Portfolio Theory of Information Retrieval

Jun Wang and Jianhan Zhu
Department of Computer Science, University College London
Malet Place, London WC1E 6BT, UK
jun_wang@acm.org, jianhan.zhu@ucl.ac.uk

ABSTRACT

This paper studies document ranking under uncertainty. It is tackled in a general situation where the relevance predictions of individual documents have uncertainty, and are dependent between each other. Inspired by the Modern Portfolio Theory, an economic theory dealing with investment in financial markets, we argue that *ranking under uncertainty* is not just about picking individual relevant documents, but about choosing the right combination of relevant documents. This motivates us to quantify a ranked list of documents on the basis of its expected overall relevance (mean) and its variance; the latter serves as a measure of risk, which was rarely studied for document ranking in the past. Through the analysis of the mean and variance, we show that an optimal rank order is the one that balancing the overall relevance (mean) of the ranked list against its risk level (variance). Based on this principle, we then derive an efficient document ranking algorithm. It generalizes the well-known probability ranking principle (PRP) by considering both the uncertainty of relevance predictions and correlations between retrieved documents. Moreover, the benefit of diversification is mathematically quantified; we show that diversifying documents is an effective way to reduce the risk of document ranking. Experimental results in text retrieval confirm the theoretical insights with improved retrieval performance.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval models, Search process, and Selection process

General Terms

Algorithms, Experimentation, Measurement, Performance

Keywords

Modern portfolio theory, Mean-variance analysis, Probability ranking principle, Ranking under uncertainty

1 Introduction

Information retrieval (IR) concerns how to retrieve documents for a user information need. The process of retrieving documents may be divided into two stages. In the first stage, the relevance between the given user information need and each of the documents in a collection is calculated. Probabilistic retrieval models that have been proposed and tested over decades are primarily focusing on this task, aiming at producing a "best guess" at a document's relevance. Examples include the RSJ model [15] (a further development of that model led to the BM25 term weighting function [16]), and the language modelling approaches [23]. The second stage focuses on how to present (normally rank) documents to the user. The probability ranking principle (PRP) [6] forms the basis in this stage, stating that the system should rank documents in order of decreasing probability of relevance; it has been shown that, following the principle, the overall effectiveness of an IR system, such as expected Precision, is maximized [13].

If we make an analogy with the field of finance, our ranking task resembles the investment problem in financial markets; for example, suppose that an investor needs to select a set (portfolio) of n stocks that will provide the best distribution of future return, given his or her investment budget – an analogy with IR is that we invest ranking positions in documents. The PRP of IR might suggest that, for optimal selection, one should first rank stocks in order of decreasing future returns and then choose the top-n most "profitable" stocks to construct the portfolio. Such a principle that essentially maximizes the expected future return was, however, rejected by an economist Harry Markowitz in his Nobel Prize winning work, the Modern Portfolio Theory (MPT) of finance, in 1952 [11]. As one of the most influential economic theories dealing with finance and investment, the MPT was motivated on the basis of the following two observations [11]. 1) The future return of a stock is unknown and cannot be calculated with absolute certainty. Investors have different preferences of the risk associated with uncertainty. Therefore, it is highly desirable to have a method of quantifying this uncertainty or risk, and reflect them and incorporate users' risk preferences when selecting stocks. 2) Since in practice the future returns of stocks are correlated, assuming independence between the returns and selecting them independently to construct a portfolio is not preferable.

Realizing the two fundamental issues, the MPT emphasizes that risk (uncertainty) is an inherent part of future return, and quantifies it by using the variance (or the standard deviation) of the return. The theory suggests that, for a *risk-averse* decision, an investor should both maximize the return as a desirable thing and minimize the variance of the return as an undesirable thing. Under such a formulation, the MPT mathematically shows that diversification, known as "not putting all of your eggs in one basket", is an effective way to reduce the risk of the portfolio.

Going back to our IR problem, we have two similar critical issues: 1) during retrieval, the relevance of documents is unknown and cannot be estimated with absolute certainty from IR models. There are many sources of uncertainty such as ambiguity in the query, specific user preferences,

SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.

and deviations between the scoring function and the "true" probability of relevance. 2) The relevance estimates of individual documents are also correlated, either positively or negatively [8]. Thus it is of great interest to see how we can follow the school of thinking in the field of finance to address the ranking problem in IR.

In this paper, we focus on the theoretical development of the portfolio theory of document ranking. We formulate the ranking problem as a portfolio selection problem. That is, in response to a user information need, a top-n ranked list (portfolio) of documents is selected as a whole, rather than ranking documents independently. To characterize a ranked list, we employ two summary statistics, mean and variance. The mean represents a best "guess" of the overall relevance of the list, while the variance summarizes the uncertainty or risk associated with the guess. Our analysis provides new insights into the way we rank documents, and demonstrates that a better and more general ranking principle is to select top-n documents and their order by balancing the overall relevance of the list against its risk (variance). An efficient ranking algorithm is then introduced to trade off between efficiency and accuracy, and leads to a generalization of the PRP, where both the uncertainty of the probability estimation and diversity of ranked documents are modelled in a principled manner. The new ranking approach has been applied to the ad hoc text retrieval and sub-topic retrieval. The experiments demonstrate that our approach can adapt to different risk preferences of evaluation metrics, and as a result significant performance gains have been achieved.

The remainder of the paper is organized as follows. We will discuss the related work in Section 2, present our theoretical development in Section 3, give our empirical investigation in Section 4, and conclude in Section 5.

2 Related Work

Gordon and Lenk have discussed the two underlying assumptions of the PRP: independent assessment of relevance by the user and certainty about the estimated probabilities of relevance [8]. To deal with the assumption of the independence, Chen and Karger in [4] argued that the PRP, which ranks documents in descending order of probability of relevance, is not always optimal for different user information needs (or risk preferences we may say). In some scenarios users would be satisfied with a limited number of relevant documents, rather than requiring all relevant documents. The authors therefore proposed to maximize the probability of finding a relevant document among the top n under the assumption of binary relevance. By treating the previously retrieved documents as non-relevant ones, their algorithms naturally introduced diversification into the probabilistic ranking.

