大學圖書館
8 卷 2 期（民國 93 年 9 月）　頁 94-123

相關排序於資訊檢索之發展與探討

Development and Issues Concerning the Relevance Ranking for Information Retrieval

張郁蔚　　**Yu-wei Chang**
國立台灣大學圖書資訊學研究所博士生
Doctoral student, Graduate School of Library and Information Science, National Taiwan University
E-mail : ywchang@archives.gov.tw

【摘　要】

面對網路資源檢索結果往往過多的問題，如何將當中最相關的資訊優先排序在最前面，讓使用者優先看到，以節省其瀏覽所有檢索結果的時間，是目前系統開發者選擇相關排序（Relevance ranking）作爲檢索結果排序的原因。目前大部分的資訊檢索系統係以相似性而非相關性作爲檢索結果排序基礎，但使用者係判斷檢索結果相關與否之唯一裁判者，爲此，相關排序的發展也開始將使用者的相關判斷及資訊使用行爲作爲改善相關排序效能之努力方向，希望能針對使用者的不同需求提供更個人化的資訊服務。

【Abstract】

The retrieval results of a web search are usually very large. Therefore, the information retrieval system usually uses the relevance ranking to ranking retrieval results, provideing users a saving-time way to look through the searching results and obtain the most relevant documents. Currently, most information retrieval systems use similarity instead of relevance to rank the searching results. However, the end-user is the one to determine the relevance of retrieval results, so many researchers consider the user's relevance requirement and information behavior are the way to improve the effectives of relevance ranking. The objective of relevance ranking is to provide personalized information service for users.

關 鍵 詞：相關；排序；相關排序
Keywords: Relevance; Ranking; Relevance ranking

壹、前言

檢索出滿足使用者資訊需求的相關文獻是資訊檢索系統的目的及研究重心，自1950年代末期以來，已陸續有不少學者對「相關」（Relavance）概念進行探討，雖然有關的討論相當多元，但是正如Saracevic所言，雖然不經過說明，大家即能直覺地了解何謂相關，但事實上，截至目前相關仍是一個尚未被清楚了解之概念。一直以來，資訊檢索的相關研究分為二大方向，一是系統導向相關，二是使用者導向相關。不過，1990年代中期以後，開始有一些研究將資訊檢索的相關視為是人機之間的認知互動，且二者之間互動的相關種類非常多元，不僅包含一種相關。（註1）事實上，能眞正判斷檢索結果與檢索問題相關與否，只有資訊需求者本身，不過受到資訊需求者本身知識或認知狀態的差異，及不斷異動的本質，資訊需求者的相關判斷基本上是相當主觀的。因此，為了讓資訊檢索系統的檢索結果能滿足資訊需求者的需求，資訊檢索系統設計者必須了解資訊檢索使用者的檢索行為特性及相關判斷過程，才能使檢索系統在使用者問題與資料之間建立一座溝通良好的橋樑。

九〇年代以後，網路及WWW的風行，促使網路資源快速成長，但網路資源並不同於一般傳統資料庫檢索系統中的資料，其資訊品質是未經控制的，每個人都可以容易地自行上載資料，成為網路資源的提供者，因此如何能快速地在龐大且品質不一的資源中找到相關資訊，實為當今資訊檢索的一大挑戰。由於目前搜尋引擎的檢索結果往往過多，因此如何節省使用者一一瀏覽所有檢索結果的時間，將當中最相關的資訊優先排序在最前面，讓使用者優先看到，是目前系統開發者選擇相關排序（Relevance ranking）作為排序檢索結果的原因。

相關排序的做法目前大部分是以文件和查詢問句之間的相關性數值高低作為排序依據，除此之外，網路資源之超連結結構也開啓新的相關排序概念，並在使用者導向相關概念的發展下，使用者的相關判斷也納入相關排序的重要考量，藉以調整資訊檢索系統的排序結果，讓檢索結果能更接近使用者的相關判斷結果。故針對相關排序的討論，本文將分四大部份，首先說明相關的定義與概念，及相關與使用者、檢索系統的關係，其次說明相關排序的意義及相關排序方式，最後則就相關排序的問題提出討論，提供讀者對相關排序的概括性了解。

貳、相關概念之發展

資訊檢索系統的目的係在找出相關文件（註2），因此打從資訊科學於1940年代被視為一個獨立之領域以來，「相關」就一直是其基本且重要的一個概念。自1945年 Vannevar Bush發表"As We May Think"提出一個非常簡單的系統方法，讓令人困惑的知識

得以具有一些次序以來，相關的定義已有相當大的改變，並在 1960 年代成爲熱門的研究主題。（註3）如從不同研究者曾使用的相關名詞，例如相關（Relevance）、主題相關（Topicality）、效用（Utility）、有用（Usefulness）、適用（Pertinence）、系統相關（System Relevance）、使用者相關（User Relevance）、滿意（Satisfaction）等，可以發現相關的名詞相當不一致，且凸顯出相關概念的模糊性及複雜性，即便經過四十多年的發展，截至目前，相關仍是一個尚未被清楚理解的模糊概念。

分析相關概念的發展，許多學者如Swanson及Saracevic，同Schamber一樣，將各種觀點的相關理論及探討分成主觀或系統導向（System-Oriented）相關及客觀或使用者導向（User-Oriented）相關二類。以 Schamber的分類爲例，系統導向相關主要包括邏輯相關（Logical Relevance）及主題相關（Topical Relevance）；使用者導向相關則包括情境相關（Situational Relevance）及心理相關（Psychological Relevance）。（註4）所謂系統導向相關主要在探討使用者所使用的查詢詞和檢索系統索引詞之間的一致性問題，與情境無關；使用者導向的主觀相關，則是以使用者對系統檢索結果的相關判斷作爲相關高低的依據，重視的是使用者對相關的解讀與心智判斷。

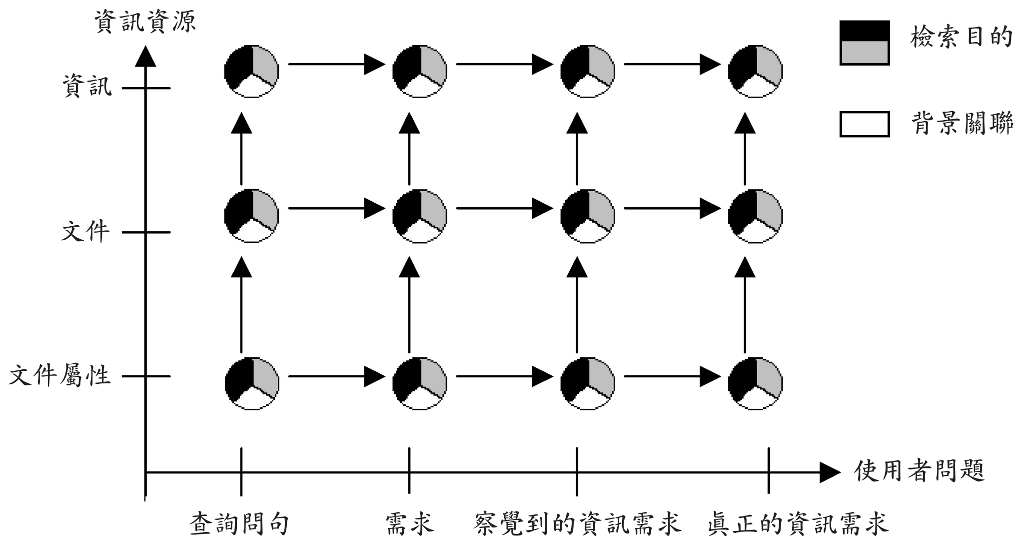
如以相關研究的發展時間來看，早期相關的研究重心在系統導向的主題相關，之後擴展對非主題相關的探討，並由系統導向轉向對使用者導向的關注，認爲資訊需求者本身才是唯一能評估資訊檢索系統能否提供相關檢索結果的關鍵。所謂主題相關，依據 Cuadra 和 Katter 所下的定義爲：「相關係輸入系統的檢索問題和文章內容之間的符合性，亦即文章所涵蓋的內容對資訊條件敘述的合適程度」， 但是 Cooper認爲符合性（Correspondence）和 合 適（Appropriate）均是相當模糊的名詞，對定義相關或釐清相關的爭議不可能有很大的貢獻，故於1971年提出「邏輯相關」的理論，認爲如果文章中有部分主題資訊能滿足資訊需求者的需求，則此文章即被視爲是相關的文章。雖然邏輯相關並不是直接就查詢詞和代表文章內容的索引詞作比對，但只要文章中包含能以邏輯推理得到答案的最小前提組（Minimal Premise Set），則可判斷該文章是否爲相關文章。1973 年，Wilson 提出「情境相關」，在邏輯相關的基礎上，加上歸納邏輯的觀點，並考慮資訊需求者個人之知識狀態及其關心的重點。情境相關中的「情境」是指資訊需求者個人本身的情境，且此情境只有資訊需求者能看到或瞭解，其他人是無法眞正得知的。之後，Harter 根據 Sperber 和 Wilson 之中心思想，於 1992 年提出「心理相關」，認爲相關的資訊就是能改變人類知識狀態的資訊，也就是能產生文字關聯效果（Contextual Effect）的資

訊，因為在一般日常對話中提及相關時都不是指「關於某主題」（On the Topic）之「關於」（About）層次，反倒是一些相互關聯的主題，加強或減弱個人的認知狀態，故心理相關和情境相關二者均是承認相關本身變化的特質，只是心理相關的涵蓋範圍更廣。（註5）

相對於系統導向的相關概念，使用者導向的相關概念認為唯有資訊需求者才能決定檢索結果的相關與否，惟使用者的相關判斷標準會隨著個人不同時間的不同需求或知識狀態的異動產生標準不一致的情形，因此使用者導向相關係屬於主觀標準，非客觀標準。除了上述系統導向相關及使用者導向相關二類外，1990年中期以後一些相關研究開始將資訊檢索視為是人類和電腦之間的認知互動，且當中互動的相關種類相當多元，非僅有一種相關。Saracevic 發展的整合型相關系統模式，指出檢索系統效能的好壞受到相關系統中各種相關表現形式（Relevance Manifestations）之間的相互作用與適應的影響，並進一步說明相關表現形式包括了主題相關、認知（Cognitive）相關或適用、情境相關、動機（Motivational）或情感（Affective）相關等四種不同層次的顯現方式（註6），其中主題相關的主題概念是指「關於」（Aboutness），而非內容；適用則和使用者本身查覺到的資訊需求有關，代表的是介於資訊需求與資訊物件之間使用者的認知狀態；情境相關則指效用或有用，表示在使用者目前認知狀態及所察覺的資訊需求下，資訊和工作任務間的關係；動機或情感相關則屬於目標導向相關，和使用者整體的意圖有關。（註7）

