# Frequentist and Bayesian Approach to Information Retrieval

Giambattista Amati

Fondazione Ugo Bordoni, Rome, Italy

**Abstract.** We introduce the hypergeometric models KL, DLH and DLLH using the DFR approach, and we compare these models to other relevant models of IR. The hypergeometric models are based on the probability of observing two probabilities: the relative within-document term frequency and the entire collection term frequency. Hypergeometric models are parameter-free models of IR. Experiments show that these models have an excellent performance with small and very large collections. We provide their foundations from the same IR probability space of language modelling (LM). We finally discuss the difference between DFR and LM. Briefly, DFR is a frequentist (Type I), or combinatorial approach, whilst language models use a Bayesian (Type II) approach for mixing the two probabilities, being thus inherently parametric in its nature.

## 1 Introduction

In a problem of statistical inference, the distribution generating the empirical data have a mathematical form and contains certain parameters, such as mean, variance or other characteristics with unknown values. In Information Retrieval (IR), statistical inference is a very complex type of inference since it involves stratified textual data and different populations, different types of information tasks and information needs, and more importantly a relevance relation is defined over the set of documents. Models for IR may therefore contain parameters whose estimation is based on relevance data.

Language Modelling (LM) [5, 7, 14] is an example of application of statistical inference to IR. According to the language modelling approach to IR [19, 6, 15] a document is a sample of the population, and language model computes the probability that a query is generated by a document. In LM we may use either the mixing or the compounding of two probability distributions, the first distribution models the document, the second one models the collection. The combination of these two probability distributions has the effect of *smoothing* the raw likelihood of occurrence of a term in a document. The statistical combination, whether it is of mixing or compounding type, contains a parameter. In general the value of this parameter is determined by a set of training data made up of a set of topics together with the complete set of relevance values made by some assessors. It is a matter of fact that the optimal value of this parameter varies according to the size, the content of the collection, as well as to the length of the queries, and thus performance may significantly change from collection to collection, and for different query-lengths.

Although DFR baseline models were originally motivated by providing parameter-free models for IR [4], recent developments of the DFR approach have shown that a refinement of the term frequency normalization component (also known as the *document*

M. Lalmas et al. (Eds.): ECIR 2006, LNCS 3936, pp. 13–24, 2006.

*length normalization* problem) may improve the performance of DFR models [2, 12]. A parameter $c$ was introduced to define how "large" is the standard document length in the collection. Term-frequencies are then resized according to the standard length. In general the standard length is the average document length in the collection, and in such a case $c$ is set to 1.

Since LM and the most general form of the DFR models use a parameter, the existence of a highly performing model of IR, easy to implement and completely free from parameters, is still an open problem. The introduction of new parameter-free models must however perform consistently well on small and very large collections, and with different query lengths.

The present investigation on parameter free models for IR thus is important from both theoretical and pragmatical perspectives. The main result of this paper is the definition of very simple but highly performing models of IR that make only use of the textual data and not of the relevance data.

There are other two well known parameter-free models of IR: the vector space model [23, 25, 24, 20] and Ponte and Croft's model [19]. Except Ponte and Croft's model, we here show that language modelling is inherently Bayesian, and it is thus based on parameter smoothing techniques.

Our analysis will start with two foundational views: frequentist and Bayesian. We revisit the information retrieval inference problem assuming these alternative positions. With the aim of producing a parameter free model for IR in mind, we finally provide a document-query matching function based on the information theoretic definition of divergence given by the hypergeometric distribution. Also, we experimentally compare the frequentist approach to language modelling, BM25 and to other DFR models.

## 2 The Metaphor of the Urn Models for IR Models

We assume that IR is modeled by a set of urns or recipients. Sampling consists in the experiment of drawing balls of different colours $\mathbf{V}$ (the vocabulary or the index) from these urns. In the urn paradigm the population of balls represent all tokens of the collection, and the colours are simply the terms listed in the index. Each urn (document) has a prior probability to be selected $\mathrm{P(d)}$, and the balls (tokens) of the same colour (term) have a prior probability $\mathrm{P(t)}$ to be drawn. A document is thus regarded as a sample of the population. In the DFR approach the matching function between a query-term and a document is the probability of extracting a number tf (term frequency) balls of the same colour out of l(d) trials (document length).

An alternative approach is used by Language Modelling. It computes the probability of the query-term in the document by smoothing the maximum likelihood estimate (MLE) of the term-frequency in the document, $\hat{\mathrm{p}} = \dfrac{\mathrm{tf}}{\mathrm{l(d)}}$, with the relative term-frequency in the collection, $\mathrm{P(t)} = \dfrac{\mathrm{TF}}{\mathrm{TFC}}$, where tf is the within-document frequency, l(d) the document length, TF is the number of tokens of that term in the collection and TFC is the overall number of tokens in the collection. Smoothing can be obtained by

either mixing these two probabilities, or extracting the MLE from the compounding of the multinomial distribution with a prior distribution, for example Dirichlet's Priors.

