行政院國家科學委員會專題研究計畫 成果報告

改善多樣性檢索之搜索結果
研究成果報告(精簡版)

計畫類別：個別型
計畫編號：NSC 100-2218-E-004-001-
執行期間：100年08月01日至101年07月31日
執行單位：國立政治大學資訊科學系

計畫主持人：蔡銘峰

計畫參與人員：碩士班研究生-兼任助理人員：陳志明
碩士班研究生-兼任助理人員：林專耀
碩士班研究生-兼任助理人員：程政康
博士班研究生-兼任助理人員：陳函斌

報告附件：國際合作計畫研究心得報告

公開資訊：本計畫可公開查詢

中華民國 101年10月31日

中文摘要：

此計劃主旨在於：如何利用機器學習的方式習得一個具有多樣性檢索結果之排列模型。此多樣化檢索之特性對於一些常見易混淆的問句，可以提供給使用者涵蓋層面較多之檢索結果，並可進一步讓使用者找尋其想要之資訊。因此，此研究計劃發展出一個可以考量到多樣化檢索結果之特性的學習式排列演算法，此演算法結合了支持向量機分類（Support Vector Classification，SVC）和支持向量機迴歸（Support Vector Regression，SVR）二種作法，因此我們稱此方法為二步式排序支持向量機（Two-step Ranking SVM）方法。此提出的方法，會針對不同的文檔對（Document Pairs）進行不同的排序學習程序。除了此演算法外，我們也對檢索模型進行研究，並且提出了一個利用財務金融領域中的後現代投資組合理論（Post-modern Portfolio Theory）衍生出來的檢索模型，此模型也具備有多樣化檢索結果的能力。在之後工作中，我們希望可以進一步加入更多值得擁有的特性，如：考量到地域性以及時間性的排列模型。除了學習式演算法的推導之外，本人也希望藉由這個研究計劃，深入探討排列問題其學習理論之基礎。而開發出來的學習排列演算法，也希望可以在往後應用到一套真實線上檢索系統中，透過如此的結合將可以利用更多使用者回饋的資料來改善既有的檢索系統。

中文關鍵詞：

資訊檢索、機器學習、多樣化檢索、支持向量機、投資組合理論

英文摘要：

In this study we focus on how to diversify ranked results making explicit use of information about the topics that the query or the documents may refer to. For this ranking problem, we assume that each document belongs to one of the categories in a taxonomy； then, we attempt to obtain as many categories within the retrieved results as possible while maintaining ranking quality. We propose a Two-step Ranking SVM approach to diversifying ranking results. In the first step of the proposed method, we conduct the Support Vector Classification (SVC) training on the pairs of documents with Different Ranks and Same Category (DRSC). Because of the margin-maximization specialty of SVM, the learned model will tend to separate the documents in the same category as far as possible, which indirectly introduces the concept of diversity into the retrieved results. Following the first step, we then continue to conduct the Support Vector Regression (SVR) training on the pairs of documents with Same Rank and Different Category (SRDC). That is, starting with the output values of the first classification model, the second step performs SVR on the SRDC pairs； therefore, this practice will result in pulling the pairs of documents in the different categories as close as possible, which directly put the concept of diversity into the learned model. In addition, we also propose a novel retrieval model based on Post-modern Portfolio Theory. The proposed model also has the ability of diversifying retrieval results, and we attempt to combine the model with the learning-to-ranking techniques for the diversity task.

英文關鍵詞： Information Machine Learning, Diversified Retrieval, Support Vector Machine, Portfolio Theory

1 The Background and Goals of the Research Project

前言、研究目的、文獻探討

1.1 Background

How to learn an effective ranking function for document retrieval has drawn much attention from the information retrieval and machine learning communities. For a given query, an effective ranking function should provide a ranking list with a comprehensive coverage of information, including any ambiguity within the query. Take "apple" as an example, this query may refer to the Apple company, the Apple emulator, and the fruits. For such ambiguous queries, if there is no further information about user's intention, an information retrieval system should better provide a ranking list of documents with all possible interpretations.

However, most of the current work related to learning to rank is mainly aimed to learn a ranking function that only consider relevance but without the consideration of diversity. Using the current learning-to-rank techniques cannot obtain a ranked list that reflects the breadth of available information and any ambiguity inherent in a query. In [12], the authors mentioned that the ideal document ordering for such ambiguous queries would properly account for the interests of the overall user population. The earlier documents might cover key concepts of each topic. Later documents would supplement the basic information, rather than redundantly repeating the same thing over and over again. Obviously, methods based on the conventional relevance-based ranking are unlikely to be optimal solutions to such a diversified ranking problem.

