# A Probabilistic Model for Information Retrieval Based on Maximum Value Distribution

Jiaul H. Paik
University of Maryland, College Park, USA
jia.paik@gmail.com

## ABSTRACT

The main goal of a retrieval model is to measure the degree of relevance of a document with respect to the given query. Probabilistic models are widely used to measure the likelihood of relevance of a document by combining within document term frequency and term specificity in a formal way. Recent research shows that tf normalization that factors in multiple aspects of term salience is an effective scheme. However, existing models do not fully utilize these tf normalization components in a principled way. Moreover, most state of the art models ignore the distribution of a term in the part of the collection that contains the term. In this article, we introduce a new probabilistic model of ranking that addresses the above issues. We argue that, since the relevance of a document increases with the frequency of the query term, this assumption can be used to measure the likelihood that the normalized frequency of a term in a particular document will be maximum with respect to its distribution in the elite set. Thus, the weight of a term in a document is proportional to the probability that the normalized frequency of that term is maximum under the hypothesis that the frequencies are generated randomly. To that end, we introduce a ranking function based on maximum value distribution that uses two aspects of tf normalization. The merit of the proposed model is demonstrated on a number of recent large web collections. Results show that the proposed model outperforms the state of the art models by significantly large margin.

## Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Search and Retrieval: Retrieval Models

## General Terms

Algorithm; Experimentation; Performance


*SIGIR'15,* August 09 - 13, 2015, Santiago, Chile.

DOI: http://dx.doi.org/10.1145/2766462.2767762.

## Keywords

Document ranking; Retrieval model; Extreme value theory

## 1. INTRODUCTION

To measure the weight of a term in a document, most well known functions combine three major components - the term frequency, the inverse document frequency and the document length. The term frequency factor is a key evidence for determining a term's salience in a document, while inverse document frequency is used for attenuating the effect of terms that occur too often in the collection to be meaningful for relevance determination. On the other hand, term frequency is closely related with document length, since long documents tend to use a term repeatedly. Thus, term frequency normalization, in accordance with the document length, is necessary to remove the advantage that the long documents have in retrieval over the short documents.

Given these three major components, the key question is then how these components can be integrated to produce a composite weight for each query term in each document and that is where one model differs from the other. Most well known weighting functions under the vector space model compute the composite weight by taking the product of the $tf$ factor and the $idf$ factor, where the $tf$ factor is some combination of $tf$ and the document length. Classical probabilistic models (for example BM25 [24]), adopt somewhat the same strategy. Although, they have the same objective, the two models have very different ways of determining the functional form of the tf factor. The nature of the tf functions under the vector space framework are generally constructed empirically, which are primarily guided by the experimental results, while BM25 formula is derived by approximating the logarithm of odds ratio of two Poisson distributions- one for relevant documents and the other for non-relevant documents. On the other hand, language models (LM) [21] differ from the above models in a fundamental way in the sense that the documents are ranked based on the likelihood that the query has been generated from the document in consideration. In addition, unlike tf.idf models, language models do not use explicit length normalization. The length of the document is an integral part of the probability estimation. Non-parametric probabilistic models are also known to be very effective in information retrieval. One of the widely used non-parametric probabilistic model is divergence from randomness (DFR) [1] based approaches, where the term weight is computed by measuring the divergence between a term distribution produced by a random process and the actual term distribution. One major deficiency with

these models is that they consider only the document length normalized $tf$ and ignore within document relative $tf$ distributions. Recent research [20] shows that integration of within document relative $tf$ into scoring model improves performance significantly. However, it is yet not clear how this variable can be added into the existing models formally.

This article describes a probabilistic retrieval model that obviates empirical way of determining a ranking function, unlike existing tf-idf models [29, 20]. The model introduces a tf factor based on the distribution of maximum values of normalized tf. The model achieves a number of important goals. First, it integrates the recent multi-aspect tf normalization schemes into a probabilistic framework. Second, the model automatically factors in the distribution of normalized tf in a term specific way, unlike many standard models. Third, it uses a mixture of two maximum value distribution to better model distributions of terms having varying heaviness of tails. To the best of our knowledge, this work is the first to address the ranking problem using the distribution of maximum values.

The effectiveness of the proposed model is evaluated on a number of recent web test collections containing millions of documents. We compare the performance of the proposed method to the state of the art representative baselines from tf-idf model, classical probabilistic model, language model and divergence from randomness model. Our primary experimental results show that the proposed model almost always outperforms the state of the art baselines by a significantly large margin. We carry out additional set of experiments to compare the performance of the proposed model against a log-logistic (LL) based model that uses multi-aspect tf normalization. Once again, the results suggest that the proposed model is often significantly better than LL model. Moreover, the results demonstrate that our model is more precise than the state of the art models, thereby making it a potential choice for web search.

We organize the article as follows. Section 2 reviews the state of the art. The proposed approach is described in Section 3. The experimental setup is detailed in Section 4. In Section 5, we present the experimental results. Finally, we conclude in Section 6.

# 2. PRIOR WORK

Modeling term weight is the central issue in an information retrieval system. Three widely used models in IR are the probabilistic models [23], the vector space model [28, 27], and the inference network based model [31]. Furthermore, probabilistic models can be broadly classified into three groups, namely the classical probabilistic model, language model and a non-parametric divergence from randomness model. A large number of instances of these models exist in the literature. In this section we mainly review the state of the art representatives from each of these categories.

## 2.1 Classical Probabilistic Model

The key part of the probabilistic models is to estimate the probability of relevance of the documents for a query. This is where most probabilistic models differ from one another. Since the introduction of full text search, a large number of weighting formulae have been developed that attempt to measure document relevance probabilistically and BM25 [22] seems to be the most effective weighting function from among them. BM25 model approximates the two Poisson model of relevance. The approximation is done using a increasing asymptotic $tf$ function. Although, structurally, BM25 and tf-idf functions are very similar (in the sense that they both use $tf$ and $idf$ factor), they differ in many respects. First, BM25 has a well grounded theory, while most of the tf-idf models have an empirical background. Second, anatomically, $IDF$ factor of BM25 discounts the collection size by the document frequency of the term, which is different from the standard IDF factor. Third, BM25 uses a different query term frequency function, unlike tf-idf models where that function is linear. The length normalization factor uses the average document length and a parameter has been introduced to control the relative length effect.

