

國家圖書館

自行研究計畫成果報告



計畫編號：

年 度：105

執行期限：105年11月20日

計畫名稱：探討圖書資料自動主題編目之可行性

研究人員：李宜容、牛惠曼、陳慧文

目 次

目 次	i
圖表目次	iv
摘要	v
第一章 緒論	1
一、研究背景	1
二、研究目的	2
三、研究方法	2
四、研究範圍與限制	3
第二章 文獻探討	4
一、關鍵詞擷取	6
二、關聯詞分析	8
三、文件歸類	9
四、機器學習	10
五、自動分類	11
第三章 研究問題與執行方法	14
一、研究問題	14

二、執行方法-----	14
第四章 研究結果-----	18
一、音樂性錄音錄影資料自動分類研究結果-----	18
二、中文圖書自動分類研究成果-----	30
第五章 研究結論與建議-----	34
一、研究結論-----	34
二、建議-----	36
附錄-----	41
一、音樂性錄音錄影資料自動分類原則-----	41
二、音樂性錄音錄影資料宗教及布袋戲出版團體表-----	43
三、音樂性錄音錄影資料無法完成分類書目資料-----	45
四、「協助編目工作的資訊技術研究計畫」會議討論內容摘要-----	55
參考文獻-----	60

圖 表 目 次

圖 3-1	音樂性錄音錄影資料自動分類開發模型圖-----	15
圖 4-1	含有「客語」時自動分類結果可為「客家民謠」及「流行歌曲」	18
圖 4-2	跨條件而給予兩組類號及主題詞-----	19
表 4-1	音樂性錄音錄影資料自動分類測試結果及結果分析-----	20
圖 4-3	「五月天」之演唱合輯「原聲帶精選—超級劇無霸」-----	22
圖 4-4	「原聲帶精選—超級劇無霸」與表演者的自動分類-----	23
圖 4-5	「長榮集團46週年慶祝音樂會」之「發行公司」及「專輯類型」	24
圖 4-6	欄位260\$b「交響」及500\$a「團體演奏專輯」有助於自動分類---	24
圖 4-7	「霹靂英雄音樂精選五十雷霆劍海錄原聲帶」之「發行公司」---	25
圖 4-8	欄位260\$b「霹靂國際多媒體」有助於自動主題分析-----	26
圖 4-9	「幸福快遞電影原聲帶」之「專輯名稱」與「有聲出版類別」---	27
圖 4-10	依據正題名及附註項之「電影原聲帶」進行自動分類-----	27
圖 4-11	「慈悲自在」之「發行公司」及「有聲出版類別」-----	28
圖 4-12	出版社「原動力」及附註項之「宗教音樂」-----	29
圖 4-13	自動主題分析開發模型圖-----	32

摘 要

圖書館資訊組織之主題分析係描述資料的內涵，相較於描述資料的形態之記述編目，需要運用更多人類的心智活動，以確保提供精確的類號與主題詞檢索。利用自動化技術協助主題分析的研究，期能協助人工主題分析工作，以節省編目人力與時間成本。

本研究採用實證研究法，發掘人為經驗之主題分析知識，將其轉換為可具體呈現之規則，提供程式撰寫與運行，並輸入特定類型與數量之音樂性錄音錄影資料，批次執行後自動產生主題分析結果，經檢測結果之可處理率與正確率，進而提供進行中之中文圖書自動分類研究參考。

由音樂性錄音錄影資料自動分類研究結果得知，題名項目之內容詞彙與主題判斷相關性低於編目員傳統經驗、出版者之出版領域與主題判斷存在高度相關、著者之專業領域與主題分析存在高度相關、利用其他可取得之資源輔助主題分析、組合題名及著者與出版者提高主題分析相關度；且測試結果之可處理率與正確率均相當高，對於節省主題分析之人力與時間具有成效。中文圖書自動分類現有研究成果包括：題名項目之內容詞彙經自動斷詞可取得相關主題分析、出版者之出版領域存在特定性、著者之專業領域出現主題集中現象、組合各項關鍵項目提高精確率。

本研究對於實行自動主題編目工作給予以下建議：以詞庫及分類表類

目與主題詞表為輔助斷詞工具、擴大題名範圍至其他相關題名項目、運用外部線上資源加強了解資料內容、引用著者權威控制資料辨別著者異同、依編目員主題分析熟悉度分配工作、加入計算主題分類工具書內詞彙之比重、找出自動主題分析關鍵項目之最佳權重、容許出現一組以上之類號及主題詞、隨時檢視自動分類結果並調整機器學習機制。

中文關鍵詞：自動主題分析，自動分類，關鍵詞擷取，關聯詞分析，機器學習

第一章、緒論

一、研究背景

圖書館之資訊組織工作主要分為記述編目及主題分析，主題分析包括類號標引及主題標引，不同於記述編目以描述資料的形體特性為主，主題分析係描述資料的內涵，因此需要運用更多人類的心智活動，以確保提供精確的類號與主題詞，方便讀者透過類號或主題詞進行資料查詢，進而獲取所需資料。

隨著資訊科技的進步，網際網路的崛起，數位資源的快速發展，各式資料或文件的加值與運用增多，使得資訊組織的需求提高，近十年來有關人工智慧（Artificial Intelligence, AI）在眾多領域的應用皆有相當大之成果，因此在資訊組織的運用可行性，值得圖書館界思考。過去國內已開始有學者進行相關研究，多數研究著重於藉由關鍵詞擷取、關聯詞分析、機器學習（Machine Learning）、主題分類、文件歸類等方式。若可藉電腦進行大批書目資料統計分析，找出人工主題分析的關鍵要素與邏輯，針對待處理資料建議出供圖書館編目員參考的主題分析結果，以協助圖書館館藏資源之主題分析工作，即可節省編目人力與時間成本。

二、研究目的

有鑒於主題分析係資訊組織工作中最為耗費編目員心力的項目，在未以人為方式進行主題分析前，僅透過書目紀錄基本著錄項目（例如：採訪登錄簡編資料、透過申請ISBN及ISRC建立之轉檔簡目）的內容，進行關鍵詞擷取與分析，並且藉由音樂性錄音錄影資料以往人為的經驗歸納出主題分析之規則的先行研究，同時做為本館與國立臺灣大學工程科學與海洋工程學系暨研究所黃乾綱副教授研究團隊有關「協助編目工作的資訊技術研究計畫」的基礎與參考，期能據此歸納與整理出主題分析的規則，進而推衍出與主題分析相關的關鍵點，提供機器學習的訓練，歸納建議自動主題分析的法則，以節省人工主題分析作業的時間與人力。因此，希望藉由本研究達到以下目的：

- （一）將專業編目人員主題分析內隱知識進行整理、歸納與應用。
- （二）將主題分析的知識移轉為機器可理解與學習。
- （三）藉由程式運作自動產生主題分析結果（包括：類號及主題詞）。

三、研究方法

本研究採用實證研究法，探索人為經驗為基礎，以實際操作作為研究之啟始，先發掘人為經驗之知識，將之轉換為可具體呈現之規則，提供程式撰寫與運行，並輸入特定類型與數量之音樂性錄音錄影資料，批次執行後自動產生主題分析結果，並檢測結果之精確性。編目員藉由前述研究，進而提供

開發機器可學習之自動主題分析功能之參考，並與臺大研究團隊進行溝通，以透過會議討論方式，共同合作及研究出可供輔助一般圖書編目之自動主題分析系統。

四、研究範圍與限制

本研究之研究範圍主要為自ISRC系統轉出之音樂性錄音錄影基本書目資料為主，推算出可由機器判定主題分析之規則，進一步延伸至一般中文圖書。

依據相關研究與文獻發現機器學習之學習資料越多，自動主題分析的結果會越好，因此有關自動主題分析的實證研究成效，仍需視學習資料的多寡而定。

但研究中之資料均為特定類型之資料，因此以ISRC系統轉出之音樂性錄音錄影之基本書目資料為研究對象，乃因ISRC系統可提供轉入之欄位，且資料範圍較小，人力於此方面資料可深入研究與探討歸納。

而預定延伸進行之國家圖書館書目資料庫之中文圖書書目，因中文語文為研究人員所熟悉，且為目前國家圖書館館藏業務的重點，可以參與討論的編目員較多，期能藉此研究後，對於中文資訊組織之主題分析工作有所助益，因此本研究暫未加入其他語文和類型資料。

第二章、文獻探討

就資訊的利用與服務而言，資訊組織工作是整體圖書館資源服務的基礎，也由於有了資訊組織，使得圖書館系統目錄與資源內容得以為讀者查詢及應用，進而創造國內的各項研究成果與學術發展。

資訊組織傳統的主題分析工作，無論是類號標引或者是主題詞標引，皆需要耗費大量的人力與人類的心智活動，而且相同主題之資源可能因為不同的人，其主題分析的結果便會有所不同。

然而國內出版品類型及數量逐年增多，圖書館之資訊組織作業的編目人力成本逐年攀升，如果以一本書自採購、分編加工、至圖書上架供閱，其間所需處理的複本控制、驗收登錄、分類編目、圖書加工，直到圖書上架供閱等作業之成本，可能高於該書的售價。

傳統式的人工主題分析，精確度較高，但是缺點是速度稍慢，而且主題分析的結果往往容易由於不同的分析者之學術素養與認知經驗而受影響，造成相同主題資料之主題分析結果不一致（曾元顯，民91年12月，頁7）。因此八〇年代興起有關自動主題分析的研究，以了解透過自動化技術能否達到輔助或取代人工的主題分析工作，從而降低圖書館技術服務工作的人事成本。

透過機器的自動主題分析，基本上結果的一致性稍高，且相對地速度較

快，因此可以節省大量的人力與時間成本，然而缺點是必須預先準備大量的訓練資料或文件，並且可能出現與資料之主題內容不相干的主題分析結果（曾元顯，民91年12月，頁7）。

圖書館之資訊組織工作，包括記述編目與主題編目，主題編目前，必須先經過主題分析，主題分析是將圖書資料依其內容、性質分門別類，並給予與主題相同的特定類號及主題詞，這是相當高階的知識組織與管理工作。因此，主題分析作業自動化之前，必須先整理出主題分析時的規則，機器或電腦方能據以學習及執行。

先驅者的研究與實驗結果提醒我們，利用不同的分類方法處理不同的分類問題，會有不同的表現與結果；使用相同的分類方法處理不同的分類問題，其結果與成效不盡相同；以單一的分類方法對於單一類型文件分類上所達成的改進成效，不保證此單一的分類方法對於另一類型文件分類上也會有相同的改進效果（曾元顯，民92年9月，頁10）。

民國95年林昕潔以書名、書籍簡介、著者、出版社等詮釋資料設計分類系統的研究中發現，除了書名及書籍簡介外，著者與類別高度相關，並透過出版品的類別比例，判斷出版社與類別的相關程度（林昕潔，民95，頁46-47）。民國97年黃嘉宏的研究指出，除了題名特徵擷取有助於自動分類外，著者資訊與類別也具有相當程度的關聯，導入著者欄位來輔助分類，可以有效增進或提升自動分類的成效（黃嘉宏，民97，頁38，44）。

然而，最佳的主題分析規則通常難以人工分析歸納方式獲取，而是在機器或電腦進行自動主題分析之前，加以統計、分析與訓練，使其自動學習出人工主題分析的知識與經驗（曾元顯，民91年12月，頁6）。

以目前自然語言的處理技術，尚無法讓機器或電腦理解任意詞句，因此機器在進行主題分析時，需將資料或文件的語句拆解成較小的單元，也就是資料或文件內所謂的「關鍵詞」或「特徵詞彙」，再自這些詞彙中找出其與類號及主題詞的對應關係（曾元顯，民91年12月，頁6）。

自動化的資訊組織和主題分析工作與關鍵詞擷取、關聯詞分析、文件歸類、機器學習、自動分類等息息相關，然而八〇年代以來，國內外相關研究也顯示：“沒有一種方法可以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”