1.2 Research Purpose

For this work, therefore, we begin with the acquirement of category information for each document of the OHSUMED dataset in the LETOR benchmark [10], by using the Medical Subject Heading (MeSH) hierarchy. Then, we apply the two-step Ranking SVM algorithm to generate the ranker that can obtain as many categories as possible within the retrieved results while maintain ranking quality. Notice that diversity is usually opposed to relevance, as mentioned in [17]. To assess performance comprehensively, we adopt several evaluation metrics including NDCG, Category Recall, α-DCG [3], and MAP. All of experiment results are compared to the conventional Ranking SVM [9]. According to the results, the proposed approach not only improves the ranking quality, but also increases diversity within the retrieved results.

1.3 Related Work

In the past, there are several studies related to the topic of diversifying retrieved results, including evaluation metrics and systematic studies. One of the influential work on diversification is that of Maximal Marginal Relevance (MMR) proposed by Carbonell and Goldstein in [1]. In their work, the MMR evaluation metric is proposed to examine the tradeoff between diversity and the relevance within retrieved results. They use two similarity functions: one measures the similarity among documents, and the other one measures the similarity between document and query. Then, the diversification is carried out by using a parameter to control these two similarity functions.

Zhai *et al.* [17] claim that the dependence among the retrieved results is also important. This work is later formalized in [18], where Zhai and Lafferty propose a risk minimization framework for information retrieval that allows a user to define a loss function over the retrieved results. This framework permits users to specify their "unhappiness" about a given set of results. In [16], Zhai discusses the problem of dependent topic retrieval, and proposes two loss functions depending on certain language models, which can lead to a selection of diverse results.

In [2], Chen and Karger consider information retrieval in the context of ambiguous queries. The basic idea of this work is that the results should be ranked sequentially according to the probability of a document being relevant conditioned on the other documents in its front. They propose an objective function capable of considering the diversification of retrieved results, and the function can also focus on retrieving at least one relevant document for all users. In [12], Radlinski *et al.* propose a learning algorithm to compute a set of retrieved results from a diverse set of orderings. By iterating through all documents in each of the positions while holding fixed the documents in the other positions, they attempt to learn a best ranking of documents using user click-through data. By means of the ability that the user click-through data can diminish the value of similar documents, their approach can naturally produces a diverse set of results. In [3], Clarke *et al.* study diversification in the context of answering questions. They focus on developing a framework of evaluation that takes into account both novelty and diversity. In this work, questions and answers are treated as sets of "information nuggets," and relevance is a function of the nuggets contained in the questions and the answers. In addition, they also propose a novel metrics, α-DCG, capable of taking into account both diversity and relevance, thereby can better reflect the ranking quality and diversity within the retrieved results. We also use this metric for the assessment of our proposed approach.

2 Methodology

研究方法

In Ranking SVM [9], the ranking problem for document retrieval can be represented as the following constrained optimization problem:

$$\begin{aligned} &\min \quad \frac{1}{2}\vec{\omega}\cdot\vec{\omega} + C\sum \xi_{i,j,k} \\ &\text{s.t.} \begin{cases} \forall (d_i, d_j) \in r_k^* : \vec{\omega}\Phi(q_k, d_i) \geq \vec{\omega}\Phi(q_k, d_j) + 1 - \xi_{i,j,k} \\ \forall i \forall j \forall k : \xi_{i,j,k} \geq 0 \end{cases}, \end{aligned} \tag{1}$$

where $\Phi(q, d)$ is a feature vector obtained from document d with respect to query q. If we rearrange the constraints as:

$$\vec{\omega}(\Phi(q_k, d_i) - \Phi(q_k, d_j)) \geq 1 - \xi_{i,j,k}, \tag{2}$$

the above optimization problem can be casted as the classification problem of SVM on all document pairs (d_i, d_j) with different ranks r^*.

Unlike the conventional Ranking SVM, the proposed two-step Ranking SVM consists of two training phases: In the first step, with the taxonomy information, Support Vector Classification (SVC) is carried on the document pairs with Different Ranks but the Same Category (DRSC) to produce a ranking model. In the second step, to diversify ranking results, the ranking model continues to be trained by Support Vector Regression (SVR) on the document pairs with the Same Rank but Different Categories (SRDC). With the aid of taxonomy information, the proposed approach focuses on different types of pairs in different steps for the goal of diversification. Below we describe the intuition behind our method in details.

Instead of conducting SVC training on all different-rank pairs, the first step performs SVC training on DRSC pairs only. This practice is due to the following benefits: First, this way indirectly includes diversity by separating documents in the same category. Because of the margin-maximization specialty of SVM, conducting SVC on DRSC pairs will permit the learned model to not only separate the documents with different ranks, but also keep the documents in the same category as away as possible. Second, this way is consistent with the nature of comparison that under the same category, objects are more reasonable for comparison. Third, this way leads to more efficient training time because the number of DRSC pairs is fewer than that of all different-rank pairs.

