

行政院國家科學委員會專題研究計畫報告

總計劃：多語言資訊檢索與擷取之研究

控制詞彙自動索引之研究(III)：資訊檢索中控制詞彙系統之研究

計畫編號：NSC90-2213-E-002-047

執行期限：90年8月1日至91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

研究助理：莊雅蓁 國立台灣大學圖書資訊學系 碩士

一、中文摘要

本計畫第一年已完成基本的控制詞彙指派機制，第二年完成分類機制的建立。第三年則利用第一年與第二年的成果，建構一個運用控制詞彙進行檢索系統。控制詞彙的使用主要為：1)取代使用者的查詢詞彙；2)擴展使用者的查詢詞彙。本計畫將著重於擴展查詢詞彙，這種作法可以一併考慮使用者的詞彙。擴展查詢詞彙的方式有：以控制詞彙擴展；以同義詞典擴展，以控制詞彙加權。除此之外，控制詞彙之間的關係亦是本研究探究的重點。系統績效的評比則是以不經任何處理的使用者查詢問句作為基線模式(Baseline Model)，比較前述擴展方式的績效，以探究控制詞彙的效用。

二、英文摘要

The first year of this project has built an indexing module; the second year has built a classification module. The report will describe the construction of a retrieval system with the capability of using controlled vocabulary. The application of controlled vocabulary is to replace or to expand the free terms submitted by users. The report focuses on the expansion of user's free terms, since this method could introduce the capabilities of free terms and controlled vocabulary. Two kind of query expansions are used in this research. The first uses controlled vocabulary only. The second uses statistical synonym first and then uses controlled vocabulary. In addition, the relations of controlled vocabulary are also considered in the research, e.g., the broad term, narrow term, and related term. A baseline model is used to compare the performance of query expansion.

三、序論

傳統上，資訊檢索的研究著眼於自然語言的自動索引，也因此索引模式的重心是如何識別出索引文件中代表主題的關鍵詞(keyword)或是關鍵片語(key phrase)，而其方法自然是透過計算的方式，將重要性最高的關鍵詞彙或關鍵片語挑出來作為索引詞彙。相對來說，以控制詞彙為主的自動索引研究就少了許多。其實上述的關鍵字與關鍵片語並不能代表檢索者真正欲檢索的「概念」，只是這些詞彙字面上的意義可能符合檢索者的檢索概念而已，因而這類研究訓練所得的知識，並

不能稱為真正的概念索引。從事控制詞彙索引時，索引者必須把文章的主題轉譯成控制詞彙所提示的概念，因此控制詞彙才是概念索引，比較符合檢索者概念檢索的需求。除此之外，自1970年代末期以來，由於資料庫的迅速發展，但求全率愈高，求準率不見得提高，因此自然語言索引所引起的求全率與求準率不平衡的問題，再度促使資訊界重視控制詞彙索引法。傳統的控制詞彙索引除了求全率不如自然語言索引外，最為人詬病的，就是不符經濟效益。Sparck Jones認為在一個檢索系統中，假如利用更多的索引詞彙來增加文件描述的完整性(exhaustivity)，文件被檢索出的機率就會增加。通常在同時擁有自然與控制詞彙索引的資料庫中，一篇文件的控制詞彙索引數量，通常遠比自然語彙的關鍵字索引數量少，所以控制詞彙索引的求全率不如自然語言索引的現象是可以理解的。如果可以透過某種方式自動產生控制詞彙索引，解決控制詞彙索引因為成本較高導致數量較少的問題，不但可以保有控制詞彙求準率較高的優點，又可以藉由提高文件索引的完整性，來解決求全率較低的問題。

本計畫第一年已完成控制詞彙的自動指派機制；第二年完成的控制詞彙的自動分類機制。本年度則將建構一個控制詞彙輔助系統，探討其對於資訊檢索績效的影響。我們將以擴展使用者查詢詞彙的方式，如以控制詞彙擴展；以同義詞典擴展，以控制詞彙加權，探討各種不同關係的控制詞彙，其影響檢索績效的程度。

