Probabilistic Models of Information Retrieval Based on Measuring the Divergence from Randomness

GIANNI AMATI
University of Glasgow, Fondazione Ugo Bordoni
and
CORNELIS JOOST VAN RIJSBERGEN
University of Glasgow

We introduce and create a framework for deriving probabilistic models of Information Retrieval. The models are nonparametric models of IR obtained in the *language model* approach. We derive term-weighting models by measuring the divergence of the actual term distribution from that obtained under a random process. Among the random processes we study the binomial distribution and Bose–Einstein statistics. We define two types of term frequency normalization for tuning term weights in the document–query matching process. The first normalization assumes that documents have the same length and measures the information gain with the observed term once it has been accepted as a good descriptor of the observed document. The second normalization is related to the document length and to other statistics. These two normalization methods are applied to the basic models in succession to obtain weighting formulae. Results show that our framework produces different nonparametric models forming baseline alternatives to the standard *tf-idf* model.

Categories and Subject Descriptors: H.3.3 [**Information Search and Retrieval**]: Retrieval models

General Terms: Algorithms, Experimentation, Theory

Additional Key Words and Phrases: Aftereffect model, BM25, binomial law, Bose–Einstein statistics, document length normalization, eliteness, idf, information retrieval, Laplace, Poisson, probabilistic models, randomness, succession law, term frequency normalization, term weighting

1. INTRODUCTION

The main achievement of this work is the introduction of a methodology for constructing nonparametric models of Information Retrieval (IR). Like the *language model* approach of IR [Ponte and Croft 1998]) in which the weight of a word in a document is given by a probability, a nonparametric model is derived in a purely theoretic way as a combination of different probability distributions.

Authors' addresses: G. Amati, Fondazione Ugo Bordoni, via B. Castiglione 59, 00142, Roma, Italy; email: gba@fub.it; C. J. van Rijsbergen, Computing Science Department, University of Glasgow, 17 Lilybank Gardens G12 8QQ Glasgow, Scotland; email: keith@dcs.gla.ac.uk.

The advantage of having a nonparametric approach is that the derived models do not need to be supported by any form of data-driven methodology, such as the learning of parameters from a training collection, or using data smoothing techniques. In addition to the feature that the constructed models do not have parameters, the second important feature that results from our approach is that different choices of probability distributions to be used in the framework generate different IR models. Our framework was successfully used in the TREC-10 (see Table III) Conference [Amati et al. 2001] and the model B_EL2 (see Table I) was shown to be one of the best performing runs at the WEB track.

There are many models of IR based on probability theory [Damerau 1965; Bookstein and Swanson 1974; Harter 1975a,b; Robertson and Sparck-Jones 1976; Cooper and Maron 1978; Croft and Harper 1979; Robertson et al. 1981; Fuhr 1989; Turtle and Croft 1992; Wong and Yao 1995; Hiemstra and de Vries 2000; Ponte and Croft 1998], but "probabilistic models" have in general denoted those models that use "relevance" as an element of the algebra of events and possibly satisfy the Probability Ranking Principle (PRP). PRP asserts that documents should be ranked in decreasing order of the probability of relevance to a given query. In this respect, our framework differs from foregoing models in that relevance is not considered a primitive notion. Rather, we rank documents by computing the gain in retrieving a document containing a term of the query. According to this foundational view, blind relevance feedback for query expansion was used in TREC-10 only to predict a possible term frequency in the expanded query [Carpineto and Romano 2000] (denoted by *qtf* in formula (43) of Section 4) and not used to modify the original term weighting.

Our framework has its origins in the early work on automatic indexing by Damerau [1965], Bookstein and Swanson [1974], and Harter [1975a,b], who observed that the significance of a word in a document collection can be tested by using the Poisson distribution. These early models for automatic indexing were based on the observation that the distribution of informative content words, called by Harter "specialty words," over a text collection deviates from the distributional behavior of "nonspecialty" words. Specialty words, like words belonging to a technical vocabulary, being informative, tend to appear more densely in a few "elite" documents, whereas nonspecialty words, such as the words that usually are included in a stop list, are randomly distributed over the collection. Indeed, "nonspecialty" words are modeled by a Poisson distribution with some mean λ.

Hence one of the hypotheses of these early linguistic models is that an informative content word can be mechanically detected by measuring the extent to which its distribution deviates from a Poisson distribution, or, in other words, by testing the hypothesis that the word distribution *on the whole document collection* does not fit the Poisson model.

A second hypothesis assumed by Harter's model is that a specialty word again follows a Poisson distribution but on a smaller set, namely, *the set of the elite documents*, this time with a mean μ greater than λ. The notion of eliteness was first introduced in Harter [1974, pp. 68–74]. According to Harter, the idea of eliteness is used to reflect the level of treatment of a word in a small set of documents compared with the rest of the collection. In the elite set a word occurs to a relatively greater extent than in all other documents. Harter defines eliteness through a probabilistic estimate which is interpreted as the proportion of documents that a human indexer assesses elite with respect to a word t. In our proposal we instead assume that the elite set of a word t is simply the set D_t of documents containing the term. Indeed, eliteness, as considered in Harter's model, is a hidden variable, therefore the estimation of the value for the parameter μ is problematic. Statistical tests have shown that his model is able to assign "sensible" index terms, although only a very small data collection, and a small number of randomly chosen specialty words, are used.

Harter used the Poisson distribution only to select good indexing words and not to provide indexing weights. The potential effectiveness of his model for a direct exploitation in retrieval was explored by Robertson, van Rijsbergen, Williams, and Walker [Robertson et al. 1981; Robertson and Walker 1994] by plugging the Harter 2-Poisson model [Harter 1975a] into the Robertson–Sparck-Jones probabilistic model [Robertson and Sparck-Jones 1976]. The conditional probabilities $p(E|R)$, $p(\overline{E}|R)$, $p(E|\overline{R})$, $p(\overline{E}|\overline{R})$, where E was the elite set, and R was the set of relevant documents, were substituted for the cardinalities of the sets in the 2×2-cell contingency table of the probabilistic model. The estimates of these conditional probabilities were derived from the 2-Poisson model by means of the Bayes theorem [Titterington et al. 1985], thus deriving a new probabilistic model that depends on the means λ and μ in the nonelite and elite sets of documents, respectively. The model has been successfully extended and then approximated by a family of limiting forms called *BM*s (*BM* for Best Match) by taking into account other variables such as the within document–term frequency and the document length [Robertson and Walker 1994].

A generalization of the 2-Poisson model as an indexing selection function, the N-Poisson model, was given by Margulis [1992].

We incorporate frequency of words by showing that the weight of a term in a document is a function of two probabilities $Prob_1$ and $Prob_2$ which are related by:

$$w = (1 - Prob_2) \cdot (-\log_2 \, Prob_1) = -\log_2 \, Prob_1^{1-Prob_2}. \tag{1}$$

The term weight is thus a decreasing function of both probabilities $Prob_1$ and $Prob_2$. The justification of this new weighting schema follows.

The distribution $Prob_1$ is derived with similar arguments to those used by Harter. We suppose that words which bring little information are randomly distributed *on the whole set of documents*. We provide different basic probabilistic models, with probability distribution $Prob_1$, that define the notion of *randomness in the context of information retrieval*. We propose to define those processes with urn models and random drawings as models of randomness. We thus offer different processes as basic models of randomness. Among them we study the binomial distribution, the Poisson distribution, Bose–Einstein statistics, the inverse document frequency model, and a mixed model using Poisson and inverse document frequency.

We illustrate the Bernoulli model of randomness by an example. Suppose that an elevator is serving a building of 1024 floors and that 10 people take the elevator at the basement floor independently of each other. Suppose that these 10 people have not arrived together. We assume that there is a uniform probability that a person gets off at a particular floor. The probability that we observe 4 people out of 10 leaving at a given arbitrary floor is

$$B(1024, 10, 4) = \binom{10}{4} p^4 q^6 = 0.00000000019,$$

where $p = 1/1024$ and $q = 1023/1024$.

We translate this toy problem into IR terminology by treating documents as floors, and people as tokens of the same term. The term-independence assumption corresponds to the fact that these people have not arrived together, or equivalently that there is not a common cause which has brought all these people at the same time to take that elevator. If F is the total number of tokens of an observed term t in a collection D of N documents, then we make the assumption that the tokens of a nonspecialty word should distribute over the N documents according to the binomial law. Therefore the probability of *tf* occurrences in a document is given by

$$Prob_1(tf) = Prob_1 = B(N, F, tf) = \binom{F}{tf} p^{tf} q^{F-tf},$$

where $p = 1/N$ and $q = (N-1)/N$.

The Poisson model is an approximation of the Bernoulli model and is here defined as in Harter's work (if the probability $Prob_1$ in formulae (1) and (2) is Poisson, then the basic model $Prob_1$ is denoted by P in the rest of the article).

Hence the words in a document with the highest probability $Prob_1$ of occurrence as predicted by such models of randomness are "nonspecialty" words. Equivalently, the words whose probability $Prob_1$ of occurrence conforms most to the expected probability given by the basic models of randomness are noncontent-bearing words. Conversely, words with the smallest expected probability $Prob_1$ are those that provide the *informative content* of the document.

The component of the weight of formula (1):

$$Inf_1 = -\log_2 Prob_1 \tag{2}$$

is defined as the *informative content* Inf_1 of the term in the document. The definition of amount of informative content as $-\log_2 P$ was given in semantic information theory [Hintikka 1970] and goes back to Popper's [1995] notion of informative content and to Solomonoff's and Kolmogorov's Algorithmic Complexity Theory [Solomonoff 1964a,b] (different from the common usage of the notion of entropy of information theory, Inf_1 was also called the *entropy* function in Solomonoff [1964a]). $-\log_2 P$ is the only function of probability that is monotonically decreasing and additive with respect to independent events up to a multiplicative factor [Cox 1961; Willis 1970]. $Prob_1$ is a function of the within document–term frequency *tf* and is the probability of having *by pure chance* (namely, according to the chosen model of randomness) *tf* occurrences of a term t in a document d. The smaller this probability is, the less its tokens are distributed in conformity with the model of randomness and the higher the informative content of the term. Hence, determining the informative content of a term can be seen as an inverse test of randomness of the term within a document with respect to the term distribution in the entire document collection.

The *second probability*, $Prob_2$ of Formula 1, is obtained by observing only the set of all documents in which a term occurs (we have defined such a set as the *elite set* of the term).

$Prob_2$ is the probability of occurrence of the term within a document with respect to its elite set and is related to the risk, $1 - Prob_2$, of accepting a term as a good descriptor of the document when the document is compared with the elite set of the term. Obviously, the less the term is expected in a document with respect to its frequency in the elite set (namely, when the risk $1 - Prob_2$ is relatively high), the more the amount of informative content Inf_1 is gained with this term. In other words, if the probability $Prob_2$ of the word frequency within a document is relatively low with respect to its elite set, then the actual amount of informative content carried by this word within the document is relatively high and is given by the weight in formula (1).

To summarize, the *first probability* $Prob_1$ of term occurrence is obtained from an "ideal" process. These "ideal" processes suitable for IR are called models of randomness. If the expected probability $Prob_1$ turns out to be relatively small, then that term frequency is highly unexpected according to the chosen model of randomness, and thus it is not very probable that one can obtain such a term frequency by accident.

The probability $Prob_2$ is used instead to measure a notion of *information gain* which in turn tunes the informative content as given in formula (1). $Prob_2$ is shown to be a conditional probability of success of encountering a further token of a given word in a given document on the basis of the statistics on the elite set.

An alternative way of computing the informative content of a term within a document was given by Popper [1995] and extensively studied by Hintikka [1970]. In the context of our framework, Popper's formulation of informative content is:

$$Inf_2 = 1 - Prob_2.$$

Under this new reading the fundamental formula (1) can be seen as the product of two informative content functions, the first function Inf_1 being related to the whole document collection D and the second one Inf_2 to the elite set of the term:

$$w = Inf_1 \cdot Inf_2. \tag{3}$$

We have called the process of computing the information gain through the factor $1 - Prob_2$ of formula (1) the *first normalization of the informative content*. The first normalization of the informative content shares with the language model of Ponte and Croft [1998] the use of a risk probability function. The risk involved in a decision produces a loss or a gain that can be explained in terms of utility theory. Indeed, in utility theory the gain is directly proportional to the risk or the uncertainty involved in a decision. In the context of IR, the decision to be taken is the acceptance of a term in the observed document as a descriptor for a potentially relevant document: the higher the risk, the higher the gain, and the lower the frequency of that term in the document, in comparison to both the length of the document and the relative frequency in its elite set.

