教育資料與圖書館學

Journal of Educational Media & Library Sciences

http://joemls.tku.edu.tw

Vol. 51 , no. 4 (Summer 2014) : 499-523

以主題模型方法爲基礎的資訊計量學領域研究主題分析

Analyses of Research Topics in the Field of Informetrics Based on the Method of Topic Modeling

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　51 : 4 (Summer 2014) : 499-523
DOI:10.6120/JoEMLS.2014.514/0633.RS.AM

以主題模型方法為基礎的資訊計量學領域研究主題分析

林頌堅

摘要

本研究利用主題模型方法發現資訊計量學研究主題的結構，探討主題的逐年與發展以及比較相關期刊。本研究利用檢索自 Web of Science 資料庫的 *Journal of Informetrics* 和 *Scientometrics* 期刊 2007 到 2013 年間的論文資料，做為主題模型方法的輸入，推算主題結構的研究資料。研究結果發現主題數量設定為 10 的主題模型具有最小的混淆度。雖資料範圍與分析方法不同，本研究產生的主題仍可和由專家分析產生的結果相容。實務的案例研究與書目計量指標的測量方式在各年度都較受重視，且整個領域日趨穩定。兩種核心期刊都廣泛地涉及資訊計量學的所有主題，但 *Journal of Informetrics* 特別著重於書目計量指標的建構與應用，而 *Scientometrics* 則關注於國家、機構、領域及期刊上的學術生產力評鑑方式與影響因素的探討。

關鍵詞：研究主題分析，資訊計量學，主題模型

緒　論

在圖書資訊學領域的研究範疇裡，應用統計或其他數值方法分析與呈現書目資料、科學記錄、網路記錄或其他資訊活動紀錄等資料，探討人類生產、傳播與使用資訊等活動的特性與規律，向來都是非常重要的一個次領域。由於分析的目的不同，致使研究人員使用不同的研究資料與研究方法，且在論文的撰寫與發表上，採用不同的名稱來稱呼，常見者有書目計量學（bibliometrics）、科學計量學（scientometrics）、資訊計量學（informetrics）（Hood & Wilson, 2001）。隨著人類在網際網路的涉入愈廣與愈深，過去也出現了專門描述與分析在網際網路上資訊行為的網路計量學（webometrics, cybermetrics）（Thelwall, 2009）。雖有各種不同的名稱，但基本上這個領域的研究者大多彼此認為具有相同的專業

世新大學資訊傳播學系助理教授
通訊作者：scl@cc.shu.edu.tw

2014/06/06投稿；2014/08/05修訂；2014/08/05接受

(specialties),進行相同主題的研究,且有專屬的研討會以及期刊,作為學術傳播的管道,形成學術社群。以下為方便行文起見,除了提及各研究時,以引用論文時使用原先的名稱稱呼之外,本研究以涵義較廣的資訊計量學統稱這個次領域。為更加了解這個研究領域近年發展的內涵,本研究嘗試以核心期刊的文字內容,分析這個研究領域內可能的主題以及其發展情形。

Noyons、Moed與Van Raan(1999)和Cobo、López-Herrera、Herrera-Viedma與Herrera(2011)都指出書目計量學的兩個主要研究方向,一為成效分析(performance analysis),也就是根據書目資料評估國家、大學、研究者等群體的表現與他們的影響力,另一為科學映射(science mapping),是指利用科學對映圖呈現科學研究領域的結構與動向,如Börner、Chen與Boyack(2003)上評論的研究。由於資訊計量學本身便是利用分析各學術領域的學術生產與引用活動,探討學術領域的認知與社會特性,因此過去有許多研究便是以資訊計量學領域本身做為分析的對象。

Schoepflin與Glänzel(2001)在探討這個領域是否已成為「硬科學」(hard science)時,認為先前Schubert與Maczelka(1993)和Wouters與Leydesdorff(1994)關於科學計量學是否具有立即效應(immediacy effect)以及經由立即效應形成的研究前沿(research front)等學術傳播現象的研究,都是將這個領域當作一個整體,利用核心期刊*Scientometrics*在一段時間範圍內所有的發表論文資料,根據論文的引用文獻,計算Price指標與平均的引用文獻相對年齡等參考文獻指標進行分析。然而Schoepflin與Glänzel(2001)發現每一年度論文的Price指標和連續出版品比率的中位數皆明顯大於期刊整體的數據,表示許多論文比整體期刊呈現的情形來得「硬」。從這種現象,他們推論在科學計量學領域裡來自不同學科背景的研究者有各自不同的傳播、引用和發表方式,使得這個科際整合的領域趨向多元發展,而顯得充滿異質(heterogeneous)。因此,研究科學計量學的發展與其在學術傳播上的特性時,有必要先確認研究領域的研究主題,再針對各種主題進行分析。

Schoepflin與Glänzel(2001)進一步將研究的論文資料分類為六個主題,統計各個主題論文數量的占有率變化以及各項參考文獻指標。六個主題分別是:1.書目計量學理論、數學模型與書目計量學定律的公式化(bibliometric theory, mathematical models and formalisation of bibliometric laws);2.案例研究與實務論文(case studies and empirical papers);3.方法學論文包含應用(methodological papers including applications);4.指標工程與資料呈現(indicators engineering and data presentation);5.書目計量學中的社會學取向,科學社會學(sociological approach to bibliometrics, sociology of science);6.科學政策、科學管理與廣泛或技術討論(science policy, science management and general or technical discussions)。

這個研究的結果發現：在研究的時間範圍內，案例與方法學有明顯與穩定的成長，但社會學取向和科學政策的論文數量減少，且各種主題類別的論文在Price指標、參考文獻為連續出版品的比率、參考文獻的平均年齡和平均引用的參考文獻比率等各項指標值上差異都很大，特別是科學政策。而各年度計算得到的參考文獻指標值，受到當時主題類別主要占有率的影響，例如科學政策在1980年是占有率最大的主題類別，當這個主題的論文數量在1989和1997年大幅減少後，所得到的期刊整體指標值便和1980年有很大不同。

Glenisson、Glänzel、Janssens與De Moor（2005）對於*Scientometrics*期刊2003年論文進行的研究，援引了Schoepflin與Glänzel（2001）的六個類別，加上近年興起的網路計量學後，重新歸類為：科學計量學先進議題（advances in scientometrics）、實務論文與個案研究（empirical papers/case studies）、數學模型（mathematical models）、政策議題（political issues）、社會學方法（sociological approaches）以及資訊計量學與網路計量學（informetrics/webometrics）等六類。用上述的分類為基礎，同時也利用詞語共現分析（co-word analysis）及書目計量指標了解各個類別的特性。科學計量學的先進議題相關論文，除了少數例外，大部分論文的平均參考文獻年齡在5到15年間，參考文獻為連續出版品的比例約佔50%到90%間；且這類論文經過詞語共現分析產生的叢集結果則分散在多個叢集，主要的兩個叢集分別傾向於書目計量學指標和資訊計量學法則的方法學研究。實務性論文可依據它們的參考文獻的連續出版品所占部分分為兩群，較低一群與政策相關研究具有類似的特徵；實務性論文的詞語共現分析所呈現的結果，除了與社會學、政策到科技創新等許多相關主題形成的一個巨大叢集，另一個叢集上的論文大多與國家和機構方面或科學領域的個案研究有關，還有一個叢集的論文數較少，則是有關於共被引分析（co-citation analysis）以及其他引用統計的分析。網路計量學和其他網路相關議題的論文集中於一個叢集，其參考文獻具有低相對年齡以及大多為連續出版品等特徵。從上述的分析，Glenisson等人（2005）認為科學計量學目前主要的兩個面向是基於科學計量學標準技術的方法學研究和擴展傳統書目計量學範圍的實務研究。

Dutt、Garg與Bali（2003）則將*Scientometrics*期刊上1978到2001年發表論文資料的主題，分為科學計量評估（scientometric assessment）、引用與叢集分析（citation and cluster analysis）、科學計量分布（scientometric distribution）、科學史（history of science）、科學合作（scientific collaboration）、科學計量學理論研究（theoretical studies on scientometrics），不在上述主題的論文則歸類為其他。然後分析這期間論文的主題分布情形。結果發現：論文數最多的主題是科學計量評估，這個現象反映出科學政策的制定逐漸運用科學計量工具的事實，其次是理論研究。在前期（1978至1986年），科學史方面的論文較多，其次是引用與

