網際網路資料搜尋

唐雨文

靜宜大學西班牙語文學系兼任講師

ywtang@pu.edu.tw

壹、前言

隨著資訊科技的日新月益，網際網路[1]日益普及，人類在步入廿一世紀的同時，也正在進行一項資訊史上最鉅大的工程：「建構一個無遠弗屆虛擬的資訊新世界」[2]。在這項浩大的工程之中，人類持續不斷地將現實世界裡自古到今各式各樣的訊息（Information），諸如書刊、雜誌、報紙、論文、古籍、歷史文件、公文、書信以至於聲音、圖片、影像、動畫等，都將之數位化（Digitalization）處理，公佈在網際網路上。簡而言之，這個不斷建立數位化資訊的活動，已經在整個網際網路上累積建構出一個前所未有的龐大資料庫。

然而，這個在資訊氾濫的數位時代裡，資訊的篩選遠比資料的取得困難，目前人類面臨最根本的問題之一就是：「如何在浩翰如海、廣大無垠的虛擬資訊世界裡，有效、快速並正確地尋得所需的資訊」。例如，尋找有關 1982 年諾貝爾文學獎得主哥倫比亞作家馬奎斯（Gabriel García Márquez）的文學評論、阿根廷共和國在 2001 年 6 月暴動德拉魯亞（Fernando de la Rúa）總統下台杜哈德（Eduardo Duhalde）總統繼任後的政經情況、北美自由貿易區本年度的貿易總額、行為主義（Behaviorism）在語言教學上的運用等正確的相關資料。一般而

---

[1] 網際網路是一種電腦網路，以 TCP/IP（Transmission Control Protocol / Internet Protocol: 傳輸控制協定/網際網路協定，最初是由美國國防部高階研究計畫署開發出來。是一種非常通用的網路通訊協定，目前乙太網路、區域網路、個人電腦、大型主機都支援此協定）通訊協定互相連線，形成的虛擬網路，用以交換信息、傳遞資料與共享各類資源。網際網路的英文是 Internet，最早是美國國防部為了國防上的需要而建立的網路，希望能透過電腦網路，有效率的傳達訊息，並且希望一旦戰爭爆發，這些電腦連線不易被切斷。後來這個網路開始開放給一些大學和學術研究機構使用，並且與世界各國連線，慢慢地形成今天的全球網際網路。只要我們的電腦與任何一部 Internet 上的主機連線，那我們也就是 Internet 上的一分子了。有別於 Internet，WWW（World Wide Web，全球資訊系統）是 1989 年 3 月，位於瑞士日內瓦的「歐洲粒子物理實驗室」（CERN, European Laboratory for Particle Physics）的 Mr. Tim Berners-Lee 所提出，以 Internet 為基礎發展出來的分散式超媒體資訊系統，它與 BBS、Gopher、News、FTP、Archie 等同為 Internet 上的資源服務；不同的是，它提供了圖形使用者介面（GUI, Graphic User Interface），以文字、圖形、聲音、影像等多媒體的表達方式，結合超連結（Hyperlink）的觀念，讓使用者輕易取得網路上的各種資源。

[2] 自 1969 年美國國防部設立 ARPANET（Advanced Research Projects Agency Network）以來，1990 年網際網路（Internet）上提供服務的網路伺服器（Web Server）約 30 萬台，1992 年約增至 50 萬台；到了 1997 年 7 月，則大幅提升至 2 千萬台，使用人數超過 1 億 9 千萬人。2001 年 7 月已有 1 億 2 千萬台終端電腦連上 Internet，2001 年 3 月估計共有 4 億 3 千萬人上網。根據英國著名的 Netcraft 公司（http://www.netcraft.com）的統計，2003 年元月連上網際網路的網路伺服器已經超過 3 千 5 百萬台，網頁（Web Pages）更超過 8 億個，而且數目還在不斷增加中。有關 Internet 發展的歷史請參照：William H. Forrester, 2000, *The Online Searcher's Companion*, Lodon: Library Association Publishing, pp. 116-123 及 Mann, Chris and Fiona Stewart, 2000, *Internet Communication and Qualitative Research: A Handbook for Researching Online*, Lodon: SAGE Publications, pp. 1-10.

言，提供此類查詢服務的資訊軟體或網路服務系統，我們稱之為「搜尋引擎」（Search Engines）。當然，我們也可以直接進入相關的網站尋找資料而不透過搜尋引擎；此時，網路瀏覽器（Web browser）中「我的最愛」（Internet Explorer 中的「Favorites」）或「書籤」（Netscape 中的「Bookmarks」）的就顯得格外的方便與重要。

不同的搜尋引擎有不同的查詢語法（Query Syntax），但網路搜詢服務經過多年的統整與發展，查詢語法已日趨統一，其間存在的差異已經越來越小。本文的目的即在介紹搜尋語法的一般規則、網路資料的蒐尋策略、資料的驗證以及目前網路上常見的幾個蒐尋引擎。然而，網際網路浩瀚無垠，成長快速，各類蒐尋引擎的數量可謂汗牛充棟，不一而足，決非筆者三言兩語所能道盡。故在此僅能略述筆者從事教學與研究工作時常用的網路資料搜集策略與工具，期能透過心得之分享，達到拋磚引玉之功效，使網際網路成為學習與研究者的實用工具之一。

貳、資料搜尋及查詢語法

在網際網路的世界裡「分類主題」（Subject Catalogues）和「目錄服務」（Directory Services）是最簡單的資料搜尋方式。「分類主題」經由主動登錄和人為篩選的方式，蒐集數以萬計的網站資料，整理成為主題式的分類目錄，讓使用者可以逐層瀏覽，直到找到符合需求的網站為止。「目錄服務」則是一種資料儲存的結構與擷取的協定，為一特殊的資料庫，其中的資料屬性特殊，大部分屬於鮮少更動的靜態資料（如政府機關的地址、電話等）。目錄服務以樹狀結構（Tree Structure）將這些資料加以組織並儲存起來，提供一個快速的搜尋機制方便使用者從中獲得所需的資訊。在日常生活中，電話簿就是目錄服務的一種；而網路中的 Domain Name Service （DNS）同樣也是一種目錄服務，它和一般的資料庫最大的不同在於查詢的速度。

網際網路上提供分類主題或目錄服務的各個入口網站通常會把各式各樣的網站加以分析歸類，建立目錄，瀏覽者依據目標網站可能的類別，點選適當的目錄項目，層層展開，直至找到目標網站為止。此種搜尋方式適合於要找總類資訊而沒有特定目標網站的搜尋者，以雅虎奇摩（http://tw.yahoo.com/）為例，若想瀏覽行政院所屬單位的網站，可以從「政治行政＞組織機構＞政府機關＞中央政府＞行政院＞」進入尋找。此種搜尋方式，必須先預測目標網站的類別，加上每個入口網站的目錄分類方式不盡相同，同一類型網站在不同的入口常被歸入不同的分類目錄中，甚至同一入口的分類也未必夠明確劃分（如在雅虎奇摩中，「文學類」網站是歸類在「藝術文化」、「社會人文」還是「休閒天地」？），所以往往要經過幾番嘗試，才能覓得其門而入。

