Score Distribution Models: Assumptions, Intuition, and Robustness to Score Manipulation

Evangelos Kanoulas* Keshi Dai† Virgil Pavlu† Javed A. Aslam†

*Department of Information Studies
University of Sheffield
Regent Court, 211 Portobello Street
Sheffield S1 4DP, UK
e.kanoulas@sheff.ac.uk

†College of Computer and Information Science
Northeastern University
360 Huntington Ave, #202 WVH
Boston, MA 02115, USA
{daikeshi, vip, jaa}@ccs.neu.edu

ABSTRACT

Inferring the score distribution of relevant and non-relevant documents is an essential task for many IR applications (e.g. information filtering, recall-oriented IR, meta-search, distributed IR). Modeling score distributions in an accurate manner is the basis of any inference. Thus, numerous score distribution models have been proposed in the literature. Most of the models were proposed on the basis of empirical evidence and goodness-of-fit. In this work, we model score distributions in a rather different, systematic manner. We start with a basic assumption on the distribution of terms in a document. Following the transformations applied on term frequencies by two basic ranking functions, BM25 and Language Models, we derive the distribution of the produced scores for all documents. Then we focus on the relevant documents. We detach our analysis from particular ranking functions. Instead, we consider a model for precision-recall curves, and given this model, we present a general mathematical framework which, given any score distribution for all retrieved documents, produces an analytical formula for the score distribution of relevant documents that is consistent with the precision-recall curves that follow the aforementioned model. In particular, assuming a Gamma distribution for all retrieved documents, we show that the derived distribution for the relevant documents resembles a Gaussian distribution with a heavy right tail.

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

General Terms: Theory, Measurement

Keywords: information retrieval, score distribution, density functions, recall-precision curve

1. INTRODUCTION

Given a user request an information retrieval system assigns scores to each document in the underlying collection according to some definition of relevance of each document to the user's request and returns a ranked list of documents to the user. In reality, this ranked list of documents is a mixture of both relevant and non-relevant documents. For a wide range of retrieval applications (e.g. information filtering, topic detection, meta-search, distributed IR), *modeling* and *inferring* the distribution of relevant and non-relevant documents over scores in a reasonable way can be highly beneficial. For instance, in information filtering, topic detection and recall-oriented retrieval, modeling the score distributions of relevant and non-relevant documents can be utilized to find the appropriate threshold between relevant and non-relevant documents [16, 17, 2, 19, 9, 15]. In distributed IR and meta-search it can be used to normalize document scores and combine different collections or the outputs of several search engines [5, 12].

Inferring the score distribution for relevant and non-relevant documents in the absence of any relevance information is an extremely difficult task, if at all possible. *Modeling* score distributions is often the basis of any possible inference. Due to this, numerous combinations of statistical distributions have been proposed in the literature to model score distributions of relevant and non-relevant documents. In the 1960s and 70s, Swets attempted to model the score distributions of non-relevant and relevant documents with two Gaussians of equal variance [16], two Gaussians of unequal variance, and two exponentials [17]. Bookstein instead proposed a two Poisson model [7] and Baumgarten a two Gamma model [5]. A negative exponential and a Gamma distribution [12] has also been proposed in the literature. The dominant model has been a negative exponential for the non-relevant documents and a Gaussian for the relevant ones [2, 12, 19]. Bennett [6] observed that when using a two-Gaussians model for text classification, document scores outside the modes of the two Gaussians (corresponding to "extremely irrelevant" and "obviously relevant" documents) demonstrated different empirical behavior than the scores between the two modes (corresponding to "hard to discriminate" documents). This motivated him to introduce several asymmetric distributions to capture these differences. Kanoulas et al. [11] recently proposed a Gamma distribution for the non-relevant documents and a mixture of Gaussians for the relevant documents.

The complexity of the underlying process that generates document scores makes it hard to theoretically argue about the actual distribution of document scores. Most of the aforementioned models were proposed on the basis of empirical fits to scores produced over different document corpora.

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

There have also been several attempts to intuitively argue about the shape of the different distributions. The starting point for most of these attempts has been some basic assumptions about the frequency of query term occurrences in documents (e.g. in Manmatha et al. [12]). Harter [10] and Bookstein and Swanson [8] used a mixture of Poisson distributions to model the distribution of words in a document, with one Poisson corresponding to the distribution of words in relevant documents and the other to the distribution of words in non-relevant documents.

In a different line of work, Arampatzis and van Hameren [2] showed that the distribution of relevant document scores rapidly converges to a Gaussian via the Central Limit Theorem as the number of query terms increases, under some basic assumptions. Further, they claimed that this is not true in the case of non-relevant documents.

Finally, Robertson [14] considered various combinations of distributions and examined whether these combinations exhibit anomalous behavior with respect to theoretical properties of precision and recall. Arampatzis et al. [1] proposed two truncated versions of the exponential-Gaussian model to overcome the theoretical problems associated with the original exponential-Gaussian model.

In this work, we model score distributions in a rather different, systematic manner. We start with a basic assumption on the distribution of terms in a document. Following the transformations applied on term frequencies by two basic ranking functions, BM25 and Language Models, we derive the distribution of the produced scores for all documents in an analytical form and illustrate that the derived distribution can be well approximated by a Gamma distribution.

