中文全文資訊檢索之效能評量初探

Exploration of Effectiveness Evaluation for Chinese Full-Text Information Retrieval

黃雲龍
國立體育學院體育管理學系

張佑任
南華大學資訊管理學系碩士班

摘　要

全文資訊檢索已成爲一個跨學域的重要研究問題，然而中文全文資訊檢索的研究，因爲中文語言的本質與特徵，所以起步比西文資訊檢索的研究較晚，目前中文全文資訊檢索研究的困難點，在於缺乏一個具有科學實驗效度研究環境與效能評量的標準。本研究因應數位化資訊檢索的需求，以及確保中文資訊檢索系統之檢索品質，其分析不同模式群集索引的中文全文資訊檢索系統，適用在同一平臺、同一文件集之效能評量。

關鍵詞：資訊檢索，效能評量，向量空間模型，群集索引模型，奇異值分解

Abstract

Full-Text Information Retrieval is becoming an interdisciplinary interest. Mandarin Chinese Full-Text Information Retrieval is facing more basic difficulties than English Full-Text Information Retrieval because of research lag and language nature. Lack of an objective test collection and a standard effectiveness evaluation for information retrieval experiments is the fundamental issue for Mandarin Chinese Full-Text information retrieval. This research will have mainly object that analysis the fitness method of Full-Text Information Retrieval in same documents set.

Keywords: Information Retrieval, Effectiveness Evaluation, Vector Space Model, Cluster Indexing Model, Singular Value Decomposition

壹、導論

資訊檢索至今有四十多年之久的發展歷史，但中文資訊檢索研究，則是最近十年內才剛興起，因中文語言本身的特性的影響，所以與西文資訊檢索有許多的差異。

然而在中文資訊檢索領域，目前缺少做爲實驗研究的基礎環境，所以許多研究議題無法突破，當系統面對實際應用複雜且大規模的環境，可能仍有許多改進之處。

傳統的資料庫資訊檢索，乃依據查詢問題的屬性值(attribute value)與資料庫內每一筆記錄的屬性值進行比對(matching)，之後將完全相同的記錄檢索出來傳給使用者。

但是資料中還存在許多非結構化的資料處理問題，例如公文、書信、技術規範、標準手冊、筆記、備忘錄、工作記錄等，這些資料並非能符合傳統資料庫資訊檢索。

全文資訊檢索則是檢索者依據查詢主題之詞彙，來擷取相關之文件，在全文文件集中，運用與內容(content)相關的屬性特徵，做爲檢索過程中文件內容的識別因子(content identifiers)，即所謂的內容檢索(content retrieval)。

而在圖書館中常見的圖書檢索多以標題、作者等做爲查詢條件，這並不能稱的上是全文檢索。早期結構化資料在儲存、檢索上，都比非結構化來的簡單，但是隨著資訊技術發展，全文檢索已成爲資訊檢索的重點。

本研究因應數位化資訊檢索的需求，以及確保中文資訊檢索系統之檢索品質，嘗試建立一套中文資訊檢索績效評量模式，針對不同的群集索引模式，在同一性質的測試文件集(documents set)內，進行績效評量，期盼能夠尋求出更合適的檢索方法，與發展成熟群集索引模式之檢索應用系統。

貳、文獻探討

下面就研究所需，針對資訊檢索的基礎概念、檢索評量與索引理論等方向進行文獻探討，文獻的探討有助於了解資訊檢索的理論，進而能從中找到改善檢索績效的方法。

本研究所探討的資訊檢索績效評量，主要是建構於量化的基礎上，實驗中儘可能控制自變數，如：檢索問題、檢索者、檢索策略等，然後在評量不同的索引模式系統績效，以便提供未來開發群集檢索系統的參考依據。

一、資訊需求

在資訊檢索的過程中，主要的起點在於資訊需求(information need)，即是指個人的內在認知與瞭解的狀態，與外在環境接觸後所產生的不確定，試圖找尋可供判斷此不確

定事物的一種功能。「資訊」是指爲減少一些不確定事物的任何刺激（Stimulus），而「需求」則是指存在於個人的任何不確定事物。

Taylor(1968)將資訊需求過程中的問題的形成分成四層次：1.內藏式需求（Visceral need）：指需求者潛意識或非潛意識的需求，需求者可能無法察覺，並且混沌曖昧，難以用言辭表達的資訊需求。2.意識化需求（Conscious need）：需求者可以察覺出來的需求，但概念仍很模糊，對問題的界定與陳述不是很清晰。3.正式化需求（Formalized need）：在此階段，需求者能對其問題有一具體、詳細的陳述。4.妥協的需求（Compromised need）：對系統提出查詢問題，因受限於系統本身的限制，所協調的折衷需求。在資訊需求的過程(圖 1)存著愈向右則問題偏離真正的資訊需求愈遠的現象。

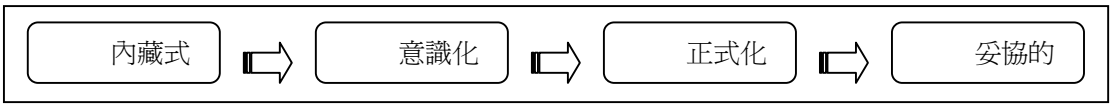

圖 1 Taylor 的資訊需求過程四階段

資料來源：廖書賢、黃雲龍，「從 TREC 的發展趨勢回顧中文全文資訊檢索關鍵議題」，第五屆三軍官校基礎學術研討會論文，民國 87 年 5 月。

Cooper(1971)認爲資訊需求是一種存在於無形的心靈、思想上，當人將其內隱之需求以文字或語言表達，此時的內隱資訊需求將被轉化成外顯的查詢(query)，換句話說，檢索者輸入的部份便是查詢 (檢索問題)。查詢即由使用者制定出之一系列規範，用以明確清楚地描述資訊需求。檢索者產生資訊需求後，須要能夠充分地向系統表達此資訊需求，倘若系統不能夠完整地瞭解到檢索者真正需求的資訊，則檢索的效果將大受影響。

因此，檢索者如果可以將資訊需求，以正確且適當的詞彙表達出來，而系統也可以比對相同概念的詞彙，則檢索者對檢索結果應該滿意，在 Iivonen(1995)的研究中，發現不同檢索者(intersearcher)對同一檢索問題概念轉換成檢索詞彙查詢時的一致性僅達 31.2%；而對同一檢索者(intrasearcher)會因檢索介面或環境的不同產生不一致狀況。事實上，檢索者可能因無法確切了解自己的需求，而找不到適當的詞彙來表示，或者檢索者所使用的檢索詞彙與系統使用的索引詞彙不同，如同義詞(synonym)，而導致不滿意的結果。

二、文件的組織與資訊檢索系統

文件是資訊的載體，在載體內可能涵蓋許多不同主題的資訊，但是無法保証文件中的資訊是否符合檢索者的資訊需求，文字是紀錄語言的符號或數字，完全相同的文字並不能確認其所代表資訊完全相同，或許也有無資訊存在的可能情況。

如何將文件中資訊以適當的形式表示，以利資訊檢索的應用，是所有資訊檢索系統都必須面對的問題。目前一個被廣泛應用的方法，就是運用所謂的索引(indexing)方法，將文件以適當的形式表示與組織，其所含的意義爲「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」。

所謂資訊檢索系統是一個具有儲存與檢索功能的系統，乃是預先將大量之文件按一定的方式組織儲存並數位化，然後使用者再依據其資訊需求問題之描述，檢索出相關文件資訊。圖 2 爲資訊檢索系統基本架構，描繪出資訊檢索的研究參考架構，其主要架構包含六個部分，分別爲文件分類與儲存(document subsystem)、使用者介面(user-system interface)、索引模式(indexing)、索引詞典(vocabulary)、檢索(searching)、比對判斷(matching)。

