Risky Business: Modeling and Exploiting Uncertainty in Information Retrieval

Jianhan Zhu, Jun Wang, Ingemar Cox
University College London, UK
jianhan.zhu@ucl.ac.uk,
jun_wang@acm.org, ingemar@ieee.org

Michael Taylor
Microsoft Research
Cambridge, U.K.
mitaylor@microsoft.com

ABSTRACT

Most retrieval models estimate the relevance of each document to a query and rank the documents accordingly. However, such an approach ignores the uncertainty associated with the estimates of relevancy. If a high estimate of relevancy also has a high uncertainty, then the document may be very relevant or not relevant at all. Another document may have a slightly lower estimate of relevancy but the corresponding uncertainty may be much less. In such a circumstance, should the retrieval engine risk ranking the first document highest, or should it choose a more conservative (safer) strategy that gives preference to the second document? There is no definitive answer to this question, as it depends on the risk preferences of the user and the information retrieval system. In this paper we present a general framework for modeling uncertainty and introduce an asymmetric loss function with a single parameter that can model the level of risk the system is willing to accept. By adjusting the risk preference parameter, our approach can effectively adapt to users' different retrieval strategies.

We apply this asymmetric loss function to a language modeling framework and a practical risk-aware document scoring function is obtained. Our experiments on several TREC collections show that our "risk-averse" approach significantly improves the Jelinek-Mercer smoothing language model, and a combination of our "risk-averse" approach and the Jelinek-Mercer smoothing method generally outperforms the Dirichlet smoothing method. Experimental results also show that the "risk-averse" approach, even without smoothing from the collection statistics, performs as well as three commonly-adopted retrieval models, namely, the Jelinek-Mercer and Dirichlet smoothing methods, and BM25 model.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H3.1Content analysis and Indexing; H.3.3 Information Search and Retrieval

General Terms

Algorithms, Experimentation, Measurement, Performance

1 Introduction

The aim of information retrieval is to find information relevant to users' needs. Probabilistic retrieval models primarily focus on building the correspondence (relevance) between users' information needs (queries) and documents. These models have led to various document ranking algorithms, including the language modelling approaches [14] and the Divergence from Randomness (DFR) model [1].

However, most estimates of the relevance of documents to queries only provide point estimation and ignore the uncertainty associated with the estimates. As such, they provide a "best guess" by maximum likelihood estimation (MLE) or maximizing a posterior probability estimation of relevancy. These approaches do not consider two fundamental research issues, namely, (i) the uncertainty associated with the match between queries and documents, and (ii) development of a utility-based ranking function that considers the corresponding uncertainty in the matches.

To investigate these issues, we have conducted both theoretical and empirical investigations. Theoretically, we return to the basic question in information retrieval: estimating the probability of relevance. We argue that this probability estimation should consider both the associated uncertainty, and the fact that these uncertain probability estimates will be used to rank documents. Our approach uses a Bayesian formulation to model the uncertainty associated with the estimation. An asymmetric loss function is used to model the risk associated with the estimation uncertainty. The resulting ranking formula incorporates both the mean and variance of the estimate and provides a single parameter that allows us to adjust the desired level of risk. By risk adjustment, our approach can adapt to a range of IR metrics that reflect different user information search strategies.

We apply the approach to a language modelling framework, and a novel risk-aware document scoring function is presented. Empirically, the proposed approach has been studied and evaluated on several TREC collections. The experiments demonstrate that significant performance gain can be achieved if we take a "risk-averse" approach. Our experiments also show that the risk-adjustment parameter can effectively adapt to different levels of risk associated with users' different retrieval strategies.

In the following, we describe previous work in Section 2, discuss our risk-aware ranking approach in Section 3, test our approach on five TREC test collections in Section 4, and conclude in Section 5.

2 Related work

Formal retrieval models have formed the basis of information retrieval research since the early 1960's. The two different *document-oriented* and *query-oriented* views on how

to assign a probability of relevance of a document to a user need have resulted in several different types of practical models [17]. The classic probabilistic model of information retrieval (the RSJ model) [18] takes the query-oriented view (or need-oriented view), assuming a given information need and choosing the query representation in order to select relevant documents. A further development of that model led to the widely-adopted term weighting function known as the BM25 formula [19]. In the document-oriented view, first proposed by Maron and Kuhn [13], the objective is to choose the appropriate document representation to match queries and judge its relevance. The language modelling approach [14] builds upon the document-oriented view. In the basic language models, a *unigram* model is estimated for each document and the likelihood of the document model with respect to the query is computed. Many variations and extensions have been proposed [6, 10, 23]. The third type of model is called the Divergence from Randomness (DFR) model [1]. In this model, the weight of a query term is calculated on the basis of the hypothesis that the more divergence there is between the within-document term-frequency and the term's frequency within the collection, the more the information is carried by the term in the document.

To resolve the uncertainty with the estimation, recent studies have focused on building a more accurate document model. Examples include smoothing from collection statistics [6, 25], the latent models [2, 7], and Dirichlet models [12]. Alternatively, a full Bayesian treatment has been introduced into the language modelling framework [23]. However, the uncertainty directly associated with the ranking problem has received much less attention.

The most relevant work can be found in [10, 26], where a risk minimization framework is proposed and documents are ranked based on an ascending order of the expected risk of a document. It has been applied to subtopic retrieval by modelling not only relevance but also redundancy, novelty, and subtopics [26]. But, nonetheless, in the resulting implementation, the studies [10, 26] still result in a point estimation, and use the mode of the posterior as opposed to integrating out model parameters. Therefore, the uncertainty of the estimation is still not fully addressed.

In previous work [27], we have conducted early study of utilizing the variance of the prediction, where we assumed the relevance score follows a Gaussian distribution. In this paper, we derive a more general and practical form of the risk-adjustable ranking function, and apply the ranking function to language models.

3 Moment-based Ranking

The Probability Ranking Principle (PRP) of information retrieval [16] implies that ranking documents in descending order by their probability of relevance produces optimal performance under a "reasonable" assumption, i.e. the relevance of a document is independent of other documents in the collection [22].

A classic solution to estimating the probability of relevance is to treat it (or the parameters of the assumed distribution) as a fixed unknown constant that does not have an associated distribution. Existing relevance-based retrieval models, such as the RSJ model [18], two-Poisson model [19], and resulting BM25 formula [20], all belong to this category. The language modelling approaches, although not designed to directly estimate the relevance of documents, also consider the model parameters as unknown fixed constants.

The main drawback of this approach is that exact measures of the uncertainty associated with the estimation are not handled in a principled manner, either for the probability of relevance or the model parameters. As a consequence, unreliably-estimated documents may be ranked highly in the ranked list, reducing the retrieval performance of the top-N returned documents.

