23. A)

圖書館事業合作發展研討會

Library Cooperation & Development Seminar

August 17-18, 1986

全文圖書系統的展望

謝清俊

中央研究院計算中心

國立中央圖書館

National Central Library

全文圖書系統的展望

謝 清 俊
中央研究院計算中心主任

摘 要

本文將由資訊處理技術的角度，特別是借重全文處理的經驗，來探討未來全文圖書系統可能的發展。文分為二部份，其一為對全文圖書系統的構思：一種以全文為對象，具有文獻登錄、處理、應用和產生等完整迴路的系統架構，其次是在前述的構思下，探討涉及的各種資訊處理技術的問題，包括：文獻的登錄問題，文獻在電腦內的表達方式和相關資料的抽取問題，文獻的檢索問題，以及檢索與應用系統的結合問題等。

由本文的分析，將顯示本文中對全文圖書系統的構思在技術上是可行的。

一、前　言

擁有全電子化的圖書系統，一直是人類的夢想。遠在1945年，不虛先生(Mr. Vannevar Bush)就曾想像了一個叫Memex的系統。它雖然是一個家庭用的系統，卻能儲存約五千冊藏書的內容，並配合以既容易使用又富有創造性的軟體充份利用其典藏做查詢及研究的工作。這個夢想，雖然迄今未能實現，然而隨著科技的進步，我們可以感覺到它一步一步地接近了。

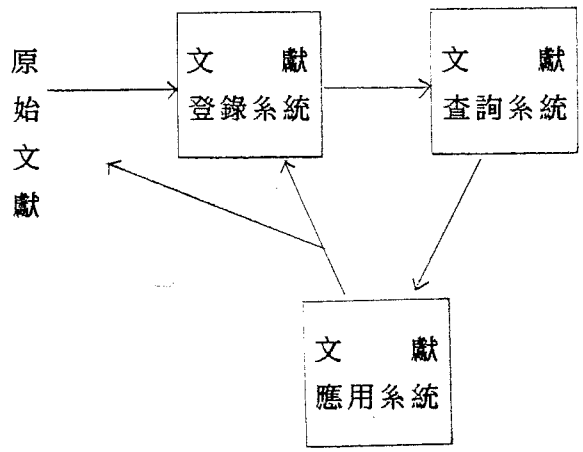
全電子化圖書系統的必要構想之一，就是要能處理典藏文獻的全部原文。我們稱之為全文處理的功能。目前，雖然還沒有全電子化的全文圖書館，可是全文資料庫的使用已日益普遍〔1〕。現有的全文資料庫之內容多集中在工具書和記錄性質的檔案上。它的查詢系統亦多沿用既有的圖書資料查詢系統〔1，2〕。雖然已發展出自由詞彙查詢(free vocaburary search)的技術，然而卻仍須借重權威檔，索引典以及對自然語文處理的能力，以提高其查詢的精確度〔1，3〕。因此，全文處理能力的開發，包括全文檢索技術，就成為全電子化圖書系統的關鍵性技術。在本文中，我們將由全文處理技術的角度，來逐項探討實施全電子化圖書系統的可行性。

全電子化圖書系統的另一個重要構想是開發相關的應用軟體，以發展為一個完整的文獻處理系統。目前，無論圖書系統或是資料庫系統,對使用者而言，它們主要的功能是查詢。以全文資料庫為例，它並不能取代現有的書籍雜誌〔1〕，它還是以一種較精密的查詢系統的身份出現的。換言之，將全文納入系統之後，應該設法發展處理全文的應用軟體，使之成為一個文獻處理的工具，而不局限於查詢。在本文中，我們將以發展文字處理功能的卡系系統(Note-card System)作案例說明。

由以上的討論我們可以明了，當全文資料納入圖書系統時，無論在系統的結構上或是系統的功能上均將做大幅度的調整與擴充。本文想探討的，就是這些問題。為便於討論，讓我們假設對圖書管理的功能－－如借還書、採購等等，暫時撇開不談，而集中於對使用者相關的問題來討論，並且這些討論是由資訊處理的角度來看的。

二． 全電子化圖書系統之構想

根據以上的構想，一個全電子化的圖書系統，應有圖一中的三大功能。



圖一：構想中的全電子圖書系統功能方塊圖

在圖一中，文獻登錄系統的功能是將原始文獻的全文，轉換為機讀形式，並能分析及認別文獻之結構，產生文獻結構之資料作為查詢及應用之所需。此外它尚能識別文獻的必要屬性資料，例如書目資料，作為系統管理、查詢或應用之需要。