叢集分析；科學計量分布在前期與中期（1987至1994年）都相當重要，但後期（1994至2001）逐漸減少；後期具有最重要地位的主題則是科學合作。

除了利用*Scientometrics*期刊為核心期刊分析資訊計量學領域以外，Egghe（2012）則分析*Journal of Informetrics*的前五卷239篇文章。將239篇文章加以分類，其中引用分析（citation analysis）和h-類型指標（h-type indices）是論文最多的兩個主題，這兩個主題便佔有一半以上的論文。此外，Egghe（2012）也將*Journal of Informetrics*最常引用的30種期刊以及最常被引用的30種期刊彼此間的引用關係繪製成網路圖。利用圖形呈現相關期刊之間的關連，使得彼此有引用關係的期刊在圖形上形成叢集。從引用關係形成的網路圖可觀察到*Journal of Informetrics*最相近的期刊是*Scientometrics*，兩者都有相當多關於引用分析的論文。其不同在於*Journal of Informetrics*有較多關於模型—理論（model-theoretic）以及網路議題（networking issue）方面的論文；*Scientometrics*則包含許多案例研究（case studies）的論文。

從上述分析可觀察到目前以論文分類進行有關資訊計量學領域的研究，大多是由專家以其對領域的知識來進行分類，然而Chen、McCain、White與Lin（2002）的主題確認結果則是完全以論文的參考資料產生，且呈現為可視的科學映射圖（scientific map）。這項研究以1981到2001年間的*Scientometrics*期刊論文為研究資料，選擇被引用次數達五次以上的參考文獻，共計403筆文獻，根據這些文獻的共被引次數計算Pearson 相關係數（Pearson's correlation coefficients）產生共被引矩陣（co-citation matrix），然後利用主成分分析（principal component analysis）對共被引矩陣進行因素分析（factor analysis），以分析出的因素代表領域的專業。因素分析的結果共計找出25個因素，較大的三個因素分別命名為科學研究中的引用（citations in science studies）、全球與國家的科學表現（world and national science performance）、研究產出的評估（evaluation research outputs）。這種根據期刊論文的資料分析領域主題的方法能夠提供較完整的觀點，以一致的統計演算法進行分析所得到的結果也較客觀。

Chen等人（2002）以論文共同被引用的情形做為關連性估測的方法，語詞的共現分析則是利用相類似研究主題的論文具有相近的語詞分布情形，測量論文之間的關連性，做為確認學術領域研究主題的依據，例如Bhattacharya 與Basu（1998）在凝態物理（Condensed Matter Physics）領域，Ding、Chowdhury與Foo（2001）在資訊檢索領域以及van den Besselaar與Heimeriks（2006）在資訊科學領域等研究。雖然Leydesdorff（1997）認為詞語的意義隨它們與其他詞語關係的頻率及其出現位置，會有所改變，也就是同一個詞語在不同的主題下具有不同的意義。另外，在某一個特定的主題下，具有相同語意的一群詞語可互相替換而不會影響內容的意義，也就是許多論文的研究主題相同但使用不同的

詞語。另外，論文的內容描述研究過程中所有相關的問題、理論、方法、技術與結果，其中自然包含多種研究主題，因此必須能夠確認分析資料中的多種研究主題。直接利用詞語在內容上的共現次數進行分析在處理上述的一詞多義（polysemy）、同義詞（synonymy）以及文件的多重主題等問題上有很大的困難。然而相較於利用參考文獻資料的共被引分析，詞語共現分析方法能應用在沒有被引用關係的論文資料。且共被引分析會因在領域的變動與趨勢以及引用者的行為而變得複雜（Noyons & van Raan, 1998），如同Lu與Wolfram（2012）所指出的，研究人員需要具備相關的領域知識才能判斷結果品質並提出合理解釋。以文字內容為研究資料的分析方法則能夠以意義較明顯的詞語特徵審視產生的結果，在解釋結果方面較具有優勢。

主題模型（topic modeling）方法能夠從文件的文字內容資料發現文件集合內隱藏的主題結構，也就是揭露文件集合中可能具有的各種主題以及每一個文件上的主題比例。主題模型使用LDA（latent Dirichlet allocation）模型描述文件產生的過程（Blei, Ng, & Jordan, 2003），這個模型假設每一筆文件都包含多種主題，可利用主題的機率混合（probabilistic mixture）來表示文件的組成，且每一個主題都是由一組詞彙（vocabulary）上的詞語依據不同機率混合組成。文件上每一個詞語的產生是先從該文件的主題機率混合上挑選一個主題，然後再以主題相對應的詞語機率混合挑選出一個詞語做為產生的詞語。由於LDA主題模型將文件視為各種主題的混合，而每一個主題則視為是詞彙上各個詞語的比例分布，主題結構只和詞語在文件資料上的分布有關，因此資訊檢索研究應用主題模型來解決一詞多義、同義詞與文件的多重主題等問題（Mimno & McCallum, 2007），也開始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）。

總結上述分析，主題模型方法在確認學術領域的研究主題上具有下列特性與優點：

1. 主題模型方法產生的結果完全由輸入的論文資料決定。

2. 主題模型方法將每一篇論文視為是由各個主題依不同的比例組成。這樣的結果不會發生每一筆具有多重主題的論文資料硬性地被歸類於某一個主題的情形，避免分類上的困難與錯誤。

3. 針對各個主題，選取主題的機率混合上機率值較大的詞語，配合在這個主題上有較大分布機率的論文，根據意義較明顯的詞語特徵以及論文資料的文字內容來解釋與分析產生的結果。

4. 主題模型方法利用機率分布值來表示產生的主題與論文，很容易應用於進一步需要進行數值計算的處理，例如主題或論文的相關性評估等。

5. 主題模型方法產生的主題呈現為詞語的分布機率，可以比較不同主題在

詞語上的分布異同，也可以透過一組相關詞語在各個主題上的分布了解主題之間的關連，解決複雜的一詞多義問題。

因此應用主題模型確認學術領域的研究主題是相當具有發展潛力與極高應用性的研究方向，然而目前的相關研究大多為方法可行性的探討，實際案例與應用的質量都有待提升，因此本研究將嘗試探討以下的問題：

1.以資訊計量學領域主要期刊的論文內容為研究資料，利用主題模型方法發現期刊論文集合的主題結構，也就是確認資訊計量學領域的主題，從主題包含的詞語了解各主題的涵義，並且根據文件對應的主題機率混合，找出各主題的典型論文。

2. Schoepflin與Glänzel（2001）、Glenisson等人（2005）和Dutt等人（2003）等先前的研究都指出，運用科學計量工具的實務案例分析與對於科學計量學標準技術的方法學研究等主題，在研究時間範圍內的論文數量有明顯而穩定的成長，反之社會學取向、科學史和科學政策等主題的論文數量減少。這些研究都在2005年之前完成，因此多只採用*Scientometrics*期刊的論文資料，本研究嘗試了解2007年*Journal of Informetrics*創刊後到2013年間的資訊計量學領域主題分布與發展情形。

3. Egghe（2012）認為「*Scientometrics*較偏向實務應用而*Journal of Informetrics*較偏向理論」，本研究嘗試根據主題模型方法確認得到的主題，比較兩種期刊主題的整體分布與各年間的發展情形。

二、研究方法

在說明本研究的進行步驟與研究方法前，首先介紹主題模型與文件內容之間的關係：主題模型假定文件集合內共有T個主題，每一個主題由不同的詞語機率組合ϕ_k來表示，ϕ_k是由一個參數為β的Dirichlet分布所產生，且這個機率組合上面的每一個數值表示從這個主題上挑選出某一個特定詞語的機率值。另一方面，如果要產生的文件為d，文件d相對應的主題機率混合稱為θ_d，θ_d上的機率值為文件d上所有主題的分布機率，由另一個參數為α的Dirichlet分布產生。w為文件d裡的某一個詞語，根據從主題混合θ_d中取樣得到的主題z以及其相對應的詞語混合ϕ_z所產生。主題模型方法的詳細定義與說明可參見Blei等人（2003）和Griffiths與Steyvers（2004）。