To address this problem, this paper takes an alternative Bayesian view point. We consider the parameters, either the probability of relevance or the model parameters (in the language models), as random variables, and calculate their posterior probability given the observed data. Without losing generality, let us denote θ as the estimation of the correspondence between a query $\mathbf{q}$ and a document $\mathbf{d}$. In the relevance models, the estimation θ is equal to $P(r|\mathbf{d},\mathbf{q})$ (r denotes relevance), while in the language models, it is equal to $p(\mathbf{q}|\theta_{\mathbf{d}})$ ($\theta_{\mathbf{d}}$ is the estimated document model for $\mathbf{d}$). Formally, the posterior probability of θ is written as:

$$p(\theta|O) \propto p(O|\theta)p(\theta), \tag{1}$$

where O denotes the observations we obtained so far. It may include the features from the document and the query, or relevance feedback from the user.

By associating an uncertainty with each document's estimate of relevance, we are able to consider more sophisticated ranking algorithms. For example, consider two documents, the first of which has the highest estimated relevancy but with a corresponding high uncertainty, and the second document having a slightly lower relevancy but with a corresponding low uncertainty. If we can only choose one document, which should we choose? If the first document is chosen, then there is a chance that it is "very relevant", but equally, there is a chance that it is much less relevant ("slightly relevant") than we expected. Herein lies the risk. If we choose the first document, the result may be "very relevant" or "slightly relevant". By choosing the second document, we reduce the variability (risk), but now the chance that the document is "very relevant" is smaller, as is the chance that it is "slightly relevant". Thus, there is a choice to be made between consistency, at the expense of relevancy, or relevancy at the expense of consistency.

To explicitly model such uncertainty with ranking, this paper introduces a loss function $L(\tilde{\theta},\theta)$. It denotes the loss when we choose $\tilde{\theta}$ as our ranking score while the true value is θ. We do not take a point estimation, e.g., maximizing a posterior probability, as in [6, 25]. Instead, we integrate out the unknown model parameters. This provides a natural and principled way to deal with the uncertainty both related to the ranking and parameter estimation. Marginalizing out the unknown true θ, we obtain the following expected loss for a given $\tilde{\theta}$:

$$E^{\theta}(L(\tilde{\theta},\theta)) = \int L(\tilde{\theta},\theta)p(\theta|O)d\theta, \tag{2}$$

where E denotes the expectation. The optimal ranking score should minimize the expected loss function in Eq. (2). We therefore derive our Bayesian optimal ranker as follows:

$$\tilde{\theta}^{B} = \operatorname*{argmin}_{\tilde{\theta}} E^{\theta}(L(\tilde{\theta},\theta)), \tag{3}$$

where $\tilde{\theta}^{B}$ is our Bayesian ranker. It is indeed the Bayes estimator of θ with respect to the loss function L. Notice that the development so far is similar in spirit to the framework introduced in [10, 26]. However, we shall see later that our full Bayesian treatment results in a completely different document scoring function. On the other hand, the aim and formulation of our Bayesian ranker also depart from those in [23]. We intend to address the ranking uncertainty, while

the authors in [23] aimed at dealing with the uncertainty when building a document model (the document model is marginalized out when generating a query).

In the remainder of this section, we will introduce an asymmetric loss function, and derive an approximation to Eq. (3). The loss function will then be applied to language models.

3.1 Asymmetric loss function

One way to model ranking uncertainty is to introduce an asymmetric utility function that penalizing under-estimating less that over-estimating the rank score, or vice versa. We introduce one such function called the linear/exponential (LINEX) asymmetric loss function [24] given by:

$$L(\tilde{\theta}, \theta) = e^{b(\tilde{\theta}-\theta)} - b(\tilde{\theta} - \theta) - 1, \tag{4}$$

where $b \in \mathbb{R}$ is the parameter to balance the loss. Roughly speaking, the loss of LINEX increases exponentially on one side of zero and increases linearly on the other side. One of the advantages of the LINEX loss function is that we can easily balance the loss on both sides by adjusting a single parameter.

Substituting Eq. (4) into the expected loss in Eq. (2) gives:

$$E^{\theta}(L(\tilde{\theta}, \theta)) = e^{b\tilde{\theta}} E^{\theta}(e^{-b\theta}|O) - b(\tilde{\theta} - E^{\theta}(\theta|O)) - 1, \tag{5}$$

where we use $\tilde{\theta}$ as the ranking score while the true value is θ. Minimizing Eq. (5) with respect to $\tilde{\theta}$ yields the optimal ranking score (For detailed information, we refer to [24]):

$$\tilde{\theta^B} = -(1/b)\ln(E^{\theta}(e^{-b\theta}|O)) \tag{6}$$

In Eq. (6), $E^{\theta}(e^{-b\theta}|O)$ is the moment generating function (MGF), which generates the moments of the probability distribution θ, and $\ln(E^{\theta}(e^{-b\theta}|O))$ is called the cumulant-generating function of θ. Based on [9, 11], we define the cumulant generating function via the characteristic function as:

$$\ln(E^{\theta}(e^{-b\theta}|O)) = \sum_{n=1}^{\infty} \kappa_n \frac{(-b)^n}{n!}, \tag{7}$$

where the cumulants κ_n are given by derivatives of the cumulant-generating function [9, 11] as: $\kappa_1 = \mu$, $\kappa_2 = \mu_2 = \sigma^2$, $\kappa_3 = \mu_3$, $\kappa_4 = \mu_4 - 3\mu_2^2$, $\kappa_5 = \mu_5 - 10\mu_2\mu_3$, ..., where μ is the mean, σ^2 is the variance, and μ_n is the n-th moment about the mean of θ.

The risk-adjustable ranking function then becomes:

$$\begin{aligned}\tilde{\theta^B} &= -(1/b)\sum_{n=1}^{\infty} \kappa_n \frac{(-b)^n}{n!} \\ &= \mu - b\sigma^2/2 + \kappa_3 b^2/6 - \kappa_4 b^3/24 + \ldots,\end{aligned} \tag{8}$$

where $\tilde{\theta^B}$ is our Bayesian optimal ranker, μ denotes the mean of the posterior probability of θ, σ^2 is its variance, κ_3 is a measure of the lopsidedness of the posterior distribution[1], and κ_4 is a measure of whether the distribution is tall and skinny or short and squat.

Eq. (8) gives a general formula to rank documents when considering asymmetric loss. It shows that to address the uncertainty, the final ranking is equal to the mean adjusted by a function of the variance (weighted cumulants). Depending on the specific probability distribution θ, one can derive different forms of Eq. (8). For example, if θ conforms to a normal distribution, the cumulants are κ_1=μ, κ_2=σ^2, and $\kappa_3 = \kappa_4 = ... = 0$.