Unlike in [4] that concerns only the dependence of documents' relevance, our proposed *mean-variance paradigm* considers that the two assumptions of the PRP are highly connected and address them together in a more general setting. One of the theoretical contributions of our paradigm is that we mathematically demonstrate that diversifying the top-n documents is a way to reduce the variance and therefore risk of the ranked list. The greedy algorithm proposed in [4], which considers only the correlation between two neighboring documents, is in fact a special case in our proposed ranking method. Our paradigm is a general one, independent of the retrieval model that is being used, and has the advantage of tuning the risk via a single parameter.

Previous studies on integrating diversity has been focused on document re-ranking. Heuristically, Carbonell and Goldstein [3] proposed the Maximal Marginal Relevance (MMR) criterion to reduce redundancy by re-ranking retrieved documents under the Vector Space setup. Lafferty and Zhai in [10] presented a risk minimization framework. In the framework, documents are ranked based on an ascending order of the expected risk of a document. The MMR criterion has also been employed in the risk framework in resolving the subtopic retrieval problem [24], by modelling not only relevance but also redundancy, novelty, and subtopics. But, nonetheless, when coming to the practical algorithm, the studies [10, 24] still resolve to take point estimation, and use mode of the posterior as opposed to integrating out model parameters. Therefore, the uncertainty of the estimation is still not properly addressed. This is different from our mean-variance paradigm where the document ranking relies on both the mean *and* variance of the probability estimation of document relevance.

Our preliminary study on collaborative filtering has demonstrated that ranking derived from the analysis of mean and variance improves recommendation performance significantly [20]. We now provide a comprehensive treatment, looking at a more general application: text retrieval. Our formulations in this paper are flexible for both users' *risk-averse* and *risk-loving* behaviors, whereas our previous work focused only on risk-averse behaviors in collaborative filtering.

3 Mean-Variance Ranking

3.1 Relevance Return of a Ranked List

The task of an IR system is to predict, in response to a user information need, which documents are relevant. Suppose, given the information need, the IR system returns a ranked list consisting of n documents from rank 1 to n – in an extreme case, all the documents need to be ordered when n equals the number of documents in the collection. Let r_i be the estimated relevance score of a document in the list during retrieval, where $i = \{1, ..., n\}$, for each of the rank positions. We intentionally keep the discussion general, while bearing in mind that the exact definition of the relevance score, either degree of relevance or probability of relevance [14], relies on the system's assumption of the relevance and adopted retrieval model.

Our objective is to find an optimal ranking that has the maximum effectiveness in response to the given user information need. There are many ways of defining the effectiveness of a ranked list. A straightforward way is to consider the weighted average of the relevance scores in the list as:

$$R_n \equiv \sum_{i=1}^{n} w_i r_i, \tag{1}$$

where R_n denotes the overall relevance of a ranked list. We assign a variable w_i, where $\sum_{i=1}^{n} w_i = 1$, to each of the rank positions for differentiating the importance of rank positions. This is similar to the discount factors that have been applied to IR evaluation in order to penalize late-retrieved relevant documents [9]. It can be easily shown that when $w_1 > w_2 ... > w_n$, the maximum value of R_n gives the ranking order $r_1 > r_2 ... > r_n$. This follows immediately that maximizing R_n, by which the document with the highest relevance measure is retrieved first, the document with the next highest is retrieved second, and so on, is equivalent to the PRP. By contrast, in finance, R_n is the overall future return of a portfolio having n stocks; r_i is the return of individual stock i, while w_i is the percentage of the budget invested in the stock i.

However, the overall relevance R_n cannot be calculated with certainty. It relies on the estimations of relevance scores r_i of documents from retrieval models. As we discussed, uncertainty can arise through the estimations. To address

such uncertainty, we make a probability statement about the relevance scores, assuming the relevance scores are random variables and have their own probability distributions. Their joint distribution is summarized by using the means and (co)variances. Mathematically, let $E[r_i]$, $i = \{1, ..., n\}$ be the means (the expected relevance scores), and let C_n be the covariance matrix. The non-diagonal element $c_{i,j}$ in the matrix indicates the covariance of the relevance scores between the document at position i and the document at position j; the diagonal element $c_{i,i}$ is the variance of the individual relevance score, which indicates the dispersion from the mean $E[r_i]$. The calculations of the mean and variance in text retrieval are discussed in Section 4.1.

Introducing $E[r_i]$ and $c_{i,j}$ gives the expected overall relevance of a ranked list and its variance as follows:

$$E[R_n] = \sum_{i=1}^{n} w_i E[r_i] \tag{2}$$

$$Var(R_n) = \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j c_{i,j}, \tag{3}$$

where $Var(R_n)$ denotes the variance of the ranked list. For the derivation of Eq. (3), we refer to [11]. By contrast, in finance, $E[R_n]$ is regarded as the expected overall return of a portfolio containing n stocks; $Var(R_n)$ is the variance of the overall return of the portfolio, a measure of the *volatility* (or risk) associated with the portfolio [7]. Notice that we adopt variance for mathematical convenience, while it is also possible to measure the risk by the standard deviation.

3.2 Expected Relevance vs. its Variance

The mean and variance summarize our belief about the effectiveness of a ranked list from the following two aspects. The mean measures the overall relevance returned from the ranked documents as a whole, and for optimal retrieval it seems intuitively obvious to maximize the mean. This is essentially what the PRP has suggested. But, on the other hand, the variance measures the likelihood that we have under- or over-estimated the expected relevance. That is it represents the level of a risky prospect if we produce an optimal rank order by maximizing the mean. If it is underestimated, the user will likely be pleased with the output, whereas if it is overestimated, the user will likely be displeased with the output. Thus, for *risk-averse* users or systems, the variance should stay as small as possible, but, for *risk-loving* users or systems, a large variance might be a preferable attribute.

For the risk-averse case, consider the following example of movie recommendation, a popular application in IR. The task is to suggest top-n ranked movie items that the user is most likely to like, given the user's past ratings (a representation of information needs). In this example, the movie items' relevance scores have multiple values 1-6, with 1 being the lowest rating and 6 being the highest one. Suppose that the system returns a top-10 ranked list of movie items as a recommendation solution. Fig. 1 plots the randomly sampled recommendation solutions, marked by circles, each of which contains top-10 ranked items. Their means and variances are calculated based on Eq. (2) and Eq. (3). The item-based model [17] was used to predict the individual items' relevance, and the covariance matrix is estimated from the historic rating data. For a risk-averse decision, the graph shows that, given a mean value (the expected relevance), one can find an efficient ranking solution that has the minimal variance (risk). Varying the mean value, we obtain a set of efficient ranking solutions; they are geometrically located on the upper left boundary. In finance, the boundary is called the *efficient frontier* [11]. In IR, it represents the set of ranking solutions that have maximal mean (the expected overall relevance) given an upper bound on the variance (risk).