主題模型方法便是希望根據文件的內容推算文件的產生模型，計算出LDA模型裡的各個未知參數ϕ_k和θ_d，發現文件集合內隱藏的主題，且進一步獲得各文件上每個詞語的出現機率。直接推算方式是使用E-M演算法（Expectation-Maximization algorithm; Hofmann, 1999）找到一組ϕ和θ，使得機率值$P(w|\phi,\theta)$最大，w是文件資料集合內所有詞語。但這個方法可能會產生局部最

大值（local maxima）。為解決這個問題，許多研究提出各種參數搜尋演算法，其中Gibbs取樣（Gibbs sampling; Griffiths & Steyvers, 2004）具有相對迅速與容易實作等優點，因此本研究便採用Gibbs取樣演算法推算LDA模型裡的參數ϕ_k和θ_d。

㈠期刊論文資料蒐集與處理

本研究首先從WoS資料庫選取*Scientometrics*以及*Journal of Informetrics*兩種期刊在分析時間範圍（2007至2013年間）的所有論文，也就是只選擇文件類型為Article的論文資料。對於取回的論文資料，先檢視TI和AB兩個欄位的資料，去除少數沒有AB欄位資料的論文，然後將兩個欄位的資料合併取出做為論文的文字內容。[1]

每一筆論文資料的文字內容都經過改換為小寫（lower case）、去除標點符號與數值，去除停用詞（stop words）等處理，取出出現的詞語以及其出現頻次，其中停用詞除了利用Smart系統建議的571個停用詞[2]，另外也加入在資訊計量學期刊論文非常高頻率出現的詞語，例如“information”、“research”、“analysis”等。為保留資訊計量學中相當重要的詞語，如“co-citation”、“h-index”等，本研究並沒有對文字內容進行詞幹化（word stemming）。此外，經過對每一篇論文內容進行統計後，常見（大於論文資料總筆數的5%）或罕見（小於論文資料總筆數的1%）的詞語均被刪除，以節省計算資源，加速運算的過程，並使得推算出的主題模型容易獲得解釋。本研究採用R統計分析軟體進行資料處理，利用tm套件（Meyer, Hornik, & Feinerer, 2008）處理文字內容。

㈡決定主題數量

主題模型的參數推算與詞語預測程式同樣以R軟體的topicmodels套件（Grün & Hornik, 2011）撰寫而成，本研究將主題模型中產生詞語機率混合與主題機率混合的參數β與α分別預設為0.01及1.0。[3]

LDA模型還需要考慮主題的個數T，主題個數的選擇會影響結果的解釋。Griffiths與Steyvers（2004）建議採用後驗機率（posterior probability）最高的主題個數。許多研究則利用混淆度（perplexity）預測模型的成效，並根據模型的成

[1] 雖目前兩種期刊都能取得電子全文，且全文包含的主題資訊更為豐富而確實。但全文的資料量相當龐大，且在進行主題模型的推算前，需先行針對不同期刊撰寫論文資料蒐集程式，以及解析檔案的PDF格式，過濾本文之外的額外訊息，鑒於本研究的目的主要在於探討主題模型方法的應用，因此僅選擇簡單扼要的題名與摘要做為文字內容，並從WoS資料庫檢索所需資料，進行適用性探討。

[2] Smart系統所列的571個停用詞，可參考http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-smart-stop-list/english.stop。

[3] 考慮單一研究領域不會有太多研究主題，將產生主題機率混合的參數α的值固定為1.0，而不採用topicmodels套件預設的50/T。

效決定個數（Blei et al., 2003），愈小的混淆度表示模型的成效愈好。本研究利用混淆度決定最佳的主題數量過程如下：

1.本研究將蒐集的所有論文隨機分為十份。

2.進行不同主題數量的混淆度分析：

2.1 從十份論文資料中，每次輪流保留一份做為測試語料，其餘九份訓練語料用來推算主題模型的參數。

2.2 應用訓練語料推算出的主題模型，針對保留的測試論文進行詞語預測，計算文件內容的混淆度。

2.3 計算十個混淆度的平均值。

3.在計算出各種不同主題數量的平均混淆度後，選取平均混淆度最小時的主題數量做為最佳的主題數量。

4.以最佳的主題數量為參數，推算最後使用的主題模型。

最後，本研究將應用上述方法所得到的主題模型解決前一節所列出的三個研究問題。

㈢確認資訊計量學主題結構

根據主題模型的理論，論文集合內隱含的每一個主題都可表示成一個詞語的機率混合，由在機率混合上的機率值代表每一個詞語出現在相對應主題上的可能性。因此，在完成主題模型推算後，本研究針對每一個主題的詞語機率混合進行排序，選取機率值最大的前十個詞語做為主題的核心詞語，藉以了解每一個主題的涵義。

除了核心詞語以外，為進一步了解主題模型方法產生的主題結構，本研究也定義了各個主題的典型論文，也就是主題有相當高機率會出現的論文，來了解論文上的主題比例。針對每一個主題，本研究以主題出現在每一筆論文資料的可能性進行排序，挑選出機率值最大的前十筆論文資料做為這個主題的典型論文。這些典型論文的題名同樣能夠看出主題的涵義。

特別要加以說明的是Schoepflin與Glänzel（2001）、Glenisson等人（2005）、Dutt等人（2003）和Egghe（2012）等研究先以專家制定主題，然後依據主題對論文進行分類，通常一筆論文資料僅指定一個主題；然而本研究所使用的主題模型方法並不限制論文僅能擁有一個主題，但每一個主題的重要性由它出現在論文的分布機率值大小來決定。相較於Schoepflin與Glänzel（2001）、Glenisson等人（2005）、Dutt等人（2003）和Egghe（2012）等先前研究都以主題包含的論文數量比較主題大小，且討論主題在不同論文集合間的數量發展情形，本研究則透過主題出現在論文集合內的分布機率值比較主題的大小，不同集合之間的比較則以機率值的變化情形進行討論。

㈣ 研究主題的分布與逐年發展情形

研究領域的主題機率混合可定義為所有論文的主題機率混合的平均，愈大的機率值表示這個領域在相對應的主題上有較多的關注與產出。

同樣的情形，每一年論文的主題機率混合可定義為該年度發表論文的主題機率混合的平均。測量不同年度之間主題分布的差異時，可利用對稱式KL差異（symmetric Kullback-Leibler divergence）（Rzeszutek, Androutsos, & Kyan, 2010）。當兩個年度的主題分布相等時，它們之間的對稱式KL差異為0；反之，兩個分布差異愈大，它們之間的對稱式KL差異所獲得的值愈大。本研究將測量不同年度之間主題分布的對稱式KL差異，找出產生變化的年度，並詳細說明主題的發展情形。

㈤ 比較期刊主題特色

最後，計算*Scientometrics*和*Journal of Informetrics*個別在每一年的主題機率混合。本研究將期刊的主題機率混合定義為所屬論文的主題機率混合的平均。比較各個主題在期刊上的出現機率，較大的機率值表示該期刊較為關注相對應的主題，有較多的產出。針對每種期刊在2007到2013年間，選擇出現機率較大的主題進行逐年分析。

三、研究結果

㈠ 論文資料概述

本研究共計選取1,755筆論文資料，*Scientometrics*和*Journal of Informetrics*各為1,374筆和381筆。各出版年度的整體與各期刊的論文數量如圖1所示。2013年*Scientometrics*和*Journal of Informetrics*兩種期刊總計出版的論文數目已超過2007年的兩倍，特別是*Journal of Informetrics*在2007年只出版31篇論文，2013年出版93篇論文，已為當年的3倍，由此可見，資訊計量學近年相當受到研究者的矚目。

