Risk-Aware Information Retrieval

Jianhan Zhu[1], Jun Wang[1], Michael Taylor[2], and Ingemar J. Cox[1]

[1] University College London, Adastral Park Campus, Ipswich, U.K.
{j.zhu,j.wang,i.cox}@adastral.ucl.ac.uk
[2] Microsoft Research, Cambridge, U.K.
mitaylor@microsoft.com

Abstract. Probabilistic retrieval models usually rank documents based on a scalar quantity. However, such models lack any estimate for the uncertainty associated with a document's rank. Further, such models seldom have an explicit utility (or cost) that is optimized when ranking documents. To address these issues, we take a Bayesian perspective that explicitly considers the uncertainty associated with the estimation of the probability of relevance, and propose an asymmetric cost function for document ranking. Our cost function has the advantage of adjusting the risk in document retrieval via a single parameter for any probabilistic retrieval model. We use the *logit* model to transform the document posterior distribution with probability space [0,1] into a normal distribution with variable space $(-\infty, +\infty)$. We apply our risk adjustment approach to a language modelling framework for risk adjustable document ranking. Our experimental results show that our risk-aware model can significantly improve the performance of language models, both with and without background smoothing. When our method is applied to a language model without background smoothing, it can perform as well as the Dirichlet smoothing approach.

1 Introduction

The well-known Probability Ranking Principle (PRP) [9,12] states that by ranking documents in descending order of their probability of relevance to a query, an information retrieval (IR) system's overall effectiveness to the query will be maximized. Probabilistic retrieval models [1,8,11] have followed the PRP in document ranking. These models consider a document's probability of relevance as a deterministic quantity, i.e., it is known with absolute certainty. In practice, we believe that the probability of relevance is better described by a distribution that models the uncertainty associated with any estimate of a document's probability of relevance. The mean of the distribution represents the true (and unknown) probability of relevance of the document to the query.

When we compute a score for the probability of relevance of the document to the query, this represents our best estimate of the mean. It can be considered as a sample value drawn from the distribution centered on the true mean. The variance of the distribution represents the uncertainty associated with the estimate. The sources of uncertainty are many, and include ambiguity in the query, specific user preferences, and deviations between the scoring function and the true

M. Boughanem et al. (Eds.): ECIR 2009, LNCS 5478, pp. 17–28, 2009.

probability of relevance. Obviously, if the variance is large, then the uncertainty in the estimate of a document's probability of relevance is also large.

Any single estimate of a document's true mean (probability of relevance) is equally likely to be less than or greater than the true mean. And the likely magnitude of the error is a determined by the variance. If the variance is large, then the magnitude of the error will likely be large. If the variance is small, then the magnitude of the error will likely be small.

Let us now consider what will happen when we apply the PRP framework if the probabilities of relevance are not deterministic. In this case, we will, of course, select the documents with the largest estimates of probability of relevance. The most relevant document is assumed to be the document with the largest probability of relevance. However, the veracity of this assumption depends on both the variance associated with the probability of relevance value, and the probabilities of relevance and variances of the other documents.

If the top-ranked document's probability of relevance has a large associated variance, then there is a much greater likelihood that we have significantly under or over estimated the probability of relevance of the document. If we underestimate the probability of relevance, then the user will likely be pleased with the choice. If we overestimate the probability of relevance, then the user will likely be displeased with our choice. If we average the user's perceived relevance over many queries, choosing top-ranked documents that have high variance will result in a user experiencing a high degree of variability in the search results. Sometimes results will appear very good and other times very poor.

In practice, the situation is more complex since some documents in the result set will have high estimated probability of relevance and low variance and other documents will have high estimated probability of relevance and high variance. Rank ordering these documents is more complicated and depends on the optimization criterion. However, given two documents with the same (or very similar) estimated probability of relevance, but one with much lower variance than the other, we should always rank the more certain document (lower variance) above the other.

