教育資料與圖書館學

Journal of Educational Media & Library Sciences

http://joemls.tku.edu.tw

Vol. 52 , no. 3 (Summer 2015) : 269-298

利用文字內容主題特徵與機器學習方法探討MIS相關期刊在ISI資料庫的主題分類

A Study of the Subject Categorization of the MIS-related Journals in the ISI Databases Using Topical Features in the Text Content and Machine Learning Methods

林 頌 堅 Sung-Chien Lin

Assistant Professor

E-mail：scl@cc.shu.edu.tw

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

教育資料與圖書館學　52 : 3 (Summer 2015) : 269-298
DOI:10.6120/JoEMLS.2015.523/0027.RS.AM

研究論文

利用文字內容主題特徵與機器學習方法探討MIS相關期刊在ISI資料庫的主題分類

林頌堅

摘要

本研究利用主題模型、期刊群集與類別預測等方法，分析與討論ISI主題類別IS&LS的MIS相關期刊中同時被賦予Management類別的情形。在期刊群集實驗裡，所有被指定到Management類別的期刊及其它同樣具有相似主題特徵的期刊都被聚集在同一個期刊群集內，「管理」是其共同且最突顯的主題。由於此群集包含的期刊和先前研究的MIS群集大多相同，因此視為本研究的MIS群集。類別預測實驗使用分類迴歸樹方法，分別以ISI的Management類別以及本研究的MIS群集做為正案例，進行期刊類別預測。兩次試驗產生的分類樹都以「管理」主題的出現機率為主要的分類規則，但後者不僅分類樹較為單純，同時預測錯誤也較少。也就是若將MIS群集內所有期刊都指定到Management類別，會使檢索的成效更為周全有效。

關鍵詞：ISI主題類別，機器學習，主題模型，期刊群集，類別預測

世新大學資訊傳播學系助理教授
通訊作者：scl@cc.shu.edu.tw

2015/07/31投稿；2015/08/14修訂；2015/08/14接受

緒　論

Thomson Reuters公司下的ISI（Institute for Scientific Information）每一年度出版的期刊索引報告（Journal Citation Reports, JCR）和引文資料庫（Web of Science, WoS）[1]都為收錄的期刊賦予一個或以上的主題類別（subject category）。例如*Journal of Information Science*被賦予的主題類別為Information Science and Library Science（IS&LS），而*Information Systems Research*的類別除IS&LS之外，同時還有Management。這是期刊在ISI資料庫上的一個重要資料，提供使用者透過學科檢索期刊（Leydesdorff & Rafols, 2009）。

主題類別是ISI資料庫上代表期刊所屬學科的重要資料，許多書目計量學研究採用ISI的主題類別定義學科的研究範疇。進行圖書資訊學領域相關期刊的群集分析時，便通常利用ISI資料庫內具有主題類別IS&LS的期刊作為分析資料，而這些研究大多發現管理資訊系統（management information systems, MIS）相關期刊明顯與其他期刊分離，自成一個群集（Ni, Sugimoto, & Cronin, 2013; Tseng & Tsay, 2013）。此外，Boyack、Klavans與Börner（2005）對2000年所有SCI與SSCI期刊在科學映射圖的結果進行群集則發現，圖書資訊學領域最相關的主題類別I&LS（Information & Library Science）的相關期刊聚集成圖書館事業（libraries and librarians and their work）和資訊科學先進議題（advances in information science）兩個群集，同一類別下MIS相關期刊[2]卻和電腦科學（Computer Science）類別下的軟體系統（software system）相關期刊聚集在一起。

上述研究的MIS相關期刊，在JCR（Social Science Edition）中除被指定為IS&LS類別以外，有部分也兼具Management類別，且大多數期刊的題名中包含information、management以及system等詞語，顯然可見這些期刊為圖書資訊學應用管理領域「資訊系統」（information system）的科際整合結果。根據Ni等（2013）和Tseng與Tsay（2013）使用的期刊相似程度評估方法推論，這些MIS相關期刊具有相似的文字內容或引用類似的文獻，所以一同被聚集成群，極可能都具有管理學方面的知識基礎。這些期刊中許多作者與編輯委員也屬於管理學領域，發表的論文也是代表某些管理學的研究前沿。也就是這些期刊的內容包含管理學主題，且被管理學研究人員認可為學科內的期刊。然而即便如Abrizah、Noorhidawati與Zainab（2015）進行期刊分類的描述偏好調查（stated preference study）時，有15位受試者認為MIS相關期刊所屬的資訊系統（information system）類不應放在IS&LS類別下，而應放在Computer Science或

[1] ISI於1960年由Eugene Garfield創立，1992年被Thomson Corporation 收購，2008年 Reuters Group併購Thomson Corporation後，目前ISI屬於Thomson Reuters的Intellectual Property & Science business，JCR與WoS都為該公司出版的資料庫。為行文方便，以下都以ISI稱呼。

[2] 在Boyack等（2005）中，以資訊管理（information management）相關期刊稱呼這些期刊。

Management類別下，但本研究也發現部分MIS相關期刊並沒有被ISI資料庫指定為Management類別，例如*Information System Journal*（inform syst j）、*Journal of the Association for Information Systems*（j assoc inf syst）、*Journal of Global Information Management*（j glob inf manag）等期刊，使用者在利用ISI資料庫查詢期刊資料時，在Management類別裡無法發現這些MIS相關期刊。因此，這種情形是否代表ISI資料庫目前對於這些MIS相關期刊的主題分類並不十分周密，收錄的期刊並未被正確地賦予所有可能的主題類別？

本研究即是分析IS&LS類別下的MIS相關期刊的主題特徵，並探討這些期刊的主題類別指定情形。本研究將以主題模型（topic modeling）方法（Blei, Ng, & Jordan, 2003）確認這些期刊在文字內容上共同具有的主題特徵，且利用兩種不同的機器學習（machine learning）方法探討MIS相關期刊的ISI主題類別指定情形。第一種方法依據期刊在主題特徵的相似程度，將IS&LS相關期刊聚集成群集，再根據各群集所屬期刊，找出MIS相關期刊的群集及其主題特徵。第二種方法則是應用分類迴歸樹方法（classification and regression tree）（Breiman, Friedman, Stone, & Olshen, 1984），分別使用被賦予Management類別的期刊以及群集分析產生的MIS相關群集做為正案例，產生分類規則，了解MIS相關期刊的關鍵主題特徵，並進行類別預測，比較兩者的分類規則與預測結果。第一種的期刊群集方法為非監督式（unsupervised）的機器學習方法，第二種的分類預測方法則屬於一種監督式（supervised）的方法。

本研究的研究問題包括：

㈠根據期刊內容的主題特徵，對IS&LS相關期刊進行群集，MIS相關期刊是否能聚集成群？ MIS相關期刊若能聚集成群，該期刊群集的主題特徵為何？本研究產生的MIS相關期刊群集與先前Tseng與Tsay（2013）等研究的MIS相關期刊群集有何異同？

㈡利用期刊的主題特徵與分類迴歸樹方法，根據期刊同時被指定到Management類別的情形以及本研究產生的MIS相關期刊群集，分別產生分類規則，進行期刊類別預測。兩次試驗的分類規則有何異同？預測結果的比較為何？

二、相關研究

㈠ISI主題類別與其統計特性

ISI的主題類別是以期刊的題名與引用樣式（citation patterns）等訊息做為分類標準，由主觀的經驗法則（heuristic）對期刊進行分類而產生（Leydesdorff & Rafols, 2009）。而且除人工的目測檢視外，ISI也利用一個未曾公布的Hayne-Coulson演算法，計算類別與期刊在引用資料與被引用資料的相似程度，來協助指定期刊的主題類別（Pudovkin & Garfield, 2002）。

Klavans與Boyack（2006）、Boyack等（2005）和Rafols與Leydesdorff（2009）等研究曾以統計描述期刊被賦予的主題類別數量，這些研究結果都呈現：在ISI資料庫中，期刊被指定到多個主題類別的情形相當常見。Klavans與Boyack（2006）和Boyack等（2005）的研究，統計2000年SCIE和SSCI的7,121種期刊資料以及當年度的205個主題類別，平均每種期刊被指定到1.59個類別，其中被指定到單一類別的期刊有4,019種，同時被指定到兩種類別的期刊有2,225種，其餘877種期刊有超過兩種以上的主題類別。Rafols與Leydesdorff（2009）的資料則來自於2006年SCI和SSCI，主題類別共有220個，7,611種期刊，平均每種期刊被指定到1.56個類別。除了許多期刊被賦予多個主題類別的情形以外，Rafols與Leydesdorff（2009）並指出，ISI類別的期刊數目分布呈現對數常態分布（log normal distribution），也就是相對少數的類別擁有大量的期刊，許多類別卻只有少量期刊。雖然ISI分類上期刊數目的分布並不平均，但相較於Rosvall與Bergstrom（2008）和Blondel、Guillaume、Lambiotte與Lefebvre（2008）等演算法產生的分類結果，ISI分類分布不平均的情形仍較為緩和，前10個最多期刊的ISI主題類別僅占所有期刊總數的15%。

