回溯性資料數位化服務之規劃與建置

Networked Information Services for Retrospective Data

曾元顯
輔仁大學圖書資訊學系
tseng@blue.lins.fju.edu.tw

資訊傳播與圖書館學, 第 9 卷, 第 2 期, 2002 年 12 月, 頁 27-39.

摘要

電腦網路的發達，使得資訊的出版、傳播、與取得過程更加便利。然而，要將過去的紙本資料數位化，以提供網路化的資訊服務並非易事。輔仁大學中國社會文化研究中心多年來一直在進行中國大陸政治、經濟、社會、文化方面消息的蒐集與分析，並對大陸、香港、台灣地區的報紙新聞進行剪報工作，以提供學者近代中國社會文化概況的研究素材。本文即在介紹社文中心館藏資料數位化與網路化的建製過程、規劃與現況，提出經驗供相關人員參考。受限於人力、時間與經費，我們嘗試將紙本資料掃瞄後的影像文件，以 OCR 軟體辨識成數位文字，再經全自動化的關鍵詞擷取、關聯詞分析、與自動索引作業，建成便於檢索利用的資料庫。各項實驗顯示，OCR 文字辨識的運用可以節省大量的人力、時間與經費，OCR 辨識錯誤率在 10%以內，對檢索效果影響不大，錯誤率達 30%時，檢索成效的下降幅度還不到 30%。

Abstract

The advent of the Internet has made publication, dissemination, and access of information more easily. However, to digitize past retrospective data to allow networked information services is not an easy task. The Socio-Cultural Research Center (SCRC) at Fu Jen Catholic University has been collecting publications in the areas of politics, economy, society, and culture about China for years. Particularly, SCRC has been collecting and clipping newspapers from Mainland China, Hongkong, and Taiwan and accumulating over 700,000 clippings to support academic analyses and studies. This article is dedicated to describing the planning, development, and current status of a networked information service for the SCRC' s collections to hopefully share our experience with others undertaking similar projects. Due to the limited resources in time, human power, and financial support, we have experimented with the use of OCR technology to convert scanned images into digital texts. Automatic ways of keyword extraction, co-occurrence thesaurus construction, and indexing were then developed for these Chinese digital texts to build retrieval systems for effective access of the collections. Previous researches and ours as well showed that the use of OCR technology can save enormous time and efforts to provide networked information services. The inevitable OCR errors affect retrieval performance barely for high quality input. Retrieval effectiveness drops less than 30% even when the error rate goes beyond 30%.

關鍵詞：數位化、新聞剪報、OCR 文字辨識、檢索、中文

Keywords: Digitization、newspaper clipping、OCR、 retrieval、Chinese

一、前言

電腦網路的發達，使得資訊的出版、傳播、與取得過程更加便利。越來越多的資訊出現在網路上，使用者可免費或付費利用。網路資源便於取用(accessibility)的特性，不僅讓使用者克服了時間、空間、與國界的障礙找到資料，取得全文文件，同時也刺激了資訊提供者加速現有資料的數位化，以提供網路化的資訊服務。

雖然未來的資料以數位形式出現是可預期的事情，然而現今紙本資料，仍然記錄了全世界非常多的資訊。要將（回溯性）紙本資料數位化，以提供網路化的資訊服務並非易事。

輔仁大學中國社會文化研究中心（簡稱「社文中心」）五十年來一直在進行中國大陸政治、經濟、社會、文化方面消息的蒐集與分析，並對大陸、香港、台灣地區的報紙新聞進行剪報工作，以提供學者近代中國社會文化概況的研究素材。這些資料，在中國大陸的報章書刊外界不易取得的文化大革命時期，尤為重要。由於原始資料逐漸老舊、館藏空間漸感不足、且紙本資料的查詢利用困難，過去數年來，社文中心積極對其五十年來蒐集的資料，進行數位化的工作。其目的在試圖對這些資料做更佳的保存，並進一步希望能提供更有效率的網路化檢索與利用。

本文即在介紹社文中心館藏資料數位化與網路化的規劃、過程、建置與現況。受人力、時間與經費的限制，在無法全面進行館藏文件人工主題分析的情況下，我們試圖發展全自動化的作業流程，即運用 OCR 文字辨識軟體將掃瞄的影像檔案轉成數位文字檔，再利用自行研發的關鍵詞自動擷取、關聯詞自動分析、及索引自動建立等技術來建構館藏文件的資訊存取系統。這樣的作法過去並不多見。在此篇文章中除介紹社文中心館藏的數位化情形，也報告我們如何利用這些自動化技術來建構系統，並說明系統成效評估的細節，最後簡要展示此系統。希望這些研究的結果與經驗，對相關人員未來進行類似的專案，能有所助益。

二、社文中心館藏介紹

中國社會文化研究中心是天主教輔仁大學的一個研究機構，1994 年接收了來自香港「中國新聞分析」（China News Analysis, CNA）的大量館藏。「中國新聞分析」為中國觀察家中知名的人物勞達一神父（Father Laszlo Ladany）所創的時事通訊（newsletter）。自 1953 年創刊起，勞達一神父便根據中國大陸的官方出版品與廣播訊息，進行系統性的分析，來撰寫並出版有關中國政治、社會、文化方面的評論。「中國新聞分析」長久以來，遂成為中國觀察家、外交官、記者、學者必讀的刊物 [1]。「中國新聞分析」主要的館藏有：中國新聞分析、新聞剪報、廣播抄稿、人名機構卡片。這些資料移轉到輔仁大學後，社文中心仍繼續充實維護，並正式更名為「中國消息分析」。各項資料簡介如下：

