電子文獻主題之自動辨識

Automatic Identification for Topics of Electronic Documents

陳光華

Kuang-hua Chen

台灣大學圖書館學系助理教授

khchen@steelman.ls.ntu.edu.tw

【摘要 Abstract】

網際網路上的電子文件數量極爲龐大，如何快速有效的進行電子文件主題標引的工作逐漸成爲一項重要的研究課題。目前有關的研究著重於名詞的行爲，期望藉由文獻中名詞的頻率或是其他統計值，求得文獻的主題分類。雖然文獻的主題是由名詞組成，但是本文認爲決定那些名詞成爲主題的因素卻不只是名詞。因爲文獻的組織是具有結構性的，是事件驅動（Event-Driven）的，而事件則是由名詞與動詞共同完成的，名詞與動詞在決定文獻主題的過程中具有重要地位。本論文考慮文獻的一般行爲，提出四項因素：1) 詞彙的重要性，2) 詞彙的重複性，3) 詞彙的共現性，4) 詞彙的距離，建構一個數學模型並進行讀者與模型的比較實驗。實驗結果顯示該模型的自動主題辨識與人工主題辨識具有相當的效能。

The volume of electronic documents in the Internet grows very quickly. How to effectively assign topics to documents becomes an important issue. In the present time, the researches based on this line focus on the behavior of nouns in documents. Although topics are composed of nouns, the constituents that determine which nouns are topics are not only nouns. We think that texts are well-organized and are event-driven. Therefore, nouns and verbs together contribute the process of topic identification. This paper considers four factors: 1) word importance, 2) word frequency, 3) word co-occurrence, and 4) word distance and constructs a mathematical model. The preliminary experiments show that the performance of the proposed model is equivalent to that of human being.

【關鍵字 Keywords】:

資訊檢索；電子文獻；主題辨識

Information Retrieval; Electronic Document; Topic Identification

一、前言

自從網際網路的蓬勃發展，電子資訊的累積極爲快速，而電腦科技的日新月異，再加上商業體系的投入，整個世界就像是縮小了一般，訊息的傳遞很快，人類享有前所未見的資訊服務。以往資訊的提供者是有組織的機構，如圖書館、博物館、行政單位、企業組織等等；今日，任何人也可以作爲資訊的提供者，藉由網際網路的連線，將個人的觀點、評述、作品傳播給讀者。知識的權力下放給一般的使用者，網際網路呈現一片生氣盎然。然而，問題隨之產生。

皮亞特斯基-薛比歐（Piatetsky-Shapiro）與弗

勞萊（Frawley）在 1991 年指出每 20 個月，資訊量會增加一倍，要注意的是，這是 1990 年的估測。（註 1）而根據統計，Usenet 的資訊，每年增加一倍；Internet 資訊流量，每月增加 12%。（註 2）至於臺灣的情形，教育部統計 TANet 對外的網路流量，1992 年 2 月爲 27,446,892Kbytes；1993 年 1 月爲 47,275,416Kbytes；到了 1995 年 11 月，增加爲 283,458,248Kbytes；吾人可以發現這期間，網路對外流量成長 10 倍以上。（註 3）資訊的浪潮吞噬了人們，使用者茫然不知所措，要從茫茫的資訊大海選取一瓢，真不是容易的事。如何協助使用者或讀者取得需要的資料，成爲一項重要的課題，爲使用者整理資源的服務應運而生。此類的服務分爲兩種：一爲主題指引（Subject Directory）；一爲搜尋引擎（Search Engine）。目前提供此類服務的機構非常多，前者如 Yahoo（註 4），後者則以迪吉多公司的 Alta Vista 爲代表。（註 5）然而，他們也都存在著一些問題。以 Yahoo 爲例，該公司聘請大量的員工，負責的工作是判定新增 WWW 首頁的主題分類。然而，同一個人對於同一份文件，今日與明日的判斷不見得相同，即所謂的 "Intra-indexer Inconsistency"問題；不同的人對同一份文件的判斷也不見得相同，亦即所謂的 "Inter-indexer Inconsistency"問題。因此，引進自動的機制協助人類，似乎是不可避免的。另外，就使用者的角度，目前各個提供檢索服務的公司，其查詢介面使用的仍然是控制式的查詢語言，這對於使用者而言非常不方便。若是有一種機制能夠決定文獻資料的主題，也能夠用於判定使用者使用自然語言查詢的主題，這樣的機制可以快速地協助人們更新服務的內容，避免分類不一致的情形，對於資料的過濾、取用、與查詢將會有很大的幫助。

目前自然語言處理與計算語言學的研究相當活躍，隨著電腦硬體的快速發展，以往無法做到的技術得以在新的計算平台上實現。雖然語言（包括書面語（Written Language）與口語（Spoken Language））的現象十分的複雜，要全盤掌握仍有待長久的努力，但是也已經有很多的研究成果。（註 6）網路資源文獻主要是以書面語的形式傳播（當然多媒體的資訊越來越多，但是仍然少不了書面語的部份，同時這些書面語通常扮演說明的角色，具有舉足輕重的功能），因此使用語言處理技術分析電子文件是很自然的作法。目前有許多研究機構從事有關文獻主題的研究，若仔細地檢視這些研究成果，可以發現它們著重於文獻中名詞的行爲，期望藉由文獻中名詞的頻率或是其他統計值，求得文獻的主題。雖然文獻的主題是由名詞組成，但是，筆者認爲決定哪些名詞成爲主題的因素卻不只是名詞而已。一般而言，文獻的組織是具有結構性的，文獻中資訊內容是由事件驅動（Event-Driven），而事件則是由名詞與動詞共同完成的。因此，名詞與動詞在決定文獻主題的過程中都具有重要地位。本文將補上目前世界各國在這方面研究的空白，加入動詞的考量，並依據四項因素：1)詞彙的重要性，2)詞彙的重複性，3)詞彙的共現性，4)詞彙的距離，模擬書面語（文獻）的一般行爲，希望據此建構一個數學模型，進而運用該模型分析文文獻資料，協助人類自動辨識文獻的主題。

二、文獻分析

圖書館爲了協助讀者取得需要的資訊，館藏皆經由一定的加值處理（Value-added Processing）。例如索引者（Indexer）與摘要者（Abstractor）爲文獻加上必要的索引詞彙與摘要，其目的是讓讀者或使用者擁有更多的訊息判斷本身的資訊需求。一般而言，各類型的文獻都有適當的 Metadata 用以描述文獻的各項訊息。傳統的機讀格式（MARC）可視爲一種 Metadata，記載極爲繁複的訊息（註 7）；也有比較簡單的 Metadata，例如都柏林核心集（Dublin Core），只有 13 個必要的欄位。（註 8）無論是那一種 Metadata 應當都會有一欄記載所描述文獻的主題，美國國會圖書館出版的 LCSH 標題表，讓圖書館員從中選取適當的詞彙，爲文件加註適當的主題（註 9）；美國醫學圖書館也有 MeSH（註 10），其功能與 LCSH 相同；我國國家圖書館