ISI資料庫期刊經常被指定到多個主題類別的情形，也影響到主題類別之間的引用資料分布。Rafols與Leydesdorff（2009）比較ISI分類以及Rosvall與Bergstrom（2008）和Blondel等（2008）等兩種演算法的期刊群集結果，發現2006年期刊引用資料發生在ISI主題類別之內與類別之間的比例（within/between categories ratio）約為3.10，較兩種演算法的結果低許多；利用餘弦指標測量兩兩類別之間引用樣式的相似性，則約有3.9%的餘弦值超過0.5，分別高出兩種演算法約6倍和10倍。顯示相較於演算法產生的期刊群集，ISI資料庫的主題分類在類別之間的邊界區分較模糊，在主題類別之間的期刊有較多引用情形，而不同主題類別的引用樣式也有較多重覆。Rafols與Leydesdorff（2009）指出，ISI分類中期刊被指定到多個主題類別，使得類別間有許多交叉連結（cross-connection）以及類別內期刊數量較平衡等情形，都是為了便利書目發現（bibliographic disclosure）的「索引者效應」（indexer effects），然而較不適用於分析科學傳播內的潛藏結構（latent structures）。

㈡ISI主題類別在書目計量學研究的應用

許多研究利用ISI資料庫的主題類別檢索相關學科的期刊，做為學科範疇，進行各種資訊計量學相關研究。例如：Ni等（2013）和Tseng與Tsay（2013）將IS&LS類別裡的期刊，依據其相似程度，聚集成群，進行圖書資訊學領域的研究主題識別（topic identification）或次領域區分（subfield delineation）。

主題類別也可應用在知識領域視覺化的研究，目前有兩種方式，一種是以主題類別為節點單位，產生科學映射圖（scientific mapping）探索科學研究的

全貌，如de Moya-Anegón等（2007）和Leydesdorff與Rafols（2009）、Zhang、Liu、Janssens、Liang與Glänzel（2010）等；另一種則是在進行期刊為節點單位的科學映射圖研究中，將ISI主題類別做為評估期刊間相似程度品質的參考基準，這方面的研究有Klavans與Boyack（2006）提出的區域準確性（local accuracy）和Boyack等（2005）的結構準確性（structural accuracy）。

也有研究利用被引用期刊（cited journals）的主題類別，做為期刊的特徵，估算期刊的知識基礎多樣性（diversity）（Porter & Rafols, 2009），或利用引用期刊（citing journals）在主題類別期刊分布計算被引用期刊之間的相似程度（Wang & Wolfram, 2015; Wolfram & Zhao, 2014）。

㈢期刊主題分類的改善建議與自動分類研究

既然ISI主題類別的設置目的是因應檢索的需要，Glänzel與Schubert（2003）特別針對研究評鑑（research evaluation）的需求，發展一種三個步驟的期刊分類方法，指定期刊論文的類別。步驟一：首先根據科學計量學者與專家的經驗，設置分類架構。步驟二：依據分類架構，對期刊進行分類。此時可以視多重指定（multiple assignment）的情況，調整分類架構。步驟三：若是各類別內核心期刊發表的論文，便可毫無疑義地將論文歸屬於期刊所屬的次類別；若期刊屬於多領域科學或無法歸類到較特定的次類別，對這些期刊上發表的論文進行分類時，便個別根據其參考文獻處理。Glänzel與Schubert（2003）在此提出15個主類別、67個次類別及一個多領域科學（multidisciplinary science）的分類架構。

除以專家主觀的意見產生類別，然後進行分類外，由於許多研究認為這些方法在客觀性與合理性等方面有所不足，因此嘗試提出主題類別的自動產生方式。這些方法大多先計算期刊在特徵上的相似程度，然後利用群集演算法（clustering algorithms）將特徵相似的期刊聚集成群，也就是使用期刊群集的方式發現學科（disciplines）、次學科（sub-disciplines）或專業（specialties）。使用的期刊特徵大多為期刊之間的交互引用（cross citations）、共被引（co-citations）與書目耦合（bibliographic coupling）等引用資料為基礎的資訊，也有利用期刊內容上出現的詞語等文字資訊做為期刊特徵，或整合引用與文字資訊，Janssens、Zhang、De Moor與Glänzel（2009）比較了上述各種資訊進行期刊群集的成效。期刊相似性的測量方式則包括著名的餘弦、Jaccard、Pearson相關係數（Pearson's correlation coefficient）等測量方式，Boyack等（2005）比較5種利用交互引用以及3種利用共被引的期刊相似性測量方式，共計8種不同方式在區域準確性、結構準確性、可擴展性（scalability）以及最後的群集品質等成效。

期刊群集應用的演算法，除了使用一般常見的k-means叢集演算法（Boyack et al., 2005），或凝聚式階層演算法（Janssens et al., 2009）之外，Leydesdorff

(2006)利用因素分析法分解彙整期刊對期刊引用關係形成的引用矩陣，發現期刊間引用關係的潛藏結構，以產生的因素視為被引用期刊的主題群集。由於近年網路分析(network analysis)的技術與應用都有大幅提升，許多研究將期刊之間的相似性關係視為網路上的連結，從各種網路型態特徵分解網路，進行群集分析，例如Leydesdorff(2004)利用Pearson 相關係數測量期刊在被引用上的相似程度，以這項資訊建構期刊網路，並將網路上位在同一個雙重連結成分(bi-connected components)上的節點，視為可能的期刊群集。Samoylenko、Chao、Liu與Chen(2006)則利用餘弦指標計算期刊引用模式的相似性，利用這項資訊建構期刊的最小生成樹(minimum spanning trees)，然後從產生的最小生成樹，逐一刪除較不相似的期刊連結，找出可能的期刊群集。Rafols與Leydesdorff(2009)在比較ISI主題分類、Glänzel與Schubert(2003)分類與期刊群集演算法的研究中，所使用的快速展開法(fast unfolding)(Blondel et al., 2008)和隨機漫步法(random walk)(Rosvall & Bergstrom, 2008)等兩種群集演算法也是根據期刊網路的型態特徵，藉由盡量使類別內的引用(within-category citation)對類別間的引用(between-category citation)的比率最大化來發現網路上密集的期刊群集。Chen(2008)則分析期刊網路上各對期刊間的相似性，利用親和性傳導法(affinity propagation method)，尋找期刊網路上的代表期刊以及其相對應的期刊群集，盡量使期刊群集內的代表期刊與其他成員之間的距離最小化。Chen(2008)使用的期刊資料包括2001年的SCI上影響係數大於1的期刊以及2005年的SSCI期刊；前者共包括1,905種期刊，後者則有1,578種期刊。親和性傳導法產生期刊群集大致符合ISI的主題類別，然而群集裡所有成員的平均距離明顯比相對應的ISI類別還要小，顯示期刊群集內成員彼此間有極高的相關性(relatedness)。

除了針對資料庫內多個學科的期刊進行群集分析以外，與本研究更相關的群集分析研究是針對圖書資訊學的相關期刊，區分出這個學科的次學科與專業，也就是學科的主題識別。如先前所述Ni等(2013)和Tseng與Tsay(2013)都利用ISI資料庫內指定於IS&LS類別的期刊為分析範圍。Ni等(2013)利用IS&LS類別裡的期刊，以期刊-作者耦合(venue-author coupling)、共被引分析、主題分析(topic analysis)、編輯委員會連鎖分析(interlocking editorial board membership)等四種方式分別進行期刊間的相似程度評估，並利用凝聚式階層群集演算法(agglomerative hierarchical clustering algorithm)區別出期刊群集。Ni等(2013)的四種方法產生的結果相當一致，都包括MIS、IS、LS及傳播相關期刊等群集，而MIS相關期刊的群集明顯與其他群集遠離。Tseng與Tsay(2013)使用的研究資料為2000到2004年的50種IS&LS相關期刊與2005到2009年的66種IS&LS相關期刊。以期刊引用的文獻做為特徵，利用Dice係數評估期刊之間的書目耦合資訊，然後以完全連結(complete linkage)的凝聚式階層群集法與側影輪廓指標(silhouette index)等方法進行多階段群集(multi-stage clustering)。

2000到2004年的群集結果有IR（資訊檢索）、MIS（管理資訊系統）、SM（科學計量學）、AL（學術圖書館）、ML（醫學圖書館）、CD（館藏發展）等次學科；2005到2009年的群集結果除了上述的次學科外，還包含兩個較小的群集：OA（公開取用）與RL（區域圖書館）。且類似於Ni等（2013）的結果，MIS相關期刊與其他期刊分離。Tseng與Tsay（2013）並進一步對群集分析的結果進行面向分析（facet analysis），以多樣性指標（diversity index）分析期刊群集的特性，揭露出圖書資訊學的部分次領域具有地區性。

三、研究方法與資料

㈠研究資料

本研究先從2013年的JCR（Social Science Edition）獲得指定到IS&LS類別的期刊資料，此一類別下共84種期刊。然後根據期刊題名檢索WoS引文資料庫，取出IS&LS的相關期刊在2007到2013年出版的論文資料。由於本研究需要利用題名與摘要等文字資料產生主題特徵，因此扣除在WoS資料庫中摘要資料過少，甚至缺乏的5種期刊後，共使用79種期刊進行研究，這些期刊如本研究附錄所示。需要說明的是，為簡化研究過程，目前本研究並不處理期刊題名更動的情形，不同題名的期刊便視為不同期刊，例如*Libraries & The Cultural Record*自2012年起改名為*Information & Culture*，在此視為兩種期刊。進行研究時，彙整各種期刊每個年度的論文題名與摘要等文字內容成為一篇文件，共計得到512篇文件。由於WoS引文資料庫收錄各期刊的起始年度並不相同，各期刊的資料年份，亦即本研究使用該期刊的文件數，同樣可參見本研究附錄。本研究附錄也呈現2013年的IS&LS類別相關期刊同時被賦予JCR（Social Science Edition）其他主題的類別，除Management外，少數IS&LS類別相關期刊也有被指定為Communication、History of Social Sciences、Social Sciences、Interdisciplinary等類別。