（曾元顯，民103，頁7）。

一、關鍵詞擷取

關鍵詞擷取是資料或文件自動處理的基礎，所謂關鍵詞擷取是一種辨識數位資料或文件內具有意義及代表性的字串、片語、詞彙或內容片斷的自動化技術（曾元顯，民103，頁9）。關鍵詞是資料或文件之主題意義的最小單元，因此大部分非結構化的資料或文件自動處理前，皆需進行關鍵詞擷取，從而理解資料或文件的內容主題，以便作為主題分析的依據。

然而關鍵詞擷取前，就資料或文件而言，會有斷詞的問題，斷詞時需輔

以同義異名詞表，或透過標準的權威控制款目進行，以避免資料的漏檢或不相對應，但是人工斷詞費工耗時，目前多數系統皆採用現有已建立的詞庫（例如：字辭典、百科全書，主題詞表……等）作為斷詞的依據。

斷詞的同時還必須考量到同義異名詞彙的處理，也就是有必要建立同義詞表，使其一之詞彙可查到其他形式的同義詞彙，在圖書館的資訊組織工作上，把這類意義相同但是用詞不同的詞彙，以一致且標準形式的紀錄記載及管理的方法稱為權威控制，而記載這些同義異名詞彙的檔案即所謂的權威檔（Olson，1998）（Hoffman & Hatch，2002）。

若不以現有詞庫作為斷詞依據，應用於自動主題分析時之另一方式是將自動主題分析的關鍵項目（例如：題名、集叢名……等）內之文字內容等字串，分割為兩個字、三個字、四個字……，以分割後的詞彙探索出現的頻率，藉此找出與主題相關之特徵資料，即所謂的「特徵選擇」。

抽取出特徵詞彙後，需把這些詞彙與類別進行對應，然而如何去界定哪些詞彙屬於某個類別？哪些詞彙不屬於某個類別？哪些詞彙對某個類別而言是重要的？哪些詞彙對某個類別來說是不重要的？或者不同詞彙對於不同的類別有不同的鑑別度？再者如何決定這些詞彙在每個類別應有的權重（Cohen & Singer，1996）（Joachims，2001），這正是之所以需要機器學習的用意。

作為主題分析根據的特徵詞彙之個數和選取方式，會直接影響文件分類

的成效。所選擇出來的特徵詞彙，短的特徵詞彙出現在未來待處理的文件中之機率較高，但是卻不易用來區分類別，反而是長的特徵詞彙出現在未來待處理的文件中之機率較低，但是其對於類別的鑑別度較高，一旦出現則比較容易用來預測類別，而且不同的特徵詞彙對於分類成效的貢獻度不同，越好的分類器，所需的特徵詞彙越少（曾元顯，民91年12月，頁7）。

二、關聯詞分析

從非結構化的資料或文件中，我們可以獲取事物或事項之間的關聯，這些關聯就像是記載了一些知識等待人們探索，有些知識可從資料或文件中的任意詞句中顯露出來，有些知識則可由事先預備好的結構化資料或文件加以補足（曾元顯，民103，頁8），就像「知識地圖」一般，人們透過這類地圖，可以逐步地發現所需的資訊，進而整理、歸納出概括性的原則。

關聯詞係由資料或文件內容先後一同出現的詞彙，透過詞彙間的共現或共伴現象而產生關聯之詞彙，即所謂關聯詞。據此所分析出的關聯詞則可經由詞彙檢索的精確率及召回率予以評估其關聯的程度。

此外，關聯詞分析並不限於同一資料欄位內詞間關聯的分析，跨欄位的關聯詞分析，尤其適用於自動主題分析，如何透過不同欄位之關鍵詞的關聯性分析，將有助於自動主題分析的機器學習與後續之類號及主題詞的預測和建議。

三、文件歸類

文件分類或文件歸類係指依文件內容主題或其相似度給予特定或相似的類別（也就是替文件貼上標籤的工作），這些類別多是事先定義，也可以是未事先定義的，其目的在於對文件分門別類後以便於查找（包括提供使用者依類別或主題查找所需的資料）及管理，而不受限於文件所採用之語言、詞彙或用語（曾元顯、莊大衛，民92年9月，頁1）。

要將文件分類自動化，須先整理出分類時的使用規則，機器或電腦方能據以執行。然而，有效的分類規則，雖然需借重人為的經驗，但是除了人為經驗外，透過電腦相關的統計與分析，則可以進一步獲取人為經驗以外的規則，這是由於效果最佳的分類規則，常常是難以人工分析與歸納而獲取（曾元顯、莊大衛，民92年9月，頁1）。

文件自動分類技術係以已分類之資料作為訓練資料，透過機器學習之演算法的學習，自動建立分類行為的類別模式（classification model），也就是文件分類的規則（陳信源、葉鎮源、林昕潔、黃明居、柯皓仁、楊維邦，民98年6月，頁3）。

然而文件自動分類之前，需有機器學習，機器學習時會分析現有資料情況，以便找出相關規則。民國98年陳信源等6位即針對圖書資訊之敘述資料的著者和作品領域與出版者和出版品類別進行相關程度的研究，並考量專家挑選特徵和各項加權的比率，進一步建立最佳的圖書分類模型（陳信源、葉

鎮源、林昕潔、黃明居、柯皓仁、楊維邦，民98年6月，頁12-19）。

文件分類或文件歸類尚須顧慮類別的變動，以計算機協會（Association for Computing Machinery, ACM）對其論文的分類政策即是讓使用者可以使用舊的類別找到過往的文件資料，也就是說被淘汰的舊的類別，仍繼續保留，但是已不用於新文件的分類，而歸檔文件也須記載新的分類資訊，以便回溯之文件，也可如同新文件一般，被同一主題新的類別所對應（Introduction to the ACM Computing Classification System [1998 version]）與檢索。

四、機器學習

機器進行自動分類前，須加以訓練，以便機器自動學習出人工分類的經驗與知識（曾元顯，民91年6月，頁4），也就是讓機器去分析一堆訓練文件，這些訓練文件必須記錄人工分類的知識，機器反覆讀取文件及標示類別後，須能自動歸納出相應的原則，以便下次機器看到類似文件時，可以建議出適當的類別（曾元顯，民91年6月，頁4）。

在機器歸納出相應的規則之前，必須提供機器訓練的文件進行學習，並提供測試的文件執行測試，而且訓練文件越多，機器學習的效果越好（曾元顯，民91年6月，頁4）。若詞彙與類別的關係與詞彙在特定類別中出現的次數相關，則機器學習時會以在特定類別中出現的詞彙次數判別該詞彙與某類別間的關係。

民國92年曾元顯利用文件自我擴展方式，將人工尚未分類的文件，以機

器自動分類後，作為訓練文件，並反覆擴展及進行機器學習，直到分類器收斂為止，藉此提高機器學習成效及分類的準確度（曾元顯、莊大衛，民92年9月，頁3），實驗結果發現，訓練文件越少時，文件自我擴展方式對於機器學習的改進效果越明顯（曾元顯、莊大衛，民92年9月，頁9）。

但是如果機器學習的資料有分類不一致的情形時，機器便無法在資料完全正確的環境下獲得完整的訓練，而且既然連人工分類答案都不一致時，機器學習的成效自然會受到影響（曾元顯，民91年6月，頁15）。

機器學習而來的分類規則通常分為兩種：第一種是符號式的，例如：「IF conditions, THEN categories」；第二種是數值式的，即一串詞彙及其權重數值，以及結合這些數字的運算方式。這些透過機器學習而得的分類規則，多半是人為難以了解，而且也無從修改的（曾元顯，民91年6月，頁5-6）。

五、自動分類

自動分類會涉及分類架構或分類表，分類架構分為單層式及階層式。圖書館圖書資料的分類法為階層式的分類架構，類別先分大類，大類下細分小類，小類下還會逐層細分，越底層的類別，在主題上是越相近。自動分類的處理目標是把資料分類至最底層，而不是中間或上層。

大部分自動分類的方式是把每個類別視為一個獨立的分類問題，再利用二元分類器（binary classifier）就每個類別單獨處理。自動分類前沿實驗結果顯示，「文件數多的類別僅佔極少數，文件少的類別佔大多數」（曾元顯，

民91年6月，頁14）。

分類不一致是普遍而且很難避免的現象，如何有效防止及自動偵測分類不一致，則是另一項研究課題。即使分類資料不完全正確，但是好的分類器仍可從中進行有效的學習，而自動分類最大的優點之一便是可能比人工分類之一致性來得更高，不過自動分類的目的並不是要來取代人工分類，而是協助降低人力負擔的成本，好讓人力資源運用在更為專精的工作上。

有關自動分類研究方面，民國104年林頌堅以期刊的題名及摘要之文字內容的主題特徵提出主題類別自動產生方式，也就是先計算期刊在主題特徵上的相似度，再將具有相似特徵的期刊聚集，使用期刊群集的方式探索學科、次學科、與專業領域（林頌堅，民104年夏季，頁273），結果發現，分類規則若加入更多其他主題特徵的分類規則時，容易因此讓許多正案例被預測為負方的錯誤判斷結果，簡言之，即是仍然會發生少數類別預測錯誤的情況，然而經由觀察這類預測結果錯誤之期刊的編輯主旨，可進一步發現這些期刊都是主題多元的跨學科整合性期刊（林頌堅，民104年夏季，頁287）。

自動分類的流程包括：特徵擷取與選取、建立文件表示式、分類法則或模型的學習與建立。「特徵擷取與選取」是指從文字資料中擷取具有語意的詞彙；「建立文件表示式」係利用特徵集合與其所對應的特徵值建立文件的表示方法，以便作為電腦運算的根據；「分類法則或模型的學習與建立」則是參考已分類的文件作為訓練資料，透過機器學習演算法，自動學習並建立

分類行為的類別模型（陳信源、葉鎮源、林昕潔、黃明居、柯皓仁、楊維邦，民98年6月，頁4）。

而自動分類的成效評估可以考量三個方面，即：正確率（accuracy）、精確率（precision）、召回率（recall）。但是此三方面其實會是相互嵌制，端視使用者的期望而訂出對於分類器的期望值。因此自動分類的最大成效受限於類別之間的區別程度，越容易區別出來的類別，人工分類會越一致，自動分類的成效就會越高，相反地，越不容易區別出來的類別，人工分類便會越不一致，自動分類的成效自然越低（曾元顯，民91年6月，頁19）。

第三章、研究問題與執行方法

一、研究問題

本研究將探討的問題分別如下：

- (一) 判斷資料主題分析之關鍵項目。
- (二) 人工判斷資料主題分析關鍵項目之優先順序與利用電腦自動判定之優先順序比較。
- (三) 電腦自動判定之可處理性及正確性。

二、執行方法

- (一) 探索編目員有關主題分析工作經驗的作業行為。
- (二) 找出待編資料主題分析關鍵項目。
- (三) 將前項作業行為及關鍵項目整理歸納為主題分析的規則。
- (四) 將該項規則轉換為機器可執行的程式。
- (五) 輸入資料進行測試。
- (六) 產出結果並進行分析。

