教育資料與圖書館學

Journal of Educational Media & Library Sciences

http://joemls.tku.edu.tw

Vol. 45 , no. 4 (Summer 2008) : 409-431

應用資訊檢索技術於論文評閱者推薦模式之評估

Evaluation of Information Retrieval Based Models for Recommendation of Paper Reviewers

魏世杰 Shih-Chieh Wei
Assistant Professor
E-mail: seke@mail.im.tku.edu.tw

羅欣瑜 Hsin-Yu Luo
Graduate Student
Department of Information Management, Tamkang University
Taipei, Taiwan

English Abstract & Summary see link at the end of this article

教育資料與圖書館學　45 : 4 (Summer 2008) : 409-431

應用資訊檢索技術於論文評閱者推薦模式之評估

魏世杰*

助理教授
淡江大學資訊管理學系
E-mail: seke@mail.im.tku.edu.tw

羅欣瑜

研究生
淡江大學資訊管理學系

摘要

隨著傳統期刊逐漸採用電子形式出刊，也帶動投稿及評閱過程愈來愈多採用電子自動化之潮流。目前一般的線上投稿暨評閱系統雖然功能逐漸齊備，但仍少有推薦評閱者功能。為評估現有可推薦評閱者技術之表現，本文分別用標題、關鍵詞、摘要、全文4種不同長度的論文表示方式，搭配7種評閱者匹配法，其中包括向量空間模式下的4種相似度匹配法，及應用於OpenConf線上投稿系統中的3種主題式匹配法，交叉組合出4×7＝28種推薦模式。測試結果顯示，向量空間模式匹配法優於主題式匹配法。又所有推薦模式中，以摘要為論文表示方式，搭配向量空間模式的餘弦相似度匹配法，其推薦效果最好。

關鍵詞：向量空間模式，論文表示方式，評閱者推薦模式

前　言

隨著時代潮流的演進，愈來愈多學術期刊逐漸從提供查詢及下載電子版論文的單向出版模式，邁向連投稿及審查過程也採用網路進行的多向互動模式（邱炯友，2003；Ware, 2005）。為滿足市場需求，已有愈來愈多的線上投稿暨評閱系統應運而生，以提供這樣一個方便可靠的自動化平台，例如國外

* 本文通訊作者

2008/04/03投稿; 2008/07/01修訂; 2008/07/04接受

ScholarOne公司的Manualscript Central系統[1]，及國內華藝公司的Aspers（Airiti Submission and Peer Review System）系統[2]（邱炯友、李怡萍，2006）。

觀察這些線上投稿暨評閱系統的功能，一般包含線上投稿、進度查詢、同儕評閱（peer review）、資料報表產製、授權書上傳，甚至作者付費機制等（顏玉茵，2004）。其相較於傳統人工作業，固然能省下不少時間、人力，及金錢，唯美中不足處是其中有關同儕評閱部分，目前仍多停留在傳統人工指派模式。這種由期刊主編根據論文性質手動指派適當評閱者的方式，若無額外輔助，對主編而言仍形成一項負擔。

在線上投稿暨評閱系統上，由於投稿採電子形式，適合要求作者輸入論文標題、摘要、關鍵詞，或直接擷取內文，以代表論文。至於評閱者資料方面，若能要求領域專家輸入其專長，或由其過去發表文獻，自動擷取標題、摘要、關鍵詞，或內文，以代表專長，則一套由論文能自動推薦評閱者排名的機制，應可大幅減輕主編找尋適當評閱者之時間。

另外，過去介紹線上投稿暨審查系統的文獻中，實際有提到其推薦評閱者方法且開放下載試用的計有CyberChair[3]、MyReview[4]、OpenConf[5]等系統。其中OpenConf系統已成功應用於數百種研討會及期刊的論文評閱作業，故適合作為測試比較之用。關於推薦評閱者方法，OpenConf系統共提供了3種主題式匹配功能，分別為無加權主題匹配法、難指派評閱者優先加權主題匹配法，及易指派評閱者優先加權主題匹配法。

因此，為客觀評估現有期刊論文推薦評閱者技術，本文將在論文採自動萃取主題及主題間無相關性之前提假設下，以常見4種不同長度的論文表示方式，搭配7種評閱者匹配法，其中包括4種向量空間相似度匹配法，及3種OpenConf 系統的主題式匹配法，合計組合出28種推薦模式作比較，俾提供未來線上投稿暨評閱系統開發類似推薦模組之參考。

二、相關文獻

以下將介紹推薦論文評閱者之過去研究，及實際應用於本文測試之論文、評閱者向量表示方式及兩者匹配法。

㈠推薦評閱者之研究

過去有關論文推薦評閱者之研究主要有表示法（representation），匹配度

[1] 網頁參見 http://www.scholarone.com/products_manuscriptcentral_aboutMC.shtml
[2] 網頁參見 http://portal.airiti.com/modules/tinyd0/index.php?id=25
[3] 網頁參見 hhttp://www.cyberchair.org
[4] 網頁參見 hhttp://myreview.intellagence.eu
[5] 網頁參見 hhttp://www.openconf.org

（similarity factor），及指派最佳化（optimization of the assignment problem）三方向。表示法方面多將論文及評閱者表現成由許多主題維度構成的向量，主題來源分為人工自訂及自動萃取兩種，前者簡便但匹配較受限，後者處理耗時但匹配較具彈性。自動萃取論文的主題成分時，有的只靠內文，有的還參考領域分類樹（Biswas & Hasan, 2007）或領域論文集（Mimno & McCallum, 2007）輔助。另外，向量內各主題成分所佔比重則依匹配所需精細程度分成布林值，或連續實數值兩種。特別是評閱者表示法方面，依其來源資料蒐集難易及多寡程度，從簡單的人工勾選布林向量法，到複雜的自動萃取實數向量法皆有。

有了論文及評閱者向量之後，兩者匹配度的計算可分成兩類，一類借用常見的向量相似度算法，假設向量各主題維度彼此獨立。兩向量若相同主題之成分比重愈高，則匹配度愈高，例如常見的Dice係數、Jaccard係數、內積、餘弦等計算法（Salton & Buckley, 1988）。另一類匹配度算法則允許向量各主題維度間具相關性，但採不同因應法，例如隱藏語意索引（latent semantic indexing）、運輸問題（transportation problem）求解法等。其中，隱藏語意索引能自動找出維度獨立的新向量空間，於其上利用傳統向量相似度計算匹配度（Dumais & Nielsen, 1992）。運輸求解法視論文向量各主題為供應站，評閱者向量各主題為需求站，企圖將各供應站之成分比重，適當分配送往高相關的需求站，以求得最大利潤，即匹配度（Hartvigsen & Wei, 1999）。經由這樣轉化成作業研究中載重量受限之運輸最大化問題求解後，兩向量若相關主題之成分比重愈高，則匹配度愈高。另外，也有一類匹配度之協同計算（collaborative computing）是不依賴論文及評閱者表示法的，其作法是由評閱者直接對少量論文評估匹配度，於算出任兩評閱者相關度後，遇評閱者未評估過之論文時，則可依其他相關度高之評閱者對該論文評估之匹配度高低，而匯整推論其匹配度（Rigaux, 2004）。

