Post-Modern Portfolio Theory for Information Retrieval

Ming-Feng Tsai\textsuperscript{a,b}, Chuan-Ju Wang\textsuperscript{c,*}

\textsuperscript{a}Department of Computer Science, National Chengchi University, Taipei 116, Taiwan
\textsuperscript{b}Program in Digital Content & Technologies, National Chengchi University, Taipei 116, Taiwan
\textsuperscript{c}Department of Computer Science, Taipei Municipal University of Education, Taipei 100, Taiwan

Abstract

Information Retrieval (IR) aims to discover relevant information according to a user’s information need. The classic Probability Ranking Principle (PRP) forms the theoretical basis for probabilistic IR models. This ranking principle, however, neglects the uncertainty introduced through the estimations from retrieval models. Inspired by the Post-Modern Portfolio Theory (PMPT), this paper proposes a mean-semivariance framework to handle the uncertainty. The proposed framework not only deals with the uncertainty but has the ability to distinguish bad surprises (downside uncertainty) and good surprises (upside uncertainty) when optimizing a ranking list. The experimental results show that the proposed method improves the IR performance over the PRP baseline in terms of most of IR evaluation metrics; moreover, the results suggest that the mean-semivariance framework can further boost the top-position ranking quality.

© 2012 Published by Elsevier B.V. Selection and/or peer-review under responsibility of Program Committee of INNS-WC 2012

Keywords: retrieval models, optimization, semivariance

1. Introduction

Information Retrieval (IR) aims to discover relevant information according to a user’s information need. In general, the process of retrieving information consists of two phases. In the first phase, probabilistic retrieval models [1] compute the relevance between a given user’s information need (query) and each of the documents in a collection. The second phase focuses on how to rank the calculated documents; the classic Probability Ranking Principle (PRP) [2] forms the theoretical basis of this phase, which ranks the documents with the order of decreasing probabilities of relevance to the query. However, the ranking principle neglects the uncertainty associated with the relevance of the documents to the query; the uncertainty may result from various sources, such as specific user preferences and ambiguity within a query. Take the query “jaguar” as an example. This query may refer to the Jaguar Cars company, the Apple Jaguar operation system, the Fender Jaguar electric guitar, or the felines. For such a query with some uncertainty, an IR system should provide a ranking list of documents with all possible interpretations, which may better meet as many information needs as possible.

To deal with the uncertainty, we draw an analogy between the ranking problem in IR and the investing problem in finance; that is, selecting a set of stocks (portfolio) resembles selecting a set of documents (ranking list). In
1952, Harry Markowitz in his Nobel Prize winning work, proposed a theory, Modern Portfolio Theory (MPT), which attempts to maximize portfolio expected return for a given amount of portfolio risk by carefully choosing the proportions of various assets [3]. Wang and Zhu first incorporated MPT into the process of IR and formulated the ranking problem as a portfolio selection problem [4]. In their framework, two summary statistics, mean and variance, are used to characterize a ranking list; the mean represents a best “guess” of the overall relevance of the list, while the variance sketches the uncertainty associated with the guess. For a risk-averse solution, the relevance of a ranking list is maximized, and in the meantime, the variance of the relevance is minimized.

However, the summary statistic “variance” cannot distinguish a bad surprise (relevance scores less than expectations) from a good surprise (relevance scores that exceed expectations). In the framework of [4], a good surprise will also be a penalty for the risk-averse approach when optimizing a ranking list. Motivated by the concept of downside risk in Post-Modern Portfolio Theory (PMPT) [5, 6, 7], this paper proposes a mean-semivariance framework that only takes bad surprises into account for the risk-averse approach, and considers only good surprises for the risk-loving approach.

Two Text REtrieval Conference (TREC) tracks, TREC08 and Robust04, are used for evaluating the proposed mean-semivariance framework. The experimental results show that the proposed method improves the performance over the baseline, PRP, and the MPT [4] on both tracks in terms of most of evaluation metrics. Moreover, the results suggest that the mean-semivariance framework can further boost the top-position ranking quality.