本研究將文件分為訓練文件及測試文件，以2007到2012年的434篇文件做為訓練資料，2013年的文件則為測試資料。由於*Libraries & The Cultural Record*改名*Information & Culture*，因此測試文件共計78篇。訓練文件用於推導主題模型的參數、產生IS&LS相關期刊的主題特徵、進行期刊群集，以及建立分類規則；在建立分類規則之後，測試文件則用於預測文件的類別。圖1是訓練文件及測試文件取得與處理的示意圖，以下分別介紹期刊的主題模型推導與特徵抽取、集群分析與類別預測等處理。

㈡期刊的主題模型推導與特徵抽取

本研究利用主題模型方法，分析期刊文字內容蘊含的主題，做為期刊的特

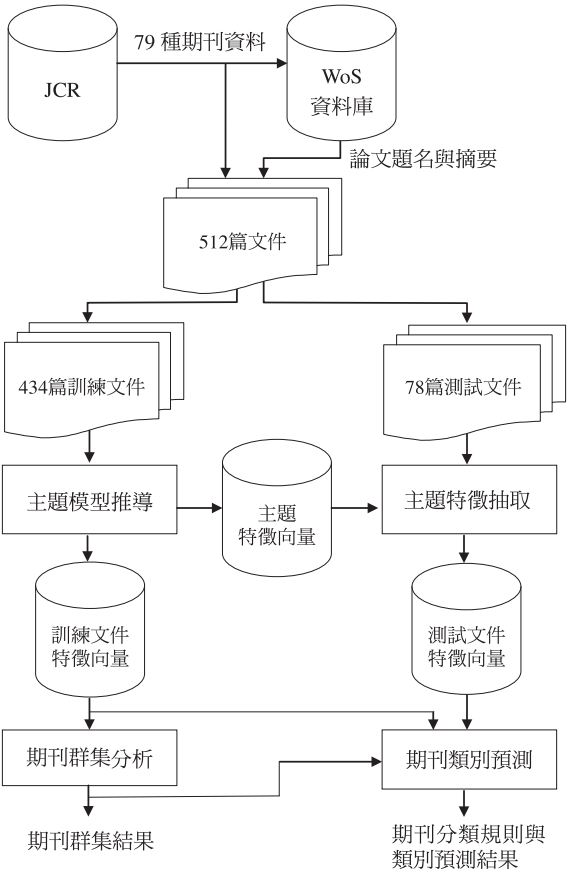

圖1 訓練文件及測試文件取得與處理示意圖

徵。主題模型方法使用LDA（latent Dirichlet allocation）統計模型描述文件產生的過程，揭露文件集合中可能出現的各個主題以及這些主題出現在每筆文件上的可能性。這個模型假設每筆文件都包含多個主題，可以利用機率混合（probabilistic mixture）來表示文件內的主題組成比例，且每一個主題都是由一組詞彙上的詞語依據不同的出現機率組成（Blei et al., 2003）。較簡易的說法是主題模型分別以一組主題特徵向量ϕ_k和文件特徵向量θ_d表示詞語出現在第k個主題的機率混合和各主題出現在文件d文字內容上的機率混合。

本研究將訓練文件輸入主題模型方法，利用Gibbs演算法（Griffiths & Steyvers, 2004）推測最佳的主題數以及主題特徵向量與文件特徵向量。決定主題數的條件是使用10次交叉驗證（10-fold cross validation）估計在不同主題數下的混淆度（perplexity）[3]平均值，獲得混淆度平均值最低的主題數便是最佳的結

[3] 在某一個模型之下的混淆度是指以該模型預測文本的平均候選詞語數，混淆度愈低者代表該模型愈好。

果。最後，再以最佳主題數做為參數，以所有的訓練文件推測主題特徵向量與文件特徵向量，判斷各個主題的內容意義以及各文件內的主題組成比例。並也以主題特徵向量推估測試文件的特徵向量，做為期刊類別預測的資料。

㈢期刊群集分析

在利用各期刊的訓練文件推導主題模型後，利用模型中的文件特徵向量，進行期刊群集。本研究假定對期刊而言，每一年度的文件都同樣重要，期刊的特徵向量便是主題出現在期刊文字內容上的機率混合，可定義為該期刊所有訓練文件特徵向量的平均，代表各主題特徵在期刊上的重要性。

然後根據期刊特徵向量測量期刊彼此間在內容主題上的相似程度。由於期刊的特徵向量為每一個主題出現在期刊文字內容上的機率混合，本研究利用對稱式Kullback-Leibler差異（symmetric Kullback-Leibler divergence）（Rzeszutek, Androutsos, & Kyan, 2010）計算每一對期刊在特徵向量上的差異，期刊之間的對稱式KL差異值愈大，在內容主題上便愈不相似。因此，以對稱式KL差異的負值做為期刊之間的相似程度。

接著利用期刊之間的相似程度進行期刊群集。由於親和性傳導演算法（Frey & Dueck, 2007）具有較低的誤差以及較高效率等優點，因此本研究選擇此演算法做為期刊群集方法。親和性傳導演算法將資料以及資料之間的相似程度視為網路的節點與連結強度，使得資料群集的問題成為在網路上尋找可代表一群節點的範例（exemplars）。某一個節點a能否被節點b代表的可能性由b做為a代表的承擔程度（responsibility）以及a可被b代表的合宜程度（availability）的總和決定。如果某一個節點a，其承擔程度和合宜程度總和最大的節點，正好為節點a本身，則這個節點可做為代表一個群集節點的範例；如果為另一個節點b，則b可做為代表a的範例。依據這樣的方式，找出期刊的群集。

利用親和性傳導演算法產生期刊群集後，計算期刊群集的側影輪廓指標。側影輪廓方法計算每一種資料與同一群集其他成員的平均差異，並也計算資料與最容易混淆其他群集的平均差異。當資料與同一群集的平均差異小於最容易混淆其他群集的差異時，該資料的側影輪廓寬度（silhouette width）為正值，反之為負值。一個群集的平均側影輪廓寬度為群集成員側影輪廓寬度。透過側影輪廓指標可以檢查群集品質的優劣，且找出兩個群集之間容易混淆的期刊。

最後，本研究再將群集內所有期刊的特徵向量進行平均，分析期刊群集的重要主題特徵。

㈣期刊類別預測

本研究進行兩次期刊類別預測試驗，分別使用IS&LS同時被賦予Management類別的期刊與期刊群集結果中判斷為MIS相關群集的成員為正案例。將訓練文件中所有正負案例的特徵向量，輸入分類迴歸樹（Breiman et al., 1984），產

生分類規則。分類迴歸樹是一種二元樹（binary tree），每個節點為一群資料的集合，上層節點的資料依據某一分類規則分為兩群，產生下層節點。一般而言，分類迴歸樹的建立分為兩階段。第一階段，首先將所有資料放置於二元樹的根節點（root node）內，選擇能盡量使資料區分為兩群後各自具有相同類別的變數做為此節點的分類規則。分成的兩群資料分別成為新的節點，如上述的方式再繼續將兩群資料各自分群。但若區分後節點內的類別相同情形改善有限、節點內的資料數量過少或階層數過大，便不再區分節點的資料。當所有節點都無法再進行區分，便完成第一階段。第二階段，為了避免過度適配（overfitting），從下而上，選擇合適的節點進行刪減，也就是當某個節點的區分可能導致未來的測試文件有較高的錯誤分類風險時，將以下階層的節點予以刪減。

然後將測試文件的特徵向量輸入產生的分類迴歸樹，預測及討論測試文件被分類為正方或負方的情形。分類迴歸樹的預測方式為從根節點開始，將輸入的資料與目前所在節點上的規則進行比對，根據比對結果進入下一層的節點，一直到達葉節點（leaf node）為止。預測的結果為葉節點上最多訓練資料所屬的類別。

最後，比較前後兩次試驗產生的分類規則與預測結果，特別著重於預測錯誤案例的分析。

四、研究結果

㈠IS&LS相關期刊的主題特徵

本研究首先利用主題模型方法分析期刊的文字資料，產生IS&LS相關期刊的主題特徵。根據2007到2012年的期刊內容資料為訓練文件，估計主題模型的各種參數。計算各種不同主題數在10次交叉驗證方法下的平均混淆度，以包含10個主題的主題模型獲得最佳的結果，因此本研究便將主題數設定為10。表1為本研究確認的10個主題上前20個出現機率最大的詞語。為了計算詞語出現的次數，本研究將題名與摘要等文字資料統一為小寫字，因此表1上出現的縮寫詞和專有名詞均為小寫，包括lis、ict、gis、spanish、european等等，並也將連字號（hyphen）去除，所以表1上的egovernment與hindex分別應是e-government與h-index。

表1的最下一列根據出現機率最大的詞語語意將各主題命名，例如T1因包含literacy、librarians、school、teaching等詞語，命名為“school library”，而T2包含firms、trust、product等管理學領域常見詞語，故將其命名為“management”。大多數的主題在命名時並不會混淆，雖然某些主題之間有相同的詞語，例如：T1和T7上都包括librarians和databases，T2和T9上都包括virtual和trust，T4和T8上則都有care、patients和intervention。這些重複出現的詞語，有

