應用基於排序相關係數之特徵向量轉換於資訊檢索排序模型學習

Feature-Level Context Transformation for Learning to Rank for Information Retrieval Based on Rank-Order Correlation

葉鎮源 *
Jen-Yuan Yeh
國立自然科學博物館營運典藏與資訊組
館前路 1 號
臺中市 404 北區
jenyuan@mail.nmns.edu.tw

林忠億
Jung-Yi Lin
鴻海精密工業股份有限公司創新數位系統事業群
基湖路 32 號
臺北市 114 內湖區
jungyilin@gmail.com

鄭培成
Pei-Cheng Cheng
健行科技大學資訊管理學系
健行路 229 號
桃園縣 320 中壢市
pccheng@uch.edu.tw

楊維邦
Wei-Pang Yang
國立東華大學資訊管理學系
大學路二段 1 號
花蓮縣 974 壽豐鄉
wpyang@mail.ndhu.edu.tw

摘要

排序模型學習一般以特徵向量表示輸入的樣本資料，運用機器學習對已正確排序(或分類)的資料分析歸納，自動建構出有效的排序模型或規則。本文基於排序相關係數(Rank-Order Correlation Coefficient)計算不同特徵在排序特性的關聯程度，並藉由特徵關聯作中介轉換，將特徵向量映射到特徵關聯空間(Feature Correlation Space)，使得特徵向量由一階表徵(First-Order Raw Representation)提升到二階表徵(Second-Order Context Representation)。同時，以二階表徵特徵向量作為資料表示，整合 RankSVM 建構一線性函式(Linear Function)形式的二元分類器，用來判別兩兩文件是否為正確排

* 本文通訊作者。

序，進而推導建立所有文件的排序序列。實驗使用 LETOR 4.0 資料集驗證本文所提方法的可行性，評估指標選用 MAP 和 MeanNDCG，並以 RankSVM 為基準，比較二階表徵特徵向量對於排序模型學習的影響，評估結果顯示本文所提方法的可行性。

關鍵詞: 文件檢索、排序模型學習、特徵關聯計算、特徵向量二階表徵轉換、排序預測與評估。

Abstract

In practice, methods of learning to rank for information retrieval typically represent training instances as vectors of features and exploit supervised learning to automatically produce an effective ranking model (or retrieval function). This paper measures relationship between features using rank-order correlation coefficients, based on which second-order context vectors are derived by projecting first-order raw vectors into the feature correlation space. A novel learning method is then proposed, based on second-order context vectors, to train a ranking model (in form of linear function) by RankSVM. The proposed learning method was evaluated using the LETOR 4.0 dataset and found to perform well, in terms of metrics of MAP and MeanNDCG.

Keywords: document retrieval, learning to rank, feature correlation extraction, feature-level second-order representation transformation, ranking prediction and evaluation

一、前言

資訊檢索處理資訊的呈現(Representation)、儲存(Storage)、組織(Organization of)和取得利用(Access to)，是根據使用者的資訊需求提供查詢的方法和過程[2]。資訊檢索的研究議題相當廣泛，其中一個非常重要且過去文獻探討過最多的是「如何決定文件與查詢條件是否相關？」。實務上這個議題被視為*文件排序*的問題，目的是計算查詢條件與文件的相關度(Relevance)或相似度(Similarity)，藉此篩選出與查詢相關的文件，並將相關度高的文件排序在查詢結果的前面位置，而將相關度低的排序在結果的後面位置。傳統資訊檢索的研究提出各種不同的檢索模型[2][25]，包括：(1) 布林模型(Boolean Model)、(2) 向量空間模型(Vector Space Model)、(3) 機率模型(Probabilistic Model)，以及(4) 語言模型(Language Model)。前列模型將文件與查詢條件表示為索引關鍵詞的集合，同時定義排序函式(Ranking Function，或稱 Retrieval Function)計算查詢條件與文件的相關度。通常，排序函式多以非監督式(Unsupervised)方法設計，例如：Okapi BM25 [34]，而函式中的參數則由經驗法則(Heuristics)或實驗測試決定，不但不易設定適當的值，也可能導致模型過適(Over-fitting)、不具通用性的窘境。

排序模型學習使用監督式(Supervised)的機器學習，對已正確排序(或分類)的訓練資料進行特徵擷取和分析歸納，自動建構出有效的排序模型或規則，且此模型被認為可將具相同性質的新資料正確排序或分類[1]。排序模型學習應用於資訊檢索，稱之為***應用機器學習於資訊檢索排序模型建構*** (***Learning to Rank for Information Retrieval; LR4IR***)，其與一般排序模型的差別在於排序的對象是文件，且須考量文件與查詢條件

的相關性。排序模型學習具有優點[23]：(1) 基於特徵向量的查詢與文件對組表示，使得各種檢索模型在轉換為單一特徵後，可輕易整合於排序學習的演算法；(2) 具鑑別學習(Discriminative Training)能力，可利用訓練資料學習得到排序模型。此外，諸如TREC[1]、CLEF[2]及NTCIR[3]等檢索效能評比的舉辦，使得越來越多根據個別查詢條件標記相關性的文件集被釋出。因此，近年來應用機器學習於資訊檢索排序模型建構的研究可說如雨後春筍般蔚為風潮。

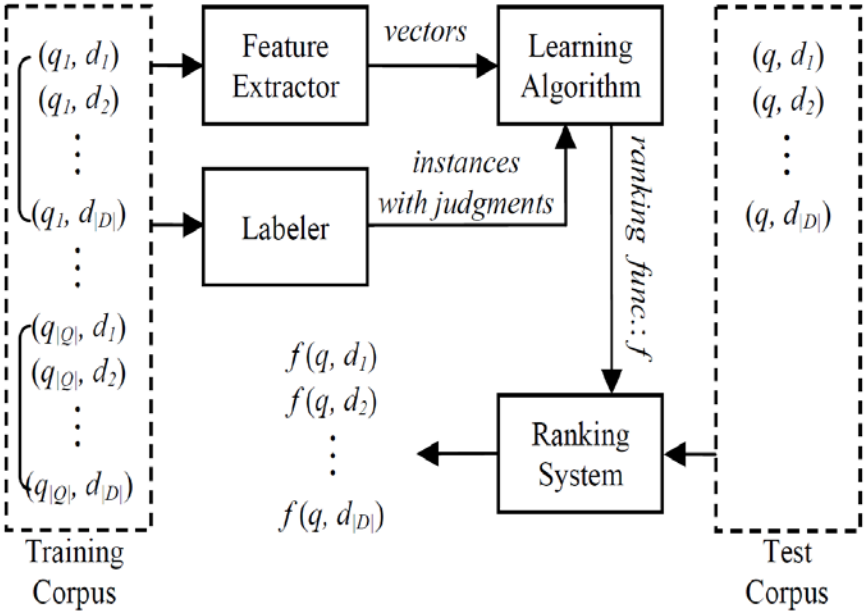
應用機器學習於資訊檢索排序模型建構的研究大多都以*特徵向量*表示輸入的樣本資料。一般來說，基於描述資料的特徵定義特徵值，並視特徵為向量空間的維度，可將輸入資料轉換成特徵向量作為學習演算法的基本處理單元。但不同於傳統資訊檢索研究將字詞視為向量空間的維度，應用機器學習於資訊檢索排序模型建構的研究則以各種排序函式作為特徵，其目的是整合不同的排序函式得到最佳的排序模型。前述的特徵包括有基於計算查詢條件與文件相關度的特徵，稱為*與查詢相關*(Query-Dependent)的特徵；或是*與查詢無關*(Query-Independent)的特徵，用來作為查詢結果排序的客觀條件，以從不同面向的考量來調整或強化與查詢相關方法的結果[4]。與查詢相關的特徵計算查詢條件與文件之索引詞的相似度當作特徵值，此做法可能因查詢條件的索引詞集合過小而導致資料稀疏性(Sparsity)問題；換言之，特徵向量中可能存在某些維度值為零(或數值失真)的情形。而對與查詢無關的特徵來說，也可能因為資訊量不足引發相同問題。再者，特徵向量的另一個缺點是假設特徵彼此獨立無關。但事實上，特徵彼此的關係並非絕對無關，例如：TF-IDF [2]和Okapi BM25 [34]都是由TF (Term Frequency)及IDF (Inverse Document Frequency)兩個元素組成。若將這兩個排序函式視為向量空間的獨立維度時，其相關性便被忽略，無法真實反應在向量表示式中。

