圖書與資訊學刊
24(Feb '98)44-68
ISSN 1023-2125

中文全文文件群集索引理論研究與實證

A Theoretic and Empirical Research of Cluster Indexing for Mandarin Chinese Full Text Document

黃 雲 龍
Yun-long Huang

國立政治大學圖書資訊學研究所
Graduate Institute of Library and Information Science, National Chengchi University

【摘要 Abstract】

當前商業應用的全文檢索系統仍以字串比對的全文檢視法，配合布林查詢介面爲主流，這種系統過於簡化電子文件檢索系統環境的形式與內容關係。本研究根據向量空間模型（VSM），探討索引詞彙的形式與文件內容關係，運用奇異值分解技術（SVD），建構中文全文文件的群集索引模型（CIM）。本文從兒童日報全文語料庫中選取醫藥新聞502篇文件，經由各項實驗設計和步驟致以下結論：CIM索引的效果優於傳統VSM，而且可以提昇其效能，達到具有權威控制機制下的索引效果。

Since most popular commercialized systems for full text document retrieval are designed with full text scanning and Boolean logic query mode. These systems use an oversimplified relationship between the indexing form and the content of document. We use Singular Value Decomposition (SVD) try to develop a Cluster Indexing Model (CIM) based on Vector Space Model (VSM) in order to explore the index theory of cluster indexing for Chinese full text document. Test corpus was selected from Children's Daily News : the medicine news (MED) with 502 documents. Under a series of experiments, the following conclusions are discovered : we find the indexing performance of CIM is better than traditional VSM, and has almost equivalent effectiveness of the authority control of index terms.

關鍵詞 Keyword

自動索引 群集索引 資訊檢索 向量空間模型 群集索引模型 奇異值分解

Automatic Indexing, Cluster Indexing, Information Retrieval, Vector Space Model, VSM, Cluster Index Model, CIM, Singular Value Decomposition, SVD

壹、前　言

由於資訊技術的應用，使得資訊的儲存、呈現、處理與交換的方式發生很大的變革。特別是文件電子化以後，對於非結構化的全文文件需要新的全文資料庫技術、方法與模型，以解決全文文件處理、應用與管理問題。

過去的全文檢索方法有可以概分爲全文檢視（full text scan）與非全文檢視。全文檢視法，以字串比對（string matching）方式做全文的檢索，通常配合傳統英文布林邏輯查詢（Boolean query）。此種方式有很大的爭議，那就是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研究證實不同檢索者（intersearcher）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（intrasearcher）也會因爲檢索介面或環境的不同而產生不一致的結果①。

非全文檢視法如特徵比對法（surrogate match）、字元反轉法（character inverted）利用建立索引方式，雖然可以提昇檢索的速度，但是索引的容量卻很龐大。新近的研究以簽名檔（signature file）方式來表達文件的特徵，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②。

雖然當前商業應用的全文檢索系統仍以字串比對的全文檢視法，或以非全文檢視法如特徵比對法、字元反轉法配合布林查詢介面爲主流。但是這種系統過於簡化電子文件檢索系統環境的形式與內容關係。因此先進的資訊檢索研究如群集法（clustering）都強調內容或概念檢索方式，這種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索引的相似衡量（similarity measure）來檢索資料，能夠提供使用者更精確的檢索結果。這種以內容或概念檢索方式是新近研究的主題，也是本文研究群集索引的動機。

回顧過去中文資訊檢索的發展，在自動索引上的研究很少。但是中央研究院資訊科學研究所的中文詞知識庫小組，長期在中文資訊處理基礎研究上所得到的成果，已經可以支援許多先進的中文資訊處理應用研究的發展。

同時中央研究院於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），目前已經有總數近一億一千萬字的文件上線③。這套中文全文檢索系統（CTP/FTMS）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④。

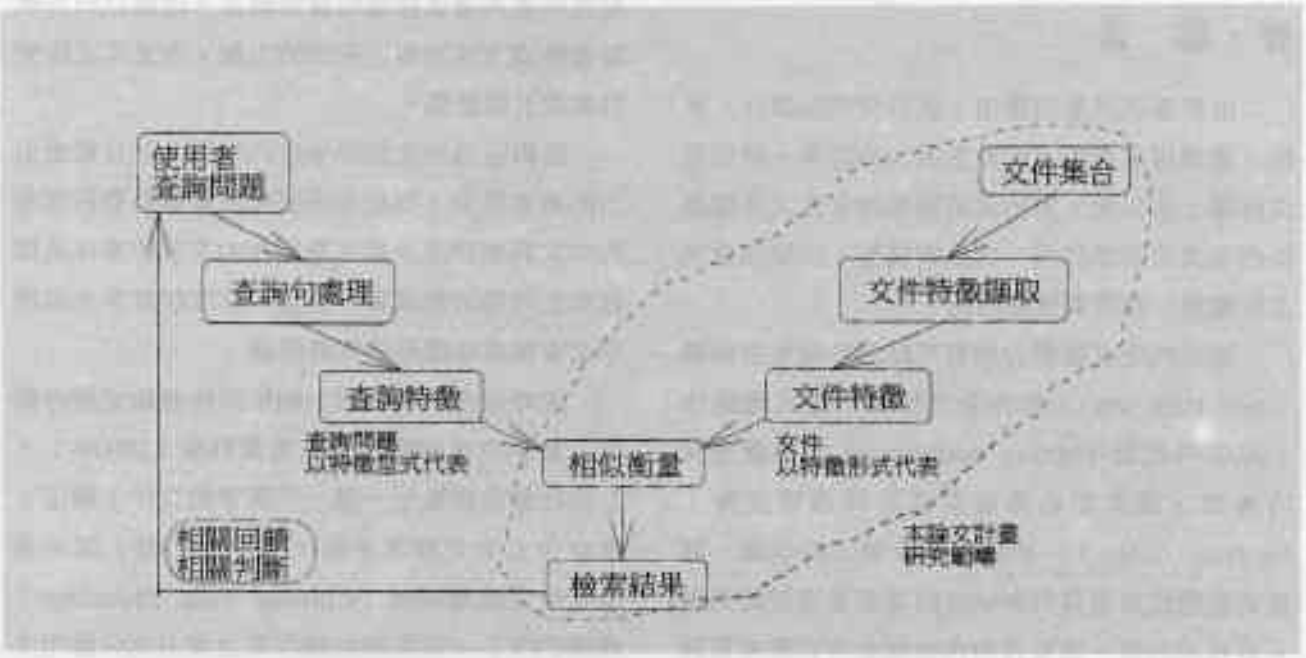
因此，在中文自動索引研究上可以藉助於中央研究院全文檢索系統，它可以提供全文儲存、檢索、語文統計以及輔助人工選取索引詞彙的各項有利工具。本文基於中文語文的特色，連結中央研究院中文資訊處理基礎與應用研究的成果，嘗試開發中文全文文件群集索引理論的研究。

貳、全文資訊檢索研究的問題

以下將分別先從資訊檢索概念界定本論文研究範疇，並說明電子文件檢索系統環境複雜的語意關係。同時從語言、文字與認知、概念的「形式與內容關係」詮釋本文的問題本質。然後說明本文嘗試建構的實證操作模型等方面討論本研究的問題。