此外，Ingwersen提出的互動式資訊檢索認知模式，指出一個資訊系統包含系統、使用者與環境三部分，其中社會的及組織的環境提供了一個情境架構，直接影響使用者的活動類型。（註8）Mizzaro 則在文件與使用者檢索問題之間加入時間及其他背景因素，以四層面的組合架構，將各種相關理論視為是其架構中之一點。其四個層面分別是：1.資訊資源（Information Resources）：由文件（Document）、文件屬性（Surrogate）（如題名、作者、書目資料、摘要、關鍵字）及資訊（Information）三部分組成；2.使用者問題（Representation of the User's Problem）：視使用者本身狀態，可能包括最原始的真正資訊需求（Real Information Need）、經察覺（Perception）變成察覺到的資訊需求（Perceived Information Need）、之後以人類自然語言所表達出來的需求（Request），以及最後將以自然語言表示的查詢問題轉譯成系統可判讀的查詢問句（Query）；3.時間（Time）：在某個特定時間點，如果使用者改變原來的資訊需求，會致使文件與使用者問題之間可能不存在任何關係，但也許稍後，因使用者的需求問題，文件與使用者問題之間產生則具有相

關關係，因此所謂時間係包括從使用者眞正的資訊需求，一直到使用者滿意爲止的各個可能時間點；4.組成要件（Components）：此層面僅部分具有次序關係，不像資訊資源、使用者問題及時間具有完全的次序關係，但除了時間以外，資訊資源及使用者問題可以再進一步分成使用者感興趣的主題（Topic）、檢索目的（Task）及不屬於主題及檢索目的之其他可能會影響檢索方式及檢索結果的背景關聯因素（Context）等三種，且導致一份文件與使用者問題產生關係的原因，可能是受到第四層面之一個或多個因素影響。從圖一 Mizzaro 的相關種類示意圖，可以看出 Mizzaro 所定義的相關是由四個層面架構而成，亦即「相關＝資訊需求×使用者問題×時間×組成要件」的關係。該架構之應用，以1959年Vickery對主題相關及使用者相關的討論爲例，可分別以 rel（Document, Query, {Topic}）及 rel（Information, RIN, {Topic, Task, Context}）來表示該相關內容的討論。（註9）



圖一　Mizzaro 的相關種類示意圖

資料來源：Stefano Mizzaro, "How Many Relevances in Information Retrieval?," Interacting with Computer 10(3) : 303-320. < http://www.dimi.uniud.it/izzaro > (15 Dec. 2003).

參、使用者與檢索系統之相關判斷

從系統導向、使用者導向及認知模式導向的相關發展過程，使用者相關判斷成了資訊檢索的重要研究核心，因爲唯有了解整個相關判斷過程，才能提高使用者對資訊檢索系統檢索結果的滿意度。有關使用者相關判斷的調查與研究不勝枚舉，從研究結果可歸納出類似的相關判斷因素，例如1967年，Cuadra和Catter之實證研究找出了三十八種影響相關判斷的變數，計歸納有文章類型、資訊需求陳述、判斷者經驗、背景及態度、判斷條件、表達方式的選擇及相關判斷尺度等五大類因素（註10）；1994年，Barry 以內容分析法找出二十三種計七大類（文件內容、讀者過去的背景經驗、讀者信仰及喜好、和其他資訊資源的關係、文件之的來源品質、文件實體部分、讀者情境）相關判斷因素（註11），並發現其中 "文件內容" 係使用者最主要的相關判斷因素；2002年，Maglaughin和Sonnenwald 發現影響使用者判斷文件屬於相關、部分相關及不相關之因素計有二十九種六大類（作者、摘要、內容、全文、期刊或出版者及個人因素）（註12），當中使用者最主要據以判斷相關的依據也是文件內容，如內容正確性、新穎性、深度、範圍、參考引用、主題等，不過實際在進行相關判斷時，是同時運用多種標準進行相關判斷，而非僅依據單一因素做相關判斷。

從研究使用者相關判斷的因素，可發展出評鑑系統的項目。研究影響使用者相關判斷因素的學者，基本的共同特徵就是對傳統不由使用者擔任評鑑者的做法感到不滿，他們將資訊需求用者視爲是系統檢索品質的最終裁判者，並認爲相關判斷也受到許多非主題相關因素的影響，文件與使用者問題之間的關係並非僅是相關或不相關的二分原則，而是相關程度上的差異。不過因受限主題相關是目前資訊系統僅能處理的相關層次，導致許多人對相關的印象也限縮在主題相關的範圍上，忽略掉非相關因素的探討。針對主題相關，Sormunen等人指出，高度相關的文件具有四項特徵：1.主題會被詳細地討論；2.從不同層面探討主題；3.作者對相同概念會以不同形式文字表達，避免同一文字之的重複使用；4.包含較多符合查詢問句的關鍵字。（註13）

目前檢索系統的運作方式，基本上也主要是以字彙比對爲前提，因爲檢索系統唯一的查詢依據就是使用者輸入的查詢關鍵字，因此檢索結果與使用者使用的關鍵字或關鍵字組成的查詢問句有很直接的關係，如果使用者提供檢索系統愈清楚詳細的線索，檢索系統也愈能清楚使用者的資訊需求，不過依據相關研究指出，使用者很少會使用進階式的查詢問句，例如布林邏輯運算子或片語檢索，使用布

林邏輯運算子的查詢比率僅大約占10%（註 14），尤其對網路使用者而言，使用單一查詢關鍵字是其典型的檢索行爲特徵，且在許多搜尋引擎的檢索畫面上，可以觀察到搜尋引擎試圖以最簡單的查詢設計，讓使用者無須去了解複雜的進階查詢，只需在查詢詞輸入欄的唯一入口提供線索即可，但相對地，面對簡單的查詢畫面，使用者僅能在查詢詞上做變化。雖然主題指引式的搜尋引擎，例如Yahoo、Argus Clearinghouse 等，提供使用者依循指定的主題路徑進行網頁文件的搜尋，解決使用者決定查詢關鍵字的問題，加上檢索引擎已先行將文件進行分類，將可能相關的類別予以限制，導致檢索精確率會較高，不過，因爲是選擇性地選擇某個主題或類別，無法全面性地去蒐尋所有可能相關資訊，致使可能會遺漏掉歸屬在其他類別的相關資料，降低資料的回收率。

相對於使用者進行相關判斷所運用的多元判斷因素，目前資訊檢索系統未具有同人類相關判斷一樣的思考能力，其所憑藉的仍是規則及演算的固定法則。雖然自然語言的研究已有一些成果，但系統僅是就依據的法則予以更複雜化與精細化，離人類的心智判斷仍有一大段距離，且就成本及其他因素考量，目前絕大多數的檢索系統仍是採取統計模式的運作方式，不管何種檢索模式，系統進行檢索的唯一線索係使用者輸入的查詢關鍵字。所以使用者如何將本身的資訊需求，以精確的查詢關鍵字和系統溝通是很重要的起點。由於每個人的學習過程、檢索經驗及認知上的差異，每個人對同一概念的詞彙表達方式並不儘相同，因此將主題概念轉譯成適當的查詢關鍵字，實爲一種非常複雜的心智判斷過程，正如同索引人員在爲文件特徵選擇索引詞時，面臨的也是概念與索引詞的配對問題。即使同一位主題專家，在不同時間或環境下對於同一份文件選擇的索引詞也會不同，更何況是不同索引者的個人差異所導致的索引詞選取不一致情形。因此我們可以想像，對無專業檢索經驗及知識的一般使用者而言，即使系統可以處理自然語言的查詢問句，但如何產生適當的查詢問句才是最困難的考驗。

此外，從使用者不斷嘗試的持續檢索行爲，反映出使用者所使用的查詢關鍵字和文件索引詞之間存在有相當的落差，這落差可能僅是不同文字形式之差異，如同義詞，也可能是意義精確程度上的差距，因此使用者唯一能調整的即是不斷地修正檢索策略，直到檢索結果能達到某種滿意程度。爲了達到個別化資訊服務的理想目標，資訊檢索系統也努力嘗試從使用者興趣檔及檢索紀錄萃取出屬於使用者的個別資訊檢索行爲及資訊需求，以提高使用者對檢索結果之滿意度。因此就相關的發展，目前仍是將研究重心放在使用者上，使用者的資

訊檢索行爲仍是研究者欲了解的重要課題。

肆、相關排序

相關排序係一種試圖依據文獻和使用者需求的相關性，將文獻排列的過程（註15），藉由事先定義好的演算法，電腦可據以自動計算每份文件的相關分數，做爲相關排序的依據，以便將檢索結果中使用者最感興趣的文件排在前面（註16），其最早應用在網際網路上係1990年代初期 Kahles介紹的WAIS。（註17）

目前的搜尋引擎常常會產生幾百筆或幾千筆或更多的檢索結果，且依相關研究指出，使用者平均只看檢索結果清單第一頁的前十至二十筆資料（註18），因此檢索結果的排序方式將會直接影響使用者對資訊檢索系統的滿意度。爲解決檢索結果過多的問題，將檢索結果依據文件的相關性高低進行排序，已是多數搜尋引擎排序功能的一種選擇，甚至作爲排序方式的預設值。此外，相關概念和資訊檢索過程是緊密結合的，因爲從資訊需求者一開始進行資料檢索，即嘗試使用相關的關鍵字進行查詢，並在判斷檢索結果的可用性及精確性後，考慮是否需要修正先前所使用的關鍵字及檢索策略，提出更能符合本身資訊需求的新檢索策略，再次進行檢索。故從檢索前、檢索過程到檢索結果後的相關回饋（Relevance Feedback）及評估過程，相關不僅包括使用者的相關判斷，也包含資訊檢索系統執行相關檢索的運作。惟目前資訊檢索系統僅能比對查詢關鍵字和代表文件內容的索引詞二者之間的一致性，因此使用者本身表達資訊需求的能力很重要，其是否能透過自然語言或布林查詢問句的形式，精確地表達出資訊需求，往往影響了檢索結果的相關性。

