教育資料與圖書館學

Journal of Educational Media & Library Sciences

http://joemls.tku.edu.tw

Vol. 51，特刊 (2014) : 3-26

自動化資訊組織與主題分析
近二十年來的研究與發展

Research and Development on Automatic Information Organization and Subject Analysis in Recent Decades

曾 元 顯 Yuen-Hsien Tseng
Research Fellow
E-mail：samtseng@ntnu.edu.tw

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

教育資料與圖書館學　51 : 特刊 (2014) : 3-26
DOI:10.6120/JoEMLS.2014.51S/0652.RV.AM

回顧評論

自動化資訊組織與主題分析近二十年來的研究與發展

曾元顯

摘要

資訊組織與主題分析是圖書資訊學探討的重點課題。隨著資訊科技進步，全球資訊網的興起，數位文件的數量越來越龐大，各種文件處理、加值、活化運用的需求增多，自動化資訊組織與主題分析的必要性越來越高。本文先簡介相關自動化技術近二十年來的發展，並以採、編、典、藏、用的觀點，看待整個自動化作業的過程；接著，以一些應用實例具體說明其效益，特別著重在關鍵詞擷取、關聯詞分析、文件自動組織、主題分類建構等應用，讓讀者具體領略其實務上的可行性；最後，說明這二十年來資訊技術的進步幅度快速、應用範圍廣泛，在此僅能探討到一小部分。透過這些介紹與案例展示，本文希望促進圖書資訊學與其他學科人員，互相交流與合作的機會，以開展創新的資訊服務。

關鍵詞：關鍵詞擷取，關聯詞分析，文件歸類，主題分類，資訊檢索

前　言

資訊組織與主題分析向來是圖書資訊學探討的重點課題。圖書資訊學以宏觀的角度，從政策、策略、社會因素到資訊技術應用等面向，探討各類文件、資訊在生產與消費的生命週期中，牽涉到的採訪、編目、典藏、存取、加值、利用與服務等議題。而資訊組織與主題分析便牽涉到其中偏向策略與技術的編目、典藏與存取的加值處理議題。然而，無論各類文件或資訊如何被組織與分析，其最終目的，是要提供使用者（或用戶）便利的資訊利用。

傳統上，資訊組織與主題分析處理的資料內容，主要是圖書館的各式館藏，包含圖書、期刊、報紙、影音檔案等各類出版品或歷史文件，且主要的服務對象為一般大眾（公共圖書館）、學校師生（學術或教育圖書館），以及特定

國立臺灣師範大學資訊中心研究員
通訊作者：samtseng@ntnu.edu.tw

2014/10/29投稿；2014/11/09修訂；2014/11/10接受

領域機構人員（專門圖書館）。近數十年來，在電腦網路普及、數位文件風行之後，資訊組織與主題分析可應用的範圍（處理的文件類型與服務的用戶類別），可擴大到任何有潛在價值的資訊利用情境。其應用案例，可包含網頁資料查找、學術文件搜尋、訴訟（或專利）的前案檢索、新聞事件歸類、網民意見追蹤調查、垃圾郵件過濾、文件自動摘要、關聯資訊擷取、生物資訊探勘、主題趨勢辨識、商業智慧分析、自動詢答系統等，都成為資訊組織與主題分析可以探討的課題。

因此，本文就上述應用方向，探討自動化資訊組織與主題分析近二十年來的發展。而傳統的資訊組織與主題分析議題，在圖書館學方面，是屬於基本的核心課程（Chan, 2007; Chowdhury, 2010; Ogden, 1977; Olson & Boll, 2001; Taylor & Joudrey, 2008），有悠久的發展歷史、眾多的教科書以及實務的標準作業流程，便不在本文探討的範圍。

本文探討的自動化發展，技術不斷創新、相關知識演變快速。雖可說是資訊科技（工程或管理）的範圍，但其主要流程、所需背景知識，以及猶待解決的問題，還是脫離不了上述：「採」（有潛在價值文件的採集與界定）、「編」（分析、組織、編目）、「典」（資料的標準化與作業的法則化）、「藏」（有效率的儲存與可長久的保存）、「用」（資訊的利用）等五項概念與流程。因此，這項看似傳統的議題，在現在的環境中，其實牽涉了多種領域，特別與近年來的資訊檢索、文字探勘、知識管理、機器學習、自然語言處理等研究，有高度的相關性。

舉例而言，在1999年，瑞士蘇黎世Fritz Kutter基金會舉辦了「自動編目與檢索競賽」（Automatic Cataloguing and Searching Contest）[1]。主辦單位提供了英文、德文、法文、義大利文等四國語言近200年來的500本書籍資料。凡跟編目有關的頁面，如封面、書名頁、版權頁、出版序、目錄、封底等，都被掃描成影像檔後，再利用光學字元辨識（Optical Character Regnition, OCR）軟體轉換成數位文字提供給參賽者。而其競賽題目，分為15道布林表達式（欄位式）查詢題，以及15道以自然語言描述的查詢題目。參賽者需根據這些資料以及題目，提交系統產出的答案。從獲獎者的技術報告看來（Tseng, 2001），自動化索引（資訊檢索）、關鍵詞擷取（文字探勘）、完整文句的詢答（自然語言處理）、進階檢索模式的運用（機器學習）等技術，都被運用到。

此項競賽揭示了下列意涵：㈠此競賽的舉辦動機，提到：近年來圖書編目的成本越來越高，甚至有高過圖書價格本身的情況（歐美地區人事成本昂貴，要做到正確的主題分析與編目，人員訓練與編目時間成本高昂；而且事先的人工編目，能否符合事後使用者的各種利用，還很難說），考量到目前資訊技術

1 該競賽網頁：http://www.kutter-fonds.ethz.ch/contest99.html已不存在，只剩歷屆競賽項目與得獎名單：http://www.kutter-fonds.ethz.ch/PreviousPrizewinners.html

進步的情況，因而舉辦此項競賽，以了解自動化技術可達到的績效程度，可否取代或輔助人工的編目，從而降低人事或時間的成本。因此，自動化資訊組織與主題分析，不是要不要的問題，而是針對某項應用，其效果如何、如何運用的問題。㈡主辦單位不看系統自動編目或索引的過程與內容，而只從使用者端看系統反應的結果。亦即，系統如何進行自動編目或索引，不是使用者關心的議題（卻是必要的資訊處理過程），系統只要好用、可靠，能輕易檢索出想要的資料即可。因此，就整體的資訊利用服務而言，資訊組織與主題分析的處理過程是幕後的基石，而資訊檢索系統則是前端重要的使用媒介。

綜上所述，自動化資訊組織與主題分析，將運用到許多不斷精進的技術。然而，為免陷於技術至上的迷思，或完全依賴工具解決問題，對各種類型文件或知識資產的資訊組織與主題分析，可以用：採、編、典、藏、用，這五項要訣來看待其整個自動化過程，以了解全貌，並讓文件處理的目的不會失焦。

特別是在全球資訊網（World Wide Web, WWW）（近二十年來）的時代，以此五項要訣的角度來看，其可發揮的潛在效益，非常巨大。以Google為例，其僅「採」集網頁，並以極其簡潔的介面，讓使用者方便快速的運「用」，即創造巨大的產值與個人的效益。當然，Google在「編、典、藏」方面的先進技術，才是其成功之處。然而，Google所創造出的效益，主要還是拜WWW網頁的內容夠多、數量夠大、資訊類型廣泛，以及長尾效應下能夠滿足各方不同需求的特性所賜。相對的，除了網頁有超連結的特性外，其他類型的文件，特別是機構、企業內的文件，少了這些特性，Google技術所能產生的效益便難以移植。

本文將先廣泛介紹上述自動化資訊組織與主題分析的相關研究，其次簡要敘述重要的技術發展趨勢，最後以一些應用實例具體說明其效益。本文一方面希望圖書資訊學相關人員，不侷限於圖書館的情境，往外看到更多新的需求、應用與進展；另一方面，也希望向不同領域的人員，介紹圖書資訊學看待文件處理、知識加值的宏觀概念與既有的智慧，以促進各領域之間相互交流與合作的機會。

