天主教輔仁大學圖書資訊學系碩士班碩士論文

指導老師：曾元顯

概念性類別標題詞自動擷取的評估

Evaluation of Generic Title Generation for Clustered Documents

研究生：陳秀涵撰

中華民國九十五年六月

目錄

第一章　緒論 4
　第一節　研究背景 4
　第二節　研究動機 4
　第三節　研究目的 5
　第四節　預期研究貢獻 5
第二章　文獻探討 6
　第一節　文件自動歸類方法 6
　第二節　標題詞選取方法 11
　第三節　WordNet 12
　第四節　InfoMap 15
　第五節　中央研究院中英雙語知識本體詞網 19
　第六節　路透社文件集 21
　第七節　FJU-CTC 文件集 23
　第八節　美國專利文件 25
第三章　研究設計 31
　第一節　實驗設計 31
　第二節　實驗工具 33
　第三節　測試文件集 39
　第四節　評估方法 40
　第五節　研究限制 41
第四章　研究結果與評估 43
　第一節　專利文件集 43
　第二節　路透社文件集 47
　第三節　FJU-CTC 文件集 49
第五章　結論與後續研究方向建議 55
　第一節　結論 55
　第二節　後續研究 58
參考書目 60

圖目錄

圖 1：階層凝聚式歸類.....11
圖 2：WordNet 使用者查詢介面示意圖 14
圖 3：WordNet 查詢結果示意圖 15
圖 4：InfoMap 相關詞彙查詢介面 17
圖 5：InfoMap 詞彙網路查詢介面 18
圖 6：InfoMap 詞彙網路查詢結果 18
圖 7：Sinica Bow 詞彙網路查詢介面.....20
圖 8：Sinica Bow 詞彙網路查詢結果.....21
圖 9：文件自動歸類流程.....32
圖 10：實驗流程圖.....32
圖 11：NSC 專利文件歸類結果視覺化 56

表目錄

表 1：REUTERS-21578（YANG 版本） 最大與最小的十類文件統計表 23
表 2：FJU-CTC 最大與最小的十類文件統計表 25
表 3：三種選詞方法用在三個文件集的評估結果..... 44
表 4：專利文件概念性標題詞評估結果..... 45
表 5：由 MI 選出的類別特徵詞 47
表 6：路透社文件集概念性標題詞評估結果..... 48
表 7：FJU-CTC 文件集標題詞與概念性標題詞評估結果 50
表 8：NSC 專利文件最後歸類結果 57

第一章 緒論

第一節 研究背景

文件歸類(Document Clustering，或稱文件聚類、文件分群)是一種應用於分析文件主題、依據文件主題關聯性歸納、組織文件的技術。與文件分類(Document Categorization)不同之處在於歸類是一種由下而上(bottom-up)的分析法，事先並無預先設計類別架構以及類別名稱，事先決定的參數只有欲切割的群數以及相似度門檻，然後由系統自動將相似的文件群聚在一起。由於這種特性，每個文件群聚的標題詞與主題必須在歸類之後才能加以判斷解釋。

通常，解釋群聚結果的方式是給每個群組標籤(label)，標籤的給定可以有許多種方式，本研究主題即為自動化產生歸類後文件群組的概念性標題的方法。

第二節 研究動機

如前所述，針對歸類後文件需要給定標籤作為協助解釋該群聚主題的輔助，一般給定類別標籤的方式是由文件群組中挑選重要詞彙作為類別標題詞，這樣的方式雖然具有適切性，但是產生的類別標題卻不一定是最具有解釋性的概念性標題，特別當文件涵蓋的主題較廣泛的時候，只取文件用詞作為標題詞不見得能夠適切表達整體宏觀的概念。在做專利或是科學文獻的分析時，這樣的問題特別明顯。

有鑑於此，本研究想嘗試透過外部階層式語料庫的輔助，以文件群聚中取出的辭彙的共同上位詞(hypernym)，找出適合文件群組的概念性標題。舉例來說，假設一類別中取出的代表詞彙為"桌子"、"椅子"、"床"，最適當的概念性

標題詞應該是"傢俱"。然而，這樣的詞彙不一定是文件用詞，因此從文件中取詞不一定可以取得真正具代表性的概念性類別標題詞。

此問題通常由專家給定標題詞來解決，若要做到自動化，就必須採用外部語料庫像是 WordNet，或是其他具備階層式概念的語料庫供我們查詢詞彙的上位概念。本研究首先先由歸類後的文件中選取文件使用的特徵詞，再透過上位詞搜尋方法將這些詞彙對應到階層式語料庫中的概念性標題詞，再由人工評估其適切度。

第三節 研究目的

1. 評估自動化產生歸類後文件的概念性標題詞的方法，以及此方法是否比直接挑選文件用詞來得具詮釋性。
2. 評估此方法用在不同文件集的成效，以了解適用範圍。

第四節 研究貢獻

1. 驗證透過外部語料庫對於標題詞的產生是否有助益。
2. 驗證此方法在不同特性的文件集實施的成效以推估未來適用範圍。
3. 高品質的概念標題詞擷取技術有助於將歸類技術應用在多方面協助使用者：例如資訊過濾或查詢結果的即時分群。

第二章 文獻探討

本研究中最重要的幾個因素為：文件自動歸類方法、文件標題詞選取方法、階層式語料庫以及測試文件集，以下將針對這四大部分作相關文獻回顧與說明。

第一節 文件自動歸類方法

文件歸類的步驟分為特徵值擷取與篩選、文件呈現以及歸類三大部分[1]。

一、 特徵值擷取與篩選

文件的最小組成元素為詞彙，特徵值即為最能代表文件的辭彙。Frakes & Baezay[2]認為文件中的高頻詞與文件主題有較高的關聯性，能作為文件的特徵值(feature)。然而若一詞彙在每篇文件都出現，則該詞彙對文件便不具有代表性[3]。由於文件歸類的結果會因為文件的多維度向量(即特徵值的數目)及資料不足導致歸類成效下降[4]，因此減少特徵值的數量便成為重要的處理程序。

從文件中擷取特徵值主要有三種模式：

1. 詞庫比對法

利用事先建立的詞庫對文件進行比對，將文件中出現在詞庫的辭彙擷取出來。其優點為簡單，且正確性高，不會擷取出無意義的辭彙;缺點是需要耗費人力與時間事先建立詞庫，且詞庫的品質，涵蓋的領域，新詞、人名、地名、機關名…等專有名稱難以窮舉，且詞庫越大比對的速

度就越慢。

2. 文法剖析法

透過自然語言處理技術的文法剖析程式，剖析出文件中的名詞片語，再運用一些方法與準則，過濾掉不適合的詞彙。其結果幾乎也都是有意義的名詞片語，但大部份的剖析程式，需要藉助已經建立的詞典或語料庫[5]，因此其缺點也和詞庫比對法一樣。除此之外，有些文法剖析法甚至只能剖析合乎文法的完整文句，使得書目、標題等資料裡的關鍵詞無法被擷取出來。

3. 統計分析法

透過對文件的分析，累積足夠的統計參數後，再將統計參數符合某些條件的片語擷取出來。最簡單的統計參數是計數詞彙發生的頻率，即詞頻，將詞頻落在某一範圍的詞彙取出。由於沒有用到詞庫或語料庫，會有擷取錯誤的情況發生，得到無意義或不合法的詞彙。此外，統計參數不足的關鍵詞無法被選到。然而其優點是較不受語文國別與句型的限制，而且可以擷取出未曾被詞庫、語料庫網羅的專業用語、新生詞彙與專有名稱等片語。

一些文獻對特徵值擷取成效有初步的探討，其中Arppe[6]以文法剖析方式試驗其擷取成效，結果發現大約80%-99% 的關鍵詞為名詞片語，而且雖然名詞片語的擷取準確率與召回率皆可達 95% 以上，然而具代表性的名詞片語不到總數的 50%，因此單純剖析出名詞片語後，仍需要依據其他特徵以過濾掉不要的詞彙。Godby[7]則比較文法剖析法與統計分析法的優劣，發現統計分析法除了可

以跟文法剖析法做得一樣好之外，亦具備簡單、不受語文國別與句法的限制、以及可同時過濾不具代表性片語的優點。

特徵值擷取在中文文件上因為語文本身的特性，與英文又有不同。中文的特色是詞彙之間沒有空白，此外中文詞彙是一開放性的集合，任意的組合都可以是新的辭彙。

國內對中文特徵值自動擷取的問題也有研究。清大自然語言處理實驗室曾嚐試擷取關鍵詞作為書後索引(book index)，其主要方法為運用電子字典協助斷出詞彙[8]，再以統計方式配合自然語言處理技術剖析名詞片語，最後再設定過濾條件，篩選索引詞彙[9]。在成效評估方面，以一本軟體使用手冊為對象，相對於人工製作的索引，其精確率與召回率可同時達到 63% 的程度。至於導致錯誤的主要來源有：斷詞錯誤（42%）、統計特徵不足（39%）、以及無法處理複雜語法結構（19%）。

中央研究院資訊科學研究所也有關鍵詞自動擷取運用在資訊檢索的研究。其主要作法乃先建構一種稱為 PAT-tree 的資料結構，再輔以詞頻等統計特徵擷取出關鍵詞[10] [11]。PAT-tree 雖然在資訊檢索上有相當優良的特性，不過其建造過程需耗費相當長的時間，例如，建構 600 Mega bytes 的資料需要一個星期的時間[12]。可以想見，此種方式的有效運用，必須要能改進 PAT-tree 的建構速度。

曾元顯[13]提出一種關鍵特徵擷取技術，運用統計分析

法斷出關鍵詞，沒有用到辭典、語料庫、或自然語言處理的技巧。因此具備擷取速度快、擷取的正確率高（82%-100%）、中英文均適用、擷取的詞彙沒有長度限制、可同時擷取廣義詞與狹義詞等特性。

