

圖書與資訊學刊論文的高頻詞語抽取和分析

Extracting High-frequency Terms from *The Bulletin of Library and Information Science*: A Result Analysis

林 頌 堅

Sung-chien Lin

世新大學資訊傳播學系助理教授

Assistant Professor, Department of Information and Communication Studies

Shih Hsin University

E-mail : scl@cc.shu.edu.tw

【摘要 Abstract】

本論文中嘗試利用自動詞語抽取的技術，對政治大學出版圖書與資訊學期刊的論文進行高頻詞語抽取與分析。在這個問題中，需要克服中文的自然語言處理問題，我們嘗試利用統計方式配合語言學知識來發展詞語抽取技術。利用這個技術，我們對近六年圖書與資訊學期刊的 164 篇論文的題名與摘要抽取出了 233 種詞語，在本論文中並對於這些詞語做了一些探討。本研究發現圖書與資訊學期刊論文的詞語以圖書館學的相關概念最多，其次資訊使用研究、資訊科技、圖書資訊學教育、圖書館利用教育和檔案學的相關詞語也相當多。

This paper employs some automatic diction extraction techniques to locate and analyze some key and high-frequency terms from articles published in the Bulletin of Library and Information Science of National Chengchi University. Faced with some difficulties in natural language processing of the Chinese language, we have developed techniques based on statistical information and linguistic heuristics. Two hundred and thirty-three types of terms are extracted from the titles and the abstracts of 164 pieces of articles published last six years. From the concepts carried in the extracted terms, we discover important subject topics in the journal. They include library science, information use studies, information technologies, education for library and information science, library user education, and archives.

關鍵詞 Keywords

詞語抽取 期刊論文 圖書與資訊學刊

Term extraction ; Journal articles ; The Bulletin of Library and Information Science



壹、緒論

在本論文中，我們嘗試利用自動詞語抽取的技術，從政治大學出版的圖書與資訊學刊中的論文抽取出這份期刊高頻而與論文主題相關的詞語。在將圖書與資訊期刊的論文題名與摘要電子檔，建立成為資料庫後，利用自然語言處理的技術，自動地抽取出在論文題名與摘要中常出現且對於論文主題相關的詞語(註 1)，並對於抽出的結果加以分析，討論這些詞語在期刊的分布情形，作為進一步探討在圖書與資訊期刊中的一些研究發展、成果發表與學術傳播的現象。本論文將簡述針對這個問題所發展的詞語抽取技術，詞語抽取的結果與目前所得到的初步分析和討論。

圖書資訊學界長久以來即對於學術論文的分析有很高的興趣，並且也發展出多種的文獻計量方法與技術 (White and McCain, 1990) (Wilson, 2001)，研究人員嘗試利用這些方法與技術從論文資料庫中發現事實，進行分析，來建立研究發展與學術傳播的相關理論。在過去的文獻計量方面的研究上，較常觀察的主體包括有作者(Budd and Seavey, 1990)(Zhang, 1997)(Egghe, Rousseau and Hooydonk, 2000)、期刊(Wormmell, 1998)(Tsay, Jou and Ma, 2000)、論文(Koehler, 2000)(Wang and Soergel, 1998)等資料項目。在 ISI 建立了龐大的引用文獻資料庫後，引用文獻及從這裡延伸的項目也是許多研究人員常利用作為統計分析的資料(Garfield, 1979)(Borgman, 1990)。

對於論文在於文獻計量的研究，除了以整篇論文作為觀察的主題以外，另一個思考角度是從資訊檢索的觀點來看，以作為論文表徵(presentation)的關鍵詞語或論文本身中內含的詞語來作為觀察的主題(Zitt, et. al., 1999)(Hood and Wilson, 2001)(Callon, Law and Rip, 1986)。

我們可以將論文視為是作者與讀者的一段意見傳播過程，作者在寫作的過程中，嘗試以文字表達將所希望影響讀者的概念傳達給讀者。而讀者則在閱讀的過程中，依據他們已有的知識結構與論文中承載的概念，重新來形塑他們的知識結構。因此，作者為了闡述說明某些特定概念，會在論文中重複用來表達這些概念的詞語，使得這些一再出現的詞語能清楚地將作者的意見傳達給讀者，達到說服讀者接受作者意見的目的。比方說，本論文的主題是討論論文中的詞語抽取方法，因此，在論文中我們便多次使用「論文」、「詞語」、「抽取」等等詞語，以表達這篇論文所希望探討的主題。讀者也能藉由這些詞語所表達的概念逐步建立起自己的相關知識結構。因此，論文中高頻出現的詞語往往相關於論文的主題，而且愈高頻的詞語對於論文的相關性愈為重要。這也是資訊檢索領域的研究人員常以詞組為基礎做為論文表徵的原因，比方說著名的 SMART 系統，便是使用詞語的頻率進行檢索的例證。相較於資訊檢索領域對論文表徵方法與技術的不斷改進以及不勝枚舉的書目計量研究中利用作者、引用文獻作為觀察主體所得到的研究成果而言，利用詞語為觀察主體所進行的文獻計量研究目前可說是屈指可數(Leydesdorff, 1997) (Wilson, 2001)，也因此可以有極大的發展空間。

要發展利用詞語為觀察主體的文獻計量研究，首先需要發展詞語抽取技術。在發展以詞語為基礎的論文表徵技術上，中文處理相對於其他語言來說相當困難。在書寫上，中文句子中的詞語間沒有空白作為界限，因此，很難從句子中將這些構成詞語的字串抽取出來。在本論文中，我們將以統計式的自然語言處理方法來發展詞語抽取技術，從期刊論文中抽取高頻而與主題相關的詞語，作為分析的基礎。



以下是本論文其餘的部分，在下一節中，我們將討論中文詞語抽取的問題，並根據本研究的需求發展合適的技術。第三節，則以近六年的圖書與資訊期刊論文為對象，從論文的摘要及題名中抽取高頻詞語，並且分析這些詞語的分布情形。最後則是本論文的結論與未來的發展方向。