二、相關研究與發展趨勢[2]

如同前述，自動化的資訊組織與主題分析，與資訊檢索、文字探勘、知識管理等議題息息相關，且幾乎都以由來已久的資訊檢索一詞，來涵蓋前述自動化資訊組織與主題分析的範圍（Salton, 1989）。因此，下面的相關研究介紹，將資訊檢索等詞彙視為「採、編、典、藏、用」的相關概念一起探討，不特別區分其細微差別。

2本節有大部分改寫自：圖書資訊學學術研究之回顧與前瞻一書中，由曾元顯撰寫的第七篇第二章「資訊檢索技術發展趨勢」。

1990年代初，美國政府單位開始資助一系列的相關研究，包括TIPSTER計畫（Harman, 1992）、MUC（Message Understanding Conference）會議（Sundheim, 1991）、TREC（Text REtrieval Conference）[3]資訊檢索評比會議，以探索從大量訊息中偵測、擷取與摘要相關資訊的技術，其原先目的雖為增進情報分析能力，後來則演變成較為學術且應用更廣的研究計畫。其中，特別是TREC評比會議，自1991年起每年舉辦，至今已有二十餘年（江玉婷、陳光華，1999）。歷年來，其根據文件類型、資訊需求的不同，舉行過各種研究任務的評比，包括：主題檢索、資訊過濾、跨語檢索、全文影像辨識與檢索、語音與視訊檢索、與使用者互動的情境檢索、巨量資料檢索、超連結之網頁檢索、新事件偵測、詢答系統、生物資訊檢索、法律文件檢索、部落格訊息檢索等，都是從「用」的角度看背後的技術可達到何種成效。（除此之外，另一種「用」的觀點是，系統的設計是否符合人類的使用習慣、直觀的認知，或創新簡易而又有趣的體驗方式，因此有另一個相關的研究，如：Human Computer Interaction International Conference，專門探討這類議題。）此評比會議的重要貢獻，在提供一套共用的測試集（包含文件集、問題集，與對應於每道問題的答案），以及相同的評估準則與程序，使得各個研究團隊可在相同的實驗環境底下，反覆的自我與互相比較，促使相關的技術能夠得到真正的進步。

由於TREC對相關研究社群的貢獻甚鉅，且有益於該國語文的資訊檢索與情報分析研究，日本、歐洲、印度紛紛起而效尤，仿照TREC形式，各自舉辦了NTCIR（NII Testbeds and Community for Information access Research，始於1999年）、CLEF（Conference and Labs of the Evaluation Forum，更早名稱為：Cross-Language Evaluation Forum，始於2000年）與FIRE（Forum for Information Retrieval Evaluation，始於2008年）等具該國語文特色的資訊檢索評比會議。這些評比，雖以資訊檢索為主軸，事實上也觸及資訊擷取、文字探勘、機器學習等相關議題，如日本的NTCIR曾有專利資訊探勘評比任務，歐洲的CLEF則有文字、語音、影像、跨語檢索等任務。由於TREC的模式，能提供大規模、可比較、能重複利用的測試集與評估機制，讓有興趣的研究團隊做有效利用，大幅降低相關研究的門檻，因而近二十年來，有各式、大量的資訊組織方式或主題分析技術被提出來。

受網際網路全文網頁以及數位出版的蓬勃發展影響，文件處理有更多的人員投入研究，早期從ACM SIGIR（ACM Special Interest Group on Information Retrieval）會議獨自探討，到現在有更多的國際研討會，如CIKM（ACM Conference on Information and Knowledge Management）、WSDM（ACM International

3 TREC會議以及後續提到的日本NTCIR、歐洲的CLEF、印度的FIRE論壇、SIGIR等會議、JASIST等期刊，都可在網路上找到其網站以及大量的相關文獻，為節省版面，在此不特別顯示其網址與參考書目。

Conference on Web Search and Data Mining）、JCDL（ACM/IEEE Joint Conference on Digital Libraries）、ECIR（European Conference on Information Retrieval）等，都將資訊檢索、文字探勘、知識管理列為主要的議題。而新興的國際期刊如*Information Retrieval*或既有的老牌期刊如*JASIST*（*Journal of the American Society for Information Science and Technology*）、*IPM*（*Information Processing and Management*）、*TOIS*（*ACM Transactions On Information Systems*）等，有關資訊組織或主題分析的議題也急速擴增。

國內的相關研究，早期主要集中於圖書館學領域的資訊儲存與檢索、使用者需求分析，與資訊系統的使用行為探討。自1990年代起，國內圖書資訊界與電腦科學界陸續有：網頁版公用目錄搜尋系統（曾元顯、林瑜一，1998）、整合代理搜尋系統（朱讚美，2000；謝欣君、張玉山、袁賢銘，1998）、中文搜尋引擎（Chien, 1995a; Chien & Pu, 1996）、模糊搜尋（曾元顯、林瑜一，1998）、音樂內容檢索（Tseng, 1999）、語音檢索（Bai, Chen, Chien, & Lee, 2002）、智慧搜尋（Chien, 1995b）、網頁資訊擷取（Chang & Lui, 2001）、中文詢答系統（Sasaki, Chen, Chen, & Lin, 2005）、文字探勘（Tseng, Lin, & Lin, 2007）、影像內容檢索（Lin, Chang, & Chen, 2005）等研究的投入。近十多年來也培養出相當多的碩、博士級人才，相關的研究成果、技術與人力資源，也都逐漸移轉至國內產業界。然而，跟國外或中國大陸等具有龐大市場的地區相比，國內研究的能量與影響力仍有待持續投入與提昇，以便使正體中文的文件處理技術能隨國外技術的演進而不斷進步，甚至能獨樹一格，成為正體中文相關研究的全球重鎮。

過去這二十年來，國內外相關技術不斷進步，但累積的知識也顯示：沒有一種方法可有效處理各式文件以及各類應用，自動化資訊組織與主題分析的技術，需隨待處理文件與應用方式的不同而加以調整。

即便如此，這段期間有多項技術的突破與進展，對世人的日常生活產生根本的影響。展望未來，此影響仍將持續，並隨新技術、新環境、新需求而更行發展。下面特別針對這方面的基礎研究與核心技術部分，簡要說明近二十年來的進展。

㈠索引建構

對文件的加值處理而言，索引建構是所有資訊組織與主題分析任務的基礎。過去學術界與實務界已發展出實用且高效率的演算法（Van Rijsbergen, 1979），以及適合索引結構的資料壓縮方法（Witten, Moffat, & Bell, 1999），可同時降低磁碟空間的需求，並提高執行效率。Google也曾提出其Google File System（Ghemawat, Gobioff, & Leung, 2003），用以對付巨量資料。近年來國際重要的研討會雖已少見此議題，但從實務面上，此議題仍極為重要。如何設計良好的索引結構，以應付即時的文件更動、容納上百種排序特徵、結合使用者權

限，使得依照個人興趣、使用情境或社會網絡集體行為而調整的精準排序，不限於巨大的商業搜尋引擎或雲端架構才能擁有，而可落實到伺服器、桌機等級的電腦設備，且為注重隱私的企業、機構，甚至個人所能採用。這些議題仍是自動化資訊組織有待突破的研究課題。

㈡檢索模型

配合索引結構，檢索模型近年來重大的發展與長足的進步，過程如下：從早期的布林邏輯精確比對，完全仰賴使用者提出精簡正確的查詢詞彙來控制整個檢索流程；進展到向量空間模型，允許查詢條件以自由文字描述；再到機率模型，獲得更有效的排序；近十年來則有語言模型，讓各種改進得以有可靠的數學邏輯推演，脫離經驗法則式的權重猜測與設定，解決字彙不匹配的問題；最近，在商業搜尋引擎採用上百種特徵據以排序檢索結果後，排序學習變成拯救人工調整上百個特徵權重的必要手段。排序學習不僅成為近年來最重要的商業智慧之一，也將與語言模型繼續成為未來檢索理論的重要研究課題。