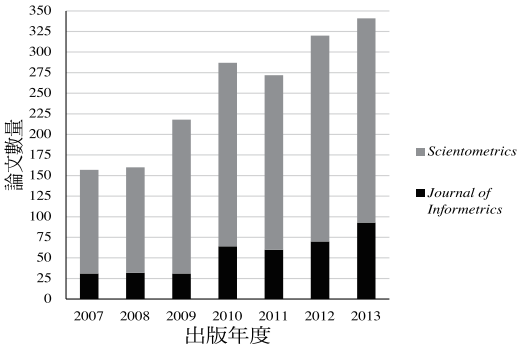

圖1 *Scientometrics*和*Journal of Informetrics*各年度論文數量

經過文字內容處理後，本研究從1,755筆論文文字資料抽取出1,147種詞語，並統計每種詞語出現在每筆論文文字資料的頻次（frequency count）。將論文集合隨機分為十份計算混淆度，根據混淆度的平均值搜尋最佳的主題數量。本研究預設的主題數量為5到100之間，共計20種數量，所得到的混淆度平均值以主題數量設定為10的683.46為最小，各種主題數量的混淆度平均值請參見圖2。因此，本研究推算資訊計量學的主題結構時，便將主題數量設定為10。

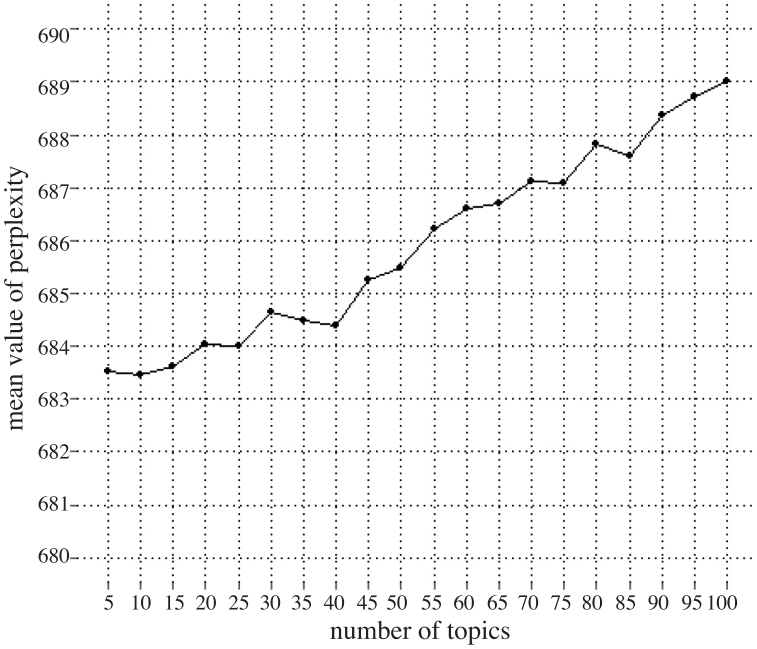

圖2 各種主題數量混淆度平均值

㈡資訊計量學研究主題

首先呈現與說明本研究獲得的主題結構。在本研究中，將每一個主題的核心詞語和典型論文的數量分別設定為10。由於論文篇幅的限制，各個主題的典型論文只列出該主題出現機率最高的兩篇論文，各個主題的核心詞語和典型論文舉例分別如表1和表2。

從表1及表2的核心詞語與典型論文的題名，大抵都能得知各主題的意義。較易明顯看出內涵的主題有：Topic 1與科技的創新發展及知識傳播網絡有關，且這類研究經常使用專利（patents）做為研究資源；Topic 2則是目前相當重要的h類型指標的理論研究與延伸應用；Topic 3為學術生產在時間上的增長與趨勢；Topic 4是有關期刊影響係數（journal impact factor）的實務研究；Topic 5有關於大學等研究機構在生產力與品質等研究成效的衡鑑與排名；Topic 6涉及有關論文審查（review）等學術出版等方面的議題；Topic 7包括影響係數（impact

表1　本研究產生的資訊計量學主題

主題	核心詞語				
Topic 1	network	knowledge	technology	patents	networks
	technological	innovation	development	model	emerging
Topic 2	index	h-index	reserved	distribution	rights
	citations	model	time	indices	citation
Topic 3	articles	publication	published	journals	bibliometric
	science	database	years	authors	period
Topic 4	citation	impact	journals	journal	citations
	cited	factor	factors	references	highly
Topic 5	universities	university	researchers	performance	academic
	productivity	ranking	quality	rankings	institutions
Topic 6	authors	author	significant	review	relationship
	differences	show	correlation	effect	gender
Topic 7	indicators	bibliometric	indicator	performance	field
	level	measures	impact	fields	evaluation
Topic 8	countries	collaboration	international	scientific	china
	country	national	chinese	output	production
Topic 9	scientific	scientists	science	social	sciences
	fields	disciplines	work	community	researchers
Topic 10	science	web	core	mapping	documents
	clustering	field	co-citation	bibliographic	topics

表2　本研究產生各主題的典型論文舉例

主題	典型論文的題名
Topic 1	**D**ivergence and convergence: Technology-relatedness evolution in solar energy industry. **I**dentification of technological knowledge intermediaries.
Topic 2	**C**onjugate partitions in informetrics: Lorenz curves, h-type indices, Ferrers graphs and Durfee squares in a discrete and continuous setting. **P**robing the h-core: An investigation of the tail-core ratio for rank distributions.
Topic 3	**T**rends of DDT research during the period of 1991 to 2005. **I**ncreasing dominance of English in publications archived by PubMed.
Topic 4	**E**mpirical study of journal impact factors obtained using the classical two-year citation window versus a five-year citation. **J**ournal weighted impact factor: A proposal.
Topic 5	**R**esearch performance and bureaucracy within public research labs. **N**ational research assessment exercises: The effects of changing the rules of the game during the game.
Topic 6	**T**he validity of staff editors' initial evaluations of manuscripts: A case study of Angewandte Chemie International Edition. **R**ejection rates for multiple-part manuscripts.
Topic 7	**A** systematic empirical comparison of different approaches for normalizing citation impact indicators. Source normalized indicators of citation impact: An overview of different approaches and an empirical comparison.
Topic 8	**T**echnological collaboration patterns in solar cell industry based on patent inventors and assignees analysis. **S**cience and scientific collaboration in South Africa: Apartheid and after.
Topic 9	**C**haracterizing a scientific elite: The social characteristics of the most highly cited scientists in environmental science and ecology. **L**ocating active actors in the scientific collaboration communities based on interaction topology analyses.
Topic 10	**U**sing "core documents" for the representation of clusters and topics. **D**o second-order similarities provide added-value in a hybrid approach?

factor）等書目計量指標（bibliometric indicator）的測量與在實務上的應用；Topic 8為有關科學研究合作的議題，特別是近年在資訊計量學領域的研究頗多在探討國際間的合作與產出之間的關連；Topic 9探討學術領域裡的科學家特性以及他們之間的社會關係；Topic 10為學術領域分析，包括運用叢集方法進行研究主題確認以及將研究主題與其間關連程度視覺化的科學映射圖等，其中共被引分析和書目耦合等共現現象是經常用來估計資料項目間關連程度的資訊。

㈢資訊計量學研究主題分布與發展情形

根據所有論的主題機率混合之平均計算資訊計量學在2007到2013年間所有研究主題的分布機率值，其小數點下兩位的概數如表3所示。從表3可看出資訊計量學在每一個主題的分布大致均勻，較大的主題是Topic 1（應用專利分析進行科技創新與知識傳播網絡的相關研究）和Topic 4（期刊影響係數的相關研究），都是屬於實務型的研究；較小的兩個主題Topic 6與Topic 9，分別是學術出版審查與科學家特性和社會關係等社會學相關的議題。從主題模型方法獲得的結果可以認為實務案例分析的相關研究較社會學取向的研究受到資訊計量學研究人員的重視，但差異並不大。

表3　2007–2013年資訊計量學研究主題分布情形（至小數點後兩位）

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
0.11	0.10	0.10	0.11	0.10	0.09	0.10	0.10	0.09	0.10

由每一年度發表論的主題機率之平均計算資訊計量學2007到2013年的逐年研究主題分布機率值，也可發現每一年的分布也相當均勻。為進一步了解資訊計量學的發展情形，利用的對稱式KL差異計算2007到2013年不同年間的研究主題分布差異。表4是計算的結果，由於產生的結果沿對角線對稱，所以只列出上半部。

表4　資訊計量學2007至2013年不同年度間研究主題分布差異（$*10^{-2}$）

	2007	2008	2009	2010	2011	2012	2013
2007		2.63	1.81	0.91	2.11	2.26	1.54
2008			1.85	2.57	3.09	4.94	3.00
2009				1.90	1.92	1.76	1.51
2010					2.10	1.68	1.42
2011						0.96	0.56
2012							0.90
2013							

