

數位文件之資訊組織與主題分析自動化之技術與應用

Automatic Information Organization and Subject Analysis for Digital Documents

曾元顯 Yuen-Hsien Tseng

輔仁大學圖書資訊學系

Department of Library & Information Science, Fu Jen Catholic University

「台北市立圖書館館訊」, 2002 年 12 月, 第 20 卷, 第 2 期, 頁 23-35

【摘要】

資訊組織與主題分析是圖書館學的理论與實務中最主要的課題之一,其目的在探討如何分析並組織文件,以提供使用者便捷、有效的資訊服務。資訊組織與主題分析需要高度的知識加工處理,傳統上有賴於訓練有素的圖書館人員進行此項工作。由於資訊科技的持續進步,使得資訊組織與主題分析探討的很多課題有自動化處理的作法。本文便是在介紹筆者發展的一些自動化的方法,並探討其中的觀念、技術、應用、與其未來的影響。文中特別介紹自動化索引、索引典自動建構以及自動分類,並展示其應用範例,期使讀者能具備自動化作法的概念,以便有效運用現有的資訊科技,更有效率的進行資訊組織與主題分析的工作。

【Abstract】

Information organization and subject analysis (IOSA) is the main concern in library science and library services. The goal of IOSA is about how to analyze and organize documents to provide effective and efficient information access. IOSA requires human knowledge involvement. Traditionally, only well-trained librarians are qualified for this task. Due to the advance of information technologies, many tasks about IOSA now have the automatic solutions. This article introduces these automatic approaches developed by the author. The concepts, techniques, applications and future impacts are explored. Specifically, this article describes the automatic ways of indexing, thesaurus construction, and text categorization. Application examples are demonstrated to allow readers better understand these approaches so that future IOSA could be achieved in a more efficient way, through the integration of human efforts and automatic methods.

關鍵詞：索引、檢索、索引典、分類、自動化

Keywords: Indexing, Retrieval, Thesaurus, Classification, Automatic.

壹、前言

資訊組織與主題分析是圖書館學科與實務中最主要的課題之一（註 1、2），其目的在探討如何分析並組織文件（Documents），以提供使用者便捷、有效的資訊接取（Information Access）服務。此課題涵蓋的範圍相當龐大，在圖書館學的學程中，有大約三分之一的課程與其直接相關。近年來常被探討的知識管理，在非結構化資料的管理方面，也與資訊組織與主題分析探討著相同或類似的問題。而數位化圖書館等計畫，更投入主要的資源，以進行數位文件資訊組織與主題分析的工作。具體而言，圍繞在這個課題的相關題目有索引、摘要、檢索、分類、編目等項目，而與這些題目相關的概念有權威檔、權威控制、索引典、主題表、分類表、編目規則以及各種作業規範與資料格式的標準等等。（註 3）

資訊組織與主題分析需要高度的知識加工處理，傳統上有賴於訓練有素的圖書館人員進行此項工作。由於資訊科技的持續進步，使得資訊組織與主題分析探討的很多課題有自動化處理的作法。本文便是在介紹這些自動化的方法，並探討其中的觀念、技術、應用、與評估其未來的影響，期使讀者能具備自動化作法的概念，以便有效運用現有的資訊科技，更有效率的進行資訊組織與主題分析的工作。

資訊組織與主題分析的自動化工作，包括自動索引與檢索、自動分類、自動歸類、自動摘要、自動過濾、資訊自動擷取、索引典自動建構、事件自動偵測與追蹤、甚至自動化參考服務（詢答系統）等。然而受限於文章長度的限制，本文只就筆者發展較完整的技術做探討與介紹，這包括自動索引、索引典自動建構以及自動分類。文章後面將介紹索引、索引典與分類在資訊組織與主題分析中扮演的角色，以及自動化技術大致的作法與成效，然後分析自動化作法的影響，最後舉範例與應用作進一步說明。希望將資訊組織與主題分析自動化的發展現況，分享予讀者，讓這些技術得以被適當的運用。

貳、自動化資訊組織與主題分析

一、自動化索引

索引 (Index) 是用來快速找尋文件的輔助資料檔，其內容記錄了哪些詞彙出現在什麼文件的訊息。(註 4) 索引的概念，普遍存在各處，舉凡字典、百科全書、甚至編輯較為用心的書籍，都有索引的例子。如書後索引，常把書中重要的詞彙及其出現的章節或頁數整理排序後顯示出來，以方便讀者利用。相對於文件記錄了詞彙，索引是以詞彙反方向記錄文件的結構，以便讓使用者查找資料時，依詞彙快速找出想要的文件。因此索引詞彙是檢索比對的依據，是影響文件檢索成效好壞的主要因素之一。

