圖書館學刊 第十二期（民國86年12月），頁127~141

資訊的組織與擷取

Organization and Extraction for Information

陳光華

Kuang-hua Chen

國立臺灣大學圖書館學系暨研究所助理教授

Assistant Professor
Department and Graduate Institute of Library Science
National Taiwan University

【摘　要】

網際網路的發展使得資訊檢索的研究進入更具挑戰性的環境，然而資訊檢索系統通常僅僅告訴使用者有哪些相關的文件，而非真正提供使用者所需要的資訊。資訊擷取的研究則是進一步分析文件，依據預先定義的樣版取出特定的資訊。參照於圖書館以機讀編目格式描述藏品，資訊擷取系統所稱的樣版與機讀編目格式都可視為一種元資料格式，亦即是用於描述資料的資料。本文說明元資料與資訊擷取的關係，並討論如何藉由自然語言處理的語言分析技術有效協助使用者擷取所需要的資訊。

【ABSTRACT】

The development of Internet makes the researches on information retrieval more changeable. Actually, the so-called "information retrieval" is "text retrieval." It is necessary for users to find out the needed information from the retrieved texts. A higher-level task is information extraction, which extracts the information based on pre-defined templates. From the viewpoint of Library Science, these pre-defined templates are the metadata, which describes the collection of libraries in common. This paper discusses the relationships between metadata and information extraction and how natural language processing helps the task of information extraction.

關鍵詞 Keywords：

資訊檢索；資訊擷取；元資料

Information Retrieval; Information Extraction; Metadata

壹、前言

知識與資訊一直是人類進步所賴以為繼的動力，一旦吾人停止對於知識的渴望，個人的發展便停滯不前；若所有人類停止對於資訊的追求，人類文明的演進也將因此而中斷。資訊的生產消費過程中，有些人是屬於上游的生產者，有些人是屬於中游的生產者與消費者，有些人是屬於下游的消費者。在網際網路風起雲湧的時代，資訊生產與消費的行為有了極大的改變，幾乎所有人都可以扮演生產者與消費者的角色。一般人也可以生產資訊，透過網際網路將之散佈，供眾人消費、取用，資訊的通路成為一條康莊大道，不再受制於少數的資訊托拉斯。因此，網路上的電子文件是多如牛毛，資訊消費者以往是不容易取得資訊，而現在卻面臨資訊氾濫的現象，人們被資訊所淹沒，不知道什麼才是真正需要的資訊。如何才能有效協助讀者或使用者取得資訊呢？圖書館存在的歷史已有五千年（註1），在如此悠久的歷史中，或許有很多經驗與方法可以提供吾人參考與借鏡的。

圖書館長久以來一直扮演知識庫的角色，以美國國會圖書館為例，其館藏量已達九千萬件（註2），我國國家圖書館館藏量為一百四十八萬餘件（註3），而臺灣大學圖書館館藏量則已超越一百九十六萬餘件（註4）。圖書館如何協助讀者從如此巨量的館藏中選擇需要的、適切的圖書呢？一般而言，圖書館典藏的圖書資訊都經過某種層次的組織與整理，因而典藏品透過組織與整理呈現一種有序的狀態，使得讀者能夠有效定位（Locate）圖書資訊的所在。前述的組織與整理正展示於圖書館學發展過程中幾個重要的里程碑，如杜威分類法（DDC）、美國國會分類法（LLC）、美國國會標題表（LCSH）、中文圖書標題表、機讀編目格式（Machine Readable Catalog，簡稱MARC）、英美編目規則（AACR2），以及衍生的中國機讀編目格式（Chinese MARC）、中國編目規則等。不同的分類法、標題表以相異的切入角度描述圖書資訊的主題編目（Subject Catalog），編目規則規範如何進行圖書資訊的記述編目（Descriptive Catalog）；而MARC則記錄著主題編目與記述編目的資料。網際網路使得實體圖書館發展出虛擬圖書館的分身，典藏品也由紙本資料走向電子資料，資訊型態的不同以及大量資訊的累積造成取用的方式亦有所不同，然而有效地滿足讀者或使用者資訊的需求卻無二致。