貳、高頻詞語抽取方法

正如在前面本論文所說明的，作者為了將某些觀念清楚地表達給讀者，會在文章一再闡述這些觀念，因此與這些概念相關的詞語在這篇文章中會較其他的詞語來得常出現。若從期刊的角度來看，期刊的編輯會根據論文的主題作為選擇論文的依據，因此同一期刊中所發表的論文具有相關的主題，同一期刊中所有論文所成的集合中勢必包含大量與期刊主題相關的詞語，而且這些主題相關的詞語相較於其他的詞語具有頻率較高、並非偶然出現等的特徵。以下我們即探討詞語在抽取上的特徵並利用這些特徵發展符合研究需求的技術。

一、詞語抽取的問題

在語言學的研究中，字(Characters)是在書寫上能辨識的最小單位，以一組具有特定順序的字串構成詞或詞組來代表一個概念(Fromkin and Rodman, 1988)。文章則可以視為是由若干主題相關的句子所構成，這些句子中則由詞與詞組依照特定的句法結構(Syntactic structure)形成。在中文自然語言處理中，有幾點中文的語言特性是特別重要而值得注意的：首先，中文字的數量相當龐大，常用的中文字總數在數千個以上，而且每一個字都是一個音節(Fromkin and Rodman, 1988)。而現代漢語裡的詞大多數是多個中文字所構成的多字詞，其中以二字詞最多。而且在書寫

上，中文句子裡詞與詞間並沒有明顯的界限。這些因素使得中文在詞的概念上較不明顯，對於以詞語為基礎的中文自然語言處理或資訊檢索，是一個相當大的難題。因此，對中文句子進行自然語言處理，首先需要經過斷詞處理(Word identification)，確認句子中構成詞語的字串(陳克健, 民 88)。

目前較為成熟的斷詞處理技術大多先採用以詞典法(Dictionary look-up)為基礎的斷詞處理(Chen and Liu, 1992)。這個技術的核心是一個事先編輯好的詞語表(詞典)，記錄所有可能的詞與詞組的字串。在對句子進行斷詞處理時，先以詞語表搜尋出輸入句子中所有可能是詞語的候選字串，再配合上字串的長度、頻率、相連強度(Association)等等語言學上的經驗知識與統計資訊，排除混淆的候選詞語，決定句子中最佳的候選詞語組合。因此，在斷詞處理技巧中，除了運用合適的語言學知識配合有效的演算法之外，相當重要的是建立一個完整的詞語表以提供所有句子中可能的候選詞語。

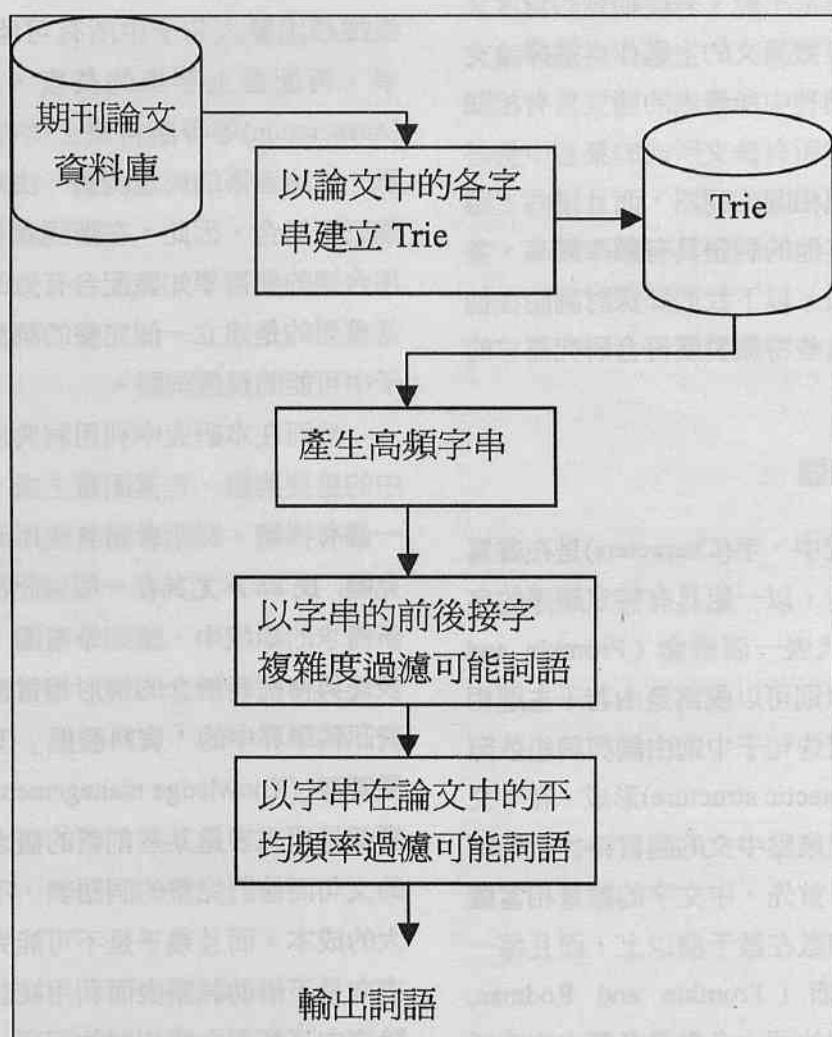
然而在本研究中利用詞典法來抽取期刊論文中的重要詞語，有其困難之處。由於語言本身是一個有機體，詞語會隨著使用而出現或消失(陳克健, 民 88)。尤其在一個資訊流通發達與具備創新需求的環境中，諸如學術圈，產生新的詞語來表達與傳播新概念的情形相當常見。比方說，在資訊科學界中的「資料發掘」(Data mining)、「知識管理」(Knowledge management)等近年出現的詞語都是用來表達某些創新的概念。跟隨所要處理的文句而修訂完整的詞語表，不僅需要耗費相當大的成本，而且幾乎是不可能完成。一個可行的方向是不藉助詞語表而利用統計資訊，抽取期刊論文中高頻而主題相關的詞語。在本研究中，只有高頻且主題相關的詞語才有研究詞語分布的價