若要找尋特定內容的資料，「分類主題」或「目錄服務」因為必須層層點選瀏覽，所以比較沒有效率。此時，必須透過「搜尋引擎」來加速搜尋工作的進行。分類主題和蒐尋引擎的主要差異在於，分類主題利用登錄或電腦輔助方式蒐集網站資料，再由人工加以分類，並由人工撰寫或自動摘要方式提供網站說明；而搜尋引擎利用電腦自（主）動到網路上蒐集網頁資料，加以分析整理，製成索引資料庫，並以自動摘要方式，提供網頁說明。由於分類主題經由人工檢視與過慮，資料品質較為一致，查詢結果的資料筆數在可以控制的範圍之內，適合對某一主題有概略的瞭解或是沒有確實的資訊需求，只是想藉網路瀏覽掌握新知的人。而搜尋引擎由電腦程式依預設的標準自動判斷搜尋的要求，查詢結果的資料品質良莠不易掌握，資料數量也常多到無法控制，因此比較適合查詢有明確定義的資料，能以關鍵字或片語組合來描述及希望由各種不同來源搜尋完整資訊的人。

一般來說，搜尋引擎的用法十分簡單，直接在搜尋視窗內填入要搜尋的「關鍵字」（Keyword）[3]即可。但是，由於單一關鍵字搜尋的結果往往動輒千萬筆，所以必須增加關鍵字或加入適當的「查尋條件」（Query Condition），來限制搜尋結果的範圍，加速達到搜尋的目的。如同分類主題或目錄服務各有各的索引分類標準一樣，各種搜尋引擎也有自己專屬的「查尋語法」（Query Syntax），只不過其中的差異並不大。在介紹網路資料搜尋策略、網路資料驗證及常用的幾個搜尋引擎之前，在此先簡介一般通用的查尋語法。

一、使用「空格」：

執行限定條件查詢時，若在數個關鍵字之間加入「空格」，通常可以尋找到更為精準的資料。例如，在雅虎奇摩中要搜尋有關「馬奎斯」的資料，若輸入關鍵字「馬奎斯」得到的搜尋結果有 963 筆，若逐筆進入檢視，將會耗費不少時間，此時就必須善用「空格」來執行更準確的搜尋。如果只是想找有關馬奎斯生平的資料，則可以輸入「馬奎斯 生平」（中間以空格分開）做為檢索的關鍵字，搜尋結果大幅縮小為 31 筆資料，更為精準。

二、使用「" "」（引號）

引號內通常是一個短句，而不只是一個關鍵字而已。搜尋的結果將會包含跟引號內完全一樣的句子（exact phrase）[4]。以雅虎（http://www.yahoo.com）為例，要查詢「SARS」（嚴重急性呼吸道症候群）的相關資訊，若直接輸入「SARS」，搜尋結果得到 519,000 筆資料，但若多加一個字並用引號括起來「"SARS Symptons"」查詢，得到 2,920 筆資料；若再多加幾個字「"SARS Symptons and treatment"」則僅得 2 筆資料。可見在搜尋條件中妥善運用引號，可以有效地限制搜尋的範圍，使搜尋的結果更加精確。所以建議讀者養成在搜尋條件使用引號的習慣。但要注意太長的句子通常會造成搜尋失敗，找不到資料。

三、使用邏輯字元「+」（加號）及「-」（減號）：

「+」號的意義是搜尋的結果一定要包含該關鍵字，而「-」號的意義則是搜尋的結果一定不可以包含該關鍵字。同樣以雅虎奇摩為例，要搜尋「精神分析」與「文學」，可輸入「+精神分析 +文學」，搜尋結果得到 1,490 筆資料。若想排除其中含有「女性」的資料，則可輸入「+精神分析 +文學 -女性」，得到 886 筆資料，範圍較為縮小。但要注意查詢語法的邏輯，若輸入「+精神分析 +女性文學 -女性」，會因為邏輯矛盾而找不到任何資料。

四、使用布林運算法（Boolean operation）[5]語法查詢：

布林運算語法可以直接採用「AND」、「OR」、「NOT」三個運算子（Operators），也可依序使用「 & 」、「 | 」、「 ! 」三個符號來代替。「AND」代表查詢的結果是兩個（或數個）關鍵字的「交集」（如圖一），「OR」是關鍵字的「聯集」（如圖二），而「NOT」則關鍵字的「差集」，即第一個關鍵字的查詢結果除去第二個關鍵字的查詢結果（如圖三）。例如，若想要查詢有關「馬奎斯」與「百年孤寂」的網站，可以輸入「馬奎斯 OR 百年

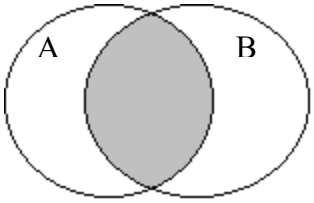
---

[3] 依據 Richard K. Belew 的定義（Richard K. Belew, 2000, *Finding Out About: A Cognitive Perspective on Search Engine Technology and the WWW*, Cambridge: Cambridge University Press, p. 10），關鍵字（Keyword）是語言的核心（atoms），通常是字（Word）、字串（Piece of Words）或句子（phrases），用來突顯文件的主題或內容。關鍵字是文件主題焦點的中樞，是連接使用者對資訊的需求和文件主題焦點的橋樑，文件透過關鍵字才容易被尋得。關鍵字可從兩個層面探討：1.在文件中如何建立關鍵字及 2.關鍵字如何被使用者運用。

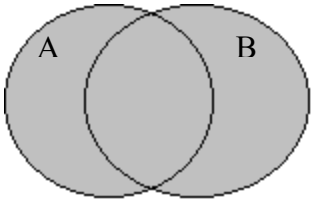
[4] 有些搜尋引擎提供尋條件選項，將"All the word"換成"Exact phrase"，其搜尋結果與加上引號相同。

[5] 布林運算法又稱為布林代數（Boolean Algebra），是英國數學家喬治布林（George Boolean）（1815-1864）於 1854 年發表處理數位邏輯的代數運算式。布林代數與一般代數不同，它只適用於處理 0 與 1 的邏輯數值。

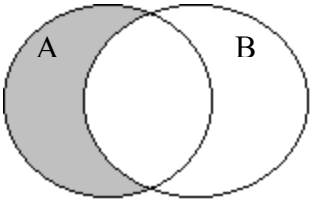
孤寂」，在雅虎奇摩中得到 1,830 筆資料。若不希望包含「諾貝爾」，那麼可以輸入「（馬奎斯 OR 百年孤寂） NOT 諾貝爾」，得到 2 筆資料。值得一提的是，由於布林運算法要鍵入運算子，多少會造成使用上的不便，因此諸如雅虎等搜尋引擎已捨棄不用，而以「+」及「-」代之。[6]



圖一：A AND B



圖二：A OR B



圖三：A NOT B

雖然使用布林運算子可以提高搜尋的準確度，但是不同的搜尋引擎對相同的布林運算式常有不同的解釋。有些搜尋引擎是從左向右解讀，有些則是由右向左解讀；有的搜尋引擎的 AND 優先順序大於 OR，有些則相反。也就是說，不同的搜尋引擎對完全相同的搜尋條件會產生不同的搜尋結果。因此依據筆者的經驗，除非其他的搜尋方式都告失敗，儘量不要使用布林運算子；若一定要使用，則應加入適當的括號（如上例）以避免混淆。