A similar idea based on risk minimization for obtaining an expansion of the original query can be found in Lafferty and Zhai [2001]. In this work some loss functions (based on relevance and on a distance-similarity measure) together with Markov chains are used to expand queries.

So far, we have introduced two probabilities: the probability $Prob_1$ of the term given by a model of randomness and the probability $Prob_2$ of the risk of accepting a term as a document descriptor. Our term weight w of formula (1) is a function of four random variables:

$$w = w(F, tf, n, N),$$

where *tf* is the within document–term occurrence frequency, N is the size of the collection, n is the size of the elite set of the term, and F is the total number of occurrences in its elite set (which is obviously equal to the total number of occurrences in the collection by definition of elite set). However, the size of *tf* depends on the document length: we have to derive the expected term frequency in a document when the document is compared to a given length (typically the average document length). We determine the distribution that the tokens of a term follow in the documents of a collection at different document lengths. Once this distribution is obtained, the normalized term frequency *tfn* is used in formula (1) instead of the nonnormalized *tf*. We have called the process of substituting the normalized term frequency for the actual term frequency *the second normalization* of the informative content.

One formula we have derived and successfully tested is:

$$tfn = tf \cdot \log_2 \left(1 + \frac{avg\_l}{l}\right), \tag{4}$$

where avg_l and l are the average length of a document in the collection and the length of the observed document, respectively.

Our term weight w in formula (1) is thus a function of six random variables:

$$w = w(F, tfn, n, N) = w(F, tf, n, N, l, avg\_l).$$

1.1 Probabilistic Framework

Our probabilistic framework builds the weighting formulae in sequential steps.

1. First, a probability $Prob_1$ is used to define a measure of informative content Inf_1 in Equation (2). We introduce five *basic models* that measure Inf_1. Two basic models are approximated by two formulae each, and thus we provide seven weighting formulae: $I(F)$ (for Inverse term Frequency), $I(n)$ (for Inverse document frequency where n is the document frequency), $I(n_e)$ (for Inverse expected document frequency where n_e is the document frequency that is expected according to a Poisson), two approximations for the binomial distribution, D (for divergence) and P (for Poisson), and two approximations for the Bose–Einstein statistics, G (for geometric) and B_E (for Bose–Einstein).

2. Then the first normalization computes *the information gain when accepting the term in the observed document as a good document descriptor*. We introduce two (first) normalization formulae: L and B. The first formula derives from Laplace's law of succession and takes into account only the statistics of the observed document d. The second formula B is obtained by a ratio of two Bernoulli processes and takes into account the elite set E of a term. Laplace's law of succession is a conditional probability that provides a solution to the behavior of a statistical phenomenon called an apparent *aftereffect* of sampling by statisticians. It may happen that a sudden repetition of success of a rare event, such as the repeated encountering of a given term in a document, increases our expectation of further success to almost certainty. Laplace's law of succession gives an estimate of such a high expectation. Therefore Laplace's law of succession is one of the candidates for deriving the first normalization formula. The information gain will be directly proportional to the amount of uncertainty $(1 - Prob_2)$ in choosing the term as a good descriptor. Another candidate can be provided similarly with the first normalization B, which computes the incremental rate of probability of further success under a Bernoulli process.

3. Finally, we resize the term frequency in light of the length of the document. We test two hypotheses.

$H1$. Assuming we can represent the term frequency within a document as a density function, we can take this to be a uniform distribution; that is, the density function of the term frequency is constant. The $H1$ hypothesis is a variant of the verbosity principle of Robertson [Robertson and Walker 1994].

$H2$. The density function of the term frequency is inversely proportional to the length.

In Section 2 we introduce the seven basic models of randomness. In Section 3 we introduce the notion of aftereffect in information retrieval and discuss how the notion is related to utility theory. We define the first normalization of the informative content and produce the first baselines for weighting terms. In Section 4 we go on to refine the weighting schemas by normalizing the term frequency in the document to the average document length of the collection. Experiments are discussed in Section 8.

2. MODELS OF RANDOMNESS

Our framework has five basic IR models for the probability $Prob_1$ in formula (1). the binomial model, the Bose–Einstein model, the *tf-idf* (the probability that combines the within document–term frequency with the inverse document frequency in the collection), *tf-itf* (the probability that combines the within document–term frequency with the inverse term frequency in the collection), and tf-expected_idf models (the probability combining the within document–term frequency with the inverse of the expected document frequency in the collection). We then approximate the binomial model by two other models, that is, the Poisson model P and the divergence model D. We also approximate the Bose–Einstein model with two other limiting formulae, the geometric distribution G and the model B_E.

2.1 The Bernoulli Model of Randomness: The Limiting Models P and D

We make the assumption that the tokens of a nonspecialty word should distribute over the N documents according to the binomial law. The probability of tf occurrences in a document is given by

$$Prob_1(tf) = B(N, F, tf) = \binom{F}{tf} p^{tf} q^{F-tf},$$

where $p = 1/N$ and $q = (N-1)/N$.

The expected relative frequency of the term in the collection is $\lambda = F/N$. Equation (5) is that used by Harter to define his Poisson model. The informative content of t in a document d is thus given by

$$Inf_1(tf) = -\log_2 \left[\binom{F}{tf} p^{tf} q^{F-tf}\right]. \tag{5}$$

The reader may notice that the document frequency n (the number of different documents containing the term) is not used in this model.

To compute the cumbersome formula (5) we approximate it assuming that p is small. We apply two limiting forms of Equation (5), each of them associated with an error. Unfortunately, in IR errors can make a nontrivial difference to the effectiveness of retrieval. IR deals with very small probabilities in everyday life equivalent to 0 and the probability from the elevator example given in the introduction would be very small but nontrivial in IR. We do not yet know to what extent errors may influence the effectiveness of the model.

The first approximation of the Bernoulli process is the Poisson process; the second is obtained by means of the information-theoretic divergence D.

Assuming that the probability p decreases towards 0 when N increases, but $\lambda = p \cdot F$ is constant, or moderate, a “good” approximation of Equation (5) is the basic model P and is given by

$$\begin{aligned} Inf_1(tf) &= -\log_2 B(N, F, tf) \\ &\sim -\log_2 \frac{e^{-\lambda}\lambda^{tf}}{tf!} \\ &\sim -tf \cdot \log_2 \lambda + \lambda \cdot \log_2 e + \log_2(tf!) \\ &\sim tf \cdot \log_2 \frac{tf}{\lambda} + \left(\lambda + \frac{1}{12 \cdot tf} - tf\right) \cdot \log_2 e + 0.5 \cdot \log_2(2\pi \cdot tf). \end{aligned} \tag{6}$$

The value λ is both the mean and the variance of the distribution. The basic model P of relation (6) is obtained through Stirling’s formula which approximates the factorial number as:

$$tf! = \sqrt{2\pi} \cdot tf^{tf+0.5} e^{-tf} e^{(12 \cdot tf + 1)^{-1}}. \tag{7}$$

For example, the approximation error for 100! is “only” 0.08%.

The formula (5) can alternatively be approximated as follows [Renyi 1969] (to obtain approximation (8) Renyi used Stirling’s formula),

$$B(N, F, tf) \sim \frac{2^{-F \cdot D(\phi, p)}}{(2\pi \cdot tf(1-\phi))^{1/2}},$$

where $\phi = tf/F$, $p = 1/N$, and $D(\phi, p) = \phi \cdot \log_2 \phi/p + (1 - \phi) \cdot \log_2((1-\phi)/(1-p))$ is called the *divergence* of ϕ from p.

The informative content gives the basic model D:

$$Inf_1(tf) \sim F \cdot D(\phi, p) + 0.5 \log_2(2\pi \cdot tf(1-\phi)). \tag{8}$$

We show in the following sections that the two approximations P and D are experimentally equivalent under the Probability Ranking Principle [Robertson 1986]. Hence, only a refinement of Stirling's formula or other types of normalizations can improve the effectiveness of the binomial model. We emphasize once again that, besides N which does not depend on the choice of the document and the term, the binomial model is based on both term frequency *tf* in the document and the total number F of term occurrences in the collection, but not on the term document frequency n.

2.2 The Bose–Einstein Model of Randomness

Now suppose that we randomly place the F tokens of a word in N documents. Once the random allocation of tokens to documents is completed, this event is completely described by its occupancy numbers: $tf_1, \ldots, tf_N$, where tf_k stands for the term frequency of the word in the kth document.

Every N-tuple satisfying the equation

$$tf_1 + \cdots + tf_N = F \tag{9}$$

is a possible configuration of the occupancy problem [Feller 1968]. The number s_1 of solutions of Equation (9) is given by the binomial coefficient:

$$s_1 = \binom{N+F-1}{F} = \frac{(N+F-1)!}{(N-1)!F!}. \tag{10}$$

Now, suppose that we observe the kth document and that the term frequency is *tf*. Then a random allocation of the remaining $F - tf$ tokens in the rest of the collection of $N-1$ documents is described by the same equation with $N-1$ terms:

$$tf_1 + \cdots + tf_{k-1} + tf_{k+1} + \cdots + tf_N = F - tf. \tag{11}$$

The number s_2 of solutions of Equation (11) is given by

$$s_2 = \binom{N-1+(F-tf)-1}{F-tf} = \frac{(N+F-tf-2)!}{(N-2)!(F-tf)!}. \tag{12}$$

In Bose–Einstein statistics the probability $Prob_1(tf)$ that an arbitrary document contains exactly *tf* occurrences of the term t is given by the ratio s_2/s_1. That is,

$$Prob_1(tf) = \frac{\binom{N-F-tf-2}{F-tf}}{\binom{N+F-1}{F}} = \frac{(N+F-tf-2)!F!(N-1)!}{(F-tf)!(N-2)!(N+F-1)!}. \tag{13}$$

Equation (13) reduces to

$$Prob_1(tf) = \frac{(F-tf+1)\cdot \ldots \cdot F \cdot (N-1)}{(N+F-tf-1)\cdot \ldots \cdot (N+F-1)}.$$

Note that both numerator and denominator are made up of a product of $tf + 1$ terms. Hence

$$Prob_1(tf) = \frac{\left(\frac{F}{N} - \frac{tf-1}{N}\right) \cdot \ldots \cdot \frac{F}{N} \cdot \left(1 - \frac{1}{N}\right)}{\left(1 + \frac{F}{N} - \frac{tf+1}{N}\right) \cdot \ldots \cdot \left(1 + \frac{F}{N} - \frac{1}{N}\right)}. \tag{14}$$

If we assume that $N \gg tf$, then $(tf - k)/N \sim 0$ and $(k+1)/N \sim 0$ for $k = 0, \ldots, tf$ and Equation (14) reduces to

$$\begin{aligned} Prob_1(tf) &\sim \frac{\frac{F}{N} \cdot \ldots \cdot \frac{F}{N} \cdot 1}{\left(1 + \frac{F}{N}\right) \cdot \ldots \cdot \left(1 + \frac{F}{N}\right)} = \frac{\left(\frac{F}{N}\right)^{tf}}{\left(1 + \frac{F}{N}\right)^{tf+1}} \\ &= \left(\frac{1}{1 + \frac{F}{N}}\right) \cdot \left(\frac{\frac{F}{N}}{1 + \frac{F}{N}}\right)^{tf}. \end{aligned} \tag{15}$$

Let $\lambda = F/N$ be the mean of the frequency of the term t in the collection D; then the probability that a term occurs tf times in a document is

$$Prob_1(tf) \sim \left(\frac{1}{1+\lambda}\right) \cdot \left(\frac{\lambda}{1+\lambda}\right)^{tf}. \tag{16}$$

The right-hand side of Equation (16) is known as the *geometric distribution* with probability $p = 1/(1+\lambda)$. The model of randomness based on (16) is called G, (G stands for Geometric):

$$\begin{aligned} Inf_1(tf) &= -\log_2 \left(\left(\frac{1}{1+\lambda}\right) \cdot \left(\frac{\lambda}{1+\lambda}\right)^{tf}\right) \\ &= -\log_2 \left(\frac{1}{1+\lambda}\right) - tf \cdot \log_2 \left(\frac{\lambda}{1+\lambda}\right). \end{aligned} \tag{17}$$

The geometric distribution is also used in the Ponte–Croft [1998] model. However, in their model λ is the mean frequency F/n of the term with respect to the number n of documents in which the term t occurs [Ponte and Croft 1998, p. 277] compared to $\lambda = F/N$ in ours. A second difference with respect to our Bose–Einstein model, is that in Ponte–Croft's model the geometric distribution $p(t)$ is used to define a correcting exponent $R_{t,d} = 1 - p(t)$ of the probability of the term in the document (compared to $-\log\ Prob_1$ in ours). Their exponent $R_{t,d}$ computes (according to their terminology) the risk of using the mean as a point estimate of a term t being drawn from a distribution modeling document d.

