WWW資訊檢索的新趨勢─欄位檢索

目前檢索引擎不分欄位式的全文檢索方式，已經因為網頁數量的快速膨脹，所導致的回覆款目過多，對使用者造成極大的困擾，因此如何過濾資料成為今後主要的研究課題。本文中作者探討欄位檢索在過濾資料上的功效，並且以一個簡單的實驗來驗證欄位檢索在過濾資料上的效果。

吳政叡

1990年代資訊傳播的重大革命是全球資訊網（WWW）和網際網路（Internet）的結合，不但大幅降低了資訊傳播的障礙，更進一步促成網頁數量的增加。為了有效來檢索網頁上的內容和資料，於是有檢索引擎（Search Engine）的產生，利用全文檢索的技術，自動將抓取到的網頁加以斷字取索引，建立資料庫，因此使用者可以很快的找到所欲查詢的關鍵字，出現在那些網頁。這種操作的特性是高速（電腦自動斷字）和一網打盡（全文檢索），在初期獲得顯著的成功，於是各個檢索引擎如雨後春筍般的產生，至今在WWW上遨遊的人，相信沒有人不曾使用過檢索引擎。但是隨著WWW上的網頁數量快速激增，檢索引擎的資料庫也隨之快速膨脹，引發回覆量過多的問題，成為現在檢索引擎使用者時刻遇到的共同夢魘，雖然檢索引擎有將回覆款目加以排序，但是排在前面的，又往往不是所需的資料，也無法一一檢視所有款目，造成有資料卻找不到的困境。

很明顯的，在此時刻如何過濾資料成為主要的研究課題，但是過濾資料必須有所依據，即須有對欲過濾資料的一些額外描述資訊。L. Dempsey 和 R. Heery 依資料記錄（Record）的有無結構性和複雜程度，將此類（資源描述性）資訊分成三種: [註 1]

(一) 使用未結構化的資料（即原始資料），如檢索引擎使用電腦來自動抓取資料（如網頁）和自動製作索引，來支援資料的查詢。

(二) 使用結構化的資料（即非原始資料），可支持欄位查詢，資料結構簡單，可由非專家或文件創造者自行著錄，如都柏林核心集等。[註 2]

(三) 使用較完整的描述格式，欄位多，資料結構複雜，通常由專家來著錄。

依據上述的架構，可以很明顯的看出檢索引擎是歸在第一類，並沒有提供過濾資料所需的資訊。至於圖書館界廣泛使用的機讀編目格式（MARC）則歸入第三類，雖然可以提供高品質的資訊，可是製作成本過高，各圖書館在應付傳統的紙本印刷資料，在人力上就已經捉襟見肘，更別提要應付數量龐大的網頁，所以也不適合用來提供過濾網頁所需的描述資訊。因此較適合的應該是第二類資訊，雖然其中包含很多種類，但是共有的特性是支持欄位檢索、結構簡單、製作成本低，如都柏林核心集（Dublin Core）即屬此類，顯然是較適合用來支持網頁過濾的功能。有關都柏林核心集的網頁過濾效果，作者在上期（109期）中已有一個小規模實驗來加以驗證 [註 3]，請讀者自行參閱。

雖然在上期的實驗中，已初步證實欄位檢索的效果，但是礙於目前並無大規模的此類型資料庫可以用來進行大規模實驗，因此作者利用Infoseek檢索引擎為實驗對象，來模擬大規模資料量時，欄位檢索可能有的效果。Infoseek在自動

斷字取索引時，有利用位置的差異，將取得的索引加以粗略的分類，其中將HTML網頁在<head></head>中，<title></title>內的字詞，特別註記其位置。由於網頁中此部分的資訊相當於「篇名」或（一般書籍的）「書名」，是檢索時經常被使用的一個欄位，因此以選修作者所開設的研究所課程「元資料概論」的研究生為實驗者，各令其使用五個專業化字詞和五個生活化字詞，來比較在使用Infoseek檢索引擎時，一般方式（即無限定位置）和限定title位置方式，兩者在資料回覆量上的差異，結果整理如下：

表1. Infoseek中一般方式和限定title位置的平均回覆款目數量比較表。

	一般方式 限定title位置
專業化字詞	1,836,749 464
生活化字詞	1,087,422 4,383

從兩者在量上的差異，讀者不難發現，即使是簡單的以位置來粗略的結構化資料（或分欄位），在過濾資料上就可以發揮巨大的功效。

在實驗中的另一個重大發現，是那些有出現在title位置字詞的網頁，其（排列）分數經常相對較低，造成的一個事實是在使用（無限定位置）一般方式檢索時，這些網頁（或回覆款目）都是排列在較後面，因此使用者幾乎不可能會找到這些網頁，這也間接證實大家熟知的事實—現在檢索引擎提供的回覆資料，混雜了太多不相干資訊，同時排序方法有問題，有用的資料常常無法排列在前面。

註釋

註 1：L. Dempsey and R. Heery, “An Overview of Resource Description Issues,” March 1997, <http://www.ukoln.ac.uk/metadata/DESIRE/overview/rev_01.htm”, p. 4。

註 2：吳政叡，「元資料實驗系統和都柏林核心集的發展趨勢」，國立中央圖書館臺灣分館館刊 4 卷 2 期（民 86 年 12 月），頁12-18。

註 3：吳政叡，「資訊的檢索失誤率探討」，中國圖書館學會會訊109 期（民 87 年 6 月）。