圖書資訊學研究 9：1 (December 2014)：171-200

資訊科學期刊的主題分布與多樣性研究

A Study of Topic Distribution and Diversity of Journals in the Field of Information Science

林頌堅

Sung-Chien Lin

世新大學資訊傳播學系助理教授
Assistant Professor
Department of Information and Communications
Shih Hsin University

【摘要 Abstract】

本研究利用資訊科學領域重要期刊的論文內容，運用主題模型方法，確認領域的主題，分析各種期刊重視的研究主題，並且將期刊在主題分布的差異視覺化。本研究使用的資料是資訊科學的六種期刊於 2007 到 2013 年發表論文的題名與摘要，運用主題模型方法所產生的 20 個主題，可以進一步歸類為資訊搜尋與檢索研究和資訊計量學兩個群組。雖然本研究的分析範圍和研究方法都與先前的研究不同，但主題確認的結果與先前研究類似，顯示資訊科學的研究主題發展大致上穩定。此外，主題確認完成後，本研究也利用各種指標測量期刊的主題多樣性。各種指標對期刊的主題多樣性測量結果相當一致，有較廣泛研究主題的 *Journal of the American Society for Information Science and Technology* 與 *Journal of Information Science* 具有較大的主題多樣性。

In this study, we used the text content of papers published in important journals of the field of Information Science as the data resource and the method of topic modeling to identify research topics in the examined field to analyze the important topics of each journal

投稿日期：2014 年 10 月 1 日；接受日期：2014 年 12 月 16 日
Email: 林頌堅 scl@cc.shu.edu.tw

and to map the journals according the disparities between their topic distribution into a graph. The input text data used in this study were the title and abstract of articles published between 2007 and 2013 in six important journals of the field of Information Science. The results of topic modeling were 20 topics, which can be further classified into two groups, "information search and retrieval" and "informetrics". Although the scope of analysis and the research methods in this study and previous studies are different, the results are similar. It means that the development of research topics in the examined field are generally stable. Based on the results of topics identification, we also measured the topic diversity of the journals using various indicators. The measure results of different diversity indicators are quite consistent. *Journal of the American Society for Information Science and Technology* and *Journal of Information Science*, which had broad distribution over research topics, both obtained a larger topic diversity among all the examined journals.

【關鍵字 keywords】

主題模型方法；研究主題確認；多樣性指標

Topic modeling; Topic identification; Index of diversity

壹、緒論

過去有許多資訊計量學研究以資訊科學領域做為分析對象，運用共被引（co-citation）資訊（Moya Anegón, Herrero Solana, & Jiménez-Contreras, 2006; Waltman, van Eck, & Noyons, 2010; White & McCain, 1998）或是內容的相似程度（Janssens, Leta, Glänzel, & De Moor, 2006; Lu & Wolfram, 2012），嘗試確認這個領域的研究主題[1]。大多數的研究選擇資訊科學做為分析領域的原因是以研究者本身較熟悉的領域做為技術發展的基礎，便於了解研究方法的可行性與優缺點，希望應用

1 Janssens et al. (2006) 和 Lu and Wolfram (2012) 使用圖書資訊學 (library and Information Science, LIS) 稱呼其分析的學科領域，但實際上這些研究所採用的期刊資料以資訊科學為主，而並不包含圖書館學 (Library Science) 的主要期刊。因此，本研究將其視為是資訊科學領域的研究。但在下文說明各研究的分析領域與研究成果時，仍以採用各自研究本身使用的名稱。

成功後，可以擴展到其他領域。雖然研究的方法與資料都有不同，並且對產生的主題使用不同的名稱，但大抵上可以發現資訊科學分為兩個次領域：資訊尋求與檢索研究和資訊計量學研究，前者包含資訊檢索技術與系統設計、資訊使用行為以及網路相關議題等主題，後者則有引用分析、研究成效評估以及科學映射等。例如，White and McCain（1998）分析了 1972 到 1995 年間 12 份資訊科學相關期刊內的作者共被引資料，產生八個資訊科學的專業（specialties），包括：1）實驗型檢索（experimental retrieval）；2）引用分析（citation analysis）；3）實務型檢索（practical retrieval）；4）書目計量學（bibliometrics）；5）一般圖書館系統理論（general library systems theory）；6）使用者理論（user theory）；7）科學傳播（scientific communication）與 8）OPAC。White and McCain（1998）進一步將這八個專業分為研究知識文獻以及其社會脈絡的領域分析（domain analysis）和研究人 - 電腦 - 文獻介面的資訊檢索等兩個次學科。Janssens et al.（2006）利用詞語共現分析（co-word analysis）以及 Ward's 階層叢集（Ward's hierarchical clustering）技術，將五種資訊科學相關期刊在 2002 到 2004 年間出版的論文區分為六個叢集，分別是資訊檢索、社會議題、網路相關議題、專利分析（patent analysis）以及一個主要為科學裡的合作關係（collaboration in science）、引用分析和國家研究成效（national research performance）的書目計量學叢集和另一個以方法學和書目計量理論相關的論文為主的書目計量學叢集。這六個論文叢集可以分為兩大群，一為資訊檢索、社會議題和網路相關議題，另一為專利分析和兩個書目計量學主題。Janssens et al.（2006）並且運用論文叢集，找出各期刊著重的主題，他們發現：*Scientometrics* 的論文主題主要為書目計量與專利分析。*Information Processing and Management (IPM)* 和資訊檢索叢集幾乎重疊，表示 *IPM* 上的論文和資訊檢索十分相關。*Journal of the American Society for Information Science and Technology (JASIST)* 有較多主題，包括資訊檢索、社會議題和網路相關議題。論文主題傾向於社會議題方面的期刊，除了有 *JASIST* 以外，還包括 *Journal of Information Science*（*J Info Scie*）和 *Journal of Documentation*（*J Doc*）。網路則是五種期刊都重視的議題。目前除了 Janssens et al.（2006）的研究，少有研究在確認主題之後，進一步分析各種主題在不同期刊上的分布，了解每一種期刊著重的主題，並且發現期刊之間在主題上的差異。

資訊科學的研究經常需要運用並融合多種不同知識來源，因此是一個在知識來源上深具多樣性（diversity）的學科。Bonnevie-Nebelong

（2006）和 Huang and Chang（2012）等研究嘗試利用論文的參考文獻來源，探討資訊科學相關期刊或次領域的知識來源多樣性。Bonnevie-Nebelong（2006）計算 *J Doc*、*JASIST* 以及 *J Info Scie* 等期刊在分析範圍內所有論文的參考文獻總數除以參考文獻來源的期刊種類，稱之為期刊的引用對被引用者比（citations/citee-ratio），用來了解與比較各期刊的知識來源多樣性。較低的引用對被引用者比表示期刊的參考文獻上出現許多不同種類的期刊，也就是使用的知識來源具有較高的多樣性。Huang and Chang（2012）利用 1978 到 2007 年間圖書館學與資訊科學各五種期刊論文引用的參考文獻，比較兩個資訊科學次領域的參考文獻來源的跨學科（interdisciplinary）特性。Huang and Chang（2012）依據美國國會圖書館的分類系統（Library of Congress classification），確認期刊論文引用的書籍與期刊等文獻所屬的學科，然後將圖書館學與資訊科學的期刊論文在各年度引用各種學科文獻的比例，輸入 Brillouin 指標（Brillouin's Index）測量跨學科性。結果發現這兩個次領域參考文獻來源的跨學科性都有逐年上升的情形，並且在每一年裡資訊科學的跨學科性都比圖書館學高。Bonnevie-Nebelong（2006）採用的引用對被引用者比，僅能表現多樣性概念中的種類數量面向（Stirling, 2007），Huang and Chang （2012）採用 Brillouin 指標做為測量方式，同時兼具來源的種類數量與平衡性（balance）等兩個多樣性面向的考量。並且 Huang and Chang（2012）的研究先確認引用文獻的來源學科，能夠深入了解的跨學科知識整合，但學科分類必須事先定義，可能產生一些問題，例如：因為學科分類可能隨時間改變，無法反應當時的學科界線；再者，研究上若是沒有適合的學科分類，將受到限制（Bache, Newman, & Smyth, 2013）。此外，上述兩個研究都利用論文的參考文獻代表知識來源，透過參考文獻的來源分布分析期刊或次領域在知識來源上的多樣性，並非從論文內容上出現的研究主題，探討期刊在研究主題上的多樣性。