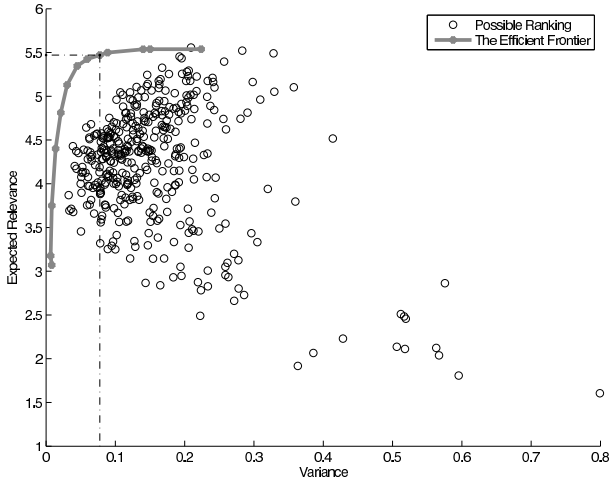

Figure 1: The relationship between the expected overall relevance and variance of the top-10 ranked list. The curve is the *Efficient Frontier*.

Therefore, mathematically, we have the following criteria for risk-averse ranking:

1. *Maximize the mean $E[R_n]$ regardless of its variance*
2. *Minimize the variance $Var(R_n)$ regardless of its mean*
3. *Minimize the variance for a specified mean t (parameter):* $$min\ Var(R_n),\ \text{subject to } E[R_n] = t \tag{4}$$
4. *Maximize the mean for a specified variance h (parameter):* $$max\ \ E[R_n],\ \text{subject to } Var(R_n) = h \tag{5}$$
5. *Maximize the mean and minimize the variance by using a specified risk preference parameter b:* $$max \quad O_n = E[R_n] - bVar(R_n) \tag{6}$$

The first two criteria provide the two simplest cases, optimizing either of the quantities; the first criterion is what the PRP has optimized, while the second one gives minimum variance solutions, which might be suitable for the most cautious users or system setup.

The important ones are the ranking criteria 3, 4, and 5, and they are mathematically equivalent [2]. Here, we focus on the formulation of Eq. (6) as it is the common objective function used in practice. For a risk-averse solution, the parameter $b > 0$. The efficient frontier plotted in Fig. 1 is a set of the solutions that maximize the objective function as b ranges from 0 (the right side) to 40 (the left side). Note that the frontier cannot tell us which one is the single best ranked list for a given user information need; it has to be dependent on the user's risk preference, and can be tuned for the specified evaluation metric, shown in Section 4.

In finance, investors are usually assumed to be risk-averse. But in IR risk-loving behaviors may be useful in some situations. For instance, pseudo relevance feedback is a risky solution since it assumes the first few retrieved documents are relevant. It is often reported to have an ability to improve MAP (mean average precision) [12, 19]. In this regard, it is beneficial to study the effectiveness of the risk-loving solutions when we set $b < 0$ in the objective function. In fact, by applying the utility theory, one can give a more general justification of the objective function in Eq. (6) [22]. For readability, its detailed derivation is given in Appendix.

3.3 Diversification vs. Uncertainty

This section discusses diversification, and formally derives its relationship with the uncertainty of a ranked list. A further decomposition of the variance in Eq. (3) gives

$$\begin{aligned} Var(R_n) &= \sum_{i=1}^{n} w_i^2 c_{i,i} + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n} w_i w_j c_{i,j} \\ &= \sum_{i=1}^{n} w_i^2 \sigma_i^2 + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n} w_i w_j \sigma_i \sigma_j \rho_{i,j}, \end{aligned} \tag{7}$$

where $\sigma_i = \sqrt{c_{i,i}}$ is the standard deviation, and $\rho_{i,j} = \frac{c_{i,j}}{\sigma_i \sigma_j}$ is the correlation coefficient. $\rho_{i,j} = 1$ means that there is an exact positive relationship between two documents, $\rho_{i,j} = 0$ means no relationship between the two documents, and $\rho_{i,j} = -1$ indicates an exact negative relationship between the two documents. As shown in Eq. (7), to reduce the uncertainty of the relevance prediction for the returned documents, we need to have small correlation coefficients (preferable negative correlations) between documents. This means diversifying the documents in the ranked list will reduce the variance and therefore the uncertainty of the expected overall relevance of the returned documents.

To understand this, consider two extreme cases: in the first case, suppose we have a ranked list consisting of two documents, where the correlation coefficient ρ between them is -1. This means that their estimated relevance scores change in the exact opposite direction in response to different information needs. The *volatility* (the change) of the documents' relevance cancels one another completely and leads to a situation where the ranked list has no volatility at all. As a result, a certain amount of relevance for any kind of user information needs is maintained. Conversely, when we have two documents that are perfectly correlated ($\rho = 1$) in the list, the relevance returns of the two documents move in the perfectly same direction in response to different information needs. In this case, the returned relevance of the list mimics that of each of the two documents. As a result, the list contains the same amount of uncertainty (risk) as each of the two documents alone. In this case, risk is not reduced.

3.4 Document Ranking - A Practical Solution

Unlike in finance, the weight w_n in IR, representing the discount for each rank position, is a discrete variable. Therefore, the objective function in Eq. (6) is no-smooth, and there is no easy solution for directly optimizing it. In this section, we present an efficient document ranking algorithm by sequentially optimizing the objective function. It is based on the observation that the larger the rank of a relevant document, the less likely it would be seen or visited by a user. An economical document selection strategy should first consider rank position 1, and then add documents to the ranked list sequentially until reaching the last rank position n. For each rank position, the objective is to select a document that has the maximum increase of the objective function. Notice that such a sequential update may not necessarily provide a global optimization solution, but it provides an excellent trade-off between accuracy and efficiency.

The increase of the objective function from position $k-1$ to k is:

$$\begin{aligned} O_k - O_{k-1} = &\sum_{i=1}^{k} w_i E[r_i] - b\sum_{i=1}^{k}\sum_{j=1}^{k} w_i w_j c_{i,j} \\ &- \sum_{i=1}^{k-1} w_i E[r_i] + b\sum_{i=1}^{k-1}\sum_{j=1}^{k-1} w_i w_j c_{i,j}, \end{aligned} \tag{8}$$

Table 1: Overview of the six tested collections.