過去文獻雖有學者分析特徵在排序序列異同的特性(例如：[14][35])，但鮮少有研究針對向量表示的資料稀疏及特徵維度並非獨立無關等問題對於排序模型學習的影響進行探討。本文提出基於排序相關係數(Rank-Order Correlation Coefficient)計算兩兩特徵關聯強度的方法，並藉由特徵的排序關聯作中介轉換，將特徵向量映射(Project)到特徵關聯空間，使得特徵向量由一階表徵(First-Order Raw Representation)提升到二階表徵(Second-Order Context Representation)，以填補原始稀疏的特徵值並改善因假設特徵獨立無關導致向量描述可能失真的問題。最後，以特徵向量二階表徵轉換為基礎，提出一套新的資訊檢索排序模型學習架構，期在訓練學習過程中有效產出更佳的排序模型。

本節接續介紹應用機器學習於資訊檢索排序模型建構的方法架構。圖一是典型的應用機器學習於資訊檢索排序模型建構架構[48]，由訓練(Training)和測試(Test)兩個階段組成。考慮查詢集合$Q = \{q_1, \ldots, q_{|Q|}\}$和文件集合$D = \{d_1, \ldots, d_{|D|}\}$，訓練資料定義為「查詢與文件對組(Query-Document Pair)」的集合，任一對組$(q_i, d_j) \in Q \times D$。標記器(Labeler)賦予每個對組一個相關性標記(Relevance Judgment)作為訓練學習的參考答案。標記通常由專家判定，可能的類型有：(1) 分類標記(Class)，例如：相關或不相關；

[1] TREC (Text REtrieval Conference), http://trec.nist.gov/.
[2] CLEF (Cross Language Evaluation Forum), http://www.clef-initiative.eu/.
[3] NTCIR (NII Test Collection for IR Systems), http://research.nii.ac.jp/ntcir/.
[4] 例如：考慮網頁連結結構的 PageRank [4]和 TrustRank [16]、考慮使用者瀏覽行為的 BrowseRank [24]。

(2) 等級標記(Rating)[5]，例如：絕對相關、可能相關，或不相關；(3) 順序標記(Order)，例如：k代表d_j在q_i的查詢結果中排在第k個位置；(4) 分數標記(Score)，例如：$sim(q_i, d_j)$表示q_i與d_j的相似度。



圖一: 應用機器學習於資訊檢索排序模型建構架構[48]

考慮特徵集合 $F = \{f_1, \ldots, f_{|F|}\}$，特徵擷取器(Feature Extractor)計算查詢與文件對組(q_i, d_j)在特徵 f_k 的條件下，q_i 與 d_j 的關係經過量化後的特徵值 $f_k(q_i, d_j)$。若將特徵視為向量空間維度，便得到一組特徵向量表示 q_i 與 d_j 的關係；換言之，定義$(q_i, d_j) = (f_1(q_i, d_j), \ldots, f_{|F|}(q_i, d_j))$。此處，特徵可以是各種檢索方法或排序模型，包括：(1) 與查詢相關的特徵，例如：TF [2]、IDF [2]和 BM25 [34]，以及(2) 與查詢不相關的特徵，例如：HostRank [47]、Feature Propagation [33][36]和 Topical PageRank [30]。

學習演算法(Learning Algorithm)接受訓練資料(即，查詢與文件對組的特徵向量和相關性標記)，經由學習過程建構排序函式 f，函數值 $f(q_i, d_j)$可視為(q_i, d_j)的標記。例如，假設 f 為相似度函式，$f(q_i, d_j)$表示 q_i 與 d_j 的相似度。以查詢 q_i 來說，文件集合 D 的所有文件藉此函式計算與 q_i 的相似度，再依函數值排序得到一文件序列作為 q_i 的查詢結果。一般認為，整個訓練學習的過程即是一個搜尋求解的過程[28]，通常須導入特定的評估指標，例如：準確度(Accuracy)、錯誤率(Error Rate)，或者資訊檢索特有指標 MAP (Mean Average Precision) [2]及 NDCG (Normalized Discount Cumulative Gain) [18]，用以評量學習過程中產出之排序模型的優劣，進而最佳化求得適當的模型。

在測試階段中，考慮新查詢 q，任一查詢與文件對組(q, d_i)經由特徵擷取器得到特徵向量，這組向量輸入排序函式 f 得到 $f(q, d_i)$。同樣地，$f(q, d_i)$可當成(q, d_i)的相關性標記。最後，預測得到一文件序列作為 q 的查詢結果。

本文共六個章節：第二節整理排序模型學習相關的研究及其在資訊檢索的應用；第三節詳述應用基於排序相關係數之特徵向量轉換於資訊檢索排序模型學習的架構；

[5] 當只定義相關和不相關兩種等級時，等級標記被視為分類標記的一種。

第四節說明實驗設計及結果討論；第五節探討所提出之排序模型學習方法的特性與貢獻；最後是結論和未來研究發展建議。

二、 相關研究

排序模型學習的研究大致上可分成三種類型[23]，包括：(1) 逐點式方法(Pointwise Approach)、(2) 成對式方法(Pairwise Approach)，及(3) 序列式方法(Listwise Approach)。在逐點式方法中，訓練資料實例(Instance)的相關性標記屬於分類或等級標記。通常，這類技術將排序問題轉換為迴歸分析(Regression)或分類問題，目的是建構一個分類模型將任一資料實例對應到可能的等級或類別，且此對應的標記被認為是正確的等級或類別。例如，Prank [9]結合感知模型(Perceptron Model)，透過映射(Projection)直接得到完整的物件排序。McRank [21]定義 5 種等級標記{0, 1, 2, 3, 4}(0 表示不相關，4 表示完全相關)，排序模型建構運用 Gradient Boosting (Friedman 2001)進行分類學習。

在成對式方法中，訓練資料被轉換為兩兩物件的對組(Pair of Objects)，並將原始的相關性標記轉換成表示兩個物件的排序關係標記。這類方法基於分類法則建構排序模型，藉以將兩個物件的對組標示成「正確排序」或「錯誤排序」。接著，由兩兩成對物件是否為正確排序可推導建立所有物件的排序序列。例如：RankSVM [19]將查詢與文件對組(q, d_i)和(q, d_j)轉換為新的資料實例(d_i, d_j)，並使用 SVM 建立兩兩物件對組的二元分類器。而排序關係標記的轉換原則：若 d_i 與 d_j 為正確排序，則將排序關係標記成+1；相反地，若 d_i 與 d_j 是錯誤排序，則將(d_i, d_j)的排序關係標記為−1。RankBoost [13]和 QBrank [54]運用 Boosting 得到統整式排序模型，可將錯誤排序的物件對組數量降到最低。RankNet [6]及 LambdaRank [5]應用類神經網路求得最佳成本函數作為排序模型，前者利用交叉熵(Cross Entropy)定義評估的成本函數(Cost Function)，後者則使用基於 NDCG 定義的 Gradient。FRank [40]提出精準損失函數(Fidelity Loss Function)，並結合廣義加法模型(Generalized Additive Model)進行學習，使得排序結果更為精準。Semi-RankSVM [31]使用 RankSVM 作最佳化學習，並利用圖形正規化達到最小化成對損失。

在序列式方法中，訓練資料不再以單一物件或兩兩物件的對組為單元，而將焦點放在所有物件的完整排序序列上，企圖建構可直接產出完整排名的排序模型。例如：ListNet [7]導入以機率法則為基礎的序列式損失評估函數(Loss Function)，運用類神經網路及 Gradient Descent 建構序列的預測模型。特別一提，成對式方法與序列式方法的最小錯誤(Least-Error)函數都以最小排序文件錯誤為估算基礎，當序列只有兩個文件時，序列式學習便退化為成對式學習。因此，成對式方法被視為是序列式方法的一個特例[1]。

近年來有許多研究將排序模型學習應用於資訊檢索，例如：[29]使用相關、不相關兩種類別作為查詢與文件對組的標記，並運用最大熵值法(Maximum Entropy Approach)與支援向量機(Support Vector Machine; SVM)建構分類模型，其文件排序的原則是將被分類為相關的文件排序在被分類為不相關的文件之前。[19]分析搜尋記錄中查詢和點擊(Click-Through)的使用者行為，提出 RankSVM 得到較佳的排序模型。[8]改良 RankSVM 的 Hinge Loss Function，將兩個影響檢索好壞的因素納入考量：(1) 排