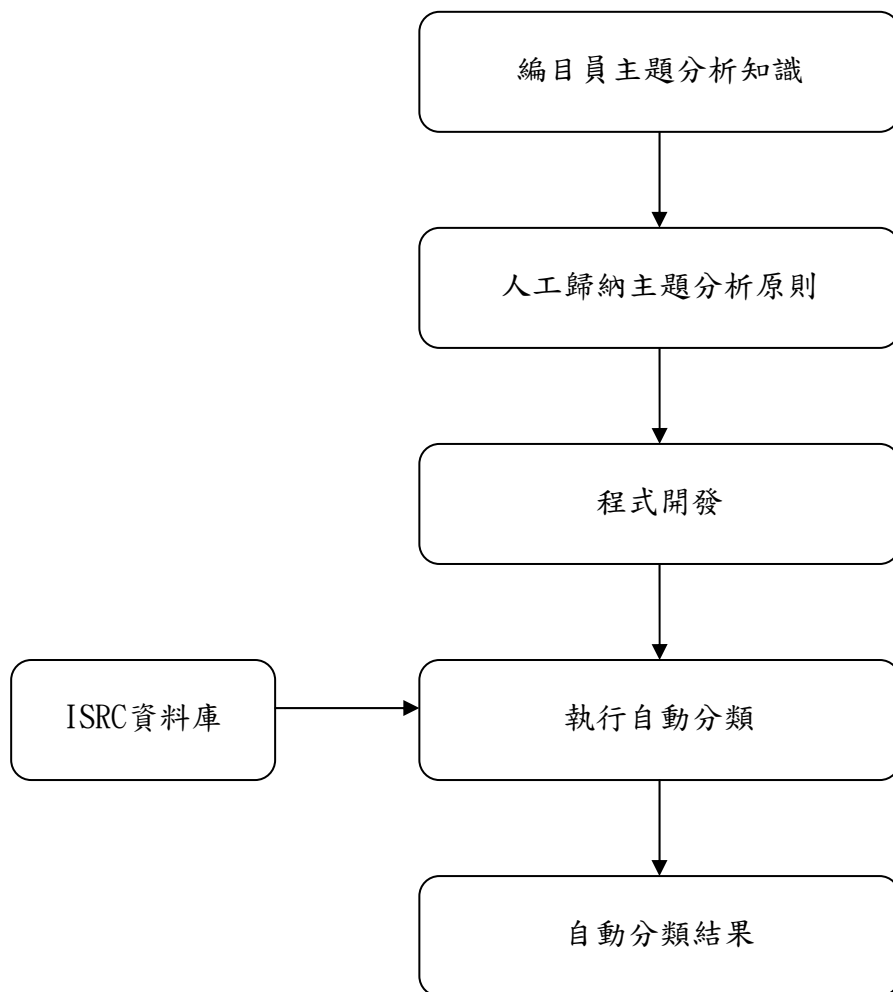


圖3-1 音樂性錄音錄影資料自動分類開發模型圖

茲將上述各項執行方法分別說明如下：

(一) 就編目員作業行為探索方面，包括：

- 1、人工斷詞及概念分類：概念分類包括同義詞、同名異義詞、廣義詞、狹義詞的對照轉換，以及關聯詞的分析。
- 2、查詢現有書目資料庫相關書目的分類。
- 3、找尋及比對主題工具：含括中文圖書分類法及中文主題詞表。

4、給定分類號及主題詞。

（二）找出待編資料主題分析關鍵項目

1、題名：包括正題名及副題名。

2、著者學科領域：除了著者之學科領域外，尚須考量著者筆名及同名不同人之著者姓名問題。

3、目次、序言、摘要：書目編製時，有關目次、序言、摘要等資料之建檔成本較高，因此考量引進及利用外部資源，以利辨識待編資源之主題性。

4、叢書名：由於集叢名多半含有學科或主題特性，其功能類似題名，雖然範疇較為廣泛，但是仍然有助於主題之辨識。

5、出版者的出版領域：除了大型出版者之外，多數出版者有其專業之學科領域，若出版者為政府機關，其出版品之內容主題更為特定，此外，某些出版者僅出版特定主題或類型之資料，因此，出版者也可做為主題分析判斷之參考。

（三）將前項作業行為及關鍵項目整理歸納為主題分析的規則。

經由人工斷詞及概念分類的經驗，以及關鍵項目的比重，比對主題工具內之學科領域與詞彙，將之交叉整理及系統條列後，產出主題分析之規則。

（四）將該項規則轉換為機器可執行的程式。

以不同關鍵項目所擷取之關鍵詞作為程式判斷條件，自動帶入類號及主題詞之方式撰寫程式，同時納入特定主題之出版團體列表，作為輔助判斷主

題的依據。

（五）輸入資料進行測試。

以ISRC系統進行轉出之音樂性錄音錄影資料139筆進行測試。

（六）產出結果並進行分析。

續前項資料測試後，將結果進行分析，以提高自動分類的比率及精確度，進而將所得結果提供進行一般圖書之自動主題分析工作參考。

第四章、研究結果

本研究結果將分為音樂性錄音錄影資料自動分類研究結果，及目前還在進行之中文圖書自動分類研究，對自動分類功能的運作已有相當成果。

一、音樂性錄音錄影資料自動分類研究結果

本項有關音樂性錄音錄影資料自動分類之實證研究，係以ISRC系統進行轉出之音樂性錄音錄影資料139筆進行測試，結果達成自動分類者計95筆，可處理率達68%以上，其他另有44筆未給予類號及主題詞。由產出自動分類的95筆中，有1筆為跨條件而給予兩個類號及主題詞（圖4-1及4-2），這類情形正好可以讓編目員得知此資料可能包含多個類號及主題詞。



圖4-1 含有「客語」時自動分類結果可為「客家民謠」及「流行歌曲」

084	—	913.6033	^
		ncsclt	
084	—	913.533	^
		ncsclt	
245	10	客家歌舞劇「義民禮讚」演唱專輯	
		[錄音資料]	
		呂錦明作曲；葉日松作詞；臺灣客家山歌團表演	
260	—	臺北市：	
		臺灣客家山歌團。	
		2014	
300	—	1張唱片(49分)：	
		數位，立體聲；	
		4 3/4吋	
		1張歌詞	
500	—	團體演唱專輯	
500	—	雷射唱片	
505	0	1. 熱血染成一枝旗 / 呂錦明作曲；葉日松作詞；	
		臺灣客家山歌團/混聲合唱演唱(3:58)--2. 情話綿綿 /	
		呂錦明作曲；葉日松作詞；姜雲玉，	
		呂錦明/二重唱演唱(5:59)--3. 萬古留名比花香 /	
		呂錦明作曲；葉日松作詞；姜雲玉，孔仁芳/領唱；	
		臺灣客家山歌團/合唱演唱(3:14)--4. 戰地明月 /	
		呂錦明作曲；葉日松作詞；陳石貴演唱(5:55)--5. 千人	
		飯送溫情 / 呂錦明作曲；葉日松作詞；孔仁芳領唱；	
		臺灣客家山歌團/女聲合唱演唱(3:44)--6. 血染大地 /	
		呂錦明作曲；葉日松作詞；	
		臺灣客家山歌團/男聲合唱演唱(3:24)--7.	
		化作火焰轟轉來看大家 / 呂錦明作曲；葉日松作詞；	
		臺灣客家山歌團/混聲合唱演唱(3:38)--8. 枋寮香火射四方	
		/ 呂錦明作曲；葉日松作詞；	
		臺灣客家山歌團/混聲合唱演唱(4:13)--9.	
		問奏曲-義民嘉年華舞曲 / 呂錦明作曲；葉日松作詞；	
		黃佩舒/鋼琴獨奏演唱(1:19)--10. 七月二十嘉年華 /	
		呂錦明作曲；葉日松作詞；臺灣客家山歌團/童聲，	
		混聲合唱演唱(3:53)--11. 萬古流芳在人間 / 呂錦明作曲	
		；葉日松作詞；姜雲玉，陳石貴/領唱；	
		臺灣客家山歌團/合唱演唱(5:29)	
		製作：臺灣客家山歌團；編曲：呂錦明	
508	—	客語	
650	7	1cstt	
650	7	流行歌曲	
		1cstt	
650	7	客家民謠	

圖4-2 跨條件而給予兩組類號及主題詞

進一步分析，達成自動分類的95筆書目資料中，由兩位15年以上編目經驗之資深編目員逐一比對，其中94筆經自動分類後書目之分類號及主題詞完全正確，正確率達98%以上。另其中有1筆為有兩組分類號及主題詞，其中一組為正確，但需人工判斷，再取其正確分類，雖然未精準提供一組正確主題分析項目，但確實達成指引館員及縮短主題分析時間之成效。

44筆未給予類號及主題詞係由於其主題並未出現於音樂性錄音錄影資料自動分類原則，然而此44筆中仍有部分可區分出其特有的主題類型，例如：歌仔戲有4筆、兒童故事有2筆、京劇1筆、「五月天」演唱專輯1筆未被給予類號及主題詞，因此可針對這些部分再給予分類原則，在未來可處理率

可望再提升。其餘完全無法進行自動分類處理之36筆書目，究其原因為其主題較為多元、出現詞彙無一致之規則性等。

表4-1 音樂性錄音錄影資料自動分類測試結果及結果分析

測 試 結 果	結 果 分 析	備 註
達成：95筆	94筆依規則完成自動分類	給予原則內規定之類號及主題詞
	1筆跨條件完成自動分類	給予兩組類號及主題詞
未達成：44筆	4筆歌仔戲未符合條件	無法自動分類，但未來於分類原則上可擴充條件
	2筆兒童故事未符合條件	無法自動分類，但未來於分類原則上可擴充條件
	1筆京劇未符合條件	無法自動分類，但未來於分類原則上可擴充條件
	1筆流行歌曲表演團體「五月天」演唱合輯未符合條件	無法自動分類，但未來於分類原則上可擴充條件
	36筆未符合條件	無法自動分類

藉由以上音樂性錄音錄影資料自動分類研究之進行，得出以下研究結果：

(一) 題名項目之內容詞彙與主題判斷相關性低於編目員傳統經驗

本研究之音樂性錄音錄影資料，由人工主題分析之經驗所歸納出之原

則，其多數來自於題名項目之內容詞彙，編目員係由題名項目獲取與主題相關之詞彙，再查尋比對分類表類名及主題詞表之主題詞，但是在本研究中發現，有別於編目員傳統經驗，此類資料之題名似乎對主題之判定影響程度不若出版者與表演者高，研究中有部份書目甚至不需從題名項目就可給予正確的主題分析。

（二）出版者之出版領域與主題判斷存在高度相關

在音樂性錄音錄影資料中，部分出版者有其特定的出版領域，在本研究中發現編目員依其經驗以宗教及戲劇類型將出版社進行分類，整理出佛教出版社、基督教出版社、和布袋戲出版社等三種出版社類型，佛教出版社及基督教出版社可搭配題名詞彙進行主題分析，而布袋戲出版社則可以出版社項目單獨作為主題分析判斷之依據。

（三）著者之專業領域與主題分析存在高度相關

音樂性錄音錄影資料，由於其著者過於繁雜（包括：作曲者、作詞者、演唱者、表演者……等），故本研究中目前並無人力進行相關規則之大量分析，因此在音樂性錄音錄影資料沒有以著者作為主題分析之判斷依據，但由未被分類之資料中發現，若可將著名之音樂表演個人及團體做不同領域之劃分與整理，則自動分類之可處理率及精準率會更為大幅提高。

研究中未符合條件之1筆流行歌曲表演團體「五月天」演唱合輯未給予建議之類號及主題詞（圖4-3及4-4），乃因「五月天」演唱合輯之題名不具

主題性，且「音樂性錄音錄影資料自動分類原則」未以個人或團體名稱訂出歸類原則，考量特定表演者對於演奏音樂或流行歌曲有其專長或特定領域，因此或可以這類表演者的個人及團體名稱建立列表，以作為自動主題分析與歸類的判別條件，如此一來，「五月天」及其他團體的演唱合輯或專輯，就會被自動分類為流行歌曲的類號及主題詞。