索引的建構，早期都以人工方式分析每一篇文章，再選擇一組適當的索引詞來代表該文件，最後再將之反向編制成索引。選擇代表某一文件的索引詞是件頗為繁瑣的工作，必須要猜測、預想使用者會以什麼樣的詞彙來搜尋該篇文章。例如：一本書名叫做「資訊檢索系統」的書，必須選擇「資訊」、「檢索」、「系統」、「資訊檢索」、「檢索系統」、「資訊檢索系統」等至少這六個詞作為該書名的索引詞(註 5)，才能讓使用者方便、合理的以這些詞彙找到該本書。

然而這還只是簡單的斷詞問題，有時為了替使用者進一步著想，還必須處理同義異名詞 (Synonym) 的問題。例如：「查詢系統」、「搜尋引擎」可能都跟「檢索系統」同義，因而必須建構一個同義詞表，讓其中一種詞彙能夠查詢到其他型態的同義詞彙。圖書館學中，把同義異名詞以統一、標準的形式紀錄並管理的方式稱為權威控制 (Authority control)，而把記錄同義異名詞的檔案稱為權威 (控制) 檔 (註 6) (Authority file) (註 7、8)。例如：文件中遇到「孫文」、「孫中山」、「國父」等用詞時，在內部的索引檔中一律都以「孫文」代之，並在檢索時，導引使用者將其他同義詞彙轉換成「孫文」後再做查詢，以避免資料漏檢或字彙不匹配的問題。由於人名、地名常有此種情形發生，因此常見到圖書館中有人名權威檔、地名權威檔等檢索輔助檔案的設計 (註 9)。然而，有時這種同義異名詞難以一一羅列，如外國譯名：「克麗絲汀」與「克莉思汀」，又如複合詞：「數位圖書館」與「數位化圖書館」、「基因研究」與「基因的研究」等，這些跟文件作者的用詞習慣比較有關，而不一定是約定俗成的習慣用詞，因此難以預測、捕捉，進而收納到同義詞表中。

早期的資訊檢索系統，尤其是中文的檢索系統，需依賴人工斷詞或人工維護的詞庫來斷詞，以提供索引詞彙，使得使用單位如圖書館等機構在運用檢索系統時，還不時要為檢索系統提供人工斷詞的協助，造成使用單位的負擔。現今的系統，則大都不需使用單位協助斷詞，而是以連詞 (n-gram) 方式 (註 10)，或由系統提供事先建好的詞庫，再輔以定期更新或系統自動擷取新詞的方式，來自動建構索引詞。

至於同義異名詞的問題，有些系統提供事先蒐集的同義詞庫供檢索使用，有些則進一步提供自動化的方式解決部分權威控制的問題，以降低資料漏檢的情形。例如，提供「同音查詢」，以解決外國譯名歧異的問題。中文同音查詢是透過注音符號表，將查詢詞彙與索引詞彙都轉成注音，在注音符號上做比對，因此像前例的「克麗絲玳」與「克莉思汀」，便因注音符號相同，而可以互相查尋找到。至於英文，透過 soundex code 對文字做編碼，也有類似的同音查詢作法（註 11）。

另外，自動擷取文件中的重要詞彙做成索引後，檢索時以近似字串或模糊搜尋的方式比對索引詞庫（註 12），不僅可以調出相關文件，也可以調出字串相近的詞彙，將之列於查詢結果中，可以提示使用者相關的查詢用詞，達到權威控制的目的。如：「數位圖書館」與「數位化圖書館」、「基因研究」與「基因的研究」、「筆記本電腦」與「筆記型電腦」，甚至「中研院」與「中央研究院」、「李遠哲院長」與「李院長遠哲」等，只要模糊比對的門檻設定得當，這些字串近似的詞彙（或稱形似詞），都可以自動互相提示出來，而不必人工做權威控制或人工維護同義詞庫。但像「閣揆」與「行政院長」、「紅樓夢」與「石頭記」等字串完全不重疊的詞彙，它們之所以成為同義詞，完全是人類強加上去的知識，因此除非告訴電腦這種知識，否則自動化系統常常沒有線索可以將它們關聯在一起。另外，像「杜鵑」（可能是花名，也可能是鳥名）或「孫悟空」（可能是人物名，也可能是迷幻藥名）等同形異義詞，甚至像「統一」（可能來自「統一中國」或「統一企業」）等長詞中的短詞，也必須仰賴檢索者提供更多的查詢線索，以解決詞義模糊（ambiguity）的問題。光看到「杜鵑」，連人都無法回答是指杜鵑花還是指杜鵑鳥時，電腦當然也無法正確決定該回應什麼樣的文件給使用者。

