Mean-Variance Analysis: A New Document Ranking Theory in Information Retrieval

Jun Wang

University College London
j.wang@adastral.ucl.ac.uk

Abstract. This paper concerns document ranking in information retrieval. In information retrieval systems, the widely accepted probability ranking principle (PRP) suggests that, for optimal retrieval, documents should be ranked in order of decreasing probability of relevance. In this paper, we present a new document ranking paradigm, arguing that a better, more general solution is to optimize top-n ranked documents as a whole, rather than ranking them independently. Inspired by the Modern Portfolio Theory in finance, we 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 maximizes the overall relevance (mean) of the ranked list at a given risk level (variance). Based on this principle, we then derive an efficient document ranking algorithm. It extends the PRP by considering both the uncertainty of relevance predictions and correlations between retrieved documents. Furthermore, we quantify the benefits of diversification, and theoretically show that diversifying documents is an effective way to reduce the risk of document ranking. Experimental results on the collaborative filtering problem confirms the theoretical insights with improved recommendation performance, e.g., achieved over 300% performance gain over the PRP-based ranking on the user-based recommendation.

1 Introduction

Information retrieval (IR) aims at retrieving documents[1] that are relevant to a user's information needs. To be able to effectively present the retrieved documents to the user, the probability ranking principle (PRP) states that [13]:

"If an IR system's response to each query is a ranking of documents in order of decreasing probability of relevance, the overall effectiveness of the system to its user will be maximized."

Despite its success in many IR applications, the principle however leaves the following fundamental issues unsolved. 1) The PRP relies on the assumption

[1] By convention, we use the term *document*. However, the discussion in this paper is generally applicable to any information items, either textual or non-textual.

M. Boughanem et al. (Eds.): ECIR 2009, LNCS 5478, pp. 4–16, 2009.
© Springer-Verlag Berlin Heidelberg 2009

that the relevance of one document to an information need is independent of the other documents in the collection. In many situations, this assumption is not realistic [4]. It is beneficial to develop a more general ranking theory that can deal with document dependency. 2) The PRP employs probability of relevance to represent the uncertainty whether a document will be judged relevant. The PRP assumes that such unknown probability of relevance is a fixed unknown constant and can be calculated with certainty [4]. But, when we estimate the probability of relevance, another type of uncertainty can also arise, for instance, due to limited sample size. Estimation errors in the probability of relevance (commonly reported as a single point estimator, such as mean or mode) may cause unreliably-estimated documents to be retrieved. Therefore, retrieval systems should have a method of quantifying this uncertainty, and address it when ranking documents.

This paper attempts to deal with these two issues by proposing a new *mean-variance* paradigm of document ranking. It is inspired by the Modern Portfolio Theory in finance, introduced in 1952 by Markowitz [10]. The theory is concerned with portfolio selection which involves many possible financial instruments. The task is to select the portfolio of securities (e.g., stocks or shares) that will provide the best distribution of future consumption, given an investment budget. Markowitz' approach is based on the analysis of the expected return of a portfolio and its variance of return, where the latter serves as a measure of risk.

Our work here is focused on the theoretical development; we examine the proper use of relevance measures for document ranking, rather than considering in detail methods to calculate the measures. We draw an analogy between the portfolio select problem in finance and the document ranking problem in information retrieval, and argue that the document ranking problem can be effectively cast as a portfolio selection problem: in response to a user information need, top-n ranked documents are selected and ordered as a whole, rather than ranking them 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 and 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 maximizing the overall relevance of the list given an upper bound on the 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 the diversity of ranked documents are modelled in a principled manner. The new ranking approach has been applied to the recommendation problem. The experiment on collaborative filtering demonstrates that significant performance gain has been achieved.

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

2 Mean-Variance Analysis for Document Ranking

2.1 Expected Relevance of a Ranked List and Its Variance

The task of an IR system is to predict, in response to a user information need (e.g., a query in ad hoc textual retrieval or a user profile in information filtering), 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 candidate set. Let r_i be the relevance measure of a document in the list, 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 measure, either degree of relevance or probability of relevance [14], relies on the system setting and adopted retrieval model.