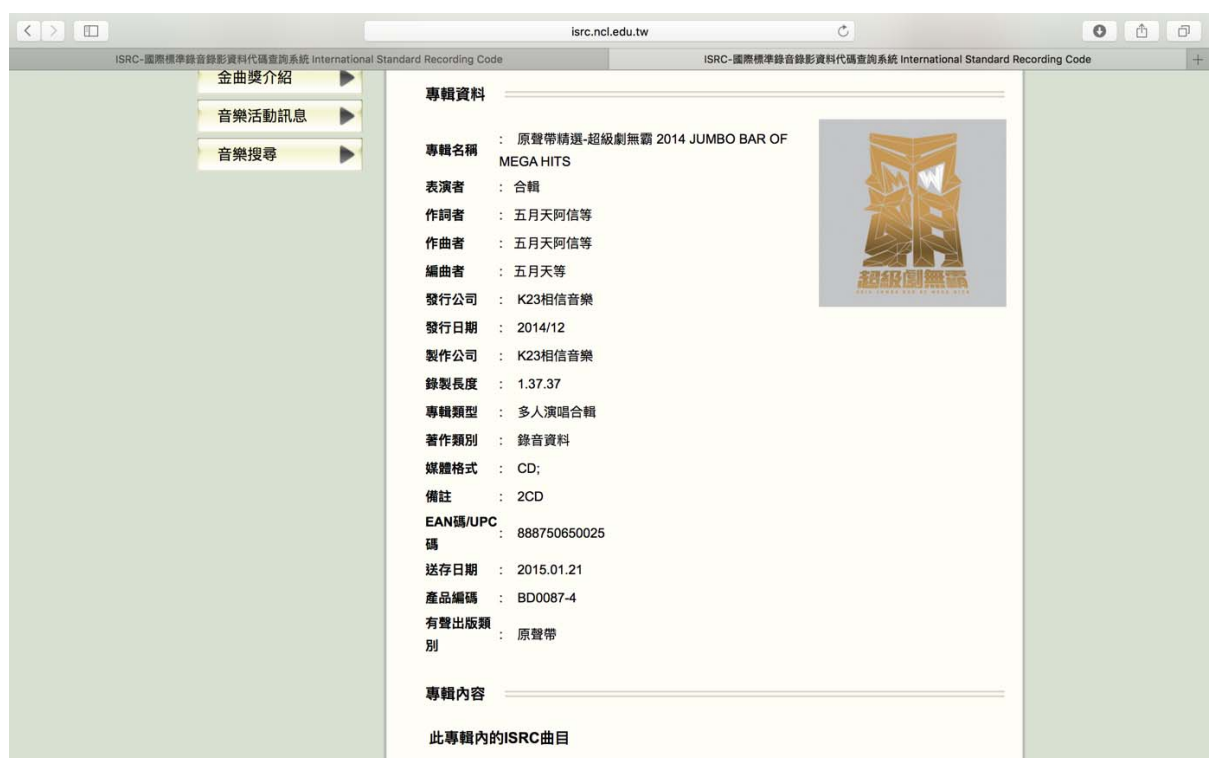


圖4-3 「五月天」之演唱合輯「原聲帶精選—超級劇無霸」

	chi
	CYT
	CYT
	ccr
	S1
066	原聲帶精選-超級劇無霸 2014 JUMBO BAR OF MEGA HITS
245	[錄音資料]
10	五月天阿信等作曲；五月天阿信等作詞；合輯表演
260	臺北市；
	相信音樂，
	2014
300	1張唱片(109分)：
	數位，立體聲；
	4 3/4吋
500	多人演唱合輯
500	原聲帶
500	888750650025
500	雷射唱片
505	0
	1. 將軍令 / 五月天阿信作曲；五月天阿信作詞；
	五月天演唱(3:28)--2. 為你平定的天下(柔情版) /
	李劍青作曲；五月天阿信作詞；李劍青演唱(5:09)--3.
	走了 / 亂彈阿翔作曲；亂彈阿翔作詞；
	亂彈阿翔演唱(4:57)--4. The Look / Per Gessle作曲
	；Per Gessle作詞；亂彈阿翔，家家演唱(3:52)--5.
	孤獨 / 山村隆太，百田留衣作曲；山村隆太，百田留衣；
	改編詞：陳沒作詞；flumpool演唱(4:58)--6.
	真實故事改編 / 五月天石頭作曲；五月天石頭作詞；
	五月天石頭演唱(5:08)--7. 我還是愛著你 / MP魔幻力量
	廷廷作曲；MP魔幻力量 廷廷作詞；
	MP魔幻力量演唱(4:49)--8. 步步 / 阿信，Cola
	Kai作曲；阿信作詞；五月天演唱(4:38)--9. 入陣曲 /
	五月天怪獸作曲；五月天阿信作詞；
	五月天演唱(3:29)--10. 由我們主宰 / Andrew
	Lawrence Bloch, Edmond Thomas Dunne, Joseph
	Lawrence Altruda, Antonina Armato, Tim
	James作曲；Andrew Lawrence Bloch, Edmond
	Thomas Dunne, Joseph Lawrence Altruda,
	Antonina Armato, Tim James; 改編詞：陳沒，
	五月天阿信作詞；歡開合唱團，亂彈阿翔，五月天，白安，
	家家，嚴爵，MP魔幻力量演唱(2:35)--11. 彩虹黑洞 /
	MP魔幻力量 廷廷作曲；MP魔幻力量 廷廷作詞；
	MP魔幻力量演唱(4:09)--12. 最後一首情歌 / 嚴爵作曲
	；黃婷，嚴爵作詞；嚴爵演唱(4:08)--13.
	有一種勇氣叫放棄 / 木蘭號AKA陳韋伶作曲；
	木蘭號AKA陳韋伶，黃婷作詞；丁噶演唱(4:05)--14.

圖4-4 「原聲帶精選—超級劇無霸」與表演者的自動分類

(四) 利用其他可取得之資源輔助主題分析

在音樂性錄音錄影資料中，自ISRC取得之其他描述性資料，也同樣進行斷詞及詞彙分析，找出其中可作為主題判斷之詞彙，進而輔助音樂性錄音錄影資料之主題歸類。

以ISRC國際標準錄音錄影資料代碼查詢系統 (<http://isrc.ncl.edu.tw>) 查詢「長榮集團46週年慶祝音樂會」，於專輯資料畫面之「發行公司」內容出現「交響」二字，「專輯類型」一欄之內容「團體演奏專輯」，有助於此一資料之主題分析（圖4-5及4-6），即類號為「912.31」、主題詞為「交響曲」或「管絃樂」。

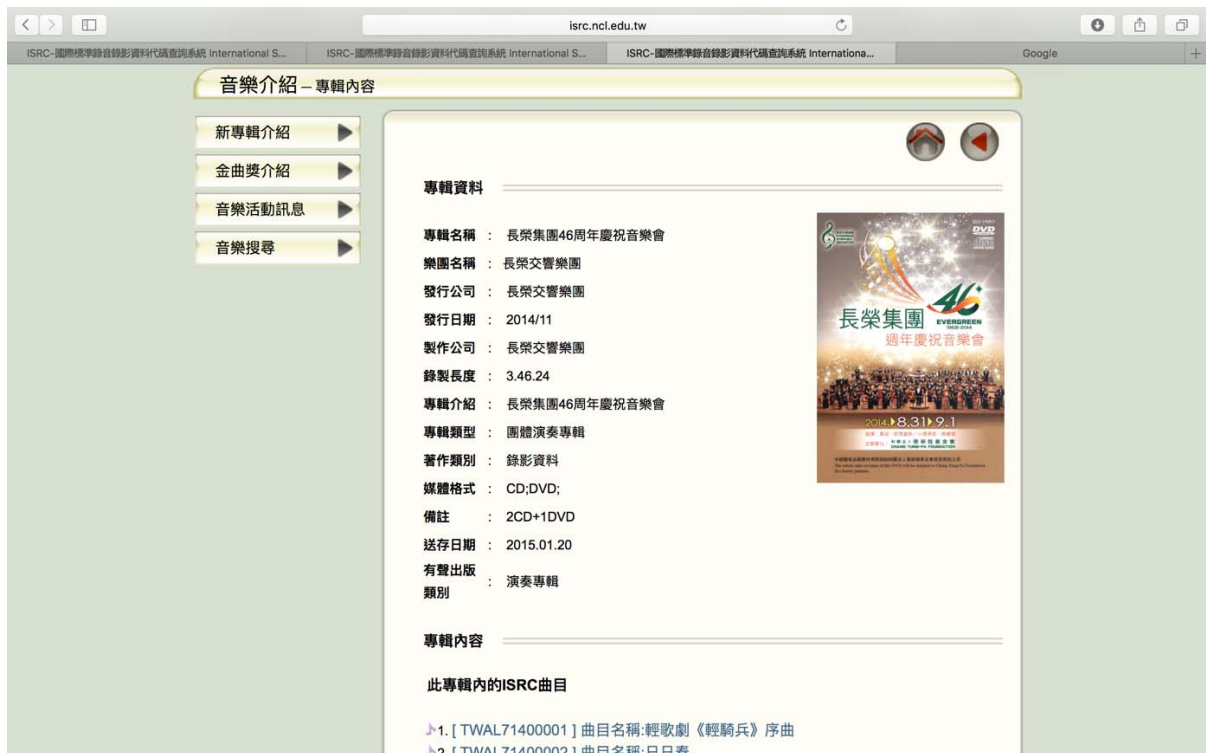


圖4-5 「長榮集團46週年慶祝音樂會」之「發行公司」及「專輯類型」

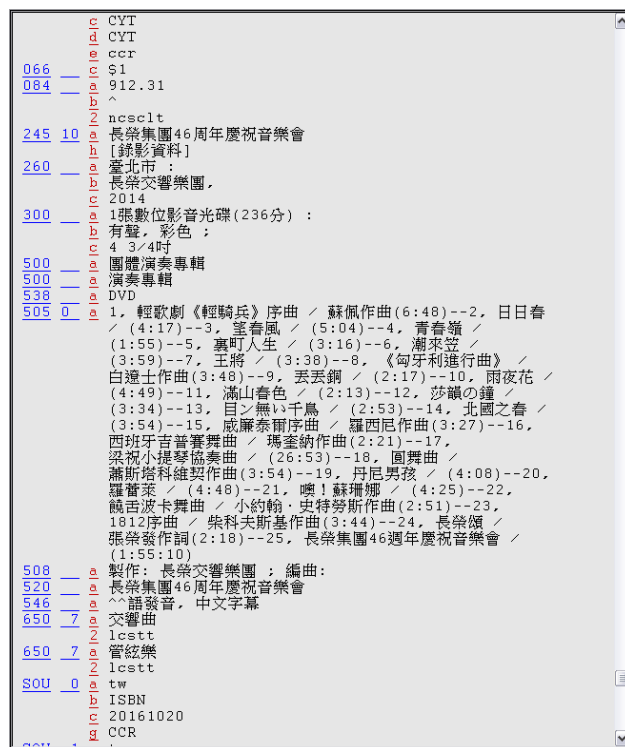


圖4-6 欄位260\$b「交響」及500\$a「團體演奏專輯」有助於自動分類

在ISRC國際標準錄音錄影資料代碼查詢系統中查詢布袋戲「霹靂英雄音樂精選五十雷霆劍海錄原聲帶」，其「發行公司」一欄之內容「霹靂國際多媒體」即代表專門出版布袋戲錄音錄影資料的出版社（圖4-7及4-8），因此有助於此一資料之主題分析。



圖4-7 「霹靂英雄音樂精選五十雷霆劍海錄原聲帶」之「發行公司」

084	a	915.18
	^	
	ncsclt	
245	10	【霹靂英雄音樂精選四十九】霹靂創海錄原聲帶
	b	[錄音資料] /
	c	浩旭等作曲
260	a	臺北市 :
	b	霹靂國際多媒體,
	c	2014
300	a	1張唱片(103分) :
	b	數位, 立體聲 ;
	c	4 3/4吋
500	a	多人錄音專輯
500	a	原聲帶
500	a	雷射唱片
505	0	1, 戰歌狂響 / 浩旭作曲(3:04)--2, 叱吒風雲 /
		丁天牧作曲(2:50)--3, 癡世白蓮 /
		風采輪作曲(2:56)--4, 秋楓暮霞惋紅曲 /
		浩旭作曲(3:02)--5, 道真雙秀 / 孫敬凡作曲(4:30)--6,
		森獄閻王 / 孫敬凡作曲(4:01)--7, 漂鈴殺響 /
		風采輪作曲(3:29)--8, 藍山姑娘 /
		黃建泰作曲(3:36)--9, 垢浪明珠 /
		孫敬凡作曲(4:15)--10, 翠環山玉波池 /
		風采輪作曲(3:34)--11, 風塵鑄曲 / 風采輪作曲 ;
		風采輪演唱(3:40)--12, 浩氣破天 /
		翁偉瀚作曲(2:56)--13, 天驕牧神 /
		丁天牧作曲(3:36)--14, 恩怨情仇江湖路 /
		丁天牧作曲(3:47)--15, 日落煙都 /
		劉昱昀作曲(2:55)--16, 蛻變黑后 /
		丁天牧作曲(3:23)--17, 玄膽太子 /
		風采輪作曲(3:39)--18, 白雪紅梅 /
		風采輪作曲(3:13)--19, 劍舞紅霞 /
		黃建泰作曲(3:17)--20, 一殉王魂 /
		黃建泰作曲(3:21)--21, 佛謎心懷鉛 /
		風采輪作曲(3:53)--22, 崩天裂地 / 荒山亮作曲 ;
		荒山亮作詞 ; 荒山亮演唱(3:52)--23, 心罪 / 孫敬凡作曲
		; 廖明治作詞 ; 阿傑演唱(5:06)--24, 天下獨行
		蔡佳瑩作曲 ; 廖明治作詞 ; 蔡佳瑩演唱(4:35)--25,
		癡情難忘 / 風采輪作曲 ; 廖明治作詞 ; 風采輪,
		荒山亮演唱(4:19)
650	7	掌中戲
	lcstt	
650	7	劇樂
	lcstt	
650	7	布袋戲