㈢查詢模式

查詢模式是屬於上述五項概念中「用」的部分，是最接近使用者經驗的流程。相較於系統底層的索引建構與系統上層的檢索模型，使用者端的應用情境、使用模式、互動介面、查詢意圖等議題，乃主題分析應用成功之關鍵。另外，查詢的表達（formulation）、精鍊（refinement）、擴增（expansion）與回饋（feedback）等研究，也是資訊查找利用時的重要議題。如何開創新的應用情境（如趨勢偵測）、設計更便利的互動介面（如語音輸入、動態感知、擴增實境、3D呈現）、解讀使用情境的查詢意圖（如根據所在位置或前後瀏覽的文件，正確解讀所需資訊），都是新一代主題分析技術有待研發之課題。

㈣延伸應用

在1990年代中期Web被開始大量運用於電子商務後，各種資訊檢索、擷取、組織、分析、探勘的需求，進一步刺激了這個領域的發展，更多相關議題被提出討論，如詢答系統、資訊擷取、主題分類、事件歸類、訊息過濾、複本偵測、多媒體資訊檢索、跨語檢索、廣告推薦、社群意見分析與惡意訊息的防範等。過去這些課題雖有相當的研究，然而針對不同文件的類型（如書目、新聞、網頁、部落格、學術文獻、社群網站討論群組、個人偏好與健康資訊等）、領域（如法條、專利、生物資訊）與樣態（如文字、語音、空間資訊），都有需要客製微調與整合之處，甚至於需要應用或開創新的處理模式，以達到更高的成效。

㈤行動搜尋

近年來無線網路普及，智慧手機、平板電腦等行動裝置的大量使用，造就行動中搜尋資訊的需求。這類需求，不僅止於資訊的擷取，常伴隨著後續的行動。例如，旅行在陌生的環境中，以語音輸入或以手機拍攝實物影像，再藉由語音辨識、文字辨識或圖像比對，自動尋找最近的相關服務（如具特定菜色的餐館，或具特定造型、設施的旅店），進而上線預約或電話預定等。這類將靜態資訊（如：地址）與動態資訊（如：路線、訂位狀況）串連，資訊流與金流（預約費用）整合，加上多模態資訊（文字、語音、影像、地圖資訊）的擷取與呈現（如：資訊視覺化、虛擬實境）之技術與服務，也是近年來重要的研究議題。

㈥整合發展

僅是前面兩項基礎研究（索引建構與檢索模型）的演進與應用，便造就了市值極高的搜尋引擎產業。其他傳統的重要議題與新興應用，則可讓資訊組織與主題分析的研究與技術，滲透到更廣泛的範圍，如數位典藏、知識管理、電子商務、市場民調、數位學習、資訊服務等。從搜尋引擎Google提供包含網誌、影音、翻譯、地圖與學術搜尋在內的各項資訊處理服務，即可知道自動化資訊組織與分析的研究與應用，在現在的時空環境，幾乎無所不在。

三、應用案例

由於相關的自動化技術，近年來不斷的演變、精進，在有限的篇幅難以窮舉，且每種新興的技術，在不同的應用場域成效各有所長。因此，本節僅就筆者過去研發的相關技術，說明其應用情境，供讀者了解其能發揮的成效，做為了解此類技術發展與應用現況（或足跡）的參考。

㈠關鍵詞擷取應用

「關鍵詞自動擷取」是一種辨認數位文件內有意義且具代表性字串（string）、片語（key phrases）、詞彙（keywords），或內容片段（key segments）的自動化技術。由於關鍵詞是呈現文件主題意義的最小單位，因此大部分對非結構化文件的自動處理，如自動索引、索引典自動建立、內容摘要、主題分類、文件歸類、資訊過濾、事件偵測與追蹤、知識探勘、資訊視覺化、相關回饋、檢索提示、關聯知識分析、自動化權威控制等，都必須先進行關鍵詞擷取的動作，再進行後續的處理。因此，關鍵詞擷取是所有文件自動處理的基礎與核心技術。

然而，關鍵詞的認定，有時頗為主觀。為能自動化處理，筆者假設關鍵詞為文件中重複的特定字串（其左接詞與右接詞，在文件的不同地方會有所不

同），而發展出一套快速、簡單、有效的規則，來擷取文件的關鍵詞彙。有趣的是，此自動擷取方法與語言文字關係不大，可運用於英文、中文文件（需過濾停用詞），以及光學自動辨識過的雜訊文件上（Tseng, 1998），甚至也可直接運用在多媒體的數位文件上，例如音樂檔案，以擷取其中的關鍵旋律（Tseng, 1999）。此項技術的優勢與應用範圍，是其他關鍵詞擷取技術，如（Chien, 1997; Witten, Paynter, Frank, Gutwin, & Nevill-Manning, 1999），沒有具備的特性。

根據我們之前針對此項技術的實驗，在沒有運用大量詞庫的情況下，書目資料的關鍵詞擷取準確度為90%，新聞全文資料的準確度為 86%。而在運用12萬詞的詞庫後，新聞全文的關鍵詞擷取準確度為96%，其中每篇新聞有33%個關鍵詞為詞庫中沒有收錄的詞彙。

此項技術，最早應用於輔仁大學的圖書書目檢索上（曾元顯、林瑜一，1998），後來試用於輔仁大學中國社會文化研究中心（簡稱社文中心）蒐藏的「中國消息分析」（China News Analysis, CNA）新聞剪報資料庫中（曾元顯，2002）。約在1953-1982年間，CNA在香港蒐集大陸各省報紙的新聞剪報，在1996-1998年時，社文中心將約60萬篇剪報掃描成影像檔，其中約有30萬份已

圖1 從OCR文件自動擷取關鍵詞提供查詢範例

人工輸入篇名、日期等欄位，讓使用者據以查詢跟調閱剪報影像檔案。為進一步研究全文搜尋的可行性，我們將其中約8,439篇經OCR光學文字辨識成數位文字檔（蔡孟竹、曾元顯，2003）。針對這些正確率平均僅70%的新聞，進行關鍵詞擷取，建成關鍵詞庫，再以模糊搜尋方式，提供使用者第一階段先搜尋關鍵詞庫，勾選適當的關鍵詞後，再進行第二階段的全文檢索。

圖1所示的範例（Tseng, 2001），在第一階段的關鍵詞模糊查詢時，可找出文件中字串相近的詞彙，其中包含近義詞，也包含辨識錯誤的同義詞。如圖1左例中「農村幹部」、「工農幹部」是查詢詞「鄉村幹部」的近義詞，而第三個詞彙「鄉村幹都」則是文件中辨識錯誤的查詢詞。圖1右例中則顯示，不必進行第二階段的全文檢索，第一階段的查詢，即可獲得文件中有用的資訊，亦即「反革命份子」，有「萬玉昌」、「王洪富」等人。此兩案例顯示，此種設計的查詢效益，不僅比單獨使用查詢詞直接檢索文件，可找到更多相關文件，而獲得較高的查全率（recall），同時也可更快了解文件中包含的訊息，而無須看到全文文件。從某種程度上，這已解決部分的「查詢不匹配」、「同義詞」、「近義詞」等需要人工維護權威控制詞的問題。

㈡關聯資訊應用

某些關鍵詞之間，具有主題上的相關（如屬於相同主題或事件），或具備某種屬性的關聯（如上、下位詞），若能善加利用，可提升檢索的成效、提示使用者更佳的檢索詞彙，或呈現文獻中隱含的知識。隨著應用場域的不同，擷取相關詞或其關係屬性的方法也相異，這方面的技術文獻相當多，如：（Chen, Yim, Fye, & Schatz, 1995; Hearst, 1992; Sanderson & Croft, 1999），惟成效不易比較。這十多年來我們也發展出兩套方法，以因應不同的應用場合。