（一）中國消息分析：1953 年問世，最初 25 年以週刊形式發行，自 1979 年起改為雙週刊，截至 1998 年停刊，共 1,625 期、約 2,500 篇文章、12,311 頁。

此份英文時事通訊曾發行 40 餘國，為研究中國大陸重要刊物，國外著名媒體曾專文介紹 [2,3]。其主要內容為中國大陸政治、經濟、社會相關新聞的分析。

（二）新聞剪報：為深入瞭解當時資訊封閉的中國大陸，於 1949 年開始蒐集大陸中央及地方(省)報紙，此外香港及台灣地區的報紙也納入蒐集的範圍，至今仍每天持續由人工檢視報紙。報紙文章經初閱、審閱兩階段挑選後，進行剪貼、標示（日期、報社、版別、類別）的工作。每篇剪報均按政治、經濟、文化三大範疇，其下再細分 100 多項專門類別，由人工標示類別後歸檔。目前大約累計 75 萬份剪報，報紙的字體包括簡體及繁體兩種，其中簡體字經歷中國大陸三次修訂，也就是不同時期的報紙，其簡體字稍有不同。

（三）廣播抄稿：由於文化大革命時期大陸印刷刊物不易取得，遂以人工抄錄當時大陸廣播的消息，這部分手抄稿涵蓋 1966 年到 1988 年間的廣播消息，共約四萬份文件。

（四）人名機構卡片：根據大陸官方發表的新聞稿，以人工整理紀錄大陸黨政機構、軍隊狀況以及重要人物生平、職務升遷之異動。每位人物、每個機構均以一張卡片的方式登錄，以便於查詢調閱。卡片上大都為人工抄錄的文字，也有部分是新聞消息的直接剪貼。

三、社文中心館藏數位化規劃

近年來資訊科技逐漸成熟普及，相對的，這些資料逐漸老舊，維護、利用不易，社文中心遂自 1997 年起開始著手館藏資料數位化的工作。其主要的目的是期望能保存原件，並做更好的館藏利用，讓使用者時空無礙的使用數位化的備份，而不必直接取用原件，以降低原件損毀的機率。

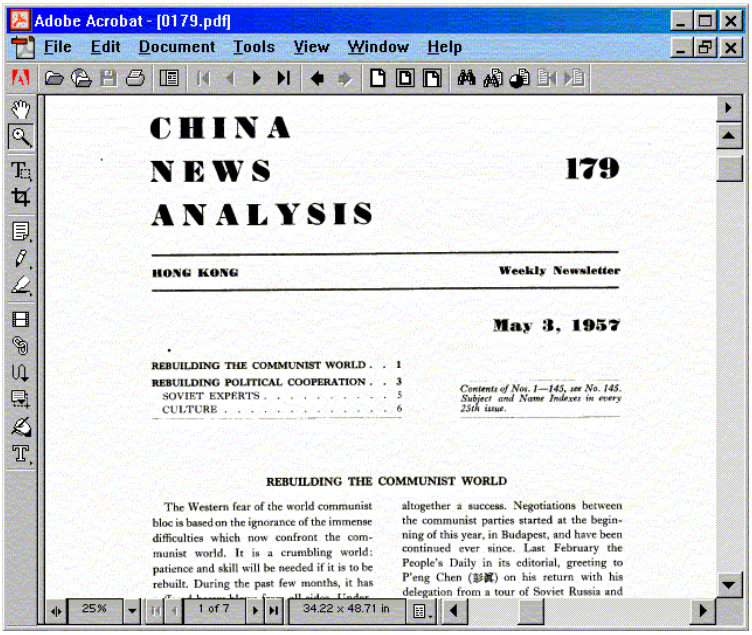
由於剪報資料量龐大，參考國內外的作法，曾考慮製作成微縮影片加以保存。但微縮影片的利用仍不便利，一旦維護不週，索引加工不全，記載在微縮片的資料反而更難取用。另一個可能的作法，是將原件掃瞄成影像檔，以便於電腦的後續處理與散佈。然而目前可以開啟、檢視影像檔案的軟體，在五年或十年後可能因新的檔案格式興起，而不一定繼續普及存在，或因電腦平台、作業系統與瀏覽軟體的更新而失效。因此有研究指出，站在永久保存的立場，數位影像儲存文件的方法不見得是最好的途徑。然而由於現今儲存影像檔案的光碟媒體使用週期有五年或十年之久，只要到時再進行檔案轉載或格式轉換，在成本與資料維護上應當是目前最可接受的選擇。

除了確定以影像檔案儲存全文資料外，將龐大文字資料數位化，另外要考慮的因素，是後續使用者如何檢索利用的問題。目前大部分全文影像檢索系統的作法，大都以人工進行主題分析，建立主題標目、關鍵詞庫及其他詮釋資料（metadata）以提供檢索。然而，索引者在使用者使用之前所做的主題分析與詮釋資料的紀錄，雖能讓使用者依此精準、快速的找到資料，但對於使用之前無法預測的使用行為與使用需求，能幫助的地

方就很有限了。因此，若能配合資訊檢索的新近技術，提供多面向的全文檢索，比較能滿足使用者各種可預測與不可預測的資訊需求。

當然，上面的描述是理想的狀況，其所耗費的人力與時間成本是相當可觀的，尤其是人工的主題分析部分。然而，以此理想，我們可以就現有的資源進行調配，將難得的資源發揮到最大的效用。就此原則，下面將就各個館藏資料數位化的規劃與過程做進一步的介紹。