傳統的資訊檢索對文件相關與否的假定，係假設一件文件的相關與其他文件是無關的，也就是說每件文件的相關性是獨立的、個別的，文件之間並不存在相互影響的關係，因此系統可以很容易地個別計算每件文件的相關分數，做爲相關排序的基礎。不過實際上，此種相關獨立的假設是不存在的，因爲使用者的相關判斷會受到已看到排序在前面的文件內容的影響，亦即如果使用者已看過足夠多的相關文件，會使排序在後面的檢索結果，即使有高相關的文件，使用者也可能不會繼續瀏覽下去，或是給排序後面的相關文件較低的相關判斷，因此文件之間事實上是存有相互依存的關係，並非是獨立不相關的個體。

一般而言，自動檢索技術是以單一字詞在文件中出現的頻率爲基礎，依照本身的計算方法算出每個文件與每個字詞的相似性做爲相關排序的依據。當自動檢索技術運用在搜尋引擎時，搜尋引擎會根據使用者輸入的查詢關鍵字，找出包含此查詢關鍵字的網頁，並依照本身查詢關鍵字的權重函數（Term Weighting Function）和向

量相似性函數（Vector Similarity Function），計算出每個網頁和此查詢關鍵字的相關性，再將這些網頁根據相關性加以排序後傳回給使用者。（註19）不過在大部分檢索系統以單字、詞或片語做為代表文件特徵的情形下，語彙特徵往往是從原文中所抽取出來的一小片斷，代表的是文件的表面資訊，並無法表現出文件主題隱含的語意，因此檢索結果和使用者的主題需求，可能是存有相當的差距。

資訊檢索模式決定了檢索結果的產生方式，各種檢索模式中，布林檢索模式（Boolean Retrieval Model），是以 AND、OR、NOT 布林邏輯運算子和關鍵字組成檢索字串，其中表示文件特徵的關鍵字可以向量的方式表示，且對於相關是採取相關與不相關二元化的概念，如果文件有包含此查詢關鍵字即為相關文件，反之，如果文件沒有包含查詢關鍵字即屬於不相關文件。此種完全以關鍵字出現與否做為相關與否的判定標準，忽略了介於相關與不相關二者之間灰色地帶的可能性詞彙，因為在人類複雜的語言中，對於相關主題概念的表達有多種可能方式，如同義詞及相關詞，因此就詞彙與主題概念的關係而言，原本就可能存有相關程度上的差異。此外，布林檢索模式對所有詞彙是採取同等重要的觀點，並未考慮對不同使用者而言，每個關鍵字可能有不同的重要性，應針對不同的資訊需求給予不同的權重。因此就系統設計及成本而言，布林檢索模式最簡單，檢索速度最快，但缺點是無法表達較複雜的檢索需求，且也無法將檢索結果進行排序。

向量空間檢索模式（Vector-Space Retrieval Model）係將文件及查詢問句以一組向量來表示，查詢問句與文件之間的相關性，係在計算二個向量之間的相似性，當二個向量之間的夾角愈小，表示二者之間的相似性及相關性愈高，反之，當二個向量之間的夾角愈大，二者之間的相似性及相關性就愈低。此檢索模式常見的相似性計算方式，係計算二個向量之間夾角的COSINE值，COSINE值愈高，代表二者之間的相似性愈大，相關性愈高，除此之外，另有 JACCARD、DICE 等多種計算方式。雖然二個向量之間的相似性有多種計算方式，但其共同特性是當二個向量中相同的屬性增加時，相似性也會增加，且二個向量的內積（Inner Product）增加時，計算所得之相似性也會隨之提升。除了計算二個向量之間的夾角大小外，也可以計算二個向量之間的距離，亦即當二個向量之間的距離愈短，表示二者之間的相似性及相關性愈高；反之，當二個向量之間的距離愈長，表示二者之間的相似性及相關性愈低。

此外，向量空間檢索模式的一大特色是認為各個關鍵字與文件之間的重要性關係是不同的，一般是以0到1之權重值範圍表示各關鍵字的重要性

差異，而不以武斷的觀點去表示某關鍵字和文件之間絕對不相關，例如0.8或0.9的權重值代表該關鍵字與文件有高度相關，0.1或0.2則相對表示該關鍵字與文件的相關性較低。一般關鍵字的權重數值常是以「詞頻」（Term Frequency）和「逆向文件頻率」（Inverse Document Frequency）的積爲依據，其中詞頻係詞彙在文件中出現的頻率，詞頻愈高表示該詞彙愈具重要性，但是以詞彙表示文件的主題特徵，除了須具備內容代表性外，也須具備可以和其他文件區別的特質。亦即當某一個詞彙包含在資料庫的大部分文件內容時，雖然該詞彙的詞頻很高，但如果不具備可以區別文件的特徵，也不適合做爲代表文件特徵的索引詞；相對地，如果某個詞彙只在文件集的少數文件中出現，則該詞彙反而具有該文件特徵的代表性，這就是逆向文件頻率的概念。逆向文件頻率的定義有多種形式，不過基本上均是以最基本簡單的公式，亦即IDF = log(1/DF)（DF爲文件頻率）作爲不同形式上的小變化。

至於另一種機率檢索模式（Probabilistic Search Model）的出現，是因爲相關是文件、使用者及資訊需求的各種變化的函數，但在不能精確地預測文件與使用者資訊需求的相關性下，改以機率的觀點來描述並運算查詢詞彙與相關文件之間的不確定性。機率檢索模式是以關鍵字出現的機率來看相關，該模式對於關鍵字權重也有不同形式的給定方式，例如log（r/R）/（n/N）、log（r/R）/（n-r/N-R）、log（r/R-r）/（n/N-n）、log（r/R-r）/（n-r/N-n-R+r）等，其中N是文件總數，R是與查詢Q有關的文件總數，n是有出現關鍵詞T的文件數，r是與查詢Q有關並出現關鍵詞T的文件數。（註20）在比較上，機率檢索模式可以做到空間向量模式的效果，惟二者不同的地方在基本的假設與運算的模式。（註21）另各種機率檢索模式的一個共同特點就是機率排序原則（Probability Ranking Principle），其強調當被檢索出來的文件依其和檢索問句之間的機率去排序時，可以達到最佳的檢索效果（Optimal Retrieval Performance）。（註22）簡言之，向量空間檢索模式及機率檢索模式均是屬於較簡單的統計檢索模式，不像自然語言檢索模式涉及人類複雜的語言結構，因此目前大多數的檢索模式仍是以統計方式作爲運作基礎。

目前一般資訊檢索系統係以相似性，而非相關性作爲比對及排序的基礎，因此與理想上的相關概念仍有相當的差距，不過，針對傳統相似性的做法，許多研究及搜尋引擎業者已另考量其它排序方式的可行性，讓相似性不再是檢索結果的唯一相關排序依據，因此相關排序在未來資訊檢索上的發展，仍有相當大的改進空間。以下則就目前所謂網頁的相關排序方式舉例說明，提供了解目前相關排序的發展現況。

伍、相關排序方式

目前大部分資訊檢索系統的檢索結果相關排序，係先比對出符合使用者所輸入的關鍵字文件，再計算關鍵字與每件文件之間的相似性數值，做爲相關性排序的基礎。文件相似性排序的概念是認爲如果我們已握有一篇相關文件，我們可以在文件集中，利用相似性的計算方式找出和該文件相似性達到某一門檻值以上的其他文件，因此相關排序的基本前提是，據以排序的檢索結果應是整個文件集中與查詢問句最相關的某個範圍的文件集合，如果檢索結果和查詢主題不符，即使相關排序的相關正確性很高，對使用者而言，仍是不具參考價值的資訊。

以檢索結果數目的多寡來看，當檢索到的文件數量很少，系統自然可以很容易地進行相關排序之處理，同時使用者也可以容易地看出相關排序結果的正確性，但如果使用者想要對某一特定主題的文件進行全面性的瀏覽，勢必需要相當的檢索結果數量，因此在此種情形下，系統的相關排序負荷不但較大，使用者也無法立即判斷大量文件的相關排序正確性，不過相對地，如果相關排序能就衆多的檢索結果予以符合使用者相關判斷的處理，則能幫助使用者從排序在最前面的核心文件開始瀏覽，有層次地了解檢索結果中各文件之間的相互關係，節省瀏覽及判斷所花費的時間與精力。不過，傳統的資料庫檢索及排序演算方式並不容易應用在網路環境，當中的主要問題在於熱門搜尋引擎之使用者數目過大，以及搜尋引擎可檢索及需排序的文件數量過多，因爲搜尋引擎無法事先預測某個時間點使用某個搜尋引擎的使用者數量，加上網路資源的數量及差異性均甚大，無法像資料量較少的一般資料庫，可以仔細地組織、儲存及索引每筆資料，方便使用者以定義好的查詢問句快速地檢索到相關資料。（註23）以下茲就主要的相關排序方法說明如下：

一、關鍵字出現次數及所在位置

比對使用者輸入的查詢關鍵字是否出現在文件內容中，或者比對關鍵字與代表文件特徵的索引詞之間的一致性，是最簡單、最普遍及最基本的相關排序方式，亦即檢索結果的所有文件內容一定會包含使用者輸入的查詢關鍵字，同時一般會再配合詞頻、逆向文件頻率及關鍵字出現位置等其他因素進行相關排序的調整，例如：行政院研究暨發展考核委員「我的 e 政府」網站（http://www.gov.tw）的搜尋結果係以輸入的關鍵字進行比對，其中最先列出的是和輸入的關鍵字完全一樣的類目和網站名稱，接下來會列出類目或網站名稱中有符合關鍵字的搜尋結果，最後如果網站的敘述內容有符合關鍵字者，也會將這些結果列出。