Name	Description	# Docs	Topics	# Topics
TREC2007 enterprise track document search	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 Robust 2004	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
TREC subtopic collection	Financial Times of London 1991-1994	210,158	TREC 6,7,8 interactive track topics	20

where $k \in \{2, ..., n\}$. The final equation is derived as

$$O_k - O_{k-1} = w_k(E[r_k] - b{w_k\sigma_k}^2 - 2b\sum_{i=1}^{k-1} w_i\sigma_i\sigma_k\rho_{i,k})$$

Since w_k is a constant for any document in rank k, dropping it gives the following ranking criterion:
select a document at rank k that has the maximum value of

$$E[r_k] - b{w_k\sigma_k}^2 - 2b\sum_{i=1}^{k-1} w_i\sigma_i\sigma_k\rho_{i,k} \tag{9}$$

3.5 Discussions

Ranking principle: Eq. (9) extends the PRP into a more general situation. It contains three components. The first component concerns the point estimate of the relevance $E[r_k]$, which is essentially equivalent to the PRP. The second component generalizes the PRP by considering the uncertainty of the point estimate. It concerns the variance of the estimates of individual documents. The third component extends it further by looking at the correlations between the estimates. A positive b produces *risk-averse* ranking where negatively correlated (with previously retrieved documents) documents should be given high ranking scores. In this case, diversification, which is quantified by the weighted average of the correlations between the ranked documents (see the second component in Eq. (7)), is effectively incorporated into the document ranking. The smaller the parameter b is, the less *risk-averse* the ranking is. When $b = 0$, it goes back to the PRP, which only considers the point estimate $E[r_k]$. When $b < 0$, the ranker intends to take more risk. The impact of b and its relations with IR metrics are studied in Section 4.

Higher moments: The discussions so far rely on a Gaussian assumption about the distribution of relevance scores. Most probabilistic retrieval models are, however, not Gaussian. Strictly speaking, using the first two moments (the mean and variance) may not be entirely adequate to describe the distribution, and the third moment might be needed to indicate the skewness (asymmetry to the mean) if any. But in practice as an approximation the analysis of the mean and variance is fair enough to trade-off between complexity and speed.

Relations with prior work Our ranking approach is a general one. When $b > 0$, the last component in Eq. (9) resembles the MMR (Maximal Marginal Relevance) re-ranking method [3]. As discussed, the MMR re-ranking, as a heuristic method, linearly combines relevance and novelty using a parameter between 0 and 1. It judges a document to have high "marginal relevance" if it is both relevant to the query and contains minimal similarity to already selected documents. Thus, our probabilistic approach provides a theoretical justification. Also, our formulation is less computationally expensive as it does not need to find minimal similarity. The empirical comparison between them is in Section 4.3.2.

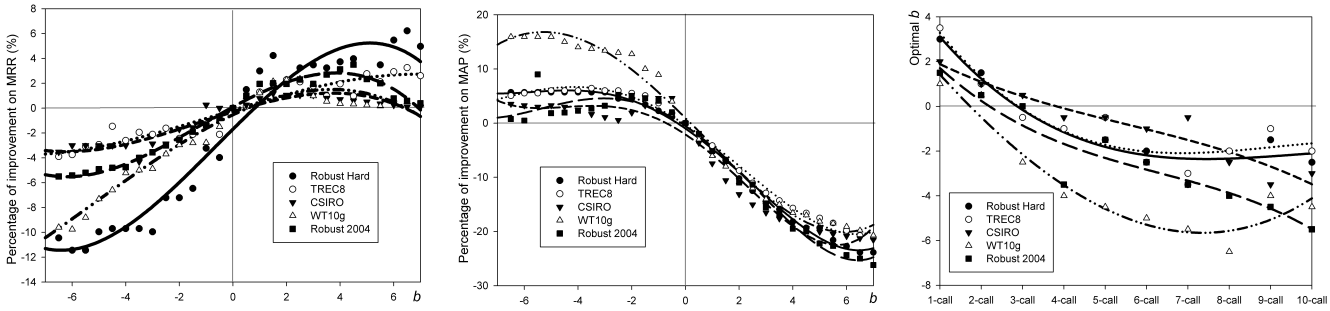

Figure 2: Relationship between: (a) MRR and b, (b) MAP and b, and (c) optimal b and k-call, $k = 1, ..., 10$.

The ranking criterion in Eq. (4) gives an alternative formulation for the objective at which Chen and Karger in [4] have aimed: fixing the amount of relevance the user intends to receive in Eq. (4) (by setting the parameter t) is similar to optimizing the number of relevant documents in the ranked list, proposed in [4]. The merit of our mean-variance formulation is that the resulting ranking principle is a general one and can be applied to any IR models, whereas the formulation in [4] is only suitable for binary relevance IR models as it explicitly relies on the assumption of binary relevance, and is coupled with the IR model during ranking.

4 Empirical Study and Evaluation

Our evaluation focuses on text retrieval, where ad hoc and subtopic retrieval [5] are studied; we report results on five TREC test collections for ad hoc retrieval and one TREC collection for subtopic retrieval. These collections are described in Table 1. Our main goal is to validate our theoretical development, and investigate the effectiveness of the various risk preference settings.

4.1 Calculation of Mean and Variance in IR

Different probabilistic retrieval models result in different estimators of $E[r_i]$ and C_n. $E[r_i]$ can be determined by a point estimate from the specific text retrieval model that has been applied. In this paper, three widely-adopted retrieval models, namely, the Dirichlet and Jelinek-Mercer smoothing language models [23], and the BM25 model [16] are used to calculate the expected relevance scores. For the two language models, we employ the posterior mean of the query-generation model as the estimator. Strictly speaking, the BM25 scores are not calculated in a probabilistic way, but it is reasonable to assume that its output scores are random variables and have uncertainty associated with them.

The covariance matrix C_n represents both the uncertainty and correlation associated with the estimations. Although they are largely missing in current probabilistic retrieval models, there are generally two ways of estimating them in practice. Formally, they should be determined by the *second moment* of the relevance scores. For instance, one can estimate the (co)variances of individual document models (parameters) by adopting the Bayesian paradigm [1, 25]. Alternatively, for given two documents, the covariance between their relevance scores can be approximated by the covariance with respect to their term occurrences. This is similar to use historic data of two stocks to calculate the correlation between their future returns in finance.