四、研究方法

為了分析控制詞彙對於檢索效益之影響，因此本研究設計以下三種不同的實驗：

- ◆基礎檢索(Baseline)
- ◆以控制詞彙擴展
- ◆以同義詞典擴展，以控制詞彙修正

其中，基礎檢索是以使用者所建構之原始查詢問句直接進行檢索，所得之檢索結果做為評比的基準，分析不同的詞彙資源以及不同的詞彙擴展設計對於檢索效益提升之成效。以控制詞彙進行擴展之實驗，可再細分為數個以不同詞彙關係擴展查詢問句之實驗模組，主要目的除了比較控制詞彙對於檢索效益提升之貢獻程度之外，亦欲了解不同詞彙關係

對於檢索結果的個別影響 以同義詞典擴展再以控制詞彙加權之實驗，是以同義詞典擴展之後，將新的查詢問句利用控制詞彙以詞彙加權的方式修正檢索詞彙，以減低同義詞典建構時可能包含之雜訊的干擾程度。

本研究設計之各項實驗，目的在於尋找最佳之查詢問句擴展模型，希望使用可能提升檢索效益的各種方法，建構整體上表現最好的查詢問句擴展之組合。對於各查詢擴展形式的檢索效益，本研究不採用常用的 11-Point Precision 評估方式，因為求全率是實際檢出的文件數目與文件集中所有真正相關文件數目的比率，故實驗語料本身必須具備標準答案，但本研究使用的語料缺乏此條件，無法計算求全率，而求準率的定義是指實際檢出的文件之中有多少比率是真正相關的文件，本研究可在取得檢索結果之相關候選文件之後，以人為判斷檢索出的文件是否與查詢問句相關，再計算求準率，因此本研究以求準率做為檢索效益的指標。

1. 基線檢索模式 (Baseline Model)

基礎檢索進行檢索之步驟為：

- (1) 將原始查詢問句(Q_o)中的檢索詞彙進行 bi-gram 處理，轉換為查詢問句向量。
- (2) 將查詢問句向量與文件向量進行相關係數 ($Sim(Q,D)$)計算。
- (3) 將檢出文件以 $Sim(Q,D)$ 相關係數值排序，選取前 50 篇進行相關判斷
- (4) 依據相關判斷結果計算求準率以評估檢索效益。

2. 以控制詞彙擴展

檢索進行時，系統首先將使用者的查詢問句與控制詞彙進行比對，再依據比對結果搜尋控制詞彙的不同關係之相關詞彙，再將該相關詞彙加入原查詢問句進行擴展。以控制詞彙擴展的實驗中，每一查詢問句以五種不同的詞彙關係進行擴展：

- ◆ 狹義詞擴展 (Narrower Term, NT)
- ◆ 等同詞擴展 (USE, USE FOR)
- ◆ 關聯詞擴展 (Related Term, RT)
- ◆ 廣義詞擴展 (Broader Term, BT)
- ◆ 聯合擴展 (NT, BT, RT, USE)

等同詞擴展的目標是包含控制詞彙中敘述同一概念的所有可能詞彙。相同地，其他以不同詞彙關係擴展的實驗依次為利用狹義詞、關聯詞以及廣義詞等詞彙關係進行擴展。最後的聯合擴展則將所有不同詞彙關係的擴展以聯集方式整合。擴展階層以兩層為限，亦即本研究只取用距離檢索詞彙之深度向上及向下各為二層以內的詞彙。對於一些檢索詞彙，有些詞彙關係並不存在，如同詞、狹義詞、廣義詞以及關聯詞，故可能有些查詢問句的某種詞彙關係擴展結果是無擴展的狀態。

