

第二十章

全國檔案電子目錄檢索與應用

Implementing the National Archives Search Engine

林皇興

Lambert Lin

達友科技技術服務部 副總經理

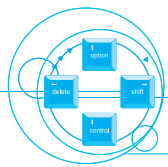
Vice President , Technical Service Dept,

Docutek Solutions, Inc.

壹、前言

依照「推動全國檔案資訊化計畫」，全國檔案資訊系統是以分年的方式建置系統及推廣使用（註 1），因此在民國九十年，主要是進行資訊系統的系統規劃、系統分析、系統開發以及系統測試等項目，而於民國九十一年進入第二階段，開始針對行政院暨所屬一級機關、直轄市及縣市政府等約五十八個機關，進行輔導上線的事宜。因此在這個階段中，最能夠被機關所屬同仁、以及全國民眾感受到國家檔案管理的重要性以及實施績效的算是全國檔案資訊系統中的相關查詢子系統了。而對政府各級單位的同仁來說，透過相關的查詢服務所帶來的便利，更能促進各單位對檔案管理相關作業的主動支持與配合，對於系統應用、輔導及推廣有很直接的助益。

「民眾查詢子系統」系提供民眾欲查詢全國各機關檔案目錄的單一窗口，其背後的資料庫係由全國各機關彙送的檔案目錄電子檔，經過檔案管理局彙整後



所建置而成的。在查詢子系統中，除了協助資訊查詢者找到資訊外，還需要整合檔案申請作業、使用者帳號登錄、權限管理等相關週邊的需求以及應用，因此規劃一個操作便利、高效能、具延展性的系統是在這個階段的目標。

雖然所有的電子檔案都是以資料庫存放，但由於資料量龐大、欄位結構多，並且需要透過網路進行多人同時且快速的查詢，因此已經無法利用資料庫管理系統內建的查詢語法來進行資料庫內容的查詢與檢索作業，而需要透過「搜尋引擎」的建置，來達到上述的要求。

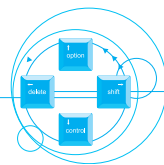
貳、認識搜尋引擎

一般目前訪問所認識的「搜尋引擎」一詞，是於網際網路陸續普及之後，有雅虎奇摩(<http://www.kimo.com.tw/>)、蕃薯藤(<http://www.yam.com.tw/>)等提供網際網路的網站資源之搜尋服務出現，這些服務也常於網路應用統計中位居網路應用的前幾名，而類似這樣的搜尋服務就被通稱為「搜尋引擎」。

因此，在後來「搜尋引擎」一詞就被泛稱為只要透過網際網路的方式，針對大量的資訊來源提供高速搜尋服務的軟體機制，就可以稱為「搜尋引擎」。所以，不論在機關、公司內部，還是開放給網際網路上的民眾或網友使用的搜尋服務，都是「搜尋引擎」可以應用的範圍。

現今的「搜尋引擎」除了根據使用者的搜尋條件找到資訊之外，還需要提供各種方便的搜尋機制，協助使用者找的更精確，或者是藉由智慧的軟體機制，引導非檔案專業的使用者找到真正的有用的資訊。而在搜尋的資訊源方面，除了資料庫之外，還可以包括網頁、各種文件格式等，達到資訊整合搜尋的目的，因此搜尋引擎的建置與應用已經在資訊業界演進成為一門專業的服務了。

以往一個系統的查詢需求，不外乎是要將資訊置於資料庫之後，透過資料庫管理系統中的資料篩選(註 2)功能，將資料找出來並進行顯示。這樣的做法針對結構性的資訊搜尋較有效率，但是對於針對一段文字進行部分符合之全文檢索需求，就會顯得沒有效率，因此必須透過「搜尋引擎」的全文檢索演算法原理，事



先針對資料庫全文的內容建立索引，以加速搜尋、檢索的效率。因此在具有大量文字內容的國家檔案資料庫中，如果需要提供完整的搜尋服務，就必須透過「搜尋引擎」中「全文檢索」的機制來達成。