擷取出文件特徵值後，需要依照一些準則篩選最能代表文件的辭彙，因為一份文件中約有 90%的辭彙不具代表性[14]。特徵值的篩選方法應用在歸類上主要為文件頻率及詞彙強度兩種模式[15]，說明如下：

1. 文件頻率(Document Frequency)

詞彙在文件集裡出現的文件篇數。可設定一門檻值，僅保留高於門檻的辭彙。

2. 詞彙強度(Term Strength)

用來評估詞彙與文件之間的相似度。根據詞彙在相似文件中出現的機率，將大於門檻值的詞彙留下。

由於每篇文件的長度不同，文件中特徵值出現的次數與分部也不同，因此必須將文件根據特徵值做權重計算，常用的方法是 Salton 提出的三種模式[16]：

i. 布林法(Boolean)：計算特徵權重最簡單的方法，詞彙在文件中有出現則權重為 1，反之則為 0。

ii. 詞頻法(Term Frequency)：直接以詞彙在文件中出現的次數作為權重。

iii. 詞頻乘以文件頻率之倒數(Term Frequency-Inverse Document Frequency,TF-IDF)：TF 為一特徵值 i 在文件 j 出現的次數，IDF 為所有文件出現特徵值 i 文件數的

倒數，最後將兩值相乘，作為特徵值 i 在文件 j 中的權重，N 文文件集文件總數。公式如下：

$$W_{ij}=Tf_{ij}*\log(N/DF_{ij})$$

二、 文件呈現

過去研究多以 Salton 提出的向量空間模式(Vector Space Model)為主。文件經過特徵值選取轉換之後，每個特徵值代表空間中一個維度分量，每篇文件由多個特徵值組成，即為多個分量所組成的多維度向量。

三、 歸類模式

一般常用的歸類方法為階層凝聚式歸類(Hierachical Agglomerative Clustering)及非階層歸類法中的 K-means 歸類[17]。前者初始時將每篇文件視為一個叢集，循序合併相似度高的叢集，其過程如一個樹狀圖形如圖 1 所示。後者則將所有的文件視為一個大叢集，依照文件相似度對叢集進行分割的動作。在所有歸類演算法中，雖然階層凝聚式歸類的運算速度較慢，然而其成效最佳[18]。其中又以華德法(Ward ‘s Method)以及完全連結法(Complete Linkage Methos)成效較佳[19]。

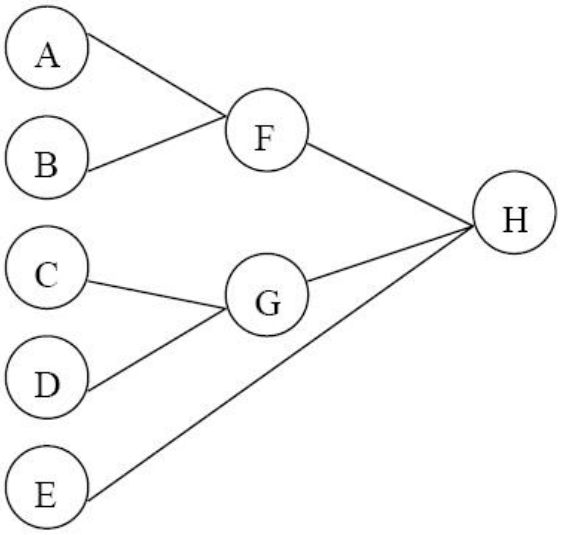

圖 1：階層凝聚式歸類

第二節 標題詞選取方法

對歸類後文件做主題分析與給定標籤是困難的任務。自動化的方法通常仰賴自文件當中擷取顯著的代表性詞彙，舉例而言，當使用向量空間模型(VSM Model)來描述文件叢集時，是以文件向量的加權總合或其中心點(centroid)來表示。而在這些文件向量中，具有最高權重值的詞彙將用來作為此文件叢集的標題。舉例來說，在 Cutting, et. al 及 Marit A, et. al 提出的叢集方法中[20] [21]，是使用正規化的出現頻率（term frequency, TF）作為文件向量中每個詞彙的權重值；而在 Yiming Yang, et al 提出的方法中[22]，則是使用 TF 與反轉文件頻率（inverse document frequency, IDF）的乘積作為權重值。Mehran Sahami，Salim Yusufali 以及 Michelle Q. W. Baldonaldo[23]針對不同公式比較成效，研究指出基本的質心公式勝過條件機率，即使條件機率的公式尚額外計算了詞彙在該類別與其他類別出現的機率。

在 Krista Lagus, et. al 所提出的 SOM 模型中[24]，類別以二維的圖形呈現，每個類別的標籤則是用類別中最高分數的詞，除以其他叢集中的相關詞總數正規化而來。

Russell Swan, et. al[25]所提出的分類詞彙以偵測事件的應用中則是基於事件主題偵測的需求，將排序在前的名詞與名詞子句作為類別標題。這些詞彙的順位是藉由將一時間間隔內出現之詞彙的卡方（chi-square）值排序而得。

Oren Zamir, et. al[26]為了處理 Web 上查詢結果的歸類問題，選用在最多文件中出現的最長句子為標題詞。

其他研究領域像是文件自動摘要或是自動翻譯的相關會議如文件理解協會(Document Understanding Conference, DUC)[27]的主要任務在於找出萃取出文章的極短摘要的方法。這些短約 10 字的摘要也有作為標題詞的潛力。然而大部分參賽者仍然採用文件中萃詞的方法[28]，而這些方法仍需藉由一個具有人工指定標題的文件集合來訓練出一個「翻譯模型」，才能夠將文件字彙對應（map）到人工指定標題[29]。此外，這些摘要對於文件集合來說，多半是傾向於事件描述導向，而非主題描述導向。

可見即使有許多技術在做類似的工作，仍然沒有研究可以解決我們所提出的問題。

第三節 WordNet[30]

美國普林斯頓大學認知心理實驗室(Princeton

University Cognitive Science Laboratory)的心理學家，語言學家和電腦工程師聯合設計的一種基於認知語言學的英文語料庫，計畫主持人為 George A. Miller。語料庫的特性在於按照詞彙的意義與詞彙之間的關係組成語意網路。收錄了英文的名詞、動詞、形容詞與副詞。

WordNet 的建立源自於 1985 年起使用 Kučera 和 Francis 的現代英文版標準語料庫（通常被稱為布朗語料庫， Brown Corpus)，因為它提供不同詞類的頻率。然而 Henry Kučera 告知這個語料庫的語意標記資料已經售予 Houghton Mifflin，因此 WordNet 取而代之使用 Richard Beckwith 在 1988 年發展的同義詞索引，同時也併入 Charles Osgood 用來發展語意學差異的形容詞詞組。由於同義詞對 WordNet 而言非常重要，因此 WordNet 採用了許多同義詞字典的資訊，包括 Laurence Urdang（1978）的「基本同義詞與反義詞」(Basic Book of Synonyms and Antonyms)，Urdang（1978）改編 Rodale 的「同義詞搜尋家」(The Synonym Finder)，以及 Robert Chapman（1977）的第四版「Roget 國際同義詞」(Roget's International Thesaurus)。1986 年，WordNet 拿到 Fred Chang 在 Naval 個人研究與發展中心所編撰的文字列表，比較結果僅有 15%重疊。因此 Chang 的列表亦被納入 WordNet。1993 年，Ralph Grishman 等在紐約大學建構的 COMLEX 通用詞彙有 39143 個字；WordNet 僅含括其中 74%。因此 WordNet 也採用了 COMLEX。

簡而言之，WordNet 包含來自多方的資源，事實是英文詞彙非常廣泛，即使到現在 WordNet 依然不斷地在收錄新

字。[31]

建構語意網路有助於協助我們判斷詞彙表達的觀念以及協助我們認知其意涵。舉例來說，當我們在 WordNet 中查詢「airplane」，可以得到一個詞義(sense)，若一詞有多種詞性 WordNet 也會加以註明。

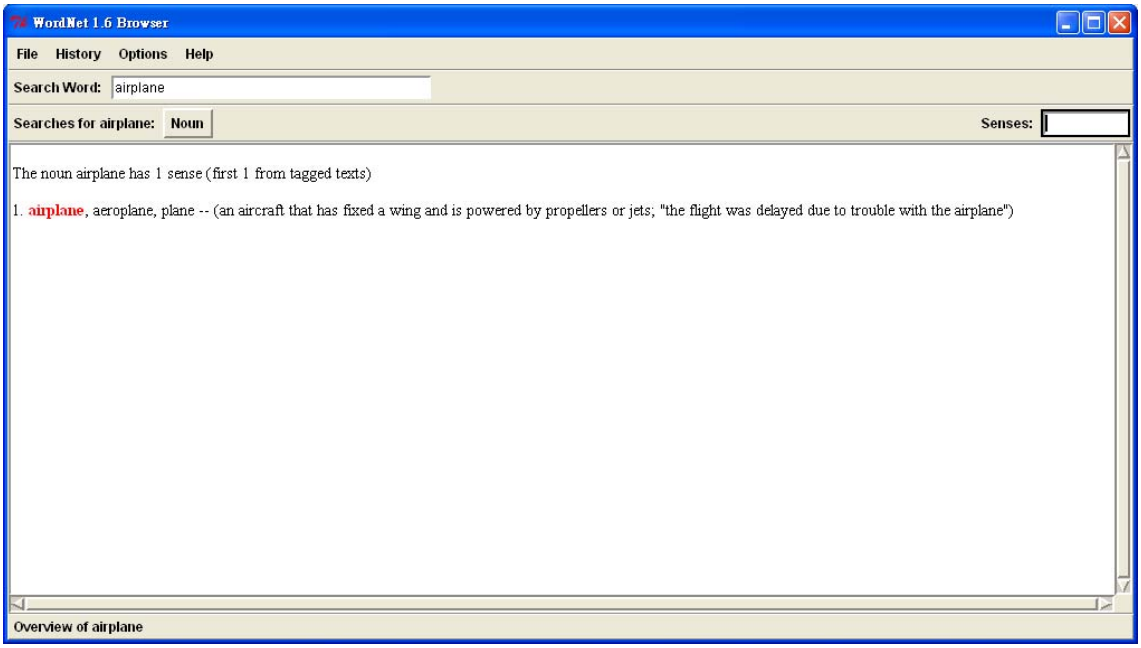

圖 2：WordNet 使用者查詢介面示意圖

我們還可以進一步查詢「airplane」的上位詞(airplane is a kind of…)，查詢結果會依照階層架構顯示，最接近的上位詞為「aircraft–a vehicle that can fly」，更高階的上位詞為「craft-- a vehicle designed for navigation in or on water or air or through outer space」。最高階則是「entity,something–anything having existence(living or nonliving)」，如圖 3所示。

