B1 12

中文全文文件自動索引研究系統規劃

黃雲龍　博士班研究生
台灣大學商學研究所資訊管理組
f9701064@ccms.ntu.edu.tw

謝清俊　研究員
中央研究院資訊科學研究所
hsieh@sinica.edu.tw

謝清佳　副教授
台灣大學資訊管理系所
cchsieh@im.ntu.edu.tw

摘要

隨著電子文件時代的來臨，資訊的儲存、呈現、處理與交換的方式發生很大的變革。對於文件數位化以後的資料處理與應用，尤其是全文文件(full text document)的檢索，更需要研究如何善用資訊技術以協助使用者在浩瀚的資料空間裡取得文件。因此，自動化文件檢索系統(automatic document retrieval system)應運而生。本文首先回顧全文儲存及檢索系統的發展歷程，以及資訊檢索研究的議題。並以美國康乃爾大學的 SMART 系統作爲例證，說明其在資訊檢索相關研究上的貢獻與其系統的設計藍圖。最後根據三個原則：(1)中文語文的特性；(2)運用現階段中文資訊處理研究與應用發展的技術，如中央研究院中文自動斷詞技術以及中文全文檢索系統(CTP)；(3)參考向量空間模型(Vector Space Model ；簡稱 VSM)理論與初期中文全文自動索引的實驗設計。依此規劃中文全文文件自動索引研究系統的雛形環境，提供日後中文資訊檢索研究環境的發展基礎。

關鍵字：系統規劃、自動索引、資訊檢索、向量空間模型、 VSM 、 SMART

A Proposal of Automatic Indexing Research System for Mandarin Chinese Full Text Document

Yun-Long Huang
Graduate Institute of Business Administration, NTU

Ching-Chun Hsieh
Institute of Information Science, Academia Sinica

Ching-Cha Hsieh
Dept. of Information Management, National Taiwan University

ABSTRACT

In the coming of information age, accessing digitized document is a critical issue of information processing. We will develop some principles as the architecture of automatic indexing research system for mandarin Chinese full text document. First, we had reviewed the development of text storage and retrieval system and the research issues of information retrieval. We introduce the SMART project in Cornell as a exemplification and discuss its contributions in information retrieval and the blueprint of its design idea. Finally, we suggest the Vector Space Model(VSM) as the underlying theory to explore the critical issues of automatic indexing research system for mandarin Chinese full text document. We will present an initial experiment design based on VSM and utilize the technique of automatic word segment and Chinese Text Processor which had been developed by Academia Sinica.

Keywords: system planning, automatic index, information retrieval, vector space model, VSM, SMART

壹、前言

人類社會的變遷，由於資訊技術所引發的第三波革命，造就了後工業化文明的新文明，使得資訊的儲存、呈現、處理與交換的方式發生很大的變革。最根本的改變就是文件電子化，文件資料以機讀文件(machine readable document)或數位文件(digitized document)的方式展現其新的風貌。再加上資訊網路的普及，透過網際網路的通訊協定，全世界的網路連成一體，更加速的促成電子文件時代的來臨。

對於這種文件電子化以後的資料處理與應用，尤其是全文資料的檢索，更需要研究發展出一些方法在浩瀚的資料空間裡及時有效的取得資料。因此，在未來數位文件愈來愈普及，人們如何在文件系統中擷取良好品質的資訊，應是資訊資源運用的重要課題。

過去商業應用的關連式資料庫，大都是處理以關連表所記錄之格式化屬性值的資料，資料形式則包括有檔案記錄、字母與數字型資料，此類資料處理問題已經獲致相當程度的解決[26]。

企業資料中還有很多非結構化的資料，例如公文、書信、會議記錄、規章辦法、技術規範、標準手冊、筆記、計畫書、契約書、備忘錄、工作記錄、公司出版品.....，還有辦公室自動化以後的各類電子郵件、電報、傳真等等，這些全文資料的處理與應用，將是未來資料管理的重點。

本文以下將先說明自動索引與資訊檢索的

關係。接著回顧過去西文全文自動索引理論的研究與發展趨勢，簡介知名的 SMART 研究系統。然後簡述中文資訊處理的研究與發展。最後根據 VSM 模型理論，討論中文全文自動索引研究的特性，研擬初期的實驗設計，據此規劃中文自動索引研究系統的雛形環境。

貳、自動索引與資訊檢索的關係

1.全文檢索系統的發展歷史

全文儲存與檢索系統(text storage and retrieval system)的自動化起源於 1945 年，開始是以微縮片(microfilm)形式儲存，提供人類的檢索。 1960 年早期才開始出現電腦機讀格式的檢索系統。但是直至 1970 年初期系統設計仍以文件索引爲主，配合微縮片的全文，以索引詞彙做爲文件的擷取控制[36]。

真正的電腦化全文儲存與檢索系統開發於 1960 年代的中期，主要的設計目標以實驗系統爲主，例如 Salton 等人所發展的 SMART。隨後美國空軍於 1967 開發的 LITE(Legal Information Thru Electronics)系統是應用於法律文件的第一個實務系統。 1960 年代的末期，最知名的系統是應用於 IBM 大型主機上的 STAIRS(Storage and Information Retrieval System)。直至 1970 年代末期，開始有新聞全文文件的連線檢索系統應用，從此線上檢索服務開啓全文資訊檢索的成長期[36]。

