大學圖書館
7卷2期（民國93年9月） 頁191-205

資訊擷取技術之探討
A Survey of Information Extraction

蔣禮芸 **Li-yun Chiang**
國立台灣大學圖書館閱覽組組員
Readers Service Department, National Taiwan University Library
E-mail: liyun@ntu.edu.tw

【摘 要】

資訊檢索技術發展至今，已走向所謂的智慧型檢索技術，而資訊擷取技術即為其中之一。本文從資訊檢索、資訊檢索技術以及資訊服務的發展趨勢，說明資訊檢索的研究方向，並深入探討資訊擷取技術的意義、目的、應用與作法等項目。

【Abstract】

With the new techniques and methods, the development of information retrieval has entered a new stage. The first part of this paper examines the development of information retrieval and information service and the second part concentrates on the purpose, methods, and applications of information extraction.

關 鍵 詞：資訊擷取；智慧型檢索技術；自然語言處理技術；資訊檢索
Keywords：Information extraction; Intelligent information retrieval; Natural language process; Information retrieval

壹、前言

在今日這個資訊社會，由於資訊科技的快速發展，人們只需坐在電腦前，並連接上網際網路，便可很便利地取得全世界各式各樣的資訊。然而，隨著資訊數量大量增加，產生所謂「資訊爆炸」的現象，造成使用者「資訊負擔」，甚至引起「資訊焦慮」。我們都曾經有使用檢索引擎查詢資料的經驗，在鍵入查詢條件後，時常會找到上千、上萬筆的「相關」(主題相關)資訊，此時要從這麼多的檢索結果中，找尋眞正符合需求的資

訊便成爲一件相當重要的事，使用者必須逐筆進行相關判斷。因此，如何提高檢索結果的精確率，或藉由智慧型檢索技術 (Intelligent Information Retrieval，簡稱IIR) 的輔助，以節省使用者閱讀不相關資訊的時間與精力，便成爲資訊科學家們努力的方向。

目前，已有許多相關研究在討論利用電腦自動對蒐集到的資訊進行內容或概念的分析，提高電腦對訊息的理解，以減低誤引的情況，並可自動擷取重要的概念與主題，資訊擷取技術 (Information Extraction，簡稱IE) 即爲這些新發展的資訊處理技術之一。本文從資訊檢索、資訊檢索技術以及資訊服務的發展趨勢，說明資訊檢索的研究方向，接著深入探討資訊擷取技術的意義、目的、應用與作法等項目。

貳、資訊檢索的意義與發展趨勢

在進入本文的主題-資訊擷取之前，有必要先說明資訊檢索 (Information Retrieval，簡稱IR) 的意義與概念。資訊檢索一詞最早由Mooers於1950年所提出(註1)，其認爲資訊檢索是資訊提供者與使用者間的一種特殊溝通過程，而資訊在四度空間裡傳播，亦即在一般的空間座標上增加一個時間座標。當資訊提供者發出資訊後，並不立即傳給使用者，而是先儲存起來，等到日後使用者提出需求時，才傳給使用者他所需要的資訊。(註2)Mooers的定義可說是資訊檢索最初的定義，此一名詞的出現，加速資訊科學脫離圖書館學的範疇，亦是資訊科學走向獨立的關鍵。

資訊檢索的目的，在於運用現代化科技對資訊儲存、分類、索引、摘要、檢索、傳播作不斷的提昇，讓使用者可以在資訊檢索系統中鍵入任一關鍵字，而由系統從一堆未處理的原始資料或已分析整理過的資訊中進行各個欄位值的比對，並經過關聯文字綜合成大量有用的「潛在資料」(註3)；換言之，資訊檢索最主要的目的，即利用電腦強大的儲存與計算能力，幫助使用者由大量缺乏結構化 (Non-structural) 的資料中，快速找到所需資料。(註4)

卜小蝶曾於〈圖書資訊檢索技術〉一書中整理資訊檢索技術的發展趨勢，比較早期與近期研究資訊檢索技術的差異。她指出早期資訊檢索技術處理的大部分是文字型資訊，如今各式各樣的資訊型態相繼出現，資訊檢索系統必須能夠處理圖片、影像、聲音等多媒體的資訊；資訊內容上亦打破學術文獻的範疇，舉凡新聞、娛樂、消費、股票等，都可以電子化的方式儲存，供使用者查詢瀏覽。(註5)