## 2.2 Language Model

Probabilistic language modeling approaches [21, 15] follow a different principle in estimating the relevance of a document, unlike classical probabilistic models. Typically, language modeling approaches compute the probability of generating a query from a document, assuming that the query terms are chosen independently. Unlike TF-IDF models, language modeling approaches do not explicitly use document length factor and the $idf$ component. It seems that the length of the document is an integral part of this formula and that automatically takes care of the length normalization issue. However, smoothing is crucial and it has very similar effect as the parameter that controls the length normalization factor and term specificity in pivoted normalization or BM25 model. Three major smoothing techniques (Dirichlet, Jelinek-Mercer and Two-stage) are commonly used in this model [32].

Although, query likelihood model is reasonably effective, one major deficiency with using a multinomial distribution as a language model is that all term occurrences are treated independently. The term-independence assumption in information retrieval is often adopted in theory and practice, as it renders the retrieval problem tractable. It is well known that once a term occurs in a document, it is more likely to reappear in the same document. This phenomenon is known as word burstiness [18] and is a type of dependency that is not modelled in the multinomial language model. Cummins et al. [8] present a Smoothed Polya Urn Document language model, which incorporates word burstiness only into the document model. They use the Dirichlet compound multinomial (DCM) to model documents in place of the standard multinomial distribution, whereas the standard multinomial is used to model query generation.

## 2.3 Divergence from Randomness Model

Amati and Rijsbergen [1] proposed a class of non-parametric probabilistic approaches to term weighting called divergence from randomness (DFR) model. The weight of a term in DFR models is the amount of divergence between a term distribution produced by a random process and the actual term distribution. The anatomy of the weighting function of DFR is defined as follows

$$w(t,d) = -log_2(Prob_1) \cdot (1 - Prob_2). \quad (1)$$

The left factor measures the information content of the term in a document based on its distribution in the entire collection, while the right factor measures the information gain of the term with respect to its occurrence in the elite set (set of documents that contains the term). $Prob_1$ is com-

puted using various well known distributions (such as Bose-Einstein statistics, Poisson distributions etc), while $Prob_2$ is measured using Laplace law of succession or the ratio of two Binomial distributions. Like other models, DFR models use the same basic components. However, the integration of various component are derived theoretically. DFR models use explicit length normalization and following standard practice, average document length is considered as the ideal document length.

## 2.4 Vector Space Models

In vector space model, the search problem is viewed in a different way. Queries and documents are represented as the vectors of terms. To compute a score between a document and a query, the model measures the similarity between the query and document vector using cosine function. The central part of the vector space model is to determine the weight of the terms that are present in the query and the documents. Salton and Buckley [26] summarize a number of term weighting approaches which use various types of normalization. It is evident that document length is an important component in effective term weighting. Singhal et al. [29] identify a number of weaknesses of cosine and maximum $tf$ normalization and they observe that a weighting formula that retrieves documents with chances similar to their probability of relevance performs better. Following this observation, they propose a pivoted normalization scheme that acts as a correction factor of old normalization and is one of the most effective term weighting schemes in the vector space framework. Typically, the term weighting functions in vector space model are constructed empirically. Several work tried to go beyond purely empirical approaches and use the data instead to learn the patterns that satisfy the data. For example, Greiff [12] uses exploratory data analysis to uncover some important relationship between the document frequency and the relevance of a document.

Most of the earlier work on vector space model normalizes the term frequency in accordance with the length of the documents. Paik [20] argued that the length based normalization alone is not sufficient to capture the different aspects of term salience and that within document distribution of the terms plays an important role. He then proposed a two-aspect normalization scheme. An asymptotic bounded increasing function (much in spirit with BM $tf$ function) is then used to transform the normalized $tf$ values. Two $tf$ components are then combined using query length information. However, the main weakness of the model is its highly empirical nature and that is where the model proposed in this article differs from [20]. The proposed model has a formal probabilistic foundation that directly produces the weighting function.

## 2.5 Other Models

In inference network, document retrieval is modeled as an inference process [31]. A document instantiates a term with a certain strength and given a query, the credit from multiple terms is accumulated to compute a relevance, which is very much equivalent to the similarity score of vector space model. From an operational angle, the strength of instantiation of a term for a document can be considered as weight of the term in a document. The strength of instantiation of a term can be computed using any reasonable formula.

Some models go beyond the use of bag of words features only and incorporates the proximity/phrases of query terms in the documents [6, 9]. Metzler and Croft [19] develop a general formal framework for modeling term dependencies via Markov Random Fields. The model allows arbitrary text features, such as occurrence of single term, ordered phrases and unordered phrases to be incorporated as the potential evidences of relevance. They explore full independence (bag of words) , full dependence (between every pair of query terms) and sequential dependence (between consecutive query terms) in the language modeling framework. Since, the model has to compute the positional information during query processing time, it is more computationally complex than our model.

Fang et al. [10] give a comprehensive analysis of four retrieval models by defining a set of constraints that needs to be satisfied for effective retrieval. Using these constraints the strengths and weaknesses of some well known models are analyzed and some of the models are modified. There are also a number of recent works that focus on the constraint based analysis of the retrieval models [4, 7].

# 3. PROPOSED WORK

In this section we describe the proposed ranking model. We first revisit the key variables used in a typical ranking model and describe the roles they play. We then describe how maximum value can be used for ranking. Finally, we turn on to present the maximum value based models and their parameter estimation.

## 3.1 TF-IDF Model: A Probabilistic View

Within document Term frequency and inverse document frequency (idf) are the two main building blocks of information retrieval models that measure query-document similarity. These two variables play a complementary role in ranking documents in response to a query. The idf factor of a term $t$ ($idf(t)$) measures the information gain of randomly picking a document that will fall in the elite set for $t$ (the set of document that contains $t$ and henceforth we denote it as $E(t)$). On the other hand, tf factor of $t$ for a document $d$, ($tff(t,d)$) measures the relative weights of documents within $E(t)$. Thus, from an operational perspective, $idf(t)$ balances the weight between different $E(t)$, while $tff(t,d)$ adjusts the relative weights of documents within the same elite set. Term frequency hypothesis suggests that $tff(t,d)$ is an increasing function of normalized term frequency. Intuitively, this means, if the rank of a document $d$ having $ntf(t,d)$ (normalized tf of $t$ in $d$) is relatively high in $E(t)$, the contribution made by $tff(t,d)$ is also high. Hence, given the distribution of normalized tf of a term in $E(t)$, a natural way to measure $tff(t,d)$ is to take the percentage of documents in $E(t)$ having normalized tf not higher than $ntf(t,d)$. Thus, $tff(t,d)$ can be defined as follows:

$$tff(t,d) \propto P(X \leq ntf(t,d)) \tag{2}$$

where $X$ is the random variable on normalized tf values in $E(t)$.