Further, we also consider the score distribution for relevant documents. We detach our analysis from particular ranking functions. Instead, we consider a simple model for precision-recall curves proposed by Aslam and Yilmaz [3], which makes some very basic assumptions about the shapes of precision-recall curves that are produced by reasonable retrieval system on average. Given this model, we present a general mathematical framework which, given any score distribution for all retrieved documents, produces an analytical formula for the score distribution of relevant documents that is consistent with the precision-recall curves that follow the aforementioned model. In particular, assuming a Gamma distribution for all retrieved documents, we show that the derived distribution for the relevant documents resembles a Gaussian distribution with a heavy right tail.

2. FROM TERM FREQUENCIES TO RETRIEVAL SCORES

Traditional retrieval models score documents based on how well their language matches the language of the user's request. Thus, the essential component of all traditional scoring functions is the number of occurrences of query terms within a document (term frequency, TF). Different retrieval models apply different transformations over the term frequencies to produce a score per query term. The final score of a document is usually an aggregate of the document scores for each individual term.

Before we consider the distribution of term frequencies and the transformation applied by ranking functions over them in an analytical manner we illustrate the evolution of the term frequency distribution for all retrieved documents (documents that contain at least one of the query terms) for a sample query from the TREC 8 ad hoc collection (*Ireland Peace Talks*) and for two different retrieval models, BM25 and Language Models, in Figure 1.

The left panel corresponds to the transformation of TF distribution by BM25, while the right panel corresponds to the transformation by the Jelinek-Mercer Language Model.[1] Each column then, in both panels, corresponds to an individual query term and each row to progressively more complex transformations of the term frequency. The bottom row plots illustrate the final score distribution by the two retrieval models.

As can be observed, for both retrieval models, there is a critical step in the term frequency transformation (from Row 2 to Row 3) after which the score distribution radically changes and appears to be closer to the final score distribution. Furthermore, the shape of the final score distribution appears to be dominated by the most frequent query term in the collection (as expected) — for the sample query this is the term *talk* — and thus our main goal will be to derive the score distribution for each individual query term.

3. DERIVING THE DISTRIBUTION OF RAW STATISTICS

For a fixed query, consider a partition of the collection into relevance classes, such that D_Q is the class of documents that satisfy the information need to a certain degree $Q{>}0$. Depending on several factors like the user, the information need, the collection of documents etc, Q can take a range of values from "completely irrelevant" (the lowest Q) to "extremely relevant" (the highest Q). Note that in test collections (such as TREC) for simplicity only two or three classes are considered. The discussion in this section assumes a fixed quality/relevance class Q, and assumes all documents in the class contain all query terms at least once.

A query term t has a certain contribution to the document quality in response to the user query. For a given document quality Q, we assume an approximately constant probability of seeing the term t at any position in a document in class D_Q; hence we can model term t occurrences in documents in class D_Q with a Poisson process with rate $\lambda = \lambda_t = f(g, Q)$, where $g = g_t$ relates to the general rarity of the term in the language. Such a model is memoryless and implies that the query term appears equally likely at any moment. We do not model the dependence f — any monotonic function can be used, depending on the class model.

Counting the occurrences of a term t when reading a random document $d \in D_Q$ is analogous to counting buses at a bus station: arrive at the station, wait for the first bus, for the second bus, etc., and leave at some point (when the document ends). It is well known that the waiting times $w_1, w_2, w_3, \ldots$ among Poisson generated events are *exponentially* distributed i.i.d. random variables

$$w_l \approx \lambda e^{-x\lambda}. \quad (1)$$

The average waiting time is $\theta = 1/\lambda$, the mean of the exponential distribution. Intuitively, θ corresponds to a notion of the expected ratio of document length to term frequency, i.e., DL/TF.

[1]The parameter values used for BM25 are k_1=1.2 and b=0.75, and $\lambda = 0.2$ for the Jelenik-Mercer Language Model.