若從資訊檢索技術與服務的層面，可發現過去的檢索技術係以文字比對爲基礎，大多利用統計的方式計算文件的相關性，最新的研究方向，則強調發展智慧型檢索技術，亦即結

臺灣大學圖書館期刊資料庫

合人工智慧技術，使檢索系統具備某種程度的智慧能力，使電腦由字串比對的層次提昇到概念分析。(註6) 此外，現今的資訊檢索技術亦反映出網際網路的特色，允許全球每日千萬人次上線檢索，並以全球廣域網路爲目標，服務的型態也轉爲對個人提供個人化 (Personalization) 的資訊服務，並整合大量的學科，如電腦科學、電機科學、語言學、認知心理學等。對使用者而言，服務的價值在於知識的溝通，使用者希望能透過資訊檢索系統得到所需的知識 (Knowledge)，而非單純的提供許多相關資訊 (Information)，此知識與資訊在層次上有所不同，雖然最終目的都在滿足使用者的需求。

參、資訊擷取技術的意義、目的與應用

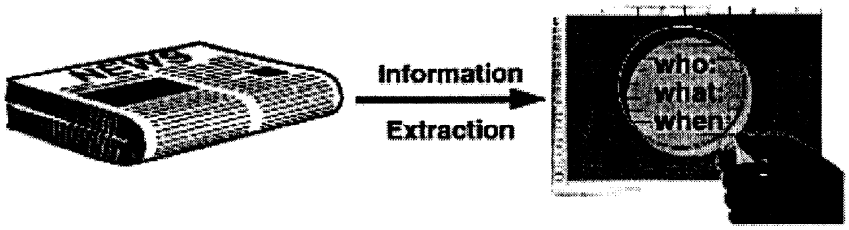
資訊擷取其實並非一個新的想法，而是一項新的技術。早在1964年 Wilks 在 " Test Searching with Templates" 一文中，便提出資訊擷取的概念，但當時並未討論到運用電腦來進行資訊的擷取。早期的研究較爲人所知的是1983年Cowie用一個標準的結構去描述植物與動物。(註7)

游基鑫指出，資訊擷取係從一堆資料中擷取出使用者想要的資訊，這些資料可以是文字、聲音、圖像或統計資料。(註8) 黃燕萍定義資訊擷取技術爲藉由轉換、過濾、淨化未處理過的資料成爲可以被分解及分析的資訊，這些資訊通常都是使用者想從原始資料中取得的，可以代表原始資料。(註9) 卜小蝶則認爲資訊擷取技術是爲了提昇資訊檢索認知而發展出的技術，是針對特定領域的文件進行深度分析，以擷取文件中的重要事實。(註10) 整合上述之定義可知，資訊擷取係從大量性質、類型各異的原始資料中，擷取出使用者想要的重要資料；由於必須一定程度地掌握文件內容及概念，因此資訊擷取技術相當依賴自然語言處理技術 (Natural Language Process) 的發展。

資訊擷取亦可稱爲訊息理解 (Message Understanding)，該詞明白揭示資訊擷取技術的意義，係從大量的文字資料中自動抽取特定訊息 (Particular Information)，以做爲資料庫存取 (Database Access) 之用的技術。圖一爲資訊擷取示意圖，該圖以報紙的新聞資訊爲例，說明資訊擷取的基本概念，係藉由對資訊的整理、組織程序，以及對訊息的理解，以萃取資訊內容的重要概念，並透過結構化的方式呈現。

由於資訊擷取技術的出現，乃是因應網路快速發展之後，線上資料的成長速度已經遠超人工的處理速度，所以有許多研究人員極力發展各種可行的方法來自動建構所需的資訊擷取系統(註11)，強調的是針對特定領域的文件進行深入的分析，因此資訊擷取系統的功能除了速度上的提昇外，還能夠掌握文件中的重要或特定訊

國家圖書館學術資訊典藏庫



圖一　資訊擷取示意圖

資料來源：Paul Jacobs, "Test Interpretation: Extracting Information," <http://cslu.cse.edu/HLTsurvey/ch7node5.html>（民國89年1月15日）。

息，將文件予以結構化，以提高檢索的效率與檢索結果的精確率，並減輕使用者的資訊負擔。如圖二所示，透過擷取網路上的新聞資訊，並以固定格式展現，使用者可以很快的掌握每篇新聞的重要人、事、物，省去使用者一篇篇去查閱所花費的時間。而經過資訊擷取系統處理過的資訊，將更容易予以分門別類，讓資訊的整理與組織更有系統，且更易於傳播與利用。

爲達到萃取資訊內容的目的，資

新聞文件：XXX
報導時間：XXX
報導地點：XXX
事件種類：XXX
人名：｛擷取出的人名｝
地點：｛擷取出新聞本文中的地名｝
時間：｛擷取出新聞本文中的時間｝

圖二　中文新聞文件資訊擷取格式

資料來源：黃燕萍，「中文社會新聞文件資訊擷取」（碩士論文，國立雲林科技大學資訊管理研究所，民國88年）<http://www.management.org.tw/MBA_WWW/8/index.htm>（民國89年1月18日）。