其次，關鍵字出現在文件的不同位置，所代表的意義是不同的，例如關鍵字出現在題名或摘要的重要性比出現在內容的正文高，而在正文中，關鍵字出現在前言與結論的重要性也是不同的，同樣地，在網路環境中，雖然網頁文件和一般文件有標籤（Tag）結構的差別，但是同一般文件一樣，一般認為關鍵字出現在<Title>標籤比出現在其他標籤位置重要。Culter 曾在1997年利用網頁文件的架構與鏈結，改進網頁文件中資訊擷取的效率，其將網頁標籤分成六大類，並分別給予不同的最佳權重，因此出現在不同類標籤的關鍵字，在網頁文件上所代表的重要性是不同的，同時 Culter 的實驗結果也顯示，利用最佳權重的網頁文件資訊擷取方法，比未採用最佳權重的方法，足足提高了約百分之四十四的精確率。（註24）

二、網頁之連結架構

利用網頁文件之間的連結關係所發展的相關排序演算法係源自引用（Citation）概念，其假定網頁本身與有連結關係的網頁之間存在某種程度的相關性，亦即被引用的資料一定對引用者有相當程度的影響與關係，因此被引用次數的多寡可反映出該網頁的重要性及相當的內容品質。以被連結的網頁而言，外來的連結網頁如為重要的或權威的網頁，則相對可提高本身網頁的重要性，因為對該外來的重要網頁而言，本身網頁因被視為是一個重要的網頁才會產生連結關係，而重要的網頁通常也會有較多的外來連結，因此此種相關概念的推導係屬於一種相互推薦產生的結果。

對需要大量且快速搜尋網路資源的搜尋引擎而言，如何將檢索結果中最相關的網頁優先排序在前面的相關排序功能係目前逐漸普遍的排序選擇，從表一可以看出許多搜尋引擎的排序選項已包含相關排序。此外，目前網頁文件的相關排序方式，基本上是在原有的詞頻、逆向文件頻率及關鍵字位置的傳統檢索模式基礎上，加入網路連結結構的特性，例如資料來源評價、更新頻率、受歡迎程度、檢索速度、權威程度（Degree of Authority）及集中程度（Degree of Hubness）等（註25），發展出個別的相關排序方式。此種利用網頁連結特性所發展的相關排序演算法，可溯自1998年先後推出的PageRank及HITS演算法，同時PageRank及HITS的產生也引發許多研究者依據對其進行相關修正及研究之熱潮，另外，如 Drori 的 DRMR（Documents Ranking by Mutual References）演算法（註 26）以及專門蒐集科學文獻電子檔的 CiteSeer 等也都是利用引用概念作為排序依據，針對此種排序概念，本單元將僅介紹最早且最重要的PageRank及HITS二種網頁相關排序法。

表一　搜尋引擎排序選項比較

輸出（Output）					
	排序指標				排序選項
	reviewed status	meta tags	link popularity	direct hit	
Alta Vista			✓		相關性 關鍵字 按網站群組
Excite			✓		相關性 按 URL 排序
FAST			✓		相關性
Go(Infoseek)	✓	✓	✓		相關性 按日期排序 按網站群組
Google			✓		相關性 按網站群組
GoTo			✓		
HotBot		✓	✓	✓	按網站群組
Lycos				✓	
Northern Light			✓		相關性 按日期排序 按網站群組
WebCrawler					

資料來源：謝寶煖，「搜尋引擎比較表」，2000 年 11 月 15 日，<http://www.lis.ntu.edu.tw/nhsieh/articles/SE.htm > (15 Dec. 2003).

1. PageRank

PageRank 是1998年 Larry Page 及 Sergey Brin 創建Google搜尋引擎所採用的排序方法，PageRank的概念係假定從一個網頁被其他網頁的連結情形，可顯示該網頁的重要性高低，當一網頁被連結的次數愈多，其重要性也相對愈高，排名順序也愈高，如果外來連結的網頁是重要性較高的網頁時，也會提高被連結網頁的 PageRank，因此PageRank並不是將所連結均視爲同等重要，而有考慮每個網頁

本身的PageRank，也就是說影響PageRank 數值高低的因素不僅僅是連結次數的多寡。

網頁之間的連結關係可分爲向外連結至其他網頁（Forward Links）及被其他外來網頁連結（Backlinks）二種連結方向，每個網頁向外連結出去的連結，其 PageRank 即爲該網頁的PageRank除以連結數，例如包含二個向外連結之網頁A之PageRank爲100，其二個向外連結出去的連結，每一個連結所包含的PagaRank即爲50（如圖二），因此就網頁被連結的情形來看，被連結的連結數愈多，本身的PageRank 就愈高，此點從下列 PageRank的計算公式（註27）也可明顯看出此種關係。

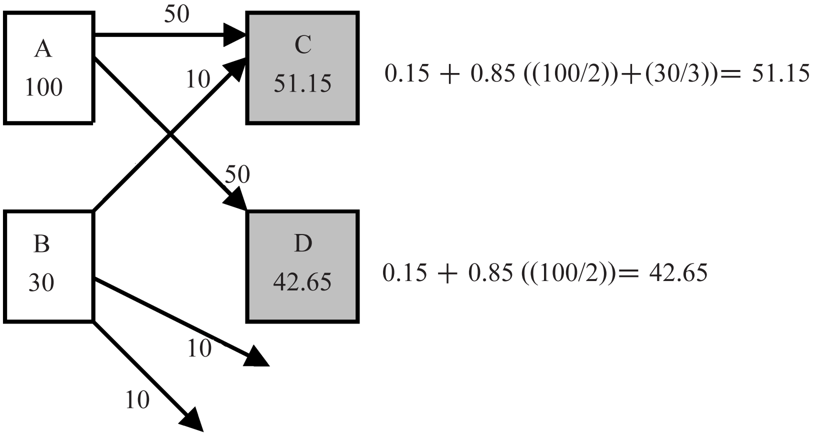
PR(A)＝(1 － d)＋ d(PR(T1) / C(T1) ＋……＋ PR(Tn) / C(Tn))

上述公式係說明當A網頁被T1、T2…Tn個網頁連結，PR(T1)爲T1網頁的 RageRank，C(T1)爲 T1網頁向外連結的連結數，參數d爲介在 0與1之間的衰退因素（Decay Factor），預設值爲0.85，其引入之目的在減小 T1、T2…Tn對A網頁的影響，因爲PageRank本身符合機率分布，因此加上 1-d 後得到的結果在計算達到平衡後，所有網頁的 PageRank 平均值爲 1（註28），故該公式可直接寫成：

PR(A)＝ 0.15 ＋ 0.85(PR(T1) / C(T1)＋……＋ PR(Tn) / C(Tn))

以圖二爲例，網頁C及網頁D的PageRank 可分別計算得出 51.15 及42.65。

PageRank的主要概念係在模擬網路的隨意漫遊（Random Surf），亦即



圖二　PageRank 演算法示意圖

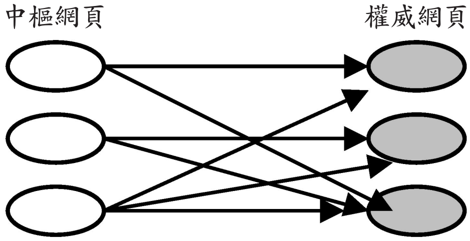
考慮使用者在網路上隨機選擇一路徑跳至另一個隨機網頁的情形，因此每個網頁PageRank也就是使用者隨機漫遊網頁的機率。（註29）至於 Google 對檢索結果的相關排序，實際上也是先以使用者輸入的查詢關鍵字進行比對，並依查詢關鍵字所在位置予以加權進行相似性排序後，再依據各網頁的PagaRank對原先排序結果進行排序次序的調整。

2. HITS

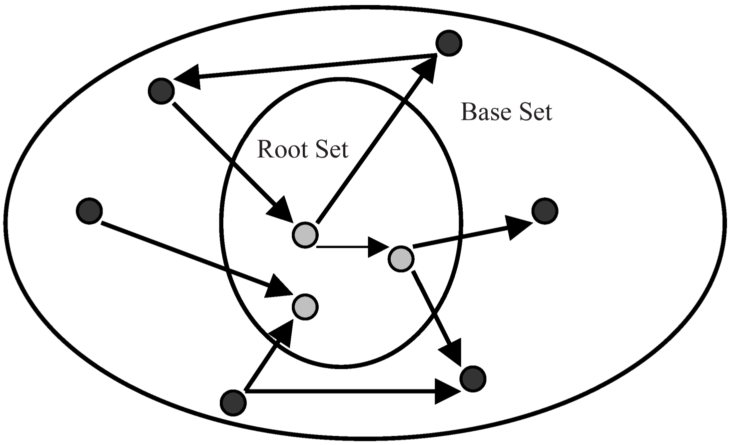
HITS（Hypertext Induced Topic Search）是IBM公司發展的Clever搜尋引擎所採用的網頁排序方法，它是由Jon Kleinberg在1998年繼PageRank推出後，另一個利用網頁連結架構來進行網頁相關排序的演算法，其將網頁分成權威網頁（Authoritative Page）與中樞網頁（Hub Page）二類（如圖三），對於一特定主題而言，權威網頁係擁有許多與該主題有關內容的外來連結網頁，中樞網頁則是擁有許多指向權威網頁連結的網頁。

以 HITS 之概念而言，如果一個網頁之外來連結數愈多，則此網頁的權威分數（Authority Score）就愈高，因此，這個網頁與連結到它本身的所有網頁的內容相關性也較強，相反地，如果一個網頁向外連結到其他網頁的連結數愈多，則該網頁的中樞分數（Hub Score）就愈高，表示其內容包含了較多的相關連結（註30），因此好的中樞網頁所連結到的權威網頁是屬於好的權威網頁，相對地，好的權威網頁其外來連結的網頁也是好的中樞網頁，二者之間的關係也是建立在相互推薦的依存關係上。

HITS 的相關排序運作方式是先找出符合查詢關鍵字的預設數量的相關網頁做為 Root Set，其次加入和 Root Set 網頁有連結關係的網頁，擴展形成Base Set，以Base Set建構出一個鄰區圖（Neighborhood Graph）（如圖四），最後再依據演算公式計算出 Base Set 中每個網頁作為權威網頁的權威分數及中樞網頁的中樞分數，據以進行相關排序。權威分數的矩陣（a[]）係A矩陣與中樞分數的矩



圖三　中樞網頁與權威網頁示意圖



圖四　Root Set 與 Base Set 關係示意圖

資料來源：Jon M. Kleinberg, "Authoritative Sources in a Hyperlinked Environment," Journal of the ACM 46:5(September 1999): 609.