Once a ranking model has been obtained in the first step, the second step continues to conduct SVR training on the ranking model by using the SRDC pairs. Specifically, starting with the output values of the first-step model, the second step performs ϵ-SVR on the SRDC pairs. This step is mainly to bring the concept of diversity directly into the learned model. That is because the SVR training will enable the learned model to keep the documents having different categories as close as possible, and not to deteriorate the ranking quality too much by bringing the documents with the same rank together. Hence, the second step can be considered as a retraining approach to the diversification purpose. The two-step Ranking SVM approach can be summarized in Algorithm 1.

Algorithm 1 Two-step Ranking SVM

```
Given: a set $S = \{(d_1, r_1, c_1), (d_2, r_2, c_2), \dots (d_n, r_n, c_n)\}$, where $d_i$ is a document, $r_i$ is the rank of the document, and $c_i$ is the category information of the document.

**Pair Construction**
**if** $r_i \neq r_j$ & $c_i = c_j$ **then**
    construct DRSC pair via $\Phi(d_i, d_j)$
**end if**
**if** $r_i = r_j$ & $c_i \neq c_j$ **then**
    construct SRDC pair via $\Phi(d_i, d_j)$
**end if**

**First Step**: conduct SVC on the DRSC pairs
    $$\min \quad \frac{1}{2}\vec{\omega} \cdot \vec{\omega} + C \sum \xi_{i,j,k}$$
    $$\text{s.t.} \begin{cases} \forall (d_i, d_j) \in r_k^* : \langle \vec{\omega}, \Phi(d_i, d_j) \rangle \geq 1 - \xi_{i,j,k} \\ \forall i \forall j \forall k : \xi_{i,j,k} \geq 0 \end{cases}$$

**Second Step**: conduct SVR on the SRDC pairs
    $$\min \quad \frac{1}{2}\vec{\omega} \cdot \vec{\omega} + C \sum \left(\xi_{i,j,k}^{+} + \xi_{i,j,k}^{-}\right)$$
    $$\text{s.t.} \begin{cases} \forall (d_i, d_j) \in r_k^* : y_{i,j} - \langle \vec{\omega}, \Phi(d_i, d_j) \rangle \leq \epsilon + \xi_{i,j,k}^{+} \\ \forall (d_i, d_j) \in r_k^* : \langle \vec{\omega}, \Phi(d_i, d_j) \rangle - y_{i,j} \leq \epsilon + \xi_{i,j,k}^{-} \\ \forall i \forall j \forall k : \xi_{i,j,k}^{+} \geq 0 \\ \forall i \forall j \forall k : \xi_{i,j,k}^{-} \geq 0 \end{cases}$$
```

3 Results and Discussions
結果與討論

3.1 Experimental Setup

In this study, the OHSUMED dataset in LETOR is used as our experimental dataset. For the taxonomy information, we employ the 2008 MeSH tree structure[1], in which there are 16 main categories in the root, such as anatomy and organisms. We classify each OHSUMED document into the 16 categories by using majority approach. By means of the majority method, each document in the OHSUMED collection thus belongs to one of 16 categories in the MeSH tree. For other details of the OHSUMED collection, please refer to [10]. For implementation, except Ranking SVM by svmlight [8], all experiments are carried out by liblinear [6].

[1] http://www.nlm.nih.gov/mesh/trees2008.html

Table 1: Performances of NDCG@15, CR@15, α-DCG@15, and MAP: All numbers are average values over 5 folds. Numbers in brackets indicate the p-value from a paired one-tailed t-test. Bold faced numbers indicate that the entry is statistically significant from the run of Ranking SVM at 95% confidence level.

Methods	**NDCG@15**	**CR@15**	**α-DCG@15**	**MAP**
Ranking SVM	0.423	0.535	4.294	0.433
SVC only	**0.446 (0.01)**	0.543 (0.21)	**4.614 (0.01)**	**0.450 (2e-4)**
SVC+SVR	**0.440 (0.05)**	**0.554 (0.04)**	**4.562 (0.04)**	**0.445 (0.01)**

3.2 Experimental Results

In addition to NDCG and MAP, we also use Category Recall (CR) and α-DCG to evaluate the diversity within the retrieved results. Among these four metrics, NDCG and MAP are mainly for the evaluation of ranking quality, CR is for diversity, and α-DCG is an in-between version. Since diversity cannot just be presented in a few top ranks, we focus on the performance at position 15. Notice that the value of CR is $\frac{5}{15}$ if there are 5 different categories within top 15 results. For α-DCG, the parameter of α is set to 0.5; for more details about α-DCG, please refer to [3].