[1]The third central moment is called the skewness. A distribution that is skewed to the left (the tail of the distribution is heavier on the left) will have a negative skewness. A distribution that is skewed to the right (the tail of the distribution is heavier on the right), will have a positive skewness.

3.2 Risk-aware language models

In this section, we present an application of the proposed document ranking approach under the language modelling framework. However, it is worth mentioning that the proposed method is generally applicable to any probabilistic retrieval model.

3.2.1 Unigram language models: In the language modelling framework, document ranking is primarily based on the following two steps. First, "choose" a generative model for the target document $P(\theta|\mathbf{d})$, and then generate the query terms using that model $p(\mathbf{q}|\theta)$ [25], where $\mathbf{d}$ and $\mathbf{q}$ can be formally represented as vectors of indexed term counts as:

$$\mathbf{q} \equiv (q_1, ..., q_i, ..., q_{|V|}),\ \mathbf{d} \equiv (d_1, ..., d_i, ..., d_{|V|}), \tag{9}$$

where q_i (d_i) is the number of times the term i appears in the query (document) and $|V|$ is the size of a vocabulary.

The *unigram* language models consider a Multinomial distribution θ for each document, where $\theta = (\theta_1, ..., \theta_i, ..., \theta_{|V|})$, and $\sum_i \theta_i = 1$. The primary task is, for each candidate document, to estimate $\theta_i = p(i|\theta)$ for each query term i. A straightforward approach is to apply the maximum likelihood estimation (MLE) method:

$$\hat{\theta}_i = \frac{d_i}{\sum_i d_i} = \frac{d_i}{|\mathbf{d}|}, \tag{10}$$

where $|\mathbf{d}| \equiv \sum_i d_i$ is the document length.

However, estimating from one single document is unreliable due to small data samples. One of the common solutions is to use the posterior probability as opposed to the likelihood function. Using the conjugate prior of the Multinomial distribution (the Dirichlet) results in the following posterior probability:

$$\begin{aligned}p(\theta|\mathbf{d}, \alpha) \propto p(\mathbf{d}|\theta)p(\theta|\alpha) &= \prod_i (\theta_i)^{d_i} \prod_i (\theta_i)^{\alpha_i - 1} \\ &= \prod_i (\theta_i)^{d_i + \alpha_i - 1},\end{aligned} \tag{11}$$

where the prior $p(\theta|\alpha)$ is deployed to incorporate prior knowledge of the model parameters, and $\alpha \equiv (\alpha_1, \ldots, \alpha_{|V|})$ is the parameter of the Dirichlet prior. Maximizing the posterior probability (taking the mode [5]) gives a general form of the language modelling approaches:

$$\hat{\theta}_i^L = \frac{d_i + \alpha_i - 1}{\sum_i (d_i + \alpha_i) - |V|} = \frac{c_i - 1}{\hat{c} - |V|}, \tag{12}$$

where, for simplicity, we denote $c_i \equiv d_i + \alpha_i$ and $\hat{c} \equiv \sum_i (d_i + \alpha_i)$. The so-called hyper-parameter α acts as pseudo-counts, and can be used to integrate collection statistics for smoothing the estimation [6, 25]. A further discussion on this can be found in Section 3.3.

3.2.2 Our risk-aware approach: Previous studies have demonstrated that introducing the pseudo-counts in Eq. (12) alleviates the problem of small data samples to some extent. Yet, using such point estimation alone to rank documents is insufficient.

In this section, we derive an alternative, risk-aware language model from Eq. (8). As an approximation, we apply Eq. (8) to the posterior probability distributions of individual terms, and leave the application of Eq. (8) to the

whole query's posterior probability distribution to future work. Therefore, the cumulants of the Dirichlet distribution can be obtained on the basis of the moments of the Dirichlet distribution as follows:

$$\kappa_1 = \bar{\theta}_i = \frac{c_i}{\hat{c}}, \quad \kappa_2 = \sigma_i^2 = \frac{c_i(\hat{c} - c_i)}{\hat{c}^2(\hat{c}+1)}, \quad ... \qquad (13)$$

where $\bar{\theta}_i$ and σ_i^2 are the mean and variance of θ_i, respectively.

For language models of documents from a reasonable large dataset, c_i is generally much smaller than $\hat{c}$. Therefore, we can easily show that the value of κ_n decreases sharply as n increases. Our experiments (not shown) also show that our risk-adjustment approach is effective when only taking into account the first and second moments, i.e., mean and variance, and the introduction of higher moments does not affect the result significantly. Therefore, an approximation that considers both the first and second moments provides an excellent trade-off between accuracy and efficiency:

$$\tilde{\theta_i^B} \approx \bar{\theta}_i - b\sigma_i^2/2 \qquad (14)$$

Assuming term independence in a unigram language model, we reach our final ranking score of a document $\mathbf{d}$ for a given query $\mathbf{q}$ as

$$\tilde{\theta^B} \approx \prod_{i=1}^{|V|}(\bar{\theta}_i - b\sigma_i^2/2)^{q_i} = \prod_{i=1}^{|V|}(\frac{c_i}{\hat{c}} - \frac{b}{2}\frac{c_i(\hat{c}-c_i)}{\hat{c}^2(\hat{c}+1)})^{q_i}, \qquad (15)$$

where $\tilde{\theta^B}$ denotes a risk-aware ranker in the language modelling framework. Again $c_i \equiv d_i + \alpha_i$ and $\hat{c} \equiv \sum_i (d_i + \alpha_i)$.

Therefore, a single parameter b adjusts the risk in ranking. A positive b produces a risk-averse (conservative) ranking where unreliably-estimated documents (with big variance) are given a lower rank (in the fear that they are less relevant than estimated). The bigger the parameter b is, the more conservative the ranking is. On the other hand, a negative b gives a risk-inclined ranking. In this case, unreliably-estimated documents are given a higher rank (in the hope that they are more relevant than estimated).

3.3 Discussions

Eq. (15) consists of three types of parameters. $\{d_i\}_{i=1}^{|V|}$ accounts for the document features; parameter b provides a natural way to address risk, and hyper-parameter α is a place to integrate subjective or prior knowledge. The discussions on the parameters, variance, and their relationships now follow.