陣的積（h[]），亦即h[] = Aa[]，其中A矩陣是Base Set各網頁的關係矩陣；而 a[]＝ A^Th[]，A^T爲 A 矩陣的轉置矩陣，（註31）詳細的演算說明請參見圖五。

從圖五可得知 p1、p2、p3三者之間的A關係矩陣爲 ：

$$A = \begin{matrix} 1 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{matrix}$$

(1表示有連結關係，0表示沒有連結關係)。

繼而可依據A矩陣得知：

$$A^T = \begin{matrix} 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{matrix}$$

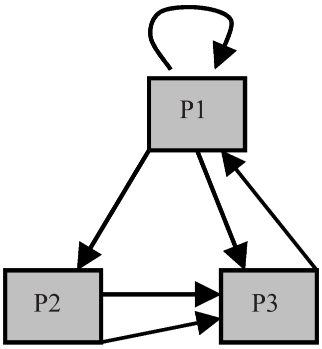
$$A\,A^T = \begin{matrix} 3 & 1 & 2 \\ 1 & 1 & 0 \\ 2 & 0 & 2 \end{matrix}$$

$$A^TA = \begin{matrix} 2 & 2 & 1 \\ 2 & 2 & 1 \\ 1 & 1 & 2 \end{matrix}$$

並假設p1、p2及p3之$h^{(0)}$及$a^{(0)}$均爲[1 1 1]，然後依據權威分數與中樞分數的公式，即可得出每個網頁的權威分數及中樞分數，作爲相關排序調整的依據。

PageRank與HITS均是利用網路連結關係作爲相關排序的依據，但二者的差異主要包括：

⑴就 Global 及 Local 的觀點，Page-



圖五　三個網頁之連結關係

資料來源：Te-Xuan Lin, "HitRank: Search Ranking Based on Mining User-Oriented Feedback," (Master Thesis, Department of Electronic Engineering, National Taiwan University of Science and Technology, July 2002), 26.

Rank 是計算文件集中所有網頁的PageRank，因此不管是相同或不同使用者使用何種查詢關鍵字，其最後調整檢索結果相關排序的 PageRank 是事先已計算好的固定數值，不受查詢主題變動而異動，而HITS則是針對一小群Base Set中每個網頁為權威分數與中樞分數的計算範圍，因此隨著使用者查詢不同主題所形成的不同Base Set，每個網頁的權威分數與中樞分數均是重新計算，並無固定的數值，故與PageRank 比較，PageRank 是屬於Global 概念，HITS 則是屬於 Local觀念。（註32）

⑵ PageRank 係就整個文件集中所有網頁彼此連結關係計算每個網頁的PageRank，因此即使使用者的查詢主題不同，每個網頁的 PageRank仍是固定的，且 PagaRank只計算外來方向的連結（Forward Direction），亦即指向權威網頁的連結關係，而 HIT則將網頁分成權威網頁與中樞網頁，考慮連結方向有本身被其他外來除連結之關係及本身向外連結至其他權威網頁之二種關係，不像 PageRank 僅考慮一個方向的連結關係。

三、使用者回饋

使用者回饋係將使用者的相關判斷結果作為相關排序的考量因素，利用使用者閱覽檢索結果後所點選的結果，作為使用者下一次使用相同查詢詞，檢索系統運作檢索結果的調整依

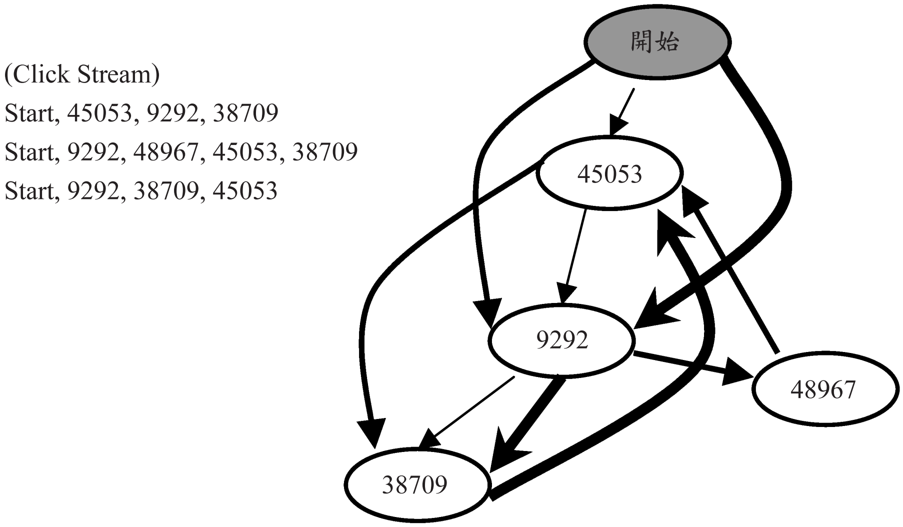
據，提供更具個別化的資訊服務。

1. Direct Hit

Ask Jeeves公司發展的Direct Hit係當搜尋引擎將查詢結果呈現給使用者時，即開始追蹤使用者對檢索結果的點選情形，如果排序在前面的網頁被使用者點選瀏覽，且瀏覽時間較長，則表示該網頁受歡迎的程度較高，且系統會增加該網頁的相關度，反之，如果使用者花很短的時間瀏覽所點選的網頁，則表示該網頁的相關度較低，因該演算法建立在使用者點選的基礎上，故又被稱爲受歡迎程度（Popularity）演算法。由於Direct Hit的排序依據是由網頁被點選（Click-through）的次數及被瀏覽的時間長短決定，加上使用者的相關判斷具有不斷變動的特質，導致在不同時間使用相同查詢詞的檢索結果相關排序判斷可能會不同，因此Direct Hit是屬於一種動態排序。（註33）

2. HitRank

HitRank 是一種以使用者導向的回饋方式進行檢索結果的排序，這種排序方法首先將使用者瀏覽檢索結果所點選的順序，以一個具方向順序的圖形表示出來，稱之為「點選順序圖」（Click Stream）（如圖六），並計算點選順序圖內每個網頁的分數，做爲網頁排序的依據，因此 HitRank



圖六　點選順序圖（Click Stream）

資料來源：Te-Xuan Lin, "HitRank: Search Ranking Based on Mining User-Oriented Feedback," (Master Thesis, Department of Electronic Engineering，National Taiwan University of Science and Technology, July 2002), 36.

除了考慮到該網頁被使用者點選的次數，也考量了使用者的眞正檢索行爲。

HitRank 包括二點假設，一是網頁被點選的次數愈多，重要性也愈高，二是點選順序圖中最後一個被點選的網頁具有最高的相關分數，另從下列的演算公式（註 34）可以看出 HitRank 採用同 PageRank 架構的計算模式。

$$HR(I) = d + (1 - d)\Sigma HR(Ii)L(Ii) / C(Ii)$$

d 預設值爲 0.15，表示原始權重值
HR(I)爲網頁 I 的 HitRank 分數
Ii 是 I 的外來連結
C(Ii)是 Ii 的對外連結數
L(Ii)是 Ii 到 I 之間的連結數

3. QueryFind

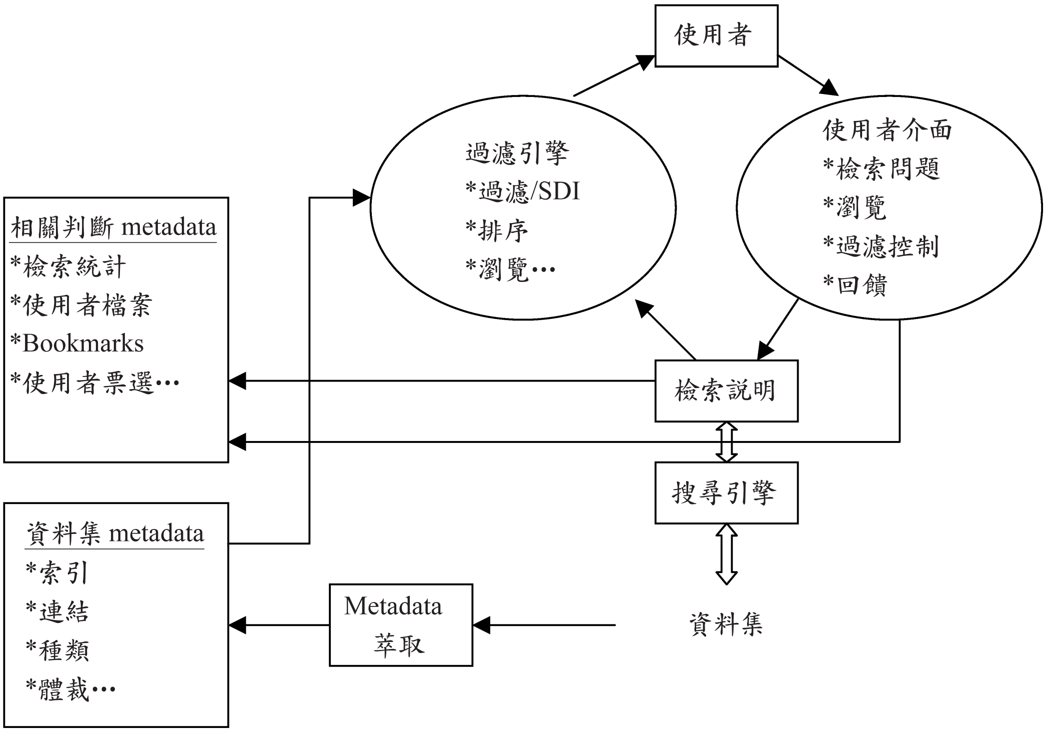
QueryFind 排序概念和 HitRank 很類似，利用分析 Query Logs 包含的檢索問句、檢索結果的點選紀錄、網頁資訊以及瀏覽網頁等豐富訊息，可據以了解使用者的檢索行爲並做爲改善檢索效果的依據，惟二者的差別在於，QueryFind 未建構使用者的點選順序圖，僅需蒐集被點選的網頁及該網頁在其原始搜尋引擎的排名次序，故可減少資料的處理及分析時間。（註35）此外，QueryFind 的基本假設也是認爲一個網頁如果被使用者點選的次數愈多，該網頁的使用者回饋排序分數（Users' Feedback Ranking Score）也愈高，同時也參考原始搜尋引擎所給予的相關排名次序，亦即所謂的內容導向排序分數（Content-Oriented Ranking Score），故其和以下Meta Search 所採用的重新合併排序方法概念很類似。