Table 1 lists the performance of three referenced approaches including Ranking SVM, SVC only, and SVC+SVR. Note that the parameters of the three methods are all tuned on the validation datasets provided in LETOR. The method of SVC only, as indicated in the table, indeed improves diversity in terms of CR@15 because of the indirect injection of diversity by separating DRSC pairs; however, the improvement is not significant. By further retraining the ranker, the method of SVC+SVR eventually enhances diversity significantly while maintaining the ranking quality in terms of all criteria. These results are consistent with the discussions in Section 2.

3.3 Summary

This work proposes a two-step Ranking SVM for diversifying ranking results. With the aid of taxonomy information, the proposed method focuses on different types of pairs in different steps for the goal of diversification. According to our experimental results, the two-step method of SVC+SVR not only improves ranking quality, but also diversifies the retrieved results with varied categories. In the following section, we attempt to use another retrieval model based on portfolio theory to deal with the problem.

4 Extension Work: Post-Modern Portfolio Theory for Information Retrieval
延伸工作：後現代投資組合理論之檢索模型

4.1 Introduction

In general, the process of retrieving information consists of two phases. In the first phase, probabilistic retrieval models [13] 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) [4] 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 [11]. Wang and Zhu first incorporated MPT into the process of IR and formulated the ranking problem as a portfolio selection problem [14]. 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.

4.2 The Mean-Semivariance Framework

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 [14], 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, $w_1 > w_2 \cdots > w_n$ [7]. 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).

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 [20], the relevance scores are assumed to follow a Gaussian distribution, and as in [19], 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 $Var(R_n)$ in [14]:

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

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

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

where $Var_-(R_n)$ ($Var_+(R_n)$) denotes the downside (upside, respectively) variance of the overall relevance. Unlike the variance defined in Eq. (3), which can be calculated exogenously, the semivariance is endogenous. As a result, we use the approximation proposed by [5] to calculate $Var_Q(R_n)$:

$$Var_Q(R_n) \approx \sum_{i=1}^{n}\sum_{i=1}^{n} w_i w_j \hat{c}_{i,j},$$

where

$$\hat{c}_{i,j} = \begin{cases} E\left[Min\left(r_i - E[r_i], 0\right) \times Min(r_j - E[r_j], 0)\right], & \text{if } Q = -, \\ E\left[Max\left(r_i - E[r_i], 0\right) \times Max(r_j - E[r_j], 0)\right], & \text{if } Q = +. \end{cases} \tag{4}$$

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

$$max\ A_n = E[R_n] - a \times Var_Q(R_n), \tag{5}$$

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. (5). Therefore, the greedy algorithm in [14] is adopted to optimize the objective function. The difference of the objective function from the position $k-1$ to k is

$$\begin{aligned} A_k - A_{k-1} &= E[R_k] - a \times Var_Q(R_k) - E[R_{k-1}] + a \times Var_Q(R_{k-1}) \\ &= \sum_{i=1}^{k} w_i E[r_i] - a \sum_{i=1}^{k}\sum_{i=1}^{k} w_i w_j \hat{c}_{i,j} - \sum_{i=1}^{k-1} w_i E[r_i] + a \sum_{i=1}^{k}\sum_{i=1}^{k-1} w_i w_j \hat{c}_{i,j} \\ &= w_k(E[r_k] - a w_k \hat{c}_{kk} - 2a \sum_{i=1}^{k-1} w_i \hat{c}_{i,k}). \end{aligned} \tag{6}$$

The i-th ranked document for $i \in \{2, 3, \cdots, n\}$ is selected to maximize the difference $A_i - A_{i-1}$ in Eq. (6). (Note that the first document ($i = 1$) is set to the document with the highest relevance score.)