有了任一組論文及評閱者匹配度之後，指派最佳化考慮如何滿足各種實際負荷限制下，作出總匹配度最高之指派結果。這些負荷限制包括每篇論文最少需要多少評閱者、每位評閱者最多不能看超過多少論文數等。這部分的作法有利用作業研究的Hungarian演算法（Rigaux, 2004）、最大轉運問題法（transshipment problem）（Hartvigsen & Wei, 1999）、詞權重殘差和法（sum of residual term weight）（Hettich & Pazzani, 2006），或其他啟發式（heuristic）指派法（Mauro, Basile, & Ferilli, 2005）等。只是這部分的考量顯然比較適用在大量批次指派需求，例如短期間的研討會或計畫案評閱等。就期刊論文少量多次的指派需求而言，匹配度毋寧是比指派最佳化更宜受重視。

㈡論文及評閱者向量表示方式

傳統資訊檢索領域常以向量形式來記錄文章特徵（Baeza-Yates & Ribeiro-Neto, 1999），文章向量常擁有數百到數千個關鍵詞特徵維度，每一維度記錄文章內某關鍵詞的成分比重，成分比重愈高表示愈能由該關鍵詞找到文章。另外，視匹配所需精細程度，有兩種表現文章維度的成分比重方法，一為較粗糙的離散布林值，表示文章有無該關鍵詞；一為較精細的連續實數值，表示文章內該關鍵詞重要度。在匹配結果需要排名時，一般採用連續實數值法，以精細分出名次，避免相同名次文章太多情形。

在採用連續實數值表現文章關鍵詞成分的方法中，最著名的要屬TFIDF表示法（Salton & McGill, 1989）。此法綜合考量關鍵詞的文章內出現度及文章集罕見度，只有兩者愈高，成分值才高。其中，文章內出現度即詞頻（term frequency），文章集罕見度則由關鍵詞出現文章數之倒數（inverted document frequency）作代表，出現某關鍵詞的文章數愈少，其罕見度愈高。

形式上來說，論文其實只是具特殊格式之文章。因此，本文視關鍵詞為主題，將一論文向量P表示成主題向量，並採用如下的連續實數主題成分計算法。

$$\mathrm{P}=(P_1, P_2, K, P_v)$$
$$P_i=Tf_i \cdot IDF_i, \quad i=1\mathrm{K}\, v$$
$$TF_i=1+\ln(tf_i)$$
$$IDF_i=\ln\left(\frac{N}{df_i}\right)$$

其中，v為論文集總主題數，P_i為論文之主題i成分，TF_i為主題i之論文內頻繁度，IDF_i為主題i之論文集罕見度，tf_i為主題i之論文內出現次數，df_i為論文集中有出現主題i之論文數，N為論文集總論文數，ln為自然對數函數。採用對數函數理由在讓出現次數的邊際效益呈遞減現象。遇tf_i或df_i為0時，P_i無法計算一律指派為0。

由於任意實體的描述資訊只要能以一群關鍵詞的多次集合（multiset）表現，最後皆可轉化成上述主題向量。因此，視描述資訊的蒐集情形，評閱者向量也可由所蒐集到的描述關鍵詞表示成類似的主題向量。

㈢基於向量相似度之匹配法

向量相似度廣泛運用於文件的分類、群聚和檢索等應用。Salton列出常用的向量相似度公式有內積、Dice係數、餘弦、Jaccard係數等算法（Saton & Buckley, 1988）。其中，Dice係數及Jaccard係數比較偏集合的交集，聯集觀點，兩向量關鍵詞交集數愈高，其相似度愈高。又相同交集數之下，Dice係數傾向

比Jaccard係數給出更高相似度值。餘弦及內積比較偏向量夾角觀點，兩向量夾角愈小，其相似度愈高。餘弦計算時向量要先作標準化變為單位向量，相較於內積計算，比較不受關鍵詞多寡不均，即文件長短之影響。

由於在論文及評閱者皆表現成主題向量時，兩主題向量之相似度可視為兩者之匹配度，因此仿照傳統向量相似度算法，本文擬測試四種論文及評閱者匹配法。給定論文向量$P=(P_1,P_2,\ldots,P_v)$，評閱者向量$R=(R_1,R_2,\ldots,R_v)$，則兩者的內積相似度$Score_{inner}$、Dice係數相似度$Score_{dice}$、餘弦相似度$Score_{cosine}$、Jaccard相似度$Score_{jaccard}$四種匹配度算法分別如下。

$$Score_{inner}(P,R)=\sum_{k=1}^{v}P_k\cdot R_k$$

$$Score_{dice}(P,R)=\frac{2\cdot\sum_{k=1}^{v}P_k\cdot R_k}{\sum_{k=1}^{v}P_k+\sum_{k=1}^{v}R_k}$$

$$Score_{cosine}(P,R)=\frac{\sum_{k=1}^{v}P_k\cdot R_k}{\sqrt{\sum_{k=1}^{v}P_k^2}\cdot\sqrt{\sum_{k=1}^{v}R_k^2}}$$

$$Score_{jaccard}(P,R)=\frac{\sum_{k=1}^{v}P_k\cdot R_k}{\sum_{k=1}^{v}P_k+\sum_{k=1}^{v}R_k-\sum_{k=1}^{v}P_k\cdot R_k}$$

㈣OpenConf主題式匹配法

OpenConf為一操作簡單且資源需求低之線上投稿暨評閱系統，主要供研討會，及期刊論文之用（OpenConf, 1997）。其特色之一是提供3種啟發式匹配法，分別為無加權主題匹配法（unweighted topic match）、難指派評閱者優先加權主題匹配法（weighted topic match with the hard reviewers assigned first），及易指派評閱者優先加權主題匹配法（weighted topic match with the easy reviewers assigned first）。使用前須先由每位論文作者勾選其論文相關主題，及每位評閱者勾選其專長相關主題，分別形成論文向量及評閱者向量。兩種向量的主題成分皆採用布林值，即只有1或0，分別代表有或無該主題。然後，由系統管理者設定每篇論文所需評閱者人數，及每位評閱者可評閱之論文數上限。

1. 無加權主題匹配法

無加權主題匹配法依論文向量P及評閱者向量R，兩者主題交集個數之多

寡來決定匹配度高低，其公式如下。下標 *nw* 表示無加權主題（no weighting）匹配法。

$$Socre_{nw}(\mathrm{P, R}) = \sum_{k=1}^{v} P_k \cdot R_k$$

此匹配法和前述的向量內積相似度類似，差異只在論文及評閱者向量兩部分皆採用布林向量表示法。就論文 P 而言，可列出所有評閱者 R 和自己的匹配度 $Score_{nw}(P, R)$，匹配度高之評閱者，其推薦排名在前。

