大學圖書館
2卷3期(民國87年7月) 頁4-26

中文全文資訊檢索研究架構與重要議題探討

A Study of Research Framework and Critical Issues Concerning Mandarin Chinese Full Text Information Retrieval

黃雲龍 Yun-long Huang

景文技術學院資訊管理技術系助理教授
Assistant Professor, Department of Information Management
Jin-Wen Institute of Technology

【摘 要】

全文資訊檢索研究已經成爲一個跨學域的研究問題，但是過去學者缺乏對於研究架構的討論，面對快速變遷的環境，新的研究議題與挑戰更多、更複雜。本文嘗試釐清全文資訊檢索研究問題的本質，並提出一個研究架構雛形，討論相關的研究範疇與理論基礎，並檢視重要的研究發展趨勢，指引中文全文資訊檢索研究的重要議題。

【 Abstract 】

Full text information retrieval becomes the focus of interest in the area of interdisciplinary studies. This paper attempts to clarify some basic questions concerning the full text information retrieval as well as the scope of theory and related research.

關 鍵 詞：研究架構；全文資訊檢索；資訊搜尋行爲；相關；全文檢索會議
Keywords：Research framework；Full text information retrieval；
Information seeking behavior；Relevance；TREC

人文學者、公私立大學教師、大學生、工程師、科學家、醫生、證券分析師等。研究結果可謂相當豐富。

圖書館是資訊的貯藏所，我們期盼提高圖書館的使用率，發揮圖書館功能，了解使用者的資訊尋求行爲是前提。即使在網際網路盛行的今日，圖書館要整理、提供網路資源予使用者，使用者的網路資源尋求及使用行爲更該有所認識。如此才能幫助圖書館設計更好的資訊系統，進而提供更完善的資訊服務，使圖書館的功能能充分發揮。在此情況下相信必能提升圖書館在使用者生活或工作上的重要性，間接地將會提升圖書館員的社會地位。因此，我們期待有更多的使用者資訊尋求行爲的研究出現。

（葉乃靜）

壹、前言

資訊技術的變革對資訊的儲存、呈現、處理與交換的方式產生很大的影響，最根本的改變就是文件電子化。再加上資訊網路的普及，透過網際網路的通訊協定（如 TCP/IP），全世界的網路連成一體，更加速的促成電子文件時代的來臨。在文件電子化愈來愈普及以後，資料處理與應用將是資訊資源運用的重要課題。尤其是全文資料 (full text) 的檢索，我們需要研究在浩瀚的資料空間裡，如何及時、有效的取得良好品質的資訊。

傳統資訊檢索 (information retrieval) 領域的發展主要集中在圖書館學，例如主題分析、分類、索引與檢索等知識的研究。由於資訊科技的快速發展，電腦具有大量儲存及快速處理的特性，應用電腦以自動檢索資訊的研究應運而生。尤其在今日，資訊爆炸、出版品大量激增，學科主題分化精細，即使圖書館提供使用者各式各樣的輔助工具，使用者在進行資訊過濾與選擇時，仍然需要一個更有效率的工具，例如自動化文件檢索系統 (automatic document retrieval system)，來幫助使用者資訊搜尋工作的進行。

由於資料形式與資料種類的不同，我們需要不同的管理技術與策略。資料種類可依目前資料處理問題概分爲結構化與非結構化資料。資料形式則包括有檔案記錄、字母與數字型資料、文章、圖表、影像、動畫及聲音等。結構化資料如過去商業應用的關連式資料庫，大都是處理以關連表來記錄格式化屬性值的資料，資料形式則包括有檔案記錄、字母與數字型資料，此類資料處理問題已經獲致相當程度的解決。（註 1）現存的企業資料中還有很多非結構化的資料處理問題，例如公文、書信、會議記錄、規章辦法、技術規範、標準手冊、筆記、計畫書、契約書、備忘錄、工作記錄、公司出版品 ..... ，還有辦公室自動化以後的各類電子郵件、電報、傳眞等等，這些全文資料的處理與應用，將是未來的資料管理的重點。

其他各行各業所面對的資料，如大衆傳播業的電子報紙 (electronic newspaper)；出版業的電子書 (electronic book)；至於法規、法院案例、醫療診斷資料、技術報告、期刊、技術維護與作業手冊、公文 ..... 等電子文件，都需要新的全文資料庫技術、方法與模型，以解決處理、應用與管理問題。同時傳統資料庫需要與全文資料庫整合，才能充分管理各種不同形式與不同種類的資料。這是本文探究全文資訊檢索研究架構與關鍵議題的動機。

貳、全文資訊檢索研究的問題本質

爲了釐清問題的本質，以下將從全文資訊檢索系統技術的發展，電子化文件檢索環境的複雜性，全文資訊檢索的概念與「形式與內容」關係的

討論，說明全文資訊檢索研究的問題本質。

一、全文資訊檢索系統技術的發展

全文資訊檢索系統的自動化起源於 1945 年，開始是以微縮片 (microfilm) 形式儲存，提供人類的檢索。1960 年早期才開始出現電腦機讀格式的檢索系統。但是直至 1970 年初期系統設計仍以文件索引爲主，配合微縮片的全文，以索引詞彙做爲文件的擷取控制（註 2 ）。

眞正的電腦化全文儲存與檢索系統開發於 1960 年代的中期，主要的設計目標以實驗系統爲主，例如 Salton 等人所發展的 SMART 。隨後美國空軍於 1967 開發的 LITE(Legal Information Thru Electronics) 系統是應用於法律文件的第一個實務系統。 1960 年代的末期，最知名的系統是應用於 IBM 大型主機上的 STAIRS(Storage and Information Retrieval System) 。直至 1970 年代末期，開始有新聞全文文件的連線檢索系統應用，從此線上檢索服務開啓全文資訊檢索的成長期（註 3 ）。

二、電子化文件檢索環境的複雜性

全文資訊檢索乃是預先將文件按一定的方式組織和儲存（如特徵與分類），然後使用者根據檢索的需求查出資訊的過程。如果以電子化文件檢索系統環境作考量，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（請參考圖一）。由此可見全文文件資訊檢索的問題涵蓋廣泛。