另一方面電腦科學與資訊科學的學者專家也由其學科領域的觀點，發展出滿

足使用者資訊需求的作法，最明顯的例子即是所謂的搜尋引擎（Search Engine）、以及自訂分類架構的主題指引（Subject Directory）。然而更具挑戰性的任務卻是資訊擷取（Information Extraction，簡稱IE）的研究，以往吾人對於資訊檢索（Information Retrieval，簡稱IR）的理解是檢索系統送回許許多多系統認為相關的文件，至於使用者需要的資訊必須閱讀文件才能得知。資訊擷取的研究則希望由文件中擷取特定的資訊，而不僅僅是檢索出文件而已。

本文以下的部份將介紹元資料（Metadata）、資訊擷取、與自然語言處理（Natural Language Processing）之間的關係，並說明語言分析技術如何運用於資訊服務系統，以有效提昇系統的服務品質。第二節討論資訊的加值並解釋何謂元資料；第三節提出資訊檢索與資訊擷取在資訊服務層次的關係，並討論目前資訊擷取系統的服務績效；第四節則說明自然語言處理的相關技術；最後是簡要的結論。

貳、資訊加值與元資料

圖書館館藏資料都經過一定程度的加值處理，館員依據編目規則、標題表、館方政策等指導原則為館藏加註詮釋性資料，讓讀者或使用者有效地檢索館藏亦即進行編目分類的工作。以道林（Dowlin）所著的"The Electronic Library"一書為例，館員加註的詮釋性資料如圖一所示。前述的分類編目可以分為兩種：一為記述編目；一為主題編目。圖一的書名/ 作者、出版項、稽核項、叢書名、附註項、ISBN/價格等屬於記述編目，主要是記載藏品實際的資料，不必經由編目館員進一步的分析。至於標題與索書號內的分類號則屬於主題編目的範圍，編目館員必須分析藏品的內容，經過思考然後加註適當的標題與分類號。一旦館藏皆加註上述的資料，圖書館的讀者或是使用者可以透過編目卡片或OPAC 線上檢索系統，有效地檢索館藏資料。

事實上，前述的詮釋性資料即是元資料（註5），所謂的元資料也就是用於描述資料的資料（data about data）。人類的日常生活中元資料幾乎無所不在。掏起皮夾裡的身份證，吾人可發現其上記載有身份證字號、姓名、出生年月日、父母、出生地、戶籍資料等等，這些資料便是用於描述我們每一個體（Entity）的元資料，各種不同的政府機構、信用機構則依據元資料檢索個人犯罪、納稅、信用等等情況。除此之外，學生證、汽機車行照、駕照、信用卡、金融卡、貴賓卡、會員卡等等都是某種形式的元資料格式，用以記載各種不同目的的元資料。因此吾人可以發現，不同的個體（Entity）可能需要不同格式的元資料；在不同的使用需求下，相同的個體

也可能需要不同格式的元資料。再以道林所著的"The Electronic Library"一書為例，可以用圖二描述這本書。圖二是圖書館使用的機讀編目格式，也可以視為一種Metadata資料格式。圖一與圖二描述相同的藏品，其目的卻有所不同，顯然圖二的機讀格式並不是給人看的，這也是它被稱為「機讀」編目格式的原因；而圖一卻是相當輕易地為人所理解。

在網際網路的虛擬世界裡，電子文件繁不勝數，可以將網際網路視為龐大的虛擬圖書館，吾人為了解決資訊檢索的難題，也必須發展適當的元資料格式描述繁雜的電子文件，讓使用者真正得到網路時代帶來的好處。目前已然運作的元資料格式有美國政府出版品使用的GILS元資料格式（註6），然而基本上GILS是政府資源指引系統，其目的是為了協助人民檢索聯邦政府資源，GILS元資料格式則是該檢索服務系統採用的元資料。此外，美國聯邦地理資訊委員會（Federal Geographic Data Committee，簡稱FGDC）的數位地理元資料標準格式，其主要使用於地理資料的交換、查詢、散佈，並透過Z39.50協定（屬於OSI參考模型的第七層通訊協定）以電腦網路為作業的環境。（註7）另外深受眾人期待的都柏林核心集（Dublin Core）則是針對網際網路上眾多的電子文件而設計，提供1 3 個核心的元資料欄位，希望能夠有效描述電子文件的特性，使得網際網路的資訊檢索及資訊擷取等服務具備更好的品質。