一、論文研究範疇

從資訊檢索概念（如圖一）可以簡單區分爲四個主要的研究領域，包括：查詢介面、檢索引擎、自動索引與相關評量。本文研究範疇界定於



圖一：全文資訊檢索概念圖

資料來源：整理自謝清俊，「中文資訊處理專題研討課程」課堂筆記，（台灣大學 民國83年11月11日）。

自動索引理論（圖中虛線部份）。

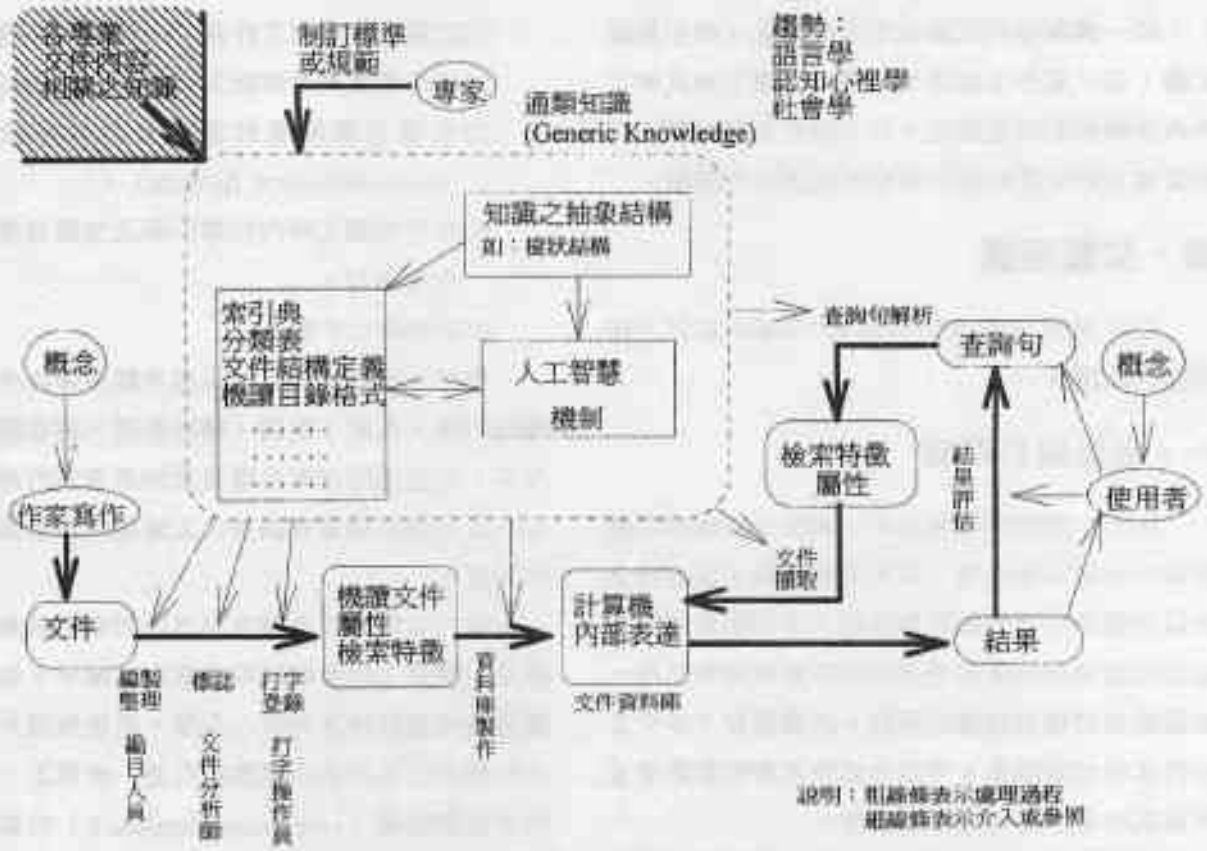
所謂索引（indexing）就是在於分析文件內容、決定文件特徵，並且將文件以特徵形式代表的整個過程。索引的目的是希望系統能夠提供使用者查詢到正確相關（relevance）的文件。因此，自動索引乃研究各種索引方法，建立良好的索引形式，將文件相關的內容以有效的索引形式表達於系統內部，以提昇資訊檢索的檢出率（recall）與精確率（precision）。

群集索引的優點是減少傳統反轉索引（inverted indexes）方式的龐大索引空間，同時提昇檢索系統的效率與效能⑤。因此群集索引是本文研究的重點，包括：群集索引構面（cluster indexing factor）的建立以及群集索引的相關評量。

二、電子文件檢索系統環境

觀察現今的電子文件檢索系統的設計、製作與使用環境（如圖二）⑥，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，以及多重的形式與內容關係。

在本文的研究中，雖然仍以建構在索引詞彙為基礎的檢索機制上，但是必須同時考量通類知識如新聞分類、索引典以及中文構詞學（morphology）的引用，建立未來中文資訊處理研究與應用的基礎環境。



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」，在當前語文問題學術研討會論文集，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。

三、形式與內容關係

認知、概念、思考和語言文字是人類藉以構成資訊檢索形式與內容關係的關鍵變數。本研究的構成概念（construct）就是基於語言文字的形式與文件內容的關係所建構而成，希望透過索引詞彙的形式衡量與組合，在操作層次能夠找出代表文件內容的群集索引構面，驗證群集索引的可行性。

廣義的認知是指所有形式的認識作用，包括：感覺、知覺、注意、記憶、推論、想像、預期、計畫、決定、問題解決及思想溝通⑦。思考可以視爲一種認知過程，它是人類最複雜的行爲方式，而概念是認知最重要的單位，也是認知後抽象信息的表達⑧。

我們所使用的語言、文字，皆爲概念的一種形式表達。由於中文是一種表意的文字，「詞彙（word）是最小的、能夠獨立運用的、有意義的語言單位」⑨。詞彙的意義代表人在思考時所呈現的概念，因此概念是詞彙的內容，而詞彙既是概念的形式，也是語言的基本形式，所以詞彙與概念有著形式與內容的對應關係。但是詞彙與概念之間並非一對一的絕對關係，而且會隨著時空改變。黃董秋歸納概念與詞彙的對應關係，有些時候一個概念可以用一個詞表示，有些概念就需要多個詞彙（如複合詞）。不同的詞可以表示

同一概念（如人名、字號或不同民族的不同表示），同一個詞也可以表示不同的概念（如引伸或比喻）⑩。基於上述研究範疇以及索引形式與文件內容關係的問題描述，以下將在文獻回顧中分別簡述本研究實証操作模型的相關研究議題。

參、文獻回顧

以下針對自動索引理論與VSM等兩個方向進行文獻探討。

一、自動索引理論

在上一節的問題陳述中，說明了所謂索引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。索引的目的是希望系統能夠提供使用者查詢到正確相關的文件。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。

爲使索引過程更爲有效，當然需要一套適用的索引理論，提供資訊檢索系統設計、發展與應用。這其中牽涉許多有關索引內容與方法的問題，而本文將專注於機器自動索引系統的問題深入探討。

就機器自動索引而言，Salton提出一個自動索引系統的藍圖，可以經由下列過程實現⑪：

㈠確認文件內容的每一個詞彙（word）。

㈡利用停止索引詞表（stop list）排去共通詞彙（common words）。

㈢去除接尾詞（suffix）以產生詞幹（word stem）。

㈣利用索引典（thesaurus）控制並替代出現頻率低的詞彙。

㈤利用詞組（term phrases）方式替代出現頻率高的的詞彙。

㈥計算單一索引詞、詞組及索引典的索引詞顯著值及加權值。

㈦計算查詢句與文件索引向量的相似值。

㈧使用者確認檢索結果與查詢之間的相關。

㈨根據相關衡量計算索引詞相關因子（term-relevance factors）。

㈩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引擎。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（relevance feedback）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。

本文嘗試解決的問題包含在前七個步驟內，這也只是部份的資訊檢索的索引理論。Salton提出文件的索引向量（index vector），以及測量索引形式優劣的方法，以索引詞顯著值（term significance）計算結果，進行實驗檢定，揭露自動索引理論的發展方向⑫。

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，目前的自動索引理論或方法都建立在統計的或機率的模型上。以下就分別討論這兩種方法。

㈠統計的模型

基於索引的目的，如何選擇合適的索引詞彙

來代表文件內容？直覺的想法是該詞彙有沒有出現在文件內，如果某個詞彙出現次數很高，它也許是一個很重要的識別因子。但是如果該詞彙同時出現在很多文件內，相對的其價值可能就不如出現在較少文件內的詞彙。因此，統計模型的索引方法分別有以下幾種。

1.詞彙頻率（term frequency；簡寫爲tf）

Salton提出文件的索引向量概念，以索引詞顯著值計算結果，進行索引方法的實驗檢定。從最簡單的二元加權（binary weight；簡寫爲BIN）方式，以衡量該詞彙有無出現在文件內來選取索引詞⑫。進一步以詞彙出現的頻率（次數）作爲判斷的標準⑬，如下列(A)式，f_i^k是索引詞k在第i個文件的出現頻率，是索引詞k在文件集合內的總次數。

$$F^k = \sum_{i=1}^{n} f_i^k \quad \cdots\cdots(A)$$

但是詞彙頻率的方法只顧及檢索的檢出率，忽略了精確率的目標。因爲頻率高的詞彙可能在每個文件中出現，則其所代表的資訊量就相對的少於出現在較少文件的詞彙，所以其精確率將受到影響。於是研究者考慮以文件頻率（document frequency；簡寫爲df）來修正此一缺失。

2.文件頻率的修正因子

文件頻率df_j定義爲：「索引詞T_j出現在總數爲N個文件集合的文件次數」。Jones提出以文件頻率的倒數（Inverse Document Frequency；簡稱爲IDF）爲修正索引詞彙顯著值的修正因子⑭。

$$IDF = \log(N / df_j) \quad \cdots\cdots(B)$$

此時整合詞彙頻率與文件頻率的加權方式，以兩者的相乘（tf×IDF）結果代表索引詞彙T_j在文件D_i的顯著值，見下列(C)式。在前述Salton實驗中，tf×IDF的索引結果，在一定檢出率水準下，精確率均優於二元加權或詞彙頻率的方式⑮。

$$w_{ij} = tf_{ij} \cdot \log \frac{N}{df_j} \quad \cdots\cdots(C)$$

另外一種常被引用的方法稱爲索引詞區別值（Term Discrimination Value；簡稱爲TDV）。

3.索引詞區別值

Bonwit & Tonsman根據索引過程的假設，如果在文件空間中，代表文件的索引向量，彼此的相似值很接近，則文件會糾纏聚集在一起⑯。因此文件之間將無法區別，這樣的結果不利於檢索；反之，如果文件是散佈在文件空間中，彼此分離的情況將有利於檢索。在索引過程中，索引詞將根據其區別文件的效果，被指定到文件的索引向量中。

也就是說，文件之間越相近則會糾纏聚集在一起，則文件空間的密度就會越高。我們假設$s(D_i, D_j)$代表文件i與j的相似值，文件空間密度（Q）則以所有文件之間相似值的平均值來計算，如(D)式。

$$Q = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ i \neq j}}^{N} s(D_i, D_j) \quad \cdots\cdots(D)$$