The PRP framework optimizes the expected relevance of documents. However, as we discussed, this can lead to a high degree of volatility in the quality of our result sets, if the estimated probability of relevance of documents have high variances/uncertainties. Of course, in the case of uncertainty we are equally likely to be right or wrong in our estimate. If we choose a more uncertain document over a less uncertain document, we risk returning a poor quality document, but the risk also offers the potential that we return a document that is more relevant than the document with less uncertainty. Thus, there is a tradeoff between risk and reward. In this paper, we introduce an asymmetric loss[1] function with an adjustable parameter that allows us to increase or decrease our risk. This loss

[1] For document ranking, the loss of under-estimating the probability of relevance may not be equal to that of over-estimating. For example, in many retrieval scenarios, particularly within the top-ranked positions, we argue that it might be favorable to take a more conservative ranking decision because the cost of over-estimating the probability of relevance might be higher than that of under-estimating.

function is described in Section 3, after a discussion of related work in Section 2. We apply the cost function to language models in Section 4. Section 5 presents experimental results, which show that a risk-averse preference can significantly out-perform a risk-taking preference. More interestingly, a risk-averse preference is also shown to significantly outperform the traditional PRP approach. Finally, we conclude in Section 6.

2 Related Work

To remedy the problem of unreliable probability estimation from limited data, e.g. maximum likelihood estimation, recent studies have focused on building more accurate language models for documents, including background smoothing based on collection statistics [18], and a Bayesian treatment of the language modelling framework [16]. By considering risk in retrieval, a risk minimization framework was proposed in [19] for ranking documents based on the expected risk of these documents. The framework has been applied to subtopic retrieval for modelling redundancy and novelty in addition to relevance. However, these approaches do not explicitly model the uncertainty associated with document ranking.

In previous work, we [20] proposed a Bayesian risk-adjustable approach to account for the uncertainty in document ranking. We derived a Bayesian ranking function, which is applied to a document's posterior distribution with the probability space in [0,1]. Based on the work in [20], our contribution in this paper is to present a more generic approach to account for the uncertainty, that uses the logit model [3] to transform the document posterior distribution into a normal distribution. We subsequently derive a cost function based on the normal distribution.

Vinay *et al.* [14] have also modeled a document's relevance score as a Gaussian random variable. They used the normal distributions to estimate the probability that one document should be ranked higher than another, and this is used as the basis for calculating the *expected ranks* of documents. However, Vinay *et al.* [14] did not consider risk in document ranking.

Webber *et al.* [15] considered topic variability in IR evaluations. Given a topic, they proposed using the mean and variance of participating systems' scores on a metric as the standardization factors, which can be used to normalize a system's score on the topic.

3 A Risk-Aware Information Retrieval Model

Suppose that we have a term q_t in a query q. We denote θ_t as the estimation of the correspondence between q_t and a document d. θ_t is equal to $p(r|d, q_t)$ (r denotes relevance) in the relevance models, and equal to $p(q_t|\theta_d)$ (θ_d is the language model for d) in language models. θ_t follows a distribution with a range from 0 to 1, e.g., a Beta distribution for language models.

We propose to transform the distribution of θ_t into a normal distribution mainly due to two reasons. Firstly, a normal distribution can be uniquely described by its mean and variance. Secondly, for a normal distribution, we can obtain an analytic solution of the LINEX loss function [13] as shown later in this section. Furthermore, normal distributions have been widely adopted in previous work, e.g., Vinay *et al.* [14] and Herbrich *et al.* [6] used normal distributions to model document relevance scores, and players' skill, respectively.

It has been shown in [3] that the distribution obtained from the logit transformation of θ_t approximately conforms to a normal distribution.

$$r_t = f(\theta_t) = \ln \frac{\theta_t}{1-\theta_t}, \tag{1}$$

where r_t is the relevance of a document to the term.

The logit function in Eq. (1) follows previous work on considering both the probability of relevance and non-relevance in document ranking [4,10]. The logit model favours documents that are highly relevant for some terms in a multi-term query[2].

The estimated mean and variance of the normal distribution obtained from Eq. (1) are given by

$$E[f(\theta_t)] \approx \ln \frac{\bar{\theta}_t}{1-\bar{\theta}_t} + \frac{2\bar{\theta}_t - 1}{2\bar{\theta}_t^2(1-\bar{\theta}_t)^2} Var(\theta_t) \tag{2}$$

$$Var[f(\theta_t)] \approx \frac{Var(\theta_t)}{\bar{\theta}_t^2(1-\bar{\theta}_t)^2}, \tag{3}$$

where the mean and variance of θ_t are $\bar{\theta}_t$ and $Var(\theta_t)$, respectively, $E[f(\theta_t)]$ is the mean of $f(\theta_t)$, and $Var[f(\theta_t)]$ is the variance of $f(\theta_t)$. Further details can be found in Appendix A.

