教育資料與圖書館學，第30卷，第4期（1993），頁351-361

圖書館中英文全文影像光碟檢索系統之設計與製作

陳瑞順　丁崑健　陳登吉

A Design and Implementation of Library Chinese Full-Text Image CDROM Retrieving System

Ruey-shun Chen

Lecturer

K. C. Ting

Director of Library

D. J. Chen

Professor
National Chiao-Tung University
Shinchu, Taiwan, R.O.C.

Abstract

Library CDROM with its enormous storage, retrieving from network compabilities, it has been gradually replacing some of its printed counterpart. But one of the disadvantage is that only English version CDROM, and can not access the full-text CDROM from Chinese environment.

This paper is proposed a new method to solve this problem. The method use hash function as data structure and document image processing technique to perform a practical library Chinese full-text CDROM retrieving system. Its advantage can reduce storage space and allow multiuser to retrieve the same CDROM from network, and allow user to retrieve and print-out full text image CDROM database under Chinese environment.

前　言

近年來，傳統檔案管理系統已逐漸被電腦媒體所取代，大量資料均可由電子媒體形式所保存，例如製作成光碟CDROM，其容量大，每片約600MByte。其目的就在減少儲存空間，延長保存年限，可透過網路存取檢索容易，製作修改效率提高等優點(註一)。美國UMI公司於1987年開始製作光碟產品以美國各大學的博碩士論文摘要行銷世界，普遍爲各大學所採用，國內各圖書館亦陸續採購使用，作爲教授、研究生及學生的研究工具。美國UMI公司更於1982年1月推出英文版的全文光碟，種類涵蓋所有學術領域，對研究者眞是一大福音，然國內正缺乏此種中文全文光碟系統技術的開發，但鑑於目前光碟具備有以下優點：

㈠體積小，目前最常用的大小爲5.25吋CDROM，易於流通與推廣。

㈡高度儲存量，每片CDROM可儲存高達600MByte的資料，不是一般硬碟所能達到。

㈢易於保存，目前已商品化的CDROM光碟壽命一般至少在10年以上。

㈣便於隨機檢索，目前光碟的平均存取時間約爲600mS，雖比硬碟慢，但與一般用來儲存大量資料的磁帶、微縮膠片相比，檢索速度要快得多。

㈤成本低廉，CDROM driver亦降爲每台約三千元，將變成PC之標準配備。

由以上分析可知，目前光碟技術發展已相當成熟，且可達到普遍化、實用化及量產的階段。於是國立中央圖書館在民國80年即完成期刊論文索引光碟片，儲存國內出版1,161種期刊上的論文索引資料共157,360筆(註二)。但是此種系統當使用者在使用時，並無法了解所搜尋到的資料是否確爲自己所需，往往只能由論文標題自行判斷文章的內容，有時辛苦找到全文時才發現與所需資料無太大關聯。有鑑於此，本人即於81年著手開發此種中文全文光碟檢索系統，以國立交通大學的博碩士論文爲對象。希望使用者在得到檢索後的結果時，可立即由電腦螢幕上看到論文全文。

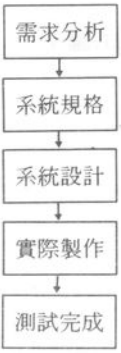
目前，本系統已設計製作完成，所有功能包含檢索及全文閱讀皆完成，關於中文檢索方面，不論單一個中文字或一串字組成的任何詞彙，都可當作

搜尋的鍵值。同時由於理工科目往往在論文中不可避免地加入一些英文專有名詞，因此，在設計時即考慮到處理英文方面的問題，所以本系統可順利地兼顧中英文各種需求，快速地找到使用者所欲檢索的資料，同時將全文清晰地顯示在螢幕上，可直接閱讀論文全文，眞正找到所欲檢索的資料，也可直接由印表機印出論文全文，如此一來，既可節省往返書架，找尋期刊的時間，並能透過校園網路查得所需資料並印出，可節省研究者甚多時間。

由於目前個人電腦技術發展成熟，486速度直通工作站。影像處理的技術已成功地商品化，因此，掃描影像存入電腦再也不是一件困難的事。關於軟體方面，資料壓縮所使用的演算法可以達到至少30%的壓縮比，因此影像資料不再是一些大而不易處理的東西。CDROM資料儲存量大，因此，中文全文檢索系統才得以順利地發展完成。本文就此一光碟系統之演算法、資料格式、系統軟體配置需求以及系統特色及功能等加以介紹，最後並提出未來努力之展望。