圖4-8 欄位260\$b「霹靂國際多媒體」有助於自動主題分析

ISRC國際標準錄音錄影資料代碼查詢系統查詢「幸福快遞電影原聲帶」，其「專輯名稱」一欄之內容「電影原聲帶」即代表此專輯的主題性，此外，「有聲出版類別」一欄之內容「電影原聲帶」，也是有助於此一資料之主題分析，據此而自動建議類號為「913.62」或「915.7」、主題詞為「電影歌曲」或「電影音樂」（圖4-9及4-10）。



圖4-9 「幸福快遞電影原聲帶」之「專輯名稱」與「有聲出版類別」



圖4-10 依據正題名及附註項之「電影原聲帶」進行自動分類

轉自ISRC國際標準錄音錄影資料代碼查詢系統的「慈悲自在」，其「發行公司」一欄之內容「原動力」即代表專門出版佛教錄音錄影資料的出版社（圖4-11及4-12），此外，「有聲出版類別」一欄之內容「宗教音樂」，則可確定建議類號為「224.53」或「913.48」、主題詞為「佛教音樂」。



圖4-11 「慈悲自在」之「發行公司」及「有聲出版類別」

二、中文圖書自動分類研究成果

目前仍進行之中文圖書自動分類研究項目中，已有以下之研究成果：

（一）題名項目之內容詞彙經自動斷詞可取得相關主題分析

進行中文圖書自動分類研究時，與資深編目員進行主題分析經驗分享中，同時發現題名項目內含詞彙為編目員進行主題分析之初步依據。編目員第一眼均以題名之詞彙來判斷該書的主題，分析題名是決定主題範疇的關鍵工作。

目前研究已將本館已編及未編中文圖書書目資料之題名（欄位245\$a）利用斷詞工具進行自動斷詞，再將未編書目資料題名斷詞結果比對已編圖書資料題名斷詞結果，比對結果相關性越高之書目的類號及主題詞，即可作為未編資料建議的類號及主題詞。目前比對所得結果之前10項高度相關之書目中，已出現可採用之類號及主題詞，但仍需參酌書目之其他關鍵項目，再度提高主題分析的精確度。

（二）出版者之出版領域存在特定性

訪談資深編目員發現出版者在其學科領域的判斷有其重要性，進而由研究團隊透過電腦進行出版者與主題關聯性的統計分析，發現政府出版品依政府機關之屬性有其特定之出版領域，一般出版者除了綜合性出版商外，多數均有其專精出版領域，將更大量之完編書目進行統計可以取得更精確之數據，增加出版者來判定主題領域之精準度，但仍無法單獨以出版者之領域判

斷該書之主題，而需搭配題名等關鍵項目。在進行自動分類時，若排序高度相關之建議類號及主題詞的書目為該出版社出版領域統計分析資料中從未出現之範圍，則可降低該書目之權重。

（三）著者之專業領域出現主題集中現象

研究分享過程發現著者有其專長之研究與著作領域，因此對於主題辨識有所助益，進而由研究團隊透過電腦進行著者與主題關聯性的統計分析，發現出版品數量最多者為某一政府機關，這些出版品之出版主題即為該機關之業務範疇。個人著者方面也發現多數之個人著者有其研究與著作領域，也就是在這些著者之著作的類號及主題詞有集中性，所以可由著者名稱來提高主題分析之精準度。雖然著者名稱對於主題判斷有所幫助，但是使用著者名稱進行主題分析仍有其限制，也就是著者名稱會遇到例如機關名稱的改名、合併，以及著者本名、筆名、俗名法號、同名不同人、一人多名……等之情形，遇此情況會影響主題分析的精確度。

（四）組合各項關鍵項目提高精確率

研究中發現當組合題名、著者，出版者等三項關鍵項目進行主題分析時，較能取得精確度更高的類號與主題詞，運用電腦運算，隨著資料測試之成果，可調整各項間之比重，取得最佳之主題分析狀態。

綜整音樂性錄音錄影資料之自動分類成果及目前同時進行中之中文圖書自動主題分析研究，此兩項實務導向的研究，發現由人工主題分析作業轉

換為自動主題分析過程中，可規劃出以下之開發模型（圖4-13）。

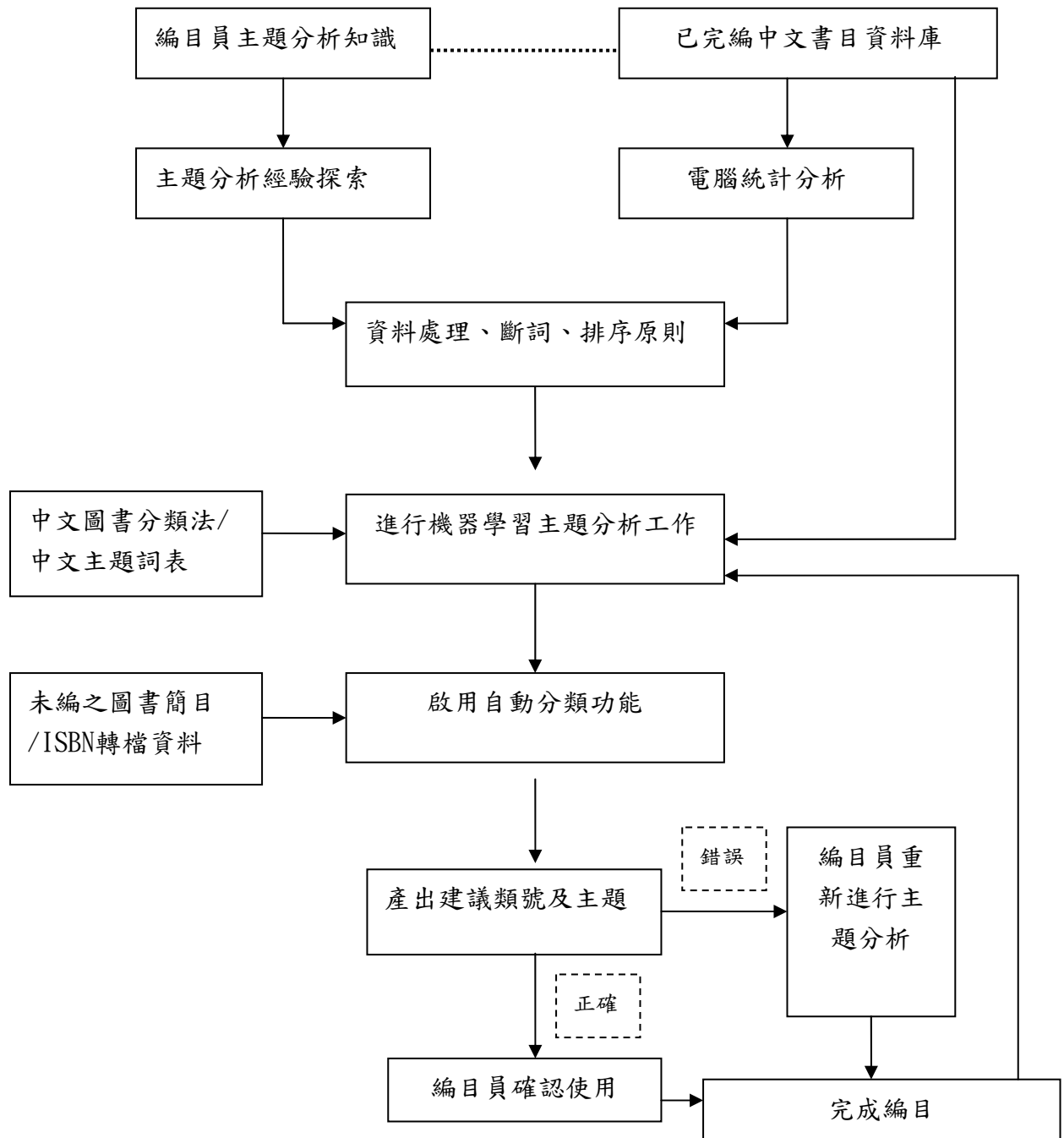


圖4-13 自動主題分析開發模型圖

編目員的主題分析工作除了來自身經驗外，同時也參考過往已完編的書目資料，而中文書目資料庫內的書目紀錄之主題分析也是編目員經驗的累積。編目員主題分析知識與已完編中文書目資料庫二者對於主題分析工作是相輔相成，所以探索其中之脈絡並對其進行分析，以便將人類的心智運作模式提供予電腦學習。

透過編目員主題分析知識與經驗的探索，以及已完編之中文書目資料的電腦統計分析，可以取得資料處理、斷詞及資料排序原則，經由程式對比已完編之中文書目資料提供機器學習進行主題分析工作，並且同時提供中文圖書分類法及中文主題詞表供機器比對其類名及主題詞，以便提高比對的精準度。另外對於中文書目資料庫中從未給予之分類號及主題詞，機器無法由已完編中文書目資料庫進行學習，因此中文圖書分類法及中文主題詞表即成為這類資源主題分析唯一比對的參考值。

藉由上述各學習來源建立自動主題分析功能，並將未編之圖書簡目（例如：題名、著者、出版者、集叢名）及ISBN轉檔資料批次執行自動主題分析，進而產出建議之類號及主題詞，若自動建議之類號及主題詞符合該筆書目資料使用，則編目員確認後即完成主題編目，若為不符合該筆書目使用之建議類號及主題詞，則由編目員重新進行主題分析。之後將藉由此兩種途徑完成之編目資料，再次提供機器學習，並針對建議錯誤之類號及主題詞，則回饋正確之類號及主題詞予機器學習作為正確之主題分析機制。

第五章、研究結論與建議

圖書館學中的資訊組織範疇，多年來已發展出相當完備的理論與實務，以便各圖書館據以進行資訊組織與主題分析的工作，無論是資訊組織或主題分析，都是知識密集且高階的知識組織與管理工作。

早期圖書館仰賴訓練有素的編目人員進行資訊組織的主題分析工作，然而出版量的大增及資訊科技的進步，以往傳統的人工作業已不足以應付與日俱增的出版品，因此期待藉由機器或電腦協助主題分析工作的研究漸次展開。

近20年來，有些研究成果已經達到一定程度的效果，甚至接近人工主題分析的水準（曾元顯，民91年12月，頁7），但是自動化方式的運用，並不是用來取代人力，而是協助或輔助現有人力，以達成高效率、高品質的作業成果。

本研究之結論與建議如下：

一、研究結論

經由前述研究結果得知自動主題分析工作確實可將人工的主題分析經驗進行轉換，以往認知必須透過人工作業的主題分析工作，可透過電腦的統計、分析與歸納後，學習而產生與人為經驗相似的作業結果，研究中發現確實可獲得高度精確率。

可取得書目資料的欄位內容越多，則其主題分析之結果會越精確，也就是書目紀錄內之資料著錄欄位同時擷取題名、著者、出版者等欄位，則其主題分析結果會比僅擷取題名欄位之內容的主題分析結果更為精確。以特定資料類型而言，可發現其與圖書資料間各有其特有判斷主題項目的欄位。

關鍵詞擷取方面，僅適用於題名或集叢名內文字進行斷詞和擷取與統計及分析，著者及出版者名稱不需再行斷詞，而係整欄擷取。同名不同人及筆名、本名等情況，以及機關團體和出版社的改名或組織變動，若未進行權威控制，都有可能因此影響自動主題分析與歸類的結果。