Our objective is to find an optimal ranked list 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 measures 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 [7]. 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 – by which the document with highest relevance measure is retrieved first, the document with next highest is retrieved second, etc. – is equivalent to the PRP.

However, the overall relevance R_n cannot be calculated with certainty. It relies on the estimations of relevance measures r_n 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 measures, assuming the relevance measures are random variables and have their own probability distributions[2]. 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 measures), and let C_n be the covariance matrix. The non-diagonal element $c_{i,j}$ in the matrix indicates the covariance of the relevance measures 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 measure, which indicates the dispersion from the mean $E[r_i]$.

[2] For instance, to measure the uncertainty associated with the estimation of probability of relevance, one might assume that the probability of relevance ($\theta \in [0, 1]$) is a random variable and follows the Beta distribution.

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 retrieval model that has been applied. For instance, in text retrieval we may employ the posterior mean of the query-generation model in the language modelling approach [12] as the estimator, or, in collaborative filtering, the expected relevance may be obtained by using the expected rating estimated from the user-based or item-based method [5,15].

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. One way is to determine them based on the *second moment* of the predictive retrieval models. For instance, one can estimate the (co)variances of individual document models (parameters) by adopting the Bayesian paradigm [1]. Alternatively, the covariance matrix can be determined from historical information of realized relevance or the features (e.g., terms) that represent documents.

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 R_n. It indicates the dispersion from the mean $E[R_n]$. The validity of Eq. (3) can be seen from the following derivation:

$$\begin{aligned} Var(R_n) =& E[(\sum_i w_i r_i)^2] - E[\sum_i w_i r_i]^2 \\ =& \big(\sum_i \sum_j w_i w_j E[r_i r_j]\big) - \big(\sum_i \sum_j w_i w_j E[r_i]E[r_j]\big) \\ =& \big(\sum_i \sum_j w_i w_j \big(E[r_i r_j] - E[r_i]E[r_j]\big)\big) \\ =& \sum_i \sum_j w_i w_j c_{i,j} \end{aligned} \tag{4}$$

2.2 Relevance vs. Variance: A New Document Ranking Strategy

The mean and variance summarize our belief about the effectiveness of a ranked list from the following two aspects[3]. The mean measures the overall relevance of 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 indicates the dispersion from the expected relevance, representing the level of a risky prospect if we produce an optimal

[3] For simplicity, we use the term mean and expected overall relevance interchangeably.

rank order by maximizing the mean. Therefore, when we optimize the ranked list, its overall variance (risk) is required to stay as small as possible.

The relationship between the expected overall relevance of a ranked list and its variance is illustrated by the relevance-variance graph in Fig. 1. In the figure, a possible top-n ranking solution is characterized by its mean and variance, and represented as a point. Fig. 1 (a) shows that possible solutions are conceptually located in four regions. For optimal retrieval, a ranking solution is preferred to be located inside the upper left region because it has high returned relevance and low variance; conversely, any solution located inside the lower right region needs to be avoided due to its low relevance and large variance. Yet, in many practical situations, it is a trade-off between the returned relevance and variance. We either take more risk (larger variance) in order to obtain more highly relevant documents in the ranked list (the upper right region), or conversely trade off relevancy for having more confidence on the ranked documents (the lower left region).