由圖書館（無論是實體圖書館或是虛擬圖書館）經營的角度，必須提供讀者或使用者各式各樣的需求，因此用以描述文獻的元資料格式有許多不同的欄位作為使用者檢索的檢索點，若是考慮各類型藏品有其不同的需求，元資料的格式隨之不同，更不必提使用者看待藏品或是資訊時採取的個人觀點。所以僅僅為了如何有效描述資源，必須處理的狀況就已經非常複雜。目前已有一些自動辨識人名、地方名、組織名的系統（註9），可以協助吾人加註文獻資料的元資料，但是困難的地方，卻是這一類的系統如何適應不同的元資料格式，如何配合元資料格式適當地變更加註的元資料。因此，各領域的學者專家也逐漸瞭解元資料的確是解決資訊需求的重要課題，也有越來越多的研究人員投入元資料的研究。

書名/作者	The electronic library : the promise and the process / Kenneth E. Dowlin
主要作者	Dowlin, Kenneth E
出版項	New York, N.Y. : Neal-Schuman Publishers, c1984
稽核項	xi, 199 p. : ill. ; 23 cm
叢書名	Applications in information management and technology series
附註	Includes bibliographical references and index
ISBN/價格	0918212758 (pbk.) : $24.95
標題	Libraries -- Automation
	Information technology
索書號	Z678.9 D68 1984

圖一 館藏的加值處理

001		83021957 //r91
005		19911024125216.4
008		831004s1984 nyua b 00110 eng cam a
010		83021957 //r91
020		0918212758 (pbk.) :|c$24.95
040		DLC|cDLC|dDLC
050	00	Z678.9|b.D68 1984
082	00	025/.04|219
090		Z/678.9/D68/1984///1410222AL/1415924CL/1453410CL/1733896CF
091		TUL|bAL|bCL|bCL|bCF
095		TUL|dZ678.9|eD68|y1984|t095|bAL|c1410222
095		TUL|dZ678.9|eD68|y1984|t095|bCL|c1415924
095		TUL|dZ678.9|eD68|y1984|t095|bCF|c1733896
095		TUL|dZ678.9|eD68|y1984|t095|bCL|c1453410
099		TUL|d|e|y|f|t091|b|c|x|z
100	10	Dowlin, Kenneth E
245	14	The electronic library :|bthe promise and the process / |cKenneth E. Dowlin
260	0	New York, N.Y. :|bNeal-Schuman Publishers,|cc1984
300		xi, 199 p. :|bill. ;|c23 cm
440	0	Applications in information management and technology series
504		Includes bibliographical references and index
650		0 Libraries|xAutomation
650		0 Information technology
910		8'93 D#139 MCL

圖二 MARC格式

參、資訊擷取

資訊擷取是由文件中擷取事先預設所需的資訊；資訊檢索則是由文件集合中檢索相關的文件。因此，「資訊檢索」這個詞彙事實上誤導了吾人對於相關研究的認識，如果「必也正名乎」，應該使用「文件檢索」代替「資訊檢索」。資訊擷取可視為比資訊檢索更深一層的資訊服務。正如訊息理解會議（Message Understanding Conference，簡稱MUC）所說的，資訊擷取不僅僅辨識重要的個體，還必須決定個體之間的關係。然而因為資訊擷取工作的特殊性，所以到底擷取何種資訊是依資訊服務系統服務的範疇而定。以MUC會議歷年的主題為例，MUC-5會議處理的文件為聯合貿易行為以及微電子產品相關的文件；MUC-6則是有關管理層級變化的新聞報導。（註10）

MUC-6會議訂定的工作項目為：辨識專有名詞（Name Identification）、照應詞解析（Coreference Resolution）、腳本樣版（Scenario Template）等三項。專有名詞的辨識正如字面上的意思，企圖擷取文件中的專有名詞；而照應詞的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。吾人可以將這三項工作視為是有層級的關係，唯有專有名詞辨識完成，才能夠進行照應詞解析，而後進行腳本樣版的記錄。事實上，前述工作中有兩項（辨識專有名詞、腳本樣版）正如圖書館編目館員進行的分編工作一般，館員首先進行記述編目然後是主題編目，將所得的資料填入元資料格式的欄位（ＭＡＲＣ），前述的腳本樣版亦即吾人所稱的元資料格式，而所謂的資訊擷取就是此處的第三項工作。腳本樣版是屬性與對應值的集合，而資訊擷取系統則是針對不同的屬性由文件擷取適當的值填入腳本樣版。