值，是可以進一步分析的對象。所以本論文將不考慮利用詞典法抽取所有的詞語，而以統計法來抽取期刊論文中的重要詞語。

可能是所要抽取詞語的候選字串具有以下的條件：(一)字串在期刊論文的頻率次數(Frequency count)高於某一個閾值(Threshold)，具備超過閾值的次數表示這些字串可能是在期刊的論文中經常被使用的詞語，即有可能是重要的詞語。(二)字串是一個完整的詞語，有些字串的頻率次數雖然很高，可是是屬於高頻詞語中的一個部分，無法

代表特定的概念，這些字串不應被抽取出來。比方說，在圖書資訊學期刊中，「資訊檢索」是出現次數很高也是很重要的相關詞語，應當被抽取出來；這個詞語的部分，如「資訊檢」或是「訊檢索」也具有相同的出現次數，但不是一個完整的詞語，不應當被抽取，然而「資訊」或「檢索」本身便是詞語，則應當也需被抽取出來。(三)本研究希望抽出與期刊論文的主題相關的詞語，作為進一步研究期刊論文主題分布的基礎。



圖一：高頻詞語抽取技術與過程

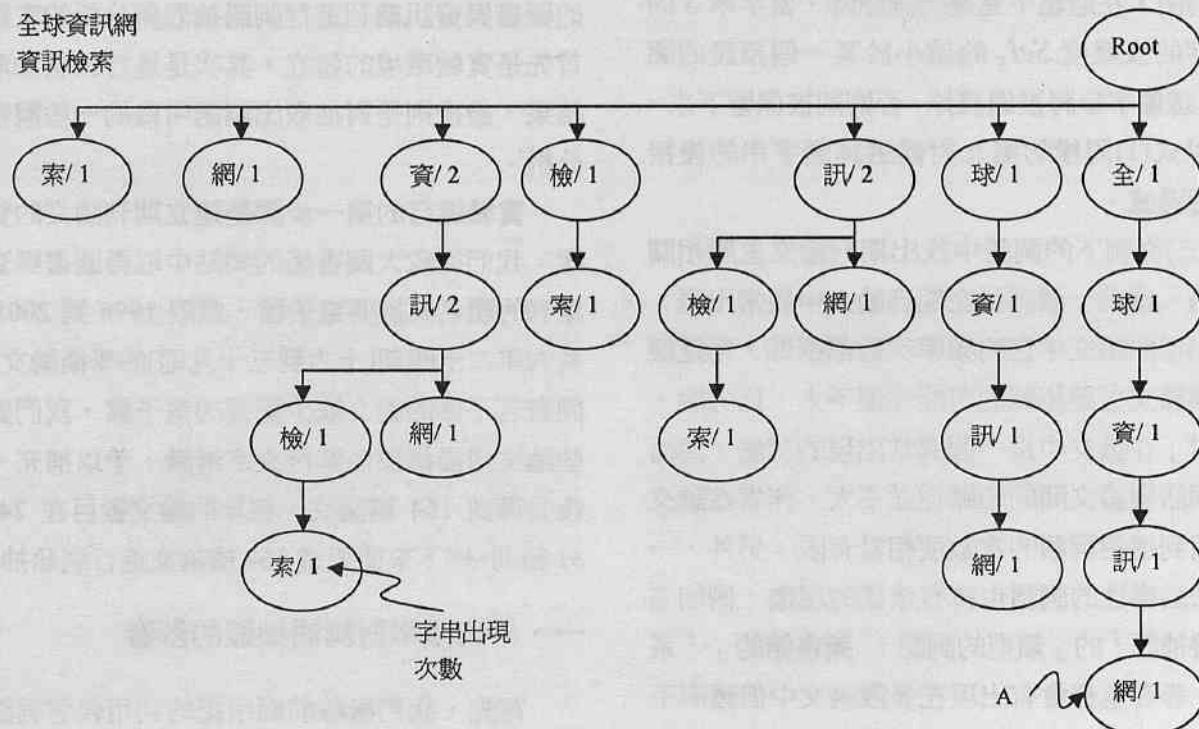


二、抽取方法

根據以上對詞語抽取的問題加以分析後，我們設計了一系列詞語抽取的技術。圖一是本研究提出的詞語抽取過程，各步驟解釋如下：

(一)從其論文中選取高頻的字串作為候選詞語。首先我們計算出現各種字串在期刊論文中的頻率次數，然而在期刊論文中存在非常多可能的字串組合，若以窮舉搜尋法(Exhausted searching)將所有可能的字串產生出來，並計算這些字串在期刊論文中的出現次數，需要相當多計算時間而不切實際，因此有必要發展更有效率的計算方式。在本研究中，我們使用 Trie 資料結構來解決

這個問題，Trie 資料結構如圖二所示。將輸入的文字依據出現的順序安排成 Trie 中的一個路徑(Path)，比方說圖二中的「全球資訊網」可以表示成從根節點(Root)到節點 A 的一條路徑，節點中另外還記錄每一個從根節點到這個節點的相對應字串在期刊論文的出現次數，如此一來可以很輕易地獲得期刊論文中所有出現字串以及它們的次數。比方說，從圖二的 Trie 中，我們很容易地可以得到字串「資訊」的次數為 2，而「檢索」的次數為 1。接下來再依據預先設定的頻率次數閾值，從產生的 Trie 中取出高頻的字串作為候選詞語。



圖二：以字串「全球資訊網」和「資訊檢索」建立起來的 Trie

(二)從候選詞語中過濾不完整的詞語。如果在期刊論文中一個字串前後連接字的可能情形不多的話，可能這個字串是詞語的一個部分，而非一個完整的詞語。如前面提到的例子，字串「訊