Let us see in details similarities and differences of these two approaches.

### 2.1 Types of Urns

We may classify IR models according to the way we interpret the stratification of the population [10]. We can imagine an ordinary sampling experiment as the selection of balls from a single urn, in general with replacement and shuffling. This is called a Type I model. We may select before a urn at a random, and then make an experiment as described by a Type I model. The urn selection generates a Type II model. Type III model is similarly defined. Translating this hierarchy to IR, we may say

- IR model of Type I. One single urn, where the urn can be either a document or a collection.
- IR model of Type II. We have several urns, which represent either a set of documents or a set of collections.
- IR model of Type III. Different urns containg other urns (set of sets of documents/collections).

Before we construct the frequentist (non-Bayesian) model, we would like to quote Good's argument on the choice of Type I or Type II model for probability estimation [10, page 5-11]:

> [...] The Bayesian will wish to ascribe different weights to different initial (or Type II) distributions. [...] Just as the non-Bayesian finds it expedient to construct mathematical models of Type I probability distributions in which he tries to minimize the number of parameters, the Bayesian will do the same, but with both the Type I and Type II probability distributions. This leads to Type II problems of estimation and significance.

If the Type II probability distribution is unknown, like with the Dirichlet priors in LM, then the Bayesian methodology necessarily leads to the parameter estimation problem.

### 2.2 IR Model of Type I: The Document as a Sample of the Collection

The natural choice for generating a language model of a document is the binomial process. The document is a finite binary sequence of Bernoulli trials whose outcome can be either *a success*, that is an occurrence of the term, or *a failure*, that is an occurrence of a different term. To be more precise, we also assume that the finite binary sequence is *random*, that is any trial is statistically independent from its preceding trials. In a Bernoulli process the probability of a given sequence is

$$P(tf|d,p) = p^{tf} \cdot (1\text{-}p)^{l(d)-tf}$$

where p is the probability of occurrence of the term.

There are $\binom{l(d)}{tf}$ of *exchangeable sequences* (in IR they are also called *a bag of words*), therefore the probability is given by the binomial

$$\mathrm{P}(\mathrm{tf}|\mathrm{d},\mathrm{p}) = \binom{\mathrm{l(d)}}{\mathrm{tf}} \mathrm{p}^{\mathrm{tf}} \cdot (1\text{-}\mathrm{p})^{\mathrm{l(d)}-\mathrm{tf}} \tag{1}$$

The best value for the parameter p in the binomial is unknown. We note that the likelihood $\mathrm{P}(\mathrm{tf}|\mathrm{d},\mathrm{p})$ is maximised when $\frac{d\mathrm{P}(\mathrm{tf}|\mathrm{d},\mathrm{p})}{d\mathrm{p}} = 0$ which is equivalent to set p to the *maximum likelihood estimate MLE* of the term in the document:

$$\hat{\mathrm{p}} = \frac{\mathrm{tf}}{\mathrm{l(d)}} \text{ (MLE)} \tag{2}$$

When the prior p is unknown, then the MLE is a good estimate for p. However we know that the prior probability of occurrence of the term t is the relative term-frequency in the collection:

$$\mathrm{P}(\mathrm{t}) = \frac{\mathrm{TF}}{\mathrm{TFC}} \tag{3}$$

But, what does happen if we substitute the prior $\mathrm{P}(\mathrm{t})$ for p in Equation 1?

Let us then substitute $\mathrm{P}(\mathrm{t})$ for p in Equation 1. For each document d the prior, $\mathrm{P}(\mathrm{t})$ of Equation 3, will be fixed, whilst $\hat{\mathrm{p}}$ of Equation 2 will vary. We have seen that the probability in Equation 1 is maximised with documents d for which $\hat{\mathrm{p}}$ goes towards the value $\mathrm{P}(\mathrm{t}) = \frac{\mathrm{TF}}{\mathrm{TFC}}$. That is, the maximum likelihood estimator coincides with the prior $\mathrm{P}(\mathrm{t})$ when the sample is selected randomly, or better, when the tokens of the term in the document occur randomly. In summary, when the document is little informative, the MLE $\hat{\mathrm{p}}$ of a term in the document approaches the prior $\mathrm{P}(\mathrm{t})$. For non informative terms, we may say that they occur *randomly*. There are words which always fit to this random behaviour. These are the functional words, and they are also called *non-specialty words* [11]. Usually these words are kept in a list of non-informative words, that constitute the so called *stop list*.