Risk has been studied in a variety of contexts. We propose to use an asymmetric loss function, LINEX, first proposed by Varian [13] in the context of financial investment.

It has been shown in [17] that if the distribution of a document's relevance score has a Gaussian form, $\phi(x|\mu, \sigma^2)$, there exists an analytic solution for the LINEX loss function, given by

$$\tilde{\phi} = \mu - b\sigma^2/2, \tag{4}$$

[2] To illustrate this, we give an example. Suppose a query consists of two terms, document A's relevance scores to the two terms are both 0.3, and document B's relevance scores to the two terms are 0.1 and 0.9, respectively. Assuming term independence, retrieval models give A and B the same relevance score, i.e., 0.09. However, it can be easily derived that the logit model gives B higher relevance score than A since B is highly relevant to one of the two terms. Our initial experiments show that the logit model performs better than traditional retrieval models on long queries, and has similar performance to traditional models on short queries. Systematic comparison is out of the scope of this paper, and will form part of our future work.

where $\tilde{\phi}$ is our risk-adjustable ranking function. Note that μ is the estimated mean, not the true mean, and σ^2 is the variance.

Eq. (4) gives a general formula, which has the advantage of adjusting the risk via a single parameter b, to rank documents when considering asymmetric loss. To address the uncertainty, the final ranking is equal to the mean of the normal distribution subtracted (or added) by a weighted variance. A positive b produces a risk-averse (conservative) ranking where the unreliably-estimated documents (with big variance) should be given a lower ranking score. The bigger the parameter b is, the more conservative the ranker is. On the other hand, a negative b gives the risk-inclined ranking.

Substituting the estimated mean in Eq. (2) and variance in Eq. (3) into Eq. (4), we get our risk-adjustable ranking function as

$$\tilde{\phi}_t == \ln \frac{\bar{\theta}_t}{1-\bar{\theta}_t} + \frac{2\bar{\theta}_t - 1 - b}{2\bar{\theta}_t^2(1-\bar{\theta}_t)^2} Var(\theta_t) \tag{5}$$

4 Risk-Aware Language Models

We apply the proposed document ranking approach in Eq. (4) under the language modelling framework. However, it is worth noting that the proposed method is generally applicable to any probabilistic retrieval models.

We formally represent a document $\mathbf{d}$ and a query $\mathbf{q}$ as vectors of term counts as $\mathbf{q} \equiv (q_1, ..., q_t, ..., q_{|V|})$ and $\mathbf{d} \equiv (d_1, ..., d_t, ..., d_{|V|})$, where q_t (d_t) is the number of times the term t appears in the query (document) and $|V|$ is the size of a vocabulary. A language model θ for the document is $\theta = (\theta_1, ..., \theta_t, ..., \theta_{|V|})$, where $\sum_t \theta_t = 1$, and the probability space of θ_t is [0,1].

To estimate θ_t, a straightforward approach is to apply maximum likelihood estimation. However, estimating from one single document is unreliable due to small data samples. A common solution 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:

$$p(\theta|\mathbf{d}, \alpha) \propto p(\mathbf{d}|\theta)p(\theta|\alpha) \propto \prod_t (\theta_t)^{d_t} \prod_t (\theta_t)^{\alpha_t - 1} \propto \prod_t (\theta_t)^{d_t + \alpha_t - 1}, \tag{6}$$

where prior $p(\theta|\alpha) \equiv (\alpha_1, \ldots, \alpha_{|V|})$ incorporates prior knowledge, e.g. collection statistics for smoothing the estimation [16,18]. For Jelinek-Mercer (or linear) smoothing, we set $\alpha_t = \frac{\lambda|\mathbf{d}|}{1-\lambda} \cdot \frac{C_t}{|\mathbf{C}|}$, where λ is a parameter, C_t is the number of occurrences of term t in the collection, $|\mathbf{d}|$ is the document length, and $|\mathbf{C}|$ is the collection size; for Dirichlet smoothing, we set $\alpha_t = \mu \frac{C_t}{|\mathbf{C}|}$, where μ is a parameter.

Since the posterior probability in Eq. (6) is a Dirichlet distribution, its mean $\bar{\theta}_t$ and variance $Var(\theta_t)$ are known analytically [5], and given by:

$$\bar{\theta}_t = \frac{c_t}{\hat{c}} \tag{7}$$

$$Var(\theta_t) = \frac{c_t(\hat{c} - c_t)}{\hat{c}^2(\hat{c}+1)}, \tag{8}$$

where, for simplicity, we denote $c_t \equiv d_t + \alpha_t$ and $\hat{c} \equiv \sum_t (d_t + \alpha_t)$.