參、全國檔案查詢之需求

爲了讓使用這個系統的民眾以及各機關同仁都可以很直覺、簡單的來利用搜尋的功能，因此在這裡列出一些該系統較主要的搜尋需求。此系統具有以下功能：

一、同音查詢

針對中文語音獨有之『同音』，開發同音查詢功能，當如輸入：『柯林頓』或『珂林頓』，則可以找到真正的『柯林頓』之內容。對於錯誤輸入或不同時期的翻譯名詞不同調的問題有很好的幫助。

二、容錯查詢

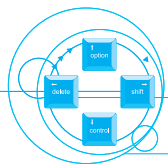
針對中文語彙中『近似』之詞彙，開發容錯(Fuzzy)查詢功能，如輸入：『教育科學』，則以容錯方式顯示幾種可能性如：『科學教育』、『教育策略與科學實驗』、『教育』、『科學』，等方式顯示。另外容錯功能亦可以用在英文的字或片語方面的搜尋應用，例如輸入：『Porposal』的錯字，可以找到正確的『Proposal』之內容。此功能另又可以稱爲模糊查詢，針對英文拼字中相近的字母組合，進行查詢，例如：輸入『Computer』可以找到『Compute』、『Computing』等內容。

三、同義詞查詢

系統內建同義詞查詢功能可透過中英文同義詞庫建立『詞彙關連』，運用同義詞檢索功能，一網打盡相近詞彙，以提昇檢索的含蓋率。另也透過同義詞的機制，將中英文翻譯詞庫置入，達到輸入英文詞找到同義的中文內容，或者輸入中文詞找到同義的英文內容，並同時找到相關的同義詞之資訊。

四、相關度排序功能

系統內建『相關度排序』功能，可依不同查詢方式，依查詢結果予以加權計



分，提供使用者識別查詢結果之相關程度。

五、截字（向右切截）功能

透過萬用字元『*』的搜尋功能來達到截字的功能，例如輸入：『Writ*』找到『Write』、『Writer』、『Writing』…等資料。

六、支援欄位間，欄位內完整布林運算功能，邏輯運算符號及'('，')'括浮運算的功能

需要支援 And, Or, Not, Near 等邏輯運算，並得以透過多欄位的層次來表達邏輯關係並進行搜尋。系統並支援多欄位併成單一欄位透過邏輯運算搜尋，也可以進行欄位間的邏輯運算。

七、Stop Word 外部設定的功能

透過純文字的外部設定檔可以進行完整的不索引英文字設定，例如 And、Or、Not、This、That 等，對檢索系統來說的虛字，是需要屏除於索引之外的。

八、支援萬用字元檢索功能

支援以?(問號)代表一個字元，*(星號)代表多個字元。如第五點所述，例如輸入：『Info*』找到『Information』、『Inform』、『Informatics』…等資料。而輸入：『8?88』可以找到『8088』、『8788』…等資料。

九、EXACT MATCH 查詢

又稱片語搜尋加上特殊符號搜尋，例如 輸入：『Search Engine』則自動啟動片語搜尋找到『Search Engine』而非傳統搜尋引擎，會找到『Search』 And 『Engine』。又『B&Q』中雖有特殊符號，但可以進行 Exact Match。可以找到『8088』、『8788』…等資料。