（一） 中國消息分析：原文件為一期 1-5 篇文章，每篇文章均為英文，並包含部分中文譯名，文長約 1 至 20 頁，相當於 200 到 14000 個英文字。像這樣長的文章，主題分析與詮釋資料（metadata）的編製與記錄並不容易，尤其具備英文內容分析的專才難覓。因此我們規劃採用光學文字辨識（Optical Character Recognition, OCR）軟體，將影像檔轉成文字檔，以取得全文做為全文檢索的依據，並自動從文字中，擷取文章的日期、期別、與標題等資料，以配合全文的搜尋。現今 OCR 英文文字的辨識能力可達 90%-95%以上，而且成本低廉。例如，社文中心曾統計，將其過去二三十年發行的 1400 期英文紙本期刊的影像檔，以 OCR 文字辨識，經兩位工讀生利用四部電腦同步批次作業，僅花六個工作天、工讀費九千元即完成 15000 頁影像檔的辨識工作。然而當原文件版面不只一欄時，OCR 軟體常跨欄辨識文句，使得文字即使辨識正確，文句亦因欄位辨識錯誤而出現段落錯亂。因此，我們利用 MS WORD 的排版工具與英文拼字校正輔助工具，進行人工校正，最後再轉存成 HTML 檔案，以便於網路環境的瀏覽與檢索。除此之外，為保留原始文件的版面與內容圖片，同一份文件的每頁影像檔，還另外用 PDF 格式合併（如圖一），讓使用者可以在查詢全文時，一併調閱。



圖一：中國消息分析影像檔合併成 PDF 檔案的範例

（二） 新聞剪報：社文中心的剪報，均剪貼於一張比 A4 稍大的白紙上，其上再用手寫註明報別、日期、版次、與分類代號，如圖二所示。但對於有些跨好幾個版面的文章，則在白紙上貼一個小袋，將屬於同一篇文章的數份剪報裝在袋中。因此，影像掃瞄時除利用機器自動送紙外，常需要手工介入一份一份調校，減緩了影像掃瞄的建檔速度。經歷三年的執行，目前已完成自 1952 年到 1999 年的影像掃瞄 60 萬份。由於資料量龐大，難以像其他新聞剪報檢索系統一樣進行人工的主題分析與關鍵詞設定，目前僅就最基本的項目如報別、日期、版次、分類代號、標題與條碼進行人工建檔，這部分也已回溯建檔了 35 萬份。為充分利用現有資訊科技，在經費有限、時程緊湊的情況下，我們也嘗試了以 OCR 軟體轉換影像成文字的自動化作業，並以檢索系統自動建構索引以提供使用者主題檢索的努力。這部分的工作，不僅具備實用的潛在價值，也具備研究的價值，是整個社文中心館藏數位化的特色之一，我們將在下面數節中做詳盡的介紹。


我政府关于中印边界問題的声明

圖二：新聞剪報範例

（三） 廣播抄稿：由於廣播抄稿都是手寫文字，除影像掃瞄、人工建檔外，無法像新聞剪報那樣進行文字轉換的自動化作業，所幸這部分資料較少，目前已完成日期、電台、條碼、以及全文輸入的人工建檔約四萬份。同樣在沒有額外人力進行主題分析的情況下，我們利用新近的檢索系統，自動進行關鍵詞、關聯詞的擷取與索引的自動建構，以提供進階的檢索服務。

（四） 人名機構卡片：不像上述的全文資料，人名機構卡片常以條列式的方式，列舉人物、機構的異動與變遷，也因此常常會提到與該卡片主題相關的其他人物或機構。為了能充分展現、連結這些關係，較好的作法，是將這些資料鍵入關聯式資料庫中，以便於分析、提取各種人物、機構之間的變遷或互動的關係，並

與上面的三種全文資料的館藏結合，整合成知識性較強的檢索系統。然而由於經費與時程限制，這部分工作還未執行。但我們預估，人名機構卡片的數位化工作，將可扮演串連整合上述三種全文資料的角色，提升其整體的資訊利用價值。

四、社文中心「全文影像 OCR 文件檢索測試集」

前面曾提到，紙本資料掃瞄成影像檔後，可利用 OCR 軟體辨識影像檔的內容取得數位文字，以提供全文的檢索利用。國外像美國的 Historic Newspapers in Digital Times 計畫，即嘗試將 1887 年之日報掃瞄成影像檔，再經 OCR 辨識成文字後，放置網路上供使用者檢索 [4]。類似的計畫還有 1851 年至 1923 年時期紐約時報的回溯性建檔、影像掃瞄、OCR 文字辨識、網頁環境的檢索利用 [5]。此外，希臘也有新聞數位化館藏與運用的研究 [6]。就我們所知，國內也有國家圖書館進行紙本資料的影像掃瞄，以及 OCR 文件檢索的試驗。事實上，現在便有商用套裝軟體像 Acrobat [7] 和 DynaDoc [8] 可支援這樣的作法。如果原始文件可產生高品質的影像和 OCR 文字，採用這個途徑將會是成本效益極佳的選擇。比起運用人工分析文件內容、建立編目資料或詮釋資料（metadata）以提供檢索與利用的作法，這套流程可提供時效更快、成本更低的文件數位化與網路化的檢索與利用。當然，若經費、時間許可，結合自動化的作業與人工分析，將可提供使用者最大的效益。