當組合題名、著者、出版者進行自動主題分析時，應考量各項之權重。例如：題名含有「佛教」及出版者為佛教團體、以及題名含有「佛教」及「基督教」且出版者雖然為佛教團體時，若題名與出版者兩項的權重不同時，其自動主題分析結果可能有所不同。

過去題名項目是編目員人工進行主題分析之首要判斷項目，再以著者與出版者輔助確認其主題範圍，有時甚至不需要運用著者與出版者兩者來判斷主題。在思考如何透過電腦運作找出規則的同時，編目員打破傳統思維，在運用音樂性錄音錄影資料小規模實驗下發現，以電腦運算時應是優先歸納統計出版者與著者之領域，再由題名之斷詞去取得更精確的主題分析。由此可知，在大規模書目資料進行運算時，電腦可以統計分析出過去編目員由分類單一書目無法看出之整體規則，目前我們藉此看到的是出版者出版領域之特

定性與著者專業領域之主題集中現象，大量書目數據的統計分析與運算，讓過去用以輔助主題分析之關鍵項目出版者與著者，在自動分類電腦運作之下卻是判斷主題分析之優先項目，再透過題名的自動斷詞，比對出更精確之主題分析資料，透過編目員主題分析經驗的整理、大量書目統計分析及傳統觀念的突破，確實已讓書目自動分類工作向前邁進了一大步。

ISRC系統進行轉出之音樂性錄音錄影資料139筆進行測試，結果達成自動分類者計95筆，可處理率達68%以上。由產出自動分類的95筆中，有1筆為跨條件而給予兩個類號及主題詞，達成自動分類的95筆書目資料中，由兩位15年以上編目經驗之資深編目員逐一比對，其中94筆經自動分類後書目之分類號及主題詞完全正確，正確率達98%以上。另有1筆為有兩組分類號及主題詞，其中一組為正確，但需人工判斷，再取其正確分類，雖然未精準提供一組正確主題分析項目，但確實是達成指引館員及縮短主題分析時間之成效。

二、建議

就本研研究所得之結論提出以下建議：

(一)以詞庫及分類表類目與主題詞表為輔助斷詞工具

有關斷詞方面，自動主題分析將關鍵項目（例如：題名、集叢名……等）內之文字內容等字串，分割為兩個字、三個字、四個字……，以分割後的詞彙探索出現的頻率，藉此找出與主題相關之特徵詞彙的方式，可能會篩選出與該資料主題不相關的特徵詞彙，因此建議應引進線上詞庫（例如：字典、辭

典、百科全書)、分類表類目、主題詞表等作為斷詞工具，或者選擇出特徵詞彙後，再與詞庫等斷詞工具的詞彙進行比對，以避免不相關的主題因文字前後的相連，而產生不屬於該資料之主題的關鍵詞，如此方能提升主題正確的斷詞比率，以及避免資料之主題誤判的機率。

(二)擴大題名範圍至其他相關題名項目

本研究中題名為主題分析之關鍵項目，但是依據編目員的經驗，副題名有時比正題名更具有主題性，乃因副題名是用來解釋正題名，因此副題名的主題資訊通常也會比正題名之主題更為明確。在編目員的經驗中發現，現在的出版品為了吸引讀者，往往正題名會取得相對華麗，這是為了有助行銷，造成正題名與主題的關係逐漸淡化，真正與主題相關之資訊，往往會是在副題名中呈現。

而待編資源若為叢書，當以整套叢書進行分編時，其正題名即為叢書名，編目員的經驗會以叢書名做為主題分析時該叢書之首要主題考量，而各冊的題名則作為次要參考。待編資源若分散單獨進行分編時，其正題名即為各冊之題名，編目員除了以各冊的題名作為主題分析的考量外，有經驗的編目員會參考叢書名以協助主題分析工作。所以於主題分析時也會增加叢書名作為重要項目。

(三)運用外部線上資源加強了解資料內容

外部資源方面，現在網路上已有很多有關各出版品的相關資訊或簡介資

料（例如：目次、摘要、書評、內容試閱），因此可以透過程式撰寫及執行來引入外部資料，以協助各待編出版品之主題分析工作。

(四)引用著者權威控制資料辨別著者異同

目前圖書館已針對同名不同人及筆名、本名等情況進行權威控制，並編製龐大的權威紀錄，這些權威紀錄對於同名不同人及筆名、本名等情況，做了不同的區別，以方便館員或讀者的辨識及有助於資料的查詢。這些權威控制工作的成果—權威紀錄，正可藉此協助資源主題分析之機器學習，方便整理出同名不同學科或研究領域的著者，進而有助於依著者名稱進行主題分析。

(五)依編目員主題分析熟悉度分配工作

有關音樂性錄音錄影資料自動分類之實證研究，以ISRC系統進行轉出之音樂性錄音錄影資料139筆進行測試，結果達成自動分類者計95筆，另有44筆未給予類號及主題詞。達成自動分類者，由於已透過自動分類給予類號及主題詞，因此這類資料可由新進編目員接續進行處理，並透過此方式讓新進編目員能更有效率且快速熟悉主題分析作業模式，至於未能達成自動分類之書目資料，由於尚無類號及主題詞，因此建議這類資料請由資深編目員進行主題分析，並且給予類號及主題詞。

(六)加入計算主題分類工具書內詞彙之比重

對於中文書目資料庫中從未給予之類號及主題詞，機器無法由已完編中

文書目資料庫進行學習，僅能對照參考中文圖書分類法及中文主題詞表，將此所建議之類號及主題詞，經由編目員確認後，再提供機器學習，後續有有關該主題之資源進行自動主題分析時，因該主題類型可供機器學習之資料量過少，建議應結合中文圖書分類法及中文主題詞表，並提高此二項工具中的類名及主題詞比對的比重，方能建議出較為精確的類號及主題詞。

(七)找出自動主題分析關鍵項目之最佳權重

由中文圖書自動分類研究中發現，自動分類在題名、著者及出版者三項關鍵項目中，缺一不可，在翻轉了人工主題分析優先以題名為判斷依據的觀念後，大量書目之統計分析數據，讓著者與出版者對各主題有清楚之脈絡，編目員需在自動分類所產生之主題分析資料中，繼續發掘此三項關鍵項目對主題分析之最佳權重，透過調整權重與資料篩選的優先順序達成提高自動主題分析之精確度。

(八)容許出現一組以上之類號及主題詞

由錄音錄影資料的自動分類作為測試，若符合自動分類原則而使得自動建議的結果包括多組類號及主題詞，如此的建議提供予編目員，以便藉此指引編目員該資料也可能可採用另一個類號及主題詞。例如：當自動主題分析所建議之多組類號及主題詞，其所得百分比差異不超過5%時，建議兩組類號及主題詞皆提供予編目員參考。

(九)隨時檢視自動分類結果並調整機器學習機制

本次錄音錄影資料測試結果發現有些資料經由自動分類後，未能自動建議類號及主題詞，究其原因係待處理資料有落於本次歸類原則之外者，例如：歌仔戲、京劇及兒童故事，皆可針對該主題找出歸類原則，並對各領域之表演者或演唱者（包括個人及團體）建立表演領域列表，即可解決像「五月天」知名團體因該筆ISRC資料無符合自動分類原則而無法被歸類問題。藉由分析未歸類之資料，找出未來應加入歸類原則之條件，同時編目員有必要定期檢視自動分類的結果，由此我們建議自動主題分析也須由電腦自動統計分析這些建議錯誤而經編目員修改之資料，並提供編目員參考及調整機器學習機制。

未來期望能將更複雜多元樣貌的待編書目應用自動分類系統進行自動主題分析工作，唯有透過不斷的測試改善，將讓電腦分析學習各種類型書目資料，才是自動分類系統更大的挑戰，但也是能強化自動分類系統的機會，藉此並讓圖書館書目主題分析更加一致、精確，進而解決資訊組織工作長期人力經費不足之問題。

附錄

一、音樂性錄音錄影資料自動分類原則

音樂性錄音錄影資料自動分類原則

條件	類號	主題詞	備註
500\$a國語流行專輯	913.602	流行歌曲	
500\$a臺語流行專輯	913.6033	流行歌曲 臺語	
500\$a客語流行專輯 或 245含有("客家" 或 "客語") 且 "歌曲"	913.6033	流行歌曲 客語	
245含有 "客家" 且 "山歌" 或 260\$b含有 "客家"	913.533	客家民謠	
500\$a電視原聲帶 或 245\$a、\$b含有 "電視原聲帶"	913.63 915.6	電視歌曲 電視音樂	
500\$a電影原聲帶 或 245\$a、\$b含有 "電影原聲帶"	913.62 915.7	電影歌曲 電影音樂	
500\$a宗教音樂 且 (245含有 "佛教" 或 "梵唄" 或 "梵唱"或 260\$b為佛教出版團體(參看附表))	224.53 913.48	梵唄 佛教音樂 佛教 歌曲	
500\$a宗教音樂 且 (245含有 "基督教" 或 "頌歌" 或 "靈歌" 或 "讚美詩" 或 "福音" 或 260\$b為基督教出版團體(參看附表))	244.4 913.418	基督教 歌曲 宗教音樂	
260\$b為布袋戲出版團體(參看附表)	915.18	掌中戲 劇樂 布袋戲	
500\$a演奏專輯 且 500\$a團體演奏專輯 且 (245或260\$b含有 "室內樂" 或 "重奏")	912.2	室內樂	含次序性
500\$a演奏專輯 且 500\$a團體演奏專輯 且 (245或260\$b含有 "交響")	912.31	交響曲 管絃樂	
500\$a演奏專輯 且 500\$a團體演奏專輯 且 (245或260\$b含有 "管絃")	912.3	管絃樂	
500\$a演奏專輯 且 500\$a團體演奏專輯 且 (245或260\$b含有 "弦樂")	912.4	絃樂合奏	

500\$a演奏專輯 且 500\$a團體演奏專輯 且 (245或260\$b含有 "管樂")	912.5	管樂合奏	
500\$a演奏專輯 且 500\$a團體演奏專輯 且 (245或260\$b含有 "擊樂")	912.6	敲擊樂	
500\$a演奏專輯 且 500\$a團體演奏專輯 且 (245或260\$b含有 "國樂")	919.8	國樂	
500\$a演奏專輯 且 500\$a團體演奏專輯 且 (245或260\$b含有 "絲竹樂")	919.82	絲竹樂	
500\$a演奏專輯 且 245含有 "小提琴"	916.1	小提琴音樂	
500\$a演奏專輯 且 245含有 "大提琴"	916.3	大提琴音樂	
500\$a演奏專輯 且 245含有 "鋼琴奏鳴曲"	917.1022	鋼琴 奏鳴曲	含次序性
500\$a演奏專輯 且 245含有 "鋼琴協奏曲"	917.1026	鋼琴協奏曲	
500\$a演奏專輯 且 245含有 "鋼琴"	917.1	鋼琴	
500\$a演奏專輯 且 245含有 "南管"	919.881	南管音樂	
500\$a演奏專輯 且 245含有 "北管"	919.882	北管音樂	