Liu, Rafols and Rousseau（2012）提出一個可以廣泛應用於多樣性研究的分析架構。在此架構下，分析的類別可以定義是參考文獻所屬的學科，或是由一組關鍵詞語所定義的主題類別。當分析的類別是參考文獻的學科時，如同 Bonnevie-Nebelong（2006）和 Huang and Chang（2012）等研究，這個架構的目的是分析領域的知識來源多樣性；但是當使用關鍵詞語的組合做為分析的類別時，這個架構便能夠從論文內容分析期刊或者領域在研究主題上的多樣性。Bache et al.（2013） 運用主題模型方法提出測量單一論文的研究主題多樣性的研究架構，根

據論文集合內的文字內容推測可能的主題以及每一個主題出現在論文上的機率，以主題的出現機率輸入 Rao-Stirling 指標，計算論文的研究主題多樣性。由於缺乏研究主題多樣性的客觀評判標準，且主觀評判需要花費大量人力成本，無法實際上對大量真實出版的論文進行可行性研究。Bache et al.（2013）根據期刊主題的差異性，設計產生兩組虛擬文件（pseudo-documents），第一組虛擬文件由來自兩篇不同主題的期刊論文串接而成，第二組虛擬文件則是由相近主題的期刊論文串接。研究結果證實了假設預期的第一組虛擬文件的多樣性具有比第二組高。Bache et al.（2013）的研究架構是針對單一論文的多樣性測量，並且主要目的是探討此一研究架構的可行性，輸入的資料不是真實的論文資料。

根據上述的討論，本研究的目的包括：

1. 利用主題模型方法，從實際的論文內容確認資訊科學的研究主題，並進一步發現這個領域的主題結構；
2. 探討主題在各種期刊上的分布情形，並比較各種期刊在主題分布的差異；
3. 測量資訊科學相關期刊的研究主題多樣性。

從論文內容確認領域研究主題的方法可參見林頌堅（2014）對資訊計量學的研究。以下首先說明多樣性測量的相關研究，然後再呈現本研究的研究方法，敘述研究的結果並進行討論。

貳、多樣性測量的相關研究

過去關於多樣性的研究主要來自於跨學科研究（interdisciplinary research）。Rafols and Meyer （2010）從知識整合的觀點提出分析跨學科性的架構。在此一架構裡，多樣性是分析的研究個體（領域、期刊、研究機構等等）使用知識主題的廣度，是跨學科性重要的觀察面向之一。Rafols and Meyer（2010）引用 Stirling （2007） 的論點，認為多樣性是描述一個系統的元素如何被分配到各類別的特性。Stirling（2007）的多樣性概念包含三個面向：1）種類數量（variety）：元素被分配到的類別數量；類別愈多，多樣性應該愈大。2）分布的平衡性（balance）：元素分配到類別上的平衡性或均勻度（evenness）；如果元素在類別上的分布較平均的話，便表示系統具有較大的多樣性。3）分布類別的差距（disparity）或相似性（similarity）：元素分配到的類別彼此間差異愈大，系統的多樣性愈大。因此，Rafols and Meyer （2010）認為跨學

科論文的參考文獻應同時具備：所屬期刊分布於較多的類別、在類別上的分布較平均、主題分布的類別彼此間具有較大的差距等特性。

Porter and Rafols（2009）、Rafols and Meyer（2010）和 Leydesdorff and Rafols（2011）等探討跨學科性的研究曾提出 Shannon 熵（entropy）值、Simpson 指數[2]、Rao-Striling 指標以及 Gini 係數（Gini coefficient）等幾種指標測量主題的多樣性，以下說明各種指標的計算方式：

1. Shannon 熵值 H 用來計算分佈上的不確定性（uncertainty），其計算如式（1）：

$$H \stackrel{\text{def}}{=} -\sum_{i=1}^{N} p_i \log(p_i) \qquad (1)$$

N 是系統中的類別數量，p_i 是元素在第 i 個類別上的數量比例。當元素分散於多個類別並且數量分布較均勻時，式（1）的計算便會得到比較大的熵值。在極端的情形下，系統上的元素僅有分布在一個類別上時，熵值將為 0；反之，若是元素平均分布在 N 個類別上，此時熵值為 log（N）。

2. 以 Simpson 指標 I 計算元素在類別上分布的多樣性，其計算如式（2）：

$$I \stackrel{\text{def}}{=} \sum_{i=1}^{N} \sum_{j=1\ (i\neq j)}^{N} p_i p_j = 1 - \sum_{i=1}^{N} {p_i}^2 \qquad (2)$$

p_i 和 p_j 分別是所有元素在類別 i 和 j 上的數量比例。當元素集中在一個類別時，Simpson 指標的值將為 0；反之，若是元素平均分布在 N 個類別上，此時 Simpson 指標的值為（N-1）/N。愈是平均分布的情形，Simpson 指標的值愈大。

3.Rao-Stirling 指標 D 為 Simpson 指標結合類別差距的計算方式，如式（3）：

$$D \stackrel{\text{def}}{=} \sum_{i=1}^{N} \sum_{j=1\ (i\neq j)}^{N} p_i p_j d_{ij} \qquad (3)$$

2 Porter and Rafols (2009) 稱為 Herfindhal 指標。

此處的 d_{ij} 是類別 i 和 j 之間的差距，p_i 和 p_j 分別是所有元素在類別 i 和 j 上的數量比例。當元素集中在一個類別時，Rao-Stirling 指標如同 Simpson 指標的值一樣皆為 0，但 Rao-Stirling 指標的最大值較為複雜，需視類別差距的計算結果而定。概念上來說，當系統上的元素分布於較多的類別、在類別上的分布較平均、分布的類別差異較大等情況時，Rao-Stirling 指標將有較大的值。

4. Gini 係數 G 可以用來計算分佈上的不相等（inequality）或不均勻（unevenness）等情形（Leydesdorff & Rafols, 2011）。計算如式（4）

$$G \stackrel{\text{def}}{=} \frac{\sum_{i=1}^{N} (2i-N-1)\,\widehat{p_i}}{N-1} \qquad (4)$$

$\widehat{p_i}$ 是以類別上的元素數量比例由小到大排列後的第 i 個數量比例。若是元素在所有的類別上分布相等，Gini 係數的值為 0；當到達最大不相等的情況，也就是系統上所有的元素都集中在一個類別上時，Gini 係數的值將為 1。因此，與其它的指標不同，元素的數量分布愈均勻，Gini 係數愈小。

關於上述指標之間的關連，Porter and Rafols（2009）和 Rafols and Meyer（2010）的研究都發現 Rao-Stirling 指標與 Simpson 指標和 Shannon 熵值彼此間都有很高的相關性。Leydesdorff and Rafols（2011）對這些指標進行因素分析，其結果顯示 Shannon 熵值、Gini 係數和 Rao-Stirling 多樣性歸屬於同一個因素，但利用 Spearman 等級相關分析 Gini 係數和 Shannon 熵值的結果為負數，表示當論文的參考文獻在主題分類上的分布愈均勻時，所得到 Shannon 熵值愈大，但 Gini 係數愈小。