目前，各國研究者提出的資訊擷取系統效能（Performance）不一，實作的方法也有其不同之處，表一摘錄各系統的精確率（Precision）與回現率（Recall）。（註11）專有名詞的辨識目前取得很好的成果，平均的回現率與精確率都在90％以上，照應詞的解析與腳本樣版仍然是相當的困難。

表一、資訊擷取系統的精確率與回現率

工作項目	回現率		精確率	
	平均	最高	平均	最高
專有名詞辨識	90%	96%	90%	97%
照應詞解析	66%	75%	76%	86%
腳本樣版	45%	47%	65%	70%

一套基本的資訊擷取系統是由分詞模組、語彙分析模組、語法分析模組所組成。當然不同的語言有其特殊的考量，而引進不同的處理模組，例如印歐語系的文件必須作字形（Morphology）的處理，而不必引入分詞模組；有時也必須引進特定範疇的知識以有效擷取特定的資訊。自然語言處理的相關研究早已發展出許多語言分析的技術，資訊檢索以及資訊擷取研究領域與自然語言研究領域交流方熾，各種的語言分析技術目前也廣泛運用於相關的資訊服務系統，下一節將初步介紹相關的分析技術。

肆、自動化技術

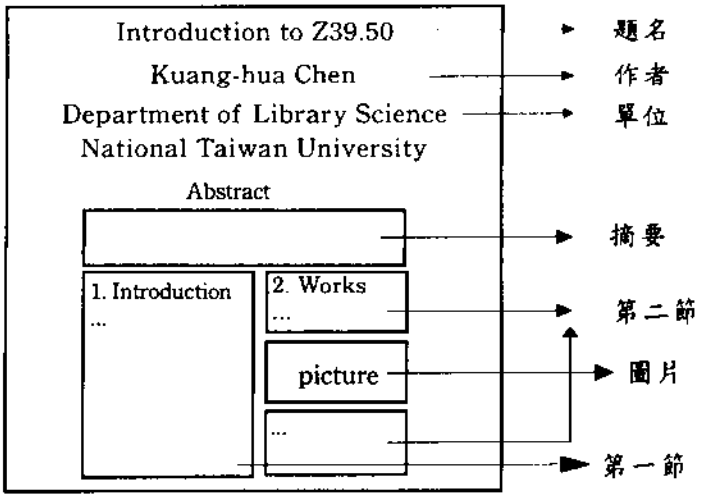
吾人已經瞭解純粹使用人力處理急遽成長的電子文件是緩不濟急、不可行的作法。如何使用電腦自動處理電子文件，或是協助吾人處理電子文件，才是面臨資訊爆炸時代比較恰當的作法。在採取任何自動處理程序之前，首先，必須先瞭解電子文件的特性。一般的電子文件主要是以書面語（Written Language）的形式出現，當然多媒體形式的文件越來越多，其中包含圖片、音訊、視訊，使得電子文件好似千面女郎，更具有吸引讀者的優勢。然而即便是多媒體的文件，其中仍然有很大的部份是文字，雖然有文獻指出「一圖賽千文」，但是文字仍具有圖片不可取代的說明功能。因之，本節主要說明如何引入自然語言處理的技術協助資訊系統的建構，以及這些技術如何有效提昇資訊系統的服務。

基本的資訊擷取系統可以包含以下幾個部份：文件版面分析模組（Layout Analysis Module）、分詞模組（Word Segmentation Module）、語彙分析模組（Lexical Analysis Module）、語法分析模組（Syntactic Analysis Module）、語義分析模組（Semantic Analysis Module），其功能分別敘述如下。