Replacing the mean and variance of the Dirichlet distribution in Eq. (5), our risk-aware language model becomes:

$$\tilde{\phi}_t = \ln \frac{c_t}{\hat{c} - c_t} + \frac{\hat{c}[2c_t - (1+b)\hat{c}]}{2c_t(\hat{c} - c_t)(\hat{c}+1)} \tag{9}$$

Finally, our ranking score of a document d for query $\mathbf{q}$ is:

$$\tilde{\phi} = \sum_{t=1}^{|V|} q_t \times (\tilde{\phi}_t) = \sum_{t=1}^{|V|} q_t \times \{\ln \frac{c_t}{\hat{c} - c_t} + \frac{\hat{c}[2c_t - (1+b)\hat{c}]}{2c_t(\hat{c} - c_t)(\hat{c}+1)}\} \tag{10}$$

5 Experimental Evaluation

We studied our approach on four TREC test collections described in Table 1. The TREC2004 robust track is evaluated with an emphasis on the overall reliability of IR systems, i.e. minimizing the number of queries for which the system performs badly. Among the TREC2004 robust track, 50 queries were identified as "difficult", which can help us understand whether our approach is effective for both "ordinary" and "difficult" queries. Documents were stemmed using the Porter stemmer, and stopping is carried out at query time. In our experiments, only the title portion of the TREC topics were used.

Table 1. Overview of the four TREC test collections

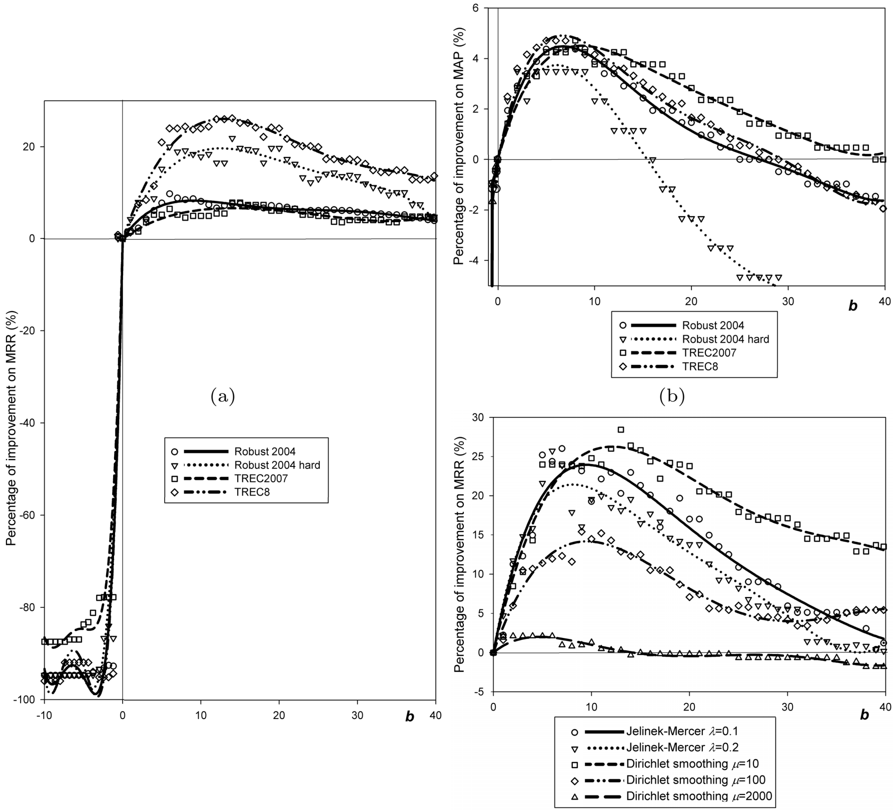
Name	Description	# Docs	Topics	# Topics
TREC2007 enterprise track document search task	CSIRO website crawl	370,715	1-50 minus 8, 10, 17, 33, 37, 38, 46, 47	42
TREC 2004 robust track (Robust2004)	TREC disks 4, 5 minus CR	528,155	301-450 and 601-700 minus 672	249
Robust2004 hard topics	TREC disks 4, 5 minus CR	528,155	Difficult Robust2004 topics	50
TREC8 ad hoc task	TREC disks 4, 5 minus CR	528,155	401-450	50

5.1 The Risk-Adjustable Parameter

We first investigate the effect of parameter b in retrieval via standard metrics, and then study relationships between the optimal b and risk-sensitive metrics.