臺灣大學學術期刊資料庫

訊擷取技術對於所分析的文件必須有某種程度的剖析理解能力，因此非常仰賴自然語言處理技術的輔助。資訊擷取技術的主流研究，起源於1987年的訊息理解會議(Message Understanding Conference，簡稱MUC)，這個會議主要提倡利用自然語言處理技術，對文字資訊作更深度的剖析，以提高資訊檢索的認知程度。(註12) MUC除了具有研討會的性質，同時也是一個比賽(Competition)，藉由參賽的方式，讓相關的技術得以交流、發展。在1993年的MUC會議當中，資訊擷取的型式得到進一步的確認，主要是從新聞文章中，擷取有關恐怖份子活動的訊息，包括恐怖活動發生在何時、何地、有多少受害者、由那一個恐怖組織發起等，並將這些訊息填入有關恐怖活動事件的樣版(Template)中。之後的資訊擷取作法大致遵循這種方式，只是將樣版應用在不同的領域當中。(註13)

目前資訊擷取技術仍有許多問題待突破，要建立一個可運作的資訊擷取系統並不容易，大多數系統仍處於實驗階段，還未進入到商業使用的階段，目前所知的應用有以下幾方面：(註14)

一、健康照顧資訊分析

這部分的應用是藉資訊擷取系統從醫師的診斷、病人的症狀、各種檢查的結果、治療對策等資料中，擷取出有關的各項資料，製成病歷檔案，以協助相關健康照顧方面的研究；或將每位病人的資料予以分類，以支援保險給付的程序。

二、技術與學術文獻內容分析

可針對某一學科領域的文獻，擷取其重要的訊息或特定的概念，並按照此樣版進一步追蹤其他相關文獻。

三、犯罪活動整合分析

這裡所稱之犯罪活動整合分析，乃是指資訊擷取系統可對線上即時新聞、廣播、電視等分散的資料進行分析，藉由辨識出恐怖活動的相關性，以掌握並追蹤恐怖份子的行動，並協助警方列出可能的罪犯。

目前台灣地區相關的研究是在1994年由中央警察大學的謝慶欽所進行之〔資訊擷取技術應用於犯罪偵查之研究—以強盜案件爲例〕，該系統主要在協助警方過濾龐大之前科資料，以縮小偵查範圍，提高辦案效率，透過資訊擷取技術，可以自動搜尋比對相關之嫌犯。該系統是由文義分析程式與檢索子系統二者所構成。文義分析程式掃描偵查報告，將每一位嫌犯特徵化，並依此建構嫌犯剖析特徵之資料庫。檢索子系統則從嫌犯剖析之資料庫內讀取資料，並與刑案現場所發現之特徵核對，以擷取出相關的嫌犯資料。(註15)

臺灣大學學術期刊資料庫

四、商業競爭分析

這裡主要是指在商業經濟與投資方面的應用，亦即透過資訊擷取系統分析報章雜誌上的文章，以辨識出相關產品與服務的合併情形、共同投資的狀況，或是合伙人的詳細資料等。

若比較資訊擷取系統與資訊檢索系統的差異，可發現資訊擷取系統的建置較資訊檢索系統更爲困難，亦需要更多的專門領域知識。因爲在牽涉的範圍上，資訊擷取系統較資訊檢索需要更廣泛、更深入的學科知識與規劃，一個資訊擷取系統必須具備所謂的知識樹 (Knowledge Tree)，以有效分析資訊內容概念，目前較可行的方式是用漸進、累積的方式擴展各領域 (Domain) 的知識，再逐漸擴展爲一完整的知識體系。而在系統的處理過程方面，資訊擷取系統較資訊檢索系統需要更大量的計算。至於系統的表現，一般來說資訊擷取可獲得較資訊檢索更高的精確率，在使用上也較有效率，因爲資訊擷取系統可減低使用者閱讀不相關資訊的機會。（註16）

另一方面，資訊擷取與資訊檢索也可以從以下幾方面加以結合：（註17）

1. 資訊檢索的技術可應用於資訊擷取系統中，對大量的文件進行前置處理，將文件轉變爲處理過的格式。
2. 資訊擷取技術對文件、字詞、語義的分析，可協助資訊檢索系統對索引詞彙 (Index Term) 的辨識。
3. 資訊擷取技術可將資訊檢索的輸出結果摘要，變成較連貫且一致的格式輸出，亦即將輸出結果結構化。