Lv and Zhai [17] argued that straightforward non-parametric (plain percentile based) way of estimating this probability does not fully factor in the main objective of tf hypothesis, since it ignores the quantum of differences of normalized tf values. Thus, they advocate the use of parametric probability distribution functions to circumvent this limitation.

They use log-logistic distribution for computing $tff(t,d)$ as follows

$$tff(t,d) = P(X \leq ntf(t,d)|c,\alpha) = F(ntf(t,d)|c,\alpha) = \frac{ntf(t,d)^{\alpha}}{c^{\alpha} + ntf(t,d)^{\alpha}} \quad (3)$$

where $c > 0$ and $\alpha > 0$ are the model parameters which can be estimated from the normalized tf values in $E(t)$. The main issue in this approach is to choose the right distribution function that captures the distribution of normalized values properly. We use maximum value distribution of two aspect normalized tf values in the above framework to measure the $tff(t,d)$. In the next two sections, we describe multi-aspect tf normalization scheme followed by the maximum value based model.

## 3.2 Term Frequency Normalization

Raw term frequencies are known to be less effective because of its correlation with the document length. Thus, a long document enjoys preference over a short document if the term frequency is used as is. A document becomes longer if it contains many unrelated contents together. Therefore, although the frequency of a term may not increase in this case, the document uses many distinct terms. Since, the chance of a random match of a term between a query and a document is approximately proportional to the number of distinct terms in the document, long documents get an additional advantage over shorter documents. On the other hand, documents also become longer if they repeat the same content, thereby resulting in higher term frequencies without giving any additional useful information.

Therefore, to enhance retrieval accuracy, it is imperative to regularize the term frequency in accordance with the document length. A standard and successful approach for doing this is to compare the length of the concerned document to the length of an ideal document (pivotal document). Both, pivoted tf-idf and BM25 effectively use this strategy where the length of the pivotal document is the average document length of the retrieval collection. Thus, the $tf$ of an average length document remain unchanged, while $tf$ of the documents longer (shorter) than average length document are punished (rewarded).

Recently, Paik [20] argued that the traditional length based normalization alone is not sufficient to capture the different aspects of term importance and proposed two normalization formulae- one is based on within document average term frequency, while the other makes use of the traditional length based approach. These two normalized $tf$s are then combined. We use the same normalization schemes as described in [20], since it gives state of the art results. For convenience, the normalization factors are called $ritf(t,d)$ (relative intra-document frequency of term $t$ in the document $d$) and $lrtf(t,d)$ (length normalized frequency of term $t$ in the document $d$). The following equations formally define the normalization schemes.

$$ritf(t,d) = \frac{log(1+tf(t,d))}{log(k+mtf(d))} \quad (4)$$

$$lrtf(t,d) = tf(t,d)\, log(1+\frac{adl}{l(d)}) \quad (5)$$

The terms $mtf$, $adl$ and $l(d)$ denote the mean term frequency of the document that contains $t$, the average document length of the collection and the length of the document $d$, and $k$ ($\geq 1$) is a smoothing parameter. The proposed model combines these frequency normalizations in a probabilistic framework.

## 3.3 Limitations of Existing Models

In the last section we have described multi-aspect tf normalization scheme. In this section we discuss the potential limitations of existing methods and the major difficulties in integrating multi-aspect tf normalization into the state of the art probabilistic models.

We start our discussion with the MATF model. We reiterate that, although, *idf* function does not vary much from one model to the other, it is the $tf$ function that often makes the main difference. In [20], the function $\frac{x}{1+x}$ is used to transform the normalized $tf$ values to enforce term coverage. However, the function has a number of notable shortcomings. First, the choice of the function is purely empirical in nature. Second, the function does not have the knowledge of the distribution of $tf$ in the elite set. Third, since the function operates on the $tf$ values having incompatible range (range of $ritf$ is much smaller than that of $lrtf$), one component overpowers the other component, thereby compromising the ultimate effectiveness.

BM25 model is a nice bridge between tf.idf and probabilistic model. Anatomically, BM25 is clearly separable into $tf$ and *idf* component, where the $tf$ function is a special case of log-logistic model and is guided by 2-Poisson model. BM25 normalizes $tf$ in accordance with the document length where average document length is used as an ideal (or pivotal) document. However, it is not clear how to integrate relative intra-document $tf$ into this model, since the notion of pivot for relative intra-document tf is hard to define. Moreover, BM25's $tf$ function is also distribution independent.

Unlike the previous two models, divergence from randomness model (DFR) takes a more principled approach in terms of factoring in the term distribution. Once again, it is yet unknown how relative intra document $tf$ ($ritf(t)$) can be added to this model that will be theoretically consistent with DFR's basic principle. Moreover, normalized $tf$ values are continuous valued random variable and thus, an attempt to integrate it into DRF will give rise to theoretical anomaly, since DFR uses discrete distributions to measure information gain.

Language model is very different from all the models discussed above primarily because it neither uses *idf* explicitly nor it uses length normalization. Thus we confine our discussion on the models that have explicit $tf$ and *idf* factors. In the next section we describe the maximum value based model and how it can be used to circumvent some of the problems outlined above, followed by the development of a model that uses two aspect $tf$ normalization in a probabilistic framework.

## 3.4 Maximum Value Model

Unlike existing ranking models, we attempt to measure $tff(t,d)$ based on the nature of some of the largest values of normalized $tf$ for that term. A natural consequence of using maximum value based ranking is that it makes the weight of a term in a document dependent upon the distribution of normalized tfs in $E(t)$.

To that direction, the simplest possible approach could be to take the maximum value of normalized tf for a term $t$

and then measure $tff(t,d)$ relative to the maximum value. Clearly, this scoring is perfectly consistent with $tf$ hypothesis, where the document having highest normalized tf gets highest weight. We can easily think of two naive approaches to measure $tff(t,d)$ that are based on maximum values. One potentially feasible approach can be percentile based scoring that we have outlined before, while the other simple approach can be to measure $tff(t,d)$ as a ratio of $ntf(t,d)$ (or some increasing function of ratio) and the maximum normalized $tf$ for that term in the collection. To understand the limitations of these two approaches, let us consider the following examples.

Let $x_1, x_2, \ldots, x_{n-1}, x_n$ be the normalized $tf$ values for a term $t$ in ascending order. As our first case, let us assume that $(x_n - x_{n-1}) \approx 0$. The percentile based method may give higher weight for $x_n$ compared to $x_{n-1}$ even if they are nearly the same. This happens because percentile based method does not factor in the magnitude of difference, which consequently violates the tf hypothesis. As a second case, if it happens that $x_n \gg x_{n-1}$, scoring based on ratio gives too much priority on the maximum value alone, which results in sharp discount of scores of other documents. As a consequence, a document even if genuinely relevant, is undesirably punished.