同樣也出版中文圖書標題表，供全國圖書館使用。（註 11）。

館員進行主題標引的工作，是屬於心智的活動，館員依據館方的政策以及控制詞彙的規範，衡諸本身對於文獻的瞭解，給定適當的主題。這樣的工作消耗許多的人力，而在電子文件越來越多的情況之下，完全由人力進行主題標引的工作幾乎不可行。因此運用電腦科技協助主題標引的工作成爲資訊檢索領域一項重要的研究課題。

自動主題標引可分爲兩個方式：

- 由標題表或索引典挑選適當詞彙
- 由文件本身挑選適當詞彙

第一種方式事實上是模擬圖書館館員進行人工標引的作法，也稱爲控制詞彙標引；而第二種方式就是主題辨識，也可稱爲自由詞彙標引，接近言談語言學家企圖擷取文件核心意念的作法。這兩種作法各有其優缺點。使用控制詞彙，一般使用者若不知道那些是控制詞彙，則無法檢索需要的資訊；使用自由詞彙會造成索引典過於龐大，難以規範詞彙與詞彙之間的關係。筆者將採用第二種方式，發展自動主題標引的程序，以減少大量人力的投入，未來將整合控制詞彙與自由詞彙。

資訊檢索技術的研究是爲了解決讀者資訊需求的問題，協助讀者以最少的時間取得最多且有用的資訊，避免讀者迷失於龐然的資訊之洋。資訊檢索相關研究的發展可以大致分爲兩個脈絡：一爲由文件本身出發；一爲由使用者的觀點出發。無論是由何種角度從事資訊檢索的研究，其目的都是希望能夠達到上述的目標。

貝爾金（Nicholas J. Belkin）從使用者檢索的角度切入，提出 16 種不同的檢索策略（Information Seeking Strategies，簡稱 ISS），將使用者可能的檢索方式分爲 16 個空間，並且描繪使用者初始狀態的空間，以及如何由一個空間轉入另一個空間。這些空間由互動方法（Method of Interaction）、互動目標（Goal of Interaction）、檢索模式（Mode of Retrieval）、使用的資源（Resource Considered）四個向度規範。四個向度各有兩種可能性，因此共計有 16 個 ISS。（註 12）

里斯柏根（C. J. Van Rijsbergen）、史派克瓊斯（K. Sparck Jones）、沙騰（Gerard Salton）等人建構了從文件詞彙入手的統計學派，著重於透過文件詞彙的分析，提出文件模型、判定文件主題或是伴隨回饋機制，建構符合使用者資訊需求的檢索系統。多數的電腦科學研究人員遵循這個脈絡進行資訊檢索的研究。

史派克瓊斯於 1972 年提出了逆向文件頻率（Inverse Document Frequency，簡稱 IDF），並進行一連串實驗，發現使用 IDF 的檢索系統能夠產生比較有效的檢索結果。（註 13）沙騰於 1973 年至 1975 年之間提出了數篇論文，進一步使用詞彙鑑別值（Term Discrimination Value，簡稱 TDV）的觀念，加強資訊檢索系統的效用。（註 14）羅契歐（J.J. Rocchio, Jr.）、艾德（E. Ide）等人則提出查詢修正（Query Modification）的觀念，希望經由第一次查詢所得的結果，經過修正原始查詢，再次送出新的查詢，以獲得更好的檢索成果。這種觀念導入回饋機制的研究，也使得統計學派的資訊檢索研究與使用者有一定程度的互動。（註 15）

從語言學的觀點，相對於口語，文獻內的文字表述屬於書面語，因而語言學家同樣對文件感到莫大的興趣，尤以從事言談分析（Discourse Analysis）研究或是語料素材（Corpus）研究的語言學家爲然。因此，另外一派以語言學的角度建構文件模型的研究也蓬勃發展（註 16），認爲純然的統計模型會忽略語言的特性，無法掌握文件的重要特質。因此，提出許多結合統計的語言模型，企圖更加合理地規範文件。在電腦硬體日新月異的情況之下，許多計算語言學的技術得以實現，因之，資訊檢索領域與計算語言學領域遂有逐漸交流的情況。

葛拉茲（B. Grosz）與賴德樂（C. Sidner）提出的修辭言談結構（Rhetorical Discourse Structure，簡稱 RDS）用以模擬言談的結構，認爲

言談是一個主題意念完整的結構，其間的遣詞用字都有一定的關係。她們採用語意近似關係（Thesaural Relation）描述言談的結構，並且定義數種不同的語意關係，然而其缺點是這些關係是二元的，並沒有強弱之別。（註 17）坎普（H. Kamp）則運用了名詞宇集（Universal）的概念，透過照應詞（Anaphora）的決定，建立言詞結構中句子與句子的關係。（註 18）然而，目前仍然沒有研究或論著提出直接運用言談分析技術於資訊檢索，資訊檢索有其特殊的使用環境，也就是時間的限制。顯而易見，前述的作法將耗用大量的計算時間，在電腦硬體或是軟體技術再次大幅邁進之前，這些作法的實用性不高。

妥協的方案也有眾多研究人員提出。賀斯特與普勞特（M. Hearst and C. Plaunt）利用名詞出現的頻率計算言談結構的範疇（Scope），運用於資訊檢索系統，證實比全文檢索系統更有效。（註 19）筆者則於 1995 年提出的計算模型結合了名詞與動詞的語言特性並計算詞頻統計特性，有效地規範言談結構與主題辨識等現象。然而，美中不足的是其計算量仍嫌過大，模型有待進一步的修正。（註 20）基於以上的說明，吾人可發現資訊檢索與計算語言學技術的結合仍有很多發展的空間。

三、模型的背景

傳統上，資訊檢索的研究通常使用詞頻（Term Frequency，簡稱 TF）作爲選擇索引詞彙的標準，認爲排除所謂的功能詞彙（Function Word）之後，文件中出現越多次的詞彙越能夠代表該文件的特性。然而，若是相同的詞彙在許多文件都出現，則其代表性會比較不可靠，因爲其鑑別性（Discriminativity）比較低。史派克瓊斯針對這個缺點，提出了逆向文件頻率（Inverse Document Frequency，簡稱 IDF）的修正作法。（註 21）IDF 可以用下列的數學式表示：

$$IDF(w) = \log((P\text{-}O(w))/O(w))$$

P 是某一文件集合的文件總數，$O(w)$是包含詞彙 w 的文件總數。當詞彙 w 出現於一半以上的文件，則其 IDF 小於等於 0，吾人可以認爲這個詞彙一點都不重要，對文件集合中的文件不具有鑑別性。引用一個 IDF 小於等於 0 的詞彙做爲文件的索引，就好似使用黃皮膚區別華人與韓人一般，這樣的檢索系統無法有效滿足使用者的檢索需求。史派克瓊斯的修正作法將檢索系統的效能往前邁進一大步，直到現在，TF 結合 IDF 的策略仍然是資訊檢索領域中極爲經典的代表。