一、設計方法

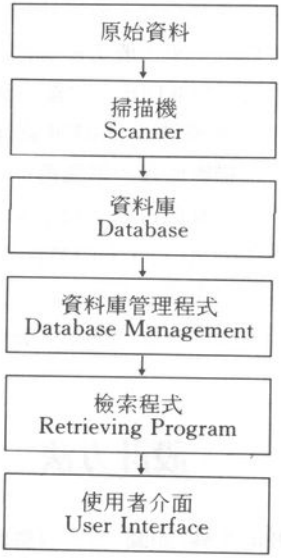
我們將由系統軟體發展的步驟(如圖一)，以整個系統爲著眼點，對此一中英文全文檢索系統作一有系統的分析，並對每一個所遇到的問題提出經濟且有效的解決方法。



圖一　系統發展流程圖

(一)需求分析

由使用者的角度來看整個系統的架構，可用圖二表示，主要功能爲資料檢索與全文閱讀，以下分別詳述之。



圖二 系統方塊圖

1. 資料檢索：

(1)檢索的目的是以篇爲單位的論文。

(2)可能的檢索項目有論文的：

a. 日期：年、月。

b. 系所。

c. 作者。

d. 指導教授。

e. 標題關鍵字。

f. 摘要關鍵字。

在以上六項資料中，前二項可視爲以數字爲鍵值的檢索項，剩下的都必須作到可以同時使用中英文作爲鍵值的檢索項(註三)，例如：中文姓名與

英文姓名、中文關鍵字與英文關鍵字。

2. 全文閱讀：

關於全文閱讀方面，必須可將論文內容以可辨識的解析度顯示在螢幕上，同時可以上、下翻頁，以利觀看論文內容。論文內容也可送到雷射印表機印出，將論文還原成原先一頁一頁的樣子。

3. 使用者介面：

一個有用的系統必須同時具有一富有親和力的使用者介面，因此在操作時，螢幕上必須顯示足夠的訊息，引導使用者，使每個人都可輕鬆地找到自己所需的資料。

㈡系統規格

根據上一節系統需求分析的結果，我們可相對地寫出所需的系統規格，達成所要需求(註四)。

1. 資料檢索方面

其有三種不同的鍵值：中文、英文及數字，以檢索項而言，每一檢索項可能以數字作爲鍵值，或是以中文和英文作爲鍵值。中文鍵值的最小單位是一個中國字，英文鍵值的最小單位也是一個字(word)，例如：「liquid」，「structure」。不但可以以某一檢索項找出與鍵值吻合的論文，也可利用and、or，將不同檢索項的結果加以組合以利使用者找尋資料。

2. 全文閱讀方面

經由掃描機掃進電腦的論文內容必須經過適當的壓縮。此處適當的壓縮表示(註五)：

⑴壓縮效率不可太差，以免浪費儲存空間。

⑵解壓縮不可太慢，以免當全文閱讀時影響顯示速度。

必須在以上二項因素之間取得一個適當的折衷，才能有最佳的效能。

關於印表輸出這方面，在考慮輸出品質與普及率的前提下，選定HP LaserJet Ⅲ作爲輸出印表機。

3. 使用者介面方面

以選單(menu)爲主的操作方式，是最適合一般使用者使用電腦的方式，故本系統也採用相同策略，再加上足夠的使用說明。

二、系統設計與製作

我們由系統規格爲藍本，設計出整個系統架構的概觀，以完成需求分析中的要求。

㈠資料檢索方面

要達到系統規格的功能，我們必須建立一資料庫，因爲欲加入資料庫的論文必已事先決定，故此資料庫只提供增添資料與檢索的功能，在增添資料時，以篇爲單位，所須執行的工作有：

1. 編號：每一篇論文都應有一獨一無二的序號，此序號即爲論文加入時的號碼。

2. 封包影像資料：影像資料由掃描機(scanner)掃入時，每頁都是一獨立的檔案，此時我們必須將這些檔案連接成一個更大的影像檔包含此篇論文所有的影像資料，即論文內容。

3. 讀入文字資料：文字資料包含前述的：標題、摘要、作者、指導教授等資訊。

4. 產生索引資料：此刻即牽涉到建立索引的方式，以下分別以三種不同鍵值詳述之，因爲所有索引項都脫離不出這三種鍵值。

⑴數值：因爲以數值爲鍵值的索引項的可能值一定是有限個的，所以可直接以陣列建立索引表，舉例如下：

論文序號	年份	索引表
1	80	80-1 2 3
2	81	81-2
3	80	
4	80	

如此一來，即可迅速地由使用者輸入的鍵值找到對應的論文序號。

⑵英文：以英文字(word)爲鍵值的搜尋法中以B-tree的效率最高，但此處我們採用雜湊法(hash)作爲索引表的建立方式，因爲

a. 容易實作，且易除錯、維護，因爲建立一個雜湊表比建立一個B-tree還要容易許多。

b. 效率折損不大，雖然B-tree的效率最高，但使用雜湊表並不會使

效率折損許多。

c. 提供更大彈性，因爲雜湊表的大小可以調整，因此可自由地選擇最適合的大小。

由以上幾點考量，再加上一個精選的雜湊函數(hash function)就可使碰撞(collision)的機率降到最低，而大大地提高索引速度(註六)。以此法建立的索引表可以涵蓋所有出現在論文中的英文單字，並且相當有效率。