We first look at the effect of the risk-adjustable parameter b on mean reciprocal rank (MRR) and mean average precision (MAP) for four test collections. When b changes between -10 and 40, Fig. 1 (a) shows the percentage of improvement[3] on MRR on four test collections for a baseline model without background smoothing. Each data point in Fig. 1 (a), 1 (b), and 1 (c) represents the percentage of improvement for a given b.

[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 $b = 0$.



(a)

(b)

(c)

Fig. 1. Plots of the percentage of gain on MRR or MAP against parameter b. (a) Percentage of gain on MRR against b for four collections. (b) Percentage of gain on MAP against b for four collections. (c) Percentage of gain on MRR against b for TREC8 collection under Jelinek-Mercer and Dirichlet smoothing.

We can see from Fig. 1 (a) that by taking a risk-aversion approach ($b > 0$), i.e., revising the relevance score downwards, the value of MRR is improved for all four collections. Generally, all four curves share a similar structure, with the percentage of gain on MRR quickly improving as b increases above zero, reaching a peak value when b is between 5 and 15, and then gradually declining. A merit of our approach is that the performance gain is robust with respect to the choice of b, and a value of b anywhere between 0 and 40 leads to positive performance gains on all four collections. The results indicate that risk-aversion is favorable for all four collections. Topic difficulty does not seem to affect our approach, since the performance gain on Robust 2004 hard topics is even more significant than that on the whole set of Robust 2004 topics.

On the other hand, by taking a risk-loving approach ($b < 0$), i.e., opportunistically overestimating the relevance score, the performance on MRR degrades for all four collections.

Table 2. Relationship between risk-sensitive metrics and the optimal b for them on four collections

(a) 1-call, 3-call, 6-call, and 10-call at 10 and their optimal b

	Optimal b			
Measures	Robust04	Robust04 hard	TREC07	TREC8
1-call	16.2	19.1	13.4	14.8
3-call	8.8	10.2	12.3	7.7
6-call	6.6	4.7	11.4	2.3
10-call	4.2	0.2	1.0	2.2

(b) NDCG at 1, 10, 50, 100, 500, and 1000 and their optimal b

	Optimal b			
Measures	Robust04	Robust04 hard	TREC07	TREC8
NDCG-1	10.2	16.1	17.4	13.3
NDCG-10	8.4	10.0	16.3	13.2
NDCG-50	5.8	5.5	5.8	5.6
NDCG-100	5.4	5.4	5.3	5.5
NDCG-500	5.3	5.6	4.4	5.8
NDCG-1000	3.0	5.8	4.0	4.0

To confirm our findings, we tested the effect of b on the MAP for the four collections. Fig. 1 (b) also shows that a risk-aversion approach can help improve the performance, while performance degrades under a risk-loving approach. Following similar trend as MRR in Figure 1 (a), curves in Fig. 1 (b) show that the MAPs increase quickly at the beginning, reach a peak for b between 0 and 10, and then declines gradually. However, the performance gain on MAP is not as significant as that for MRR. The exact reason for the variation between MRR and MAP will be the subject of future work.

We next compare to models using background smoothing. Fig. 1 (c) shows the performance gain on MRR for both Jelinek-Mercer and Dirichlet smoothing with different parameters for the TREC8 collection. We can see from Fig. 1 (c) that a risk-aversion approach is effective for all five different background smoothing methods. However, increasing the influence of background smoothing via large values of λ or μ reduces the effectiveness of risk adjustment. We believe this is because background smoothing plays a similar role to our approach in relevance score adjustment. Similar results were obtained on the other three collections.

We now investigate how our risk-aversion approach behaves under a risk-sensitive metric called n-call at 10 [2]. Given a ranked list, n-call is one if at least n of the top 10 documents returned for a query are relevant. Otherwise, n-call is zero. Averaging over multiple queries yields the mean n-call. The two extremes are 10-call, an *ambitious* metric of perfect precision, returning only relevant documents, and 1-call is a *conservative* metric that is satisfied with only one relevant document. Therefore, a *risk-aversion* approach, which can reliably find one relevant document with small variance, is preferred for 1-call, while a *risk-loving* approach, which gives small weight to the variance, is favored for 10-call.