Calculation of Means and Semicovariance Matrix

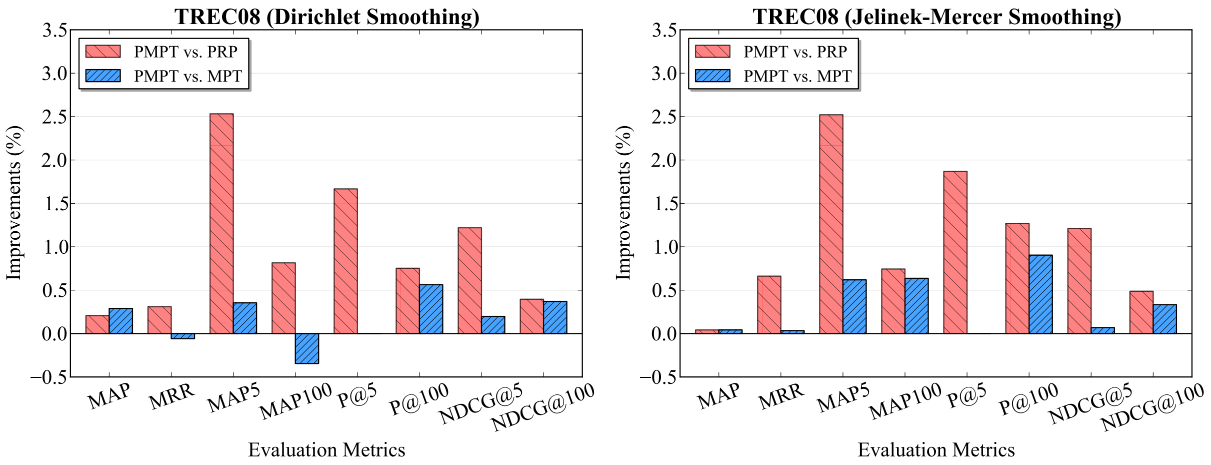
Different retrieval models generate different estimators of $E[R_n]$ and $Var_Q(R_n)$. In the following experiments, the probabilistic language models with Dirichlet and Jelinek-Mercer (JM) smoothing are adopted [15]. In a multinomial language model with parameter $\mathbf{y} = (y_1, \cdots, y_i, \cdots, y_{|V|})$, given a document $\mathbf{d} \equiv (d_1, \cdots, d_i, \cdots, d_{|V|})$, the posterior probability can denoted as $p(\mathbf{y} \mid \mathbf{d}, \alpha)$, and

$$\begin{aligned} p(\mathbf{y} \mid \mathbf{d}, \alpha) \propto p(d \mid \mathbf{y}) p(\mathbf{y} \mid \alpha) &= \prod_i (y_i)^{d_i} \prod_i (y_i)^{\alpha_i - 1} \\ &= \prod_i (y_i)^{d_i + \alpha_i - 1} \\ &\sim Dir(\mathbf{d} + \alpha), \end{aligned} \tag{7}$$

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

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

$$\hat{y}_i = \frac{d_i + \alpha_i}{\sum_{i=1}^{n} d_i + \alpha_i}. \tag{8}$$



(a)

(b)

Figure 1: **Comparison of our approach (PMPT) against the MPT and the PRP on TREC2008 ad hoc task.**

According to Eq. (8), given a query $\mathbf{q} \equiv (q_1, \cdots, q_i, \cdots, q_{|V|})$, $E[r_i]$ can be estimated by

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

On the other hand, we draw $\mathbf{y}$ as samples from the distribution in Eq. (7) to obtain randomized relevance scores; then the semicovariance defined in Eq. (4) can be easily calculated.

4.3 Experiments

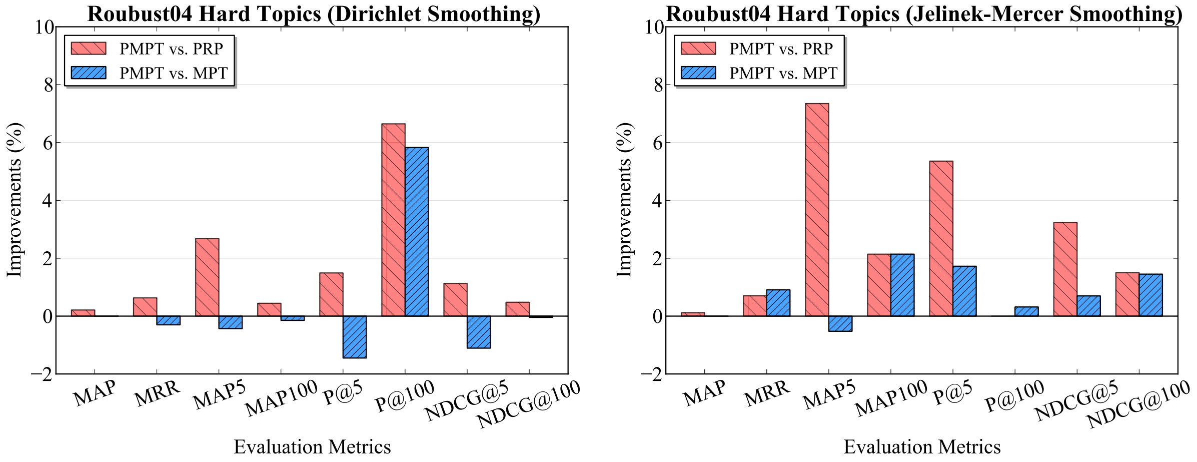
This section first describes the experimental datasets and evaluation metrics. With respect to different datasets and smoothing techniques, there are four sets of experiments conducted in this paper. We then present the experimental results, and conclude this section by providing some discussions and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

In our experiments, two TREC tracks are used for evaluating the proposed method, including TREC08 and Robust04. Table 2 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.