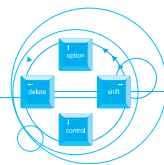
十、日期查詢功能

當資料欄位為日期型態，或檔案的日期屬性，則系統可以將其當作搜尋條件之範圍，並可以指定某日期之前，或之後，或日期區間。

十一、數值檢索的功能

當數值欄位的資料與當成檢索條件時，則系統可以將其當作搜尋條件之範圍，並可以指定某小於、大於該數值，數值區間。

十二、提供使用者可自訂選擇查詢結果每頁顯示筆數



主要提供使用者可於輸入搜尋條件時，自行選擇搜尋結果的畫面中，每頁要顯示幾筆，以滿足特殊的顯示需求。

十三、查詢結果計有條列式（約四到六個欄位）與詳目式（所有顯示欄位）

預設的搜尋結果是以條列式的方式將符合的紀錄顯示部分欄位資訊，而且可以勾選部分的筆數進行詳目式的資料顯示。

在條列式的搜尋結果中，除了可以排序之外，還需要可以分頁呈現，並且提供上下頁之分頁功能，也允許快速的跳至其中一頁。

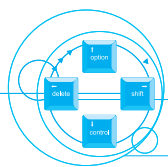
十四、搜尋結果標示符合的關鍵詞

針對搜尋的結果畫面部分，需要針對使用者在閱覽多筆或單筆搜尋結果資料時，可以將符合查詢的結果之關鍵詞標示起來顯示成不同的顏色，提供識別。

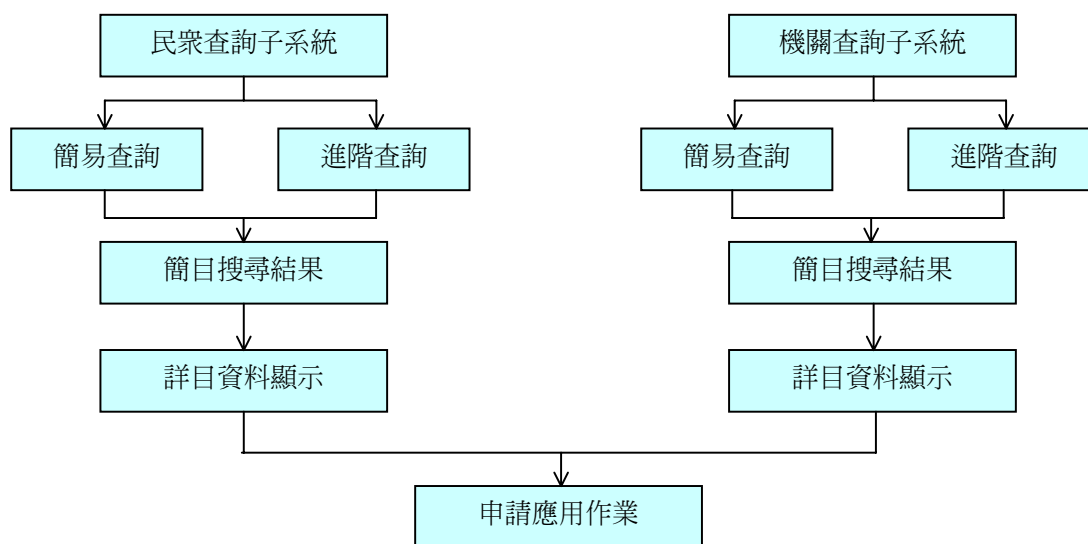
肆、全國檔案查詢系統之建構

基於前面談到的查詢系統功能上之相關需求，另考量到使用者的對象包含了民眾及機關同仁兩類，而兩者對於查詢的廣度與需求不盡相同，因此有必要將介面上拆分成「民眾查詢子系统」以及「機關查詢子系统」，前者供一般民眾查詢已開放的文件資訊，並提供申請調閱功能，而後者提供各政府機關的行政同仁來使用，其可以檢索到的範圍更完整，並提供相關的檔案申請調閱等作業。整個系統取名為「NEAR(National Electronic Archives Retrieve)」，從縮寫來看，有作為民眾接近政府檔案之窗口之意，也是這個系統的定位，因此網域定為 near.archives.gov.tw。

從檢索查詢的角度來看，全國檔案資訊系統檢索的相關的功能可以整理表示如圖一。一般民眾的檢索需求會從「民眾查詢子系统」入口近來搜尋，而政府各機關的使用者將透過權限審核後，進入「機關查詢子系统」的入口。每個查詢入口預設是提供簡易查詢的畫面，讓操作者輸入簡單的條件即可進行跨資料庫、跨欄位的檢索。而為了滿足進階使用者的更精密搜尋需求，系統亦提供進階搜尋的入口畫面供切換，透過多條件欄位的條件設立，讓搜尋的結果更為精確。在搜尋