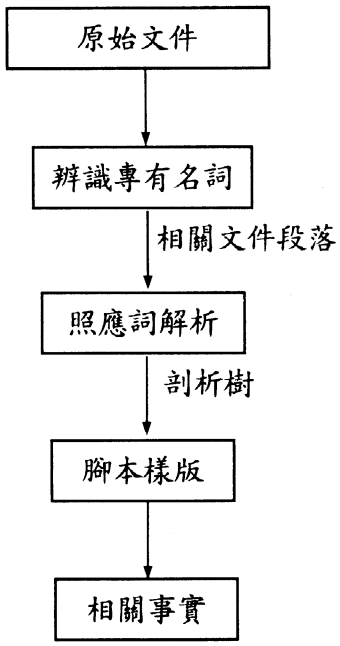
肆、資訊擷取的作法

在討論資訊擷取技術的作法之前，必須先說明文件的組成分子。一份文件的組成份子包括內容 (Content)、格式 (Format) 與結構 (Structure)。內容是一份文章所要表達的眞正含意，格式是文件呈現的方式，結構則表示了這份文章內容順序上的關連性。對於同一類的文件而言，內容可能不盡相同，格式也不一定會一樣，但卻會有類似，甚至完全相同的結構，而這個結構會把整份文件區分爲一個個獨立而有意義的資料區塊，利用這樣的結構特性，便可協助系統去擷取同一類文件中有意義的資料。（註18）

根據訊息理解會議所訂定的工作項目，我們可將資訊擷取的工作分爲辨識專有名詞 (Name Identification)、照應詞解析 (Coreference Resolution) 與腳本樣版 (Scenrio Template) 三項工作項目，其關係如圖三所示。專有名詞辨識是企圖擷取文件中的專有名詞；照應詞解析是串連專有名詞及其對應的代名詞；腳本樣版則是依照預先訂定的樣版，由文件中擷取相關的資訊填入樣版的欄位。這三項工作具有層級關係，必須先辨識專有名詞，才能進入照應詞解析的部分，而後進行腳本樣版的記錄。（註19）此外，這三項工作皆必須依賴自然語言處理

臺灣大學學術期刊資料庫

技術，才能夠正確地擷取資訊，文後將針對自然語言處理技術中與資訊擷取相關的技術加以說明。



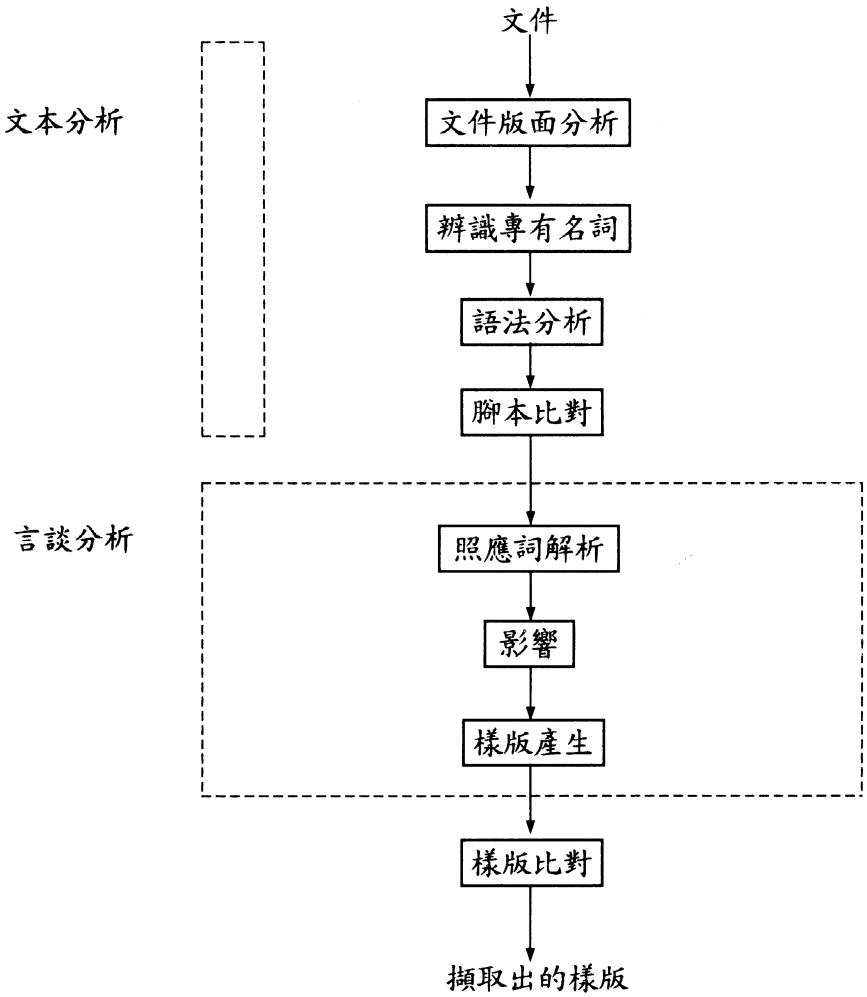
圖三　資訊擷取技術的作法
資料來源：作者自行整理

伍、資訊擷取的相關技術

根據前述資訊擷取的作法，系統必須要能夠理解或辨識個別的事實，因此便需要藉助自然語言處理的相關技術，包括文件版面分析(Layout Analysis)、分詞模組(Word Segmentation Module)、語彙分析(Lexical Analysis)、語法分析(Syntactic Analysis)以及語義分析(Semantic Analysis)等五大部分(註21)，茲針對上述各項技術加以說明。