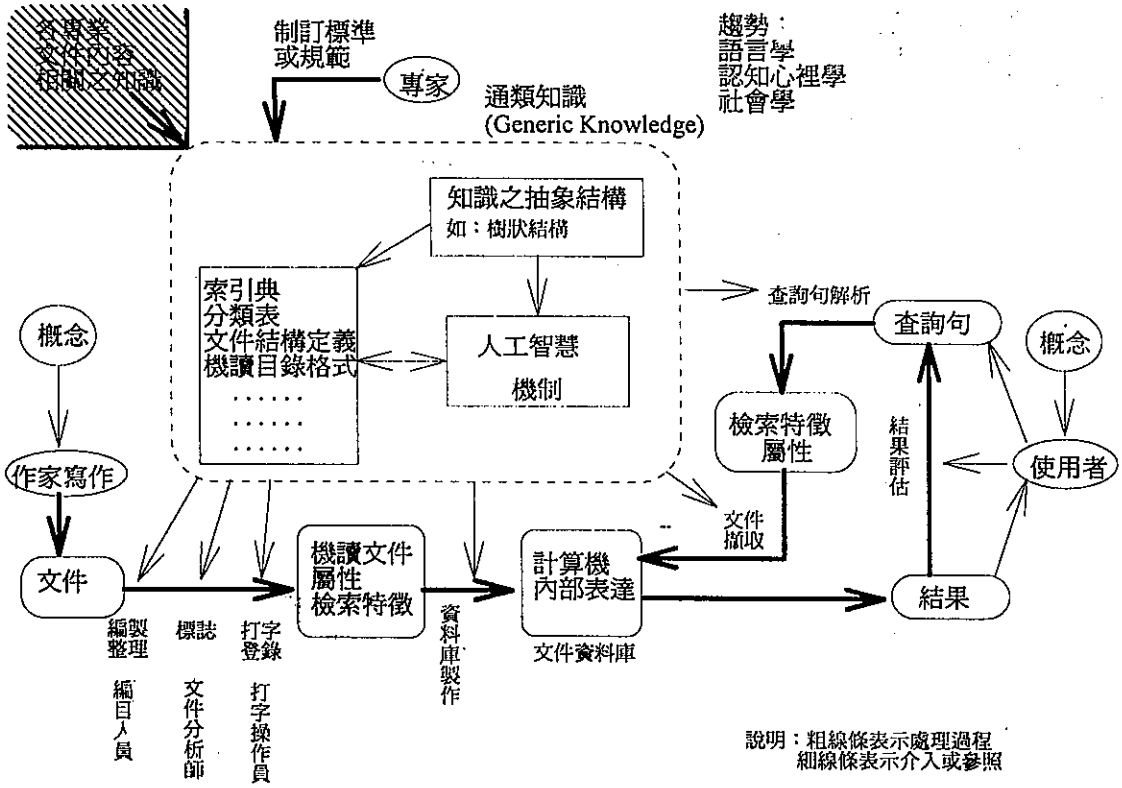
三、全文資訊檢索的概念

基本上，全文資訊檢索研究的問題是：系統如何擷取含有「符合使用者需求」的「相關資訊」的文件。（註 5 ）其中牽涉到不同使用者主觀的資訊需求，以及系統如何客觀的表達文件的內容訊息。 Lancaster 的實驗證實，在 700,000 個文件的集合，300 個查詢實驗中，主要的失敗來自於四項因素：(1)索引語言使用不當；(2)文件索引不當；(3)檢索問題陳述不當；(4)人機檢索介面互動不足。（註 6. ）

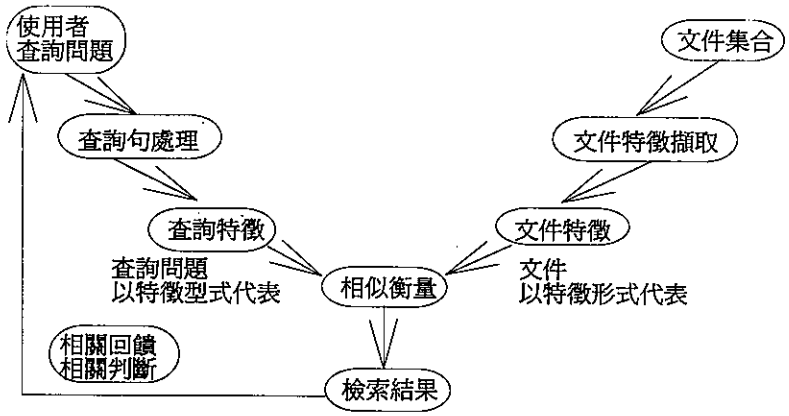
如果我們把問題的範疇概念化，專注於全文資訊檢索的概念上（請參考圖二），問題的重點就在於「形式與內容」的關係了。因爲，全文資訊檢索系統是一個結構化的制式系統，文件集合的內容主題都要轉換成檢索系統的內部形式代表；而使用者資訊需求也必須先轉換成查詢問題的描述，然後進一步轉換成系統可處理的查詢句形式。所以，在問題的概念層次上，全文資訊檢索問題的本質，就是「形式與內容」的關係。

四、形式與內容關係

爲了充分了解全文資訊檢索系統所涉及的「形式與內容」關係的問題，以下將從人類認知、概念、思考和語言文字的意義與彼此的關聯先做



圖一、電子文件檢索系統環境示意圖（註 4 ）



圖二、全文資訊檢索概念圖（註 7 ）

探討。

1. 認知與概念的內容意義

所謂認知 (cognition)，狹義而言是指認識或知道，Guilford & Hoepfner 指出「認知是一種察覺 (awareness) 直接發現，重新發現或是認得 (recognition) 各種形式的訊息」。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。（註 8 ）

思考可以視爲一種認知過程，它是人類最複雜的行爲方式。能夠具有使用符號來表示物體或事件的特性，當一個符號代表一組具有共同特性的事物時，我們說他指示一個概念。我們使用概念來整理及分類環境的事物及經驗，同時進行思考，而概念是認知最重要的單位。（註 9 ）

所謂概念 (concepts) ，係指人類大腦將感官所經歷或認知的事物，以一種抽象的信息表達，並以此抽象的信息，聯想、分析、歸納與判斷，從而認知到事物的存在與關係，此種抽象信息的表達即爲概念。（註10）這些概念經由文字的描述或語言的溝通，就是我們想要傳達的意義與內容。

我們所使用的語言、文字，皆爲概念的一種形式表達。文字的形、音、義雖然蘊含有概念的信息，但文字不一定是概念，它只是概念的代表。我們也可能以繪畫、雕塑、音樂或其他的形式藝術，來表達抽象的概念。概念是人類得以互相溝通、思維、反應事物本質屬性的方式。

*2.*概念的形成與發展

對於人類所具有的概念化能力，就是我們的思想與心智，透過不斷的學習發展而來。概念形成過程是一種假設測試的過程，需要利用現在已知和儲存的信息提出假設，人類根據這些假設庫，對任一刺激做出反應，如果反應正確則假設繼續應用；如果反應錯誤則假設更換。經由這樣的過程直到取得某個正確的假設，即形成某個概念。（註11）

藉著概念與概念之間的關係，我們可以區別客觀存在的事物，同時建立彼此之間的關連。圖書館所用的索引典正是一套顯現知識概念結構的辭典，爲資訊的儲存與檢索提供標準化的語彙，以確保使用者對同一主題分析處理資料與檢索時所用語彙的一致性。（註12）

*3.*語言文字的形式與思考

語言與思考開始時是分離並且獨立的。在人的成長過程思考通常先於語言的發展。當成長到能夠具體應用語言思考的時候，語言的思考功能內化爲人本身的思維。（註13）我們創造的語言，就是一套與概念相對應的符號。所以語言是一個人將內心的思想與意識表達，並與人溝通的工具。因此，語言會限制我們思考的範圍，也會影響我們思考的內容。如果語言的組織和形式有些缺陷或盲點，我們的思考可能也會有某些缺陷或盲點。所以我們說語言的世界不一定等於我們的經驗世界。（註14）

文字是記錄語言的書寫符號系統。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」。（註15）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。雖然概念與詞彙有著形式