第一套方法，乃依據主題上相關的詞彙常會一起出現在同一句子的現象，找出關聯詞彙，以便用於查詢提示或文獻內容摘要。

如圖2所示（Tseng, 2002），在一萬多篇的產業新聞資料庫中，使用者輸入「mp3」，系統立刻回應相關的詞彙，如：「音樂」、「CD」、「播放」等跟mp3主題上直接相關的詞彙；另外也回應了：「中環」、「國碩」等跟mp3播放器有關的公司。對於此項產業的初入門者，也許一開始不了解「中環」、「國碩」跟mp3的詳細關係，但從圖2點選「中環」即可顯示兩者的密切關係：文件顯示中環公司是mp3播放器的大廠，一年出貨量可達100萬台。另外，系統中若還包含預先建立的結構化資料，也可立即顯示中環公司的各項數據與資訊，讓使用者只下達一個查詢詞「mp3」、點選「中環」，即可獲取（access）領域專家等級的知識。這對初入門者可說是一項極為便利的知識取用工具，而其背後又無須進行太多的人工知識管理作業（只有廠商資料庫需要維護，但此項資料亦可透過其他公開或官方網站，以程式代理人自動抓取獲得）。

圖2 關聯詞應用於產業新聞檢索案例

圖3與圖4是第一套方法的另一項應用。國科會（現為科技部）在2013年頒發傑出研究獎給70位國內的優秀學者，並邀請其各自撰寫感言一篇，除了簡介自己的研究貢獻，篇末也有受獎人的致謝文字。受獎人之一的臺師大科教中心張俊彥主任獲邀代表70位受獎人上台致詞。但由於張主任不懂其他領域的研究內容，也不知大家在得知受獎後的情意面向（affective aspect），從過去跟筆者合作的經驗中（Tseng, Chang, Rundgren Chang, & Rundgren, 2010），他靈機一動，遂跟國科會事先索取這70篇感言，再交由筆者進行關鍵詞擷取與詞彙關聯分析，終能不負應邀代表致詞的托付。

圖3顯示這70位得獎學者的貢獻領域，左上角顯示半導體、光電、奈米以及細胞研究等領域，右邊則顯示醫學、物理、資訊、農業生技等領域，左下角則顯示勞工、觀光、產業績效等領域。綜合而言，社會科學受獎者比例較低，生、醫、農、理、工等領域的受獎人較多，反映出國內在這方面的研究成果，能見度較高。

圖4顯示這70位得獎者致謝感言的常見詞彙，從中可知國科會當然是最被感謝的對象，其後是學生、家人、師長。頒獎典禮上，張教授上台報告這兩張圖，沒有受到質疑，顯示其大略反映了大家的認知。無須詳讀並自己摘要這70篇感言，一圖勝千言的意境，不言而喻。

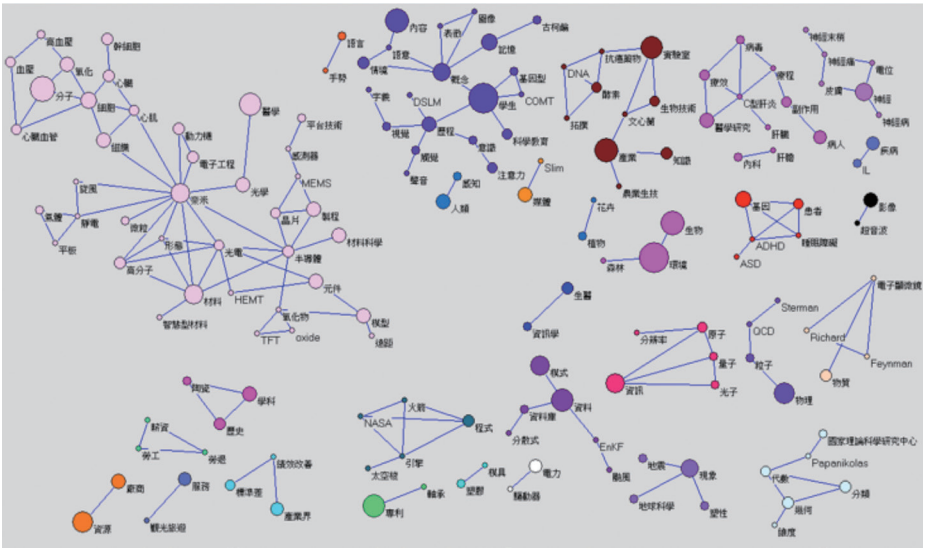

圖3　2013年70篇國科會傑出研究獎得獎
感言研究貢獻段落關鍵詞彙關聯圖（彩圖請見電子檔）

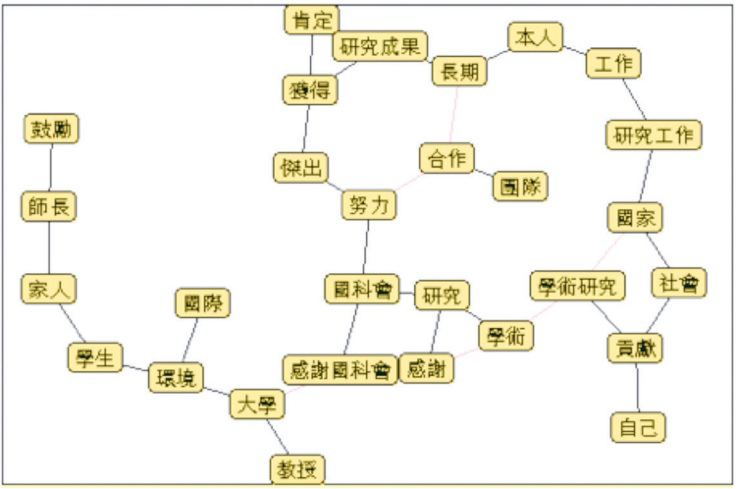

圖4　2013年70篇國科會傑出研究獎得獎
感言致謝段落關鍵詞彙關聯圖

第一套方法，可在圖2看到詞彙關聯圖的效益。然而該圖並沒有顯示「mp3」跟「音樂」與「中環」分別是何種關係。若能將詞彙間的關係釐清，則可據以進一步篩選資訊。例如，要求系統只顯示mp3播放器的「製造」廠商，那麼圖2就可去除無關的詞彙，讓出更多空間來顯示除了中環、國碩之外的可能廠商。

第二套方法，便是為了獲取詞彙之間明確的關係屬性而開發。我們參考相關的研究（Fader, Soderland, & Etzioni, 2011），並運用中研院資訊科學研究所

陳克健研究員的句法剖析程式（Hsieh, Bai, Chang, & Chen, 2012），以自然語言處理技術，發展一套中文開放領域關係擷取（Chinese Open Relation Extraction, CORE）系統（Tseng et al., 2014）。此套系統，可擷取文句中的三元素，如：（詞彙1，關係，詞彙2），此乃最基本的知識表達方式之一。例如，輸入一句：「愛迪生發明了燈泡。」，此系統可輸出：（愛迪生，發明了，燈泡）。

此種知識三元素，除可補強前述的關聯詞彙圖外，亦可直接運用於知識性的搜尋服務。我們將聯合報系2002-2009年約200萬篇的新聞，逐句剖析擷取後，建立上述三元素的資料庫，提供如圖5所示的運用。若使用者想查詢「哪些事務源自中國」，則可在關係詞中輸入「源」，在第三個輸入框中輸入「中國」，系統即可比對出符合這兩個欄位的結果。圖5可看出系統有機會把相關的答案列舉出來（雖非全部正確），而成為一個圖書館線上問答服務（參考服務）的自動化系統。

列號	主詞	關係詞	受詞	句子原文
1	SARS	起源於	中國大陸	SARS起源於中國大陸
2	拼布	最早發源	於中國古代	拼布最早發源於中國古代
3	膠彩畫	其實源自	中國唐宋的金碧山水和工筆	膠彩畫其實源自中國唐宋的金碧山水和工筆
4	魔術	源起	於古老的中國	魔術源起於古老的中國
5	公雞圖騰	源自	中國陝西地區的剪紙藝術	公雞圖騰源自中國陝西地區的剪紙藝術
6	資金	仍源源不絕匯入	中國大陸	資金仍源源不絕匯入中國大陸
7	漢醫	源自	二千年前的中國	漢醫源自二千年前的中國
8	納豆	起源於	中國的豆鼓	納豆起源於中國的豆鼓
9	鹿港火龍祭	源起	於中國古代每年中秋有一個火龍祭	鹿港火龍祭源起於中國古代每年中秋有一個火龍祭
10	牛肉麵	並非源自	中國大陸	牛肉麵並非源自中國大陸