表1　IS&LS相關期刊10個主題，每個主題前20個出現機率最大詞語

T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
literacy	firms	spatial	literacy	egovernment	retrieval	librarians	clinical	virtual	citation
librarians	trust	geographic	cancer	mobile	classification	collections	medical	production	citations
behaviour	security	gis	care	ict	query	publishing	care	capital	papers
school	firm	urban	risk	broadband	semantic	scholarly	patients	participation	publications
lis	product	map	patients	telecommunications	document	preservation	patient	political	publication
seeking	intention	algorithm	news	sector	text	document	medication	action	indicators
teaching	capabilities	land	messages	divide	queries	books	records	theories	bibliometric
faculty	risk	maps	women	political	relevance	records	physicians	argue	index
student	virtual	objects	campaign	competition	engines	materials	record	cognitive	universities
reading	customer	accuracy	adults	citizens	searching	publishers	informatics	interactions	scientists
respondents	outsourcing	geographical	message	diffusion	task	repositories	healthcare	discourse	cited
instruction	team	scale	behaviors	regulatory	image	bibliographic	hospital	privacy	patent
universities	companies	modeling	attitudes	infrastructure	tasks	metadata	biomedical	intellectual	fields
teachers	customers	distance	seeking	regulation	algorithm	archives	measurements	ethical	sciences
searching	satisfaction	uncertainty	respondents	rural	indexing	book	alerts	collective	hindex
schools	teams	treatment	age	governments	ontology	print	drug	construction	output
job	acceptance	simulation	smoking	investment	engine	spanish	disease	meaning	patents
librarian	enterprise	space	prevention	providers	clustering	format	intervention	european	productivity
satisfaction	employees	error	older	wireless	ranking	databases	objectives	trust	indicator
databases	erp	geospatial	intervention	markets	metadata	copyright	errors	concerns	distribution
school library	management	geographic information	health information	e-government and telecommunications policy	information retrieval	publications and collections	medical informatics	communities and social networks	scientometrics and informetrics

部分詞語則代表這個領域上經常出現的概念，例如document、searching和metadata等，有部分是因主題之間彼此相關，所以使用相同詞語。

㈡期刊群集結果

以親和性傳導演算法產生期刊群集，結果共分為8個期刊群集。各期刊所屬的期刊群集，可參見附錄一。最大的期刊群集為第5個期刊群集（簡稱為C5，以下類推），共包含20種期刊，最小的期刊群集只包含3種期刊，共有2個群集，分別是C3和C7。對產生的期刊群集計算側影輪廓指標，結果如圖2。圖2上，每一條橫線代表一種期刊的側影輪廓寬度，也就是它與最容易混淆群集的平均差異。如果橫線在原點（0）右邊，期刊的側影輪廓寬度為正值，表示該期刊與同一群集的平均差異小於最容易混淆其他群集的差異，也就是該期刊比較靠近被歸類的群集；在原點左邊的橫線，其對應的期刊較接近其他群集。期刊的側影輪廓寬度值愈大，愈不容易與其他群集混淆。這些側影輪廓寬度並按照期刊所屬的群集以及它們的大小排列，例如第一群的9條橫線便是被歸類於C1的期刊的側影輪廓寬度，而圖2上也呈現C1的平均側影輪廓寬度為0.09。比較各群集的平均側影輪廓寬度，最大者為C7與C5。圖2的結果也顯示絕大多數期刊的側影輪廓寬度大於0，也就是大多數期刊比較靠近所屬的群集，只有少數期刊較接近其他群集，這些期刊分布在C1、C2與C8。所有期刊整體的平均側影輪廓寬度則為0.45。

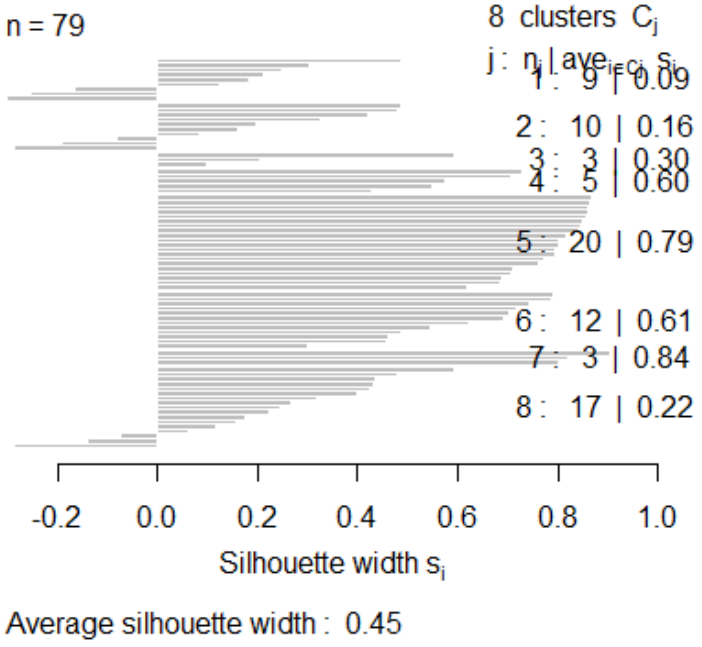

圖2 本研究期刊群集結果的側影輪廓分析

觀察每一個期刊群集內的期刊，在2013年的JCR（Social Science Edition）裡IS&LS類別下同時被指定為Management類別的期刊全部都在群集C5內。群集C5共有20種期刊，期刊題名可參見表2，其中同時被指定到Management類別的11種期刊，列於表2的左欄，未被指定的期刊則列於右欄。此外，群集C5內期刊的題名大多包含“information”、“system”、“management”、“technology”等關鍵詞。

表2　本研究期刊群集結果C5群集所包含期刊

同時指定到IS&LS與Management的期刊	只指定到IS&LS的期刊
inform manage-amster	data base adv inf sy
inform organ-uk	eur j inform syst
inform syst res	inform syst j
inform technol manag	inform technol peopl
j inf technol	int j inform manage
j knowl manag	j assoc inf syst
j manage inform syst	j glob inf manag
j organ end user com	j glob inf tech man
knowl man res pract	j strategic inf syst
mis q exec	
mis quart	

參考圖2，C5具有很高的平均側影輪廓寬度（0.79），且C5內每種期刊的側影輪廓寬度都在0.6以上，也就是這個群集的成員彼此間的差異比它們和群集外其他期刊的差異來得小，C5大部分的期刊最容易混淆的群集為C2，共有15種，其餘5種期刊則容易與C4混淆。

查看群集外訓練文件中其他期刊的側影輪廓寬度，容易與C5混淆的期刊全都在C2內，表3列出這些期刊與它們的側影輪廓寬度，其中以*Ethics and Information Technology*（ethics inf technol）的側影輪廓寬度值最小，其次為*Social Science Information*（soc sci inform）。兩種期刊的側影輪廓寬度值都為負數，代表它們與C5期刊平均上更接近於原先指定C2群集。

表3　訓練文件中易與C5群集混淆的期刊及其側影輪廓寬度

易與C5混淆的期刊	側影輪廓寬度
ethics inf technol	–0.29
gov inform q	0.33
inform soc	0.49
inform technol dev	0.16
soc sci comput rev	0.42
soc sci inform	–0.08
telecommun policy	0.08
telemat informat	0.48

將每一群集下所包含期刊特徵向量進行平均，呈現出各主題在期刊群集的出現機率。8個期刊群集的主題出現機率分布情形，如圖3所示。圖3可發現某些期刊群集偏重在某一主題上，例如：C4、C5、C7和C8；某些期刊的主題分布則集中在少數主題上，例如C1、C2、C3和C6。主題T9在所有的期刊群集上都是重要主題，顯然communities and social networks（社群和社會網絡）是目前大多數IS&LS相關期刊關注的主題。以C5來說，最為重要的主題毫無意外地是T2 management，其次是T9 communities and social networks，其他主題的出現機率都不高。而C5期刊容易混淆的群集C2和C4，其重要的主題分別是T5 e-government and telecommunications policy（電子化政府和電信傳播政策）以及T6 information retrieval（資訊檢索）。

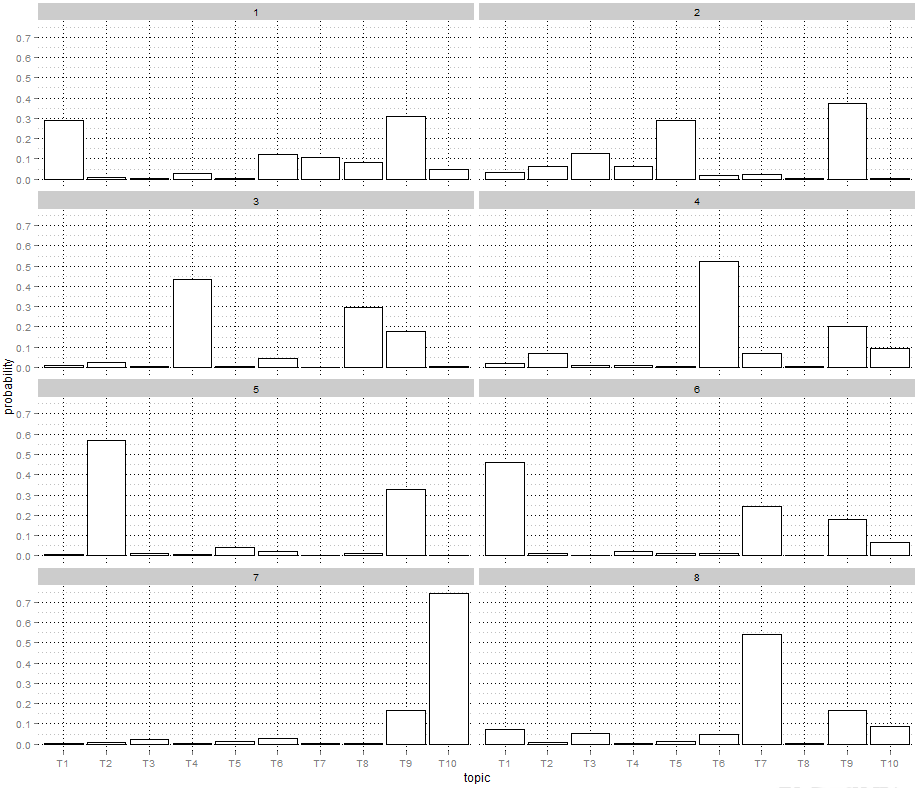

圖3 各期刊群集的主題出現機率分布情形

㈢群集結果比較

本研究產生的8個群集中，因被指定為Management類別的期刊都在C5內，可視為與MIS最為相關，以下將C5的期刊與先前研究的MIS相關期刊群集進行比較。考慮Ni等（2013）、Tseng與Tsay（2013）和Abrizah等（2015）等針對