索引詞區別值則以空間密度的改變來計算，假設Q_j與Q分別代表詞彙T_j被指定爲索引詞及刪除之後的空間密度，則詞彙T_j的區別值如(E)式：

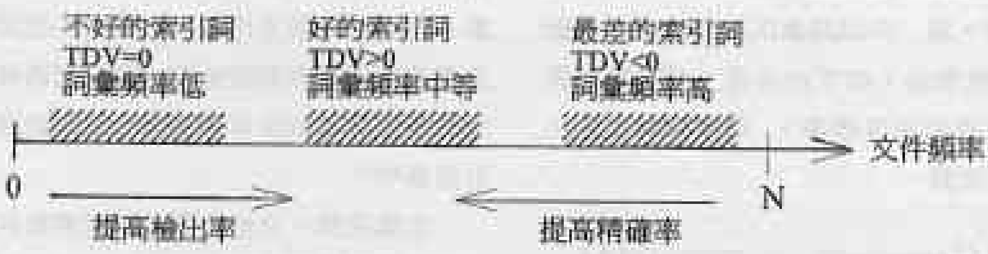
$$TDV_j = Q - Q_j \quad \cdots\cdots(E)$$

如果$TDV_j > 0$，代表沒有詞彙T_j的時候文件糾纏聚集，空間密度高。而指定詞彙T_j爲索引詞的時候文件分離，空間密度低。因此，詞彙T_j是好的索引詞；反之$TDV_j < 0$，詞彙T_j是不好的索引詞。因此，索引詞彙T_j在文件D_i的顯著值計算可以修正如下：

$w_{ij} = tf_{ij} \cdot TDV_j$ ……………………………(F)

根據Salton實驗，雖然索引的結果稍微低於(C)式的效果⑲。但是，詞彙的區別值、詞彙頻率與文件頻率之間有一些關係存在。隨著詞彙頻率與文件頻率的增加，詞彙區別值從零到正值，然後逆轉爲負值。

詞彙的區別值與詞彙頻率之間的現象，提供我們更深入於語言學構詞分析的思考方向。而文件頻率的增加也不是全然代表內容訊息的增加。因此Salton建議，將頻率低的詞彙以詞組（phrase）方式，改善系統的檢出率。然後將頻率高的詞彙以索引典（thesaurus）⑳的方式，改善系統的精確率（如圖三）。



圖三：索引詞特性與文件頻率關係

資料來源：G. Salton, 'A Theory of Indexing', Regional Conference Series in Application Mathematics (Society for Industrial and Applied Mathematics, 1975), p.43.

Salton透過實驗證實，根據索引詞區別值索引方法⑳，以詞組㉑加上索引典的索引效能，比詞彙頻率或文件頻率的修正結果高出20%。

(二)機率的模型

上述有關索引詞顯著值的計算，提供我們選擇好的索引詞的一些方法。但是，這些方法並未提及或區分索引詞彙在相關文件與不相關文件中的特性與差異。機率模型的索引方法就是要闡明索引詞彙在相關文件中出現，以及在不相關文件中出現的意義。機率模型在資訊檢索的應用，從資訊量的衡量開始。

1.資訊量的衡量

Shannon最早將entropy概念引進傳播理論中，它是以隨機的事件機率（ p_i ）估計傳輸過程中接收端可能獲得的平均資訊量。原來只是衡量傳輸過程中各種符號（symbol）的統計結構，作爲傳輸系統壓縮、預測、編碼的設計參考㉒。定義如下：

$$E(p_1, p_2, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log_2 p_i \quad \text{(bits)} \cdots\cdots\cdots (G)$$

根據entropy衡量資訊概念，也可以導出類似IDF加權方式，稱之爲訊號雜訊比值（Signal－Noise Ratio；簡寫爲S/N）。S/N與IDF的基本觀念很接近，假設索引詞只出現在少數的文件內，呈現不均勻的分佈狀況，則該詞彙應該是具有鑑別文件能力的好索引詞。Salton定義如下㉓：

$$S^k = \log F^k - N^k \quad \cdots\cdots (H)$$

S^k 表示索引詞 k 在文件中所擁有的內容訊息，F^k 是索引詞 k 在文件集合內的總次數，N^k 代表索引詞 k 在索引文件內呈現平均分佈時的雜訊值，其定義與entropy衡量的㈠式相同：

$$N^k = \sum_{i=1}^{n} \frac{f_i^k}{F^k} \log \frac{F^k}{f_i^k} \quad \cdots\cdots (I)$$

當索引詞正好出現在每一個文件內，且 $f_i^k = 1$，$F^k = n$，則雜訊最大 $N^k = \log n$，$S^k = 0$；反之，若是索引詞 k 只出現在一個文件內，則雜訊值爲零 $S^k = \log F^k - N^k = \log F^k$ 所以，訊號雜訊比值可以修正詞彙頻率的索引效果：

$$w_{ik} = tf_{ik} \cdot S^k \quad \cdots\cdots (J)$$

但是，Wong & Yao指出S/N的計算如㈠式，受到詞彙頻率影響太大，總頻率高者可以得到高的加權結果㉔。如果是以標題或摘要爲主的索引系統中，此種計算方式更突顯出其缺失。所以目前的實驗結果，S/N的索引效果不如IDF方法。

2.相關機率的加權

假設有一個文件的索引向量是 $x = (x_1, x_2, \ldots, x_t)$，$x_i$ 表示個別出現在文件 x 的詞彙，根據Bookstein & Swanson以及Cooper & Maron對文件排序的函數定義㉕：

$$g(x) = \log \frac{p_r(x|rel)}{p_r(x|nonrel)} + \log \frac{p_r(rel)}{p_r(nonrel)} \quad \cdots\cdots (K)$$

$p_r(x|rel)$ 及 $p_r(x|nonrel)$ 分別表示詞彙 x_i 在相關與不相關文件集合的事前出現機率（priori probabilities）。假設詞彙出現在相關或不相關文件是一種獨立的情況，我們可以下式來估計：

$$p_r(x|rel) = \prod_{i=1}^{t} p_r(x_i|rel)$$

$$p_r(x|nonrel) = \prod_{i=1}^{t} p_r(x|nonrel) \quad \cdots\cdots (L)$$

如果文件的索引向量是以加權方式索引，$D = (d_1, d_2, \ldots, d_t)$，$d_i$ 表示詞彙 x_i 在文件內的權重（若 $d_i = 0$，亦即詞彙 x_i 不在文件內）。則機率相關的加權（term-relevance weight；簡寫爲 tr_j）計算如下：

$$tr_j = \log \frac{p_j(1-q_j)}{q_j(1-p_j)}$$

$$= \log \frac{p_r(x_j = d_j|rel)\, p_r(x_j = 0|nonrel)}{p_r(x_j = d_j|nonrel)\, p_r(x_j = 0|rel)} \quad \cdots\cdots (M)$$

所以，參考前述有關IDF、TDV、S/N的加權方式，索引詞 T_j 在文件 D_i 的顯著值計算如下：

$$w_{ij} = tf_{ij} \cdot tr_j \quad \cdots\cdots (N)$$

雖然機率模型區分索引詞彙在相關文件與不相關文件中的特性與差異，但是，機率相關的加權方法也有其致命傷，其一是代表母體的文件集合在哪裡？沒有母體的資訊只有靠抽樣與估計；其二是如何在一群文件集合樣本中，能夠合理而精確的估計索引詞在相關文件與不相關文件中的特性。

因此，根據不同的假設可發展出不同的加權公式。Robertson & Jones從一個文件集合（N個文件）中區分相關與不相關文件數（R；N-R），然後分別估計相關且含有索引詞 T_j 的機率（p_j），無關且含有索引詞 T_j 的機率（q_j），以及文件中可能含有索引詞 T_j 的機率（s_j）㉖。導出以下四種加權公式：

$$w_{ij}^1 = \log p_j - \log s_j \quad \cdots\cdots (O)$$

$$w_{ij}^2 = \log p_j - \log q_j \quad \cdots\cdots (P)$$

$$w_{ij}^3 = \log \frac{p_j}{1-p_j} - \log \frac{s_j}{1-s_j} \quad \cdots\cdots (Q)$$

$$w_{ij}^4 = \log \frac{p_j}{1-p_j} - \log \frac{q_j}{1-q_j} \quad \cdots\cdots (R)$$

其中若以 r_j 表示相關且含有索引詞 T_j 的文件數，n_j 表示文件中含有索引詞 T_j 的文件數。則可以 r_j/R，$(n_j-r_j)/(N-R)$，n_j/N 分別估計 p_j、q_j 及 s_j 的機率。經過Robertson & Jones 實驗證實索引效果 $w_{ij}^4 > w_{ij}^3 > w_{ij}^2 > w_{ij}^1$。

即使如此，Wong & Yao提出一些索引環境的假設和推論，例如在大型文件集合中，$r_j \ll N$，則 $q_j \approx s_j = n_j/N$，以及在 $p_j \approx 1$ 和 $p_j \approx 0.5$的情形下，索引效果較差的加權方式 w_{ij}^1, w_{ij}^2 都可以導出如(B)式的IDF正確衡量；相反的，較能精確反應相關機率加權的 w_{ij}^3, w_{ij}^4 只能導出近似的IDF衡量結果⑫。雖然這些假設未經實證，但是理論上的討論已經有缺陷。這是機率模型的主要弱點。