圖5　什麼「源」自「中國」的詢問範例（彩圖請見電子檔）

關聯詞的應用還有很多。如新聞語料中有關國、高中生活與科技教科書概念的關聯圖，可運用於大眾素養量表（問卷）的設計，以了解大眾日常生活所需科學知識（Tseng et al., 2010）。如歷年來爭議不斷的核四是否續建、餿水油長期滲入到各類食品的事件。當碰到這類問題時，民眾有否足夠知識進行決策（選對民意代表與其政策）、平時更加注意食安問題，以及誤食有害食物後知道如何自行調處身體照顧行動等，都需要去了解現行的國、高中教科書，能否傳授足夠相關生活知識給予國民，以促進國家整體的良性發展。另外一例，則是協助犯罪偵察（Tseng, Ho, Yang, & Chen, 2012）。關聯圖的這類運用已超乎資訊組織與主題分析固有的應用範圍。

㈢產業知識活化應用

如前所述，資訊組織與主題分析，對圖書館的實務相當重要。但更多單位，如私人機構，也有類似的需求，只是傳統上沒有圖書館或知識管理單位的編制。以最近接觸產業界的需求為例，企業在設計、製造、生產、行銷、客服的過程中，累積了許多知識資產，如各項過程中的文件、規範、日誌、答客問、分析報告、問題討論與解決記錄等，包含了有形、無形、整理過與未整理過的各項知識資產。為了讓上述的資產活化、便利後續的利用與加值、幫助新人快速進入狀況、降低人員異動的訓練成本，遂有運用資訊組織與主題分析的技術，來活化其知識資產的需求（曾元顯，2014）。

在幫助企業滿足其需求的過程中了解到，從實務面來看，企業的使用者只需關心「用」的部分，「採、編、典、藏」則為機構內部的圖書館、資訊中心或知識管理單位要處理與建立相關機制的部分。

圖6所示，是以各項技術處理資料後，呈現的使「用」範例[4]。圖中被分析、整理的資料，為聯合報系2002到2009共八年約200萬篇的新聞報導。透過如：關鍵詞擷取、關聯詞分析、文件自動摘要等自動化處理後，使用者如要調閱相關文件，僅需輸入適當的關鍵詞，系統便能列出相關的文件與其摘要、對應的相關詞彙，以及該關鍵詞的時間篇數序列，進而達到檢索、摘要、關聯、分類統計的目的。

例如，圖6的輸入詞彙為「凌華」，從檢索結果右上角的詞彙關聯圖以及左邊的文件摘要，可大略看出凌華為一家工業電腦廠商，且跟：新普、瀚宇博、威達電等公司有產業上的關係（事實上他們都是工業電腦相關廠商）。對此領域不熟的用戶，僅需輸入一個簡單詞彙，便能了解資料庫中隱含的各種資訊。此外，圖中的時間序列，若運用在客戶問題的反應上，可得知哪些產品的失效（或成熟）的週期；若運用在常見問題上，則可讓公司的回應人員更快找到類似的問題與相關的答案，提昇問題解決的時效，建立良好的客戶關係。透過此類系統，讓初學者或入門者，即時且便利地存取專家等級的知識，而不怕職務異動造成的知識斷層，是資訊組織與主題分析最重要意義，以及最直接有效的用處。

然而，當文件資料越來越多時，數位資料需要進一步分門別類，以便快速縮小搜尋範圍、統計各類別（如產品的某種瑕疵）的發生次數、並降低同義異名或同名異義的問題。也就是說，完整而理想的資訊組織與主題分析境界，應能完成下列各項任務，對產業界才有最大的助益：

- 詞彙控制：聚合同義異名、區別同名異義詞彙；

4為了不透露過多的企業資料，圖6的內容僅為示意圖，但已是該企業第一階段使用的系統介面。

圖6 自動化資訊組織與主題分析後的資料運用範例之一（彩圖請見電子檔）

- 詞彙關聯：蒐集並建立上、下位詞，或廣義、狹義、相關詞；
- 內容摘要：擷取文件的重要內容（動態查詢導向摘要或靜態重點摘要）；
- 主題歸類：將主題相關的文件聚集成類；
- 文件分類：建構分類架構，之後將文件按類歸檔；
- 資訊檢索：知識架構的分類瀏覽、關鍵詞的資源查詢、資訊片段或全文的取用、類別或時間的交叉分析、預警性的提示，或意外的知識發現。

上述問題都是學術研究一直在解決的課題，特別是詞彙控制、詞彙關聯與文件分類的部分，以往只能依賴人工處理，需要投入大量人力。所幸我們過去發展的技術，已觸及這些問題的解決方案。雖然無法完全解決，但以自動化或半自動化方式，已可降低人工處理這些問題所需的極大成本。

圖7的範例，是查詢「intel」的結果，從詞彙關聯圖中，可看出資料庫記載的同義異名詞為：「英特爾」，也可看出其主要的產品（具有隸屬關係的詞彙），有：處理器、晶片組等。

前端的系統使「用」方式與上述範例，被企業接受後，我們才開始著手系統後端「採、編、典、藏」的處理工作。其中資料採訪蒐集的範圍，由企業指定；而文件館藏、歸檔、儲存的部分，也由其資訊部門自行處理。我們協助的，是其中的「編」（分析、組織、編目）與「典」（資料標準化與作業法則化）部分。由於細節無法全盤揭露（合約所限），下文僅就部分資料，進行範例式的說明。

圖7　自動化資訊組織與主題分析後的資料運用範例之二（彩圖請見電子檔）

圖8與圖9範例[5]，是以企業內部的客訴資料，進行的自動化文件歸類的結果。從中文範例，可看出：「無法開機」、「不開機」、「重複開機」等，在企業的資料庫裡，幾乎是同義異名詞，且根據客訴的日期，此問題有可能在將近半年時間內會再出現。英文範例也有類似情形，有關Cold boot（冷啟動）的fail（失效）案件，都被歸類在一起。這對於文件主題的分門別類、依照文獻保證原則按類給目（資料中有該類文件才訂定該分類名稱）的知識分類與管理，具有極大的輔助作用。

從歸類結果，可建議該有的類別名稱，然後依照企業的領域範圍，組織成其需要的分類架構，據以將文件分門別類，方便後續依類找文的應用，或結合關鍵詞查詢，了解該關鍵詞在每個類別的分布狀況，做查詢篩選或資產分析的依據。

- 68 Docs. : 0.1269（開 機:94.9213）
  - 9235 : 2 Docs. : 0.77（重覆:4.90, 重覆開機:2.12, 會重覆 **resent:**2.12, **resent:**2.12, 開機時無法到 window 畫面:2.12）
    - 20xy/4/16: 無法開機(會重覆 Resent)
    - 20xy/5/27: 開機時無法到 Windows 畫面，重覆開機
  - 65108 : 66 Docs. : 0.26（開機:89.54）
    - 20xy/2/21: 無法開機
    - 20xy/2/27: 開機有救護車聲
    - 20xz/12/7: 無法開機。
    - 20xy/4/12: 開機不正常，時而開機時而不開機
    - 20xy/6/14: 無法開機
    - 20xy/5/6: 使用兩個禮拜後，突然冒煙，之後無法開機
    - …

圖8　主題歸類結果範例（中文）

5為免透漏過多細節，實際的資料都經模糊化處理，如：西元年代皆換成20xy等四個數字與字母形式，而產品編號則以XYZ等代號取代，或經過噴霧處理。

- 5 Docs. : 0.45 (**cold boot**:16.95, **degree**:3.97, **test**:1.93, **fail**:1.80)
  - 20xx/10/3:When cold boot test at -5 degree, It is failed. Test count:1500 Passed count:480 === Cold boot @ -5 degree ...
  - 20xw/6/2:Cold boot(XY) / -25 degree test fail & XY boot machine count times not mach. AT boot machine count ...
  - 20xw7/7:Cold boot(XY mode) / -40 degree test Fail. Systems hang on "00" in the -40 degree.
  - 20xw/2/3:Cold boot(XYZ) at 75 degree failed. System hangs on debug code "00" or "F0".
  - 20xx/7/4:Cold boot(XYZ) -45 degree fail 08/02 test two phase temp: -40 degree 1000 cycle is pass.