2. 難指派評閱者優先加權主題匹配法

加權主題匹配法從供需觀點，依照評閱者及論文選擇一主題數量之多寡，來計算該主題本身的指派容易度。一般而言，若選擇某主題之評閱者人數愈多，表示供給多，因此該主題愈好指派。同樣的，若選擇某主題之論文篇數愈少，表示需求少，因此該主題也愈好指派。一主題 t 之指派容易度 $Easy_{topic}(t)$ 算法如下。

$$Easy_{topic}(t) = \frac{Count_{reviewer}(t)}{Count_{paper}(t)}$$

其中，$Count_{reviewer}(t)$ 代表選擇主題 t 之評閱者人數，$Count_{paper}(t)$ 代表選擇主題 t 之論文篇數。

有了各主題的指派容易度之後，再依各主題向量實際擁有之主題，加總匯整出每篇論文及每位評閱者的指派容易度。給定論文向量 P 及評閱者向量 R，則論文指派容易度 $Easy_{paper}$，及評閱者指派容易度 $Easy_{reviewer}$，兩者計算公式分別如下。

$$Easy_{paper}(P) = \sum_{k=1}^{v} P_k \cdot Easy_{topic}(t_k)$$

$$Easy_{riviewer}(R) = \sum_{k=1}^{v} R_k \cdot Easy_{topic}(t_k)$$

其中，v 為論文集總主題數，t_k 為第 k 個主題，P_k 為論文 P 有無主題 t_k，R_k 為評閱者 R 有無主題 t_k。

最後，為了讓較難指派的論文優先指派，加權主題匹配法將論文依照論文指派容易度由低到高排序，循序從指派容易度最低分的論文開始指派。指派過程首先針對論文找出至少有一主題重疊之評閱者集合，然後依照評閱者指派容易度由低到高排序，以優先挑選較難指派（指派容易度較低）之評閱者。因此，論文 P 及評閱者 R 之匹配度公式如下。

$$ReviewerSet(P) = \left\{ R \mid \sum_{k=1}^{v} P_k \cdot R_k > 0 \right\}$$

$$Score_{whf}(P,R)=\begin{cases}\dfrac{1}{Easy_{reviewer}(R)} & \text{for } R\in ReviewerSet(P)\\ 0 & \text{otherwise}\end{cases}$$

其中，$ReviewerSet(P)$為至少和論文P有一主題重疊（即有資格）之評閱者集合，下標whf表示加權主題指派評閱者過程是亦採難指派評閱者優先（weighted hard first）原則。

3. 易指派評閱者優先加權主題匹配法

前述難指派評閱者優先加權主題匹配法首先採用難指派論文優先指派原則，遇評閱者集合有多人以上，才採用難指派評閱者優先指派原則。易指派評閱者優先加權主題匹配法與之類似，首先採用難指派論文優先指派原則，但遇評閱者集合有多人以上，則採用易指派評閱者優先指派原則。其目的為讓較易指派評閱者優先指派給難指派論文，以求得平衡，避免初始不易指派情形。因此，論文P及評閱者R之匹配度公式如下。

$$Score_{whf}(P,R)=\begin{cases}Easy_{reviewer}(R) & \text{for } R\in ReviewerSet(P)\\ 0 & \text{otherwise}\end{cases}$$

其中，下標wef表示加權主題指派評閱者過程是採易指派評閱者優先（weighted easy first）原則。

三、研究方法

為了解推薦論文評閱者過程，採用不同長度的論文表示方式及評閱者匹配法對推薦結果之影響，本文參考傳統資訊檢索作法，設計四種論文表示方式，分別為標題、關鍵詞、摘要、全文。另外，搭配七種評閱者匹配法，其中包括四種向量空間相似度匹配法，及三種OpenConf系統所採用的獨特主題匹配法，合計交叉組合成28種推薦模式。評估表現方面，本文以自行收集之論文及評閱者資料為測試資料集，依平均精確率為評估指標，觀察哪些論文表示方式和評閱者匹配法的組合，其推薦結果表現較佳。整個實驗流程如圖1所示。

㈠測試資料集的蒐集及前處理

測試資料集分成論文資料集及評閱者資料集。論文資料集方面，本研究從ICIM 2004[6]及ICIM 2006[7]兩屆國際資訊管理研討會（International Conference on Information Management）論文集中任意挑選100篇中文論文，擷取每篇論文

[6] 會議資料參見 第十五屆國際資訊管理學術研討會論文集，中原大學，台北，2004年5月29日。

[7] 會議資料參見 第十七屆國際資訊管理學術研討會論文集，義守大學，高雄，2006年5月27日。

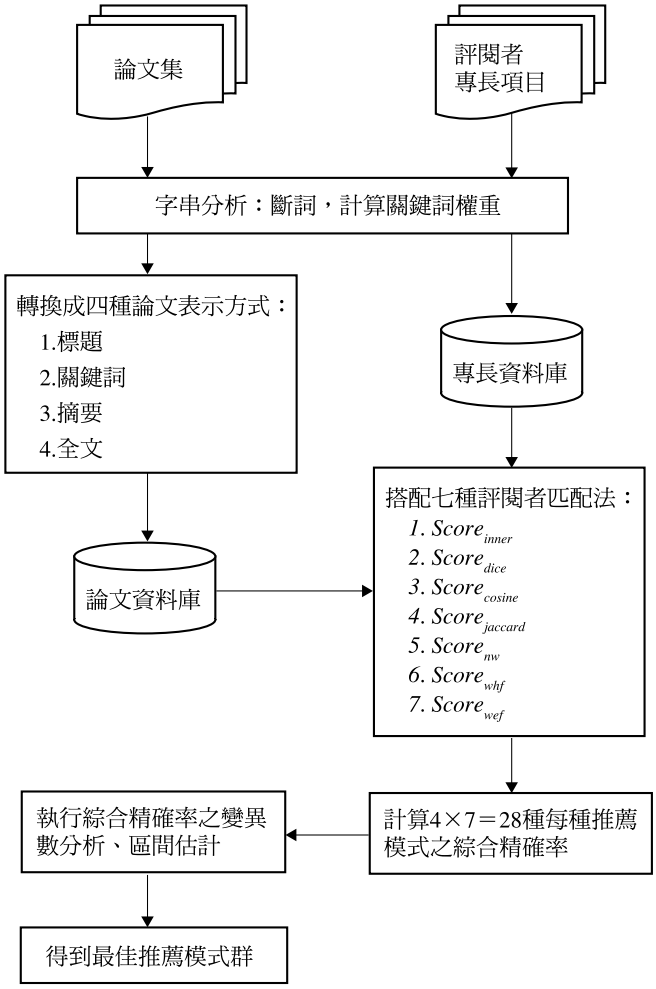

圖1 實驗流程圖

的作者姓名、標題、關鍵詞、摘要、內文等欄位。除作者姓名外，各欄位內容皆經過AutoTag[8]中文斷詞，儲存成關鍵詞串列，只有兩字元以上關鍵詞才保留，關鍵詞可重複出現，以利計算串列內詞頻。有關此100篇論文斷詞前後之全文統計資料如表1所示。其中，關鍵詞根（word type）指不重複計算之關鍵詞種類。