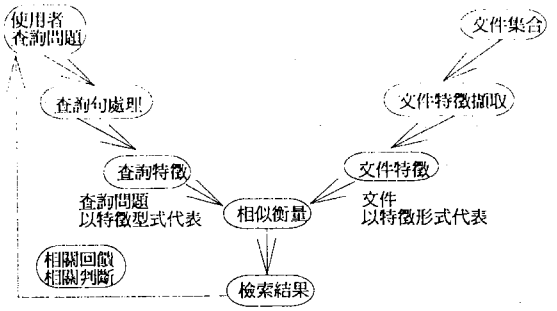
然而全文資訊檢索的問題不同於傳統資料的屬性檢索方式，它是以使用者的查詢問題與文件的相似衡量(similarity measure)結果提供資訊。因此，以下將簡述資訊檢索概念以及電子文件檢索環境，以說明自動索引與資訊檢索的關係，同時界定全文資訊檢索的研究問題。

2.資訊檢索概念與研究問題

資訊檢索乃是預先將文件按一定的方式組織和儲存(如特徵與分類)，然後使用者根據檢索的需求查出資訊的過程。因此包含了資訊的儲存與檢索。

傳統資料庫的資訊檢索，乃根據查詢問題的屬性值與系統內每一筆記錄的屬性值比對，系統將比對結果完全相同的記錄檢索出來給使用者。在全文的資料庫系統中，我們要決定與文件內容相關的屬性特徵，作爲資訊檢索過程中文件內容的識別因子(content identifiers)，例如：關鍵詞、索引詞或描述語。檢索過程系統將查詢問題轉換爲屬性特徵的代表，並且與文件特徵代表進行相似衡量，在一定的界限值(threshold)標準下，提供檢索結果。全文資訊檢索的概念如圖一所示。

所謂索引(indexing)就是在於分析文件內容、決定文件特徵，並且將文件以特徵形式代表的整個過程。而自動索引就是將上述文件儲存及索引過程自動化，希望藉由系統能夠提供使用者查詢到正確相關(relevance)的文件。因此，自動索引乃研究各種索引方法，建立良好的索引形式，將文件相關的內容以有效的



圖一 全文資訊檢索概念
資料來源：整理自Salton, G., Automatic Text Processing, ,1989.

索引形式表達於系統內部，以提昇資訊檢索的檢出率(recall)與精確率(precision)。由此可知，全文資訊檢索必須仰賴於自動索引的效果，而自動索引的效果則受到系統內部表達的效度與外部查詢者需求不確定性的限制[32]。

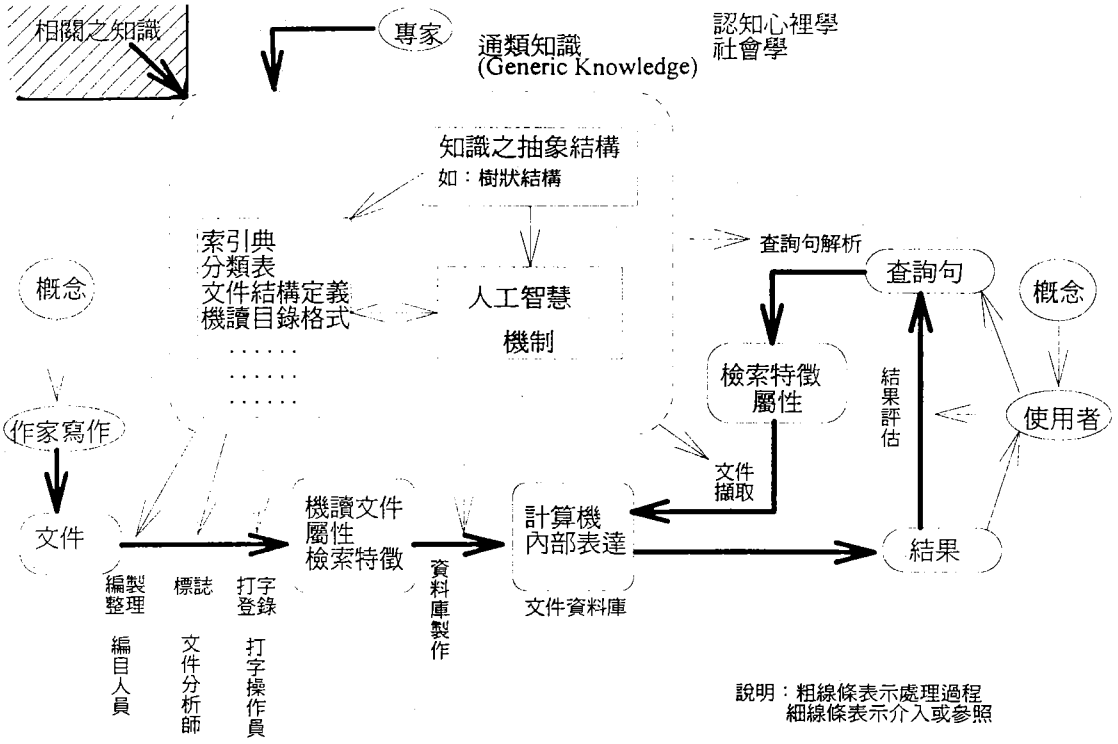
基本上，資訊檢索研究的問題是：系統如何擷取含有「符合使用者需求」的「相關資訊」的文件[43]。其中牽涉到不同使用者主觀的資訊需求，以及系統如何客觀的表達文件的內容訊息。 Lancaster(1968;引自 Salton,1986,p.651)的實驗證實，在 700,000 個文件的集合， 300 個查詢實驗中，主要的失敗來自於四項因素：(1)索引語言使用不當；(2)文件索引不當；(3)檢索問題陳述不當；(4)人機檢索介面互動不足。