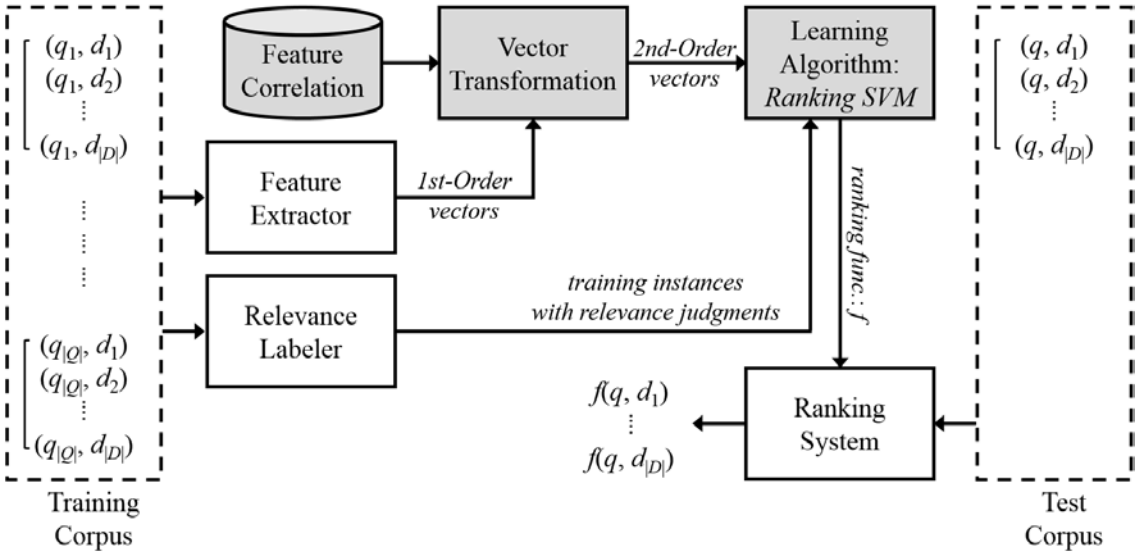
序在越前面的文件有較高的正確性，及(2) 訓練學習應避免因某些查詢有較多相關的文件訓練實例而導致排序模型學習的偏差。[45]將 SVM 與 RankSVM 整合應用於詞彙定義的搜尋，針對使用者查詢的詞彙提供較精確的詞彙解釋。[50]延伸 SVM 於分類的樣本挑選(Sampling)技術來解決排序問題。[27]利用 RankNet 提出具遞迴結構的多層式排序模型學習架構，可將查詢結果中位於前面位置的文件重新排列得到更好的結果。SVM-MAP [52]延伸 RankSVM，直接對評估指標 MAP [2]進行檢索排序模型的最佳化。AdaRank [46]改良 RankBoost，以評估指標 MAP 和 NDCG [18]作最佳化得到排序模型。有別於 RankSVM 使用查詢與文件對組的向量差作為訓練資料表示，RV-SVM [51]改良 RankSVM，直接使用單一查詢與文件對組作輸入，而輸出的排序模型為一非線性函數。相關的研究另有：[3][39][42][43][44][55]。

除此之外，過去文獻另有研究應用基因演算法(Genetic Algorithm; GA)和基因規劃法(Genetic Programming; GP)進行排序模型的學習，例如：[11]和[12]將排序函式表示成樹狀結構(其中，節點為運算元、葉節點為變數)，透過 GP 產出最佳化的排序模型。[10]探討 GP 在演化過程中選擇不同評估函數(Fitness Function)對於排序模型學習的影響。RankGP [48]使用單族群 GP (Single Population GP)，族群中每個個體為一線性函式，每個函式計算查詢與文件對組的相似度作為排序依據，並以評估指標 MAP 作最佳化學習。RankMGP [22]提出多層式(Layered)、多族群(Multi-Population)的 GP 架構，每個階層中的各個族群獨立演化，並以每個族群的最佳個體作為向量維度重新定義輸入的訓練資料，藉此加速演化速度、提高學習的準確度。

三、 研究架構與方法設計

本文延伸圖一應用機器學習於資訊檢索排序模型建構架構，提出「應用基於排序相關係數之特徵向量轉換於資訊檢索排序模型學習」方法，詳見圖二，而與圖一的差異以灰色標示：

- 特徵關聯(Feature Correlation)：導入排序相關係數計算兩兩特徵的關聯強度，以建構特徵關聯矩陣；
- 特徵轉換(Vector Transformation)：藉由特徵關聯矩陣的中介轉換，將特徵擷取器輸出的一階表徵特徵向量映射到特徵關聯空間，得到二階表徵特徵向量作為學習演算法輸入；
- 基於 RankSVM 的學習演算法(Learning Algorithm: RankSVM)：整合特徵轉換得到的二階表徵特徵向量，使用 RankSVM [19]作為學習演算法進行排序模型學習。



圖二: 應用基於排序相關係數之特徵向量轉換於資訊檢索排序模型學習的架構

表一定義符號以方便各章節中說明。表二整理圖二在訓練和測試階段的工作流程。

表一: 符號定義與說明

符號	定義與說明
F	特徵集合 $F = \{f_1, \ldots, f_k\}$, $\|F\| = k$
Q	查詢集合 $Q = \{q_1, \ldots, q_m\}$, $\|Q\| = m$
D	文件集合 $D = \{d_1, \ldots, d_n\}$, $\|D\| = n$
$D(q_i)$	查詢 q_i 的文件集合 $D(q_i) = \{d_{i,1}, \ldots, d_{i,n_i}\}$, $D(q_i) \subseteq D$, $\|D(q_i)\| = n_i$
f	排序模型 $f(\cdot)$。$f(q, d)$表示對查詢 q 而言，文件 d 的相關性標記

表二: 圖二應用基於排序相關係數之特徵向量轉換於資訊檢索排序模型學習的工作流程

	訓練階段	測試階段
輸入	$\{(q_1, D(q_1)), \ldots, (q_m, D(q_m))\}$[6]	$\{(q_{m+1}, D(q_{m+1}))\}$
輸出	排序模型 f	$\{f(q_{m+1}, d_{m+1,1}), \ldots, f(q_{m+1}, d_{m+1,n_{m+1}})\}$
工作流程	1. 資料標記 2. 特徵擷取 (1) 一階表徵特徵向量擷取 (2) 特徵關聯矩陣建構 (3) 特徵向量二階表徵轉換 3. 排序模型學習	1. 資料標記 2. 特徵擷取 (1) 一階表徵特徵向量擷取 (2) 特徵關聯矩陣建構 (3) 特徵向量二階表徵轉換 3. 排序預測 4. 效能評估

[6] 為了簡化說明，本文以 $D(q_i)$替代圖二的 D。

1. 資料標記

訓練階段的資料標記賦予機器學習時的參考答案；測試階段的資料標記則提供效能評估時用來比較機器標記的正確答案。本研究將「查詢與文件對組(q, d)」視為標記的最小有意義單元，以查詢 q 與文件 d 的相關度判別，依「等級標記」原則標記為：(1) 絕對相關(Definitely Relevant)、(2) 可能相關(Possibly Relevant)，以及(3) 不相關(Not Relevant)。

2. 特徵擷取與轉換

特徵擷取藉由特徵的定義和量化轉換得到一組特徵向量表示查詢 q_i 與文件 $d_{i,j}$ 的關係。但，特徵向量表示至少存在有：(1) 向量表示的資料稀疏性，及(2) 特徵維度並非獨立無關兩個問題。本文提出基於排序相關係數計算特徵關聯強度的方法，並藉由特徵的排序關聯作中介轉換，將特徵向量映射到特徵關聯空間，以填補原始稀疏的特徵值，同時改善因假設特徵獨立無關可能造成向量描述失真的問題。

整體而言，特徵擷取與轉換步驟，包含：(1) 一階表徵特徵向量擷取、(2) 基於排序相關係數之特徵關聯矩陣建構，以及(3) 特徵向量二階表徵轉換。

(1) 一階表徵特徵向量擷取

以特徵集合 F 的所有特徵構成向量空間維度，將查詢與文件對組(q_i, $d_{i,j}$)表示為特徵向量，定義查詢 q_i 與文件 $d_{i,j}$ 的關係：

$$\vec{d}_{i,j} =< w_{i,j,1}, ..., w_{i,j,k} > \tag{1}$$

$w_{i,j,k}$ = f_val(f_k, q_i, $d_{i,j}$)表示在特徵 f_k 的條件下，q_i 與 $d_{i,j}$ 的關係量化後的特徵值。此處，$\vec{d}_{i,j}$ 稱為「一階表徵特徵向量」，以便與之後說明的二階表徵特徵向量有所區別。

f_val 是特徵擷取函式，定義如下：

$$\text{f\_val}(\cdot): \{(f_k, q_i, d_{i,j}) \mid f_k \in F, q_i \in Q, d_{i,j} \in D(q_i)\} \rightarrow [0, 1] \tag{2}$$