與內容的對應關係，但並不是一對一的絕對關係，而且會隨著時空改變。黃蕙株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）。（註16）

參、全文資訊檢索研究的參考架構

對於任何一個學域而言，都需要完整的研究架構 (Research Framework)，研究架構的功能主要在於界定研究的範疇、引導研究方向，這樣才能累積研究成果、建立理論典範。從上述問題本質的討論可知，資訊檢索研究是一個跨學域的問題，思考一個明確的研究架構將有助於引導未來的研究發展。

一、研究架構的建議

從前述技術發展歷程可知，全文資訊檢索研究已經四十餘年，但是始終未見學者提出一個研究參考架構。本文嘗試參考資訊管理學域中 Ives, Hamilton, and Davis 在 1980 年所提的「電腦化管理資訊系統研究架構」，（註17）同時參考全文資訊檢索的概念（如圖二）（見頁 7），以及 Kemp(1974) 提出的「相關、適當、資訊需求與檢索問題的關係」架構，（註18）建構全文資訊檢索的研究參考架構（圖三）。

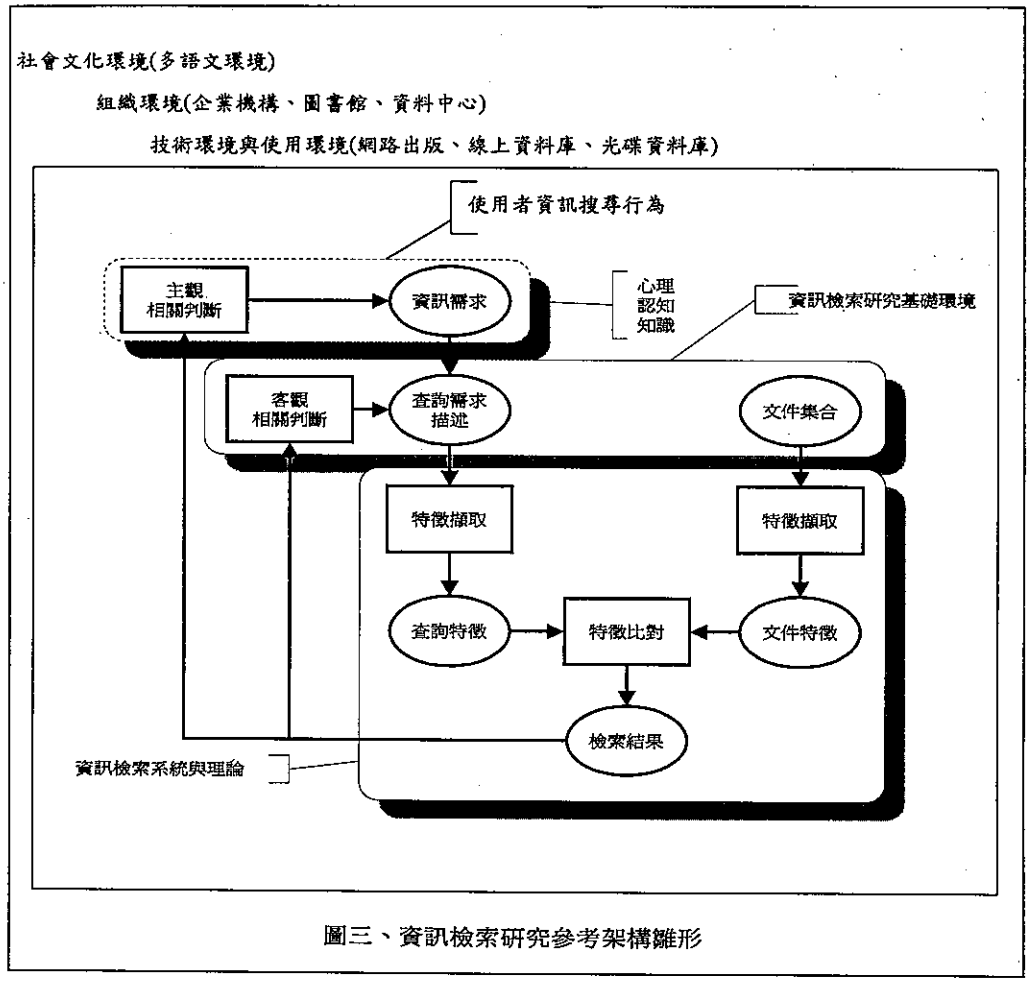
Ives, Hamilton, and Davis 所提的架構涉及三種主要變數：環境、績效與資訊子系統。環境變數中又分爲五類：外部環境、組織環境、使用者環境、資訊系統開發環境與運作環境；績效變數則包含：發展、運作與使用績效；資訊子系統則討論個別子系統的特質與發展上、應用上的差異。（註19）

從圖三可知，本文嘗試從環境、使用者與系統三個主要變數上描述全文資訊檢索的研究架構雛形。這個架構略爲簡化了 Ives, Hamilton, and Davis 所提的變數，而焦點仍在於使用者資訊需求與全文資訊檢索的概念。

從研究架構上可以區分出一些不同的研究範疇 (scope)，包括：使用者資訊搜尋行爲 (Information Seeking Behavior)、資訊檢索研究基礎環境、資訊檢索系統與理論部份。如果研究範疇設定的越明確，這些範疇可以再區隔爲更小的議題，例如在資訊檢索系統與理論範疇內，從文件集合的途徑出發就是自動索引理論的研究議題；從查詢需求途徑可以探討人機介面、自然語言查詢與查詢句擴張等議題。如果研究範疇超越一個領域，那麼所需的實驗設計就更複雜，例如檢索效能評估，就要同時考慮檢索品質、檢索效率、檢索系統本身與檢索者檢索技巧等不同觀點。（註20）

如果再進一步延伸到技術環境的變項，例如探討網路上網頁的搜尋、

線上資料庫檢索等議題，那麼即使是原來單一範疇內最基本、最單純的議題也有新的理論發展空間，例如智慧型代理系統 (Intelligent Agent) 的開發問題，原本只是單純的搜尋引擎問題，現在就變成必須同時考慮兩個以上的研究範疇問題了。因此，若再加上組織環境的變異，問題的複雜度就越高，研究者只有逐步的放寬研究限制，取得更大的研究資源，或者植基於大量的研究成果上向前推進理論的疆域。



圖三、資訊檢索研究參考架構雛形

由研究架構來檢視過去的相關研究，雖然早期的研究偏重於自動化系統，但是檢索系統的規劃、開發、實施與運作的議題並未受重視。由於受限於研究資源如資訊技術的處理能力、電子文件資料庫不易取得等因素，學者也只能進行小型的實驗室研究。直到 Salton 從 1961 年起展開 SMART(System for Mechanical Analysis and Retrieval Text；簡稱 SMART) 研究計畫，利用 Cleverdon 在 1950 中期至 1960 中期完成的 Cranfield 研究的實驗文件爲基礎。（註21）自此才揭露自動索引理論的發展方向與藍圖（如圖三資訊檢索系統與理論區塊），Salton 建立的向量空間模型在近三十年的資訊檢索研究上是最簡單、也最有彈性的模型。