3.電子文件檢索環境

過去的全文檢索方法有可以概分爲全文掃瞄與非全文掃瞄[10]。全文掃瞄法，以字串比對方式做全文的檢索，通常配合傳統布林邏輯查詢(Boolean query)，早期常見的檢索系統都利用此種方式。此種方式最大的爭議，假設使用者具有對檢索內容相當清楚的認知，而且可以用一致且代表內容概念的詞彙來發展其檢索策略。 Iivonen[27]研究證實不同檢索者對同一檢索問題轉換成布林查詢陳述時的一致性僅達 31.2 % ；而同一檢索者也會因爲檢索介面或環境的不同產生不一致的結果。

非全文掃瞄法如特徵比對法(surrogate match)、字元反轉法(character inverted)等利用建立索引的方式，可以更快速的檢索文件。目前商業化的全文檢索系統以全文掃瞄、詞彙反轉索引、配合布林查詢的設計爲主[1][24]。新近的研究以簽名檔(signature file)方式來表達文件的特徵，此法可以減少索引容量，而且結合近似查詢的方式與第二階段的全文掃瞄，達到更快速的檢索效率[14]。

另外， Jardine & Van Rijsbergen[28]根據群集假設(cluster hypothesis)：「文件之間的關連傳達了資訊需求與文件之間的相關訊息("the associations between documents convey information about the relevance of documents to requests")」，提出群集索引方法。假設使用者對於想要檢索的內容是一群相關文件的集合，透過檢索問題與群集的相似衡量檢索資料。群集法強調內容檢索方式， Voorhees[44]研究證實



圖二 電子文件檢索系統環境示意圖

資料來源：謝清俊，「語文工作與資訊發展－－從電子文件的發展談對語文研究的期盼」，當前語文問題學術研討會，民國 83 年 6 月 26 日。

群集索引與檢索是一種精確導向的設計(precision-oriented)，提供使用者更精確的檢索結果。

謝清俊[12]以圖二描繪現今的電子文件檢索系統的設計、製作與使用環境。由此系統環境可以瞭解，從文件到使用者檢索結果之間，需要涉及許多專業知識的利用與整合，反映了文件檢索系統中複雜的語意情境，同時也指引了未來檢索系統發展的趨勢，使檢索能深入文件內容知識，並與文件處理合而爲一。

因此，在自動索引研究中，必須同時考量通類知識如圖書分類、索引典以及中文構詞學(morphology)的引用，以圖二爲理想系統的設計目標，建立研究與應用的基礎環境。

參、西文全文自動索引的研究與發展

1. 自動索引理論與向量空間模型

前一節簡述了自動索引在資訊檢索過程中的角色，其重點在於經過這樣的過程，到底保存了多少文件內容的相關訊息，可以作爲使用者的資訊需求與資訊內容之間的良好的橋樑。因此，爲了使用者的資訊需求與資訊內容之間有個良好的橋樑，我們需要一套適用於各種索引系統的索引理論，提供有效的索引過程，建立電子文件檢索系統的基礎。

最早的西文索引理論起源於 1957 年 Fredrick Jonker 的研究[4]。同期有 Cleverdon 在 1950 中期至 1960 中期完成的 Cranfield 研究，針對索引系統的評估，期望找出最佳的索引語言與索引方法，其研究方法與研究結果對後來的影響相當深遠[6]。

爾後， Salton 也利用 Cranfield 研究的實驗文件爲基礎，從 1961 年起展開 SMART(System for Mechanical Analysis and Retrieval Text ；簡稱 SMART)研究計畫[23]。接著 Salton[37])提出文件的索引向量，以及測量索引形式優劣的方法，以索引詞顯著值計算結果，進行實驗檢定，揭露自動索引理論的發展方向。 Salton 建立的 VSM 在自動索引與資訊檢索研究上因此被廣泛引用[35]。

VSM 是應用矩陣代數的數學模型，在資訊檢索研究上，以上述文件索引向量表達的「索引詞--文件矩陣」爲模型的核心。 Can & Ozkarahan[18]指出 VSM 的五個主要組成包括：索引方法、文件集合、索引詞集合、索引詞加權原則以及索引詞--文件矩陣。

因此，本文在未來的系統規劃上也將以 VSM 理論爲基礎，然後配合中文自動索引的實驗設計，建立未來研究系統的雛形環境。以下將以美國康乃爾大學的 SMART 系統作爲例證，說明其在資訊檢索相關研究上的貢獻與其系統的設計藍圖。