The remainder of the paper is organized as follows. Section 2 presents the theoretical developments of our mean-semivariance framework. Some experimental results are investigated in Section 3. Section 4 concludes.

2. The Mean-Semivariance Framework

2.1. Overall Relevance Scores

Given a query, suppose an IR system returns a ranking list composed of \( n \) documents from rank 1 to \( n \) with corresponding estimated relevance scores from \( r_1 \) to \( r_n \). By following [4], we define the effectiveness of a ranking list via the weighted average of the relevance scores in the list as

\[
R_n = \sum_{i=1}^{n} w_i r_i.
\]

Above, \( R_n \) denotes the overall relevance of a ranking list, \( w_i \) denotes the weight of the \( i \)-th ranked position, \( \sum_{i=1}^{n} w_i = 1 \), and, in general, \( w_1 > w_2 \cdots > w_n \) [8]. In this case, obviously, \( R_n \) can be maximized with \( r_1 > r_2 \cdots > r_n \) (i.e., ordering the documents according to their estimated relevance scores).

2.2. Uncertainty of Relevance Scores and Risk Measures

The uncertainty can be introduced through the estimations from retrieval models. To cope with such uncertainty, the relevance scores \( r_i \) are assumed to be random variables. The distribution of the relevance scores can be varied; for example, as in [9], the relevance scores are assumed to follow a Gaussian distribution, and as in [10], the uncertainty of the relevance scores is introduced from the underlying probabilistic language models (conjugate prior of the Multinomial distribution).

The uncertainty of the overall relevance is characterized with its variance \( \text{Var}(R_n) \) in [4]:

\[
\text{Var}(R_n) = \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j c_{i,j},
\] (1)

where \( c_{i,j} \) denotes the covariance of the relevance scores between the \( i \)-th ranked document and the \( j \)-th ranked one. However, this variance cannot distinguish a bad surprise from a good surprise. Motivated by the concept of downside risk in PMPT, we only take downside variance into account for the risk-averse approach and by contrast, consider only upside variance for the risk-loving approach. In other words, we use semivariance as the indicator of risk (uncertainty), which is mathematically defined as

\[
\text{Var}_-(R_n) = E \left[ \left( \text{Min}(R_n - E[R_n], 0) \right)^2 \right],
\]

\[
\text{Var}_+(R_n) = E \left[ \left( \text{Max}(R_n - E[R_n], 0) \right)^2 \right].
\]
where \( \text{Var}_-(R_n) \) (\( \text{Var}_+(R_n) \)) denotes the downside (upside, respectively) variance of the overall relevance. Unlike the variance defined in Eq. (1), which can be calculated exogenously, the semivariance is endogenous. As a result, we use the approximation proposed by [11] to calculate \( \text{Var}_Q(R_n) \):

\[
\text{Var}_Q(R_n) \approx \sum_{i=1}^{n} \sum_{i=1}^{n} w_i w_j \hat{c}_{i,j},
\]

where

\[
\hat{c}_{i,j} = \begin{cases} 
E\left[\min(r_i - E[r_i], 0) \times \min(r_j - E[r_j], 0)\right], & \text{if } Q = - , \\
E\left[\max(r_i - E[r_i], 0) \times \max(r_j - E[r_j], 0)\right], & \text{if } Q = + . 
\end{cases}
\]

2.3. Optimization for the Ranking List

To optimize the effectiveness of a ranking list with the summary statistics, mean and semivariance, the objective function for the optimization is

\[
\text{max } A_p = E[R_0] - a \times \text{Var}_Q(R_n),
\]

where \( a \) denotes the risk preference parameter and \( Q \equiv \text{sgn}[a] \). Note that for a risk-averse solution, \( a < 0 \) and for a risk-loving solution, \( a > 0 \); additionally, with \( a = 0 \), documents are ranked by the PRP.