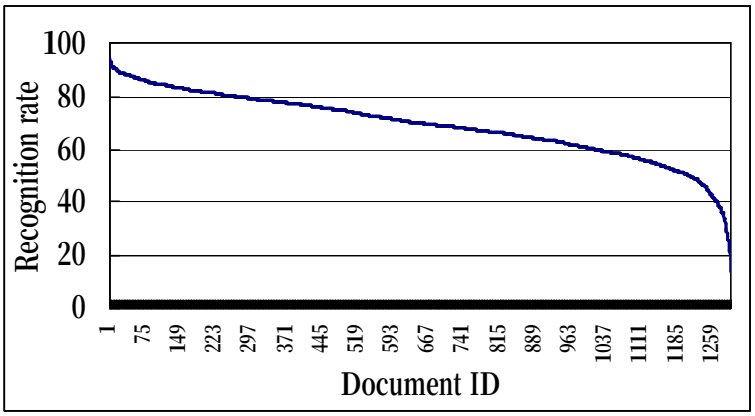
此自動化流程的問題在於 OCR 文件常常含有辨識錯誤的詞彙與段落，導致其提供的檢索與利用的品質可能降低。過去的研究顯示，OCR 辨識錯誤的情形，對影像品質良好的文件，並不嚴重，對檢索成效的影響也不大 [9]，因此大部分品質較好、年代較新的紙本資料可適用此套流程。然而圖書館的館藏，常常是年代較為久遠、印刷或紙質較差的紙本資料，其 OCR 的結果，常常是詞彙錯誤率較高的數位文件，其對檢索結果的影響也較顯著。因此，一個值得研究的課題，是如何降低 OCR 的辨識錯誤、提升 OCR 文件的檢索成效。

過去數年，國外的資訊檢索研究機構與學者，曾就西方語言的文件，進行 OCR 錯誤對檢索結果影響的研究 [10-17]。這些研究，大多仰賴一套檢索測試集（test collection）以評估檢索成效，從而提出各種檢索技術，以試圖降低 OCR 錯誤對檢索結果的影響。然而對中文 OCR 文件而言，相關的研究則相對較少。

為了研究中文 OCR 文字檢索的成效，我們乃在社文中心的協助之下，自行發展了一套「中文 OCR 文件檢索測試集」。一套檢索測試集通常包含三個部分：待檢索的文件、測試成效的查詢主題，以及描述哪些文件與哪些查詢主題有關的相關判斷（即查詢問題的「標準」答案）。社文中心的測試集一共包含 8438 篇 OCR 文件，以及 30 道以自然語言表達的查詢主題。這些文件來自於 1950 到 1976 年間中國大陸及台灣方面有關外交、軍事的剪報，因此大部分是簡體字，少數是繁體字。事實上我們原先選用了 11,108 份全文影像進行 OCR 辨識，並從我們自行比較的三種 OCR 軟體中，選出較好的軟體進行辨識，但最後僅得出 8438 篇有效的 OCR 文件，其他的 2670（=11,108-8438）篇文件

則由於影像品質太差或 OCR 軟體無法讀取其檔案格式，導致 OCR 軟體無法產生出有效的文字。從 1300 份 OCR 文件的樣本得知，這些 OCR 文件的平均辨識率約 69%，其辨識率分佈情況如圖三所示。

至於 30 道查詢主題的來源，由於實際使用者的需求難覓，遂自 1950 到 1976 年間的相關學術期刊中，摘錄 100 篇論文的標題做為擬定查詢主題的參考，再以人工改寫成符合 TREC 競賽規格的查詢格式而成 [18]。這裡假設如果某個作者針對某個題目撰寫論文，則至少當時有那位作者對該題目有某種程度的資訊需求，或者說當時已出現某些資訊足夠支撐某些人就該問題撰寫論文。因此嘗試從學術期刊的論文標題中尋求查詢主題應該是個合理蒐集使用者需求的作法。表一中展示一個查詢主題的範例。這些中文查詢主題，都進一步經由社文中心的研究人員改寫成英文，以提供未來英、中文跨語言檢索的測試，擴展此測試集後續的應用。



圖三：OCR 文件辨識率分佈狀況，橫軸為文件序號，文件依辨識率由高而低排序

<top><num> Number: 02
<title> 麥克馬洪線
<desc> Description:
中共與印度間對於麥克馬洪線之爭論
<narr> Narrative:
相關文章內容包含中共或印度方面對於麥克馬洪線的看法或主張,
若文章內容僅是對於中印邊境間之戰爭情形做報導則視為完全不相關。
</top>

表一：查詢主題範例

至於查詢主題的相關判斷工作，則由三位分別具備歷史與圖書資訊學背景的大學

生與研究生依三種相關程度作判斷。他們針對每一道查詢主題，檢視每份剪報的全文影像，再判斷其為相關、部分相關、或不相關，依此分別給予 2 分、1 分或 0 分的分數。最後將這三位判斷者給每份文件的分數加總起來，得到每份文件的相關判斷分數，如表二所示。全部 8438 篇文件中，共有 899 份文件的相關分數不是 0 分，亦即有 899 份文件被判定與某一查詢主題有相關或部分相關，平均每道查詢主題的相關文件有 30 篇，最少的有 4 篇，最多的有 125 篇。如此，在兩個月內，這三位判斷者共做了 3 x 30 x 8438 = 759,420 次相關判斷。