以控制詞彙進行查詢問句擴展之步驟為：

- (1) 將原始查詢問句(Q_o)中之檢索詞彙，一一與控制詞彙進行字串比對。

- (2) 於搜尋相符之控制詞彙之後，依據實驗設計，搜尋與該詞彙具某種詞彙關係之相關詞彙，將之加入原查詢問句，形成以控制詞彙擴展後之查詢問句(Q_T)。
- (3) 將擴展後之查詢問句(Q_T)中的檢索詞彙進行 bi-gram 處理，轉換為查詢問句向量。
- (4) 將查詢問句向量與文件向量進行相關係數 ($Sim(Q,D)$) 計算。
- (5) 將檢出文件以 $Sim(Q,D)$ 相關係數值排序，選取前 50 篇進行相關判斷。
- (6) 依據相關判斷結果計算求準率以評估檢索效益。

3. 以同義詞典擴展，以控制詞彙加權

本實驗利用控制詞彙修正同義詞典擴展之後可能產生的雜訊，其假設為，若由同義詞典擴展後之查詢問句中的檢索詞彙亦為控制詞彙，則可保證該詞彙之正確性，而不屬於控制詞彙的其他檢索詞彙有可能是同義詞典自動建構中所包含的雜訊，因此應對正確的檢索詞彙加權以加強其重要性，相對地也減低雜訊可能的干擾程度。對於詞彙的加權值須如何設定，目前仍未有可供依循的標準，故本研究將檢索詞彙之權重初步設定為：原查詢問句與擴展後查詢問句之所有檢索詞彙權重皆為 1，可於索引典中對映成功之檢索詞彙加權後之權重為 2。以同義詞典進行查詢問句擴展，再以控制詞彙修正之步驟為：

- (1) 將原始查詢問句(Q_o)中之檢索詞彙，一一於同義詞典中進行字串比對。
- (2) 於同義詞典中搜尋相符詞彙之後，將該詞彙所屬之詞彙群中其他成員全部加入原查詢問句，形成以同義詞典擴展後之查詢問句(Q_s)。
- (3) 將 Q_s 中所有檢索詞彙，一一與控制詞彙進行字串比對。
- (4) 若 Q_s 中之檢索詞彙亦為控制詞彙中，則該檢索詞彙之權重為 2。
- (5) 將以同義詞典擴展並以控制詞彙修正後之查詢問句(Q_{SM})中的檢索詞彙進行 bi-gram 處理，轉換為查詢問句向量。
- (6) 將查詢問句向量與文件向量進行相關係數 ($Sim(Q,D)$)計算。
- (7) 將檢出文件以 $Sim(Q,D)$ 相關係數值排序，選取前 50 篇進行相關判斷。
- (8) 依據相關判斷結果計算求準率以評估檢索效益。

五、實驗材料與評估準則

本研究使用的文件為國科會科資料中心提供之「中華民國科技研究報告摘要資料庫」，收錄國內大專院校、研究機構與公民營企業等單位資助之研究計劃成果報告摘要。所使用的資料內容為民國七十一年起至民國八十年間提出並執行完成之研究計畫成果報告之書目摘要，共計 11,875 筆。文

件長度約 200 至 500 字不等，文件類別分為數理、工程、醫學、農業及人文社會等五大類。

為了進行效能的評估，實驗使用的查詢問句均經特別設計。目前無論是網際網路上的搜尋引擎使用者，或是專門資料庫檢索系統之使用者，大部分以多個檢索詞彙來表示查詢的主題。本研究則以多詞組合方式表達查詢問句，以充分表達使用者資訊需求的三至五個檢索詞彙，來建構檢索實驗之原始查詢問句。為了避免查詢問句涵蓋類別之文件數量過低，導致無效的檢索結果，我們依據實驗文件資料庫中文件類別分佈情形由各學科領域之大學部及研究所學生提供特定檢索主題，再以三個以上的詞彙表達檢索主題，組成一原始查詢問句。表一為原始查詢問句之檢索詞彙與主題敘述。

由於實驗環境並非既有的測試文件資料，除了缺乏實驗所需的查詢問句外，亦缺乏評量檢索效益的標準答案，因此無法計算求全率。故本研究在取得檢索結果之相關候選文件之後，以人為判斷檢索出的文件是否與查詢問句相關，再計算求準率做為檢索效益評估的依據。本研究以主題相關為相關判斷之原則，採用多位相關判斷者執行相關判斷。評估方法是將檢索實驗後之檢索結果，以人工判斷的方式挑選出相關的文件，亦即實驗檢出的文件中，有多少文件是相關判斷者認為相關之文件。相關判斷者對每篇文件需給予分數，以表示文件與查詢問句的相關程度，再依此計算各檢索結果的求準率。