2. SMART 的發展與貢獻

SMART 的發展可以參考表一所整理的記事年表。 SMART 並非一個單一大型的程式，而是許多組件所構成，包括文件輸入、向量輸入、搜尋程式、群集索引與輸出程式等。系統設計的早期目的是提供研究人員研究與實驗的環境； 1985 年以後則朝向多元的設計目標，以提供研究者、資料管理者與一般使用者的使用環境。

從 SMART 的例證我們可以發現以下幾個事實：(1)系統的發展時程很長；(2)技術環境的變遷很大；(3)基礎研究的規劃對長程發展的重要性；(4)研究系統的發展對於理論研究的貢獻與累積。

表一 SMART 發展的簡略年表

年代	重要記事
1961	SMART 計畫開始
1964	Time - Sharing Design
1970	IBM 360 系統建置，以 FORTRAN IV、 Assembly 、 PL/ I 爲主
1974	IBM 370 批次作業系統完成，最原始的完整版本
1980	SMART 主計畫完成，包括 索引、群集、搜尋與評估等組件
1980	開始技術革新，以 C 語言改寫系統，建置 UNIX 作業系統
1981	利用 INGRES 關連資料庫系統整合資料管理
1982	S Statistical 資料分析與統計軟體開發
1985	開放性系統架構，迄今改寫至 11 版，置於 FTP server 可自由取用

資料來源：整理自 Fox, A. E.,Technical Report 83-560 ； Buckley, C., Technical Report 85-686, < http://cs-tr.cs.cornell.edu/>,(Accessed Nov. 26,1996).

SMART 對西文資訊檢索研究發展的貢獻在於研究的累積與資源共享，從而促進理論的快速發展。利用 SMART 系統的相關研究非常多，從早期以美國哈佛大學與康乃爾大學爲主的自動索引、文件檢索、文件動態更新，到先進的索引方法如群集索引，擴充布林模型，相關回饋等[37]。同時在知名的資訊檢索研究期刊如 ： IPM(Information Processing & Management)、 JASIS (Journal of The American society for Information Science)；或資訊檢索研究論壇與研討會如 ： ACM-SIGIR Forum(Association for Computing Machinery Special Interest Group in Information Retrieval)以及新近以大型語料爲主的 TREC(Text Retrieval Conference)實驗[25]，都可以見到 SMART 的研究發表。

三十餘年來資訊檢索研究， VSM 從 1970 年代初期的 SMART 計畫中建立[40]，根據 VSM 發展的資訊檢索研究也不斷的整合新的技術與方法，加入通類知識的應用，建立系統的人工智慧機制。例如群集索引與自動分類[19][21][31]，以類神經網路處理使用者的相關回饋[45]、自然語言處理[29][33][49]、自動建立系統索引典[20][42]，以及強調以概念爲基礎的資訊檢索系統設計[47][48]。這些研究在突破傳統 VSM 的限制上有一些新觀念，因此不斷的推進資訊檢索理論發展的疆域。

3.自動索引系統的藍圖

就機器自動索引而言，Salton(1989,p.307)提出一個自動索引系統的藍圖，可以經由下列過程實現：

(1)確認文件內容的每一個詞彙(word)。

(2)利用停止索引詞表(stop list)篩選共通詞彙。

(3)去除接尾詞(suffix)以產生詞幹(word stem)。

(4)利用索引典(thesaurus)控制並替代出現頻率低的詞彙。

(5)利用詞組方式替代出現頻率高的的詞彙。

(6)計算單一索引詞、詞組及索引典的索引詞顯著值及加權值。

(7)計算查詢句與文件索引向量的相似值。

(8)使用者確認檢索結果與查詢之間的相關。

(9)根據相關衡量計算索引詞相關因子。

(10)利用相關文件內的索引詞及加權值重新建立查詢句。

(11)重複第七步驟。

從以上程序可知，前五個步驟牽涉許多語言學的問題，包括：斷詞、構詞原則、詞彙語意關係等，而且還隱含對各種專業知識領域的應用需求。這五個步驟還有很多人工處理與自動處理上的問題。

第六步驟涉及各種索引方法對索引詞彙形式衡量的理論，提供索引詞彙選取的標準。第七步驟是處理資訊檢索的核心模組。最後幾個步驟則是有關索引品質與系統效能的進一步修正，以使用者相關回饋的資訊，提供系統在一定的檢出率水準下，改善檢索結果的精確率。

肆、中文全文自動索引研究系統基礎架構

本文所擬的系統基礎架構是根據 Richardson, Jackson & Dickson[34]的想法，基礎架構 (architecture) 是一組長期發展的原則 (principles)，它是一種機制可以確認並且解決衝突，同時可以建立技術發展策略的共識。因此，本文所擬研究系統的基礎架構是一組規劃的原則，可以歸納爲以下幾點：(1)根據中文語文的特性；(2)運用現階段中文資訊處理研究與應用發展；(3) 初期中文自動索引研究規劃與 VSM 理論基礎等。

1.中文語文的特性

語言是人類進行思考與交流的工具，文字則是記錄語言的工具。但是隨著文字進一步的抽象化，文字形式會隨著語言的分化而分化，文字形式會適應各自語言的特點而發展。所以，中文直接表意文字與西文拼音文字，不論