表1 測試資料集100篇中文論文斷詞前後之全文統計資料

斷詞前總字數	單篇論文最多字數	單篇論文最少字數	斷詞後總關鍵詞個數	斷詞後總關鍵詞根個數	單篇論文最多關鍵詞個數	單篇論文最少關鍵詞個數
1,015,884	10,984	8,997	480,052	22,459	1,488	887

[8] *AutoTag*中文斷詞工具，中央研究院詞庫小組，詳細資料參見http://godel.iis.sinica.edu.tw/CKIP/ws/

評閱者資料集方面，為方便客觀評估精確率，本研究視前述論文資料集所擷取之作者為評閱者候選人，並以人工蒐集每位評閱者在全國博碩士論文網[9]上曾指導或發表過有關於資訊或管理相關領域的學位論文，取其條列關鍵詞作為該評閱者的專長項目，如圖2。其中，領域乃依系所名稱作判別。每位評閱者所發表或指導過的學位論文越多，專長項目的累積將越多。將這些累積的專長項目經過AutoTag斷詞，儲存成關鍵詞串列，並計算其內關鍵詞出現次數，便形成評閱者專長資料庫，如圖3。其中，reviewerid欄位是評閱者代號，每位評閱者蒐集到的關鍵詞經過斷詞後存進term欄位，tf欄位為其串列內詞頻。

圖2　利用全國博碩士網蒐集評閱者專長項目的方法

←T→	reviewerid	reviewername	term	tf
☐ ✎ 🗑	1	丁明勇	架構	1
☐ ✎ 🗑	1	丁明勇	委外	1
☐ ✎ 🗑	1	丁明勇	系統	1
☐ ✎ 🗑	1	丁明勇	網絡	1
☐ ✎ 🗑	1	丁明勇	關係	2
☐ ✎ 🗑	1	丁明勇	資訊	2
☐ ✎ 🗑	1	丁明勇	界定	2
☐ ✎ 🗑	2	方清宏	碰撞	1
☐ ✎ 🗑	2	方清宏	資料	1
☐ ✎ 🗑	2	方清宏	頻道	3

圖3　評閱者專長資料庫內容

[9] 相關資料參見 全國博碩士論文網，國家圖書館，http://etds.ncl.edu.tw/theabs/

另外，於人工蒐集評閱者專長過程，若遇姓名相同者，本研究是依其所屬單位判別是否同一人，若遇混淆無法辨別狀況，也會全蒐集進來視為同一人。因此，同一位評閱者在各領域的論文關鍵詞都會蒐集在其專長項目之下。若評閱者在某領域擁有的論文較多，將累積較多該領域的關鍵詞，利於推薦其為該領域論文之評閱者人選。本評閱者資料集共蒐集到194 名評閱者專長項目，其關鍵詞統計資料如表2。

表2 測試資料集194名評閱者專長項目關鍵詞統計資料

	最少	最多	平均
斷詞後評閱者專長關鍵詞個數	2	190	69

㈡論文及評閱者表示方式

基於前面的論文及評閱者關鍵詞串列，本文考慮將之轉換成向量表示方式，以供後續兩者匹配比較之用。論文方面，一篇論文常見分成標題、關鍵詞、摘要、內文等四個欄位，各欄位所含資訊量多寡不一。若取少者作為論文代表，例如標題或關鍵詞欄位，則可能因為作者用詞不同或用詞個數受限制，無法完整表達而有匹配誤差。若取多者作為代表，例如摘要或內文欄位，除了處理時間變長，也可能因為雜訊太多而產生匹配誤差。為求作一客觀比較，本文設計將論文資料集的100篇論文，依不同欄位來源，分別轉換成四種論文表示方式，如下：

1. 標題 P^0_{title} 及 P^t_{title}
2. 關鍵詞 $P^0_{keyword}$ 及 $P^t_{keyword}$
3. 摘要 $P^0_{abstract}$ 及 $P^t_{abstract}$
4. 全文 $P^0_{fulltext}$ 及 $P^t_{fulltext}$

其中，每種表示方式的轉換過程主要在將設定來源的關鍵詞串列分別轉換成一個二值權重關鍵詞向量 P^0，及一個TFIDF權重關鍵詞向量 P^t，以供後續不同評閱者匹配法使用。就論文的二值權重關鍵詞向量 P^0，每一獨立關鍵詞維度只以1/0記錄論文有/無該關鍵詞出現。就論文的TFIDF權重關鍵詞向量 P^t，每一獨立關鍵詞維度值則記錄論文能以該關鍵詞作代表之程度，其值愈高表示愈能代表（即辨識或檢索）該論文。

至於評閱者方面，由於獨立來源的資料量不多，故本文只採一種關鍵詞表示方式，將原始累積關鍵詞串列分別轉換成一個二值權重關鍵詞向量 $R^0_{keyword}$，及一個TFIDF權重關鍵詞向量 $R^t_{keyword}$，以供後續不同評閱者匹配法使用。其中，在計算TFIDF權重時，因為評閱者資料量少，所以採用前面統計論文資料集全文所得之IDF值（關鍵詞辨識力）作輔助。

㈢評閱者匹配法

有了論文及評閱者向量，論文即可依據向量相似度進行評閱者匹配，並依匹配程度高低作排名及推薦評閱者。為了解不同向量相似度及OpenConf系統的匹配表現，本文共設計比較七種評閱者匹配法，如表3。其中⑴-⑷取自傳統資訊檢索領域常用的四種向量相似度（Salton & Buckley, 1988），分別為內積、Dice係數、餘弦、Jaccard係數匹配法。⑸-⑺則取自OpenConf系統所提供的三種獨特主題匹配法，分別為無加權主題、難指派評閱者優先加權主題、易指派評閱者優先加權主題匹配法。七種匹配法中向量相似度匹配法皆使用到論文及評閱者的TFIDF權重向量，OpenConf匹配法則使用二值權重向量。

表3　七種評閱者匹配法

匹配法公式	匹配法描述
⑴ $Score_{inner}(P^t, R^t)$	內積相似度匹配法
⑵ $Score_{dice}(P^t, R^t)$	Dice相似度匹配法
⑶ $Score_{cosine}(P^t, R^t)$	餘弦相似度匹配法
⑷ $Score_{jaccard}(P^t, R^t)$	Jaccard相似度匹配法
⑸ $Score_{nw}(P^0, R^0)$	OpenConf無加權主題匹配法
⑹ $Score_{whf}(P^0, R^0)$	OpenConf難指派評閱者優先加權主題匹配法
⑺ $Score_{wef}(P^0, R^0)$	OpenConf易指派評閱者優先加權主題匹配法

在使用OpenConf系統時，若論文所需評閱者人數已滿，甚至找不齊，後面評閱者就不列入排名。同理，若評閱者可評閱論文數已達上限，就停止列入指派。若考慮這些負荷限制，將造成每篇論文無法求得和所有評閱者之匹配排名。由於本文重點只在比較推薦模式的專長符合度，不考慮實際人力負荷限制，故實驗時的系統參數設定如下：