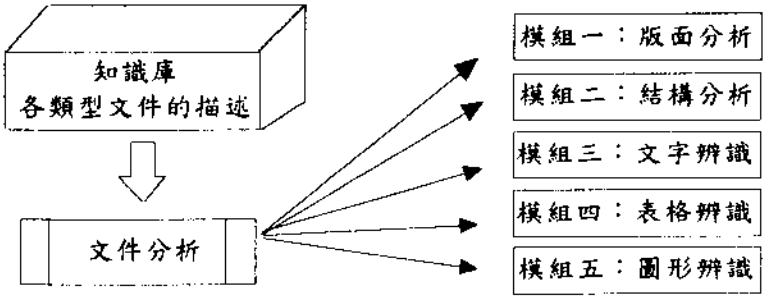
一、版面分析模組

文件通常由文字、標題、表格、圖形等等組成，圖三是學術論文版面構成的

一個例子。處理這類文件時，文件版面分析模組必須區分文件的結構區塊，然後串聯文字部份構成書面語，將其交由後續的語言處理模組；表格部份交由表格處理程序；圖形則交由圖形處理程序。由於文件的形式變化很大，期刊論文、會議論文、雜誌、報紙各有不同的形式，文件版面分析模組經常是以知識為基礎（Knowledge-based）的自動化程序，隨著不同類型的文件，採用相對應的知識，以適應性系統（Adaptive System）的方式進行文件版面分析的工作。（註12）圖四說明一個適應性文件分析模組可能具有的次模組。



圖三 學術論文版面結構



圖四 適應性文件分析

二、分詞模組

分詞模組主要用於將中文句子分成一個個詞彙，由於中文沒有詞間標記，分詞模組便成為任何以詞彙為基礎（Word-based）的自動化中文處理系統不可或缺的前處理程序。分詞並不如想像的簡單，舉個例子說明如下：（註13）

將劉大目的確實行動作了解釋

這個句子包含很多可能的二字詞（Two-character Words），例如：目的、的確、確實、實行、行動、動作、了解、解釋，但是只有一種分詞結果是正確的，如下所示。

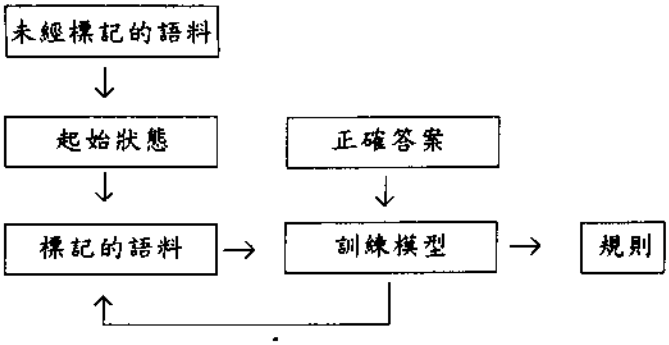
將◆劉大目◆的◆確實◆行動◆作◆了◆解釋

前述的例子還有一個困難的問題必須處理，亦即如何辨識劉大目是一個人名，而非三個單字詞（One-character Words）。國內研究自然語言處理的重要機構都已經研發中文分詞系統，以中央研究院詞庫小組的分詞系統而言，若不處理人名、地名、機構名，其正確率達99.77%（註14）；台灣大學資訊工程學系自然語言處理實驗室則將外文譯名與中文專有名詞（包括人名、機構名等等）的處理結合，也有很高的正確率。（註15）一般而言，分詞系統必須使用辭典作為分詞的依據，而辭典則依分詞系統的作法而有不同的格式。例如，採用長詞優先的分詞系統，僅需記載詞彙本身的資訊；而採用統計作法的分詞系統，則必須記載詞彙頻率、甚至雙連（bigram）、三連（trigram）的資訊。（註16）