目前自動索引在某些檢索需求上，如同音詞、形似詞、全文檢索方面可以做到比人工索引的效果還好。但如果仍有額外人力建構自動化系統無法處理的同義詞庫，如前述的「閣揆」與「行政院長」等，對某些檢索需求而言，仍然會有所幫助。

二、索引典自動建構

因字彙不匹配而產生檢索遺漏的問題不只會發生在各式各樣的同義異名詞上，還會發生在近似詞或主題詞上。例如：「資訊系統」就包含「資訊檢索系統」、「資訊出版系統」等次系統，當「資訊檢索系統」用前述六個詞彙建索引時，並無法讓「資訊系統」這個查詢詞比對到「資訊檢索系統」這樣的資料。事實上，以「資訊系統」為查詢詞，可否查到「資訊檢索系統」是個見仁見智的問題，有的使用者會希望可以，以便找到較完整的資料，有的使用者會希望不可以，以便找到較精確的資料。最好的方式，是由使用者自己決定。系統若能從查詢詞「資訊系統」，提示出「資訊檢索系統」，反之亦然，

則使用者就可自己決定要不要選用另一個詞做進一步的查詢。

跟權威檔類似，圖書館學中，也有「索引典」(Thesaurus)的概念，來解決主題近似的詞彙問題(註 13-16)。「索引典」裡列舉各種檢索詞彙以及詞彙之間彼此的關係，如「廣義詞」、「狹義詞」、「相關詞」等，可用於查詢詞的互相推薦，以擴大或縮小查詢範圍，或提示相關概念的不同查詢用語，使檢索從原本的字串比對層次，提升到以語意做比對的層次。

索引典的建構，必須在詞彙的層次上進行主題分析與資訊組織的工作。為了建構此種詞彙之間語意上的關係，往往需要人力高度的知識加工。人工製作的索引典有正確性高的優點，但缺點則是成本大、建構速度慢、維護不易、以及事先選用的詞彙可能與後續新進的文件無關。過去資訊檢索實驗的研究亦指出，一般目的(general-purposed)的索引典運用在特定領域的文件檢索上(註 17)，會出現無法提升檢索效能的情形。(註 18)只有當索引典的用詞與文獻的用詞密切相關時，索引典的運用才會顯現出成效。因此，根據文獻本身的主題，自動且即時產生索引典的方法，是值得研究的重要課題。

自動化的方法，大抵都倚賴相關的詞彙在文件中常常一起出現的線索，來建構索引典。(註 19、20)此種方式建構出來的索引典，可稱為「共同出現索引典」(Co-occurrence thesaurus)，或簡稱「共現索引典」。過去的作法，大都是針對英文文件發展而來，直接運用於中文文件時，會碰到中文關鍵詞擷取的問題。藉由過去相關作法的改進，筆者自行研發出一種效率極高的關聯詞分析方法，將原先需要超級電腦運算的計算量，降低成只需個人電腦即可負擔的計算量。其擷取出來的關聯詞，可運用於檢索系統當中，作為查詢提示、檢索摘要、或知識地圖等進階功能。我們就自動建構出來的共現詞彙，曾評估其關聯程度，以精確率與召回率的方式評估，其平均精確率達 0.5284；若以相關比例評估，有 69%的提示詞被判定與查詢詞相關。和過去的類似研究相較，其成效在相同的水平上，顯示此自動方法不僅快速、適用於中英文，而且自動建構出來的索引典，有目前所知最好的品質。(註 21、22)

如同前述，人工製作的索引典準確度高，但召回率低、成本大、建構速度慢。相對的，自動化建構索引典，成本低、速度快、召回率高、與館藏文件用詞一致，但準確率較低、功能性較少(如難以提供參見詞、反義詞、新舊詞對照等功能)。兩種方法恰可互補不足、相輔相成。自動建構的索引典，經由人工不定時的篩選、補充、調整等加工後，也許還不能百分之百正確無誤，卻可以及時反應文獻的內容。即便沒有人工的修飾，運用在查詢提示，可透露出埋藏在文件中各種知識的關聯，對便利使用者探索、發掘文件中記載的知識，提供了宛如專家在旁指導的檢索輔助。