五、使用網址查詢：

某些搜尋引擎提供網址查詢功能（如雅虎奇摩）。例如，想要查詢有關兒童文學的資料，而且只想在學術網路上尋找，那麼可以輸入「兒童文學 edu.tw」（以空格分開）。或者想查詢有關歷史的資料而且是政府相關機關的網站，則可以輸入「歷史 org.tw」。

六、使用萬用字元（Wild card symbol）「*」、「?」或「%」：

萬用字元查詢也稱為「切截查詢」（Truncation Searching），是利用萬用字元來幫助我們克服查詢時所面臨的關鍵字單複數、字形變化、縮寫或略語的問題，來查詢同一個字根但不同變形的字。萬用字元「*」可以代表「任何」字元（不限字元數），「?」則代表「任何單一」字元。當我們不能確定關鍵字如何拼寫時，萬用字元就變得十分有用。例如，「bar*」代表 bar, barbarian, barbican 等；而「barc?」則代表 barcarole, barchan 等。又如，輸入「國*科*會」，在 Gais 網路資訊系統（http://gais.cs.ccu.edu.tw/）中不僅可以找到「國科會」，還可以找到「行政院國家科學委員會」、「中華民國管理科學學會」、「中華民國婦產科學會」、「中華民國產業科技發展協進會」等相關資訊。有些搜尋引擎則以「%」代替萬用字元「*」與「?」，使用的方法相同。

參、網路資料搜尋策略

現今網路世界的軟硬體設備發展已日益成熟，個人電腦連接上網十分普遍，甚至已經邁入全方位的寬頻時代。網路人人會上，但入寶山卻可能常常徒勞而返。要在浩瀚無垠廣闊無邊的網路虛擬世界中找到想要的資料，是需要一些技巧和策略的。以下列舉筆者多年來在網路上搜尋資料所慣用的策略與心得，提供給大家參考。

---

[6] 有關布林運算法，請參照：William H. Forrester, 2000, *The Online Searcher´s Companion*, Lodon: Library Association Publishing, pp. 26-31 及 Alison Cooke, 2001, *A Guide to Finding Quality Information on the Internet: selectoin and evaluation strategies*, Lodon: Library Asociation Publishing, pp. 20-22.

一、確定搜尋的目的：

在網路上搜尋特定的資料與在網路上瀏覽閒逛不同，後者可以漫無目標，隨興所至，自我放逐，是以「休閒」為目的；而前者是為了因應擷取資訊的需要而從事的活動，必須以「尋獲」為目的。然而在網路上漫遊成習的人，很容易在搜尋的特定資料時分心。在搜尋活動開始之前，搜尋者應該明確地知道搜尋的目的為何以及要尋找的標的為何，尤其應該判斷所要搜尋的資料是否有其確切的意義與需要，否則搜尋後的資料遭到閒置，不能加以運用，徒然浪費了寶費的時間和金錢。

二、預定搜尋時限：

每次搜尋的結果必須靠搜尋者自己親身的閱讀，以判斷是否準確可用。而一個人長期處在電腦螢光幕及鍵盤前面，眼睛與身體均很容易疲勞，注意力因而降低，搜尋的效率因而減緩。所以應該事先預定搜尋的時限，時限一到，不論資料是否尋獲，都應停止搜尋活動的進行，做一下短暫的休息。我們可以利用這一段休息的時間思考並檢討搜尋的過程，修改關鍵字及查詢的語法，以利後續搜尋活動的進行。

三、善用搜尋引擎，選擇適當的關鍵字和查詢語法：

良好的網路搜尋是絕對離不開搜尋引擎的。搜尋引擎能夠幫助我們在最短的時間內找到最大量的資料，而搜尋的成功與否通常取決於正確的關鍵字組合和適當的查詢語法。關鍵字可能是專有名詞、技術的正式名稱、有關的描述文字、出現率極高的詞彙、作者、書名、機構名稱等。使用關鍵字配合搜尋引擎在網路上尋找，比較容易得到準確有效的結果。若要搜尋的資料是屬於陌生的領域，對其專有名詞尚不熟悉，此時可以儘量想一些較接近的字彙，先去做初步的搜尋，在搜尋的結果內可能會找到幾篇主題類似或相符的文章。接下來，便可以利用這些文章的內容找到更適切的關鍵字，配合查詢語法，縮小搜尋的範圍，以取得更精確的搜尋結果。這種「無中生有」的搜尋方式，常常能成功地「開疆闢土」，「小兵立大功」。另外，值得一提的是若關鍵字是外來語或是外文中譯，則必須以原文為關鍵字才容易獲得正確的資料。如，Tourette's Syndrome 中文譯為「托氏症」，但若以「托氏症」為關鍵字，則可能搜尋不到任何資料；而以原文為關鍵字則可以成功找到資料。

不同種類的資料應使用不同的尋引擎，不同語言的資料也最好使用支援該語言的搜尋引擎。例如，若是要找的是西班牙文的資料，用國內的搜尋引擎就不容易找得到。通常著名的搜尋引擎都有支援不同語言的版本，如雅虎就有中文、英文、西班牙文、法文、德文等多種語言專屬的版本。

四、適度修正查詢策略：

初步的查詢可能產生下列三種結果：1）查詢到的資料完全不符合需求、2）查詢到的資料數量多但符合需求的少以及 3）查詢到的資料數量少但符合需求的多。以上這三種情況都需要適度地修正查詢策略，再做進一步的查詢，使查詢資料的「質」與「量」都符合需求。如果查詢的結果是零或是完全不符合需求，應檢查關鍵字是否拼寫正確、更換關鍵字或更換的搜尋引擎再查。若是資料數量很多，但符合需求的少，則應縮小查詢的範圍。大部份的搜尋引擎都有提供「收斂查詢」（Sub-Searching）的功能，可以根據第一次查詢的結果再輸入查詢條件，以進行第二次查詢，使結果更為精確。若是資料數量很少，但符合需求的多，則有兩種可能：1）可能漏掉了某些重要的關鍵字，應該修正查詢語法，擴大查詢的範圍；2）該種資料在網路上的實際數量有限，此時應採取傳統的資料搜集方式，如求助於圖書館或師長等。

五、專心致志，心無旁務：

網路虛擬的世界花花綠綠，容易讓人瀏連忘返。在資料搜尋過程中，瀏覽器自動碰出千奇百怪的「彈出視窗」（Pop-up Windows）或在搜尋引擎上各式各

樣的廣告常會吸引並轉移我們的注意力，使我們在不知不覺中離開了原有的搜尋路徑，偏離了搜尋的主題。所以網路資料搜尋如同其他工作一樣，都必須專心致志，心無旁務。邊聽網路即時廣播，邊收發電子郵件，又同時在網路上閒聊的資料搜尋者，常常就是「入寶山卻空手而返」的人。

六、善用「我的最愛」或「書籤」

有些寶貴的資料，往往需要費好多功夫才能在網路上找到，此時可以將含有該資料的網址儲存在瀏覽器的「我的最愛」或書籤中，以便日後使用。要注意的是，應該將這些網址做適當的組織歸類，建立各個領域不同主題的資料匣（如「文學」、「語言學」、「拉丁美洲研究」等），日後使用時便可以節省搜尋的時間。也可以把這些組織好的「我的最愛」或「書籤」當作資源交換的載體，當別人有需要的時候，可以將整個領域資料匣拷貝過去即可，十分方便。