IS&LS類別下期刊群集與分類的研究，選擇期刊資料時間範圍較接近的Tseng與Tsay（2013）的Set 2（2005~2009年）群集結果，以及Abrizah等（2015）針對2011年JCR的IS&LS類別下的期刊進行分類的描述偏好調查，做為比較對象。

Tseng與Tsay（2013）的群集結果中，Set 2的期刊群集結果共分為9個群集。根據從各期刊群集論文題名與摘要抽取的描述語，第2個群集的期刊與MIS最為相關，因此本研究將Tseng與Tsay（2013）Set 2的第2個群集與C5進行比較，且為解說方便，以下稱為Tseng與Tsay（2013）MIS群集。C5與Tseng與Tsay（2013）MIS群集兩者皆出現的期刊共有10種，Tseng與Tsay（2013）MIS群集獨有者有6種期刊，只出現在本研究的C5者則有10種期刊，兩個群集的差異如表4所示。

表4　Tseng與Tsay（2013）MIS群集與本研究C5群集差異比較

Tseng與Tsay（2013）MIS群集獨有	本研究C5獨有
gov inform q	data base adv inf sy
inform soc	eur j inform syst
int j geogr inf sci	inform organ-uk
j comput-mediat comm	inform technol peopl
soc sci comput rev	j glob inf tech man
telecommun policy	j knowl manag
	j organ end user com
	j strategic inf syst
	knowl man res pract
	mis q exec

Abrizah等（2015）的期刊分類描述偏好調查（stated preference study）將IS&LS類別下的83種期刊分為圖書館學、資訊科學與資訊系統等三類。MIS相關期刊被多數受訪者指定在「資訊系統」類下。該類期刊共有21種，在以下的比較，將它們稱為Abrizah等（2015）ISys。C5的期刊與Abrizah等（2015）ISys相比較，兩者相同的期刊共有15種，只出現在Abrizah等（2015）ISys而為C5所無的期刊共有6種，C5則有5種期刊未出現在Abrizah等（2015）ISys名單內，兩者的差異可參見表5。

表5　Abrizah等（2015）的資訊系統類期刊與本研究C5群集差異比較

Abrizah等（2015）ISys獨有	本研究C5獨有
ethics inf technol	int j inform manage
inform technol dev	j glob inf tech man
int j comp-supp coll	j knowl manag
j comput-mediat comm	j organ end user com
soc sci comput rev	knowl man res pract
telecommun policy	

單獨出現在C5而沒有在Tseng與Tsay（2013）MIS群集或Abrizah等（2015）ISys內的期刊大多是較晚收入ISI資料庫，而不在Tseng與Tsay（2013）與Abrizah等（2015）的研究範圍，例如：*Journal of Global Information Technology Management*（j glob inf tech man）和*Journal of Organizational and End User Computing*（j organ end user com）。只出現在Tseng與Tsay（2013）MIS群集與Abrizah等（2015）的期刊，大多都歸類在C2，例如：*Social Science Computer Review*（soc sci comput rev）和*Telecommunications Policy*（telecommun policy），由側影輪廓分析得知，這些期刊最容易混淆的群集也大多都是C5。

㈣類別預測結果

將訓練文件中IS&LS主題下同時被賦予Management類別的58個正案例與其餘期刊文件，輸入分類迴歸樹，產生的結果如圖4：首先，在第一層的節點，依據T2的出現機率值，將376個負案例與58個正案例的訓練文件分為兩群：第二層的左邊節點包含T2的機率值小於0.2257的文件，共有327個負案例與1個正案例；右邊節點的文件，其T2的機率值大於或等於0.2257，共有49個負案例與57個正案例。接著，第二層的右邊節點仍然再以T2的出現機率值大小區分為兩群：第三層的左邊節點為T2的機率值小於0.6835的文件，共46個負案例與32個正案例，右邊節點T2的機率值大於或等於0.6835的文件，則有3個負案例與25個正案例。第三層的左邊節點則以T5的出現機率值做為區分條

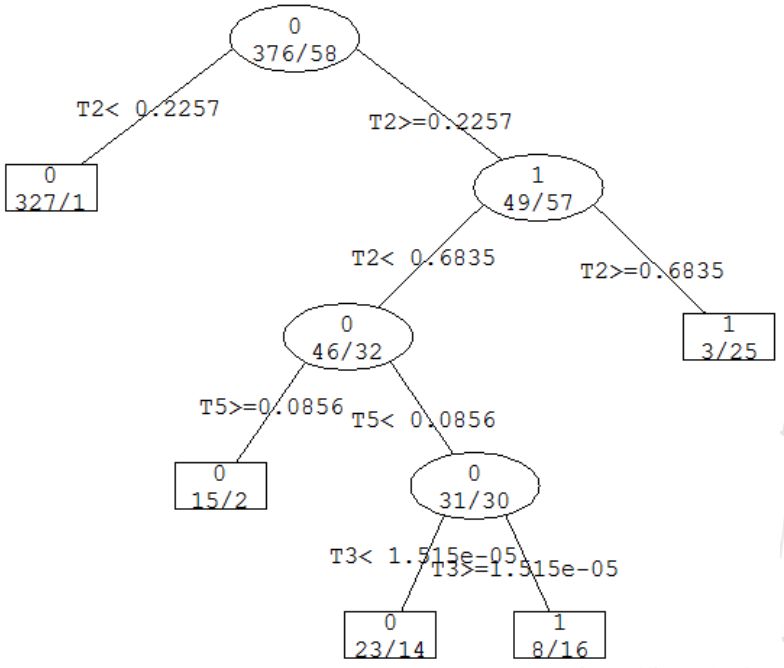

圖4　以被賦予Management類別期刊為正案例產生的分類迴歸樹

件：第四層的左邊節點為T5的機率值大於或等於0.0856的文件，共15個負案例與2個正案例，右邊節點T5的機率值小於0.0856的文件，則有31個負案例與30個正案例。最後，第四層的右邊節點以T3的出現機率值區分：第五層的左邊節點為T3的機率值小於$1.515*10^{-5}$的文件，共23個負案例與14個正案例，右邊節點T3的機率值大於或等於$1.515*10^{-5}$的文件，則有8個負案例與16個正案例。

以上述分類迴歸樹的分類規則預測78篇測試文件，共有10篇預測錯誤。表6是這些預測錯誤的結果，左邊的4種期刊為負案例卻預測為正方，也就是這些期刊並沒有被賦予Management類別，但分類系統根據分類規則判斷它們應被賦予Management類別；右邊的6種期刊則為正案例卻預測為負方。

表6　以被賦予Management類別期刊為正案例產生分類迴歸樹的預測錯誤結果

負案例卻預測為正方期刊	正案例卻預測為負方期刊
data base adv inf sy	inform organ-uk
int j inform manage	inform syst res
j glob inf manag	inform technol manag
j strategic inf syst	j inf technol
	knowl man res pract
	mis q exec

接下來，利用C5內的期刊為正案例，其餘期刊則為負案例，產生分類迴歸樹，進行期刊類別預測，與使用ISI的Management類別為正案例的分類迴歸樹與預測結果相比較。以C5內期刊為正案例，在訓練資料中共有105個正案例與329個負案例。此一分類迴歸樹僅有一條以T2的出現機率值為判斷標準的分類規則。T2小於0.2257，在訓練文件共有327個負案例與1個正案例，為負方；T2大於等於0.2257，在訓練文件中則共有2個負案例與104個正案例，為正方。將2013年的測試文件輸入此一分類迴歸樹，在78個測試文件中，僅有4個預測錯誤，表7是預測錯誤的結果。負案例卻預測為正方有2種，在表7的左邊。表7的右邊為正案例卻預測為負方的2種期刊。

表7　以C5內期刊為正案例產生分類迴歸樹的預測錯誤結果

負案例卻預測為正方期刊	正案例卻預測為負方期刊
inform dev	inform organ-uk
online inform rev	j inf technol

比較前後兩次類別預測試驗的分類規則與類別預測結果：前者並不容易區分IS&LS主題下的期刊是否同時被賦予Management類別；後者則能以相當單純的分類規則區分期刊是否被歸類於C5，且預測錯誤的情形較少。而表6的負案例卻預測為正方的期刊，在前面的期刊群集分析中都被歸類於C5，這也說明這些期刊與被賦予Management類別的期刊有相似的主題特徵，因此在類別預測試驗中相當容易混淆。