In this paper, for the two language models, the relevance scores are assumed to follow the Dirichlet distribution, and their variances σ^2 are thus conveniently calculated [25]. Since the BM25 is not a probability model, we set the variances as a constant for all documents. This allows us to study the effectiveness of the correlations solely when using the BM25 scores. The correlation ρ is approximated by using the Pearson's correlation coefficient between each pair of documents' term vectors. Ranking is based on the sequential update formulated in Eq. (9), and only the correlations with the previously retrieved documents are needed. Thus, the computational load of calculating covariances has been reduced significantly. The weights of rank positions w_i are chosen according to the discount factors in [9].

4.2 Ad Hoc Text Retrieval

4.2.1 Parameter: As studied by Thom and Scholer in [18], the IR evaluation metrics generally have two distinct categories: those strongly biased towards early-retrieved documents, such as Mean Reciprocal Rank (MRR), and those trying to capture a broader summary of retrieval performance, including Mean Average Precision (MAP). Let us first evaluate the impact of the risk preference parameter b toward the two categories.

Fig. 2 (a) and (b) plot the percentage of improvements against a varying b from -6 (risk-loving) to 6 (risk-averse). The fitted curves are based on the data points, and the percentage of improvement on the MRR and other metrics is based on the improvement over the setting where $b = 0$ (equivalent to the PRP). In this experiment, the Dirichlet smoothing language model (where μ=2000, a typical setting) is adopted for obtaining the relevance scores. From Fig. 2 (a), we can see that positive values of b, i.e, diversifying search results, helps improve the MRR metric. This explains that by "investing" into different kinds of documents, the actual chance of returning the first relevant documents as early as possible can be actually increased.

By contrast, for a metric capturing a broader summary of retrieval performance such as MAP, Fig. 2 (b) shows that negative values of b, which emphasize a document positively correlated with the early-retrieved documents, help improve the performance. "Investing" in the same "type" of documents is a risky action (big variance), and might hurt the MRR metric. But, on average, it does increase the performance of the entire ranked list (in this setting, $n = 1000$). This is similar to the effectiveness of pseudo relevance feedback in ad hoc retrieval, i.e., the top ranked documents are generally likely to be relevant, and to find other documents similar to these top ranked ones will help improve MAP [19].

To further understand these risk behaviors, we then study how the parameter behaves under a risk-sensitive metric called k-call at 10, or k-call for simplicity, proposed in [4]. Given a ranked list, k-call is one if at least k of the top-10 documents returned 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. Thus, a *risk-averse* approach, which can reliably find one relevant document, is preferred for 1-call, while a *risk-loving* approach is favored for 10-call [4].

Table 2: Comparison of our approach against the PRP via three retrieval models. For a retrieval model, three lines in each cell are performance of our approach and the PRP, and performance gain of our approach over the PRP, respectively. A Wilcoxon signed-rank test is conducted and statistically significant improvements are marked with $*$.

(a) Our approach vs. the PRP via the Dirichlet smoothing language model

Measures	CSIRO	WT10g	Robust	Robust hard	TREC8	Measures	CSIRO	WT10g	Robust	Robust hard	TREC8
MRR	0.774	0.587	0.612	0.427	0.635	Prec@10	0.684	0.382	0.387	0.227	0.433
	0.765	0.574	0.596	0.402	0.615		0.653	0.333	0.379	0.211	0.407
	+1.18%	+2.26%	+2.68%	+6.22%*	+3.25%		+4.75%	+14.71%*	+2.11%	+7.58%*	+6.39%*
MAP	0.404	0.225	0.232	0.092	0.226	Prec@100	0.448	0.196	0.173	0.129	0.213
	0.388	0.202	0.228	0.089	0.223		0.432	0.178	0.169	0.124	0.204
	+4.12%	+11.39%*	+1.75%	+3.37%	+1.35%		+3.70%	+10.11%*	+2.37%	+4.03%	+4.41%
NDCG	0.664	0.499	0.501	0.317	0.493	1-call	0.98	0.902	0.877	0.8	0.94
	0.651	0.477	0.483	0.312	0.484		0.98	0.88	0.819	0.74	0.88
	+2.01%	+4.55%	+3.53%	+1.61%	+1.98%		0.0%	+2.50%	+7.08%*	+8.11%*	+6.82%*
NDCG@10	0.170	0.169	0.183	0.083	0.162	6-call	0.74	0.34	0.278	0.08	0.32
	0.162	0.152	0.179	0.077	0.154		0.66	0.202	0.261	0.04	0.28
	+4.66%	+11.22%*	+2.31%	+7.75%*	+5.28%		+12.12%*	+68.32%*	+6.51%*	+100.0%*	+14.29%*
NDCG@100	0.382	0.318	0.341	0.180	0.326	8-call	0.52	0.16	0.151	0.02	0.2
	0.367	0.295	0.331	0.173	0.315		0.38	0.11	0.129	0.005	0.16
	+4.02%	+7.63%*	+2.98%	+3.89%	+3.65%		+36.84%*	+45.45%*	+17.05%*	+300.0%*	+25.00%*
Prec@1	0.147	0.064	0.056	0.046	0.072	10-call	0.28	0.057	0.042	0.02	0.04
	0.145	0.062	0.054	0.046	0.072		0.2	0.02	0.036	0.0	0.02
	+1.38%	+3.23%	+3.70%	0.0%	0.0%		+40.00%*	+185.0%*	+16.67%*	–*	+100.0%*