Pazienza指出一個資訊擷取系統可以概分爲兩大部分，第一部分是透過所謂的「本文分析」(Local Text Analysis)直接從文件中擷取出單獨的事實(facts)，第二部分「言談分析」(Discourse Analysis)則是整合這些事實成爲更大量的事實或新的事實；最後則是將相關的事實轉變爲系統要求的輸出格式(註20)，其結構圖請見圖四。

臺灣大學圖書館 期刊資料庫



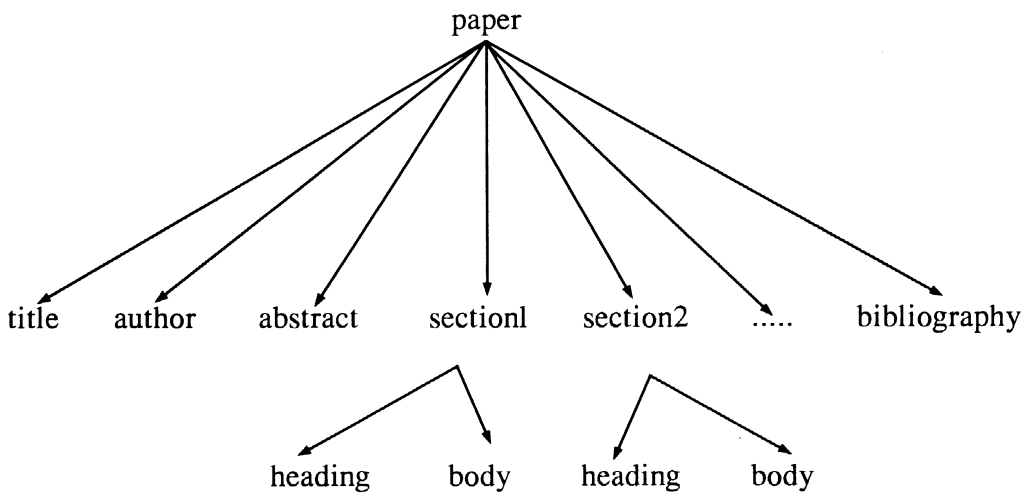
圖四　資訊擷取系統結構圖

資料來源：Ralph Grishman, "Information Extraction: Techniques and Challenges," in Information Extraction, ed. Maria Teresa Pazienza (New York: Springer-Verlag Berlin Heidelberg,1997), 11.

一、文件版面分析(Layout Analysis)

一份文件通常由文字、標題、表格、圖型等組成，以學術論文爲例，其結構如圖五所示，具有標題、作者、摘要、與章節等部分。

國家圖書館學術期刊資料庫



圖五 學術論文的結構

資料來源：莊水龍，「助理型軟體資訊擷取技術」(碩士論文，國立台灣大學資訊工程學研究所，民國88年)，頁7。

此外，學術論文在版面配置上，也常有一定的格式。如圖六所示，通常一篇學術文獻的開頭一定有文章的題名、作者、服務單位、摘要等資訊，而正文的排列方式雖因不同期刊而有不同，但大多以二欄的型式爲主。因此，在處理這類文件時，文件版面分析模組必須能夠區分文件的結構區塊，辨別文字、表格與圖形，並串聯文字的部分，交由後續的語言處理模組。

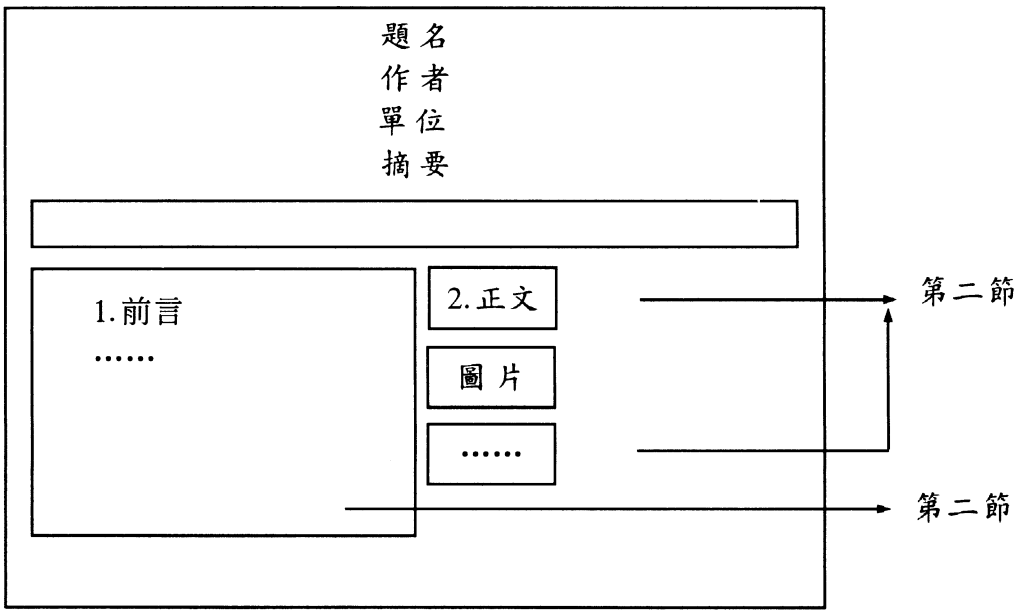
由於文件的版面格式變化很大，期刊論文、會議論文、雜誌、報紙等各有不同的形式，因此文件版面分析必須針對不同類型的文件，採用相對應的知識，故必須依賴描述各類型文件的知識庫爲基礎。(註22)

