行政院國家科學委員會專題研究計畫報告

資訊檢索中索引典系統建置之研究

計畫編號：NSC 89-2213-E-002-112
執行期限：89 年 8 月 1 日至 90 年 7 月 31 日

主 持 人：陳光華　國立臺灣大學圖書資訊學系 副教授

一、中文摘要

從 WWW 上資訊檢索應用而言，資源必須經過有系統地組織與整理，才能得到滿意的檢索結果。控制詞彙的角色在前述的背景下，扮演越來越關鍵的角色。控制詞彙通常是透過索引典或是主題表體現，提供詞彙之間的等同關係、層級關係、關連關係。藉由索引典，使用者可以瞭解資訊系統的知識結構，可以選擇適切的檢索詞彙，以獲致比較滿意的檢索結果。然而，索引典的建構與管理卻是一項困難工作，若沒有設計良好的系統，則無法有效地發揮索引典的長處。目前也沒有任何的資訊檢索系統或是數位圖書館系統整合索引典子系統，使得資訊檢索系統的效能始終侷限於求準率（Precision）不高的窘況。本計畫探討索引典的建置，發展索引典系統的功能模式，建立索引典建置與管理之模式，並實作雛型系統。

二、英文摘要

From the viewpoint of WWW-based IR researches, the metadata and resource description framework become much more important than ever before. Therefore, the application of the controlled vocabularies increasingly plays an important role. The recent literatures also report that controlled vocabularies significantly improve IR effectiveness. Thesauri and Subject Headings consist of controlled vocabularies and represent the knowledge structures in details and in various relations. However, the construction and management of thesauri is not an easy task and the integration of thesaurus subsystem into IR systems is not considered in the current IR systems or DL projects. The project will investigate the applications of thesauri and subject heading, propose a model to handle the thesauri or the subject headings in a systematic way, and implement a practical system.

三、序論

網際網路的時代使得吾人可以輕易地取得資訊，然而資訊品質則是令人質疑的重要問題，為了提升資訊的品質，各種網路的服務應運而生，包括原生型服務、加值型服務、訊息性服務。近年來，數位圖書館研究逐漸受到大家的重視，其目的是希望提供綜合型的服務，讓使用者可以透過數位圖書館系統滿足資訊的需求。因而，一個嚴謹的數位圖書館系統，在文獻資料在典藏之前，必須經過分類編目的過程，編目人員通常進行兩種不同的分析方式：一為實體分析（Physical Analysis），處理的是文獻的作者（Author）與題名（Title）等資料；另一為內容分析（Content Analysis），處理的是文獻的分類以及為該文獻設定若干標題（有時必須複分）。文獻的作者與題名是沒有任何爭議的，然而文獻的分類與標題則是必須經由編目人員一番思考，才能妥善處理的。為了達到較為一致的處理方式，圖書館文獻資料的分類編目是採權威控制的方式進行的，也就是說分類有既定的分類法，如國會圖書分類法（LCC）、杜威十進位分類法（DCC）、中文圖書分類法；標題則有既定的標題表，如國會標題表（LCSH）、醫學標題表（MeSH）、中文圖書標題表。圖書館編目館員根據前述權威控制的方式進行文獻資料的編目，而每一筆文獻資料的編目資料（是一種 Metadata，也就是詮釋資料）就成為該文獻資料可能的檢索點（Access Point），至於所使用的詞彙則被稱為索引詞彙。透過詮釋資料，可以檢索各種類型的資料，因而，詮釋資料的著錄益形重要。圖書資訊學界對於資料的著錄、組織與整理，已有長久而完整的作法，如何將之運用於資訊檢索系統，將是重要的研究課題。本計畫擬研發的索引典系統，不僅可以定義詞彙關係，輸入控制詞彙，驗證詞彙關係，瀏覽控制詞彙，查詢檢索詞彙，還可以整合於資料著錄系統與資訊檢索系統，有效提升著錄與檢索的一致性，

索引詞彙事實上扮演兩個角色，其一是讓資料著錄者選擇適當的詞彙，以表達所處理的文獻資料的主題；其二是讓資料檢索者下達適當的詞彙，以檢索經適度處理的文獻資料。索引典或是標題表承載者控制詞彙，因此必須與這二類人有效地互動，才能取得令人滿意的檢索結果。然而，除了製作精良的光碟資料庫之外，目前網際網路上的檢索系統使用索引典的情形並不多見，本計畫擬探討索引典實際應用的情形，分析索引典系統應具有的功能與規格，並實際發展一套索引典原型系統。