These problems are addressed using a sampling based technique which exclusively focuses on maximum values of samples. Rather than relying on a single value, we attempt to measure the distribution of values at the right tail where some of the largest values fall. Hence, our main goal is to model the nature of the right tail of $ntf(t,d)$. We hypothesize that the most potentially relevant documents for a term fall on that part of the distribution. Quite clearly, this hypothesis is consistent with the standard tf hypothesis. Thus, the main challenge is to model the nature of the right most tail as accurately as possible. In other words, this model measures the likelihood that $ntf(t,d)$ will fall on the right most tail. Thus, if the probability is higher, likelihood of $d$ being relevant will also be higher.

We now focus on the models for maximum values. We reiterate that in order to avoid the influence of a single quantity (maximum value), the following sampling based approach is taken to derive maximum value distributions. Let us assume that $N$ samples, each of size $n$ are drawn from the same population. From each sample we can get the largest value. Thus in $nN$ observations we have $N$ largest values corresponding to each random sample. The distribution of the largest values in $nN$ observations will tend to follow the same asymptotic expression as the distribution of the largest value in samples of size $n$. Consequently, the asymptote must be such that the largest value of a sample of size $n$ taken from it must have the same asymptotic distribution. Formally, the maximum value distribution is defined as follows. Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed random variable with distribution $F$.
Let $M_n = max(X_1, X_2, \ldots X_n)$. Then,

$$Pr(M_n \leq x) = Pr(X_1 \leq x, X_2 \leq x \ldots X_n \leq x) \quad (6)$$
$$= F^n(x) \quad (7)$$

Since a linear transformation does not change the form of the distribution, the probability that the largest value is less than $x$ should be equal to the probability of a linear function

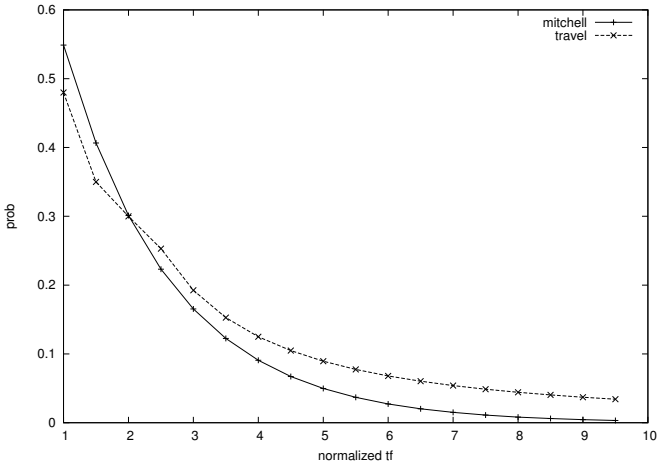


**Figure 1: Distributions of random samples of normalized elite set term frequency of *mitchell* and *travel*.**

of $x$. Thus, the above equation is equivalent to

$$Pr(\frac{M_n - b_n}{a_n} \leq x) = F^n(a_n x + b_n). \quad (8)$$

**Fisher-Tippett-Gnedenko theorem** [11] states that if a pairs of real numbers $(a_n, b_n)$ ($a_n$ and $b_n$ must be functions of $n$) exist such that $a_n > 0$ and

$$\lim_{n\to\infty} F^n(a_n x + b_n) \to D(x) \quad (9)$$

for a distribution $F$, then $D(x)$ can be Type I or Type II distribution defined below.

The type I distribution [13] (known as Gumbel distribution) is defined as

$$F_g(x) = \exp(-\exp(-\frac{x-\mu}{\alpha})),\ \mu \in R;\ \alpha > 0. \quad (10)$$

while type II distribution [13] (Frechet distribution) for positive random variable is defined as

$$F_f(x) = \exp(-\left(\frac{\mu}{x}\right)^{\alpha}),\ x \geq 0;\ \mu > 0;\ \alpha > 0. \quad (11)$$

Having defined the maximum value distribution, our next major goal is to verify that the maximum value distributions satisfy the mandatory preconditions in order to be applicable in our task. Specifically, the data must be coming from a distribution $F$ that satisfies **Fisher-Tippett-Gnedenko theorem**. Thus, our primary goal is to fix the underlying distribution function from which the data have been supposedly generated. In order to guess $F$, we first examine the distributions of normalized frequencies for a few randomly chosen terms. We noticed that the density graphs near the extreme right tail are not monotonically decreasing and it happens primarily because of the presence of random noise or extreme outliers. We empirically (by plotting) identify the points at which the density graphs violate this smoothness for the first time and ignore all the data larger than this particular point. On Clueweb collections, our analysis suggests that normalized tf values between 70-80 seem to be

a reasonable cut-off point and thus, in our experiments we set it to 75 empirically (but that value may depend on the nature of the collection). We then plot the distributions of the truncated data. As an example, Figure 1 shows distributions for two selected terms. To better understand the relationship between the pattern of distributions and term's collection level occurrence, we choose two terms (*"mitchell"* and *"travel"*) of varying specificity. Figure 1 clearly shows that both the terms seem to be following long tail distributions with monotonically decreasing density functions. We consider two such long tail distributions – namely, exponential distribution and Pareto distribution. Note that the nature of the tails are different in these two cases.

### *Case 1.*

Suppose the data have been distributed from exponential distribution. Thus, $F(x) = 1 - \exp(-x/\alpha)$, $\alpha > 0$. If we choose $a_n = 1$ and $b_n = \ln n$, Then

$$F^n(a_n x + b_n) = \left(1 - \exp(-\frac{-x - \ln n}{\alpha})\right)^n \quad (12)$$

$$= \lim_{n\to\infty} \left(1 - \frac{\exp(-x/\alpha)}{n}\right)^n \quad (13)$$

$$= \exp(-\exp(-x/\alpha)). \quad (14)$$

Thus, if the data is generated from exponential distribution, for $a_n = 1$ and $b_n = \ln n$, maximum value distribution converges to Gumbel distribution.

### *Case 2.*

Suppose now the data have Pareto tail. Thus, $1 - F(x) = cx^{-\alpha}$ as $x \to \infty$, with $c > 0$ and $\alpha > 0$. Again if we set $a_n = n^{\frac{1}{\alpha}}$ and $b_n = 0$, then for $x > 0$ we have

$$F^n(a_n x) = \left(1 - c(a_n x)^{-\alpha}\right)^n \quad (15)$$

$$= \lim_{n\to\infty} \left(1 - c\frac{x^{-\alpha}}{n}\right)^n \quad (16)$$

$$= \exp(-(\frac{\mu}{x})^{\alpha}) \,(\text{setting}\, c = \mu^{\alpha}) \quad (17)$$

which turned out to be Frechet distribution. Hence, the above results provide us the necessary evidence that the maximum value distributions can be applied on our data.