二、分詞模組(Word Segmentation Module)

分詞模組主要是進行斷句的工作。以中文而言，在斷句時常會遭遇到下列的困難：(註23)

1. 中文詞與詞之間不像西文有明顯的分隔符號。
2. 中文原本沒有詞的概念，也沒有詞的確切定義。
3. 中文構詞及句法千變萬化，使斷詞更顯困難。
4. 缺乏一完整、確切，且適用於電腦

臺灣大學學術期刊資料庫



圖六　學術論文版面結構

資料來源：陳光華，「資訊的組織與擷取」，圖書館學刊12期（民國86年12月），頁134。

系統的中文辭典。

5. 文字用語無法涵蓋口語的資訊。

6. 新詞不斷增加中。

除了上述的困難外，分詞模組還必須能夠辨識專有名詞。解決的辦法可分爲三大類，其一是長詞優先系統，其二是以規則式系統，第三類則是以統計爲本的機率式系統；一般而言，無論那一種分詞系統，都必須使用辭典作爲分詞的依據。(註24) 目前中文系統較常使用的辭典有中國大陸編的現代漢語詞頻統計，以及中央研究院詞庫小組所編的「中研院平衡語料庫」。

㈠ 長詞優先的系統

這三種斷詞方法中，以長詞優先最爲簡單易懂。其施行步驟如下：

1. 必須要有一個專門爲長詞優先所製作的辭典。這樣的辭典在製作時，

臺灣大學圖書館

要考慮出現頻率與長度最長兩項條件。

2. 斷詞時，一次取出固定字數的字組，並記錄目前的指標。
3. 將此字組與長詞優先辭典比對，以最長的詞爲優先考慮對象，若是符合則將此字組視爲一個詞，並取出下一個固定字數字組；若不符合，則將此字數減一，繼續比對到字數只剩下一，並改變目前取出字組的指標。
4. 取出下一個固定字數字組，繼續步驟2。

這個方法的缺點是忽略了長詞在句子集合中的重要性，例如「小明從前門進來」，最正確的斷詞結果應該是小明 從 前門 進來，但如果依照長詞優先斷詞的規則，系統會處理成小明 從前 門 進來。(註25)

㈡ 規則式系統

規則式系統是根據語言的文法規則而設計，只要將文法規則建立完備，就可依照一般電腦科學中編譯器設計概念，將每個詞取出來。但這種方式的限制在於文法規則很多，且常會有歧異的情形出現，要從衆多的文法規則中找出適合目前句子的規則並不容易。(註26)

㈢ 機率式系統

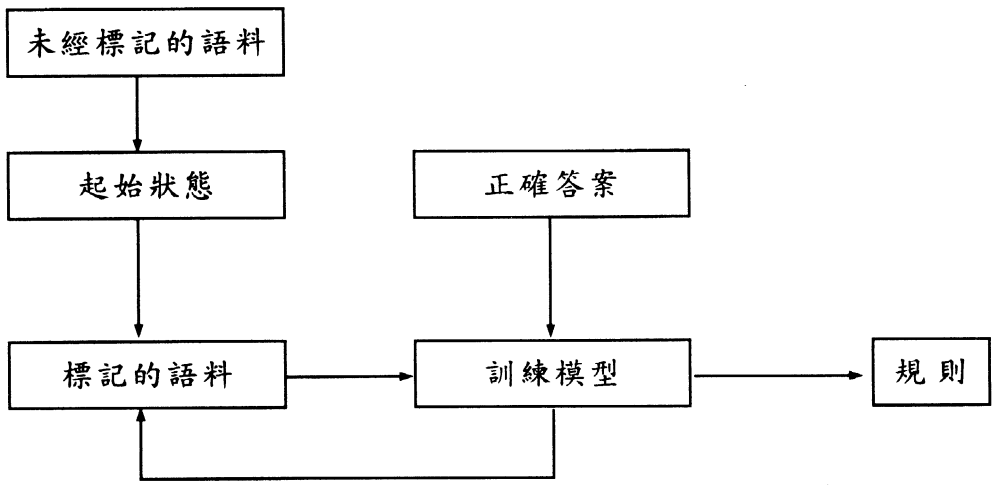
機率式系統最主要的精神在於轉化一般口語規則成爲語料機率分佈資訊。在斷詞時，必須不斷地判斷各種斷詞情形的可能性，綜合各種情形後，出現機率最高的字組，便被視爲是最符合一般口語能接受的條件。