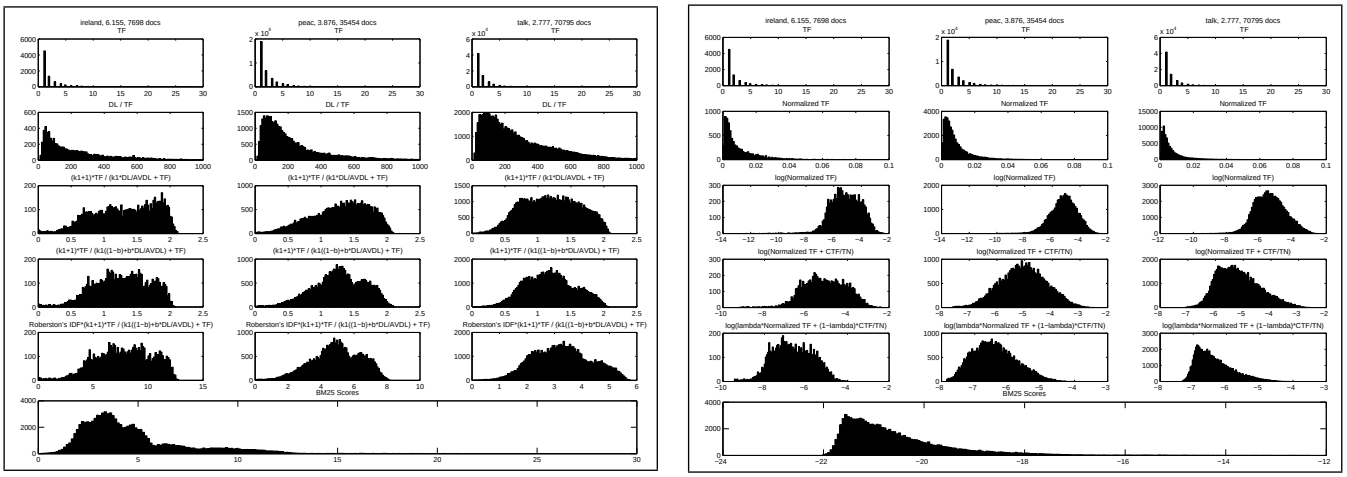

Figure 1: The empirical histograms of term frequencies evolving and resulting to the final scores for a sample query (*Ireland Peace Talks*) over the TREC 8 Ad Hoc Track collection for both BM25 (left) and JM Language Model (right). Each column corresponds to single query term while the rows correspond to progressively more complex transformations of the term frequency (TF) up to the final score for the two ranking functions. DL is the document length, ADL is the average document length, CTF is the collection term frequency, and TN is the number of terms in the collection.

Our purpose is to model the distribution of the random variable DL/TF for documents in class D_Q. We will do so separately for each frequency and then express the general distribution as a mixture.

Let us now fix a term frequency $k = 1, 2, 3, \ldots$ and denote $D_{Qk} = \{d \in D_Q \mid TF(t,d) = k\}$ the set of documents in D_Q that contain term t exactly k times. Here, we make the approximation that the document ends exactly after the k-th occurrence, and so we can write the document length DL as the sum of k waiting times $\sum_{l=1}^{k} w_l$, which immediately implies that DL is Gamma distributed (and more specifically Erlang-distributed), with shape k and scale $\theta = 1/\lambda$:

$$DL_{Qk} \sim Gamma(k, \theta). \tag{2}$$

Since k is a constant for the subclass D_{Qk}, the waiting time X_{Qk} is also Gamma distributed:

$$X_{Qk} = \frac{DocLength}{TermFrequency} = \frac{DL}{k} \sim Gamma(k, \theta/k). \tag{3}$$

Since the quality class D_Q is partitioned into the classes D_{Qk} for $k = 1, 2, 3, \ldots$, the waiting time X on D_Q follows a mixture of Gamma distributions with a constant mean θ, while DL on D_Q follows a mixture of Gamma distributions with a constant scale θ:

$$DL_Q \sim \sum_k P_Q(k) \cdot Gamma(k, \theta) \tag{4}$$

$$X_Q \sim \sum_k P_Q(k) \cdot Gamma(k, \theta/k) \tag{5}$$

where $P_Q(k) = Pr[TF(d,t) = k \mid d \in D_Q]$ denotes the probability that a document in class D_Q contains the term t exactly k times.

Assuming a constant probability p that a term occurrence gives quality Q, $P_Q(k)$ can be expressed as probability of $k-1$ failures (term occurrences that do not imply quality Q) followed by one success (term occurrence when quality Q is reached); therefore we model the mixture probabilities $P_Q(k)$ with a geometric distribution (equivalent to a negative binomial distribution with $\beta = 1$),

$$P_Q(k) = p(1-p)^{k-1} \tag{6}$$

where $p = p_t = \theta/ADL_Q$ expresses the correlation between the term and the information need on the class D_Q (the average document length, the general rarity of the term t, and the quality Q). For example, $p = 0.5$ implies that there are twice as many documents containing k terms than documents containing $k+1$ terms in the class D_Q. Intuitively p can be thought as a notion of inverse term frequency:

$$p = \theta/ADL_Q \approx avg(DL/TF)/ADL_Q \approx avg(1/TF).$$

Note that a number of different mixtures could be used, perhaps based on the query type. For instance, an informational query could use a negative binomial or a Poisson mixture. For the particular case of a geometric mixture however, an interesting result follows: Neuts and Zachs [13] show that under certain conditions similar to ours, a negative binomial mixture of Gamma distributions with constant scale is actually itself a Gamma distribution. With a different notation, their result is

$$\sum_k p_k \cdot Gamma(\beta + k, \theta) = Gamma(\beta, \theta/p) \quad \text{when} \tag{7}$$

$$p_k = NegBinomial(p, \beta) = \binom{k+\beta-1}{\beta-1} p^{\beta}(1-p)^k \tag{8}$$

Applying this on DL (with β=1) implies that DL is exponentially distributed on D_Q with mean θ/p. Of course this must hold for all query terms, not only for t, which requires a proportionality $\theta/p = constant = ADL_Q$. In practice, for a given quality class, the document length variable will not be exactly exponentially distributed for two reasons: (1) relevance judgments cover a range of qualities inducing an average effect, (2) our Poisson process model for query term occurrence works reasonably well for frequent terms, but can fail on rare terms. However, this model is fairly accurate in that DL can be modeled well by a Gamma distribution

with a small shape parameter (the exponential distribution is Gamma with shape = 1.)