(2) 特徵關聯矩陣建構

依特徵值 f_val(f_k, q_i, $d_{i,j}$)將 $D(q_i)$的所有文件排序得到一文件序列 $\overset{\vee}{R}_{i,k}$，定義如下：

$$\overset{\vee}{R}_{i,k}(D(q_i)) = [r_{i,k}(d_{i,j_1}) \succ ... \succ r_{i,k}(d_{i,j_{n_i}})] \tag{3}$$

$r_{i,k}$是一序列函數，可將 $D(q_i)$的文件對應到 $\overset{\vee}{R}_{i,k}$ 的相對位置，例如：$r_{i,k}(d_{i,j_l}) = l$ 表示 d_{i,j_l} 排在 $\overset{\vee}{R}_{i,k}$ 中第 l 的位置。$r_{i,k}(d_{i,p}) \succ r_{i,k}(d_{i,t})$，若且唯若存在 p 和 t，f_val(f_k, q_i, $d_{i,p}$) ≥ f_val(f_k, q_i, $d_{i,t}$)；換句話說，f_val(f_k, q_i, $d_{i,p}$) ≥ f_val(f_k, q_i, $d_{i,t}$)，$r_{i,k}(d_{i,p}) \succ r_{i,k}(d_{i,t})$ 表示 $d_{i,p}$ 在 $\overset{\vee}{R}_{i,k}$ 中排在 $d_{i,t}$ 的前面。

給定 q_i 和 $D(q_i)$，所有特徵依 Eq. (3)得到 k 組文件序列{ $\overset{\vee}{R}_{i,1}(D(q_i))$, …, $\overset{\vee}{R}_{i,k}(D(q_i))$ }。接著，利用排序相關係數計算兩兩序列在排序特性上的相關度(或稱，

排序關聯程度)。前述排序關聯程度乃指在 q_i 的條件下，不同特徵的排序關聯程度，基於此，本文提出 Macro-Correlation 法則，使得考慮查詢集合 Q 的所有查詢時，仍可計算得到特徵 f_i 與特徵 f_j 的整體性排序關聯程度 $c_{i,j}$：

$$c_{i,j} = \frac{1}{|Q|}\sum_{t}^{|Q|} F\_RCorr(\overset{\vee}{R}_{t,i}(D(q_t)), \overset{\vee}{R}_{t,j}(D(q_t))) \tag{4}$$

F_RCorr 是排序相關係數函數，用來計算 $\overset{\vee}{R}_{t,i}(D(q_t))$ 與 $\overset{\vee}{R}_{t,j}(D(q_t))$ 的排序關聯程度：

$$F\_RCorr(\cdot): \{(\overset{\vee}{R}_{t,i}(D(q_t)), \overset{\vee}{R}_{t,j}(D(q_t))) \mid q_t \in Q \text{ and } f_i, f_j \in F\} \rightarrow [0,1] \tag{5}$$

最後，Eq. (6)定義特徵關聯矩陣 C：

$$C = \begin{array}{c} \\ f_1 \\ \vdots \\ f_k \end{array} \begin{array}{c} \begin{array}{ccc} f_1 & \cdots & f_k \end{array} \\ \begin{bmatrix} c_{1,1} & \cdots & c_{1,k} \\ \vdots & \ddots & \vdots \\ c_{k,1} & \cdots & c_{k,k} \end{bmatrix} \end{array} \tag{6}$$

由 Eq. (4)可知，$c_{i,j} = c_{j,i}$。換言之，特徵關聯矩陣 C 是對稱矩陣(Symmetric Matrix)，滿足條件：

$$C_i = \hat{C}_i, \text{ i.e., } \forall p, \; c_{i,p} = \hat{c}_{i,p} \tag{7}$$

其中，$C_i = <c_{i,1}, \ldots, c_{i,k}>$是 C 的第 i 列列向量(Row Vector)，$\hat{C}_i = <\hat{c}_{i,1}, \ldots, \hat{c}_{i,k}>$是 C 的第 i 行行向量(Column Vector)。值得一提的是，C 的列向量(或行向量)同時定義單一特徵的向量表示，以特徵 f_i 來說，其向量表示式為：

$$\begin{aligned} \vec{f}_i &= C_i = < c_{i,1}, ..., c_{i,k} > \\ &= \hat{C}_i = < \hat{c}_{i,1}, ..., \hat{c}_{i,k} > \end{aligned} \tag{8}$$

表三條列本研究採用的 F_RCorr 排序相關係數函數。舉例來說，Kendall's *tau* (τ) 的值域是 [–1, +1]，–1 表示完全反相關(Perfect Inversion)，+1 表示完全相關(Perfect Agreement)，0 表示無關(No Association)。為了分析比較的方便，本文利用正規化(Normalization)將函數值的值域對應到新的值域[0, +1]，0 表示無關，+1 表示完全相關。

表三: 本研究採用的 F_RCorr 排序相關係數函數

ID	排序相關係數函數	函數值域	正規化方法 [7]	正規化後的值域

[7] 正規化方法捨棄反相關，用意是為確保 Eq. (6)中特徵關聯矩陣 C 的每個元素值大於或等於零，避免在特徵向量二階表徵轉換過程產生偏差。事實上我們曾經嘗試將完全反相關正規化為0.0，無關正規化

AP	APCorrelation [49]	[–1, +1] where –1: perfect inversion 0 : no association +1: perfect agreement	Set values in [-1, 0) to 0	[0, +1] where 0 : no association +1: perfect agreement
G	Goodman and Kruskal's Gamma (*G*) [15]			
K1	Kendall's *tau* (τ) [20]			
P	Pearson's *r* [32]			
SD	Somer's *d* [37]			
SR	Spearman's *rho* (ρ) [38]			
K2	Kendall tau distance [20]	[0, +1] where 0: perfect agreement +1: perfect inversion	1. Set value *x* to a new value 1 – *x* 2. Set values in [0, 0.5) to 0 3. Normalize values in [0.5, 1] to [0, 1] by min-max normalization [17]	[0, +1] where 0 : no association +1: perfect agreement

實作時觀察到在相同 *F_RCorr* 排序相關係數函數的條件下，以不同查詢計算特徵 f_i 與特徵 f_j 的排序關聯程度，其值的分佈不集中且差異大，可能導致 $c_{i,j}$ 失真。為了克服這個問題，本文利用盒鬚圖(Box-and-Whisker Plot; Boxplot) [41]進行異常值偵測(Outlier Detection)，藉此過濾可能的異常值。首先將所有查詢的排序關聯程度由低到高排序，接著選取排在第 25%位置的值為 Q_1、排在第 75%位置的值為 Q_3 計算 IQR = $Q_3 - Q_1$，最後保留位於[Q_1 – 1.5 × IQR, Q_3 + 1.5 × IQR]區間的值計算平均值 $c_{i,j}$。

(3) 特徵向量二階表徵轉換

Eq. (1)定義查詢與文件對組(q_i, $d_{i,j}$)的一階表徵特徵向量 $\vec{d}_{i,j} = < w_{i,j,1}, ..., w_{i,j,k} >$。$\vec{d}_{i,j}$ 的*二階表徵特徵向量*，定義如下：

$$d_{i,j}^{(2)} = < w_{i,j,1}^{(2)}, ..., w_{i,j,k}^{(2)} > \quad (9)$$

其中，$w_{i,j,k}^{(2)}$ 是 $w_{i,j,k}$ 藉由 Eq. (6)的特徵關聯矩陣 *C* 作中介轉換得到的二階表徵特徵值。表四整列本研究提出的特徵向量二階表徵轉換方法。