But, documents are not built randomly, and thus documents cannot be regarded as they were random samples of the population of the collection. Frequencies of words are biased by some content or semantics. The more a document model diverges from a random behaviour, the more informative it is. In such a case, if the MLE $\hat{\mathrm{p}}$ of a term and its prior $\mathrm{P}(\mathrm{t})$ diverge, the binomial probability diminishes, and the term conveys information. We may assume that the divergence given by the binomial can be used as a measure of the significance of the term (the smaller the binomial , the more significant the term). The mechanism of the DFR models, but also of the 2-Poisson and BM25 models (see a formal derivation of the BM25 from a DFR model [4]), encompasses explicitly such a divergence measure. Then, following our intuition on the divergence from randomness, it would be very natural to use the probability

$$\mathrm{P}(\mathrm{tf}|\mathrm{d}) = \binom{\mathrm{l(d)}}{\mathrm{tf}} \mathrm{P}(\mathrm{t})^{\mathrm{tf}} \left(1 - \mathrm{P}(\mathrm{t})\right)^{\mathrm{l(d)}-\mathrm{tf}} \tag{4}$$

to define a measure of divergence of the probabilities $\hat{\mathrm{p}}$ and $\mathrm{P}(\mathrm{t})$[1].

[1] The same formula is used for query expansion by merging top-ranked documents into a single sample.

*Document ranking is thus obtained by ordering the documents which minimize Equation 4*. As we already observed Equation 4 is maximised when the MLE is equal to $\mathrm{P(t)}$, and in such a case the term distributes randomly (non informative terms), but Equation 4 is minimised when the two probabilities $\hat{\mathrm{p}}$ and $\mathrm{P(t)}$ diverge (informative terms). In other words the probability of Equation 4 is *inversely* related to a measure of informativeness of the term. We may soon regard Equation 4 as primitive. However, we want to derive Equation 4 by using aType I model. Doing this we see that the DFR approach is thus frequentist, since it comes from a Type I model, and in contraposition we see that LM employes a Bayesian Type II model. Never the less, both LM and DFR share the same basic probabilistic space. Let us explore these aspects in details.

## 3 Type I Model: The Hypergeometric Model

We said that a DFR model assumes *a high divergence between MLEs and prior probabilities* as a measure of a high informativeness of the term. In other words $\mathrm{P(tf|d, p = P(t))}$ of Equation 4 and information content are inversely related. We need a function which is additive on independent events (terms), and the logarithmic function is the only function which satisfies such a condition:

$$\mathrm{Inf(tf||d)} = -\log_{\mathbf{2}} \mathrm{P(tf|d, p = P(t))}$$

We now want to show Equation 4 with a direct derivation from a frequentist approach. The frequentist approach to IR yields the system of probabilities using the paradigm of the occupancy numbers, or with a less sophisticated terminology, transforming the IR inference problem into a combinatorial form. A well known combinatorial problem is the following: in a population of TFC balls there are TF red balls. What is the probability that in a sample of cardinality l(d) there is exactly a number tf of red balls? There are $\binom{\mathrm{TF}}{\mathrm{tf}}$ ways to choose a red ball, and there are $\binom{\mathrm{TFC}-\mathrm{TF}}{\mathrm{l(d)}-\mathrm{tf}}$ to choose a ball of different colour. All possible configurations are $\binom{\mathrm{TFC}}{\mathrm{l(d)}}$. Therefore the probability is

$$\mathrm{P(tf|d)} = \frac{\binom{\mathrm{TF}}{\mathrm{tf}} \cdot \binom{\mathrm{TFC}-\mathrm{TF}}{\mathrm{l(d)}-\mathrm{tf}}}{\binom{\mathrm{TFC}}{\mathrm{l(d)}}} \tag{5}$$

The probability distribution of Equation 5 is called the *hypergeometric distribution*. An equivalent formula can be obtained by swapping l(d) with TF:

$$\mathrm{P(tf|d)} = \frac{\binom{\mathrm{l(d)}}{\mathrm{tf}} \cdot \binom{\mathrm{TFC}-\mathrm{l(d)}}{\mathrm{TF}-\mathrm{tf}}}{\binom{\mathrm{TFC}}{\mathrm{TF}}}$$

A limit theorem for the hypergeometric distribution is (see [9, page 59]):

$$\binom{\mathrm{l(d)}}{\mathrm{tf}} \left(\mathrm{P(t)} - \frac{\mathrm{tf}}{\mathrm{TFC}}\right)^{\mathrm{tf}} \left(1 - \mathrm{P(t)} - \frac{\mathrm{l(d)} - \mathrm{tf}}{\mathrm{TFC}}\right)^{\mathrm{l(d)}-\mathrm{tf}} < \mathrm{P(tf|d)} < \binom{\mathrm{l(d)}}{\mathrm{tf}} \mathrm{P(t)}^{\mathrm{tf}} \left(1 - \mathrm{P(t)}\right)^{\mathrm{l(d)}-\mathrm{tf}} \left(1 - \frac{\mathrm{l(d)}}{\mathrm{TFC}}\right)^{-\mathrm{l(d)}}$$