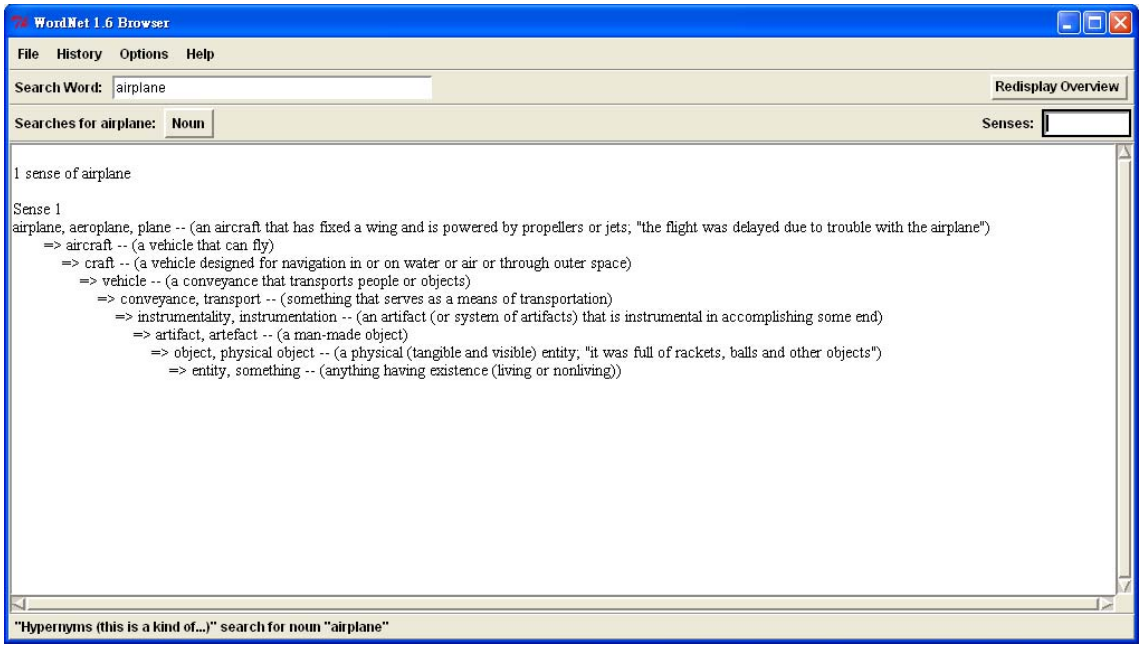

圖 3：WordNet 查詢結果示意圖

因此，WordNet 非常適合作為產生概念性標題的輔助。WordNet 提供了應用程式介面(API)供外界使用，因此可以直接下載安裝，透過程式輸入查詢詞，得到查詢的結果。也提供如圖 2、圖 3所示的使用者介面供使用參考。

WordNet 目前已經發展到 2.1 版，本研究基於與 Perl 套件的相容性採用的是較舊的 1.6 版本。

第四節 InfoMap[32]

InfoMap 自然語言處理軟體(The Infomap NLP Software，以下簡稱 InfoMap)是由史丹福大學語言學習中心(Stanford University's Center for the Study of Language and Information)的計算語言實驗室(Computational Semantics Lab)發展的專案計畫所建構。專案稱為「Information Mapping project」。計畫主持人為 Stanley Peters 教授。

這個專案的目的在於以「語用」(詞彙在文句中如何使用)的角度來理解詞彙的意義。這對語言學習、機器翻譯以及智慧型資訊檢索有極大的幫助。現今大部分的鉅量文件搜尋像是網際網路的搜尋引擎、圖書館查詢系統或是歷史新聞檢索，都是採用關鍵字比對方法。使用者的查詢以一連串關鍵詞表達，資料庫中的文件只要符合任何關鍵詞就會被回傳給使用者。然而，如果我們不單將關鍵詞視為單純字串，而是針對概念(concept)做查詢，查詢結果將更加貼近使用者的需要。因此 InfoMap 的建立就是希望能夠分析字彙的涵義。

與 InfoMap 相關的技術說明如下。

一、 向量空間模式(Vector Space Models)

Hinrich Schütze 是 InfoMap 向量空間模式(或稱 WORDSPACE)的發展先驅 。此模式式的運作是透過計算不同詞彙在同一篇文章的共現頻率(co-occurrence)，將詞彙對應到一個多維度的向量空間。單一詞彙與其他相同主題的關鍵詞共現狀態的分布成為一個變量(profile)，用以呈現詞彙的「語用」，並且能夠精確地將同義詞連結在一起。藉由這樣的運算，我們可以用一個主題範圍的內的多重詞彙表達單一關鍵詞，使用者可以依照需求挑選、理解其意涵。總結來說，透過轉換，我們可以將查詢詞轉換為該詞彙的 profile，並將這個 profile 與文件形成的 profile 做比對，以使查詢結果更符合概念上的相關，而不僅是字面上出現與否。

二、 連結分析模式(Link Analysis Models)

連結分析在現今是一種常用的方法,特別是用於網際網路上查詢結果的排序以及分析社群網絡。這項技術對於

將相似詞彙歸類相當有幫助，並且能夠辨別字彙的歧義。

三、　字典、領域知識以及醫學領域(Dictionaries, Knowledge Bases, and the Medical Domain)

為了建立語義模型，這個專案執行了一項開創性的研究，結合自動化模型以及傳統的字典以及領域知識。舉例來說，我們可以直接採用醫學標準語言系統(Unified Medical Language System，UMLS)作為自動化辨別語義的工具，使用在醫學搜尋引擎上。[33]

上面提到的相關技術在 InfoMap 的網站上都有雛型展示系統。舉例來說，我們可以選用 New York Times 作為來源語料庫查詢「train」，系統會提示相關詞彙供選擇，如圖 4。