表四: 本研究提出的特徵向量二階表徵轉換方法

ID	轉換方法	轉換公式	說明
D	向量內積轉換 (Dot-Product)	$w^{(2)}{}_{i,j,k} = \vec{d}_{i,j} \cdot \vec{C}_k$	將 *C* 的列向量 $\vec{C}_k$ 視為特徵 f_k 的向量，定義 $w^{(2)}{}_{i,j,k}$ 為 $\vec{d}_{i,j}$ 與 $\vec{C}_k$ 的向量內積
L	隱含語義索引轉換 (Latent Semantic Indexing Based)	步驟包含[25][26]： 1. 特徵關聯矩陣奇異質分解 $C = USV^T$	1. 特徵關聯矩陣 *C* 作奇異值分解 (Singular Value Decomposition, SVD) 轉換成 *U*、*S* 與 *V* 的乘積 2. 假設 *C* 的秩值(Rank)為 *p*，維度約化

為 0.5，完全正相關正規化為 1.0，但實驗結果顯示其效果不彰。

		2. 維度約化(Dimension Reduction)[8] $C \approx C_z = U_z S_z V_z^T$ 3. 摺疊(Folding-in) $d_{i,j}^{(2)} = d_{i,j}^T U_z S_z^{-1}$	選定一整數 z ($z < p$)，保留 U 與 V 的前 z 個行向量及 S 的對角線{s_1, …, s_z}，得到一個近似於 C 且維度 $z \times z$ 的新矩陣 C_z 3. 將一階表徵特徵向量 $\vec{d}_{i,j}$ 摺疊到維度 $z \times z$ 的隱性關聯空間 S_z 得到二階表徵特徵向量 $d_{i,j}^{(2)}$

3. 排序模型建構學習

圖二使用二階表徵特徵向量作為資料表示，整合RankSVM [19]進行排序模型學習。RankSVM將任兩個查詢與文件對組的二階表徵特徵向量轉換成一個新的物件對組向量，並以物件對組(即，任兩個查詢與文件對組的對組)為單位，使用支援向量機(Support Vector Machine; SVM)得到二元分類器，藉此將物件對組分類為正確或錯誤排序。本研究使用程式工具SVM^{rank}進行實作 [9]，產出的排序模型為一線性函式(Linear Function)形式的分類器，用來判別兩兩文件是否為正確排序，進而推導建立所有文件的排序序列。

特別說明的是，本文選擇 RankSVM 的原因在於過去文獻已驗證其可行性及成效，是一個具代表性且有參考價值的基準。但，本文提出的應用排序相關係數進行特徵向量轉換之資訊檢索排序模型學習架構，其核心演算法並非只限定 RankSVM，也可用各種基於特徵向量表示樣本資料的學習演算法替代。

4. 排序預測與效能評估

訓練學習得到的排序模型被認為可輸出正確的相關性標記，若將具相同性質的新資料輸入此模型，即預測得到一文件序列。換句話說，考慮新查詢 q_{m+1} 與文件集合 $D(q_{m+1}) = \{d_{m+1,1}, \ldots, d_{m+1,n_{m+1}}\}$，集合{ $f(q_{m+1}, d_{m+1,1}), \ldots, f(q_{m+1}, d_{m+1,n_{m+1}})$}是排序模型 f 的標記輸出，藉此將 $D(q_{m+1})$的文件排序作為 q_{m+1} 的查詢結果，此一過程即所謂的排序預測。

排序模型學習過程通常導入特定的評估指標以最佳化求得適當的排序模型。同樣地，排序預測結果也須經由評估過程檢驗排序模型的效能。考量 MAP [2]和 NDCG [18]是資訊檢索常用的評估指標，再加上文獻中有許多研究數據可供比較，因此，本文採用 MAP 與 MeanNDCG (即，所有查詢的 NDCG 平均值)作為排序模型的效能評估指標。

四、 實驗結果

以下依序介紹實驗使用的測試資料、實驗設定和評估結果。

1. LETOR 資料集

微軟釋出的資料集LETOR 4.0[10]，包含MQ2007 和MQ2008 兩個子集，提供資訊檢

[8] 因運算成本考量，通常不會實際計算 C_z 矩陣，而是使用 U_z 與 S_z 矩陣將一階表徵特徵向量摺疊到隱性關聯空間 S_z。

[9] http://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html.

[10] http://research.microsoft.com/en-us/um/beijing/projects/letor/.

索排序模型學習研究在相同評比資料及評估指標下進行比較。本研究使用的資料集是MQ2008 子集 [11]，該子集約有 800 個查詢，共 15,211 個查詢與文件對組。每個查詢與文件對組用 46 個特徵表示為向量且已標註相關性(特徵清單詳列於表五；表六提供查詢與文件對組的資料格式範例，每列為一查詢與文件對組，由相關性、查詢代碼、特徵向量表示式及文件代碼構成)。MQ2008 事先被切割成 5 個集合，以進行 5-Fold交叉驗證(Cross Validation)。每個集合有：(1) Training Dataset用來訓練學習，平均有 9,127 個對組、(2) Validation Dataset用來調整模型參數或避免過適學習，平均有 3,042 個對組，及(3) Test Dataset用來評估排序模型的效能，平均有 3,042 個對組。LETOR同時提供MAP和MeanNDCG評估工具及多個評比基準(Baselines)的結果供比較。

由表五得知，46 個特徵主要由 TF (Term Frequency)、IDF (Inverse Document Frequency)、TF*IDF、DL (Document Length)、BM25 [34]和 LMIR [53]組成，因分析文件範圍差異而衍生出不同形式的特徵。以特徵相關性來說，舉例而言，TF*IDF of whole document (ID: 15)和 BM25 of whole document (ID: 25)都是 TF 與 IDF 構成，此兩特徵間存在相關性，且與 TF 和 IDF 存在相依性。另外，分析 MQ2008 的 Fold 1 子集合發現，訓練資料中共有 9,630 筆查詢與文件對組，以 46 個特徵計算共有 442,980 (9630 × 46 = 442980)個特徵值，其中特徵值為零所占的比例是 47.23%。同樣的現象在其他子集合中也可觀察得到(Fold 2 約略 47.16%、Fold 3 約略 47.07%、Fold 4 約略 47.41%、Fold 5 約略 46.91%)，平均而言每個子集合特徵向量約有 47.16%的特徵值為零，即存在資料稀疏性。綜合前述兩項觀察結果，MQ2008 資料集符合本文所提方法的特性。

表五: MQ2008 用來表示查詢與文件對組向量的 46 個特徵

1: TF (Term frequency) of body	13: TF*IDF of title	25: BM25 of whole document	37: LMIR.JM of anchor
2: TF of anchor	14: TF*IDF of URL	26: LMIR.ABS of body	38: LMIR.JM of title
3: TF of title	15: TF*IDF of whole document	27: LMIR.ABS of anchor	39: LMIR.JM of URL
4: TF of URL	16: DL (Document length) of body	28: LMIR.ABS of title	40: LMIR.JM of whole document
5: TF of whole document	17: DL of anchor	29: LMIR.ABS of URL	41: PageRank
6: IDF (Inverse document frequency) of body	18: DL of title	30: LMIR.ABS of whole document	42: Inlink number
7: IDF of anchor	19: DL of URL	31: LMIR.DIR of body	43: Outlink number
8: IDF of title	20: DL of whole document	32: LMIR.DIR of anchor	44: Number of slash in URL
9: IDF of URL	21: BM25 of body	33: LMIR.DIR of title	45: Length of URL
10: IDF of whole document	22: BM25 of anchor	34: LMIR.DIR of URL	46: Number of child page
11: TF*IDF of body	23: BM25 of title	35: LMIR.DIR of whole document	
12: TF*IDF of anchor	24: BM25 of URL	36: LMIR.JM of body	

[11] 前置實驗分析發現 MQ2007 中所有查詢的個別特徵關聯數值分佈較分散、差異大，雖已導入異常值過濾機制，但仍導致整體性特徵關聯可能失真，使得轉換後的二階表徵特徵向量不具意義。而 MQ2008 中所有查詢的特徵關聯數值分佈較集中，較符合本文所提方法的特性。因此，本文實驗中僅使用 MQ2008 子集來驗證所提方法的優劣。

表六: 查詢與文件對組的資料格式範例

Relevance Label	Query	f_1	f_2	…	f_{46}	DocID
2 (definitely relevant)	qid:10032	1:0.056537	2:0.000000	…	46:0.076923	#docid: GX029-35-5894638
0 (not relevant)	qid:10032	1:0.279152	2:0.000000	…	46:1.000000	#docid: GX030-77-6315042
0 (not relevant)	qid:10032	1:0.130742	2:0.000000	…	46:1.000000	#docid: GX140-98-13566007
1 (possibly relevant)	qid:10032	1:0.593640	2:1.000000	…	46:0.000000	#docid: GX256-43-0740276