檢索」的前接字只有一個可能，因此不是一個完整的詞語。同樣在字串「資訊檢」的情形，它的後接字選擇可能也不多，所以不是一個完整的詞語。但是「資訊檢索」、「資訊」、「檢索」都是一



個完整的詞語，前後連接字的可能情形非常多。在本研究中嘗試以式(1)來計算字串 S 的前接字的複雜度 Sel_s 。

$$Sel_s \stackrel{\text{def}}{=} -E[\log p_{as}] = -\sum_{pas} p_{as} \log p_{as} \quad (1)$$

在式(1)中， p_{as} 是指字串 S 的前接字為 a 時的對於字串 S 的條件機率，其值為字串 aS 與 S 在期刊論文中出現次數的比率。在式(1)的定義下， $Sel_s \geq 0$ ，當字串 S 的前接字只有一種情形時，字串 S 的複雜度 Sel_s 的值為 0，表示 S 極可能是詞語的一部分；如果 Sel_s 的值愈大，表示其前接字的種類愈多且分布愈平均，則字串 S 很可能是一個詞語。在過濾不完整的詞語時，當字串 S 的前接字的複雜度 Sel_s 的值小於某一個預設的閾值時，這個字串將被過濾掉，否則則被保留下來。我們以式(1)同樣的概念對候選詞語字串的後接詞進行過濾。

(三)從剩下的詞語中找出期刊論文主題相關的詞語。當某一個詞語在期刊論文中經常出現，但在出現的論文中它的頻率次數都很低，則這個詞語與論文主題相關的可能性便不大，比方說，「探討」在論文中是一個經常出現的詞語，然而這個詞語與論文間的相關性並不大，作者在論文中使用到這個詞語的次數便相當有限。另外，一些語法結構性的詞語也會有這樣的現象，例如名詞後接補語「的」類型的詞語，「圖書館的」、「系統的」等等也都會有出現在多數論文中但頻率不高的情形。反之，如果一個詞語與某一些論文的主題間有很高的相關性，這個詞語將在這些論文間出現多次。換言之，我們可以觀察到相關詞語的出現並非隨機(Random)，在某些論文中會出現的頻率高，而在主題不相關的論文中出現的可能性較低。因此我們可以利用這項特性來選取與論文主題相關的詞語。我們以詞語出現在期刊論文

中的總次數與出現的論文篇數的比值，也就是每個詞語在出現論文的平均頻率，作為詞語選取的標準，作為相關詞語的選取標準，當這個比值超過某一個閾值時，表示這個詞語在出現的論文中的平均頻率很高，我們便認為這個詞語可能是與某些論文的主題相關，因此需要加以保留；反之，如果詞語在論文中的平均頻率低於閾值時，則認為這個詞語可能與論文主題相關較小或是結構性的詞語而去除。

參、實驗結果與討論

本節報告利用上述詞語抽取方法對政大出版的圖書與資訊期刊進行詞語抽取與分析的實驗。首先是實驗環境的建立，其次是進行詞語抽取的結果，最後則是對抽出詞語所做的一些觀察與分析。

實驗進行的第一步驟是建立期刊論文的資料庫，我們從政大圖書館的網站中取得圖書與資訊期刊的題名與摘要電子檔，選取 1996 到 2001 年共六年二十四期(十六到三十九期)的學術論文。期間有若干期的論文缺少摘要的電子檔，我們對這些論文掃描摘要並進行文字辨識，予以補充。最後共得到 164 篇論文，每年的論文數目在 24 到 31 篇間。接下來便對這 164 篇論文進行詞語抽取。

一、平均頻率對詞語抽取的影響

首先，我們檢驗前面所提的利用候選詞語在論文中的平均頻率作為詞語與論文相關程度的測量方法。在這個實驗中我們所設定的詞語出現次數、詞語前後接字的複雜度等閾值分別為 5 次和 1.0，經過前兩個詞語抽取步驟後，共得到 873 種候選詞語。表一的結果是以不同的平均頻率閾值對 873 種候選詞語進行過濾所得到的詞語數目，當我們設的平均頻率閾值愈大，所得到的詞語數



目便愈少。對這些抽取出來的候選詞語，仔細觀察它們的平均頻率，可以發現，如同前面所說明的，平均頻率最低的是一些論文中經常一起出現的字所形成的片段，比方說像是「的現」、「的現」是「的現象」、「的現況」、「的現在」等等的一個部分，這些字串不是詞語，雖然符合出現次數和詞與前後接字的閾值，但容易被平均頻率去除。其次是一些論文寫作中常用但概念較為廣泛的詞語，比方說像是「推展」或是「形式」等，這些詞語出現的論文篇數多但是被使用的次數並不高。平均頻率愈高，愈有可能是一個具有特定概念的詞語，而且所表示的概念與論文主題愈相

關。平均頻率最高的詞語，比方說像是「口述歷史」或是「館際合作」等詞語，它們所表示的概念非常特定且與論文的主題十分相關，這些詞語只有出現在少數的論文中，但是一旦出現它的頻率就非常高。從上面的觀察中，我們可以將平均頻率視為是一種檢驗詞語抽取的有效性以及與論文主題相關程度的指標，當平均頻率的值愈大，抽取出來的字串愈不可能是在論文中偶然發生的詞語片段，而愈有可能是一個與論文主題相關的詞語。表二則是將所有的候選詞語依照平均頻率排序，每隔 30 個詞語取樣的結果，從這裡我們可以觀察到上述的現象。

表一：以不同的平均頻率閾值所得到的詞語數目及各種詞語長度的分布

| 平均頻率 閾值 | 抽出的詞語 總數 | 二字詞語 數目 | 三字詞語 數目 | 四字詞語 數目 | 五字詞語 數目 | 六字詞語 數目 |
|------------|-------------|------------|------------|------------|------------|------------|
| 1.5 | 379 | 227 | 53 | 67 | 25 | 7 |
| 2.0 | 233 | 125 | 34 | 49 | 19 | 6 |
| 3.0 | 111 | 53 | 12 | 32 | 10 | 4 |
| 5.0 | 52 | 21 | 5 | 17 | 7 | 2 |