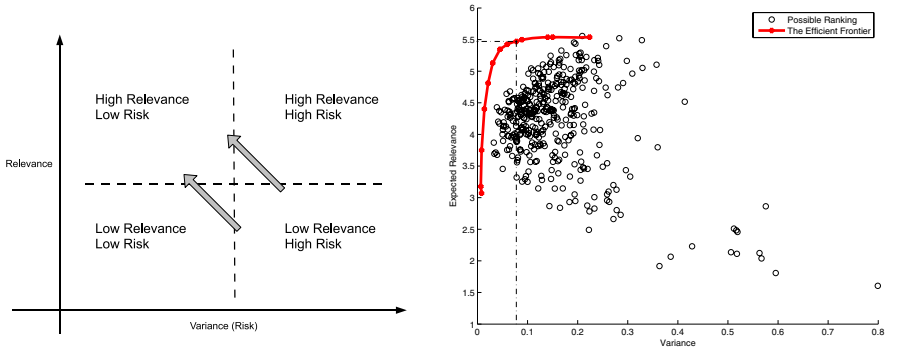
Fig. 1 (b) further demonstrates this in the application of recommendation [5], where the task is to suggest items that the user is most likely to like on the basis of the user's past ratings (a representation of user information needs). In this example, information items are movies, and their relevance has multiple values 1-6, with 1 being the lowest rating (no star) and 6 being the highest one (five stars). Suppose, in response to a recommendation request, the system returns a top-10 ranked list of movie items as a recommendation. Fig. 1 (b) plots the randomly sampled recommendation solutions, 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 [15] was used to predict the individual items' relevance, and the covariance matrix is estimated from the historic rating data. From the graph, we can see that, for a given variance value (risk), one can find an efficient ranking solution that has the highest expected relevance. Varying the variance, we obtain a set of efficient ranking solutions; they are geometrically located on the upper left boundary. Following the same terminology in finance, we name the boundary the *efficient frontier* [10]. From the figure, we can see that the efficient frontier presents the set of ranking solutions that have maximal expected relevance given an upper bound on the variance.

Based on the analysis of mean and variance, we are ready to express our hypothesis about generating a top-n ranked document list as:

> *In response to a given user information need, a retrieval system should generate a ranked document list in such a way that the overall expected relevance of the list is maximized given an upper bound on the risk that the user/system is willing to take, where the risk is expressed by the variance of the list.*

Mathematically, it can be expressed as the following optimization problem:

$$\begin{aligned} &\max\ E[R_n] \\ &\text{subject to } Var(R_n) \leq t \end{aligned} \tag{5}$$



(a) (b)

Fig. 1. The trade-off between relevance and variance in the top-n list of documents. (a) The four regions. (b) Feasible solutions and efficient frontier.

where t denotes an upper bound of risk that needs to meet. As shown in Fig. 1 (b) (see the vertical and horizontal dotted lines), maximizing the relevance given the upper bound of variance is equivalent to minimizing the variance given the lower bound of expected relevance. In practice, it is convenient to formulate the problem by maximizing the following unconstrained objective function:

$$O_n = E[R_n] - \alpha Var(R_n) \tag{6}$$

where α, similar to t, is a parameter adjusting the risk level[4]. The efficient frontier plotted in Fig. 1 (b) is a set of the solutions that maximize the objective function as α ranges from 0 (the right side) to 40 (the left side). It is worth noticing 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 upon the user's risk preference, and can be tuned by the specific performance measure.

2.3 Diversification – A Way to Reduce Uncertainty

This section continues with a discussion of diversification, another importance criterion, for document ranking. A further derivation from Eq. (3) gives

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

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

[4] Alternatively, the objective function in Eq. (6) can be derived formally by Utility theory [19]. The utility parameter a represents the user's risk preference. When $\alpha > 0$, the ranking is risk-averse, while when $\alpha < 0$, it is risk-loving.

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 relevance measures of the returned documents.

To understand this, let us consider two extreme cases. In the first case, suppose we have a ranked list containing two documents, where the correlation coefficient between them is -1. This means that they move in the exact opposite direction in response to different information needs. The *volatility* of the documents (as to whether they are relevant or not relevant) cancels one another completely and this leads to a situation where the ranked list has no volatility at all. As a result, a certain amount of relevancy for any kind of user information needs is maintained. Conversely, when we have two documents that are perfectly correlated (the correlation coefficient equals to 1) in the list, the relevance returns of the two documents move in perfect same direction in response to different information needs. In this case, the relevance return of the list mimics that of the two documents. As a result, the list contains the same amount of uncertainty (risk) as those of the two documents. In this case, risk is not reduced.