圖9 主題歸類結果範例(英文)

依上述建構的分類架構進行文件分類，不論自動式，或人工作業(人工確認)，經過一段時間後，多少會出現分類不一致或類別混淆的情形，這時再引入分類一致性自動偵測與處理(曾元顯、王峻禧，2007)，使分類架構持續地符合企業的需求。

上述各種處理步驟與技術(如：關鍵詞擷取、文件歸類、分類架構建立)，都沒有標準答案，因此難以事先預估成效。僅能以文獻保證原則，擷取出適當的關鍵詞、關聯詞，與分類描述詞，再透過人工的分析、組織、整理與加工，以獲得品質較佳、成效較好的結果，來活化既有的知識資產，增進使「用」的效果。

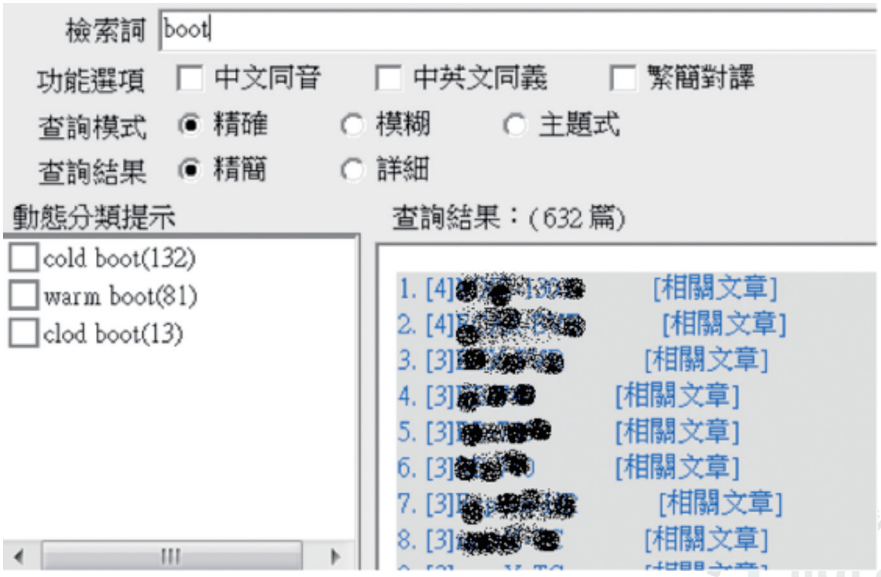

圖10 以半自動化方式抓取出雜訊資料範例(查詢結果經過噴霧處理)

此外，這些自動化技術，也需要高品質的輸入才能獲得高品質的輸出。若最前端的資料蒐集階段沒有管控好資料的品質，則後續的自動處理效果將大打折扣。例如，圖10顯示企業實際的資料中，的確有雜訊的問題。例如cold boot

有132篇文件，但輸入錯誤的clod boot也有13篇，在輸入的資料中，約有高達10%的雜訊或錯誤。

關於此項協助企業活化其知識資產的案例，難以再多做說明。但從上面的實例中（圖6到圖10），我們過去十多年來發展的技術，到現在仍可運用於實務的情境，顯示自動化資訊組織與主題分析，有其歷久彌新的重要角色。

四、結　語

圖書館學長年研究的資訊組織與主題分析課題，在理論與實務上產出的知識成果，珍貴而實用。資訊技術的進步，讓數位文件大量暴增，不僅引發更多需求，也帶來更多的應用機會。我們從過去接觸的各類文件處理案例，以圖書館學的「採、編、典、藏、用」角度，進行各種自動化技術的研發與應用，發現確能帶來具體的效益，不僅可盤點既有的知識資產，讓其活化，更可幫助建立可長可久的文件處理流程與運用模式，除了解決既有的沉痾，甚至可進一步創造新的價值。

這二十年來資訊技術的進步幅度驚人，相關的文獻成長快速，各種理論與技術百家爭鳴，即便我們在此領域進行理論研究、技術發展、觀摩考察與實務應用，具備了十多年的經驗，也僅能探討到一小部分，並列舉一、二實例，供讀者具體領略其相關的研究與發展。更完整的資訊組織與主題分析自動化技術與應用，則有待更多的介紹與探討。

誌　謝

本文感謝教育部「邁向頂尖大學計畫」、科技部「跨國頂尖研究中心計畫」NSC 103-2911-I-003-301與MOST 103-2221-E-003-013-MY3，以及國立臺灣師範大學「華語文與科技研究中心」之支持。

參考文獻

朱讚美（2000）。*Z39.50*協定伺服端之研究與實作（未出版之碩士論文）。國立中正大學資訊工程研究所，嘉義縣。

江玉婷、陳光華（1999）。TREC現況及其對資訊檢索研究之影響。圖書與資訊學刊，*29*，36-59。

曾元顯（2002）。回溯性資料數位化服務之規劃與建置。資訊傳播與圖書館學，*9*(2)，27-39。

曾元顯（2014）。知識探勘於知識資產活化的運用。台北：國立臺灣師範大學。

曾元顯、王峻禧（2007）。分類不一致之自動偵測：以農資中心資料為例。圖書館學與資訊科學，*33*(2)，20-32。

曾元顯、林瑜一(1998)。模糊搜尋、相關詞提示與相關詞回饋在OPAC系統中的成效評估。中國圖書館學會會報，*61*，103-125。

蔡孟竹、曾元顯(2003)。中文OCR文件檢索測試集之製作與應用。教育資料與圖書館學，*40*(3)，325-344。

謝欣君、張玉山、袁賢銘(1998)。異質性搜尋引擎代理人之設計與實作。1998台灣區網際網路研討會發表之論文，花蓮縣。

Bai, B.-R., Chen, C.-L., Chien, L.-F., & Lee, L.-S. (2002). Intelligent retrieval of dynamic networked information from mobile terminals using spoken natural language queries. *IEEE Transactions on Consumer Electronics*, *44*(1), 62-72.

Chan, L. M. (2007). *Cataloging and classification: An introduction* (3rd ed.). Lanham, MD: Scarecrow Press.

Chang, C.-H., & Lui, S.-C. (2001). IEPAD: Information extraction based on pattern discovery. In *Proceedings of the 10th International Conference on World Wide Web* (pp. 681-688). New York, NY: ACM.

Chen, H., Yim, T., Fye, D., & Schatz, B. (1995). Automatic thesaurus generation for an electronic community system. *Journal of the American Society for Information Science and Technology*, *46*(3), 175-193.

Chien, L.-F. (1995a). Q(Csmart)-A high-performance Chinese document retrieval system. In *Proceedings of the 1995 International Conference on Computer Processing of Oriental Languages* (pp. 176-183). Bethesda, MD: Chinese Language Computer Society.

Chien, L.-F. (1995b). Fast and quasi-natural language search for gigabytes of Chinese texts. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 112-120). New York, NY: ACM. doi:10.1145/215206.215345

Chien, L.-F. (1997). PAT-tree-based keyword extraction for Chinese information retrieval. *ACM SIGIR Forum*, *31*(SI), 50-58.

Chien, L.-F., & Pu, H.-T. (1996). Important issues on Chinese information retrieval. *Computational Linguistics and Chinese Language Processing*, *1*(1), 205-221.

Chowdhury, G. G. (2010). *Introduction to modern information retrieval* (3rd ed.). New York, NY: Neal-Schuman.