仔細觀察表4不同年度間的研究主題分布差異，可發現兩個現象：首先，2008年的主題分布與其他各年度間有較大差異，除了與2009年的對稱式KL差異值為$1.85*10^{-2}$以外，其餘的對稱式KL差異值都超過$2.5*10^{-2}$，與2012年的差異值$4.94*10^{-2}$更是所有不同年度間的最大值。從對稱式KL差異分析2008年的

主題分布與其他各年度間有較大差異的原因，發現主要差異來自於Topic 1（應用專利分析的科技創新與知識傳播網路研究）與Topic 2（h類型指標理論與應用研究）在2008年的分布機率值與其他年度比較不同。圖3顯示Topic 1和Topic 2在2007到2013年間主題分布機率值的變化情形。Topic 1在多數的年度都獲得所有主題中最高機率值，2008年是各年中最低的機率值，同時也是所有主題中獲得最低機率值的主題。Topic 2在2008年則是所有主題中獲得最高機率值的主題。與上述相反的情形，發生在2012年：Topic 1和Topic 2分別是所有主題中獲得最高與最低機率值的主題。

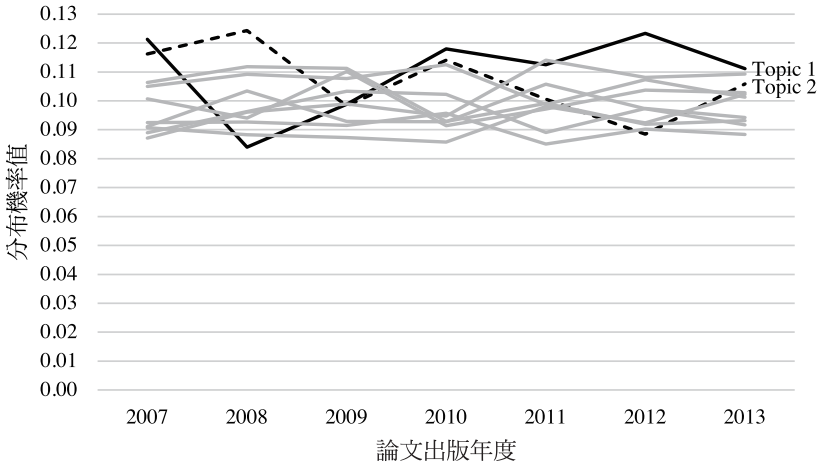

圖3　2007–2013年Topic 1和Topic 2主題分布機率值變化

其次，相對於先前各年度，資訊計量學在最近三個年度（2011至2013年）之間主題分布的差異較小，與先前年度有較大差異。除前述的Topic 1和Topic 2，圖4顯示另外三個在前期（2007至2009年）與後期（2011至2013年）較有改變的

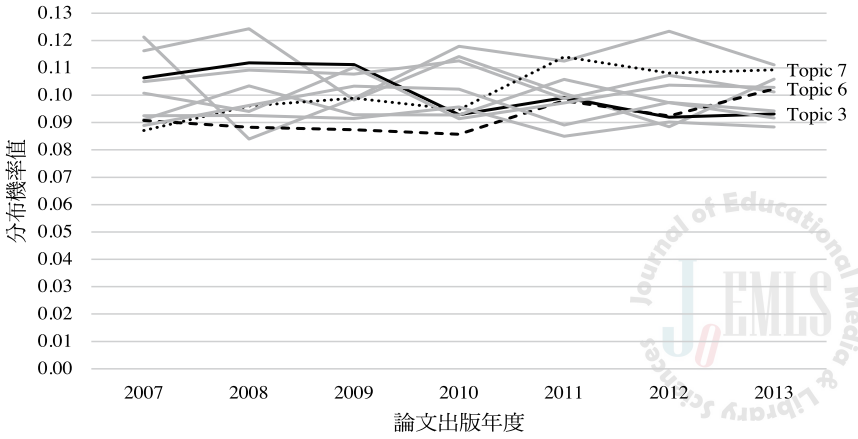

圖4　2007–2013年Topic 3、Topic 6和Topic 7主題分布機率值變化

主題Topic 3（學術生產的趨勢分析）、Topic 6（學術出版審查）和Topic 7（書目計量指標）的主題分布機率值變化情形。Topic 3在前期的分布機率值都在0.11左右，但後面四年則僅有0.10以下，可看出有明顯下降趨勢。相反的，Topic 6和Topic 7在前四年分別在0.09與0.10以下，後期則提升到0.10與0.11。

㈣資訊計量學期刊主題分布與比較

以下首先探討本研究由主題模型方法得到的*Journal of Informetrics*與*Scientometrics*期刊主題分布結果，然後討論兩種期刊在各主題具有的典型論文數量。表5為本研究產生的主題在*Journal of Informetrics*與*Scientometrics*論文上分布機率的概數。在表5可觀察到兩種期刊在所有主題上都有不低的機率，表示*Journal of Informetrics*與*Scientometrics*期刊上的論文在資訊計量學的各種主題上均有涉及；*Journal of Informetrics*在Topic 2、Topic 7和Topic4三個主題上有很明顯地較高機率，其次是Topic 10；*Scientometrics*在每個主題上的分布較平衡，稍高的機率落於Topic 1和Topic 8上，同時也略為傾向Topic 3、Topic 4和Topic 5等主題。根據前面的分析，Topic 2、Topic 4和Topic 7分別是有關於h類型指標、期刊影響係數和其他書目計量指標的測量方式與在實務上的應用，且許多指標的計算方式都與引用現象有關，Topic 10為有關學術領域分析的研究主題。*Scientometrics*較高的主題：Topic 1為應用專利分析進行科技創新發展與知識傳播的相關研究，Topic 8的主題為科學合作研究，Topic 3為學術生產的增長與趨勢分析，Top 5是對於研究機構生產力與品質的評鑑，上述的主題大多在於考量機構、期刊、國家或領域等學術研究與科技發展成效的分析與應用。特別值得一提的是，此外，Topic 4的期刊影響係數是兩種期刊都關心的主題。

表5 本研究主題於*Journal of Informetrics*與*Scientometrics*論文的分布

主題	*Journal of Informetrics*	*Scientometrics*
Topic 1	0.09	0.12
Topic 2	0.19	0.08
Topic 3	0.06	0.11
Topic 4	0.12	0.10
Topic 5	0.07	0.10
Topic 6	0.09	0.09
Topic 7	0.13	0.10
Topic 8	0.06	0.11
Topic 9	0.08	0.09
Topic 10	0.10	0.09

註：至小數點後兩位

圖5呈現*Scientometrics*期刊較重要的主題Topic 1、Topic 3和Topic 8的分布機率值變化情形。很明顯的可發現在大多數年度上，這三個主題的分布機率都比其他主題高出一些。Topic 1的分布機率經常是各主題中最高的，顯而易見

*Scientometrics*期刊一直都很重視利用專利分析等科學計量工具發現科技創新與學術界和產業之間的關連與傳播，然而這個主題在2008至2009年卻有相當大的下降，值得進一步探究與了解。表示學術生產趨勢分析主題的Topic 3在分析時間範圍的前期（2007至2009年）相當受到重視，2011年後的分布機率則有稍微下降的情形，但仍比大多數主題來得高。Topic 8的分布機率則是在後期（2010至2013年）逐漸提昇，說明了科學合作是*Scientometrics*期刊論文愈來愈重視的研究主題。