七、提升語言能力，增進閱讀速度：

網路資料搜尋的結果，不論是螢幕瀏覽或是列印成書面資料，都必須經過搜尋者以閱讀的方式加以消化，才能判定資料的可用性。網路上包含用各種語言寫成的資料，而許多具有實際價值的資料都是以外文撰寫而成，因此搜尋者的外語能力以及閱讀的速度就顯得十分重要。當然，此時或許可以運用「網頁翻譯器」（Web Page Translator）或「網路護照」（Web Passport）等工具來協助瞭解資料的內容，但是該等線上翻譯服務通常都只是做機械式的逐字翻譯，並不會做意義上的判斷，所以翻譯出來的文章可讀性不高。因此，提升自己的語言及閱讀能力才是根本的解決之道。[7]

肆、網路資料驗證

搜尋引擎是最常被用來尋找網路資料的工具，它的索引標記通常都是由系統自動維護（新增、刪除、修改），能對網路內容的更迭及時反應及更新，因而它所包含的資料範圍也十分廣泛。例如，截至 2002 年 10 月為止，著名的搜尋引擎 Google 至少已經囊括了 20 億筆的網頁索引，數量十分龐大驚人。但是水能載舟亦能覆舟，任何人於任何時間都可以在網路上公佈任何資訊，而搜尋引擎幾乎對所有的網頁都照單全收，對其中的資料缺乏「品質鑑別」及「內容控管」的機制。索引的網頁太過龐大，會造成搜尋結果資料筆數太多，內容差異太大，無法有效地達到搜尋的目的。這個缺點可以透過前述使用適當的查詢條件來獲得改善。搜尋引擎的另一個限制是所能搜尋的資料格式有限，不符合所支援的格式者將無法被順利找到。例如，網路上最常見的資料格式是 HTML 格式[8]、TXT 純文字格式及 PDF 格式[9]，若網頁資料不是用這些格式撰寫的就不容

---

[7] 有關網路資料搜尋其他的技巧可參閱：Alfred Glossbrenner and Emily Glossbrenner, 1998, *Search engines for the World Wide Web*, Berkeley: Peachpit Press, pp. 11-29; Mary McGuire et al., 2000, *The Internet handbook for writers, researchers, and journalists*, New York: Guilford Press, pp. 70-71; Cynthia B Leshin, 1998, *Student Resource Guide to the Internet: Student Success Online*, New Jersey: Prentice Hall, pp. 79-85 及 Chowdhury, G. G and Sudatta Chowdhury, 2001, *Information Sources and Searching on the World Wide Web*, Lodon: Library Associaton Publishing, 10-14.

[8] HTML（HyperText Markup Language），即「超文本標記語言」，是一種簡單的文本格式控制語言，用於 WWW（全球資訊網）網站的建構，為編寫網頁的主要語言，正因它可將文字、聲音、圖片等不同屬性的物件整合在一起，成為 HTML 超文件，並且能展現每一種物件的原有特性，因而成為當今網路上的主流語言；也就是說，現今大多數全球資訊網上的網站都是以 HTML 建置的。

[9] PDF（PortableDocument Format，即「電子文件交換格式」，是由 Adobe 公司開發出來的文字格式。PDF 的開發主要是為了解決不同平台間電子文件格式交換所引發的問題，即文件收受者若是沒有安裝建立該文件的應用程式，就無法開啟該文件。然而，只要有 Acrobat Reader，任何人都可以打開 PDF 文件，並且不會因為平台、軟體與版本之間相容性的限制，而損害文

易被尋獲，數位影音資料就是一例。依據 Alison Cooke 在 2001 年的估計，網際網上的「隱形網站」（Invisible web）是搜尋引擎所能搜尋到的「公開索引網站」（Publicly indexable web）的 500 倍[10]。若以先前所提 Google 所能搜尋到的 20 億筆網頁為公開索引網站，則隱形網站高達 10,000 億之多；也就是說，我們能尋得而加以運用的網路資料，只有實際的五百分之一而已。這是搜尋引擎的另一個限制[11]。

雖然如此，我們能透過搜尋引擎找到的資料仍然是十分驚人的，如何對這些龐大的資料加以「鑑別」與「驗證」是資料搜尋者最重要的工作之一。一般來說，網際網路上的網站可區分為官方組織網站（Organizational Sites）、學術網站（Academic Webs）、商業網站（Commercial Webs）及個人網站（Personal Home Page）。通常官方組織網站或學術網站較具公信力，在此取得的資料也較為正確；從私人或商業網站上取得的資料則因資料錯誤的機率很高，因此要特別多做查證，小心判斷。網路資料若與書面資料（如書籍、期刊）有所矛盾，通常要以書面資料為準。

然而，近年來似乎越來越難區分來自官方、學術、商業或私人網站的資料，所有的資料看起來都有一定程度的相似性，卻又或多或少有些許差異。例如，九一一事件後，美國總統布希對打擊恐怖主義的宣誓，就可以在網路新聞上找到數十個版本。因此網路資料的評鑑與查證就成為網路資料運用者的一大挑戰。當我們找到一筆初步認為可資運用的資料時，我們首先要確定的是它的可信度（Reliability）。最好的驗證方式是運用離線資料（Offine sources）[12]查證，或向另一個網路資料查詢。以下綜合 Mary McGuir[13]的觀點，整理出六個確認網路資料可信度的方法：

一、評鑑發佈資料的組織或個人在該領域的權威性：

「查證誰是資料的公佈者，並判斷其在該領域的權威性」是一件非常重要的工作，因為任何人在任何時間都可以從任何地點在網路上公佈資料，而這個資料可能只是作者的誤解或偏見。我們可以利用幾個線索推測資料的出處，例如從網站上的電話或電子郵件信箱推斷該網站的贊助者，而該網站的資料可能與這位贊助者的利益有關，而影響資料的公正性。我們也可以從該站的網址查證資料的公正性。例如，若一個網站的網址是 http://www.harvard.edu/library/~bmaritin，則我們可以知道該網站是架設在美國哈佛大學（Harvard University）的網路伺服器下，而「~」符號通常代表的是個人，所以/library/~bmaritin 所指的應該是該校圖書館有關的人員，我們可以藉由電子郵件、以「bmaritin」為關鍵字至 WhoWhere （http://www.whowhere.com）查詢或直接以電話向該圖書館查證作者的姓名、身份、著作等相關背景資料，判斷其在所屬領域中的地位，以做為評估資料可信度的依據。要強調的是，很難決定建位在官方網站下個人網頁的公正性，因為該網頁可能只是官方網站提供給個人的一個網頁空間，只是表達個人觀點和意見的地方，其言論通常不代表該官方網站的立場。