2. 實驗設定

(1) 資料設定

表七說明實驗使用的資料設定。E1 使用 MQ2007 建構特徵關聯矩陣，訓練、驗證及測試資料使用 MQ2008，目的是評估以不同性質資料集得到的特徵關聯對於排序模型學習的影響；E2 使用 MQ2008 的 Training Dataset 建構特徵關聯矩陣，訓練、驗證及測試資料也使用 MQ2008，目的是評估以相同性質資料集得到的特徵關聯對於排序模型學習的影響。

表七：實驗資料設定

ID	特徵關聯矩陣	訓練資料	驗證資料	測試資料
E1	MQ2007 (All)	MQ2008 (Training)	MQ2008 (Validation)	MQ2008 (Test)
E2	MQ2008 (Training)			

(2) 參數設定

參數設定包括：(1) 隱含語義索引轉換之維度約化 z 值，及(2) SVM^{rank} 的參數。本文提出以維度約化比例設定 z 值。舉例來說，假設特徵關聯矩陣 C 的秩值為 p、維度約化比例為 20%，z 值設定為 $0.2 \times p$。實驗中列舉可能的維度約化比例{10%, 20%, …, 90%}，利用驗證資料(Validation Dataset)求得最佳 MAP 時的維度約化比例，並以該值為測試時設定 z 值的依據。SVM^{rank} 依循 LETOR 規範設定參數為「-c <C> -e 0.001 -l l」，假設<C>的可能值{0.00001, 0.00002, 0.00005, 0.0001, 0.0002, 0.0005, 0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10}，同樣使用驗證資料求得最佳 MAP 時的<C>值，並以該值作為測試時的設定。

3. 評估結果

排序模型的效能評估指標採用 MAP 與 MeanNDCG (即，所有查詢的 NDCG 平均值)，且數值越高者表示其效能較佳。考量過去文獻多以 MAP 的比較為主，本節中所列實驗數據僅針對 MAP 進行討論，但仍列出 MeanNDCG 供比較。

實驗結果整理於表八和表九。表中，RankSVM是評比基準，餘為本文方法在不同設定的組合，以E{1, 2}-{AP, G, K1, P, SD, SR, K2}-{D, L}識別。E{1, 2}表示表七的 2 種實驗資料設定，{AP, G, K1, P, SD, SR, K2}對應表三的 7 種排序相關係數函數，{D, L}代表表四的 2 種特徵向量二階表徵轉換方法。括號中的百分比數是該方法的評估結果

相對於RankSVM的提升比例 [12]。除此之外，L.UB是隱含語義索引轉換對於測試資料集進行最佳化設定 [13]的評估結果，用來當作L設定下的上限值(Upper Bound)參考。

表八顯示在固定排序相關係數函數的條件下，對於 AP、G、K1、P、SD、SR 及 K2 來說，最佳組合依序是：E1-AP-L、E1-G-D、E1-K1-L、E1-P-D、E1-SD-D、E1-SR-L 及 E1-K2-D，這些組合都優於 RankSVM，提升比例是 0.75%、0.69%、0.39%、0.32%、0.56%、1.64%及 1.04%。在固定特徵向量二階轉換的條件下，對於 D 和 L 來說，最佳組合依序是：E1-K2-D 和 E1-SR-L，兩者都優於 RankSVM，提升比例是 1.04%和 1.64%。當考慮 D 和 L 時，14 組設定(即，E1-{*}-{D, L})僅 E1-P-L 比 RankSVM 差，其餘都比 RankSVM 優。最後，考慮 L.UB 時，7 組設定(即，E1-{*}-L.UB)都比在 D 的條件下表現佳，且各種組合的表現都高於 RankSVM，提升比例是：E1-AP-L.UB (+1.49%)、E1-G-L.UB (+1.78%)、E1-K1-L.UB (+1.40%)、E1-P-L.UB (+1.45%)、E1-SD-L.UB (+0.96%)、E1-SR-L.UB (2.12%)和 E1-K2-L.UB (+1.59%)。

表九顯示在固定排序相關係數函數的條件下，對於 AP、G、K1、P、SD、SR 及 K2 來說，最佳組合依序是：E2-AP-D、E2-G-D、E2-K1-L、E2-P-L、E2-SD-L、E2-SR-L 及 E2-K2-L，這些組合都優於 RankSVM，提升比例是 1.05%、0.61%、0.91%、0.20%、0.35%、0.82%及 1.22%。在固定特徵向量二階轉換的條件下，對於 D 和 L 來說，最佳組合依序是：E2-AP-D 和 E2-K2-L，相較於 RankSVM 的提升比例是 1.05%和 1.22%。當考慮 D 和 L 時，14 組設定(即，E2-{*}-{D, L})除了 E2-P-D 外，其餘都優於 RankSVM。最後，考慮 L.UB 時，7 組設定(即，E2-{*}-L.UB)都比在 D 的條件下表現佳，所有組合都優於 RankSVM，提升比例是：E2-AP-L.UB (+1.66%)、E2-G-L.UB (+1.74%)、E2-K1-L.UB (+1.47%)、E2-P-L.UB (+1.35%)、E2-SD-L.UB (+1.46%)、E2-SR-L.UB (1.77%)和 E2-K2-L.UB (+2.49%)。

表八：E1 設定下 MQ2008 的評估結果

Model	MAP	MeanNDCG	Model	MAP	MeanNDCG
RankSVM	0.4696	0.4832	E1-P-L	0.4692 (-0.08%)	0.4798 (-0.71%)
E1-AP-D	0.4722 (+0.55%)	0.4855 (+0.47%)	E1-P-L.UB	**0.4764 (+1.45%)**	0.4852 (+0.41%)
E1-AP-L	0.4731 (+0.75%)	0.4871 (+0.81%)	E1-SD-D	0.4722 (+0.56%)	0.4843 (+0.23%)
E1-AP-L.UB	**0.4766 (+1.49%)**	0.4878 (+0.96%)	E1-SD-L	0.4719 (+0.50%)	0.4857 (+0.51%)
E1-G-D	0.4728 (+0.69%)	0.4873 (+0.85%)	E1-SD-L.UB	0.4741 (+0.96%)	0.4862 (+0.61%)
E1-G-L	0.4710 (+0.29%)	0.4836 (+0.08%)	E1-SR-D	0.4714 (+0.37%)	0.4853 (+0.43%)
E1-G-L.UB	**0.4780 (+1.78%)**	**0.4888 (+1.16%)**	E1-SR-L	**0.4773 (+1.64%)**	0.4863 (+0.64%)
E1-K1-D	0.4698 (+0.03%)	0.4839 (+0.14%)	E1-SR-L.UB	**0.4796 (+2.12%)**	0.4842 (+0.20%)
E1-K1-L	0.4714 (+0.39%)	0.4839 (+0.15%)	E1-K2-D	**0.4745 (+1.04%)**	0.4850 (+0.38%)
E1-K1-L.UB	**0.4762 (+1.40%)**	**0.4887 (+1.13%)**	E1-K2-L	0.4710 (+0.30%)	0.4839 (+0.14%)
E1-P-D	0.4711 (+0.32%)	0.4807 (-0.52%)	E1-K2-L.UB	**0.4771 (+1.59%)**	0.4875 (+0.88%)

表九: E2 設定下 MQ2008 的評估結果

Model	MAP	MeanNDCG	Model	MAP	MeanNDCG
RankSVM	0.4696	0.4832	E2-P-L	0.4705 (+0.20%)	0.4839 (+0.14%)

[12] b 比 a 的相對提升比例，計算方式：(b – a) / a × 100%。

[13] 實驗測試不同參數 *z* 與<C>的設定，由前一節「實驗設定」描述知道，可能的 *z* 值有 9 個、<C>值有 19 個，共 171 種組合，從中挑選出擁有最佳 MAP 的組合作為最佳化設定。