與目前圖書的登錄系統比較，此系統有下列的特徵：

一、功能擴大。如：包括了全文，增加了分析的工作。

二、自動化程度提高。如：原始之書目資料寄望以機器自動地由全文中摘取。

三、操作之方式變化。如：不必用人工整理原始書目資料，但需制定標準的全文輸入或排印的規範，以便機械化的處理。

從資訊處理的角度來看，原始文獻經登錄系統處理之後，即變成計算機內部文獻表達的方式建立成為計算機內部之文獻檔案。而這些檔案，就是以下文獻查詢系統所處理之對象。

至於文獻查詢系統的功能將包含目前圖書或資料庫查詢系統的功能並將會增強。其主要原因是目前的查詢系統中沒有文獻結構的資料。有了文獻結構的資料將有助於查詢的工作。譬如：有了文獻結構的資料就可以提供更佳的搜尋範圍選擇功能(proximity functions)，而關鍵字(key words)的查詢索引亦將引起變化。

關於文獻應用系統部份，將與現有的系統差異極大。理由是很單純的，系統中多了全文資料和文獻結構資料，可做的事自然大為增加。在此中，是重要的最利用已查出之文獻資料，作為產生新文獻的基礎。此系統若能作為產生新文獻之工具，那麼就構成了圖一中之迴路。有此迴路，就可生生不息地作成處理文獻的循環。這將是全電子化圖書系統帶來的革命性改變。也是全自動化圖書系統的肇始。

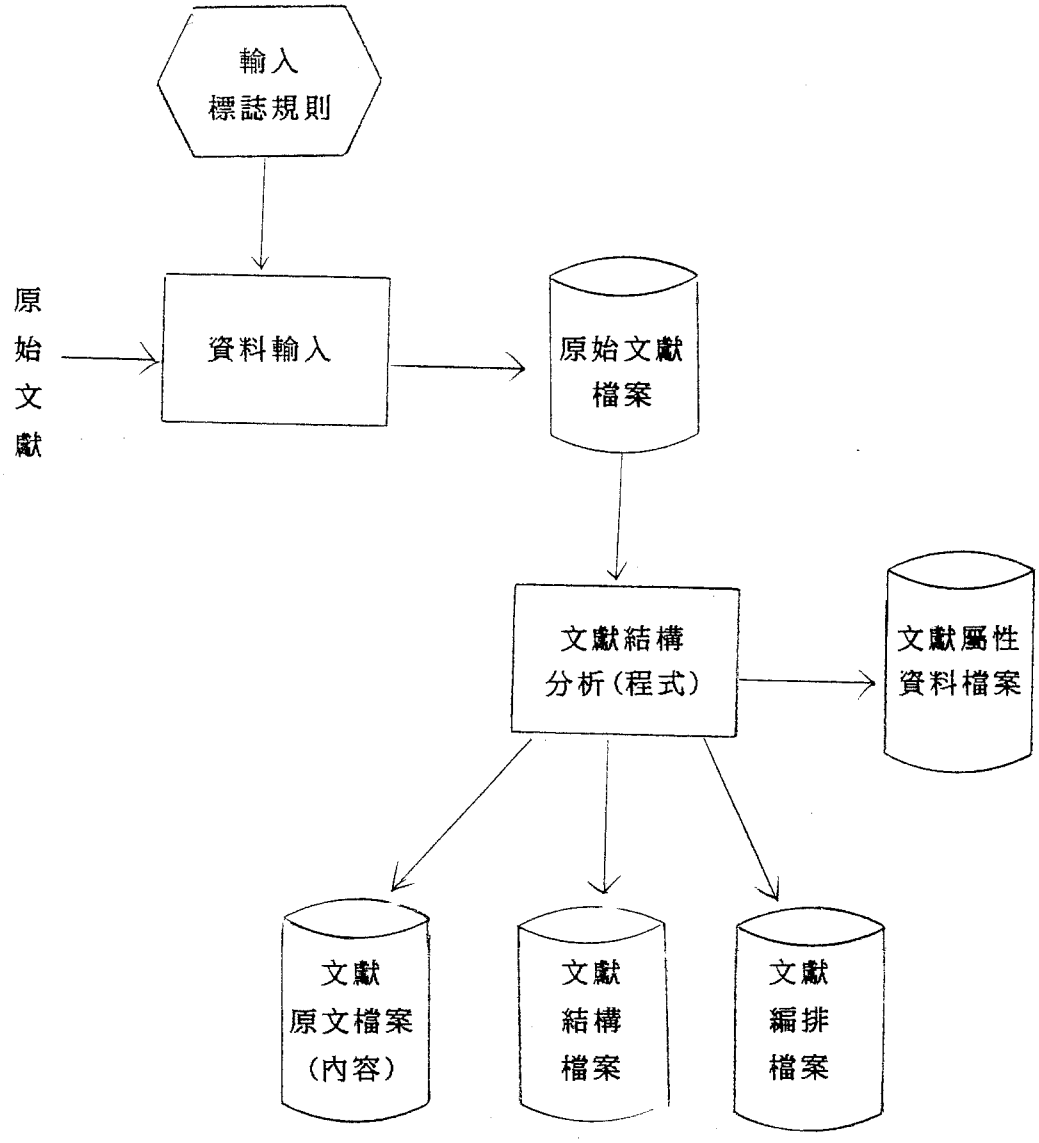
除了新文獻產生的功能以外，其他的應用也相當廣闊，例如文獻之分析、統計、索引之產生、編排、修纂等等不勝枚舉。此方面之發展，以其趨勢而言，必將與目前之文字處理、資料處理等等相結合，而走向整體資訊處理的途徑。