參、研究方法

一、應用主題模型方法的研究領域主題確認

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

題出現在每一筆文件上的可能性。這個模型假設每一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，並且每一個主題則都是由一組詞彙（vocabulary）上的詞語依據不同的機率組成（Blei, Ng, & Jordan, 2003）。較簡易的說法是主題模型假定每一筆文件及每一個主題都可以表示成一組特徵向量（feature vector）。文件特徵向量的每一個元素表示某一個主題出現在文件上的機率，在主題模型方法裡，通常以 θ_d 表示文件 d 的特徵向量，特徵向量 θ_d 上的第 i 個元素 θ_{di} 代表第 i 個主題在文件 d 上的機率值，元素的數值愈大，表示相對應的主題出現在文件上的可能性愈大。主題特徵向量的每一個元素則是表示當文件挑選到此一主題時每一個詞語的出現機率。如果以 ϕ_k 表示第 k 個主題的特徵向量，特徵向量 ϕ_k 上的第 j 個元素 ϕ_{kj} 便是對應詞彙中第 j 個詞語出現在第 k 個主題的機率。

主題模型方法便是根據文件內容的詞語出現情形，估計模型裡的每一組文件特徵向量 θ_d 與主題特徵向量 ϕ_k。在估計出主題的特徵向量後，針對各個主題，可以從機率值較大的詞語，判斷這個主題的內容；對於某一筆文件則可以根據較大出現機率的主題，了解這個文件包含的主題。目前已經有將主題模型應用於研究主題確認的研究發表（Griffiths & Steyvers, 2004; Zheng, McLean, & Lu, 2006）。林頌堅（2014）也從 WoS 引文資料庫檢索 2007 到 2013 年的 *Scientometrics* 和 *Journal of Informetrics*（*Informetrics*）期刊的 1755 筆論文資料。利用主題模型方法的 Gibbs 演算法（Griffiths & Steyvers, 2004），從論文的文字內容推導資訊計量學的重要研究主題，分析兩種期刊個別重視的主題，並且了解主題在分析時間範圍內的增減情形。

本研究以林頌堅（2014）為基礎，將研究範圍擴大為資訊科學的六種重要期刊在 2007 到 2013 年出版的論文。六種期刊包括：*IPM*、*J Doc*、*J Info Scie*、 *Informetrics*、*JASIST* 和 *Scientometrics*。選擇這六種期刊的原因是 *Informetrics* 雖然是相當新的期刊，但擁有相當高的期刊影響係數（journal impact factor），其餘五種期刊也都是 Janssens et al.（2006）、Bonnevie-Nebelong（2006）、Huang and Chang（2012） 和 Lu and Wolfram（2012）等先前研究曾經使用的資訊科學代表性期刊，並且有不低的期刊影響係數。以 2007 年至 2013 年為論文資料蒐集的時間範圍的目的，則是為了了解資訊科學以及各期刊近期的研究主題並且考慮充足的分析資料。從 WoS 資料庫上檢索這六種期刊出版期間為 2007 年至 2013 年並且文件類型為 Article 的論文資料，共 4210 筆，去除沒有 AB 欄位的 10 筆資料後，共 4200 筆。取出 TI 和 AB 欄位的

資料做為論文的文字內容。在統一大小寫、過濾數字與停用詞、刪除罕用詞與常用詞後，共得 4932 種詞語。將文件 - 詞語矩陣（document-term matrix）輸入 Gibbs 演算法推導文件特徵向量與主題特徵向量等模型參數。除了估計文件特徵向量與主題特徵向量以外，主題模型方法還需要決定主題的數目。在最佳主題數的選取時，本研究採用機器學習（machine learning）與資料探勘（data mining）常用的 k 次交叉驗證（k-fold cross validation）方法，並依循常例將 k 設為 10，使用 10 次交叉驗證估計混淆度（perplexity）的平均值。獲得混淆度平均值最低的主題數便是最佳的結果。

二、主題之間的差距與結構

主題模型方法的優點之一是利用機率分布值來表示產生的主題，能夠很容易應用於需要進行數值計算的處理（林頌堅，2014）。為了瞭解整個領域的主題結構，本研究利用能夠測量兩個分布之間差異性的對稱式 KL 差異（symmetric Kullback-Leibler divergence）（Rzeszutek, Androutsos, & Kyan, 2010）估計主題之間的差距。假定第 i 個和第 j 個主題的特徵向量分別為 ϕ_i 和 ϕ_j，它們之間的對稱式 KL 差異 $sKL(\phi_i,\phi_j)$ 如式（5）：

$$sKL(\phi_i,\phi_j)=\sum_{k=1}^{V}\left(\phi_{ik}\log\frac{\phi_{ik}}{\phi_{jk}}+\phi_{jk}\log\frac{\phi_{jk}}{\phi_{ik}}\right) \qquad (5)$$

ϕ_{ik} 和 ϕ_{jk} 是詞彙中的第 k 個詞語分別在第 i 個和第 j 個主題上的出現機率，V 是詞彙的大小。當兩個主題的機率分布完全相同時，也就對所有的詞語而言，$\phi_{ik}=\phi_{jk}$，此時對稱式 KL 差異 $sKL(\phi_i,\phi_j)=0$；當兩個主題的機率分布差異愈大，$sKL(\phi_i,\phi_j)$ 的值便愈大。

以式（5）計算主題之間的差距之後，本研究利用階層式叢集分析，將主題依據它們之間的差距分群，進一步了解整個領域上的主題結構。

三、期刊的主題出現機率

運用論文的主題出現機率，可以推測期刊的主題出現機率。假定每一筆論文在其發表的期刊上的重要性都是相同的，每一個主題在期刊上的出現機率便可以定義為各主題在期刊發表的論文上出現機率的平均值。例如式（6）表示第 i 個主題在期刊 J 上的出現機率 θ_{Ji}：

$$\theta_{Ji} \stackrel{\text{def}}{=} \sum_{d \in J} \theta_{di} \Big/ |J| \qquad (6)$$

d 代表期刊 J 上發表的某一筆論文，θ_{di} 為第 i 個主題在論文 d 上的出現機率。$|J|$ 是集合內的論文數量。依據式（6）的定義，如果某一個主題在期刊 J 上許多論文上都具有很高的出現機率，這個主題在期刊 J 上也將會有很高的出現機率。

四、期刊之間的差距

期刊之間的差距也可以視為是它們的主題機率分布的差異。在每一種期刊的主題出現機率估計完成後，便可以運用對稱式 KL 差異來測量期刊之間的差距。例如期刊 A 與期刊 B 之間的差距的計算方式為：

$$sKL(\theta_A, \theta_B) = \sum_{i=1}^{T} \left(\theta_{Ai} \log \frac{\theta_{Ai}}{\theta_{Bi}} + \theta_{Bi} \log \frac{\theta_{Bi}}{\theta_{Ai}} \right) \qquad (7)$$

θ_{Ai} 與 θ_{Bi} 分別為第 i 個主題在期刊 A 與期刊 B 上的出現機率，T 是主題數目。如同測量主題之間差距的情形，當兩種期刊的主題出現機率分布相等時，它們之間對稱式 KL 差異的值為 0；當兩種期刊的主題出現機率分布差異愈大，對稱式 KL 差異的值便愈大。