Fader, A., Soderland, S., & Etzioni, O. (2011). Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 1535-1545). Stroudsburg, PA: Association for Computational Linguistics.

Ghemawat, S., Gobioff, H., & Leung, S.-T. (2003). The Google file system. *ACM SIGOPS Operating Systems Review*, *37*(5), 29-43. doi:10.1145/1165389.945450

Harman, D. (1992). The DARPA TIPSTER project. *SIGIR Forum*, *26*(2), 26-28. doi:10.1145/146565.146567

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th Conference on Computational linguistics-Volume 2* (pp. 539-545). Stroudsburg, PA: Association for Computational Linguistics. doi:10.3115/992133.992154

Hsieh, Y.-M., Bai, M.-H., Chang, J. S., & Chen, K.-J. (2012). Improving PCFG Chinese Parsing with Context-Dependent Probability Re-estimation. In *Proceedings of the Second CIPS-*

SIGHAN Joint Conference on Chinese Language Processing (pp. 216-221). Tianjin, China: Association for Computational Linguistics.

Lin, W.-C., Chang, Y.-C., & Chen, H.-H. (2005). From text to image: Generating visual query for image retrieval. In C. Peters et al. (Eds.), *Multilingual information access for text, speech and images* (pp. 664-675). Berlin, German: Springer. doi:10.1007/11519645_65

Ogden, T. H. (1977). *Subjects of analysis* (Reissue ed.). New York, NY: Jason Aronson.

Olson, H. A., & Boll, J. J. (2001). *Subject analysis in online catalogs* (2nd ed.). Englewood, CO: Libraries.

Salton, G. (1989). *Automatic text processing: The transformation, analysis, and retrieval of information by computer*. Reading, MA: Addison-Wesley.

Sanderson, M., & Croft, B. (1999). Deriving concept hierarchies from text. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 206-213). New York, NY: ACM. doi:10.1145/312624.312679

Sasaki, Y., Chen, H.-H., Chen, K.-h., & Lin, C.-J. (2005). Overview of the NTCIR-5 cross-lingual question answering task (CLQA1). In *Proceedings of NTCIR-5 Workshop Meeting*. Tokyo, Japan: National Institute of Informatics - Research Organization of Information and Systems.

Sundheim, B. M. (1991). Overview of the third message understanding evaluation and conference. In *Proceedings of the 3rd Conference on Message Understanding* (pp. 3-16). Stroudsburg, PA: Association for Computational Linguistics. doi:10.3115/1071958.1071960

Taylor, A. G., & Joudrey, D. N. (2008). *The organization of information* (3rd ed.) Westport, CO: Libraries.

Tseng, Y.-H. (1998). An approach to retrieval of OCR degraded text. *National Taiwan University Journal of Library Science*, *13*, 153-168.

Tseng, Y.-H. (1999). Content-based retrieval for music collections. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 176-182). New York, NY: ACM. doi:10.1145/312624.312675

Tseng, Y.-H. (2001). Automatic cataloguing and searching for retrospective data by use of OCR text. *Journal of the American Society for Information Science and Technology*, *52*(5), 378-390. doi:10.1002/1532-2890(2001)9999:9999<::AID-ASI1080>3.0.CO;2-A

Tseng, Y.-H. (2002). Automatic thesaurus generation for Chinese documents. *Journal of the American Society for Information Science and Technology*, *53*(13), 1130-1138. doi:10.1002/asi.10146

Tseng, Y.-H., Chang, C.-Y., Rundgren Chang, S.-N., & Rundgren, C.-J. (2010). Mining concept maps from news stories for measuring civic scientific literacy in media. *Computers & Education*, *55*(1), 165-177. doi:10.1016/j.compedu.2010.01.002

Tseng, Y.-H., Ho, Z.-P., Yang, K.-S., & Chen, C.-C. (2012). *Mining term networks from text collections for crime investigation*. *Expert Systems with Applications*, *39*(11), 10082-10090. doi:10.1016/j.eswa.2012.02.052

Tseng, Y.-H., Lee, L.-H., Lin, S.-Y., Liao, B.-S., Liu, M.-J., Chen, H.-H., ... Fader, A. (2014). Chinese open relation extraction for knowledge acquisition. In *Proceedings of the 14th*

Conference of the European Chapter of the Association for Computational Linguistics, Volume 2: Short Papers (pp. 12-16). Gothenburg, Sweden: Association for Computational Linguistics.

Tseng, Y.-H., Lin, C.-J., & Lin, Y.-I. (2007). Text mining techniques for patent analysis. *Information Processing and Management: An International Journal*, *43*(5), 1216-1247. doi:10.1016/j.ipm.2006.11.011

Van Rijsbergen, C. J. (1979). *Information retrieval* (2nd ed.). Boston, MA: Butterworth-Heinemann.

Witten, I. H., Moffat, A., & Bell, T. C. (1999). *Managing gigabytes: Compressing and indexing documents and images*. San Francisco, CA: Morgan Kaufmann.

Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C., & Nevill-Manning, C. G. (1999). KEA: Practical automatic keyphrase extraction. In *Proceedings of the Fourth ACM Conference on Digital Libraries* (pp. 254-255). New York, NY: ACM. doi:10.1145/313238.313437

曾元顯 ORCID 0000-0001-8904-7902

Journal of Educational Media & Library Sciences 51: Special Issue (2014) : 3-26
DOI:10.6120/JoEMLS.2014.51S/0652.RV.AM

Review Article

Research and Development on Automatic Information Organization and Subject Analysis in Recent Decades

Yuen-Hsien Tseng

Abstract

Information organization and subject analysis (IOSA) is an important issue in the field of library and information science (LIS). As the fast advance in information technology, more and more digital documents are emerging in a pace such that automated IOSA become inevitable. This article firstly introduces the development of related automatic techniques in recent decades and promotes a tranditional viewpoint based on the workflow of: (1) data collection and aggregation, (2) cataloguing, (3) regulation, (4) archving, and (5) usage, to regulate the whole process when applying automated techniques to any IOSA task. Some application examples are then described to let the readers have a feel of the feasibility of these techniques; specifically the applications of keyword extraction, association analysis, document clustering, and topic categorization are mentioned. We conclude that the related techniques and applications are still developing in a quick pace such that only a few percentages of them can be mentioned. This article is intended to promote the mutual cooperation among the LIS and other fields.

Keywords: *Keyword extraction; Association analysis; Document clustering; Topic categorization; Information retrieval*

ROMANIZED & TRANSLATED REFERENCE FOR ORIGINAL TEXT

朱讚美[Chu, Chan-Mei]（2000）。*Z39.50*協定伺服端之研究與實作（未出版之碩士論文）[*Implementing the server of Z39.50 protocol* (Unpublished master's thesis)]。國立中正大學資訊工程研究所，嘉義縣[Institute of Information Engineering, National Chung Cheng University, Chiayi, Taiwan]。

江玉婷、陳光華[Chiang, Yu-Ting, & Chen, Kuang-Hua]（1999）。TREC現況及其對資訊檢索研究之影響[The TREC and its impact on IR researches]。圖書與資訊學刊，*29*，36-59 [*Bulletin of Library and Information Science*, *29*, 36-59]。

曾元顯[Tseng, Yuen-Hsien]（2002）。回溯性資料數位化服務之規劃與建置[Networked information services for retrospective data]。資訊傳播與圖書館學，*9*(2)，27-39 [*Journal of Information, Communication, and Library Science*, *9*(2), 27-39]。

Research Fellow, Information Technology Center, National Taiwan Normal University, Taipei, Taiwan
E-mail: samtseng@ntnu.edu.tw

曾元顯[Tseng, Yuen-Hsien]（2014）。知識探勘於知識資產活化的運用[*Zhishi tankan yu zhishi zichan huohua de yunyong*]。台北：國立臺灣師範大學[Taipei: National Taiwan Normal University]。