常見的作法是將文句中相鄰的字，可以是雙連（Bigram）或三連（Trigram）的字組，在辭典內查詢其機率值，機率值乘積最大者，此字組在機率的觀點上便是最佳的組合。(註27)

三、語彙分析(Lexical Analysis)

語彙分析主要是爲詞彙加上詞類標記，簡單地說便是分辨該詞是屬於名詞、動詞、形容詞還是副詞等，以便進行更高階的處理。語彙分析的作法有規則式、統計式與混合式三種作法。規則式作法以卜利爾(Eric Brill)的系統最爲著名，他使用一份已經加上詞類標記的語料庫作爲訓練語料庫(Training Corpus)，並作爲建構模型之用，其程序如圖七所示。第一步先計算語料庫中每個詞彙的詞類，及其相應的詞類頻率，接著便進入起始階段，以最常出現的詞類作另一份語料詞彙的詞類，由於部分語料也已經經過標記，因此可以比較起始階段的正確率，並產生錯誤型態。這些錯誤型態其實可視爲一種規則，在下次遇到同樣狀況時，可依照錯誤型態加以修正。最後不斷重覆前述的程序，直到沒有新的錯誤型態產生，或正確率達到事先設定的標準爲止。一旦規則建立完成，便可依據同樣的方法處理陌生的句子。(註28)

臺灣大學學術期刊資料庫



圖七　卜利爾的訓練程序

資料來源：陳光華，「資訊的組織與擷取」，圖書館學刊12期(民國86年12月)，頁136。

統計式的作法以喬齊(Kenneth Church)最早提出，其使用詞類的機率與詞類的雙連機率，決定句子中每一個詞彙的詞類，其計算模型可用以下的數學式表示：

這些機率可以由訓練語料庫取得，並依統計模型計算所有可能的組合機率，然後決定機率最大者爲解答。混合式的作法則是融合規則式與統計式的優點，以塔帕奈勒與凡弟萊勒(Tapanainen and Voutilainen)所提出的作法最爲出名，正確率可達到98%，但缺點則是要付出較長的計算時間。(註29)

$$\prod_{l=1}^{n} P(w_i|t_i) \times P(t_{i+1}|t_i)$$

$P(w_i|t_i)$：為已知詞類為 ti 的情形下，其詞彙為 wi 的機率。

$P(t_{i+1}|t_i)$：為已知當前詞彙為 ti，下一個詞類為 $ti-1$ 的機率。

臺灣大學學術期刊資料庫

㈠ **語法分析** (Syntactic Analysis)

語法分析會產生所謂的剖析樹(Parsing Tree)，是對文句作語法的分析，找出動詞、主詞、受詞等。文章的作者在撰寫文章時，用出現在該文章中的所有詞彙，來表達他的意念，而讀者也利用文章中的詞繪來重建作者創作時的心智活動。然而，並不是所有的詞彙都一樣重要。例如，若是將文件中的冠詞、副詞、介詞等刪除，我們仍然能夠知道這份文件大概的意思，這便說明了上述詞彙並不那麼重要。另一方面，動詞與名詞就扮演了相當重要的角色。此外，在中文的文件中，具有主詞-述詞-受詞的結構，而名詞往往扮演主詞和受詞，動詞則扮演述詞的角色，所以透過名詞和動詞的比對，即可看出句子是否談論相同內容，例如：

例 1 美國 (Nc) 期中 (Na) 選舉 (Na) 揭曉 (VE)

例 2 美國 (Nc) 期中 (Na) 改選 (VE) 與州長 (Na) 選舉 (Na) 的 初步 結果 揭曉 (VE)

以上兩句皆有"美國"、"期中"、"選舉"等名詞及"揭曉"這個動詞，經由這兩種詞性，我們便可以判斷這兩個句子是否在談論相同內容，其它詞性似乎較不爲重要。(註30)

這種抽取名詞和動詞的作法，便是所謂的部分剖析 (Partial Parsing)。這種部分剖析的策略相較於完全剖析 (Complete Parsing)，能夠辨識句子中的重要成分，卻不需花費太多的時間處理各種語法間結構上的關係，因此逐漸受到重視。(註31)

㈡ **語義分析** (Semantic Analysis)