另外，在表一中，我們也觀察了在各種平均頻率閾值下所抽取出來各種長度的詞語個數。從這裡可以觀察到長度愈短的字串愈容易由提高平均頻率的閾值所過濾，長度大於或等於 4 的字串，則較容易保留。造成這種現象是由於通常較短的字串大多是詞語的片段或是表示較廣泛概念的詞語，比方說前面曾提到的「的現」、「推展」

與「形式」等等，這些字串可能在多篇論文中出現，但在論文中的出現頻率不高，所以當平均頻率的閾值提高時，很容易被過濾。但是長度較長的字串則往往表示較為特定的概念，出現的論文數目不多，但在這些論文中出現的頻率則相當高，比方說，如「口述歷史」或是「館際合作」等都是這類具特定意義且主題相關的詞語。

表二：將候選詞語依照平均頻率排序，每隔 30 個詞語取樣的結果

| 序號 | 詞語 | 平均頻率 | 序號 | 詞語 | 平均頻率 | 序號 | 詞語 | 平均頻率 |
|----|------|------|-----|------|------|-----|----|------|
| 1 | 口述歷史 | 18 | 301 | 推展 | 1.67 | 61 | 由於 | 1.17 |
| 31 | 詞表 | 5 | 331 | 資料庫的 | 1.67 | 631 | 進行 | 1.13 |
| 61 | 館際合作 | 4.4 | 361 | 形式 | 1.5 | 661 | 為主 | 1.1 |



| | | | | | | | | |
|-----|-----|------|-----|-----|------|-----|-----|---|
| 91 | 修改 | 3.33 | 391 | 議題 | 1.43 | 691 | 調整 | 1 |
| 121 | 新聞 | 2.75 | 421 | 業務 | 1.4 | 721 | 重要性 | 1 |
| 151 | 電子 | 2.48 | 451 | 時代 | 1.33 | 751 | 相互 | 1 |
| 181 | 的檔案 | 2.25 | 481 | 年級 | 1.25 | 781 | 在探討 | 1 |
| 211 | 數位化 | 2 | 511 | 相當 | 1.25 | 811 | 同時 | 1 |
| 241 | 處理 | 1.86 | 541 | 式的 | 1.2 | 841 | 以分 | 1 |
| 271 | 地區 | 1.78 | 571 | 的圖書 | 1.19 | 871 | 的現 | 1 |

從表一和表二的結果中，在進行詞語抽取時，如果將平均頻率的閾值設定得較高，可以得到較佳的正確性，而且詞語代表表示的概念也較為特定；但如果設定得較低，我們可以得到較多的且概念較為廣泛的詞語，能涵蓋的論文主題較廣。將平均頻率設定為合適的閾值，詞語的正確性和主題涵蓋率可以達到最佳的組合。因此，為了得到較好的涵蓋性與正確性，各種閾值需要加以權衡。在接下來的實驗中，我們將詞語出現次數、詞語前後接字的複雜度和詞語出現在論文的平均頻率等各項閾值分別設為 5 次、1.0 和 2.0。

二、抽取出來的詞語與論文關鍵詞的比較

為了驗證抽取出來詞語的有效性，我們將抽取出來的詞語與圖書與資訊期刊論文中作者或編輯所給的關鍵詞相比。在關鍵詞部分，164 篇論文中，關鍵詞共 458 種，總次數為 578。在一篇論文中，最多關鍵詞共有 9 種，最少的只有 1 種，大部分論文的關鍵詞數目在 3 到 5 種間。由 578 個關鍵詞只有 458 種，可見得大部分的關鍵詞只出現過一次(394 個，佔全體關鍵詞數的 68.17%)，出現次數最多的關鍵詞是「網際網路」，總共出現於 9 篇論文中。大部分關鍵詞的長度在 4 字到 6 字間；最長為 25 字，最短為 2 字。

如果假定論文中的關鍵詞與論文主題間有很高的相關性，而且詞語又是從與論文主題十分相

關的題名和摘要中抽取出來，抽出的詞語和關鍵詞間應該有很多相同。但在比較關鍵詞與抽出的詞語，所抽出的詞語中只有 83 種關鍵詞，在論文中的次數為 150 次，本研究進行詞語抽取的檢全率(Recall rate)為 25.95%。從這個角度看來，本論文的詞語抽取法似乎並沒有達到良好的效果，下面我們便分析造成這個結果的原因。首先，若以關鍵詞出現在論文題名或摘要的頻率來看，事實上大部分的關鍵詞未曾出現在論文的題名或摘要中，或者出現頻率過低，造成無法進行抽取的情形。在所有的關鍵詞只有 324 種，曾經出現於 164 篇論文的題名或摘要中，出現頻率超過 5 次的關鍵詞更只有 118 種。許多代表重要概念的詞語，如「資訊需求」，在全部的論文中只出現在 3 篇論文中，而它的頻率只有 4 次。「檔案管理」只出現在 2 篇論文中，頻率也只有 2 次。這種關鍵詞在論文題名或摘要中出現頻率過低的現象，值得進一步研究。其次，在進行詞語抽取時，我們假定有效的詞語長度在 2 到 6 字間，然而一些關鍵詞的長度超過這個預設長度，所以也無法抽取出來，如「圖書館自動化系統」、「圖書資訊學教育」等等。最後有些重要的關鍵詞，雖然它們的出現頻率高，但許多作者在論文摘要中皆會出現這些語語，造成它們的平均頻率太低，所以也容易被過濾掉，比方說像「網際網路」在圖書與資訊期刊的論文中相當廣被使用，但在多數論