If instead we were to assume $\lambda = F/n$ as was done in the Ponte–Croft model, then Equation (16) could still be considered as a limiting form of the Bose–Einstein statistics, as in general tf is small with respect to n.

The second operational model associated with the Bose–Einstein statistics is constructed by approximating the factorials by Stirling's formula. Starting from formula (13) we obtain the model B_E:

$$\begin{aligned} \mathit{Inf}_1(\mathit{tf}) &= \log_2 \frac{(N+F-\mathit{tf}-2)!F!(N-1)}{(F-\mathit{tf})!(N+F-1)!} \\ &= -\log_2(N-1) - \log_2(e) \\ &\quad + f(N+F-1, N+F-\mathit{tf}-2) - f(F, F-\mathit{tf}) \end{aligned} \tag{18}$$

where

$$f(n,m) = (m+0.5) \cdot \log_2 \left(\frac{n}{m}\right) + (n-m) \cdot \log_2 n.$$

We show that in our experiments B_E and G are indistinguishable under different types of normalization. This confirms that G and B_E are good approximations of the Bose–Einstein statistics, since G and B_E have been derived by two completely different approximation hypotheses of the Bose–Einstein statistics.

2.3 The *tf-idf* and *tf-itf* Models of Randomness

The probability $\mathit{Prob}_1(\mathit{tf})$ is obtained by first computing the *unknown* probability p of choosing a document at random and then computing the probability of having *tf* occurrences of t in that document.

Using a Bayesian approach we assume that there is some true or a priori distribution (prior) of probabilities over documents with unknown probability p of occurrence. The Bayes Rule provides a way of calculating the a posteriori probability distribution. If $P(X = p|N)$ is the prior and n is the number of documents out of N containing the term, then we obtain the a posteriori probability

$$P(X = p|n, N) = \frac{P(X = p|N)P(n|N, p)}{\sum_p P(X = p|N)P(n|N, p)}.$$

Let the probability $P(n|p, N)$ of obtaining n out of N when $X = p$ be

$$\binom{N}{n} p^n q^{N-n}.$$

The a posteriori probability distribution depends heavily on the prior. However, if N is large, then the a posteriori probability, independently of the prior, condenses more and more around the maximum likelihood estimate (n/N) and also the relative document frequency n/N maximizes the a posteriori probability. The prior distribution becomes less and less important as the sample becomes larger and larger.

If the prior is a uniform prior probability independent of p, then the a posteriori probability according to Bayes' Rule is given by Laplace's so-called Law of Succession:

$$\frac{n+1}{N+2}.$$

If the prior is assumed to be of the beta form, that is, with the density function proportional to $p^{\alpha} q^{\beta}$ [Good 1968; van Rijsbergen 1977], where $\alpha, \beta > -1$, a posteriori probability according to Bayes' Rule is given by

$$\frac{n+1+\alpha}{N+2+\alpha+\beta}.$$

In the INQUERY system the parameter values α and β are set to -0.5. In the absence of evidence (i.e., when the collection is empty and $N = 0$), for $\alpha = \beta = -0.5$ the a posteriori probability p has the maximum uncertainty value 0.5. In this article we also set the values of α and β to -0.5. The probability of randomly choosing a document containing the term is thus

$$\frac{n+0.5}{N+1}.$$

We suppose that any token of the term is independent of all other tokens both of the same and different type, namely, the probability that a given document contains *tf* tokens of the given term is

$$Prob_1(tf) = \left(\frac{n+0.5}{N+1}\right)^{tf}.$$

Hence we obtain the basic model $I(n)$:

$$Inf_1(tf) = tf \cdot \log_2 \frac{N+1}{n+0.5}. \tag{19}$$

A different computation can be obtained from Bernoulli's law in Equation (5). Let n_e be the expected number of documents containing the term under the assumption that there are F tokens in the collection. Then

$$n_e = N \cdot Prob(tf \neq 0) = N \cdot (1 - B(N, F, 0)) = N \cdot \left(1 - \left(\frac{N-1}{N}\right)^F\right).$$

The third basic model is the tf-Expected_idf model $I(n_e)$:

$$Inf_1(tf) = tf \cdot \log_2 \frac{N+1}{n_e+0.5}. \tag{20}$$

Now, $1 - B(N, F, 0) \sim 1 - e^{-F/N}$ by the Poisson approximation of the binomial, and $1 - e^{-F/N} \sim F/N$ with an error of order $O((F/N)^2)$. By using this approximation, the probability of having one occurrence of a term in the document can be given by the term frequency in the collection assuming that F/N is small; namely,

$$Prob_1(tf) = \left(\frac{F}{N}\right)^{tf}.$$

Again from the term independence assumption, we obtain with a smoothing of the probability, the *tf-itf* basic model $I(F)$,

$$Inf_1(tf) = tf \cdot \log_2 \frac{N+1}{F+0.5} \quad \frac{F}{N} \text{ small or moderate }. \tag{21}$$

A generalization of the $I(F)$ was given by Kwok [1990] with the ICTF Weights (the Inverse Collection Term Frequency Weights), in the context of the standard probabilistic model using relevance feedback information [Robertson and

Sparck-Jones 1976]. Kwok reported that the ICTF performed much better than Salton's IDF model [Salton and Buckley 1988]. We show that in our experiments $I(F)$ and $I(n_e)$ behave similarly and independently with different types of normalization.

3. FIRST NORMALIZATION *N*1: RESIZING THE INFORMATIVE CONTENT BY THE *AFTEREFFECT* OF SAMPLING

Suppose that we are searching for tokens of a term and after a long unsuccessful search we find a few of them in a portion of a document. It is quite likely that we have finally reached a document in which we expect increased success in our search. The more we find, the higher is the expectation. This expectation is given by $Prob_2(tf)$ in formula (1), and has been called by statisticians an apparent *aftereffect* of future sampling [Feller 1968, pp. 118–125]. There are several models for the aftereffect: one of these is the law of succession of Laplace [Good 1968]. The intuition underlying the aftereffect in IR is that the greater the term frequency *tf* of a term in a document, the more the term is contributing to discriminating that document.

If *tf* is large then the probability that the term may select a relevant document is high. The fact that *tf* is large depends on the length of the document. Moreover, relevant documents may have different lengths and we cannot predict the size of a relevant document. Therefore we assume for the moment that the length of a relevant document is of arbitrary and large size. In Section 4 we show how to normalize the actual document length l to a given length. When enlarging the actual size of a relevant document to an arbitrary large size, the chance of encountering a new token of the observed term increases in accordance with the size *tf* of already observed tokens.

We thus assume that the probability that the observed term contributes to select a relevant document is high, if the probability of encountering one more token of the same term in a relevant document is similarly high. We reason that a high expectation of encountering one more occurrence is due to some underlying semantic cause and should not be simply accidental. The probability of a further success in encountering a term is thus a conditional probability that approaches 1 as *tf* increases and becomes large. On the contrary, if successes were brought about by pure chance, then the conditional probability would rather approach 0 as *tf* increases and becomes large. We need, however, a method to estimate this conditional probability.

We assume that the probability $Prob_2(tf)$ is related only to the "elite set" of the term, which is defined to be the set D_t of all documents containing the term. We also assume that the probability $Prob_2(tf)$ in formula (1) is obtained by a conditional probability $p(tf+1|tf,d)$ of having one more occurrence of t in the document d and that $p(tf+1|tf,d)$ is obtained by an aftereffect model.

This probability is computed in the next two sections.

3.1 The Normalization *L*

The first model of $Prob_2(tf)$ is given by Laplace's law of succession. The law of succession in this context is used when we have no advance knowledge of how many tokens of a term should occur in a relevant document of arbitrary large size. The Laplace model of aftereffect is explained in Feller [1968]. The probability $p(tf+1|tf, d)$ is close to $(tf+1)/(tf+2)$ and does not depend on the document length.

Laplace's law of succession is thus obtained by supposing the following.

The probability $Prob_2(tf)$ modeling the aftereffect in the elite set in formula (1) is given by the conditional probability of having one more token of the term in the document (i.e., passing from tf observed occurrences to $tf+1$) assuming that the length of a relevant document is very large.

$$Prob_2(tf) = \frac{tf+1}{tf+2}. \tag{22}$$

Similarly, if $tf \geq 1$ then $Prob_2(tf)$ can be given by the conditional probability of having tf occurrences assuming that $tf-1$ have been observed. Equation (22) with $tf-1$ instead of tf leads to the following equation,

$$Prob_2(tf) = \frac{tf}{tf+1}. \tag{23}$$

Equations (1) and (23) give the normalization L:

$$weight(t, d) = \frac{1}{tf+1} \cdot Inf_1(tf). \tag{24}$$

In our experiments, which we do not report here for the sake of space, relation (23) seems to perform better than relation (22), therefore we refer to formula (24) as the First Normalization L of the informative content.

3.2 The Normalization *B*

The *second* model of $Prob_2(tf)$ is slightly more complex than that given by relation (23). The conditional probability of Laplace's law computes directly the aftereffect on future sampling. The hypothesis about aftereffect is that any newly encountered token of a term in a document is not obtained by accident. If we admit that accident is not the cause of encountering new tokens then the probability of encountering a new token must increase. Hence, the aftereffect on the future sampling is obtained by a process whose probability of obtaining a newly encountered token is *inversely related* to that which would be obtained by accident. In other words, the aftereffect of sampling *in the elite set* yields a distribution that departs from one of the "ideal" schemes of randomness we described before. Therefore, we may model this process by Bernoulli. However, a sequence of Bernoulli trials is known to be a process characterized by a complete lack of memory (lack of aftereffect): previous successes or failures do not influence successive outcomes. The lack of memory does not allow us to use Bernoulli trials, as, for example, in the ideal urn model defined by Laplace, the conditional probability would be $p(tf+1|tf, d)$, and this conditional probability would be constant.

To obtain the estimate $Prob_2$ with Bernoulli trials we use the following urn model. We add a new token of the term to the collection, thus having $F+1$ tokens instead of F. We then compute the probability $B(n, F+1, tf+1)$ that this new token falls into the observed document, thus having a within document–term frequency $tf + 1$ instead tf. The process $B(n, F+1, tf+1)$ is thus that of obtaining one more token of the term t in the document d out of all n documents in which t occurs when a new token is added to the elite set. The comparison $(B(n, F+1, tf+1))/(B(n, F, tf))$ of the new probability $B(n, F+1, tf+1)$ with the previous one $B(n, F, tf)$ tells us whether the probability of encountering a new occurrence is increased or diminished by our random urn model.