備註：084^^\$a類號\$b^\$2ncsclt

650^7\$a主題詞\$2lcstt

資料來源：國家圖書館館藏發展及書目管理組（民105）。音樂性錄音錄影資料自動分類原則。未出版。

二、音樂性錄音錄影資料宗教及布袋戲出版團體表

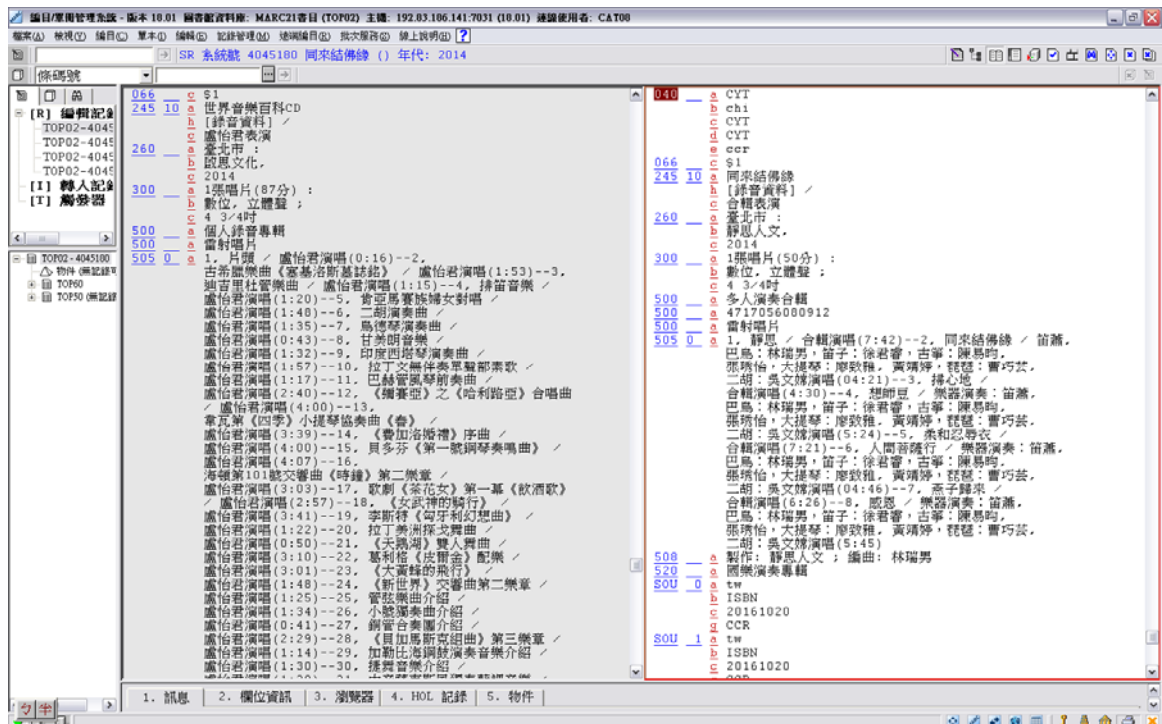
宗教及布袋戲出版團體表

佛 教 出 版 社	基 督 教 出 版 社	布 袋 戲 出 版 社
聖韻音樂	士林靈糧堂	霹靂國際多媒體
原動力	天韻有聲出版社	無非文化
佛光山文教基金會	台北靈糧堂	義興閣掌中劇團
中華國際大悲咒水功德會	校園	
梵音創作表演工作坊	福音證主協會	
臺灣內觀禪修基金會	基督教福音證主協會	
國際法身禪修中心	天聲傳播	
法爾禪修中心	基督教天聲傳播協會	
佛緣國際出版社	台北真道教會	
舊譯講修林	基督教台北真道教會	
香光莊嚴雜誌社	撒拉弗音樂事工	
彌勒大道總會基金會	基督教希奇全人復興發展協會	
祈善行	希奇全人復興發展協會	
生命基金會	真理天地國度企業號	
中華印經協會	基督教中國主日學協會	
達摩	雅歌	
慈濟	宇宙光	
如是我聞	基督教宇宙光全人關懷機構	
佛光文化	臺灣基督長老教會傳播基金會	
佛典	上智	
妙蓮華	聖保祿孝女會附設上智文化事業	
全佛文化	學園傳道會	
佛文化	諦聽文化	
光正萬教殿	和散那	
玄揚禪坊	道聲	
一貫道崇正基金會	佳音廣播電臺	
般若佛韻	異象工場	
妙地風光	臺灣福音書房	
愛華音樂	房角石音樂工作室	
青海無上師世界	蒲公英希望基金會	
諦聽文化	讚美之泉	

慈光圖書館	大安基督長老教會	
伽耶山基金會	臺灣利未支派事奉協會	
發一天元長聖道場	光啟文化	
富仕音樂	同心國際傳播	
諦聽文化	救世傳播協會	
愛華	得音	
香光書鄉	聚藝音樂	
香光尼眾佛學院	磐石製作	
貴族唱片	台灣利未支派	
彌勒皇教文化	伊甸社會福利基金會	
人乘	火把行道會	
般若海		
法界出版社		
龍閣文化		
財團法人玄奘文教基金會		
靜思人文		
梵音創作表演工作坊		
禪天下		
功德山		
普音文化公司		

資料來源：國家圖書館館藏發展及書目管理組（民105）。音樂性錄音錄影資料自動分類原則。未出版。

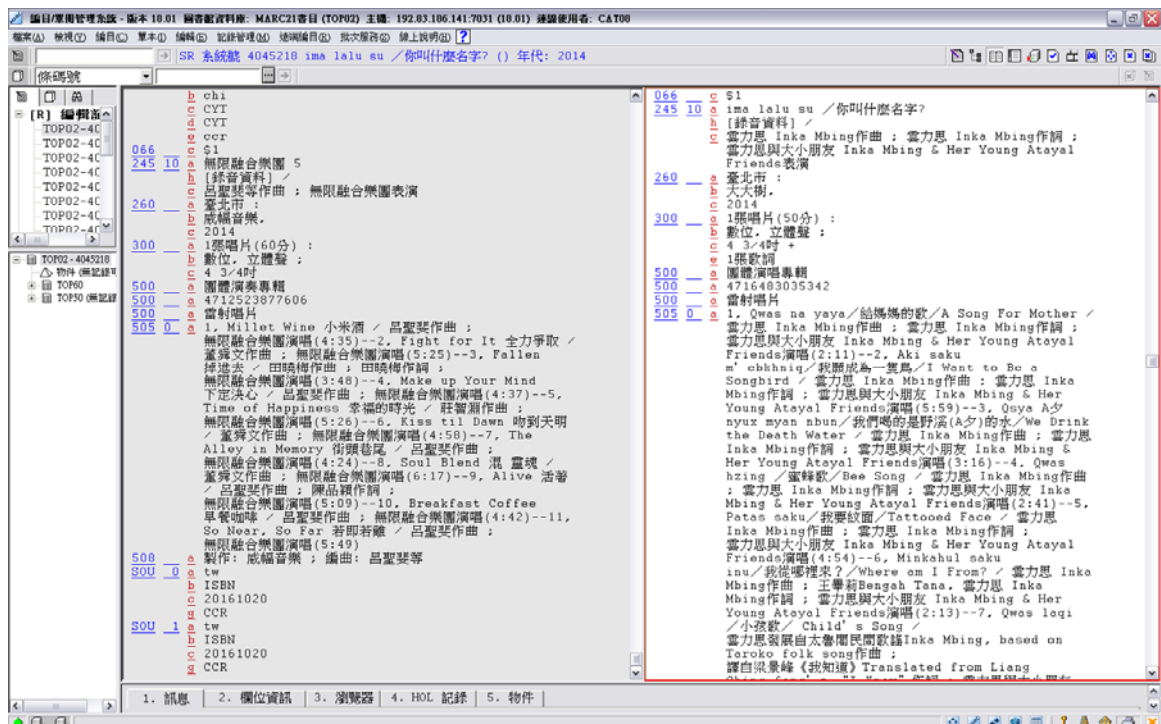
三、音樂性錄音錄影資料無法完成分類書目資料

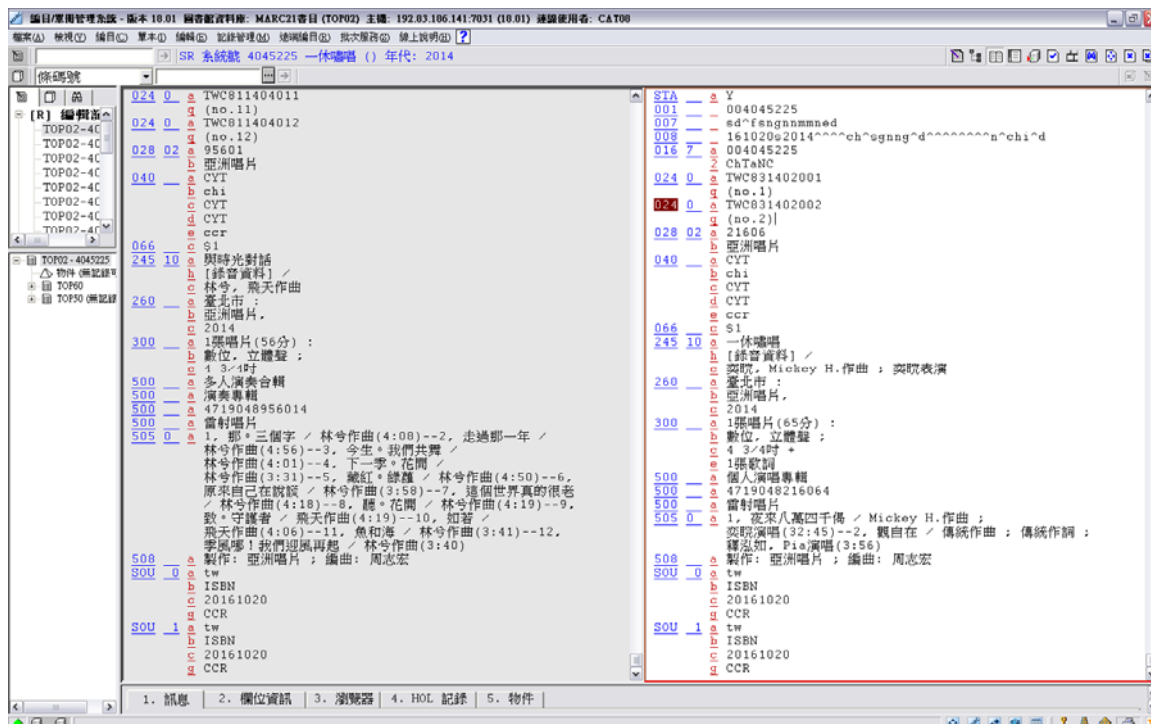


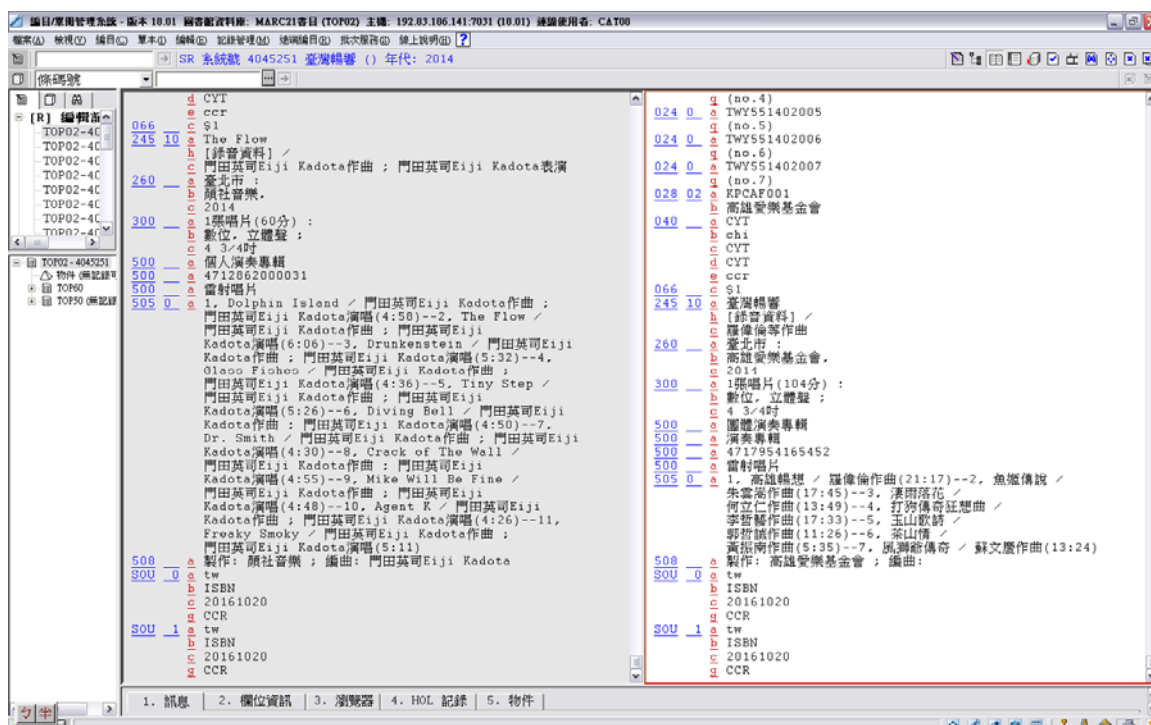
















四、「協助編目工作的資訊技術研究計畫」會議討論內容摘要

討論日期：民國105年6月13日、7月5,12,19日、8月2,9,23,30日、9月13日下午1時30分至4時整

討論地點：國家圖書館178會議室

出席人員：黃乾綱、李家名、黃仁壕、李宜容、呂寶桂、牛惠曼、陳慧文、馬千惠、吳海威

重點紀錄：

(一) 目前約210萬筆中文圖書書目，臺大已將之轉為MODS格式，唯自訂欄位OWN及029於轉為MODS，仍以OWN及029標示。

(二) 書目資料：由於本館書目資料前後採用之規範有所不同，因此為利機器學習與類號和主題詞建議的一致性，敬請以下列規範採用之時間點進行區分：

1.<<中文圖書分類法2007年版>>：自2008年1月1起採用。

2.<<中文主題詞表>> (即代碼為lcstt)：自2007年7月1日起採用。

若為<<中文圖書標題表>> (即代碼為csht) 則為舊有的 (資料網址為http://catweb.ncl.edu.tw/portal_e3_cnt.php?button_num=e3&folder_id=72)。