where $\mathrm{P(t)}$ is the frequency $\frac{\mathrm{TF}}{\mathrm{TFC}}$ of the term in the collection. Therefore, the binomial distribution of Equation 4

$$\mathcal{B}(\mathrm{l(d)}, \mathrm{tf}, \mathrm{P(t)}) = \binom{\mathrm{l(d)}}{\mathrm{tf}} \mathrm{P(t)}^{\mathrm{tf}} \left(1 - \mathrm{P(t)}\right)^{\mathrm{l(d)}-\mathrm{tf}}$$

is obtained as a limiting form of the hypergeometric distribution when the population TFC is very large and the size of the sample is very small, that is when both $\frac{\mathrm{l(d)}}{\mathrm{TFC}} \sim 0$ and $\frac{\mathrm{tf}}{\mathrm{TFC}} \sim 0$. *Thus, we have formally derived the Equation 4*:

$$\begin{aligned}\mathrm{Inf(tf||d)} &= -\log_{\mathbf{2}} \mathrm{P(tf|d, p = P(t))} = -\log_{\mathbf{2}} \mathcal{B}(\mathrm{l(d)}, \mathrm{tf}, \mathrm{P(t)}) \\ &= -\log_{\mathbf{2}} \left[\binom{\mathrm{l(d)}}{\mathrm{tf}} \mathrm{P(t)}^{\mathrm{tf}} \left(\mathbf{1} - \mathrm{P(t)}\right)^{\mathrm{l(d)}-\mathrm{tf}}\right]\end{aligned}$$

We need to simplify relation 4 to have a workable model of IR. To obtain this, we start with a very useful relation that relates the binomial distribution to the information theoretic *divergence* $\mathcal{D}$ of $\phi$ from $\psi$ (also called the symmetric Kullback-Leibler divergence):

$$\mathcal{D}(\phi, \psi) = \phi \cdot \log_2 \frac{\phi}{\psi} + (1 - \phi) \cdot \log_2 \frac{(1 - \phi)}{(1 - \psi)} \tag{6}$$

Renyi [21] indeed proves the following relation:

$$\mathcal{B}\left(\mathrm{l(d)}, \mathrm{tf}, \mathrm{P(t)}\right) \sim \frac{2^{-\mathrm{l(d)} \cdot \mathcal{D}\left(\hat{\mathrm{p}}, \mathrm{P(t)}\right)}}{\left(2\pi \cdot \mathrm{tf}(1 - \hat{\mathrm{p}})\right)^{\frac{1}{2}}} \tag{7}$$

where $\hat{\mathrm{p}}$ is the MLE of the probability of the term in the document d of Equation 2. We may delete the contribution of $(1 - \hat{\mathrm{p}}) \cdot \log_2 \frac{(1 - \hat{\mathrm{p}})}{(1 - \mathrm{P(t)})}$ in Equation 7 because it is very small. Using the asymmetric Kullback-Leibler divergence

$$\mathbf{KL}(\hat{\mathrm{p}}||\mathrm{P(t)}) = \hat{\mathrm{p}} \cdot \log_{\mathbf{2}} \left(\frac{\hat{\mathrm{p}}}{\mathrm{P(t)}}\right)$$

we can further simplify the information content:

$$\begin{aligned}\mathrm{Inf(tf||d)} &\sim \mathrm{l(d)} \cdot \mathcal{D}\left(\hat{\mathrm{p}}, \mathrm{P(t)}\right) + 0.5 \log_2 \left(2\pi \cdot \mathrm{tf} \cdot (1 - \hat{\mathrm{p}})\right) \\ &\sim \mathrm{l(d)} \cdot \mathbf{KL}(\hat{\mathrm{p}}||\mathrm{P(t)}) + 0.5 \log_2 \left(2\pi \cdot \mathrm{tf} \cdot (1 - \hat{\mathrm{p}})\right) \\ &\sim \mathrm{tf} \cdot \log_2 \left(\frac{\hat{\mathrm{p}}}{\mathrm{P(t)}}\right) + 0.5 \log_2 \left(2\pi \cdot \mathrm{tf} \cdot (1 - \hat{\mathrm{p}})\right)\end{aligned}$$