Therefore, we may talk in this case of an *incremental rate* α of term occurrence in the elite set rather than of probability $Prob_2$ of term occurrence in the elite set, and we suppose that the incremental rate of occurrence is

$$\alpha = \frac{B(n, F, tf) - B(n, F+1, tf+1)}{B(n, F, tf)} = 1 - \frac{B(n, F+1, tf+1)}{B(n, F, tf)}, \qquad (25)$$

where $(B(n, F+1, tf+1))/(B(n, F, tf))$ is the ratio of two Bernoulli processes. If the ratio

$$\frac{B(n, F+1, tf+1)}{B(n, F, tf)}$$

is smaller than 1, then the probability of the document receiving at random the new added token increases. In conclusion, the larger tf, the less accidental one more occurrence of the term is, therefore the less risky it is to accept the term as a descriptor of a potentially relevant document. $(B(n, F+1, tf+1))/(B(n, F, tf))$ is a ratio of two binomials given by Equation (5) (but using the elite set with $p = 1/n$ instead of $p = 1/N$):

$$\alpha = 1 - \frac{B(n, F+1, tf+1)}{B(n, F, tf)} = 1 - \frac{F+1}{n \cdot (tf+1)}. \qquad (26)$$

The Equations (26) and (31) give

$$weight\,(t, d) = \frac{B(n, F+1, tf+1)}{B(n, F, tf)} \cdot Inf_1(tf) = \frac{F+1}{n \cdot (tf+1)} \cdot Inf_1(tf). \qquad (27)$$

The relation (27) is studied in the next sections and we discuss the results in the concluding sections.

3.3 Relating $Prob_2$ to $Prob_1$

In this subsection we provide a formal derivation of the relationship between the elite set and statistics of the whole collection; that is, we show how the two probabilities $Prob_2$ and $Prob_1$ are combined. We split the informative content of a term into a 2×2 contingency table built up of the events *accept/not accept*(-ing) the term as document descriptor, and *relevance/not relevance* of the document. Let us assume that a term t belongs to a query q. We assume that if the term t also occurs in a document then *we accept it as a descriptor* for a potentially relevant document (relevant to the query q). A gain and a loss are thus obtained by accepting the term query t as a descriptor of a potentially relevant document. The gain is the amount of information we really get from the fact that the document will be actually relevant. The gain is thus a fraction of $Inf_1(tf)$; what is not gained from $Inf_1(tf)$ is the loss in the case that the document will turn out to be not relevant. This translates into the equation

$$gain + loss = Inf_1(tf). \tag{28}$$

We weight the term by computing only the expected gain; namely,

$$weight\,(t, d) = gain.$$

The conditional probability $Prob_2(tf)$ of occurrence of the term t is related to the odds in the standard way (the higher its probability the smaller the gain):

$$Prob_2(tf) = \frac{loss}{gain + loss}. \tag{29}$$

From Equation (29) the *loss* is

$$loss = Prob_2(tf) \cdot Inf_1(tf). \tag{30}$$

For scoring documents we use only the gain, which from (28) and (30) is

$$\begin{aligned} weight\,(t, d) = gain \ &= \ Inf_1(tf) - loss \\ &= \ (1 - Prob_2(tf)) \cdot Inf_1(tf). \end{aligned} \tag{31}$$

As an example, let us consider the term "progress" which occurs 22,789 times in a collection containing 567,529 documents. Let us use the Poisson model P for computing the amount of information Inf_1 and use Laplace's law of succession to compute the loss and gain of accepting the term as a descriptor for a potentially relevant document. We have two cases: the term frequency in the document is equal to 0 or not. In the second case suppose $tf = 11$ as an example. We have the following contingency table.

	Accept ($tf = 11$)	Not Accept ($tf = 0$)	
Rel	$gain_1 = 6.9390$	$loss_0 = 0.04015$	
Not Rel	$loss_1 = 69.3904$	$gain_0 = 0$	
	$Inf_1 = 76.3295$	$Inf_0 = 0.04015$	

First we compute the amount of information $Inf_1 = 76.3295$ as given by formula (6) with $tf = 11$ and $1 - Prob_2(tf) = 1 - (10/11) = 0.0909$ from (23); then $gain_1$ is obtained by multiplying these two values. Similarly, $loss_1 = 0.9090 \cdot 76.3295 = 69.3904$.

When $tf = 0$ we reject the term; that is, the term is assumed to be not a descriptor of a potentially relevant document. In other words, by rejecting the term we have a gain when the term "progress" is not important for predicting the relevance of the document. According to Laplace's law of succession the gain is 0, and the loss is very small.

4. SECOND NORMALIZATION *N*2: RESIZING THE TERM FREQUENCY BY DOCUMENT LENGTH

Taking into account document length, the average document length has been shown to enhance the effectiveness of IR systems. Also, document length was shown to be dependent on relevance [Singhal et al. 1996]. According to the experimental results contained in Singhal et al. [1996] a good score function should retrieve documents of different lengths with their chance of being retrieved being similar to their likelihood of relevance. For example, the *BM* 25 matching function of Okapi:

$$\sum_{t\in Q}\frac{(k_1+1)tf}{(K+tf)}\cdot\frac{(k_3+1)\cdot qtf}{(k_3+qtf)}\log_2\frac{(r+0.5)(N-n-R+r+0.5)}{(R-r+0.5)(n-r+0.5)}, \tag{32}$$

where

—R is the number of documents known to be relevant to a specific topic;

—r is the number of relevant documents containing the term;

—qtf is the frequency of the term within the topic from which Q was derived;

—l and avg_l are, respectively, the document length and average document length;

—K is $k_1((1-b)+b(l/(avg\_l)))$;

—$k1$, b, and k_3 are parameters that depend on the nature of the queries and possibly on the database; and

—k_1 and b are set by default to 1.2 and 0.75, respectively; k_3 is often set to 1000 (effectively infinite). In TREC 4 [Robertson et al. 1996] k_1 was in the range $[1, 2]$ and b in the interval $[0.6, 0.75]$, respectively.

By using the default parameters above ($k_1 = 1.2$ and $b = 0.75$), the baseline unexpanded BM25 ranking function, namely, without any information about documents relevant to the specific query, ($R = r = 0$) is:

$$\sum_{t\in Q}\frac{2.2\cdot tf}{0.3+0.9\dfrac{l}{avg\_l}+tf}\cdot\frac{1001\cdot qtf}{1000+qtf}\log_2\frac{N-n+0.5}{n+0.5}. \tag{33}$$

The INQUERY ranking formula [Allan et al. 1996] uses the same normalization factor of the baseline unexpanded BM25 with $k_1 = 2$ and $b = 0.75$, and $qtf = 1$:

$$\frac{tf}{tf+0.5+1.5\dfrac{l}{avg\_l}}\cdot\frac{\log_2\dfrac{N+0.5}{n}}{\log_2(N+1)}. \tag{34}$$

Hence, the BM25 length normalization *tfn*,

$$tfn = T\cdot tf, \tag{35}$$

where T is

$$T = \frac{1}{tf+0.3+0.9\cdot\dfrac{l}{avg\_l}}, \tag{36}$$

is a simple but powerful and robust type of normalization of *tf*. The BM25 length normalization is related to Equations (24).

Indeed, $tf + 0.3 + 0.9\cdot(avg\_l/l) = tf + k_1$ when the document has the length $l = avg\_l$.

Table I. Models are Made Up of Three Components[a]

	Basic Models	Formula
P	Poisson approximation of the binomial model	(6)
D	Approximation of the binomial model with the divergence	(8)
G	Geometric as limiting form of Bose–Einstein	(17)
B_E	Limiting form of Bose–Einstein	(19)
$I(n_e)$	Mixture of Poisson and inverse document frequency	(20)
$I(n)$	Inverse document frequency	(19)
$I(F)$	Approximation of $I(n_e)$	(21)
	First Normalization	
L	Laplace's law of succession	(23)
B	Ratio of two Bernoulli processes	(26)
	Second (Length) Normalization	
$H1$	Uniform distribution of the term frequency	(41)
$H2$	The term frequency density is inversely related to the length	(42)

[a] For example, $B_E B2$ uses the limiting form of Bose–Einstein formula (19), normalized by the incremental rate of the Bernoulli process of formula (26) and whose within document–term frequency is normalized by formula (42).

The normalization factor T in Equation (36) generates models

$$weight(t, d) = T \cdot Inf_1(tf). \tag{37}$$

Moreover, if the basic model $Inf_1(tf)$ in (37) is $I(n)$ or $I(F)$ of relations (19) or (21) as shown in Table I, then from the normalization T we obtain the randomness model given in Equation (24) up to the parameter K_1. We provide a formal derivation of BM25 in Section 6.

Our next concern is to introduce an additional methodology that can normalize the random variables *tf* to a given length of the document. In other words, we would like to obtain the expected number of tokens of a term in a document and in the collection if the lengths of the documents in the collection were equal to a fixed value, for example, to their average length.

The probabilistic models of randomness are based on the term-independence assumption. When tokens of the same term occur densely within a portion of text we may detect term dependence by a monotonically decreasing function of the probability of obtaining this density by randomness. We can explain and measure formally how *im*probable that density is by chance, but this formal model does not give us any insight (as far as we know) into the distribution of document lengths. It is thus difficult to express how improbable or why we obtain a specific length of observed document. How to compare *tf* tokens in a document of length l_1 with *tf* tokens in a document of length l_2 is not yet derivable in our framework. Hence, for the moment we can only make some hypotheses on how to compare different term frequencies and test them. We use a Bayesian methodology to choose the hypothesis for term frequency correction which is best from the empirical point of view.

We make two initial assumptions on how to resize term frequencies according to the length of the documents and we evaluate them. This assumption is similar to the "verbosity hypothesis" of Robertson [Robertson and Walker 1994], which states that the distribution of term frequencies in a document of length l is a 2-Poisson with means $\lambda \cdot (l/avg\_l)$ and $\mu \cdot (l/avg\_l)$, where λ and μ are the original means related to the observed term as discussed in the Introduction and avg_l is the average length of documents.

We define a density function $\rho(l)$ of the term frequency, and then for each document d of length $l(d)$ we compute the term frequency on the same interval $[l(d), l(d)+\Delta l]$ of given length Δl as a normalized term frequency. Δl can be chosen as either the median or the mean avg_l of the distribution. The mean minimizes the mean squared error function $\sum_{i=1}^{N}(\Delta l - l(d))^2/N$, and the median minimizes the mean absolute error function $\sum_{i=1}^{N}(\Delta l - l(d))/N$. Experiments show that the normalization with $\Delta l = avg\_l$ is the most appropriate choice.

$H1$. The distribution of a term is uniform in the document. The term frequency density $\rho(l)$ is a constant ρ

$$\rho(l) = c \cdot \frac{tf}{l} = \rho, \tag{38}$$

where c is a constant.

$H2$. The term frequency density $\rho(l)$ is a decreasing function of the length l.

We made two assumptions $H1$ and $H2$ on the density $\rho(l)$ but other choices are equally possible. We think that this crucial research issue should be extensively studied and explored. According to hypothesis $H1$ the *normalized term frequency tfn* is

$$tfn = \int_{l(d)}^{l(d)+avg\_l} \rho(l)dl = \rho \cdot avg\_l = c \cdot tf \cdot \frac{avg\_l}{l(d)}, \tag{39}$$

whereas, according to the hypothesis $H2$,

$$tfn = \int_{l(d)}^{l(d)+avg\_l} \rho(l)dl = c \cdot tf \cdot \int_{l(d)}^{l(d)+avg\_l} \frac{dl}{l} = c \cdot tf \cdot \log_e \left(1 + \frac{avg\_l}{l(d)}\right). \tag{40}$$

To determine the value for the constant c we assume that if the effective length of the document coincides with the average length, that is, $l(d) = avg\_l$, then the normalized term frequency *tfn* is equal to *tf*. The constant c is 1 under the hypothesis $H1$ and $c = 1/\log_e 2 = \log_2 e$ under the hypothesis $H2$:

$$tfn = tf \cdot \frac{avg\_l}{l(d)} \tag{41}$$

$$tfn = tf \cdot \log_2 e \cdot \log_e \left(1 + \frac{avg\_l}{l(d)}\right) = tf \cdot \log_2 \left(1 + \frac{avg\_l}{l(d)}\right). \tag{42}$$

We substitute uniformly *tfn* of Equations (41) or (42) for *tf* in $weight(t,d)$ of Equations (24) and (27).