2.4 Document Ranking – A Practical Solution

Directly optimizing the objective function in Eq. (6) is computationally expensive. 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 the user. Therefore, 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 positions $k-1$ to k is:

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

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

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

Because w_k is a constant for all documents, we can safely drop it when using the increase to rank documents. This gives the following ranking criterion:

$$\begin{aligned} &\textit{select a document at rank } k \textit{ that has the maximum value of} \\ &E[r_k] - \alpha w_k {\sigma_k}^2 - 2\alpha \sum_{i=1}^{k-1} w_i \sigma_i \sigma_k \rho_{i,k} \end{aligned} \tag{10}$$

Eq. (10) provides a general principle of the document ranking. It contains three components. The first component concerns the relevance of a document, which is essentially equivalent to the PRP. The second component indicates that the relevance estimation should be subtracted by a weighted variance (when $\alpha > 0$); it is a generalization from the PRP, and has an ability to address the uncertainty of the point estimate of the relevance $E[r_k]$. A positive α produces *risk-aversion* ranking where an unreliably-estimated document (with big variance) should be given a large rank. The smaller the parameter α is, the less *risk-aversion* the ranking is. When $\alpha = 0$, it goes back to the PRP, which only considers the point estimate $E[r_k]$. In this case, the ranker intends to take more risk regardless of the uncertainty associated with the relevance estimation. The last component shows that the ranking prefers the documents that have small correlations (preferably negative correlations) with the already retrieved documents in the lower rank positions. Therefore, diversification, which is quantified by the weighted average of the correlations between the ranked documents, is effectively incorporated into the document ranking. In [2], a heuristic reranking criterion, the MMR (Maximal Marginal Relevance), is proposed by employing both the query-document and document-document similarities. The last component resembles the MMR, providing a theoretical justification.

3 Related Work and Discussion

In information retrieval, the most relevant work can be found in [3], where Chen and Karger argued that the classic probabilistic ranking principle (PRP) [13], which ranks documents in descending order of probability of relevance, is not always optimal for different user information needs. In some scenarios users would be satisfied with a limited number of relevant documents, rather than all relevant documents. The authors therefore proposed to maximize the probability of finding *a* relevant document among the top n. By considering the retrieved lower ranked documents as non-relevant ones, their algorithm introduced diversification into the probabilistic ranking. Their experiments on the specific metric that reflects above different user behaviors show that the approach designed for directly optimizing the metric outperforms the PRP. Another related work can be found in [2], where Carbonell and Goldstein proposed to re-rank retrieved documents, and use the Maximal Marginal Relevance (MMR) criterion to reduce redundancy. The criterion has been applied to the recommendation problem in [21]. In text retrieval, the MMR criterion has also been employed in a risk framework to address the subtopic retrieval problem [9,20]. But nonetheless,

when coming to the practical algorithms, these studies in [9,20] still resolve to take a point estimate, and use mode of the posterior without considering the uncertainty of the point estimate.

Our work can be regarded as research along this direction, but set out for more ambitious goals. We argue that ranking documents by examining their expected relevance is insufficient. A new point of view that focuses on evaluating the documents' relevance under conditions of risk is presented. By introducing variance as a measure of risk, diversification is naturally integrated into the probabilistic ranking. We demonstrate that it will play a central role in reducing the risk of document ranking. Our probabilistic ranking principle in Eq. (10) is independent of any retrieval models, and has the advantage of tuning the risk via a single parameter.

4 Empirical Study on Collaborative Filtering

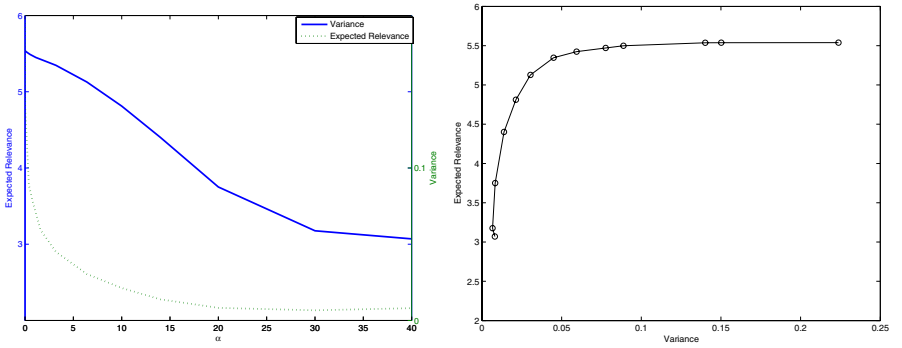
In this section, we present our empirical study on the recommendation problem, while leaving the evaluation on other retrieval applications such as text and multimedia retrieval, expert search, content-based filtering, and advertising ranking for future work. The reason to study recommendation first is due to the fact that the recommendation problem is generally formulated as rating prediction, while we believe a better view of the task is to regard it as a ranking problem [18]; our main goal is to validate our theoretical development, and investigate the impact of the parameter.