E2-AP-D	**0.4745 (+1.05%)**	0.4863 (+0.65%)	E2-P-L.UB	**0.4760 (+1.35%)**	0.4868 (+0.75%)
E2-AP-L	0.4728 (+0.67%)	**0.4886 (+1.12%)**	E2-SD-D	0.4709 (+0.28%)	0.4855 (+0.48%)
E2-AP-L.UB	**0.4774 (+1.66%)**	0.4875 (+0.89%)	E2-SD-L	0.4713 (+0.35%)	0.4868 (+0.75%)
E2-G-D	0.4725 (+0.61%)	0.4850 (+0.37%)	E2-SD-L.UB	**0.4765 (+1.46%)**	0.4867 (+0.73%)
E2-G-L	0.4707 (+0.24%)	0.4855 (+0.48%)	E2-SR-D	0.4729 (+0.70%)	0.4858 (+0.53%)
E2-G-L.UB	**0.4778 (+1.74%)**	**0.4895 (+1.31%)**	E2-SR-L	0.4734 (+0.82%)	0.4858 (+0.54%)
E2-K1-D	0.4715 (+0.40%)	0.4851 (+0.40%)	E2-SR-L.UB	**0.4779 (+1.77%)**	0.4876 (+0.92%)
E2-K1-L	0.4739 (+0.91%)	0.4866 (+0.71%)	E2-K2-D	0.4721 (+0.54%)	0.4864 (+0.67%)
E2-K1-L.UB	**0.4765 (+1.47%)**	**0.4882 (+1.04%)**	E2-K2-L	**0.4753 (+1.22%)**	0.4866 (+0.70%)
E2-P-D	0.4675 (-0.46%)	0.4837 (+0.10%)	E2-K2-L.UB	**0.4813 (+2.49%)**	**0.4903 (+1.46%)**

整體而言，本文方法的各種組合相較於 RankSVM 來說都具有效能提升，此點說明本文所提應用基於排序相關係數之特徵向量轉換於資訊檢索排序模型學習方法的可行性。再者，表八和表九是使用 MQ2007 與 MQ2008 (Training Dataset)計算特徵關聯的結果。比較在多數設定下，後者的結果普遍比前者佳。由此可知，特徵關聯計算的品質直接影響到排序模型學習的效能，且使用相同性質的資料集計算特徵關聯有較佳的表現。

五、 討論

本文提出的「應用基於排序相關係數之特徵向量轉換於資訊檢索排序模型學習」方法具有以下特性：(1) 此方法屬於監督式機器學習，具有自動設定最佳函式參數值、有效結合單一排序模型成為整體最適的排序模型，以及利用規則化法則避免模型過適的優勢[23]；(2) 方法設計已模組化，具有彈性可依需要進行個別模組的替換與修正。例如：導入更多類型的排序相關係數函數計算特徵關聯、使用不同特徵定義特徵值與特徵向量、整合各種基於特徵向量表示樣本資料的學習演算法；(3) 隱含語義索引轉換的維度約化將 k 個特徵對應到 z 個群組，以達到資料平滑化(Data Smoothing)和隱性關聯萃取的目的；(4) 本文方法的限制在於特徵關聯計算不夠精準時，可能導致轉換後的二階表徵特徵向量不具意義，直接影響到排序模型學習效能。

本研究的貢獻，包括：(1) 利用排序相關係數計算兩兩特徵的關聯強度，而特徵關聯矩陣的列向量(或行向量)同時定義單一特徵的向量表示，可用來分析不同特徵在排序序列的特性、基於向量相似度設計特徵分群原則，或者發展特徵選取(Feature Selection)技術；(2) 基於特徵關聯矩陣提出的二階表徵映射轉換，可移植於資訊檢索或機器學習領域中同樣以向量表示為核心的演算法，藉此解決向量表示稀疏性及假設特徵維度獨立造成失真的問題；(3) 以特徵向量二階表徵為基礎，整合 RankSVM [19]提出新的資訊檢索排序模型學習架構(詳見圖二)；(4) 使用 LETOR 4.0 資料集驗證本文方法的可行性及成效，相關結果可供其他研究參考。

六、 結論

本文基於排序相關係數計算不同特徵在排序特性的關聯強度，並藉由特徵的排序關聯作中介轉換，將特徵向量映射到特徵關聯空間，使得特徵向量由一階表徵提升到二階表徵，藉此填補原始稀疏的特徵值並改善因假設特徵獨立無關導致向量描述可能失真的問題。同時，以特徵向量二階表徵轉換為基礎，擴展圖一應用機器學習於資訊

檢索排序模型建構的架構，提出圖二應用基於排序相關係數之特徵向量轉換於資訊檢索排序模型學習的方法。實作時採用 7 種排序相關係數：(1) APCorrelation、(2) Goodman and Kruskal's Gamma (G)、(3) Kendall's *tau* (τ)、(4) Pearson's r、(5) Somer's d、(6) Spearman's *rho* (ρ)，及(7) Kendall tau distance。另提出 2 種特徵向量二階表徵轉換方法：(1) 向量內積轉換，與(2) 隱含語義索引轉換。最後，使用二階表徵特徵向量作為資料表示，整合 RankSVM 進行排序模型學習，產出的排序模型為一線性函式形式的二元分類器，用來判別兩兩文件是否為正確排序，進而推導建立所有文件的排序序列。

實驗使用 LETOR 4.0 資料集驗證本文方法的可行性及成效，評估指標採用 MAP 和 MeanNDCG，並以 RankSVM 為基準，比較二階表徵特徵向量對於排序模型學習的影響。評估結果顯示：(1) 在 MQ2008 的測試資料上，本文方法的各種組合相較於 RankSVM 都有效能提升，說明本文所提應用基於排序相關係數之特徵向量轉換於資訊檢索排序模型學習方法的可行性；(2) 特徵關聯的品質直接影響到排序模型學習的效能，且使用相同性質的資料集計算特徵關聯有較佳的表現。

未來可能的研究方向有：(1) 導入更多類型的排序相關係數函數計算特徵關聯、使用不同特徵定義特徵值與特徵向量，以及整合各種基於特徵向量表示樣本資料的學習演算法，以驗證特徵向量二階表徵轉換在不同設定下的成效及建議適當組合；(2) 利用特徵關聯矩陣的列向量(或行向量)作為單一特徵的向量表示，以分析不同特徵在排序序列的特性、基於向量相似度設計特徵分群原則，或者進行特徵選取技術的研發。

[謝啟]

本研究由科技部專題研究計畫（MOST102-2218-E-178-001: 基於排序相關係數進行特徵向量二階表徵轉換之資訊檢索排序模型建構研究）提供經費補助，特此致謝。

參考文獻

[1] 游斯涵，使用機器學習方法於語音文件檢索之研究，碩士論文，臺北：國立臺灣師範大學資訊工程研究所，2009。

[2] Baeza-Yates, R. and Ribeiro-Neto, B., *Modern information retrieval*, Addison-Wesley, New York, NY, 1999.

[3] Bian, J., Liu, T.-Y., Qin, T. and Zha, H., "Ranking with query-dependent loss for web search", *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining (WSDM 2010)*, 2010: pp. 141-150.

[4] Brin, S. and Page, L., "The anatomy of a large-scale hypertextual Web search engine", *Computer Networks and ISDN Systems* (30:1-7), 1998: pp. 107-117.

[5] Burges, C.J., Ragno, R. and Le, Q.V., "Learning to rank with nonsmooth cost functions", in Schölkopf, B., Platt, J.C. and Hoffman, T. (Eds.), *Advances in Neural Information Processing Systems (Vol. 19)*, The MIT Press, Cambridge, MA, 2007: pp. 193-200.

[6] Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N. and Hullender, G., "Learning to rank using gradient descent", *Proceedings of the 22nd International Conference on Machine Learning (ICML 2005)*, 2005: pp. 89-96.

[7] Cao, Z., Qin, T., Liu, T.-Y., Tsai, M.-F. and Li, H., "Learning to rank: from pairwise approach to listwise approach", *Proceedings of the 24th International Conference on Machine Learning (ICML 2007)*, 2007: pp. 129-136.

[8] Cao, Y., Xu, J., Liu, T.-Y., Li, H., Huang, Y. and Hon, H.-W., "Adapting ranking SVM to

document retrieval", *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006)*, 2006: pp. 186-193.

[9] Crammer, K. and Singer, Y., "Pranking with ranking", in Dietterich, T.G., Becker, S. and Ghahramani, Z. (Eds.), *Advances in Neural Information Processing Systems (Vol. 14)*, The MIT Press, Cambridge, MA, 2002: pp. 641-647.

[10] Fan, W.G., Fox, E.A., Pathak, P. and Wu, H., "The effects of fitness functions on genetic programming-based ranking discovery for web search", *Journal of the American Society for Information Science and Technology* (55:7), 2004: pp. 628-636.

[11] Fan, W.G., Gordon, M.D. and Pathak, P., "A generic ranking function discovery framework by genetic programming for information retrieval", *Information Processing & Management* (40:4), 2004: pp. 587-602.