五、結　論

以往許多研究討論ISI資料庫的主題分類做為資訊計量研究或科學評鑑的應用，本研究從資訊檢索的目的分析與討論主題類別IS&LS的MIS相關期刊中同時被賦予Management類別的情形，藉由期刊文字內容蘊含的主題以及機器學習方法，一方面探討由這些主題特徵形成的期刊群集以及其重要主題，另一方面則討論沒有被指定到Management類別的MIS相關期刊是否也應賦予該類別，以提升檢索的成效。

本研究彙整期刊論文的題名與摘要做為期刊的文字內容，利用主題模型方法計算IS&LS類別的主題特徵以及每一主題在各期刊內容的出現機率。相較於以論文為單位獲得的主題特徵（林頌堅，2014a，2014b），以期刊為單位獲得的主題特徵更適合用於分析整個領域的主題分布情形。從IS&LS類別的主題特徵可發現其涵蓋的主題範圍相當廣泛，包括圖書館學相關的school library（學校圖書館）和publication and collections（出版與館藏），屬於資訊科學的information retrieval（資訊檢索）和scientometrics and informetrics（科學計量學與資訊計量學），以及和其他領域進行整合研究產生的management（管理）、e-government and telecommunications policy（電子化政府與電信傳播政策）、communities and social networks（社群與社會網路）、geographic information（地理資訊）、health information（健康資訊）和medical informatics（醫學資訊學）等主題。

在期刊群集方面，本研究與Ni等（2013）和Tseng與Tsay（2013）使用的期刊特徵、相似程度估算和群集方法皆有不同，並著重在分析MIS相關期刊在文字內容上的主題特徵以及在ISI的主題類別指定情形。本研究利用期刊的主題特徵和親和性傳導演算法，能將IS&LS類別下主題相近的期刊聚集成群。在形成的群集結果中，所有被指定到Management類別的期刊都被聚集在同一個期刊群集內，因此將這個群集視為MIS相關的期刊群集。這個群集的期刊和Tseng與Tsay（2013）與Abrizah等（2015）等研究的MIS群集大多相同。「管理」是這個MIS相關群集最突顯的內容主題。本研究MIS群集沒有包含但出現其它研究的MIS群集的期刊，則大多被歸類於另一個以「電子化政府和電信傳播政策」主題相關的群集，這意味著在引用的文獻或專家印象中，這兩個主題之間有某種程度的關連，但在文字內容上有所差異。

另一方面，目前監督式機器學習方法在圖書資訊學領域的應用則仍屬少見。本研究應用分類迴歸樹方法，分別使用ISI的Management類別以及本研究的MIS群集做為正案例，進行期刊類別預測，討論沒有被指定到Management類別的MIS相關期刊是否也應賦予該類別。兩次試驗產生的分類迴歸樹分別都以正案例最為明顯的主題特徵「管理」的出現機率值做為主要的分類規則。但前者為進一步區別在「管理」主題有較高的出現機率，但未被賦予Management類

別的期刊，在分類迴歸樹上需要加入更多其他主題特徵的分類規則，但也因此產生許多正案例卻被預測為負方的誤判；後者因群集分析能將所有「管理」主題特徵具有高出現機率的期刊聚集成MIS群集，所以用較簡單的分類規則便可區別屬於正負案例，使得預測錯誤的情況較少。也就是若將所有MIS群集內的期刊都指定到Management類別，在檢索時將能獲得所有在內容上有明顯「管理」主題特徵的期刊，會增加檢索MIS相關期刊時的效能與完整性。

以MIS群集內的期刊做為正案例產生的分類迴歸樹，仍會發生少數的類別預測錯誤。觀察預測錯誤期刊的編輯主旨，可發現這些期刊都是主題相當多元的科際整合期刊。負案例卻預測為正方的2種期刊：*Information Development*（inform dev）的收錄範圍包括資訊系統、服務與技能的發展以及資訊在個人與國家發展上的角色，涵蓋資訊傳播科技的提供、管理與使用的現況發展（SAGE, 2015）；*Online Information Review*（online inform rev）則提供資訊科學與資訊科技等相關領域的研究人員交流有關各種脈絡下的線上資訊研究，著重在線上系統、服務與資源等議題，特別是資訊的產生、管理、利用、傳佈與重組等過程與程序（Emerald, 2015）。從期刊主題出現機率的分布情形，也的確可發現這些期刊具有多樣的主題特徵。但從它們在2013年「管理」主題的出現機率明顯增加，甚至*Online Information Review*有一期特刊專門討論決策支援系統（decision support systems），MIS相關研究在2013年成為這些期刊較重視的論文。反之，正案例卻預測為負方的期刊，*Information and Organization*（inform organ-uk）的目的在因應資訊與通訊科技廣泛，且範圍逐漸增加的社會影響，探討資訊科技和社會組織（social organization）之間的關係（Elsevier B. V., 2015），*Journal of Information Technology*（j inf technol）同樣關注於資訊科技的策略、變革、基礎建設等組織、社會與管理相關的議題（Palgrave Macmillan, 2015）。從這2種期刊的主題出現機率分布情形觀察，2013年的主題都偏重於「社群與社會網路」。由於研究資料的限制，這些期刊研究主題改變的情形是否持續，需要再繼續觀察。

參考文獻

林頌堅（2014a）。以主題模型方法為基礎的資訊計量學領域研究主題分析。*教育資料與圖書館學*，*51*(4)，499-523。doi:10.6120/JoEMLS.2014.514/0633.RS.AM

林頌堅（2014b）。資訊科學期刊的主題分布與多樣性研究。*圖書資訊學研究*，*9*(1)，171-200。

Abrizah, A., Noorhidawati, A., & Zainab, A. N. (2015). LIS journals categorization in the Journal Citation Report: A stated preference study. *Scientometrics*, *102*(2), 1083-1099. doi:10.1007/s11192-014-1492-3

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *The Journal of Machine Learning Research*, *3*, 993-1022.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10). doi:10.1088/1742-5468/2008/10/P10008

Boyack, K. W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. *Scientometrics*, *64*(3), 351-374. doi:10.1007/s11192-005-0255-6

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. Belmont, CA: CRC Press.

Chen, C.-M. (2008), Classification of scientific networks using aggregated journal-journal citation relations in the Journal Citation Reports. *Journal of the American Society for Information Science and Technology*, *59*(14), 2296-2304. doi:10.1002/asi.20935

de Moya-Anegón, F., Vargas-Quesada, B., Chinchilla-Rodríguez, Z., Corera-Álvarez, E., Munoz-Fernández, F. J., & Herrero-Solana, V. (2007). Visualizing the marrow of science. *Journal of the American Society for Information Science and Technology*, *58*(14), 2167-2179. doi:10.1002/asi.20683

Elsevier B. V. (2015). *Information and Organization*. Retrieved from http://www.journals.elsevier.com/information-and-organization/

Emerald. (2015). *Online Information Review*. Retrieved from http://www.emeraldgrouppublishing.com/products/journals/journals.htm?id=oir

Frey, B. J., & Dueck, D. (2007). Clustering by passing messages between data points. *Science*, *315*(5814), 972-976. doi:10.1126/science.1136800

Glänzel, W., & Schubert, A. (2003). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics*, *56*(3), 357-367. doi:10.1023/A:1022378804087

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(Suppl. 1), 5228-5235. doi:10.1073/pnas.0307752101

Janssens, F., Zhang, L., De Moor, B., & Glänzel, W. (2009). Hybrid clustering for validation and improvement of subject-classification schemes. *Information Processing and Management*, *45*(6), 683-702. doi:10.1016/j.ipm.2009.06.003

Klavans, R., & Boyack, K. W. (2006). Identifying a better measure of relatedness for mapping science. *Journal of the American Society for Information Science and Technology*, *57*(2), 251-263. doi:10.1002/asi.20274

Leydesdorff, L. (2004). Clusters and maps of science journals based on bi-connected graphs in Journal Citation Reports. *Journal of Documentation*, *60*(4), 371-427. doi:10.1108/00220410410548144

Leydesdorff, L. (2006). Can scientific journals be classified in terms of aggregated journal-journal citation relations using the Journal Citation Reports? *Journal of the American Society for Information Science and Technology*, *57*(5), 601-613. doi:10.1002/asi.20322

Leydesdorff, L., & Rafols, I. (2009). A global map of science based on the ISI subject categories. *Journal of the American Society for Information Science and Technology*, *60*(2), 348-362. doi:10.1002/asi.20967

Ni, C., Sugimoto, C. R., & Cronin, B. (2013). Visualizing and comparing four facets of

scholarly communication: Producers, artifacts, concepts, and gatekeepers. *Scientometrics*, *94*(3), 1161-1173. doi:10.1007/s11192-012-0849-8

Palgrave Macmillan. (2015). About the journal. Retrieved from http://www.palgrave-journals.com/jit/about.html

Porter, A. L., & Rafols, I. (2009). Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. *Scientometrics*, *81*(3), 719-745. doi:10.1007/s11192-008-2197-2

Pudovkin, A. I., & Garfield, E. (2002). Algorithmic procedure for finding semantically related journals. *Journal of the American Society for Information Science and Technology*, *53*(13), 1113-1119. doi:10.1002/asi.10153

Rafols, I., & Leydesdorff, L. (2009). Content-based and algorithmic classifications of journals: Perspectives on the dynamics of scientific communication and indexer effects. *Journal of the American Society for Information Science and Technology*, *60*(9), 1823-1835. doi:10.1002/asi.21086