吾人若仔細探究目前的研究取向，只要是採取統計方法的研究，基本上是遵循史派克瓊斯、沙騰等人開創性研究的脈絡。當然，還有其它重要的研究取向，例如貝爾金提出的使用者導向（User-oriented）的作法。（註 22）本文則仍然是由計算與統計的角度，觀察資訊檢索這項重要的研究課題。依循統計方式的研究，研究人員都將注意力投注於名詞性的詞彙，嘗試由文件的名詞篩選出具有代表性的名詞做爲文件的特徵。上述的研究程序有一個盲點，無論研究人員如何做，都是試圖在名詞群找關係。例如，尤曼（Youman）、莫瑞斯與赫斯特（Morris and Hirst）、瑞那（Reynar）等人的研究。（註 23）雖然最終的索引詞彙是由名詞構成，然而並不代表這個過程中不可以引進其他的重要因素。吾人若觀察索引者做索引的方式，應當對這個過程有進一步的瞭解。

索引者閱讀文件，通過個人的理解，依據標題表或是索引典，給予適當的控制詞彙做爲索引詞彙；如果是自由詞彙索引（Free-text Indexing），索引者則由文件中挑選名詞做爲索引詞彙。雖然吾人無法完全斷定索引者的心智活動爲何，但是，顯然不只是閱讀文件中的名詞，從而決定應該使用那些詞彙。從事言談（Discourse）研究的語言學家認爲有意義的文件必定有某些結構，因此提出各種不同的理論，試圖規範言談的結構。其中比較有名的是言談展現結構（Discourse Representation Structure，簡稱 DRS）（註 24），以及修辭言談結構（Rhetorical Discourse Structure，簡稱 RDS）。（註 25）但是，由處理電子文件的角度看文件索引詞彙

的判定或是主題的挑選，前述言談結構之計算量太大，無法快速地處理大量且急遽成長的網路電子文件。因此，綜合前述的討論可以歸納爲以下兩點：

- 僅使用名詞進行自動索引或主題辨識的模型並不完整。
- 將言談結構帶入文件模型並不適用於大量的電子文件。

爲了解決前述的問題，本文提出的模型是規範文件中名詞與動詞以及名詞與名詞之間的關係，用以自動決定文件主題。

四、模型的數學架構

組織完善，意念完整的文件，其名詞與名詞以及名詞與動詞的關係相當密切，筆者建構的模型是基於下列的假設：

名詞與動詞共存於述語參數結構（註 26）；
而名詞間關係是建構於言談層次。

自動辨識電子文獻主題的第一步是必須瞭解構成書面語的要素，也就是一般人撰寫文章的過程。透過大規模語料庫資料的蒐集與分析，使用統計學的模型，可以用電腦技術模擬這種過程。由於目前電腦尚未達到具有智慧的能力，僅能夠透過定量的觀察與模擬，期望當數量到了某一個數字後，定量的模擬能夠逐漸逼近定性的瞭解。筆者使用四種詞彙的數學統計值：

- 詞彙的重要性
- 詞彙的重複性
- 詞彙的共現性
- 詞彙的距離

作爲建構整個模型的基礎，以下分別討論此四種統計值。

詞彙的重要性代表的是，當它出現於文獻時，做爲作者意念核心的機會，也就是當索引者重建作者創作時的心智活動，由文件挑選詞彙做爲文件主題的機會。並不是所有的詞彙都一樣重要。例如，若是將文獻中的冠詞、副詞、以及介系詞等詞彙刪除，仍然能夠知道這份文獻的梗概，這說明了上述的詞彙並不十分重要。反觀之，名詞與動詞就十分重要了。詞彙的頻率常常可以代表某種程度的重要性，這種情形，尤以一般的資訊檢索系統爲最。然而，詞彙的重要性無法由 TF 完全顯示，因爲所謂的重要性是針對文獻而言，並非詞彙本身重要與否。因此 IDF 才能代表詞彙對文獻的重要程度。當文獻的數目夠大時，IDF 值就具有相當高的穩定性，可據以作爲詞彙重要性的計算標準。

意念一致的文獻資料，作者使用的詞彙必然趨向某一個語意範疇。從統計的觀點，這表示該語意範疇的詞彙一起出現的機率比較大。判斷那些詞彙屬於同樣的語意範疇是相當困難的工作，但是由大規模的語料庫計算詞彙的共現的程度就很簡單了。可以使用共容訊息（Mutual Information，簡稱 MI）計算詞彙的共現，其數學式分別如下所示：（註 27）

$$MI(t_i,t_j)=\log\frac{P(t_j\mid t_i)}{P(t_j)}=\log\frac{P(t_i,t_j)}{P(t_i)P(t_j)}$$

共容資訊的意義是，當詞彙 t_i 與詞彙 t_j 經常一起在語料庫出現，聯合機率 $P(t_i,t_j)$ 會甚大於 $P(t_i)\times P(t_j)$，因此 $MI(t_i,t_j)$會甚大於 0；當 t_i 與 t_j 出現的方式是背道而馳時，$MI(t_i,t_j)$會甚小於 0；當彼此沒有什麼關係時（以機率論的術語而言，也就是互相獨立），因此 $P(t_i,t_j)\cong P(t_i)\times P(t_j)$，所以 $MI(t_i,t_j)$接近於 0。

詞彙的位置也很重要。基於文獻是有生命的文字組合的觀點，相關的詞彙其出現的距離必定不會太長。因爲，一旦相隔太遠，彼此之間的相乘效果就大打折扣，這不會是一般作者的用意。引入距離的因素，比較能夠忠實反應寫作的行爲。距離的計算可採用如下的方式，首先爲每一個名詞與動詞設定一個編號，以下面這一段文字爲例：

蘇聯 1 許多 製造 2 民生 3 日用品 4 的 工業 5 得到 6 政策性 7 的 補貼 8，其 目的 9 是 保持 10 物價 11 的 平穩 12。但 補貼 13 勢 難 普及 14 於 各行各業 15，因此 又 造成 16 某 些 日用品 17 不足 18 或 完全 缺乏 19 的 後遺症 20 。 現在 既然 要 引進 21 市場 22 經濟 23，補貼 24 政策 25 又勢 難 繼續 26，一旦，放棄 27，許多 民生 28 物資 29 的 價格 30 必然 上漲 31，於是 又 引出 32 民間 33 屯積 34 物資 35 與 通貨膨脹 36 的 壓力 37。