資訊檢索過程是由檢索者對資訊有所需求開始，便產生了檢索者的資訊搜尋行爲，而所謂的資訊搜尋行爲是一系列的資訊需求轉換及滿足檢索者需求之過程。另一方面大量文件集合（document set），經由適當的組織，選取出足以描述文件的特徵加以呈現；同時將檢索詞句進行查詢分析（query analysis），並將查詢映射（mapping）至文件空間（document space）中，以便進一步進行比對，最後檢索出相關資訊給檢索者，若檢索者對系統檢索的結果不滿意，則可透過互動式的人機介面（interactive men-machine interface）對系統進行相關回饋，以便修正檢索策略，得到最佳結果。

因此，資訊檢索系統設計的良窳，便成爲使用者是否使用的關鍵因素。換言之，資訊檢索系統應是一種中介體，它可以幫助使用者獲取所需要的資訊，並滿足需求。任何資訊檢索系統存在的目的都是希望擷取需求者真正所需要的資訊或概念上接近文件，同時過濾、篩選對需求者無用的資訊。

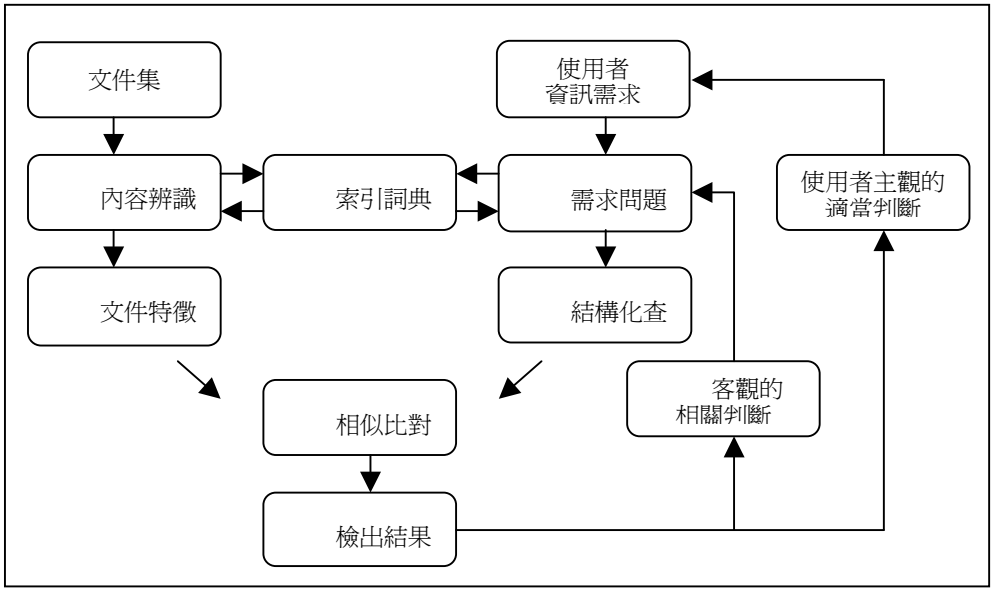

圖 2 資訊檢索系統基礎架構概念 資料來源：本研究

三、「相關」的概念與相關判斷

Mooers、Perry、Taube 和 Luhn 首先以系統的觀點(system's view)來定義相關，將相關定義爲檢索詞彙和文件索引特徵詞彙之適當性的關係，認爲文件與查詢句關係只有二元模式決策，而資訊檢索系統主要目標應在檢索使用者相關資訊，竟量避免檢索到不相關資訊。1958 年 Vickery 認爲相關概念可大略劃分爲主題相關(relevance to a subject)與使用者相關(user relevance)二個層面，其中主題相關屬於以系統爲出發點，即檢索問題之主題詞彙和文件特徵詞彙之間適當性關係，至於使用者相關則大多涉及主觀的因素，乃指使用者追尋此資訊程度，由使用者自行去衡量。

在系統的觀點盛行年代中，所討論的相關多爲主題相關，其中最常使用的主題相關定義是 Cuadra 和 Katter(1967)所描述的，「相關是介於資訊需求陳述與文件內容上的一致性及適合程度」，所以主題相關顧名思義，就是從主題的觀點來探討相關，其假設是主題相關的資訊應該能滿足檢索者的資訊需求，因此主題相關是一種客觀相關。

1971 年 Cooper 認爲 Cuadra 和 Katter 所提定義其中一致性(conclusion)與適合程度(premises)都太過於模糊，因此他提出邏輯相關概念，將相關定義爲「如果一個文件內容中有一句或一部份跟資訊需求陳述相關，即該文件則應被歸類爲相關文件」，其中邏輯相關的定義必須在三個限制前提內，1.檢索問題必須是 Yes-No 問題、2.資料的敘述方式必須是以正式語言的句子、3.資訊系統必須具有推理的功能，能依循脈絡直接回答問題。黃慕萱 (1997)認爲，邏輯相關也是一種主題相關，但其判斷主題相關的方法，並不是將代表資訊需求的詞彙和代表文章內容的詞彙互相配對比較，而是先尋找該文件中是否包含能滿足資訊需求之最小前提組(minimal premise set)。[廖書賢，民國 87 年]

Van Rijsbergen(1975)認爲 Cooper 所提出的邏輯相關，對資訊檢索系統是很重要的概念，且具有一定程度的客觀性。1973 年 Wilson 以 Cooper 的邏輯相關爲基礎，認爲單從一個推論的角度來看相關是不夠的，必須兼具歸納的角度，也就是說，文件中的資訊能加強某一概念或假設，則這篇文件應該被視爲相關文件。從此，對於非主題相關的研究也漸漸在擴展之中，如情境相關、心理相關、效用相關等，都是試圖從各種不同的角度探討相關概念。

一般資料庫測試，通常使用單一相關判斷，Lesk 和 Saltion (1968)以二元尺度分類法(100%相關與不相關)做相關判斷研究，指出作者與非作者之間相關判斷的一致性只有 30%，不同判斷者所產生的相關判斷結果，通常有某種程度的差異存在，許多學者認爲

在相關判斷程度上採取二元尺度分類法，最大的缺點是忽視文件之間的差異性，有些文件可能與查詢問題之間的關連性較大、或是提供了較直接的回答，這樣的文件應該給予較高的相關加權，若只將文件分成相關與不相關兩種，則對系統績效評量的準確性上將有所影響，然而 Burgin(1992)、Kazhdan(1979)及 Lesk 和 Salton(1968)均對不同的相關判斷對系統績效的影響作了研究，最後的結果均一致性的表示相關判斷的差異並不會影響系統效益優劣排序的穩定度。

在實際測試環境中，因測試文件集龐大，在實行文件相關判斷卻有困難點，早期 Cranfield 實驗採用逐一比對的方式，並由 5 位相關背景之研究生，進行了 50 萬次以上的十分詳盡相關判斷，但是 Swanson 和 Harter 均認爲 Cleverdon 在 Cranfield 實驗中，並沒有完全找出所有相關文件，推斷仍可能遺漏了 7000 多篇的相關文件。[Harter 1996]

對於較大型的實驗，要將每一個文件進行詳盡的相關判斷，所需耗費的工程更是浩大，因此 TREC 採用了 pooling 的方式，來找出相關文件，即針對每一個主題，從測試系統所回送的結果中，取出 100 篇文件，形成一個 pool，再去除重複才進行相關判斷，平均每個主題會遺漏約一篇相關的文件。

四、評量標準與結果呈現

績效評估是一組標準的指令用來比較不同系統的相對績效。一個資訊系統的效能評量，一般分爲量化評估或質化評估兩部分，一爲效益，另一爲效率；在資訊檢索系統上，效率意指檢索系統能檢索到相關資訊並過濾掉非相關資訊的程度，其中牽涉到相關性、檢出率與精確率；效益指的是系統上的經濟成本與系統檢索回應時間。