We are now ready to provide the retrieval score of each document of the collection with respect to a query. The query is assumed to be a set of independent terms. Term-independence translates into the additive property of *gain* of Equation (31) over the set of terms occurring both in the query and in the observed document. We obtain the final matching function of relevant documents under the hypothesis of the uniform substitution of *tfn* for *tf* and the hypothesis $H1$ or $H2$:

$$R(q,d) = \sum_{t \in q} weight\,(t,d) = \sum_{t \in q} qtf \cdot (1 - Prob_2(tfn)) \cdot Inf_1(tfn), \tag{43}$$

where *qtf* is the multiplicity of term-occurrence in the query.

5. NOTATIONS

The normalizing factor $N1$ of $Inf_1(tf)$ in Equation (24) is denoted L (for Laplace), and that in Equation (27) is denoted B (for Binomial). Models of IR are obtained from the basic models P, D, $I(n)$, $I(F)$, and $I(n_e)$, B_E and G applying either the first normalization $N1$ (L or B) and then the second normalization $N2$ (i.e., substituting in (27) *tfn* for *tf*). Models are represented by a sequence *XYZ* where X is one of the notations of the basic models, Y is one of the two first normalization factors, and Z is either 1 or 2 according the second normalization $H1$ or $H2$. For example, $PB1$ is the Poisson model P with the normalization factor $N1$ of (27) with the uniform substitution *tfn* for $tf(t,d)$ according to hypothesis $H1$, and $B_E L2$ is the Bose–Einstein model B_E in (19) with the first normalization factor $N1$ of (24) with the uniform substitution *tfn* for $tf(t,d)$ according to hypothesis $H2$.

6. A DERIVATION OF THE UNEXPANDED RANKING FORMULA BM25 AND OF THE INQUERY FORMULA

The normalization of the term frequency of the ranking formula BM25 can be derived by the normalization $L2$, and therefore both BM25 and INQUERY [Allan et al. 1996] formulae are strictly related to the model $I(n)L2$:

$$I(n)L2 : \frac{tfn}{tfn + k_1} \log_2 \frac{N+1}{n+0.5}, \tag{44}$$

where

$$tfn = tf \cdot \log_2 \left(1 + \frac{avg\_l}{l}\right) \quad \text{and} \quad k_1 = 1, 2.$$

Let $k_1 = 1$ and let us introduce the variable $x = l / avg\_l$. Then

$$\frac{tfn}{tfn+1} = \frac{tf}{tf + \dfrac{1}{\log_2(x+1) - \log_2 x}}.$$

Let us carry out the Taylor series expansion of the function

$$g(x) = \frac{1}{\log_2(x+1) - \log_2 x}$$

at the point $x = 1$. Its derivative is

$$g'(x) = \frac{\log_2 e \cdot g^2(x)}{x(x+1)}.$$

Table II. The Probability $\Phi(\beta)$ Is the Probability Computed by the Binomial Distribution that a Random Document Has Length $|(l/avg\_l) - 1| < 1$ in a Collection with Mean avg_l and Variance σ^2

						Documents :
Collection	TREC	*avg*_l	σ	$\beta = avg\_l/\sigma$	$\Phi(\beta)$	$\lvert(l/avg\_l) - 1\rvert < 1$
Disks 1,2	1,2,3	209.6	776.2	0.27	0.61	0.89
Disks 4,5	6	265.5	1149.4	0.23	0.59	0.91
Disks 4,5 (no CR)	7,8	246.5	707.2	0.35	0.64	0.90

From $g(1) = 1$ and $g'(1) = \log_2 e \cdot 0.5$ we obtain

$$\begin{aligned}\frac{tfn}{tfn+1} &= \frac{tf}{tf + 1 + \log_2 e \cdot 0.5 \cdot \left(\frac{l}{avg\_l} - 1\right) + O\left(\left(\frac{l}{avg\_l} - 1\right)^2\right)} \\ &= \frac{tf}{tf + 0.2786 + 0.7213 \cdot \frac{l}{avg\_l} + O\left(\left(\frac{l}{avg\_l} - 1\right)^2\right)}. \end{aligned} \tag{45}$$

The expansion of (45) in $tfn/(tfn + 1)$ with error $O((l/avg\_l - 1)^3)$ gives

$$\begin{aligned}&\frac{tf}{tf + 1 + \log_2 e \cdot 0.5 \cdot \left(\frac{l}{avg\_l} - 1\right) - \frac{1}{8}\log_2 e \cdot (3 - 2\log_2 e)\left(\frac{l}{avg\_l} - 1\right)^2} \\ &= \frac{tf}{tf + 0.2580 + 0.7627 \cdot \frac{l}{avg\_l} - 0.0207 \cdot \frac{l}{avg\_l}^{2}}.\end{aligned}$$

The INQUERY normalization factor of formula (34) is obtained with the parameter $k_1 = 2$ which corresponds to the application of Laplace's law of succession as stated in formula (22) (with coefficients 0.5572 and 1.4426 instead of 0.5 and 1.5).

The $O(((l/avg\_l) - 1)^2)$ in (45) is small when $|(l/avg\_l) - 1| < 1$. It is interesting to estimate the probability that the length l of a random document satisfies such a relation. By applying the Central Limit Theorem to the random variable l with mean avg_l and variance σ^2, the discrepancy $l - avg\_l < \sigma \cdot \beta$ for every fixed value β converges to the value $\Phi(\beta)$ given by the normal distribution Φ. If we set $\beta = avg\_l/\sigma$ the relation $|(l/avg\_l) - 1| < 1$ is satisfied. Thus the approximation (45) should hold when the standard deviation σ is close to the mean avg_l. In practice, the expected number of documents satisfying the constraint $|(l/avg\_l) - 1| < 1$, given by the Central Limit Theorem, is smaller than the actual number, as shown in Table II. The effectiveness of the approximation is also confirmed by our experiments, not reported here, that have shown that the $BM25$ formula with its parameters set as in formula (45) has the same performance of $I(n)L2$.

7. EXPERIMENTS

We used two test collections of TREC (Text REtrieval Conference). The first test collection is on disks 1 and 2; the second collection is on both disks 4 and 5. For the first test collection we used the topics of TREC-1 through TREC-3 (50 topics each), and for the second collection we used the topics of TREC-6 through TREC-8 (50 topics each).

Disks 1 and 2 for TREC-1 through TREC-3 experiments consist of about 2 Gbytes of data, of about 528,000 documents from the Department of Energy Abstracts, the Federal Register, the Associated Press Newswire, and the Ziff-Davis collections. Disks 1 and 2 contain (after the use of the stop list) 138,743,975 pointers (a pointer is the unit piece of information of the inverted file that contains the pair "term–document" information and the relative within document term frequency). We used the compression techniques of Witten et al. [1999] to represent the inverted file in a compressed format. The space required by the compressed inverted file for disks 1 and 2 is 96 Mbytes, that is, 11.4 bits per pointer. The average length of a document from disks 1 and 2 is 210 tokens (tokens from the stop list were not computed).

The TREC-6 test collection consists of about 2.1 Gbytes of data, of about 556,000 documents, from the Congressional Record, *Financial Register*, *Financial Times*, Foreign Broadcast Information Service, and *LA Times* collections. Differently from TREC-6, in TREC-7 and TREC-8, the collection CR (about 28,000 transcripts from the Congressional Record) was not indexed. Disks 4 and 5 contain 147,625,088 pointers. The space occupied by the compressed inverted file for disks 4 and 5 is 103 Mbytes; that is, the inverted file needs 11.2 bits per pointer. The average length of a document on disks 4 and 5 is 265 tokens. This average length decreases to 246 without indexing the CR collection. Indeed, the CR document length average is much longer than the document average length of other collections (624 tokens per document).

The text in the fields that was human-assigned was not indexed for use in the experiments.

Each of the 50 topics consists of three fields: a title (from one to three words), a description (one or two sentences), and a narrative (a paragraph listing specific criteria for accepting or rejecting a document). In our experiments we used all these three fields. We used Porter's stemming algorithm and a stop list of 235 words.

We tested the basic models with first and second normalization and compared them with model $BM25$ of Okapi as defined by formula (33). To find the noninterpolated average measure of precision (Chris Buckley proposed this measure which was first used in TREC-2 [Harman 1993]) for each query and for each ith retrieved relevant document the exact precision $Prob_i$ is first computed (i.e., i/r, where r is the document position in the rank); then the average precision for the query is obtained (i.e., $\sum_i Prob_i/R$, where R is the number of relevant documents in the collection) and finally one obtains the mean of the average precision over all topics. The noninterpolated average precision for the 11 levels of recall is shown in Tables IV through VII, X, and XI by AvgPr, the precision at 5, 10, 30, 100, and R (R-precision) retrieved documents,

Table III. Comparison of Models with TREC-10 Data[a]

Method	Official run	AvPrec	Prec-at-10	Prec-at-20	Prec-at-30
		Model Performance Without Query Expansion			
$B_E L2$		0.1788	0.3180	0.2730	0.2413
$I(n)L2$		0.1725	0.3180	0.2740	0.2353
$I(n_e)L2$	fub01ne	0.1790	0.3240	0.2720	0.2440
$B_E B2$		0.1881	0.3280	0.2980	0.2487
$I(n)B2$	fub01idf	0.1900	0.3360	0.2880	0.2580
$I(n_e)B2$		0.1902	0.3340	0.2860	0.2580
		Model Performance with Query Expansion			
$B_E L2$	fub01be2	0.2225	0.3440	0.2860	0.2513
$I(n)L2$		0.1973	0.3200	0.2730	0.2380
$I(n_e)L2$	fub01ne2	0.1962	0.3280	0.2760	0.2507
$B_E B2$		0.2152	0.3400	0.2870	0.2527
$I(n)B2$		0.2052	0.3380	0.2970	0.2680
$I(n_e)B2$		0.2041	0.3360	0.2990	0.2660

[a] The first normalization $L2$ is superior to $B2$ only if combined with model B_E and query expansion. Model B_E performs in general very well in combination with the query expansion technique.

where R is the number of relevant documents for each query, denoted by Pr5, Pr10, Pr30, Pr100, and R-Pr, respectively. We use l and avg_l as the length of a document and the average number of tokens in a document in the collection, respectively.

We submitted at TREC-10 four runs as shown in Table III to compare retrieval with or without query expansion.

Because of the size of the collection (10 Gbytes for about 1,600,000 Web documents), and as we had very limited storage capabilities, we reduced the size of the inverted files and performed some document and word pruning. Specifically, we indexed with single terms only, ignoring punctuation and case. The whole text was indexed except for HTML tags, which were removed from documents. *Pure single keyword indexing was performed, and link information was not used*. We removed 2897 documents with more than 10,000 words and 57,031 documents with less than 10 words. Also, we removed 86,146 documents containing more than 50% of unrecognized English words. In all, we removed 118,087 documents. Words contained in less than 11 documents, that were apparently exclusively misspelled words, were not included for the indexing. Words containing more than 3 consecutive equal characters or longer than 20 characters were also deleted. In this way, the number of distinct words in the collection was only 293,484. We used a very limited stop list and did not perform word stemming at all.