Table 2 (a) illustrates the relationship between the optimal values of b and n-call on the four collections. Table 2 (a) demonstrates that when n decreases, the optimal value of b tends to increase. This demonstrates how the risk adjustment

Table 3. Performance comparison on six metrics. Three lines in each cell are performance of a language model and our risk-aware approach, and the percentage of gain of our approach over the language model, respectively. Positive and statistically significant improvements are in bold, and in bold and marked with "*", respectively. Where stated, we tested statistical significance with t tests, one-tail critical values for significance levels α=0.05.

(a) LM with Dirichlet smoothing ($\mu = 2000$) vs. $b = 5$ for the background-independent LM

Measures	Robust04	Robust04	TREC07	TREC8	Measures	Robust04	Robust04	TREC07	TREC8
		hard					hard		
MRR	0.604 **0.605** +0.2%	0.441 **0.45** +2.0%	0.819 **0.849** +3.7%	0.613 **0.615** +0.3%	NDCG-100	0.238 0.232 -2.5%	0.166 0.166 0%	0.503 **0.513** +2.0%	0.282 **0.285** +1.1%
2-call	0.743 0.735 -1.1%	0.58 **0.6** +3.4%	0.98 0.92 -6.1%	0.82 0.78 -4.9%	Prec-10	0.387 **0.389** +0.5%	0.233 **0.247*** +6.0%	0.662 **0.68** +2.7%	0.418 0.411 -1.7%
NDCG-10	0.399 0.398 -0.3%	0.244 **0.247** +1.2%	0.678 **0.696** +2.7%	0.424 0.421 -0.7%	Prec-100	0.203 0.197 -2.9%	0.148 0.146 -1.3%	0.472 **0.479** +1.5%	0.248 **0.252** +1.6%

(b) LM with linear smoothing ($\lambda = 0.1$) vs. $b = 5$ for the LM with linear smoothing ($\lambda = 0.1$)

Measures	Robust04	Robust04	TREC07	TREC8	Measures	Robust04	Robust04	TREC07	TREC8
		hard					hard		
MRR	0.544 **0.609*** +11.9%	0.375 **0.424 *** +13.1%	0.804 **0.846** +5.2%	0.488 **0.611*** +25.2%	NDCG-100	0.235 **0.244** +3.8%	0.16 **0.164** +2.5%	0.497 **0.522** +5.0%	0.287 **0.297** +3.5%
2-call	0.723 **0.747** +3.3%	0.56 **0.6*** +7.1%	0.96 0.96 0%	0.76 **0.82*** +7.9%	Prec-10	0.382 **0.407*** +6.5%	0.236 0.236 0%	0.669 **0.693** +3.6%	0.413 **0.447*** +8.2%
NDCG-10	0.386 **0.415*** +7.5%	0.228 **0.235** +3.1%	0.672 **0.702** +4.5%	0.404 **0.449*** +11.1%	Prec-100	0.205 **0.209** +2.0%	0.145 **0.149** +2.8%	0.465 **0.488** +4.9%	0.26 **0.261** +0.4%

parameter, b, controls how much risk we are prepared to take when ranking documents, and the effect this has on the result set. For large values of b, i.e., risk-aversion (conservative ranking) we have a much greater chance that at least one document will be relevant, but the chance that many of the documents will be relevant is diminished. Conversely, for a risk-loving (aggressive ranking), we have a much greater chance that many of the documents will be relevant, but at the expense that some searches produce no relevant documents. This supports our discussion in Section 1 in which we described how a risk-loving strategy will lead to more volatility in our search results, but that the benefit of this volatility is that for some searches, we will display more relevant documents.

Next we study the effect of ranking positions on b. Table 2 (b) shows the optimal b value for the Normalized Discounted Cumulated Gain (NDCG) at different cut-off points on the four collections. Table 2 (b) illustrates that the optimal value of b for each collection decreases when the cut-off point increases. Such behavior suggests that lower rank position favors more conservative ranking (bigger b), but higher rank position favors more aggressive ranking (smaller b).