Rosvall, M., & Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, *105*(4), 1118-1123. doi:10.1073/pnas.0706851105

Rzeszutek, R., Androutsos, D., & Kyan, M. (2010). Self-organizing maps for topic trend discovery. *Signal Processing Letters, IEEE*, *17*(6), 607-610. doi:10.1109/LSP.2010.2048940

SAGE. (2015). *Information Development*. Retrieved from http://idv.sagepub.com/

Samoylenko, I., Chao, T.-C., Liu, W.-C., & Chen, C.-M. (2006). Visualizing the scientific world and its evolution. *Journal of the American Society for Information Science and Technology*, *57*(11), 1461-1469. doi:10.1002/asi.20450

Tseng, Y.-H., & Tsay, M.-Y. (2013). Journal clustering of library and information science for subfield delineation using the bibliometric analysis toolkit: CATAR. *Scientometrics*, *95*(2), 503-528. doi:10.1007/s11192-013-0964-1

Wang, F., & Wolfram, D. (2015). Assessment of journal similarity based on citing discipline analysis. *Journal of the Association for Information Science and Technology*, *66*(6), 1189-1198. doi:10.1002/asi.23241

Wolfram, D., & Zhao, Y. (2014). A comparison of journal similarity across six disciplines using citing discipline analysis. *Journal of Informetrics*, *8*(4), 840-853. doi:10.1016/j.joi.2014.08.003

Zhang, L., Liu, X., Janssens, F., Liang, L., & Glänzel, W. (2010). Subject clustering analysis based on ISI category classification. *Journal of Informetrics*, *4*(2), 185-193. doi:10.1016/j.joi.2009.11.005

附 錄

本研究分析的期刊

期刊縮寫	資料年份	論文數	該期刊於ISI資料庫同時被賦予主題類別	期刊群集結果
afr j libr arch info	7	98		6
aslib proc	7	251		1
aust acad res libr	6	102		6
aust libr j	5	99		6
can j inform lib sci	7	91		1
coll res libr	7	216		6
data base adv inf sy	5	84		5
electron libr	7	383		8
ethics inf technol	5	126	Ethic	2
eur j inform syst	7	257		5
gov inform q	7	371		2
health info libr j	7	203		1
inform cult	2	38	History of Social Sciences	8
inform dev	5	85		2
inform manage-amster	7	352	Management	5
inform organ-uk	6	72	Management	5
inform process manag	7	559		4
inform res	7	403		1
inform soc	7	144		2
inform soc-estud	6	149		8
inform syst j	7	139		5
inform syst res	7	301	Management	5
inform technol dev	5	88		2
inform technol manag	7	141	Management	5
inform technol peopl	5	90		5
int j comp-supp coll	7	141	Education & Educational Research	1
int j geogr inf sci	7	601	Geography	2
int j inform manage	7	390		5
interlend doc supply	7	189		8
investig bibliotecol	7	159		8
j acad libr	7	349		6
j am med inform assn	7	904		3
j am soc inf sci tec	7	1,265		4
j assoc inf syst	7	199		5
j comput-mediat comm	7	249	Communication	3
j doc	7	281		1
j glob inf manag	7	105		5
j glob inf tech man	4	48		5
j health commun	7	566	Communication	3
j inf sci	7	338		4
j inf technol	7	165	Management	5
j informetr	7	381		7
j knowl manag	5	293	Management	5
j libr inf sci	7	133		6

j manage inform syst	7	265	Management	5
j med libr assoc	7	189		1
j organ end user com	4	70	Management	5
j scholarly publ	7	125		8
j strategic inf syst	7	130		5
knowl man res pract	6	181	Management	5
knowl organ	7	168		4
learn publ	7	192		8
libr collect acquis	7	100		8
libr cult rec	2	15	History of Social Sciences	8
libr hi tech	7	317		8
libr inform sc	7	67		6
libr inform sci res	7	209		1
libr quart	7	104		6
libr resour tech ser	7	102		8
libr trends	7	304		6
libri	7	187		6
malays j libr inf sc	7	142		6
mis q exec	6	95	Management	5
mis quart	7	270	Management	5
online inform rev	7	353		4
portal-libr acad	7	148		6
prof inform	7	507		8
program-electron lib	7	174		8
res evaluat	7	218		7
restaurator	7	113		8
rev esp doc cient	6	159		8
scientometrics	7	1,374		7
serials rev	7	156		8
soc sci comput rev	7	250	Social Sciences, Interdisciplinary	2
soc sci inform	7	190	Social Sciences, Interdisciplinary	2
telecommun policy	7	441	Communication	2
telemat informat	3	99		2
transinformacao	6	113		1
z bibl bibl	7	74		8

Journal of Educational Media & Library Sciences 52 : 3 (Summer 2015) : 269-298
DOI:10.6120/JoEMLS.2015.523/0027.RS.AM

Research Article

A Study of the Subject Categorization of the MIS-related Journals in the ISI Databases Using Topical Features in the Text Content and Machine Learning Methods

Sung-Chien Lin

Abstract

In this study we analyzed and discussed that the MIS-related journals under the ISI subject category of IS&LS are simultaneously given with subject category Management, using methods of topic modeling, journal clustering and subject category prediction. In the experiment of journal clustering, all journals under subject category Management and other journals also having similar topical features can be gathered into a cluster, and "management" is their common and the most distinct topic. Because the journals belonged to this cluster are almost same to those in the MIS clusters generated by the previous studies, we considered it as the MIS cluster in this study. In the second experiment, we used the classification and regression tree (CART) technique to predict assignment of subject category with that the journals in the original subject category Management and in the MIS cluster produced in this study as positive examples, respectively. The trees generated by the two tests both used the occurring probabilities of the topic "management" as the main classification rule. However, in the latter test, we did not only obtain a simpler classification tree but also had a result with less predicting errors. This means that if all journals in the MIS cluster could be given with subject category Management, the retrieval results can be more effective and complete.

Keywords: *ISI subject category, Machine learning, Topic modeling, Journal clustering, Category prediction*

SUMMARY

Introduction

In the previous studies about cluster analysis of journals related to the field of Library and Information Science (LIS), such as the studies by Ni, Sugimoto, and Cronin (2013) and Tseng and Tsay (2013), the researchers usually used the journals under the subject category Information Science and Library Science (IS&LS) in the Institute for Scientific Information (ISI) databases as data for analysis. Most of them found that there were a few journals grouped into a unique

Assistant Professor, Department of Information and Communications, Shih Hsin University, Taipei, Taiwan
E-mail: scl@cc.shu.edu.tw

cluster apart from other journals and the most common theme of these journals was Management Information Systems (MIS). Several of the MIS-related journals were simultaneously given with another subject category Management in the ISI databases but a few of them were not. Thus when users request data of MIS journals from the ISI databases, they can't retrieve the whole set of data by using only the subject category Management as queries. Does it mean that the assignment of subject category Management to these MIS-related journals in the ISI databases were not comprehensive?

From the point of information retrieval, two experiments took place in this study to analyze and discuss the assignment of subject category Management to the MIS-related journals under the subject categories of IS&LS. The present study used two different machine learning techniques and it was based on topical features extracted from the text content in journals. The first experiment was the cluster analysis of IS&LS journals according to the topical features contained in the journals to explore the cluster structure of the examined journal set and important topical features emerging in each of the clusters. Cluster analysis is known as a kind of unsupervised learning methods and it had been widely used in the studies of LIS. In the second experiment, we used classification and regression tree (CART) (Breiman, Friedman, Stone, & Olshen, 1984), a technique for supervised learning, to predict the assignment of subject category Management to IS&LS journals. We then examined the MIS-related journals that are not currently categorized as Management journals and discussed if these journals should be in the Management category in order to improve retrieval effectiveness.

Methods

The research data of this study were bibliographic data of papers published in IS&LS journals retrieved from the Web of Science database with the search criteria that a) any journal title that is in the list of subject category IS&LS in 2013 JCR (Social Science Edition) and b) the publication year is between 2007 and 2013. Text data in the Title and the Abstract fields of the articles in the same year and in the same journal were combined into a document. The IS&LS journals without the Title and Abstract data in the database were dropped in this study. The documents between 2007 and 2012 were used as training data to estimate parameters of topic models, to generate the topical features for each journal, to perform cluster analysis, and to create classification trees for predicting the assignment of subject category Management to the journals. The remainders of the documents were then used as test data in the experiment of subject category prediction.

After preparing text data for training and testing, documents in the training data were firstly input to the method of topic modeling (Blei, Ng, & Jordan, 2003; Griffiths & Steyvers, 2004) to compute the topical features for each journal in the dataset. To each document a feature vector was assigned, which was composed of the estimated probabilities of every topics occurring in the corresponding document. Each topic in the study was also represented by a feature vector consisting of occurring probabilities of all word tokens when the topic appeared in documents.

In the experiment of journal clustering, a feature vector for each journal was computed by averaging the feature vectors of documents corresponded to the journals in the training data. Dissimilarity between any pairs of two journals was estimated by the symmetric Kullback-Leibler divergence (Rzeszutek, Androutsos, & Kyan, 2010) of the corresponded feature vectors. Clustering algorithm used in this study grouped journals with similar topical features based on the estimated dissimilarities between journals was the affinity propagation algorithm (Frey & Dueck, 2007). We also used the silhouette index to evaluate clustering quality and identify journals that were ambiguous between two clusters. Finally, the topical features of each cluster were obtained by averaging the feature vectors of journals belonged to the cluster.