Figure 1 and Figure 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 [14]. As shown in Figure 1, when operating on the TREC08 dataset, the proposed PMPT method improves most of evaluation metrics

Table 2: **Overview of the two TREC test collections.**

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



(a) (b)

Figure 2: **Comparison of our approach (PMPT) against the MPT and the PRP on Robust2004 hard topics.**

over the baseline, PRP, and MPT; this improvement shows that 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 Figure 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 Figure 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 [19], which suggests that it might be more preferable to apply the “risk-sensitive“ approach with the JM smoothing technique.

5 Conclusions and Future Work

This extension work proposes a general mean-semivariance framework to study document ranking under uncertainty, which also implies the diversity of retrieved results. 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-to-rank techniques to find out the optimal parameters of the proposed framework, and how to adapt the framework for diversified information retrieval. We will also study how to combine the two-step ranking SVM technique with the mean-semivariance framework.

References

[1] Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336, New York, NY, USA, 1998. ACM.

[2] Harr Chen and David R. Karger. Less is more: probabilistic models for retrieving fewer relevant documents. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 429–436, New York, NY, USA, 2006. ACM.

[3] Charles L.A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. Novelty and diversity in information retrieval evaluation. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 659–666, New York, NY, USA, 2008. ACM.

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

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

[6] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874, 2008.

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

[8] T. Joachims. Making large-Scale SVM Learning Practical. *Advances in Kernel Methods-Support Vector Learning, B. Schölkopf and C. Burges and A. Smola*, 1999.

[9] Thorsten Joachims. Optimizing search engines using clickthrough data. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142, New York, NY, USA, 2002. ACM.

[10] T.-Y. Liu, J. Xu, T. Qin, W. Xiong, and H. Li. LETOR: Benchmark Dataset for Research on Learning to Rank for Information Retrieval. *Proceedings of SIGIR 2007 Workshop on Learning to Rank for Information Retrieval*, 2007.

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

[12] Filip Radlinski, Robert Kleinberg, and Thorsten Joachims. Learning diverse rankings with multi-armed bandits. In *ICML '08: Proceedings of the 25th international conference on Machine learning*, pages 784–791, New York, NY, USA, 2008. ACM.

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

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

[15] C. Zhai and 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*, pages 334–342. ACM, 2001.

[16] ChengXiang Zhai. *Risk Minimization and Language Modeling in Information Retrieval*. PhD thesis, Carnegie Mellon University, 2002.

[17] ChengXiang Zhai, William W. Cohen, and John Lafferty. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 10–17, New York, NY, USA, 2003. ACM.

[18] ChengXiang Zhai and J. Lafferty. A risk minimization framework for information retrieval. *Information Processing and Management*, 42(1):31–55, 2006.

[19] J. Zhu, J. Wang, I.J. Cox, and M.J. 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*, pages 99–106. ACM, 2009.

[20] J. Zhu, J. Wang, M. Taylor, and I. Cox. Risk-aware information retrieval. *Advances in Information Retrieval*, pages 17–28, 2009.

國科會補助專題研究計畫項下國際合作研究計畫
國外研究報告

日期：2012 年 10 月 20 日

計畫編號	NSC　100 － 2218 － E － 004 － 001 －		
計畫名稱	改善多樣性檢索之搜索結果		
出國人員姓名	蔡銘峰	服務機構及職稱	國立政治大學資訊科學系 助理教授
合作國家	西班牙	合作機構	龐貝法布拉大學
出國時間	2012 年 3 月 31 日至 2012 年 4 月 8 日	出國地點	西班牙，巴塞隆納

一、國際合作研究過程

主要是應西班牙龐貝法布拉大學 Mari-Carmen Marcos 教授之邀擔任英國電腦學會（British Computer Society，BCS）所主辦之歐洲資訊檢索大會（European Conference on Information Retrieval，ECIR）之議程委員，希望能夠透過參加會議的方式，建立起國際合作的機會。本次會議於 2012 年 4 月 1 日至 4 月 5 日在西班牙巴塞隆納龐貝法布拉大學（University of Pompeu Fabra，UPF）舉行。