(3)中文：與英文的字相當的應是中文的詞，但若以類似英文的作法來建立中文的索引表，第一個面臨的問題就是在一個句子中，如何判定中文的詞。舉例來說：「第一個」「面臨」的「問題」或是「第」「一」「個」「面臨的問題」。因爲中文不像英文以空白作爲詞的分隔字元，同時，目前電腦也沒有足夠智慧來判斷中文的「詞」。經過長久思考後，總算想出一個可有效解決的方式，就是以中國字作爲索引表的鍵值，同時存下此中文出現在句子中的位置，如此一來，不論中文的詞是由那些字組成的，都可順利地查到。舉例來說，一段本文及索引表如下：

用 這 個 小 飾 品 來 裝 飾 這 個 禮 物 是 最 好 不 過 的
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19

索引表

中文字	位置值
小	4
飾	5 9
品	6
裝	8
好	16

假若檢索的鍵值爲「裝飾」，則「裝」的位置在「飾」的前一個字，於是「裝」的位置值必須比「飾」少一個才是所求。由上例可知，裝—8，飾—9，恰滿足此狀況，故找到一個欲搜尋的資料了。

5. 寫入資料，建立資料庫。在此步驟必將以前所讀入的文字資料與建立的索引表，一併寫入資料庫中，以利日後檢索工作。

以上所述是有關完成資料檢索所需的方法，接下來看看有關全文閱讀的

問題。

㈡全文閱讀

達成全文閱讀的主要技術就是影像處理，所以第一個面對的問題就是資料格式的選擇，在考慮過壓縮比與解碼速度之後，我們選擇標準的PCX格式。此種格式雖然不能提供最好的壓縮比，但是在解碼速度、記憶體使用方面，有著不錯的表現(註七)。

對反壓縮輸出到螢幕與印表機其實是沒有什麼不同的，只不過當輸出到印表機時，必須加入適當的印表控制碼，控制印表輸出。

㈢使用表介面

以人爲出發點考慮使用者的立場，是使用者介面的重點，主要以menu-driven方式使用者易於了解，並可過濾使用者的輸入，防止程式不正常中止。圖三所示爲本系統所用的使用者介面。

國立交通大學圖書館中英文全文光碟檢索系統

〔A〕 光碟資料庫檢索系統簡介
〔B〕 系統使用說明
〔C〕 資料查詢
〔D〕 離開系統

圖三 Menu-driven之使用者介面

㈣實際製作

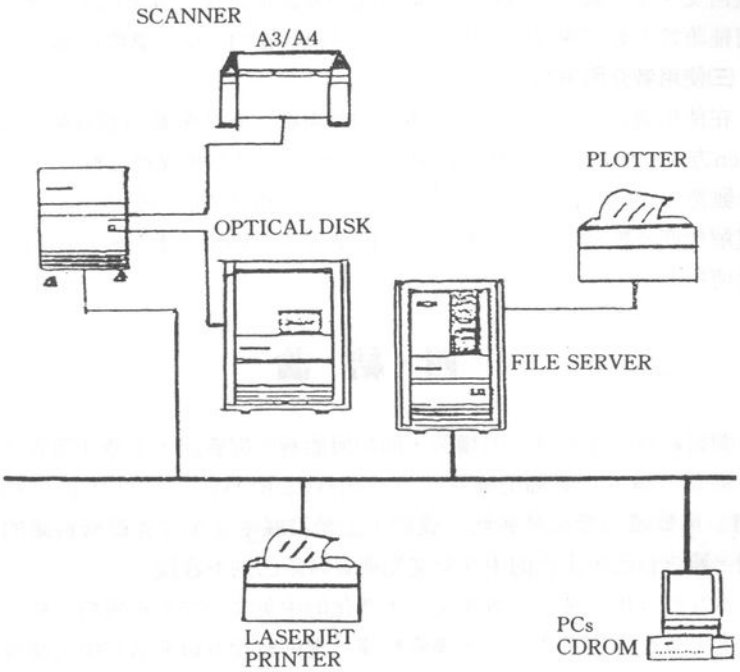
在此我們採用目前最爲廣泛使用的IBM PC作爲程式發展平台，並使用大家熟悉且具有可攜性的C語言，中文則架構在倚天中文系統下，製作出第一個可以使用中英文檢索的中英文全文光碟查詢系統。其系統架構如圖四所示：主要爲scanner、雷射印表機及個人電腦含光碟driver。

