Post-Modern Portfolio Theory for Information Retrieval

Ming-Feng Tsai
Department of Computer Science
National Chengchi University

Chuan-Ju Wang
Department of Computer Science
Taipei Municipal University of Education

INNS-WC 2012, October 3, 2012

Outline

Introduction

- The process of retrieving information consists of two phases:
  1. Compute the relevance between a given user's information need and each of the documents in a collection.
  2. Rank the documents according to the computed relevance scores.
- The classic Probability Ranking Principle (PRP) forms the theoretical basis of the 2nd phase.
  - Rank the documents with the order of decreasing probabilities of relevance to the query.

Uncertainty

- However, the PRP neglects the uncertainty associated with the relevance of the documents to the query.
- Examples of sources of uncertainty:
  - Specific user preferences.
  - Ambiguity within a query.
- Take the query "jaguar" as an example.
  - The Jaguar Cars company.
  - The Apple Jaguar operation system.
  - The Fender Jaguar electric guitar.
- An ideal Information Retrieval (IR) system should provide a ranking list of documents with all possible interpretations.

Modern Portfolio Theory

- In 1952, Harry Markowitz in his Nobel Prize work, proposed the Modern Portfolio Theory (MPT).
  - Attempt to select a set of stocks (portfolio) that maximize its total return for a given amount of risk.
- An analogy between the ranking problem in IR and the investing problem in finance.
  - Selecting a set of stocks (portfolio) resembles selecting a set of documents (ranking list).
  - The risk resembles the uncertainty.

Modern Portfolio Theory

- Wang and Zhu (2009)[1] first introduced MPT into the process of IR and formulated the ranking problem as a portfolio selection problem.
- Two statistics, mean and variance, are used to characterize a ranking list.
  - Mean: A best "guess" of the overall relevance of the list
  - Variance: The uncertainty associated with the guess
- For a risk-averse user, the relevance of a ranking list is maximized, and in the meantime, the variance of the relevance is minimized.

[1] J. Wang, J. Zhu, Portfolio theory of information retrieval, *Proceedings of the 32nd international ACM SIGIR*, (2009), 115-122.

Our Approach

- However, the "variance" cannot distinguish a bad surprise (relevance score less than expectation) from a good surprise (relevance score more than expectation).
- Motivated by the concept of Post-Modern Portfolio Theory (PMPT), this paper proposes a mean-semivariance framework:
  - Only take bad surprises into account for risk-averse users.
  - Only consider good surprises for the risk-loving users.

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.
- The effectiveness of a ranking list is defined as

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

  - In general, $w_1 > w_2 \cdots > w_n$
  - Then, R_n can be maximized with $r_1 > r_2 \cdots > r_n$.

Uncertainty of Relevance Scores

- The relevance scores r_i are assumed to be random variables.
- The uncertainty of the overall relevance is characterized with its variance $Var(R_n)$:

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

  - $c_{i,j}$ denotes the covariance of the relevance scores between the i-th ranked document and the j-th ranked one.

Semivariance

- As mentioned, however, this variance cannot distinguish a bad surprise from a good surprise.
- We use semivariance as the indicator of uncertainty, which can be defined as follows:

$$
\begin{aligned}
Var_{-}(R_n) &= E\left[(Min(R_n - E[R_n], 0))^2\right], \\
Var_{+}(R_n) &= E\left[(Max(R_n - E[R_n], 0))^2\right],
\end{aligned}
$$

  - $Var_{-}(R_n)$: the downside variance of the overall relevance scores.
  - $Var_{+}(R_n)$: the upside variance of the overall relevance scores.
- We use an approximation method to calculate these two indicators.[2]

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

Optimization for the Ranking List

- To optimize the effectiveness of a ranking list, we define the objective function as

$$max\ E[R_n] + a \times Var_Q(R_n),$$

  - where a denotes the risk preference parameter and $Q \equiv \text{sgn}[a]$.
  - Risk-averse: $a < 0$.
  - Risk-loving: $a > 0$.
  - When $a = 0$, documents are ranked by the PRP.
- A greedy algorithm is adopted to optimize the objective function.

Settings

- Two NIST Text REtrieval Conference (TREC) tracks are used for evaluating the proposed method, including TREC08 and Robust04.

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

Table : **Overview of the two TREC test collections.**

- Evaluation metrics: Precision, Mean Average Precision (MAP), Mean Reciprocal Rank (MRR), and Normalized Discounted Cumulative Gain (NDCG).

Results

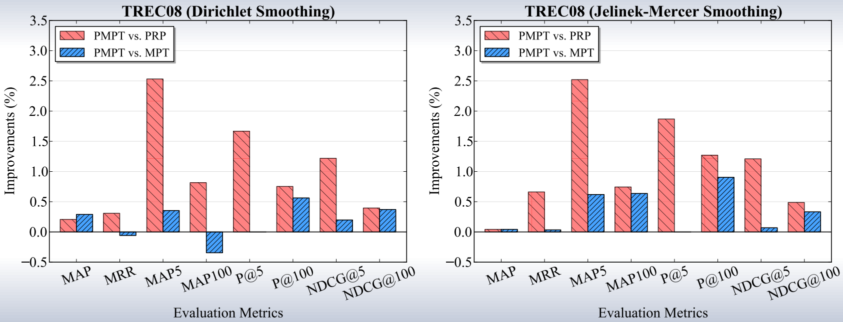

Figure : **Comparison of our approach (PMPT) against the MPT and the PRP on TREC2008 ad hoc task.**

Results

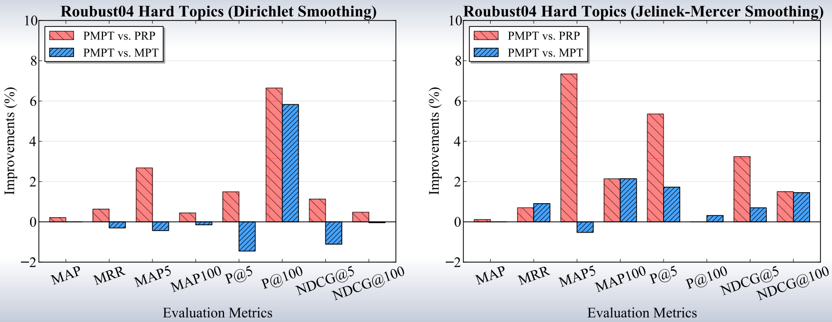

Figure : **Comparison of our approach (PMPT) against the MPT and the PRP on Robust2004 hard topics.**

Conclusions and Future Work

- This paper proposes a mean-semivariance framework to study document ranking under uncertainty.
- The downside uncertainty can be distinguished with the upside uncertainty when optimizing a ranking list.
- The experimental results validate that the proposed framework improves the ranking quality over the PRP baseline and the MPT approach.
  - The proposed framework obtains about 1%-7% improvements over the PRP baseline in terms of MAP5, P@5, and NDCG@5.
- Future directions:
  1. How to use learning techniques to find out the optimal parameters of the proposed framework.
  2. How to adapt the framework to diversified information retrieval.