三、自動分類

資訊接取 (Information Access) 的服務中，除了提供使用者以詞彙比對主動查找文件外，另一個重要的方式就是要能讓使用者瀏覽主題、發現文件。這是因為有時使用者根本不知道自己要找的特定文件是什麼，要看到後才能確認，因此就難以用任何詞彙來檢索文件。這種按類瀏覽、選擇文件的方式，也可視為是一種最簡單的檢索，因為使用者只要重複的瀏覽、點選主題目錄，即可找到資料。大部分的入口網站，像「奇摩」、「雅虎」都聘請大量的人員進行文件分類，以提供網站或網頁分類目錄的服務。以「雅虎」而言，其搜尋系統甚至委外建置，過去數年來從 Infoseek 換成 Altavista 再換成 Google，但是其分類目錄則由其本身持續不斷的維護，足見分類服務的重要性有時更超過檢索服務。

分類 (Classification) 是圖書館學科重要的課題。將圖書、文件依其內容、性質分門別類，是圖書館中重要的日常工作。為了有效管理實體文件，圖書館常常採用某種圖書分類系統，如杜威十進分類法，對文件進行內容分析、類別標示，然後再依類上架，以便於歸檔、調閱、瀏覽與分享。此種為了便於管理實體文件的分類方式，常對文件做單一分類 (註 23)。單一分類的缺點是分類者的認知與使用者的認知可能不一致，以致於出現使用者依類索文，卻找不到想要文件的情況。

圖書館學中另一種組織資訊的方式，是以標題表 (Subject Headings) 的形式出現。標題的意義即是標出文件的主題，因此可以用多種角度、多個標題詞彙來描述文件的主題，從而減少標題給定者與使用者之間的認知不一致問題。通常這些標題詞之間也按序排列，進而組織成一套標題表。另外，一份文件通常也會給定數個關鍵詞，以彰顯該文件的內容與用詞。就這點而言，關鍵詞與標題詞有些類似，但差異是，標題詞是受控制的詞彙 (Controlled Vocabulary)，以減少如同義異名詞帶來的差異，而且不一定是文件的用詞；但關鍵詞則是不受控制、沒有組織的，而且常是出現在文件內的詞彙，或是標題表裡還未收納的新生詞彙或領域用詞 (註 24)。

文件分類，需要瞭解文件的主題大意，才能給定類別，因此是相當高階的知識處理工作。要將文件分類自動化，必須先整理出分類時的規則，電腦才能據以執行。然而，有效的分類規則通常難以用人工分析歸納獲得。因此，機器在做自動分類之前，還必須加以訓練，使其自動學習出人工分類的經驗與知識。

現今自然語言理解的技術，還無法讓電腦瞭解任意的自由文句。因此機器在做文件分類時，常將文件分解成一個個語意較小的單位，通常為文件的關鍵詞彙，或稱「特徵詞彙」，再從這些詞彙與類別中找出對應的關係。有時分類的問題，簡單到只要文件中出現什麼特徵詞，就分到什麼類別去。但大部分的情況都沒那麼簡單。例如，「文化大

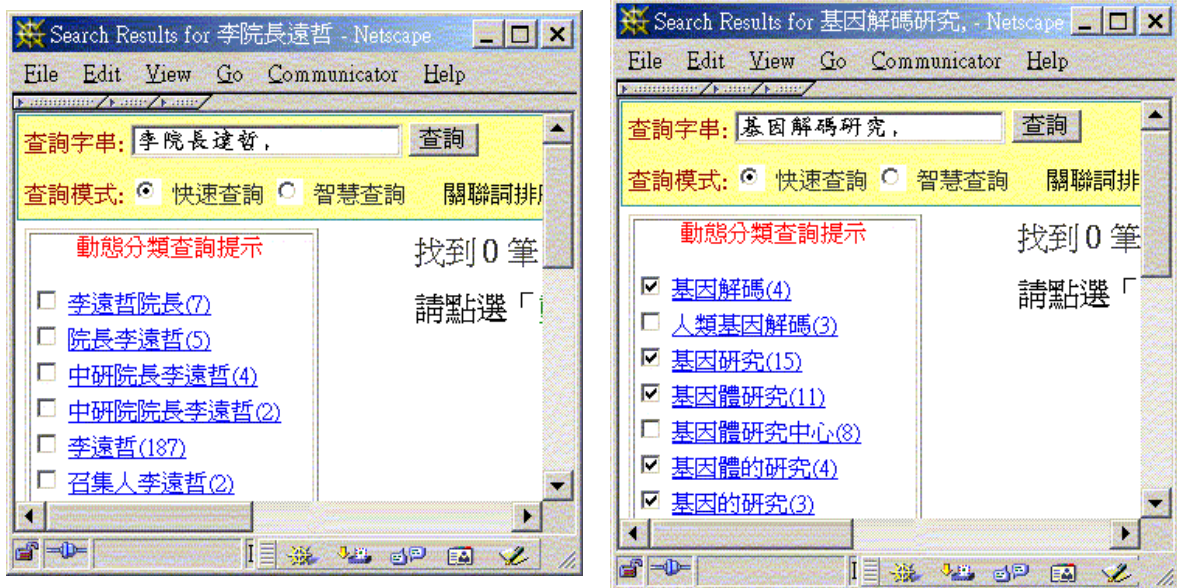
革命」這個類別，如何界定哪些詞彙一定是屬於這個類別，哪些不是？顯然某些詞彙對這個類別較重要（比較有鑑別力），其他的則較不重要（比較不具鑑別力），甚至某些詞彙對不同的類別有不同的鑑別力。如何決定這些詞彙在每個類別的權重，正是機器學習（Machine Learning）可以派上用場的地方（註 25、26）。