(b) Our approach vs. the PRP via the Jelinek-Mercer smoothing language model

Measures	CSIRO	WT10g	Robust	Robust hard	TREC8	Measures	CSIRO	WT10g	Robust	Robust hard	TREC8
MRR	0.869	0.558	0.592	0.393	0.589	Prec@10	0.729	0.384	0.399	0.242	0.444
	0.843	0.492	0.549	0.352	0.472		0.653	0.309	0.371	0.229	0.398
	+3.08%	+13.41%*	+7.83%*	+11.65%*	+24.79%*		+11.64%*	+24.27%*	+7.55%*	+5.68%*	+11.56%*
MAP	0.41	0.182	0.204	0.084	0.212	Prec@100	0.432	0.167	0.156	0.125	0.219
	0.347	0.157	0.185	0.078	0.198		0.406	0.143	0.148	0.122	0.209
	+18.16%*	+15.92%*	+10.27%*	+7.69%*	+7.07%*		+6.40%*	+16.78%*	+5.41%	+2.46%	+4.78%
NDCG	0.633	0.433	0.421	0.271	0.452	1-call	1.0	0.92	0.865	0.81	0.94
	0.587	0.398	0.396	0.252	0.422		0.98	0.86	0.831	0.78	0.86
	+7.88%*	+8.82%*	+6.25%*	+7.55%*	+7.05%*		+2.04%	+6.98%*	+4.09%	+3.85%	+9.30%*
NDCG@10	0.185	0.157	0.175	0.081	0.149	6-call	0.74	0.28	0.297	0.12	0.32
	0.170	0.141	0.169	0.078	0.140		0.62	0.18	0.241	0.06	0.28
	+8.96%*	+11.23%*	+3.80%	+3.90%	+6.36%*		+19.35%*	+55.56%*	+23.24%*	+100.0%*	+14.29%*
NDCG@100	0.377	0.286	0.314	0.169	0.305	8-call	0.64	0.14	0.181	0.04	0.22
	0.355	0.262	0.292	0.159	0.287		0.44	0.08	0.133	0.02	0.2
	+6.25%*	+9.27%*	+7.55%*	+6.58%*	+6.34%*		+45.45%*	+75.00%*	+36.09%*	+100.0%*	+10.00%*
Prec@1	0.133	0.052	0.049	0.038	0.063	10-call	0.38	0.06	0.064	0.02	0.12
	0.13	0.048	0.044	0.037	0.062		0.26	0.0	0.032	0.0	0.02
	+2.31%	+8.33%*	+11.36%*	+2.70%	+1.61%		+46.15%*	–*	+100.0%*	–*	+500.0%*

(c) Our approach vs. the PRP via the BM25 model

Measures	CSIRO	WT10g	Robust	Robust hard	TREC8	Measures	CSIRO	WT10g	Robust	Robust hard	TREC8
MRR	0.906	0.614	0.619	0.448	0.602	Prec@10	0.776	0.404	0.438	0.267	0.447
	0.893	0.602	0.544	0.442	0.579		0.718	0.353	0.416	0.26	0.431
	+1.46%	+1.99%	+13.79%*	+1.36%	+3.97%		+8.08%*	+14.45%*	+5.29%	+2.69%	+3.71%
MAP	0.434	0.211	0.249	0.101	0.231	Prec@100	0.486	0.179	0.184	0.137	0.233
	0.415	0.191	0.231	0.096	0.225		0.463	0.169	0.177	0.133	0.228
	+4.58%	+10.47%*	+7.79%*	+5.21%	+2.67%		+4.97%	+5.92%*	+3.95%	+3.01%	+2.19%
NDCG	0.683	0.491	0.516	0.332	0.498	1-call	1.0	0.912	0.883	0.78	0.904
	0.667	0.469	0.497	0.322	0.480		1.0	0.86	0.876	0.76	0.88
	+2.33%	+4.60%	+3.87%	+3.03%	+3.85%		0.0%	+6.05%*	+0.80%	+2.63%	+2.73%
NDCG@10	0.193	0.181	0.204	0.089	0.157	6-call	0.8	0.298	0.349	0.103	0.322
	0.184	0.162	0.191	0.086	0.150		0.74	0.24	0.297	0.1	0.32
	+4.83%	+11.94%*	+6.87%*	+3.84%	+4.44%		+8.11%*	+24.17%*	+17.51%*	+3.00%	+0.63%
NDCG@100	0.413	0.317	0.360	0.183	0.325	8-call	0.72	0.182	0.189	0.06	0.284
	0.401	0.297	0.345	0.181	0.314		0.62	0.141	0.161	0.04	0.22
	+3.05%	+6.87%*	+4.44%	+1.15%	+3.69%		+16.13%*	+29.08%*	+17.39%*	+50.00%*	+29.09%*
Prec@1	0.151	0.063	0.058	0.049	0.077	10-call	0.4	0.03	0.076	0.02	0.098
	0.149	0.062	0.057	0.049	0.076		0.26	0.02	0.036	0.0	0.02
	+1.34%	+1.61%	+1.75%	0.0%	+1.32%		+53.85%*	+50.0%*	+111.11%*	–*	+390.0%*

The relationship between the optimal value of b and k-call (k=1,...,10) is plotted in Fig. 2 (c). The figure shows that when k is small such as 1 and 2, the optimal b is positive for all collections. This means that diversifying top-10 search results reduces the risk of not returning any relevant documents. When k increases, the optimal b becomes negative. This shows that a risk-loving approach will increase the chance of finding many relevant documents.

4.2.2 Performance: We now test the performance against various setups and metrics. 5-fold cross validation is carried out on the four ad hoc test collections. Queries in each collection were randomly partitioned. For each partition, model parameters were trained with all the other partitions and performance for the partition is evaluated with the trained parameters. We evaluated the concatenated ranked lists from all 5 partitions, and report the results in Table 2. When compared with the PRP via the Dirichlet smoothing language model in Table 2 (a), out of the 60 reported results, 57 improvements are positive, and 27 improvements are statistically significant. When compared with the PRP via the Jelinek-Mercer smoothing language model in Table 2 (b), out of the 60 reported results, all the improvements are positive, and 48 improvements are statistically significant. When compared with the PRP via the BM25 model in Table 2 (c), out of the 60 reported results, 58 improvements are positive, and 22 improvements are statistically significant.

Overall, our approach largely outperformed the PRP in our experiments. As different IR metrics may reflect different risk-taking preferences, e.g., risk-loving or risk-averse, by tuning the parameter, our approach provides an effective way for optimizing different IR metrics.

4.3 Subtopic Text Retrieval

Subtopic retrieval is concerned with finding documents that cover many different subtopics of a general query topic. In subtopic retrieval, the utility of a document is dependent on other documents in the ranking. To study the effectiveness of our ranking approach in this task, we compare our approach with the PRP and the MMR ranking method [3]. We also study the relationship between the parameter b and a range of subtopic specific metrics. We used the TREC interactive track subtopic collection, which, to our knowledge, is the only publicly available subtopic collection. The collection consists of 20 topics adapted from TREC ad hoc retrieval topics. The number of subtopics for these topics ranges from 7 to 56 with an average length of 20. For a topic, the relevance judgment for each document is a vector, whose length is the number of subtopics. The vector consists of 1 and 0, which represents relevant and not relevant for a subtopic, respectively.