3.索引詞群集索引的顯著值計算

從entropy衡量資訊的概念，陳淑美、楊允言分別在財經新聞自動分類研究中，引用集中度與廣度的索引詞選取條件⑧⑨。所謂集中度的意義，簡言之，一個具有分類價值的索引詞，應該集中出現在某幾類文件中，而不是散佈在各類中。所謂廣度的意義，一個具有分類價值的索引詞，應該盡可能分佈在某一分類的各篇文件內，而不是只集中在幾篇文件內。

本研究將根據上述的想法，從影響索引效果的兩個主要因素：索引的詳盡性（indexing exhaustivity）與索引詞的明確性（term specificity），給予集中度與廣度這兩個概念明確的定義。

(1)索引的詳盡性

索引的詳盡性是指索引能夠反應某一文件內容主題的詳盡程度，索引愈詳盡則使用愈多的索引詞彙，以描述文件內容主要及次要主題。在VSM文件--索引詞矩陣的表達中，每一個文件索引向量內的權數加總，如以binary的索引方法，只要把橫列（row）的元素加總，即可代表該文件的索引詳盡性。

前述的文獻回顧對於TF、IDF、TDV的計算已有討論。本文爲了在索引詞選取時有一個好的評量標準，並考量引進人工分類的知識，建立CIM在索引的詳盡性的評量標準。因此參考Entropy的計算公式，建立詞彙在群集內的（區域性）分佈資訊量，稱之爲索引詞群集索引的廣度（uniformity；H_i）。

$H_{jk} = -\sum_{i=1}^{n} p_{ij} \log p_{ij}$ 代表文件編號，k 代表某一類別……(S)

$p_{ij} = \frac{tf_{ij}}{\sum tf_{ij}}$；$tf_{ij}$ 表示第 i 個文件中，索引詞出現的頻率……(T)

由上述定義可知，每一個索引詞彙可能在不同的類別內會有一個資訊量，根據該詞分佈的廣度可以決定是否在該類別內使用該索引詞。

(2)索引詞的明確性

索引詞的明確性則是反應索引詞的廣義或狹義程度，當我們使用較廣義的索引詞，就比較不容易辨識相關與不相關的文件差異。在VSM文件--索引詞矩陣的表達中，以每一個索引詞被指定索引文件的次數來代表，如果以BIN方式索引，在VSM矩陣中直行(column)元素的加總就是代表索引詞的明確性，亦即索引詞的文件次數。

本文參考Entropy的計算公式，建立詞彙在群集上的分佈資訊量（全域），稱爲索引詞群集索引的集中度（conformity；ICF）⑲。

$$ICF_j = -\sum_{i=1}^{n} p_{ij} \log p_{ij} \quad , \ i \text{代表類別} \cdots\cdots (U)$$

$$p_{ij} = \frac{d_{ij}}{\sum d_{ij}} \quad , d_{ij} \text{ 表示第 } i \text{ 類群集中，出現的文件數} \cdots\cdots (V)$$

二、向量空間模型

前述索引理論不管是統計的或機率的模型，都是根據索引向量的形式來表示，相關的資訊檢索模型也是以此爲基礎。目前主要的資訊檢索模型包括有：向量空間模型、機率模型、擴充布林模型等⑳。所謂擴充布林模型只是把傳統的布林模型加上查詢句的索引詞彙加權，檢索文件輸出的相關排序，並引用索引向量的處理以及模糊集概念來加強檢索的效能。所以VSM可以說是最簡單也是最有生產力的模型㉑。

(一)向量空間模型基本概念

VSM是應用矩陣代數的數學模型，五個主要組成包括：索引方式（indexing methodology；M）、文件集合（document collection；C）、索引詞集合（index term；T）、索引詞加權架構（weighting schema；W）以及索引詞--文件矩陣㉒㉓。以下將介紹此一模型相關的基本概念。

1.索引詞向量

索引詞向量是構成文件索引向量的基礎。文件的內容是以索引詞向量所構成的空間來表達，文件索引向量就是以索引詞向量空間的線性組合方式，做爲索引形式上的代表。向量內的屬性可以加權，加權代表該索引詞（屬性）在文件內容上的顯著性（重要性）。文件索引向量的形式如下所示：

$$D = (t_1, t_2, \ldots\ldots, t_n) \cdots\cdots\cdots\cdots (W)$$

$$D_i = (w_{i1}, w_{i2}, \ldots\ldots, w_{in}) \cdots\cdots\cdots\cdots (X)$$

(W)式表示所有文件以索引詞向量 t_i 爲代表的形式。(X)式則表示某一文件 D_i 的索引向量，w_{ik} 代表索引詞 t_k 在 D_i 中的加權結果（亦即顯著性）。

2.相似衡量

Luhn最早建議設計自動化檢索系統，以比較文件內容與使用者查詢句之間的識別因子爲基礎⑭。常見的正規化相似衡量公式如表一所列，其計算的複雜度與所需的資源成本由小到大依序爲：Dice Coefficient、Cosine Coefficient、Jaccard Coefficient。本文採用Cosine Coefficient定義。

3.索引詞--文件矩陣

VSM以文件的索引向量爲基礎，因此索引詞--文件矩陣是VSM的核心，因爲他提供了文件在機器內的一種表達方式。

索引詞--文件矩陣如圖四所示，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：$\mathbf{D}_{m \times n}$。

$$\begin{bmatrix} w_{11} w_{12} \cdots\cdots w_{1n} \\ w_{21} w_{22} \cdots\cdots w_{2n} \\ \\ \\ w_{m1} w_{m2} \cdots\cdots w_{mn} \end{bmatrix}_{m \times n}$$

圖四、索引詞--文件矩陣 D

表一：常見正規化相似衡量公式

Similarity Measure sim(X,Y)	Evaluation for Binary Term Vectors	Evaluation for Weighted Term Vectors
Inner product	$\lvert X \cap Y \rvert$	$\sum_{i=1}^{t} x_i \bullet y_i$
Dice coefficient	$2\frac{\lvert X \cap Y \rvert}{\lvert X \rvert + \lvert Y \rvert}$	$\frac{2\sum_{i=1}^{t} x_i y_i}{\sum_{i=1}^{t} x_i^2 + \sum_{i=1}^{t} y_i^2}$
Cosine coefficient	$\frac{\lvert X \cap Y \rvert}{\lvert X \rvert^{1/2} \bullet \lvert Y \rvert^{1/2}}$	$\frac{\sum_{i=1}^{t} x_i y_i}{\sqrt{\sum_{i=1}^{t} x_i^2 \bullet \sum_{i=1}^{t} y_i^2}}$
Jaccard coefficient	$\frac{\lvert X \cap Y \rvert}{\lvert X \rvert + \lvert Y \rvert - \lvert X \cap Y \rvert}$	$\frac{\sum_{i=1}^{t} x_i y_i}{\sum_{i=1}^{t} x_i^2 + \sum_{i=1}^{t} y_i^2 - \sum_{i=1}^{t} x_i y_i}$

資料來源：G. Salton, Automatic Text Processing (Addison-Wesley Publishing Company, 1989), p.318.

向量空間模型的研究發展

Salton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⑤。從早期的自動索引、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⑥。

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⑦，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。例如群集檢索、索引與自動分類⑧⑨⑩，以類神經網路處理使用者的相關回饋⑪、自然語言處理⑫⑬⑭、自動建立系統索引典⑮，以及強調以概念爲基礎的資訊檢索系統設計⑯⑰。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。

基於上述相關研究在VSM上的累積，可以引用的分析技術很多，其中有早期的因素分析⑱⑲⑳，以及當代群集檢索研究應用的群集分析技術㉑。又如Can發展的包含係數概念的群集方法（C^3M）㉒。此外，新近的研究也有一些創新的模型，但是仍然合併VSM的基礎，例如Yang利用線性最小平方配適（LLSF）與奇異值分解技

衡（SVD）的結合[9]；Lang合併潛在語意模型（LSI）在類神經網路模型上的應用等[5]。上述各種方法在其過去的相關研究中可以證實VSM是最具有彈性的研究模型。

而且1992年以後的資訊檢索研究開始朝向大型語料發展，一方面是先進的資訊檢索技術一直都只在實驗室內研究，一旦面對實際應用需求的複雜環境仍然束手無策[9]。當前的研究仍然將VSM視爲解決大型語料檢索研究的中心支柱。雖然當前主宰實務應用優勢的系統仍以全文檢視的字串比對或字元反轉索引爲主，只要VSM突破大型語料的困境，同時結合智慧型自然語言查詢介面與關鍵詞自動擷取技術的發展，這樣一個先進的整合模型指日可待。

㈢群集索引模型的建構

1.群集索引的意義

群集索引所探討的重點是在索引詞所構成的原始資料空間，如何找到可以縮減索引空間的群集索引構面，使得原本存在許多線性相依關係，或者是繁雜無序的原始資料空間，藉由群集索引構面縮減空間維度，以萃取索引詞之間潛在的共同因素結構。其主要目的是以較少的構面來表示原先的資料結構，而又能保存原有資料結構所提供的大部份資訊。

本文將應用線性代數的奇異值分解方法（Singular Value Decomposition；簡稱SVD）作爲群集索引理論的數值分析技術[9]。在索引詞所構成的原始向量空間中，SVD可以將原始的索引詞--文件矩陣分解，建立一個較小維度且相互獨立（直交）的群集索引構面，經由SVD運算將索引詞維度縮減，在奇異值向量空間（singular vectors）找出索引向量的共同因素，這些共同因素代表索引詞之間潛在的群集索引構面。