結果的顯示上，預設先提供簡目式的摘要卡顯示，當使用者已經發現真正要的一筆或數筆資訊時，則可以調閱出詳細目錄的資訊，以作為調閱申請的參考。整個系統介面的設計則是以簡單、直覺為訴求，減少使用者的適應與學習障礙。



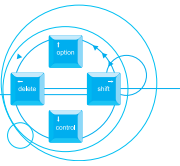
圖一 全國檔案檢索系統功能規劃圖

民眾簡易查詢的入口畫面，只需要輸入關鍵字即可進行搜尋，另外也可以選擇搜尋結果的排序方式，而民眾進階搜尋的畫面，提供各種欄位作為搜尋條件，以便更精確的進行搜尋。

而對於各機關的同仁來說，由於可以搜尋的資料範圍較廣，且經常為公務需要，需要調閱檔案，因此提供更多的功能與欄位來協助搜尋更精確。各機關的同仁使用之簡易查詢畫面，可以輸入簡單的搜尋條件，而機關查詢子系統中的進階搜尋畫面，可以透過輸入多項條件來找到更精確的搜尋結果內容。以案名欄位中輸入「總統 or 總動員」為條件，而於發／來文者欄位輸入「台北縣」為例來進行搜尋。

伍、檢索系統之系統架構

全國檔案的檢索查詢系統之系統架構規劃上，主要需要考量的包括大資料

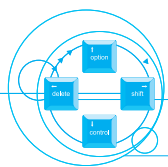


量、正確性以及檢索效能等三個項目。在資料量的考量方面，由於每年會以二至三千萬筆記錄進行歸檔開放，因此資料庫的容量規劃，以及儲存空間的建置是此檢索查詢系統所面臨的考驗之一，在這樣大的資料量中，如何維持系統的容錯能力以及長期維持正確運轉，則對硬體與資料庫軟體方面，需要較特別的選擇與規劃。

至於資料庫資料結構的規劃，對於系統的最佳化，甚至日後檢索系統的效能與檢索需求息息相關，如果資料結構規劃時僅考慮到資料儲存的方便性而忽略檢索的需求時，屆時檢索部分就需要額外的程序來達到要求，相對的也可能需要犧牲效能，因此資料庫資料結構的規劃，在系統建置之前，是需要非常審慎的規劃的。

資料庫建置起來之後，接下來就需要開始規劃檢索系統的架構了。檢索系統為了解決循序逐筆搜尋時搜尋速度緩慢的問題，因此建立全文索引是必須的。全文索引就像書籍後面所附的索引頁的資料類似，協助我們很快的可以索引到該書的一些重要字詞。因此全文檢索系統可以根據設定，將資料庫的內容事先掃描一遍，並進行全文演算法處理，於資料庫外部建置索引檔，或稱為索引資料庫，以便使用者進行搜尋時可以在最短的時間內得到答案，而且搜尋時與後端的資料庫是分開的，不會增加後端資料庫的任何負載，讓資料庫可以更專心於處理應用程式的資料匯入、移轉、檢調等作業，也讓資料庫可以保持更高、更穩定的效能，而非花時間在處理高消耗、無法預測需求量的檢索需求上，這一點是很重要的，可以讓資料庫保持在最健康狀態。

在全國檔案檢索系統中所採用的搜尋引擎程式，可以將資料庫的多個欄位的內容進行跨欄位搜尋，以達到簡易搜尋中的單一欄位條件輸入，但搜尋的範圍涵蓋指定的數個欄位的內容，因此只要決定哪些欄位的值是要涵蓋進來的就可以了。而檢索系統也可以支援做多欄位多條件的搜尋應用，唯在條件上分成兩類，一類是對指定的欄位進行全文檢索，例如：案由、機關名稱...等文字欄位的內容；另外一類是可以用來進行狀態過濾，或區間過濾的內容，例如：密等、檔案狀況、檔案日期...等。因此，需要將資料庫中的相關欄位進行歸類，以便與搜尋入口畫面的檢索條件匹配。