### *Mixture Model.*

Although, $F_g$ and $F_f$ are the asymptotic approximations to maximum value models, the shapes of their distributions are not identical. Frechet distribution has longer right tail (Pareto tail) than Gumbel. This has some interesting correlation with the distribution of term frequencies in a large collection. If a term is more general (but not really stopwords), the frequency distribution for that term likely to have a longer tail than that of more specific term. Figure 1 illustrates this point clearly: the density curve of "*mitchell*" (which is a rare term) touches the x-axis much before that of "*travel*" (which is a more general term). Thus, an attempt to model the distribution of a term using only one of Gumbel and Frechet may lead to lower accuracy. Any real query contains terms having varying collection frequency and this motivates us to use a weighted mixture of the two distributions. Thus, our resulting distribution is defined as

$$G(x) = p \cdot F_g(x) + (1 - p) \cdot F_f(x), \quad 0 < p < 1 \quad (18)$$

where $p$ can be considered as prior of $F_g(x)$. A straightforward way to estimate $p$ is to use a standard method such as gradient ascent method that directly optimizes a target retrieval metric (such as NDCG@20). Indeed, we adopt such an approach, but not directly on $p$. As we have discussed earlier, $F_g$ (Gumbel) distribution seems better in modeling the distribution of a term having relatively smaller $df$ values (more specific). Thus, instead of optimizing the value of $p$ independently, we make the value of $p$ dependent on $df$. Specifically, if a term has low $df$ (high $idf$) we give higher weight to $F_g(x)$. In other words, $p$ should be higher for high $idf$ terms. We formalize this intuition using the following well known linear model

$$\frac{p}{1-p} = \beta \cdot idf \quad (19)$$

which gives the following solution for $p$

$$p = \frac{\beta \cdot idf}{1 + \beta \cdot idf}. \quad (20)$$

where $\beta$ ($> 0$) is a free parameter.

## 3.5 Scoring Function

We are now ready to define our final scoring function. Our scoring function uses two aspect tf normalization in maximum value distribution framework. Formally, if $X$ and $Y$ be the random variables corresponding to $ritf(t)$ and $lrtf(t)$ in $E(t)$ respectively, then $tff(t,d)$ is defined as

$$tff(t,d) = \alpha \cdot P(X \leq ritf(t,d)) + (1-\alpha) \cdot P(Y \leq lrtf(t,d))$$
$$= \alpha \cdot G(ritf(t,d)) + (1-\alpha) \cdot G(lrtf(t,d)) \quad (21)$$

where $0 < \alpha < 1$, is the interpolation parameter. Consequently, the final scoring function for a query $Q = q_1 q_2 \ldots q_n$ and a document $d$ is defined as

$$S(Q,d) = \sum_{q \in Q} tff(t,q) \cdot idf(q) \quad (22)$$

where $idf(t) = \log(N/df(t))$. The parameter $\alpha \in (0,1)$ in Equation 21 is set empirically.

## 3.6 Model Parameter Estimation

In this section we detail our method for estimating the parameters of the two maximum value distribution models described in the last section. These parameters play important role in determining the actual shape of the distributions which in turn make them term dependent. There are many methods for parameter estimation including maximum likelihood estimation (MLE), which perhaps is an obvious choice. However, in our case, MLE does not seem to be a good choice for the reason we detail next.

We explain the difficulty with Gumbel distribution only (similar argument holds for Frechet). The log-likelihood function of Gumbel based on random sample $x_1, x_2, \ldots, x_n$ is given by

$$L(\alpha, \mu) = -\sum_{i=1}^{n} \frac{x_i - \mu}{\alpha} - n \ln \alpha - \sum_{i=1}^{n} exp(-\frac{x_i - \mu}{\alpha}). \quad (23)$$

The system of differential equations (used for MLE)

$$\frac{\partial L}{\partial \mu} = \frac{\partial L}{\partial \alpha} = 0 \quad (24)$$

yields the following estimates for $\mu$ and $\alpha$

$$\mu = \alpha(\ln n - ln \sum_{i=1}^{n} exp(-\frac{x_i}{\alpha})) \quad (25)$$

and

$$\bar{x} = \alpha + \frac{\sum_{i=1}^{n} x_i exp(-\frac{x_i}{\alpha})}{\sum_{i=1}^{n} exp(-\frac{x_i}{\alpha})}. \quad (26)$$

Clearly, Equation 26 shows that $\alpha$ does not have closed form expression. Thus, we need to apply iterative numerical methods to find value of $\alpha$. Iterative methods may take substantial amount of time for very large collection such as Clueweb, since it needs to iterate over the set of maximum values from each random sample for each distinct term in the collection. This is precisely the reason we use point estimates (with a somewhat empirical transformation) of central tendencies for these models.

### 3.6.1 Parameter Estimation for Gumbel

The mean of Gumbel distribution is

$$\mu + 0.57 \cdot \alpha \quad (27)$$

while the standard deviation is

$$\frac{\pi}{\sqrt{6}}\alpha. \quad (28)$$

To estimate the values of $\alpha$ and $\mu$ we equate them with corresponding sample mean and standard deviation, which finally gives the following estimates.

$$\alpha = \frac{\sqrt{6}}{\pi}s \quad \text{and} \quad \mu = \bar{x} - 0.58 \cdot \frac{\sqrt{6}}{\pi}s \quad (29)$$

where $\bar{x}$ and $s$ are sample mean and standard deviation respectively. Since our data is positive random variable and originates from exponential distribution we use Equation 14 for final ranking. Thus, we do not need to worry about the parameter $\mu$. Our only concern is the parameter $\alpha$. Surprisingly, point estimate of $\alpha$ as is does not perform well in practice. Thus, in practice, we use a linear transformation, $\alpha = z_1 + z_2 \cdot s$, where $z_1$ and $z_2$ are set empirically to 2.5 and 0.04 respectively.