2.CIM的操作定義

CIM只是VSM的延伸，所有VSM的操作定義皆適用於CIM。CIM不同於VSM的最大特色是具有群集索引構面（共同因素；或稱爲潛在結構），同時也改善了VSM無法滿足直交（orthogonal）的理論缺失。而群集索引構面的產生來自於原始資料的索引詞_文件矩陣經由SVD運算後的結果。

根據SVD的公式，索引詞_文件 $D = UWV^T$，D、U、W、V^T的維度分別$t \times d$、$t \times r$、$r \times r$、$r \times d$，其中t爲所有索引詞的數目，d爲所有文件數目，r爲D的秩（rank）。

經由SVD分解後，可以將原始較大維度的文件索引空間縮減爲一個具有共同因素爲k的群集索引空間，使得原來的$D = UWV^T$改變爲 $D_k = U_k W_k V_k^T$。如此一來，不但可以簡化自動索引的空間與檢索效率，同時建構了文件的群集關係與索引詞的群集索引結構。

在CIM模型中 U_k、W_k、V_k^T 的維度各縮小爲$t \times k$、$k \times k$、$k \times d$且 D_k 的秩亦縮小爲k，而共同因素k是一個比原始索引詞向量空間小的數值，W_k 則保留了索引詞_文件矩陣中較大的k個奇異值，U_k 稱爲詞彙向量（term vector），V_k 稱爲文件向量（document vector）。因此，原來繁雜的資料空間轉換爲具有群集索引的共同因素空間。

根據上述文件索引的過程，本文將設計以下實驗分別從檢出率、精確率、索引構面及相關係數界限值（threshold）等四個角度去綜合評量CIM索引的效果。

肆、研究方法與實驗設計

本研究之性質屬於科學理論發展的研究，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，研究索引詞彙的形式與文件內容關係，建構中文全文文件的群集索引理論。

在研究的概念層次上，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（statistical relationship），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CIM爲個設理論的測試模型，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。

實驗語料從兒童日報全文資料庫中選取醫藥新聞502篇文件（表二），參考並修訂中央通訊社新聞分類索引典，進行實驗語料人工分類。根據分類結果，設計五個查詢句，進行各項實驗。並由台大圖書館學系高年級學生以人工選詞方式，選取原始詞集2564詞。最後以人工進行索引詞的節縮語同義詞權威控制，共有索引詞2369詞。各項選詞結果並經專家審查確認。

本文設計以下幾個實驗進行各項研究：

1.CIM最適群集索引構面的實驗：以人工選詞、權威控制以及人工索引後進行此項實驗評量。利用SVD技術建立IDF加權後的索引形式，探討群集索引構面縮減的各項性質。

2.CIM與傳統VSM的比較：根據VSM原始定義，分別以BIN、TF、IDF三種索引形式，比較傳統VSM利用原始索引詞--文件矩陣進行評量，並與CIM不同索引構面的評量進行比較，以瞭解CIM與傳統VSM的差異。

3.索引詞群集索引形式評量：根據前述對索引的詳盡性與索引詞的明確性兩個索引因素的定義，進行索引詞群集索引形式（集中度與廣度）的評量。根據評量篩選後的索引詞，以IDF加權方式比較詞彙篩選前後的系統效能，以瞭解良好的索引詞群集索引形式的性質。

前二項實驗以測試CIM的基本性質爲主，以最適索引構面區間、最佳索引加權方式以及CIM與傳統VSM的比較來觀察。最後利用CIM並配合群集索引形式的衡量結果，檢定原始詞集、權威控制詞集與篩選詞集的索引效能比較。

伍、實驗結論

爲了對下述實驗結果有更清楚、客觀的認知與討論，以下將針對實驗的評量方式與過程、以及實驗查詢句設計與查詢句基本的性質先作描述

表二：兒童日報新聞實驗語料基本性質

新聞類別	文件數	總字數	每篇平均字數	人工選詞詞數	權威控制詞數	每篇平均詞數（人工）
醫藥	502	179450	357	2564	2369	5

資料來源：本研究整理

，然後再說明實驗結果。

一、評量方式與過程

參考過去西文的研究報告可知檢出率及精確率是主要的評量指標。但是傳統上以這兩個指標各爲縱軸與橫軸，展現系統或模型在兩個指標上的消長關係。通常用十等分法將檢出率水準訂在10%、20%、30%……90%、100%找出對應的精確率水準，然後成對的繪製於圖上，以平滑的折線圖方式呈現。

這樣的方式在實驗室環境下有些問題實待克服，例如檢出率要以十等分法繪製，其實只能用一個估計值，而不是一個實驗的實際值，因爲實際的衡量不太可能正好有十等分的實驗值出現。尤其在查詢句與相關文件數較少的情形下，可能造成的誤差更大。

一般的評量方式過程如表三所展示。假設某一查詢結果，文件集合內共有五篇文件相關。例子中將如何來決定檢出率爲30%、50%、70%及90%的精確率呢？同時在相關文件數只有五篇的情形下，檢出率爲80%，而精確率水準從29%～67%存在很大的變異[8]。

因此本文利用新的評量過程與呈現方式，除了考量本研究所提的CIM在不同的群集索引構面下有不同特質，同時也兼顧了呈現檢出率與精確率的成對關係，把不同的群集索引構面下檢出率與精確率的特性能夠呈現。

同時，本研究假設在檢索系統實際運作情形下，系統會根據相關係數（簡寫爲sim.）的界限值，提供檢索結果。例如參考表三的評量過程，界限值0.5以上的共有十篇文件，其中四篇文件相關，所以檢出率是80%(4/5)，而精確率則是40%（4/10）。另一種情形在界限值0.2以上的共有十五篇文件，其中五篇相關，所以檢出率是100%（5/5），而精確率則下降爲33%（5/15）。這樣的評量方式可以模擬最接近實際系統運作情境，而且免除上述因估計可能產生的誤差問題。

界限值的決定仍待未來進一步研究，本研究從諸多實驗值中判定，在群集索引構面較大的時候（例如K=100時），所有文件與查詢句的相關係數較低，有時候最大的只有sim=0.5，所以界限值設爲0.5。附錄中所有圖表以這個界限值基礎下，呈現各種實驗的評量結果。

二、查詢句設計

本研究查詢句的設計是從使用者查詢的觀點來考慮，使用者可能以多個檢索詞來代表查詢的主題。同時參考國外研究所用的查詢句平均約七個以上的詞來設計[9]。因此本研究的查詢句也是多詞的組合方式，以能夠充分表達查詢主題的三～五個檢索詞爲主，而且爲了簡化此次研究的複雜性，檢索詞一律以TF=1未經任何加權情形下進行實驗。

同時受限於實驗語料不足，實驗的查詢句是根據人工對實驗文件分類後，選擇特定類別進行主題分析，每句以三個以上的詞彙來表達所要檢索的主題。附錄以表列實驗語料的查詢句設計，包括每個查詢句的相關文件數、查詢平均值以及每個檢索詞在實驗語料中各種索引形式評量的屬性（參考附錄表五）。

三、實驗結論

根據前一節實驗設計的前提下，這一節針對主要的實驗結果，作一個綜合的陳述。

(一)CIM最適群集索引構面區間的決策

此項實驗以人工選詞、權威控制以及人工索引後進行實驗評量，並利用SVD技術建立IDF加權後的索引形式，探討群集索引構面縮減的各項性質。在相關係數界限值的控制條件下，實驗結論如下（參考附錄表六、七；圖八、九）：

表三：查詢句相關評量表

相關係數排序	相關係數	文件編號	相關與否	檢出率	精確率
1	0.9876	289	相關	20%(1/5)	100%(1/1)
2	0.9654	45	相關	40%(2/5)	100%(2/2)
3	0.9432	201		40%(2/5)	67%(2/3)
4	0.9000	7	相關	60%(3/5)	75%(3/4)
5	0.8765	498		60%(3/5)	60%(3/5)
6	0.8210	261	相關	80%(4/5)	67%(4/6)
7	0.7543	270		80%(4/5)	57%(4/7)
8	0.6432	18		80%(4/5)	50%(4/8)
9	0.6098	192		80%(4/5)	44%(4/9)
10	**0.5543**	**332**		**80%(4/5)**	**40%(4/10)**
11	0.4543	46		80%(4/5)	36%(4/11)
12	0.4321	77		80%(4/5)	33%(4/12)
13	0.3201	89		80%(4/5)	31%(4/13)
14	0.2100	62		80%(4/5)	29%(4/14)
15	**0.2000**	**456**	相關	**100%(5/5)**	**33%(5/15)**

1.檢出率隨著群集索引構面增加而降低。

2.精確率隨著群集索引構面增加而上升。

3.檢出率與精確率存在消長關係，檢出率上升，精確率則下降。

4.界限值越低則檢出率相對提高，但是精確率都下降。

5.根據因素分析方法因素個數選取的經驗法則，CIM最適群集索引構面區間的決策準則：「文件數與群集索引構面之間的比例至少在五～十倍」。所以醫藥語料應以50≦K≦100爲準。