Word and document pruning together with the absence of stemming has obviously produced data characteristics largely different from those which would have been obtained had we followed the same indexing process as with the previous TREC data. As a consequence, we have introduced a parameter c in order to correct the resulting average length of the collection:

$$tfn = tf \cdot \log_2\left(1 + \frac{c \cdot avg\_l}{l}\right) \quad (with\ c = 7). \tag{46}$$

Table IV. The Best Precision Values Are in Bold. $I(n_e)B2$ and Its Approximation $I(F)B2$ Have the Best Average Precision and Precision at 5 Documents Retrieved. The Two Limiting Forms of Bose–Einstein Model, $GB2$ and B_EB2, Have Best Precision at 10. BM25 Has Best Precision for High Recall

Disks 1 and 2 of TREC 1, Topics 51–100. Relevant Documents: 16386							
Models	AvegPr	Pr5	Pr10	Pr30	Pr100	R-Pr	Rel Ret
$I(F)B1$	0.1989	0.6200	0.5660	0.4973	0.3886	0.2813	7128
$I(F)L1$	0.1933	0.5760	0.5760	0.4853	0.3814	0.2751	6993
$I(F)B2$	**0.2103**	0.6400	0.5740	**0.5333**	0.4038	0.2878	**7396**
$I(F)L2$	0.2068	0.6200	0.5700	0.5127	0.3978	0.2843	7300
$I(n)B1$	0.1911	0.6040	0.5740	0.5027	0.3798	0.2675	6928
$I(n)L1$	0.1968	0.5920	0.5600	0.5013	0.3908	0.2787	7034
$I(n)B2$	0.2003	0.6280	0.5900	0.5200	0.3964	0.2781	7123
$I(n)L2$	0.2077	0.6200	0.5800	0.5193	0.4030	0.2863	7267
$I(n_e)B1$	0.1985	0.6240	0.5660	0.4987	0.3882	0.2795	7109
$I(n_e)L1$	0.1946	0.5800	0.5420	0.4907	0.3856	0.2764	7006
$I(n_e)B2$	0.2098	**0.6440**	0.5860	0.5327	0.4054	0.2865	7395
$I(n_e)L2$	0.2073	0.6200	0.5720	0.5153	0.4004	0.2852	7307
$GB1$	0.1984	0.6120	0.5820	0.5093	0.3934	0.2782	7144
$GL1$	0.1968	0.5920	0.5560	0.4953	0.3878	0.2771	7093
$GB2$	0.2041	0.6320	**0.5980**	0.5193	0.3974	0.2816	7274
$GL2$	0.2047	0.6280	0.5660	0.5107	0.3952	0.2856	7232
B_EB1	0.1984	0.6120	0.5820	0.5093	0.3934	0.2782	7144
B_EL1	0.1968	0.5920	0.5560	0.4953	0.3878	0.2771	7093
B_EB2	0.2042	0.6320	**0.5980**	0.5193	0.3974	0.2816	7276
B_EL2	0.2047	0.6280	0.5660	0.5107	0.3952	0.2856	7232
$PB1$	0.1696	0.5360	0.5020	0.4587	0.3536	0.2517	6404
$PL1$	0.1741	0.5360	0.5300	0.4593	0.3562	0.2572	6442
$PB2$	0.2003	0.6000	0.5900	0.5127	0.3970	0.2755	7094
$PL2$	0.2065	0.6360	0.5780	0.5087	0.4056	0.2861	7124
$DB1$	0.1695	0.5360	0.5000	0.4587	0.3536	0.2513	6404
$DL1$	0.1741	0.5360	0.5300	0.4587	0.3562	0.2572	6442
$DB2$	0.2003	0.6000	0.5900	0.5127	0.3970	0.2755	7094
$DL2$	0.2065	0.6360	0.5780	0.5087	0.4056	0.2861	7124
$BM25$	0.2091	0.6240	0.5740	0.5260	**0.4080**	**0.2882**	7307

8. RESULTS

Our results show that all these models are robust with respect to different data sets. Notwithstanding the fact that we do not have parameters, models are shown to have a performance in most TREC experiments better than BM25 (TREC-10 included). In the following we discuss the results shown in Tables III through XI.

1. There is no convincing evidence or argument in favor of either normalization B or L. The results of TREC-7 (Table X) are confirmed on TREC-8 (Table XI) and similarly, the relative performance of the models in TREC-1

Table V. The Best Precision Values are in Bold. $I(n_e)L2$ and Its Approximation $I(F)L2$ Have the Best Average Precision and Precision at 5 Documents Retrieved. The Standard *idf-tf* Model with Laplace's Law of Succession $I(n)L2$ Has the Best Precision at 30. BM25 Has the Best Precision at High Recall Values and the Highest Precision at 10

Disks 1 and 2 of TREC 2, Topics 101–150. Relevant Documents: 11645							
Models	AvegPr	Pr5	Pr10	Pr30	Pr100	R-Pr	Rel Ret
$I(F)B1$	0.2320	0.5640	0.5180	0.4800	0.4090	0.3069	6356
$I(F)L1$	0.2333	0.5720	0.5420	0.4853	0.4026	0.3116	6322
$I(F)B2$	0.2413	0.5640	0.5440	0.4960	0.4134	0.3142	6464
$I(F)L2$	**0.2456**	0.5880	0.5540	0.5087	0.4160	0.3208	6497
$I(n)B1$	0.2225	0.5480	0.5160	0.4780	0.4028	0.3006	6261
$I(n)L1$	0.2364	0.5680	0.5440	0.5047	0.4130	0.3148	6380
$I(n)B2$	0.2262	0.5600	0.5200	0.4907	0.4086	0.3037	6258
$I(n)L2$	0.2439	0.5560	0.5420	**0.5147**	0.4224	0.3187	6472
$I(n_e)B1$	0.2325	0.5560	0.5260	0.4873	0.4110	0.3093	6410
$I(n_e)L1$	0.2348	0.5720	0.5460	0.4920	0.4050	0.3137	6349
$I(n_e)B2$	0.2406	0.5600	0.5420	0.4993	0.4154	0.3155	6483
$I(n_e)L2$	**0.2456**	**0.5960**	0.5540	0.5087	0.4176	0.3219	6503
$GB1$	0.2329	0.5440	0.5280	0.4833	0.4112	0.3094	6392
$GL1$	0.2379	0.5800	0.5540	0.4980	0.4074	0.3178	6392
$GB2$	0.2336	0.5400	0.5220	0.4947	0.4106	0.3089	6320
$GL2$	0.2417	0.5800	0.5440	0.5120	0.4142	0.3177	6391
B_EB1	0.2329	0.5440	0.5280	0.4833	0.4112	0.3094	6392
B_EL1	0.2379	0.5800	0.5540	0.4980	0.4074	0.3179	6392
B_EB2	0.2336	0.5400	0.5220	0.4947	0.4106	0.3089	6321
B_EL2	0.2418	0.5800	0.5440	0.5120	0.4144	0.3181	6391
$PB1$	0.1951	0.5280	0.5060	0.4667	0.3772	0.2780	5769
$PL1$	0.2089	0.5640	0.5260	0.4700	0.3836	0.2892	5924
$PB2$	0.2223	0.5760	0.5420	0.4940	0.4144	0.3039	6232
$PL2$	0.2383	0.5880	0.5540	0.5000	0.4194	0.3223	6402
$DB1$	0.1951	0.5280	0.5060	0.4660	0.3772	0.2776	5769
$DL1$	0.2089	0.5640	0.5260	0.4693	0.3836	0.2892	5924
$DB2$	0.2223	0.5760	0.5420	0.4940	0.4144	0.3039	6232
$DL2$	0.2383	0.5880	0.5540	0.5000	0.4196	0.3223	6403
$BM25$	0.2455	0.5720	**0.5560**	0.5087	**0.4252**	**0.3230**	**6523**

through TREC-3 (see Tables IV, V, VI) shows similar trends. In TREC-1 through TREC-3, $L2$ is in general superior to $B2$ independently of the basic model used, whereas in TREC-7, TREC-8, and TREC-10 (see Tables X, XI, III), $B2$ is in general superior to $L2$ independently of the basic model used. The notable exception is the Poisson model P: $L1$ and $L2$ perform in general better than $B2$.

It is interesting to observe that results of TREC-6 (Table VII) (whose test bed uses the additional collection CR containing long documents) are significantly different from all other TREC experiments. This allows us to conjecture but not to assert that the statistics of the collection (e.g., number of unique terms, mean and variance of document length) may have more effect on the relative

Table VI. The Best Precision Values Are in Bold. $I(n_e)L2$ and Its Approximation $I(F)L2$ Have the Best Average Precision. The Two Approximations of the Bernoulli Model, $PL2$ and $DL2$, Have the Highest Precision at 5 Documents Retrieved. The Standard *idf-tf* Model with Laplace's Law of Succession $I(n)L2$ Has the Best Precision at 30. BM25 Has the Best Precision at High Recall Values and the Highest Precision at 10

Disks 1 and 2 of TREC 3, Topics 151–200. Relevant Documents: 9805							
Models	AvegPr	Pr5	Pr10	Pr30	Pr100	R-Pr	Rel Ret
$I(F)B1$	0.2565	0.6960	0.6520	0.5320	0.3776	0.3217	5437
$I(F)L1$	0.2675	0.6960	0.6560	0.5367	0.3832	0.3336	5460
$I(F)B2$	0.2644	0.7160	0.6620	0.5380	0.3846	0.3254	5516
$I(F)L2$	0.2765	0.7440	0.6660	0.5540	0.3902	**0.3390**	5524
$I(n)B1$	0.2439	0.6800	0.6400	0.5193	0.3694	0.3100	5320
$I(n)L1$	0.2669	0.7080	0.6740	0.5367	0.3870	0.3329	5535
$I(n)B2$	0.2480	0.7000	0.6540	0.5307	0.3714	0.3114	5315
$I(n)L2$	0.2716	0.7280	0.6720	0.5500	0.3926	0.3325	5524
$I(n_e)B1$	0.2569	0.7000	0.6540	0.5313	0.3820	0.3223	5454
$I(n_e)L1$	0.2682	0.6880	0.6580	0.5420	0.3826	0.3348	5483
$I(n_e)B2$	0.2637	0.7080	0.6680	0.5400	0.3848	0.3258	5514
$I(n_e)L2$	**0.2767**	0.7320	0.6720	0.5533	0.3906	0.3379	5543
$GB1$	0.2548	0.6880	0.6580	0.5227	0.3746	0.3182	5436
$GL1$	0.2681	0.6960	0.6800	0.5393	0.3842	0.3343	5495
$GB2$	0.2527	0.7040	0.6520	0.5260	0.3750	0.3165	5373
$GL2$	0.2682	0.7120	0.6680	0.5447	0.3818	0.3303	5446
B_EB1	0.2548	0.6920	0.6580	0.5220	0.3746	0.3182	5436
B_EL1	0.2681	0.6960	0.6780	0.5393	0.3840	0.3343	5495
B_EB2	0.2527	0.7040	0.6520	0.5260	0.3750	0.3165	5373
B_EL2	0.2683	0.7120	0.6680	0.5447	0.3820	0.3303	5446
$PB1$	0.2107	0.5800	0.5400	0.4667	0.3330	0.2821	4990
$PL1$	0.2314	0.6280	0.5800	0.4873	0.3466	0.3056	5092
$PB2$	0.2459	0.7120	0.6660	0.5267	0.3744	0.3093	5336
$PL2$	0.2705	**0.7520**	0.6780	0.5573	0.3934	0.3274	5490
$DB1$	0.2107	0.5800	0.5400	0.4667	0.3330	0.2821	4990
$DL1$	0.2314	0.6280	0.5800	0.4873	0.3466	0.3056	5092
$DB2$	0.2459	0.7120	0.6660	0.5273	0.3744	0.3093	5336
$DL2$	0.2706	**0.7520**	0.6780	0.5573	0.3934	0.3274	5490
$BM25$	0.2754	0.7320	**0.6840**	**0.5587**	**0.3960**	0.3352	**5586**

performance of models than the content of the submitted topics. However, we tried a small experiment which begins to corroborate such an hypothesis. We used the topics of TREC-6 on the collection used in TREC-7 and TREC-8 (without indexing the collection CR). In order to compare the two Tables VII and VIII we considered the means of different precision values and of the number of retrieved documents in Table VIII and computed the variation rates with respect to the values of Table VII and then normalized to the mean values. Results show that the normalization B increases average precision and more significantly the early precision, whereas L slightly increases the precision for

Table VII. The Best Precision Values Are in Bold. $I(n_e)L2$ and Its Approximation $I(F)L2$ Have the Best Average Precision. The Standard *tf-idf* Model with Laplace's Law of Succession $I(n)L2$ Has the Highest Precision at 5 Documents Retrieved. $I(n_e)B1$, Namely, the *idf* and Poisson Mixture Model Together with the Uniform Distribution Hypothesis on Term Frequency $H1$ and the Bernoulli Normalization B, Has the Best Performance at Higher Recall Values