Figure 2 illustrates the empirical histogram of DL/TF for the query term *system*. As can be observed, a Gamma distribution appears to be a good approximation of the empirical score distribution, offering empirical evidence that the assumptions and approximations in our theory are reasonable.[2]

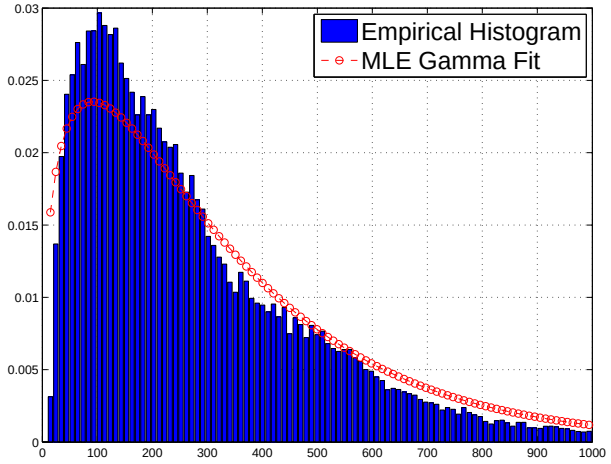

Figure 2: The empirical histogram and the Gamma density function fit over the $\frac{DL}{TF}$ scores for term *system* in TREC8.

4. DERIVING THE SCORE DISTRIBUTION FROM SCORING FUNCTIONS

In this section, we derive the score distribution of the retrieved documents in a systematic manner. We consider the transformation applied on the distribution of the elementary statistics described in the previous section by two scoring functions, BM25 and Jelinek-Mercer Language Model. The derivations presented here can be applied in the case of other retrieval models, such as TF-IDF and Divergence From Randomness (DFR).

4.1 Score Transformations

Consider a transformation of the random variable X by a monotonic, differentiable function r, $Y = r(X)$. The probability density function (pdf) of Y, $f_Y(y)$, can then be computed as a function of the pdf of X, $f_X(x)$ [4]. Let $F_Y(y)$ and $F_X(x)$ be the cumulative density function (cdf) of Y and X, respectively. Without loss of generality let r be a non-decreasing function. Then,

$$\begin{aligned} F_Y(y) &= Pr\{Y \leq y\} = Pr\{r(X) \leq y\} \\ &= Pr\{X \leq r^{-1}(y)\} = F_X(r^{-1}(y)) \text{ and} \end{aligned}$$

$$f_Y(y) = \frac{d}{dy}F_Y(y) = \frac{d}{dy}F_X(r^{-1}(y)) = \frac{\partial r^{-1}(y)}{\partial y} \cdot f_X(r^{-1}(y))$$

In the general case of a monotonic function r,

$$f_Y(y) = |\frac{\partial r^{-1}(y)}{\partial y}| \cdot f_X(r^{-1}(y))$$

[2]Some fits will be better than others, depending on the example. No theoretical model will fit all empirical examples, of course.

A rudimentary transformation of interest is just the inverse of $X = DL/TF$, which gives the normalized term frequency TF/DL. According to the previous section, $X = DL/TF \sim f_X = \sum_{k\geq 1} P_Q(k) * Gamma(k, \frac{\theta}{k})$. It is known that a mixture of Gamma can approximate any smooth function [18]. By approximating $P_Q(k)$ with a geometric distribution inverting TF/DL has the effect displayed in Figure 3. A relevant class of documents (high Q) implies:

- the geometric rates $1 - p = 1 - 1/(\lambda \cdot ADL)$ for query terms are higher, which means the mean $1/p$ is higher, or the mixture P_Q will have non-negligible coefficients for higher scale parameters k. This will make the mixture look more "hill"-like due to more effective components.

- for each query term, the Poisson generating process will be governed by a higher rate, $1/\theta$, which dictates a lower mean to all Gamma components of the mixture, or a "light" right-side tail. When the inverse transformation is performed (see below), the result distribution will have a heavier tail.

Conversely, a lower quality Q implies a mixture with effectively significant coefficients only for the lower k values, and also that the components of the mixture are less skewed towards the left-side, overall producing a more exponential-like distribution (after inversion).

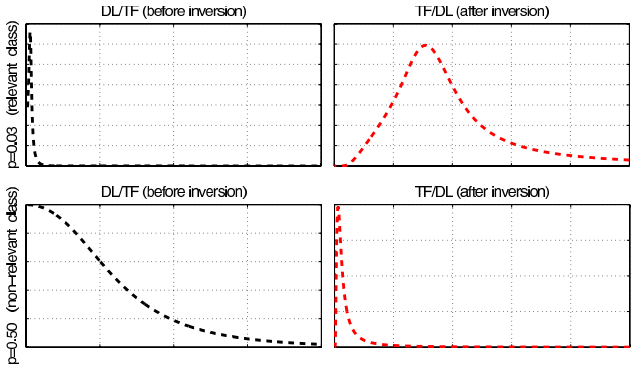

Figure 3: Mixture of gamma before and after the inversion, for different quality classes

Note that in practice fitting a Gamma, an inverse Gamma or an inverse Gaussian distribution in the TF/DL scores of existing collections/judgments (like TREC) are likely to differ in goodness-of-fit mostly due to random effects than other theoretical reasons - this is primarily due to complex score manipulations, and due to the sparsity and inaccuracy of the judgment process.