四、研究方法

網際網路的發展使得吾人對於資源取得的管道與以往有極大的不同，同時各種資源型式的比重也隨之不同。數位圖書館的研究在前述的背景之下受到各國政府、學術、商業等機構的重視，紛紛投入大量的人力物力。由於數位圖書館是科際性（Interdisciplinary）的研究，各領域的專家紛紛由不同的角度審視數位圖書館，例如有些學者認為數

位圖書館是資訊檢索系統的延伸，而有些學者認為數位圖書館是實體圖書館虛擬化的延伸，更有些學者認為數位圖書館就只是網路化的光碟資料庫。對於將數位圖書館視為是實體圖書館延伸的研究者而言，我們經過多次的調查與研究，得知使用者使用 OPAC 或 WebPAC 的檢索模式。例如，根據美國圖書館學會的調查，公共圖書館的讀者多數使用標題檢索，學術圖書館與專門圖書館的學者多數使用題名或著者檢索；而耶魯大學的研究顯示，56%是題名或著者檢索，而 33%是主題檢索（包括標題與分類號）。然而，隨著圖書館作業逐漸的自動化，線上查詢也成為檢索作業的主流，讀者檢索的方式也有所改變，根據台灣大學黃慕萱教授 1996 年所做的調查，有 82.5%的檢索是屬於主題檢索。由上述的數據顯示，主題檢索是資訊檢索的主要方式。數位圖書館系統透過網際網路提供各類型的資訊服務，對於資訊檢索的功能而言，主題檢索的重要性將益形重要，因為使用者更加地多元化，無法預期使用者具有何種背景知識，因而，數位圖書館的研究者與建置者必須強化主題檢索的功能，提供額外的服務子系統，有效協助使用者主題檢索的需求。

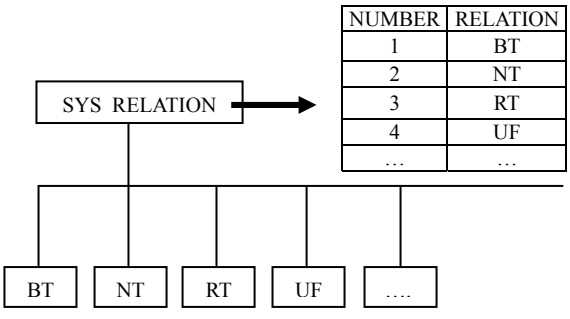
紐西蘭學者 Alastair Smith 曾經檢視 11 個數位圖書館，仔細比較各數位圖書館提供的檢索功能，多數的系統提供布林運算功能，但僅有五個提供控制詞彙的功能，而且僅有二個系統提供相關詞彙的功能。由這些情況顯示，目前的數位圖書館系統的建置，僅著重於將典藏品數位化，提供主題式的數位收藏環境，設計良好的展示網頁，結合網際網路上既有的搜尋引擎或是分類目錄，然而，並沒有實質上提升檢索服務的層次。對於透過習慣於以檢索方式取得資源的使用者而言，如何強化數位圖書館系統的檢索功能是最為關切的課題。實體圖書館對於文獻資料的主題有一套系統化的處理方式，無論是索引典或是標題表代表的是系統內部的知識結構，而個別的詞彙則代表特定的知識或概念。因此，若能使用自動、辦自動或人工的方式，賦予文獻資料適當的敘述語（Descriptor，亦即索引典或標題表使用的詞彙），而檢索系統又能有效地運用，則數位圖書館服務的層次必定能夠提升。