6.最適群集索引構面區間的決策在實務應用時，仍須考量文件集合的性質、大小，索引語言的形式，查詢語言的使用與其特性，以及系統設計的目標選擇可以作最適的決策。

㈡CIM與傳統VSM的比較

根據VSM原始定義，分別以BIN、TF、IDF三種索引形式，比較傳統VSM利用原始索引詞－文件矩陣進行評量，並與CIM不同索引構面的評量進行比較，以瞭解CIM與傳統VSM的差異。在相關係數界限值的控制條件下，醫藥語料的實驗結論如下（參考附錄圖五～圖七）：

1.由於傳統VSM的索引詞--文件矩陣是一個結構比較不嚴密的矩陣，所以文件與查詢句的評量結果相關係數都比較低，實驗提報的結果只有sim=0.2的情形。

2.傳統VSM在TF及BIN方式下，檢出率與精確率均明顯不如CIM。

3.在sim=0.2，IDF的索引下，VSM的檢出率遠低於CIM，但精確率則高於CIM索引構面K<80的情形。

4.如果sim=0.5，IDF索引方式下，由於VSM無法有效評量文件與查詢的相關性，CIM最適索引構面區間內，效能顯然優於VSM。

(二)索引詞群集索引形式評量

根據前述索引的詳盡性與索引詞的明確性兩個索引因素的定義，進行索引詞群集索引形式（集中度與廣度）的評量。根據評量篩選後的索引詞，以IDF加權方式比較詞彙篩選前後的系統效能，以瞭解良好的索引詞群集索引形式的性質。

實驗根據篩選後的詞集、權威控制詞集與原始選詞集等三個詞集，分別進行評量結果的交叉分析，以瞭解這三個不同詞集在索引上的差異。（參考附錄表六、七；圖八、九）

1.原始詞集2564詞，依群集索引形式篩選結果縮減爲1149詞。
2.在高界限值（sim=0.5）下，篩選1149詞集的平均精確率60.15%分別略低於原始2564詞集的63.49%，與權威控制2369詞集的64.78%。

總結三項的實驗結果與上述不同詞集的交叉分析可知，CIM在索引的效果上優於傳統VSM，而且可以改善或者提昇其效能，達到具有權威控制機制下的索引效果。同時根據人工分類知識的引用，在索引詞群集索引形式的評量上，可以提供選詞的標準。在初期的實驗下，CIM證實能以較小而具有分類價值的索引詞詞集來索引文件，但是對索引的效能沒有影響。

陸、未來研究議題

由於過去中文自動索引研究的缺乏，本文雖然在CIM建構上引進SVD技術，並且在索引詞群集索引形式的評量上有新的突破。但是就中文全文的自動索引理論而言，本文仍在建立理論典範（Paradigm）的初期探勘階段。經由上述實驗結論與問題討論可知，未來在中文自動索引理論的研究上，尚有許多重要的議題需要解決。以下分別列舉具體研究議題，提供未來研究建議。

一、中文自動索引應用研究基礎環境的規劃

觀察西文自動索引或資訊檢索相關研究的發展歷程，除了在研究方法與研究結果提供我們研究的啓發以外，其中還有一點相當重要，就是建立理論與應用研究的基礎環境。我們可以從SMART研究計畫個案中觀察到這個重要的啓示。

SMART的發展可以參考表四所整理的記事年表[49]。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。

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。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）以及新近以大型資料爲主的TREC（Text Retrieval Conference）實驗，都可以見到SMART的研究發表。

因此，現階段中文資訊檢索研究的困難，在

於缺乏一個具有科學實驗效度的研究環境，而且先進的實驗室模型無法與實務應用連結。如果中文自動索引或資訊檢索研究也能規劃、建立一個應用研究的基礎環境，相信對於未來的研究發展有一個研究累積、突破創新與科技整合的契機。

從本研究的經驗可知，運用已有的中文資訊處理基礎與應用研究成果，現階段可以先從相關評量模組開始，建立標準的測試語料庫，設計足夠代表評量系統效能的查詢句，模擬實際使用者環境下系統效能評量的新標準，建立中文資訊檢索研究在國際競爭的研究環境下的新展望與貢獻。

二、以中文字或詞彙為索引形式的理論研究

就中文索引問題而言，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。因此，由於語文本身的差異，索引的形式自然不同。

本文從索引的觀點發現，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個。

所以在研究設計上，先考慮以詞彙及詞組的形式作為先導的試驗，除了在語意上可以有更大的包容，也可以進一步參考人工系統處理資訊的通類知識，例如索引典的機制。可以使得計算

表四：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 FORTRANIV、Assembly、PL/I 為主
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 資料分析與統計教體開發
1985	開放性系統架構，逐步改寫至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/>（26 Nov, 1996）。

機系統因爲人工智慧的介入，突破其制式系統的限制。本文在CIM索引詞群集索引形式衡量上的實驗發現，即是突破傳統VSM無法有效詮釋形式與內容關係的瓶頸。

雖然說以詞彙爲主的理論研究途徑一直主導過去的研究，但是就中文本身的特色而言，中文字在自動索引研究上所具有的特質仍未被闡明。相對於中文詞彙索引形式在構詞上的開放性，中文字具有封閉系統的特性。在過去中文電腦基本用字的統計研究上，林樹（1971）的研究是一個代表。該研究共收集了8523字，其中「最常用字」有1857字，在202萬2604字次的樣本下，出現頻度高達97.34%⑭。

因此，以中文字作爲索引形式的理論研究，在簡立峰所建立的Csmart智慧型文件檢索系統，充分掌握中文字的特性，以字元索引爲基礎，發展文件識別特徵檔的檢索技術，不但文件索引空間縮小，搜尋速度極快，已經有初步的研究成果⑮。但是，Nie、Brisebois & Ren也提出中文詞彙在正確斷詞的情形下，能夠避免並解決以中文字爲索引形式可能引發的缺失⑯。

所以就中文資訊檢索研究而言，從這兩個不同的途徑著手，都有很大的挑戰與突破的空間，這也正是中文在理論研究上的優勢。

（收稿日期：1997年12月29日）

註　釋

註①：Iivonen, M., "Consistency in Selection of Search Concepts and Search Terms", Information Processing & Management, 31:2 (1995), pp.173-190.

註②：簡立峰，「尋易系統（Csmart）與中文智慧型資訊檢索」，在21世紀資訊科學與技術的展望國際學術研討會論文集，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。

註③：謝清俊、林頌，「中央研究院古籍全文資料庫的發展概要」，中央研究院資訊科學研究所文獻處理實驗室技術報告，1997年3月。

註④：謝清俊，「從二十五史全文資料庫的經驗談中文文件檢索系統設計的考量」，第三屆中文信息處理國際會議，專題技術報告抽印本（北京，1992年10月16－28日）。

註⑤：G. Salton, Automatic Text Processing (Addison-Wesley Publishing Company, 1989), p.328.

註⑥：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」，在當前語文問題學術研討會論文集，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。

註⑦：饒孟侃，認知心理學，初版四刷（台北市：心理出版社，民國82年），頁1-2。

註⑧：朱邦復，「概念網路」，李國鼎先生科技政策與管理講座，1993年12月。

註⑨：我們的日常應用的話語大致可以分成四級，句子（sentence）、詞組（phrase）、詞彙（word）、詞素（morpheme），詞素應是最小的、有意義的單位，卻不是可以獨立運用的單位。這是詞彙與詞素最大的不同。詞素也有人稱爲語位或語素。方師鐸，國語詞彙學構詞篇，（益智書局，民國59年），頁19、20。

註⑩：黃慕萱，「索引典的基礎理論」，索引典理論與實務研討會論文集（台北市：中國圖書館學會，民83年），頁20-34。

註⑪：同註⑤, p.307.

註⑫：G. Salton, "A Theory of Indexing", Regional Conference Series in Application Mathematics (Society for Industrial and Applied Mathematics, 1975)

註⑬：同註⑫。

註⑭：最早提出索引詞頻率衡量的是H. P. Luhn, 1957，引自G. Salton, "A Theory of Indexing," Rogional Conference Series in Application Mathematics (Society for Industrial and Applied Mathematics, 1975), p.55.

註⑮：同註⑤, p.280.

註⑯：同註⑫, pp.28-29.

註⑰：同註⑫, p.8.

註⑱：同註⑫, p.43.

註⑲：同註⑫, pp.43-44.

註⑳：同註⑫, p.51.

註㉑：同註⑫，詞組在此實驗中包括有單元詞彙（single terms）、雙詞彙（pairs）詞組和三詞彙（triples）詞組。

註㉒：C.E. Shannon., "A Hathematical Theory of Communication," Bell System Technical Journal, 27 (July Oct, 1948), p.379-423, p.623-656.

註㉓：G Solton & McGill, M. J., Introduction to Modern Information Retrieval (McGraw Hill Book Co., 1983), p.65.

註㉔：S.K.M. Wong & Yao, Y.Y., "An Introdaction-Theoretie Measure of Term Specificity," JASIS, (January, 1992), p.59.

註㉕：同註⑫, p.285.

註㉖：同註⑲, p.55.

註㉗：同註㉑, pp.55-56.

註㉘：陳淑美，「財經新聞自動分類研究」（碩士論文，台灣大學圖書館研究所，民國81年）。

註㉙：楊允言，「文件自動分類及其相似性排序」（碩士論文，清華大學資訊科學研究所，民國82年6月）。

註㉚：ICF (Inverted Cluster Frequency) 的命名是參考前述文件頻率修正因子 (IDF) 概念。

註㉛：D. M. Everett, & Cater, S. C., "Topology of Document Retrival Systems," Journal of the American Society for Information Science, 43：10 (1992), p.659.