1. 每篇論文所需評閱者人數：194，即所有評閱者都列入評比。

2. 每位評閱者可評閱論文數上限：100，即所有論文都列入評比。

其目的是為了得到所有評閱者對於所有論文的推薦排名位置。有了完整的評閱者排名清單，才能和向量相似度模式站在相同立足點上評估推薦結果優劣。

㈣推薦模式之評估

依據前面的四種論文表示方式，搭配七種評閱者匹配法，共交叉組合成28種推薦模式。為方便大量客觀評估不同推薦模式優劣，本文採用原作者專長理應為論文正確評閱者專長之觀點，參考傳統資訊檢索評估作法，改以論文召回評閱者之排名清單，來計算推薦模式之精確率。因此，第i篇論文推薦評閱者的平均精確率Avg_P_i計算公式如下：

$$Avg\_P_i = \frac{\sum_{j=1}^{n_i} P_i(j)}{n_i}$$

其中，$P_i(j)$表示第i篇論文於召回第j位作者時之精確率，n_i表示第i篇論文召回作者總數。原作者若在論文召回評閱者排名清單中的位置愈前面，其召回點位置的精確率愈高。若論文召回原作者不只一位，則取各作者召回點位置精確率之平均值。此平均精確率值介於0～1之間，愈高表現愈好，當所有作者皆未召回時其值為0，當所有召回作者皆名列前茅時其值為1。

利用每篇論文的平均精確率Avg_P_i，再就論文資料集所有100篇論文求平均，即得每個推薦模式之綜合精確率P，此值亦介於0～1之間，愈高表現愈好，其計算公式如下：

$$P = \frac{\sum_{i=1}^{100} Avg\_P_i}{100}$$

得到28個推薦方法的綜合精確率後，再以統計方法作ANOVA變異數分析，檢定結果是否具顯著差異性，最後以區間估計找出實驗中最佳結果之區間範圍，以找出效果相等之最佳推薦模式群。

四、實驗結果及統計分析

本研究實驗環境計使用Microsoft Windows XP Professional作業系統、MySQL 4.0.24-nt資料庫、J2SE 5.0_06及PHP 4.3.11程式語言、OpenConf 2.1線上投稿暨評閱系統、AutoTag中文斷詞器、SPSS統計軟體等工具。

㈠實驗結果

以下為四種論文表示方式及七種評閱者匹配法交叉組合下的28個推薦模式結果。將實驗結果分成兩大部分來看，在四種論文表示方式下，表4為搭配傳統向量四種相似度匹配法所得的綜合精確率推薦結果，表5則為搭配OpenConf系統三種獨特主題匹配法所得的綜合精確率推薦結果。

表4　搭配傳統向量四種相似度匹配法所得綜合精確率

匹配法 / 論文表示方式	$Score_{inner}$	$Score_{dice}$	$Score_{cosine}$	$Score_{jaccard}$	平均
標題	0.25	0.31	0.31	0.31	0.3
關鍵詞	0.29	0.35	0.36	0.35	0.34
摘要	0.24	0.33	**0.37**	0.33	0.32
全文	0.18	0.20	0.34	0.20	0.23
平均	0.24	0.3	0.35	0.3	0.3

由表4可知，採用到傳統向量空間相似度匹配法的16種推薦模式，其綜合精確率表現介於0.18至0.37之間，平均值0.3。最高值0.37為摘要論文表示方式，搭配餘弦相似度匹配法的推薦模式。就表4之匹配法而言，以餘弦平均值0.35最高，其次Dice、Jaccard係數平均值皆0.3，最低為內積平均值0.24。顯然餘弦相似度公式中，比內積相似度公式多了一個用來標準化的分母項，將不同論文長度因素也考慮進去，結果有比較好。

就表4之論文表示方式來比較，關鍵詞表示法平均值0.34最高，其次為摘要0.32、標題0.3、全文0.23。這顯示使用全文表示方式可能雜訊多，效果反而差。

表5　搭配OpenConf系統三種獨特主題匹配法所得的綜合精確率

評分方法 / 論文表示方式	$Score_{nw}$	$Score_{whf}$	$Score_{wef}$	平均
標題	0.13	0.02	0.04	0.06
關鍵詞	**0.20**	0.03	0.05	0.09
摘要	0.06	0.02	0.02	0.03
全文	0.04	0.02	0.02	0.03
平均	0.11	0.02	0.03	0.05

由表5可得知，採用到OpenConf系統獨特主題匹配法的12種推薦模式，其綜合精確率表現都不高，介於0.02至0.20之間，有10種推薦模式表現低於0.1，有9種推薦模式表現低於0.05。其平均值0.05，比表4傳統向量相似度匹配法的平均值0.3小相當多。表5最高值0.20為關鍵詞論文表示方式，搭配無加權主題匹配法的推薦模式，其值也比表4最高值0.37小很多。就表5之匹配法而言，只有無加權主題匹配法的平均值0.11較好。顯然考慮指派難易度在追求匹配效果上不見得有利。就表5之論文表示方式而言，表現最好的仍為關鍵詞平均值0.09，其次標題0.06，摘要及全文皆0.3。

綜合而言，在28種推薦模式中，就七種評閱者匹配法來看，使用餘弦相似度匹配法平均結果0.35最佳；就四種論文表示方式來看，使用關鍵詞論文表示方式平均結果0.34最佳；就交叉組合的推薦模式來看，使用摘要論文表示方式，搭配餘弦相似度匹配法的推薦模式結果0.37最高，表現最好。

㈡統計分析

針對28種推薦模式的綜合精確率結果，以SPSS統計軟體作ANOVA變異數分析，以檢驗其是否具顯著差異性。在每種推薦模式樣本數100，$\alpha = 0.05$之下，分析結果的p-value為0，小於α，代表28種推薦模式具顯著差異。

接著為找出統計上表現等同於最佳的推薦模式群，遂以28種推薦模式中表

現最佳，綜合精確率為0.37的摘要搭配餘弦推薦模式為中心，在樣本數100及$\alpha = 0.05$之下，取得其95%的信賴區間（0.32, 0.43）。依此找出落在區間內的其他推薦模式，共有7種，即為論文評閱者最佳推薦模式群，如表6。其中，匹配法方面，餘弦相似度出現3次，Dice及Jaccard相似度次之，出現2次，可見餘弦相似度、Dice相似度、Jaccard相似度三者能力接近，而未作標準化的內積匹配法及其他主題式匹配法效果並不好。表示法方面，關鍵詞及摘要出現3次，全文出現1次，可見資訊量中等的關鍵詞、摘要表示法兩者能力接近，而資訊最少的標題表示法能力略差，資訊太多的全文表示法能力則最不好。

表6 綜合精確率落於（0.32 , 0.43）最佳區間內論文評閱者最佳推薦模式群

論文表示方式	匹配法	綜合精確率*P*
(1) 摘要	餘弦相似度匹配法	0.37
(2) 關鍵詞	餘弦相似度匹配法	0.36
(3) 關鍵詞	Dice相似度匹配法	0.35
(4) 關鍵詞	Jaccard相似度匹配法	0.35
(5) 全文	餘弦相似度匹配法	0.34
(6) 摘要	Dice相似度匹配法	0.33
(7) 摘要	Jaccard相似度匹配法	0.33