在此之後，圖書館學與資訊科學開始密切的互動與交流。但是彼此的研究範疇仍侷限於個別專注的焦點，而缺乏跨學域與整合的研究。資訊科學研究者專注於解決資訊檢索系統與理論問題，例如自動索引理論、文件自動分類、群集索引與檢索、自動產生索引典、相關回饋、模糊查詢、人工智慧的類神經網路自動學習以及自然語言處理技術的應用如口語輸入、自然語言查詢介面等；（註22）而圖書館學者則專注於資訊技術對圖書資訊的組織理論與使用者資訊尋求行爲的衝擊與影響。（註23）

隨著數位技術與使用環境的變遷，如資訊網路的普及，以及組織環境與外部文化環境的變異，許多新興的研究議題紛紛出籠，例如跨語言資訊檢索、多媒體資訊檢索、智慧型資訊檢索以及網路資源檢索等。（註24）除此之外，自然語言處理技術的引進則開啓了資訊擷取、文件摘要（註25）等更先進的研究；而強調先進檢索技術應用與使用者資訊尋求行爲特質的研究，則開啓了資訊分送(Routing)、資訊過濾與個人化資訊服務（類似資訊選粹 SDI 服務）的議題。（註26）

從這些先進研究議題可以發現，問題所涉及的學域更廣泛、更複雜，不僅是圖書館學與資訊科學，更包括語言學、認知心理學、社會學、大衆傳播學等，形成一個大型的科際組合學科。因此本文希望藉由研究架構的提出，使得各相關學域的學者有一個整合的思考方向，彼此相互的激盪，把彼此的研究範疇與疆域向外延伸，尋求一個理論發展的方向、累積研究成果、建立理論典範。

二、研究架構的基礎

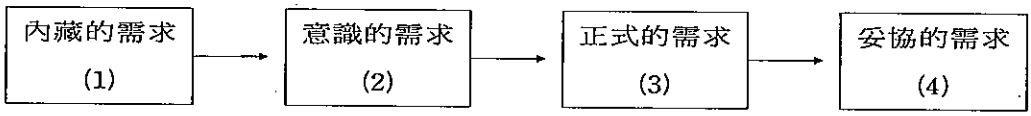
經由上述研究問題的本質描述與研究架構的檢視可知，全文資訊檢索研究有兩個關鍵因素：人類的資訊需求與評估檢索效能的相關判斷。以下將分別討論這兩個因素在資訊檢索研究上的意義，做爲本研究架構的基礎。

㈠資訊需求的意義

資訊搜尋行爲是描述使用者從察覺資訊需求到滿足資訊需求之間的一連串活動。Cooper(1971) 認爲所謂的

資訊需求是一種心理的狀態，只存在於人的心靈、思想之中。當人將其內隱的資訊需求以語言、或文字表達，此時內隱的資訊需求轉換爲外顯的查詢問題 (query)。（註27）爲了將問題輸入系統，必須要再經過一次轉換，此時稱爲查詢要求 (request)。而不管是 query 或是 request，都無法充份、完整地表達一個人的資訊需求。就程度上而言，這種定義資訊需求的方式是一種狹義的觀點。

Taylor(1968) 將資訊需求劃分爲四個轉換階段（圖四）：*1.*內藏的需求 (the visceral need)、*2.*意識的需求 (the conscious need)、*3.*正式的需求 (the formalized need)、*4.*妥協的需求 (the compromised need)。（註28）當使用者對資訊有所需求時，一開始這個需求是模糊不清、難以說明的，甚至根本無法查察其存在；隨著需求愈來愈強烈，他開始查察到存在著一種對於資訊的需求，並逐漸地加強對需求的了解與解釋；到了第三個階段，使用者能夠以某種符號（言語、文字）將他所認知到的資訊需求加以描述；最後，使用者將資訊需求的描述加以修正，以符合系統輸入的規格。這是資訊需求廣義的觀點。



圖四、使用者資訊需求形成的四個轉換階段

隨著階段的演進，資訊需求愈來愈清楚、愈來愈明確－由晦暗不明到能以符號表示，但在每一個轉換的過程中，無可避免地會對原始資訊需求產生一些改變，而這個改變有可能偏離了使用者眞正的資訊需求。例如圖書館專業的參考晤談目的在於幫助使用者適切地表達其資訊需求，但若是使用者表達錯誤，或是在溝通上雙方認知有所差距，都會使檢索結果發生誤差。此外，由於受限於系統的規格，當使用者將其資訊需求正式地表達出來後，還需要轉換爲系統可以接受的格式（妥協的需求）或是檢索策略 (search strategy)，在這段過程中，若是系統所能接受的輸入無法確切地描述使用者的正式需求，同樣也會使得系統輸出的結果無法被使用者接受。

從以上的觀點可以整理出表現資訊需求的三種形態，一是存在於人心中的資訊需求，二是以人類語言（或文字）表達的資訊需求，三是符合系統規範的資訊需求。第一種形態由於難以觀察、測量，因此資訊檢索系統領域的研究應該以第二種形態的資訊需求爲研究的起點。

㈡相關 (relevance) 的意義

相關判斷一直都是資訊檢索效能評估的基礎。在資訊檢索的研究中，

「相關」指的是文件與查詢之間的關連性 (relatedness)。檢出的文件經過評判以確定該文件與查詢是否相關，據此計算出檢索效能。依循著這種方式，研究者可以比較各種不同的檢索策略或是不同的檢索系統之間的優劣。由此可知，「相關」是資訊檢索中極爲重要的研究課題。

Schamber 等 (1990) 認爲提出一個清楚、周全的「相關」定義，對於研究上的意義如下：（註29）

1. 相關是所有資訊系統（包括全文、多媒體、問答、資料庫、知識庫等）檢索效能評量的準則之一。隨著各種新興資訊系統開發問世，系統的複雜度有增無減，而這些系統在檢索效能上的衡量都必須要經過以人爲最終價值依歸的相關判斷。
2. 資訊檢索系統較新的發展是在檢索過程中納入使用者的相關判斷做爲相關回饋 (relevance feedback)。在這樣的系統中，使用者也是系統的一部份，於是「相關」不再是回應刺激做出反應的概念 (reactive concept)，而成爲一種主動的概念 (active concept)，對於系統本身的運作有重大的影響，若是無法瞭解相關對使用者的意義，系統設計根本無從依循。
3. 資訊科學領域必須要在理論上、實證上建立完整的相關定義才能推進整個領域朝其他的研究主題繼續探尋。

根據 Saracevic(1975) 的研究指稱 S. C. Bradford 是最早在資訊科學領域使用「相關」一詞的人（約在1930 年代），之後隨著研究的發展，學者紛紛提出不同觀點的「相關」理論，並試圖給予「相關」合適的定義。（註 30）經過 Swanson(1986) 的整理，將各種觀點的「相關」理論區分成客觀相關以及主觀相關兩類。（註31）

對於「相關」的理論與定義，儘管有許多學者提出其見解，但截至目前仍是衆說紛云，未有定論。綜括各種「相關」的理論與定義，包括主題相關、邏輯相關、證據相關 (evidential relevance)、情境相關、以及心理相關 (psychological relevance)。（註32）就資訊檢索研究而言，主觀相關的理論尚停留在初期理論建構的階段，更遑論由自動化系統來處理主觀相關的問題。因此，現階段仍以處理客觀相關爲主，而其中又以主題相關最常被引用。