由於文字充滿了各種歧異的現象，例如：同義字、一詞多義等問題，使用者在閱讀時，能夠從上下文判斷該詞眞正所指的意義爲何，但對電腦系統來說，就必須對詞意加以分析，才能決定到底使用者所指爲何。目前最常使用的作法是以辭典爲本，進行語義的分析。(註32) 在實做上英文最常使用的同義辭典爲朗文的同義辭典或 ROGET；中文則採用大陸學者所編的同義詞詞林。

陸、結語

從本文對資訊擷取技術的介紹，可發現資訊擷取技術用樣版萃取文件內資訊，與圖書館的分類編目工作相當類似，兩者皆是對資訊進行組織與整理，以便於傳播與利用。它們的差別在於由人工或是電腦進行概念萃取的工作。對於資訊的組織與整理，一直以來都是圖書館所秉持的專業知識，但是，由於出版品的大量發行，以及電子資源的出現，圖書館界已有探討利用電腦的協助來進行分類與編目的可行性。就筆者個人的看法，以人工方式對資訊進行分類、編目的品質是不容否定的，但是在資訊量過大的情況下，用人工的方式勢必難以追上圖書的出版量，更遑論電子資源的

部分了。但用電腦自動處理與組織資訊，現階段仍有許多待克服的問題，且成本亦會相對提高。

目前已有許多期刊都已發行電子版，網路上亦有相當多可供利用的電子文獻，圖書館可考慮藉助資訊擷取技術對這些電子資訊進行組織與整理的工作。雖然此技術仍在發展當中，但隨著技術愈趨成熟，以及系統表現的提昇，欲應用於圖書館將不再是問題，只是圖書館方面必須基於成本、效率等各方面進行考量，以決定是否採用。

本文從資訊檢索的概念進入探討資訊擷取技術，可以發現既具有層屬關係卻又各自發展。目前資訊檢索的領域有許多專門的研討會，各領域的研究人員似乎只專注於提昇該領域的層次，但要知道整個資訊檢索技術的突破，是需要融合、交流與激發的，而最重要的關鍵點，通常是在最根本的地方。因此，筆者認爲目前各項資訊檢索技術最終將走向整合，不僅是檢索技術的整合，還是各學科的整合，尤其是圖書館學對資訊的組織與整理概念，應該要能融入資訊檢索的研究之中；還有自然語言處理所依賴的辭典，其實與圖書館的索引典與標題表非常類似，而圖書館界對這方面的研究已有很長一段歷史，其經驗對資訊檢索領域的研究應有相當大的助益。

註釋

註1：Hans Wellisch, "From Information Science to Informatics : Terminological Investigation," Journal of Librarianship 4:3 (July 1972): 131.

註2：王崇德，資訊科學原理（台北市：農資中心，民國80年），頁11-12。

註3：黃燕萍，「中文社會新聞文件資訊擷取」（碩士論文，國立雲林科技大學資訊管理研究所，民國88年），<http://www.management.org.tw/MBA_WWW/8/index.htm> (18 Jan. 1990)。

註4：卜小蝶，圖書資訊檢索技術（台北市：文華，民國85年），頁9。

註5：同前註，頁8。

註6：同前註，頁24。

註7：Maria Teresa Pazienza, Information Extraction (New York: Springer-Verlag Berlin Heidelberg, 1997), 2-3.

註8：游基鑫，「中文資訊擷取環境建構與同指涉問題之研究」（碩士論文，國立台灣大學資訊工程研究所，民國89年），頁2。

臺灣大學學術期刊資料庫

註9：同註3。

註10：同註4，頁9。

註11：莊水龍，「助理型軟體資訊擷取技術」(碩士論文，國立台灣大學資訊工程學研究所，民國88年)，頁6。

註12：同註4，頁216-217。

註13：同註8，頁3。

註14：Wendy Lehnert, "Information Extraction," <http://www-nlp.cs.umass.edu/> (5 Jan. 2000)。

註15：謝慶欽，「資訊擷取技術應用於犯罪偵查之研究—以強盜案件爲例」(碩士論文，中央警察大學警政研究所，民國83年)。

註16：同註7，頁118。

註17：同前註，頁119。

註18：同註11，頁4。

註19：陳光華，「資訊的組織與擷取」，圖書館學刊12期(民國86年12月)，頁132。

註20：同註7，頁12-13。

註21：同註19，頁133。

註22：同前註，頁133-134。

註23：陳淑美，「財經新聞自動分類之研究」(碩士論文，國立台灣大學圖書館學研究所，民國81年)，頁29。

註24：顧皓光，「網路文件自動分類」(碩士論文，國立台灣大學資訊管理研究所，民國86年)，頁20-22。

註25：同前註，頁21。

註26：同前註，頁22。

註27：同前註。

註28：同註19，頁135-136。

註29：同前註，頁136。

註30：黃聖傑，「多文件自動摘要方法研究」(碩士論文，國立台灣大學資訊工程學研究所，民國88年)，頁27。

註31：同註19，頁137。

註32：同前註。

臺灣大學學術期刊資料庫