給分的標準為：相關判斷者認為相關的文件，分數為檢索系統所計算的相關係數值 $\text{Sim}(Q,D)$ ，因為相關係數反映出文件與查詢問句的相關程度；反之，判斷者認為不相關的文件，分數一律為 0。

對於某一查詢問句以某種形式擴展後之檢索結果的 50 篇文件，三位相關判斷者所認為相關的文件不盡相同，可能有意見一致的部分，亦有不一致的部分。然而每位相關判斷者對於各檢索結果的檢索效益分數有相等的貢獻，故取三位相關判斷者之判斷結果的平均數，做為各檢索結果之檢索效益分數。以下為求準率 (P) 的計算公式：

$$P = \frac{[R_1] + [R_2] + \dots + [R_m]}{n \times m}$$

其中，n 表示進行判斷之相關候選文件數，m 表示相關判斷人數， R_m 表示每位相關判斷者給予每一篇相關候選文件分數的總和。

六、結果與分析

經由分析前述各項實驗結果，各種查詢問句擴展的方式的效能說明如下。

(一) 字串比對方式與檢索效益分析

本研究進行查詢問句擴展實驗，必須以搜尋的方式找出與原始查詢問句相同之詞彙，再擷取與該詞彙相關聯之其他詞彙加入原查詢問句以進行擴展。為考量不同的字串比對方式對查詢問句擴展之檢

表一、原始查詢問句

QUERY 1	主題敘述	探討有關建築物結構之防震或耐震設計
	檢索詞彙	建築物,結構體,防震,耐震
QUERY 2	主題敘述	染色體分析對於診斷先天畸形,智能不足,習慣性流產,惡性腫瘤等疾病之研究
	檢索詞彙	染色體,癌,流產,畸形
QUERY 3	主題敘述	探討防洪工程與洪水預報作業所必須之降雨量與河川流量變化觀測,逕流分析等技術
	檢索詞彙	洪水,預報,降雨,河川流量,逕流
QUERY 4	主題敘述	探討有關紅樹林沼澤生態環境之研究,包含紅樹林區中各別生物生態之研究
	檢索詞彙	紅樹林,沼澤,生態
QUERY 5	主題敘述	有關以射出成型法製作塑膠製品之相關材料及技術的研究
	檢索詞彙	射出成型,塑膠,塑膠加工,模具,拔模
QUERY 6	主題敘述	有關電力系統的控制與穩定技術研究,包括軟硬體
	檢索詞彙	電力系統,穩定器,控制器
QUERY 7	主題敘述	有關電腦輔助教學多媒體的應用與軟體設計
	檢索詞彙	電腦輔助,教學,多媒體,軟體,
QUERY 8	主題敘述	有關台灣地區颱風行進路徑及強度等預報之研究
	檢索詞彙	颱風,天氣預報,風速,路徑
QUERY 9	主題敘述	有關廢水及重金屬污泥之處理技術
	檢索詞彙	廢水,重金屬,污泥,危害性廢棄物,
QUERY 10	主題敘述	有關流動流體之對流熱傳遞,增強熱傳性能之研究,任何對流系統組成因素(包括流體種類,流動型態,界面形狀,方向性等)對提升熱傳遞性能的研究。
	檢索詞彙	流動流體,熱傳遞,對流
QUERY 11	主題敘述	機器人或機械手臂控制系統設計之研究
	檢索詞彙	機器人,座標系,自由度,伺服
QUERY 12	主題敘述	食物營養成分對動物體內脂質含量(如血清膽固醇量,血清脂質,肝脂質等)或膽固醇代謝之影響
	檢索詞彙	膽固醇,脂質,脂質含量