本文參考 Cuadra 與 Katter 對主題相關的定義：「相關是資訊需求陳述與文章兩者之間在內容上的一致性，亦即文章所涵蓋的內容對資訊需求陳述的適合程度。」（註33）

主題相關考慮的是主題之間的關聯性，當文件內容所包含的主題與資訊需求的主題之間有相當程度的重疊時就可以視爲相關，這是一種系統導向 (system oriented) 的相關。也就是說，不同的人對檢索系統輸入相同的檢索問題應該會得到相同的結果。一直以來，圖書館對文件分類、製作索引、分類號，或是檢索系統以詞彙

形式代表文件特徵來檢索文件，所處理的相關問題都是在於主題相關的範圍，而這也是截至目前爲止最清楚、最爲人所接受的相關理論（註34）。

更進一步來看客觀相關與主觀相關兩種觀點，雖然兩種相關判斷都仍然以人爲判斷相關的最終標準，但採用客觀相關無疑地排除了人與人之間的差異，而保留了其中的共通點，因而較適合做爲現行的資訊檢索系統的效能評量。至於主觀相關，其論點的價值在於將每個人的資訊需求與個人的特點結合，強調系統應該滿足各人不同的資訊需求，這將是未來個人化資訊檢索或是個人化資訊服務的重點，需要在心理、認知、知識分類等各方面研究範疇的整合。

肆、西文全文資訊檢索研究的發展現況

目前主要的資訊檢索系統與理論的主要模型包括有布林模型 (Boolean Model) 、向量空間模型 (VSM) 、機率模型、擴充布林模型 (Extended Boolean Model) 等。（註35）這些模型起碼早在三十年前就已陸續被提出，但探討這些模型性質的研究仍持續在進行。當代主要的資訊檢索研究期刊包括：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) 以及新近以大型語料與檢索效能評量爲主的 TREC(Text Retrieval Evaluate Conference)，都可以見到這些研究模型的蹤跡。

以下將分別從回顧 TREC 的特色、歷屆研究議題與趨勢，以及全文資訊檢索研究平臺的發展。

一、TREC 的特色

回顧過去一至五屆 TREC 的會議論文，參與 TREC 的研究仍舊是以上述研究模型爲基礎。（註36）由於 TREC 是歷來規模最大、參加者最多的資訊檢索實驗，相較於以往的研究，TREC 的特色在此特別做一簡介：

1. 數量龐大（不管是辭彙或是文件數量）
2. 所包含的文件多爲全文，而非摘要
3. 文件來自多個學科領域
4. 查詢句設計較長且較具結構性
5. 查詢句與文件的相關性有較嚴格的標準，增加相關判定的一致性
6. 加入資訊分送實驗設計 (routing experiment)
7. 加入多種語言的語料

在 DARPA (the Defense Advanced Research Projects Agency) 與 NIST(National Institute of Standards and Technology) 的贊助下，第一屆 TREC 在 1992 年 11 月於 Gaithersburg, Maryland 舉行。其主要目的在於提供一個共同的資訊檢索實驗測試環境，讓研究者能相互討論

比較研究成果。

TREC 的實驗環境主要有二部份，一是文件，二是查詢需求主題(topic)。由於 TREC 的目的在於提供一致的實驗環境，必須考量到各檢索系統對輸入的需求不同，因此不必針對特定的檢索系統制定查詢句，只須要對查詢需求說明清楚，則各檢索系統的發展者可以依此查詢需求主題描述用於實驗設計中。

TREC 中所使用的查詢資訊需求描述爲 topic。TREC-1 與 TREC-2 採用了較複雜的結構來描述 topic，包含了數個區段(field)以及與該 topic 相關的概念(concept)，因此每個 topic 的長度都較長。到了 TREC-3 的 ad hoc topic，不但長度變短，而且所採用的文件結構也簡化許多，但參與者卻仍覺得這樣的 topic 仍然較使用者一般在使用檢索系統時所提供的查詢需求長。因此 TREC-4 就用了更短的 topic，其中只有一個區段，以一句話來描述使用者的需求。不過這樣的改變卻造成了系統處理上的麻煩，於是在 TREC-5 則又加入 title 與 narrative 兩個欄位(field)。

在 TREC-3、4、5 三屆，各個 ad hoc topic 的制定與其對應的相關文件判別都是由同一個人來負責，在 TREC-5，使用 pre-search 以取得較一致的 topic。pre-search 的程序是先使用 NIST's ZPRISE 系統來擷取可能與 topic 相關的文件，再由制定該查詢句的人針對這份文件集合來進行相關判定。最初約有 150 個 topic，在 pre-search 後，去除掉有太多相關文件或是相關文件太少的 topic 而得到 50 個 topic。在相關判定方面，由於 TREC 的文件數量極多，若是要一一由人工判別相關，非但不容易，甚至是一件不可能的事。因此 TREC 採用了聚合方法(pooling method)，由所有與會的檢索系統對各個 topic 所檢出的文件中各取前 N 份文件，混合在一起，再進行相關判定。

表一 TREC 各屆 track 列表

	TREC-1	TREC-2	TREC-3	TREC-4	TREC-5	TREC-6
Interactive			•	•	•	
Database Merging			•	•	•	
Multilingual			•	•	•	
Confusion				•	•	
Filtering				•	•	
NLP				•	•	
Very Large Corpus						•
Cross Language						•
Efficiency			•			

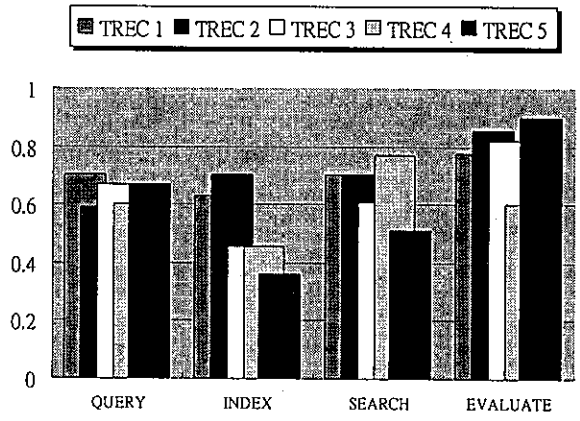
二、歷屆研究議題與趨勢

各屆 TREC 尚包含數個不同的主題，稱爲 track 。除了 routing 與 ad hoc 兩個主要 task 外，表一（見 15 頁）整理各屆 TREC 中所包含的主題。（註37）

根據 TREC 發起之主要目的，並參考圖三所提研究架構，本文從資訊檢索系統與理論範疇、資訊檢索研究基礎環境範疇確認以下四個主要議題：索引、查詢、檢索與評量。

所謂索引是指分析文件內容、決定文件特徵，並且將文件以特徵形式代表的整個過程；查詢則包括使用者介面、查詢句設計、 相關回饋、以及查詢句擴張；檢索強調查詢與索引之間的比對，如何更快速、更準確。包括相關文件排序以及相似性 (similarity) 計算等；評量則針對系統內部效能，以及兩種或多種的模組（方法）的實驗，比較其檢索效能上的差異。