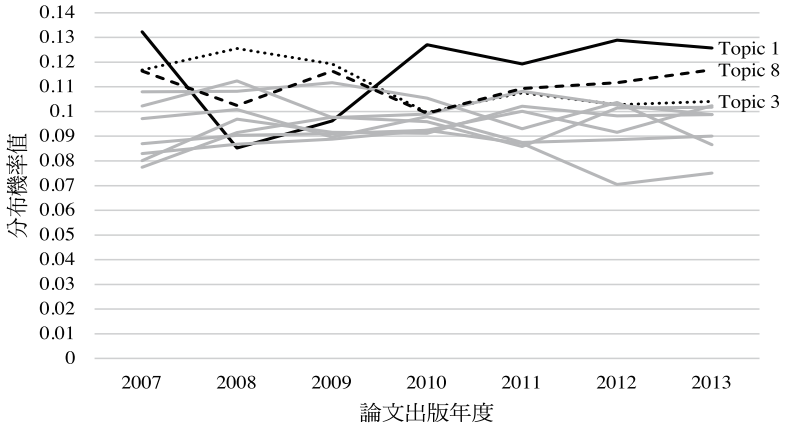

圖5　Topic 1、Topic 3和Topic 8在*Scientometrics*期刊分布機率值變化

書目計量指標的測量方式與應用有關的主題Topic 2、Topic 4和Topic 7是*Journal of Informetrics*期刊最關注的主題，這三個主題在*Journal of Informetrics*的分布機率值變化呈現於圖6。從圖6可看出在分析的時間範圍，這三個主題的分布機率比大多數的主題高，且Topic 2與Topic 4分布機率在各年度上的起伏相當類似，但Topic 7似乎與它們相反。造成這個現象的原因需要進一步分析

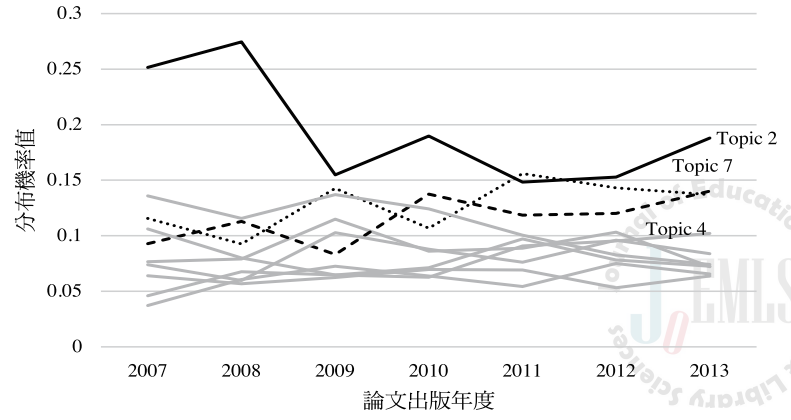

圖6　Topic 2、Topic 4和Topic 7在*Journal of Informetrics*期刊分布機率值變化

研究。另外，圖7呈現*Journal of Informetrics*期刊的另一個重要主題Topic 10在2007至2013年的分布機率值變化情形。從圖7，可觀察到這個主題在2011年前都受到相當重視，近兩年則有減少趨勢。

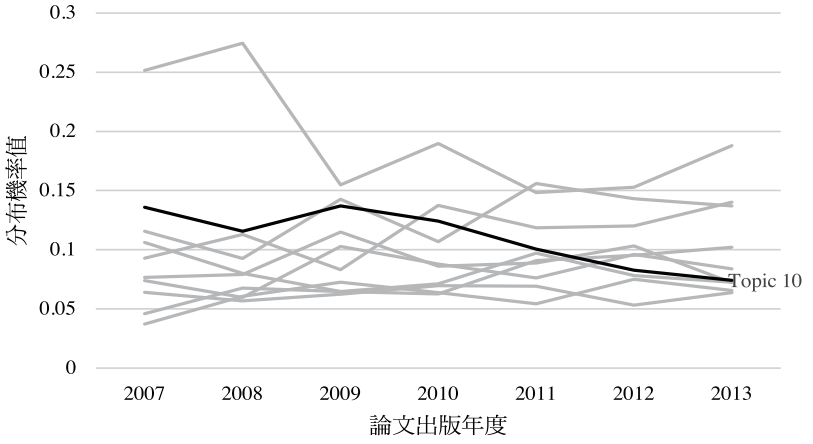

圖7 Topic 10在*Journal of Informetrics*期刊分布機率值變化情形

四、結 論

本研究利用期刊論文的題名與摘要等文字內容為研究資料，應用主題模型方法來探討資訊計量學的重要研究主題。Cobo等人（2011）除了指出成效分析與科學映射為書目計量學的兩個主要研究方向以外，他們也認為過去對於成效分析的研究，其方法著重於對群體的表現，較少以認知方式對研究領域內的特定主題以及主題在時間上的演變進行生產力與影響力的評估。換言之，便是較少整合上述兩個研究方向。因此本研究的目的除了發現學術領域上的重要研究主題外，另外便是應用主題模型方法探討研究主題的分布，在時間上的發展以及期刊的主題傾向。相較於過去的研究大多利用論文上的引用現象做為分析的資訊，本研究從論文的內容資料上獲取統計訊息，可較不受出版與引用過程的時間限制，同時產生的結果具有很豐富的語意訊息，也容易加以解釋。此外，由於主題模型方法產生的主題結構，都以機率分布的形式呈現，未來可利用統計學方法深化研究的成果並整合資訊視覺化的應用。

本研究目前獲得以下結論：

㈠雖然本研究採用的資料範圍與分析方法與先前的研究不同，但前人的研究結果可用來了解主題模型方法產生的主題結構，探討這個方法的適用性。例如：Topic 3「學術生產的趨勢分析」可歸入Schoepflin與Glänzel（2001）提出的「書目計量學理論與數學模型」；Topic 2「h類型指標」、Topic 4「期刊影響係數」和Topic 7「書目計量指標」等主題屬於「方法學與應用」和「指標工程與資料呈

現」；Topic 6「學術出版審查」和Topic 9「科學家特性與社會關係」都為「社會學取向」的一部分；Topic 1「科技的創新發展及知識傳播網絡」、Topic 5「研究機構成效的衡鑑」和Topic 8「科學研究合作」可納入「科學政策與科學管理」和「案例研究」。

㈡以主題模型方法的結果而言，在每一個主題上都受到相當的關注，但大抵而言，這段期間仍可發現實務型的研究較受到重視，包括應用專利分析的科技創新研究以及h類型指標與期刊影響係數的相關研究，另外，論文審查與科學家特性和社會關係等社會學相關的議題，雖然比案例研究受到較少重視，但差異並不大。此外，本研究也發現最近三個年度（2011至2013年）之間主題分布的差異較小，表示這個領域近期較趨於穩定。

㈢在將產生的主題結構運用於較相關期刊的主題特色時，可觀察到兩種期刊都廣泛地涉及資訊計量學上的所有主題，但*Journal of Informetrics*特別著重於書目計量指標的建構與應用，而*Scientometrics*則關注於國家、機構、領域及期刊上的學術生產力評鑑方式與影響因素的探討。

未來關於主題模型方法的研究，除了可將研究的論文來源與時間範圍擴大，對研究的領域進行更全面與深入的分析以及進一步探討主題模型方法的可行性以外，從本研究的結果也指出一些未來需要結合不同研究方法的研究方向，例如：2008年的主題分布與其他各年度之間有較大差異、2011至2013年主題分布變化較小、h類型指標與期刊影響係數等書目計量指標相關主題的關連性等現象。開發主題命名或摘要技術以及比較主題之間的差異與關連也是重要研究發展方向。此外，特別要加以說明的是目前本研究的研究資料採用WoS引文資料庫上的論文資料。由於目前大多數期刊都已建置論文資料庫，在進行後續研究時，可直接從期刊本身建置的論文資料庫取得研究資料，以避免文獻資料庫登載錯誤的情形。

誌　謝

感謝2014年海峽兩岸圖書資訊學學術研討會和本期刊的審稿者給予寶貴的修改意見。

參考文獻

Bhattacharya, S., & Basu, P. K. (1998). Mapping a research area at the micro level using co-word analysis. *Scientometrics*, *43*(3), 359-372. doi:10.1007/BF02457404

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

Börner, K., Chen, C., & Boyack, K. W. (2003). Visualizing knowledge domains. *Annual Review of Information Science and Technology*, *37*(1), 179-255. doi:10.1007/BF02457404

Chen, C., McCain, K., White, H., & Lin, X. (2002). Mapping *Scientometrics* (1981-2001). *Proceedings of the American Society for Information Science and Technology*, *39*(1), 25-34. doi:10.1002/meet.1450390103

Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2011). An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the fuzzy sets theory field. *Journal of Informetrics*, *5*(1), 146-166. doi:10.1016/j.joi.2010.10.002

Ding, Y., Chowdhury, G. G., & Foo, S. (2001). Bibliometric cartography of information retrieval research by using co-word analysis. *Information Processing & Management*, *37*(6), 817-842. doi:10.1016/S0306-4573(00)00051-0

Dutt, B., Garg, K. C., & Bali, A. (2003). Scientometrics of the international journal *Scientometrics*. *Scientometrics*, *56*(1), 81-93. doi:10.1023/A:1021950607895

Egghe, L. (2012). Five years "*Journal of Informetrics*". *Journal of Informetrics*, *6*(3), 422-426. doi:10.1016/j.joi.2012.02.003

Glenisson, P., Glänzel, W., Janssens, F., & De Moor, B. (2005). Combining full text and bibliometric information in mapping scientific disciplines. *Information Processing & Management*, *41*(6), 1548-1572. doi:10.1016/j.ipm.2005.03.021

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

Grün, B., & Hornik, K. (2011). Topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, *40*(13), 1-30.