3.3.1 The hyper-parameter: Various choices for the hyper-parameter α lead to different ways to smooth the estimation [23]. The proposed ranker has the same ability to incorporate collection statistics into the estimation. Jelinek-Mercer smoothing [8, 25] uses a single coefficient λ to linearly interpolate the maximum likelihood model with the collection model. It is straightforward to adopt Jelinek-Mercer smoothing into our model if we set $\alpha_i = \frac{\lambda|\mathbf{d}|}{1-\lambda} \cdot \frac{n(i,\mathbf{D})}{|\mathbf{D}|}$, where $n(i,\mathbf{D})$ denotes the number of occurrences of term i in the collection, and $|\mathbf{D}|$ is the collection size.

Alternatively, the Dirichlet smoothing approach [25] adjusts the parameter λ on the basis of document length $|\mathbf{d}|$ by setting $\lambda = \frac{\mu}{|\mathbf{d}|+\mu}$, where μ is the Dirichlet smoothing parameter, and the parameter α_i for the Dirichlet smoothing is $\alpha_i = \mu\frac{n(i,\mathbf{D})}{|\mathbf{D}|}$.

3.3.2 Factors influencing variance: In previous work, the relationship between the variance of parameters and collection statistics such as IDF (inverse document frequency) [4] has been studied. In this section, we show that the variance is also related to the document length. Incorporating

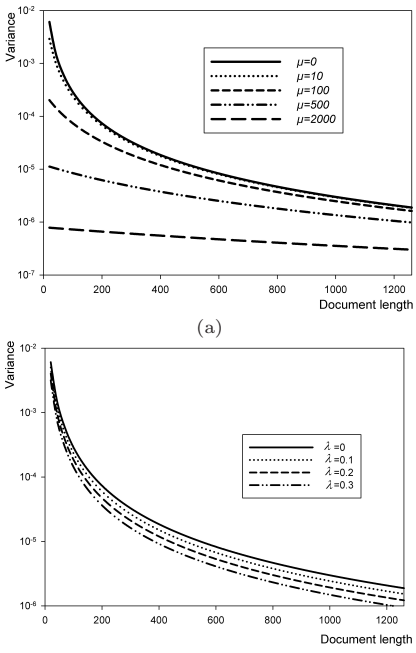


Figure 1: Effect of background smoothing on variance. (a)Under Dirichlet smoothing parameter μ. (b)Under Jelinek-Mercer smoothing parameter λ.

variance in our framework provides an alternative justification for using the document length. To understand this, let us set $\alpha_i = 0$ and substitute $c_i = d_i$ and $\hat{c} = |\mathbf{d}|$ into Eq. (13), to obtain:

$$\sigma_i^2 = \frac{d_i(|\mathbf{d}| - d_i)}{|\mathbf{d}|^2(|\mathbf{d}|+1)} = \frac{\bar{\theta}_i(1-\bar{\theta}_i)}{(|\mathbf{d}|+1)} \qquad (16)$$

This equation illustrates that the variance monotonically decreases with respect to the document length. In a risk-averse setting ($b > 0$), the ranker will favor longer documents over shorter ones, if both documents have the same estimated mean. This is a good feature because short documents may yield unreliable estimates.

To further study the effect of document length on the variance σ_i^2, and the influence of the background smoothing parameters, we plot the relationship between document length and variance in Fig. 1 (a) and (b) for the Jelinek-Mercer and Dirichlet smoothing language models, respectively.

For simplicity, we set $d_i = 3$ and $\frac{n(i,|\mathbf{D})}{|\mathbf{D}|} = 0.0001$, which are typical values for a dataset. We plot the variance for different values of Dirichlet parameter μ as shown in Fig. 1 (a). We see that large values of μ significantly reduce variance, especially for short documents. And the longer a document, the smaller the variance. When μ is typically set around 2000 [25], the variance remains relative small irrespective of the document length. Conversely, for the Jelinek-Mercer smoothing method (shown in Fig. 1 (b)), we see that the variance has a different trend when λ is typically set between 0.1 and 0.3 [25].

This explains why the Dirichlet smoothing method that utilizes the document length generally outperforms the Jelinek-Mercer method for document ranking. Since the variance under the Dirichlet smoothing method remains small irrespective of document length, it seems more preferable to apply our approach to the Jelinek-Mercer smoothing language model. By combining our risk-aware approach with

Table 1: Overview of the five TREC test collections.

Name	Description	# Docs	Topics	# Topics
TREC2007 enterprise track document search task	CSIRO website crawl	370,715	1-50 minus 8, 10, 17, 33, 37, 38, 46, 47	42
TREC2001 web track	WT10g web collection	1,692,096	501-550	50
TREC 2004 robust track	TREC disks 4, 5 minus CR	528,155	301-450 and 601-700 minus 672	249
Robust2004 hard topics	TREC disks 4, 5 minus CR	528,155	Di¡cult Robust2004 topics	50
TREC8 ad hoc task	TREC disks 4, 5 minus CR	528,155	401-450	50

the Jelinek-Mercer language model, we provide an alternative way to consider both collection statistics such as IDF[2] and document length. Our experimental results in Section 4 demonstrate that the combined approach can generally outperform the Dirichlet smoothing language model.

3.3.3 The risk-adjustable parameter: In Eq. (15), the final ranking score relies on not only the estimated mean, but, equally importantly, the variance associated with the estimate. To the best of our knowledge, this has not been previously studied. By setting the parameter b we adjust how much risk we are willing to take when ranking documents. When subtracting a function of the weighted variance from the mean ($b > 0$), the ranker takes a more conservative approach which effectively reduces the overall uncertainty (variability) in the top-N ranking list. This is at the expense of dropping documents that are potentially "very relevant" but also very uncertain. Conversely, when adding a function of the weighted variance to the mean ($b < 0$), the ranker takes more risk, and thereby increasing the uncertainty (variability) in the top-N ranking list. In this case, it is hoped that a relevant document is "very relevant", but it is just as likely to only be "slightly relevant" document. When b is set to 0, the variance is ignored and the ranking score is equivalent to conventional language models. Certainly, there are many factors influencing the optimal setting of b. In Section 4, we limit our study to investigating what is the best ranking strategy for the TREC evaluation. An investigation into formal optimization is left for future work.

4 Empirical Study and Evaluation

We examined five TREC collections described in Table 1. TREC2004 robust track is evaluated with an emphasis on the overall reliability of IR systems, i.e. minimizing the number of queries for which a system performs badly. 50 "difficult" topics among the TREC2004 robust track topics can help us understand whether our approach is effective for both "ordinary" and "difficult" topics.