4. 價值過濾（Value Filtering）

大部分的資訊檢索系統係利用文件與使用者問題之間的相似性作爲相關排序依據，各個系統的差異僅在於相似性演算方法的效能，且此種相似性的演算主要是針對文字型的資料，不過，網路資源除了文字型資料外，還包括聲音、影像、圖片等非文字的有參考價值的資訊，因此提出以資訊價值爲基礎的查詢及排序方法，突破相似性的傳統做法及限制。價值過濾技術目前分爲內容式（Content-Based）及行動式（Action-Based）二類，前者以資源內容爲過濾範圍，依據各種抽取的 metadata 進行過濾或排序檢索結果；後者則是觀察使用者的資訊使用行爲，分析大規模抽取的使用者相關判斷資訊，因此價值過濾的概念也是在內容基礎上，加入使用者的使用行爲及相關判斷的回饋資訊，價值過濾系統的概念架構如圖七所示。（註36）

四、合併排序法

Meta Search 之出現解決了使用者爲提高檢索精確性及完整性而須分別去使用不同搜尋引擎的問題，因爲每個搜尋引擎的特性不同，在沒有任何一個搜尋引擎能完整包含所有網路資



圖七　價值過濾系統的概念架構

資料來源：Andreas Paepcke, Hector Garcia-Molina, Gerard Rodriguez-Mula, and Junghoo Cho, "Beyond document Similarity: Understanding Value-Based Search and Browsing Technologies," ACM SIGMOD Record 29:1 (March 2000): 82.

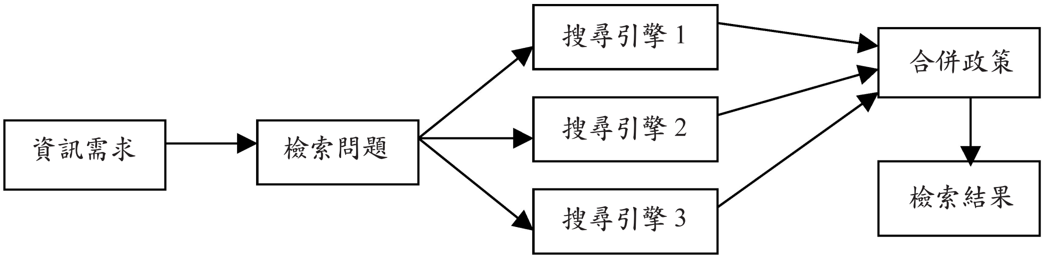
源的情形下，使用者有同時查詢二個或二個以上搜尋引擎的實際需求，針對上述問題，Meta Search的設計是當它接收到使用者所下達查詢詞後，會自動將查詢詞傳送到幾個不同的搜尋引擎，由這些搜尋引擎進行查詢，之後再整併所有搜尋引擎查詢的結果並加以排序，提供使用者參考，其本身並不直接收錄網頁。不過對於各個搜尋引擎回傳之檢索結果，Meta Search不一定會採用最適合使用者的理想的重排序計算方法，可能僅是依回傳結果的早晚來進行合併，或是依照搜尋引擎所提供的網頁排名來處理，或是依據網頁的一些特性加以分類，充其量僅是將不同搜尋引擎檢索到的重複網頁刪除後，再全部呈現出來。因為相同網頁在不同的搜尋引擎對相同檢

索問句的相關性並不相同，且每個搜尋引擎的相關排序演算法均為商業機密，故對於網頁合併問題，許多Meta Search 也開始各自發展屬於自己的排序演算法，並利用搜尋引擎傳回的網頁相關資料（如名次、原始分數等），重新計算各網頁的相關性，最後再加以重新排序呈現給使用者參考，只是在無法取得原始搜尋引擎的排序方法下，此種重新計算每個網頁相關性的合併方式，也僅是一種猜測式的合併方式。（註37）

為改進傳統合併排序法之缺失，另一種漸進式合併排序法係以相關名次做為合併排序之依據，其認為不同的搜尋引擎對網頁相關分數的計算方式均不同，因此相同的網頁在不同搜尋引擎的相關分數是有差異的，無法直接以網頁的原始相關分數直接比較，因為分數高不代表相關性也高，但是相關性大的網頁一定會排在前面，因此可以利用相關名次比較相關性的大小。漸進式合併排序法在每一次查詢時，會給予搜尋引擎適當的權重，在經過一次又一次的合併查詢，可將使用者認為最相關的網頁漸進地往前排序，其做法是將各搜尋引擎傳回的網頁，根據各搜尋引擎的參考權重除以該網頁在原搜尋引擎的排名次序進行網頁合併後，讓使用者瀏覽並從中擇選出相關網頁，並根據使用者選出的網頁計算下一次使用者使用相同查詢詞時搜尋引擎應有之權重，透過這種漸進式的合併方法，可針對每個人的不同需求呈現出不同之搜尋結果，達成有意義的合併搜尋及個別化的資訊服務。（註38）（註39）

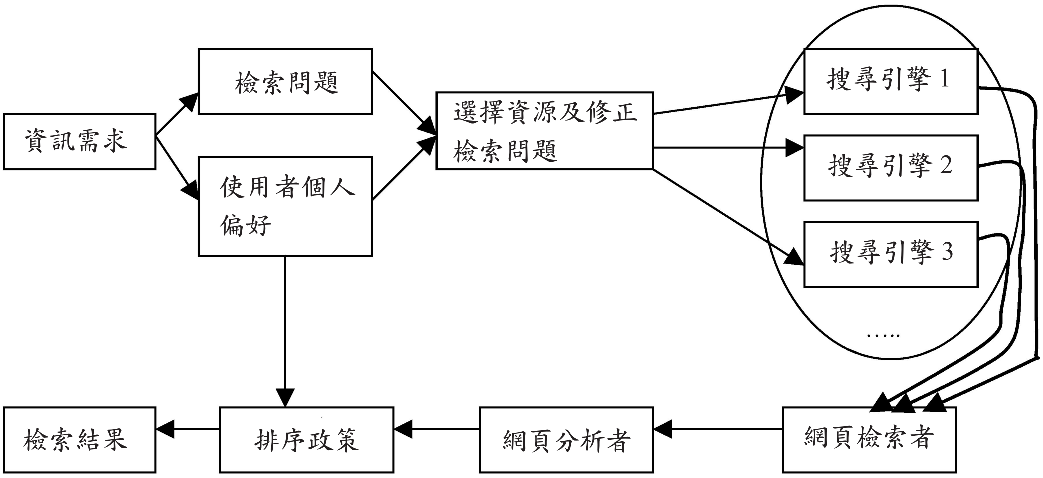
此外，NEC 研究機構研發的 Inquirus 2 Meta search 結合查詢問句及資訊需求類別，進行個別化的檢索結果排序。先前 Inquirus 之設計是將使用者的查詢問句分送十幾個搜尋引擎進行檢索，並依據對該查詢問句的預測相關進行檢索結果之合併，但新一代的Inquirus 2進一步將相關概念擴大包括使用者偏好（User Preference），因此不同的使用者即使使用相同的查詢問句，所得到的檢索結果並不相同，而是較符合個人需求的個別化建議。從圖八與圖九的比較，可以明顯看出Inquirus2和一般Meta Search架構的區別在於網頁檢索者（Page Retriever）、網頁分析者（Page Analyzer）及使用者偏好三部分，其中使用者偏好影響檢索結果的排序結果，而智慧型代理人（Intelligent Agent）可以針對使用者的個人資訊需求，每天從網路上抓取目前最新的相關網頁給使用者，在使用者閱覽過網頁後，把相關回饋傳送給智慧型代理人計算相關回饋，作為下一次代理人系統調整抓取網頁的依據。（註40）（註41）雖然智慧型代理人可以帶給使用者有用的網頁資料，但缺點是所抓取的資料量是相當有限的。

上述介紹的三大類相關排序方法，以使用者的角度來看，第一類以關鍵字比對及關鍵字所在位置來計算



圖八　Meta Search 架構

資料來源： Eric J. Glover, Steve Lawrence, Michael D. Gordonn, William P. Birmingham and C. Lee Giles, "Recommending Web Documents Based on User Preferences," <http://citeseer.nj.nec.com/cache/papers/cs/10692/http:zSzzSzwww.cs.umbc.eduzSz~ianzSzsigir99-reczSzpaperszSzglover_e.pdf/glover99recommending.pdf>



圖九　Inquirus 2 搜尋引擎架構

資料來源： Eric J. Glover, Steve Lawrence, Michael D. Gordonn, William P. Birmingham and C. Lee Giles, "Recommending Web Documents Based on User Preferences," <http://citeseer.nj.nec.com/cache/papers/cs/10692/http:zSzzSzwww.cs.umbc.eduzSz~ianzSzsigir99-reczSzpaperszSzglover_e.pdf/glover99recommending.pdf>

相關性高低的基本方法，是最容易被使用者理解該某文件被檢索且被排序在最前面的原因，因爲許多檢索系統會將檢索到的文件書目紀錄或全文內

容中，符合使用者輸入的關鍵字者以明顯之顏色標示出來，透過被明顯標示的關鍵字及標示區塊之多寡，可以方便使用者了解該文件被系統視爲相關性高低與否的原因。至於其他相關排序考量的因素，例如逆向文件頻率、文件相似性、PageRank等，使用者則較不容易理解其排序原則。