三、語彙分析模組

語彙分析模組主要是為詞彙加上詞類標記，進行更高階的處理。若是以下列的句子為例，「蘇聯 總統 戈巴契夫 宣佈 ， 蘇聯 將 在 短期 內 自 古巴 撤出 一 支 為數 約 一萬一千 人 的 訓練 旅」，加上詞類標記後為「蘇聯(Nc) 總統(Na) 戈巴契夫(Nb) 宣佈(VE) ，(COMMACATEGORY) 蘇聯(Nc) 將(D) 在(P) 短期(Nd) 內(Ng) 自(P) 古巴(Nc) 撤出(VC) 一(Neu) 支(Nf) 為數(Na) 約(Da) 一萬一千(Neu) 人(Na) 的(DE) 訓練(Na) 旅(Na)」，其中括弧內為該詞彙的詞類。（註17）語彙分析的作法有規則式、統計式、混合式等三種作法，分別簡述如下。規則式的作法以卜利爾（Eric Brill）的系統最為著名，卜利爾使用一份已經加上詞類標記的訓練語料庫（Training Corpus）作為建構模型之用。（註18）首先計算語料庫中每個詞彙的詞類種類及其相應的詞類頻率，接著進入起始階段，以最常出現的詞類作為另一份語料詞彙的詞類，由於

該份語料也已經過標記，因此可以比較起始階段的正確率，同時產生錯誤型態。這些錯誤型態事實上可視為規則，亦即下次遇到相同的狀況應該依照錯誤型態修正。最後不斷重複前述的程序一直到沒有新的錯誤型態產生或是正確率達到預設的水準為止。整個程序可以用圖五表示。一旦規則建立完成，就能夠依據同樣的作法處理陌生的句子（Unseen Sentence）。



圖五 卜利爾的訓練程序

統計式的作法則以喬齊（Kenneth Church）最早提出（註19），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，其計算模型如以下數學式所示。

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

其中 $P(w_i|t_i)$ 為已知詞類為 t_i 的情形下，其詞彙為 w_i 的機率； $P(t_{i+1}|t_i)$ 為已知當前詞彙的詞類為 t_i ，下一個詞類為 t_{i+1} 的機率，這些機率值可以由訓練語料庫取得。這種統計模型計算所有可能的組合機率，然後決定機率最大者為解答。

混合式的作法則融合規則式與統計式的優點，這一類的系統以塔帕奈勒與凡弟萊勒（Tapanainen and Voutilainen）提出者最為著名，其正確率可達98%（註20），但是付出的代價是計算時間過長。

四、語法分析模組

語法分析（剖析，Parsing）會產生所謂的剖析樹（Parsing Tree），其目的在於瞭解各詞彙扮演的語法功能。分析使用者查詢問句的研究一直是資訊檢索、資訊

擷取等領域努力的目標，然而資訊檢索系統或是資訊擷取系統必須在很短的時間反應使用者的查詢，因此使用於前述系統的剖析策略必須非常快速。但是從事剖析技術的學者專家都瞭解，剖析自然語言事實上是非常困難的，一個十幾個字的句子很有可能會有上百個可能的剖析樹，進行完全的剖析（Complete Parsing）常常無法做到，因此部份剖析（Partial Parsing）的策略逐漸受到重視。辛島（Donald Hindle）於1983年提出的規則式語法分析系統Fidditch所產生的剖析結果經常不是剖析樹，而是所謂的剖析森林（Parsing Forest），也可以將Fidditch視為部份剖析系統。（註21）筆者於1993年提出的機率式部分剖析系統，則是將句子切分為一段段的「單層剖析樹」，例如 "When we are about to read a sentence, we usually read it chunk by chunk."會切分為[When we] [are about to] [read a sentence,] [we usually read it] [chunk by chunk]。（註22）這種部份剖析的策略能夠辨識句子中的重要成分，卻不必花用太多的時間處理各語法成分間結構上的關係，非常適用於資訊檢索或擷取系統的應用。

五、語義分析模組

文字充滿了各種歧義（Ambiguity）的現象，但是讀者通常都能夠瞭解所指為何，就以英文的bank為例，很可能是銀行或是河岸的意思，讀者可由句子中其他的文字或是前後文判斷。若是使用者進行檢索時使用了bank這個詞彙，檢索系統必須決定到底所指為何，然而目前上線的資訊檢索系統並沒有進行類似的處理，否則使用者應當可以得到更好的服務。