### 3.6.2 Parameter Estimation for Frechet

Mean and variance for Frechet are defined respectively as

$$\mu\Gamma(1 - 1/\alpha), \quad \alpha > 1 \quad (30)$$

and

$$\mu^2(\Gamma(1 - 2/\alpha) - \Gamma^2(1 - 1/\alpha)), \quad \alpha > 2. \quad (31)$$

Once again, the above two expressions are not very convenient to use since the improper integral $\Gamma(.)$ needs to be evaluated in order to compute the parameter. Fortunately, median and mode of Frechet distribution have much manageable expressions. Median is defined as

$$\mu 0.69^{-1/\alpha} \quad (32)$$

and mode is defined as

$$\mu(1 + \frac{1}{\alpha})^{-1/\alpha}. \quad (33)$$

As in Gumbel, we can equate these two expressions to sample median and mode to estimate the model parameters. However, unlike Gumbel, the parameters do not have closed form solution, which can be achieved by using any standard numerical method. Note that, in this case we do not need to iterate over the sample of maximum values, instead mode and median computed once for a term is enough. It is also important to note that although median for a sample is easy to determine, we need to do a little processing to compute mode from a set of real numbers. To compute mode of a sample, we create non-overlapping bins of numbers having 0.5 as the interval. We then take the median of the bin having highest frequency as our sample mode. We have adopted computationally efficient parameter estimation methods. However, a large number of other methods exist in the literature. Thus, it may be interesting to see whether other estimation strategies can improve the retrieval results without sacrificing efficiency too much.

## 4. EXPERIMENT SETUP

In this section, we describe the experiment setup used to evaluate the proposed model. Our experiments have the following two major objectives.

1. To compare the performance of the model against the state of the art probabilistic models (Section 5.1).

2. To compare against a recently proposed multi-aspect tf-idf weighting scheme [20] (Section 5.2).

**Table 1: Summary of the test collections and topics used in our experiments. 'M' stands for million.**

| Collection | # doc | topics | # topics |
|---|---|---|---|
| Clueweb.B-09 & 10 | 50M | 1-100 | 100 |
| Clueweb.B-11 & 12 | 50M | 101-200 | 100 |
| MQ-2009 | 50M | 20001-30000 | 684 |
| Clueweb.A-09 & 10 | 500M | 1-100 | 100 |
| Clueweb.A-11 & 12 | 500M | 101-200 | 100 |

We summarize the test collections used in our experiments in Table 4. The test collections are taken from TREC web tasks of recent years (2009-2012) as well as from million query 2009 (MQ-2009). The collections contain web documents and real web queries sampled from a search engine log. The documents are crawled from web and hence they have variety of content quality. Clueweb.B collection contains nearly 50 million documents, while ClueWeb.A collection contains approximately 500 million web pages. In MQ-2009 collection, although many queries available, not all queries have been judged. Thus, we use 684 queries for which judgments are available. All the collections have graded relevance assessment. It is important to note that, MQ-2009 queries have incomplete relevance assessment. Therefore, our evaluation methodology skips the unjudged documents from the ranked lists in order to compute the values of well known metrics following the recommendation made in [25].

Documents and queries are stemmed via Porter stemmer. Stopwords are removed from documents and queries. Statistically significant performance differences are determined using a paired $t$-test at 95% confidence level ($p < 0.05$). All our experiments are done using *title* field of the topics.

**Table 2: Retrieval effectiveness of the proposed method (MVD) compared to probabilistic models. Statistically significant improvements are indicated using the first letter of the less effective method. The highest value per column is boldfaced. The numbers in parenthesis indicate relative improvement over LM, PL2 and BM25, respectively.**

| | | Clueweb.B-09 & 10 | Clueweb.B-11 & 12 | MQ-2009 | Clueweb.A-09 & 10 | Clueweb.A-11 & 12 |
|---|---|---|---|---|---|---|
| ERR@20 | LM | 0.309 | 0.264 | 0.367 | 0.254 | 0.219 |
| | PL2 | 0.312 | 0.263 | 0.373 | 0.256 | 0.219 |
| | BM25 | 0.306 | 0.253 | 0.372 | 0.248 | 0.221 |
| | MVD | **0.337**$^{\text{lpb}}$ | **0.286**$^{\text{lpb}}$ | **0.408**$^{\text{lpb}}$ | **0.286**$^{\text{lpb}}$ | **0.257**$^{\text{lpb}}$ |
| | | (8.9, 7.9, 9.9) | (8.2, 8.8, 12.9) | (11.2, 9.5, 9.8) | (12.7, 11.4, 15.0) | (17.8, 17.5, 16.4) |
| NDCG@10 | LM | 0.282 | 0.228 | 0.395 | 0.200 | 0.194 |
| | PL2 | 0.285 | 0.231 | 0.393 | 0.205 | 0.196 |
| | BM25 | 0.284 | 0.222 | 0.391 | 0.208 | 0.191 |
| | MVD | **0.332**$^{\text{lpb}}$ | **0.268**$^{\text{lpb}}$ | **0.422**$^{\text{lpb}}$ | **0.261**$^{\text{lpb}}$ | **0.231**$^{\text{lpb}}$ |
| | | (17.9, 16.5, 16.8) | (17.3, 16.1, 20.4) | (7.0, 7.4, 7.9) | (30.7, 27.0, 25.3) | (19.0, 17.8, 21.3) |
| NDCG@20 | LM | 0.275 | 0.228 | 0.459 | 0.193 | 0.196 |
| | PL2 | 0.278 | 0.228 | 0.458 | 0.195 | 0.198 |
| | BM25 | 0.280 | 0.225 | 0.453 | 0.208 | 0.186 |
| | MVD | **0.325**$^{\text{lpb}}$ | **0.265**$^{\text{lpb}}$ | **0.479**$^{\text{b}}$ | **0.248**$^{\text{lpb}}$ | **0.228**$^{\text{lpb}}$ |
| | | (18.4, 17.0, 16.4) | (15.9, 16.2, 17.8) | (4.5, 4.5, 5.8) | (28.7, 27.4, 19.0) | (16.4, 15.2, 22.7) |

## 4.1 Baselines

The performance of the proposed model is compared to a number of state of the art retrieval models from different families. BM25 [24] is chosen as the representative baseline from the classical probabilistic model. From language model, we choose Dirichlet smooth version [32], since it is known to be the most effective among the language models [10]. From divergence from randomness family, we choose PL2 [1] as the baseline, following recent work [10, 14].

Pivoted document length normalization is chosen as a basic TF-IDF baseline. MATF [20] is chosen as another state of the art tf-idf model. Note that, MATF is a highly effective empirical tf-idf model and one of the major objectives of the proposed model is to advance the multi-aspect TF model using a probabilistic foundation. Finally, since our model attempts to capture the distribution of normalized tf, we also compare to multi-aspect TF normalization with a log-logistic distribution which has similar purpose. Thus, our set of baselines contains members from all state of the art families.