近年來，文件自動分類的研究相當豐富（註 27-30），有些成果已接近人工分類的水準。人工分類的準確度高，但缺點是速度慢，分類結果易受分類者的學養、認知影響，以及不同分類者的分類一致性低。機器分類則是一致性高、速度快、節省大量人力，缺點則是需要事先準備相當數量的訓練文件，而且分類錯誤時有可能會出現與文件主題毫不相關的錯誤類別。相較於人工分類，既使有錯，也只是主題認知上的差異，不至於出現毫不相關的錯誤類別。因此，自動分類經常配合人工校正分類結果，透過機器與人力的努力，一起達到高精確、高效率的分類成效。

參、範例與應用

以下我們就上面提到的自動化技術作展示。這些展示系統都是筆者自行研發的成果，目前也都有實際的應用，如輔仁大學的書目檢索系統。

圖一展示自動化索引應用於檢索的例子。兩個例子都顯示使用者下達一個非文件用詞的情形。比起系統只回應找不到文件而讓使用者不知所措的情形，此系統以模糊比對方式，將字串近似的關鍵詞提示出來，以協助正確選擇查詢詞彙來檢索文件。此種提示，不僅可以解決部分權威控制的問題，也可導引使用者有效的查找資料。通常，使用者一開始查詢時，常輸入少許的詞彙，就想找到想要的文件。但對系統而言，只獲得一點點輸入，是很難瞭解使用者真正的需求。透過提示，可以引導使用者告訴系統多一點訊息。



圖一：查詢詞彙與文件用詞不一致時，系統自動提示近似詞彙的情形，解決部分權威控制的問題。

圖二展示另外一個查詢提示的範例。當使用者輸入 PDA 後，系統如前例在最左欄提示字串近似的詞彙，以及其出現篇數。這些篇數可以讓我們瞭解某個詞在文件資料庫內分佈的情形。若視每一個詞彙為一個類別的話，那麼使用者可以清楚知道屬於「PDA 市場」類別的文件有 6 篇，屬於「PDA 面板」類別的文件有 2 篇等等。使用者可點選篇數較少的詞彙或較小的類別，以快速縮小範圍。此外，這些出現篇數也可以協助使用者在不必進一步做查詢的情況下，就可得知用此詞彙做查詢時會找回多少篇資料，讓使用者僅作一次查詢，就可得知好幾個查詢結果。其展現出來的效果宛如一個分類目錄，只不過此分類目錄會隨查詢詞彙的不同而變化，而且也會隨文件資料庫的不同而不同，因此稱為「動態分類目錄」。

此外，圖二中間那一欄，即是運用關聯詞庫建構出來的提示詞。對於查詢詞 PDA 的每一個關聯詞，其後的圓括號顯示兩個數據，第一個同樣是出現篇數，第二個是該詞與查詢詞的關聯強度。在這個例子中，PDA 的關聯詞是按關聯強度由大到小排序而成。對於想要探索 PDA 的使用者而言，這些提示具有相當程度的摘要作用。使用者如想進一步的瞭解為何某個詞與查詢詞有所關聯，可勾選該詞彙，再與查詢詞一起查詢以調出相關的文件，從其記載的描述瞭解詳情。例如：「林文彬」是 PDA 的關聯詞，對於不是很清楚為何兩者有所關聯的使用者，勾選該詞彙進一步查詢後，可獲得「林文彬：明年 PDA 有兩千萬台規模」的文件，從此文件片段中：「碧悠總經理林文彬表示，該公司明年預計交貨的 PDA 面板數量近九百萬片，若以此比重推估，明年全球 PDA 市場規模

將超過兩千萬台」，如此可輕易得知「PDA」、「碧悠」與「林文彬」的關係。這些知識可能只是該領域專家累積多年的常識，但對初入門者而言卻是重要的必備消息。有這樣的系統輔助，透過簡單的滑鼠點擊，這些領域知識，很容易就顯露出來，被非專業人員隨手取得。