在測量期刊之間的差距後，本研究進一步將測量得到的資料輸入 MDS 演算法，將期刊之間的差距呈現於二維圖形上，了解以研究主題為基礎的資訊科學領域重要期刊分布情形。

五、期刊主題多樣性測量

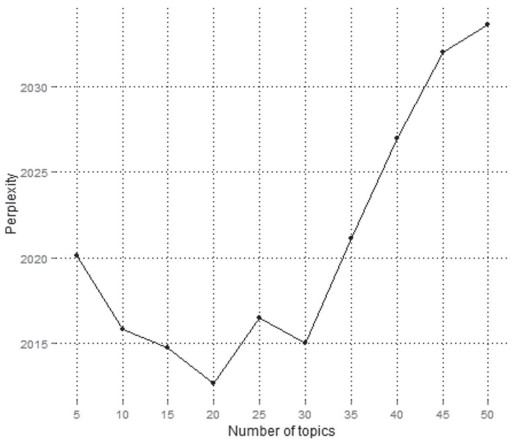
根據 Porter and Rafols（2009）、Rafols and Meyer（2010）和 Leydesdorff and Rafols（2011）的研究，Shannon 熵值、Simpson 指標、Rao-Stirling 指標和 Gini 係數等多樣性指標可以獲得一致的測量結果。計算各種多樣性指標時，可以將各個主題在期刊上的出現機率 θ_{Ji} 將分別代入式（1）到（4）的類別數量比例 p_i 進行計算。計算 Rao-Stirling 指標時，除了將主題的出現機率代入式（3）的類別數量比例以外，還需要計算類別差距 d_{ij}。本研究以對稱式 KL 差異測量主題之間的差距，也就是將式（5）的 $sKL(\phi_i, \phi_j)$ 代入式（3）的 d_{ij}。為了瞭解此一方法對 Rao-Stirling 指標造成的影響，本研究測量 Shannon 熵值、Simpson 指標和 Rao-Stirling 指標，並針對三種指標進行比較。

肆、研究結果

一、資訊科學領域的重要研究主題

將資訊科學的六種重要期刊在 2007 到 2013 年出版的 4200 筆論文內容資料輸入主題模型進行參數估計，經由 10 次交叉驗證的估計，以包含 20 個主題的主題模型獲得最佳的混淆度，因此本研究便將主題數設定為 20。從主題數為 5 到主題數為 50 的各種主題數所獲得的混淆度平均值如圖 1 所示。

圖 1　各種主題數下的混淆度平均值



資料來源：本研究繪製

表 1 列舉出 20 個主題上出現機率最大的前 10 個詞語。

表 1
20 個主題上出現機率最大的前 10 個詞語

主題編號	詞語				
1	systems	system	user	evaluation	users
	design	retrieval	search	support	models
2	review	peer	funding	abstracts	journal
	scientific	manuscripts	publication	process	published
3	classification	text	documents	features	proposed
	learning	algorithm	document	clustering	model

表 1
20 個主題上出現機率最大的前 10 個詞語（續）

主題編號	詞語				
4	h-index	index	citation	citations	impact
	distribution	rights	reserved	measures	-index
5	universities	performance	indicators	university	scientific
	quality	evaluation	researchers	ranking	academic
6	scientists	researchers	gender	scientific	authors
	productivity	academic	publication	differences	authorship
7	network	networks	structure	knowledge	scientific
	maps	social	centrality	field	topics
8	knowledge	social	factors	sharing	model
	online	management	communication	community	communities
9	library	digital	students	findings	libraries
	literacy	purpose	public	health	seeking
10	journals	citation	impact	journal	citations
	articles	science	factor	cited	published
11	language	terms	words	text	question
	retrieval	english	answer	chinese	arabic
12	image	semantic	metadata	tags	tagging
	ontology	images	indexing	social	digital
13	retrieval	model	documents	document	ranking
	proposed	performance	relevance	rights	reserved
14	countries	science	publications	scientific	publication
	articles	international	china	growth	output
15	patent	patents	technology	technological	innovation
	knowledge	development	industry	patenting	nanotechnology
16	science	citation	articles	journals	field
	author	social	sciences	scientific	fields
17	collaboration	international	scientific	countries	collaborative
	researchers	co-authorship	impact	knowledge	national
18	web	pages	sites	content	online
	wikipedia	links	users	page	site
19	search	users	web	queries	searching
	task	relevance	engines	user	tasks
20	knowledge	science	purpose	findings	theory
	work	concept	article	understanding	framework

資料來源：本研究整理

為了以下的分析方便，本研究根據表 1 在各主題出現機率值最高的前 10 個詞語，參酌 White and McCain（1998）、Moya Anegón, Herrero Solana and Jiménez-Contreras（2006）和 Janssens et al.（2006）等先前研究，將各主題命名，其結果呈現於表 2。大部分的主題都能夠明顯地從出現機率較高的詞語了解主題的意義，並進而命名。例如 Topic 1 上出現機率較高的詞語有“system(s)”、“user(s)”、“evaluation”、“design”、“relevance”等等，因此將 Topic 1 命名為「檢索系統設計與評估」。又如 Topic 2，根據出現核心詞語“review”、“peer”、“funding”、“abstracts”等等，可以了解這個主題和論文投稿的審查機制有關，因此命名為「論文審查與學術出版」。

表 2
以主題模型方法推導的 2007 到 2013 年資訊科學研究主題

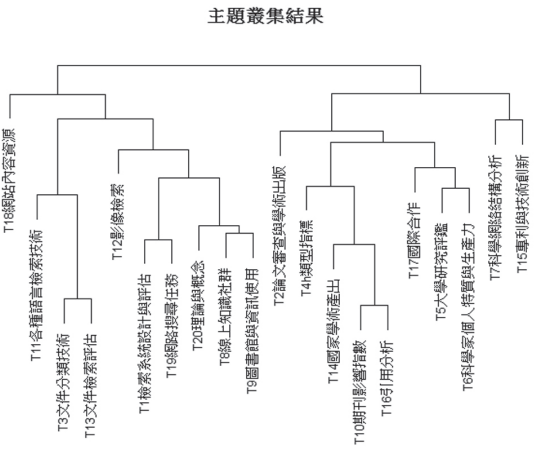
編號	主題	編號	主題
1	檢索系統設計與評估 (design and evaluation of retrieval systems)	11	各種語言檢索技術 (language-specific information retrieval technology)
2	論文審查與學術出版 (paper review and scholarly publishing)	12	影像檢索 (image retrieval technology)
3	文件分類技術 (text classification technology)	13	文件檢索評估 (evaluation of text retrieval)
4	h 類型指標 (h-type indices)	14	國家學術產出 (scientific output by countries)
5	大學研究評鑑 (research evaluation of universities)	15	專利與技術創新 (patents and technology innovation)
6	科學家個人特質與生產力 (demographic characteristics of scientists and their productivity)	16	引用分析 (citation analysis)
7	科學網絡結構分析 (structure analysis of scientific networks)	17	國際合作 (international collaboration)
8	線上知識社群 (online knowledge community)	18	網站內容資源 (web content resources)
9	圖書館與資訊使用 (library and information use)	19	網路搜尋任務 (web search tasks)
10	期刊影響指數 (journal impact factor)	20	理論與概念 (theory and concept)

資料來源：本研究整理

在少數的主題上，出現機率較高詞語的語意較廣泛，較不容易從這些詞語看出主題的意義，例如 Topic 16 與 Topic 20。Topic 16 的詞語有“science(s)”、“citation”、“articles”、“journals”等等，可能與各領域的論文引用行為有關，因此命名為引用分析。Topic 20 包含“purpose”、“findings”、“work”、“understanding”等經常出現於論文內容的詞語，根據“theory”和“concept”等詞語，將這個主題命名為「理論與概念」。此外，部分主題上的核心詞語彼此重疊，例如在 Topic 4、Topic 10 與 Topic 16 等主題中，“citation(s)”、“journal(s)”、“science(s)”、“impact”等詞語都有很高的出現機率。很明顯的，這些主題之間的關連性比其他不在這個圈子裡的主題來得高。