The task of recommendation is to suggest to a user information items that he or she might be interested in; collaborative filtering is one of the common techniques to generate a ranked list of relevant items. The covariance matrix of documents is calculated by users' ratings. We experimented with the EachMovie data set (`http://www.grouplens.org/taxonomy/term/14`), and adopted a subset described in [8], which contains 2,000 users. Each user has rated as least 40 items. The rating scale is indicated as a value between 1 and 6, with 1 being the lowest rating and 6 being the highest one. In our study, only the rating values 5 and 6 were regarded as relevant.

For testing, we assigned the users in the data set randomly to a training user and a test user set. Users in the training set were used as the basis for making predictions, while our test users were considered as the ground truth for measuring prediction accuracy. Each test users ratings have been split into a set of observed items and one of held-out items. The ratings of the observed items were input and represent user information needs (interests). Based on the user interests, the task is to generate a ranked item list. The held-out items (the test items) were used as the ground truth to evaluate the ranking accuracy.

4.1 Impact of the User's Risk Preference

Recall in Eq. (6) and Eq. (10), we have introduced parameter α to balance the expected overall relevance and its risk. This section empirically studies the



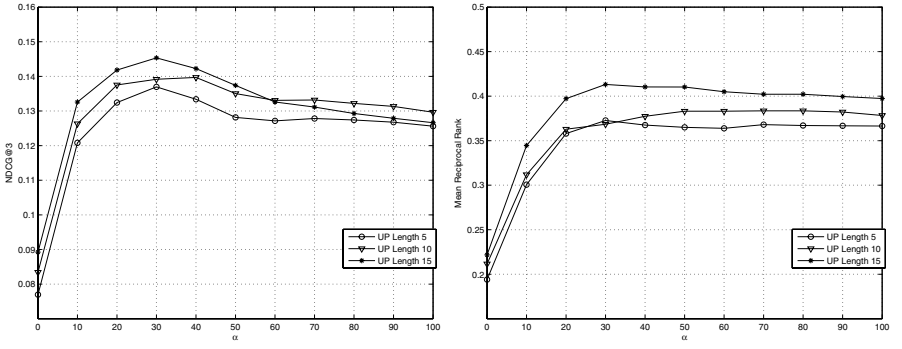
(a) (b)

Fig. 2. The behaviors of parameter α. (a) The impact of the parameter α on the relevance and variance. (b) Efficient Frontier.

behavior and impact of the parameter α. Fig. 2 (a) plots the expected overall relevance and variance against the different values of the parameter α, where the left y axis corresponds to the expected relevance of top-10 ranked documents and the right y axis shows the variance (uncertainty) of the list. The graph demonstrates that when we set a small α, the optimal rank list will have a relatively large variance (risk). As a reward of taking such risk, the expected relevance of the list stays high. But as the parameter α increases, the ranking becomes more risk-aversion. As a result, it tends to select a rank list whose variance is smaller, and subsequently the expected relevance of the list is reduced. We can thus conclude that the relevance ranking is risk-sensitive. This is similar to the observation in the financial market, i.e., any investment is a trade-off between return and risk.

We plot the efficient frontier in Fig. 2 (b) to further study optimal ranked lists under different risk preferences. The efficient frontier is calculated by applying our ranking principle in Eq. (10). The region above the frontier is unachievable by any rank order, while points below the frontier are suboptimal. The curves confirms our observation that high risk means high relevance return while low risk gives low relevance return.

To show the impact of the parameter α (and therefore the user's risk preference) on the retrieval performance, we plot the value of the parameter against two rank-based measures in Fig. 3. Since the low rank positions are crucial to a retrieval system, we report NDCG (Normalized Discounted Cumulative Gain) at 3 and the Mean Reciprocal Rank (MRR). (For the detailed definitions of the two measures, we refer to [7] and [17], respectively.) To study the influence of the user profile length, we vary the length as 5, 10, and 15 (denoted as UP Length 5, 10, and 15 in the figures). From the two graphs, we can see that the optimal ones are located around $\alpha = 30$, and significant performance gain has been achieved if we compare them to the PRP-based ranking (where $\alpha = 0$).