經由分析五屆 TREC 會議論文集的 175 篇論文，扣除其中不適合的論文 14 篇（包括無法取得的、主辦單位的總評與說明、以及沒有結果的），共有 161 篇論文。依據以上四個議題的定義，找出各篇論文的研究重點（採非互斥分類），依各屆比例繪製圖表如下。



圖五、 TREC 各屆研究重點（註38）

由上圖可以看出， TREC 所提供的測試與評量的平臺，讓研究者能測試大規模語料實驗環境下的檢索理論以及檢索系統效能，甚至更進一步地找出較爲適合的系統參數以及檢索方法。平均各屆有接近八成比例的研究進行系統內部的檢索效能評量。此外，由圖五大致可看出將研究重點放在索引的比例逐漸降低，原因可能是目前的索引模型已達到某種極限，很難再加以突破；亦或是已經得到滿意的解答。

相對而言，在評量與查詢的研究不但數量較多，也持續成長。此一現象正說明了過去先進檢索模型在小規模語料實驗室研究環境下的缺失，當系統面對實際應用複雜且大規模的環境，實驗室系統可能仍有許多改進之處。最後值得一提的就是使用者資訊搜尋行爲範疇的研究在 TREC 中仍未受重視，雖然第三、四、五屆都有互動檢索的實驗主題，但是跨學域整合研究仍不多見。

三、全文資訊檢索實驗平臺的發展

從前述有關研究架構的兩個基礎理論的討論可知，使用者資訊需求與相關判斷是實驗設計中最難處理的部份。因此爲了科學研究效度上的考量，自 1960 年的 Cranfield I 實驗之後，在全文資訊檢索研究中開發了許多測試語料 (corpus ；或稱實驗平臺， test collection) ，這些實驗平臺的開發都需要投注大量資源，以下匯

表二　西文資訊檢索常用的語料列表

測試語料	文件數量	查詢句數量	平均相關文件數
TIME	425	83	3.9
KEEN	800	63	14.9
MED	1,033	30	23.2
CF	1,239	100	31.9
CRAN	1,400	225	8.2
CISI	1,460	76	41
HARDING	2,472	65	22.6
EVANS	2,542	39	23.1
CACM	3,204	52	15.3
LISA	6,004	35	10.8
NPL	11,429	93	22.4
INSPEC	12,684	77	33
UKCIS	27,361	182	58.9
TREC-3-WSJ	173,252	50	78.3
TREC-2-WSJ	173,256	50	91.1
TREC-3-full	741,856	50	196.1
TREC-2-full	742,611	50	232.9

資料來源：修訂自 W. M. Shaw Jr, Robert Burgin, and Patrick Howell, "Performance tandards and Evaluations in IR test Collections: Vector-Space and Other Retrieval Models," Information Processing & Management 33(1997): 19.

總過去實驗所使用的實驗平臺的統計資料，並以膀胱纖維化病變全文資料庫（見表二(17 頁)CF 測試語料）的發展歷程做一簡介，一窺實驗平臺開發的困難。（註39）

膀胱纖維化病變全文資料庫(The Cystic Fibrosis Database)的全文文件取自美國國家醫學圖書館(National Library of Medicine; NLM)的 MEDLINE 資料庫，收錄 MEDLINE 資料庫中出版於 1974 年到 1979 年間並且以膀胱纖維化病變(Cystic Fibrosis, 以下簡稱 CF)爲索引詞的文件，共計 1239 篇，其中大部份是學術文章(scholarly article)，有部份則爲書籍的章節或是寫給期刊編輯反應意見的信札。絕大部份所收錄的文件都有十一個欄位記錄其相關內容（包括文件編號、作者、標題、來源、副主題(minor subject)、摘要、參考文獻列表等），但有少數文件例外。

查詢句則是先由一位 CF 專家創造出 100 個查詢句，再由數位相同領域的專家評閱後修改了其中一小部份。匯總之後將這些查詢句以較廣的主題類別(broad subject category)分成 12 類。

在相關判斷方面，共有 14 位 CF 專家參與相關判斷，該語料庫可說是到目前爲止在相關評量實驗控制上最爲詳盡的語料庫。原查詢句的設計者分別就 100 個查詢句對 1239 篇文件判斷文件與查詢句之間的相關；9 位專家各別依其專長領域針對查詢句對所有的文件進行相關判斷；4 位博士後研究員依其專長就 100 個查詢句對部份的文件進行相關判斷；另外一位具有豐富線上檢索經驗的醫藥目錄學者(medical bibliographer)則就所有查詢句對所有的文件進行相關判斷。（註40）

伍、中文全文資訊檢索研究的關鍵議題

語言是人類進行思考與交流的工具，文字則是記錄語言的工具。文字的最初形式都是象形的、表意的，直接表達人類對於大自然的認知。但是隨著文字進一步的抽象化，文字形式會隨著語言的分化而分化，文字形式會適應各自語言的特點而發展（註41）。所以，中文（或稱漢字）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。而中文語文的特性，也正是中文全文資訊檢索研究的挑戰與機會。

從資訊檢索的角度而言，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（註42）。因此，由於語文本身的差異，索引與檢索的形式自然不同。

特別是在詞彙部份，中文與英文有著類似的單位，但是構成的形式與內容結構則是完全不同。最明顯的差

別是在自然語言的處理中，英文詞彙的界線很容易區別，而中文則因爲字與字相連而不容易確認詞彙的界線。其他的不同如：英文的詞彙具有詞類的標示，中文則無；英文有詞幹(stem)的特殊結構，中文則無；中文有複雜的構詞規則，英文則無。

相較於上述西文的研究發展，中文的全文資訊檢索研究起步較晚，加上中文的特性，中文全文資訊檢索領域中仍有不少問題有待解決，其中關鍵的問題包括：標準的測試平臺、網路資源檢索工具、關鍵詞索引、文件自動分類、資訊過濾以及跨學域整合研究與互動等。（註43）這些問題也正是前述研究架構討論不同研究範疇的重點。

回顧中文資訊處理與資訊檢索相關聯的研究，在中文資訊處理基礎研究有關中文自然語文處理部份，以及中文全文資訊檢索應用研究的中文全文檢索系統發展上，已經有十餘年歷史。

例如中央研究院資訊科學研究所的中文詞知識庫小組，從民國 75 年開始結合計算機與語言學的中文詞知識庫計畫。目前的研究現況與應用發展方向以中文詞知識庫爲核心，發展中文語句分析、語音辨識、資訊檢索及語言學研究語料庫等。（註44）但是就解決全文資訊檢索的問題而言仍然很少。但是在長期的基礎研究上所得到的成果，已經可以在中文全文資訊檢索研究上支援許多先進的研究。（註45）

另外，在中文全文資訊檢索應用研究上，中研院於 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（註46）。這套中文全文檢索系統 (CTP/FTMS) 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。因此，目前在中文全文資訊檢索研究上可以藉助於上述系統，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。

現階段中文全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。如果中文全文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。過去三年作者嘗試從基礎研究的途徑建立標準測試平臺。目前可用的全文語料庫以兒童日報新聞報導資料爲主，從82年 1 月起至82年12月止，包括文教、兒童福利、醫藥、環保及專欄等類別，建置在中央研究院計算機中心的全文檢索系統上。表三列舉全文語料的基本性質，以及人工選詞與自動斷詞結果在詞的組成上的差異。