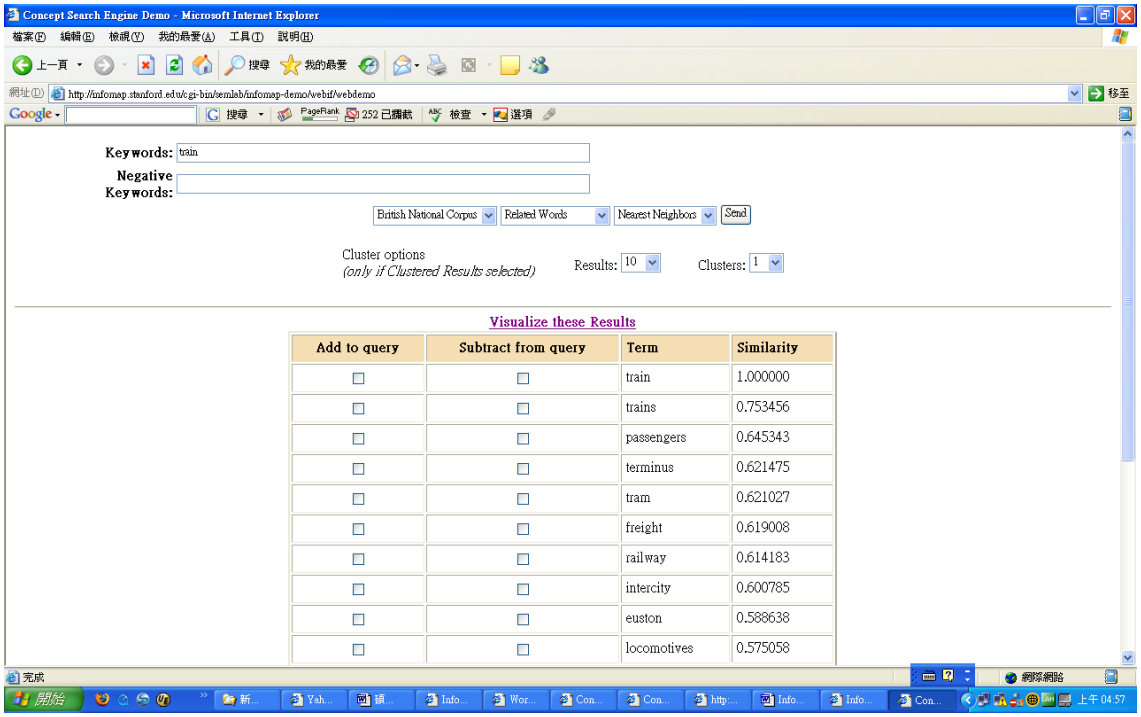

圖 4：InfoMap 相關詞彙查詢介面

亦可選擇查詢字彙的語義網路，一樣以「train」為例。

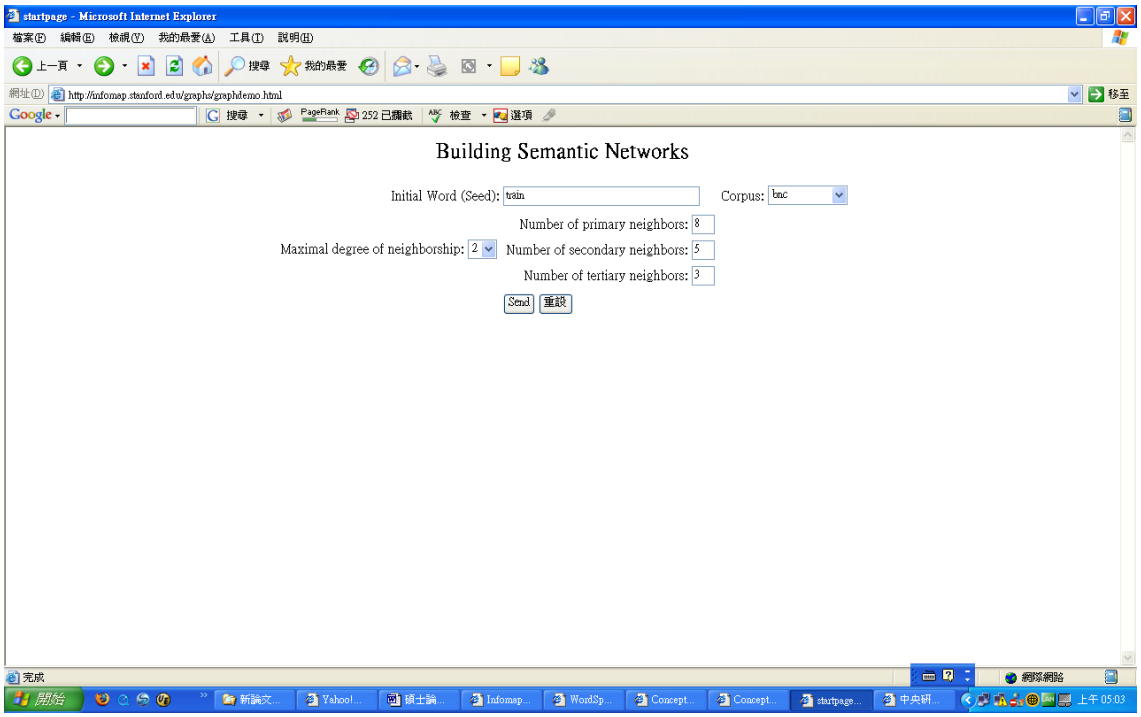

圖 5：InfoMap 詞彙網路查詢介面

查詢結果展開為星狀圖，呈現了字彙間的關聯。

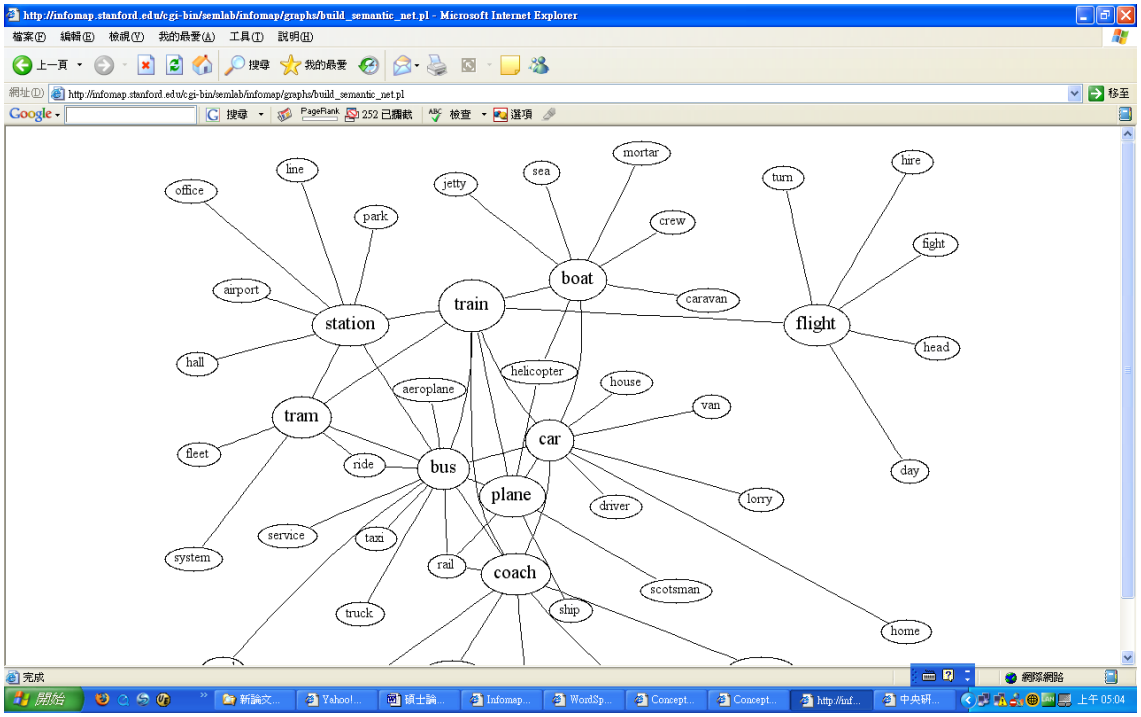

圖 6：InfoMap 詞彙網路查詢結果

透過 InfoMap，我們可以送出查詢詞，並得到查詢詞的上位詞。。

第五節 中央研究院中英雙語知識本體詞網[34]

中英雙語知識本體詞網(Sinica BOW)以英文 WordNet 架構為基礎，並以以台灣地區的語言使用為經驗基礎。提供的訊息包含中英雙語跨語言資訊轉換、語言資訊與概念架構(知識本體）的連結、詞義的區分與詞義關係的連結以及使用領域，在使用語言與詞彙資料的基礎上，提供了知識運籌的基本架構(infrastructure)。讓不同來源的典藏知識內容，可以轉換成互通的 (inter-operable) 訊息。所引用的資料主要為中央研究院語言所文獻語料庫（http://corpus.ling.sinica.edu.tw/)，資訊所詞庫小組（http://ckip.iis.sinica.edu.tw/CKIP/）開發的資料外，另外引用了 IEEE 批准執行的 SUMO（Suggested Upper Merged Ontology，http://www.ontologyportal.org）知識本體及普林斯頓大學的 WordNet（http://www.cogsci.princeton.edu/~wn/)，以及遠見科技股份有限公司與中研院共同開發資料。

此系統有幾個特性：首先，這是一個雙語系統，因此允許選擇以中文或是英文進行查詢，同時查詢結果也包含中英文上位詞(中文由 SUMO 提供)。第二，使用者能夠在一個介面上同時查得 WordNet 或是 SUMO 的語言結構。第三，系統建立了多語索引以供其他應用，第四，領域知識提供詞彙多面向的解讀與運用。

簡而言之，Sinica BOW 是一個以結構化方式呈現知識本體的資料庫，對研究語意網路以及計算語言學的任務提供了相當大的幫助。

如前所述，我們可以在 Sinaca BOW 提供的 Web 介面輸入查詢詞，我們以中文做範例，輸入「火車」，見圖 7：

圖 7：Sinica Bow 詞彙網路查詢介面

查詢結果提供多種語料庫的說明，假設我們選擇 WordNet 1.6，與 WordNet 類似，系統將傳回這個詞彙的領域、詞性、解釋、翻譯、同義詞、上位詞、以及在 SUMO 語料庫中的概念說明(concept)見圖 8。

圖 8：Sinica Bow 詞彙網路查詢結果

系統的缺點是沒有提供 API，受限於此，實驗只能用人工的方式處理查詢，無法自動化。

第六節 路透社文件集

根據 Lewis 的描述[35]，此測試集來自於 1987 年由 Reuters Ltd.與 Carnegie Group, Inc.的專家進行整理與分類的路透社新聞稿。1990 年此份資料提供麻州大學(University of Massachusetts)電腦與資訊科學系資訊檢索實驗室使用。David D. Lewis 和 Stephen Harding 對其進行資料格式與相關資料檔的建置工作。1991 與 1992 年間，在芝加哥大學資訊與語言研究中心 David D. Lewis 與 Peter Shoemaker 進一步的整理其格式與檔案。此時的版本稱之為「Reuters-22173, Distribution 1.0」，並於 1993 到 1996 年間置於麻州大學的 anonymous FTP Server 上供人下載。

1996 年 8 月在 ACM SIGIR 的會議中，一組研究團體談論到 Reuters-22173 在不同研究中的結果難以比較的問題後，大家決定應該要制定新版的測試集，希望降低其格式上的模糊問題，提供文件清楚說明其標準的使用方法，並降低許多打字、分類及格式上的錯誤。

Steve Finch 和 David D. Lewis 隨後在 1996 年的 9 月至 11 月中完成了該項工作，並大量引用了 SGML 標籤，同時除去了 595 篇重複的文件（根據文件的時間標籤，去除連秒數都相同的重複文件），最後形成了有 21578 篇文件的新版測試集「Reuters-21578, Distribution 1.0」。

在此測試集中，有五棵「分類樹」（或稱「分類架構」或「類別集合」（category set)），分別是 EXCHANGES, ORGS, PEOPLE, PLACES 與 TOPICS。每篇文件都根據其內容給定每個分類樹內適當的類別，而且可以給定多個類別。例如“nasdaq” (EXCHANGES), “gatt” (ORGS), “perez-de-cuellar” (PEOPLE), “canada” ,” japan” (PLACES), “gold” (TOPICS)。基本上，除了 TOPICS 外，其他分類樹的分類法則大體上可以根據文件內是否有提到其類別名稱即可分類，如文章中提到 australia, nasdaq 等，當然這些名稱最好是該文章的主題焦點。至於 TOPICS 的分類法則就沒那麼明顯，因此是文件自動分類最常用的分類樹[36]。

針對 TOPICS，此測試集將其分成訓練組與測試組的分割作法（split)共有三種，稱為：ModLewis、ModApte、ModHayes。其中最常用的是 ModApte 分割，其訓練組有 9603 篇文件，測試組有 3299 篇文件，類別總數約 120 類。然而 Yang 在其

研究中[37]，根據 ModApte 分割，將沒有類別的文件刪除，同時也限制只有在訓練組與測試組都有文件的類別才留下來。如此類別總數剩下 90 類，訓練組文件則有 7769 篇，測試組有 3019 篇，平均每篇文件有 1.3 個類別。本研究使用的 Reuters-21578 測試集，都是根據 Yang 的修改原則而來。