1953 年 Kent 提出以檢出率(recall ratio)和精確率(precision ratio)作爲檢索結果測量準則，直到現在，仍被廣泛接受使用中，檢索系統很難全部檢出所有相關文件，實際上只能達到某百分比，既然檢出率一般只能達到某百分比，相對就會有一部份相關文件並未能被檢出，未被檢出的相關文件百分比通常使用漏檢率代表，然而在所有檢索系統總是力求提高檢出率或者降低漏檢率。

在檢索中通常檢出一些無關的文件，要完全避免檢出無關文件是一件能難達成的事，因此精確率也只能達到某百分比，換句話說，會存在一定誤檢率(fallout ratio)，相對精準率和誤檢率也是同一問題兩面相對應的概念。

表 1 測量值之矩陣

	相關	不相關	總數
檢索到	a(hits)	b(noise)	a+b
未檢索到	c(misses)	d(rejected)	c+d
總數	a+c	b+d	a+b+c+d

資料來源：G. G. Chowdhury, Introduction to modem information retrieval, Library Association, 1999. p206.

$$recall = \frac{a}{a+c} \qquad \text{(公式 1)}$$

$$precision = \frac{a}{a+b} \qquad \text{(公式 2)}$$

由表 1 可以一目瞭然，a 表示相關文件被系統檢索到的數目，b 表示不相關文件被系統檢索到的數目，c 表示相關文件未被系統檢索到的數目，d 表示不相關文件未被系統檢索到數目，而 a+c 與 a+b 分別表示相關文件數目和系統檢出文件數目。所謂的檢出率(公式 1)是指「一查詢句，經由系統檢出相關文件之數目與文件集中的相關文件之總數目的比」；精確率(公式 2)是指「一查詢句，經由系統檢出相關文件之數目與系統檢出之文件數目的比」。

理論上，檢出率是測試系統檢索到所有相關文件的完整性；而精確率是測試系統檢索的結果之正確性。在檢索時，保證高檢出率與精準率，是檢索系統主要訴求，但經過許多無數的實驗測試，證明檢出率和精確率之間存在一種互相制衡的現象，提高檢出率的同時往往會降低精確率，反之亦然，理想檢索系統都期望同時有高的檢出率與精確率，然而這並不可能存在，因此只能在控制允許範圍內，求得高的檢出率與高的精準率。

檢索系統檢出文件後，再經由計算得出檢出率與精確率，然而所得數據該如何判斷與呈現，是一個很大的議題。績效的評量主要是探討不同系統的相對績效，或者同系統在不同控制下的相對績效，在評定不同系統的相對績效，必須用一個客觀性的測量值來評量標準，然而對於測量值的呈現則也必須公正表示。

在中文全文文件群集檢索上，文件與查詢之間可以用向量內積計算兩者的相似性，其檢出結果呈現是依據相似值大小排序。以相似值爲切截點，將不同索引構面數的檢出

率與精確率表示於同一個二維座標系上，運用這種方式可以找出系統適合的索引構面與界限值，相對於二元檢索系統[1]的評量，以相似值切截點是沒法適用的，所以相似性切截點只適用於在向量索引理論基礎的條件下，探討系統參數調整對檢索效能的影響。

目前相似性切截點(similarity cut-off) 的呈現方式有兩種，1.在相同相似值(界限值)下，探討不同構面對檢出率與精準率的關係。2.在同一索引構面構面下，探討不同相似值(界限值)對檢出率與精確率的關係。

五、群集索引模型

以向量空間爲基礎的資訊檢索實驗大約起於 1983 年的 SMART 實驗。群集索引模型主要是以向量空間爲概念，再結合奇異值分解(Singular Value Decomposition，以下簡稱 SVD) ，假設使用者對於想要檢索的內容是一群相關文件的集合，透過檢索問題與文件的相似(similarity)來衡量檢索資料，強調內容檢索的方式，提供使用者更精確的檢索結果。

首先必須先了解群集概念，所謂的群集就是依據文件之分布，讓文件自動叢聚爲數個類別，從圖 3 可以理解，其中每一文件都事先分類，相似文件會聚集在一起，形成一個小群，文件與文件之間距離(相似值)是使用向量來表示。

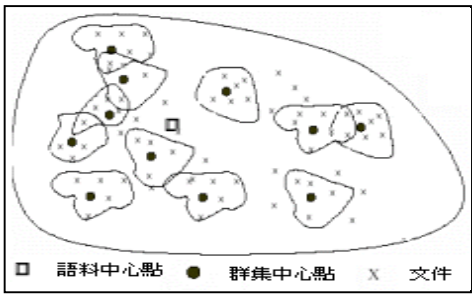

圖 3 群集分類

資料來源：G. G. Chowdhury, Introduction to modem information retrieval, Library Association, 1999. p206.

群集索引是以向量空間爲概念的索引方式，因此主要重心在文件或查詢句如何表示在向量空間中，其中索引詞--文件矩陣是由文件向量所構成，也是向量空間模型核心，所謂文件向量，即文件的內容是由索引詞向量所構成的空間來表示，而基本的元素是索引詞。

假設 t 爲索引詞，則文件內容(d)以 t 表示，在向量空間內，亦可對索引詞在文件內

[1] 本文所稱二元檢索系統即爲傳統的布林邏輯檢索系統，檢索出來的文件必須包含檢索詞彙。

容上的顯著性加權，因此文件內容可以寫成爲下面公式 3。其中一個具有 M 篇文件和 N 個索引詞的索引詞--文件矩陣表達方式爲 D_{mn} 如下公式 4。

$$D_i = (w_{i1}, w_{i2}, \ldots\ldots, w_{in}) \qquad (\text{公式 3})$$

$$D_{m\times n} = \begin{bmatrix} w_{11} w_{12} \ldots\ldots\ldots w_{1n} \\ w_{21} w_{22} \ldots\ldots\ldots w_{2n} \\ \bullet \\ \bullet \\ \bullet \\ w_{m1} w_{m2} \ldots\ldots\ldots w_{mn} \end{bmatrix}_{m\times n} \qquad (\text{公式 4})$$

（一）索引方式

根據 Salton(1989)的定義，索引(indexing)是利用文件識別因子(identifier)來建構文件特徵的過程，整個過程可以分成三個部份：1.分析文件內容、2.決定文件特徵、3.表現文件特徵。其重點在於保存文件內容的相關訊息，以作爲使用者的資訊需求與資訊內容之間的良好橋樑，其目的在於提供使用者查詢正確相關的文件，並加速非全文檢視資訊檢索的進行。[黃雲龍，民國 86 年]

1.詞彙頻率(Term Frequency；TF)

Salton(1975)提出文件的索引向量概念，以索引詞顯著值計算結果，利用二元加權(binary weight)方式，衡量該詞彙有無出現在文件內來選取索引詞，進一步以詞彙出現的頻率或次數作爲判斷的標。

詞彙頻率主要以區域性比重爲考量，而忽略全域性比重的影響，當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(entropy)就相對的少於出現在較少文件的詞彙，詞彙頻率的方法只顧及檢索的檢出率，忽略了精確率，所以其精確率將受到影響。

詞彙頻率公式如下，f_i^k 是索引詞 k 在第 i 個文件的出現頻率，F^k 是索引詞 k 在文件集合內的總次數。

$$F^k = \sum_{i=1}^{n} f_i^k \qquad (公式 5)$$

2.文件頻率倒數修正(Inverse Document Frequency；IDF)