陸、相關排序問題

從上述各種相關排序方法，可以發現相關排序的演進是在原有的傳統檢索基礎上再加入其他的補強因素，並朝向利用使用者判斷及使用紀錄的探勘研究，針對不同使用者的知識狀態差異，提供個別化的資訊服務，但就目前而言，相關排序方式所遇到的問題包括：

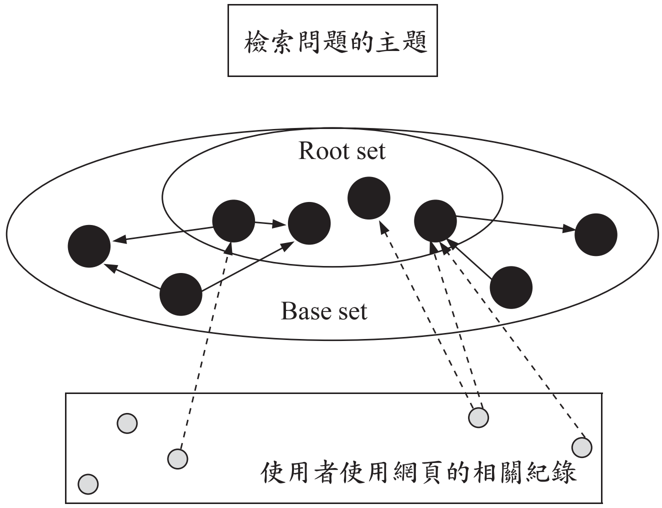
一、使用者查詢詞的權重

目前大部分檢索系統所進行的相關排序方法是以布林檢索模式爲基礎，亦即每個查詢關鍵字和文件之間的關係是相同的，並未考慮不同使用者會使用不同查詢關鍵字來表示相同概念的差異。因以文件相似度來定義相關，故詞頻愈高的文件，重要性也愈高。針對此點，檢索者必須要注意是否對同一主題概念使用一致性的關鍵字，以免因使用不同關鍵字導致代表同一主題概念的強度被減弱，影響相關文件被檢索到的機會，且可以推知，如果要刻意提高文件的相關序次，只要於文件內容中儘量重複，甚至濫用特定關鍵字出現的次數，則會在不正確的文件主題誤導下，讓使用者檢索到不相關的文件，影響檢索的精確率及使用者對檢索系統的評價。

二、網頁間連結的架構問題

以PageRank計算方式而言，如果某一網頁很早就出現，或是已經被很多網頁連結，則其PageRank會繼續增加，產生大者恆大的現象，因此對新網頁而言，因爲尚未或還很少與其他網頁有連結關係，導致即使是屬於品質很好且重要性很高的網頁，也不太可能在剛產生時就成爲好的權威網頁。此外，就HITS而言，HITS是就符合查詢關鍵詞之少數節點所擴展的Base Set的一小部分相關網頁進行權威分數及中樞分數的計算，因此根據每一次檢索所形成的不同Base Set，都要立即重新計算網頁的相關數值，不同於PageRank係就文件集合中全部網頁計算好的固定數據，導致排序結果的回應時間也會較久。（註42）

事實上，不論是PageRank或是HITS，網頁分數都是一連串重複的複雜計算，因爲網頁之間的關係一種複雜的網狀圖，需對全部有連結關係的網頁一一進行計算，直到收斂狀態爲止，故每個網頁的計算是很耗費時間的。PageRank係事先已計算好數值，不會讓使用者感覺到系統所需花費的時間，但HITS因是當場計算，



圖十　專家發現演算法

資料來源：Jidong Wang, Zheng Chen, Li Tao, Wei-Ying Ma and Liu Wenyin, "Ranking User's Relevance to a Topic through Link Analysis on Web Logs," in Proceeding of the Fourth International Workshop on Web Information and Data Management, (November 8, 2002), 51.

故在實際應用上，系統所需花費的時間是一大問題。此外，HITS 只計算Base Set 小範圍的各網頁分數，因此只要有新的網頁連結增加至 Base Set範圍內，即會明顯影響原本整個Base Set 中各網頁的權威分數及中樞分數，這也意謂著 HITS 的權威分數和中樞分數比PageRank更容易受到網頁作者的操控。

因為Base Set是依據和Root Set有連結關係的網頁所擴展形成的，因此非Root Set的其他Base Set中的網頁內容，並無包含使用者的查詢關鍵字，雖然Base Set和Root Set有連結關係，但也有可能是和使用者檢索主題無關的網頁，不具主題相關性，產生所謂「主題漂流」（Topic Drift）的問題。針對主題漂流問題，中國大陸學者 Wang 等人依據 HITS，提出專家發現（Expert Finding）演算法，將網頁的重要性及網路使用者的專業性作為演算法計算的考量因素。網頁使用者的專業水準係依據其所到訪的網頁品質作判斷，如果使用者到訪過的網頁品質愈高，則使用者的專業水準也愈高，同時，網頁的重要性是受到其他

網頁的引用及使用者拜訪次數的影響。（註 43）找出專家的程序如同HITS於形成Base Set後，再加入使用者使用網頁的到訪紀錄，利用演算法將專業水準最高的專家找出來，然後藉由分析專家經驗，將之應用在資訊檢索系統上，此方法因加入網頁內容分析的因素，因此可提高檢索結果的可靠性。

三、使用者回饋

使用者回饋係將使用者使用過的查詢關鍵字予以紀錄，並計算相對的各搜尋引擎最佳權重，作爲下一次該使用者使用相同查詢關鍵字的參考，提升檢索結果的相關性。但爲了讓使用者使用新查詢關鍵字時也能獲得較合理的合併排序結果，可以將使用者的查詢關鍵字予以適當分類，讓同一類別的查詢關鍵字除了有自己的搜尋引擎權重，另外還擁有共同的搜尋引擎權重，那麼當使用者使用一個從未使用的新查詢關鍵字時，系統也可以提供較佳的合併排序方式。此外，也可以善用使用者的查詢經驗，例如假使 A、B 二個使用者的查詢紀錄顯示二人對相同的查詢詞具有類似的相關回饋時，當A使用者使用一個B使用者曾使用過的新的查詢詞時，系統可以利用B使用者的回饋資料，來預測對A使用者較合理的合併排序結果。（註44）

目前 Inquirus 2、SavvySearch、MetaSEEK、ProFusion 等 Meta search具有資源選擇功能，例如SavvySearch提供使用者選擇符合資訊需求之特定資源類別，例如使用者如要找尋新聞，則可限定搜尋引擎只檢索相關新聞網站即可，不過此種方法的缺點可能會遺漏掉一般性檢索引擎可搜尋到的資料。Inquirus 2 則利用和各個資源類別對應好的固定搜尋引擎清單，配合查詢問句修改功能，提供使用者也可依自己的需求去修正查詢問句，針對特定需求指定查詢一般性的搜尋引擎，因此以前面使用者找尋新聞的例子，雖然不會使用一般性的搜尋引擎去搜尋，但使用者仍可指定某個一般性搜尋引擎，並要求檢索結果依日期先後排序，或限定只要某段時間範圍的新聞資訊即可。（註45）此種考慮使用特定資訊需求的檢索選擇，可讓使用者所使用的查詢問句能更符合個人的資訊需求，且 Inquirus 2在檢索結果的排序上還納入使用者偏好的考量，提供個別化的資訊服務。

目前許多Meta Search系統係採各自設計的排序演算法，並利用搜尋引擎傳回的網頁相關資料（如名次、原始分數等）重新計算各網頁的相關性，最後再加以重新排序呈現給使用者（註46），不過，因各搜尋引擎傳回有關網頁的資訊並不完整，系統必須花較多時間去執行重新計算網頁內容的工作，以獲得正確的網頁排名。NEC 研發的 Inquirus 及後續改良之Inquirus2，其架構和一般Meta search

最大的不同特點，是它下載網頁內容完整的資訊，將「死的」連結及不相關的網頁先過濾排除後，再進行檢索結果排序，大幅改善精確率（註47），不像傳統的Meta Search僅就接收到的標題、URL等不完整的網頁資訊，據以計算並排序。

針對個別資訊化的需求，相關排序的考慮因素也將使用者於瀏覽檢索結果後所作的決定、網頁停留時間，拜訪次數等使用紀錄列入研究範圍，希望將使用者的資訊行爲或資訊需求加入系統處理檢索結果的考慮因素。因系統要計算分析的訊息更多更複雜，而目前有關使用者回饋的研究只停留在少量的研究實驗，尚未應用在眞實的大量資訊環境中，且面對使用者知識狀態不斷改變的動態情形，個別化資訊服務有相當大的困難，但爲了朝此理想邁進，研究者勢必須對使用者進行更深入的研究。

四、人爲操控

其實目前相關排序的最大問題係屬於所謂人爲操控的問題。目前各搜尋引擎受到商業利益的操控，推出付費優先排序的服務，亦即只要花錢就可以讓搜尋引擎公司依其本身相關排序方式，以人爲方式操控網頁排序的次序，成爲另一種廣告的行銷方式。例如針對以關鍵字比對排序的搜尋引擎，則於本身網頁、Mata Tag 或 Image Tag中充斥重複與自己網站內容無關但是流行的關鍵字，以增加特定關鍵字的詞頻；如果是以超鏈結關係作爲相關排序依據，則製造大量網站產生大量連結，提升排序名次；如果是以使用者點選次數的多寡作爲相關排序依據，則很容易增加點選紀錄，因此人爲操控的競價排序法（Pay Per Click）直接影響了搜尋引擎相關排序的可信度，嚴重侵害使用者獲取資訊的精確率。

目前一些著名的搜尋引擎均有提供優先排序之服務，造成此種市場需求的原因，主要是爲了提高網站的曝光效果，欲將排名位置放在最前面，讓網路使用者可優先看到該網站之資料，以達到廣告效果，因爲在使用者無耐心看完全部檢索結果的情形下，排序名次是否名列前矛，尤其出現在第一頁的顯示位置是很重要的，尤其對商業網站而言，搜尋引擎的訊息是具有廣告及行銷功能之工具。

柒、結語

理想上，檢索系統相關判斷的結果及排序應該和使用者相關判斷結果一致，但目前檢索系統對不同資訊需求者如使用相同的查詢關鍵字，其檢索結果是一樣的。檢索系統進行的工作，即是將資訊需求者的查詢關鍵字和系統代表文件的索引詞進行比對，不管查詢問句或是文件，最後系統顯示的檢索結果，係以查詢詞比對的結果再加入其他因素據以調整的結果。受到時間的影響，詞彙與概念之間並