在構字規則、字形、字音、字義、構詞規則、語法及字詞的數量上有著很大的差異。

從資訊索引的角度而言，語文形式(form)所包含的意義界定在相關事物的概念上。而語文中最小的、有意義的形式是詞素，再上一層有詞彙、詞組、句子、最後集句成話(discourse)，隨著語文形式的尺寸越長所包含的意義越多[8]。因此，由於語文本身的差異，索引的形式自然不同。

特別是在詞彙部份，中文與英文有著類似的單位，但是構成的形式與內容結構則是完全不同。最明顯的差別是在自然語言的處理中，英文詞彙的界線很容易區別，而中文則因爲字與字相連而不容易確認詞彙的界線。其他的不同如：英文的詞彙具有詞類的標示，中文則無；英文有詞幹的特殊結構，中文則無；中文有複雜的構詞規則，英文則無。對中文而言，特別是構詞規則[2]，它是中文自然語文處理以及自動索引必須參考的通類知識，並且是我們自己要建立中文資訊檢索的通類知識。

2.運用現階段中文資訊處理的研究與應用發展

從民國六十年開始的計算機中文處理的研究，由於涉及中文資訊在語言與文化上的獨特性質，同時牽涉計算機科學、語文學、認知心理學以及許多的參考學域，也是一個大型的科際組合學科[11]。以下就兩個方向作簡要說明。

(1) 中文資訊處理基礎研究

由於中文的特性，電腦斷詞是電腦能夠正確辨識人類文字意義的基礎。因爲詞彙是表達語言意義的最基本單位，因此要使電腦具有語言能力，首先電腦要能自動斷詞。自動斷詞的重點是詞的辨識，必須建立詞典與斷詞標準。斷詞問題則包括：數量詞、重疊詞、派生詞(derived word)、複合詞、歧義詞、專有名詞等如何有效自動斷詞。除了上述基本的斷詞問題，尚有不同學域或應用領域問題，仍待建立專業詞庫。

中央研究院資訊科學研究所的中文詞知識庫小組，從民國 75 年開始，由謝清俊教授發起結合計算機與語言學的中文詞知識庫計畫。目前的研究現況與應用發展方向以中文詞知識庫爲核心，發展中文語句分析、語音辨識、資訊檢索及語言學研究語料庫等。

因此，隨著資訊檢索研究與應用發展，具有自然語言處理的智慧型檢索研究將是未來的重點趨勢[14][32]。就自動索引的研究而言，在索引詞選取過程仍須參考前述 Salton 所提系統藍圖，作適度的調整。

(2) 中文全文資訊檢索應用研究

中文的全文處理包括：文件元素的辨認、文件結構的表達、全文資料庫設計、全文檢索技術以及超文件表達方式等研究。

中研院於 1984 年開始推動史籍自動化，最早完成的是二十五史資料庫(1990 年)，目前已經有總數近八千萬字的文件上線[13]。這套全文檢索系統原名爲中文全文處理系統(Chinese Text Processor ；簡稱 CTP)，本系統在儲存與檢索上充分應用文獻的結構訊息，提供自由詞檢索機制，以及多詞同時檢索等創新的貢獻[3]。

因此，目前在中文自動索引研究上可以藉助於中研院全文檢索系統，它可以提供全文儲存、檢索、語文統計以及輔助人工選取索引詞彙的各項有利工具。

3. 初期實驗設計的研擬與 VSM 理論

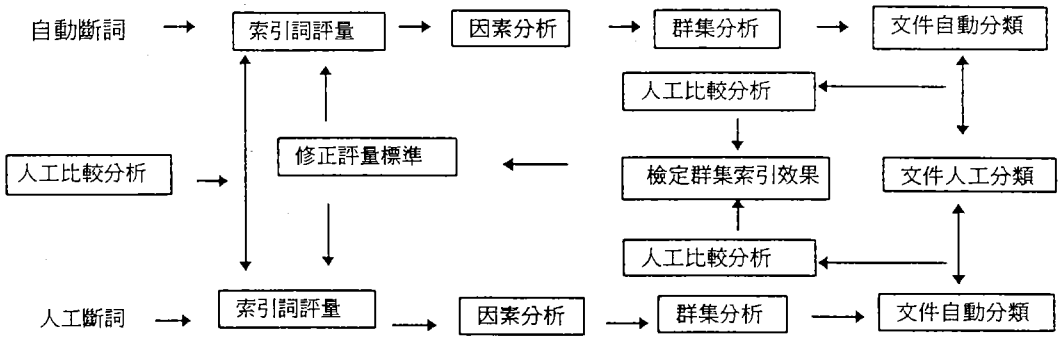
初期的實驗目的將界定於科學理論發展的研究[9]，採用文件分析與實驗室實驗等實證研究方法，探討資訊檢索的索引過程，以 VSM 爲核心，研究文件內容與索引詞彙的形式關係，建構中文全文文件的群集索引理論。

在研究過程的概念層次上，是以詮釋內容與形式的關係做深入的探索。在實證操作層次上，則分析文件內容所包含的相關資訊，以衡量索引詞彙在形式上的統計關係，來代表文件內容的主題和概念，建立文件與索引詞彙的索引關係，以 VSM 爲假設理論的測試模型，同時引用語言文字學、文獻學、新聞傳播與圖書分類學等通類知識，建立初步可接受的中文文件群集索引理論。