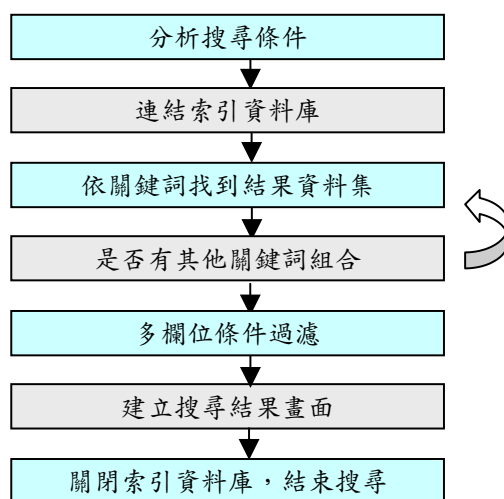


由於索引資料庫的存在也需要透過儲存空間來運作，而本系統又需要針對民眾與機關分別建置搜尋的畫面與流程，因此爲了節省硬碟空間的佔用以及整體預算的關係，必須將兩個搜尋系統共用同一份索引，所以該索引的建置也要將此需求考慮進來，也要顧慮不要因此影響太多搜尋的效能。

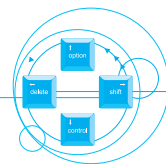
陸、檢索效能有關的因素分析

檢索系統的價值中，搜尋的效能扮演重要的地位，然而檢索功能與檢索效能往往是背其道而行，想要加入更多的檢索功能，勢必一定會對檢索的效能造成影響，因此如何需要如何兼顧功能與效能則是檢索系統建置時的重要挑戰。在分析與檢索效能有關的因素之前，讓我們來了解一下，當使用者輸入搜尋的條件之後，檢索系統處理的流程有哪些，並且可以從這些流程出發，來說明影響搜尋效能的因素爲何？

如圖一，當使用者輸入欲搜尋的關鍵詞並且按「開始查詢」按鈕之後，系統會先進行搜尋條件的分析，以便了解該次搜尋需要啓動哪些功能，條件爲何，以及要顯示哪一頁的結果資料。之後系統會去連結、開啓索引資料庫，配合全文檢索演算，快速找到符合的筆數，如果有啓動額外的功能，則也會在這個階段進行處理。之後開始處理多欄位的條件過濾，之後再將結果畫面需要的顯示資訊組合完成，並且將結果送出，顯示於使用者的瀏覽器中。



圖一 搜尋系統內部流程



影響檢索系統效能的主要因素，經分析並說明如下九個項目，分別說明之：

一、索引資料庫的大小

索引資料庫一般與資料來源的大小成直接的正比關係，也就是說，資料的來源筆數愈多，則建立起來的索引資料庫也會越大。而索引資料庫越大，在進行搜尋時則需要花費越多的記憶體(Memory)以方便進行相關過濾與排序運算，因為符合筆數將依據機率原則而成正比方式成長，而相對的對於 CPU 的資源消耗也會成正比。

二、索引資料庫的同時搜尋數量，與效能的關係

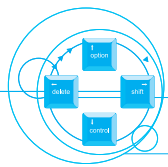
每一個索引資料庫是由數個實體的檔案來組成，這些檔案儲存於實體的硬碟中。因此在進行搜尋時，搜尋引擎就會開啓相關的索引資料庫檔案，此時的瓶頸在於檔案的存取是實體的 I/O(輸入輸出硬體作業)動作，因此會花較多的時間，如果搜尋需求上，一次需要同時搜尋多個索引資料庫時，則 I/O 更多，則造成更多的瓶頸。