本研究將檢視索引典的效益，探討索引典的建置，發展索引典系統的功能模式，建立索引典建置與管理之工具，研究索引典系統與檢索系統的整合架構。具體之研究方法如下文所示。

1. 文獻分析法
   檢視索引典與標題表的相關論文，分析目前索引典相關標準，以及索引典相關的應用。
2. 系統分析法
   調查線上索引典的使用需求以及現有之瀏覽模式，研究並評估與檢索系統整合之模式。
3. 系統實作法
   基本上我們認為數位圖書館是由索引典子系統、詮釋資料著錄子系統、資訊檢索子系統、資訊典藏子系統等四個核心子系統，以及其他的子系統（如人機介面子系統）整合而成。索引典子系統必須提供控制詞彙給資料著錄子系統使用，以確定詮釋資料著錄之一致性。索引典子系統必須提供各資訊檢索子系統使用的控制詞彙，以增進資料檢索的有效性。索引典子系統必須提供系統領域知識架構，以加強使用者對數位圖書館典藏品的瞭解。在前述的考量之下，並配合當前電腦的使用環境以及網際網路以成形的標準，索引典子系統的功能與規格如下。
   - 執行於 Windows 作業系統，透過網路與其他異質性系統互動。
   - 檔案儲存格式：以 MS Access 檔案格式為內部檔案的儲存格式。
   - 檔案瀏覽型式：以首頁展現層級架構的第一層詞彙，每一詞彙皆可點選，以展現其下之第二層詞彙，同時展現詞彙的完整路徑，讓使用者藉由瀏覽方式瞭解系統的知識架構。
   - 詞彙關係的定義：除了層級關係、等同關係、關聯關係之外，權威檔案建置者可自訂詞彙關係。
   - 詞彙的輸入與編輯：輸入詞彙，並依據定義的詞彙關係，建構輸入詞彙關係。
   - 詞彙關係的驗證：檢查詞彙之間的關係是否相互衝突。
   - 多語詞彙的處理：因應網際網路的應用需求及其跨越國界的特性，必須處理多種語言的詞彙。
   - 詞彙的自動對映：提供檢索詞彙自動轉換為索引詞彙（即系統使用的權威控制詞彙）的功能。

五、研究成果

為了考量使用者的系統環境，本計畫使用*.MDB 檔案格式，使用者可以使用 ACCESS 資料庫軟體開啟本系統建構的索引典;但是使用本系統並不需要具有 ACCESS 軟體。詞彙關係的定義為資料庫的 schema，其架構方式如圖一所示。



圖一：資料庫詞彙關係架構圖

本計畫以 SYS_RELATION 資料表，紀錄該索引典為使用者所定義的關係名稱和種類(可讓使用者自

訂關係)。在詞彙關係的定義方面，有層級關係、等同關係、關連關係、權威詞彙幾種種類，以下分別討論之：

1. 層級關係：在系統中預設存在 BT（Board Term 廣義詞）、NT（Narrow Term 狹義詞）屬於層級關係，在系統中的定義，存在以下的邏輯：

- A 詞彙為 B 詞彙的 BT <=> B 詞彙為 A 詞彙的 NT
- 其關係意義為 HAS-A 的關係（A Has a BT B）（<=>：若且為若，C <=> D，若 C 存在，則必存在 D；若 D 存在，C 也必定存在）

2. 等同關係：在系統中預設存在 RT（Relational Term 等同詞），存在以下邏輯：

- A 詞彙為 B 詞彙的 RT <=> B 詞彙為 A 詞彙的 RT

3. 關連關係：在系統中沒有預設存在的關係屬於關連關係，在關連關係的資料中，系統只記錄單向的關係，也就是說關連關係存在以下的邏輯：

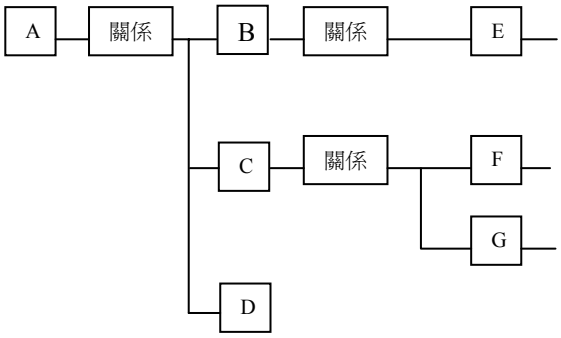
- 存在 A 詞彙為 B 詞彙的關連詞，B 詞彙不必須為 A 詞彙的關連詞

4. 權威詞彙：系統中存在 UF（Use For 權威詞）、UI（Use 見），來記錄權威詞彙，然而在系統中這兩中關係存在以下的邏輯：

- A 詞彙為 B 詞彙的 UF <=> B 詞彙為 A 詞彙的 UI
- 其關係意義為 IS-A 的關係（A Is a UF B）
- A 詞彙為 B 詞彙的 UF => A 詞彙對 B 詞彙存在 1 對多的映射關係