4.2 BM25 and Jelinek-Mercer LM

Assuming that query terms appear only once within a query the BM25 for a single query term can be calculated as:

$$\text{BM25 score} = \frac{(k_1+1)\,TF}{k_1((1-b)+b\frac{DL}{ADL})+TF} \cdot IDF \qquad (9)$$

where TF is the term frequency, IDF is the BM25 inverse document frequency, DL is the document length, and ADL is the average document length in the collection. By setting

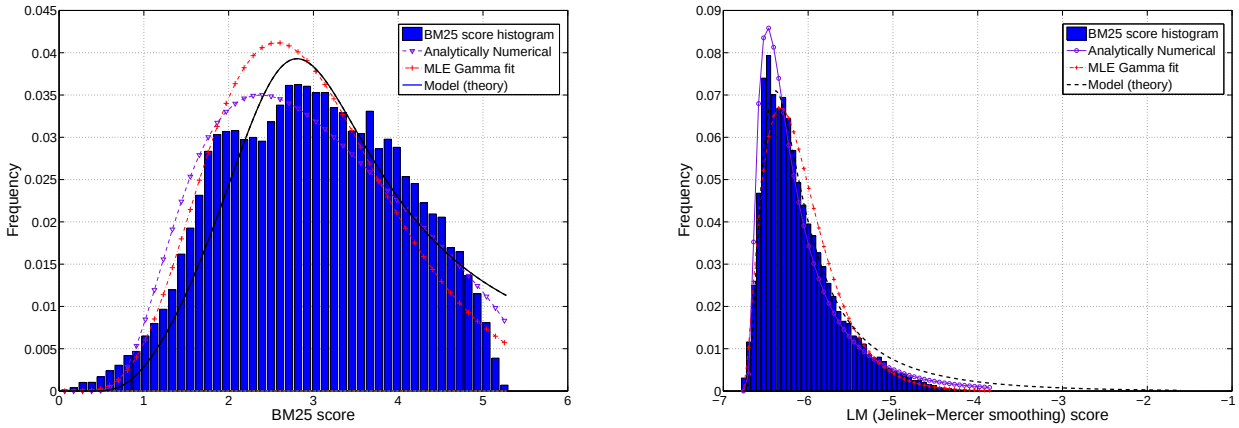

Figure 4: The empirical histograms, analytically numerical plot, and MLE Gamma fitting of the distribution of approximated BM25 scores and JM language model scores for term *system* in TREC8.

the parameter b equal to 1 (fixing the document length normalization) and defining the variable $X = DL/TF$, BM25 can be approximated by,

$$Y = r(X) = \frac{IDF(k_1+1)}{CX+1} \quad , \quad X > 0 \tag{10}$$

where $C = k_1/ADL$. Given Equation 10 it can be shown that $r^{-1}(Y) = \frac{IDF(k_1+1)-Y}{CY}$. Now, let $f_X(x)$ be the pdf of X and $f_Y(y)$ the pdf of Y. Since function r is a monotonic and differentiable when X is positive, based on the principle of function transformations of random variables [4], we can calculate the pdf of Y as a function of the pdf of X,

$$f_Y(y) = \frac{-IDF(k_1+1)}{Cy^2} f_X(\frac{IDF(k_1+1)-y}{Cy}) \tag{11}$$

when $0 < y < IDF(k_1+1)$ and 0 otherwise.

In other words we can model the pdf of an approximation of BM25 as a function of the density function of the reverse relative term frequency. Essentially, one can plug in the above formula any distribution for the relative term frequency and get an analytical form distribution of BM25.

Based on the previous section DL/TF approximately follows a Gamma distribution. Let $\hat{k}$ and $\hat{\theta}$ are estimated parameters of the Gamma distribution from X via maximum likelihood estimation (MLE) for all retrieved documents (see Figure 2). Then, the approximated pdf of BM25 score for a single term can be reached as follows,

$$f_Y(y) = \frac{-IDF(k_1+1)}{Cy^2} Gamma(\frac{IDF(k_1+1)-y}{Cy}; \hat{k}, \hat{\theta}) \tag{12}$$

We repeat the exact same derivation in the case of language models with Jelinek-Mercer smoothing. The score for each term is computed as,

$$\text{JMLM score} = \log(\lambda \frac{TF}{DL} + C(1-\lambda)) \tag{13}$$

where $C = CTF/TN$. CTF is collection term frequency and TN is the number of unique terms in the collection. As before, we let $X = DL/TF$, then the LM score can be written as,

$$Y = r(X) = \log(\frac{\lambda}{X} + C(1-\lambda)) \tag{14}$$

Using the previous assumption that DL/TF is modeled by a Gamma distribution and since the function r is a monotonic and differentiable, after the random variable transform over X we get the pdf of the LM scores as a function of the Gamma distribution that models the reverse relative term frequency.

$$f_Y(y) = \frac{-\lambda e^y}{(e^y - C(1-\lambda))^2} Gamma(\frac{\lambda}{e^y - C(1-\lambda)}; \hat{k}, \hat{\theta}) \tag{15}$$

Figure 4 shows the comparison among the empirical histogram, the analytical model derived from the distribution of DL/TF, and the Gamma distribution obtained by MLE over BM25 and JM language model scores all retrieved documents for query *system* in TREC8 collection. As illustrated on the plots, the analytical model has more freedom than the Gamma distribution, but the Gamma is still a reasonable approximation to the term score distribution. Further, the mixture model presented in the previous section with the best-fit λ is also shown on Figure 4 (black line denoted as "Model (theory)" in the legend).

Remark on the Shape of the Distribution

Most term frequency weighting functions are nonlinear monotonically increasing functions of the raw term frequency. In BM25 Roberston's TF grows fast when the raw term frequency is small and gets gradually saturated. The parameter k_1 controls the speed of the saturation. The logarithm function in Language Models also has this saturation property but without the power of controlling the saturating speed. Therefore, the JM language model scoring function has a similar to BM25 impact on transforming the distribution of low level statistics, such as DL/TF or normalized TF to the final score distribution.