五、結　論

隨著愈來愈多期刊邁入電子化多向互動出版模式，其旺盛的需求也促使線上投稿暨評閱系統的開發出現蓬勃發展。但是現有評閱系統多著重在投稿、查詢、報表等基本自動化功能上，而有關自動推薦評閱者功能方面，仍未受重視，多數從缺。其實投稿系統本身就擁有論文電子檔，萃取其文字部分作為論文代表並不困難，只要再輔以適當評閱者資料，兩者作比對來推薦評閱者，即可為期刊編輯省下不少找尋評閱者時間。

為了解不同論文表示方式及匹配法之評閱者推薦能力，本文將論文斷詞後，依照學術論文架構，內容由少到多分別萃取標題、關鍵詞、摘要、全文形成4種論文表示方式。匹配法部分則利用傳統向量空間模式的內積、Dice係數、餘弦、Jaccard係數4種相似度匹配法，及OpenConf系統的3種獨特主題匹配法，合計7種匹配法，共組合出28種評閱者推薦模式作測試。

經過蒐集的100篇論文測試集比較，依平均精確率為評估指標，本文得出以下觀察結論：

1. 28種推薦模式中最佳評閱者推薦模式為摘要論文表示方式，搭配餘弦相似度匹配法之組合。統計上，和此最佳組合不分軒輊的其餘6種推薦模式群可參考表6結果。

2. 4種長度的論文表示法中以關鍵詞表現最佳，其次為摘要及標題，唯兩

者一般表現差距不大，最差則為全文表示法。可見論文表示法選取的長度的確對匹配結果產生不同影響。長度太長的全文表示法不但處理費時，而且容易帶來雜訊，最不利於匹配。

3. 7種匹配法中，以餘弦表現最佳，其次為Jaccard及Dice，兩者表現接近，再次為內積，然後為無加權主題，最差為兩個加權主題模式。可見匹配方法對匹配結果亦具重大影響力。採用多維向量的匹配會比採用一維匯整主題式的匹配表現好。

4. 採用向量空間模式的4種相似度匹配法，搭配任意論文表示方式，其結果都比OpenConf系統的3種獨特主題匹配法更好。顯示偏重指派功能的OpenConf匹配法較不適合在無負荷限制下的期刊論文評閱者推薦上。而OpenConf會輸於多數向量模式，理由可分為兩部分：在兩加權主題模式方面，OpenConf於匹配時，只籠統依照一維的加總主題指派容易度來代表論文及評閱者，再依照論文易指派者優先，然後評閱者難或易指派者優先，來作匹配，其匹配過程遠不如向量模式採用多維主題匹配來得面面俱到，故準確率偏低。至於OpenConf無加權主題模式方面，雖然採用多維主題匹配，但是使用布林向量，其匹配過程也不如向量模式使用實數向量來得細膩，故準確率較差。

最後，經過實際資料蒐集及匹配測試，本文提供建議如下，以供未來建置評閱者自動推薦模組之參考：

1. 本研究發現在評閱者採用關鍵詞表示法下，論文若採用關鍵詞表示法，其表現亦毫不遜色。但如此一來則提供兩點額外好處，一為關鍵詞表示法的儲存空間及處理時間都不大，適合大量資料庫之建立。二為評閱者及論文皆採用一致表示法，可方便自過去論文直接萃取作者為評閱者。

2. 蒐集評閱者專長資料庫方面，國科會研究人才專長資料也是一個獨立的專長來源。但是受限於許多人才設定資料不公開，或填寫的是英文專長，實務上資料取得不若博碩士論文資料庫的關鍵詞取得來得方便。另外，國科會專長資料庫會有增刪情況而只留最新一份，但是博碩士論文的關鍵詞則具累積性，研究多的領域其累積之關鍵詞份量會更重，較符合TFIDF精神。基於上述兩理由，本文建議以博碩士論文資料庫作為主要專長來源。

3. 但是，純粹以評閱者過去論文所累積之關鍵詞當作其專長時，不易及時反映其最新延伸興趣，故輔以類似國科會專長庫之自填專長欄位，可以彌補此時差問題。特別是若出現一個全新領域，過去研究者本來就不多，純關鍵詞的專長匹配就容易成空。這時，若能引入給定專長領域分類樹（Biswas & Hasan 2007）或自動計算關鍵詞相關矩陣（Baeza-Yates & Ribeiro-Neto, 1999）也許就可以增加匹配機會。

4. 於本研究蒐集資料過程，以獨立之評閱者專長資料最難取得，故暫時人

工取自碩博士論文網作者自訂之關鍵詞。實用上，若能將這部分查詢予以網路服務（web service）自動化，提供一個專長人材查詢服務，相信於建構論文評閱者推薦模組或其他相關應用上將更利於往前跨出一大步。

參考文獻

邱炯友（2003）。學術電子期刊同儕評閱之探析。*教育資料與圖書館學*，*40*(3)，309-323。

邱炯友、李怡萍（2006）。學術電子期刊編輯整合平台市場與個案：以教育資料與圖書館學季刊為例。*教育資料與圖書館學*，*43*(3)，327-345。

顏玉茵（2004）。*台灣學術期刊電子化同儕評閱系統建構之評析*。未出版之碩士論文，南華大學出版與文化事業管理研究所，嘉義縣。

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval*. New York: Addison Wesley Longman Limited.

Biswas, H. K., & Hasan, M. M. (2007). Using publications and domain knowledge to build research profiles: An application in automatic reviewer assignment. In *International conference on information and communication technology* (pp. 82-86).

Dumais, S. T., & Nielsen, J. (1992). Automating the assignment of submitted manuscripts to reviewers. In *International ACM SIGIR conference on research and development in information retrieval* (pp. 233-244).

Hartvigsen, D., & Wei, J. C. (1999). The conference paper-reviewer assignment problem. *Decision Sciences*, *30*(3), 865-976.

Hettich, S., & Pazzani, M. J. (2006). Mining for proposal reviewers: Lessons learned at the national science foundation. In *International conference on knowledge discovery and data mining* (pp. 862-871).

Mauro, N. D., & Basile, T. M. A., & Ferilli, S. (2005). GRAPE: An expert assignment component for scientific conference management systems. *Lecture Notes in Computer Science*, *3533*, 789-798.

Mimno, D., & McCallum, A. (2007). Expertise modeling for matching papers with reviewers. In *International conference on knowledge discovery and data mining* (pp. 500-509).

OpenConf (1997). *The OpenConf conference management system*, Retrieved June 30, 2008, from http://www.zakongroup.com/technology/openconf.shtml

Rigaux, P. (2004). An iterative rating method: Application to web-based conference management. In *Proceedings of the 2004 ACM symposium on applied computing* (pp. 1682-1687).

Salton, G., & McGill, M. J. (1989). *Introduction to modern information retrieval*. New York: McGraw-Hill.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic retrieval. *Information Processing and Management*, *24*(5), 513-523.

Ware, M. (2005). Online submission and peer review systems. *Learned Publishing*, *18*, 245-250.