### 3.1 DLH and DLLH: Parameter-Free Models of IR

To obtain the matching function we use the *average amount of information* of the term. Instead of using the raw average information carried by a term, that is $\frac{\text{Inf}(\text{tf}||\text{d})}{\text{tf}}$, we use the *cross-entropy function*. With cross-entropy the average information is a *smoothed* with the Laplace normalization $L$ [4]. The Laplace smoothing is similar to Robertson and Walker's normalization used for the family of BM models [22]. Briefly, we derive the model DLH (DFR model based on the Hypergeometric distribution and the Laplace normalization) as:

$$\begin{aligned}\text{weight} = \frac{\text{Inf}(\text{tf}||\text{d})}{\text{tf}+1} = \frac{-\log_2 \mathcal{B}(\text{l(d)}, \text{tf}, \text{P(t)})}{\text{tf}+1} = \\ = \frac{\text{tf} \cdot \log_2 \left(\frac{\hat{\text{p}}}{\text{P(t)}}\right) + 0.5 \cdot \log_2 \left(2\pi \cdot \text{tf} \cdot (1-\hat{\text{p}})\right)}{\text{tf}+1} \, (\mathbf{DLH})\end{aligned} \tag{8}$$

Instead of the average information we may also use the product of two information contents:[2]

$$\text{weight} = \log_2 \left(1+\frac{1}{\text{tf}}\right) \cdot \left(\text{tf} \cdot \log_2 \left(\frac{\hat{\text{p}}}{\text{P(t)}}\right) + 0.5 \cdot \log_2 \left(2\pi \cdot \text{tf} \cdot (1-\hat{\text{p}})\right)\right) (\mathbf{DLLH}) \tag{9}$$

Since the first addendum of Equation 8 is related to the asymmetric Kullback-Leibler divergence as follows:

$$\text{l(d)} \cdot \mathbf{KL}(\hat{\text{p}}||\text{P(t)}) = \text{tf} \cdot \log_{\mathbf{2}} \left(\frac{\hat{\text{p}}}{\text{P(t)}}\right)$$

This suggest to use a further simplified parameter-free model of IR, called KL:

$$\text{weight} = \frac{\text{l(d)} \cdot \mathbf{KL}(\hat{\text{p}}||\text{P(t)})}{\text{tf}+1} = \frac{\text{tf}}{\text{tf}+1} \cdot \log_2 \left(\frac{\hat{\text{p}}}{\text{P(t)}}\right) (\mathbf{KL}) \tag{10}$$

where $\hat{\text{p}}$ is the MLE as defined in Equation 2 and $\text{P(t)}$ is the prior given by Equation 3. The use of KL divergence is also used in LM [26, 15]. The query expansion weighting function as used in language modeling approach is obtained by *minimizing* the KL-divergence between the document language model and the feedback set of returned documents.

[2] Now, $\hat{\text{p}} = \frac{\text{tf}}{\text{l(d)}}$ and $\text{P(t)} = \frac{\text{TF}}{\text{TFC}}$, and also $\text{TFC} = N \cdot \mathbf{avg\_length}$, where $N$ is the number of documents in the collection and $\mathbf{avg\_length}$ is the average length. Thus the ratio $\frac{\hat{\text{p}}}{\text{P(t)}} = \left(\frac{\text{tf}}{\text{l(d)}}\right) \setminus \left(\frac{\text{TF}}{N \cdot \mathbf{avg\_length}}\right)$ contains very small probability factors. In the implementation these small factors might lead to errors. We suggest to associate the statistics contained in the formula differently, to avoid the appearance of very small numbers, as follows:

$$\log_2 \left(\frac{\hat{\text{p}}}{\text{P(t)}}\right) = \log_2 \left(\left(\text{tf} \cdot \frac{\mathbf{avg\_length}}{\text{l(d)}}\right) \cdot \left(\frac{N}{\text{TF}}\right)\right)$$

## 4 Type II Model: Language Models

Let us take again the binomial distribution of Equation 1 as the *likelihood* probability with parameter p unknown. Bayes' Theorem compounds the likelihood distribution with Type II priors P(p|d) over the document collection. The Dirichlet distribution can be used to assign the priors. In such a case, the compound generates the *generalised hypergeometric distribution*:

$$\mathrm{P(d|tf)} = \frac{\binom{\mathrm{l(d)}}{\mathrm{tf}}\, \mathrm{p}^{\mathrm{tf}} \cdot (1\text{-}\mathrm{p})^{\mathrm{l(d)}-\mathrm{tf}} \cdot \mathrm{P(p|d)}}{\int_0^1 \binom{\mathrm{l(d)}}{\mathrm{tf}}\, \mathrm{P(tf|d,p)} \cdot \mathrm{P(p|d)} d\mathrm{p}} \tag{11}$$

Dirichlet priors has a set of parameters $A_1, \ldots, A_V > 0$, one parameter for each term $\mathrm{t}_i$ of the vocabulary of size $V$. The term-frequencies obviously satisfy the condition $\mathrm{tf}_1 + \ldots + \mathrm{tf}_V = \mathrm{l(d)}$. The Dirichlet priors are:

$$\mathrm{P}(\mathrm{p}_1, \ldots, \mathrm{p}_V|\mathrm{d}, A_1, \ldots, A_V) = \frac{\Gamma(A)}{\Gamma(A_1)\cdots\Gamma(A_V)} \mathrm{p}_1^{A_1-1} \cdots \mathrm{p}_V^{A_V-1}$$
$$A = \sum_{i=1}^{V} A_i \text{ and } \sum_{i=1}^{V} \mathrm{p}_i = 1$$