As it is illustrated in Figure 2 the typical shape of the distribution for the DL/TF tends to have a long right tail but a fast rising-up left tail. After applying a transforma-

tion function with the saturation property, the imbalance between two tails of the original distribution is alleviated, so the peak of the new distribution is right shifted, and with a shorter right tail compared to the original one. The amount of difference is dominated by the parameter controlling the saturating speed. This can be viewed in Figure 5. As k_1 becomes larger and the weighting function more linear the empirical histograms of BM25 looks more similar to the distribution of DL/TF in Figure 2. This implies that the term score distribution can also be approximated by a Gamma distribution by adjusting the shape and the scale parameters.

4.3 Summation over Query Terms

In this paper we have considered scoring functions with the following property: $score(d,query) = \sum_{t \in \text{query}} r(X_t)$, where $X_t = DL/tf(t,d)$. This class of scoring functions includes BM25, TF-IDF, some Language Models etc, but does not include scores like PageRank. Assuming term independence, the intuition for the summation $score = \sum_t r(X_t)$ is as follows:

- For non-relevant documents (low quality Q) each $r(X_t)$ will be distributed approximately as a Gamma(low shape, low scale). If the scales are approximately equal their sum follows a Gamma distribution with the same scale (gamma distribution exhibits infinite divisibility).
- For relevant documents, the mixture for each term has more effective components, thus making the sum a rich mixture, usually Gaussian like (or Gamma-like with higher scale and shape).

Thus, the distribution of the summation of several term scores could also be modeled using a Gamma distribution if we use a Gamma distribution to model the term score distribution. Figure 6 shows this summation process.

5. INFERRING THE SCORE DISTRIBUTION OF RELEVANT DOCUMENTS

In this section, we relate the score distributions for relevant and non-relevant documents with precision-recall curves. That the score distributions for relevant and non-relevant documents are related to precision-recall curves is well known and unsurprising: Given the two score distributions, one can easily infer a precision-recall curve [14], and we shall do so below as part of the treatment that follows. More interestingly, we demonstrate that one can infer the score distribution for relevant documents given a score distribution for non-relevant documents and a precision-recall curve, and we use the technique described to show that the score distributions for relevant documents will tend to have a Gaussian-like form, with a heavy right tail.

Let $f_R(s)$ and $f_N(s)$ be the score distributions for relevant and non-relevant documents, respectively. For any score threshold t, consider the set of documents whose scores are t or higher. The *recall* and *fallout* associated with this document set are easily defined in terms of $f_R(s)$ and $f_N(s)$ as follows:

$$r(t) = \int_t^\infty f_R(s)\,ds \quad (16)$$

$$fo(t) = \int_t^\infty f_N(s)\,ds. \quad (17)$$

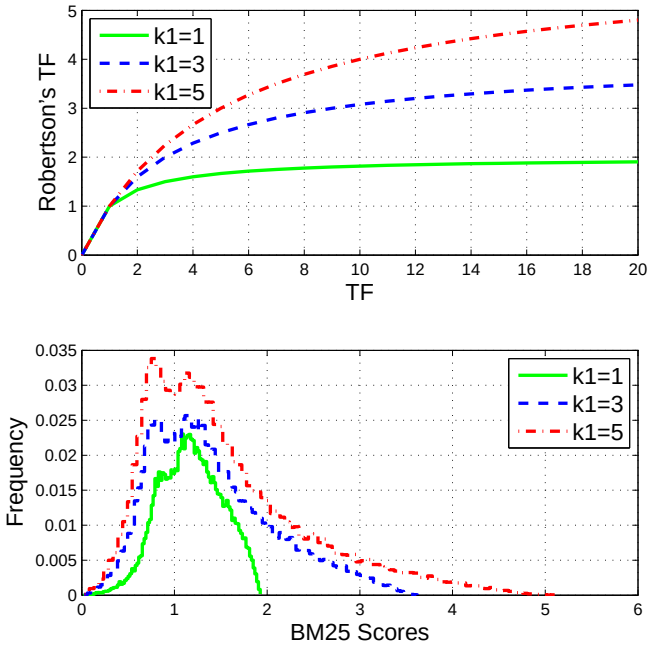

Figure 5: Roberston's TF and empirical histograms of BM25 scores with different k_1 for term *system* in TREC8

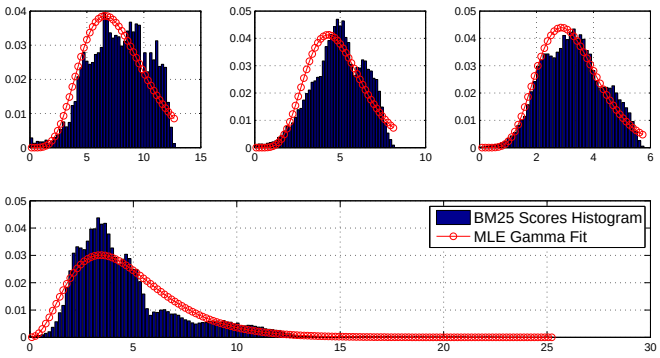

Figure 6: MLE Gamma fitting over scores of all retrieved documents for all query terms and query "Ireland Peace Talks"

Now let C be the size of the collection and let γ be the fraction of the collection that is relevant to a given query. Then there are $R = \gamma C$ total relevant documents and $N = (1-\gamma)C$ total non-relevant documents. At score t or above, there are

$$R \cdot r(t) = \gamma C \cdot r(t)$$

relevant documents and

$$N \cdot fo(t) = (1-\gamma)C \cdot fo(t)$$

non-relevant documents. Thus, the *precision* associated with this document set is simply

$$p(t) = \frac{\gamma C \cdot r(t)}{\gamma C \cdot r(t) + (1-\gamma)C \cdot fo(t)} = \frac{r(t)}{r(t) + O \cdot fo(t)} \quad (18)$$

where $O = (1-\gamma)/\gamma$ is the *odds* of non-relevance in the collection. Equations 16 and 18 are parametric equations

defining a precision-recall curve: Given the score distributions $f_R(s)$ and $f_N(s)$ (and γ), one can vary the score threshold t in Equations 16 and 18 to obtain the precision-recall curve. (A substantially similar treatment can be found in Robertson [14].)