Journal of Educational Media & Library Sciences 45 : 4 (Summer 2008) : 409-431

Evaluation of Information Retrieval Based Models for Recommendation of Paper Reviewers

Shih-Chieh Wei*

Assistant Professor
E-mail: seke@mail.im.tku.edu.tw

Hsin-Yu Luo

Graduate Student
Department of Information Management, Tamkang University
Taipei, Taiwan

Abstract

As more e-journals appear and the e-review process becomes more popular, the demand for automatic recommendation of a good peer reviewer has been ever increasing. To automate the process of paper reviewer recommendation, this work evaluates four kinds of paper representations, which include full text, abstract, title, and author defined keywords. To match reviewers with papers, this work evaluates seven scoring methods including three topic-based methods from OpenConf, a popular online submission system with source, and four similarity-based methods from the vector space model of traditional information retrieval. The results of the 28 experiments show that recommendation methods based on the vector space model are better than the three topic-based methods of OpenConf in most document representations. Among them, the abstract paper representation combined with cosine similarity matching measure has the highest average precision.

Keywords: *Vector space model; Paper representation; Reviewer recommendation*

SUMMARY

To address the needs for e-journal development, more and more online submission systems have come into the market, which include ScholarOne Manualscript Central[1] and Aspers (Airiti Submission and Peer Review System)[2]. While most of these systems provide such functions as online submission, progress check, peer review, report generation, authorization form upload, and even publication fee payment, few of them support the automatic recommendation of

* Principal author for all correspondence.

[1] For accounts of the system, see http://www.scholarone.com/products_manuscriptcentral_aboutMC.shtml

[2] For accounts of the system, see http://portal.airiti.com/modules/tinyd0/index.php?id=25

reviewers for peer review. For journal editors, assigning appropriate reviewers manually remains a long term burden.

Ideally, to recommend paper reviewers, an online submission system should be able to represent a paper by its title, keywords, abstract, or full text, which can be obtained by either manual input or automatic extraction from the uploaded text. The system should also be able to obtain the expertise of reviewers by either manual input or automatic extraction from reviewers' past publications. By having some kind of matching mechanism between a paper and its reviewers, an online submission system with the capability of reviewer recommendation should save a lot of time for journal editors in search of appropriate reviewers.

In the literature on online submission systems, there are few systems which are able to recommend reviewers and available for trial use download. These include CyberChair[3], MyReview[4], and OpenConf[5]. Among these systems, the OpenConf online submission system is the most popular system and has been adopted in hundreds of conference and journal reviewing tasks. Thus it can serve as a benchmark for comparison. For reviewer recommendation, OpenConf provides three kinds of topic matching: unweighted topic matching, weighted topic matching with hard reviewers assigned first, and weighted topic matching with easy reviewers assigned first.

To evaluate current techniques for reviewer recommendation, this work will assume that papers are represented by automatic extracted topics and that topics are independent of each other. Four kinds of paper representation with different lengths will be tested with seven kinds of scoring methods which include four similarity-based methods from the traditional vector space model and three topic-based methods from OpenConf. Therefore, a total of 28 modes of recommendation are evaluated for implementation of reviewer recommendation in future online submission systems.

Paper and Reviewer Representations

Given a paper, we prepared the four paper representations of varying length in Table 1 which include the title, keywords, abstract, and full text. For each representation, two versions of Boolean and TFIDF representations are available. $\boldsymbol{P}^{\boldsymbol{0}}=(P_1, P_2,..., P_v)$ denotes the Boolean representation which uses $P_i=1$ to denote the presence of a term i and 0 otherwise. v is the total number of terms.

[3] For accounts of the system, see http://www.cyberchair.org

[4] For accounts of the system, see http://myreview.intellagence.eu

[5] For accounts of the system, see http://www.openconf.org

Table 1 The Four Paper Representations of Varying Length

	Paper Representations	Description
(1)	P^0_{title} and P^t_{title}	Title
(2)	$P^0_{keyword}$ and $P^t_{keyword}$	Keywords
(3)	$P^0_{abstract}$ and $P^t_{abstract}$	Abstract
(4)	$P^0_{fulltext}$ and $P^t_{fulltext}$	Full text

P^t denotes the TFIDF representation which uses the following equations for computing term frequency (TF) and inverse document frequency (IDF).

$$P^t=(P_1, P_2, \ldots, P_v)$$
$$P_i=Tf_i \bullet IDF_i,\ i=1\ldots v$$
$$TF_i=1+ln(tf_i)$$
$$IDF_i=\ln(\frac{N}{df_i})$$

where v denotes the number of terms, P_i the weight of term i, TF_i the intra-document frequency of term i, IDF_i the inter-document rarity of term i, tf_i the occurrence count of term i within the document, df_i the occurrence count of documents having term i, N the total number of documents, and ln the natural logarithm. The logarithm function is used to have a decreasing marginal effect on increasing counts. When tf_i or df_i is 0, P_i cannot be computed and would be assigned zero.

For automatic ranking of the recommendation result, we consider the authors in the paper dataset as candidate reviewers and hope that reviewers having the same expertise as the author would be selected as recommended reviewer of the paper. Therefore, to avoid unfair favor over paper authors, an independent source of expertise other than the paper dataset is consulted for collecting reviewer expertise. From the Electronic Theses and Dissertations System (ETDS)[6], we filter out those publications whose advisor or author has appeared in the paper dataset. Then for each candidate reviewer, terms in the keyword field of all his publications (as advisor or author) are accumulated as the reviewer's expertise. Duplicate terms in different publications are not deleted as we want to use TFIDF representation for reviewers. However, for reviewer's TFIDF representation, the IDF value of terms would use those statistics from the paper database since a small size of reviewers would make the IDF value unreliable.

Scoring Methods for Matching Papers and Reviewers

A total of seven scoring methods for matching papers and reviewers are listed in Table 2. The first four come from the traditional vector space model of

[6] For accounts of the system, see http://etds.ncl.edu.tw/theabs/

information retrieval and the last three are adopted from OpenConf. The equations for computing the seven scoring methods are as follows.

Table 2 The Seven Scoring Methods Evaluated for Matching Papers and Reviewers

	Scoring Method	Description
(1)	$Score_{inner}(P^t,R^t)$	Inner product
(2)	$Score_{t}(P^t,R^t)$	Dice coefficient
(3)	$Score_{n}(P^t,R^t)$	Cosine measure
(4)	$Score_{jaccard}(P^t,R^t)$	Jaccard coefficient
(5)	$Score_{nw}(P^0,R^0)$	OpenConf unweighted topic matching
(6)	$Score_{whf}(P^0,R^0)$	OpenConf weighted topic matching with hard reviewers assigned first
(7)	$Score_{wef}(P^0,R^0)$	OpenConf weighted topic matching with easy reviewers assigned first

$$Score_{inner}(P, R)=\sum_{k=1}^{v} P_k \cdot R_k,$$

$$Score_{dice}(P, R)=\frac{2\cdot\sum_{k=1}^{v} P_k \cdot R_k}{\sum_{k=1}^{v} P_k+\sum_{k=1}^{v} R_k},$$

