.3*	25.5	51.8	52.6	**53.4**	51.0	51.0
	5	26.3*	25.8*	25.7*	**27.0**	25.5*	54.2	53.2	52.8	**55.2**	53.4
	10	25.6*	25.3*	25.0*	**26.7**	24.8*	53.6	52.2	51.0*	**53.8**	51.4†
TREC3-d	0.5	28.4	**28.9**	**28.9**	25.8*	28.4	**62.8**	59.8*	58.0*	57.0*	59.8
	1	28.5†	28.6	27.9*	28.7	**29.4**	62.4	58.6*	58.2*	**63.4**	60.4*
	5	24.7*	24.5*	23.0$*$	**28.3**	25.5*	54.6*	53.6*	51.6*	**62.4**	54.2*
	10	21.9*	21.7*	20.3*	**26.0**	22.5*	49.4*	48.4*	45.2*	**54.2**	47.2*
ROB-t	0.5	23.3*	23.8*	**24.0**	20.0*	22.2*	41.7	41.9	**42.0**	38.9*	40.8
	1	23.8*	24.2	**24.3**	22.2*	23.6†	**42.6**	42.5	42.5	41.4	**42.6**
	5	24.3*	**24.7**	**24.7**	**24.7**	24.5	43.5†	43.6	43.3*	**44.5**	42.9*
	10	24.1*	24.5†	24.5	**24.8**	24.4	43.4†	43.8	43.7	**44.3**	43.0
ROB-d	0.5	26.6*	27.1*	**27.4**	23.0*	25.5*	**46.3**	45.5	45.6	43.5*	46.0
	1	26.5	**26.9**	**26.9**	25.4*	**26.9**	45.8	45.3*	45.6†	45.7†	**47.0**
	5	24.5*	24.6*	24.2*	**26.5**	25.5*	44.3*	42.9*	42.0*	**46.1**	42.8*
	10	23.1*	23.1*	22.8*	**25.6**	23.8$*$	41.2*	40.7*	40.1*	**45.0**	40.2*
CLEF03	0.5	47.8	48.8	**49.3**	44.0*	47.3	34.3	**35.0**	34.8	32.3†	33.8
	1	47.7	48.4	48.2	46.2*	**49.3**	33.6	34.3	**34.5**	33.5	34.33
	5	42.7*	45.2*	44.0*	**47.0**	46.2	32.2†	31.8†	32.2	**33.8**	32.7
	10	39.5*	41.0*	39.7*	**45.0**	43.8	31.5†	31.3	30.7*	**32.8**	31.0
GIRT	0.5	40.2*	40.4*	**42.1**	35.0*	39.8*	**67.8**	67.1	67.5	62.5*	66.5
	1	41.0*	41.4*	**42.3**	38.5*	41.5	**69.5**	67.8*	68.9	65.5*	67.0*
	5	41.3	41.7	41.6	**41.8**	40.5*	**70.4**	68.4*	68.7*	69.7	66.1*
	10	41.0*	41.2†	41.2	**42.0**	39.7*	**70.0**	68.0*	67.8*	69.6	65.2*
TEL-BL	0.5	31.5*	**33.0**	**33.0**	21.6*	24.8*	47.2	**47.8**	47.6	35.6*	41.0*
	1	31.6*	32.5	**33.3**	26.8*	29.5*	47.6	48.4	**49.0**	43.0*	45.8
	5	32.0†	32.5	33.0	31.3*	**33.4**	49.8	49.4	**50.0**	47.4†	48.6
	10	31.8*	32.4*	32.9*	31.8*	**33.6**	49.0	49.6	50.0	48.8	**50.2**