同樣的，此測試集的類別分佈也很不平均，表 1 列出最大與最小的十個類別的文件篇數。

表 1：Reuters-21578（Yang 版本） 最大與最小的十類文件統計表

篇數最多的十個類別				篇數最少的十個類別			
編號	類名	訓練	測試	編號	類名	訓練	測試
1	earn	2877	1087	81	palladium	2	1
2	acq	1650	719	82	palmkernel	2	1
3	money-fx	538	179	83	rand	2	1
4	grain	433	149	84	castor-oil	1	1
5	crude	389	189	85	cotton-oil	1	2
6	trade	369	118	86	groundnut-oil	1	1
7	interest	347	131	87	lin-oil	1	1
8	wheat	212	71	88	nkr	1	2
9	ship	197	89	89	rye	1	1
10	corn	182	56	90	sun-meal	1	1

第七節 FJU-CTC 文件集[38]

FJU CTC 測試集的來源為 1966 年到 1982 年之間中國大陸各電台的廣播內容，由當時在香港的中國消息分析（China News Analysis, CNA）天主教機構，為蒐集當時難以獲得的大陸消息請人每天收聽大陸廣播轉寫文字而成，此資料集簡稱「廣播抄稿」。西元 1994 年，CNA 的資料（以及人員）由香港移轉到輔仁大學社文中心。2000-2001 年間，由國科會計畫補助，將廣播抄稿由人工打字成電子檔案，共計 42371 篇廣播稿，其中只有 30710 篇文件有人工標示的類別。曾元

顯[39]根據下列自原則來篩選文件，以製作成此分類測試集：

（一） 盡可能使用所有可用的文件，以充分運用既有的資源，包括人工的分類與電子化的檔案。

（二） 為了訓練並測試分類器的成效，文件集分成兩部分：訓練組（training set）與測試組（testing set)。所有分在訓練組裡的文件其日期必須早於所有分在測試組裡的文件，如此才能反映一般分類器的實際使用情況。

（三） 為了有效訓練與測試分類器，每一個類別必須在訓練組與測試組中都有文件。若不符合此要求的文件與類別，則不納入分類測試集中。

（四） 分類不一致的同一主題文件，最好能夠被指出來或隔離出來，以降低成效評估的不可靠性。

根據上述原則，約 70%的文件為訓練文件、30%文件為測試文件。最後符合前三個原則的文件在訓練組有 19901 篇，日期範圍為 1966/01/01 到 1976/12/31，而測試組有 8110 篇，日期範圍為 1977/01/01 到 1982/12/26，共計 28011 篇。由於文件來自於廣播手抄稿以及人工打字，因此錯字、漏字，甚至句子或段落錯漏的情況，不算少見。但這並非缺點，任何文件都有這些情況發生，只是程度不同。此項特色，甚至可以考驗分類器容忍這些雜訊的能力。

香港天主教 CNA 機構（以及之後的輔大社文中心）根據自行發展的分類表，對其館藏（包含七八十萬篇新聞剪報以及此廣播抄稿）進行人工分類。符合上述原則的類別共計 82 類，類別分佈呈現高度的偏向，亦即少數類別有很多文件，而大多數類別只有少數文件。表 2 列出最大與最小的十個類

別的文件個數。

表 2：FJU-CTC 最大與最小的十類文件統計表

篇數最多的十個類別				篇數最少的十個類別			
編號	類名	訓練	測試	編號	類名	訓練	測試
1	P52	3985	1073	73	E39	9	2
2	P10	1903	788	74	E83	7	18
3	Cu4	1115	236	75	E1a	7	7
4	E34	1085	88	76	Cu73	5	2
5	E3	950	248	77	HK	4	10
6	P5	854	536	78	E1-3	4	2
7	E6	797	234	79	E31	4	1
8	P4	736	501	80	E1-4	3	74
9	E5	569	149	81	P9	3	1
10	E36	548	310	82	P53	2	9

在類別標示方面，大部分文件只給一類，少數給超過一類，最多到四類，但多重分類的比例很少，因此其每篇文件的平均類別數，在訓練組為 1.035 類，在測試組為 1.007 類。由於此文件集橫跨 17 年之久，類別標示的一致性恐難保持[40]。但任何分類測試集都有此問題，只是程度不同。

第八節 美國專利文件

美國專利全文為 HTML 格式，內容記載專利號（Patent Number）、申請日期（Filing Date）、公告日期（Date of Patent）、發明人（Inventor）、申請人（Assignee）、美國分類號（UPC）、國際分類號（IPC）、引用參考資料（References Cited）、相關的美國專利（U.S. Patent Documents）、專利名稱（Title）、專利摘要（Abstract）、專利宣告（Claim）、專利說明（Description）等項目。其中專利說明以文字詳細描述該發明創作，通常包含下列項目：發明領域(Field of the Invention)、發明背景（Background of the Invention）、

發明摘要（Summary of the Invention)、圖式簡述（Brief Description of the Drawings)、發明細節（Detailed Description of the Invention）等。這些資料都必須從半結構化的 HTML 檔裡剖析出來，尤其專利宣告與說明的部份，更必須從自由文字(free text)中，分段擷取如發明領域、發明摘要等段落，瞭解其待解問題及解決方法，以便進行後續如摘要、分類、歸類等處理。

美國專利全文沒有用 XML 標示，導致程式無法輕易剖析理解其內容，所幸其 HTML 的標示對短欄位資料的部份還算規則，可以用特殊的網頁剖析器（wrapper）透過正規表達式（regular expression）的字串比對方式，擷取出內容。至於專利說明的全文部份，比較沒有規則，但還是可以在自由文字中比對大寫關鍵字的文句，擷取出發明領域、發明背景、發明摘要、圖式簡述、發明細節等段落，而自動產生其專利說明的內容目次(Table of Content)。至於專利宣告的部份，雖然也是自由文字，但分項說明，且寫法很傳統：獨立宣告項先寫，其後緊跟著依附宣告項。獨立宣告項為主要的專利標的，依附宣告項通常用來補充、說明獨立宣告項的細節，因此可以根據宣告寫作的習慣，自動擷取出獨立宣告項。

注釋

[1] Wei,C.P.,Hu P.J.&Dong,Y.X.,"Managing Document Categories in E-Commerce Environments:an Evolution-Based Approach,,"*European Journal of Information Systems*,Vol.11,No.3,2002,pp.208-222

[2] Frakes, W. B. & Baezay, R., "Information Retrieval: Data Structures and Algorithms," New Jersey: Prentice-Hall, 1992.

[3] Salton, G. & Buckley, C., "Term Weighting Approaches in Automatic Information Retrieval," Journal of Information Proceeding and Management, Vol. 24, No. 5, 1988, pp. 513-524.

[4] Aggrawal, C. C., & Yu, P. S., "Finding Generalized Projected Clusters in High Dimensional Spaces," Proceedings of the 2000 ACM SIGMOD international conference on Management of data, Dallas, New York: ACM Press, 2000, 70-81.

[5] 陳光華，"資訊檢索查詢之自然語言處理"，中國圖書館學會會報，第 57 期，85 年 12 月，頁 141 - 153 。

[6] Antti Arppe, "Term Extraction from Unrestricted Text," http://www.lingsoft.fi/doc/nptool/term-extraction.html, 1995.

[7] Jean Godby, "Two Techniques for the Identification of Phrases in Full Text," http://www.oclc.org/oclc/research/publications/review94/part1/twotech.htm .

[8] Jen-Nan Chen, Jyun-Sheng, Chang and Huey-Chyun Chen, "Using Word Segmentation Model for Compression of Chinese Text" http:// nlplab.cs.nhtu.edu.tw/~mathis/own/html/PAPER/JNL/95/cpcol/ CPCOL95.htm

[9] Mathis H. C. Chen, Tsong-Yi Tseng, Jason J. S. Chang, "Automatic Generation of Indices for Chinese Books," http://nlplab.cs.nthu.edu.tw/~mathis/own/html/ PAPER/JNL/96/cpcol/BookIdx.htm

[10]簡立峰，"尋易系統（Csmart）與中文智慧型資訊檢索"，資訊傳播與圖書館學，3 卷 2 期， 85 年 12 月，頁 28-37。

[11] Lee-Feng Chien, "PAT-Tree Based Keyword Extraction for Chinese Information Retrieval" ACM SIGIR 1997.

[12] William B. Frakes and Ricardo Baeza-Yates, *Information Retrieval: Data Structure and Algorithms*, Prentice Hall, 1992.

[13] Yuen-Hsien Tseng, "Fast Keyword Extraction of Chinese Documents in a Web Environment," Information Retrieval Workshop for Asia Languages - 1997, Oct. 8-9, Japan, pp. 81-87.

[14] 林傑斌、劉明德、陳湘，「資料採掘與 OLAP 理論與實務」，台北：文魁書局，2002。

[15] Liu, T., Liu, S. & Chen, Z., 2003, "An Evaluation on Feature Selection for Text Clustering," Proceedings of the Twentieth International Conference on Machine Learning, Washington, CA: AAAI Press, pp. 488-495.

[16] Salton, G., "Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer," New York: Addison-Wesley, 1989.

[17] Steinbach, M., Karypis, G. & Kumar, V., "A Comparison of Document Clustering Techniques," Technical Report 00-034, Computer Science and Engineering, University of Minnesota, 2000.

[18] Dubes, R. C. & Jain, A. K., "Algorithms for Clustering Data," New Jersey: Prentice Hall, 1988.

[19] Griffith, A., Luckhurst, H. C. & Willet, P., "Using Inter-Document Similarity Information in Document Retrieval Systems," Journal of the American Society for Information Sciences, Vol. 37, No. 1, 1986, pp. 3-11.

[20] Cutting, D. R. Karger, J. O. Pedersen, and J. W. Tukey. “Scatter/gather: A cluster-based approach to browsing large document collections,” Proceedings of the 15th ACM-SIGIR Conference, 1992, pp. 318-329.