以目前情況而論，文獻查詢系統發展得較快速，這是與目前發展全文資料庫所需之查詢技術有關聯的。至於文獻登錄和應用二系統有待開發之處仍多，距以上所談之全電子化圖書系統之構想尚遙。然而由目前相關的研究顯示，發展全電子化圖書系統所必需之技術已逐漸成形。

以下，我們將討論重點放在文獻登錄系統上，因為此系統涉及的新技術最多；而目前的查詢系統已可做全文查詢，今後的問題只在改良，至於應用系統則變化較大，我們只以卡系作例子說明。

三．文獻之結構化

一個文獻登錄系統的方塊圖如圖二所示。



圖二：一個文獻登錄系統方塊圖

在資料輸入方面，雖然目前已在發展光學閱讀設備，但是在本文中仍以討論人工鍵入之方式為主。光學閱讀設備若發展成功，當可取代部份之人工輸入工作。

當人工輸入資料時，需遵照一輸入標誌規則，以利以下之自動化處理。此標誌規則之目的，在標準化各種文獻之表達方式以協助文獻結構化之分析工作。國際標準組(ISO)已著手於有關標準之設立〔7〕。在國內，文建會支援的史籍自動化計劃對我國正史文獻部份亦有類似之工作〔8，9〕。

當原始文獻經過資料輸入而做成原始文獻檔案之後，就可經過文獻結構分析的程式將原始文獻中所隱含之資訊包括文章之原文句、文章之架構、排版之架構，以及相關屬性資料等，個別分離出來而產生結構化的資料如圖二中的四個檔案。

在文獻原文檔案中，只有該文獻內容的資料。而這些內容資料包括原文的文字串(character/word string)，和說明資料(illustrative strings)如：圖、表、公式等。通常，原文之文字串是以段(paragraph)為最小之處理單位，在段以內之結構，如句子、詞彙則在需用時再臨時處理。

在文獻結構檔案中所存放的是該文獻的結構資料。譬如，一本書可能有序、跋、目錄、索引、章、節、小節、版權頁等等架構。在文獻編排檔案中存的是該文獻的編排資料，譬如：分頁、分行、字體、字形、圖表留的空位，每頁之上下左右邊緣等等。由這些說明可知，原始文獻經此分析後，已將原文內容、文章結構以及排版資料澈底的分離，而將原文獻模式化。此模式化最重要的作用是將文獻結構的共同部份抽取出來而做成公用的邏輯架構。若能做到這一步，許多同類型的文獻就可以共享同一個分析的程式。

文獻屬性資料的識別也是正在研究中的問題。譬如：文章標題、作者、出版處、出版年代等等之自動識別。這些識別項目中，有的容易，有的困難。若是這些資料不含在原文中，仍須以人工輸入。

總而言之，本節所討論的各問題目前仍在研究之中〔3，4，5，6，8，9，12〕。目前雖無商用的系統可以做這些事，但根據研究的結果而言，此構想的可行性是相當高的。

四． 全文的檢索技術

前文已述，目前已有可用的全文檢索技術〔1，2〕。可是這些技術所施的對象並非結構化之後的文獻。因此，若上節中文獻結構資料可資利用，則檢索之索引可指向文獻之結構，再參照到原文而不是只指向原文資料。這樣的做法將使索引更具彈性，更能配合索引典的使用。此外，文獻之結構可提供更佳之搜尋範圍的限制，亦可提供以文獻目錄或架構方面之查詢功能。