文中與論文主題較不相關，所以它的平均頻率不高。上述兩種情形，在日後在改進詞語抽取方法可以加強。

表三則是檢視所抽取出的詞語是否為關鍵詞的結果，我們找出平均頻率大於或等於 6 的詞語，並依據它們平均頻率加以排序。從表中，平均頻率高的詞語約有半數曾被使用為關鍵詞，而

剩下的詞語事實上也和論文的主題有很高的相關性，比方說，「圖書採訪」（黃宗忠，民 86）、「紙張」（陳瑞文，民 89）、「視障」（吳美美，民 90）、「行銷」（謝寶媛，民 87）、「焦慮」（林麗娟、鄭靜欣，民 87）（鍾思瑩，民 87）等都和論文主題有很高的相關性。

表三：圖書與資訊期刊相關詞語是否為關鍵詞的結果

| 詞語 | 平均頻率 | 是否為關鍵詞 | 詞語 | 平均頻率 | 是否為關鍵詞 |
|--------|------|--------|------|------|--------|
| 口述歷史 | 18 | 是 | 閱讀 | 6.8 | 否 |
| 圖書採訪 | 13 | 否 | 檔案 | 6.58 | 是 |
| 報紙 | 10 | 是 | 資訊素養 | 6.25 | 是 |
| 口述歷史計劃 | 10 | 否 | 元資料 | 6 | 是 |
| 閱選訂購 | 9 | 是 | 外包 | 6 | 是 |
| 紙張 | 8 | 否 | 政府文獻 | 6 | 是 |
| 視障 | 8 | 否 | 成本 | 6 | 否 |
| 中學生 | 7 | 是 | 活字 | 6 | 否 |
| 行銷 | 7 | 否 | 印刷 | 6 | 否 |
| 圖書採訪學 | 7 | 是 | 報紙資料 | 6 | 否 |
| 焦慮 | 7 | 否 | 視障資源 | 6 | 是 |

從上面的分析裡我們可以得到如下的結論：雖然在本研究所利用的詞語抽取方法中，有許多關鍵詞由於出現頻率次數太低、字數太長或是平均頻率太低等因素，無法順利抽取出來。但我們所抽取出來的詞語與論文主題極為相關。

二、出現頻率較高的詞語

表四是圖書與資訊期刊前二十個出現頻率較高的詞語與它們的出現頻率。由於長度較短的詞語所代表的概念較廣泛，較常在論文中出現，所以在表四所呈現的高頻詞語中絕大部分是這類的詞語。從表中，我們可以觀察出在圖書與資訊期

刊中，論文作者、期刊編輯與讀者三方面所關心的問題與研究。從所抽取出來詞語主要包括了一類圖書館學的相關概念，如「圖書館」、「讀者」、「期刊」、「館員」、「組織」、「大學圖書館」，反映了圖書與資訊期刊在圖書館管理與營運方面的重視，這也是反映了圖書資訊學以圖書館為中心的高度實務導向。其次，諸如代表資訊服務方面概念的詞語也常出現於論文中，如「資訊」、「使用」、「資料」、「檢索」、「資源」等等詞語，顯示了圖書資訊學從以圖書館典藏與管理，走向重視讀者獲取與利用資訊的趨勢。再者，近年來圖書資訊學的研究問題與資訊科技的進展有相當密切的關



係，所以在論文中出現了許多「系統」、「網路」和「電子」等等資訊科技方面的詞語。特別提出說明的是圖書與資訊期刊中，還包含了教育相關概念的詞語，如「教育」和「課程」，在圖書與資訊期刊的論文中這些詞語用來表示兩種與圖書資訊學和圖書館相關的教育活動，一類是關於圖書資訊學領域本身的教育問題與研究，出現這些詞語的論文多以各國的圖書資訊學教育與課程設計為參考，反思國內的現況（王梅玲，民 86）（林素甘，民 90a）（林素甘，民 90b），這顯示了圖書與資訊學期刊內的作者所認為圖書資訊學正面臨了轉型與變革的時機，可以從其他各國得到借

鏡。另一類有關圖書資訊教育概念詞語的使用則與提升學生與民眾資訊素養相關的圖書館利用教育與課程設計相關，顯示了圖書館作為一種重要教育場所的思維（陳超明、鍾雪珍，民 86）（于第，民 90）。最後，「檔案」的高度被使用（薛理桂，民 86）（薛理桂，民 88）（薛理桂，民 90）（韋慶遠，民 85a）（韋慶遠，民 85b）（莊樹華，民 86）（賴麗雯，民 89），也表示了政治大學圖書資訊學刊對檔案學相關研究的重視。當然以上的說明只是針對抽取出的詞語作大致的觀察，若要得到更進一步的推論，需要針對這些詞語進行更詳細更嚴謹的分析。

表四：圖書與資訊期刊前二十個出現頻率較高的詞語與它們的出現頻率

| 序 號 | 詞 語 | 出 現 頻 率 | 序 號 | 詞 語 | 出 現 頻 率 |
|-----|-----|---------|-----|-------|---------|
| 1 | 圖書館 | 456 | 11 | 檔案 | 79 |
| 2 | 資訊 | 299 | 12 | 電子 | 72 |
| 3 | 研究 | 218 | 13 | 教育 | 71 |
| 4 | 服務 | 135 | 14 | 資源 | 61 |
| 5 | 使用 | 122 | 15 | 讀者 | 58 |
| 6 | 資料 | 108 | 16 | 期刊 | 54 |
| 7 | 系統 | 103 | 17 | 館員 | 51 |
| 8 | 網路 | 99 | 18 | 課程 | 49 |
| 9 | 檢索 | 92 | 19 | 組織 | 42 |
| 10 | 大學 | 87 | 20 | 大學圖書館 | 39 |