二、確認資料網站的擁有者：

有時，要查證誰是網站的贊助者並不容易，但是我們可以透過 WHOIS 系統

件的格式，也不會因為軟體或印表機的限制而無法正確的列印文件。基於這些優點，近年來 PDF 格式被網頁製作者大量使用，以方便瀏覽者下載文件。

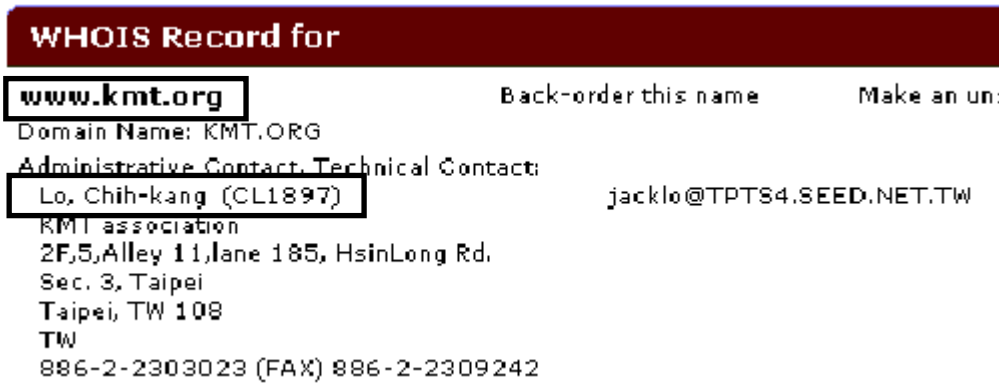
[10] Alison Cooke,2001, *A Guide to Finding Quality Information on the Internet: selectoin and evaluation strategies*, Lodon: Library Asociation Publishing, pp. 25-28.

[11] 有關搜尋引擎的利與弊請參閱：Alison Cooke,2001, *A Guide to Finding Quality Information on the Internet: selectoin and evaluation strategies*, Lodon: Library Asociation Publishing, pp. 25-28.

[12] 離線資料指的是網路以外的資料，如報刊、雜誌、字典、書籍、百科全書等平面媒體所記載的資料。

[13] McGuire, Mary et al., 2000, *The Internet handbook for writers, researchers, and journalists*, New York: Guilford Press, pp. 118-120.

查詢誰是該網站「網域名稱」（Domain Name）[14]的擁有（註冊）者（台灣網路資料中心的 WHOIS 查詢系統網址是 http://www.whois.twnic.net/；美國的 Whois 查詢系統網址是 http://www.whois.net/）。由於「網域名稱」是唯一的，又開放給社會大眾自由申請，所以有許多看似官方的網址卻是由私人所擁有的。舉一個例子，http://www.kmt.org 這個看似中國國民黨的網址（KMT 是中國國民黨的英文正式縮寫名稱），但實際上卻是由私人所擁有。我們可以透過 WHOIS 系統查出該網址的主人是一位地址登記在台北市興隆路名叫 Jack Chih-Kang Lo 的台灣人（如圖四）。也就是說在該網站上找到有關中國國民黨的資料，可能並不是由該黨發佈的，其可信度值得懷疑。在此狀況下，我們可以依照記錄上的電子郵件或連絡電話直接向 Mr. Lo 本人做進一步的查證。



圖四：WHOIS 查詢範例

三、檢查資料的格式：

檢查資料的格式以判斷它是否經過編輯。一份沒有經過編輯的資料，通常會有較多的拼字、文法或格式上的錯誤，內容的可靠性也比較受到質疑。而一份正式的文件或學術論文應該都會經過嚴謹的編輯過程，錯誤會因此減到最低。所以我們可以籍由檢查資料編輯、校訂的情形，判斷該資料的可信度。

四、檢驗資料的效期：

一份有時間性的資料，即使資料本身確實可信，但是一旦過了期限，就可能變得毫無意義，而不堪引用了。例如，用 1960 年代的人口數據來預測 2010 年的出生率，或用三個月前的美元匯率計算本月的進出口總額，都會造成錯誤的推算結果，直接影響研究的結論。因此，經過搜尋取得的資料，應該確認該資料建立的日期，以做為判斷該資料價值的依據之一。

五、分析資料的來源背景：

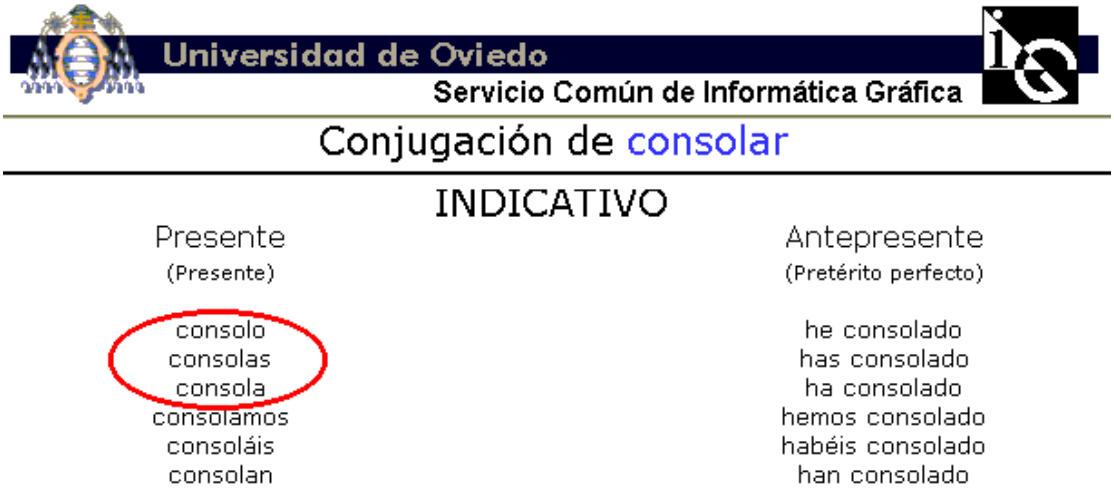
將資料所在的網址每次去掉一層，以分析該網址的來源背景，可以幫助我們確認該資料的可信度。例如，我們利用搜尋引擎找到一筆有關陳水扁總統於民國九十一年八月召開「國安會議」的資料，網址為 http://www.gio.gov.tw/info/

[14] 在網路上真正用來表示網路位址的是 IP Address（Internet Protocol Address，即數字表示的網址），是由四個小於 256 的數字所組成。如靜宜大學的 WWW 全球資訊網主機的 IP 是 104.128.48.1。因為 IP 位址是由一大串的數字組成所以很難記憶，因此我們常利用一組較有意義的英文縮寫來代表 IP Address，這就是所謂的「網域名稱」，其組成可分成四部份：主機名稱.機構名稱.機構類別.地區名稱。主機名稱指的是此主機所提供的服務種類，例如 www 為全球資訊網、bbs 為電子佈告欄等。機構名稱為此主機所代表的公司行號或機關的英文簡稱，例如 pu 為靜宜大學、nsysu 為中山大學、ncnu 為暨南大學。機構類別是此主機代表的公司行號或機關的種類代號，例如 edu 為教育學術單位、com 為商業網站、gov 為政府單位、org 為財團法人或基金會等組織、net 為網路管理機構或網路服務提供者、mil 是軍事網路。地區名稱是此主機的所在地區簡稱，例如 tw 為台灣、jp 為日本、此部份省略時表示為美國。例如，www.nsysu.edu.tw 表示在台灣的學術單位中山大學所提供 www 服務的主機。www.microsoft.com 表示在美國的微軟公司所提供 www 服務的主機。

noticia97/2002/24/p1.htm，我們可以把網址從右邊逐一去掉，檢查該資料的上層網址，如此層層檢驗，最後剩下 http://www.gio.gov.tw，是中華民國行政院新聞局的官方網站，顯示該資料的可信度很高。要注意的是，使用這個方法分析資料來源背景時必須一次只去掉網址的最右邊一層，直至找到來源為止，因為有些網站只是利用其它網站的空間設置網頁，而不隸屬於其下，一次就刪除至最底層，可能會得到不正確的來源背景資料。