為了了解整個領域的主題結構，本研究利用對稱式 KL 差異測量主題之間的差距，進行階層式叢集分析。圖 2 呈現階層式叢集分析的結果。在圖 2 上，可以發現 20 個主題分為兩個群組，左邊群組的主題都和資訊搜尋與檢索研究有關，右邊的主題則都為資訊計量學相關的主題。資訊搜尋與檢索研究群組相關的主題中的 Topic 3、Topic 13 和 Topic 11 有較強的關連，這三個主題都與文件檢索有關，大致符合過去研究所提出的演算法型資訊檢索的範圍。Topic 1 和 Topic 19 等主題主要探討系統評估和搜尋任務，則是可以視為認知型資訊檢索。另外，資訊搜尋與檢索研究部分還包含 Topic 8、Topic 9 和 Topic 20 等與資訊使用相關的主題。右邊的資訊計量學群組相關主題中的 Topic 4、Topic 10、Topic 14 和 Topic 16 可歸類為以引用為基礎的評估指標和在國家研究成效的應用。Topic 5、Topic 6 和 Topic 17 等主題則為機構與個人的學術成效以及研究合作的相關研究，這三個主題很明顯地都屬於實務型的案例研究。另外，探討科學傳播的網絡結構並將學科領域的研究主題關係映射為圖形的 Topic 7 和運用專利分析探討技術創新的 Topic 15，雖然距離評估指標研究和資訊計量學的實務型案例等主題較遠，但也屬於資訊計量學研究的範疇。

圖 2　對主題進行階層式叢集的結果，各主題名稱前的數值為其編號

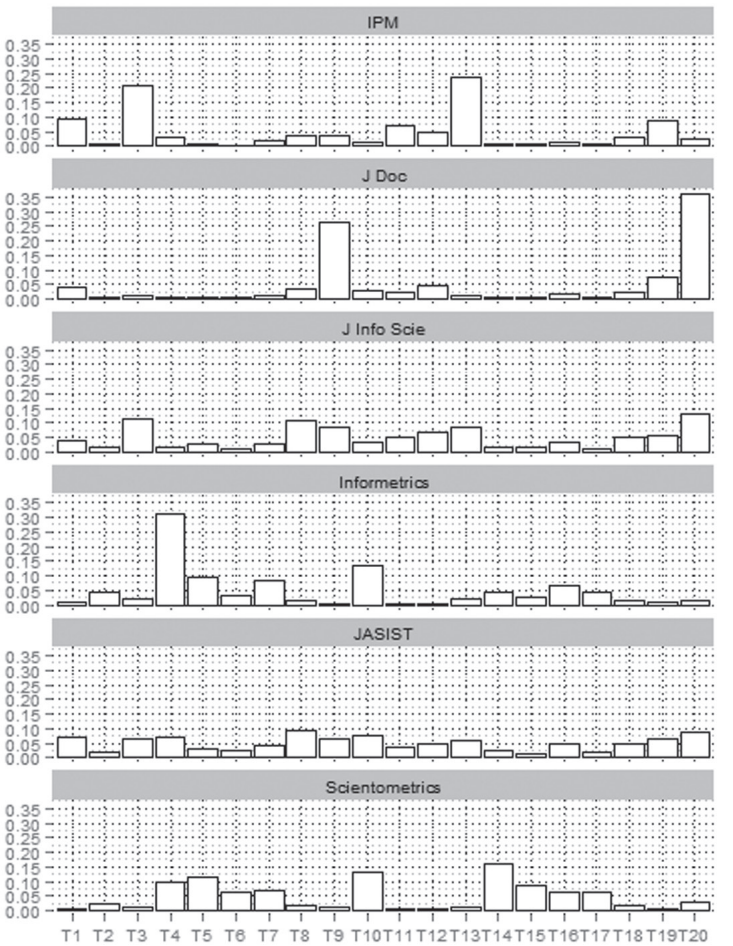


資料來源：本研究繪製

二、資訊科學重要期刊的主題分布情形

為了分析資訊科學各期刊上的主題分布情形，首先計算各期刊上每一個主題的出現機率。圖 3 上每一種期刊的長條圖，呈現出 20 個主題在該期刊上的分布機率高低。以圖 3 最上面的 *IPM* 為例，圖型上可觀察到分布機率最高的主題有 Topic 3（文件分類技術）和 Topic 13（文件檢索評估），兩個主題在 *IPM* 的出現機率都大於 0.2。另外出現機率較高的主題還有 Topic 1（檢索系統設計與評估）和 Topic 19（網路搜尋任務）。在前面的分析中，前兩者為演算法型資訊檢索，而後兩者為認知型資訊檢索。可見 *IPM* 相當關注在資訊搜尋與檢索的研究，並且更偏向於演算法型資訊檢索。

圖 3　資訊科學重要期刊的主題出現機率分布情形



資料來源：本研究繪製

圖 3 的 *J Doc* 部分，可以觀察到出現機率最高的主題分別是 Topic 9（圖書館與資訊使用）和 Topic 20（理論與概念）。另外 Topic 1（檢索系統設計與評估）、Topic 12（影像檢索技術）和 Topic 19（網路搜尋任務）等與資訊檢索相關的主題也有比其他主題稍微高一些的出現機率。

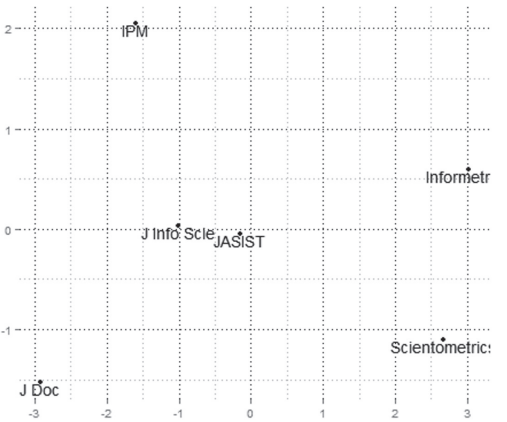
相較於其它期刊，*J Info Scie* 和 *JASIST* 的主題分布較為平均。*J Info Scie* 的重要主題有 Topic 3（文件分類技術）、Topic 8（線上

知識社群）、Topic 9 （圖書館與資訊使用）、Topic 13 （文件檢索評估）和 Topic 20（理論與概念）等多種資訊搜尋與檢索研究相關主題。*JASIST* 則是幾乎平均地分布於資訊搜尋與檢索和資訊計量學的各個主題上。

Informetrics 最重要的主題為 Topic 4（h 類型指標）和 Topic 10（期刊影響指數），其次還有 Topic 5 （大學研究評鑑）、Topic 7（科學網絡結構分析）和 Topic 16（引用分析）等。*Scientometrics* 的重要主題為 Topic 4（h 類型指標）、Topic 5（大學研究評鑑）、Topic 10（期刊影響指數）、Topic 14（國家學術產出）和 Topic 15 （專利與技術創新）等，另外 Topic 6（科學家個人特質與生產力）、Topic 7（科學網絡結構分析）、Topic 16（引用分析）和 Topic 17（國際合作）也都是 *Scientometrics* 重視的主題。*Informetrics* 和 *Scientometrics* 的重要主題都包括 Topic 4 與 Topic 10，很明顯的以引用為基礎的評估指標研究是這兩種期刊共同的重要研究方向。