詞彙 X 與 Y 的距離 $D(X,Y)$可以用以下的方式計算：

$$D(X,Y) = \mathrm{ABS}(C(X)-C(Y))$$

ABS 爲絕對值函數，$C(X)$代表詞彙 X 的編號，如 C(政策性) = 7，而 C(目的) = 9，所以 D(政策性, 目的) = 2。

綜合以上因素，筆者提出的計算模型爲：

$$Score(n) = PN \times SNN(n) + PV \times SNV(n)$$

$Score(n)$爲名詞 n 作爲主題的強度；$SNV(n)$爲名詞 n 與其他動詞的強度；SNN 爲名詞 n 與其他名詞的強度；PN 與 PV 分別爲 SNN 與 SNV 的權重。$SNN(n)$與 $SNV(n)$的計算方式如下：

$$SNN(n_i) = \sum_j \frac{IDF(n_i) \times IDF(n_j) \times f(n_i, n_j)}{f(n_i) \times f(n_j) \times D(n_i, n_j)}$$

$$SNV(n_i) = \sum_j \frac{IDF(n_i) \times IDF(v_j) \times f(n_i, v_j)}{f(v_i) \times f(v_j) \times D(n_i, v_j)}$$

$f(w)$爲詞彙 w 的頻率，$f(w_i,w_j)$爲詞彙 w_i 與 w_j 共同出現的頻率；$D(w_i,w_j)$爲 w_i 與 w_j 之間的距離。可以看出整合了前述的四項考量因素，事實上，$f(w_i,w_j)/(f(w_i)\times f(w_j))$即爲計算詞彙共現的程度，與 MI 具有相同的型式，可稱之爲共容頻率（Mutual Frequency，簡稱 MF）。

五、實驗與分析

筆者以中文文件爲處理的對象，因之如何取得大量且高品質的中文電子文件資料，也是一項重要的課題。雖然網路上可以蒐集大量的中文資料，但是從訓練模型的角度出發，經過整理且受到一定程度控制的文件，才能夠建立有效的模型。而且中文有以下的特性，使得直接由網路取得訓練語料的作法不可行。

- 中文沒有詞間標記，亦即詞與詞之間沒有空格，極易造成詞彙歧異（Ambiguity）的現象。
- 中文的詞類變化極大，同一個詞彙可能具有多種詞類。

若是採用一般的電子文件，必須先經由分詞程序（Segmentation），各種自動分詞程序所得的結果，全然正確的情形也不多見。爲了讓模型不受這些因素影響，筆者採用中央研究院平衡語料庫（Sinica Corpus）1.0 版。該語料庫每一份文件皆有分類資訊，標示各種與文件相關的資料；每一個句子都有編號，句子與句子之間以 47 的星號隔開；詞彙則已經分開，並且加上詞類標記（總計有 46 個詞類標記），或其他特徵標記（共有 8 個特徵標記）。（註 28）該語料庫收錄報導、評論、廣告圖文、信函等等不同類型的文件，收錄的媒體橫跨報紙、雜誌、學術期刊、教科書、工具書等等。（註 29）

整個實驗程序共分爲訓練階段、實驗階段、評估階段等三個階段，如下所示：

- 訓練階段
  1. 計算中央研究院平衡語料庫中所有名詞與動詞的 IDF
  2. 計算中央研究院平衡語料庫中兩兩詞彙的 MF
  3. 計算中央研究院平衡語料庫各詞彙的頻率。
- 實驗階段
  1. 隨機選取中央研究院平衡語料庫十篇新聞報導，作爲測試語料。
  2. 使用計算模型決定文件中名詞作爲主

題的優先順序，亦即根據計算所得的 Score 排列名詞。

3. 八位讀者閱讀上述十篇報導，自行決定文件主題，並依自定的優先順序列出主題。

- 評估階段

1. 評估計算模型與讀者選定的主題。

本實驗從中央研究院平衡語料庫收錄的報紙媒體語料中，隨機抽取十篇短篇新聞報導，其出處如表一所示，出處標記之格式爲「檔案名：起始句編號-結束句編號」，句數表示該篇文章的句子總數，而詞數則是該篇文章的詞彙總數。實驗語料以報導記敘、評論論說爲主，尚有散文描寫類的語料。 使用系統模型處理這十篇語料，實驗結果如表二所示，表二第二欄記載模型認爲充當文章主題的優先順序。由於模型計算文章中所有名詞的主題強度，而且文章中有甚多的名詞，第二欄並沒有將所有的名詞列入，只列出前三分之一的名詞。此外，有八位讀者參與主題辨識的工作，讀者閱讀文章後，根據自己對於文章的理解，排列主題的優先順序。至於列入排列的名詞總數並無任何限制，也就是讀者可自行認定那些名詞是否是該篇文章的主題，因此各位讀者所認定的主題總數，可能會有很大的出入。表三詳列讀者的閱讀結果。

在索引與摘要的研究中，常提及"Inter-indexer Inconsistency"與"Intra-indexer Inconsistency"的現象，實驗的結果也有類似的情形。只要稍微比較表三就能夠發現讀者進行閱讀活動後，經由思考咀嚼後的產物還是有相當的差異。表四分析了讀者給定的主題數以及相關的統計數字，AVG 代表平均數，STDEV 代表標準差。讀者做實驗之前所獲的指示是：「依據自己閱讀的結果，決定文章的主題，且依優先順序排列，但並沒有主題數目的限制」，表四的統計數字顯示讀者認定的主題數目變化相當大。就同一文章而言，標準差最高爲 5.63，最低爲 1.36；若從讀者的角度檢視表四，標準差最高爲 4.01，最低爲 0.70。讀者本身的變異程度相對較小

如果檢視讀者認定的主題彼此之間重複的情形，可以用表五顯示超過 1 位讀者以上選擇的主題。其中爲八位讀者共同認定的主題僅有 7 個，而總共有 57 個主題僅有一位讀者認爲足以代表該篇文章。圖一更易看出讀者之間變異程度實在是不小。