就目前文獻提出的作法可分為辭典為本（Dictionary-based）、案例為本（Example-based）與統計為本（Statistics-based）等策略。有些系統就以辭典（機讀辭典，Machine Readable Dictionary）中詞義排名第一者做為詞彙的詞義，然而這種處理方式等於是沒有進行任何處理。由於詞彙在不同的語境（Context）有其不同的意思，有學者使用詞彙前後各50個詞彙作為該詞彙的特徵向量（Feature Vector），由語料庫訓練而得各詞彙不同詞義的特徵向量，藉以判別相同詞彙再次出現時其詞義為何。筆者則是於1994年提出使用共容訊息（Mutual Information，簡稱MI）（註23）藉由句子中各個詞彙詞義的相互限制決定詞義的作法。（註24）現在更有學者嘗試建立加註詞義標記（Semantic Tag）的語料庫，提供研究人員進行辨識詞彙歧義的研究。（註25）

前述的自動化技術分別代表五種不同層次的文件處理程序。由於現今的文件型

式愈趨多元化，版面分析可以確立文件各部份文字結構上的關係，如果系統容許副主題（Subtopic）檢索或擷取（註26），版面分析是不可或缺工作。其餘幾項技術則視資訊系統建構者希望系統提供服務的層次而定。對於資訊系統建構階段，前述的五種技術只需要使用一次即可，然而，一旦系統建構完成，提供使用者檢索或擷取重要的資訊時，到底使用者鍵入的查詢必須處理到什麼層次，端視政策與時間以及當初系統建構時所處理的層次而定，如果系統建立時僅處理至語法分析階段，對於使用者的查詢處理至語義分析階段就沒有意義了。

伍、結語

無論是資訊檢索或是資訊擷取，其目的皆為滿足使用者的資訊需求，然而處於網際網路如此開放的環境，文件的數量迅速的增加而且文件型態變化極大，如何運用恰當的元資料描述文件，並有效提供資訊服務的品質是一項重要的課題。以目前的發展而言，元資料的研究逐漸受到大家的重視，今年（1997）的數位圖書館研討會（DL'97）也特別舉辦了「索引典與元資料」的會後會（Post-conference）。（註27）另一方面資訊擷取的研究也成為熱門的研究領域，今年3月於美國舉辦的第五屆應用自然語言處理研討會也特別開辦「建構資訊擷取系統」講習班。（註28）如何結合元資料與資訊擷取的研究，相信是圖書館學界與電腦科學界未來努力的目標。

這幾年來自然語言處理的研究已經展現其對於資訊檢索的重大影響，在原來偏向統計模式處理檢索的研究取向，加入了語言特性的元素，這可由SIGIR學術會議中有關語言處理的學術論文所佔的比重看出。（註29）對於資訊擷取系統而言，語言分析的技術更為重要，因為資訊的擷取不僅要辨識文件中的專有名詞，同時要解決照應詞問題，然後再建構專有名詞之間的關係，在這個過程中必須使用很多的自然語言技術。本文並就這些語言分析技術做了簡要的說明，提供有興趣的讀者作為初步的參考，至於更深入的細節可以閱讀相關的文獻。

註 釋

註 1：胡述兆、吳祖善合著，*圖書館學導論*（台北市：漢美，民國78年），頁3。

註 2：國立編譯館主編，*圖書館學與資訊科學大辭典*（台北市：漢美，民國84年），頁1515。

註 3：國立中央圖書館主編，*臺閩地區圖書館統計名錄*（台北市：國立中央圖書館，民國82年），頁1。

註 4：國立中央圖書館主編，*臺閩地區圖書館統計名錄*（台北市：國立中央圖書館，民國82年），頁82。

註 5：由於Metadata並沒有統一的翻譯詞彙可茲使用，目前可見的對應中文詞彙有元資料、超資料、詮釋性資料。「詮釋性資料」比較清楚地界定何謂Metadata，然而卻比較像一般的名詞組而非專有名詞；「超資料」則容易與Hyperdata混為一談。在沒有統一使用的翻譯詞彙出現之前，筆者暫時地將Metadata翻譯為元資料。

註 6：GILS. "Guidelines for the Preparation of GILS Entries." 1995, (URL: http://gopher. nara.gov:70/0/managers/gils/guidance/gilsdoc.txt).

註 7：FGDC. "Content Standards for Digital Geospatial Metadata -- FGDC." 1994, (URL: http://fgdc.er.usgs.gov/fgdc.html).