Hofmann, T. (1999). Probabilistic latent semantic analysis. In *Proceedings of the fifteenth conference on uncertainty in artificial intelligence* (pp. 289-296). San Francisco, CA: Morgan Kaufmann.

Hood, W. W., & Wilson, C. S. (2001). The literature of bibliometrics, scientometrics, and informetrics. *Scientometrics*, *52*(2), 291-314. doi:10.1023/A:1017919924342

Leydesdorff, L. (1997). Why words and co-words cannot map the development of the sciences. *Journal of the American society for information science*, *48*(5), 418-427. doi:10.1002/(SICI)1097-4571(199705)48:5<418::AID-ASI4>3.0.CO;2-Y

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

Meyer, D., Hornik, K., & Feinerer, I. (2008). Text mining infrastructure in R. *Journal of Statistical Software*, *25*(5), 1-54.

Mimno, D., & McCallum, A. (2007). Mining a digital library for influential authors. In R. Larson, E. Rasmussen, S. Sugimoto, & E. Toms (Eds.), *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 105-106). New York, NY: ACM.

Noyons, E. C. M., Moed, H. F., & Van Raan, A. F. J. (1999). Integrating research performance analysis and science mapping. *Scientometrics*, *46*(3), 591-604. doi:10.1007/BF02459614

Noyons, E. C. M., & Van Raan, A. F. J. (1998). Advanced mapping of science and technology. *Scientometrics*, *41*(1/2), 61-67. doi:10.1007/BF02457967

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

Schoepflin, U., & Glänzel, W. (2001). Two decades of "Scientometrics". An interdisciplinary field represented by its leading journal. *Scientometrics*, *50*(2), 301-312. doi:10.1023/A:1010577824449

Schubert, A., & Maczelka, H. (1993). Cognitive changes in scientometrics during the 1980s, as reflected by the reference patterns of its core journal. *Social Studies of Science*, *23*(3), 571-581.

Thelwall, M. (2009). *Introduction to webometrics: Quantitative web research for the social sciences*. San Raphael, CA: Morgan & Claypool.

van den Besselaar, P., & Heimeriks, G. (2006). Mapping research topics using word-reference co-occurrences: A method and an exploratory case study. *Scientometrics*, *68*(3), 377-393. doi:10.1007/s11192-006-0118-9

Wouters, P., & Leydesdorff, L. (1994). Has Price's dream come true: Is scientometrics a hard science? *Scientometrics*, *31*(2), 193-222. doi:10.1007/BF02018560

Zheng, B., McLean, D. C., Jr., & 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

Journal of Educational Media & Library Sciences 51 : 4 (Summer 2014) : 499-523
DOI:10.6120/JoEMLS.2014.514/0633.RS.AM

Research Article

Analyses of Research Topics in the Field of Informetrics Based on the Method of Topic Modeling

Sung-Chien Lin

Abstract

In this study, we used the approach of topic modeling to uncover the possible structure of research topics in the field of Informetrics, to explore the distribution of the topics over years, and to compare the core journals. In order to infer the structure of the topics in the field, the data of the papers published in the Journal of Informetrics and Scientometrics during 2007 to 2013 are retrieved from the database of the Web of Science as input of the approach of topic modeling. The results of this study show that when the number of topics was set to 10, the topic model has the smallest perplexity. Although data scopes and analysis methods are different to previous studies, the generating topics of this study are consistent with those results produced by analyses of experts. Empirical case studies and measurements of bibliometric indicators were concerned important in every year during the whole analytic period, and the field was increasing stability. Both the two core journals broadly paid more attention to all of the topics in the field of Informetrics. The Journal of Informetrics put particular emphasis on construction and applications of bibliometric indicators and Scientometrics focused on the evaluation and the factors of productivity of countries, institutions, domains, and journals.

Keywords: *Analyses of research topics; Informetrics; Topic modeling*

SUMMARY

Introduction

This study used the method of topic modeling to analyze important research topics in the field of Informetrics and the development in recent years. Using texts of a collection of documents as input and the topic modeling method, we generated sets of statistical parameters consisting of a set of possible topics as well as topic mixtures for all documents in the collection. In this study, the texts of papers published from 2007 to 2013 in *Scientometrics* and *Journal of Informetrics*, which are core journals in the field of Informetrics, were used as data source. The topic model resulted from the study was provided for analysis of the examined field. This study investigated the following problems:

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

1. Identify important topics studied in the field from 2007 to 2013. Discuss the feasibility of applying the topic modeling method to this kind of studies;

2. Understand the distributions and developments of topics in the field of Informetrics;

3. Compare the distributions and developments of topics in *Scientometrics* and *Journal of Informetrics*.

Methods

The method of topic modeling which describes the word generation process of text in a document, is based on a statistical model (Blei, Ng, & Jordan, 2003; Griffiths & Steyvers, 2004). It assumes that every word w in a document d is generated by a topic z, which is sampled from the topic probabilistic mixture θ_d of the document d, as well as the corresponding word probabilistic mixture ϕ_z of the topic z, which can be used to select a word from a vocabulary. Furthermore, to describe the possible word generation process in a document collection, we need two sets of parameters: the first set of parameters (θ_d) is the probability determining the selection of topics for each document. The second set (ϕ_k) determines which word will occur when a specific topic is chosen. Several algorithms have been proposed to infer those unknown parameters of topic models. The parameter inference algorithm used in this study is Gibbs sampling proposed by Griffiths & Steyvers (2004).

To infer parameters of topic models, texts were collected and prepared first before being applied with the algorithm. In this study, data of the papers were retrieved from the citation database WoS (Web of Science). The search criteria were 1) paper type was "Article", 2) publication year was from 2007 to 2013, and 3) the source was either "*Scientometrics*" or "*Journal of Informetrics*". Texts in the fields TI (title) and AB (abstract) of a paper were merged into on piece of text for topic modeling. All punctuations, numbers and stop words in the texts were removed and all words in the remaining texts were transformed into lower case using the functions provided by the R statistical analysis software package "tm" (Meyer, Hornik, & Feinerer, 2008). We also counted the occurrences of all word tokens in the document collection, and the common words (appearing in more than 5% of the documents) and rare words (appearing in less than 1% of the documents) were deleted to reduce computational resources. Finally, the frequency counts of word tokens occurring in each document were used as input for another R package "topicmodels" (Grün & Hornik, 2011) to infer the statistical parameters of topic models.

In terms of interpreting the inferred topic model, we defined "core words" and "typical papers" for each topic in this study. The core words of a topic were

those words with the highest probability of this topic. Typical papers for a topic were those papers which the examined topic had the highest probability among all topics. When the inference of topic model for the document collection completed, both the word probability for each topic, ϕ_k, and the topic probability for each document, θ_d , were sorted in descending order. The core words and the typical papers for each topic were obtained at this point. In this study, 10 core words and 10 typical papers were selected for each topic.

The probability of topics in a set of papers can be calculated by averaging the probabilities of the corresponding topics in all the papers in the set. For example, to analyze topics occurred in the field of Informetrics from 2007 to 2013, the average of topic probabilities of all papers in the entire collection was used. If a specific topic had the highest average probability in the entire collection, the topic could then be considered as the most important one in the field. Therefore, we would rank all topics occurred in the whole field by comparing the average probabilities in the total collection. In the similar way, we could calculate the probability distributions of topics in a certain journal or in publications from a certain year by averaging the topic distributions of all papers published in the journal or in the certain year.