以下根據未來群集索引實驗設計的流程，以及前述第三節有關 VSM 的組成說明系統規劃的原則。

(1) 實驗設計的流程

本實驗依據上述實驗目的，研擬實驗流程如下圖。經由人工斷詞與自動斷詞兩個途徑，分析索引詞的差異並建立最佳索引形式的的評量標準。然後經由文件自動分類與人工分類的比較分析，檢定群集索引的效果，找出群集索



圖三 自動索引實驗設計流程

引形式的評量標準。

(2) 實驗使用的語料

從過去研究使用的實驗語料可以瞭解，利用現成的語料有相當多的利益，例如常在SMART研究中引用的Cranfield實驗文件、CACM摘要、INSPEC資料庫摘要、MEDLINE書目資料庫等[41][49]。因爲現成的語料經過專業分類具有客觀性與正確性，同時索引語言完備，而且累積了很多相關研究的知識，因此能夠完成更具研究效度的實驗。。

但是中文自動索引研究相對不足，缺乏實驗用的中文全文語料庫。雖然中研院在中文資訊處理研究上累積了很多語料，但是，有很多是古籍或圖書文件，有些斷詞語料則未經適當的分類，不適用於自動索引研究。而現成的電子文件專業語料如中央通訊社的新聞全文資料庫或國科會科資中心的論文摘要，應該是很好的大型實驗語料來源，但是涉及智慧財產權，取得成本很高，也不適用於初期的實驗。如果將來能夠有合作研究的機會，將有助於大型語料與專業語料的實驗。

本文基於在實驗中建立未來中文實驗語料的目的，經過授權取得兒童日報電子文件作爲實驗語料。而且兒童日報語料具有最基本、最常用的語料，利用這些語料進行本研究相關之實驗，可以避免太過抽象或複雜觀念的詞彙意義，有利於分析最基本語料的索引性質。

實驗語料的選擇以兒童日報新聞報導資料爲主，從82年1月起至82年12月止，選擇文教、醫療、環保及專欄等類別。表二列舉實驗語料的基本性質，以及人工選詞與自動斷詞結果在詞的組成上的差異，例如詞數與詞的長度。將來人工選詞還可以進一步作詞彙的控制，例如同義詞的權威控制。這些語料的準備仍待進一步整理。

表二 兒童日報新聞語料基本性質

新聞類別	文件數	總字數	每篇平均字數	人工選詞詞數	自動斷詞詞數	自動斷詞詞類篩選	每篇平均詞數（人工）	每篇平均詞數（自動）	自動斷詞/人工選詞比
文教	1070	391474	366	6698	17356	9871	6	16	3
醫藥	502	179450	357	2562	10300	5732	5	21	4
環保	368	141247	384	2959	9959	6033	8	27	3
專欄	393	314876	801	4915	19346	10866	13	49	4

資料來源：本研究整理

(3) 索引方法

因爲文件內容是以自然語言所構成的，索引應該是基於語言學上的分析，特別是從語意的關係來萃取文件的訊息，索引系統當然要根據語言學的分析結果來設計。但是語言學的分析方法要應用於大量的電子文件有相當的困難，因此，目前的索引方法都建立在統計的或機率的模型上。

本文將採用從統計模型中最簡單的二元加權（binary weight）方式，以索引詞彙出現或未出現在文件中爲標準，進一步分別以索引詞頻率、文件頻率、索引詞區別值[38][39]等評量方法，計算索引詞顯著值。另外，根據文件分類的結果，以機率模型中最常用的資訊量(entropy；熵)的衡量，研擬並檢定群集索引的評量標準：索引詞集中度與廣度，此一評量方法曾經在陳淑美[5]與楊允言[7]的自動分類研究中使用。

本文將經由這些索引方法的實驗，一則建立 VSM 模型的核心模組：索引詞--文件矩陣，再者建立相關研究的工具，提供未來研究的基礎環境。

(4) 索引詞--文件矩陣的數值分析

VSM模型的核心是索引詞--文件矩陣。以文件的索引向量爲基礎，提供文件在機器內的一種表達方式，利用矩陣向量的數值分析，衡量索引詞彙與文件索引向量彼此的相似性，提供資訊檢索系統設計的參考模型。例如，一個具有M篇文件和N個索引詞的VSM索引詞--文件矩陣表達方式爲：$\mathbf{D}_{m\times n}$，矩陣內的元素 w_{ij} 代表上述索引方法的加權結果。

本文建議採用 VSM 的主要原因，在於相關研究累積的索引詞--文件矩陣的數值分析技術很多，其中以統計方法的因素分析[15][30][46]、群集分析[17][22]應用爲主。此外，新近的研究如 Deerwester et al.[21]採用奇異值分解技術(Singular Value Decomposition；簡稱 SVD)，建立潛在語意索引方法(Latent Semantic Analysis)； Yang & Chute[47][48]利用線性最小平方配適的估計方法(Linear Least Squares Fit ；簡稱 LLSF)建立文件分類模型； Can[19]發展一種包含係數概念的群集方法(Cover Coefficient concept-base Clustering Methodology ；簡稱 $\mathbf{C^3M}$)。各種方法在其過去的相關研究中證實具有群集索引的效率與效果。