[21] Marti A. Hearst and Jan O. Pedersen, "Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results," Proceedings of the 19th ACM-SIGIR Conference, 1996, pp. 76-84.

[22] Yiming Yang, Tom Ault, Thomas Pierce and Charles W. Lattimer, “Improving Text Categorization Methods for Event Tracking,” Proceedings of the 23rd ACM-SIGIR Conference,, 2000, pp. 65-72.

[23] Mehran Sahami, Salim Yusufali, and Michelle Q. W. Baldonaldo, “SONIA: A

Service for Organizing Networked Information Autonomously," Proceedings of the 3rd ACM Conference on Digital Libraries, 1998, pp. 200-209.

[24] Krista Lagus, Samuel Kaski, and Teuvo Kohonen, "Mining Massive Document Collections by the WEBSOM Method," Information Sciences, Vol 163/1-3, pp. 135-156, 2004.

[25] Russell Swan and James Allan, "Automatic Generation of Overview Timelines," Proceedings of the 23rd ACM-SIGIR Conference, 2000, pp. 49-56.

[26] Oren Zamir and Oren Etzioni, "Web Document Clustering: a Feasibility Demonstration," Proceedings of the 21st ACM-SIGIR Conference, 1998, pp. 46-54.

[27] Document Understanding Conferences, http://www-nlpir.nist.gov/projects/duc/.

[28] Michele Banko, Vibhu O. Mittal, and Michael J. Witbrock, "Headline Generation Based on Statistical Translation," ACL 2000.

[29] Paul E. Kennedy, Alexander G. Hauptmann, "Automatic title generation for EM," Proceedings of the 5th ACM Conference on Digital Libraries, 2000, pp.

[30] http://wordnet.princeton.edu/

[31] Wordnet: An Electronic Lexical Database , pp. xviii-xix

[32] http://infomap-nlp.sourceforge.net/

[33] http://infomap.stanford.edu/index.html#papers

[34] http://bow.sinica.edu.tw/

[35] David D. Lewis, "Reuters-21578 text categorization test collection, Distribution 1.0" README file (v 1.2), 1997, http://www.research.att.com/~lewis/

[36] Franca Debole and Fabrizio Sebastiani, "An Analysis of the Relative Hardness of Reuters-21578 Subsets" to appear in Journal of the American Society for Information Science and Technology.

[37] Yiming Yang and Xin Liu, "A Re-Examination of Text Categorization Methods," Proceedings of the 22nd Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, 1999, Pages 42 – 49.

[38] http://blue.lins.fju.edu.tw/~tseng/Collections/Chinese_TC.html

[39] 曾元顯，"分類不一致對文件自動分類效果的影響"，大學圖書館，9 卷 1 期，2005 年 3 月，頁 11-13

[40] Yuen-Hsien Tseng and William John Teahan, "Verifying a Chinese Collection for Text Categorization," Proceedings of the 27th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '04, July 25 - 29 Sheffield, U.K., 2004, pp.556-557.

第三章 研究設計

本研究採用實驗法，實驗法是指在控制情境下，有系統地操弄獨變項，使其依照預定計畫而改變，然後觀察紀錄依變項是否也會對應地產生有系統的改變。實驗法欲驗證的是獨變項(原因)與依變項(效果)之間的因果關係。

其一般執行步驟如下：決定實驗目的→提出假設→以操作定義界定變項→準備實驗工具→控制無關干擾變項→選擇實驗設計→進行實驗→資料統計分析→根據結果撰寫報告。

如第一章所述，本研究的目的在找出一種概念性類別標題自動擷取方式並評估成效，有別於以往研究直接由文件中取詞，我們假設透過外部語料庫可以取得更具詮釋性的標題詞，然而是否具詮釋性牽涉主觀認定，因此必須由人工方式評估。研究設計的細節將由後面幾個章節做細部陳述。

第一節 實驗設計

由自動歸類的流程可以清楚看出實驗的設計。

如圖 9，多個非結構化文件，經過步驟一的歸類演算法，將主題相似的文件聚合在一起成為數個叢聚，之後再透過步驟二，標題詞擷取方法，給定每個叢聚的類別名稱，有了類別名稱之後，就可以對歸類結果做出解讀。

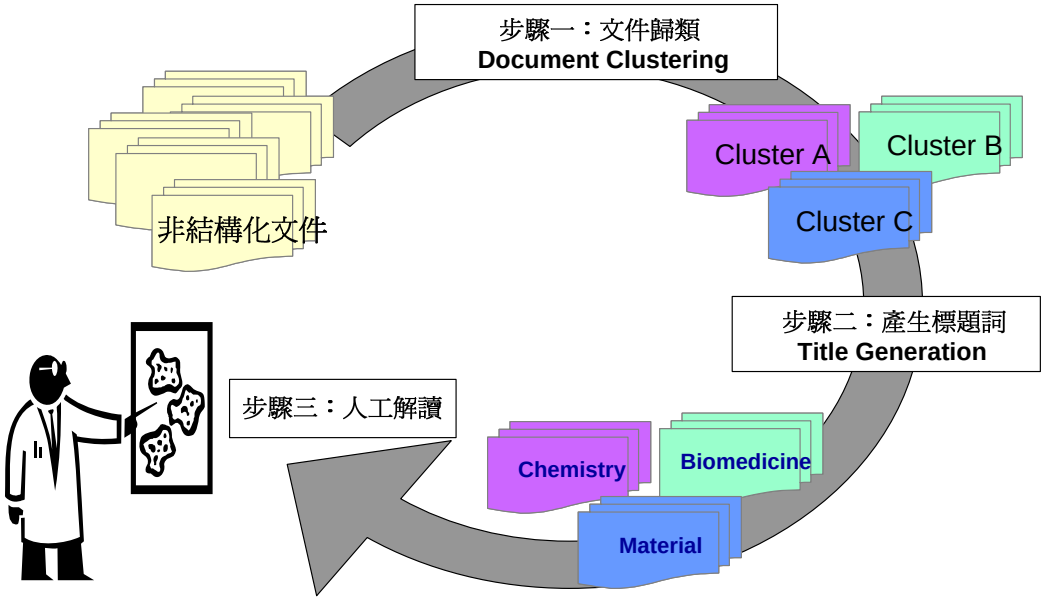

圖 9：文件自動歸類流程

從文獻分析中可知，現有的研究對標題詞擷取採用的方式，大都是從文件叢聚中擷取重要的辭彙。本研究要做的就是在此之後加上第三個步驟，以文件叢聚中挑選出來的重要辭彙為基礎，運用外部語料庫查詢這些詞彙的共同上位詞，以期能夠找出更具詮釋性的概念性標題詞。其流程如圖 10 所示。

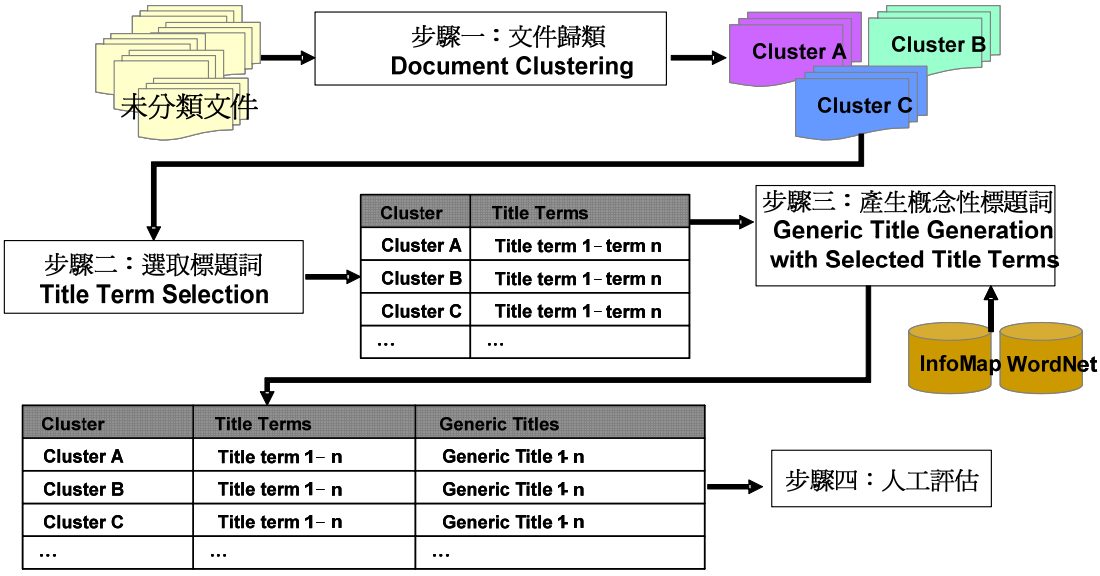

圖 10：實驗流程圖

由圖中可以清楚看出實驗的六大變因為：測試文件集、文件自動歸類工具、類別詞擷取工具、外部語料庫、上位詞查詢方法、以及成效評估方法。由於本研究欲探討的是外部語料庫對產生標題詞的幫助，因此實驗設計會在控制其他變因的情況下單獨比較"自文件叢聚中擷取的標題詞"與"標題詞的共同上位詞"何者較具適切性與解釋性。

實驗採用兩種模式：

一、當文件集沒有類別標籤，就將文件做自動歸類，之後由人工判定兩階段產生出來的標題詞何者較具適切性。

二、當文件集本身具有類別標籤(例如 Reuters 21578)，就省略歸類的步驟，直接以原先的標籤將文件分類，取出類別的代表詞，以代表詞查詢共同上位詞，之後人工需判定兩個部分：原始類別詞與取出的代表詞之優劣，以及代表詞與上位詞何者較具適切性。

第二種實驗模式是輔助本研究的實驗。這個實驗可以檢驗類別詞擷取技術的成效。假設評估的結果上位標題詞不夠好，有可能證明我們的假設不正確，也有可能是因為關鍵詞擷取方式選出來的詞彙品質不佳影響上位詞。因此透過第二種模式，我們可以先比較類別特徵擷取與原先人工給定詞彙之間的差異，以此驗證先前研究提出的 CC、TFC 或 TFxIDF 實際實施的成效。不過我們在現實世界要解的問題通常文件本身並不具備原始類別標籤。