從上述的分析，可以獲得一些值得進一步探討的地方：1）*JASIST* 與 *J Info Scie* 具有較廣泛的主題，與其它期刊的主題都有相關，並且 *J Info Scie* 重視的主題與資訊搜尋與檢索研究較相關。2）*IPM* 和 *J Doc* 的主題都著重在資訊搜尋與檢索，但是 *IPM* 較重視資訊檢索相關的主題，而 *J Doc* 則以資訊使用的研究為主。3）*Informetrics* 和 *Scientometrics* 以資訊計量學研究為主，兩者在 h 類型指標和期刊影響指數等主題上都有很高的出現機率，但兩種期刊在其他主題上有稍微不同的偏重傾向，*Scientometrics* 在專利與技術創新以及國家學術產出與國際合作等主題上有較高的出現機率，而 *Informetrics* 則是著重在科學網絡結構分析。因此下面將分別就各期刊的主題差異進行分析：本研究首先以對稱式 KL 差異測量期刊在主題分布上的差異，然後將測量得到的資料輸入 MDS 演算法，將期刊之間的關係呈現於圖 4。

圖 4　依據主題分布差異產生的資訊科學期刊映射圖



資料來源：本研究繪製

在圖 4 上可以看到，*JASIST* 映射於圖形的中央，左方為偏重於資訊搜尋與檢索主題的期刊，右方則為資訊計量學研究相關的期刊，*JASIST* 的位置反映出該期刊具有多樣化的主題，且在每個主題上分布平均。*J Info Scie* 的映射位置也相當接近圖形中央，但是更偏向於資訊搜尋與檢索方面一些。*IPM* 和 *J Doc* 為資訊搜尋與檢索相關的期刊，而它們被映射在圖形上代表資訊搜尋與檢索的左半邊，這兩種期刊之間的距離凸顯出它們在資訊搜尋與檢索的研究上偏重很不同的主題。*Informetrics* 和 *Scientometrics* 的映射位置在代表資訊計量學研究的右方，它們之間較小的距離表示這兩種期刊分布的主題大致相同，但仍然有些不同的主題。

三、資訊科學重要期刊的主題多樣性情形

表 3 是利用各種多樣性指標對資訊科學重要期刊進行測量的結果。各種指標有相當一致的結果：*JASIST* 的主題多樣性最高，主題多樣性次高的期刊為 *J Info Scie*。這兩種期刊的主題分布相當平均，特別是 *JASIST*。對照先前在期刊主題分布情形的分析，*JASIST* 對於資訊科學的各種主題都有關注，*J Info Scie* 重視的主題則是在資訊搜尋與檢索的各種主題上。*Scientometrics* 在各種指標的測量上都顯示比 *Informetrics* 更具有多樣性。因此，雖然兩種期刊都著重在資訊計量學相關的主題上，但 *Scientometrics* 所關注的主題範圍比 *Informetrics* 較廣一些。如

期刊的題名本身所揭示的，*IPM* 的主題集中在各種類型的資訊檢索技術上，因此主題多樣性較低，*J Doc* 則是主題最集中的期刊。

表 3
資訊科學領域與重要期刊的各種多樣性指標測量結果

	Entropy	Simpson	Rao-Stirling
IPM	3.47	0.87	7.58
J Doc	2.98	0.79	6.78
J Info Scie	3.97	0.92	8.56
Informetrics	3.44	0.86	7.70
JASIST	4.16	0.94	8.92
Scientometrics	3.71	0.91	8.19

資料來源：本研究整理

伍、結論

本研究利用資訊科學六種重要期刊於 2007 到 2013 年發表的論文內容，運用主題模型方法，確認領域的主題。本研究產生 20 個主題，這些主題分為兩個群組：一個群組是包括演算法型資訊檢索、認知型資訊檢索、資訊使用以及網路資源等主題的資訊搜尋與檢索研究，另一個群組則是包括以引用為基礎的評估指標研究、實務型案例研究、科學傳播網絡以及專利與技術創新等主題的資訊計量學。雖然本研究的分析範圍和研究方法都與先前的研究不同，但產生的主題與先前研究有一致的結果，顯示雖然容易受到資訊科技的影響（Åström, 2007），資訊科學的研究主題大致上相當穩定。

以下根據各期刊網站上的主旨說明與收錄範圍，討論本研究的結果：

JASIST 上發表的研究主要以資訊的生產、發現、記錄、儲存、再現、檢索、表現、操作、傳播、使用與評估以及上述過程相關的工具與技術，探討的方向也十分廣泛，包含實徵、實驗、人類文化、概念、歷史、社會技術（socio-technical）、政策分析（policy-analytic）與批判理論（critical-theoretical）等[3]。本研究的結果發現 *JASIST* 在資訊科學的

3 http://onlinelibrary.wiley.com/store/10.1002/(ISSN)2330-1643/asset/homepages/JASIST_Infogram.pdf?v=1&s=4c92c825c99f75288f8b9aa8f7234d18a1d9e35a&isAguDoi=false

每一個主題上分布平均，並且是研究範圍內主題多樣性最高的期刊。

J Info Scie 的主題包括資訊科學與知識管理[4]。在本研究的主題確認結果中，*J Info Scie* 與 *JASIST* 同樣具有較廣泛的研究主題，但 *J Info Scie* 較重視與資訊搜尋與檢索研究相關的主題。因此，其主題多樣性的測量結果比 *JASIST* 稍微低一些。

IPM 的研究範圍包含三個主軸：1）資訊科學、電腦科學、認知科學等基礎與應用研究；2）資訊檢索、數位圖書館、知識組織與分配（knowledge organization and distribution）、數位內容以及人機介面等實驗性與先進處理；3）資訊資源、服務、系統與網路以及數位圖書館的管理[5]。在本研究的結果則顯示 IPM 的研究主題確實如期刊本身的題名以及在網站上揭示的主旨，分布在資訊搜尋與檢索群組的各個研究主題上，並且稍為著重於演算法型資訊檢索。

J Doc 的主旨與收錄範圍特別著重在理論、概念、模型、架構和哲學[6]。雖然 *J Doc* 在圖書館與資訊使用以及理論與概念等主題有最高的出現機率，符合該期刊的主旨；然而本研究的結果也發現該期刊的多樣性測量結果最低、主題最集中，並且研究主題偏向於認知型資訊檢索。

Scientometrics 主要考慮科學及科學研究的量化特徵與特性，特別著重在數理統計方法應用於科學的發展與機制上[7]。本研究的研究結果發現，除了重視應用文獻引用或專利資訊進行科學與技術的研究績效評估以外，*Scientometrics* 還特別著重在對於科學家個人與團隊、國家以及國際合作等學術生產力的案例研究。

Informetrics 的論文收錄以應用於書目計量、科學計量與網路計量以及各種計量問題，特別的是該期刊指出案例研究並不適合發表在該期刊上，除非是內容包含創新的研究方法的論文[8]。本研究發現 *Informetrics* 與 *Scientometrics* 同樣重視 h 類型指標和期刊影響指數以及科學網絡結構分析等研究方法的創新與改良。並且相較於 *Scientometrics*，*Informetrics* 較少有論文的主題分布在科學家個人生產

4 http://jis.sagepub.com/

5 http://www.journals.elsevier.com/information-processing-and-management/

6 http://www.emeraldgrouppublishing.com/products/journals/journals.htm?id=JD

7 http://link.springer.com/journal/11192

8 http://www.journals.elsevier.com/journal-of-informetrics/

力、國家學術產出以及國際合作等案例分析相關的主題上。這個原因使得 *Informetrics* 在研究主題多樣性的測量結果較 *Scientometrics* 低。

此外，在利用 Shannon 熵值、Simpson 指標和 Rao-Stirling 指標測量各種重要期刊的研究主題多樣性時，可以發現各種多樣性指標的測量結果相當一致。雖然 Rao-Stirling 指標加入了主題差距的權重，但從圖 2 的階層式分析所展現的主題結構，也發現 Shannon 熵值、Simpson 指標測量值較低的期刊具有的主題都屬於同一群組，彼此間的差距不大，使得加入主題差距的 Rao-Stirling 指標測量結果與其他兩種多樣性十分相近。