在初期實驗設計中，將先以統計軟體 SAS 的因素分析與群集分析軟體試驗，同時以 SVD 方法作比較。

伍、結論

在西文的研究中，常常把資訊檢索、自然語言處理以及文件自動索引等問題糾纏再一起，研究者希望提出一個統一的模型，包括在使用者介面上具有自然語言的處理能力，在機器內具有文件自動分類的索引功能，以及在檢索機制上把查詢與文件群集索引之間做最好的

對映。但是在實際的研究重點上，仍然在三者之間有所選擇。因此，本文先從自動索引出發，將來再逐步整合資訊檢索和自然語言處理的新技術，使得系統能夠具有索引、查詢、檢索以及評量等模組的完整的研究環境。

歸納本文前述的討論，建議以 VSM 理論爲基礎，考量中文語文的特性，配合中文全文自動索引研究的設計，並且善加引用中研院在自動斷詞與 CTP 的研究環境，作爲研究系統規劃的原則，據此規劃中文自動索引研究系統的雛形環境，提供日後中文資訊檢索研究環境的設計基礎。

經由上述研究系統基礎環境的建立與實驗的設計，總結未來研究應努力的方向：

1.運用各種索引方法，衡量以詞彙爲主的索引形式，建立中文全文資訊檢索設計所需的基本語文統計資料，以及索引詞的相關性質。

2.以索引詞--文件矩陣爲核心建立 VSM，運用因素分析技術，或運用其他縮減向量空間矩陣構面的數值分析模型，例如奇異值分解技術，以分析索引詞彙的關係，萃取群集索引的通類知識。

3.以索引詞--文件矩陣爲核心，同時運用群集分析技術，分析文件之間的關係，萃取分類概念結構與文件群集的知識。

陸、參考文獻

[1] 卜小蝶，圖書資訊檢索技術，文華圖書館管理資訊公司，民國 85 年 11 月。

[2] 方師鐸，國語詞彙學構詞篇，益智書局，民國 59 年。

[3] 林晰，「文獻層級結構應用於全文處理研究」，中央研究院計算機中心，技術報告，民國 79 年。

[4] 陳昭珍，「主題索引問題初探」，美國資訊科學學會台北學生分會會訊，年刊第五期，美國資訊科學學會台北學生分會，民國 81 年 6 月，頁 14-35。

[5] 陳淑美，財經新聞自動分類研究，台灣大學圖書館學研究所碩士論文，台北，民國 81 年 12 月。

[6] 黃慕萱，資訊檢索中「相關」概念之研究，台灣學生書局，民國 85 年 4 月初版。

[7] 楊允言，文件自動分類及其相似性排序，清華大學資訊科學研究所碩士論文，新竹，民國 82 年 6 月。

[8] 趙元任，「語言成分裡意義有關的程度問題」，中國現代語文學的開拓與發展：趙元任語言學論文集，袁毓林主編，清華大學出版社，1992 年 10 月，北京。

[9] 謝清佳，資訊概念架構的芻議及其應用，國立交通大學管理科學研究所博士論文，民國 78 年 1 月。

[10] 謝清俊，全文資料庫專輯，科學月刊第十九卷第四期，民國 77 年 4 月。

[11] 謝清俊，「論中文資訊處理系統的發展」，民國 78 年。改寫自「中文資訊處理學域研究現況與展望」，收錄於資訊學門，行政院國家科學委員會，民國 76 年。

[12] 謝清俊，「語文工作與資訊發展－－從電子文件的發展談對語文研究的期盼」，當前語文問題學術研討會，行政院國家科學委員會，國立台灣大學中國文學系主辦，民國 83 年 6 月 26 日，台北市。

[13] 謝清俊、林晰，「中央研究院(台北)古籍全文資料庫的發展概要」，中國古籍整理研究出版現代化國際會議，1995 年 7 月 22-24 日，北京。

[14] 簡立峰，「尋易系統(Csmart)與中文智慧型資訊檢索」，21 世紀資訊科學與技術的展望國際學術研討會，世界新聞傳播學院圖書資訊學系，國家圖書館主辦，民國 85 年 11 月 7-9 日，台北市。

[15] Borkr, H. & Bernick, M., "Automatic Document Classification", Journal of Association of Computing Machinery, Vol. 11, 1963, pp151-162.

[16] Buckley, C.,"Implementation of the SMART Information Retrieval System", Technical Report 85-686, Cornell University Department of Computer Science, Ithaca, New York, May. 1985.< http://cs-tr.cs.cornell.edu/>,(Accessed Nov. 26,1996).

[17] Burgin, R.,"The Retrieval Effectiveness of Five Clustering Algorithms as a Function of Indexing Exhaustivity", JASIS, Vol. 46, No. 8, Sep 1995, pp562-572.

[18] Can, F. & Ozkarahan, E. A., "Computation of Term/Document Discrimination Values by Use of the Cover Coefficient Concept", JASIS, Vol. 38, No. 3, May 1987, pp171-183.

[19] Can, F., "On The Efficiency of Best-Match Cluster Searches", Information Processing & Management, Vol. 30, No. 3, 1994, pp343-361.

[20] Crouch, C. J., "An Approach to The Automatic Construction of Global Thesauri", Information Processing & Management, Vol.26, No.5, 1990, p.632.