查詢主題代號	文件代號	第一位判斷者	第二位判斷者	第三位判斷者	總分
01	0053487	1	1	0	2
01	0053489	1	2	1	4
…					
02	0054425	2	2	2	6
02	0054452	1	1	1	3
…					

表二：相關判斷範例

五、OCR 文件檢索成效評估

有了這樣的測試集後，筆者便進行中文 OCR 文件的檢索實驗，我們運用向量檢索模式，並參考過去相關研究的作法，試驗 12 種不同的檢索策略，並以 TREC 的評估程式衡量檢索成效，其平均查準率由 0.3157 進步到 0.4757。為了獲得更多的結果，並比較不同檢索模式的效能，我們與美國馬里蘭大學合作，利用當地國內沒有的研究資源，進行檢索試驗。我們以機率檢索模式聞名的 INQUERY 檢索系統做實驗，在六種檢索策略中，獲得最佳的平均查準率為 0.4692，此結果與我們自行發展的系統在此 OCR 測試集中的成效（0.4757）相差不多。表三中顯示各個檢索策略的成效，以最後一欄的平均查準率表示。至於表三中詳細的檢索策略說明，可參考 Tseng 與 Oard 的論文 [19]。

System	Field	Indexing Method	Weighting scheme	Run ID	Ave. P
Inquery	title only	1-gram		I1s	0.3612
Inquery	title only	2-gram		I2s	0.4083
Inquery	title only	1-gram and 2-gram		Ins	0.4397
Inquery	all fields	1-gram		I1	0.3472
Inquery	all fields	2-gram		I2	0.4621
Inquery	all fields	1-gram and 2-gram		In	0.4692
Crystal	title only	1-gram	log(tf)*log(IDF)*ByteSize, tf*Cosine	Gb1s	0.3509

Crystal	title only	1-gram	log(tf)*log(IDF)*Cosine, tf*Cosine	Gc1s	0.3059
Crystal	title only	2-gram	log(tf)*log(IDF)*ByteSize, tf*Cosine	Gb2s	0.4044
Crystal	title only	2-gram	log(tf)*log(IDF)*Cosine, tf*Cosine	Gc2s	0.4000
Crystal	title only	1-gram and 2-gram	log(tf)*log(IDF)*ByteSize, tf(3w-1)*Cosine	Gbs	0.4164
Crystal	title only	1-gram and 2-gram	log(tf)*log(IDF)*Cosine, tf(3w-1)*Cosine	Gcs	0.4098
Crystal	all fields	1-gram	log(tf)*log(IDF)*ByteSize, tf*Cosine	Gb1	0.3963
Crystal	all fields	1-gram	log(tf)*log(IDF)*Cosine, tf*Cosine	Gc1	0.3157
Crystal	all fields	2-gram	log(tf)*log(IDF)*ByteSize, tf*Cosine	Gb2	0.4582
Crystal	all fields	2-gram	log(tf)*log(IDF)*Cosine, tf*Cosine	Gc2	0.4344
Crystal	all fields	1-gram and 2-gram	log(tf)*log(IDF)*ByteSize, tf(3w-1)*Cosine	Gb	0.4757
Crystal	all fields	1-gram and 2-gram	log(tf)*log(IDF)*Cosine, tf(3w-1)*Cosine	Gc	0.4459

表三：運用 Inquery 查詢系統以及 Crystal 檢索系統 [20] 得到的檢索成效

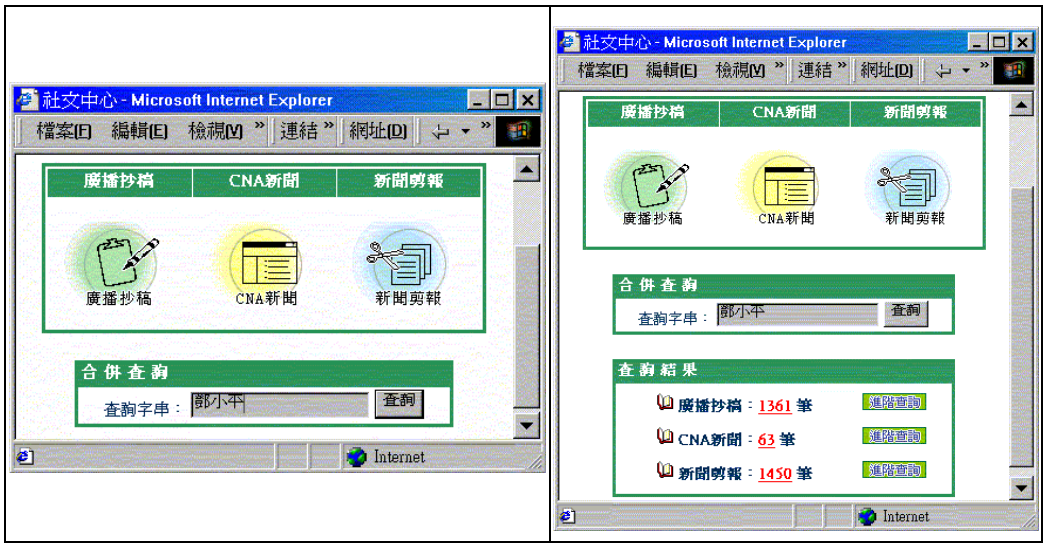
上述結果僅能得知在 OCR 正確率約 0.69 的情況下，檢索成效約 0.475。但這 0.4757 的結果到底好不好，要跟完全乾淨的相同測試集比較後才能知道。為了做這樣的比較，在沒有經費將這 8438 份剪報重新輸入與校正的情況下，我們選擇將部分的剪報以人工重新輸入，以獲得沒有錯誤的乾淨文件，共 899 篇。這 899 篇都是在這 30 道查詢主題中被判定為相關的文件，將此 899 篇乾淨文件取代原 899 篇 OCR 文件，我們得到另一組 8438 篇「部分乾淨測試集」，用以模擬全部文件都是乾淨的測試集。檢索試驗的結果顯示，此「部分乾淨測試集」的最佳平均查準率為 0.6588。這個測試集中，相關的文件是乾淨的，而不相關的文件都是有辨識錯誤的文字，所以其平均查準率事實上會比 8438 篇文件全部都是乾淨的測試集還要高（因為不相關的文件，比較沒有正確的詞彙來影響查詢結果）。雖然只是近似值，這 0.6588 的數值可以做為此測試集檢索效能的上限（upper bound），讓我們 OCR 文件的檢索試驗有比較的依據。