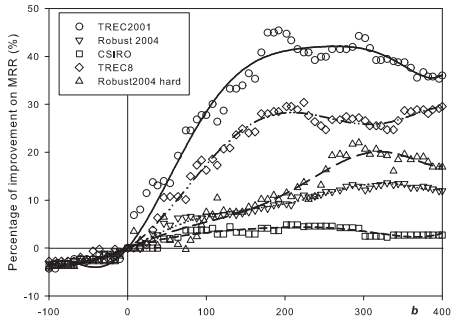
Documents were stemmed using the Porter stemmer, but not stopped at indexing time. Instead, stopping is carried out at query time using a standard list of 421 stopwords. In all our experiments, only the title portion of the TREC topics are used as queries.

Recall that in Fig. 1 (a), for large values of μ in the Dirichlet smoothing, the variance becomes relatively small, and consequently, our risk adjustment becomes less effective. Therefore, it is more favorable to combine our approach with the Jelinek-Mercer smoothing language model. Therefore our experiments in Section 4 focus on the evaluation of the risk-averse approach combined with the Jelinek-Mercer smoothing language model.

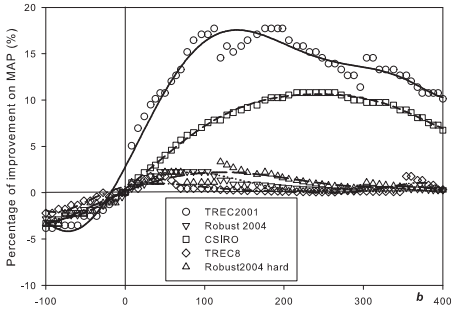
4.1 The risk-adjustable parameter

In this section, we first investigate whether a risk-averse or risk-inclined approach is more helpful in improving retrieval performance. We then study what the key factors are that affect the optimal choice of the risk adjustment parameter.

[2]Language models provide theoretical justification for the use of IDF in document ranking [25]



(a)



(b)

Figure 2: Plots of the percentage of gain on metrics against parameter b on five collections (under Jelinek-Mercer smoothing). (a) MRR (b) MAP

4.1.1 Risk-averse or risk-inclined? Let us first look at the effect of the risk adjustment parameter over different TREC collections. The performance is measured by MRR (Mean Reciprocal Rank) and MAP (Mean Average Precision), the two main performance measures in TREC. We vary the value of b between -100 and 400, and the percentage improvement[3] for these b values and fitted curves based on the data points, are shown in Fig. 2. The results are reported on the risk-adjusted Jelinek-Mercer smoothing language model where $\lambda = 0.1$. Similar results were obtained for λ=0.2, 0.3, and 0.4. When we apply our approach to a model without background smoothing (assign 0.5 to the zero counts), we also obtain similar results, and the improvements for MRR and MAP for $b > 0$ are even greater.

We can see from Fig. 2 that risk aversion applies to all five collections. When taking a risk-inclined approach, i.e. $b < 0$, the MAP and MRR for all five collections degrade. In contrast, a risk-averse approach where $b > 0$ can improve the MRR and MAP on all five collections. The improvements on MRR for four of these collections are statistically significant [4], and on MAP are statistically significant for two collections. It is worth noting that the improvements on MRR for the WT10g collection can be above 40%, and on MAP for the WT10g collection can be above 15%. This suggests that for a larger collection like the WT10g, risk adjustment is even more favorable than for smaller collections.

Note that our risk adjustment approach is robust/stable for a value of b anywhere between 100 and 400, for all collec-

[3]The percentage of improvement (or gain) on MRR and other metrics is based on the improvement of the risk adjusted model over the model where $b = 0$.

[4]We tested statistical significance with t tests (one-tail critical values for significance levels α=0.05).

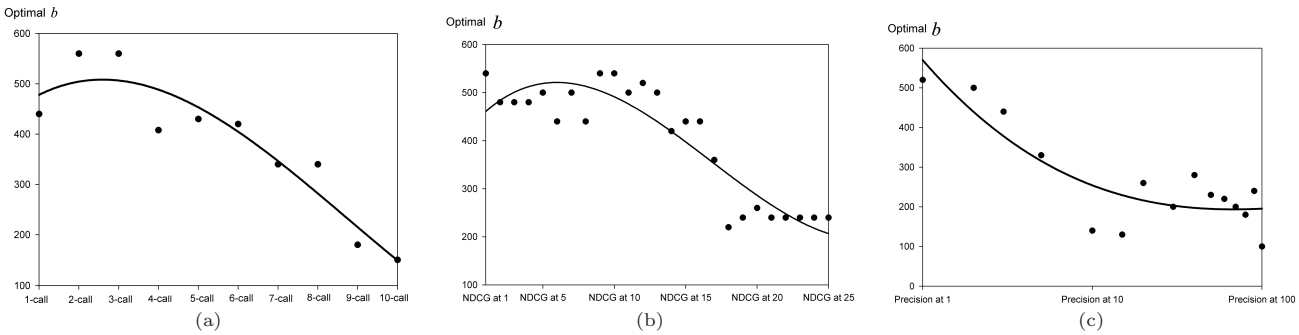


Figure 3: Effect of risk adjustable parameter b on k-call, and NDCG and Precision at m, respectively (under the Jelinek-Mercer smoothing language model) (a) Optimal b for 1 to 10-call. (b) Optimal b for NDCG at 1 to 25. (c) Optimal b for Precision at 1 to 100, where we take the log of the Precision at m axis.

tions. Our approach is effective for the Robust2004 and Robust2004 50 difficult topics since the improvements on MRR for these two can be above 10% and 20%, respectively.

4.1.2 Key factors affecting the ranking risk: We now investigate how our model behaves under a risk-sensitive metric called k-call at 10, or simply k-call, proposed in [3]. Given a ranked list, k-call is one if at least k of the top-10 documents for a query are relevant. Otherwise, k-call is zero. Averaging over multiple queries yields mean k-call. The two extremes are 10-call, an *ambitious* metric of perfect precision: returning only relevant documents, and 1-call as a *conservative* metric that is satisfied with only one relevant document. Therefore, a *risk-averse* approach, which can reliably find one relevant document with small variance, is preferred for 1-call, while a *risk-inclined* approach, which gives small weight to the variance, is favored for 10-call [3].

Fig. 3 (a) illustrates the relationship between the optimal values of b and 1 to 10-call on the WT10g collection. The data points are the optimal b values versus these metrics, and the curve is fitted based on the data points. The figure demonstrates that when k decreases, the optimal value of b tends to increase. It confirms that the risk adjustment parameter b controls how much risk we are going to take when ranking documents. A bigger value of b gives a more risk-averse (conservative) ranking.