The *a posteriori* probability distribution after conditionalizing on the Type II distribution $\mathrm{P}(\mathrm{p}_1, \ldots, \mathrm{p}_V|\mathrm{d}, A_1, \ldots, A_V)$ takes the same form of Equation 4, that is:

$$\mathrm{P}(\mathrm{d}|\mathrm{tf}_1, \ldots, \mathrm{tf}_V, A_1, \ldots, A_V) =$$
$$= \frac{\Gamma(A + \mathrm{l(d)})}{\Gamma(A_1 + \mathrm{tf}_1)\cdots\Gamma(A_V + \mathrm{tf}_V)} \mathrm{p}_1^{\mathrm{tf}_1+A_1-1} \cdots \mathrm{p}_V^{\mathrm{tf}_V+A_V-1}$$

Setting $A_{\mathrm{t}} = \mu \cdot \mathrm{P(t)}$ with $\mu$ an unkown parameter, the MLE of the compound of the likelihood with probability P as defined by Equation 11 or 4 is:

$$\hat{\mathrm{p}}_{LM} = \frac{\mathrm{tf} + \mu \cdot \mathrm{P(t)}}{\mathrm{l(d)} + \mu}$$

Using additivity on independent events of the logarithmic function, we have:

$$\mathrm{p}(Q|\mu, \mathrm{d}) \propto \frac{1}{|Q|} \sum_{i=1}^{|Q|} \log_2 \left( \frac{\mathrm{tf}_i}{\mu \mathrm{P}(\mathrm{t}_i)} + 1 \right) - \log_2(\mathrm{l(d)} + \mu) \quad (\text{ LM }).$$

## 5 Comparison of the Frequentist with the Bayesian Approach

We have seen that the frequentist approach defines a parameter-free model of IR, while the Bayesian approach leads to the construction of a parameter based model of IR. The main difference between the two approaches are

**Table 1.** Short queries (Title) of the Robust Track of TREC 2004 (250 queries)

| Model | MAP | R Prec. | Prec. at 5 | Prec. at 10 |
|---|---|---|---|---|
| DLLH | 0.2483 | 0.2887 | 0.4651 | 0.4281 |
| DLH | 0.2438 | 0.2858 | 0.4843 | 0.4373 |
| KL | 0.2343 | 0.2765 | 0.4763 | 0.4289 |
| Ponte & Croft | 0.2383 | 0.2847 | 0.4297 | 0.3972 |
| LM ($\mu = 600$) | 0.2519 | 0.2939 | 0.4803 | 0.4313 |
| BM25 (b=0.75, k=1.2) | 0.2418 | 0.2858 | 0.4731 | 0.4273 |
| PL2 (c=6) | **0.2563** | **0.2979** | **0.4876** | **0.4430** |

1. DFR approach computes the probability of observing two probabilities, while LM smoothes the MLE of a term in the document.
2. DFR approach weights terms according to the *im*probability of observing the MLE of a term in the document given the prior, and it is based on information theoretic notions, such as amount of information and uncertainty. LM instead weights the probability of observing the term in a document given a prior distribution.
3. In DFR approach there are no *non-zero probabilities*, that is when a term does not occur in a document it does not contribute at all to the document score. On the contrary, a term that does not appear in a document plays an important role in LM approach. This requires extra computational costs either in terms of additional index or retrieval structures.
4. The basic DFR models (such as Formulas 10, 8 and 9) can be used as they are for query expansion. A parameter free model of query expansion can be also defined [3]. Also Kullback-Leibler divergence based techniques for query expansion [8, 26], as it was here shown, are approximations of the hypergeometric model and the binomial model.
5. With DFR approach we can combine LM with DFR models or BM25 into a single model, with the advantage of not having non-zero probabilities [1, 13].
6. On the other hand, Bayesian approach is flexible and easy to be applied with a stratified population and in presence of other parameters, while frequentist approach requires a major attention to model complex combinatorial problems.