Disks 4 and 5 of TREC 6, Topics 301–350. Relevant Documents: 4611							
Models	AvegPr	Pr5	Pr10	Pr30	Pr100	R-Pr	Rel Ret
$I(F)B1$	0.2457	0.5160	0.4580	0.3427	0.2162	0.2885	2667
$I(F)L1$	0.2557	0.5400	0.4420	0.3293	0.2074	0.2979	2640
$I(F)B2$	0.2482	0.5240	0.4840	0.3367	0.2092	0.2863	2651
$I(F)L2$	0.2597	0.5400	0.4600	0.3267	0.2058	0.2962	2595
$I(n)B1$	0.2381	0.5280	0.4620	0.3413	0.2144	0.2794	2607
$I(n)L1$	0.2560	0.5520	0.4480	0.3327	0.2090	0.3017	2654
$I(n)B2$	0.2362	0.5440	0.4640	0.3327	0.2062	0.2730	2546
$I(n)L2$	0.2544	**0.5760**	0.4840	0.3333	0.2126	0.2887	2594
$I(n_e)B1$	0.2479	0.5280	0.4640	**0.3487**	**0.2182**	0.2940	**2689**
$I(n_e)L1$	0.2557	0.5560	0.4700	0.3427	0.2164	0.2950	2654
$I(n_e)B2$	0.2488	0.5480	**0.4860**	0.3393	0.2112	0.2855	2638
$I(n_e)L2$	**0.2600**	0.5480	0.4620	0.3313	0.2086	0.2931	2595
$GB1$	0.2458	0.5480	0.4700	0.3473	0.2124	0.2883	2653
$GL1$	0.2567	0.5400	0.4620	0.3367	0.2116	**0.3051**	2623
$GB2$	0.2414	0.5320	0.4720	0.3333	0.2058	0.2797	2566
$GL2$	0.2548	0.5400	0.4560	0.3253	0.2074	0.2879	2538
$B_E B1$	0.2452	0.5480	0.4680	0.3467	0.2120	0.2878	2652
$B_E L1$	0.2562	0.5400	0.4620	0.3353	0.2114	0.3045	2622
$B_E B2$	0.2410	0.5320	0.4720	0.3327	0.2058	0.2791	2565
$B_E L2$	0.2546	0.5400	0.4560	0.3253	0.2072	0.2879	2537
$PB1$	0.2032	0.4600	0.4140	0.3100	0.1878	0.2445	2307
$PL1$	0.2243	0.4760	0.4260	0.3247	0.2000	0.2642	2452
$PB2$	0.2183	0.5040	0.4440	0.3113	0.1870	0.2509	2373
$PL2$	0.2424	0.5320	0.4560	0.3300	0.2010	0.2778	2497
$DB1$	0.2027	0.4600	0.4120	0.3100	0.1878	0.2440	2306
$DL1$	0.2238	0.4760	0.4260	0.3240	0.1998	0.2636	2451
$DB2$	0.2178	0.5040	0.4440	0.3107	0.1868	0.2503	2372
$DL2$	0.2421	0.5320	0.4560	0.3300	0.2008	0.2778	2496
$BM25$	0.2440	0.5600	0.4700	0.3233	0.2032	0.2834	2511

high values of recall (R-precision included). Model G is the most sensitive to the effect of the normalization process.

2. The Poisson model $PL2$ has a good performance for early precision early in the ranking (precision at five documents retrieved). As for the average precision, the performance is good in TREC-1 through TREC-3 (see Tables IV through VI), less satisfactory in TREC-6 and TREC-7 (see Tables VII, X), and unsatisfactory in TREC-8 (Table XI) (but in TREC-8 $PL2$ has the best performance for precision at five documents retrieved). Instead, the normalization $B2$ seems to work poorly with P.

Table VIII. The Best Precision Values Are in Bold. Removing Long Documents from the Collection Has Positive Effects on the Approximation G of the Bose–Einstein Model and on the Term Frequency Normalization B

Disks 4 and 5 Without CR Collection, Topics 301-350 of TREC 6. Rel. Doc.: 4290							
Models	AvegPr	Pr5	Pr10	Pr30	Pr100	R-Pr	Rel Ret
$I(n)B1$	0.2550	0.5240	0.4600	**0.3420**	0.2130	0.2906	2535
$I(n)L1$	0.2689	0.5320	0.4540	0.3380	0.2112	0.3089	2568
$I(n)B2$	0.2581	0.5560	0.4680	0.3320	0.2036	0.2866	2470
$I(n)L2$	0.2705	0.5560	**0.4840**	0.3267	0.2088	0.3004	2510
$I(n_e)B1$	0.2648	0.5400	0.4480	0.3393	**0.2176**	0.3025	**2615**
$I(n_e)L1$	0.2711	0.5320	0.4500	0.3320	0.2058	0.3154	2545
$I(n_e)B2$	0.2662	**0.5680**	0.4680	0.3373	0.2100	0.2991	2566
$I(n_e)L2$	**0.2751**	0.5440	0.4620	0.3213	0.2044	0.3129	2493
$GB1$	0.2615	0.5400	0.4500	0.3407	0.2118	0.2997	2576
$GL1$	0.2714	0.5400	0.4540	0.3327	0.2070	**0.3169**	2527
$GB2$	0.2605	0.5560	0.4740	0.3340	0.2038	0.2893	2502
$GL2$	0.2707	0.5440	0.4540	0.3247	0.2028	0.3018	2444
$PB1$	0.2170	0.4640	0.4060	0.3073	0.1842	0.2566	2271
$PL1$	0.2373	0.4600	0.4220	0.3187	0.1960	0.2750	2373
$PB2$	0.2338	0.5160	0.4400	0.3073	0.1868	0.2653	2318
$PL2$	0.2569	0.5160	0.4480	0.3213	0.1972	0.2882	2417
$BM25$	0.2584	0.5200	0.4560	0.3167	0.1978	0.2943	2420

Table IX. Best Performing Models for Each Test Collection and for Different Precision Measures. The Basic Probability Models $I(F)$, D, and B_E Are Not Considered Here, as They Do Not Differ Significantly from Their Alternative Approximations $I(n_e)$, P, and G, Respectively

TREC	AvegPr	Pr5	Pr10	Pr30	Pr100	R-Pr	Rel Ret
1	$I(n_e)B2$	$I(n_e)B2$	$GB2$	$I(n_e)B2$	$BM25$	$BM25$	$I(n_e)B2$
2	$I(n_e)L2$	$I(n_e)L2$	$BM25$	$I(n)L2$	$BM25$	$BM25$	$BM25$
3	$I(n_e)L2$	$PL2$	$BM25$	$BM25$	$BM25$	$I(n_e)L2$	$BM25$
6	$I(n_e)L2$	$I(n)L2$	$I(n_e)B2$	$I(n_e)B1$	$I(n_e)B1$	$GL1$	$I(n_e)B1$
7	$I(n_e)B2$	$I(n_e)B2$	$I(n_e)B2$	$I(n_e)B2$	$GB1$	$I(n_e)B2$	$I(n_e)B2$
8	$I(n_e)B2$	$PL2$	$I(n_e)B2$	$I(n_e)B2$	$I(n_e)B2$	$I(n_e)B2$	$I(n_e)B2$

3. Model G with both normalizations $B2$ and $L2$ has a good performance in all TREC experiments. G's performance depends on the choice of the normalization $B2$ (better in TREC-7 and TREC-8; see Tables X and XI) and $L2$ (better in TREC-1 through TREC-3, TREC-6, and TREC-10; see Tables IV through VII and III). Surprisingly, our experiments with TREC-10 show that B_EL2 is the model which best combines with the query expansion technique. Indeed, B_EL2 with query expansion was the best performing run at TREC-10.

4. The model $I(n_e)$ works well with both normalizations $B2$ and $L2$. We observe also that $I(n_e)$ performance depends on the choice of the normalization $B2$ which is better in TREC-1, TREC-7, TREC-8, and TREC-10 (see Tables IV, X, XI, and III) or $L2$ which is better in TREC-2, TREC-3, and TREC-6 (see Tables V through VII).

Table X. The Best Precision Values Are in Bold. $I(n_e)L2$ and Its Approximation $I(F)L2$ Have the Highest Precision at Different Recall Levels

Disks 4 and 5 of TREC 7, Topics 351–400. Relevant Documents: 4674							
Models	AvegPr	Pr5	Pr10	Pr30	Pr100	R-Pr	Rel Ret
$I(F)B1$	0.2352	0.5720	0.4960	0.3700	0.2370	0.2785	2876
$I(F)L1$	0.2180	0.5320	0.4780	0.3553	0.2170	0.2586	2777
$I(F)B2$	**0.2484**	**0.5800**	**0.5200**	**0.3813**	0.2374	0.2869	**2883**
$I(F)L2$	0.2312	0.5400	0.5000	0.3647	0.2158	0.2711	2796
$I(n)B1$	0.2191	0.5240	0.4720	0.3413	0.2116	0.2625	2531
$I(n)L1$	0.2225	0.5520	0.4920	0.3620	0.2230	0.2659	2828
$I(n)B2$	0.2337	0.5520	0.4840	0.3467	0.2164	0.2700	2540
$I(n)L2$	0.2360	0.5400	0.4960	0.3687	0.2278	0.2763	2845
$I(n_e)B1$	0.2352	0.5680	0.4960	0.3700	0.2382	0.2778	2861
$I(n_e)L1$	0.2184	0.5440	0.4760	0.3553	0.2176	0.2601	2782
$I(n_e)B2$	0.2482	**0.5800**	0.5100	**0.3813**	0.2386	**0.2874**	2881
$I(n_e)L2$	0.2320	0.5400	0.4980	0.3613	0.2174	0.2717	2810
$GB1$	0.2364	0.5720	0.5000	0.3760	**0.2390**	0.2787	2859
$GL1$	0.2196	0.5360	0.4720	0.3527	0.2166	0.2640	2770
$GB2$	0.2463	0.5720	0.5100	0.3753	0.2350	0.2847	2858
$GL2$	0.2315	0.5520	0.4880	0.3587	0.2174	0.2713	2780
B_EB1	0.2361	0.5720	0.5000	0.3760	**0.2390**	0.2787	2859
B_EL1	0.2196	0.5360	0.4720	0.3527	0.2166	0.2640	2770
B_EB2	0.2462	0.5720	0.5100	0.3753	0.2350	0.2847	2858
B_EL2	0.2315	0.5520	0.4880	0.3580	0.2174	0.2713	2780
$PB1$	0.1914	0.4840	0.4300	0.3407	0.2126	0.2434	2526
$PL1$	0.1944	0.4640	0.4480	0.3440	0.2092	0.2465	2584
$PB2$	0.2194	0.5200	0.5020	0.3533	0.2208	0.2624	2669
$PL2$	0.2212	0.5120	0.4880	0.3607	0.2194	0.2634	2743
$DB1$	0.1914	0.4840	0.4300	0.3407	0.2126	0.2434	2526
$DL1$	0.1944	0.4640	0.4480	0.3440	0.2092	0.2465	2584
$DB2$	0.2194	0.5200	0.5020	0.3533	0.2206	0.2624	2669
$DL2$	0.2212	0.5120	0.4880	0.3607	0.2194	0.2634	2743
$BM25$	0.2274	0.5320	0.4880	0.3540	0.2152	0.2643	2676

5. The model $I(n)$ works similarly to $I(n_e)$ but always performs slightly less well than $I(n_e)$.

6. By comparing the results from the models that are approximations or limiting forms of one theoretical basic model, we may observe that they are indistinguishable. We do not need to distinguish between the models P and D for the binomial basic model nor between the models G and B_E for the Bose–Einstein basic model. Similarly, we may observe that $I(F)$ and $I(n_e)$ do not differ significantly in the experiments. Since $I(F)$ can be considered as an approximation of $I(n_e)$, the experiments show that we may reduce the seven basic models (P, D, G, B_E, $I(n_e)$, $I(F)$, and $I(n)$) to four: P, G, $I(n_e)$, and $I(n)$.