圖二：關聯詞查詢範例。

上面的關聯詞是以一維空間的文字展示，我們也可以將關聯詞應用到二維空間的環境，使其具備另一番使用上的趣味。此外，如果我們能辨別詞彙的性質，在顯示這些詞彙時，還可以根據其性質主動提示出更多相關的訊息。如圖三所示，使用者在查詢 MP3 後，發現有很多詞彙跟 MP3 有關係，其中「中環」像似一家公司名稱，使用者點選後，可以找出 MP3 與中環相關的文章，瞭解其詳細的關係。此時，系統在廠商資料庫中，也比對到中環這家公司，因此也將該公司的詳細資料顯示出來。如此，從非結構化的文件中，可以獲得事物之間的關聯，而這些關聯猶如記錄了某些知識等待探索，有些知識由文件中的自由文句中透露出來，有些知識則由事先準備好的結構化資料加以補充。在這樣的二維空間中一步步探索時，猶如在地圖上一步步發現所需的資訊或知識。因此，這樣的系統，可簡略稱為「知識地圖」。

廠商資料

公司名稱	中環股份有限公司
資本額	173.2 億
成立日期	1978/12/02
上市日期	1992/02/17
股票代號	2323
董事長	翁明顯
總經理	翁明顯
主要產品或業務	光碟產品 磁碟片 版權收入

新聞摘要

- [中環、國碩9月營收分別較8月略有增減](#)
- [中環致力多角化，以維持中長期穩定成長](#)
- [中環推出資訊家電“太空小子”](#)
- [中環預估PDA系列產品明年出貨量可大幅成長至200萬台](#)
- [中環的太空男孩可望推升營收](#)
- [中環8月將推出隨身智慧型秘書](#)
- [中環預計今年MP3出貨量可達100萬台](#)

新聞內容

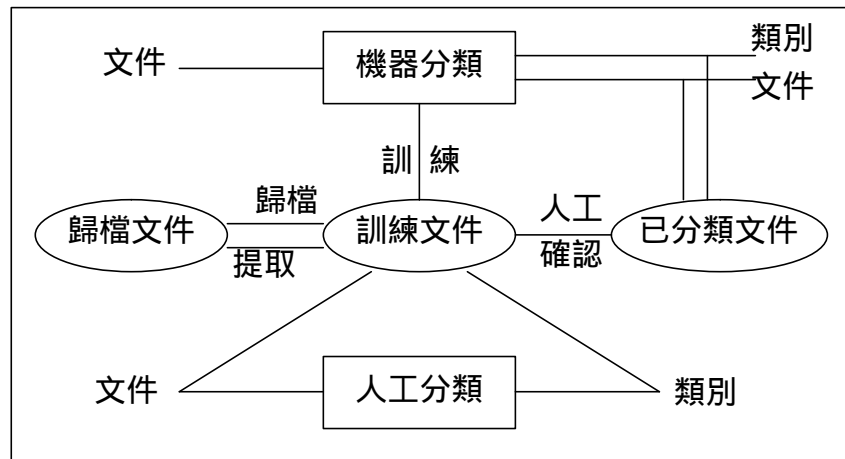
<2000-03-28> **中環預計今年MP3出貨量可達100萬台**
 (記者張珮琳)中環今年以推出CD-R、DVD-R光碟片、MP3隨身聽、以及網路為產品主軸，預計MP3出貨量可達100萬台，營收達30億，產品毛利25%，而中環只負責設計研發，再交由台灣、韓國等廠商量產製造，中環表示未來市場需求量大，不排除在大陸設廠的可能。

MP3 Player一推出時市價約5,000元一台，但中環研發出可置於MP3中，直徑5公分或6公分的CD-R光碟片，而光碟片佔生產成本約60%，此種規格生產成本比原有規格之光碟片低，因此造成MP3 Player價格下降，目前市價一台約3,000元台幣左右。