表三　兒童日報新聞全文語料基本性質

新聞類別	文件數	總字數	每篇平均字數	人工選詞詞數	自動斷詞詞數	自動斷詞詞類篩選	每篇平均詞數（人工）	每篇平均詞數（自動）	自動斷詞／人工選詞比
文教	1070	391474	366	6698	17356	9871	6	16	3
兒福	350	126845	362	2730	8225	4462	8	24	3
醫藥	502	179450	357	2562	10300	5732	5	21	4
環保	368	141247	384	2544	9959	6033	8	27	3
專欄	393	314876	801	4915	19346	10866	13	49	4

資料來源：（註47）

基於研究的客觀性、實驗設計與研究資源的考量，上述語料整理參考中央通訊社新聞全文資料庫的分類索引典，選擇文件性質較接近的環保及醫藥新聞爲主。根據這兩類實驗語料的特性，修訂、建立本實驗在環保及醫藥語料的分類索引典。另外由於專欄文件是根據各專欄主題分類，未參照任何分類架構。

有關查詢句設計與相關判斷的進行方式，在上述醫療、環保與專欄三類語料中，分別由三位台大圖書資訊學系四年級同學依據互斥分類結果，先進行各小類文件的主題分析，然後設計查詢主題與相關判斷，初步整理結果如表四所列。相對於前述膀胱纖維化病變全文資料庫的發展過程，表四所列的測試平臺只是初具雛形，仍須進行更嚴謹的實驗設計與效度驗證。

表四　兒童日報新聞全文語料測試平臺描述

測試語料類別	文件數量	查詢句數量	平均相關文件數
醫藥	502	32	7.1
環保	368	22	3.9
專欄	393	27	2.9

陸、結論與建議

本文圖三所描繪的研究架構只是一個雛形，尚未提出完整而明確的變數定義，因此也缺少更嚴謹的實證研究，以便把過去相關論文做一個分類，以驗證研究架構的完整性，同時可以檢視過去研究的重點與缺失，或發現未來研究趨勢。但是就已列舉的研究範疇而言，也已經可以提供研究者思考研究問題的概念雛形。本文期盼未來有更多整合不同研究範疇的計畫發展，從根本的研究環境做起，以期建立中文全文資訊檢索的理論典範。

註釋

註 1 ：R. Hull, & R. King, "Semantic Database Modeling: Survey, Application, and Research Issues," ACM Computing Surveys 19:3 (Sep. 1987): 201-260.

註 2 ：W. Saffady, Text Storage and Retrieval Systems: A Technology Survey and Product Directory (Meckler Corporation, 1989), 4.

註 3 ：同前註。

註 4 ：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」，當前語文問題學術研討會，（台北：行政院國家科學委員會，國立台灣大學中國文學系，民國 83 年），附錄。

註 5 ：C. J. Van Rijsbergen, "Information Retrieval: New Directions: Old Solutions," Proceedings of The ACM SIGIR(1983), 264.

註 6 ：G. Salton, "Another Look at Automatic Text-Retrieval Systems," Communications of the ACM 9:7 (July 1986): 651.

註 7 ：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），頁 7。

註 8 ：鍾聖校，認知心理學，初版四刷（台北：心理出版社，民國 82 年），頁 1-2。

註 9 ：同前註，頁 139。

註 10 ：同註 8，頁 139-150。

註 11 ：黃蕙株，「索引典的基礎理論」，索引典理論與實務研討會論文集，（台北：美國資訊科學學會台北分會，農業科學資料服務中心，國立中央圖書館，民國 83 年），頁 20。

註 12 ：同前註，頁 20-25。

註 13 ：同註 8，頁 165-181。

註 14 ：同註 11，頁 26-27。

註 15 ：方師鐸，國語詞彙學構詞篇，（益智書局，民國 59 年），頁 19、頁 29。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)，詞素雖是最小的、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。

註 16 ：同註 11，頁 32。

註 17 ：B. Ives, S. Hamilton, and G. B. Davis, "A Framework for Research in Computer-Based Management Information Systems," Management Science 26:9(1980): 910-934.

註 18 ：D. A. Kemp, "Relevance, Pertinence, and Information System Development," Information Storage & Retrieval 10:2(1974): 38.

註 19 ：同註 17。

註 20 ：黃慕萱，「檢索系統評估之方

法--理論與實務」，中國圖書館學會會報五十九期，（民國86年），頁115。

註21：A. E. Fox, "Some Considerations for Implementing the SMART Information Retrieval System Under UNIX" Technical Report 83-560, (Cornell University Department of Computer Science, Ithaca, New York, Sep. 1983.) <http://cs-tr.cs.cornell.edu/> (26 Nov. 1996).

註22：G. Salton, The SMART Retrieval System, Experiments in Automatic Document Processing (Prentice Hall, Inc., Englewood Cliffs, N. J., 1971).

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

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

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

R. Wilkinson, & P. Hingston, "Using The Cosine Measure in A Neural Network for Document Retrieval," in Proceedings of ACM-SIGIR(1991): 202-210.

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

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

Lang, Sheau-Dong（郎小棟），"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，（新竹：國立清華大學，Aug. 1996).

註23：吳美美，「言談分析和資訊檢索互動研究」，教育資料與圖書館學三十卷四期，（民國82年9月），頁340-350。

吳美美，「試論資訊檢索理論」，在當代圖書館事業論集：慶祝王振鵠教授七秩榮慶論文集，初版（臺北市：正中，民國83年），頁731-751。

吳美美，「中文檢索詞彙初探」，21世紀資訊科學與技術的展望學術研討會論文集，（桃園縣：世界新聞傳播學院圖書資訊學系，民國87年），頁167-193。

林珊如，「資訊行爲研究方法之探討」，21世紀資訊科學與技術的展望國際學術研討會論

文集，（台北市：世界新聞傳播學院圖書資訊學系，國家圖書館，民國85年），頁377-395。

黃慕萱，「終端使用者之線上資訊尋求行為分析」，21世紀資訊科學與技術的展望國際學術研討會論文集，（台北市：世界新聞傳播學院圖書資訊學系，國家圖書館，民國85年），頁351-376。

註24：簡立峰，「尋易系統(Csmart)與中文智慧型資訊檢索」，21世紀資訊科學與技術的展望國際學術研討會論文集，（台北市：世界新聞傳播學院圖書資訊學系，國家圖書館，民國85年），頁299-314。

張俊盛，柯淑津，陳惠群，「跨語言資訊檢索的網路資源探勘」，21世紀資訊科學與技術的展望學術研討會論文集，（桃園縣：世界新聞傳播學院圖書資訊學系，民國87年），頁153-166。

曾元顯，「多媒體資訊檢索技術之探討」，21世紀資訊科學與技術的展望國際學術研討會論文集，（台北市：世界新聞傳播學院圖書資訊學系，國家圖書館，民國85年），頁281-298。

Chen, Hsin-His, "Cross-Language Information Retrieval", 電子詞典、機器翻譯與資訊擷取研討會，（台北市：中華民國計算語言學會，民國86年）。