Since the weight \( w_i \) for each document is a discrete variable, it is hard to directly optimize the objective function in Eq. (3). Therefore, the greedy algorithm in [4] is adopted to optimize the objective function. The difference of the objective function from the position \( k - 1 \) to \( k \) is

\[
A_k - A_{k-1} = E[R_k] - a \times \text{Var}_Q(R_k) - E[R_{k-1}] + a \times \text{Var}_Q(R_{k-1})
= \sum_{i=1}^{k} w_i E[r_i] - a \sum_{i=1}^{k} \sum_{i=1}^{k} w_i w_j \hat{c}_{i,j} - \sum_{i=1}^{k-1} w_i E[r_i] + a \sum_{i=1}^{k-1} \sum_{i=1}^{k-1} w_i w_j \hat{c}_{i,j}
= w_k (E[r_k] - aw_k \hat{c}_{kk} - 2a \sum_{i=1}^{k-1} w_i \hat{c}_{ik}).
\]

The \( i \)-th ranked document for \( i \in \{2, 3, \ldots, n\} \) is selected to maximize the difference \( A_k - A_{k-1} \) in Eq. (4). (Note that the first document \( (i = 1) \) is set to the document with the highest relevance score.)

2.4. Calculation of Means and Semicovariance Matrix

Different retrieval models generate different estimators of \( E[R_n] \) and \( \text{Var}_Q(R_n) \). In the following experiments, the probabilistic language models with Dirichlet and Jelinek-Mercer (JM) smoothing are adopted [12]. In a multinomial language model with parameter \( y = (y_1, \ldots, y_i, \ldots, y_{|y|}) \), given a document \( d \equiv (d_1, \ldots, d_i, \ldots, d_{|y|}) \), the posterior probability can denoted as \( p(y \mid d, \alpha) \), and

\[
p(y \mid d, \alpha) \propto p(d \mid y) p(y \mid \alpha) = \prod_{i} (y_i)^{d_i} \prod_{i} (y_i)^{\alpha_i - 1}
= \prod_{i} (y_i)^{d_i + \alpha_i - 1}
\sim \text{Dir}(d + \alpha),
\]

where \( p(w \mid \alpha) \) is the Dirichlet prior on \( y \) with parameter \( \alpha = (\alpha_1, \ldots, \alpha_i, \ldots, \alpha_{|y|}) \).

The mean of \( y_i \) is chosen to be the estimator of \( y_i \); i.e.,

\[
\hat{y}_i = \frac{d_i + \alpha_i}{\sum_{i=1}^{n} d_i + \alpha_i}.
\]
According to Eq. (6), given a query $q \equiv (q_1, \cdots, q_i, \cdots, q_{|V|})$, $E[r_i]$ can be estimated by

$$
\hat{r}_i = \prod_{i=1}^{|V|} \hat{y}_i^{q_i}.
$$

On the other hand, we draw $y$ as samples from the distribution in Eq. (5) to obtain randomized relevance scores; then the semicovariance defined in Eq. (2) can be easily calculated.

3. Experiments

This section first describes the experimental datasets and evaluation metrics. With respect to different datasets and smoothing techniques, there are four sets of experiments conducted in this paper. We then present the experimental results, and conclude this section by providing some discussions and analyses.

In our experiments, two TREC tracks are used for evaluating the proposed method, including TREC08 and Robust04. Table 1 lists the details of the used datasets for the two tracks. We report experimental results on the datasets for ad-hoc retrieval. Therefore, the following metrics are calculated for evaluating the effectiveness of the proposed approach: Mean Average Precision (MAP), Mean Reciprocal Rank (MRR), Precision, and Normalized Discounted Cumulative Gain (NDCG); in addition to the overall performance, we also examine the performance at top-5 and top-100 positions, in order to study the effect of risk-averse and risk-loving approaches on a retrieved-document list at different positions.