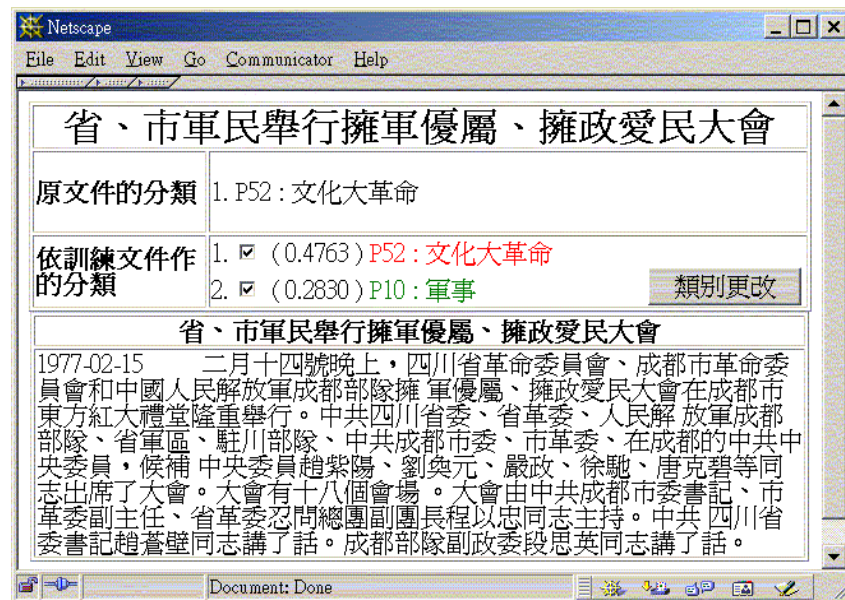
圖三：知識地圖範例。

在自動分類方面，圖四顯示自動分類的運作流程。經過訓練後，分類器即可對新進的文件做出分類。圖五顯示一篇文件被分類的情形，其中分類器分出「文化大革命」與「軍事」兩類，而且其分類的信心度分別為 0.4763 與 0.2830。這篇文件事前已經由人工分類了，而且類別為「文化大革命」。這個例子可以看出，機器分類有時可以分得很好，但有時也會被文件中的其他次要詞彙混淆，像這篇因為有「成都部隊」、「人民解放軍」等詞，就多出了「軍事」這一類。不過只要不太離譜，這篇文件多了「軍事」這一類似乎無傷大雅。

從這個例子也可看出，假如沒有機器分類，那麼人工就要仔細閱讀、分析該篇文件，然後為了標出類別，還要熟悉各個類別的意義及其在分類架構中的相對範圍。假若使用的分類架構很大，例如有上百個類別，一一記住每個類別的內容與範圍，以比對出文件的類別，常常需要熟練且具備耐心的人員才能勝任。有了機器分類後，自動分類的結果對人力而言，可以視為是一種分類提示，對減輕人力分析文件、決定類別的努力，有相當大的助益。尤其對新手而言，此種分類提示猶如有專家在旁協助，分類工作將更快能步上軌道。



圖四：自動分類流程。



圖五：單篇文件自動分類範例。

肆、結語

圖書館學中發展了相當完備的理論與實務，來進行資訊組織與主題分析的任務。然而這些都是知識密集的工作，早期除了倚賴訓練有素的人員外，很難能從機器獲得協助。然而資訊科技不斷的進步，逐漸有自動化的作法出現，有些已達到相當好的效果。本文就筆者過去數年來持續研究發展的自動化技術，做簡要的說明，並展示其可行性。這些自動化方法不是要用來取代人力，而是要與現有人力互補互助，一起達到高效率、高水準的資訊服務。

自動化的資訊組織與主題分析，還有其他的項目，受限於本文的篇幅及筆者的研究範圍，無法一一介紹。然而資訊服務者應當多方蒐集相關訊息，思考服務內容與方式如何運用新的方法，在資源有限但數位文件成長無限的情況下，使資訊服務的功能不斷加深加廣。

本文提到自動化索引與索引典建構，已應用於一些數位化圖書館計畫當中(註31)，另外自動化分類系統也已應用於如中國時報等單位，每天就數百個類別進行上千篇文件的分類。未來的工作將持續深入瞭解圖書館學的理論與實務、發展更多技術或系統，以協助資訊組織與主題分析的工作，並就已經應用的系統，從使用者的回饋中，持續的加強與改進。

誌謝

本文承藍文欽博士潤稿、加註，特此致謝。

本文由國科會計畫補助，計畫編號：NSC 91-2413-H-030-012。

附註

註1：Chan, Lois Mai, *Cataloging and Classification: An Introduction*, 2nd ed. New York: McGraw-Hill, 1994.

註2：Olson, Hope A., and John J. Boll, *Subject analysis in online catalogs*, 2nd ed. Englewood, Colorado: Libraries Unlimited, 2001.