一個概念不能同時擁有兩個不同的權威詞，也就是說兩個權威詞所權威代表的詞群不能有交集。

除了詞彙關係的建立之外，為了確保使用者建立的詞彙與詞彙關係不可以相互矛盾，本計畫也實作詞彙與詞彙關係驗證的功能。基本上，詞彙與詞彙之間的關係，可由圖二表示。

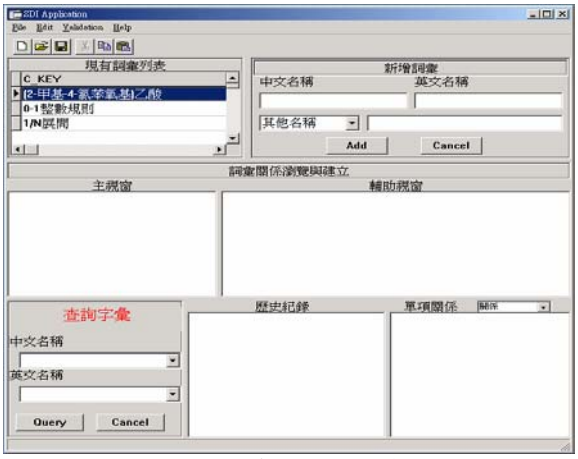


圖二：詞彙關係示意圖

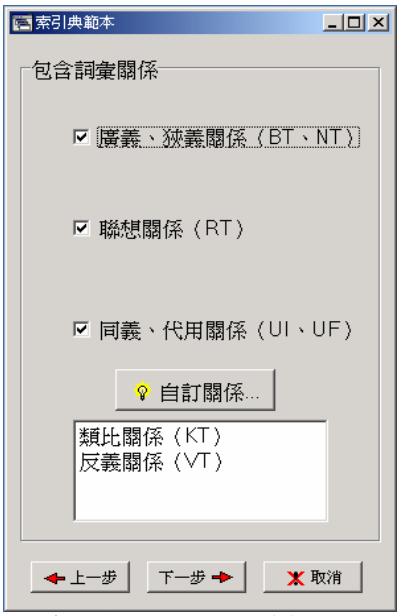
詞彙關係的驗證方式，所依循的邏輯主要來自於不同種類詞彙關係的定義，可以分為層級性的驗證、等同性的驗證、權威關係驗證三個部分：

- **層級性的驗證：**驗證關係包括：BT、NT、UF、UI。在 A 關係（例如 BT 或 NT）中，A 不能在出現在他以下的詞群中，也就是說 B、C、D、E、F、G 中不能再出現 A 詞彙。
- **等同性的驗證：**驗證關係包括：(RT、RT)、(BT、NT)、(UF、UI)，在兩兩一組的等同性驗證中必須存在以下的組合：關係甲和關係乙為等同的一對關係，詞彙 A、B 存在：A-甲-B，則必同時存在 B-乙-A 的紀錄。
- **權威性的驗證：**驗證關係包括：UF（控制詞）、UI，具備有一對控制詞和一般詞的關係，而一個一般詞最多只能擁有一個控制詞。