實驗模型判定主題的品質，首先可以和讀者的實驗結果比較，若是以多數讀者認定的主題爲基礎，則模型實驗結果與讀者閱讀結果的重複性可用表六表示。表六的第一欄「無」表示模型選定的主題並無任何一位讀者認爲是主題；第二欄「一位」表示有一位讀者選定的主題與模型選定的相同，其他欄依此類推，系統模型選定的主題若是越多讀者選定的主題，代表模型越能夠模擬閱讀行爲。檢視表五與表六可以發現，八位讀者一致同意的主題，系統模型大致也都找出來，如 TEXT01、TEXT04、TEXT05、以及 TEXT10 等文章。然而，TEXT07 的實驗結果卻不好，仔細閱讀該篇文章，發覺該篇文章的意念較爲發散，並不容易做好，再比較讀者閱讀該篇文章後給定的主題，可由表五發現各讀者的閱讀結果，也是以 TEXT07 最不一致。圖二繪製系統模型實驗結果與讀者閱讀結果的相似程度，總共有 38 個主題並沒有任何讀者認爲是主題；另外有 42 個主題至少有一位讀者選定；被八位讀者共同認定的主題有 4 個。若比較圖一與圖二，也能夠輕易發現它們的分佈情形相當類似。

表一 實驗文件之統計資料

	出處標記	句 數	詞 數
TEXT01	_T-SA.TAG:1-13	13	102

TEXT02	_T-SA.TAG:32-50	19	230
TEXT03	_T-SY.TAG:1-12	12	132
TEXT04	_T-SY.TAG:89-112	24	222
TEXT05	_T-SY.TAG:212-244	33	374
TEXT06	_T-SL.TAG:60-88	29	345
TEXT07	_T-SL.TAG:490-531	42	406
TEXT08	_T-SL.TAG:680-705	26	453
TEXT09	_T-SL.TAG:721-745	25	268
TEXT10	_T-SL.TAG:797-829	33	353

表二 實驗結果

	名詞擔任主題之優先順序
TEXT01	考試、優待、條文、標準
TEXT02	考試、優待、草案、淪陷區、學歷、部長、標準、辦法
TEXT03	制度、背後、思想、體系、行爲
TEXT04	憲法、普莉揚卡、被選舉權、第三世界、復仇者、身分、政權、後起之秀、桑妮雅、時間
TEXT05	變化、馬列主義、群眾、死路一條、馬克思、教廷、共產黨人、祖國、口號、卡斯楚
TEXT06	史氏、配樂家、ＨＥＲＢＥＲＴＳＴＯＴＨＡＲＴ、色彩、民謠、風味、音樂、包機
TEXT07	考試、虫子、肢腳、老頭、夢魘、金杏枝、磚頭書、查泰萊、習慣
TEXT08	人間、烽火、Ｊａｓｐｅｒｓ、時代、特徵、實質、診斷、商業
TEXT09	編輯、點子、朋友、觀光客、過客、市民、刊、感情、城市
TEXT10	改革、轉機、政變、軍方、葉爾辛、短文、改革派、機率、保守派

表三 讀者選取的主題

	讀者一	讀者二
TEXT01	優待、入學考試、標準	入學考試、優待、標準、條文
TEXT02	大陸、學籍、優待、標準	學籍、學歷、大陸、教育
TEXT03	蘇聯、經改、開放、矛盾	改革、經改、體制、蘇聯、特權
TEXT04	桑妮雅、普莉揚卡、拉吉夫	桑妮雅、普莉揚卡、印度、政權
TEXT05	卡斯楚、馬列主義	卡斯楚、共產黨人、馬列主義
TEXT06	亂世、佳人	亂世、佳人、電影
TEXT07	飄、心得	小說、飄、亂世、佳人
TEXT08	傳播、媒介、實質	實質、傳播、商業、書報、雜誌
TEXT09	人間、稿、台北	台北、稿、都會
TEXT10	蘇聯、經濟、改革、難題	蘇聯、經濟、改革、補貼、難題

表三 讀者選取的主題 （續）

	讀者三	讀者四
TEXT01	大陸、入學考試、優待、高中、五專、大學、教育部	入學考試、優待
TEXT02	大陸、教育、學籍、學歷、學生、考試、優待、國統綱領、草案、淪陷區	學歷、條件、大陸、教育局
TEXT03	蘇聯、經改、轉型期、體制	開放、改革、思想、行爲、模式、矛盾、轉型期、特權、既得利益、體系
TEXT04	印度、普莉揚卡、桑妮雅、甘地、拉吉夫、第三世界、政權	桑妮雅、普莉揚卡、拉吉夫、復仇者、煩惱
TEXT05	蘇聯、古巴、哈瓦那、戈巴契夫、卡斯楚、第三世界、馬列主義	卡斯楚、共產黨人、蘇聯、戈巴契夫、馬列主義、第三世界
TEXT06	亂世、佳人、郝思嘉、蓋博、費雯麗、密契兒、霍華、賽茨尼克、電影、音樂、影業、票房、記錄、金像獎、配樂家、史氏、亞特蘭大、熱潮、焦點	亂世、佳人、美國、霸權、藝術
TEXT07	亂世、佳人、書、小說、愛亞、白瑞德、郝思嘉、電影	心得、小說、興趣
TEXT08	商業性、出版界、商品、人文性、時代、知識、書報、雜誌、商業、Jasper、廣告	實質、媒介、商業性、消費者、知識、訊息、特徵、出版物
TEXT09	台北、都會、城市、市民、台北人、形貌、感情、專輯	台北、專輯
TEXT10	蘇聯、經濟、改革、戈巴契夫、民生、保守派、葉爾辛、政變	蘇聯、戈巴契夫、經濟、改革、補貼、物價、中小企業

表三 讀者選取的主題 （續）

	讀者五	讀者六
TEXT01	入學考試、優待、標準	大陸、入學考試、考試、優待、標準
TEXT02	學歷、學籍、辦法、草案	大陸、台灣、學籍、學歷、學生、國統綱領、淪陷區、考試、優待
TEXT03	轉型期、制度、體制、利益	蘇聯、經改、制度、思想、行爲、體制、轉型期、特權
TEXT04	普莉揚卡、印度、政治、生活	桑妮雅、印度、普莉揚卡、婚姻、政治、拉吉夫
TEXT05	共產黨、古巴、蘇聯、卡斯楚	卡斯楚、古巴、共產黨、馬列主義、馬克斯、蘇聯、戈巴契夫、哈瓦那
TEXT06	電影、文學、金像獎	電影、亂世、佳人、文學、藝術、美國、時代、文化、好萊塢
TEXT07	小說、書	小說、飄、電影、亂世、佳人、愛亞
TEXT08	出版物、價值、商業、廣告	書報、雜誌、文字、知識、訊息、價值、時代、特徵、商業
TEXT09	稿、感情、點子、城市	台北、城市、都市、感情、市民
TEXT10	蘇聯、經濟、改革、政策	蘇聯、經濟、改革、物價、市場、企業、政治、派系