六、比對離線資料來源：

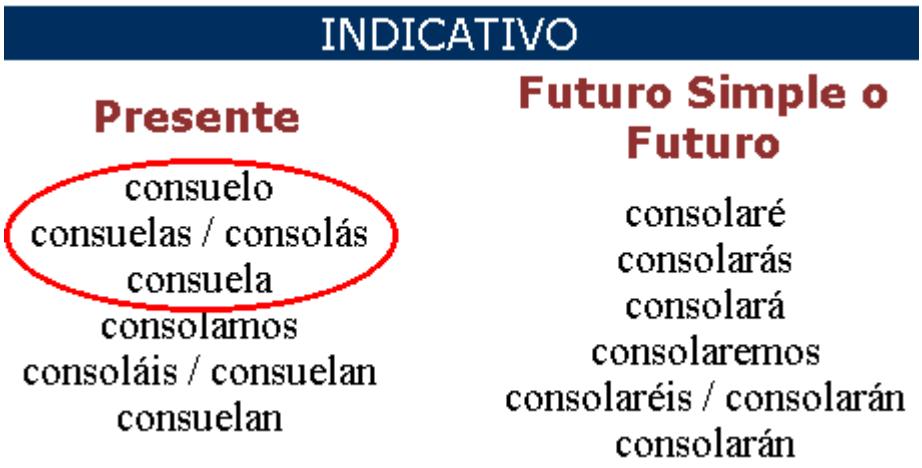
網路資料必須進一步與離線（傳統）資料加以比對檢驗，才能確認其可靠性。尤其是撰寫學術性論文或研究報告所需的資料，應該以離線資料做為主要的資料來源。隨便使用未經查證的網路資料，容易影響論述的客觀性和正確性，造成研究結果的偏頗與謬誤。舉例來說，西班牙文的動詞變化十分複雜，除了專屬的紙本動詞變化字典（Dictionary of Spanish Verb Conjugation）外，亦有著名的學術研究單位在網路上建置西班牙文動詞變化查詢資料庫，提供使用者線上即時查詢，十分方便好用。西班牙歐比耶多大學（Oviedo University，成立於 1608 年）提供的西班牙文動詞變化線上資料庫是西班牙文學習者常用的學習工具。我們進入該資料庫（http://tradu.scig.uniovi.es/conjuga.html）查詢西班牙文動詞「consolar」（安慰）的式時變化，得到圖五的結果：



圖五：利用西班牙 Oviedo 大學動詞資料庫查詢 consolar

查詢的結果顯示，「consolar」是一個規則動詞，其陳述式現在時的各人稱變化為：consolo, consolas, consola ...；對於這個結果，西班牙文初學習者也許不會質疑其正確性。但是我們若再利用西班牙皇家學院[15]（Royal Spanish Academy）的線上動詞變化字典查詢（http://dlc.rae.es/verba/），得到的結果如圖六：

[15] 西班牙皇家學院（Real Academia Española）成立於 1713 年，主要的任務在於統一並且維護西班牙語的純正、確切和輝煌（purity, propriety and splendor），並且隨著時代潮流的脈動，修正或新增西班牙語的詞彙、發音及語法。該學院所編的西班牙文字典（簡稱為 DRAE, Diccionario de la Real Academia Española，1726 年初版，目前為第 22 版）在學界享有權威的地位，是其它字典編纂的依據。



圖六：利用西班牙皇家學院動詞變化字典查詢 consolar

查詢的結果顯示，「consolar」是一個不規則動詞，其陳述式現在時的各人稱的變化為：consuelo, consuelas, consuela ...，與 Oviedo 大學的查詢結果不同。最後，我們查閱傳統紙本字典，發現西班牙皇家學院線上查詢的結果才是正確的。這個例子告訴我們，雖然是著名（權威）的學術機構（如 Oviedo 大學），但其網路上的資料也不能保證百分之百正確，必須進一步與傳統資料交互比對，檢驗確認。

另外，值得一提的是，網路資料若遇到有心的駭客，十分容易被篡改，原來應該是正確無誤的資料，也因此變得荒誕不經。所以使用網路資料從事學術研究工作，必須時時提醒自己：「懷疑、懷疑、再懷疑；查證、查證、再查證」。

伍、常用的搜尋引擎

前面已經提過，良好的網際網路搜尋是絕對離不開搜尋引擎的。然而，搜尋引擎為數眾多，經過時間潮流的統整，其功能及包含的資料的範圍已經沒有多大的差異，反而是其所運用的使用者界面（User´s Interface）設計的良窳，會直接影響到資料搜尋的進行。以下簡略介紹四個筆者認為設計優良且人氣鼎盛的搜尋引擎，供各位讀者參考：

一、雅虎（Yahoo!）

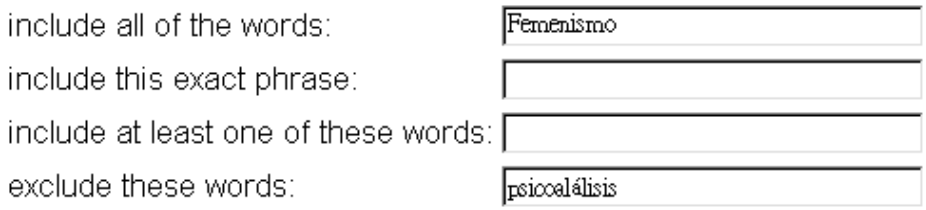
雅虎成立於 1995 年，由楊致遠先生及 Mr. David Filo 所創辦，為全球第一個提供網路導覽服務的網站，也是全球使用人口最多的網站，是居世界領導地位之一的入口網站。雅虎總部設在美國加州聖克拉克市（Santa Clara），網址為 http://www.yahoo.com/，目前在全球共有二十四個分網，使用者只要改變適當的區域名稱即可進入各分網。例如，台灣分網為 http://tw.yahoo.com/、西班牙分網為 http://es.yahoo.com/、墨西哥分網為 http://mx.yahoo.com/ 餘此類推。雅虎總站及各分網的搜尋方式有些許差異，簡介如后：

（一）雅虎總站 http://www.yahoo.com/

雅虎總站的特色是提供進階搜尋功能（Advenced Search），使用者不必記憶任何搜尋語法，只要直接把關鍵字（詞）鍵入進階搜尋窗格即可。雅虎總站的進階搜尋窗格有四個。第一個是「包含所有文字」（Include all of the words），搜尋的結果將包含所有的關鍵字，類似以布林運算子 AND 串連關鍵字或在關鍵字之前加上「+」號。第二個是「包含一模一樣的句子」（Include this exact phrase），搜尋的結果將包含與關鍵詞一模一樣的句子，類似在關鍵詞外加上引號。第三個是「至少包含一個關鍵詞」，搜尋結果為關鍵詞的聯集，類似以布林運算子 OR 串連關鍵字。第四個是「不包含這些關鍵字」，搜尋的