Next we study the effect of ranking positions on b. In Fig. 3 (b) and (c), we plot the optimal b value for NDCG at 1 to 25, and Precision at 1 to 100 on the Robust2004 collection, respectively. The curves are fitted based on the data points. The figures illustrate that when the cut-off point is under 14 for NDCG, and 5 for Prec, respectively, the optimal values of b for the two metrics are both large, i.e. around 500, which estimates ranking scores conservatively by weighting variance highly. However, when the cut-off point is above 17 for NDCG, and 10 for Precision, respectively, the optimal values of b for the two metrics are both relatively small, i.e. around 200, which subtracts a lower function of variance from the mean. Such behavior suggests that lower rank position favors more conservative ranking (bigger b).

4.2 Performance

4.2.1 Risk adjustment with Jelinek-Mercer smoothing: We compare our risk-averse approach with both the Jelinek-Mercer and Dirichlet smoothing language models. We carried out 5-fold cross validation on each of the five collections. For each collection, we randomly divided the topics into five partitions. For each partition, topics in the other partitions were used to estimate the parameters on the basis of the MAP, and performance for the partition is evaluated with the trained parameters. Results were averaged over the 5 partitions, and reported in Table 2.

The results on a number of metrics including MRR, MAP, NDCG, Precision, and k-call are reported in Table 2. When comparing our approach with the Jelinek-Mercer smoothing method, 57 out of the 60 reported improvements are positive, and 45 of these improvements are statistically significant. Note that 17 of these improvements exceed 20%. When compared with the Dirichlet smoothing method, 52 out of the 60 reported improvements are positive, 28 of these improvements are statistically significant, and 13 of these improvements exceed 15%. We therefore conclude that a combination of our risk adjustment approach with Jelinek-Mercer smoothing can largely outperform both the Jelinek-Mercer and Dirichlet smoothing methods on all collections.

4.2.2 Risk adjustment without background smoothing: The current best performing ranking algorithms, such as BM25 and the language models, all use collection statistics, such as the IDF of query terms, the collection size, and the average document length. However, the proposed ranker reveals that incorporating the moments, e.g. variance, into document ranking provides an alternative way to "correct" the estimation without using collection data. Independence from collection statistics is beneficial when access to the entire data set is prohibited in situations such as peer-to-peer network environments [15], and meta-search engines.

In this section, we compare our approach without background smoothing with the state-of-the-art language models and the BM25 model. We used 5-fold cross-validation, and the results are reported in Table 3.

Table 3 shows that when comparing our approach without smoothing to the Jelinek-Mercer smoothing language model, the improvements for 53 out of the 60 reported results are positive, 43 of the improvements are statistically significant, and 14 improvements are above 20%. When comparing our approach with the Dirichlet smoothing language model, 46 out of the 60 reported improvements are positive, 26 of these improvements are statistically significant, and 13 improvements are above 15%. When comparing our approach with the BM25 model, 25 out the 60 reported improvements are positive, 10 of these improvements are statistically significant, and 8 improvements are 9% or above.

4.2.3 Discussion: The strong performance of our approach even without background smoothing demonstrates the effectiveness of our risk-adjustment approach in document ranking.

Our risk-averse approach conservatively estimates document relevance score, therefore, putting documents with both high relevance and low variance higher up in the ranked list. Therefore, this risk-averse approach effectively improves

Table 2: Risk-adjusted Jelinek-Mercer smoothing language model vs. Jelinek-Mercer and Dirichlet smoothing language models. Five lines in each cell are performance of risk-adjusted model, Jelinek-Mercer smoothing, gain of our approach over the Jelinek-Mercer smoothing, Dirichlet smoothing, and gain of our approach over the Dirichlet smoothing, respectively. Statistically significant improvements are marked with "*".

Measures	CSIRO	WT10g	Robust	Robust hard	TREC8	Measures	CSIRO	WT10g	Robust	Robust hard	TREC8
MRR	0.89	0.634	0.618	0.455	0.621	Prec@10	0.738	0.364	0.41	0.256	0.433
	0.849	0.436	0.544	0.373	0.474		0.653	0.3	0.371	0.231	0.404
	+4.8%	+45.4%*	+13.6%*	+22.0%*	+31.0%*		+13.0%*	+21.3%*	+10.5%*	+10.8%*	+7.2%*
	0.782	0.55	0.596	0.393	0.606		0.667	0.338	0.381	0.211	0.413
	+13.8%*	+15.3%*	+3.7%	+15.8%*	+2.5%		+10.6%*	+7.7%*	+7.6%*	+21.3%*	+4.8%
MAP	0.41	0.193	0.226	0.092	0.23	Prec@100	0.456	0.162	0.178	0.131	0.231
	0.37	0.158	0.22	0.09	0.226		0.406	0.147	0.174	0.127	0.222
	+10.8%*	+22.2%*	+2.7%	+2.2%	+1.8%		+12.3%*	+10.2%*	+2.3%	+3.2%	+4.1%
	0.398	0.202	0.221	0.089	0.225		0.44	0.18	0.169	0.126	0.205
	+3.0%	-4.5%	+2.3%	+3.4%	+2.2%		+3.6%	-10.0%	+5.3%	+4.0%	+12.7%*
NDCG	0.655	0.480	0.470	0.317	0.490	1-call	1.0	0.9	0.9	0.82	0.96
	0.587	0.398	0.396	0.252	0.422		0.98	0.82	0.831	0.8	0.88
	+11.5%*	+20.5%*	+18.6%*	+25.6%*	+16.0%*		+2.0%	+9.8%*	+8.3%*	+2.5%	+9.1%*
	0.651	0.477	0.483	0.312	0.484		0.98	0.86	0.847	0.74	0.88
	+0.5%	+0.5%	-2.8%	+1.5%	+1.1%		+2.0%	+4.7%	+6.3%*	+10.8%*	+9.1%*
NDCG@10	0.186	0.154	0.184	0.085	0.158	6-call	0.78	0.24	0.301	0.14	0.34
	0.170	0.141	0.169	0.078	0.140		0.62	0.2	0.273	0.08	0.3
	+9.5%*	+9.5%*	+8.6%*	+9.6%*	+12.9%*		+25.8%*	+20.0%*	+10.3%*	+75.0%*	+13.3%*
	0.162	0.152	0.179	0.077	0.154		0.66	0.22	0.273	0.04	0.3
	+14.9%*	+1.6%	+2.5%	+11.0%*	+2.6%		+18.2%*	+9.1%*	+10.3%*	+250.0%*	+13.3%*
NDCG@100	0.381	0.308	0.335	0.185	0.325	8-call	0.6	0.16	0.157	0.02	0.22
	0.355	0.262	0.292	0.159	0.287		0.44	0.08	0.129	0.02	0.2
	+7.4%*	+17.5%*	+14.6%*	+16.6%*	+13.4%*		+36.4%*	+100.0%*	+21.7%*	0.0%	+10.0%*
	0.367	0.295	0.331	0.173	0.315		0.44	0.14	0.108	0.005	0.18
	+3.9%	+4.4%	+1.1%	+7.2%*	+3.3%		+36.4%*	+14.3%*	+45.4%*	+300.0%*	+22.2%*
Prec@1	0.148	0.064	0.056	0.046	0.075	10-call	0.34	0.04	0.048	0.0	0.06
	0.14	0.057	0.055	0.046	0.074		0.26	0	0.036	0.0	0.04
	+5.7%	+12.3%*	+1.8%	0.0%	+1.4%		+30.8%*	–*	+33.3%*	–	+50.0%*
	0.148	0.064	0.055	0.047	0.075		0.2	0.02	0.024	0.0	0.02
	0.0%	0.0%	+1.8%	-2.1%	0.0%		+70.0%*	+100.0%*	+100.0%*	–	+200.0%*