從本研究的結果可以觀察到關注多個主題的期刊，具有較高的主題多樣性，當以 MDS 技術呈現期刊在主題分布的差異時，這些期刊被映射圖形的中央；著重於資訊搜尋與檢索研究及資訊計量學的期刊，主題較集中，多樣性較低，則分別被映射在圖形的左邊與右邊。似乎期刊的主題多樣性與它們科學映射圖形上的位置存在一定程度的關係？由於本研究僅對資訊科學的六種重要期刊進行分析，目前的資料有限並不足以進行嚴謹的統計推論。未來將可以擴大分析範圍，蒐集更多的期刊，並且應用到更多的研究領域，藉以了解期刊的主題多樣性與科學映射結果之間的關係。

參考文獻

林頌堅（2014）。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

Åström, F. (2007). Changes in the LIS research front: Time-sliced cocitation analyses of LIS journal articles, 1990–2004. *Journal of the American Society for Information Science and Technology, 58*(7), 947-957. doi: 10.1002/asi.20567

Bache, K., Newman, D., & Smyth, P. (2013). Text-based measures of document diversity. In R. Grossman & R. Uthurusamy (Eds), *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 23-31). New York, NY: ACM. doi: 10.1145/2487575.2487672

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

Bonnevie-Nebelong, E. (2006). Methods for journal evaluation: journal citation identity, journal citation image and internationalization. *Scientometrics, 66*(2), 411-424. doi: 10.1007/s11192-006-0029-9

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

Huang, M. H., & Chang, Y. W. (2012). A comparative study of interdisciplinary changes between information science and library science. *Scientometrics, 91*(3), 789-803. doi: 10.1007/s11192-012-0619-7

Janssens, F., Leta, J., Glänzel, W., & De Moor, B. (2006). Towards mapping library and information science. *Information Processing & Management, 42*(6), 1614-1642. doi: 10.1016/j.ipm.2006.03.025

Leydesdorff, L., & Rafols, I. (2011). Indicators of the interdisciplinarity of journals: Diversity, centrality, and citations. *Journal of Informetrics, 5*(1), 87-100. doi: 10.1016/j.joi.2010.09.002

Liu, Y., Rafols, I., & Rousseau, R. (2012). A framework for knowledge integration and diffusion. *Journal of Documentation, 68*(1), 31-44. doi: 10.1108/00220411211200310

Lu, K., & Wolfram, D. (2012). Measuring author research relatedness: A comparison of word-based, topic-based and author co-citation approaches. *Journal of the American Society for Information Science and Technology, 63*(10), 1973-1986. doi: 10.1002/asi.22628

Moya Anegón, F., Herrero Solana, V., & Jiménez-Contreras, E. (2006). A connectionist and multivariate approach to science maps: the SOM, clustering and MDS applied to library and information science research. *Journal of Information Science, 32*(1), 63-77. doi: 10.1177/0165551506059226

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

Rafols, I., & Meyer, M. (2010). Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience. *Scientometrics, 82* (2), 263-287. doi: 10.1007/s11192-009-0041-y

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

Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society Interface, 4*(15), 707–719. doi: 10.1098/rsif.2007.0213

Waltman, L., van Eck, N. J., & Noyons, E. (2010). A unified approach to mapping and clustering of bibliometric networks. *Journal of Informetrics, 4*(4), 629-635. doi: 10.1016/j.joi.2010.07.002

White, H. D., & McCain, K. W. (1998). Visualizing a discipline: An author co-citation analysis of Information Science, 1972–1995. *Journal of the American Society for Information Science, 49*(4), 327-355. doi: 10.1002/(SICI)1097-4571(19980401)49:4<327::AID-ASI4>3.0.CO;2-4

Zheng, B., McLean, D. C., & Lu, X. (2006). Identifying biological concepts from a protein-related corpus with a probabilistic topic model. *Bmc Bioinformatics, 7*(58). doi:10.1186/1471-2105-7-58

A Bibliometric Study of Topic Distribution and Diversity of Journals in Information Science

Sung-Chien Lin

Assistant Professor
Department of Information and Communications
Shih-Hsin University

Introduction

Many bibliometric studies have identified important research topics in Information Science, for example Janssens, Leta, Glänzel, and De Moor (2006), Lu and Wolfram (2012), Moya Anegón, Herrero Solana, and Jiménez-Contreras (2006), Waltman, van Eck and Noyons, (2010), and White and McCain, (1998). Bonnevie-Nebelong (2006) and Huang and Chang (2012) used references of Information Science papers to discuss the diversity of knowledge sources of related journals or subfields. However, only Janssens et al. (2006) tried to find the major topics for each related journal and no study to date measures diversity based on the topics distributed in journals. Thus, the goal of this study includes:

1. to identify important topics of Information Science from text data of papers using the method of topic modeling, and furthermore, to reveal topical structure of the field;
2. to learn distributions of the topics in the related journals and compare differences among the topic distributions of the journals;
3. to measure topical diversity of the journals and to compare the results from different diversity indicators.

Data and Methods in this Study

The analytical data used in this study were papers published in the six important journals of Information Science, including *Information Processing and Management (IPM), Journal of Documentation (J Doc), Journal of Information Science (J Info Scie), Journal of Informetrics (Informetrics), Journal of the Association for Information Science and Technology (JASIST)*, and *Scientometrics*. Bibliographic data of papers in these journals were retrieved from the database of the Web of Science with the criteria that the papers were published between 2007 and 2013 and their type was

as “Article”. Texts in the field TI and AB of the bibliographic data were extracted and combined as documents. After converting all words into their lower case, removing digits and stop words, and filtering out rare words and common words, computational linguistics concordance software counted the frequency of each word occurring in each documents.

We used the frequency information of word occurrences to estimate parameters of the topic model which corresponds to the text generation process of the documents (Blei, Ng, & Jordan, 2003; Griffiths & Steyvers, 2004). One set of parameters are the probabilities that represents the frequencies of topics occurring in the documents. For each document, we can use the probabilities of topics to find the most possible topics appearing in the document, and to estimate the probability of every topic occurring in a certain journal by averaging the probabilities of the topic in all the documents of papers published in the journal. Another set of parameters are also probabilities and each of them describes how frequently a particular word occurred when documents picked up a certain topic. In this study, the words in the rank of the top-10 highest probabilities of a topic are called core words of the topic. We selected core words of each topic to determine the meaning of the topic.

We then estimate dissimilarities between topics and disparities between journals. The estimation method used was the symmetric Kullback-Leibler divergence which can measure the difference between two probabilistic distributions (Rzeszutek, Androutsos, & Kyan, 2010). The dissimilarities between topics were used as information input to the method of agglomerative hierarchical clustering to group all topics in order to demonstrate the topical structure of Information Science. The disparities between journals were input to the algorithm of multidimensional scaling (MDS) to generate a graph for visualizing the position of each journal in the field.

Finally, we measure the diversity of topics in each journal. Diversity is characteristic for describing how a set of elements are distributed among categories in a system, which is composited of the variety of categories, the balance of distributions of elements over categories, and the disparity between categories (Rafols & Meyer, 2010; Stirling, 2007). In this study, we used three different indicators to understand how the topics are distributed in the journals, including Shannon’s entropy, Simpson index, and Rao-Stirling index.

Results

We set the topic number for the resulting model to 20 due to the minimal average perplexity computed by the method of 10-fold cross-validation. The title of topics were named by not only considering their core words but also referring to topic titles used in previous studies. These topics and their titles are shown in Table 1.