第二節 實驗工具

一、 文件自動歸類與類別詞擷取工具

本研究採用 WebGenie Text Mining Tool(以下簡稱 WGTM)作為文件自動歸類的工具。此工具為一台灣軟體廠商威知資訊股份有限公司自行研發的產品，具備本研究所需的功能，同時提供完整的應用程式介面(API)供外界呼叫。

1. 索引(Indexing)：可剖析中、英文文件，做斷詞處理紀錄在索引檔。WGTM 的斷詞採用中華民國第 153789 號發明專利「數位文件關鍵特徵之自動擷取方法」，不需建立詞庫就可斷出文件中的中文專有名詞及關鍵詞。是可跨領域的文字處理工具。
2. 自動歸類(Clustering)：WGTM 提供 API，外界可設定歸類模式為 complete link 或 single link。本研究採用 complete link。
3. 特徵辭彙選取(Term Selection)：WGTM 特徵詞選取採用相關係數(corelation coefficient)公式。

二、 外部語料庫

研究採用了兩個不同的英文語料庫以及一個可輸入中文查詢的語料庫，避免以單一結果斷言整體結論。

1. WordNet：由普林斯頓大學心理研究室(Cognitive Science Laboratory at Princeton University)建立的詞彙知識庫。包含了將近 20 萬個英文字義及其語意關係，將同義詞集合成同義詞集（synset)，並標示同義詞集之間的關係，如上位詞、下位詞、部分、包含等語意關係，可公開、免費的使用。
2. InfoMap：史丹佛大學計算語言實驗室(Computational Semantics Laboratory at Stanford University)建立的上位詞查詢系統。

3. 中央研究院中英雙語知識本體詞網：以英文 WordNet 架構為基礎，並以台灣地區的語言使用為經驗基礎。提供的訊息包含中英雙語跨語言資訊轉換、語言資訊與概念架構（知識本體）的連結、詞義的區分與詞義關係的連結以及使用領域，在使用語言與詞彙資料的基礎上，提供了知識運籌的基本架構(infrastructure)。

三、 概念性標題自動擷取方法

利用自行發展的方法取得叢集中的代表性詞彙，找出詞彙的共同上位詞，再依具權重決定最後的概念性標題詞。

此方法流程敘述如下：

首先從多篇文件中摘錄出多個特徵詞彙。其中，摘錄特徵詞彙的方法包括先利用斷詞策略將所有文件進行斷詞處裡，而得到許多候選詞彙，接著再從這些候選詞彙中選出合適的詞彙作為特徵詞彙。

為了避免在篩選特徵詞彙時，受到不同的歸類方法的影響，本研究採用計算各個詞彙 T 與類別 C 之間的相關係數（Correlation Coefficient，CC），來決定特徵詞彙的取用與否，相關係數 CC 的計算公式如下：

$$CC(T,C) = \frac{(TP \times TN - FN \times FP)}{\sqrt{(TP + FN)(FP + TN)(TP + FP)(FN + TN)}}$$

其中，TP（true positive）、FP（false positive）、FN（false negative）、以及 TN（true negative）分別代表屬於類別 C 且包含詞彙 T 的文章總數、不屬於類別 C 但包含詞彙 T 的文章總數、屬於類別 C 但不包含詞彙 T 的文章總數、以及不屬於類別 C 且不包含詞彙 T 的文章總數。

在計算所有候選詞彙與類別的相關係數之後，即可根據每個候選詞彙對應的相關係數的大小或其排序，例如判斷其是否大於某一特定值（例如 0.7）或排序在前面，來決定是否選擇這個候選詞彙作為特徵詞彙。

另一方面，較佳的作法為再計算每一個候選詞彙在此些文件中的出現頻率（Term Frequency in the Cluster，TFC），並將每一個候選詞彙的相關係數 CC 與其出現頻率 TFC 相乘以取得一乘積（CCxTFC），最終再選取乘積大於某個特定值或排序在前面的候選詞彙作為特徵詞彙。

除了上述計算候選詞彙之相關性的作法外，我們還包括計算在文件叢集中，有出現該候選詞彙之文件的數量，當此數量大於某個特定值（例如佔全部文件的百分之五十）時，即可選擇此候選詞彙作為特徵詞彙。

總括上述篩選特徵詞彙的方法，相關係數 CC 適於計算作為從大量短篇文章中摘錄出特徵詞彙的參考。反之，當文章篇幅過長時，則適合用相關係數 CC 與出現頻率 TFC 的乘積 CCxTFC 來作為選取特徵詞彙時的依據。

在決定出這些文件的特徵詞彙之後，則接著在一個階層式知識結構中，找尋對應每個特徵詞彙的所有概念詞，並計算出各個概念詞的階層深度與選取次數。其中，選取次數是指在以這些特徵詞彙找尋概念詞時，每一個概念詞被選取到的次數總和。

階層式知識結構本身可視為是由一個根節點與許多子節點所構成的一種樹狀結構，其中無論是根節點亦或是子節點皆用來表示一組同義詞彙，而此根節點與子節點間或是各個子節點間的連結，則是透過一語義關係來建立。以 WordNet

為例：其知識結構中包括了許多同義詞彙（即英文單字），並且會根據例如上位詞關係（hypernym）、下位詞關係（hyponym）、整體關係（holonym）、或是附屬關係（meronym）等語義關係來建立同義詞彙彼此之間的關係。這些同義詞彙又可以被歸類到例如名詞集合、動詞集合、形容詞集合、或是副詞集合等不同的同義詞集合之中。在上述的階層式知識結構中，子節點越接近根節點代表其語義關係更為上位，而其所對應的同義詞彙也越廣泛通用。

在階層式知識結構中，一個詞彙通常會具有一個以上的概念詞（hypernym），而一個詞彙可能也是其他多個詞彙的概念詞。比如說，dog 的概念詞例如包括 animal、creature；而 gesture 則是 nod、shrug、以及 hug 的概念詞。本研究採取的方法為從每個對應特徵詞彙的同義詞彙其所代表的子節點往根節點之路徑上，所有遇到的子節點均會被選做為這些特徵詞彙的概念詞。

值得注意的是，在階層式知識結構當中，從一個子節點通往根節點的路徑可能不只一個。因此，在計算概念詞的選取次數時，無論這些路徑會經過代表同一個概念詞的節點多少次，對於同一個特徵詞彙來說，此概念詞的選取次數都只能計算一次。另外，除了選取次數之外，還必須計算每個概念詞的階層深度。我們假設根節點之階層深度為 0，而其下一層子節點之階層深度為 1，以此類推。因此，藉由計算每個節點距離根節點的階層數目，即可得到該節點所代表之概念詞的階層深度。

在階層式知識結構中，越上層（即階層深度越低）詞彙的詞義越廣，因此在選擇概念詞為文件的概念性標題時，若

選擇較為廣義的概念詞，將無法明確的表示這些文件的內容；而若選擇較為狹義的概念詞，則無法概括所有文件的內容。因此，最合適的方法就是在能涵蓋所有特徵詞彙的概念詞中，選擇具有最高階層深度的概念詞作為這些文件的概念性標題。

根據上述原則，下一步驟則是根據這些概念詞在階層式知識結構內的階層深度與選取次數，計算出這些概念詞的權重值。其中，假設概念詞的權重值正比於階層深度之 S 型函數（Sigmoid function）值，以及正比於選取次數，而以下列公式計算出每個概念詞的權重值 weight：

$$weight(hypernym) = \frac{f}{nt} \times 2 \times \left(\frac{1}{1+\exp^{-c \times d}} - 0.5 \right)$$

其中，f 表示概念詞的選取次數，d 表示概念詞所在之階層深度，而 nt 是用來表示所有特徵詞彙的總數，c 則是一個常數（例如 0.125）。

最後，將每個概念詞之權重值大小作排序，並選取具有最大權重值的概念詞作為這些文件的概念性標題。藉由在階層式知識結構找出同時能夠盡可能概括所有特徵詞彙的概念詞作為概念性標題，而不僅限於文章中既有的文字，較之先前技術，所獲得的標題更能夠概括及代表這些文件的內容。

舉例來說，假設從多個文件中摘錄出的特徵詞彙包括 table、chair、bed，而在 WordNet 1.6 版這個階層式知識結構中，經由上述計算方法，所獲得之最高權重值的概念詞是 furniture（其權重值為 0.3584）。顯然地，furniture 將比 table、chair、或 bed 更適合用來作為這些文件之概念性標題。

第三節 測試文件集

本研究採用了下面幾個文件集，文件集特性描述如下：

一、 612 篇屬於國科會的美國專利文件：在 2005/6/15 由 USPTO 查詢下載。查詢條件設定專利權人為”National Science Council”(NSC)。NSC(國家科學委員會)是台灣一個贊助科學研究的機構。國內的協會、大學或研究中心，無論公立或私立均可向國科會申請研究補助。一旦研究申請專利，智慧財產權即屬於國科會，亦即專利權人為國科會。(然而此項政策在 2000 年已經改變，因此自 2000 年之後國科會的每年專利件數就不如以往。) 因為這樣的背景，這六百多篇專利的特性是進階技術的知識內容包羅萬象，涵蓋各種領域，且每篇都是 2000 字以上的長文件。要分析這些專利文件的主題便非常困難，因為很難有人能夠了解所有領域的知識。雖然專利文件本身有 IPC 或 UPC 類別編碼，但是這些標籤都太一般性或太特殊，跟我們預期中的主題分析不盡相同。所以使用這些類別標籤無法滿足分析師的需求。就此而論，自動化建議的類別標籤在協助人解讀文件的作業上就扮演重要的角色。