5.2 Performance

Based on the study of parameter b in Section 5.1, we fix b as 5 and evaluate the effectiveness of our risk-aware approach on four collections. Note that $b = 5$ may not be optimal for different collections, language models, and metrics as shown in Fig. 1 and Table 2. However, we want to show that even by applying a universal value of b, the performance of a number of metrics on four collections can all be significantly improved. If b is optimized for individual collections, language models, or metrics, the performance can be improved even further. Table 3 (a) and (b) report the results on a number of metrics including MRR, 2-call, NDCG at 10, NDCG at 100, Precision at 10, and Precision at 100.

Table 3 (a) compares our risk-aware approach without background smoothing with the state-of-the-art language modelling Dirichlet smoothing approach with μ=2000, which was reported to have outstanding performance on a number TREC collections [18]. Table 3 (a) shows that even without any background smoothing, our risk-aware approach can perform as well as, and sometimes even better than the Dirichlet smoothing approach. Our approach outperforms the Dirichlet smoothing approach on MRR for all four collections, 14 out of 24 improvements are positive in Table 3 (a), and one improvement is statistically significant. In addition, our approach has similar performance to the Dirichlet smoothing approach on MAP for all four collections.

Table 3 (b) reports the improvements of applying our approach to the Jelinek-Mercer (linear) smoothing approach over the linear smoothing approach where we adopted the typical settings of λ=0.1 [18]. We can see that our approach can significantly improve the linear smoothing approach, i.e., 9 out of 24 results are statistically significant, 22 out of 24 improvements are positive, and the highest improvement on MRR is over 25%, showing that risk adjustment can dramatically increase the chance of returning one relevant document close to the top of a ranked list. Our approach also outperforms the linear smoothing approach on MAP for all four collections. Comparing Table 3 (b) with Table 3 (a), we can see that our approach combined with the linear smoothing performs better than our approach without background smoothing in 18 out of 24 results, and better than the Dirichlet smoothing in 18 out of 24 results. Therefore, our risk adjustment complements background smoothing in performance improvement.

6 Conclusion and Future Work

Uncertainty is an intrinsic part of document ranking, but has not generally been considered in current IR models. Current models usually provide a scalar estimate of the mean of a document's posterior probability distribution. However, the probability distribution is better described by both its mean and variance.

As discussed in the Introduction, the variance or uncertainty can introduce a level of volatility in our retrieved results, i.e. some results may be very good while others may be very poor. In the light of this, we proposed a risk-aware information retrieval model that allows us to control this volatility. That is, we can reduce the variability across searches, albeit at the expense of reducing the overall relevance of documents in the retrieved set. This was experimentally demonstrated by adjusting the risk preference parameter, b, for the risk sensitive metrics of n-call and NDCG.

Our approach uses an asymmetric risk function, LINEX, developed in the context for financial portfolio theory [7]. The LINEX cost function has an analytic solution for random variables with a Gaussian distribution. We used the logit transformation to transform the posterior distribution of the probability of relevance into a normal distribution. Under these conditions, a single risk preference parameter, b, allows us to adjust the level of risk we wish to accept.

Experimental results compared our method with a variety of language modelling approaches. Our experiments on four TREC collections showed that a risk-aversion approach ($b > 0$) helps improve the performance on MAP and MRR, but a risk-loving approach ($b < 0$) degrades performance. By adjusting b, our approach has effectively optimized a range of risk-sensitive metrics (n-call at 10 [2]) and metrics of different ranking positions (NDCG at n) that reflect different levels of risk in search.

Performance is comparable with the Dirichlet smoothing approach. However, we note this was achieved without the need for background smoothing. Our approach can also complement the Jelinek-Mercer smoothing approach. Experimental results demonstrated significant improvements when our model was used in conjunction with Jelinek-Mercer smoothing.

Since term dependence is not fully taken into account in current unigram language models [8], future work will consider the joint posterior probability distribution across multiple terms. The challenge is that the variance of the joint distribution is influenced by not only the variance of each term's posterior distribution but also the correlation between the terms.