Table 1
Research Topics of Information Science Generated by the Method of Topic Modeling

Topic	Title	Topic	Title
1	design and evaluation of retrieval systems	11	design and evaluation of retrieval systems
2	paper review and scholarly publishing	12	paper review and scholarly publishing
3	text classification technology	13	text classification technology
4	h-type indices	14	h-type indices
5	research evaluation of universities	15	research evaluation of universities
6	demographic characteristics of scientists and their productivity	16	demographic characteristics of scientists and their productivity
7	structure analysis of scientific networks	17	structure analysis of scientific networks
8	online knowledge community	18	online knowledge community
9	library and information use	19	library and information use
10	journal impact factor	20	journal impact factor

In most topics, the titles were easily determined by examining the corresponding core words. However, for a few topics their meaning was difficult to ascertain from their core words, for example Topic 16 and 20, because the concepts of those words are very wide. In addition, there were high degree of overlays between the core words in some sets of topics. The relations between the topics within the same sets are obviously stronger than those between the topics which are not in the same sets.

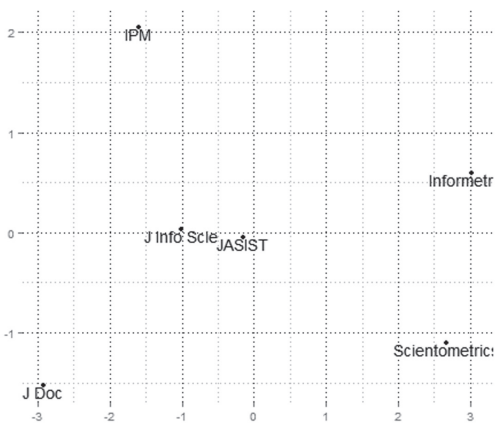
Exploring the topical structure of Information Science, we found that the 20 topics can be divided into two major classes. The first class of topics are related to “Information Search and Retrieval (ISR)”, which includes topics about algorithmic information retrieval, cognitive information

retrieval, and information use. The other class consists mainly of topics related to "Informetrics", such as bibliometric indices for evaluating universities and researchers, empirical case studies, scientific mapping, and studies of technology innovation based on patent analysis.

The major topics of each journal were described as follows: *JASIST* owned the widest scope of topics than others did and all topics of Information Science were distributed evenly in the journal. *J Info Scie* also gained most of the topics in the field, but it paid more attention on the topics related to ISR. The topics of *IPM* and *J Doc* also focused on ISR, but the former focused on information retrieval and the latter's emphasis was on information use, respectively. Just as their titles suggest, *Informetrics* and *Scientometrics* both paid more attention to topics related to Informetrics, and h-type indices and journal impact factor frequently occurred in them. Furthermore, *Scientometrics* was more interested in empirical case studies, while *Informetrics* had a slight emphasis on structure analyses of scientific networks.

The result of MDS is shown in Figure 1. In the graph, the journal with the most broad topics, *i.e. JASIST*, was mapped on the central of the map. The journals on the left side of the map are those with topics highly related to ISR, which are *J Info Scie, IPM, and J Doc*. The journals with the topics related to Informetrics, i.e. *Scientometrics* and *Informetrics*, are mapped on the right side.

Figure 1 the MDS map of the Information Science journals



The measure results of topical diversity for the six journals produced by the three different indicators were consistent to one another. In all the indicators, *JASIST* obtained the highest scores among all the journals, followed by *J Info Scie*. This signifies that these two journals have the most balanced distributions on topics of Information Science. Although *Scientometrics* and *Informetrics* both focused on topics related to Informetrics, the former paid attention to more topics than the latter did, such that the scores for *Scientometrics* are higher than those for *Informetrics*. Finally, *J Doc* is the journal in which the topics were the most concentrated and therefore it obtained the lowest scores in all diversity indicators.

Conclusions

Comparing the results of this study with those of previous studies, we observed that the topics of Information Science were rather stable in recent years, even though Åström (2007) claimed that the field was easily affected by information technology at that time.

We found that the topics that frequently occurred in the journals with low scores of diversity usually belonged to the same class in the topical structure of the field. Disparities between the topics were very little, and thus, the measure results obtained by the Rao-Stirling index were similar to the other two indicators, although the former is weighted with the disparities between topics.

We also found that the positions of journals mapped on the MDS map might associate with their diversity scores. However, evidence for the claim is deficient to make a rigid induction in this study because we only analyzed six journals. In the future, we will extend the scope of analysis by collecting more journals and applying the method to more fields to understand the association between the topical diversity of journals and their positions in the examined fields.

References

Åström, F. (2007). Changes in the LIS research front: Time-sliced cocitation analyses of LIS journal articles, 1990–2004. *Journal of the American Society for Information Science and Technology, 58*(7), 947-957. doi:

10.1002/asi.20567

Bache, K., Newman, D., & Smyth, P. (2013). Text-based measures of document diversity. In R. Grossman & R. Uthurusamy (Eds), *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 23-31). New York, NY: ACM. doi: 10.1145/2487575.2487672

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

Bonnevie-Nebelong, E. (2006). Methods for journal evaluation: journal citation identity, journal citation image and internationalization. *Scientometrics, 66*(2), 411-424. doi: 10.1007/s11192-006-0029-9

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

Huang, M. H., & Chang, Y. W. (2012). A comparative study of interdisciplinary changes between information science and library science. *Scientometrics, 91*(3), 789-803. doi: 10.1007/s11192-012-0619-7

Janssens, F., Leta, J., Glänzel, W., & De Moor, B. (2006). Towards mapping library and information science. *Information Processing & Management, 42*(6), 1614-1642. doi: 10.1016/j.ipm.2006.03.025

Leydesdorff, L., & Rafols, I. (2011). Indicators of the interdisciplinarity of journals: Diversity, centrality, and citations. *Journal of Informetrics, 5*(1), 87-100. doi: 10.1016/j.joi.2010.09.002

Lin, S. C. (2014). 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: doi:10.6120/JoEMLS.2014.514/0633.RS.AM [Text in Chinese].

Liu, Y., Rafols, I., & Rousseau, R. (2012). A framework for knowledge integration and diffusion. *Journal of Documentation, 68*(1), 31-44. doi: 10.1108/00220411211200310

Lu, K., & Wolfram, D. (2012). Measuring author research relatedness: A comparison of word-based, topic-based and author co-citation approaches. *Journal of the American Society for Information Science and Technology, 63*(10), 1973-1986. doi: 10.1002/asi.22628

Moya Anegón, F., Herrero Solana, V., & Jiménez-Contreras, E. (2006). A connectionist and multivariate approach to science maps: the SOM, clustering and MDS applied to library and information science research. *Journal of Information Science, 32*(1), 63-77. doi: 10.1177/0165551506059226

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

Rafols, I., & Meyer, M. (2010). Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience. *Scientometrics, 82* (2), 263-287. doi: 10.1007/s11192-009-0041-y

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

Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society Interface, 4*(15), 707–719. doi: 10.1098/rsif.2007.0213

Waltman, L., van Eck, N. J., & Noyons, E. (2010). A unified approach to mapping and clustering of bibliometric networks. *Journal of Informetrics, 4*(4), 629-635. doi: 10.1016/j.joi.2010.07.002

White, H. D., & McCain, K. W. (1998). Visualizing a discipline: An author co-citation analysis of Information Science, 1972–1995. *Journal of the American Society for Information Science, 49*(4), 327-355. doi: 10.1002/(SICI)1097-4571(19980401)49:4<327::AID-ASI4>3.0.CO;2-4

Zheng, B., McLean, D. C., & Lu, X. (2006). Identifying biological concepts from a protein-related corpus with a probabilistic topic model. *Bmc Bioinformatics, 7*(58). doi:10.1186/1471-2105-7-58