索效益可能的影響，故分別以「部分字串比對」及「完全字串比對」兩種不同的方式進行控制詞彙擴展實驗。部分字串比對的原則為被搜尋詞彙之部分字串與原詞彙相同者，即被視為比對成功。

舉例說明，原詞彙為「塑膠」，以部分字串比對方式則「工程塑膠」、「塑膠薄膜」、「塑膠」皆視為比對成功。完全字串比對的方式則是指被搜尋詞彙與原詞彙必須完全相同才視之為比對成功，以相同的例子來看，只有「塑膠」才會被擷取。由檢索結果大致趨勢分析可知，完全字串比對較部分字串比對有較高的檢索效益。

(二) 控制詞彙擴展

我們使用控制詞彙做了 10 種不同形式的查詢問句擴展。在不同擴展模組檢索結果的相對比較之下，以整體大致上的趨勢分析，部分字串比對以廣義詞的擴展表現最佳，於 12 個查詢問句中有 9 題的檢索效益最高，狹義詞、等同詞及關聯詞擴展之檢索效益相同，較廣義詞略差，最後為聯合擴展。

完全字串比對以聯合擴展之檢索效益較高，於 12 個查詢問句中有 5 題的檢索效益最高，其次依序為關聯詞、廣義詞、狹義詞，最後為等同詞。以完全字串比對的方式擴展則會出現無擴展的情形，其中以等同詞較多無擴展的結果，狹義詞則皆有擴展。

部分字串比對的結果，出現多種詞彙關係擴展的檢索效益相同的情形。12 題查詢問句中，有 7 題四種以上的詞彙關係擴展檢索效益完全相同，此可能因為以部分字串比對的方式擴展的詞彙數量較為龐大，且如之前所提，於擴展的第一步驟就把第二步驟以詞彙關係擷取出來的詞彙全都先取出，亦即部分字串比對的擴展不只擴展某單一詞彙關係，而是已包含各種不同詞彙關係之詞彙，使得擴展後的結果近似於聯合擴展。再分別觀察各詞彙關係所擴展的檢索詞彙，最少有 1 個詞彙，最多有 51 個詞彙與其他詞彙關係的擴展結果不同，占擴展後檢索詞彙數量的 5%~50%。雖然不同詞彙關係擴展後的詞彙組合差異很大，但因為這些與其他詞彙關係擴展結果不同的詞彙，並不存在於文件資料庫，在檢索時文件與查詢問句相關係數的計算上，也被忽略不予處理，所以這些不同的詞彙即使數量多卻沒有影響檢索結果的能力，因此會發生檢出的文件相同、文件相關係數相同、檢出文件排序相同等情形，使得檢索效益亦相同。表二是控制詞彙各詞彙關係擴展查詢問句之檢索結果。

(三) 同義詞典擴展，控制詞彙加權

以完全字串比對方式利用同義詞典第二層詞彙群，將原始查詢問句擴展後，每個檢索詞彙之權重皆相等，設定權重為 1，將擴展後之查詢問句以完全字串比對方式搜尋控制詞彙，比對成功之詞彙即將其權重調整為 2，再以加權後之查詢問句進行檢索。

同義詞典擴展再利用控制詞彙加權之實驗，目的在探討控制詞彙是否能降低同義詞典所包含的

雜訊對檢索的不良影響，故以第四組實驗結果與第二組實驗中第二層擴展的實驗結果相互比較分析：12 題查詢問句其中 8 題加權之後的檢索效益有所提升。

與基礎檢索比較，則 12 題查詢問句中，有 8 題高於基礎檢索之檢索效益。相較於以同義詞典第二層詞彙群擴展之實驗結果，12 題中只有 5 題的檢索效益比基礎檢索為佳，經由控制詞彙的加權、調整檢索詞彙權重之後，即增加為有 8 題的檢索效益高於基礎檢索。由該角度分析得知，利用控制詞彙修正以同義詞典擴展可能的雜訊，對檢索效益有正面的提升。表三是以同義詞典擴展、再以控制詞彙加權之檢索結果，以求準率比較檢索效益。