因此解決的方式一般是規劃將 I/O 分散在不同的硬碟上，或者規劃分散在不同的主機上，如果部分索引檔常態的放置於記憶體中來運作運作的話，而不依賴實體硬碟的動作，如此搜尋能量估計可以提昇數十倍甚至百倍的程度，但投資的硬體設備較多。而在目前僅建立一個索引資料庫，所以沒有這些問題，但未來有必要將索引資料庫拆分成數個，以避免單一索引資料庫的物理限制問題。

三、單一欄位搜尋與多欄位多條件的搜尋比較

「多欄位多條件的搜尋」較「單一欄位搜尋」對於處理器(CPU)與記憶體(Memory)的需求約有 1.5-16 倍的資源需求，因此影響甚鉅，而條件的欄位數目越多，則情況影響越大，因此即使有多欄位條件的需求，則建議適當的規劃與控制，以保持適當的效能。

而多欄位中的「全文欄位」與「屬性欄位」也必須適當的規劃，只將必要的欄位納入「全文欄位」索引中，也只將必要的欄位納入「屬性欄位」索引中，以達到最佳化的要求與效果。由於以上的原因，因此幾乎所有的搜尋引擎入口網站，



幾乎沒有網站提供多欄位的條件搜索。

四、排序功能對於搜尋效能的影響

動態排序，亦即無預警的臨時希望對某個資訊進行排序，是對記憶體與處理器資源運用的一大傷害，尤其當資訊源龐大的情況下更是如此。因此唯一的解決方法是針對可能的排序欄位，仔細的慎選，並且事先建立排序索引，以成立排序輔助索引檔，加速排序效率，但選項越多，則建立索引越耗時，而且也會佔用硬碟空間，並且讓原本的 I/O 負載惡化，因此要小心運用。

五、特殊搜尋功能(Ex 同音、容錯…)的啟動對於搜尋效能的影響

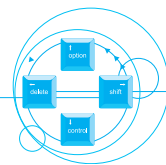
不同的搜尋特殊功能的啟動，將根據演算法或方法的不同，而對資源有不同的要求，進而產生不同的效能影響。茲舉幾個搜尋功能對演算法的影響來進行相關說明。

「同音」搜尋為例，所耗費的資源將會與輸入條件的字數，以及每個字的同音展開後字數的所有乘績做倍數的關係。例如以輸入「知識」兩個字來看，則必須將「知」的所有同音字展開，例如有「6 個」，而「識」的同音字展開有「5 個」，則啟動同音功能，比沒有啟動同音功能，需要花費 $2*6*5$ 約為 60 倍的系統資源花費。

因此以耗費系統資源的程度來看，目前的特殊搜尋功能對於系統資源佔用的排行順序會依次是，「形似詞」>「相關詞」>「近似概念詞庫」>「中文同音」>「邏輯運算」>「萬用字元」>「中英文容錯」，因此除了必須慎選一個搜尋引擎系統需要開放何種搜尋功能之外，也需要去預估使用者勾選這些功能的機率，以考量主機端負載的預估。

六、主機端 CPU 等級與數量，對於搜尋效能的影響

CPU 的等級與數量將直接影響搜尋的效能，尤其有啟動排序時，對於 CPU 資源的要求更大。CPU 等級與數量對於搜尋效能的關係可以透過下面的例子來進行說明。



擁有一顆 CPU 的 Server，如果一個搜尋花費 0.8 秒，且佔用 CPU 20% 的尖峰負載，並持續 0.8 秒，則在這個 0.8 秒的區間中，約可以同時承受三至五個搜尋需求，在此假設記憶體資源沒有遇到瓶頸，而硬碟 I/O 也可以支持，如果有雙 CPU 則約可以承受概估四至八個搜尋的需求。因此如果有十六個人同時按「開始搜尋」按鈕時，則有可能每個人都要等 2 秒以上才會看到搜尋結果。