## 4.2 Free parameters and evaluation metrics

All the baseline models (except MATF) and the proposed model contain one or more free parameters. It is important to note that the parameters of these models often influence the performance to a statistically significant degree. Hence, for the sake of reliable and competitive comparison, the parameters are optimized using 5-fold cross validation with the corresponding evaluation measure (ERR@20 [2], NDCG@10 [16] or NDCG@20) as the target metric.

We choose expected reciprocal rank (ERR), NDCG@10, and NDCG@20 as our evaluation measures. ERR has been the primary evaluation metric for recent TREC web tracks [3]. NDCG@k leverages graded relevance and also has a position wise discounting. Thus, it reflects the overall quality of the documents at top $k$. On the other hand, ERR@k is a precision bias metric that leverages graded relevance assessment. Thus, ERR is more suitable metric for web search.

Clueweb collections contain substantial number of spam documents. Thus, following previous work [5], we have filtered out spam documents from the collections. Specifically, documents assigned by Waterloo's spam classifier [5] with a score below 70 were filtered out from the initial corpus. The score indicates the percentage of all documents in ClueWeb that are presumably "spammier" than the document at hand. The models are then run on the residual corpus to produce final ranked lists.

# 5. RESULTS

In this section we summarize retrieval performance of the proposed method and the baseline methods. Throughout the result section MVD denotes the proposed model.

## 5.1 Comparison to Probabilistic Models

Table 2 compares the performance of MVD to that of the three probabilistic models, namely, language model with Dirichlet prior, BM25 and PL2. First, we compare the performances measured by ERR@20. Table 2 shows that, on two Clueweb.B collections, MVD outperforms LM, PL2 and BM25 by a margin of 8% to 12% and all the differences are statistically significant. On MQ-2009 collection, MVD is once again always statistically significant compared to all the baselines with a margin more than 9%. Similarly, on two Clueweb.A datasets, MVD is unequivocally superior to the baselines and quite clearly the performance differences are even larger than that on Clueweb.B and MQ-2009. The baseline methods seem to be performing nearly equally and in none of the cases, the performance differences among the baselines found to be statistically significant.

Our next goal is to analyze the results measured in terms of NDCG@10. Once again, MVD gives consistent performance improvement over LM, BM25 and PL2 on Clueweb.B collections. The performance differences are always statistically significant with more than 15% relative improvements. Results on MQ-2009 collection also show that MVD is significantly more effective than all the baselines, however the relative differences are smaller compared to Clueweb collec-

**Table 3: Retrieval effectiveness of the proposed method (MVD) compared to tf-idf models. Statistically significant improvements are indicated using the first letter of the less effective method. The highest value per column is boldfaced. The numbers in parenthesis indicate relative improvement over PIVOT, MATF and LL, respectively.**

| | | Clueweb.B-09 & 10 | Clueweb.B-11 & 12 | MQ-2009 | Clueweb.A-09 & 10 | Clueweb.A-11 & 12 |
|---|---|---|---|---|---|---|
| ERR@20 | PIVOT | 0.263 | 0.234 | 0.367 | 0.169 | 0.196 |
| | MATF | 0.283 | 0.282 | 0.388 | 0.227 | 0.251 |
| | LL | 0.290 | 0.275 | 0.391 | 0.244 | 0.240 |
| | MVD | **0.337**$^{\mathbf{pml}}$ | **0.286**$^{\mathbf{pl}}$ | **0.408**$^{\mathbf{pml}}$ | **0.286**$^{\mathbf{pml}}$ | **0.257**$^{\mathbf{pl}}$ |
| | | (27.9, 18.8, 16.0) | (22.0, 1.5, 3.9) | (11.2, 5.1, 4.4) | (69.6, 26.1, 16.9) | (31.2, 2.7, 7.1) |
| NDCG@10 | PIVOT | 0.219 | 0.196 | 0.381 | 0.177 | 0.175 |
| | MATF | 0.276 | 0.240 | 0.402 | 0.197 | 0.213 |
| | LL | 0.287 | 0.234 | 0.418 | 0.207 | 0.206 |
| | MVD | **0.332**$^{\mathbf{pml}}$ | **0.268**$^{\mathbf{pml}}$ | **0.422**$^{\mathbf{pm}}$ | **0.261**$^{\mathbf{pml}}$ | **0.231**$^{\mathbf{pml}}$ |
| | | (51.6, 20.5, 15.8) | (36.6, 11.4, 14.5) | (10.8, 5.0, 1.0) | (47.5, 32.3, 25.8) | (32.5, 8.7, 12.5) |
| NDCG@20 | PIVOT | 0.212 | 0.200 | 0.442 | 0.181 | 0.171 |
| | MATF | 0.286 | 0.243 | 0.466 | 0.202 | 0.209 |
| | LL | 0.284 | 0.235 | 0.477 | 0.201 | 0.198 |
| | MVD | **0.325**$^{\mathbf{pml}}$ | **0.265**$^{\mathbf{pml}}$ | **0.479**$^{\mathbf{pm}}$ | **0.248**$^{\mathbf{pml}}$ | **0.228**$^{\mathbf{pml}}$ |
| | | (53.5, 13.5, 14.4) | (32.0, 9.0, 12.7) | (8.3, 2.8, 0.4) | (37.2, 22.7, 23.0) | (33.2, 9.4, 15.0) |

tions. One reason for smaller difference is that the baseline NDCG@10 numbers are very high, which makes the relative improvements smaller. The effectiveness of MVD on Clueweb.A collections is even more encouraging. MVD surpasses the baselines on Clueweb.A-09 & 10 collection by more than 20% margin which is clearly highly significant. We observe similar trend on the other Clueweb.A collection. As in ERR@20, the baselines seem to be performing with equal effectiveness.

We notice very similar (as in ERR@20 and NDCG@20) behavior of MVD on Clueweb.B collections measured by NDCG@20. Once again, MVD is consistently and significantly better than all the baselines with noticeably large margin of relative improvement. The picture is slightly different on MQ-2009 collection. Although, MVD is better than all the baselines, difference against BM25 only found to be significant. We suspect that sparser relevance judgements of MQ-2009 collection is a possible reason behind smaller differences. Finally, MVD beats the baselines by a convincingly large margin thereby maintaining its consistency as in the previous cases.

Overall, the results indicate that the proposed model based on distribution of maximum values yields consistent and significant retrieval performance improvement over the three state of the art probabilistic baselines from different categories measured by NDCG measures. We conclude that the proposed model is significantly more precise than the baselines on all the collections, thereby making it a very suitable for web search. The experiments also reveal that the performance of the baselines are very similar to each other, irrespective of the collection, which corroborates earlier findings that if parameters of the models are properly optimized, language model, BM25 and divergence from randomness model are closely comparable.