非是一對一的絕對關係，因此一個概念的表示，可能是一個單一字彙或是二個或二個以上單一字彙組合之複合詞，且相同概念可以不同的詞彙表示，而相同詞彙也可能有不同概念的意義，因此，基本上，概念與詞彙的轉譯過程是很複雜且困難的工作，需要檢索經驗的累積與訓練才能有一定的檢索水平。

好的排序方法，除了提供相關的資料給資訊需求者外，還要依相關性高低排序減少資訊需求者找尋最相關資訊的時間，尤其面對資訊快速的成長，相關排序的需求愈是凸顯其重要性。為了讓系統的檢索結果能更貼近使用者個別化的資訊需求，使用者回饋資訊及使用者檢索紀錄均是資料探勘的研究範圍，只是分析使用者紀錄的困難性，以及所能眞正反映使用者資訊需求的正確性問題，使得目前包含使用者回饋的相關排序方法，僅應用在範圍較小的個人化系統上，尙未應用在大範圍的資料檢索上。

理想上，相關排序的考慮因素應如同使用者進行相關判斷一樣，因此同時採取多種因素進行複雜的計算，將不同相關排序特性加以整合，提高相關排序的效能是相關排序所要努力之方向，例如 Openfind 的多元排序法（PolyRank），強調其不同於一般搜尋引擎將連結次數愈高的網站排在前面的結果，因為這樣將存在大者恆大以及無論什麼時候查詢都會有相同結果的缺點，故 PolyRank 排序依據係以連結統計、關鍵字出現的位置、頻率、與內容之吻合度、網頁更新時間、網頁型態、格式等多項指標共同決定，改善傳統搜尋引擎的缺失，並可讓使用者以不同方法進行檢索結果的排序（註 48），因此正如同相關概念看似簡單又複雜多變的特性，相關排序就目前檢索系統之現況來看，仍有許多待發展及改進的空間。（本文筆者在此感謝陳教授光華及二位匿名審查者之悉心指正）

註釋

註 1：Erica Cosijn and Peter Ingwersen, "Dimensions of Relevance," Information Processing and Management 36:4 (2000): 533-534.

註 2：Stefano Mizzaro, "How Many Relevances in Information Retrieval?" Interacting with Computer 10:3 (1998): 303-320.< http://www.dimi.uniud.it/izzaro>(29 Dec. 2003).

註 3：同註 1，頁 535。

註 4：Kelly L. Maglaughlin and Diane H. Sonnenwald, "User Perspectives on Relevance Criteria : a Comparison among Relevant, Partially Relevant, and Not-Relevant Judgment," Journal of the american Society for Information Science and Technology 53:5 (2002): 328.

註 5：黃慕萱，「資訊資源與檢索」，在圖書資訊學概論，賴鼎銘等編著，初版（台北縣：空中大學，民國 90 年），頁 196、198-201。

註 6：同註 1，頁 535。

註 7：Pia Borlund and Peter Ingwersen, "Measures of Relative Relevance and Ranked Half-Life," in Proceedings of the 21st annual international ACM SIGIR conference on Research and Development in Information Retrieval, (August 1998), 325.

註 8：同註 1，頁 535。

註 9：同註 2。

註 10：黃慕萱，「影響相關判斷之因素探討」，資訊傳播與圖書館學 4 卷 3 期（民國 87 年 3 月），頁 40-41。

註 11：同註 10，頁 43。

註 12：同註 4，頁 327。

註 13：Eero Sormunen, Jaana Kekalainen, Jussi Koivisto and Kalervo Jarvelin, "Document Text Characteristics Affect the Ranking of the Most Relevant Documents by Expanded Structured Queries," Journal of Documentation 57:3 (May 2001): 371.

註 14：Caroline M. Eastman and Bernard J. Jansen, "Coverage, Relevance, and Ranking: The Impact of Query Operators on Web Search Engine Rusults," ACM Transactions on Information System 21:4 (October 2003): 384.

註 15：Yuval Z. Feinsten, Claudia V. Goldman, Yishay Mor, and Jeffrey S. Rosenschein, "Relevancy Ranking of Web Pages Using Shallow Parsing," <http://citeseer.nj.nec.com/cache/papers/cs/966/ftp:zSzzSzftp.cs.huji.ac.ilzSzuserszSzclagzSzpadd97zSzyuval.pdf/relevancy-ranking-of-web.pdf>(2 Dec. 2003).

註 16："Relevancy Ranking," <http://www.lextek.com/manuals/onix/ranking.html> (2 December 2003).

註 17：D. Scott Brandt, "Relevancy and Searching the Internet," Computers in Libraries 16:8 (September 1996): 38.

註 18：Offer Drori, "Algorithm for Documents Ranking: Idea and Simulation Results," in Proceeding of the 14th International Conference on Software Engineering and Knowledge Engineering, (15-19 July 2002, Ischia, Italy), 100.

註 19：蔡政容，「從多重搜尋引擎中選取最前 K 名搜尋結果策略之設計與分析」（碩士論文，國立台南師範學院資訊教育研究所，民國 89 年），頁 10。

註 20：楊允言，「文件自動分類及相似性排序」（碩士論文，國立清華大學資訊科學研究所，民國 82 年），頁 106。

註 21：曾元顯，林瑜一，「模糊搜尋、相關詞提示與相關詞回饋在 OPAC 系統中的成效評估」，中國圖書館學會會報 61 期（民國 87 年 12 月），< http://www.lins.fju.edu.tw/seng/papers/crystal/crystal.htm > (12 Nov. 2003).

註 22：Fabio Crestani, Mounia Lalmas, Cornelis J. van Rijsbergen, and Iain Campbell, "Is This Document Relevant?‧‧‧ Probably: a Survey of Probabilistic Models in Information Retrieval," ACM Computing Surveys 30:4 (December 1998): 532.

註 23：Mei Kobayashi and Koichi Takeda, "Information Retrieval on the Web," ACM Computing Surveys 32:2 (June 2002): 150.

註 24：吳宜松，「應用之晰概念於個人化網頁排序之研究」（碩士論文，國立中央大學資訊管理研究所，民國 91 年），頁 8。

註 25：Micheangelo Diligenti, Marco Gori, and Marco Maggini, "Web Page Scoring Systems for Horizontal and Vertical Search," in Proceedings of the Eleventh International Conference on World Wide Web, (May 2002), 508.

註 26：同註 18，頁 99-102。

註 27：Sergey Brin and Lawrence Page, "The Anatomy of a Large-Scale Hypertexual Web Search Engine," < http://www7.scu.edu.au/programme/fullpapers/1921/com1921.htm >(10 Nov. 2003).

註 28：常璐，「搜尋引擎的幾種常用排序算法」<http://lib.sdx.js.cn/libjywk/displayarticle.asp?ID=170>(1 Dec. 2003).

註 29：Pavel Calado, Berthier Ribeiro-Neto, Nivio Ziviani, Edlleno Moura, and Ilmerio Silva, "Local Versus Global Link Information in the Web," ACM Transactions on Information System 21:1 (January 2003): 46.

註 30：同註 24，頁 5-6.

註 31：Te-Xuan Lin, "HitRank: Search Ranking based on Mining User-Oriented Feedback," (Master Thesis, Department of Electronic Engineering, National Taiwan University of Science and Technology, July 2002), 25.

註 32：Behnak Yaltaghian and Mark Chignell, "Re-ranking Search Using Network Analysis: a Case Study with Google," in Proceedings of the 2002 Conference of the Centre for Advanced Studies on Collaborative Research, (September 2002), 2.

註 33：同註 28。

註 34：同註 31，頁 34。

註 35：Po-Hsang Wang, "QueryFind: Search Ranking based on User's Feedback and Expert's Agreement," (Master Thesis, Department of Electronic Engineering, National Taiwan University of Science and Technology, July 2002), 26.

註 36：Andreas Paepcke, Hector Garcia-Molina, Gerard Rodriguez-Mula, and Junghoo Cho, "Beyond Document Similarity: Understanding Value-Based Search and Browsing Technologies," ACM SIGMOD Record 29:1 (March 2000): 80-92.

註 37：同註 19，頁 14-15。

註 38：周明仁，「異質性網路資源搜尋結果的合併排序法之研究與設計」（碩士論文，國立台南師範學院資訊教育研究所，民國 88 年），頁 3。

註 39：李建億等，「適應個別網路學習者的漸進式合併搜尋系統之研究」，第八屆國際電腦輔助教學研討會，三月十八日至二十日（台中市：逢甲大學，民國 88 年），<http://paper.

nt1.isst.edu.tw/data01/acbe/iccai8/11/11.htm> (13 Nov. 2003).

註 40：同註 38。

註 41：同註 39。

註 42：同註 35，頁 18、22。

註 43：Jidong Wang, Zheng Chen, Li Tao, Wei-Ying Ma and Liu Wenyin, "Ranking User's Relevance to a Topic through Link Analysis on Web Logs," in Proceeding of the Fourth International Workshop on Web Information and Data Management," (November 8, 2002), 53.

註 44：同註 19，頁 71-72。

註 45：Eric J. Glover, Steve Lawrence, Michael D. Gordonn, William P. Birmingham and C. Lee Giles "Recommending Web Documents Based on User Preferences," <http://citeseer.nj.nec.com/cache/papers/cs/10692/http:zSzzSzwww.cs.umbc.eduzSzanzSzsigir99-reczSzpaperszSzglover_e.pdf/glover99recommending.pdf>(13 Nov. 2003).

註 46：同註 19，頁 14-15。

註 47：同註 45。

註 48：陳鳳蘭，「國內自行研發 Openfind 躍居全球最大搜尋引擎」，91 年 7 月，<http://www.google.com.tw/search? q=cache:Hdj3VLiJo_UJ:www.cs.ccu.edu.tw/course/2002-07-01__01.html+%E6%8E%92%E5%BA%8F*%E6%90%9C%E5%B0%8B%E5%BC%95%E6%93%8E&hl=zh-TW&ie=UTF-8> (2 Dec. 2003).