**Table 2.** Short queries (Title) with DFR Query Expansion of Robust Track 2004 (250 queries) with 40 most informative terms from 8 topmost-retrieved documents

| Model | MAP | R Prec. | Prec. at 5 | Prec. at 10 |
|---|---|---|---|---|
| DLLH | **0.3052** | **0.3303** | 0.5012 | 0.4538 |
| DLH | 0.2912 | 0.3181 | 0.4980 | 0.4514 |
| KL | 0.2821 | 0.3096 | 0.4948 | 0.4462 |
| LM ($\mu$=400) | 0.2968 | 0.3245 | 0.4867 | 0.4562 |
| BM25 (b= 0.75, k=1.2) | 0.2950 | 0.3182 | 0.4956 | 0.4482 |
| PL2 (c=6) | 0.2984 | 0.3253 | **0.5052** | **0.4622** |

**Table 3.** Title and DFR Query Expansion (QE) - Terabyte Track 2004 (GOV2). In order to make a comparison with DLH, we here display the best baseline run. It is relative to the same system that obtained the best TREC run, but using query expansion and single keyword approach only.

| Model | MAP |
|---|---|
| DLH | 0.277 |
| best TREC | 0.284 |
| baseline (with QE) of the best TREC | 0.253 |

## 6 Experiments

We used two test collections of TREC (Text REtrieval Conference). The first collection is from disks 4 and 5 of TREC minus the CR collection and consists of about 2 Gbytes of data, with 528,107 documents and 91,088,037 pointers. The second collection is the terabyte collection GOV2 and consists of about 426 GB Gbytes of data, with about 25 million documents. We used 250 queries (queries 300-450 and 600-700) of the Robust (ad hoc) track of TREC 2004 with the 2GB collection. These queries are ad hoc topics used i since TREC 7. We used also 50 topics of the Terabyte track of TREC 2004. The optimal performance value of $c$ of the DFR models depends on either the query-length (short or long, with or without query expansion) or the collection. The length of the query with query expansion can be regarded short in the case of the Terabyte collection because only 10 additional terms were added to the topics, while it must be considered long in the case of the 2GB collection, because 40 additional terms were added to the topics. We have compared the new models KL, DLH and DLLH with BM25, LM, Ponte and Croft's parameter free model of LM , and the Poisson model PL2, that was shown to have an excellent performance in both .GOV and the terabyte collection GOV2 at the TREC conference [17, 18]. We have used a default value $c = 6$ for PL2. On the other hand, we have used different optimal values of $\mu$ for the model LM. The same query expansion techniques as described in [3] has been applied to all models.

## 7 Discussion of the Results and Conclusions

We have derived from the frequentist approach of IR some very simple document-ranking models which we have shown to perform very well with two different collection sizes (a collection of 2 GB and a collection of 426 GB). These models are free from parameters and can be used with any collection without tuning parameters. The problem of parameter tuning is instead important in the language modelling approach. Zhai and Lafferty report [27] that an inadequate smoothing may hurt the performance more heavily in the case of long and verbose queries. Also, the optimal value of the LM parameter $\mu$ tends to be larger for long queries than for short queries. They observe that smoothing plays a more important role for long queries than for short queries. They also observe that Dirichlet prior performs worse on long queries than title queries on the web collection. In particular, for each subcollection contained in the 2GB collection the optimal value of $\mu$ varies from 500 to 4000 for the short queries and from 2000 to 5000 for the long queries. They conclude that the optimal value of $\mu$ depends both on the collection and the verbosity of the query.

The hypergeometric models have very good performance. In particular, they have the best MAP for short queries with query expansion on the 2GB collection. As for the Terabyte collection, they have better performance than the best TREC run that uses query expansion and single keyword approach, and a close performance to the best run, which however uses additional document and term structures.

## Acknowledgments

This paper was given as oral contribution at the "IR & Theory workshop", held in Glasgow the 25th July 2005. We thank Keith van Rijsbergen, Iadh Ounis for infinitely long discussions on DFR approach, and University of Glasgow for support. Special gratitude goes to Ben He who ran the experiments with the GOV2 collection. All experiments of this paper have been conducted using the Terrier version 1.0.2 [16]. The second addendum of Equation 6, implemented in DLH version of Terrier, was not here used.