結果將不包含指定的關鍵字，類似以布林運算子 NOT 串連關鍵字或在關鍵字之前加上「-」號。

上述四個搜尋窗格可以混合使用，如要查詢「Femenismo」（女性主義）但不要包含「psicoalálisis」（精神分析）則可依圖七的方式輸入關鍵字，搜尋的結果得到 166 筆資料（如圖八）。

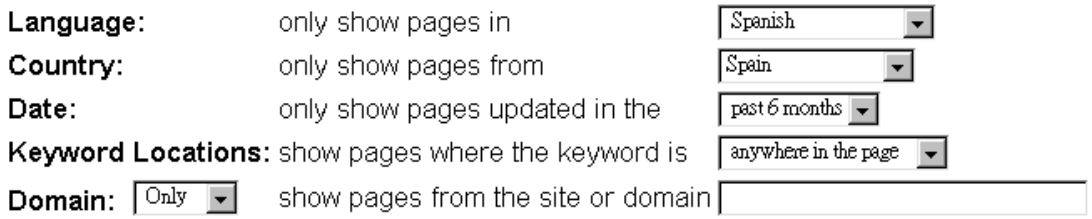


圖七：www.yahoo.com 進階查詢



圖八：進階查詢結果

雅虎總站也提供了「多樣選項」（More options），可以限定查詢資料的語言、地點、時間、關鍵字的位置等，能有效縮小搜尋的範圍，得到較精確的搜尋結果（如圖九）。



圖九：雅虎總站的「多樣選項」

另外，值得一提的是，現今網路各大搜尋引擎彼此之間大多都有建立合作關係，用以簡化搜尋過程，帶給搜尋者最大的便利。雅虎也不例外，當雅虎在本身的資料庫中找不到要求的資料時，它會自動連結到另一個著名的搜尋引擎也是雅虎的合作夥伴「Google」繼續尋找，而原先在雅虎內設定的查詢條件也會自動轉入 Google，十分方便[16]。

（二）雅虎奇摩 http://tw.yahoo.com

1997 年台灣華文搜尋引擎「奇摩站」（http://www.kimo.com.tw）成立，2000 年 11 月被雅虎購併，「雅虎奇摩」正式誕生，成為全世界第一大華文入口網站

[16] 這種同時自動連結至多家搜尋引擎的搜尋方式稱為「整合式搜尋引擎」（Meta-search Engines），只要在其查詢框內輸入關鍵字，就可以同時送到數個搜尋引擎去查詢，而得到各個搜尋引擎的查詢結果。大部份的整合式搜尋引擎並沒有建立獨立的資料庫，只是提供一個界面，將檢索條件送到各個搜尋引擎所建立的資料庫做查詢。著名的整合式搜尋引擎有：http://www.copernic.com 、 http://www.dogpile.com/ 、 http://profusion.com 、 http://www.infind.com、http://ixquick.com 及 http://www.metacrawler.com 等。其中 Copernic 可涵蓋搜尋引擎的數目達 300 餘個，堪稱整合式搜尋引擎之最。

。

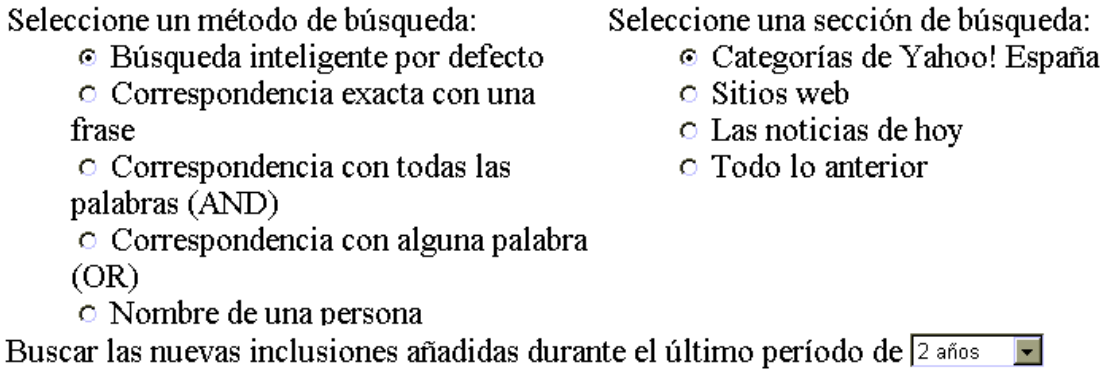
雅虎奇摩沒有提供類似雅虎總站的「進階查詢」功能，而是直接在搜尋窗格內鍵入查詢條件，可以點選「說明」查閱其查詢語法（如圖十）。該引擎搜尋的範圍以華文網站為主，適合尋找華文網路資料。



圖十：雅虎奇摩搜尋視窗

（三）雅虎西班牙 http://es.yahoo.com

提供類似雅虎總站的進階查詢功能，雖然搜尋視窗的使用者介面不同，但功能卻相同（如圖九）。此外，該引擎允許萬用字元查詢，還增加了「t:」及「u:」兩個查詢指令。「t:」表示只尋找文件的標題，「u:」則表示在文件的網址裡搜尋。其實這兩指令在雅虎總站中也有，只是將之以窗格方式選取（即圖九中的「Show pages where the keyword is」及「Show pages from the site or domain」兩個選項）。



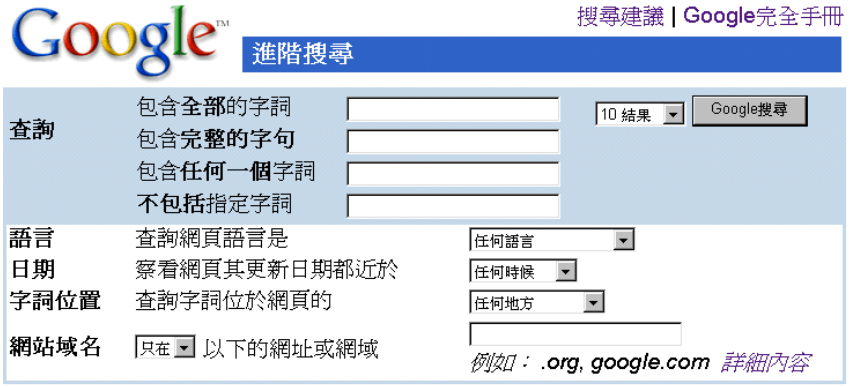
圖十一：雅虎西班牙的搜尋視窗

二、Google （http://www.google.com/）

Google[17]是另一個著名的網路搜尋引擎，由美國史丹佛大學（Standford University）兩位博士生 Larry Page 和 Sergy Brin 在 1998 年開發成功的第二代搜尋引擎。該搜尋引擎包含超過 20 億筆網頁資料，具創新、超強的演算法，以搜尋時間快速見長，每天提供的查詢服務超過上億次。與雅虎不同的是，Google 雖然也提供分類目錄，但分屬不同的網址（http://directory.google.com/）。這個分類目錄是以人工的方式篩選網頁，應用的檢索技術加以整理，再將結果依重要性排序呈現，是結合主觀的人工判斷和客觀的排序演算法所得的結果，所以口碑不錯。Google 搜尋引擎的介面相當簡潔、使用者介面也很人性化，提供「進階搜尋」、「使用偏好」和「語言選項」三個基本功能，可以直接限定查詢特定語言的資料，不必轉換至不同語言版本，使用上比雅虎方便。因其簡明的使用者介面設計及快捷的查詢速度，Google 也成為雅虎等其他多個搜尋引擎的協力