Jones(Salton, 1989)提出以文件頻率的倒數爲修正索引詞彙顯著值的修正因子，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」(公式 6)。另外還有整合詞彙頻率與文件頻率的倒數加權方式，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，這類型比重值爲區域性與全域性比重之結合(公式 7)，Salton(1975)實驗中，TF×IDF 的索引結果。

$$IDF = \log(N / df_j) \qquad (公式 6)$$

$$w_{ij} = tf_{ij} \cdot \log \frac{N}{df_j} \qquad (公式 7)$$

Salton 認爲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。文件頻率與詞彙的區別值之間存在一些現象，1.當詞彙頻率剛剛好，則能直接地區別文件內容。2.當詞彙頻率較高，則必須使用詞組(phrase)方式降低詞彙頻率，以改善系統的精確率。3.當詞彙頻率較低，則必須使用索引典(thesaurus)的方式來提高詞彙頻率，改善系統的檢出率(圖 4)。

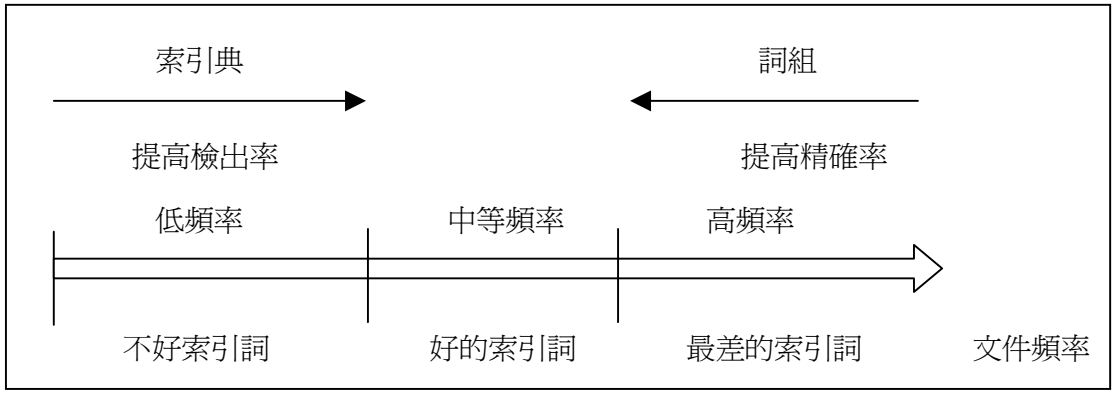

圖 4 索引詞特性與文件頻率關係

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

(二) 相似性計算

在向量空間中，系統檢出文件是以相似性值做爲依據，一般查詢句也是以向量表示，並被視爲一個虛擬文件(圖 5)，藉由餘弦測量(cosine measure)，對查詢句與文件索引向量之間的相似性比較，而常用的計算相似性衡量方式共有四個分別爲 inner product、dice coefficient、cosine coefficient、jaccard coefficient(表 2)，其值介於 0 與 1 之間。

使用相似評量方式的檢索系統其優點有三點：1.文件可以按照與查詢句相似大小排序，以顯示文件的重要性。2.使用者可以自行設定檢出數量，如果欲求高檢出率效能需降低相似界限值。3.相關回饋重新建立查詢條件，檢索系統可以重新調整查詢向量中各個不同維度上的權重，因而達成修正查詢條件(query modification)。

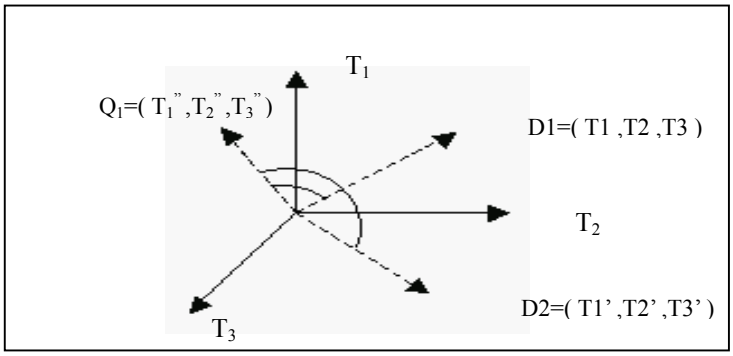

圖 5 向量空間文件表示

表 2 相似性衡量方式

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：Salton, G., A Theory of Indexing, Regional Conference Series in Application Mathematics, Society for Industrial and Applied Mathematics,1975, p.318

(三) 奇異值分解

在線性代數上常利用矩陣來作數值分析的工具(解聯立方程式組)，但是只有在矩陣是非奇異矩陣(nonsingular matrix)的情況下，才會有唯一解，碰到原始資料矩陣是一個奇異矩陣(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。因此 SVD 運用到特徵值(eigenvalue)與特徵向量(eigenvector)求解的過程，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，並將數據壓縮(過濾不相關資訊)。

在原始文件向量空間中，存在許多線性相依關係，或是繁雜無序。因此，群集檢索模型藉由縮減原始文件向量空間維度，以萃取索引詞彙之間潛在的共同因素結構，矩陣分解技術是使用 SVD 分解，將原始的索引詞--文件矩陣分解，建立成一個較小維度且相互直交(orthogonal)的群集索引構面，找出共同因素，代表索引詞之間潛在的群集索引構面，使的不同類別之下的詞彙完全不相關。

假如索引詞--文件矩陣 D 爲 t×d 矩陣，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)，則 D 之 SVD 定義爲 $D = U\Sigma V^T$ ，其中 Σ 爲 t×d 對角矩陣，且其值爲 D 的所有奇異值，若設 p=min{t,d}， 奇異值表示爲 $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_p \geq 0$ ；U 、V 則分別爲 t×t 及 d×d 的直交矩陣(圖 5)。

經由 SVD 分解後，可將原始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 $D = U\Sigma V^T$ 改變爲 $D_k = U_k \Sigma_k V_k^T$ ，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 SVD 模型中 U_k 、Σ_k 、V_k^T 的維度各縮小爲 t×k、k×k、k×d，且 D_k 的秩亦縮小爲 k，共同因素 k 是一個比原始索引詞向量空間小的數值，Σ_k 則保留了文件--索引詞矩陣中較大的 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)，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。

查詢句是由數個檢索詞彙所組成的，所以把它當作當成一個虛擬文件，每個詞彙的加權爲 TF=1。以矩陣 Q 表示，配合前面將 t 個詞彙，縮減成 k 個共同因素，也就是將原來的矩陣映射到較小的空間，轉換的公式爲 $Q_k = U_k \Sigma_k Q_k^T$ ，而轉換出的矩陣 Q_k ，將相當於文件矩陣 V_k 的一列，在使用 cosine coefficient 計算出文件與查詢句的相似值。

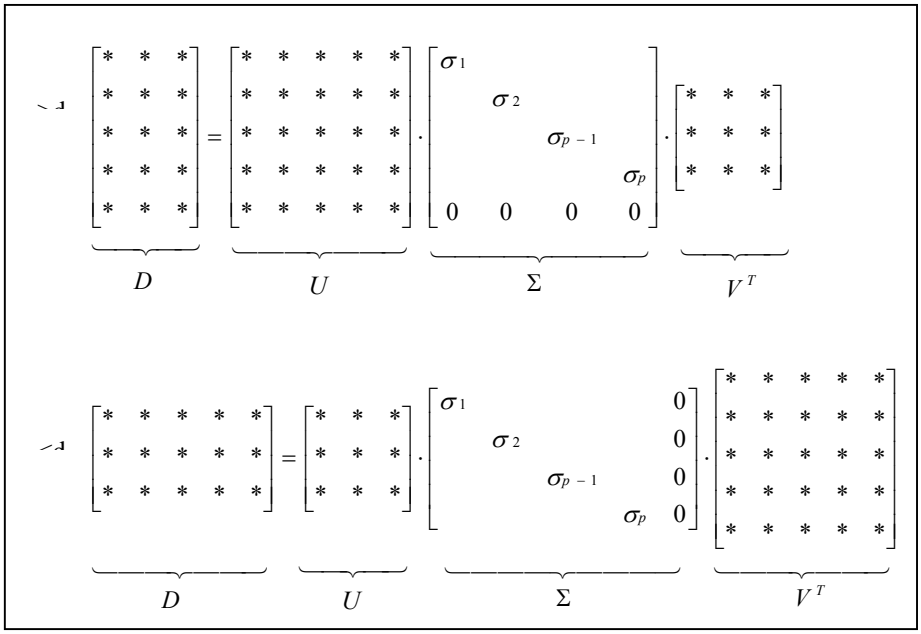

圖 6 奇異值與奇異值矩陣

資料來源：Michael W. Berry, Ztatko Drmac, Elizabeth R. Jessup, “Matrices, Vector Spaces, and Information Retrieval”, SIAM REVIEW, Vol.41, No. 2, 1999, p348.