註㉜：同註⑫, p.313.

註㉝：F. Can, & Ozkarahan, E. A., "Computation of Term/Document Discrimination Values by Use of the Cover Coefficient Concept," JASIS, 38:3 (May, 1987), p.171.

註㉞：同註⑤, p.513.

註㉟：A. E. Fox, "Technical Report 83-560," <http：//cs-tr.cs.cornell.edu/> (26 Nov, 1996).

註㊱：G. Salton, The SMART Retrieval System, Experiments in Automatic Document Processing (Prentice Hall, Inc.,: Englewood Cliffs, N. J., 1971.)

註㊲：G. Salton, "The Smart Document Retrieval Project," ACM-SIGIR, (1993), pp.357-358.

註㊳：D. D. Lewis, "An Evaluation of Phrasal and Clustering Representations on a Text Categorization Task," ACM-SIGIR, (1992), pp.37-50.

註㊴：S. Deerwester, Dumais, S.T., Furnas, G. W., Landauer, T. K. & Karshman, R., "Indexing by Latent Semantic Analysis," JASIS, 41:6 (September, 1990), pp.391-407.

註㊵：F. Can, "On The Efficiency of Best-Match Cluster Searches," Information Processing & Management, 30:3 (1994), pp.343-361.

註㊶：R. Wilkinson, & Hingston, P., "Using the Cosine Measure in A Neural Network for Document Retrieval," ACM-SIGIR, (1991), pp.202-210.

註㊷：X. Lu, "Document Retrieval : a Structural Approach," Information Processing & Management, 26:2 (1990), pp.209-218.

註㊸：J. Kristensen, "Expanding End-User's Query Statements for Free Text Searching with a Search-Aid Thesaurus," Information Processing & Management, 29:6 (1993), pp.733-744.

註㊹：Y. Yang, & Wilbur, J., "Using Corpus Statistics to Remove Redundant Words in Text Categorization," JASIS, 47:5 (1996), pp.357-369.

註㊺：C. J. Crouch, "An Approach to the Automatic Construction of Global Thesauri," Information Processing & Management, 26:5 (1990), p.632.

註㊻：I. Syu, 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).

註㊼：Y. Yang, & Chute, C. G., "An Application of Least Squares Fit Mapping to Text Information Retrieval," ACM-SIGIR, (1993), pp.281-290.

註㊽：Y. Yang, & Chute, C. G., "An Example-Based Mapping Method for Text Categorization and Retrieval," ACM Transaction on Information Systems, 12:3 (July, 1994), pp.252-277.

註㊾：H. Borko, & Bernick, M., "Automatic Document Classification," Journal of Association of Computing Machinery, 11 (1963), pp.151-162.

註㊿：M. Kurfeerst, & Asher, J. W., "A Factor Analysis of the Education Laws of Pennsylvania," Information Storage & Retrieval, 4 (1968), pp.257-270.

註(51)：S. K. M. Wong, Ziarko, W. & Wong, P. C. N., "Generalized Vector Space Model In Information Retrieval," ACM-SIGIR, (1985), pp.18-25.

註(52)：R. Burgin, "The Retrieval Effectiveness of Five Clustering Algorithms as a Function of Indexing Exhaustivity," JASIS, 46:8 (September, 1995), pp.562-572.

註(53)：同註㊱。

註(54)：Y. Yang, "Noise Reduction in a Statistical Approach to Text Categorization," ACM-SIGIR, (1995), pp.256-263.

註㉔：Sheau-Dong, Lang, "Tutorial on Text Retrieval Techniques and Their WWW Applications," 資訊擷取技術及其在WWW之應用研討會（國立清華大學，1996年 8 月13日）。

註㉕：A. E. Fox, & Koll, M. B, "Practical Enhanced Boolean Retrieval: Experiences with the SMART and SIRE Systems," Information Processing & Management, 24:3 (1988), pp.257-267.

註㉖：線性代數上常用矩陣來作數值分析的工具（解聯立方程式組），只有在非奇異矩陣（nonsingular matrix）時，才會有唯一解。若原始資料矩陣是奇異矩陣（singular matrix）時，矩陣的行列式值爲 0，則無法應用常態的矩陣求解過程。SVD因此常用來解決線性最小平方估計（linear least-squares）的問題，以克服奇異矩陣的困境。W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. Numerical Recipes in C 2nd. Edition. (Cambridge University Press, 1992.) <http://cfatab.harvard.edu/nr/bookc.html> (27 Mar, 1997)

註㉗：同註㉓, p.6.

註㉘：同註㉓。

註㉙：A. E. Fox, Technical Report 83-560 ; Buckley, C., Technical Report 85-686, < http://cs-tr.cs.cornell.edu/> (26 Nov, 1996)

註㉚：趙元任，「斷言成分裡意義有關的程度問題」，見袁毓林主編，中國現代語文學的開拓與發展：趙元任語言學論文集（北京市：清華大學出版社，1992年10月）。

註㉛：J. Y. Nie, Brisebois, M., & Ren, X., "On Chinese Text Retrieval," ACM-SIGIR, (1996), pp.225-233.