(a) (b)

Fig. 3. The impact of the parameter α. (a) NDCG at 3 and (b) Mean reciprocal rank.

4.2 Performance Evaluation

This section compares our ranking principle with the PRP-based ranking. For our approach, we set $\alpha = 30$. Notice that the setting of the parameter is not optimal for all the configurations. But we intend to investigate the performance of ranking method by using a universal value of α. Given this, it is expected that the performance can be improved even further when α is optimized with respect to individual collections or metrics.

Three popular recommendation algorithms were adopted to predict item ratings, namely the user-based algorithm [5], the item-based algorithm [15], and the method based on Probabilistic Latent Semantic Analysis (PLSA) [6]. The

Table 1. Comparison with the other approaches in the EachMovie data set. A Wilcoxon signed-rank test is conducted and all the improvements are significant.

Metrics	NDCG at 3			Precision at 3			Mean Reciprocal Rank		
Results	Basic	MV-Rank	Gain(%)	Basic	MV-Rank	Gain(%)	Basic	MV-Rank	Gain(%)
UPSize 5	0.077	0.137	77.916	0.090	0.180	100.692	0.194	0.373	92.063
UPSize 10	0.083	0.139	66.819	0.096	0.184	92.280	0.212	0.368	74.163
UPSize 20	0.089	0.145	62.864	0.100	0.199	99.659	0.222	0.413	86.319

(a) The item-based algorithm is used as the basis.

Metrics	NDCG at 3			Precision at 3			Mean Reciprocal Rank		
Results	Basic	MV-Rank	Gain(%)	Basic	MV-Rank	Gain(%)	Basic	MV-Rank	Gain(%)
UPSize 5	0.025	0.121	376.239	0.033	0.182	458.384	0.074	0.381	413.100
UPSize 10	0.034	0.133	291.192	0.034	0.202	494.827	0.092	0.416	351.368
UPSize 20	0.032	0.141	344.521	0.030	0.209	597.931	0.087	0.436	398.788

(b) The user-based algorithm is used as the basis.

Metrics	NDCG at 3			Precision at 3			Mean Reciprocal Rank		
Results	Basic	MV-Rank	Gain(%)	Basic	MV-Rank	Gain(%)	Basic	MV-Rank	Gain(%)
UPSize 5	0.045	0.106	136.489	0.059	0.158	168.476	0.139	0.326	134.868
UPSize 10	0.054	0.127	134.122	0.075	0.177	136.918	0.178	0.372	109.405
UPSize 20	0.065	0.125	91.434	0.091	0.186	104.466	0.202	0.399	97.709

(c) The PLSA algorithm is used as the basis.

results and comparisons are shown in Table 1. The performance gain was measured by the percentage of the improvement over the PRP ($\alpha = 0$). From the table, we can see that for the three basic algorithms, our ranking method outperforms the method using the PRP in all configurations. We also conducted a significance test (the Wilcoxon signed-rank test) applied to each configuration, indicating that the improvements are significant. In particular, the results are at least 3 times better when we use the user-based approach as the basic prediction method. This may be due to the fact that the user-based approach explores the correlations between users, while our mean-variance ranking addresses the correlations between items. A combination between them would generate much better performance gain than other combinations.

The relatively unsatisfied performance of the user-based, item-based, and PLSA approaches confirms the observation that rating-prediction based approaches are not ideal solutions for item ranking [11]. To address this, our ranking method, which analyzes the mean and variance of rank lists, is found to be effective in improving the ranking accuracy of recommendation consistently.

5 Conclusion and Future Work

In this paper, we have presented a new theory for document ranking by adopting the *mean-variance analysis*. We argued that an optimal document ranking strategy is to cast the ranking problem as a portfolio selection problem, where an optimal decision is to rank documents by maximizing their expected overall relevance given a risk (variance) level. It suggests that an optimal ranker should not only consider the expected relevance of the documents, but equally importantly understand the uncertainty associated with the estimation and the correlations between the documents. We have quantified the benefits of diversification and showed that it effectively reduces the risk of document ranking.