參、實驗環境概要簡介

Tague-Sutcliffe(1996)界定「實驗研究(exoerunebtak research)」和「實際測試(operational test)」，所謂的實驗研究對「自變數」做某一些成程度控制，進行觀察，決定依變數是否因自變數不同而改變。而實際測試在於實際檢索環境中進行，以了解使用者、系統、以及人機互動之情形[吳美美，民國 90 年]，下面將針對實驗環境必要性自變數如文件集、查詢詞彙、查詢句設計與相關判斷做描述性介紹。

一、實驗語料

進行績效評估之前，第一步驟就是收集資料，目前中文資訊檢索研究的阻礙在於無完整的全文實驗文件集，因此許多研究學者通常必須自建，或者經由國家圖書館等相關單位授權取得文件集等。本實驗文件集的選擇，主要是以八十二年兒童日報新聞報導資料爲主，初期整理分類成兒童福利、醫藥、環保及專欄等類別的文件集，經考量選取其中的醫藥新聞語料做爲實驗對象，表 3 列舉本實驗文件集的基本性質。

表 3 兒童日報新聞文件集基本性質

新聞類別	文件數	總字數	每篇平均字數	人工選詞詞數	每篇平均詞數
醫藥	502	179450	357	2564	5

資料來源：本研究

Salton(1975)的研究，認爲時報週刊(TIME)語料比空氣動力學和生物醫學來的較好，因此實驗文件集對象的選擇會涉及內容相關的性質，而可能影響結果的分析。實驗文件集大小對於資訊檢索系統績效評估是否有影響？Salton(1986)認爲實驗文件集的大小對於檢索效能並沒有多大的影響，在西文資訊檢索中，主要具有規範的實驗文件集，都均提供文件數量、查詢句數量、與相關文件編號，下表 4 爲本研究與西文資訊檢索實驗文件集之比較。

表 4 本研究與西文常用資訊檢索實驗文件集之比較

測試文件集	主題性質	文件數量	查詢句數量	文件/查詢句比	平均相關文件數
INSPEC	電機工程學	12,684	77(2)	164.7	33
NPL	電機工程學	11,423	93(1)	122.8	22.4
LISA	圖書館學	6,004	35	171.5	10.8
CACM	ACM 通訊	3,204	64	50	15.3
CRANFIELD	航空學	1,400	225	6.2	8.2
Cystic Fibrosis	醫學	1,239	100	12.4	31.9
MEDLARS	醫藥	1,033	30	34.4	23.2
兒童日報醫療	一般性醫學	502	10	50.2	8.2
TIME	一般性	425	83	5.1	3.9

資料來源：本研究

說明：(1)NPL 文件集之全部查詢句為 100 (2)INSPEC 文件集之全部查詢句為 84
(3)CISI 文件集之全部查詢句為 111

二、索引詞彙

兒童日報是提供給兒童閱讀爲主的讀物，所使用的詞彙應是日常生活基本、常用的，應該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。

索引詞彙的選取與中文自動斷詞技術密不可分，於選取索引詞彙作業方面採用人工，其主要準則是依據中國國家標準的「西文單一語文索引典編制標準」(CNS-13224)與中華人民共和國國家標準的「漢語敘詞表編制規則」(GB-13190-91)。

在人工選取下，由於索引詞的選取認定涉及到個人的主觀判斷，相對可能相同索引詞在不同的主題內容下，可能會有不同的認定與選取，故本實驗對索引詞的索取原則有下列三點：

1. 從研究目的及整體考量選取可能的索引詞彙。
2. 從使用者觀點考量是否做爲檢索的屬性（檢索點）。
3. 從局部的文件內容（段落）的前、後文語意脈絡（context）下考量選取有索引意義的詞彙。

三、查詢句設計與相關判斷

查詢句的設計具有間接影響相關判斷與評量結果，一般而言，檢索者在檢索時都是以多個檢索詞彙來代表查詢的主題，因此本實驗之查詢句也是多詞組合方式呈現，主要以三到五個檢索詞彙爲主，且爲了簡化複雜性，其檢索詞彙並未經過任何加權。

查詢句相關文件高低，是否真影響檢索系統？許多研究學者均認爲查詢句之相關文件過低，可能會影響結果，從表 4 可以理解在西文語料庫中，專業性語料的平均的相關文件較多。下表 5 主要選出 CRANFIELD、TIME、CACM 與本實驗語料做相關文件數量探討，分析四個語料庫發現除 CACM 語料庫外，其他相關文件在八篇內的查詢句數量幾乎佔全部 60%。

另外查詢句相關文件的判斷，在客觀第三者對語料進行人工分類完成後，研究者根據查詢句設計所要檢索的主題，由研究者事先在該小類內容中確認與查詢句相關的文件，判定相關的過程是在實驗進行之前已經確認，實驗進行時是依據事先判定的相關文件註記來評量。

表 5 CRANFIELD、TIME、CACM 與本實驗語料的相關文件數量分析

相關文	兒童日報醫療		CRANFIELD		TIME		CACM	
件數量	次數	累積%	次數	累積%	次數	累積%	次數	累積%
1	0	0%	0	0%	23	27.71%	3	5.77%
2	0	0%	6	2.67%	25	57.83%	1	7.69%
3	0	0%	29	15.56%	6	65.06%	5	17.31%
4	0	0%	19	24.00%	1	66.27%	2	21.15%
5	2	20.00%	26	35.56%	9	77.11%	3	26.92%
6	2	40.00%	28	48.00%	3	80.72%	1	28.85%
7	2	60.00%	21	57.33%	3	84.34%	0	28.85%
8	0	60.00%	15	64.00%	4	89.16%	3	34.62%
9	0	60.00%	14	70.22%	3	92.77%	1	36.54%
10	0	60.00%	15	76.89%	0	92.77%	2	40.38%
11	2	80.00%	8	80.44%	1	93.98%	4	48.08%
12	2	100.00%	7	83.56%	1	95.18%	5	57.69%

資料來源：本研究

肆、實驗結果分析

實驗主要是探討不同群集索引模式檢索的效能評量與尋求各模式的最適構面與界限值，查詢句設計方面，針對相關文件數量篩選[2]，擷取 10 個查詢句做實驗測試，下面爲實驗結果分析比較。

一、TF 索引模式結果

以 TF 索引模式的群集檢索，實驗結果檢出率的值，大約落在 40%至 95%之間，平均落在 63%左右。索引構面縮小，使的檢出率提昇；界限值提高，檢出率跟著下降，其中對檢出率的影響，界限值比索引構面較具影響力。

精確率的值，大約落在 20%至 70%之間，平均落在 45%以上，索引構面縮小，使的精確率下降；界限值提高，精確率跟著上昇，在界限值 0.4 以上，不論構面大小，其精確率約在 40%以上。