二、研究成果

ECIR 會議為資訊檢索領域之間交流的主要會議之一， 此次與會人數約有 200 多人，其中包括各個國家之資訊工程與電腦科學等研究人員，更有許多搜索引擎與社群網路等國際級大公司人士參與，如：Google、Microsoft、Yahoo! 等。在會議過程中，結識了來自各國的資訊檢索研究人員，如：美國伊利諾大學香檳分校（University of Illinois at Urbana-Champaign，UIUC）的 ChengXiang Zhai 教授、西班牙 Yahoo! 研究院的 Ricardo Baeza-Yates 研究員、英國倫敦大學學院（University College London，UCL）的 Jun Wang 教授等，和

這些研究人員們，亦有很深入的討論，他們也對於資訊檢索領域給予本人一些他們個人的看法、以及本人之前在 Learning to Rank 上研究的一些建議，對於本人往後對於資訊檢索領域的相關延伸研究有了相當大的幫助和啟發。此外，ChengXiang Zhai 教授和 Ricardo Baeza-Yates 研究員更是邀請本人往後若有機會，可以到 UIUC 或是 Yahoo! 西班牙研究院訪問研究。這些難得的機會，更是讓本人有種不虛此行的感覺！

此次會議中的 Invited Keynote Speakers 的演講皆十分精采，部分列舉如下：

1) Paolo Boldi, University of Milano
   Studying Network Structures for IR: the Impact of Size
   - Prof. Boldi 演講內容十分精采，主要介紹如何將社群網路結構（Social Network Structure）與資訊擷取（Information Extraction）作結合，並搭配 Graph Algorithms 來處理當資料量變大時的檢索排序問題（Ranking Problem）。

2) Yoelle Maarek, Yahoo! Labs
   The Surprising Role of Users in Web Search
   - Yoelle Maarek 研究員主要介紹如何利用使用者的使用量資訊（Usage Data Information）來改善使用者在網路搜索上的使用者經驗（User Experience），無論是搜索結果排序或是搜索介面的設計。在此場演講中，也提到了利用使用量資訊來改善網路搜索，可能面臨到的限制和風險等。

除了以上演講，會議中其他學者的論文報告亦十分精采，讓本人此行收穫良多。以下列舉幾項印象比較深刻的演講：

1) Interactive Search Support for Difficult Web Queries
   Abdigani Diriye, Giridhar Kumaran, Jeff Huang
   - 此篇論文提出一個新的搜索介面（searchAssist），針對網路上比較難檢索的問句，可以提供互動支援。

此一搜索介面的特色在於，其可以自動縮短使用者的長問句，讓搜索引擎比較不擅長於處理的長問句或是困難問題，可以有效地降低。根據其實驗結果顯示，其介面可以有效改善約 40%的困難問句，讓使用者更有效率地在網路上檢索資訊。

2) Result Disambiguation in Web People Search
Richard Berendsen, Bogomil Kovachev, Evangelia-Paraskevi Nastou, Maarten de Rijke, Wouter Weerkamp

- 此篇論文主要是在探討如何改善網路人名搜索之多樣性。在網路中，有時我們需要找尋某個人時，大都會利用人名的方式來搜索，但若有多人同名的情況時，將不易找尋到需要的結果。此篇論文中，提出如何利用社群多媒體（Social Media）的資訊並且結合分群（Clustering）技術，來提昇提到的網頁中提及的人名之準確度。雖然此篇論文主要是針對人名搜索，但對於本次計畫中的多樣性搜索也提供很多相關價值可供參考，會後本人也向演講者請益許多問題，可說是獲益良多。

三、建議

本國資訊檢索領域在此次重要會議中無任何論文發表，相較其他國家，資訊檢索領域研究人員參與此類會議略顯不足。

四、其他

在會議期間也遇到的一些學者，對於建立合作研究很有興趣，回來後本人也會繼續加強聯繫，期待有朝一日與他國研究人員的合作能夠開花結果。同時，也非常感謝國科會給予支持。

ECIR 2012
34th European Conference on Information Retrieval
Barcelona, April 1-5 2012
http://ecir2012.upf.edu

Department of Information Technology
and Communications,
Pompeu Fabra University,
Tanger, 122-140
08018 Barcelona, Spain.
Tel: +34 93 542 25 00
Fax: + 34 93 542 25 17
Email: mcmarcos@upf.edu

March 5, 2012

To: Ming-Feng Tsai
Assistant Professor
Department of Computer Science
National Chengchi University,
Taiwan.

Dear Dr. Tsai,

On behalf of the Organizing Committee, I would like to extend a formal invitation for you to attend the upcoming **ECIR 2012 Conference** as Program Committee Member**, 34th European Conference on Information Retrieval**, to be held in Barcelona, Spain, 1-5 April 2012.