## 5.2 Comparison to TF-IDF Models

The experiments in this section are designed to compare the proposed method to a number of tf-idf models. By this set of experiments, we intend to achieve the following major goals.

1. How does the proposed model perform compare to a basic tf-idf model that uses only pivoted document length normalization?

2. Since MVD is based on multi-aspect term frequency normalization in a new probabilistic framework, how does it compare against a recent tf-idf model (MATF) that introduced multi-aspect tf normalization? We reiterate that this is the main issue we sought to address using maximum value based model.

3. We mentioned before that MATF combines the two normalized tf using an empirical tf function that transforms normalized tf values. Moreover, it does not factor in the distribution of normalized tf in the elite set for the particular term. Thus, in this section we compare the performance of MVD to a method that uses log-logistic probability distribution of two normalized tfs. The method is denoted as LL in the table. The parameters of this model is estimated using the method detailed in [17].

Table 5.2 compares the performance of tf-idf methods and MVD. First, it is clear from the table that MVD is highly significantly better than PIVOT. This holds for all collection and measured by all three evaluation measures. The performance differences are unequivocally statistically significant. On Clueweb (both A and B) collections, MVD gives upto 50% relative improvement over PIVOT. Second, MATF, which is based on relative intra-document tf normalization and length based normalization (which we call multi-aspect tf normalization), is always poorer than MVD and the differences are almost always statistically significant. More importantly, the margin of improvement by MVD is often noticeably high.

Thus, we conclude that maximum value distribution has large impact on retrieval performance. Finally, we compare the proposed method to log-logistic distribution based method denoted as LL in the table. Note that LL uses distribution of multi-aspect tf normalization for estimating rel-

evance and thus has probabilistic interpretation. Table 5.2 once again shows that MVD often significantly surpasses LL.

# 6. CONCLUSION

In this paper we introduce a probabilistic information retrieval model. The proposed model is guided by the principle that given the normalized frequency of a term in a document, the score is proportional to the likelihood that the normalized tf is maximum with respect to its distribution in the elite set for the corresponding term. We use a mixture of two maximum value distribution, that factors in varying specificity of query terms. The proposed model, integrates multi-aspect tf normalization scheme proposed recently in a probabilistic framework. Unlike many existing models, the proposed model takes into account the term specific distribution in the elite set. However, the unique contribution is that the model measures the likelihood of relevance focussing on the maximum values of the distribution, which we believe the first such effort to view ranking problem from this perspective. An empirical evaluation on large web collections containing millions of documents and hundreds of real world web queries demonstrates that the model significantly outperforms the state of the art probabilistic models from different families. As a future work, we plan to incorporate term proximity (ordered and un-ordered bigram) information into our model.

## Acknowledgments

I thank Doug Oard for useful discussions and suggestions. Without his support and advice this work would not have been possible. This research was supported in part by DARPA contract HR0011-12-C-0015 and NSF award 1065250.

# 7. REFERENCES

[1] G. Amati and C. J. Van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst.*, 2002.

[2] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *ACM CIKM, 2009.*

[3] C. L. A. Clarke, N. Craswell, I. Soboroff, and E. M. Voorhees. Overview of the trec 2011 web track. In *TREC*, 2011.

[4] S. Clinchant and E. Gaussier. Retrieval constraints and word frequency distributions a log-logistic model for ir. *Inf. Retr.*, 2011.

[5] G. V. Cormack, M. D. Smucker, and C. L. Clarke. Efficient and effective spam filtering and re-ranking for large web datasets. *Information retrieval*, 2011.

[6] W. B. Croft, H. R. Turtle, and D. D. Lewis. The use of phrases and structured queries in information retrieval. In *ACM SIGIR, 1991.*

[7] R. Cummins and C. O'Riordan. A constraint to automatically regulate document-length normalisation. In *ACM CIKM, 2012*

[8] R. Cummins, J. H. Paik, and Y. Lv. A polya urn document language model for improved information retrieval. *ACM Trans. Inf. Syst.*, 2015.

[9] J. Fagan. Automatic phrase indexing for document retrieval. In *ACM SIGIR, 1987.*

[10] H. Fang, T. Tao, and C. Zhai. Diagnostic evaluation of information retrieval models. *ACM Trans. Inf. Syst.*, 2011.

[11] R. A. Fisher and L. H. C. Tippett. Limiting forms of the frequency distribution of the largest or smallest member of a sample. In *Mathematical Proceedings of the Cambridge Philosophical Society*, 1928.

[12] W. R. Greiff. A theory of term weighting based on exploratory data analysis. In *ACM SIGIR, 1998.*

[13] E. J. Gumbel. *Statistics of extremes*. Courier Dover Publications, 2012.

[14] B. He and I. Ounis. A study of the dirichlet priors for term frequency normalisation. In *ACM SIGIR, 2005.*

[15] D. Hiemstra, S. Robertson, and H. Zaragoza. Parsimonious language models for information retrieval. In *ACM SIGIR, 2004.*

[16] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4), Oct. 2002.

[17] Y. Lv and C. Zhai. A log-logistic model-based interpretation of tf normalization of bm25. In *ECIR, 2012*

[18] R. E. Madsen, D. Kauchak, and C. Elkan. Modeling word burstiness using the dirichlet distribution. In *ICML, 2005.*

[19] D. Metzler and W. B. Croft. A markov random field model for term dependencies. In *ACM SIGIR, 2005.*

[20] J. H. Paik. A novel tf-idf weighting scheme for effective ranking. In *ACM SIGIR, 2013.*

[21] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *ACM SIGIR, 1998.*

[22] S. Robertson and H. Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4), Apr. 2009.

[23] S. E. Robertson. Readings in information retrieval. chapter The probability ranking principle in IR. 1997.

[24] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *ACM SIGIR, 1994.*

[25] T. Sakai. Alternatives to bpref. In *ACM SIGIR, 2007.*

[26] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, 24(5), Aug. 1988.

[27] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., 1986.

[28] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11), Nov. 1975.

[29] A. Singhal, C. Buckley, and M. Mitra. Pivoted document length normalization. In *ACM SIGIR, 1996.*

[30] K. Sparck Jones. Document retrieval systems. chapter A statistical interpretation of term specificity and its application in retrieval. 1988.

[31] H. Turtle and W. B. Croft. Evaluation of an inference network-based retrieval model. *ACM Trans. Inf. Syst.*, 9(3), July 1991.

[32] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2), Apr. 2004.