表 6 TF 索引模式檢索結果

	構面 40		構面 60		構面 80		構面 100	
界限值	檢出率	精確率	檢出率	精確率	檢出率	精確率	檢出率	精確率
0.2	91.40%	24.47%	85.26%	26.54%	76.58%	31.04%	70.52%	35.55%
0.3	83.98%	35.27%	74.75%	36.36%	64.02%	40.85%	63.18%	48.54%
0.4	70.97%	43.15%	62.19%	45.06%	57.44%	49.36%	52.36%	48.36%
0.5	62.24%	51.76%	57.27%	51.77%	50.69%	54.74%	50.10%	55.19%
0.6	56.57%	61.49%	44.45%	59.37%	44.17%	65.38%	41.67%	69.67%

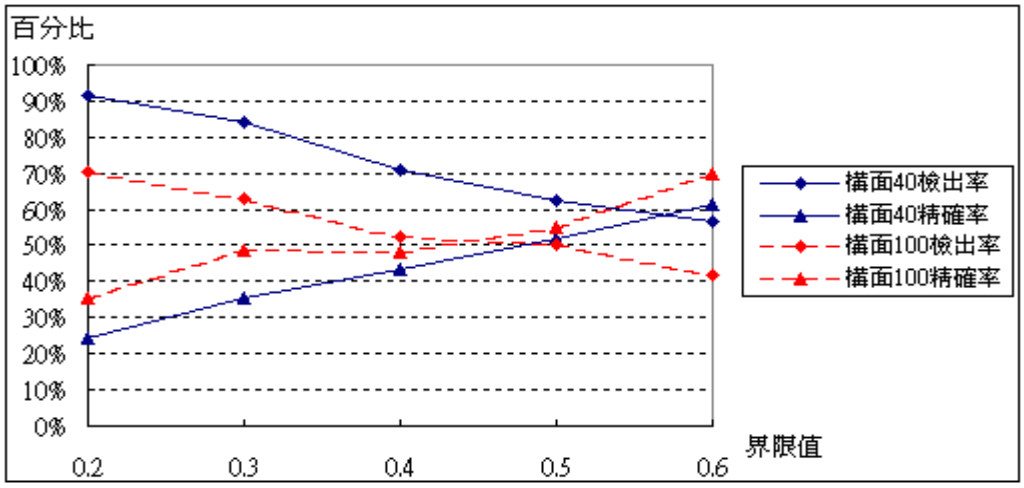

圖 7 TF 索引模式，構面 40 與 100，各界限值之檢出率與精確率

[2] 本研究針對相關文件數量篩選，擷取相關文件數量 5-12 篇之查詢句。

二、IDF 索引模式結果

從表 7 可以看出總體檢出率的平均值落在 80%以上，在索引構面 100、界限值 0.6 時，檢出率爲 57.36%，此時爲最低點。當索引構面縮小時，檢出率將會提昇，另外，界限值的提高，檢出率會跟隨著下降，檢出率在索引構面 40 時，不論界限值高低，均保持在 85%以上。

精確率部分總體平均值在 35%以上。索引構面縮小，使的精確率下降；界限值提高，精確率跟著上昇，就精確率在索引構面 100 而言，不論界限值高低，都保持在 35%以上，界限值在 0.6 時，不論索引構面大小，也保持在 35%以上。

綜合上面檢出率與精確率變化情況來看，檢出率與索引構面、界限值之間的關係成反比狀況；反之，精確率卻與索引構面、界限值成正比狀況，另外當界限值大於 0.5 以後，會有少數無法檢索到文件情況發生。

表 7 IDF 索引模式檢索結果

	構面 40		構面 60		構面 80		構面 100	
界限值	檢出率	精確率	檢出率	精確率	檢出率	精確率	檢出率	精確率
0.2	100.00%	17.99%	96.83%	22.28%	92.36%	27.16%	87.84%	35.62%
0.3	99.17%	23.82%	92.36%	28.67%	84.51%	36.65%	81.10%	47.66%
0.4	96.90%	28.80%	88.35%	33.92%	74.61%	43.96%	72.87%	50.07%
0.5	94.18%	33.51%	82.33%	36.36%	71.44%	48.81%	63.55%	55.37%
0.6	87.67%	39.42%	77.49%	42.78%	65.69%	54.51%	57.36%	60.87%

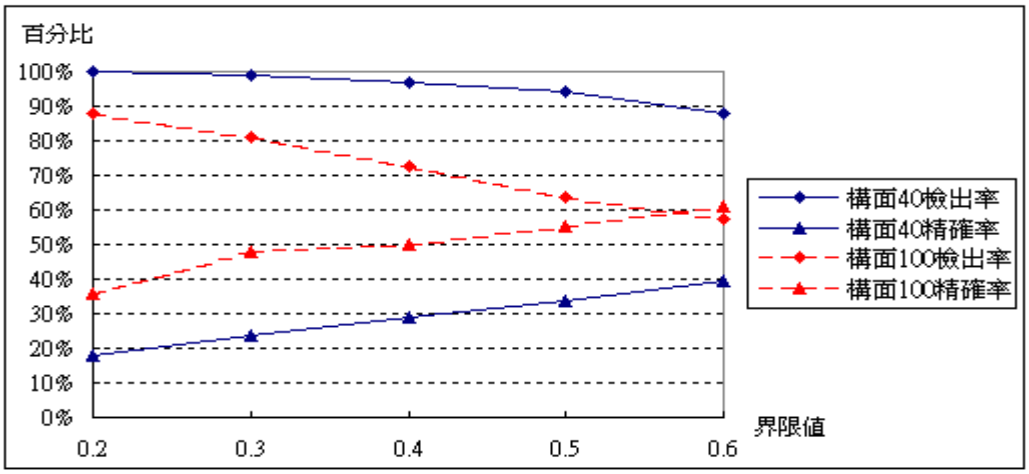

圖 8 IDF 索引模式，構面 40 與 100，各界限值之檢出率與精確率

三、IDF 與 TF 索引模式結果分析

針對 IDF 與 TF 索引模式檢索的結果，兩者的檢出率確實差距很大，而在精確率上差距就蠻小，Sparck Jones(1974)認爲不同檢索系統效能差距在 5%-10%，即有引人注目性(noticeable)的不同差異，若差距在 10%以上，則有重要性(material)的不同差異。對於兩者的差異，實驗仍運用 T-test 檢定的 P 值，在顯著水準 0.1 與 0.5 下，探討兩模式檢出率與精確率之間差距是否顯著，以辨識 IDF 索引模式與 TF 索引模式的績效。

IDF 與 TF 索引模式檢索評量比較檢出率方面，令 H_1：IDF 索引模式平均檢出率＞TF 模式平均檢出率，H_0：IDF 索引模式平均檢出率≦模式平均檢出率，發現在同樣索引構面與界限值下，除構面 100 外，其餘的 P 值均小於顯著水準 0.3，故否決 H_0，表示本實驗以 IDF 索引模式的群集索引的檢出率確實高於 TF 索引模式的群集索引之跡象。

從精準率差距表(表 9)看來 IDF 索引模式檢索的值略低於 TF 索引模式，經檢定結果 P 值小於顯著水準 0.3，分別在界限值 0.3、索引構面 40 與 80，其餘 P 值均高於顯著水準 0.3，表示 TF 索引模式的精確率高於 IDF 索引模式並沒有很顯著。

從上面實驗結果描述上，可得到兩點結果：

1. 不論 IDF 索引模式或 TF 索引模式，其檢出率隨者索引構面增加而降低；而精確率卻呈現相反現象，亦符合檢出率提昇，精確率降低的理論。
2. 界限值與檢出率呈現反比現象，刀界限值越低則檢出率越高，相反界限值越高則檢出率越低；而界限值與精確率呈現正比現象。