從上述的結果得知，我們最多還有 0.6588-0.4757=0.1831 的檢索效能還需要改善。或者從另一個角度看，69%正確率的 OCR 文件，其檢索效能至少為乾淨文件的 72.2% =0.4757/0.6588。雖然我們已經試驗了多種自動化檢索技術以改善檢索效能，我們還有其他技術還沒有運用上，其中我們覺得最重要的一項是自動偵測與更正 OCR 錯誤的技術。自動化檢索技術不管多精巧，最後還是在做字串的比對，如果字串因 OCR 辨識有錯，便會造成比對不到的情況。然而由於文件的主題詞彙常有重複出現的特性，某個詞在某地方錯了，若在某個地方對了，雖然比對的分數也許下降了，還是會有比對正確的情形，因此傳統的檢索方法還可以有部分的檢索成效。然而如果辨識的情況太差，造成大部分的主題詞彙都錯的時候，非得更正這些錯誤詞彙不可了。

因此，未來的工作將發展相關的技術，自動偵測或更正中文 OCR 文件的錯誤詞彙，以進一步提升 OCR 文件的檢索成效，讓即便是雜訊程度非常高的 OCR 文件，仍可檢索而且有效，從而加快回溯性資料的數位化服務進程，使其能以低廉的成本提供利用。

六、系統整合與展示

除人名機構卡外，其他資料的掃瞄、建檔已接近完成，並轉進關聯式資料庫中以方便管理。其中全部的影像檔約 70 萬筆，有編制書目的影像檔約 40 萬筆，約佔 45 GB（Giga Bytes），而文字部分包含中國消息分析的 1625 期、新聞剪報的 35 萬筆書目檔及 8438 份全文檔、以及廣播抄稿的 4 萬份全文，總計不超過 500MB（Mega Bytes）。三、四年前原規劃需要七、八十片的光碟櫃才能容納的資料，在儲存設備進步飛速的情況下，現在只需成本不到一萬元台幣的兩部 40GB 硬碟即可解決。此外，原先規劃需要大型主機才能負荷的索引、查詢與系統管理的工作，也由於計算機速度快速提升的情況下，現在只需市面上主流的個人電腦即可勝任。在資料掃瞄、建檔期間，我們也著手發展結合全文影像且適合 OCR 雜訊文件的檢索系統，以全自動化的方式擷取文件的關鍵詞、關聯詞，並結合資料庫管理系統的結構化 SQL（Structural Query Language）欄位查詢，提供使用者便利的網頁使用環境。



圖四：社文中心館藏整合查詢，左欄：查詢介面，右欄：查詢範例。

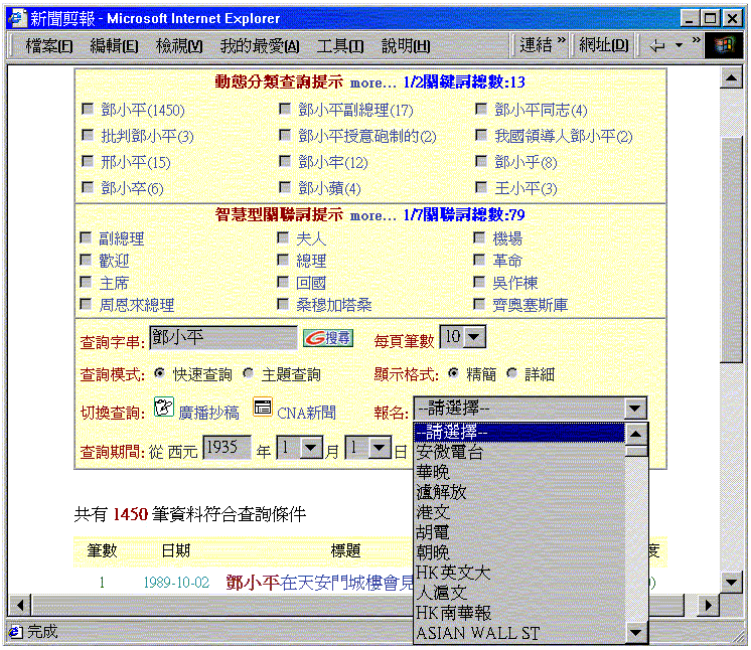
整個系統整合後的網頁查詢介面如圖四所示。使用者下達查詢字串後，可得知每個資料庫的查詢筆數，據以選擇適當的資料庫，進行檢索結果的檢視。圖五為檢視新聞剪報的畫面，系統除顯示全文檢索的比對結果外，也提示與查詢字串近似的文件用詞，例如：鄧小卒、鄧小牢，鄧小乎、邢小平等，並顯示其出現篇數，以增強模糊搜尋容錯能力，加強檢索的召回率。如前所述，在新聞剪報中我們實驗了 OCR 文字辨識全文的檢索效果，因此這個資料庫中除人工輸入的新聞標題外，還有八千多篇 OCR 文字，所以才會有鄧小卒、鄧小牢等辨識錯誤的文件詞彙。除了辨識錯誤、資料誤植外，像史達林、斯大林、史大林等外國譯名不一的情況，也可由此功能顯示出來，自動提示使用者更多相關查詢用語，找到更完整的資料。這個動態提示詞的功能，不僅可以解決部分同義異名詞外，也可以提示範圍較窄的搜尋詞彙，如「鄧小平副總理」，讓使用者可以快