註 8：Weibel, S., J. Godby, and E. Miller. "OCLC/NCSA Metadata Workshop Report." 1995, (URL: http://gopher.sil.org/sgml/metadat.html).

註 9：專有名詞的辨識一直是計算語言學與自然語言處理領域重要的研究課題，中文的專有名詞比英文更複雜、更具挑戰性。有關討論專有名詞辨識的論文如下所示：

Chen, K.H. and H.H. Chen. "Extracting Noun Phrases from Large-Scale Texts: A Hybrid Approach and Its Automatic Evaluation." *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL94)*, 1994, 234-241.

Chen, H.H. and G.W. Bian. "Proper Name Extraction from Web Pages for Finding People in Internet." *Proceedings of ROCLING X International Conference*, 1997, 143-158.

註10：Appelt, D.E. and Israel, D. Tutorial on Building Information Extraction Systems, 1997, Washington, DC, p.4.

註11：同註10。

註12：所謂的適應性系統指的是系統能夠透過某種學習的程序，適應不同類型的環境，並處理不同類型的工作。

環境，並處理不同類型的工作。

註13：這個例子是由清大張俊盛教授於1992年發表於應用自然語言處理研討會的論文改寫而來，原句是「把劉顯仲的確實行動作了分析」。
Chang, J.S. et al. "A Corpus-Based Statistical Approach to Automatic Book Indexing," *Proceedings of the Third Conference on Applied Natural Language Processing*, ACL, 1992, 147-151.

註14：Chen, K.J. and Liu, S.H. "Word Identification for Mandarin Chinese Sentences," *Proceedings of the 15th International Conference on Computational Linguistics*, 101-107.

註15：Chen, H.H. and J.C. Lee. "Identification and Classification of Proper Nouns in Chinese Texts." *Proceedings of the 15th International Conference on Computational Linguistics* (COLING96), 1996, 222-229.

註16：Chiang, T.H. et al. "Statistical Models for Word Segmentation and Unknown Word Resolution," *Proceedings of R.O.C. Computational Linguistics Conference V (ROCLING V)*, 1992, 121-146

註17：這個句子是從中央研究院資訊科學研究所詞庫小組建構的漢語語料庫取出的，詞類標記也是由中央研究院資訊科學研究所詞庫小組制訂的，以N開頭的詞類如Na、Nb、Nc為名詞；以V開頭的詞類如VC、VB為動詞。

註18：Brill, Eric. "A Simple Rule-based Part of Speech Tagger," *Proceedings of the Third Conference on Applied Natural Language Processing*, 1992, .

註19：Church, K.W. "A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text," *Proceeding of the Second Conference on Applied Natural Language Processing*, 1988, 136-143.

註20：Tapanainen, P. and Voutilainen, A. "Tagging Accurately - Don't Guess If You Know," *Proceedings of the 4th Conference on Applied Natural Language Processing*, 1994, 47-52.

註21：Hindle, D. *User Manual for Fidditch: A Deterministic Parser*, Naval Research Laboratory Technical Memorandum 7590-142, Naval Research Laboratory, Washington, D.C., 1983.

註22：Chen, K.H. and Chen, H.H. "A Probabilistic Chunker," *Proceedings of the 6th ROCLING*, 1993, 99-117.

註23：Church, K.W. and Hanks, P. "Words Association Norms, Mutual Information, and Lexicography," *Computational Linguistics*, 16(1), 1990, 22-29.

註25：Wilks, Y. and Stevenson, M. "Sense Tagging: Semantic Tagging with a Lexicon," cmp-lg/9705016, 1997.

註26：Hearst, M. and Plaunt, C. "Subtopic Structuring for Full-Length Document Access," *Proceedings of the 6th International ACM SIGIR Conference on Research and Development on Informaiton Retrieval*, 1993, 59-68.

註27：Post-Conference Workshops, URL: http://www.sis.pitt.edu/~diglib97/Workshops.htm.

註28：同註10。

註29：計算機學會（Association for Computing Machinery，簡稱ACM）設有許多特殊興趣小組（Special Interesting Group，簡稱SIG），其中SIGIR是由資訊檢索的學者專家組成，該小組每年舉辦ACM SIGIR Conference，會中發表的論文代表的是資訊檢索領域中最新的發展趨勢。