We report the metric called α-NDCG (Normalized Discounted Cumulated Gain) proposed by [5], which takes into account both novelty and relevance of documents. A parameter α between 0 and 1 balances novelty and relevance in α-NDCG, and when $\alpha = 0$, α-NDCG is equivalent to standard NDCG [9]. The larger the α value, novelty is rewarded more over relevance, and vice versa. We fixed α as 0.5 for a balance between novelty and relevance.

We also extended the traditional Recall at n and MRR metrics to define two new subtopic retrieval metrics, namely, subtopic Recall (sub-Recall) at n and subtopic MRR (sub-MRR). These two new metrics emphasize novelty, and have simpler definitions than α-NDCG, therefore, will likely help us gain a more direct view of the effect of parameter b on subtopic retrieval. Suppose there are N subtopic for a topic, we define sub-Recall at n as the number of different subtopics covered by the top n documents divided by N. Given a topic, we define sub-MRR as the inverse of the rank of the first position where documents covering all the subtopics have been retrieved. Therefore, sub-MRR awards a system which can retrieve all subtopics as close to the top of a ranked list as possible. We average sub-Recall at n and sub-MRR over a number of topics to get their means, respectively.

4.3.1 Parameter: We plot the relationship between sub-Recall at n and the corresponding optimal value of b in Fig. (3). In Fig. (3), when the cut-off points are beyond 20, the optimal b is around 0.0, i.e., little or no diversification is employed. This tells us that for top 20 or more documents, the PRP can perform as well as our risk-aware approach, i.e., a sufficient number of relevant documents retrieved by the PRP can cover different sub-topics well. However, for lower cut-off points from 2 to 15, the optimal b is always between 4.0 and 12.0, showing that a *risk-averse* approach helps choose documents on different aspects of a topic.

4.3.2 Performance: We compared our approach with the PRP and MMR [3] ranking criterion. We again used 5-fold cross-validation for subtopic retrieval on the TREC subtopic collection to optimize the parameters, and the results are shown in Table 3.

We can see from Table 3 that our approach can largely outperform both the PRP and the MMR method. Compared with the PRP, out of 30 reported results, all the performance gains by our approach are positive, and 15 performance gains are statistically significant. Compared with the MMR method, out of 30 results, all the performance gains by our approach are positive, and 12 performance gains are statistically significant.

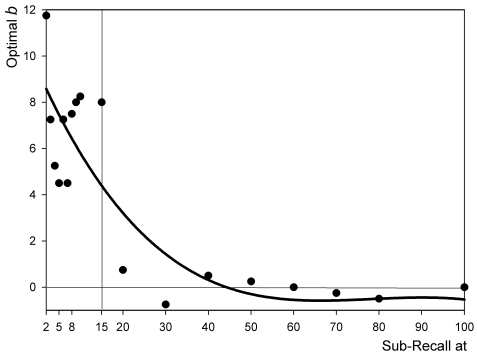

Figure 3: Relation between optimal b and sub-Recall@n.

The MMR method can slightly outperform the PRP when the cut-off points of sub-Recall and α-NDCG are below 10, but performed worse than the PRP when the cut-off points are above 10; while our approach consistently outperformed the PRP.

We think the good performance of our approach over the MMR method is due to the reason that our approach provides a more principled way for taking into account both variance and diversification in document ranking. Besides, in our approach, correlations between a new document and all top ranked documents are considered, while the MRR method only considers the maximum similarity between a new document and one top ranked document. We think that the use of only one pair of documents' similarity in the MMR method may result in unstable results when the ranked list is long.

5 Conclusion and Future Work

To address the ranking uncertainty, we have followed the school of thinking from the Modern Portfolio Theory in finance, and presented the *mean-variance* paradigm for document ranking in text retrieval. The analysis of the mean and variance of a ranked list led to a new and generalized document ranking principle.

Handling uncertainty is critical for IR modelling. There are fruitful avenues for future investigations into the proposed mean-variance paradigm, including 1) the analysis of mean and variance of IR evaluation metrics. 2) Variance as an indicator of the risk does not distinguish a bad surprise from a good surprise. It is worthwhile investigating "downside risk" in finance that considers only bad surprises. 3) Large numbers of documents make the estimation of correlations between all documents a great challenge. How to effectively and efficiently calculate the variance (risk) and correlation of the estimation remains an open question. 4) It is of great interest to study the mean-variance analysis in other IR applications such as filtering [21], multimedia retrieval, and advertising.

6 References

[1] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[2] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[3] J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*, 1998.

[4] H. Chen and D. R. Karger. Less is more: probabilistic models for retrieving fewer relevant documents. In *SIGIR*, 2006.

Table 3: Comparison of our approach, the PRP, and the MMR method via three retrieval models. A Wilcoxon signed-rank test is conducted and statistically significant improvements are marked with $*$.

(a) Our approach vs. the PRP. In each cell, the first line shows the performance of our approach, and the second line shows the performance of the PRP and gain of our method over the PRP.

Models	Dirichlet	Jelinek-Mercer	BM25	Models	Dirichlet	Jelinek-Mercer	BM25
sub-MRR	0.014	0.011	0.009	α-NDCG@5	0.417	0.372	0.367
	0.013 (+7.69%*)	0.01 (+10.00%*)	0.008 (+12.50%*)		0.404 (+3.22%)	0.289 (+28.72%*)	0.358 (+2.51%)
sub-Recall@5	0.324	0.255	0.275	α-NDCG@10	0.461	0.434	0.418
	0.298 (+8.72%*)	0.225 (+13.33%*)	0.271 (+1.48%)		0.453 (+1.77%)	0.372 (+16.67%*)	0.407 (+2.70%)
sub-Recall@10	0.381	0.366	0.352	α-NDCG@15	0.494	0.462	0.465
	0.354 (+7.63%*)	0.352 (+3.98%)	0.332 (+6.02%*)		0.492 (+0.41%)	0.412 (+12.14%*)	0.449 (+3.56%)
sub-Recall@20	0.472	0.458	0.464	α-NDCG@20	0.517	0.482	0.492
	0.444 (+6.31%*)	0.427 (+7.26%*)	0.452 (+2.65%)		0.509 (+1.57%)	0.425 (+13.41%*)	0.476 (+3.36%)
sub-Recall@100	0.563	0.582	0.577	α-NDCG@100	0.587	0.555	0.569
	0.556 (+1.26%)	0.55 (+5.82%*)	0.558 (+3.41%)		0.583 (+0.69%)	0.499 (+11.22%*)	0.551 (+3.27%)

(b) Our approach vs. the MMR method. In each cell, the first line shows the performance of our approach, and the second line shows the performance of the MMR method and gain of our method over the MMR method.