7. The term frequency normalization $H2$ of formula (42) seems to be superior to the term frequency normalization $H1$ of formula (41). Indeed, given any

Table XI. The Best Precision Values Are in Bold. Similarly to TREC-7, $I(n_e)L2$ and its Approximation $I(F)L2$ Have the Highest Precision at Different Recall Levels, Except for the Poisson Model $PL2$ Which has the Highest Precision at 5

Disks 4 and 5 of TREC 8, Topics 401–450. Relevant Documents: 4728							
Models	AvegPr	Pr5	Pr10	Pr30	Pr100	R-Pr	Rel Ret
$I(F)B1$	0.2734	0.5400	0.4820	0.3820	0.2496	0.3135	3135
$I(F)L1$	0.2645	0.5280	0.4860	0.3700	0.2416	0.3103	3067
$I(F)B2$	0.2833	0.5520	0.5060	**0.3967**	0.2528	0.3280	**3189**
$I(F)L2$	0.2767	0.5240	0.4860	0.3840	0.2448	0.3179	3095
$I(n)L1$	0.2681	0.5120	0.5000	0.3787	0.2444	0.3164	3046
$I(n)B1$	0.2664	0.5240	0.4740	0.3880	0.2524	0.3221	3000
$I(n)B2$	0.2763	0.5520	0.4980	0.3900	0.2528	0.3235	3038
$I(n)L2$	0.2792	0.5360	0.5040	0.3927	0.2492	0.3233	3073
$I(n_e)B1$	0.2735	0.5320	0.4960	0.3807	0.2504	0.3286	3142
$I(n_e)L1$	0.2664	0.5240	0.4840	0.3707	0.2420	0.3114	3061
$I(n_e)B2$	**0.2841**	0.5520	**0.5080**	**0.3967**	**0.2532**	**0.3295**	3178
$I(n_e)L2$	0.2769	0.5200	0.4940	0.3887	0.2452	0.3171	3067
$GB1$	0.2757	0.5360	0.4800	0.3880	0.2494	0.3292	3142
$GL1$	0.2667	0.5120	0.4840	0.3727	0.2416	0.3146	3031
$GB2$	0.2826	0.5440	0.5040	0.3960	0.2514	0.3290	3153
$GL2$	0.2757	0.5280	0.4860	0.3887	0.2438	0.3183	3032
B_EB1	0.2757	0.5400	0.4800	0.3880	0.2494	0.3292	3142
B_EL1	0.2669	0.5120	0.4860	0.3727	0.2416	0.3146	3031
B_EB2	0.2827	0.5440	0.5040	0.3960	0.2514	0.3290	3153
B_EL2	0.2758	0.5280	0.4880	0.3887	0.2438	0.3183	3032
$PB1$	0.2379	0.5240	0.4800	0.3520	0.2246	0.2905	2838
$PL1$	0.2350	0.5120	0.4700	0.3553	0.2232	0.2898	2829
$PB2$	0.2559	0.5560	0.4980	0.3847	0.2360	0.3060	2948
$PL2$	0.2562	**0.5680**	0.4880	0.3780	0.2374	0.3044	2923
$DB1$	0.2379	0.5240	0.4800	0.3520	0.2246	0.2905	2839
$DL1$	0.2350	0.5120	0.4700	0.3553	0.2232	0.2898	2829
$DB2$	0.2559	0.5560	0.4980	0.3840	0.2358	0.3060	2948
$DL2$	0.2562	0.5680	0.4880	0.3780	0.2374	0.3044	2923
$BM25$	0.2716	0.5400	0.4980	0.3827	0.2464	0.3181	3083

model $X \in \{P, G, I(n), I(n_e)\}$ and any normalization $Y \in \{L, B\}$, the model $XY2$ performs better than its analogous $XY1$. There are some exceptions, especially in the experiment of TREC–6 for high values of recall (Pr30, Pr100, R-Pr, and for the relevant number retrieved) as shown by Tables IX and VII.

9. CONCLUSIONS

We create a framework for generating nonparametric information retrieval models. We construct a weighting formula that is a combination of three different probabilities. The first and basic probability models are obtained from urn models with random drawings. We compute a second probability, the probability of relevance of a term in its "elite set." This provides a normalization factor on the weighting formula. Finally, a probability related to the length of a document is constructed to resize the cardinality of the term frequency in the document. Two initial hypotheses about the distribution of document length are tested.

We use the basic probability models to derive for IR, a Bernoulli model, the *tf-idf* model $I(n)$, the *tf-itf* model $I(F)$, and the model $I(n_e)$ which is a mixture of the Poisson and the *idf* models. Two workable approximations of Bernoulli's model are introduced: the Poisson model P and the information-theoretic approximation model D. These two approximation models perform equally under all normalizations ($L1$, $B1$, $L2$, and $B2$).

The other basic model is Bose–Einstein. Two approximations of the Bose–Einstein model are also introduced: the geometric model G and B_E. These two approximation models perform equally under all normalizations ($L1$, $B1$, $L2$, and $B2$).

All models are compared to the BM25 formula, which is frequently used by many participants of TREC. $I(n_e)B2$ and $I(n_e)L2$ are shown to be superior at many recall levels and in average precision. Experiments show that the model $I(n_e)$ and $I(F)$ perform similarly. $I(n_e)$ is shown to perform better than the standard idf model $I(n)$ under all normalizations. We conclude that the document frequency can be replaced by the term frequency in the collection in general in any weighting formula.

$B2$, $L2$ are shown to be universal normalization factors, in the sense that the normalization works independently of models and independently of variation in document length. $L2$ is less sensitive to the variation of document length. On the other hand, when the variation is moderate $B2$ seems to perform better. The normalization factor $B2$, containing both the document frequency and the term frequency, derives formally from the Bernoulli process and from the standard axioms of utility theory.

Our models are formally derived and they do not contain parameters that must be learned from the actual data.

Future work will investigate the relative strengths and weaknesses of each model with query expansion. Moreover, further experiments should be performed to assess the effect on performance of word stemming, document pruning, and word pruning and to include these factors as explicit variables within the framework.

ACKNOWLEDGMENTS

We thank Juliet van Rijsbergen for helping to improve our English grammar.

REFERENCES

ALLAN, J., CALLAN, J. P., CROFT, W. B., BALLESTEROS, L., BROGLIO, J., XU, J., AND SHU, H. 1996. INQUERY at TREC-5. In *Proceedings of the Fifth Text REtrieval Conference (TREC-5)*. NIST Special Publication 500-238, Gaithersburg, Md., 119–132.

AMATI, G., CARPINETO, C., AND ROMANO, G. 2001. FUB at TREC 10 web track: A probabilistic framework for topic relevance term weighting. In *Proceedings of the Tenth Text Retrieval Conference (TREC-10)*. NIST Special Publication 500-250, Gaithersburg, Md.

Bookstein, A. and Swanson, D. 1974. Probabilistic models for automatic indexing. *J. Am. Soc. Inf. Sci. 25*, 312–318.

Carpineto, C. and Romano, G. 2000. Trec-8 automatic ad-hoc experiments at fub. In *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*. NIST Special Publication 500-246, Gaithersburg, Md., 377–380.

Cooper, W. and Maron, M. 1978. Foundations of probabilistic and utility-theoretic indexing. *J. ACM 25*, 67–80.

Cox, R. T. 1961. *The Algebra of Probable Inference*. Johns Hopkins Press, Baltimore, Md.

Croft, W. and Harper, D. 1979. Using probabilistic models of document retrieval without relevance information. *J. Doc. 35*, 285–295.

Damerau, F. 1965. An experiment in automatic indexing. *Am. Doc. 16*, 283–289.

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

Fuhr, N. 1989. Models for retrieval with probabilistic indexing. *Inf. Process. Manage. 25*, 1, 55–72.

Good, I. J. 1968. *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*, Vol. 30. MIT Press, Cambridge, Mass.

Harman, D. 1993. Overview of the Second Text REtrieval Conference (TREC–2). In *Proceedings of the TREC Conference*. NIST Special publication 500-215, Gaithersburg, Md, 1–20.

Harter, S. P. 1974. A probabilistic approach to automatic keyword indexing. PhD Thesis, Graduate Library, The University of Chicago, Thesis No. T25146.

Harter, S. P. 1975a. A probabilistic approach to automatic keyword indexing. Part I: On the distribution of specialty words in a technical literature. *J. ASIS 26*, 197–216.

Harter, S. P. 1975b. A probabilistic approach to automatic keyword indexing. Part II: An algorithm for probabilistic indexing. *J. ASIS 26*, 280–289.

Hiemstra, D. and de Vries, A. 2000. Relating the new language models of information retrieval to the traditional retrieval models. Res. Rep. TR–CTIT–00–09, Centre for Telematics and Information Technology.

Hintikka, J. 1970. On semantic information. In *Information and Inference*, J. Hintikka, and P. Suppes, Eds., Synthese Library. D. Reidel, Dordrecht, Holland, 3–27.

Kwok, K. 1990. Experiments with component theory of probabilistic information retrieval based on single terms as document components. *ACM Trans. Inf. Syst. 8*, 4, 363–386.

Lafferty, J. and Zhai, C. 2001. Document language models, query models, and risk minimization for information retrieval. In *Proceedings of ACM SIGIR* (New Orleans), ACM, New York, 111–119.

Margulis, E. 1992. N-Poisson document modelling. In *Proceedings of ACM–SIGIR 92 Conference* (Denmark), ACM, New York, 177–189.

Ponte, J. and Croft, B. 1998. A language modeling approach in information retrieval. In *Proceedings of the 21st ACM SIGIR Conference on Research and Development in Information Retrieval*, (Melbourne, Australia), B. Croft, A. Moffat, and C. van Rijsbergen, Eds., ACM, New York, 275–281.

Popper, K. 1995. *The Logic of Scientific Discovery* (The bulk of the work was first published in Vienna in 1935; this reprint was first published by Hutchinson in 1959, new notes and footnotes in the present reprint). Routledge, London.

Renyi, A. 1969. *Foundations of Probability*. Holden-Day, San Francisco.

Robertson, S. 1986. On relevance weight estimation and query expansion. *J. Doc. 42*, 3, 288–297.

Robertson, S. and Walker, S. 1994. 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), Springer-Verlag, New York, 232–241.

Robertson, S., Walker, S., Beaulieu, M., Gatford, M., and Payne, A. 1996. Okapi at Trec-4. In *NIST Special Publication 500-236: The Fourth Text REtrieval Conference (TREC-4)*, D. Harman, Ed., Department of Commerce, National Institute of Standards and Technology, Gaithersburg, Md., 182–191.

Robertson, S. E. and Sparck-Jones, K. 1976. Relevance weighting of search terms. *J. Am. Soc. Inf. Sci. 27*, 129–146.

ROBERTSON, S. E., VAN RIJSBERGEN, C. J., AND PORTER, M. 1981. Probabilistic models of indexing and searching. In *Information Retrieval Research*, S. E. Robertson, C. J. van Rijsbergen, and P. Williams, Eds., Butterworths, Oxford, UK, Chapter 4, 35–56.

SALTON, G. AND BUCKLEY, C. 1988. Term-weight approaches in automatic text retrieval. *Inf. Process. Manage. 24*, 5, 513–523.

SINGHAL, A., SALTON, G., MITRA, M., AND BUCKLEY, C. 1996. Document length normalization. *Inf. Process. Manage. 32*, 5, 619–633.

SOLOMONOFF, R. 1964a. A formal theory of inductive inference. Part I. *Inf. Control 7*, 1 (March), 1–22.

SOLOMONOFF, R. 1964b. A formal theory of inductive inference. Part II. *Inf. Control 7*, 2 (June), 224–254.

TITTERINGTON, D. M., SMITH, A. F. M., AND MAKOV, U. E. 1985. *Statistical Analysis of Finite Mixture Distributions*. Wiley, Chichester.

TURTLE, H. AND CROFT, W. 1992. A comparison of text retrieval models. *Comput. J. 35*, 3 (June), 279–290.

VAN RIJSBERGEN, C. 1977. A theoretical basis for the use of co-occurrence data in information retrieval. *J. Doc. 33*, 106–119.

WILLIS, D. 1970. Computational complexity and probability constructions. *J. ACM 17*, 2, 241–259.

WITTEN, I. H., MOFFAT, A., AND BELL, T. C. 1999. *Managing Gigabytes*, second ed. Morgan Kaufmann, San Francisco.

WONG, S. AND YAO, Y. 1995. On modeling information retrieval with probabilistic inference. *ACM Trans. Inf. Syst. 16*, 38–68.

Received January 2002; revised May 2002; accepted June 2002