曾元顯、王峻禧[Tseng, Yuen-Hsien, & Wang, Chun-Shi]（2007）。分類不一致之自動偵測：以農資中心資料為例[Automatic inconsistency detection for the ASIC categorization collection]。圖書館學與資訊科學，*33*(2)，20-32 [*Journal of Library and Information Science*, *33*(2), 20-32]。

曾元顯、林瑜一[Tseng, Yuen-Hsien, & Lin, Yu-Yi]（1998）。模糊搜尋、相關詞提示與相關詞回饋在OPAC系統中的成效評估[Evaluation of fuzzy search, term suggestion, and term relevance feedback in an OPAC system]。中國圖書館學會會報，*61*，103-125 [*Bulletin of the Library Association of China*, *61*, 103-125]。

蔡孟竹、曾元顯[Tsai, Mung-Chu, & Tseng, Yuen-Hsien]（2003）。中文OCR文件檢索測試集之製作與應用[Construction and application of a Chinese OCR test collection for information retrieval]。教育資料與圖書館學，*40*(3)，325-344 [*Journal of Educational Media & Library Sciences*, *40*(3), 325-344]。

謝欣君、張玉山、袁賢銘（1998）。異質性搜尋引擎代理人之設計與實作[*Yizhixing souxun yinqing dailiren zhi sheji yu shizuo*]。1998台灣區網際網路研討會發表之論文，花蓮縣[Paper presented at the Taiwan Area Network Conference, Hualien, Taiwan]。

Bai, B.-R., Chen, C.-L., Chien, L.-F., & Lee, L.-S. (2002). Intelligent retrieval of dynamic networked information from mobile terminals using spoken natural language queries. *IEEE Transactions on Consumer Electronics*, *44*(1), 62-72.

Chan, L. M. (2007). *Cataloging and classification: An introduction* (3rd ed.). Lanham, MD: Scarecrow Press.

Chang, C.-H., & Lui, S.-C. (2001). IEPAD: Information extraction based on pattern discovery. In *Proceedings of the 10th International Conference on World Wide Web* (pp. 681-688). New York, NY: ACM.

Chen, H., Yim, T., Fye, D., & Schatz, B. (1995). Automatic thesaurus generation for an electronic community system. *Journal of the American Society for Information Science and Technology*, *46*(3), 175-193.

Chien, L.-F. (1995a). Q(Csmart)-A high-performance Chinese document retrieval system. In *Proceedings of the 1995 International Conference on Computer Processing of Oriental Languages* (pp. 176-183). Bethesda, MD: Chinese Language Computer Society.

Chien, L.-F. (1995b). Fast and quasi-natural language search for gigabytes of Chinese texts. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 112-120). New York, NY: ACM. doi:10.1145/215206.215345

Chien, L.-F. (1997). PAT-tree-based keyword extraction for Chinese information retrieval. *ACM SIGIR Forum*, *31*(SI), 50-58.

Chien, L.-F., & Pu, H.-T. (1996). Important issues on Chinese information retrieval. *Computational Linguistics and Chinese Language Processing*, *1*(1), 205-221.

Chowdhury, G. G. (2010). *Introduction to modern information retrieval* (3rd ed.). New York, NY: Neal-Schuman.

Fader, A., Soderland, S., & Etzioni, O. (2011). Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 1535-1545). Stroudsburg, PA: Association for Computational Linguistics.

Ghemawat, S., Gobioff, H., & Leung, S.-T. (2003). The Google file system. *ACM SIGOPS Operating Systems Review*, *37*(5), 29-43. doi:10.1145/1165389.945450

Harman, D. (1992). The DARPA TIPSTER project. *SIGIR Forum*, *26*(2), 26-28. doi:10.1145/146565.146567

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th Conference on Computational linguistics-Volume 2* (pp. 539-545). Stroudsburg, PA: Association for Computational Linguistics. doi:10.3115/992133.992154

Hsieh, Y.-M., Bai, M.-H., Chang, J. S., & Chen, K.-J. (2012). Improving PCFG Chinese Parsing with Context-Dependent Probability Re-estimation. In *Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing* (pp. 216-221). Tianjin, China: Association for Computational Linguistics.

Lin, W.-C., Chang, Y.-C., & Chen, H.-H. (2005). From text to image: Generating visual query for image retrieval. In C. Peters et al. (Eds.), *Multilingual information access for text, speech and images* (pp. 664-675). Berlin, German: Springer. doi:10.1007/11519645_65

Ogden, T. H. (1977). *Subjects of analysis* (Reissue ed.). New York, NY: Jason Aronson.

Olson, H. A., & Boll, J. J. (2001). *Subject analysis in online catalogs* (2nd ed.). Englewood, CO: Libraries.

Salton, G. (1989). *Automatic text processing: The transformation, analysis, and retrieval of information by computer*. Reading, MA: Addison-Wesley.

Sanderson, M., & Croft, B. (1999). Deriving concept hierarchies from text. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 206-213). New York, NY: ACM. doi:10.1145/312624.312679

Sasaki, Y., Chen, H.-H., Chen, K.-h., & Lin, C.-J. (2005). Overview of the NTCIR-5 cross-lingual question answering task (CLQA1). In *Proceedings of NTCIR-5 Workshop Meeting*. Tokyo, Japan: National Institute of Informatics - Research Organization of Information and Systems.

Sundheim, B. M. (1991). Overview of the third message understanding evaluation and conference. In *Proceedings of the 3rd Conference on Message Understanding* (pp. 3-16). Stroudsburg, PA: Association for Computational Linguistics. doi:10.3115/1071958.1071960

Taylor, A. G., & Joudrey, D. N. (2008). *The organization of information* (3rd ed.) Westport, CO: Libraries.

Tseng, Y.-H. (1998). An approach to retrieval of OCR degraded text. *National Taiwan University Journal of Library Science*, *13*, 153-168.

Tseng, Y.-H. (1999). Content-based retrieval for music collections. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 176-182). New York, NY: ACM. doi:10.1145/312624.312675

Tseng, Y.-H. (2001). Automatic cataloguing and searching for retrospective data by use of OCR text. *Journal of the American Society for Information Science and Technology*, *52*(5),

378-390. doi:10.1002/1532-2890(2001)9999:9999<::AID-ASI1080>3.0.CO;2-A

Tseng, Y.-H. (2002). Automatic thesaurus generation for Chinese documents. *Journal of the American Society for Information Science and Technology*, *53*(13), 1130-1138. doi:10.1002/asi.10146

Tseng, Y.-H., Chang, C.-Y., Rundgren Chang, S.-N., & Rundgren, C.-J. (2010). Mining concept maps from news stories for measuring civic scientific literacy in media. *Computers & Education*, *55*(1), 165-177. doi:10.1016/j.compedu.2010.01.002

Tseng, Y.-H., Ho, Z.-P., Yang, K.-S., & Chen, C.-C. (2012). *Mining term networks from text collections for crime investigation*. *Expert Systems with Applications*, *39*(11), 10082-10090. doi:10.1016/j.eswa.2012.02.052

Tseng, Y.-H., Lee, L.-H., Lin, S.-Y., Liao, B.-S., Liu, M.-J., Chen, H.-H., ... Fader, A. (2014). Chinese open relation extraction for knowledge acquisition. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, Volume 2: Short Papers* (pp. 12-16). Gothenburg, Sweden: Association for Computational Linguistics.

Tseng, Y.-H., Lin, C.-J., & Lin, Y.-I. (2007). Text mining techniques for patent analysis. *Information Processing and Management: An International Journal*, *43*(5), 1216-1247. doi:10.1016/j.ipm.2006.11.011

Van Rijsbergen, C. J. (1979). *Information retrieval* (2nd ed.). Boston, MA: Butterworth-Heinemann.

Witten, I. H., Moffat, A., & Bell, T. C. (1999). *Managing gigabytes: Compressing and indexing documents and images*. San Francisco, CA: Morgan Kaufmann.

Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C., & Nevill-Manning, C. G. (1999). KEA: Practical automatic keyphrase extraction. In *Proceedings of the Fourth ACM Conference on Digital Libraries* (pp. 254-255). New York, NY: ACM. doi:10.1145/313238.313437