在此年限之前書目資料所用為舊版類號及標題詞。

(三) 依現有編目人員作業流程可將書目來源資料區分為：

1.CIP書目 (已有類號及主題詞)：欄位084、240、245、260\$b、490、6XX內容可納入機器學習。有些書目可能包括欄位505，其內容也可納入機器學習。

2.非CIP轉入書目 (無類號及主題詞，包括ISBN書目及採訪簡編資料)：僅欄位245及260\$b可做為主題分析預測之參考欄位。

(四) 判別來源書目的方式如下：

1. CIP書目：欄位029\$aCIP。

2. 非CIP轉入書目：欄位029\$a無CIP。

(五) 非CIP轉入書目（無類號及主題詞，包括ISBN書目及採訪簡編資料）：僅欄位245及260\$b可做為主題分析預測之參考欄位，自明年度起要求採訪簡編資料鍵入欄位490，因此自明（106）年度起，採訪簡編資料之欄位490可做為主題分析預測之參考欄位。

(六) 資料擷取分析前應留意欄位029之指標1是否皆為「1」。

(七) 可納入機器學習的書目資料判別方式如下：

1. 已總校書目：有欄位OWN。

(1) 欄位OWN\$b內容為20070701以後(<<中文主題詞表>> (即代碼為lcstt)) 【相關參考欄位為欄位6XX，並且可以\$2內容是否為lcstt作為判別】

(2) 欄位OWN\$b內容為20080101以後(<<中文圖書分類法2007年版>>) 【相關參考欄位為欄位084\$a】

2. CIP書目：有欄位029\$aCIP。

(八) 機器學習前應排除學位論文、篇目、子目等書目紀錄。學位論文書目紀錄的排除方式為008/24-27=m者；篇目及子目等書目紀錄的排除方式為Leader/07=a且含有欄位490及830者。

(九) 請臺大比對<<中文主題詞表>>及<<中文圖書標題表>>，以便產出同義詞表之不同用詞。

(十) 以分類法之類號及主題詞表之詞目比對210萬筆中文圖書書目，若有書目內未曾出現的主題詞，則由家名教導千惠及海威學習如何操作軟體工具來產生，慧文及惠曼旁聽，靜芬主任也表示有興趣參與學習。

(十一) 書目內不曾用到的類號或主題詞，對於機器學習是會有影響的，因此仍然建議CIP的書目與舊版類號與標題皆納入機器學習。

(十二) 臺大實驗性開發之主題分析預測建議工具與ALEPH系統書目資料介接方式：即時或批次。

1. 第一法線上：編修過程呼叫外部程式、匯出資料再行匯入、界面可以介入，採用Windows介面。

2. 第二法透過API：利用查詢方式、編完資料、再透過API轉回ALEPH。

資料出去→加值（存檔加入類號及主題詞）→再編修

流程若採線上，則立刻給號，即1筆出1筆進。

無論線上或批次皆為ISO出、ISO進。

(十三) ALEPH系統端

1. 經由修訂程序去call外部處理程序，基本上建議方式仍會透過API，資料自ALEPH系統匯出至臺大實驗性開發之主題分析預測建議工具需先行「存檔」，並加好類號或主題詞後，再將資料回寫至ALEPH系統，方能確保書目原有資料及新加的類號與主題詞皆會回寫至ALEPH系統。

2. 類號及主題詞等回存至ALEPH系統書目紀錄時，採「附加」方式，而非「覆蓋」方式。

3. ALEPH系統使用者帳號可設定允許執行update權限。

(十四) 編目流程數位化。

(十五) 資料篩選條件：無OWN或無029\$aCIP且系統號大於前一批或005異動時間大於某日期時間（註：編目者清單所記錄之編目日期與人員代碼由於log檔過大不適宜索引與查詢，通常多數系統也非如此設計為索引或查詢目的）。

(十六) 書目紀錄上標示是否已經由臺大實驗性開發之主題分析預測建議

工具給予類號及主題詞，可用自訂欄位標示（例如：使用欄位69X），惟相關客製程式需予調整。

（十七）所用為舊版類號及標題詞之書目資料是否回溯增加新版類號及主題詞？若增加的話，館藏組無人力逐一審核，但又怕開放的書目紀錄內有不適當的類號或主題詞，會讓讀者對於書目品質有所疑慮。

【因此要不要加，是否先不加入ALEPH系統書目紀錄內，但可納入機器學習，並僅加於提供本項研究之書目紀錄內，且主題詞加於欄位653或69X，類號考慮是否加於084但不加\$2，本項尚待考量。】

（十八）類號及主題詞皆會切年代，並透過書名、著者、主題相關欄位內容的分析給予類號及主題詞。不同時期的資料，考慮用程式調整成新時期的資料，讓資料一致。

（十九）回溯資料加類號及主題詞，主題詞可考慮用欄位69X。因此，主題分析預測建議工具欄位處理：主題詞部分

學習欄位	欄位定義	建議欄位
600	標題附加款目-個人名稱	695
610	標題附加款目-團體名稱	696
611	標題附加款目-會議名稱	697
630	標題附加款目-劃一題名	690
650	標題附加款目-主題用語	691
651	標題附加款目-地理名稱	693

（二十）若機器學習無法學習到欄位，則預測建議的主題詞入欄位690。

（二十一）建議的類號及主題詞應給予運算後的分數供參；建議的類號及主題詞最後為編目員所選用時，系統需有記錄。

(二十二) 動態主題詞表：

1. 考量書目紀錄內未曾出現過的類號及主題詞。
2. 新產生已總校之書目如何重新納入機器學習?全部作業重新run過一次。
3. 主題詞更新時，是否書目內的主題詞一起變動？

參考文獻

- Cohen, W. W., & Singer, Yoram. (1996). Context-sensitive learning methods for text categorization. *Proceedings of the 19th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, 1996, 307-315.
- Hoffman, S. L., & Hatch, D. (2002). *Web world of authority control*. Retrieved December 8, 2002, from <http://www.lib.byu.edu/dept/catalog/authority/>
- Introduction to the ACM Computing Classification System [1998 version]*. Retrieved October 15, 1999, from <http://www.acm.org/class/1998/ccs98-intro.html>
- Joachims, T. (2001). 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, 128-136.
- Kwok, K. I. (1975). The use of title and cited titles as document representation for automatic classification. *Journal of Information and Management*, 11, 201-206.
- Olson, Hope A. (1998). Authority control in a global environment. In Kathleen de la Pema McCook, ed., *Global reach/Local touch* (pp. 210-218). Chicago: American Library Association.
- Sebastian, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys (CSUR)*, 34(1), 1-47.
- 王泰裕、江蕙民（民97）。模糊支援向量機應用於多類別的文件分類。成大

研發快訊，3:6，1-2。

王稔志、張俊盛（民90）。適應性文件分類系統。第十四屆計算語言學研討會論文集，99-121。

杜海倫（民88）。以標題進行新聞自動分類（未出版之碩士論文）國立清華大學，新竹市。

林昕潔（民95）。以SVM與詮釋資料設計書籍分類系統（未出版之碩士論文）。國立交通大學，新竹市。

林政男（2004）。以共現語詞為基礎的特徵選取在文件自動分類上之研究（未出版之碩士論文）。銘傳大學，臺北市。

林頌華（民87）。新聞標題自動分類（未出版之碩士論文）。國立清華大學，新竹市。

林頌堅（民87）。自動化文件分類在資訊服務上的應用。資訊傳播與圖書館學季刊，5:2，87-102。

林頌堅（民103a）。以主題模型方法為基礎的資訊計量學領域研究主題分析。教育資料與圖書館學，51:4，499-523。

林頌堅（民103b）。資訊科學期刊的主題分布與多樣性研究。圖書資訊學研究，9:1，171-200。

林頌堅（民104夏季）。利用文字內容主題特徵與機器學習方法探討MIS相關期刊在ISI資料庫的主題分類。教育資料與圖書館學，52：3，269-298。

柯淑津、陳振南（民88）。階層式文件自動分類之特徵選取研究。第十二屆

計算語言學研討會論文集，137-149。

徐軍、丁宇新、王曉龍（2007年11月）。使用機器學習方法進行新聞的情感

自動分類。中文信息學報，21(6), 95-100。

許雅芬（2002）。新聞文件自動分類之研究（未出版之碩士論文）。東吳大

學，臺北市。

陳彥呈、蔣榮先（民90）。基於階層式類神經網路之自動新聞文件分類方法。

第十四屆計算語言學研討會論文集，89-97。

陳俊凱（民84）。利用類神經網路作文件自動分類之研究（未出版之碩士論

文）。淡江大學，新北市。

陳智偉（民87）。文件分類的方法與分析（未出版之碩士論文）。國立清華

大學，新竹市。

陳信源、葉鎮源、林昕潔、黃明居、柯皓仁、楊維邦（民98年6月）。結合

支援向量機與詮釋資料之圖書自動分類法。資訊科技國際期刊，3:1,

1-21。

陳智偉（民86）。文件分類的方法與分析（未出版之碩士論文）。國立清華

大學，新竹市。

曾元顯（民91年6月）。文件主題自動分類成效因素探討。中國圖書館學會

會報，68，62-83。

曾元顯（民91年12月）。數位文件之資訊組織與主題分析自動化之技術與應用。臺北市立圖書館館訊，20:2，23-35。

曾元顯（民94年3月）。分類不一致對文件自動分類之效果的影響。大學圖書館，9:1，2-19。

曾元顯、莊大衛（民92年9月）。文件自我擴展於自動分類之應用。第十五屆計算機語言學研討會，2003年9月18-19日，129-141。

曾元顯（民96年10月）。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（民103）。自動化資訊組織與主題分析近二十年來的研究與發展。教育資料與圖書館學，51，特刊，3-26。

黃政偉（民87）。具語句特徵選取能力的類神經網路文件分類器（未出版之碩士論文）。國立臺灣科技大學，臺北市。

黃森原（民84）。中文文件自動分類（未出版之碩士論文）。國立交通大學，新竹市。

黃嘉宏（民97）。基於自動分類為基礎的圖書題名特徵擷取之研究－以輔助圖書分類系統為例。（未出版之碩士論文）。輔仁大學，新北市。

劉自誠（民88）。能有效變更之模組化階層式文件分類（未出版之碩士論文）。國立臺灣科技大學，臺北市。

錢愛兵、江嵐（2008）。基於標題的中文新聞網頁自動分類。現代圖書情報

技術，2008(10)，59-68。

顧皓光（民85）。網路文件自動分類（未出版之碩士論文）。國立臺灣大學，
臺北市。