以下是本系統的各項功能畫面。



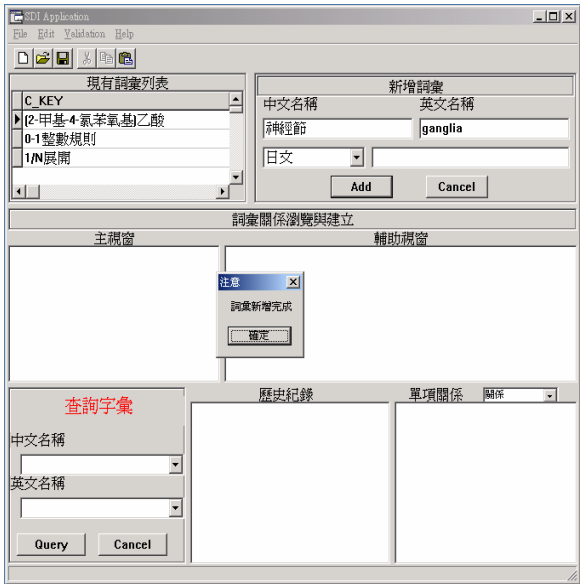
圖三：本系統的啟始畫面



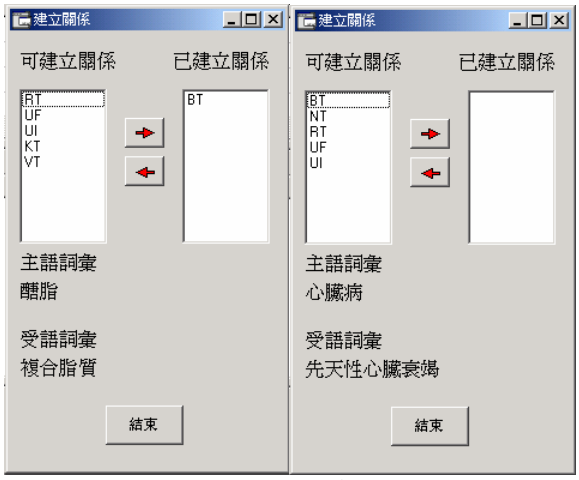
圖四：定義詞彙關係的畫面（系統預設關係）

圖五：自訂詞彙關係

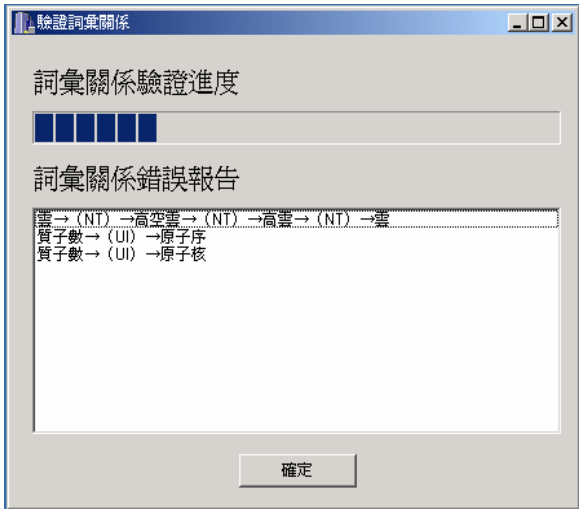
圖六：定義詞彙的語言



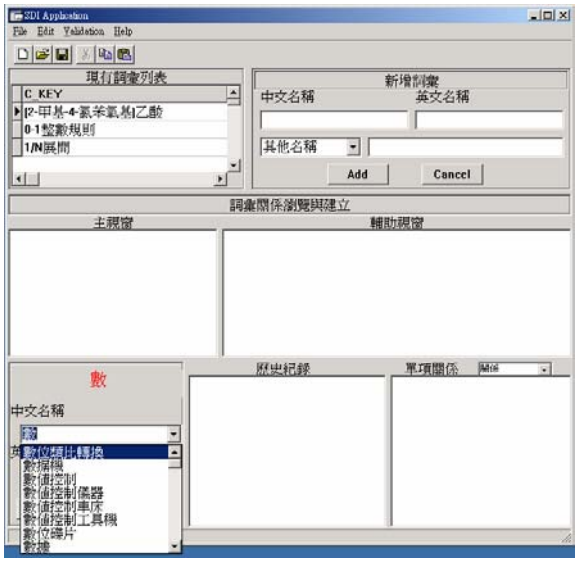
圖七：建立詞彙



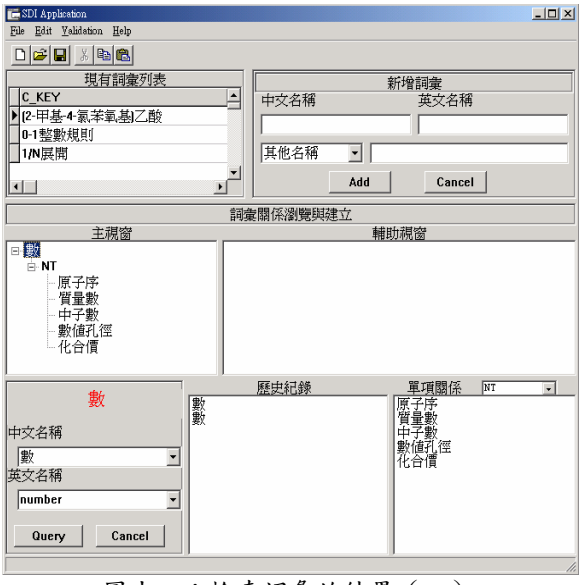
圖八：建立詞彙關係



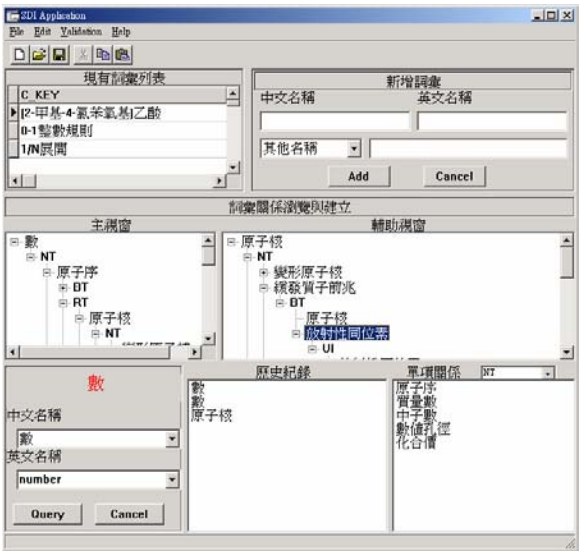
圖九：驗證詞彙關係



圖十：檢索詞彙的畫面



圖十一：檢索詞彙的結果（一）



圖十一：檢索詞彙的結果（二）

六、結論

索引詞彙事實上扮演兩個角色，其一是讓資料著錄者選擇適當的詞彙，以表達所處理的文獻資料的主題；其二是讓資料檢索者下達適當的詞彙，以檢索經適度處理的文獻資料。索引典或是標題表承載者控制詞彙，因此必須與這二類人有效地互動，才能取得令人滿意的檢索結果。本計畫已建立一個索引典建置與管理雛型系統，不僅可以讓使用者定義詞彙關係，建立詞彙，建立詞彙關係，驗證詞彙關係，也提供雙語功能，讓使用者輸入中、英語對應詞彙，甚至二種以上的語言。本計畫建立的雛型系統，可以與檢索系統結合，提供詞彙擴展的功能。

參考文獻

何光國。圖書資訊組織原理。台北市：三民，民 79 年。

陳光華。「電子文獻主題之自動辨識」。中國圖書館學會會報第 59 期(民國 86 年 12 月)，頁 43-58。

陳光華，伍健廷。「控制詞彙之自動索引」。中國圖書館學會會報第 61 期（民國 87 年 12 月），頁 81-102。

Borko, Harold and Charles L. Bernier. Indexing Concepts and Methods. New York: Academic Press, Inc., 1978.

Burgin, Robert and Dillon, Martin. " Improving disambiguation in FASIT," Journal of the American Society for Information Science 43:2 (March 1992): 101-114.

Chen, Kuang-hua. “Topic indentification in dascourse,” Proceedings of the Seventh Conference of the European Chapter of the Association for Computational Linguistics (Ireland, Dublin: Association for Computational Linguistics, 1995), 267-271.

Cheong, T. L. and Lip T. S. "A statistical approach to automatic text extraction," Asian Libraries 3:1(March 1993): 46-54.

Clarke, D. C. and Bennett, J. L. "An experimental framework for observing the indexing process," Journal of the American Society for Information Science 24:1(January/February 1973): 9-24.

Cleveland, Donald B. Introduction to Indexing and Abstracting. Littleton, Colorado: Libraries Unlimited, Inc., 1983.

Cohen, Jonathan D. " Highlights: language- and domain-independent automatic indexing terms for abstracting," Journal of the American Society for Information Science 46:3(April 1995): 162-74.

Dillon, Martin and Grar, Ann S. " FASIT: a fully automatic syntactically based indexing system," Journal of the American Society for Information Science 34:2(1983): 99-108.

Dillon, Martin and McDonald, Laura K. " Fully automatic book indexing," Journal of Documentation 39:3(September 1983): 135-154.

Dillon, Martin. "Thesaurus-based automatic book indexing," Information Processing & Management 18:4(1982): 167-178.

Dym, Eleanor D., ed. Subject and Information Analysis. New York: Marcel Dekker, Inc., 1985.