表三 讀者選取的主題 （續）

	讀者七	讀者八
TEXT01	入學考試、優待、標準、條文	入學考試、大陸、優待、大學、高中、五專、標準
TEXT02	教育、學籍、學歷、大陸、台灣、辦法、草案	大陸、學籍、學歷、學生、教育部、考試、優待、草案
TEXT03	體制、制度、經改、蘇聯、特權、既得利益、利益、問題	蘇聯、經改、制度、體制、人民、特權、轉型期
TEXT04	普莉揚卡、桑妮雅、政權、印度、婚姻、豪門	印度、普莉揚卡、桑妮雅、拉吉夫、第三世界、政權
TEXT05	卡斯楚、共產黨、馬列主義、古巴、蘇聯、戈巴契夫	古巴、哈瓦那、蘇聯、卡斯楚、戈巴契夫、馬克斯、共產黨
TEXT06	亂世、佳人、電影、亞特蘭大、內戰、文化、焦點、文學、藝術	亂世、佳人、飄、電影
TEXT07	愛亞、亂世、佳人、心得、書、學校	亂世、佳人、小說、印象
TEXT08	印刷物、商業性、知識性、人文性、診斷、書報、雜誌、廣告、實質	傳播、出版物、書報、雜誌、功能、商品、商業性、知識性
TEXT09	台北、台北人、市民、感情、專輯	台北、形貌、感覺、感情
TEXT10	經濟、改革、蘇聯、改革派、保守派、勢力、戈巴契夫、補貼、政策	蘇聯、經濟、改革、民生、工業、補貼、政治、勢力

表四 讀者判定主題數目之比較

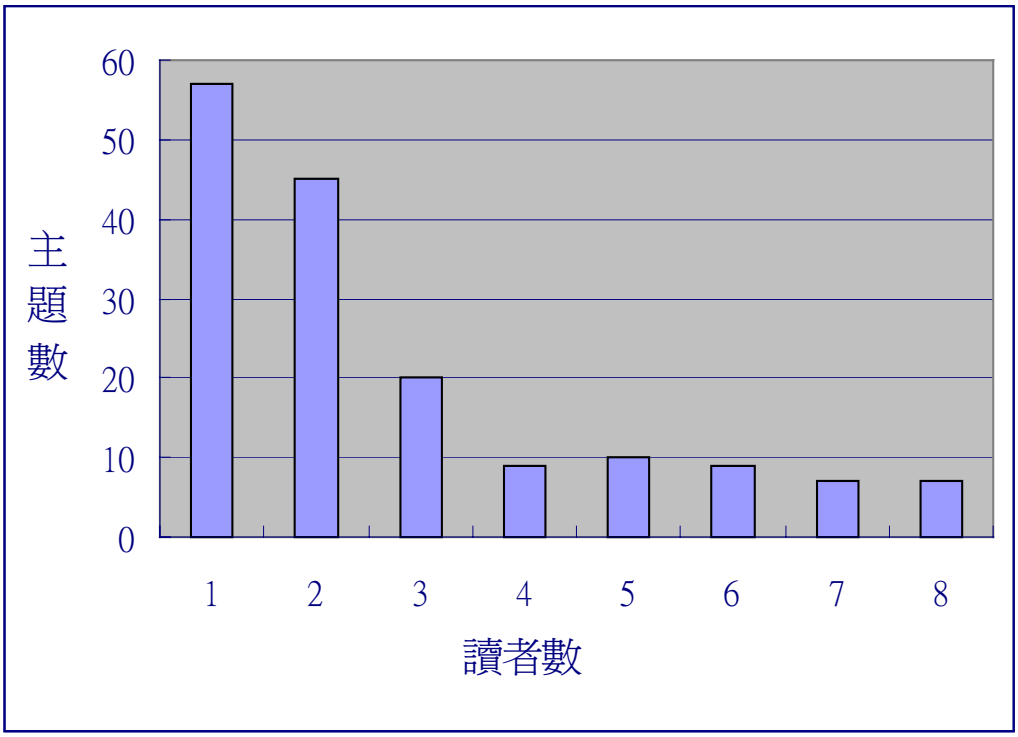
	讀者一	讀者二	讀者三	讀者四	讀者五	讀者六	讀者七	讀者八	AVG	STDEV
TEXT01	3	4	7	2	3	5	4	7	4.38	1.85
TEXT02	4	4	10	4	4	9	7	8	6.25	2.55
TEXT03	4	5	4	10	4	8	8	7	6.25	2.31
TEXT04	3	4	7	5	4	6	6	6	5.13	1.36
TEXT05	2	4	7	6	4	8	6	7	5.50	2.00
TEXT06	2	3	19	5	3	9	9	4	6.75	5.63
TEXT07	2	4	8	3	2	6	6	4	4.38	2.13
TEXT08	3	5	11	8	4	9	9	8	7.13	2.80
TEXT09	3	3	8	2	4	5	5	4	4.25	1.83
TEXT10	4	5	8	7	4	8	9	8	6.63	2.00
AVG	3.0	4.1	8.9	5.2	3.6	7.3	6.9	6.3		
STDEV	0.82	0.74	4.01	2.62	0.70	1.64	1.79	1.70		

表五 讀者判定主題的重複性

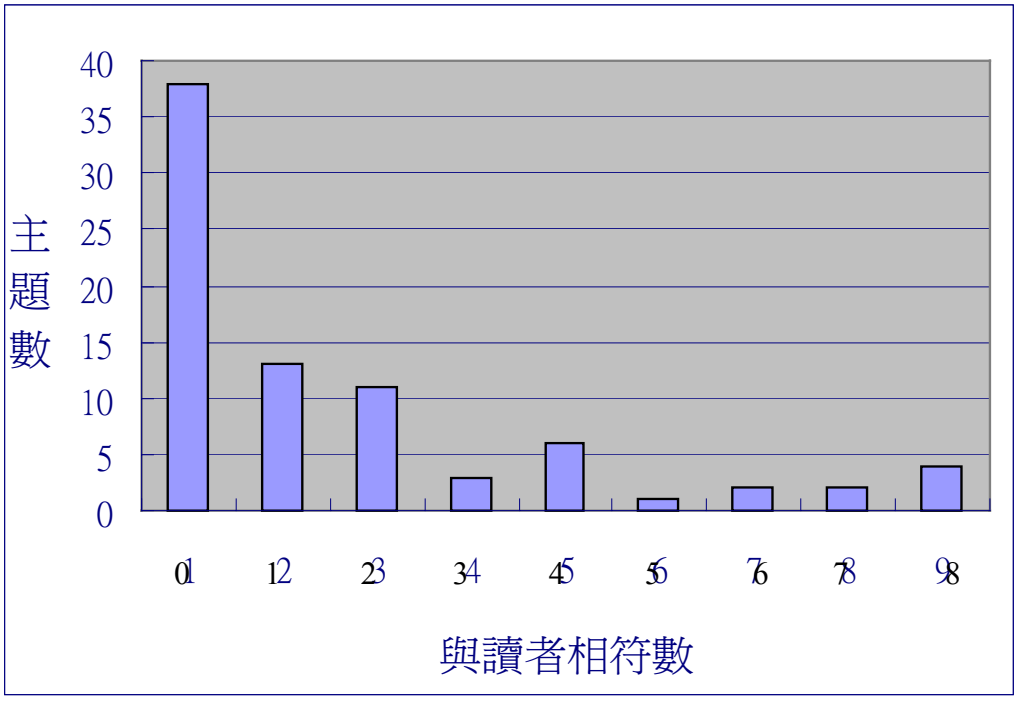
	一位	二位	三位	四位	五位	六位	七位	八位
TEXT01	2	4	1	0	0	1	0	2
TEXT02	4	4	3	2	0	0	3	0
TEXT03	4	7	0	1	2	3	0	0
TEXT04	5	3	0	1	1	1	1	1
TEXT05	0	3	1	1	2	2	0	1
TEXT06	18	5	2	0	0	1	2	0
TEXT07	5	1	4	0	2	1	0	0
TEXT08	7	8	4	3	2	0	0	0
TEXT09	4	3	4	0	1	0	1	0
TEXT10	8	7	1	1	0	0	0	3
SUM	57	45	20	9	10	9	7	7