速縮小查詢範圍到鄧小平當副總理時代的文件。另外，每個提示詞都標示其出現篇數，讓使用者查一個詞，就同時可得知其他好幾個詞的查詢筆數，節省使用者一個詞一個詞查詢的的力氣。由於這些提示詞都是文件中的語彙，如果把每個提示詞當作一個類別，使用者不僅可從其出現篇數即可得知該類別的大小，也可同時選用不同的詞彙作類別之間交叉查詢。

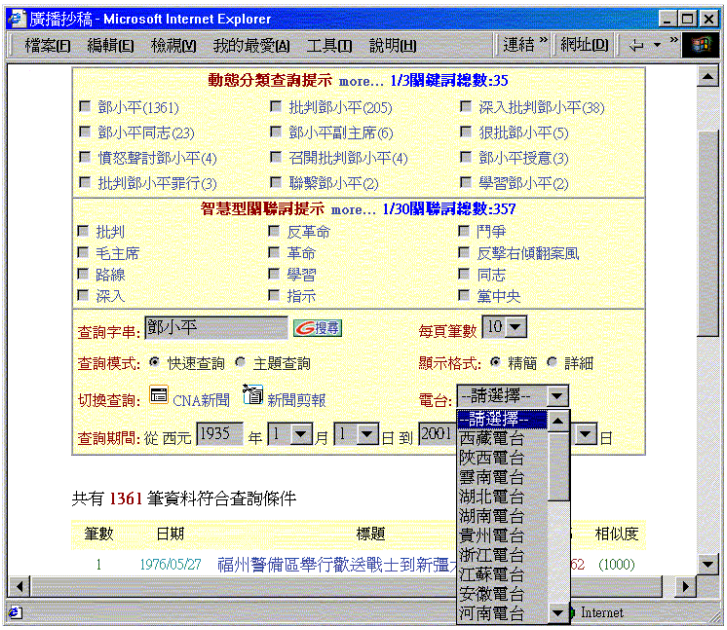
除動態分類提示詞外，系統也提示在主題上近似查詢詞的關聯詞。這些關聯詞都是與該查詢詞常常一起在文件中出現的詞彙，這種共同出現的詞彙（co-occurring terms）常能顯示出事物之間的關聯，而這些關聯猶如記錄了某些知識。因此關聯詞具有相當程度的摘要作用，使用者想做進一步的瞭解，可同時點選適當的詞彙，調出相關的文件，從其記載的描述中瞭解詳情。然而自動得出的關聯訊息，雖然成本低廉，但畢竟只從「共同出現」的線索得來，難免會有錯誤。我們的評估顯示，69% 的關聯提示詞被判定與查詢詞相關 [21]。如果再經由人工知識判斷的加工，提示出來的關聯詞就可以達到更高的的品質。

各個資料庫的檢索畫面中，還提示適合各資料庫性質的檢索點，如新聞剪報中的日期、報別，廣播抄稿中的日期、電台等等，以限制查詢的範圍。各資料庫也都提供「精確比對」做多詞交集、快速比對，以及「主題查詢」，以便運用前述的檢索模式與策略，依詞頻、文長等因素對文件加權排序。如果該資料庫有全文影像，檢索結果中會有超連結，允許點選調閱。瀏覽全文時（廣播抄稿），系統會顯示該文的所有關鍵詞，以供進一步參考與檢索。

此系統雛形大略已建構完成，目前正在進行使用者驗證與意見調查。未來會將系統轉移至國科會人文處，提供經常性的檢索利用服務。



圖五：新聞剪報的檢索範例。



圖六：廣播抄稿的檢索範例。

圖七：中國消息分析的檢索範例。

七、結語

從歷史的角度看，報紙是記載當時社會、政治、文化的草稿，雖然不是最終的歷史版本，仍有保存的必要與價值，只要用心整理，可以是未來探索現在的極佳材料。本文以社文中心的館藏數位化為例，描述其問題、需求、規劃、建置與研發過程，以及整合後的系統。這中間具有學術研究價值的部分在於全文影像 OCR 文字檢索的部分。此課題具有實用的價值卻少有人研究，尤其是中文 OCR 文件的資訊檢索部分，若沒有此次數位化專案的機會，將不知何時有中文 OCR 檢索文件測試集可用。國外及我們的研