References

1. Amati, G., Rijsbergen, C.J.V.: Probabilistic models of information retrieval based on measuring the divergence from randomness. ACM Trans. Inf. Syst. 20(4), 357–389 (2002)
2. Chen, H., Karger, D.R.: Less is more: probabilistic models for retrieving fewer relevant documents. In: Proc. of SIGIR 2006, pp. 429–436. ACM, New York (2006)
3. Cramer, J.: The origins and development of the logit model. Cambridge University Press, Cambridge (2003)
4. de Vries, A.P., Roelleke, T.: Relevance information: A loss of entropy but a gain for idf? In: Proc. of SIGIR (2005)
5. Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B.: Bayesian Data Analysis. Chapman and Hall, Boca Raton (2003)
6. Herbrich, R., Minka, T., Graepel, T.: Trueskill$^{\text{tm}}$: A bayesian skill rating system. In: Proc. of NIPS, pp. 569–576 (2006)
7. Markowitz, H.: Portfolio selection. Journal of Finance (1952)
8. Ponte, J.M., Croft, W.B.: A language modeling approach to information retrieval. In: Proc. of SIGIR 1998, pp. 275–281. ACM Press, New York (1998)
9. Robertson, S.E.: The probability ranking principle in IR. Readings in information retrieval, 281–286 (1997)
10. Robertson, S.E., Sparck Jones, K.: Relevance weighting of search terms. Journal of the American Society for Information Science 27(3), 129–146 (1976)

11. Robertson, S.E., Walker, S.: Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In: Proc. of SIGIR 1994, pp. 232–241. Springer, New York (1994)
12. van Rijsbergen, C.J.: Information Retrieval. Butterworths, London (1979)
13. Varian, H.: A bayesian approach to real estate assessment. In: Fienberg, S.E., Zellner, A. (eds.) Studies in Bayesian Econometrics and Statistics in Honour of Leonard J. Savage, pp. 198–205 (1975)
14. Vinay, V., Milic-Frayling, N., Cox, I.: Estimating retrieval effectiveness with rank distributions. In: Proc. of the Conference on Information and Knowledge Management (CIKM) (2008)
15. Webber, W., Moffat, A., Zobel, J.: Score standardization for inter-collection comparison of retrieval systems. In: SIGIR, pp. 51–58 (2008)
16. Zaragoza, H., Hiemstra, D., Tipping, M., Robertson, S.E.: Bayesian extension to the language model for ad hoc information retrieval. In: Proc. of SIGIR 2003, pp. 4–9. ACM Press, New York (2003)
17. Zellner, A.: Bayesian Estimation and Prediction Using Asymmetric Loss Functions. Journal of the American Statistical Association 81(394), 446–451 (1986)
18. Zhai, C., Lafferty, J.D.: A study of smoothing methods for language models applied to ad hoc information retrieval. In: Proc. of SIGIR 2001, pp. 334–342 (2001)
19. Zhai, C., Lafferty, J.D.: A risk minimization framework for information retrieval. Inf. Process. Manage. 42(1), 31–55 (2006)
20. Zhu, J., Wang, J., Cox, I., Taylor, M.: Risky business: Modeling and exploiting uncertainty in information retrieval. Technical Report. University College London (2008)

Appendix A

Since $f(\theta_t) = ln\frac{\theta_t}{1-\theta_t}$ is infinitely differentiable in the neighborhood of the mean of θ_t, the mean of $f(\theta_t)$ can be approximated as the mean of a Taylor series as:

$$\begin{aligned} &E[f(\theta_t)] \\ &\quad = E[f(\bar{\theta}_t)] + E[(\theta_t - \bar{\theta}_t)f'(\bar{\theta}_t)] + E[\frac{1}{2}(\theta_t - \bar{\theta}_t)^2 f''(\bar{\theta}_t)] + \cdots \\ &\quad = f(\bar{\theta}_t) + 0 + \frac{1}{2}f''(\bar{\theta}_t)Var(\theta_t) + \cdots \\ &\quad \approx f(\bar{\theta}_t) + \frac{f''(\bar{\theta}_t)}{2}Var(\theta_t) = \ln\frac{\bar{\theta}_t}{1-\bar{\theta}_t} + \frac{2\bar{\theta}_t - 1}{2\bar{\theta}_t^2(1-\bar{\theta}_t)^2}Var(\theta_t) \end{aligned} \tag{11}$$

$f(\theta_t)$ can be approximated by a first order Taylor series as $f(\theta_t) \approx f(\bar{\theta}_t) + (\theta_t - \bar{\theta}_t)f'(\bar{\theta}_t)$. Therefore, the variance of $f(\theta_t)$ is approximated as:

$$Var[f(\theta_t)] \approx 0 + Var[(\theta_t - \bar{\theta}_t)f'(\bar{\theta}_t)] = [f'(\bar{\theta}_t)]^2 Var(\theta_t) = \frac{Var(\theta_t)}{\bar{\theta}_t^2(1-\bar{\theta}_t)^2} \tag{12}$$