[12] Fan, W.G., Gordon, M.D. and Pathak, P., "Discovery of context-specific ranking functions for effective information retrieval using genetic programming", *IEEE Transactions on Knowledge and Data Engineering* (16:4), 2004: pp. 523-527.

[13] Freund, Y., Iyer, R., Schapire, R.E. and Singer, Y., "An efficient boosting algorithm for combining preferences", *Journal of Machine Learning Research* (4:6), 2004: pp. 933-969.

[14] Geng, X., Liu, T.-Y., Qin, T. and Li, H., "Feature selection for ranking", *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2007)*, 2007: pp. 407-414.

[15] Goodman, L.A. and Kruskal, W.H., "Measures of association for cross classification", *Journal of the American Statistical Association* (49:268), 1972: pp. 732-764.

[16] Gyöngyi, Z., Garcia-Molina, H. and Pedersen, J., "Combating web spam with TrustRank", *Proceedings of the 30th International Conference on Very Large Data Bases (VLDB 2004)*, 2004: pp. 576-587.

[17] Han, J., Kamber, M. and Pei, J., *Data mining: concepts and techniques (3rd edition)*, Morgan Kaufmann, Waltham, MA, 2011.

[18] Jarvelin, K. and Kekalainen, J., "Cumulated gain-based evaluation of IR techniques", *ACM Transactions on Information Systems* (20:4), 2002: pp. 422-446.

[19] Joachims, T., "Optimizing search engines using clickthrough data", *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2002)*, 2002: pp. 133-142.

[20] Kendall, M.G., "A new measure of rank correlation", *Biometrika* (30:1/2), 1938: pp. 81-93.

[21] Li, P., Wu, Q. and Burges, C.J., "McRank: learning to rank using multiple classification and gradient boosting", in Platt, J.C., Koller, D., Singer, Y. and Roweis, S.T. (Eds.), *Advances in Neural Information Processing Systems (Vol. 20)*, Curran Associates, Red Hook, NY, 2008: pp. 897-904.

[22] Lin, J.-Y., Yeh, J.-Y. and Liu, C.-C., "Learning to rank for information retrieval using layered multi-population genetic programming", *Proceedings of the 2012 IEEE International Conference on Computational Intelligence and Cybernetics (CyberneticsCom 2012)*, 2012: pp. 45-49.

[23] Liu, T.-Y., *Learning to rank for information retrieval*, Springer, Heidelberg, Germany, 2011.

[24] Liu, Y., Gao, B., Liu, T.-Y., Zhang, Y., Ma, Z., He, S. and Li, H., "BrowseRank: letting web users vote for page importance", *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008)*, 2008: pp. 451-458.

[25] Manning, C.D., Raghavan, P. and Schütze, H., *Introduction to information retrieval*, Cambridge University Press, New York, NY, 2008.

[26] Manning, C.D. and Schütze, H., *Foundations of Statistical Natural Language Processing*, The

MIT Press, Cambridge, MA, 1999.

[27] Matveeva, I., Burges, C., Burkard, T., Laucius, A. and Wong, L., "High accuracy retrieval with multiple nested ranker", *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006)*, 2006: pp. 437-444.

[28] Mitchell, T.M., *Machine learning*, McGraw-Hill, 1997.

[29] Nallapati, R., "Discriminative models for information retrieval", *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004)*, 2004: pp. 64-71.

[30] Nie, L., Davison, B.D. and Qi, X., "Topical link analysis for web search", *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006)*, 2006: pp. 91-98.

[31] Pan, Z.B., You, X.G., Chen, H., Tao, D.C. and Pang, B.C., "Generalization performance of magnitude-preserving semi-supervised ranking with graph-based regularization", *Information Sciences* (221), 2013: pp. 284-296.

[32] Pearson, K., "Note on regression and inheritance in the case of two parents", *Proceedings of the Royal Society of London*, 1895: 240-242.

[33] Qin, T., Liu, T.-Y., Zhang, X.-D., Chen, Z. and Ma, W.-Y., "A study of relevance propagation for web search", *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005)*, 2005: pp. 408-415.

[34] Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M.M. and Gatford, M., "Okapi at TREC-3", *Proceedings of the 3rd Text Retrieval Conference (TREC-3)*, 1995: pp. 109-127.

[35] Scheel, C., Neubauer, N., Lommatzsch, A., Obermayer, K. and Albayrak, S., "Efficient query delegation by detecting redundant retrieval strategies", *Proceedings of the SIGIR 2007 Workshop on Learning to Rank for Information Retrieval (LR4IR 2007)*, 2007.

[36] Shakery, A. and Zhai, C.X., "Relevance propagation for topic distillation UIUC TREC 2003 web track experiments", *Proceedings of the 12th Text REtrieval Conference (TREC 2003)*, 2003: pp. 673-677.

[37] Somers, R.H., "A new asymmetric measure of association for ordinal variables", *American Sociological Review* (27:6), 1962: pp. 799-811.

[38] Spearman, C., "The proof and measurement of association between two things", *American Journal of Psychology* (15:1), 1904: pp. 72-101.

[39] Szummer, M. and Yilmaz, E., "Semi-supervised learning to rank with preference regularization", *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM 2011)*, 2011: pp. 269-278.

[40] Tsai, M.-F., Liu, T.-Y., Qin, T., Chen, H.-H. and Ma, W.-Y., "FRank: a ranking method with fidelity loss", *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2007)*, 2007: pp. 383-390.

[41] Tukey, J.W., *Exploratory data analysis*, Pearson, 1977.

[42] Wang, L., Lin, J. and Metzler, D., "A cascade ranking model for efficient ranked retrieval", *Proceedings of the 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2011)*, 2011: pp. 105-114.

[43] Xia, F., Liu, T.-Y. and Li, H., "Statistical consistency of top-k ranking", in Bengio, Y., Schuurmans, D., Lafferty, J.D., Williams, C.K.I. and Culotta, A. (Eds.), *Advances in Neural Information Processing Systems (Vol. 22)*, Curran Associates, Red Hook, NY, 2009: pp. 2098-2106.

[44] Xu, J., Cao, Y., Li, H. and Huang, Y., "Cost-sensitive learning of SVM for ranking", *Proceedings of the 17th European Conference on Machine Learning (ECML 2006)*, 2006: pp. 833-840.

[45] Xu, J., Cao, Y., Li, H. and Zhao, M., "Ranking definitions with supervised learning methods", *Proceedings of the 14th International Conference on World Wide Web (WWW 2005)*, 2005: pp. 811-819.

[46] Xu, J. and Li, H., "AdaRank: a boosting algorithm for information retrieval", *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2007)*, 2007: pp. 391-398.

[47] Xue, G.-R., Yang, Q., Zeng, H.-J., Yu, Y. and Chen, Z., "Exploiting the hierarchical structure for link analysis", *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005)*, 2005: pp. 186-193.

[48] Yeh, J.-Y., Lin, J.-Y., Ke, H.-R. and Yang, W.-P., "Learning to rank for information retrieval using genetic programming", *Proceedings of the SIGIR 2007 Workshop on Learning to Rank for Information Retrieval (LR4IR 2007)*, 2007: pp. 41-48.

[49] Yilmaz, E., Aslam, J.A. and Robertson, S., "A new rank correlation coefficient for information retrieval", *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008)*, 2008: pp. 587-594.

[50] Yu, H., "SVM selective sampling for ranking with application to data retrieval", *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining (SIGKDD 2005)*, 2005: pp. 354-363.

[51] Yu, H., Kim, J., Kim, Y., Hwang, S. and Lee, Y.H., "An efficient method for learning nonlinear ranking SVM functions", *Information Sciences* (209), 2012: pp. 37-48.

[52] Yue, Y., Finley, T., Radlinski, F. and Joachims, T., "A support vector method for optimizing average precision", *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2007)*, 2007: pp. 217-278.

[53] Zhai, C. X. and Lafferty, J., "A study of smoothing methods for language models applied to Ad Hoc information retrieval", *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001)*, 2001: pp. 334-342.

[54] Zheng, Z., Zha, H., Zhang, T., Chapelle, O., Chen, K. and Sun, G., "A general boosting method and its application to learning ranking functions for Web search", in Platt, J.C., Koller, D., Singer, Y. and Roweis, S.T. (Eds.), *Advances in Neural Information Processing Systems (Vol. 20)*, Curran Associates, Red Hook, NY, 2007: pp. 1697-1704.

[55] Zhou, K., Xue, G.-R., Zha, H. and Yu, Y., "Learning to rank with ties", *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008)*, 2008: pp. 275-282.