Fig. 1 and Fig. 2 illustrate the experimental results, in which we plot the corresponding improvements in terms of different metrics. In the following discussions, PMPT denotes the proposed method; the two compared methods are the PRP and MPT methods [4]. As shown in Fig. 1, when operating on the TREC08 dataset, the proposed PMPT method improves most of evaluation metrics over the baseline, PRP, and MPT; this improvement shows that

| Name            | Description           | # Docs | Topics | # Topics |
|-----------------|-----------------------|--------|--------|----------|
| TREC8 ad hoc task | TREC disks 4, 5 minus CR | 528,155 | 401-450 | 50       |
| Robust2004 hard topics | TREC disks 4, 5 minus CR | 528,155 | Difficult Robust2004 topics | 50 |

Fig. 1. Comparison of our approach (PMPT) against the MPT and the PRP on TREC2008 ad hoc task.
the PMPT method can better rank retrieved documents via the mean-semivariance framework. For the performance at top positions, the PMPT method, especially, can gain about 2.5%, 1.5%, and 1.0% improvements in terms of MAP5, P@5, and NDCG@5, respectively; this leap demonstrates that the PMPT method can boost the top-position ranking quality even further, no matter with either the Dirichlet or JM smoothing techniques. Similarly, as shown in Fig. 2, the proposed PMPT method improves over the PRP baseline in terms of most of IR evaluation metrics on the Robust04 dataset. However, since the topics we used for the Robust04 track are hard topics only, in terms of some metrics, the PMPT method can only get minor improvements, or even cannot outperform over the MPT method. This phenomenon shows that, when operating on hard topics, these two approaches based on Portfolio Theory might obtain similar performance. As shown in Fig. 2(b), compared with the PRP baseline, the PMPT method with the JM smoothing still gains about 7%, 5%, and 3% improvements in terms of MAP5, P@5, and NDCG@5, respectively. Furthermore, with respect to the different smoothing techniques, we observe that the JM smoothing generally performs better than the Dirichlet smoothing does; this phenomenon may be due to the big variances caused by the JM smoothing in our experiments. This observation is also consistent with that in [10], which suggests that it might be more preferable to apply the “risk-sensitive” approach with the JM smoothing technique.

4. Conclusions and Future Work

This paper proposes a general mean-semivariance framework to study document ranking under uncertainty. In the framework, the downside uncertainty can be distinguished with the upside uncertainty when optimizing a ranking list. Experiments on two TREC datasets with the different smoothing techniques validate that the proposed framework improves the ranking quality over the PRP baseline and the MPT approach. In particular, the proposed framework obtains about 1%-7% improvements over the PRP baseline in terms of MAP5, P@5, and NDCG@5. Future directions include how to use learning techniques to find out the optimal parameters of the proposed framework, and how to adapt the framework to diversified information retrieval.

Acknowledgments

Research of this paper was partially supported by National Science Council, Taiwan, under the contract NSC100-2218-E-004-001.
References

[1] S. Robertson, K. Jones, Relevance weighting of search terms, Journal of the American Society for Information science 27 (3) (1976) 129–146.

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

[3] H. Markowitz, Portfolio selection, The Journal of Finance 7 (1) (1952) 77–91.

[4] J. Wang, J. Zhu, Portfolio theory of information retrieval, in: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, ACM, 2009, pp. 115–122.

[5] A. Damodaran, Investment philosophies: successful strategies and the investors who made them work, Vol. 185, John Wiley & Sons Inc, 2003.

[6] B. Rom, K. Ferguson, Post-modern portfolio theory comes of age, The Journal of Investing 3 (3) (1994) 11–17.

[7] F. Sortino, S. Satchell, Managing downside risk in financial markets: theory, practice and implementation, Vol. 1, Butterworth-Heinemann, 2001.

[8] K. Järvelin, J. Kekäläinen, Cumulated gain-based evaluation of ir techniques, ACM Transactions on Information Systems (TOIS) 20 (4) (2002) 422–446.

[9] J. Zhu, J. Wang, M. Taylor, I. Cox, Risk-aware information retrieval, Advances in Information Retrieval (2009) 17–28.

[10] J. Zhu, J. Wang, I. Cox, M. Taylor, Risky business: modeling and exploiting uncertainty in information retrieval, in: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, ACM, 2009, pp. 99–106.

[11] J. Estrada, Mean-semivariance optimization: a heuristic approach, Journal of Applied Finance 18 (1) (2007) 57–72.

[12] C. Zhai, J. Lafferty, A study of smoothing methods for language models applied to ad hoc information retrieval, in: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, 2001, pp. 334–342.