表六 模型與讀者實驗結果的重複性

	無	一位	二位	三位	四位	五位	六位	七位	八位
TEXT01	0	1	1	0	0	0	1	0	1
TEXT02	1	1	2	1	2	0	0	1	0
TEXT03	1	1	2	0	1	0	0	0	0
TEXT04	5	1	1	0	1	0	0	1	1
TEXT05	6	0	2	0	0	0	1	0	1
TEXT06	5	3	0	0	0	0	0	0	0
TEXT07	9	0	0	0	0	0	0	0	0
TEXT08	2	2	2	0	2	0	0	0	0
TEXT09	5	1	0	2	0	1	0	0	0
TEXT10	4	3	1	0	0	0	0	0	1
SUM	38	13	11	3	6	1	2	2	4



圖一　讀者相似程度



圖二 系統模型與讀者的相似程度

六、結語與討論

本文探討電子文件主題自動辨識之可行性，藉由自動辨識模型之提出，實際進行實驗並與人類之主題辨識行爲比較，證實自動辨識模型的效能可與人類辨識相匹配。該模型引入人類閱讀與創作行爲的三個重要因素：

- 詞彙的重複性：重要意念必定重複出現，文章整體的意義才能夠收斂，讓讀者得到作者傳遞的訊息；重複出現的詞彙會讓讀者留下深刻的印象，讀者認定其做爲文章主題的可能性隨之提高。
- 詞彙的共現性：文章核心的意義是由周遭詞彙的逐步加強而越顯清晰，作者必須使用相關聯的詞彙以強調文章的主要意義；讀者亦經由詞彙彼此的關聯，逐步理解文章傳達的訊息。
- 詞彙的距離：隨著文章的長度加強，作者的意念會隨之稀釋，讀者可能會逐步忘卻之前文章傳達的訊息。作者必須在一定的文本範圍內再次強調文章的主要意念；讀者亦經由如此的強化過程，維持閱讀文章的整體性。

另外一個重要的因素--詞彙的重要性，則與文章的類型相關（Domain Dependency）。一般而言，不同的詞彙在不同類型的文章有其不同的重要性，這可由不同類型的文章使用詞彙的情形得知。衡量詞彙的重要性必須由收集的文章語料評估，如此才能反應詞彙的分佈情形，筆者使用史派克瓊斯提出的 IDF 衡量詞彙的重要性。

事實上綜合以上的因素，筆者提出的模型有一個隱而不現的前提，意即「作者是基於有意義的心智活動，將之轉化爲有組織結構的文章。」如果作者是隨意胡謅，基本上，系統模型也無法有效找出作者到底在說些什麼。

受限於時間與人力，筆者僅邀請八位讀者參與人工辨識的實驗，但是也有 80 份的實驗樣本。以實驗結果而言，讀者之間的分散程度（變異程度）確實不低，這個結果與眾多文獻提及圖書館員進行圖書文獻之主題標引工作時，常有的“Indexer Inconsistency”的現象一致。（註 30）至於模型與讀者的實驗結果相較，眾多讀者選定的主題，系統模型都能夠有效辨識；讀者傾向於分歧的文章，系統模型也有同樣的現象，顯示模型近似於讀者的閱讀行爲。

基本上，自動主題辨識的模型可以消除耗費大量人力的問題。然而由圖書館（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 Metadata 欄位不僅僅只有主題這一個欄位，尚有其他重要的欄位。若是考慮各類型不同藏品有其不同的需求，Metadata 的格式隨之不同，以美國爲例，研製完成的 Metadata 有其適用的範圍，如 GILS 用以描述美國政府公文（註 31）；FGDC 則用以描述地理資訊（註 32）；尚待完成的 Dublin Core（註 33）則被圖書館、資訊科學、電腦網路等領域的專家寄以厚望，希望能夠有效描述網路上的電子文件。（註 34）所以僅僅爲了如何有效描述資源，必須處理的狀況就已經非常複雜，而自動主題標引僅解決其中一項工作。目前已有一些自動辨識人名、地方名、組織名的系統（註 35），可以協助吾人加註文獻資料的 Metadata，但是困難的地方，卻是這一類的系統如何適應不同的 Metadata 格式，如何配合 Metadata 格式適當地變更加註的 Metadata。

爲了因應電子文件累積量成長越來越快的現象，再衡諸目前國內的各項相關研究，筆者認爲必須先制訂適用於中文文獻的 Metadata。目前國內除了機讀格式以外仍然沒有一套 Metadata，一些學者專家也看到這個問題。國立台灣大學積極進行的大型研究計畫「電子圖書館與博物館--文獻藏品數位化計畫」中（註 36），有一個研究小組正進行相關的研究，初步計畫是參考 CIMI（註 37）與 Dublin Core 的格式，並配合使用者需求的研究，發展適用於平埔族文物的 Metadata 格式，未來將推展到適用於其他文獻的 Metadata。

當然，電腦進行前述的自動程序所得結果品質的良莠，憑恃系統模型是否真正能夠有效模擬人類的行爲，學者專家是否能將隱而不見的知識轉化爲機器能夠使用的形式。在爲人類建立更好、更有用的資訊服務系統，仍然有許多困難有待克服，也存有很大的空間讓研究人員發揮想像力，在資訊檢索的研究領域上仍有很長一段路要走。

致謝

本論文部份成果歸功於中華民國國家科學委員會研究計畫 NSC-86-2621-E-002-025T 之補助。筆者感謝中央研究院詞庫小組建構的「中央研究院平衡語料庫」，提供實驗所需的語料。筆者還必須感謝研究助理以及參與實驗的八位同學，若沒有他們的協助，本論文是無法完成的。

【註釋】

註 1： Piatetsky-Shapiro, G. and W.J. Frawley. editors. Knowledge Discovery in Databases. Cambridge, MA: MIT Press, 1991.