## References

1. AMATI, G. *Probability Models for Information Retrieval based on Divergence from Randomness*. PhD thesis, University of Glasgow, June 2003.
2. AMATI, G., CARPINETO, C., AND ROMANO, G. FUB at TREC 10 web track: a probabilistic framework for topic relevance term weighting. In *In Proceedings of the 10th Text Retrieval Conference TREC 2001* (Gaithersburg, MD, 2002), E. Voorhees and D. Harman, Eds., NIST Special Pubblication 500-250, pp. 182–191.
3. AMATI, G., CARPINETO, C., AND ROMANO, G. Fondazione Ugo Bordoni at TREC 2004. In *In Proceedings of the 13th Text Retrieval Conference TREC 2001* (Gaithersburg, MD, 2004), E. Voorhees and D. Harman, Eds., NIST Special Pubblication 500-261.
4. AMATI, G., AND VAN RIJSBERGEN, C. J. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems (TOIS) 20*, 4 (2002), 357–389.
5. BAHL, L. R., JELINEK, F., AND MERCER, R. L. A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-5*, 2 (Mar. 1983), 179–190.
6. BERGER, A., AND LAFFERTY, J. Information retrieval as statistical translation. In *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (New York, NY, USA, 1999), ACM Press, pp. 222–229.
7. BROWN, P. F., COCKE, J., DELLA PIETRA, S. A., DELLA PIETRA, V. J., JELINEK, F., LAFFERTY, J. D., MERCER, R. L., AND ROOSSIN, P. S. A statistical approach to machine translation. *Computational Linguistics 16*, 2 (June 1990), 79–85.
8. CARPINETO, C., DE MORI, R., ROMANO, G., AND BIGI, B. An information theoretic approach to automatic query expansion. *ACM Transactions on Information Systems 19*, 1 (2001), 1–27.
9. FELLER, W. *An introduction to probability theory and its applications. Vol. I*, third ed. John Wiley & Sons Inc., New York, 1968.
10. GOOD, I. J. *The Estimation of Probabilities: an Essay on Modern Bayesian Methods*, vol. 30. The M.I.T. Press, Cambridge, Massachusetts, 1968.
11. HARTER, S. P. *A probabilistic approach to automatic keyword indexing*. PhD thesis, Graduate Library, The University of Chicago, Thesis No. T25146, 1974.

12. HE, B., AND OUNIS, I. A study of parameter tuning for term frequency normalization. In *Proceedings of the twelfth International Conference on Information and Knowledge Management* (2005), Springer.
13. HE, B., AND OUNIS, I. A study of the Dirichlet priors for term frequency normalisation. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval* (New York, NY, USA, 2005), ACM Press, pp. 465–471.
14. JELINEK, F., AND MERCER, R. Interpolated estimation of markov source parameters from sparse data. In *Pattern Recognition in Practice* (Amsterdam, Netherlands, 1980), North-Holland, pp. 381–397.
15. LAFFERTY, J., AND ZHAI, C. Document Language Models, Query Models, and Risk Minimization for Information Retrieval. In *Proceedings of ACM SIGIR* (New Orleans, Louisiana, USA, September 9-12 2001), ACM Press, New York, NY, USA, pp. 111–119.
16. OUNIS, I., AMATI, G., V., P., HE, B., MACDONALD, C., AND JOHNSON, D. Terrier Information Retrieval Platform. In *Proceedings of the 27th European Conference on IR Research (ECIR 2005)* (2005), vol. 3408 of *Lecture Notes in Computer Science*, Springer, pp. 517 – 519.
17. PLACHOURAS, V., HE, B., AND OUNIS, I. University of Glasgow at TREC2004: Experiments in Web, Robust and Terabyte tracks with Terrier. In *Proceedings of the 13th Text REtrieval Conference (TREC 2004)* (Gaithersburg, MD, 2004), NIST Special Pubblication 500-261.
18. PLACHOURAS, V., AND OUNIS, I. Usefulness of hyperlink structure for query-biased topic distillation. In *Proceedings of the 27th annual international conference on Research and development in information retrieval* (2004), ACM Press, pp. 448–455.
19. PONTE, J., AND CROFT, B. A Language Modeling Approach in Information Retrieval. In *The 21st ACM SIGIR Conference on Research and Development in Information Retrieval* (Melbourne, Australia, 1998), B. Croft, A. Moffat, and C. Van Rijsbergen, Eds., ACM Press, pp. 275–281.
20. RAGHAVAN, V. V., AND WONG, S. K. A critical analysis of the vector space model for information retrieval. *Journal of the American Society for Information Science 37*, 5 (1986), 279–287.
21. RENYI, A. *Foundations of probability*. Holden-Day Press, San Francisco, USA, 1969.
22. ROBERTSON, S., AND WALKER, S. Some simple approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval. In *Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval* (Dublin, Ireland, June 1994), Springer-Verlag, pp. 232–241.
23. SALTON, G. *The SMART Retrieval System*. Prentice Hall, New Jersey, 1971.
24. SALTON, G., AND MCGILL, M. *Introduction to modern Information Retrieval*. McGraw–Hill, New York, 1983.
25. SALTON, G., WONG, A., AND YANG, C. S. A vector space model for automatic indexing. *Commun. ACM 18*, 11 (1975), 613–620.
26. ZHAI, C., AND LAFFERTY, J. Model-based Feedback in the Language Modeling Approach to Information Retrieval. In *CIKM 2001* (Atlanta, Georgia, USA, November 5-10 2001), ACM Press, New York, NY, USA, pp. 334–342.
27. ZHAI, C., AND LAFFERTY, J. A Study of Smoothing Methods for Language Models Applied to Information Retrieval. *ACM Transactions on Information Systems 22*, 2 (April 2004), 179214.