以上所說的都是待開發的研究工作，相信計算機能了解文獻結構之後，會對目前之查詢功能作更佳更有效之改良。

五． 應用之發展——個電子卡系系統的個案

當使用者找到原文資料時，目前的全文資料庫系統只能顯示其資料。若要列印則將涉及版權問題，並非都可以這麼做。就算撇開版權問題不談，列印之後對於文獻之利用還是需要電腦化的工具。為解決此問題，就非發展後繼的應用系統不可。

在本節的例子裡，我們以一個電子化的卡系系統來做說明〔9〕。此電子卡系的構想是以電腦化的檔案來模擬我們研究時常用的讀書心得卡片〔10，11〕。

假設查詢系統已經為我們找到一頁原文資料顯示在螢幕上，此時，可以經由程式將欲摘錄之文句自螢幕移到一張卡片上。電腦將自動地將文摘之出處標誌於卡片上，並將該段文字有關之關鍵詞亦隨之轉移到卡片上，以便以後作卡片之關鍵詞查詢。此外，使用者亦可將自己的心得輸入到卡片上，或將私人用的索引詞彙在卡片上建立。建好一張卡片後，電腦自動地將之儲存在卡片檔案中，而可以回到執行原文查詢之系統續繼工作。

當一疊卡片建好以後，使用者可以在螢幕上查詢卡片上之資料。必要時，仍可叫出摘錄資料之原始文獻的全文。此時，有應用軟體可以將卡片上的任何資料移到一個文字處理系統(word processor)上去，使用者可利用此文字處理系統、卡系、以及原始文獻來寫文章——作文獻產生(text generation)的工作。而這樣產生之文獻，即可自動地回　至此電子化圖書系統中，作為新的文獻。

以上所述的只是一個例子，然而由此，相信已可想像其發展之潛力。

六. 結　語

以目前研究發展的情況而言，上述之各種關鍵技術在研究室中均有良好的成效和進展，一個全電子化的圖書系統的雛型已可描述。換言之，全電子化的圖書館已具發展的潛力。以資訊立場觀之，圖書系統中轄有的典藏、文獻的資料與智識將日益增多。隨著這資訊之增加，其應用的範圍將日益擴大，將朝著整體資訊處理的方向前進。屆時圖書系統將不再像今日的系統這般孤立，與各種文獻處理的應用系統將會密切的結合。

參考資料

1. Carlo Tenopir, "Full-text database", Annual Review of Information Science and Technology, vol 19, P.215-246, ASIS, 1984.
2. Christor Faloutsos, "Access methods for text", Computing Surveys, ACM vlo 17, No.1, March 1985.
3. Arno J. H. M. Peels, Norbert J.M. Janssen, and Wop Nawijn, "Document Architecture and Text Formatting". <ACM Transactions on Office Information Systems>. vol 3, No.4, P.347-369, October 1985.
4. Kimura, Gary Dean, P. H. D. Thesis："A Structure Editor and Model for Abstract Document Objects", University of Washington, 1984.
5. List Margo, P. H. D. Dissertation "The DP Tree－A Data Structure for Multikey Retrival" University of Texas at Arlington, 1982.
6. James Clifford. "A Logical Framework for the Temporal semantics and Netual-Language Querying of Historical Database", State Univ. of N. Y. at Stony Brook.
7. ISO Draft International Standard DIP 8879. ISO TC97/SC18/WG8 , International Organization for Standardization, Geneva, Switzerland, Sept. 1984.
8. 毛漢光，"史籍自動化食貨志輸入電腦第一年總報告"，中央研究院歷史語言研究所，七月，1985.
9. "史籍自動化計畫報告"（第二期），中央研究院歷史語言研究所，八月，1986.
10. 林瑞華，"卡系與資料整理"，中央卡系統推廣中心，鳳山，高雄，台灣，中華民國，1978.
11. 梅棹忠夫原著，余阿勳、劉焜輝譯 "知識誕生的奧秘"，晨鐘出版公司，台北市，中華民國，1976.

12. Ching-Chun Hsieh, "Full Text Processing of Chinese Language—an experimental system for studying Chinese History Literatures", Annual Conference of the Association for Asian Studies, SIG Panel, Chicago, U.S.A. March 21-23, 1986.