[17] Google 這個字是指 10 的 100 次方，即 1 的後加上 100 個零，代表天文數字。Larry 與 Sergy 就是用這個字表示網際網路資訊的浩瀚無邊（謝寶煖，民 89，頁 109）。

合作夥伴。



圖十二：Google 的進階搜尋

三、Ask Jeeves （http://www.ask.com/ , http://www.pregunta.com/）

Ask Jeeves 是一個「全文搜尋引擎」（Full Text Search Engine），西班牙文版本稱為 Pregunta。使用者可以直接輸入要問的問題，Ask Jeeves 會自動選取問題中的關鍵字，自動產生最適當的查詢語法，對於不明白如何使用查詢語法的使者而言，Ask Jeeves 是一個十分方便的選擇。



圖十三：Ask Jeeves 的查詢視窗

四、GAIS 網路資訊系統（http://gais.cs.ccu.edu.tw/）

GAIS（Global Area Information Servers）網路資訊系統是中正大學吳昇教授於 1995 年，鑑於歐美大部份的搜尋引擎無法正確搜尋亞洲國家雙位元組（Doble Byte）[18]編碼的資訊，為求達到多語言、無國界的檢索功能，而發展出之多用途（General-Purposed）、可調式（Scalable）的網路資源搜尋系統。



圖十四：GAIS 的查詢視窗

[18] 位元組是資訊的記憶單位，一個位元組等於 8 個位元（bits）。一個英文字母、數字或是特殊符號字元等於 1 個位元組，而一個中文字等於 2 個位元組，即「雙位元組」。歐美國家大部份搜尋引擎的設計只針對單位元組的字元，對於亞洲國家雙位元組的編碼則無法正確搜尋。

GAIS 支援五種查詢類別：簡單關鍵字查詢、片語查詢、布林運算式查詢、多項目查詢（包含使用「,」分隔號及使用「+」、「-」符號）及自然語言查詢，功能與前述搜尋語法大致相同。值得一提的是自然語查詢（Natural Language）允許直接輸入查詢語句，找出最適合的結果，輸出的結果將會以分數高低順序排列，類似 Ask Jeeves。例如，可以輸入口語化的問題：「阿根廷的經濟現況？」，會得到圖十五的 68 筆結果。



圖十五：GAIS 搜尋「阿根廷的經濟現況」的結果。

值得注意的是，所有支援自然語查詢的搜尋引擎都有一個共通的現象，就是找出來的資料常常不甚符合我們的需要，甚至毫無關聯。這是因為自然語查詢並不是真正瞭解我們要找的東西，而只是把字串中一些常用的字去除掉，如「嗎、了、的、得、呢」等虛字，然後再用利剩下來的字去做查詢的動作。胡嘉璽（民 89，頁 249）就建議儘量不要使用自然語言的方法來找資料，而且就算所使用的搜尋引擎有支援自然語言，最好也不要使用，因為搜尋的結果可能會得到一大堆無關的資料；這些資料會傳回的原因可能是因為自然語言中沒用的圾垃字所造成的。

陸、結語

有句話說：「網路非萬能，但是沒有網路卻萬萬不能」。也許有很多人並不這麼想，因為他們經常在網路上找不到資料，因而大聲疾呼「網路無用論」。其實，網路上並不是沒有他們所要的資料，而是他們不知道如何進行正確的搜尋。

本文原為九十一學年度上學期銜恩師靜宜大學西班牙語文學系系主任何國世博士之命，為本系碩士班同學簡介網際網路資料搜尋的方法及限制時所撰。因反應不惡，偶見同學間有影印相傳者，故而決定增修其中部份內容，藉國內學術網路上傳遞知識之重要通路－「暨南大學電子雜誌」一隅，對網際網路資料搜尋做一概略性的介紹，希望能對讀者有所助益。

最後，筆者仍要強調，由於網路資料具有「易得」與「易變」的雙重特性，資料的內容並無一定的標準，因此必須經過進一步的查證方能運用。故而學術性質的研究參考資料，還是應該以傳統的圖書館書目系統做為主要的資料搜尋方式。也就是說，網際網路並不適合做為學術研究的主要參考資料來源，而是僅能做為輔助性的參考資料搜集管道，從事學術研究時不應該過度依賴。已經習慣以網際網路做為研究主要參考資料來源的讀者們看完本文，是否已經開始對自己先前從網路下載的資料產生懷疑與焦慮？這是本文的潛在目的。

參考書目

Belew, Richard K., 2000, *Finding Out About: A Cognitive Perspective on Search Engine Technology and the WWW*, Cambridge: Cambridge University Press.

Casanova, Lourdes, 1998, *Internet para Profesores de Español*, Madird: Edelsa.
Chowdhury, G. G and Sudatta Chowdhury, 2001, *Information Sources and Searching on the World Wide Web*, Lodon: Library Associaton Publishing.
Cooke, Alison, 2001, *A Guide to Finding Quality Information on the Internet: selectoin and evaluation strategies*, Lodon: Library Asociation Publishing.
Forrester, William H., 2000, *The Online Searcher´s Companion*, Lodon: Library Association Publishing.
Glossbrenner, Alfred and Emily Glossbrenner, 1998, *Search engines for the World Wide Web*, Berkeley: Peachpit Press.
Keating, Victoria, 1998, *Guía de Navegación: Literatura en Internet*, Madrid: Anaya.
Leshin, Cynthia B., 1998, *Student Resource Guide to the Internet: Student Success Online*, New Jersey: Prentice Hall.
Mann, Chris and Fiona Stewart, 2000, *Internet Communication and Qualitative Research: A Handbook for Researching Online*, Lodon: SAGE Publications.
McGuire, Mary et al., 2000, *The Internet handbook for writers, researchers, and journalists*, New York: Guilford Press.
Millán, José Antonio, 2001, *Internet y el Español*, Madrid: Fundación Retevisión.
Stout, Rick et al., 1996, *The Internet science, research, and technology yellow pages*, Berkeley, Calif: Osborne McGraw-Hill.
侯俊耀，民 85，*INTERNET 資訊搜尋指南*，台北市：第三波文化。
劉士暉、譚偉遠，民 86，*WWW 資源大搜尋：教育篇*，台北市：松崗。
程秉輝、程秉煌，民 85，*網際網路 Internet：WWW 實用寶典*，台北市：旗標。
網擎資訊，民 89，*OpenFind 網路開門：WWW 網站地圖*，台北市：第三波資訊。
網路狗仔，民 87，*中文 WWW 站另類速查寶典*，台北市：文魁資訊。
胡嘉璽，民 89，*網路資料搜尋攻略*，台北市：學貫行銷。
謝寶煖，民 89，*網站搜尋 e 網打盡：當滑鼠遇上知更鳥*，台北縣：華文網。