根據檢出率與精確率之 T 檢定之後，就檢出率而言，以 IDF 索引模式較 TF 索引模式佳，而精確率上，於界限值 0.5 以上，IDF 索引模式就較 TF 索引模式略遜一籌，其主要原因是 IDF 索引模式與 TF 索引模式，在界限值大於 0.5 以上時，有許多查詢句是無法檢索到文件，其中 IDF 索引模式數量較 TF 索引模式多，因此影響到 IDF 索引模式的檢出率與精確率。

如何尋求最適索引構面的區間，群集索引的一大問題，因爲影響索引構面的因素包含：文件集的性質、大小，索引語言的形式，查詢語言的使用與其特性。在統計學上利用特徵值的大小，或者依特徵值由大而小排列的陡階圖來決定因素的選取標準。

在西文檢索研究中，Deerwester(1990)在其 5823 詞、1033 篇文件的環境下，建議維度(構面)在 50 至 100 之間，當維度大於 20 以後，精確率會提高，當維度大於 100 之後就逐漸下降。

許多研究中嘗試利用因素分析選取因素個數的作法，以特徵值的陡階圖觀察，在特徵值突然降低且趨於平緩之處爲界線，選取前面特徵值較大共同因素。但是在本實驗中發現一個問題，奇異值陡階圖沒有出現陡階狀況(突然降低)，而是很平緩的遞減(圖 9)。這樣的情形只能以因素分析決策標準提出一些數值作爲參考。

表 8 IDF 與 TF 索引模式檢索之檢出率差距與 P 值

界限值	IDF40-TF40		IDF60-TF60		IDF80-TF80		IDF100-TF100	
	差距	P 值	差距	P 值	差距	P 值	差距	P 值
0.2	8.60%	0.068	11.57%	0.062	15.78%	0.122	17.32%	0.123
0.3	15.19%	0.021	17.61%	0.070	20.49%	0.092	17.92%	0.184
0.4	25.93%	0.009	26.16%	0.019	17.17%	0.266	20.51%	0.209
0.5	31.94%	0.002	25.06%	0.062	20.75%	0.214	13.45%	0.414
0.6	31.10%	0.007	33.04%	0.028	21.52%	0.182	15.69%	0.325

表示小於顯著水準 0.3

表 9 IDF 與 TF 索引模式檢索之精確率差距與 P 值

界限值	IDF40-TF40		IDF60-TF60		IDF80-TF80		IDF100-TF100	
	差距	P 值	差距	P 值	差距	P 值	差距	P 值
0.2	-6.48%	0.365	-4.26%	0.577	-3.88%	0.641	0.07%	0.994
0.3	-11.45%	0.186	-7.69%	0.414	-4.20%	0.690	-0.88%	0.948
0.4	-14.35%	0.158	-11.14%	0.264	-5.40%	0.690	1.71%	0.897
0.5	-18.25%	0.097	-15.41%	0.111	-5.93%	0.665	0.18%	0.990
0.6	-22.07%	0.064	-16.59%	0.162	-10.87%	0.489	-8.80%	0.589

表示小於顯著水準 0.3

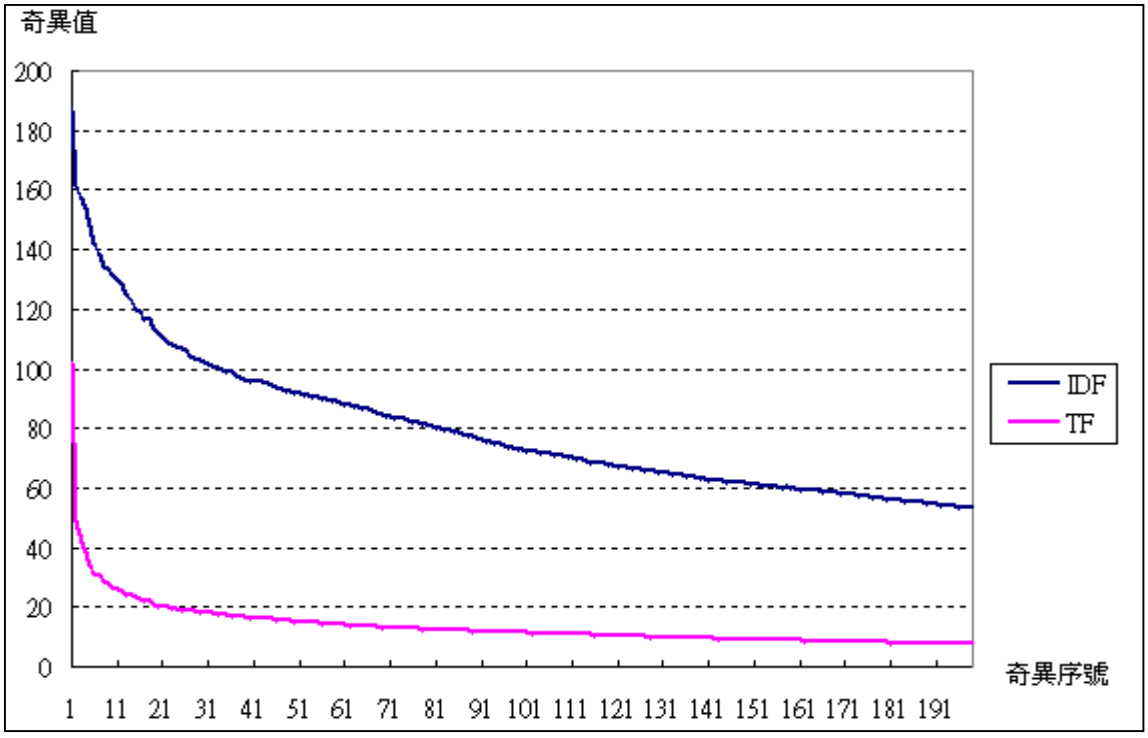

圖 9 IDF 與 TF 索引模式奇異值陡階圖

以奇異值爲索引構面，是否能表示奇異值與檢出率、精確率之間存在一定關係？從圖 9 可以發現，當 IDF 索引模式檢出率較 TF 索引模式較佳時，相對 IDF 的奇異值曲線也較 TF 的奇異值曲線高；當各奇異值曲線以陡峭遞減時，各索引模式的精確率也呈現遞減狀況。

伍、結論與建議

下面將針對整個實驗研究過程與結果做一個歸納結論，最後提出幾點未來研究上建議。

一、結論

實驗結果在兒童日報醫療文件集中，IDF 索引模式的檢出率確實比 TF 索引模式要來的佳，但是精確率卻略遜於 TF 索引模式，事實上，檢索者對於系統的檢出率與精確率的要求水準卻沒有一定的答案，一般而言，倘若系統的檢出率與精確率能同時達到 50% 以上，就能滿足檢索者的需求。

以上述檢出率與精確率同時達到 50%以上的標準來看，在 IDF 索引模式，最適索引構面應該是 100、界限值爲 0.4，而 TF 索引模式在界限值 0.5 下，任何構面都還令人滿意。

對於系統的效能評量仍以檢出率及精確率爲主，但是過去西文檢索的研究，所呈現方式是以估計十等分檢出率水準下精確率的值，但是無法剛好取得適當檢出率點，因此必須以外推法或內推法來求得值，除非查詢句的相關文件數量爲十的倍數，否則勢必會產生誤差值。

本研究所研擬的評量方式，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。同時考量觀察群集索引模型在不同索引構面下的系統效能，使得在系統評量上，必須從檢出率、精確率、索引構面及相關係數界限值等四個角度去綜合評論。所以我們觀察的系統效能不只是檢出率與精確率的傳統觀點，還包括系統在最適群集索引構面區間的選擇，以及在合適的索引構面區間內的穩定性與一致性。