註3：胡述兆，吳祖善。圖書館學導論。漢美圖書有限公司，民78年。

註4：索引中的用語，並不一定是出現在文件中的詞彙，索引者分析文件內容後，可能由索引典或自然語言中選擇適當的詞彙做為索引詞彙。

註5：前組合索引法與後組合索引法的作法可能有所不同。前組合法依賴索引者事先選用複合詞（即複合概念，如「檢索系統」）做為索引詞，以滿足檢索者較特定的查詢；後組合法則依賴檢索者自行就數個單一概念組合成較為特定的檢索條件來查找文件，如（「檢索」and「系統」），索引者只需索引單一概念詞即可。

註6：嚴謹的權威記錄每一筆資料大抵包括三個項目：選用的正式用語、資料來源、其他相關的詞彙。

- 註7：Olson, Hope A. (1998). Authority control in a global environment. In Kathleen de la Pena McCook, ed., *Global reach/Local touch* (pp. 210-218). Chicago: American Library Association.
- 註8：Shannon L. Hoffman and Deborah Hatch, "WEB WORLD OF AUTHORITY CONTROL," <http://www.lib.byu.edu/dept/catalog/authority/>, accessed 2002/11/29.
- 註9：另有劃一提名及主題的權威檔。
- 註10：Yuen-Hsien Tseng and Douglas W. Oard, "Document Image Retrieval Techniques for Chinese" Proceedings of the Fourth Symposium on Document Image Understanding Technology, Columbia Maryland, April 23-25th, 2001, pp. 151-158.
- 註11：Yuen-Hsien Tseng, "Automatic Cataloguing and Searching for Retrospective Data by Use of OCR Text", *Journal of the American Society for Information Science and Technology*, Vol. 52, No. 5, 2001, pp. 378-390.
- 註12：曾元顯、林瑜一。「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 期 (民 87 年 12 月), 頁 103-125。
- 註13：蔡明月。線上資訊檢索--理論與應用 (台北市：臺灣學生, 民 80 年), 頁 166-169。
- 註14：美國資訊科學學會臺北分會編。索引典理論與實務 (台北市：編者, 民 83 年)。
- 註15：黃慕萱。資訊檢索 (台北市：臺灣學生, 民 85 年)。
- 註16：莊雅蓁。「資訊檢索之索引典研究」, 中國圖書館學會會報 63 期 (民 88 年 12 月) 頁 77-89。
- 註17：索引典通常是針對特定學科領域編定的。這裡所謂「一般目的」與「特定領域」是相對的概念, 如：社會科學索引典相對於法律方面的文件, 可說是「一般目的」索引典與「特定領域」文件之例。
- 註18：Rila Mandala, Takenobu Tokunaga and Hozumi Tanaka, "Combining multiple evidence from different types of thesaurus for query expansion," Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, 1999, Pages 191 - 197.
- 註19：Gerard Salton, *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*, Addison-Wesley, 1989.
- 註20：Hsinchun Chen, Tak Yim, David Fye, and Bruce Schatz, "Automatic Thesaurus Generation

for an Electronic Community System," Journal of the American Society for Information Science, 46 (3) : 175-193, April 1995.

註21 : Yuen-Hsien Tseng, "Automatic Thesaurus Generation for Chinese Documents", Journal of the American Society for Information Science and Technology, Vol. 53, No. 13, 2002, pp. 1130-1138.

註22 : 曾元顯。「共現索引典之自動建構、評估與應用」。台灣大學圖書資訊學系四十週年系慶學術研討會，民 90 年 11 月。

註23 : 實體上，一書只有一分類號，是因為受上架陳列、收藏的限制，從檢索的角度言，一書有多個分類號並非不可能，MARC 中就允許一書多分類號。

註24 : 作者自己給的關鍵詞或許是不受控制的，但有些系統是由索引者選定描述詞 (descriptor)，則索引典就成為詞彙控制工具。

註25 : Thorsten Joachims, "A Statistical Learning Model of Text Classification for Support Vector Machines," Proceedings of the 23rd Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, 2001, pp. 128-136.

註26 : William W. Cohen and Yoram Singer, "Context-Sensitive Learning Methods for Text Categorization," Proceedings of the 19th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, 1996, Pages 307 - 315.

註27 : Hwee Tou Ng, Wei Boon Goh and Kok Leong Low, "Feature Selection, Perception Learning, and a Usability Case Study for Text Categorization," Proceedings of the 20th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, 1997, Pages 67 - 73.

註28 : Wai Lam, Kwok-Yin Lai, "A Meta-Learning Approach for Text Categorization," Proceedings of the 23rd Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, 2001, pp.303-309.

註29 : Yiming Yang, Tom Ault, Thomas Pierce and Charles W. Lattimer, "Improving Text Categorization Methods for Event Tracking," Proceedings of the 23rd Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, 2000, Pages 65 - 72.

註30 : 曾元顯。「文件主題自動分類成效因素探討」。中國圖書館學會會報 68 期 (民 91 年 6 月)，頁 62-83。

註31：曾元顯。「回溯性資料數位化服務之規劃與建置」，二十一世紀資訊科學與技術國際學術研討會 2001年11月29-30日，頁255-274。