queries, containing only few keywords. The PL2 model seems to give better results with $c = 5$, whereas the log-logistic ones tend to prefer $c = 0.5$ and InL2 $c = 1$. A two-sided T-test as well as a Wilcoxon test were computed between the results obtained with the best performing model and the results of other models with the same parameter setting. A $*$ in table 2 indicates that the difference between the two models considered is significant with both tests ($p = 0.05$), whereas a † indicates that the difference is significant with only one test (again with $p = 0.05$). For the MAP, out of 28 runs, log-logistic models perform slightly better than standard DFR models 13 times, and slightly worse 15 times. In most cases, there is no significant difference between the first and second best results (which are often obtained with two different models: standard DFR or

log-logistic). Moreover, over all parameter settings, there is no significant difference between the best standard DFR models and the best log-logistic models on all collections but TREC3-t, for which the PL2 model outperforms the other ones. For the precision at 10 documents, log-logistic models are slightly better 15 times out of 28 (and slightly worse 13 times). Over all parameter settings, there is no significant difference between standard DFR models and log-logistic ones on all collections considered here.

We also tested three different variants of the language model (LM) on these collections. Table 3 shows the performance of these variants, where the first sub-column corresponds to a 0.5 jelinek mercer smoothing, the second sub-column to a $\mu = 1000$ Dirichlet Smoothing, and the third sub-column to a leave-one-out likelihood Dirichlet smoothing estimation [10]. As one can note, the best results obtained with LM models are either equal or slightly lower than the best results obtained with standard DFR and log-logistic models, on all collections.

Table 3. Mean Average Precision and Precision at 10 documents for language models with 3 differents settings: first sub-column corresponding to a 0.5 Jelinek-Mercer smoothing, second to a $\mu = 1000$ Dirichlet prior, and third to a Dirichlet prior estimated with the leave-one-out likelihood method

	TREC3-t			ROB-t			ROB-d			CLEF03			GIRT			TEL-BL		
MAP	23.0	26.8	26.2	22.7	24.5	24.8	26.4	27.0	26.9	48.9	46.2	48.5	39.5	39.5	40.9	31.8	27.4	32.0
P10	43.8	53.4	54.4	39.3	43.9	44.2	44.3	45.7	45.5	35.0	32.5	34.0	65.6	67.3	68.7	47.8	44.4	49.0

It is interesting to note that model LGD outperforms model LG overall, which means that the estimation based on the document frequency (n_w) is better than the one based on the collection frequency (F_w). Language models are however unable to directly use document frequency information, since there is no direct way to convert this information into a probability distribution to be used as a collection model.

6 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. Lastly, we have presented experimental results obtained on several standard collections which validate the simplification and the model we have introduced.

References

1. Airoldi, E.M., Cohen, W.W., Fienberg, S.E.: Bayesian methods for frequent terms in text: Models of contagion and the δ² statistic
2. Amati, G., Rijsbergen, C.J.V.: Probabilistic models of information retrieval based on measuring the divergence from randomness. ACM Trans. Inf. Syst. 20(4), 357–389 (2002)
3. Church, K.W., Gale, W.A.: Poisson mixtures. Natural Language Engineering 1, 163–190 (1995)
4. Clinchant, S., Gaussier, É.: The BNB distribution for text modeling. In: Macdonald, et al. (eds.) [7], pp. 150–161.
5. Airoldi, S.F.E.M., Cohen, W.W.: Statistical models for frequent terms in text. In: CMU-CLAD Technical Report - `http://reports-archive.adm.cs.cmu.edu/cald2005.html` (2004)
6. Fang, H., Tao, T., Zhai, C.: A formal study of information retrieval heuristics. In: SIGIR 2004: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval (2004)
7. Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., White, R.W.: ECIR 2008. LNCS, vol. 4956. Springer, Heidelberg (2008)
8. Na, S.-H., Kang, I.-S., Lee, J.-H.: Improving term frequency normalization for multi-topical documents and application to language modeling approaches. In: Macdonald, et al. (eds.) [7], pp. 382–393.
9. Singhal, A., Buckley, C., Mitra, M.: Pivoted document length normalization. In: SIGIR 1996: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 21–29. ACM, New York (1996)
10. Zhai, C., Lafferty, J.: A study of smoothing methods for language models applied to information retrieval. ACM Trans. Inf. Syst. 22(2), 179–214 (2004)