二、建議

相對於國外常用的實驗文件集，本研究查詢句的數量稍嫌不足，最少都有 30 個查詢句，從表 4 的比較可以發現各文件集和查詢句比值差異很大，至於，到底實驗的測試文件集與查詢句應該要多少？才具有實驗的效度，仍然是一個開放性的問題，另外，奇異值與檢出率、精確率之間是否存在的關係？也需更多實驗加以證明，希望在未來研究上能有更清楚的探討。

再來因中文文字的特性是字字相連，在斷詞上需要特別處理，雖然中央研究院詞庫小組已有良好的斷詞技術，未來以自動選詞爲實驗對象，可能有助於簡潔複雜的人工選詞作業。

中文資訊檢索的研究涉及了許多議題，過去有陳淑美、蔣俊霞、楊允文、黃雲龍、顧皓光、廖書賢等人做過相關理論研究，因此，對於中文資訊檢索的研究發展，建構一個標準的評估環境，如大型文件集、標準評量模式等，這是學術研究上有其必要性，未來研究若能在標準的評估環境下實驗，將有助於系統機制的發展與改善檢索效能。

誌謝

感謝中央研究院資訊科學研究所提供實驗環境。

陸、參考文獻

1. 吳美美，「中文資訊檢索系統使用研究」，台灣學生書局，台北，民國 90 年 4 月。
2. 張琪玉，「情報語言學基礎」，武漢大學出版社，湖北，民國 97 年 10 月。
3. 黃雲龍，「中文全文文件群集索引理論研究與實證」，圖書與資訊學刊，第 24 期，民國 87 年 3 月，pp44-68。
4. 黃雲龍，「中文全文文件群集索引理論研究--向量空間模型(Vector-Space Model)的建構」，國立台灣大學商學研究所博士論文，民國 86 年 6 月。
5. 黃慕萱，「資訊檢索中『相關』概念之研究」，臺灣學生書局，台北，民國 85 年 4 月。
6. 廖書賢，「中文全文資訊檢索研究實驗平臺規劃與建置」，國立台灣大學資訊管理學研究所碩士論文，民國 87 年 6 月。
7. 廖書賢、黃雲龍，「從 TREC 的發展趨勢回顧中文全文資訊檢索關鍵議題」，第五屆三軍官校基礎學術研討會論文，民國 87 年 5 月。
8. Blair, D. C. & Maron, M. E., “An Evaluation of Retrieval Effectiveness for A Full-Text Document- Retrieval System”, Communications of the ACM, Vol. 28, No. 3, Mar. 1985, pp.289-299.
9. Blair, D. C. & Maron, M. E., “Full-Text Information Retrieval: Further Analysis and Clarification”, Information Processing & Management, Vol. 26, No. 3, 1990, pp.437-447.
10. Chowdhury, G. G., “Introduction to Modem Information Retrieval”, Library Association, London, 1999.
11. Cooper, W. S., “A Definition of Relevance for Information Retrieval”, Information Storage & Retrieval, 1971, Vol:7, pp.19-37.
12. Ellis, D., “The Dilemma of Measurement in information Retrieval Research”, Journal of the American Society for Information Science, Vol. 47, No.1, 1996, pp.23-36.
13. Iivonen, M., “Consistency in Selection of Search Concepts and Search Terms”, Information Processing & Management, Vol. 31, No. 2, 1995, pp173-190.
14. Jean M. T., “Some Perspective on the Evaluation of Information Retrieval System”, Journal of the American Society for Information Science, Vol. 47, No.1, 1996, pp.1-3.
15. Johanson G., “Information, knowledge and research”, Journal of Information Science, Vol. 23, No. 2, 1997, pp.103-109.

16. Louise, T. Su, “The Relevance of Recall and Precision in User Evaluation”, Journal of the American Society for Information Science, Vol. 45, No. 3, 1994, pp.207-217.
17. Michael, B. & Fredric, G., “The Relationship between Recall and Precision”, Journal of the American Society for Information Science, Vol. 45, No. 1, 1994, pp.12-19.
18. Michael, W. B. & Ztatko, D. & Elizabeth, R. J., “Matrices, Vector Spaces, and Information Retrieval”, SIAM REVIEW, Vol.41, No. 2, 1999, pp.335-362.
19. Salton, G., “The SMART Retrieval System, Experiments in Automatic Document Processing”, Prentice Hall, Inc., Englewood Cliffs, N. J., 1971.
20. Salton, G., “A Theory of Indexing ”, Regional Conference Series in Application Mathematics, Society for Industrial and Applied Mathematics, 1975.
21. Salton, G., “Automatic Text Processing: the transformation, analysis, and retrieval of information by computer ”, Addison-Wesley Publishing Company, New York, 1989.
22. Van Rijsbergen, C. J. & Croft, W. B., “Document Clustering: An Evaluation of Some Experiments With The Cranfield 1400 Collection”, Information Processing & Management, Vol. 11, 1975, pp171-182.
23. Voorhees, E & Harman, D., “Overview of the Ninth Text Retrieval Conference (TREC-8) ”, The Ninth Text Retrieval Conference (TREC-8),NIST Special publication. http://www.trec.org/
24. Voorhees, E & Harman, D., “Overview of the Ninth Text Retrieval Conference (TREC-9) ”, The Ninth Text Retrieval Conference (TREC-9),NIST Special publication. http://www.trec.org/

柒、附錄

文件集：兒童日報醫療　　文件集大小：502 篇

檢索模式：IDF 索引模式　構面：100　界限值：0.4

編號		相關文件數	檢出數	檢出率	精確率
1	愛滋病 帶原者 毒品 針頭 輸血	5	8	100.00%	62.50%
2	無菌性腦模炎 發燒 頭痛 嘔吐 脖子僵硬	5	15	100.00%	33.33%
3	視力不良 視力保健 近視 護眼	11	9	27.27%	33.33%
4	蟯蟲 寄生蟲 蛔蟲 小學生	6	2	83.33%	100.00%
5	感冒 病毒 發燒	7	18	100.00%	38.89%
6	痢疾 飲用水 法定傳染病	12	31	75.00%	29.03%
7	便當 學校 食物中毒 營養午餐 餐盒業	11	9	45.45%	55.56%
8	小兒痲痺 脊髓灰質炎疫苗 沙賓口服疫 病毒	7	7	14.29%	14.29%
9	減肥 減重 運動	12	11	83.33%	90.91%
10	幼稚園 托兒所 健康檢查	6	14	100.00%	42.86%
平均		8.2	12.4	72.87%	50.07%

文件集：兒童日報醫療　　文件集大小：502 篇

檢索模式：TF 索引模式　構面：100　界限值：0.4

編號		相關文件數	檢出數	檢出率	精確率
1	愛滋病 帶原者 毒品 針頭 輸血	5	7	100.00%	71.43%
2	無菌性腦模炎 發燒 頭痛 嘔吐 脖子僵硬	5	12	100.00%	41.67%
3	視力不良 視力保健 近視 護眼	11	5	18.18%	40.00%
4	蟯蟲 寄生蟲 蛔蟲 小學生	6	5	83.33%	100.00%
5	感冒 病毒 發燒	7	15	85.71%	40.00%
6	痢疾 飲用水 法定傳染病	12	8	8.33%	12.50%
7	便當 學校 食物中毒 營養午餐 餐盒業	11	7	36.36%	57.14%
8	小兒痲痺 脊髓灰質炎疫苗 沙賓口服疫 病毒	7	6	0.00%	0.00%
9	減肥 減重 運動	12	6	41.67%	83.33%
10	幼稚園 托兒所 健康檢查	6	8	50.00%	37.50%
平均		8.2	7.9	52.36%	48.36%