Models	Dirichlet	Jelinek-Mercer	BM25	Models	Dirichlet	Jelinek-Mercer	BM25
sub-MRR	0.014	0.011	0.009	α-NDCG@5	0.417	0.372	0.367
	0.012 (+16.67%*)	0.009 (+22.22%*)	0.007 (+28.57%*)		0.407 (+2.46%)	0.293 (+26.96%*)	0.355 (+3.38%)
sub-Recall@5	0.324	0.255	0.275	α-NDCG@10	0.461	0.434	0.418
	0.304 (+6.58%*)	0.234 (+8.97%*)	0.27 (+1.85%)		0.454 (+1.54%)	0.367 (+18.26%*)	0.411 (+1.70%)
sub-Recall@10	0.381	0.366	0.352	α-NDCG@15	0.494	0.462	0.465
	0.362 (+5.25%)	0.351 (+4.27%)	0.344 (+2.33%)		0.489 (+1.02%)	0.394 (+17.26%*)	0.451 (+3.10%)
sub-Recall@20	0.472	0.458	0.464	α-NDCG@20	0.517	0.482	0.492
	0.455 (+3.74%)	0.41 (+11.71%*)	0.446 (+4.04%)		0.509 (+1.57%)	0.411 (+17.27%*)	0.469 (+4.90%)
sub-Recall@100	0.563	0.582	0.577	α-NDCG@100	0.587	0.555	0.569
	0.558 (+0.90%)	0.55 (+5.82%*)	0.558 (+3.41%)		0.583 (+0.69%)	0.486 (+14.20%*)	0.542 (+4.98%)

[5] C. L. A. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *SIGIR*, 2008.

[6] W. S. Cooper. The inadequacy of probability of usefulness as a ranking criterion for retrieval system output. *University of California, Berkeley*, 1971.

[7] E. J. Elton and M. J. Gruber. *Modern portfolio theory and investment analysis*. J. Wiley and Sons, 2006.

[8] M. D. Gordon and P. Lenk. A utility theoretic examination of the probability ranking principle in information retrieval. *JASIS*, 42(10):703–714, 1991.

[9] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 2002.

[10] J. Lafferty and C. Zhai. Document language models, query models, and risk minimization for information retrieval. In *SIGIR*, 2001.

[11] H. Markowitz. Portfolio selection. *Journal of Finance*, 1952.

[12] D. Metzler, T. Strohman, Y. Zhou, and W. B. Croft. Indri at TREC 2005: Terabyte track. In *TREC*, 2005.

[13] S. E. Robertson. The probability ranking principle in IR. *Journal of Documentation*, pages 294–304, 1977.

[14] S. E. Robertson and N. Belkin. Ranking in principle. *Journal of Documentation*, pages 93–100, 1978.

[15] S. E. Robertson and K. Spärck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3):129–46, 1976.

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

[17] B. Sarwar, G. Karypis, J. Konstan, and J. Reidl. Item-based collaborative filtering recommendation algorithms. In *WWW*, 2001.

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

[19] S. Tomlinson. Early precision measures: implications from the downside of blind feedback. In *SIGIR*, 2006.

[20] J. Wang. "Mean-variance analysis: A new document ranking theory in information retrieval. In *ECIR*, 2009.

[21] J. Wang, A. P. de Vries, and M. J. T. Reinders. Unifying user-based and item-based collaborative filtering approaches by similarity fusion. In *SIGIR*, 2006.

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

[23] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *SIGIR*, 2001.

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

[25] J. Zhu, J. Wang, M. Taylor, and I. Cox. Risky business: Modeling and exploiting uncertainty in information retrieval. In *SIGIR*, 2009.

APPENDIX

In Section 3.2, we have given Eq. (6) on the basis of our mean-variance analysis. Here we present an additional justification from a Bayesian view point. The intuition is that the loss function for estimating the returned relevance of a ranked list is asymmetric. To model this, we adopt the LINEX asymmetric loss function [22]:

$$L(\hat{R_n}, R_n) = e^{b(\hat{R_n} - R_n)} - b(\hat{R_n} - R_n) - 1, \qquad (10)$$

where b is the parameter to balance the loss. When $b > 0$, the loss of over-estimate is larger than that of under-estimate, and when $b < 0$, otherwise.

From the Bayesian point of view, the returned overall relevance of a top-n ranked document list is a random variable. The posterior probability of the R_n can be written as $p(R_n|r_i, ..., r_n)$. Integrating out the unknown hidden variable R_n gives the expected loss as:

$$\begin{aligned} E^{R_n}[L(\hat{R_n}, R_n)] &= \int L(\hat{R_n}, R_n) p(R_n|O) dR_n \\ =& e^{b\hat{R_n}} E^{R_n}(e^{-bR_n}|r_i, ..., r_n) - \\ & b(\hat{R_n} - E^{R_n}(R_n|r_i, ..., r_n)) - 1, \end{aligned} \qquad (11)$$

where E denotes the expectation. $\hat{R_n}$ is the Bayes estimator of R_n with respect to the cost function L. The optimal estimator of R_n should minimize the expected loss function. Minimizing Eq. (11) gives the optimal Bayesian estimator as follows (for detailed information, we refer to [22]):

$$\hat{R_n^B} = -(1/b) \ln(E^{R_n}(e^{-bR_n}|r_i, ..., r_n)) \qquad (12)$$

If the overall relevance R_n is assumed to be a normal distribution, one can derive the estimation analytically as follows:

$$\hat{R_n^B} = E[R_n] - \frac{b}{2} Var(R_n), \qquad (13)$$

where $E[R_n]$ is the posterior mean and $Var(R_n)$ is the posterior variance. Replacing $b/2$ with b gives Eq. (6). Our derivation shows that, for selecting an optimal top-n ranked list, maximizing the objective function in Eq. (6) is equivalent to the Bayesian estimator of returned overall relevance that minimizes the asymmetric loss.