Fagan, Joel L. " The effectiveness of a nonsyntactic approach to automatic phrase indexing for document retrieval," Journal of the American Society for Information Science 40:2 (March 1989): 115-132.

Garfield, E. "The relationship between mechanical indexing, structural linguistics and information retrieval," Interlending and Document Supply 18:5(1992): 343-354.

Ginsberg, A. "A unified approach to automatic indexing and information retrieval," IEEE Expert 8(1993): 46-46.

Harter, Stephen P. "A probabilistic approach to automatic keyword indexing," Journal of the American Society for Information Science 26:4(September/October 1975): 280-289.

Hoppe, Alfred. " Communicative grammar and machine-assisted text contents analysis," International Classification 11:1(1984): 9-12.

Humphrey, Susanne M. and Miller, Nancy E. "Knowledge-based indexing of the medical literature: the indexing aid project," Journal of the American Society for Information Science 38:3(1987): 184-196.

Jones, Kevin P. "Toward a theory of indexing [Documentation notes]," Journal of Documentation 32:2(June 1976): 118-125.

Jones, Leslie P., Gassie , Edward W. and Radhakrishnan, Sridhar. "INDEX: the statistical basis for an automatic conceptual phrase-index system," Journal of the American Society for Information Science 41:2(1990): 87-97.

Leung, Chi-hong and Kan, Wing-kay. “A statistical learning approach to automatic indexing of controlled index terms,” Journal of the American

Society for Information Science 48:1 (January 1997): 55-65.

Meadow, Charles T. Text Information Retrieval Systems. San Diego: Academic Press, 1992.

O'Kane, Kevin C. "Generating hierarchical document indices from common denominators in large document collections," Information Processing & Management 32:1(1996): 105-115.

Rosenberg, V. "A study of statistical measures for predicting terms used to index documents," Journal of the American Society for Information Science 22:1(January/February 1971): 41-50.

Sabourin, C. F. "Computational linguistics in information science: information retrieval (full-text or conceptual), automatic indexing, text abstraction, content analysis, information extraction, query languages, bibliography," Journal of the American Society for Information Science 47:3 (March 1996): 247-249.

Salton, Gerard and Michael J. McGill Introduction to Modern Information Retrieval. New York: McGraw-Hill, Inc., 1983.

Salton, Gerard. " Term weighting approaches in automatic text retrieval," Information Processing & Management 24:5 (1988): 513-523.

Salton, Gerard. Automatic Text Processing: the transformation, analysis, and retrieval of information by computer. New York: Addison-Wesley Publishing Company, Inc., 1989.

Schuegraf, E. J. and Bommel, F. van. "An automatic document indexing system based on cooperating expert systems: design and development," Canadian Journal of Information and Library Science 18:2(July 1993): 32-50.

Silvester, J. P. and Klingbiel, P. H. "An operational system for subject switching between controlled vocabularies," Information Processing & Management 29:1(Jan/Feb 1993): 47-59.

Sparck Jones, Karen. "A statistical interpretation of term specificity and its application in retrieval," Journal of Documentation 28:1(1972): 11-21.

Sridhar, A. and Sreelatha, G. "Generation of descriptors for the text of a technical paper: a case study," Library Science with a Slant to Documentation 30:1(March 1993): 25-35.

Van Rijsbergen, C. J. Information Retrieval. London: Butterworth & CO Ltd, 1975.

Veenema, F. " To index or not to index," Canadian Journal of Information and Library Science 21:2(July 1996): 1-22.

Vleduts-Stokolov, Natasha. "Concept recognition in an automatic text-processing system for the life sciences," Journal of the American Society for Information Science 38:4(1987): 269-287.

Wagner, M. M. and Cooper, G. F. "Evaluation of Meta-l-based atuomatic indexing method for medical documents," Computers and Biomedical Research 25(1992): 226-350.

Wan, T.-L. et. al. “Experimetns with Automatic Indexing and a Relational thesaurus in a Chinese Information Retrieval System,” Journal of the American Society for Information Science, 48:2(1997): 1086-1096.

Wang, Y.-C., Vandendorpe, J.R. and evens, M. “Relational Tesauri in Information Retrieval,” Journal of the American Society for Information Science, 36(1985): 15-27.