This letter can be used by the person stated in the letter, Dr. Ming-Feng Tsai, who is an Assistant Professor in the Department of Computer Science at the National Chengchi University, Taiwan, for a Visa application to cover the stay in Barcelona, Spain, for the conference, in the stated period.

Dr. Ming-Feng Tsai
Assistant Professor
Department of Computer Science
National Chengchi University,
Taiwan.

Yours sincerely,

Mrs. Mari-Carmen Marcos
Chair, Local Organizing Committee

國科會補助計畫衍生研發成果推廣資料表

日期:2012/10/27

國科會補助計畫	計畫名稱: 改善多樣性檢索之搜索結果
	計畫主持人: 蔡銘峰
	計畫編號: 100-2218-E-004-001- 學門領域: 自然語言處理與語音處理

無研發成果推廣資料

100年度專題研究計畫研究成果彙整表

計畫主持人：蔡銘峰 計畫編號：100-2218-E-004-001-

計畫名稱：改善多樣性檢索之搜索結果

成果項目			量化			單位	備註（質化說明：如數個計畫共同成果、成果列為該期刊之封面故事...等）
			實際已達成數（被接受或已發表）	預期總達成數(含實際已達成數)	本計畫實際貢獻百分比		
國內	論文著作	期刊論文	0	0	100%	篇	
		研究報告/技術報告	0	0	100%		
		研討會論文	0	0	100%		
		專書	0	0	100%		
	專利	申請中件數	0	0	100%	件	
		已獲得件數	0	0	100%		
	技術移轉	件數	0	0	100%	件	
		權利金	0	0	100%	千元	
	參與計畫人力（本國籍）	碩士生	3	0	100%	人次	
		博士生	1	0	50%		
		博士後研究員	0	0	100%		
		專任助理	0	0	100%		
國外	論文著作	期刊論文	0	1	100%	篇	研討會論文之延伸期刊版本仍在進行中。
		研究報告/技術報告	0	0	100%		
		研討會論文	1	1	100%		International Neural Network Society Winter Conference (INNS-WC2012), Thailand 2012.
		專書	0	0	100%	章/本	
	專利	申請中件數	0	0	100%	件	
		已獲得件數	0	0	100%		
	技術移轉	件數	0	0	100%	件	
		權利金	0	0	100%	千元	
	參與計畫人力（外國籍）	碩士生	0	0	100%	人次	
		博士生	0	0	100%		
		博士後研究員	0	0	100%		
		專任助理	0	0	100%		

其他成果 (無法以量化表達之成果如辦理學術活動、獲得獎項、重要國際合作、研究成果國際影響力及其他協助產業技術發展之具體效益事項等，請以文字敘述填列。)	無

🟥	成果項目	量化	名稱或內容性質簡述
科教處計畫加填項目	測驗工具(含質性與量性)	0	
	課程/模組	0	
	電腦及網路系統或工具	0	
	教材	0	
	舉辦之活動/競賽	0	
	研討會/工作坊	0	
	電子報、網站	0	
	計畫成果推廣之參與（閱聽）人數	0	

國科會補助專題研究計畫成果報告自評表

請就研究內容與原計畫相符程度、達成預期目標情況、研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）、是否適合在學術期刊發表或申請專利、主要發現或其他有關價值等，作一綜合評估。

1. 請就研究內容與原計畫相符程度、達成預期目標情況作一綜合評估

■達成目標

□未達成目標（請說明，以 100 字為限）

□實驗失敗

□因故實驗中斷

□其他原因

說明：

2. 研究成果在學術期刊發表或申請專利等情形：

論文：□已發表 □未發表之文稿 ■撰寫中 □無

專利：□已獲得 □申請中 ■無

技轉：□已技轉 □洽談中 ■無

其他：（以 100 字為限）

本計劃相關之會議論文初步結果已於 International Neural Network Society Winter Conference（INNS-WC2012), Thailand, 2012 發表。延伸版本之期刊論文，目前正在進行中。

3. 請依學術成就、技術創新、社會影響等方面，評估研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）（以 500 字為限）

此研究計劃主要目的為發展出一個可以考量到多樣化檢索結果之特性的學習式排列演算法，並且進一步增加更多值得擁有的特性，如：考量到地域性以及時間性的排列模型。除了學習式演算法的推導之外，本人也希望藉由這個研究計劃，深入探討排列問題其學習理論之基礎。而所發展出來的學習排列演算法，也希望可以應用到一套真實線上檢索系統中，透過如此的結合將可以利用更多使用者回饋的資料來改善既有的系統。最後，也希望能藉有此研究計劃中所發展出來的技術，繼續推動國內學術界及業界對於資訊檢索和機器學習之結合新興領域的相關研究。