七、主機擁有記憶體多寡，對於搜尋效能的影響

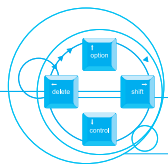
與第六點說明相似，只是從記憶體的觀點來考量。但隨著「索引資料庫」的空間越大，則單一搜尋會花掉的記憶體也會不同，另外輸入條件字數、詞數、甚至過濾條件、特殊功能等，對於記憶體的需求也持續增加。因此記憶體對於搜尋效能的關係可以透過下面的例子來進行說明。

假設每個使用者都輸入相同的條件，啟動一樣的搜尋功能，今天一個人在搜尋時，主機上面觀察到佔用 20MB 的記憶體，且耗時 1.8 秒才完成搜尋，此時如果同時有多人進行相同條件的搜尋，則記憶體成倍數需求外，耗時將成正比的方式成長，但是一旦記憶體用量已經超過作業系統中的剩餘實體記憶體時，則搜尋時間將成可怕的倍數成長(因為虛擬記憶體透過硬碟置換的關係)。因此記憶體必須經過詳細的計算，並配備適當的數量，避免碰到臨界點，造成數分鐘沒有結果，甚至產生「CGI Timeout」(註 3)的錯誤訊息。

因此，為了避免一些錯誤的情況發生，大部分的服務系統都會設定同時能夠服務的「Request Concurrent」數量，以保障系統操作品質，遇到超過臨界點時，則導引到禮貌拒絕之說明網頁，告知目前系統滿載，無法接受該搜尋請求，請稍後再來進行搜尋，以避免全部的人都不能用。

八、同一時間搜尋請求多寡，對於搜尋效能的影響

對於同一時間搜尋請求數目較多時的解決方法，一般的是採用負載平衡做法 (Load Balancing)，藉由安裝更多的主機方式來達到資源擴充的效果。解決此一問題的可行法是將整套搜尋引擎分別安裝在不同的機器內，隨同索引檔的複製，用多部主機來分散前來的搜尋之需求，而於系統前端透過軟體或硬體來達到



流量導引的功能。

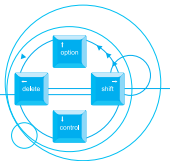
九、搜尋結果筆數的多寡，對於搜尋效能的影響

當搜尋引擎的核心已經找到完整的結果後，要將結果套成網頁顯示出來的話，往往也是需要花許多的時間與資源來組成網頁，並將網頁輸出，因此搜尋結果的筆數設定越多的話，則該部分所需的資源與所花時間，將成倍數成長，而且網頁越大，資料量越大，透過網路傳輸到使用者的瀏覽器完整顯示所需花的時間就越多。

柒、結語

檢索查詢相關系統是一般民眾與機關同仁面對檔案管理作業所接觸的第一個窗口，因此值得花一些心思將此系統規劃得更為簡便、更有效率。在未來，除了因應資料量的擴展以及檢索需求成長而需持續擴增軟硬體設備來因應之外，也需要來思考隨著技術的進步，階段性的提供更完整的功能，協助使用者可以搜尋的更精確、更有效率，甚至在未來系統還可以化被動為主動，讓各機關的同仁可以訂閱所需的相關條件及其搜尋結果，並且在同樣的條件中，有新的資料開放時，系統可以主動的進行一對一的通知服務。

【原刊載於檔案管理局出版之檔案季刊（九十一年六月）第一卷第二期】

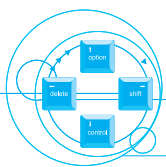


註釋：

註 1： 林秋燕。「全國檔案資訊系統之規劃」，檔案季刊創刊號，頁 29，民國 90 年 12 月。

註 2： 「篩選」指的是資料庫的 SQL 語法之 Select 子句。

註 3： 「CGI Timeout」指的是 Web Server(網頁伺服器程式)爲了避免有些程式無止盡的執行，因此通常會設定 CGI 程式的執行超時限制，例如兩分鐘。



檔案資訊資源管理