從上面的推論與觀察中，所抽取出的詞語中，雖然長度較短的詞語較容易抽取且出現頻率較高，但長度較長的詞語所表示的概念較為特定

且與論文主題較相關，更容易分析出作者在論文想表達的概念。因此，在表五中，我們列出了長度為四字或以上的前二十個高頻詞語。



表五：圖書與資訊期刊前二十個出現頻率較高的較長詞語與它們的出現頻率

| 序號 | 詞語 | 出現頻率 | 序號 | 詞語 | 出現頻率 |
|----|-------|------|----|--------|------|
| 1 | 大學圖書館 | 39 | 11 | 自動化系統 | 17 |
| 2 | 圖書館學 | 35 | 12 | 電子期刊 | 17 |
| 3 | 圖書資訊 | 32 | 13 | 圖書資訊學 | 15 |
| 4 | 資訊科學 | 27 | 14 | 數位圖書館 | 15 |
| 5 | 資訊素養 | 25 | 15 | 檔案描述 | 14 |
| 6 | 圖書館員 | 22 | 16 | 台灣地區 | 14 |
| 6 | 館際合作 | 22 | 17 | 研究方法 | 14 |
| 8 | 公共圖書館 | 21 | 18 | 圖書採訪 | 13 |
| 9 | 資訊檢索 | 19 | 18 | 都柏林核心集 | 13 |
| 10 | 口述歷史 | 18 | 20 | 利用教育 | 12 |

從表五的結果中，我們可以更明顯地觀察到上面所分析的現象。在這裡，屬於圖書館管理與營運概念的詞語有「大學圖書館」、「圖書館員」、「館際合作」、「公共圖書館」、「圖書採訪」；在資訊服務及資訊科技方面的相關詞語則有「資訊檢索」、「自動化系統」、「電子期刊」、「數位圖書館」、「都柏林核心集」等等；有關圖書館利用教育的是「資訊素養」、「利用教育」等等詞語。其中「口述歷史」是比較特殊的詞語，因為這個詞語只出現在一篇論文中（陳秀慧，民 88），但是在這篇論文的題名和摘要便出現了十八次，顯然這篇論文的主題與口述歷史間有非常高的相關性。

四、詞語所出現的論文篇數

表六為統計各抽取出來的詞語所出現論文篇

數的結果。這個結果與表四所列出的詞語相當接近，表示這些出現論文篇數較多的詞語也往往是高頻的詞語。再這裡值得注意的是「調查」、「圖書館學」、「歷史」等雖然沒有出現在前二十個高頻的詞語，但出現論文篇數較多的詞語。調查是論文中常使用的詞語，但並非與論文主題的相關性很高。圖書館學則是圖書資訊學學科領域內的重要術語之一，而它所表示的概念較其他詞語來得廣泛。歷史這個詞語的出現除了與檔案學相關的研究（韋慶遠，民 85a）（韋慶遠，民 85b）（莊樹華，民 86），顯然與圖書與資訊期刊中的論文重視圖書館學（楊美華，民 88）（宋雪芳，民 88）（盧秀菊，民 89）（陳百齡，民 90）或資訊科技應用（陳亞寧，民 90）的歷史沿革，以借古鑑今有關。



表六：圖書與資訊期刊前二十個出現篇數較多的詞語與它們的出現篇數

| 序 號 | 詞 語 | 出 現 篇 數 | 序 號 | 詞 語 | 出 現 篇 數 |
|-----|-----|---------|-----|-------|---------|
| 1 | 圖書館 | 114 | 11 | 電子 | 29 |
| 2 | 資訊 | 76 | 12 | 教育 | 26 |
| 3 | 研究 | 74 | 13 | 資源 | 25 |
| 4 | 服務 | 55 | 14 | 讀者 | 24 |
| 5 | 使用 | 50 | 15 | 組織 | 20 |
| 6 | 資料 | 47 | 16 | 調查 | 17 |
| 7 | 網路 | 42 | 17 | 館員 | 16 |
| 8 | 系統 | 31 | 18 | 大學圖書館 | 15 |
| 9 | 檢索 | 30 | 18 | 圖書館學 | 15 |
| 9 | 大學 | 30 | 20 | 歷史 | 14 |

肆、結 論

本論文利用出現頻率次數、前後接字的複雜度與平均頻率等等統計資訊來抽取論文中的詞語，並利用抽取出來的詞語進行期刊內容的初步分析。在對政治大學圖書與資訊期刊於 1996 到 2001 年所出版的 164 篇論文進行詞語抽取後，我們可以得到以下的結論。首先，利用本研究所提出的方法可以有效地從論文的題名與摘要中抽取詞語，所抽取出的詞語不僅是在圖書與資訊期刊論文中的高頻詞語，同時它們所代表的概念與論文主題有極高的相關性。在進行詞語抽取後，可

以發現所抽取出的詞語很多是圖書館相關概念，其次資訊使用研究、資訊科技、圖書資訊學教育、圖書館利用教育和檔案學等等，也是圖書與資訊學期刊論文的研究主流。

對本研究的進一步的相關探討，除了持續改進詞語抽取的方法之外，如何利用詞語抽取的結果作為基礎，進行圖書資訊學領域論文發表情形的探討，提出更有效的分析方法，甚至擴及其他領域的研究，也正是目前我們所努力的目標。

(收稿日期：2002 年 5 月 6 日)

註 釋

註 1：在語言學上，詞(words)是構成句子的最小意義單位，亦即在句子的構成中，句中的每個詞表示了一個概念或構成句子的功能，對於一些意義上較為複雜的概念，我們則以多個詞組成一個詞組來表達。由於在中文裡，詞的界限不明顯，使得詞與詞組間不易區分。雖然詞與詞組的區別在語言學的研究上有其重要性，但在本論文的研究中，兩種語言單位都是句子中代表了某種特定意義的符號，無須再加以區分，因此本論文以用詞語來代表詞和詞組兩種語言學上的單位。