[21] Deerwester, S., Dumals, S.T., Furnas, G. W., Landauer, T. K. & Karshman, R., "Indexing by Latent Semantic Analysis", JASIS, Vol. 41, No. 6, Sep 1990, pp391-407.

[22] Everitt, B., Cluster Analysis, Halsted Press, 2nd edition, 1980.

[23] Fox, A. E., "some Considerations for Implementing the SMART Information Retrieval System Under UNIX", Technical Report 83-560, Cornell University Department of Computer Science, Ithaca, New York, Sep. 1983.< http://cs-tr.cs.cornell.edu/>,(Accessed Nov. 26,1996).

[24] Fox, A. E. & Koll, M. B., "Practical Enhanced Boolean Retrieval:Experiences with the SMART and SIRE Systems", Information Processing & Management, Vol. 24, No. 3,

1988, pp257-267.

[25] Harman, D. K.,"A Special Conference Report: The First Text Retrieval Conference(TREC-1) Rockville, MD, U.S.A., 4-6 Nov. 1992", Information Processing & Management, Vol. 29, No. 4, 1993, pp411-414.

[26] Hull, R. & King, R., "Semantic Database Modeling: Survey, Application, and Research Issues", ACM Computing Surveys, Vol. 19, No. 3, Sep. 1987, pp201-260.

[27] Iivonen, M., "Consistency in Selection of Search Concepts and Search Terms", Information Processing & Management, Vol. 31, No. 2, 1995, pp173-190.

[28] Jardine, N. & Van Rijsbergen, C. J., "The Use of Hierarchic Clustering in Information Retrieval", Information Storage & Retrieval, Vol. 7, 1971, pp217-240.

[29] Kristensen, J., "Expanding End-User's Query Statements for Free Text Searching with A Search-Aid Thesaurus", Information Processing & Management, Vol. 29, No. 6, 1993, pp733-744.

[30] Kurfeerst, M., & Asher, J. W., "A Factor Analysis of The Education Laws of Pennsylvania", Information Storage & Retrieval, Vol. 4, 1968, pp257-270.

[31] Lewis, D. D., "An Evalution of Phrasal and Clustering Representations on a Text Categorization Task", ACM-SIGIR, 1992, pp37-50.

[32] Lewis, D. D. & Sparck Jones, K.,"Natural Language Processing for Information Retrieval", Communications of the ACM, Vol. 39, No. 1, 1996, pp92-101.

[33] Lu, X., "Document Retrieval: A Structural Approach", Information Processing & Management, Vol. 26, No. 2, 1990, pp209-218.

[34] Richardson, G. L., Jackson, B. M. & Dickson, G. W.,"A Principles-Based Enterprise Architecture: Lessons From Texaco and Star Enterprise", MIS Quarterly, Dec. 1990.

[35] Riloff, E. & Hollaar, L.,"Text Databases and Information Retrieval", ACM Computing Surveys, Vol. 28. No. 1, Mar. 1996.

[36] Saffady, W., Text Storage and Retrieval Systems:A Technology Survey and Product Directory, Meckler Corporation, 1989.

[37] Salton, G., The SMART Retrieval System, Experiments in Automatic Document Processing, Prentice Hall, Inc., Englewood Cliffs, N. J., 1971.

[38] Salton, G., A Theory of Indexing, Regional Conference Series in Application Mathematics, Society for Industrial and Applied Mathematics,1975.

[39] Salton, G., Automatic Text Processing, Addison-Wesley Publishing Company,1989.

[40] Salton, G., "The Smart Document Retrieval Project", ACM-SIGIR, 1993, pp357-358.

[41] Srinivasan, P.,"Optimal Document-Indexing Vocabulary for MEDLINE", Information Processing & Management, Vol. 34, No. 5, 1996, pp503-514.

[42] Syu, I., Lang, S. D. & Deo, N.,"Incorporating Latent Semantic Indexing into a Neural Network Model for Information Retrieval", The 5th International Conference on Information and Knowledge Management, Nov. 1996.

[43] Van Rijsbergen, C. J., "Information Retrieval: New Directions: Old Solutions", ACM SIGIR, 1983, p.264.

[44] Voorhees, E. M., "The Cluster Hypothesis Revisited", ACM SIGIR, 1985, pp 188-196.

[45] Wilkinson, R. & Hingston, P., "Using The Cosine Measure in A Neural Network for Document Retrieval", ACM-SIGIR, 1991, pp202-210.

[46] Wong, S. K. M., Ziarko, W. & Wong, P. C. N., "Generalized Vector Space Model In Information Retrieval", ACM SIGIR Forum, 1985, pp18-25.

[47] Yang, Y. & Chute, C. G., "An Application of Least Squares Fit Mapping to Text Information Retrieval", ACM-SIGIR,1993, pp281-290.

[48] Yang, Y. & Chute, C. G., "An Example-Based Mapping Method for Text Categorization and Retrieval", ACM Transaction on Information Systems, Vol. 12, No. 3, jul. 1994, pp252-277.

[49] Yang, Y. & Wilbur, J., "Using Corpus Statistics to Remove Redundant Words in Text Categorization", JASIS, Vol. 47, No. 5, 1996, pp357-369.