七、結論與建議

(一) 結論

查詢問句擴展之實驗，係針對具有較多文件數量的文件類別設計查詢問句，但由於實驗文件資料庫文件類別分佈過於分散，且涵蓋範圍廣大，即使某一類別的文件數量到達二三百篇，相對於同一主題的文件數量仍嫌太少。故查詢問句的限制較大，導致查詢問句數量略嫌不足，只能針對個別情形探討可能因素，而無法以足夠的樣本進行實驗結果的整體性趨勢分析。

本研究以字串比對的方式搜尋控制詞彙，故分別以「部分字串比對」及「完全字串比對」兩種不同的方式進行控制詞彙擴展實驗，由檢索結果大致趨勢分析可得，完全字串比對較部分字串比對有較高的檢索效益。

以各種詞彙關係擴展查詢問句，各種詞彙關係擴展之檢索效益並沒有顯著的差異，但以整體的趨勢分析，部分字串比對以廣義詞的擴展表現最佳，完全字串比對以聯合擴展表現最佳。以狹義詞擴展的檢索結果，與其他詞彙關係比較之下都較差，因狹義詞的主題概念更分歧，反而模糊了擴展後查詢問句的檢索主題。使用部分字串比對方式擴展後查詢問句的詞彙數量皆十分龐大。實驗結果顯示，擴展後詞彙數量過多，其檢索效益有明顯地降低情形。

(二) 建議

各種詞彙資源需要建立適當的整合模式才能發揮輔助檢索的功能，本研究探討詞彙資源整合的形式以及詞彙擴展的模式，若能進一步深入研究，尋求最佳的整合方式，可做為未來擴展查詢問句的重要參考。一般整合不同詞彙資源，通常先建構分屬不同資源的詞彙間互動與連結的方式，可能的方式有聯合(Union)連結(Chaining)與對映(Mapping)聯合為將原始查詢問句分別在不同資源進行查詢問句擴展後，再整合所有出現於各組合中的所有詞彙，形成擴展後的查詢問句。聯合有兩種方法可結合不同資源：加權與不加權。加權指各組合皆附帶權重，兩組詞彙交錯者其權重相加。連結策略則依次使用不同的資源，即原始查詢問句先利用一資源

擴展,將已進行一次擴展的查詢問句利用另一資源再次進行擴展 對映則將原始查詢問句分別利用不同資源進行擴展,再擷取二組擴展後之查詢問句重

疊的部分,以完成查詢問句的擴展。以上述三種方式為基礎,還能發展出更多不同的整合模式,值得進一步探討。

表二、以索引典擴展之檢索結果

	基礎檢索	部分字串比對					完全字串比對				
		狹義詞	等同詞	關聯詞	廣義詞	聯合	狹義詞	等同詞	關聯詞	廣義詞	聯合
QUERY 1	0.0801	0.0583	0.0583	0.0414	0.1056	0.0583	0.0529	無擴展	0.0917	0.0889	0.0635
QUERY 2	0.0931	0.0541	0.0541	0.0541	0.0541	0.0125	0.0969	0.0991	0.0988	0.1070	0.0979
QUERY 3	0.0864	0.0944	0.0944	0.0944	0.0944	0.0944	0.0930	無擴展	0.0942	0.0861	0.0984
QUERY 4	0.0345	0.0338	0.0338	0.0439	0.0450	0.0424	0.0460	無擴展	0.0479	0.0462	0.0387
QUERY 5	0.0399	0.0407	0.0415	0.0415	0.0407	0.0070	0.0415	0.0372	0.0357	0.0391	0.0416
QUERY 6	0.0456	0.0460	0.0460	0.0460	0.0460	0.0447	0.0479	0.0455	無擴展	0.0469	0.0483
QUERY 7	0.0511	0.0451	0.0455	0.0455	0.0451	0.0461	0.0386	0.0448	0.0436	無擴展	0.0386
QUERY 8	0.0530	0.0408	0.0408	0.0408	0.0408	0.0408	0.0453	0.0422	0.0452	0.0447	0.0455
QUERY 9	0.0512	0.0386	0.0386	0.0386	0.0386	0.0235	0.0396	0.0345	0.0478	0.0482	0.0429
QUERY 10	0.0610	0.1056	0.1056	0.1056	0.1056	0.1029	0.1006	0.0936	0.0839	0.0778	0.0998
QUERY 11	0.0806	0.1127	0.1107	0.1107	0.1126	0.1107	0.0880	無擴展	無擴展	0.0806	0.0884
QUERY 12	0.1361	0.0822	0.0822	0.0822	0.0822	0.0822	0.0837	0.0841	0.0902	0.0868	0.0859