參考書目

- Borgman, C.L. (1990). Scholarly Communication and Bibliometrics. Newbury Park : Sage.
- Budd, J. M. and Seavey, C. A.. Characteristics of Journal Authorship by Academic Librarians. College and Research Libraries, 51, 463-470.
- Callon, M., Law, J. and Rip, A. (1986). Mapping the Dynamics of Science and Technology. London : Macmillan Press.
- Chen, K. J. and Liu, S. H. (1992). Word Identification For Mandarin Chinese Sentences. Proceedings of the 14th International Conference on Computational Linguistics.
- Egghe, L., Rousseau, R. and Hooydonk, G.. Methods for accrediting publications to authors or countries: Consequences for evaluation studies. Journal of the American Society for Information Science, 51(2), 145-157.
- Fromkin, V. and Rodman, R. (1988). An Introduction to Language (4th ed.), New York : Holt, Rinehart and Winston.
- Garfield, E. (1979). Citation Indexing—Its Theory and Application in Science, Technology and Humanities. New York : John Wiley & Sons.
- Hood, W. W. and Wilson, C. S.. The literature of bibliometrics, scientometrics, and informetrics. Scienometrics, 52(2), 291-314.
- Koehler, W. et. al., A profile in statistics of journal articles: Fifty years of American Documentstion and the Journal of the American Society for Information Science. Cybermetrics <<http://www.cindoc.csie.es/cybermetrics/articles/v4i1p3.html>>
- Leydesdorff, L.. Why words and co-words cannot map the development of the sciences. Journal of the American Society for Information Science, 48(5), 418-427.
- Tsay, M. Y., Jou, S. J. and Ma, S. S.. A Bibliometric Study of Semiconductor Literature. Scientometrics, 49(3), 527-545.
- Wang, P. L. and Soergel, D.. A cognitive model of document use during a research project. Study I. Document Selection. Journal of the American Society for Information Science, 49(2), 115-133.
- White, H. D. and McCain, K. W. Bibliometrics. Annual Review of Information Science and Technology, 24, 119-185.
- Wilson, C. S.. Informetrics. Annual Review of Information Science and Technology, 34, 107-249.
- Wormell, I.. Informetric analysis of the international impact of scientific journal: How ‘international’ are the international journals?. Journal of Documentation, 54(5), 584-605.
- Zhang, H. Q.. More authors, more institutions, and more funding sources—Hot papers in biology from 1991 to 1993. Journal of the American Society for Information Science, 48(7), 662-666.
- Zitt, M. et al. Territorial concentration and evolution of science and technology activities in the European Union:



- A descriptive analysis. *Research Policy*, 28(5), 545-562.
- 林麗娟、鄭靜欣（民 87）。圖書館自動化與讀者焦慮類型。圖書與資訊學刊，25 期，頁 16-23。
- 林素甘（民 90a）。澳洲圖書資訊學教育。圖書與資訊學刊，37 期，頁 79-94。
- 林素甘（民 90b）。紐西蘭圖書資訊學教育。圖書與資訊學刊，39 期，頁 65-77。
- 賴麗雯（民 89）。檔案描述編碼格式在中文檔案之應用--以國史館及中研院近史所檔案館為例。圖書與資訊學刊，35 期，頁 55-81。
- 盧秀菊（民 89）。英美編目規則原則之探討。圖書與資訊學刊，32 期，頁 16-44。
- 黃宗忠（民 86）。論圖書採訪學。圖書與資訊學刊，22 期，頁 26-38。
- 謝寶煥（民 87）。行銷圖書館與資訊服務。圖書與資訊學刊，27 期，頁 40-54。
- 薛理桂（民 86）。檔案鑒定模式：「黑盒子鑒定模式」淺探。圖書與資訊學刊，20 期，頁 39-45。
- 薛理桂（民 88）。檔案描述編碼格式（EAD）之發展與實施。圖書與資訊學刊，28 期，頁 49-62。
- 薛理桂（民 90）。中文檔案描述規則之擬訂--基於國際檔案描述標準(ISAD(G))。圖書與資訊學刊，36 期，頁 1-14。
- 莊樹華（民 86）。中央研究院近代史研究所檔案館管理概述。圖書與資訊學刊，22 期，頁 86-92。
- 鍾思瑩（民 87）。政大學生圖書館焦慮之探討。圖書與資訊學刊，25 期，頁 73-91。
- 陳百齡（民 90）。新聞媒體裡的圖書館。圖書與資訊學刊，36 期，頁 15-27。
- 陳克健（民 88）。Lexical analysis of Chinese-- Difficulties and Possible Solutions。中國工程學刊，22 卷 5 期，頁 561-571。
- 陳秀慧（民 88）。「創造資訊」的圖書館：口述歷史之應用。圖書與資訊學刊，30 期，頁 82-93。
- 陳超明、鍾雪珍（民 86）。圖書館在「研究方法與論文寫作」課程中的角色。圖書與資訊學刊，22 期，頁 10-19。
- 陳瑞文（民 89）。圖書館界與無酸紙的發展。圖書與資訊學刊，35 期，頁 93-103。
- 陳亞寧（民 90）。從資訊科技的應用發展探討圖書館資訊服務的策略管理--以中央研究院計算中心為個案研究。圖書與資訊學刊，36 期，頁 38-46。
- 宋雪芳（民 88）。網路時代 House Journal 變革之研究。圖書與資訊學刊，31 期，頁 24-35。
- 楊美華（民 88）。美國圖書館暨資訊科學委員會與我國教育部圖書館事業委員會之比較。圖書與資訊學刊，28 期，頁 37-48。
- 吳美美（民 90）。視障資源整合的必要。圖書與資訊學刊，38 期，頁 18-31。
- 韋慶遠（民 85a）。大陸地區現存明清檔案的分佈及其史料價值。圖書與資訊學刊，17 期，頁 13-19。
- 韋慶遠（民 85b）。簡介大陸地區的檔案事業和與民國史編纂的關係。圖書與資訊學刊，18 期，頁 16-24。
- 王梅玲（民 86）。圖書館與資訊科學專業教育之探討。圖書與資訊學刊，21 期，頁 38-56。
- 于第（民 90）。技術學院圖書館利用教育之現況調查研究。圖書與資訊學刊，38 期，頁 69-87。