the MRR metric as evidenced by the fact that our method outperforms all three state-of-the-art models on all collections in terms of the MRR metric as demonstrated in both Table 2 (when combined with the Jelinek-Mercer smoothing language model) and Table 3 (when without background smoothing).

On the other hand, we got mixed results when comparing our risk-averse approach with the Dirichlet smoothing method and BM25 model in terms of the MAP and NDCG metrics. This is due to the fact that our method is more effective to retrieve the first relevant document, while metrics such as MAP and NDCG aim to reflect the overall retrieval performance [21] by taking into account the top-N documents, where N is set as 1000 in our experiments.

The above findings confirm our assumption in Section 3.3.2 that improvements can be made by naturally integrating variance to address uncertainty without using collection statistics. As we discussed, this is especially beneficial for retrieval applications where the information about the entire dataset is inaccessible.

In our cross-validation for our risk-adjustable approach both with and without background smoothing reported in Section 4.2.1 and 4.2.2, respectively, the range of the trained risk-adjustment parameter b is always between 100 and 400, showing that a risk-averse approach is preferable.

5 Conclusions and future work

We have presented an approach to document ranking in information retrieval that explicitly models the associated uncertainties. The proposed approach led to a risk-aware retrieval model. One of the merits of our approach is that such uncertainty is addressed in a principled way. Further, by using an asymmetric loss function, based on a LINEX cost function, only a single parameter is needed to adjust the risk preference. We have applied our method to the language modelling framework, and a practical risk-aware document scoring function was derived.

The empirical evaluations on five TREC collections have shown that a risk-averse approach significantly improves the retrieval performance over a number of IR metrics. The proposed approach dramatically improves the performance of the Jelinek-Mercer smoothing language model on a number of metrics; a combination of our approach and the Jelinek-Mercer smoothing approach can outperform the Dirichlet smoothing approach on all five collections. We also demonstrated that adding a function of the variance into document ranking provides an alternative way to improve the performance without smoothing from collection data, and the performance of our approach without background smoothing was comparable with that of the state-of-the-art models, namely, Jelinek-Mercer and Dirichlet smoothing language models, and the BM25 model.

Since our approach is more effective to the top-ranked positions, our approach achieved significant overall performance gain in terms of the MRR metric over the current models. Furthermore, by varying the risk adjustment parameter, our approach can effectively adapt to different user search behaviors, such as risk-inclined or risk-averse.

We hope that our analysis will increase the awareness of using the *second moment* for information retrieval modelling. The research challenges that need to be addressed in the future include: optimizing risk adjustment on the basis of IR metrics and ranking positions, incorporating term or document correlation in the analysis, and investigating how the correlation can reduce our ranking risk.

6 References

[1] G. Amati and C. J. V. Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst.*, 20(4):357–389, 2002.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. In *Proc. of NIPS*, pages 601–608, 2001.

[3] H. Chen and D. R. Karger. Less is more: probabilistic models for retrieving fewer relevant documents. In *Proc. of SIGIR '06*, pages 429–436, New York, NY, USA, 2006. ACM.

Table 3: Risk-adjusted method without background smoothing vs. Jelinek-Mercer and Dirichlet smoothing LMs, and the BM25 model. Seven lines in each cell are performance of our approach, Jelinek-Mercer smoothing, gain of our approach over the Jelinek-Mercer smoothing, Dirichlet smoothing, gain of our approach over the Dirichlet smoothing, BM25 model, and gain of our approach over the BM25 model, respectively. Statistically significant improvements are marked with "*".