註 2： Witten, I.H., A. Moffat and T. Bell. "Compression and Full-Text Indexing for Digital Libraries." Digital Libraries: Current Issues. Eds. N.R. Adam, B.K. Bhargava and Y. Yesha. Berlin: Srpinger-Verlag, 1995, 181-201

註 3： 陳雪華，圖書館與網路資源，台北市：文華圖書，民國 85 年。

註 4： Yahoo. (URL: http://www.yahoo.com/).

註 5： Alta Vista. (URL: http://altavista.digital.com/).

註 6： 計算語言學的研究在國外一直相當的活躍，國內雖然僅有少數的研究機構進行相關研究，但一直有很好的研究成果，如中央研究院的詞庫小組，是國內計算語言學研究的重鎮，在陳克健教授、黃居仁教授主持之下，建構了許多重要的語料庫，也發表眾多學術論文；國立清華大學電機工程學系蘇克毅教授、國立台灣大學資訊工程學系陳信希教授、國立清華大學資訊科學系蘇豐文教授、張俊盛教授也有極爲優異的研究成果。

註 7： 國家圖書館，中國機讀編目格式，台北市：國家圖書館，民國 86 年 6 月。

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

註 9： Library of Congress subject headings. 20th ed. Library of Congress, Cataloging Distribution Services, 1997.

註 10： McKinnon, Emma Jean, Carolyn Anne Reid. MeSH for searchers. Chicago: Medical Library Association, 1992.

註 11： 國立中央圖書館，中文圖書標題表，台北市：國立中央圖書館，民國 82 年四月。

註 12： Belkin, Nicholas J. Tutorial for Information Retrieval: Information Retrieval as Interaction, 1994.

註 13： Sparck Jones, K. "A Statistical Interpretation of Term Specificity and Its Application in Retrieval." Journal of Documentation 28.1 (1972): 11-21.

註 14： Salton, G. and C.S. Yang. "On the Specification of Term Values in Automatic Indexing." Journal of Documentation 29.4 (1973): 351-372.

Salton, G. "A Theory of Indexing." Proceedings of Regional Conference Series in Applied Mathematics, No. 18, Society for Industrial and Applied Mathematics, Philadelphia, PA,

1975.

Salton, G., C.S. Yang, and C.T. Yu. "A Theory of Term Importance in Automatic Text Analysis." Journal of the ASIS 26.1 (1975): 33-44.

註 15： Rocchio, J.J., Jr. "Relevance Feedback in Information Retrieval." The SMART System -- Experiments in Automatic Document Processing. Ed. G. Salton. New Jersey: Prentice-Hall Inc., 1971, 313-323.

Ide, E. "New Experiments in Relevance Feedback." The SMART System -- Experiments in Automatic Document Processing. Ed. G. Salton. New Jersey: Prentice-Hall Inc., 1971, 337-354.

註 16： 即下文所提及的葛拉茲、賴德樂、坎普等人。

註 17： Grosz, B. and C. Sidner. "Attention, Intentions, and the Structure of Discourse." Computational Linguistics 12.3 (1986): 175-204.

註 18： Kamp, H. "A Theory of Turth and Semanitc Representation." Formal Methods in the Study of Language. Eds. J. Groenendijk, T. Janssen, and M. Stokhof. Vol. 1. Mathematische Centrum, 1981.

註 19： Hearst, M. and C. Plaunt. "Subtopic Structuring for Full-Length Document Access." Proceedings of SIGIR-93, 1993, 59-68.

註 20： Chen, K.H. "Topic Identification in Discourse." Proceedings of th 7th Conference of the European Chapter of ACL, 1995, 267-271.

註 21： 同註 13。

註 22： 同註 12。

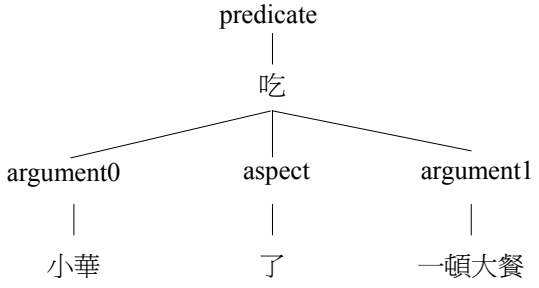
註 23： Youmans, G. "A New Tool for Discourse Analysis: The Vocabulary-Management Profile." Language 67 (1991): 763-789.

Reynar, J. "An Automatic Method of Finding Topic Boundaries." Proceedings of the 32nd Annual Meeting of ACL, 1994, 331-333.

註 24： 同註 18。

註 25： 同註 17。

註 26： 所謂的述語參數結構（Predicate-Argument Structure）指的是，句子中的動詞扮演述語（Predicate）的角色；而主詞與受詞扮演參數（Argument）的角色。若以「小華吃了一頓大餐」爲例，其述語參數的關係結構如下圖所示：



註 27： Church, K.W., and P. Hanks. "Word Association Norms, Mutual Information, and Lexicography." Computational Linguistics 16.1 (1990): 22-29.

註 28： 中央研究院詞庫小組，中央研究院平衡語料庫的內容與說明，技術報告 95-02。

註 29： 同註 28。

註 30： Hodge, Gail. Automated Support to Indexing. Philadelphia: The National Federation of Abstracting and Information Services, 1992.

註 31： GILS. "Guidelines for the Preparation of GILS Entries." 1995, (URL: http://gopher.nara.gov:70/0/managers/gils/guidance/gilsdoc.txt).

註 32： FGDC. "Content Standards for Digital Geospatial Metadata -- FGDC." 1994, (URL: http://fgdc.er.usgs.gov/fgdc.html.

註 33： 同註 8。

註 34： 有關 Metadata 更詳細的資訊，可以拜訪國際圖書館協會聯盟（IFLA）的網站 (http://www.nlc-bnc.ca/ifla/II/metadata.htm)，該站擁有極爲詳盡的 Metadata 資源，非常具有參考價值。

註 35： 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 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.

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 36： 台灣大學，電子圖書館與博物館--文獻與藏品數位化計畫，民國 86 年。

註 37： Consortium for the Interchange of Museum Information (CIMI), (URL: http://www.cimi.org/).