Now suppose that one has a candidate score distribution for either relevant or non-relevant documents and one has a candidate form for a precision-recall curve: Can one *derive* a form for the other score distribution? In what follows, we show how this can be accomplished, and using the score distributions described in Section 4 and a simple form for precision-recall curves, we *infer* a form for the score distributions of relevant documents.

Consider the simple model for precision-recall curves described by Aslam and Yilmaz [3] and shown in Figure 7.

This family of precision-recall curves is defined by the following equation, implicitly parameterized by the value of R-precision rp:

$$p(r) = \frac{1-r}{1+\alpha \cdot r}. \tag{19}$$

(Here $\alpha = (1/rp-1)^2 - 1$ governs the "shape" of the curve.) While it is certainly the case that "real" precision-recall curves are never this "clean", this simple model captures many properties found in real precision-recall curves, such as high precisions at low recall levels, low precisions at high recall levels, and so on. Furthermore, Aslam and Yilmaz show that this simple model allows one to explicitly and accurately relate average precision, R-precision, precision-at-cutoff, and other seemingly disparate measures of retrieval performance.

Using such a model for precision-recall curves, we can relate the score distributions for relevant and non-relevant documents as follows. We first parameterize Equation 19 by the score threshold t, obtaining

$$p(t) = p(r(t)) = \frac{1-r(t)}{1+\alpha \cdot r(t)}. \tag{20}$$

We now equate Equations 18 and 20

$$\frac{r(t)}{r(t)+O \cdot fo(t)} = \frac{1-r(t)}{1+\alpha \cdot r(t)}$$

and solve for $r(t)$ as a function of $fo(t)$

$$r(t) = \frac{-O \cdot fo(t) + \sqrt{(O \cdot fo(t))^2 + 4(1+\alpha)O \cdot fo(t)}}{2(1+\alpha)} \tag{21}$$

Differentiating Equation 21 by t immediately establishes a closed-form relationship between the score distributions for relevant and non-relevant documents, since by Equations 16 and 17 and the Fundamental Theorem of Calculus, we have

$$\begin{aligned} r'(t) &= -f_R(t) \\ fo'(t) &= -f_N(t). \end{aligned}$$

As an example of this methodology, let us assume that the score distribution for all documents follows a Gamma distribution, as we argued in Section 4. Since the overwhelming majority of documents are non-relevant, the score distribution for non-relevant documents will then tend to follow a Gamma distribution as well. Now consider the Gamma that fits the non-relevant documents for the TREC8 query "Estonia Economy". Using this Gamma distribution for the non-relevant documents, together with a precision-recall curve[3] from the family show in Equation 19, and employing the method described above, we obtain the score distribution for relevant documents shown in Figure 8.

While Figure 8 gives just one such example, the form of this curve is quite consistent across all tested input distributions from the Gamma family (which includes the negative exponential distribution) and all tested precision-recall curves from the family defined by Equation 19: The distribution is roughly Gaussian in form, but with a heavy right tail. That the score distribution is "Gaussian-like" is much assumed (as discussed in the introduction), but the heavy right tail is also necessary to avoid problems with a simple Gaussian, such as those described by Manmatha et al. [12] and others. Figure 9 shows the typical form of the relevant document score distribution we obtained in TREC 8. We here for the first time *derive* such a form, given reasonable forms for non-relevant score distributions and precision-recall curves.

Our results in this section are descriptive rather than prescriptive, and as such, we conclude the following:

> The tendency of the score distributions for relevant documents to look Gaussian with a heavy right tail is a natural and inevitable consequence of the facts that (1) the score distributions of non-relevant documents tend to look Gamma and (2) precision-recall curves tend to have the form shown in Figure 7.

6. CONCLUSIONS

In this work, we attempt to model score distributions in a rather systematic manner. We start with a basic assumption that query terms are generated via a Poisson process and induced that the distribution the relative term frequency in a document is a inverse Gamma distribution. Following the mathematical transformations applied on the relative term frequencies by two basic ranking functions, BM25 and Language Models, we derived the distribution of the produced scores, in an analytical form and illustrate that the derived distribution can be well approximated by a Gamma distribution. Further, we also considered the score distribution for relevant documents by relating score distributions with precision-recall curves. In particular, we adopted a precision-recall curve model that has previously been proposed and given this model we presented a general mathematical framework under which given any score distribution for all retrieved documents we can derive an analytical formula for the score distribution of relevant documents. The framework is general enough such that the same derivations can be repeated for different models of precision recall curves. Finally, under the assumption that non-relevant documents follow a Gamma distribution for all retrieved documents, we show that there is a tendency of the derived distribution for the relevant documents to look Gaussian with a heavy right-hand tail.