究顯示，OCR 文字辨識的運用可以節省了大量的人力、經費，其錯誤率若在 10%以內，對檢索成效影響不大，其錯誤率達 30%時，檢索成效下降幅度還不到 30%。

整個數位化過程，由於文件數量龐大，在數位掃瞄的外包作業過程，曾發生原件被廠商遺失、廠商將自己轉好的數位資料覆寫損毀、前後送回的資料不一致、人工建檔品質非常不穩定、數位化歷程太久資料追蹤掌控不易等情形。這些經驗，可供未來從事相關工作的人員參考，以加強廠商與資料的管理。

社文中心還有人名機構卡片還未數位化，由於它是精心整理的名錄，可以作為其他資料庫的輔助資訊或交叉參考資料，在縝密的資料與結構分析後，將其建成數位資料庫，應該可以把其他三個資料庫做更緊密的結合，發揮整合的效益。

最後，此系統雛形已大略完成，待資料全部整理、校正完成，雛形展示系統將移轉成常態服務的系統。後續可能遭遇的問題，是新聞剪報的影像檔在網路上沒有限制使用時，可能會碰上智財權的爭議。因此，此常態服務應該會規劃成會員形式的網路化服務，以保護各方的權利。其細節，還有待未來的詳細探究與規劃。

誌謝

感謝社文中心關秉寅前主任、狄神父主任，以及康芳菁、李青玲、鄭明賢、蔡孟竹等人多年的協助，本文才得以完成。

本文由國科會計畫補助，計畫編號：NSC 88-2418-H-001-011-B8908 及 NSC 88-2418-H-001-011-B9003。

參考文獻

[1] Simon Leys, “The Art of Interpreting Nonexistent Inscriptions: Written in Invisible Ink on a Blank Page,” The New York Review, Oct. 11, 1990.

[2] Joseph Treen, “Ending a China Watch,” NEWSWEEK, Jan. 10, 1983, pp. 14.

[3] “A New View of China,” THE ECONOMIST, Nov. 12, 1994, pp. 90.

[4] “Historic Newspapers in Digital Times”, http://www.colosys.net/pathfinder/AboutPathfinder/HistoricNewspapers.htm

[5] Robert Thibadeau, Chris DeWan, joel Young, and Denis Marous, “The CMU-Seagate Historical New York Times Project,” Proceedings of Fourth Symposium on Document Image Understanding Technology, Columbia Maryland, April 23-25th, 2001, pp. 115-118.

[6] S. L. Mantzaris, B. Gatos and N. Gouraros, “Creating a Digital Library From News Archives,” Proceedings of Fourth Symposium on Document Image Understanding Technology, Columbia Maryland, April 23-25th, 2001, pp. 285-287.

[7] "Adobe Acrobat 4.0" http://www.adobe.com/prodindex/acrobat/main.html

[8] "DynaDoc" http://www.asiaserve.com.tw/product/dynadoc/doc_index.htm (in Chinese)

[9] W. B. Croft, S. M. Harding, K. Taghva, and J. Borsack, "An Evaluation of Information Retrieval Accuracy with Simulated OCR Output," The 3rd Symposium of Document Analysis and Information Retrieval, 1994, pp.115-126.

[10] Kazem Taghva, Julie Borsack and Allen Condit, "Evaluation of model-based retrieval effectiveness with OCR text," ACM Transactions on Information Systems, Vol. 14 , No. 1, 1996, pp. 64-93.

[11] Kazem Taghva, Julie Borsack and Allen Condit, "Results of applying probabilistic IR to OCR text," Proceedings of the seventeenth annual international ACM-SIGIR conference on Research and development in information retrieval July 3 - 6, 1994, Dublin Ireland, pp. 202-211.

[12] K. Taghva, J. Borsack, and A. Condit, "Effects of OCR Errors on Ranking and Feedback Using the Vector Space Model," Information Processing and Management, Vol. 32, No.3, 1996, pp. 317-327.

[13] Amit Singhal, Gerard Salton, and Chris Buckley, "Length Normalization in Degraded Text Collections," Proceedings of Fifth Annual Symposium on Document Analysis and Information Retrieval, April 15-17, 1996, pp. 149-162.

[14] Elke Mittendorf, Peter Schauble and Paraic Sheridan, "Applying probabilistic term weighting to OCR text in the case of a large alphabetic library catalogue", Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval July 9 - 13, 1995, Seattle, WA USA, pp. 328-335.

[15] Daniel Lopresti and Jiangying Zhou, "Retrieval Strategies for Noisy Text," Proceedings of Fifth Annual Symposium on Document Analysis and Information Retrieval, April 15-17, 1996, pp. 255-269.

[16] Claudia Pearce and Charles Nicholas, "TELLTALE: Experiments in a Dynamic Hypertext Environment for Degraded and Multilingual Data," Journal of the American Society for Information Science, 47(4), 1996, pp.263-275.

[17] Harding, W. B. Croft, and C. Weir, "Probabilistic Retrieval of OCR Degraded Text Using N-Grams," in Research and Advanced Technology for Digital Libraries, Carol Peters and Costantino Thanos, Editors, 1997. pp. 345-359. http://ciir.cs.umass.edu/info/psfiles/irpubs/ir-115.ps.gz

[18] 蔡孟竹，中文 OCR 文件檢索測試集之製作與應用，輔仁大學碩士論文初稿，民 90 年。

[19] 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.

[20] Yuen-Hsien Tseng, "Automatic Cataloguing and Searching for Retrospective Data by

Use of OCR Text", Journal of American Society for Information Science and Technology (Previously known as Journal of the American Society for Information Science, JASIS), Vol. 52, No. 5, 2001, pp. 378-390.

[21] **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**。