註25：陳光華，「電子文件主題之自動辨識」，中國圖書館學會會報五十九期，（民國86年），頁43-58。

曾元顯，「關鍵詞自動擷取技術與相關回饋」，中國圖書館學會會報五十九期，（民國86年），頁59-64。

註26：卜小蝶，「網路資訊過濾技術與個人化資訊服務」，21世紀資訊科學與技術的展望國際學術研討會論文集，（台北市：世界新聞傳播學院圖書資訊學系，國家圖書館，民國85年），頁339-350。

卜小蝶，「提供個人化服務的線上公用目錄檢索系統初探」，中國圖書館學會會報五十九期，（民國86年），頁127-134。

陳光華，「新資訊時代的啓發性資訊服務」，21世紀資訊科學與技術的展望學術研討會論文集，（桃園縣：世界新聞傳播學院圖書資訊學系，民國87年），頁195-208。

林頌堅，「自動化文件分類在資訊服務上的應用」，21世紀資訊科學與技術的展望學術研討會論文集，（桃園縣：世界新聞傳播學院圖書資訊學系，民國87年），頁255-275。

註27：W. S. Cooper, "A Definition of

Relevance for Information Retrieval," Information Storage & Retrieval 7(1971): 21

註 28 : R. S. Taylor, "Question Negotiation and Information Seeking in Libraries," College and Research Libraries (1968): 182-183.

註 29 : L. Schamber, M. B. Eisenberg, & M. S. Nilan, "A Re-examination of Relevance: Toward a Dynamic, Situational Definition," Information Processing & Management 26(1990): 756.

註 30 : T. Saracevic, "Relevance: a Review of and a Framework for the Thinking on the Notion in Information Science," Journal of the American Society for Information Science 26(1975): 321-343.

註 31 : 黃雪玲，資訊需求者與次判斷者相關判斷之比較研究，（碩士論文，國立臺灣大學圖書館學研究所，民國 84 年），頁 18。

註 32 : 黃慕萱，資訊檢索中「相關」概念之研究，初版（台北市：臺灣學生書局，民國 85 年），頁 62-71，頁 214-215。
呂春嬌，「相關概念在資訊檢索中之發展與趨勢」，圖書與資訊學刊十六期，（民國 85 年 2 月），頁 21-32。

註 33 : 同註 27，頁 20。「Relevance is the correspondence in context between an [information] requirement statement and an article, i.e. the extent to which the article covers material that is appropriate to the requirement statement」。

註 34 : 同註 32，頁 215。

註 35 : D. M. Everett, & S. C Cater, "Topology of Document Retrieval Systems," Journal of The American Society For Information Science 43:10(1992): 659.

註 36 : D. Harman, "Overview of the First Text Retrieval Conference (TREC-1)," Proceedings of The First Text Retrieval Conference (TREC-1), ed. D. K. Harman (NIST Special Publication 500-207, March 1993) : 1-20.
D. Harman, "Overview of the Second Text Retrieval Conference (TREC-2)," Proceedings of The Second Text Retrieval Conference (TREC-2), ed. D. K. Harman(NIST Special Publication 500-215, March 1994) : 1-20.
D. Harman, "Overview of the Third Text Retrieval Conference (TREC-3)," Proceedings of The Third Text Retrieval Conference (TREC-3), ed. D. K. Harman (NIST Special Publication 500-225, April 1995) : 1-20.
D. Harman, "Overview of the Fourth Text Retrieval Conference (TREC-4)," Proceedings of The

Fourth Text Retrieval Conference (TREC-4), ed. D. K. Harman (NIST Special Publication 500-236, Oct. 1996) : 1-24.

E. Voorhees, & D. Harman, "Overview of the Fifth Text Retrieval Conference (TREC-5)," Proceedings of The Fifth Text Retrieval Conference (TREC-5), 1997, http://trec.nist.gov/pubs/trec5/t5_proceedings.html(1 Mar. 1998).

註 37 ：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集，（高雄市：國立海軍軍官學校，民國 87 年），頁 1.21-1-1.21-6。1. Ad hoc：針對某一查詢需求，找出所有相關的文件，並依照相關程度排序。2. Routing：某一查詢需求，並已知部份的相關文件，以這些相關文件爲訓練資料 (training data) 訓練檢索系統，由檢索系統從測試資料 (test data) 中找出相關文件。與剪報性質類似。3. Filter：與 routing 類似，有訓練資料與測試資料。針對某一查詢需求，找出所有相關的文件，但並不加以排序。4. Corruption：在資料有誤的情況下（例如剛經過光學字元辨識的文件資料），仍能找出正確相關的文件。5. Fusion：針對某一查詢需求，將語料分成數個子集合，或是分數次以不同的查詢句檢索，亦或是由數個不同的系統檢索，再將檢索結果融合。6. Multilingual：多語文的資訊檢索。自第三屆起加入西班牙文語料；第五屆加入中文語料。7. Interaction：探討在使用者與系統互動而非批次的環境下如何增進檢索效能。

註 38 ：同前註。

註 39 ：W. M. Shaw Jr, Robert Burgin, and Patrick Howell, "Performance Standards and Evaluations in IR test Collections: Vector-Space and Other Retrieval Models," Information Processing & Management 33(1997): 19.

註 40 ：同前註，頁 347-366。

註 41 ：董琨，漢字發展史話，（台北市：台灣商務印書館，1993 年），頁 12-18。

註 42 ：趙元任，「語言成分裡意義有關的程度問題」，中國現代語文學的開拓與發展：趙元任語言學論文集，袁毓林主編，（北京市：清華大學出版社，1992 年）。

註 43 ：Chien, Lee-Feng, Pu, Hsiao-Tieh, "Important Issues on Chinese Information Retrieval", Computional Linguistics and Chinese Language Processing 1:1(Aug. 1996):207-208.

註44：中文詞知識庫小組，「中文自然語言研究環境之建立」，詞庫・語料庫應用研討會，（台北：中華民國計算語言學學會，中研院資訊科學研究所，1996年）。

註45：Chien, Lee-Feng, "PAT-Tree-Based Keyword Extraction for Chinese Information Retrieval," Proceedings of ACM-SIGIR(1997),:50-58.
Chen, Hsin-His, & Guo-Wei Bian, "Proper Name Extraction From Web Pages for Finding People in Internet," Proceedings of The Research on Computational Linguistics (ROCLING X) International Conference(1997),:143-158.
Guo-Wei Bian, & Chen, Hsin-His, "An MT Meta-Server for Information Retrieval on WWW," Proceedings of The AAAI Spring Symposium on Natural Language Processing for the World Wide Web(USA：1997): 10-16.
簡立峰，「尋易系統(Csmart)與中文智慧型資訊檢索」，21世紀資訊科學與技術的展望國際學術研討會論文集，（台北市：世界新聞傳播學院圖書資訊學系，國家圖書館，民國85年），頁299-314。
楊允言，文件自動分類及其相似性排序，（碩士論文，國立清華大學資訊科學研究所，民國82年）。

註46：謝清俊、林晰，中央研究院古籍全文資料庫的發展概要，（台北：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），頁2-3。

註47：同註7，頁158。