表三、以同義詞典擴展再以控制詞彙加權之檢索結果

	基礎檢索	同義詞典擴展 (完全字串比對)	以同義詞典擴展 再以控制詞彙加權
QUERY 1	0.0801	0.0791	0.0514
QUERY 2	0.0931	0.1081	0.0877
QUERY 3	0.0864	0.0957	0.0773
QUERY 4	0.0345	0.0605	0.0559
QUERY 5	0.0399	0.0413	0.0609
QUERY 6	0.0456	0.0450	0.1001
QUERY 7	0.0511	0.0448	0.0986
QUERY 8	0.0530	0.0501	0.1124
QUERY 9	0.0512	0.0471	0.1107
QUERY 10	0.0610	0.0946	0.1067
QUERY 11	0.0806	0.0692	0.0942
QUERY 12	0.1361	0.0816	0.0893

參考文獻

- Bodner, R. C. and Song, Fei. "Knowledge-Based Approaches to Query Expansion in Information Retrieval" In Advances in Artificial Intelligence (New York: Springer, 1996): 151-152.
- Borko, Harold and Bernier, Charles L. 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, Hsinchun; Ng, Tobun D.; Martinez, Joanne and Schatz, Bruce R. "A Concept Space Approach to Addressing the Vocabulary Problem in Scientific Information Retrieval: An Experiment on the Worm Community System." Journal of the American Society for Information Science 48, no.1 (1997): 17-31.
- 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, M. and McDonald, L. 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.
- 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.
- Gauch, Susan and Smith, John B. "An Expert System for Automatic Query Reformation," Journal of the American Society for Information Science 44, no.3 (1993) : 133.
- Ginsberg, A. "A unified approach to automatic indexing and information retrieval," IEEE Expert 8(1993): 46-46.
- Harman, D. "Towards Interactive Query Expansion," In Proceedings of the 11th Annual International ACM-SIGIR Conference on Research & Development in Information Retrieval, Grenoble, France, 1988, 322-323.
- 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.
- Magennis, Mark. "Expert Rule-based Query Expansion."(1995) <<http://www.dcs.gla.ac.uk/ir/publications/papers/Proscript/magennis95.p.s.gz>> (Accessed: 30 May 2000).
- Meadow, Charles T. Text Information Retrieval Systems. San Diego: Academic Press, 1992.
- Mitra, Mandar; Singhal, Amit and Buckley, Chris. "Improving Automatic Query Expansion," In Proceedings of the 21st Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval Melbourne, Australia, August 24-28, 1998, 206-207.
- O'Kane, Kevin C. "Generating hierarchical document indices from common denominators in large document collections," Information Processing & Management 32:1(1996): 105-115.
- Peat, H.J. and Willett, P. "The Limitations of Term Co-occurrence Data for Query Expansion in Document Retrieval Systems," Journal of the American Society for Information Science 42, no.5(1991): 379-380.
- Qiu, Yonggang and Frei, H.P. "Concept Based Query Expansion," In Proceedings of the 16th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval Pittsburgh, PA, USA, Jun 27-Jul 1 1993 , 160-161.
- 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.
- 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 Van Bommel, F. "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-I-based automatic indexing method for medical documents," Computers and Biomedical Research 25(1992): 226-350.
- 陳光華，江玉婷。「中文資訊檢索測試集之設計與製作」。資訊傳播與圖書館學 6 : 3 (民國 89 年 3 月): 頁 61-80。