A Analytical Characterization of IR Heuristics

We show here that the conditions we have presented in section 2 are closely related to the formal definitions of heuristic retrieval constraints given by Fang et al. [6]. The notation is the same as the one used before. In particular, a retrieval function, RSV, is defined by $RSV(d,q) = \sum_{w \in q \cap d} h(x_w^d, y_d, z_w, \theta)$. We first briefly recall the definitions of [6]:

TFC1: Let $q = w$ be a query with only one term w. Assume $y_{d1} = y_{d2}$. If $x_w^{d1} > x_w^{d2}$, then $RSV(d1,q) > RSV(d2,q)$.

TFC2: Let $q = w$ be a query with only one term w. Assume $y_{d1} = y_{d2} = y_{d3}$ and $x_w^{d1} > 0$. If $x_w^{d2} - x_w^{d1} = 1$ and $x_w^{d3} - x_w^{d2} = 1$, then $RSV(d2,q) - RSV(d1,q) > RSV(d3,q) - RSV(d2,q)$.

LNC1: Let q be a query and $d1$, $d2$ two documents. If for some word $w' \notin q$, $x_{w'}^{d2} = x_{w'}^{d1} + 1$ but for any query term w, $x_w^{d2} = x_w^{d1}$, then $RSV(d1,q) \geq RSV(d2,q)$.

LNC2: Let q be a query. $\forall k > 1$, if $d1$ and $d2$ are two documents such that $y_{d1} = k \times y_{d2}$ and for all terms w, $x_w^{d1} = k \times x_w^{d2}$, then $RSV(d1,q) \geq RSV(d2,q)$.

TF-LNC: Let $q = w$ be a query with only one term w. If $x_w^{d1} > x_w^{d2}$ and $y_{d1} = y_{d2} + x_w^{d1} - x_w^{d2}$, then $RSV(d1, q) > RSV(d2, q)$.

TDC: Let q be a query and $w1$, $w2$ be two query terms. Assume $y_{d1} = y_{d2}$, $x_{w1}^{d1} + x_{w2}^{d1} = x_{w2}^{d1} + x_{w2}^{d2}$. If *idf(w1)* $\geq$ *idf(w2)* and $x_{w1}^{d1} \geq x_{w1}^{d2}$, then $RSV(d1, q) \geq RSV(d2, q)$.

An interesting special case of TDC corresponds to the situation where $w1$ only occurs in $d1$ and $w2$ only in $d2$. With this setting, the constraint can be formulated as:

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

The following property provides an analytical characterization of the above constraints (s.c. stands for *sufficient condition*).

Property 1. Let:

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

(i) TFC1 $\iff$ a_n increasing. A s.c. is: $\forall(y, z, \theta),\ \frac{\partial h(x,y,z,\theta)}{\partial x} > 0$

(ii) TFC2 $\iff$ c_n decreasing. A s.c. is: $\forall(y, z, \theta), \frac{\partial^2 h(x,y,z,\theta)}{\partial x^2} < 0$

(iii) LNC1 $\iff$ b_n decreasing. A s.c. is: $\forall(x, z, \theta), \frac{\partial h(x,y,z,\theta)}{\partial y} < 0$

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

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

(vi) *speTDC* $\iff \forall(x, y, \theta), \frac{\partial h(x,y,z,\theta)}{\partial z} < 0$

As mentioned before, conditions (i), (iii) and (vi) directly state that h should be increasing with the term frequency, and decreasing with the document length and the document/collection frequency. Note that condition (vi) only represents a necessary condition for the constraint TDC, as we have considered here a restricted form (speTDC) of this contraint. Condition (iv) and (v) directly regulate the interaction between the term frequency and the document length. Lastly, conditions (i) and (ii) 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.