$$Score_{cosine}(P, R)=\frac{\sum_{k=1}^{v} P_k \cdot R_k}{\sqrt{\sum_{k=1}^{v} P_k^2}\cdot\sqrt{\sum_{k=1}^{v} R_k^2}},$$

$$Score_{jaccard}(P, R)=\frac{\sum_{k=1}^{v} P_k \cdot R_k}{\sum_{k=1}^{v} P_k+\sum_{k=1}^{v} R_k-\sum_{k=1}^{v} P_k \cdot R_k},$$

$$Score_{nw}(P, R)=\sum_{k=1}^{v} P_k \cdot R_k,$$

$$Score_{whf}(P, R)=\begin{cases}\dfrac{1}{Easy_{reviewer}(R)} & \text{for } R\in ReviewerSet(P)\\ 0 & \text{otherwise}\end{cases},$$

$$Score_{whf}(P, R)=\begin{cases}Easy_{reviewer}(R) & \text{for } R\in ReviewerSet(P)\\ 0 & \text{otherwise}\end{cases}$$

with

$$ReviewerSet(P) = \left\{ R \mid \sum_{k=1}^{v} P_k \cdot R_k > 0 \right\},$$

$$Easy_{paper}(P) = \sum_{k=1}^{v} P_k \cdot Easy_{topic}(t_k),$$

$$Easy_{riviewer}(R) = \sum_{k=1}^{v} R_k \cdot Easy_{topic}(t_k),$$

$$Easy_{topic}(t) = \frac{Count_{reviewer}(t)}{Count_{paper}(t)}$$

where $ReviewerSet(P)$ is the set of all candidate reviewers having some intersection of terms with the paper P, $Easy_{paper}(P)$ the ease-of-assignment value for paper P, $Easy_{reviewer}(R)$ the ease-of-assignment value for reviewer R, $Easy_{topic}(t)$ the ease-of-assignment value for term t, $Count_{reviewer}(t)$ the number of reviewers having the expertise of term t, and $Count_{paper}(t)$ the number of papers containing topic term t.

Results and Discussions

This study evaluates the performance of different paper representations and scoring methods in reviewer recommendation. For paper representation, word segmentation by AutoTag[7] is first applied to Chinese papers, and based on the paper structure, four kinds of representation of different length are generated for each paper, which are the title, keywords, abstract, and full text. For scoring methods, there are four similarity-based methods from the traditional vector space model of information retrieval which include the inner product, Dice coefficient, cosine measure, Jaccard coefficient, and three topic-based methods from OpenConf. A total of 28 modes of recommendation are evaluated.

Using a dataset of 100 Chinese papers from ICIM 2004[8] and ICIM 2006[9], comparison of the 28 modes of reviewer recommendation is made based on the performance index of average precision. Based on the experimental results, the following observations are made.

1. Among the 28 modes of reviewer recommendation, the best mode of recommendation is the abstract paper representation combined with the cosine

[7] AutoTag, a Chinese word segmentation tool, Academia Sinica, see http://godel.iis.sinica.edu.tw/CKIP/ws/

[8] For accounts of the conference, see Proceedings of the 15th International Conference on Information Management, Chung Yuan Christian University, Taipei, Taiwan, May 29, 2004.

[9] For accounts of the conference, see Proceedings of the 17th International Conference on Information Management, I-Shou University, Kaoshiung, Taiwan, May 27, 2006.

measure scoring method. Statistically, six other modes of recommendation which are equally good as the best are listed in Table 3.

Table 3 The Best Modes of Reviewer Recommendation

	Paper representation	Scoring method	Average precision
(1)	Abstract	Cosine measure	0.37
(2)	Keywords	Cosine measure	0.36
(3)	Keywords	Dice coefficient	0.35
(4)	Keywords	Jaccard coefficient	0.35
(5)	Full text	Cosine measure	0.34
(6)	Abstract	Dice coefficient	0.33
(7)	Abstract	Jaccard coefficient	0.33

2. Among the four paper representations of varying length, the best is the keywords, followed by the abstract and the title, having only minor difference between the two, and the worst is the full text. It can be seen that paper representations of varying length does make a difference in matching results. Using the full length of the paper, the full text representation consumes the highest processing time and also brings in high amount of noise that makes it unfit for matching.

3. Among the seven scoring methods, the best is the cosine measure, followed by Jaccard and Dice coefficients, both being near, and then the inner product and unweighted topic matching. The worst is the two weighted topic matching methods. It can be seen that scoring methods also make a difference in matching results. Also, matching by multi-dimensional vectors performs better than that by one-dimensional topic summation scores.

4. The four similarity-based matching methods from the traditional vector space model perform better than the three topic-based matching methods from OpenConf, irrespective of combination with any paper representations. It shows that the assignment-oriented OpenConf is not suitable for journal reviewer recommendation where precision is more important than fulfilling assignment constraints. The reasons that OpenConf performs worse than the traditional vector space model can be attributed as follows. For the two weighted topic matching methods, OpenConf represents papers and reviewers by a simplified one-dimensional ease-of-assignment value which is computed from summation of all ease-of-assignment values of the constituent topics. OpenConf's matching method in one dimension is not as precise as the vector space model's matching method in multiple dimensions. For the unweighted topic matching method, OpenConf uses multi-dimensional Boolean matching whose precision is also not as good as the multi-dimensional matching in real numbers used by the vector

space model.

With the experiences in data collection and matching experiments, some suggestions can be made as follows for future implementation of reviewer recommendation systems.

1. It is found in this work that based on the reviewer representation by keywords, the paper representation by keywords performs almost as good as the best paper representation by abstract when combined with the cosine measure scoring method. If the paper and the reviewer both adopt the succinct keywords representation, there would be two obvious merits. One is much saving in storage space and processing time which would be beneficial for setting up large paper and reviewer databases. The other advantage is easy transfer of past paper authors to future paper reviewers because of the adoption of the same representation.

2. For collecting reviewer expertise, in addition to the Electronic Theses and Dissertations System (ETDS)[10] at National Central Library that was used in this work, there is also a good independent source, i.e., the Research Experts Search (RES)[11] at National Science Council. However, due to the restricted access or English expertise preferred by many RES researchers, it is in practice not as easy to retrieve Chinese researchers' expertise from RES as from ETDS. Furthermore, the expertise listed in RES is subject to change and only the most recent one can be queried, while ETDS keeps keywords of all past research with higher importance for more frequent keywords, thus better fits the TFIDF model. It is therefore recommended that ETDS be used as the main source of collecting reviewer expertise.

3. To alleviate the problem of matching new research fields, the RES source can be used to reflect the most recent extension of research interest of reviewers. When new keywords occur, which cannot be matched with previous ones, using an expert domain classification tree or keyword co-occurrence matrix might be of help.

4. Collecting reviewer expertise has been the most laborious part of this work, which is currently done by manual lookup from ETDS. For practical use, the web-based ETDS interface for human lookup can be wrapped up as an expert finding web service for machine lookup, which will greatly benefit future development of automatic reviewer recommendation systems.

[10] For accounts of the system, see http://etds.ncl.edu.tw/theabs/

[11] Details see https://nscnt07.nsc.gov.tw/WRS/