註㉜：蘇培成，「電子古籍中的缺字問題」，第一屆中國文字學會學術討論會（天津，1996年8月25日），頁3。

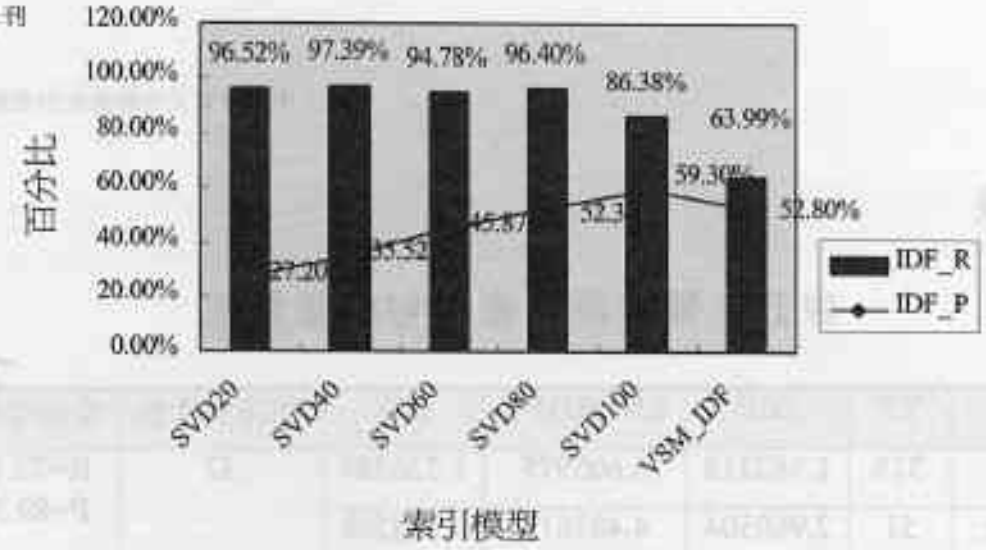
註㉝：同註㉒。

註㉞：同註㉙。

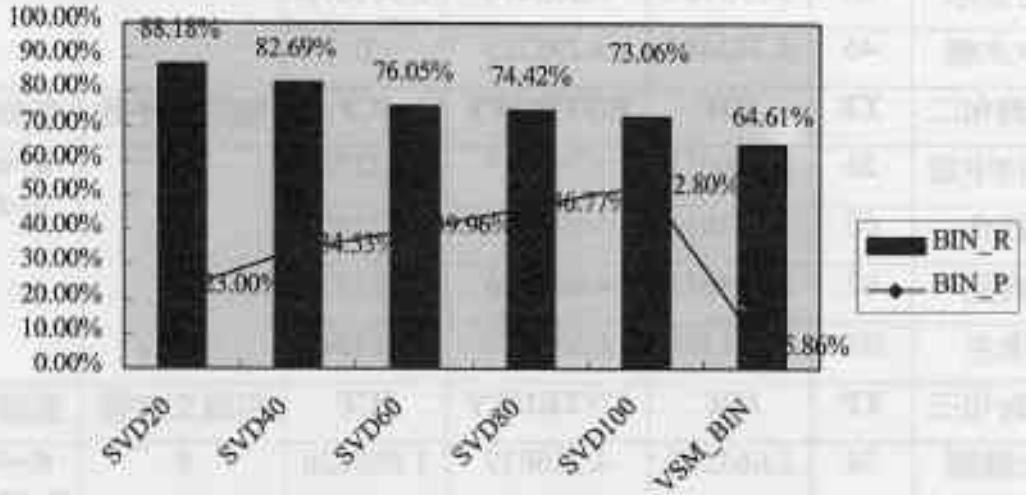
附 錄

表五：醫藥語料查詢句相關屬性

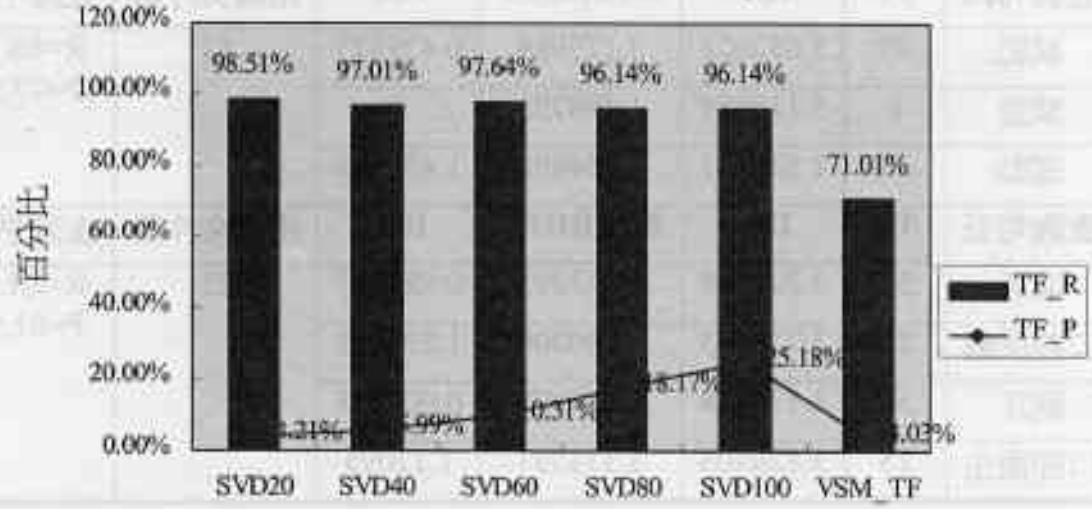
查詢句一	TF	IDF	ENTROPY	ICF	相關文件數	查詢平均值
學校	315	1.382318	6.605975	1.526388	32	R=73.10% P=89.37%
飲用水	51	2.960504	4.483618	0.324508		
衛生	209	1.431108	6.595822	1.524374		
自來水	88	2.999724	4.269671	0.550573		
飲水機	46	3.733693	3.295225	0		
查詢句二	TF	IDF	ENTROPY	ICF	相關文件數	查詢平均值
食物中毒	26	3.446011	3.796217	0.233792	23	R=40.58% P=43.54%
飲食	65	2.581014	5.028072	1.373809		
食品	62	2.784613	4.627764	0.928776		
衛生	209	1.431108	6.595822	1.524374		
查詢句三	TF	IDF	ENTROPY	ICF	相關文件數	查詢平均值
幼稚園	74	2.663252	4.766639	1.689026	6	R=94.44% P=31.49%%
托兒所	60	2.922763	4.486886	1.562718		
健康檢查	59	2.851304	4.509429	1.913386		
查詢句四	TF	IDF	ENTROPY	ICF	相關文件數	查詢平均值
減肥	49	3.653651	3.129984	0.429323	12	R=86.11% P=77.94%
減重	8	5.119988	1.298795	0		
運動	68	2.529721	4.884896	1.459906		
查詢句五	TF	IDF	ENTROPY	ICF	相關文件數	查詢平均值
潔牙	35	3.328228	3.883293	0.729843	23	R=80.19% P=81.98%
口腔保健	23	3.733693	3.360306	1.236685		
刷牙	52	3.174078	4.04047	0.901951		
口腔衛生	15	3.820705	3.373557	1.12095		



圖五 醫藥語料群集索引模型與VSM比較(IDF;sim=0.2)



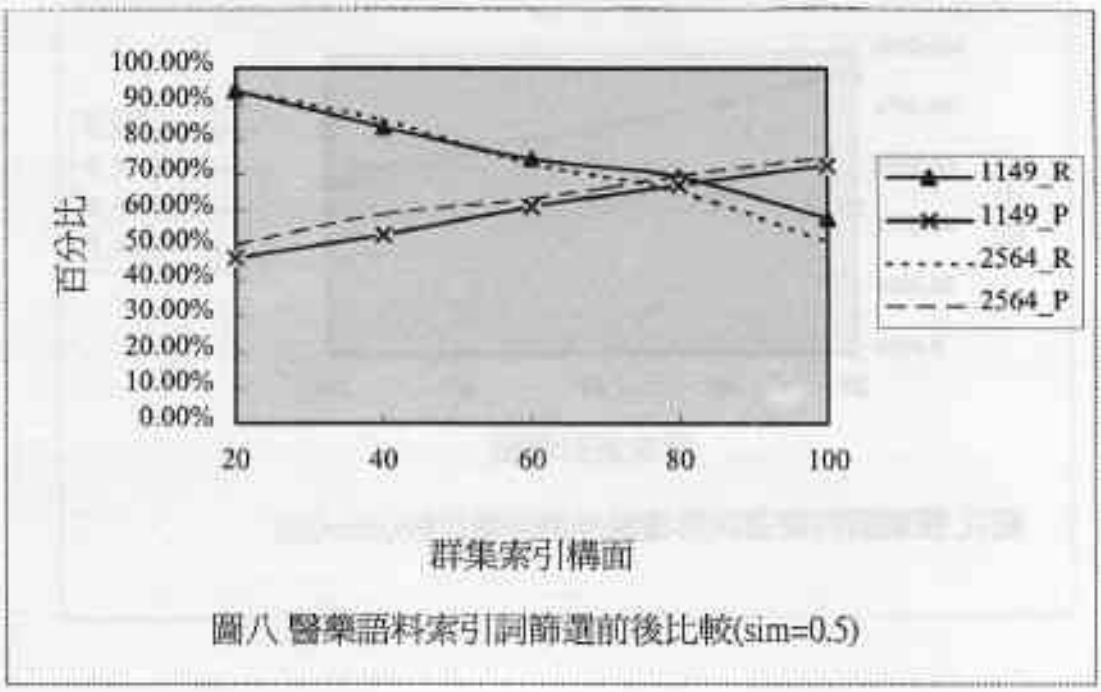
圖六 醫藥語料群集索引模型與VSM比較(BIN; sim=0.2)



圖七 醫藥語料群集索引與傳統VSM評量比較(TF; sim=0.2)

表六：醫藥語料索引詞篩選與原始選詞評量比較(sim=0.5)

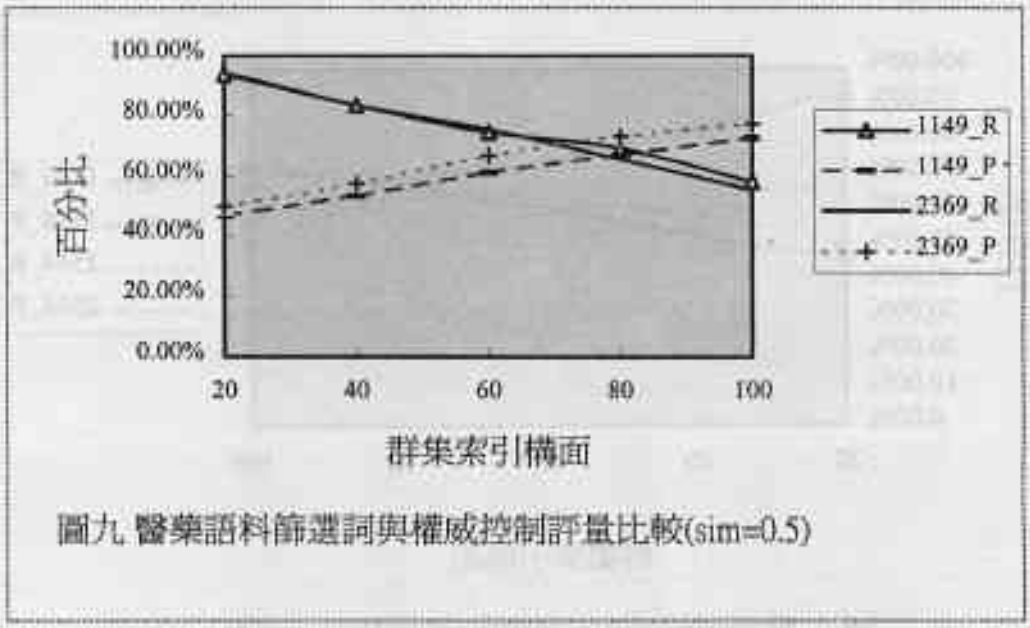
MED	1149 詞		2564 詞		精確率差距
群集索引構面	Recall	Precision	Recall	Precision	
20	93.68%	46.70%	93.61%	50.55%	3.85%
40	83.45%	53.22%	85.40%	59.00%	5.78%
60	74.56%	61.07%	73.20%	63.18%	2.11%
80	69.28%	67.07%	65.31%	69.64%	2.57%
100	57.78%	72.70%	51.24%	75.10%	2.40%
平均值	75.75%	60.15%	73.75%	63.49%	3.34%



圖八 醫藥語料索引詞篩選前後比較(sim=0.5)

表七：醫藥語料索引詞篩選與權威控制評量比較(sim=0.5)

MED	1149 詞		2369 詞		精確率差距
群集索引構面	Recall	Precision	Recall	Precision	
20	93.68%	46.70%	94.40%	49.58%	2.88%
40	83.45%	53.22%	83.41%	57.22%	4.00%
60	74.56%	61.07%	75.32%	66.80%	5.73%
80	69.28%	67.07%	65.59%	73.04%	5.97%
100	57.78%	72.70%	55.11%	77.27%	4.57%
平均值	75.75%	60.15%	74.77%	64.78%	4.63%



圖九 醫藥語料篩選詞與權威控制評量比較(sim=0.5)