7. ACKNOWLEDGEMENTS

We gratefully acknowledge the support provided by the NSF grants IIS-0533625 and IIS-0534482 and by the European Commission grant FP7-ICT-248347 (Accurat project) and the Marie Curie Fellowship FP7-PEOPLE-2009-IIF-254562.

[3]We set γ and $\alpha = (1/rp-1)^2-1$ to match those parameters from the BM25 run on that query.

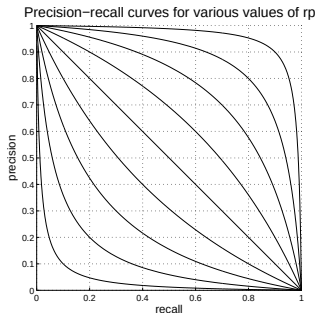

Figure 7: A family of precision-recall curves fit through the points $\{(0,1),(rp,rp),(1,0)\}$ for $rp = 0.1, 0.2, \ldots, 0.9$.

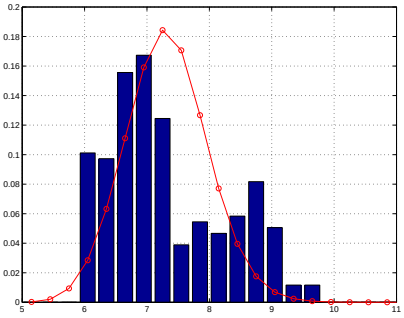

Figure 8: Inferred relevant document score distribution and empirically histogram for the TREC8 query "Estonia, economy".

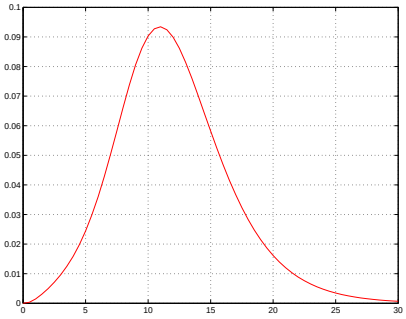

Figure 9: Typical form of the relevant document score distribution in TREC8.

8. REFERENCES

[1] A. Arampatzis, J. Kamps, and S. Robertson. Where to stop reading a ranked list?: threshold optimization using truncated score distributions. In *SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 524–531, New York, NY, USA, 2009. ACM.

[2] A. Arampatzis and A. van Hameren. The score-distributional threshold optimization for adaptive binary classification tasks. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 285–293, New York, NY, USA, 2001. ACM.

[3] J. A. Aslam and E. Yilmaz. A geometric interpretation and analysis of R-precision. In *Proceedings of the Fourteenth ACM International Conference on Information and Knowledge Management*, pages 664–671. ACM Press, October 2005.

[4] R. D. Barr and W. P. Zehna. *Probability: Modelling Uncertainty*. Addison-Wesley, 1983.

[5] C. Baumgarten. A probabilistic solution to the selection and fusion problem in distributed information retrieval. In *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 246–253, New York, NY, USA, 1999. ACM.

[6] P. N. Bennett. Using asymmetric distributions to improve text classifier probability estimates. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 111–118, New York, NY, USA, 2003. ACM.

[7] A. Bookstein. When the most "pertinent" document should not be retrieved—an analysis of the swets model. *Information Processing & Management*, 13(6):377–383, 1977.

[8] A. Bookstein and D. R. Swanson. Probabilistic models for automatic indexing. *Journal of the American Society for Information Science*, 25(5):312–318, 1974.

[9] K. Collins-Thompson, P. Ogilvie, Y. Zhang, and J. Callan. Information filtering, novelty detection, and named-page finding. In *In Proceedings of the 11th Text Retrieval Conference*, 2003.

[10] S. P. Harter. A probabilistic approach to automatic keyword indexing: Part i. on the distribution of specialty words in a technical literature. *Journal of the American Society for Information Science*, 26(4):197–206, 1975).

[11] E. Kanoulas, V. Pavlu, K. Dai, and J. A. Aslam. Modeling the score distributions of relevant and non-relevnat documents. In *In Proceedings of the 2nd International Conference on the Theory of Information Retrieval*, September 2009.

[12] R. Manmatha, T. Rath, and F. Feng. Modeling score distributions for combining the outputs of search engines. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 267–275, New York, NY, USA, 2001. ACM.

[13] M. F. Neuts and S. Zacks. On mixtures of χ^2- and f-distributions which yield distributions of the same family. *Annals of the Institute of Statistical Mathematics*, 19(1):527–536, 1966.

[14] S. Robertson. On score distributions and relevance. In G. Amati, C. Carpineto, and G. Romano, editors, *Advances in Information Retrieval, 29th European Conference on IR Research, ECIR 2007*, volume 4425/2007 of *Lecture Notes in Computer Science*, pages 40–51. Springer, June 2007.

[15] M. Spitters and W. Kraaij. A language modeling approach to tracking news events. In *In Proceedings of TDT workshop 2000*, pages 101–106, 2000.

[16] J. A. Swets. Information retrieval systems. *Science*, 141(3577):245–250, July 1963.

[17] J. A. Swets. Effectiveness of information retrieval methods. *American Documentation*, 20:72–89, 1969.

[18] M. Wiper, D. R. Insua, and F. Ruggeri. Mixtures of gamma distributions with applications. *Journal of Computational and Graphical Statistics*, 10(3):440–454, September 2001.

[19] Y. Zhang and J. Callan. Maximum likelihood estimation for filtering thresholds. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 294–302, New York, NY, USA, 2001. ACM.