In the experiment of subject category prediction, two tests were conducted. In the first test we used the IS&LS journals in the original subject category Management as positive examples input to the **classification and regression tree** (CART) algorithm. In the second test, the journals of the MIS cluster generated in the clustering experiment were used as positive examples. The generated classification trees as well as the prediction results of both tests were compared, with particular emphasis on the analysis of predicting errors.

Results

From the result of topic modeling in this study, we observed that the IS&LS journals cover a wide range of topics. There were two topics, "school library" and "publication and collections", belong to the Library Science discipline. There were two other topics, "information retrieval" and "scientometrics and informetrics", are known as important specialties in Information Science discipline. The remainders were the results created by the integration of Library Science and/or Information Science with other disciplines, such as "management", "e-government and telecommunications policy", "communities and social networks", "geographic information", "health information", and "medical informatics".

Ni et al. (2013) and Tseng and Tsay (2013) had also conducted experiments

of journal clustering in their study of topic identification or subfield delineation to the field of LIS. However, the features for representing journals, the methods for (dis)similarity estimation between journals, and the clustering algorithms used in this study and the previous two studies are different. In addition, the goals of this study were not only to identify the cluster composed of MIS journals, but also to expose the topical features emerging in the content of the journals and to analyze the journals' assignment of subject category Management. Using the affinity propagation algorithm, the IS&LS journals with similar topical features can be divided into groups. All the journals simultaneously that were assigned the subject category Management were sorted into the same cluster, and therefore, this cluster was considered as the MIS cluster in this study. The journals in this cluster and those in the MIS cluster in Tseng and Tsay (2013) and Abrizah, Noorhidawati, and Zainab (2015) were almost the same. The most distinct topical feature in the content of the journals in the cluster was "management". Those journals, which were included in the MIS cluster in the two previous studies but not in this study, were classified in another cluster related to the topic "e-government and telecommunications policy". It could mean that there were some relations between the two topics in both citation data and experts' images, but the texts in which the two topics appeared were very different.

Nowadays researches using supervised learning methods in the field of LIS are still rare. In this study, we used the techniques of CART to discuss whether the MIS journals which currently are not assigned the subject category Management should also be given with the subject category or not. The classification trees generated in the two experiments both used occurring probabilities of the topic "management", which was the most distinct topical feature in the positive examples of the training data, as the main classification rule to predict the assignment of subject category "Management" to the IS&LS-related journals. However, in the test of using the journals in the original subject category Management as positive examples, it needed to add some classification rules consisting of other topical features in order to exclude the journals which had also a higher occurring probability for the topic "management" but without the subject category "Management". These added rules introduced many predicting errors which resulted in the positive examples were predicted to be negative. In the test of using the journals in the MIS cluster generated in this study, the generated classification tree was much simpler and also brought less predicting errors. It was because the journals with higher occurring probability on the topic "management" were sorted into the MIS cluster. In summary, if the MIS cluster is used rather than the category of Management in the databases, the retrieval result of MIS journals will be more effective and complete.

ROMANIZED & TRANSLATED REFERENCE FOR ORIGINAL TEXT

林頌堅（2014a）。以主題模型方法為基礎的資訊計量學領域研究主題分析。教育資料與圖書館學，*51*(4)，499-523。doi:10.6120/JoEMLS.2014.514/0633.RS.AM【Lin, Sung-Chien (2014a). Analyses of research topics in the field of informetrics based on the method of topic modeling. *Journal of Educational Media & Library Sciences*, *51*(4), 499-523. doi:10.6120/JoEMLS.2014.514/0633.RS.AM (in Chinese)】

林頌堅（2014b）。資訊科學期刊的主題分布與多樣性研究。圖書資訊學研究，*9*(1)，171-200。【Lin, Sung-Chien (2014b). A study of topic distribution and diversity of journals in the field of information science. *Journal of Library and Information Science Research*, *9*(1), 171-200. (in Chinese)】

Abrizah, A., Noorhidawati, A., & Zainab, A. N. (2015). LIS journals categorization in the Journal Citation Report: A stated preference study. *Scientometrics*, *102*(2), 1083-1099. doi:10.1007/s11192-014-1492-3

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *The Journal of Machine Learning Research*, *3*, 993-1022.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10). doi:10.1088/1742-5468/2008/10/P10008

Boyack, K. W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. *Scientometrics*, *64*(3), 351-374. doi:10.1007/s11192-005-0255-6

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. Belmont, CA: CRC Press.

Chen, C.-M. (2008), Classification of scientific networks using aggregated journal-journal citation relations in the Journal Citation Reports. *Journal of the American Society for Information Science and Technology*, *59*(14), 2296-2304. doi:10.1002/asi.20935

de Moya-Anegón, F., Vargas-Quesada, B., Chinchilla-Rodríguez, Z., Corera-Álvarez, E., Munoz-Fernández, F. J., & Herrero-Solana, V. (2007). Visualizing the marrow of science. *Journal of the American Society for Information Science and Technology*, *58*(14), 2167-2179. doi:10.1002/asi.20683

Elsevier B. V. (2015). *Information and Organization*. Retrieved from http://www.journals.elsevier.com/information-and-organization/

Emerald. (2015). *Online Information Review*. Retrieved from http://www.emeraldgrouppublishing.com/products/journals/journals.htm?id=oir

Frey, B. J., & Dueck, D. (2007). Clustering by passing messages between data points. *Science*, *315*(5814), 972-976. doi:10.1126/science.1136800

Glänzel, W., & Schubert, A. (2003). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics*, *56*(3), 357-367. doi:10.1023/A:1022378804087

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(Suppl. 1), 5228-5235. doi:10.1073/pnas.0307752101

Janssens, F., Zhang, L., De Moor, B., & Glänzel, W. (2009). Hybrid clustering for validation

and improvement of subject-classification schemes. *Information Processing and Management*, *45*(6), 683-702. doi:10.1016/j.ipm.2009.06.003

Klavans, R., & Boyack, K. W. (2006). Identifying a better measure of relatedness for mapping science. *Journal of the American Society for Information Science and Technology*, *57*(2), 251-263. doi:10.1002/asi.20274

Leydesdorff, L. (2004). Clusters and maps of science journals based on bi-connected graphs in Journal Citation Reports. *Journal of Documentation*, *60*(4), 371-427. doi:10.1108/00220410410548144

Leydesdorff, L. (2006). Can scientific journals be classified in terms of aggregated journal-journal citation relations using the Journal Citation Reports? *Journal of the American Society for Information Science and Technology*, *57*(5), 601-613. doi:10.1002/asi.20322

Leydesdorff, L., & Rafols, I. (2009). A global map of science based on the ISI subject categories. *Journal of the American Society for Information Science and Technology*, *60*(2), 348-362. doi:10.1002/asi.20967

Ni, C., Sugimoto, C. R., & Cronin, B. (2013). Visualizing and comparing four facets of scholarly communication: Producers, artifacts, concepts, and gatekeepers. *Scientometrics*, *94*(3), 1161-1173. doi:10.1007/s11192-012-0849-8

Palgrave Macmillan. (2015). About the journal. Retrieved from http://www.palgrave-journals.com/jit/about.html

Porter, A. L., & Rafols, I. (2009). Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. *Scientometrics*, *81*(3), 719-745. doi:10.1007/s11192-008-2197-2

Pudovkin, A. I., & Garfield, E. (2002). Algorithmic procedure for finding semantically related journals. *Journal of the American Society for Information Science and Technology*, *53*(13), 1113-1119. doi:10.1002/asi.10153

Rafols, I., & Leydesdorff, L. (2009). Content-based and algorithmic classifications of journals: Perspectives on the dynamics of scientific communication and indexer effects. *Journal of the American Society for Information Science and Technology*, *60*(9), 1823-1835. doi:10.1002/asi.21086

Rosvall, M., & Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, *105*(4), 1118-1123. doi:10.1073/pnas.0706851105

Rzeszutek, R., Androutsos, D., & Kyan, M. (2010). Self-organizing maps for topic trend discovery. *Signal Processing Letters, IEEE*, *17*(6), 607-610. doi:10.1109/LSP.2010.2048940

SAGE. (2015). *Information Development*. Retrieved from http://idv.sagepub.com/

Samoylenko, I., Chao, T.-C., Liu, W.-C., & Chen, C.-M. (2006). Visualizing the scientific world and its evolution. *Journal of the American Society for Information Science and Technology*, *57*(11), 1461-1469. doi:10.1002/asi.20450

Tseng, Y.-H., & Tsay, M.-Y. (2013). Journal clustering of library and information science for subfield delineation using the bibliometric analysis toolkit: CATAR. *Scientometrics*, *95*(2), 503-528. doi:10.1007/s11192-013-0964-1

Wang, F., & Wolfram, D. (2015). Assessment of journal similarity based on citing discipline analysis. *Journal of the Association for Information Science and Technology*, *66*(6), 1189-1198. doi:10.1002/asi.23241

Wolfram, D., & Zhao, Y. (2014). A comparison of journal similarity across six disciplines using citing discipline analysis. *Journal of Informetrics*, *8*(4), 840-853. doi:10.1016/j.joi.2014.08.003

Zhang, L., Liu, X., Janssens, F., Liang, L., & Glänzel, W. (2010). Subject clustering analysis based on ISI category classification. *Journal of Informetrics*, *4*(2), 185-193. doi:10.1016/j.joi.2009.11.005