三、系統評估與討論

目前以國立交通大學博碩士論文爲樣本，已順利地寫入CDROM中，並可由一般CDROM driver來查詢，本系統具有下列優良特性，分別討論之。

㈠資料檢索方面

1. 包含所有可能的鍵值



圖四　實際中英文全文影像光碟系統硬體架構

在本系統中，使用的三種基本檢索方式：數字、英文及中文，可說是包含所有可能的檢索方式，所有須要檢索的資料都脫離不了這三種基本檢索方式的組合，因此若要將本系統應用在不同的領域資料時，只要作少許修改即可。

2. 彈性的檢索功能

在論文方面，查詢項有：系所、日期、作者、指導教授、標題關鍵字及摘要關鍵字。不但可作單項查詢，而且提供多項組合的交集查詢，倘若使用者確定查詢的資料，可鍵入完整的查詢鍵，檢出相關資料，若不確定時，也可鍵入已知的部份，仍可查出相關資料。

㈡全文閱讀方面

由於本系統對於論文本文的處理方式是以影像處理之，故不論論文中包

含任何文、字、圖、表或符號，都可完全寫眞地重現於螢幕或印表紙上，不論何種語言，如何複雜的數學式子，都可包括在內，切合實際的需求。

㈢使用者介面方面

在使用者介面上，可以說考慮的相當周延。查詢螢幕的設計使用menu-driven方式，使用者很容易學會使用，且使用上更方便找尋資料。

雖然本系統的設計已順利製作完成，但是相關的試測仍嫌不足，仍待再改進解析度及壓縮度，使達到高品質的境界，提供使用者有較佳的品質，以滿足使用者需求。

四、結 論

圖書館自動化是目前的趨勢，而如何能夠掌握資訊，有效運用資訊的工具，便是一個成功必備的條件。目前國外CDROM如：日本的硬體和美國UMI公司軟體的產品發展均已成熟，正是發展中文全文光碟資料庫的大好時機，建立自己本土化的中文全文光碟資料庫已刻不容緩。

本文設計出一種可以普遍化及大衆化的中英文全文光碟資料庫檢索系統，並進一步製作出可實用之光碟資料庫，提供教授及研究者在中文環境下檢索中英文資料。我們並提出該系統的設計方法、分析模式及實際系統製作分析。

所設計及製作出中英文全文光碟資料庫檢索系統，其優點：

㈠光碟密度高，容量大，可存大量資料。

㈡便於長久資料使用與保存，穩定性高。

㈢便於隨機彈性檢索，易於透過網路使用。

㈣資訊儲存成本低，使用範圍廣，可應用在不同領域資料之儲存。

㈤包含所有可能之查詢鍵值，易於使用。

㈥允許在中文及英文環境下全文檢索，這是中文全文光碟資料庫系統的一大突破。

由以上的優點可知，對整體中文資料的儲存、檢索均有甚大的助益。對進一步的研究，我們將再提高中文光碟的解析度品質，以較佳之壓縮技術使儲存容量增大，使檢索速度增加，以建立軟體之中文全文光碟資料庫的整體

環境，對中文資料庫的推展交流將有甚大的助益。

附 註

註一 賀立維，〈光碟與縮影的展望〉，*The Symposium on Optical Disk and Micrographic Systems*，81年12月，頁227。

光碟在政府機關應用策略之研究，行政院研考會，79年4月。

註二 吳碧娟，〈國立中央圖書館期刊論文索引光碟系統〉，**國立中央圖書館館刊**，第24卷2期(80年12月)，頁13-24。

註三 宋玉，〈自動化資訊檢索系統中索引的設計〉，**中國圖書館學會會報**，第45期(80年12月)，頁27-31。

註四 黃景星，〈光碟儲存系統簡介〉，**光電資訊**，78年3月。

陳慧芬，〈最有潛力的儲存裝置—光碟機〉，**0與1科技**，92期(78年1月)，頁253-259。

註五 同註四。

註六 Digital Equipment Corp., "Introduction to VAX/11 780 Record Management Services," order no. AA-D024B-TE.

Gio Wiederhold Inc, US Data Management System Reference, 800-1124-01.

註七 Patti-Myers, "Graphic Environment Operations for CDROM," *Optical Information System*, Sep, 1989, pp. 399-404.

R. B. Mulvany, "Engineering Design of a Disk Storage Facility with Data Modules," *IBM J. Res. Develop*, Nov, 1974, pp. 489-505.