The disparity of two topic distributions in difference paper collections can be measured with the symmetric Kullback-Leibler divergence (Rzeszutek, Androutsos, & Kyan, 2010). When the two distributions are equal, the measure of symmetric Kullback-Leibler divergence between them is 0. If there are big differences between the two distributions, the measure will be larger than 0.

Results

In this study, a total of 1,755 papers were selected, including 1,374 papers from *Scientometrics* and 381 papers from *Journal of Informetrics*. The texts (the titles and abstracts of the papers) were prepared and 1,147 word tokens were extracted. Estimating through the leave-one-out method, the minimal average perplexity is 683.46 while the number of topics was assigned to 10. Therefore, we set the number of topics to 10 in this study. The followings are these 10 inferred topics:

- Topic 1: studies of technology innovation and knowledge communication networks based on the techniques of patent analysis;
- Topic 2: theoretic studies and applications of the h-type indices;
- Topic 3: growth and trends of academic productivity;
- Topic 4: empirical studies related to journal impact factor;
- Topic 5: performance evaluation and ranking for research institutions, including productivity and quality;

- Topic 6: issues of scholar publishing, such as paper review;
- Topic 7: measures and applications of bibliometric indicators;
- Topic 8: studies of scientific collaboration;
- Topic 9: influence of demography and characteristics of scientists on their performance and the social relationships among scientists;
- Topic 10: domain analysis, including the studies of topic identification and scientific maps.

Some findings from analysis of the resulting topic model show in follows:

The topic distributed in the entire field was fairly balanced from 2007 to 2013. The probabilities of all topics in the entire collection were almost the same. However, Topic 1 and Topic 4 were slightly larger than the others.

The topic distribution in 2008 was different from those measured by the symmetric Kullback-Leibler divergence in other year. It was because the probabilities of Topic 1 and Topic 2 in 2008 were significantly different from those in other years.

The topic distributions in the last 3 years were very similar and they were different from the topic distributions in the first 4 years. The changes of Topic 3, Topic 6 and Topic 7 contributed to the different topic distributions. The probability distribution of the Topic 3 in the earlier 3 years was about 0.11, but in later 4 years it was only 0.10 or less. There was a clear downward trend. Conversely, the probability distributions of the Topic 6 and the Topic 7 were less than 0.09 and 0.10 in the first 4 years and were raised to 0.10 and 0.11 in the later years.

Journal of Informetrics and *Scientometrics* both published various topics in the field of Informetrics. But the *Scientometrics* published a wider range of topics. *Scientometrics* had slightly higher probability distributions on Topic 1 and Topic 8 and it also preferred Topic 3, Topic 4 and Topic 5. The *Journal of Informetrics* had obviously higher probability distributions on Topic 2, Topic 7, Topic 4, and Topic 10.

Discussion

According to the results, this study makes the following conclusions:

1. Although the data scope and the methods used in the analyses are not the same, the results of this study are still consistent with those of previous studies. For example, Topic 3 can be assigned into the class "bibliometric theory, mathematical models and formalisation of bibliometric laws" provided by Schoepflin & Glänzel (2001), and Topic 6 and 9 are parts of "sociological approach to bibliometrics".

2. Each topic in the field of Informetrics has received considerable attentions,

especially practice-based research such as the applications of patent analysis and journal impact factor.

3. Because there is only little difference between the topic distributions in last three years, this field was more stable than before.

4. Through comparing the topic distributions of the two core journals, we can observe that these journals published all topics in the field of Informetrics. However, *Journal of Informetrics* has special focus on the construction and applications of various bibliometric indicators. *Scientometrics* has put attention on the evaluation of academic productivity of countries, institutions, fields and journals.

ROMANIZED & TRANSLATED REFERENCE FOR ORIGINAL TEXT

Bhattacharya, S., & Basu, P. K. (1998). Mapping a research area at the micro level using co-word analysis. *Scientometrics*, *43*(3), 359-372. doi:10.1007/BF02457404

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

Börner, K., Chen, C., & Boyack, K. W. (2003). Visualizing knowledge domains. *Annual Review of Information Science and Technology*, *37*(1), 179-255. doi:10.1007/BF02457404

Chen, C., McCain, K., White, H., & Lin, X. (2002). Mapping *Scientometrics* (1981-2001). *Proceedings of the American Society for Information Science and Technology*, *39*(1), 25-34. doi:10.1002/meet.1450390103

Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2011). An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the fuzzy sets theory field. *Journal of Informetrics*, *5*(1), 146-166. doi:10.1016/j.joi.2010.10.002

Ding, Y., Chowdhury, G. G., & Foo, S. (2001). Bibliometric cartography of information retrieval research by using co-word analysis. *Information Processing & Management*, *37*(6), 817-842. doi:10.1016/S0306-4573(00)00051-0

Dutt, B., Garg, K. C., & Bali, A. (2003). Scientometrics of the international journal *Scientometrics*. *Scientometrics*, *56*(1), 81-93. doi:10.1023/A:1021950607895

Egghe, L. (2012). Five years "*Journal of Informetrics*". *Journal of Informetrics*, *6*(3), 422-426. doi:10.1016/j.joi.2012.02.003

Glenisson, P., Glänzel, W., Janssens, F., & De Moor, B. (2005). Combining full text and bibliometric information in mapping scientific disciplines. *Information Processing & Management*, *41*(6), 1548-1572. doi:10.1016/j.ipm.2005.03.021

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

Grün, B., & Hornik, K. (2011). Topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, *40*(13), 1-30.

Hofmann, T. (1999). Probabilistic latent semantic analysis. In *Proceedings of the fifteenth*

conference on uncertainty in artificial intelligence (pp. 289-296). San Francisco, CA: Morgan Kaufmann.

Hood, W. W., & Wilson, C. S. (2001). The literature of bibliometrics, scientometrics, and informetrics. *Scientometrics*, *52*(2), 291-314. doi:10.1023/A:1017919924342

Leydesdorff, L. (1997). Why words and co-words cannot map the development of the sciences. *Journal of the American society for information science*, *48*(5), 418-427. doi:10.1002/(SICI)1097-4571(199705)48:5<418::AID-ASI4>3.0.CO;2-Y

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

Meyer, D., Hornik, K., & Feinerer, I. (2008). Text mining infrastructure in R. *Journal of Statistical Software*, *25*(5), 1-54.

Mimno, D., & McCallum, A. (2007). Mining a digital library for influential authors. In R. Larson, E. Rasmussen, S. Sugimoto, & E. Toms (Eds.), *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 105-106). New York, NY: ACM.

Noyons, E. C. M., Moed, H. F., & Van Raan, A. F. J. (1999). Integrating research performance analysis and science mapping. *Scientometrics*, *46*(3), 591-604. doi:10.1007/BF02459614

Noyons, E. C. M., & Van Raan, A. F. J. (1998). Advanced mapping of science and technology. *Scientometrics*, *41*(1/2), 61-67. doi:10.1007/BF02457967

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

Schoepflin, U., & Glänzel, W. (2001). Two decades of "Scientometrics". An interdisciplinary field represented by its leading journal. *Scientometrics*, *50*(2), 301-312. doi:10.1023/A:1010577824449

Schubert, A., & Maczelka, H. (1993). Cognitive changes in scientometrics during the 1980s, as reflected by the reference patterns of its core journal. *Social Studies of Science*, *23*(3), 571-581.

Thelwall, M. (2009). *Introduction to webometrics: Quantitative web research for the social sciences*. San Raphael, CA: Morgan & Claypool.

van den Besselaar, P., & Heimeriks, G. (2006). Mapping research topics using word-reference co-occurrences: A method and an exploratory case study. *Scientometrics*, *68*(3), 377-393. doi:10.1007/s11192-006-0118-9

Wouters, P., & Leydesdorff, L. (1994). Has Price's dream come true: Is scientometrics a hard science? *Scientometrics*, *31*(2), 193-222. doi:10.1007/BF02018560

Zheng, B., McLean, D. C., Jr., & 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