Our study is intended to increase the awareness of the mean-variance analysis for the relevance estimation and ranking. There are fruitful avenues for future investigations:

- We have adopted variance to measure the risk of document ranking. Variance cannot distinguish between "good" and "bad" dispersion. But in document ranking, the concept of risk is only associated with the latter. It will of great interest to investigate alternative measures for the risk. For instance, measures focusing on "downside risk" in finance might be beneficial.
- We have used historic rating data to calculate the covariance of items. But nonetheless, how to effectively and efficiently calculate the variance (risk) and correlation between the relevance predictions remains an open question. A large number of documents makes the estimation of correlations between documents a great challenge. A possible future direction would be to apply factor models [16] to reduce the computation complexity.
- Direct optimization of the objective function is expensive. It is worth investigating a global yet efficient optimization solution.
- It is expected that our ranking method will have many practical applications. We are currently studying its effectiveness in ad hoc text retrieval; we shall

also explore more opportunities in other information retrieval fields such as multimedia retrieval, content-based filtering, and advertising.
- The parameter a represents the risk preference of the user. It is highly beneficial to study its relationship with retrieval metrics and derive a learning method that can directly tune the parameter from the data.

References

1. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, Heidelberg (2006)
2. Carbonell, J., Goldstein, J.: The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: SIGIR 1998 (1998)
3. Chen, H., Karger, D.R.: Less is more: probabilistic models for retrieving fewer relevant documents. In: SIGIR 2006 (2006)
4. Gordon, M.D., Lenk, P.: A utility theoretic examination of the probability ranking principle in information retrieval. JASIS 42(10), 703–714 (1991)
5. Herlocker, J.L., Konstan, J.A., Borchers, A., Riedl, J.: An algorithmic framework for performing collaborative filtering. In: SIGIR 1999 (1999)
6. Hofmann, T.: Latent semantic models for collaborative filtering. ACM Trans. Info. Syst. 22(1), 89–115 (2004)
7. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. ACM Trans. Inf. Syst. 20(4), 422–446 (2002)
8. Jin, R., Si, L., Zhai, C.: A study of mixture models for collaborative filtering. Inf. Retr. 9(3), 357–382 (2006)
9. Lafferty, J., Zhai, C.: Document language models, query models, and risk minimization for information retrieval. In: SIGIR 2001 (2001)
10. Markowitz, H.: Portfolio selection. Journal of Finance (1952)
11. McLaughlin, M.R., Herlocker, J.L.: A collaborative filtering algorithm and evaluation metric that accurately model the user experience. In: SIGIR 2004 (2004)
12. Ponte, J.M., Croft, W.B.: A language modeling approach to information retrieval. In: SIGIR 1998 (1998)
13. Robertson, S.E.: The probability ranking principle in IR. Readings in information retrieval, 281–286 (1997)
14. Robertson, S.E., Belkin, N.: Ranking in principle. Journal of Documentation, 93–100 (1978)
15. Sarwar, B., Karypis, G., Konstan, J., Reidl, J.: Item-based collaborative filtering recommendation algorithms. In: WWW 2001 (2001)
16. Sharpe, W.F.: A simplified model for portfolio analysis. Management Science 9(2) (1963)
17. Voorhees, E.M.: The TREC-8 question answering track report. In: TREC-8, pp. 77–82 (1999)
18. Wang, J., Roberston, S.E., de Vries, A.P., Reinders, M.J.T.: Probabilistic relevance models for collaborative filtering. Journal of Information Retrieval (2008)
19. Zellner, A.: Bayesian estimation and prediction using asymmetric loss functions. Journal of the American Statistical Association 81(394), 446–451 (1986)
20. Zhai, C., Lafferty, J.D.: A risk minimization framework for information retrieval. Inf. Process. Manage. 42(1), 31–55 (2006)
21. Ziegler, C.-N., McNee, S.M., Konstan, J.A., Lausen, G.: Improving recommendation lists through topic diversification. In: WWW 2005 (2005)