Measures	CSIRO	WT10g	Robust	Robust hard	TREC8	Measures	CSIRO	WT10g	Robust	Robust hard	TREC8
MRR	0.87	0.642	0.618	0.454	0.631	Prec@10	0.733	0.36	0.399	0.253	0.424
	0.849	0.436	0.544	0.373	0.474		0.653	0.3	0.371	0.231	0.404
	+2.5%	+47.3%*	+13.6%*	+21.7%*	+33.1%*		+12.3%*	+20.0%*	+7.6%*	+9.5%*	+5.0%
	0.782	0.55	0.596	0.393	0.606		0.667	0.338	0.381	0.211	0.413
	+11.3%*	+16.7%*	+3.7%	+15.5%*	+4.1%		+9.9%*	+6.5%*	+4.7%	+19.9%*	+2.7%
	0.863	0.606	0.609	0.442	0.579		0.718	0.353	0.401	0.251	0.431
	+0.8%	+6.0%*	+1.5%	+2.7%	+9.0%*		+2.1%	+2.0%	-0.5%	+0.8%	-1.6%
MAP	0.402	0.183	0.215	0.088	0.222	Prec@100	0.452	0.167	0.171	0.129	0.227
	0.37	0.158	0.22	0.09	0.226		0.406	0.147	0.174	0.127	0.222
	+8.7%*	+15.8%*	-2.2%	-2.2%	-1.7%		+11.3%*	+13.6%*	-1.7%	+1.6%	+2.3%
	0.398	0.202	0.221	0.089	0.225		0.44	0.18	0.169	0.126	0.205
	+1.0%	-9.4%	-2.7%	-1.1%	-1.3%		+2.7%	-7.2%	+1.2%	+2.4%	+10.7%*
	0.415	0.19	0.223	0.092	0.225		0.462	0.169	0.177	0.133	0.228
	-3.1%	-3.7%	-3.6%	-4.4%	-1.3%		-2.1%	-1.2%	-3.4%	-3.0%	-0.4%
NDCG	0.652	0.467	0.473	0.306	0.487	1-call	1.0	0.9	0.896	0.821	0.963
	0.587	0.398	0.396	0.252	0.422		0.98	0.82	0.831	0.8	0.88
	+11.1%*	+17.4%*	+19.4%*	+21.6%*	+15.3%*		+2.0%	+9.8%*	+7.8%*	+2.6%	+9.4%*
	0.651	0.477	0.483	0.312	0.484		0.98	0.86	0.847	0.74	0.88
	+0.2%	-2.0%	-2.1%	-1.8%	+0.5%		+2.0%	+4.7%	+5.8%*	+11.0%*	+9.4%*
	0.667	0.469	0.497	0.322	0.480		1	0.86	0.876	0.76	0.88
	-2.2%	-0.4%	-4.9%	-4.8%	+1.4%		0.0%	+4.7%	+2.3%	+8.0%*	+9.4%*
NDCG@10	0.184	0.154	0.184	0.085	0.155	6-call	0.78	0.24	0.289	0.141	0.322
	0.170	0.141	0.169	0.078	0.140		0.62	0.2	0.273	0.08	0.3
	+8.4%*	+8.9%*	+8.8%*	+9.1%*	+10.8%*		+25.8%*	+20.0%*	+5.9%*	+76.3%*	+7.3%*
	0.162	0.152	0.179	0.077	0.154		0.66	0.22	0.273	0.04	0.3
	+13.8%*	+1.0%	+2.7%	+10.5%*	+0.7%		+18.2%*	+9.1%*	+5.9%*	+252.5%*	+7.3%*
	0.184	0.162	0.191	0.086	0.150		0.74	0.24	0.297	0.1	0.32
	+0.2%	-5.2%	-3.7%	-1.0%	+3.4%		+5.4%	0.0%	-2.7%	+41.0%*	+0.6%
NDCG@100	0.379	0.306	0.332	0.182	0.321	8-call	0.6	0.16	0.153	0.02	0.22
	0.355	0.262	0.292	0.159	0.287		0.44	0.08	0.129	0.02	0.2
	+6.9%*	+16.8%*	+13.7%*	+14.3%*	+11.9%*		+36.4%*	+100.0%*	+18.6%*	0.0%	+10.0%*
	0.367	0.295	0.331	0.173	0.315		0.44	0.14	0.108	0.005	0.18
	+3.4%	+3.7%	+0.3%	+5.1%	+2.0%		+36.4%*	+14.3%*	+41.7%*	+300.0%*	+22.2%*
	0.401	0.297	0.345	0.181	0.314		0.62	0.14	0.161	0.04	0.22
	-5.4%	-3.0%	-3.8%	+0.4%	+2.3%		-3.2%	+14.3%*	-4.9%	-50.0%	0.0%
Prec@1	0.144	0.059	0.055	0.045	0.073	10-call	0.34	0.04	0.048	0.0	0.062
	0.14	0.057	0.055	0.046	0.074		0.26	0	0.036	0.0	0.04
	+2.9%	+3.5%	0.0%	-2.2%	-1.3%		+30.8%*	–*	+33.3%*	–	+55.0%*
	0.148	0.064	0.055	0.047	0.075		0.2	0.02	0.024	0.0	0.02
	-2.7%	-7.8%	0.0%	-4.3%	-2.7%		+70.0%*	+100.0%*	+100.0%*	–	+210.0%*
	0.149	0.062	0.057	0.049	0.076		0.26	0.02	0.04	0.0	0.02
	-3.4%	-4.8%	-3.5%	-8.2%	-3.9%		+30.8%*	+100.0%*	+20.0%*	–	+210.0%*

[4] K. Church and W. Gale. Poisson mixtures. *Journal of Natural Language Engineering*, 1995.

[5] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall, 2003.

[6] D. Hiemstra. *Using language models for information retrieval*. Doctoral thesis, University of Twente, 2001.

[7] T. Hofmann. Latent semantic models for collaborative filtering. *ACM Trans. Info. Syst.*, Vol 22(1):89–115, 2004.

[8] F. Jelinek and R. Mercer. Interpolated estimation of markov source parameters from sparse data. *Pattern Recognition in Practice*, pages 381–402, 1980.

[9] M. Kendall and A. Stuart, editors. *The Advanced Theory of Statistics Volume 1, 3rd Edition (Section 3.12)*. Griffin, London, 1969.

[10] J. D. Lafferty and C. Zhai. Document language models, query models, and risk minimization for information retrieval. In *Proc. of SIGIR '01*, pages 111–119, 2001.

[11] E. Lukacs, editor. *Characteristic Functions, 2nd Edition (Page 27)*. Griffin, London, 1970.

[12] R. E. Madsen, D. Kauchak, and C. Elkan. Modeling word burstiness using the Dirichlet distribution. In *Proc. of ICML*, pages 545–552, 2005.

[13] M. E. Maron and J. L. Kuhns. On relevance, probabilistic indexing and information retrieval. *J. ACM*, 7(3):216–244, 1960.

[14] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Proc. of SIGIR '98*, pages 275–281, 1998.

[15] J. Risson and T. Moors. Survey of research towards robust peer-to-peer networks: Search methods. *Computer Networks*, 50(17):3485–3521, 2006.

[16] S. E. Robertson. The probability ranking principle in IR. *Readings in information retrieval*, pages 281–286, 1997.

[17] S. E. Robertson, M. E. Maron, and W. Cooper. Probability of relevance: a unification of two competing models for document retrieval. *Information Technology: Research and Development*, 1(1):1–21, 1982.

[18] S. E. Robertson and K. Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3):129–146, 1976.

[19] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proc. of SIGIR '94*, pages 232–241, 1994.

[20] S. E. Robertson, S. Walker, M. Hancock-Beaulieu, M. Gatford, and A. Payne. Okapi at trec-4. In *Text REtrieval Conference (TREC)*, 1995.

[21] J. A. Thom and F. Scholer. A comparison of evaluation measures given how users perform on search tasks. In *Australasian Document Computing Symposium*, pages 100–103, 2007.

[22] C. J. van Rijsbergen. *Information Retrieval*. Butterworths, London, London, UK, 1979.

[23] H. Zaragoza, D. Hiemstra, M. Tipping, and S. E. Robertson. Bayesian extension to the language model for ad hoc information retrieval. In *Proc. of SIGIR '03*, 2003.

[24] A. Zellner. Bayesian estimation and prediction using asymmetric loss functions. *Journal of the American Statistical Association*, 81(394):446–451, 1986.

[25] C. Zhai and J. D. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proc. of SIGIR '01*, pages 334–342, 2001.

[26] C. Zhai and J. D. Lafferty. A risk minimization framework for information retrieval. *Inf. Process. Manage.*, 42(1):31–55, 2006.

[27] J. Zhu, J. Wang, I. Cox, and M. Taylor. Risk-aware information retrieval. In *Proc. of the European Conference on Information Retrieval (ECIR)*, pages 17–28, 2009.