二、 Reuters-21578 前十大類 6018 篇：在文件自動分類領域中被廣為採用的路透社文件集。我們採用文件集的一部分作為實驗對象。此文件集的特性是短文

章，(每篇大約 133 字)內容為經濟相關領域。每篇文章都被給定了一個以上的類別標籤以表示其類別，根據 Yang[1]的前處理結果，文件集包含有 90 個類別。本研究採用前十大類作為評估的對象。

三、 FJU-CTC 28011 篇：中文新聞廣播抄稿。這個文件集起源自一項進行多年數位典藏計畫，由國科會補助，輔仁大學社會科學院進行。文件集內容來自於大陸中央廣播電臺自 1966 年至 1982 年的新聞廣播節目。這些廣播內容在當時由人工的方式抄錄下來並給定類別標籤。當時的目的是藉此揭示文化大革命期間大陸內部的狀況。在 2000-2001 年間，輔大社文中心藉由數位典藏計畫將 42371 分手稿以人工方式打字建檔以利保存與後續應用。其中有 30710 篇文件有類別標籤與日期。依照下面的原則，這 30710 篇文件被整理為 FJU-CTC 文件集供自動分類實驗使用：

- 每個類別都必須有訓練文件以及測試文件，對自動分類器來說才有幫助。
- 所有的訓練文件日期都比測試文件早，以反映現實世界中處理的問題。

第四節 評估方法

研究邀請兩位主修圖書資訊學的學生來評估標題詞與文件主題的相關性。評估人員判定最具解釋性的標題詞就得到

一分，可複選，並且假若難以判斷亦可不選。標題詞產生的方式法被編碼過，因此評估者不知道他們看到的標題詞是用哪一種方式產生，亦不知道採用何種外部語料庫。

第五節 研究限制

1. 受限於歸類演算法的時間複雜度(O(n3))，歸類文件集文章篇數必須小於 8000 篇，此為 WGTM 系統限制。
2. 外部語料庫目前僅有英文工具(WordNet、InfoMAP)提供介面工外部程式呼叫，中文的部分需人工處理。

注釋

[1] Yiming Yang and Xin Liu, “A Re-Examination of Text Categorization Methods,” Proceedings of the 22nd ACM-SIGIR Conference, 1999, pp. 42-49.

第四章研究結果與評估

第一節專利文件集

一、 文件歸類方法

612 篇 NSC 的專利文件應當切割為幾個類別最適當很難有定見，因此本研究嘗試多種方法以不同面向獲得多種歸類結果。具體而言併用了文件歸類以及詞彙歸類兩種分析方式。

首先，對這 612 篇專利文件做分析、過濾、取詞幹、以及摘要的處理。專利文件中具備的結構化資訊像是專利權人、發明人以及 IPC 都被移除，不做分析處理。其餘的文字敘述部分依照專利文件的格式切割成幾個主要的段落。由於這些段落的長度迥異，因此我們透過 extraction-based methods[1]從中各選取六個最重要的句子。以這些句子替代原始文件，做詞彙選取、詞彙共現分析、索引以及歸類的對象。因為專利的 Claim 本身具備有法律特性，其用詞較為特殊，因此我們沒有將這部分加入處理。

從上述句子中萃取出 19,347 個關鍵詞，過濾條件為詞彙 DF >=2(詞彙至少出現在兩篇文件)，利用這些詞彙及其關聯詞做詞彙共現分析，以找出哪些詞彙具有共同的關聯詞對。分析結果僅有 2714 個詞出現在兩篇以上的文件並且有至少一個關聯詞。這些詞彙依照其關聯詞切割成 353 個小叢集。重複這個步驟，再將 353 個叢集合併為 101 個中等規模的叢集，再合併變為 33 個大叢集，最後合併得到共 10 個文件叢集。之所以需要分階段處理是因為即使一開始將門檻值設的很低，也無法切割出 10 個叢集。必須要循序合併。藉由這樣的過程，文件依序被合併為關鍵詞、概念、主題、領域。這

種方式的優點在於能對大量文件集做快速且有效的歸類。

另一個主題分析法是直接採用選出的重要句子，利用 complete link 模式對文件做歸類。 最後歸出 6 個大類。

二、 類別特徵詞選取方法

為了做比較，我們採用第二章敘述的三種權重模式對選出的辭彙做排序，即 TFC，$CC_{0.5}$ 以及 CCxTFC。這三種權重模式分別用在三個歸類結果。亦即上述的三個步驟產生的文件叢集。每個叢集都選出最多五個詞彙作為標題詞。

兩位評估人員評估標題詞與文件主題的相關性。三種方式中被評為最具解釋性的標題詞就得到一分。可複選，並且假若難以判斷亦可不評估。評估結果可見表二。我們任評估人員可以瀏覽整個階層架構，並依照自己的喜好選擇主題，因此不同人員評估的叢集數目略有不同。

儘管有這樣的不同點，初步仍可看出評估者偏好 $CC_{0.5}$ 或 CCxTFC 選出的詞，且都不選擇 TFC 選出的詞。這對我們有相當的幫助。因為過去研究多使用 TFC。

表 3：三種選詞方法用在三個文件集的評估結果

排序方法 / 評估對象	評估人員編號	評估樣本數	TFC		$CC_{0.5}$		CCxTFC	
第一步驟歸類結果	1	73	5	7%	52	**71%**	20	27%
	2	19	6	32%	12	63%	13	**68%**
第二步驟歸類結果	1	13	0	0%	9	**69%**	6	46%
	2	7	1	14%	3	43%	5	**71%**
第三步驟歸類結果	1	16	5	31%	12	**75%**	9	56%
	2	10	5	50%	4	40%	8	**80%**

三、　產生概念性標題及成效評估

最後一個步驟產生的類別詞以前面所敘述的概念性標題詞演算法處理，得到其共同上位詞。文件以不同方式分別被切成成 10 類以及 6 類，以 CCxTFC 的方式選出最多五個類別代表詞。見表 3。我們採用 WordNet 及 InfoMap。InfoMap 似乎也採用 WordNet 作為參考，因為回傳的詞彙許多都與 WordNet 相同。不過 InfoMap 並沒有公佈技術細節，所以本研究無法自行實作共同上位詞的演算。但是 InfoMap 提供了代理程式，讓我們可以送出查詢詞，得到回傳的結果。我們只比較 WordNet 及 InfoMap 產生的前三個詞彙。見表 4 的後兩欄列出各自的詞彙與權重分數。

用粗體字標示的是被斷定與主題相關的詞。由表中可見兩種方式成效差不多，在兩種歸類結果中，成效都至少有 50%。

表 4：專利文件概念性標題詞評估結果

ID	Cluster's Descriptors	WordNet	InfoMap
1	acid, polymer, catalyst, ether, formula	1:substance, matter:0.1853 2:drug:0.0980 **3:chemical compound:0.098**	**1:chemical compound:1.25** 2:substance, matter:1.062 3:object,physical object:0.484
2	silicon, layer, transistor, gate, substrate	1:object,physical object:0.1244 2:device:0.1211 3:artifact, artefact:0.1112	1:object,physical object:0.528 2:substance, matter:0.500 3:region, part:0.361
3	plastic, mechanism, plate, rotate, force	1:device:0.1514 2:base, bag:0.1155 3:cut of beef:0.1155	1:device:0.361 2:entity, something:0.236 3:chemical process:0.0
4	output, signal, circuit, input, frequency	**1:communication:0.1470** **2:signal,signaling, sign:0.1211** 3:relation:0.0995	**1:signal,signaling, sign:1.250** **2:communication:1.000** 3:abstraction:0.268
5	powder, nickel, electrolyte, steel, composite	1:substance, matter:0.1483 **2:metallicelement, metal:0.1211** 3:instrumentation:0.0980	**1:metallicelement, metal:0.500** 2:substance, matter:0.333 3:entity, something:0.203
6	gene, protein, cell, acid,	1:substance, matter:0.1112	1:entity, something:0.893

	expression	2:object,physical object:0.0995 3:chemical compound:0.0980	2:chemical compound:0.500 3:object,physical object:0.026
1	resin, group, polymer, compound, methyl	1:substance, matter:0.1853 **2:chemical compound:0.098** 3:whole:0.0717	1:substance, matter:2.472 2:object,physical object:0.736 **3:chemical compound:0.5**
2	circuit, output, input, signal, voltage	**1:communication:0.1470** **2:signal,signaling, sign:0.1211** 3:production:0.0823	**1:signal,signaling, sign:1.250** **2:communication:1.000** 3. round shape:0.000
3	silicon, layer, material, substrate, powder	1:substance, matter:0.1483 2:object,physical object:0.1244 3:artifact, artefact:0.1112	1:substance, matter:2.250 2:artifact, artefact:0.861 3:object,physical object:0.833
4	system, edge, signal, type, device	1:artifact, artefact:0.1483 **2:communication:0.1470** 3:idea, thought:0.0980	1:instrumentality:1.250 **2:communication:0.861** 3:artifact, artefact:0.750
5	solution, polyaniline, derivative, acid, aqueous	1:communication:0.1633 2:legal document,:0.1540 3:calculation, computation:0.1372	1:drug of abuse, street drug:0.000 **2:chemical compound:0.0** 3:set:0.000
6	sensor, magnetic, record, calcium, phosphate	1:device:0.1514 2:object,physical object:0.1244 **3:sound/audio recording:0.12**	1:device:0.312 2:fact:0.000 3:evidence:0.000
7	gene, cell, virus, infection, plant	1:structure, construction:0.1225 2:contrivance, dodge:0.1155 3:compartment:0.1029	1:entity, something:0.790 **2:life form,living thing:0.5** 3:room:0.000
8	density, treatment, strength, control, arrhythmia	1:property:0.1112 2:economic policy:0.1020 3:attribute:0.0995	1:power, potency:1.250 2:property:0.674 3:condition, status:0.625
9	force, bear, rod, plate, member	1:pistol, handgun, side arm:0.1020 2:unit, social unit:0.0980 3:instrumentation:0.0980	1:unit, social unit:1.250 2:causal agent, cause,:0.625 3:organization:0.500
10	transistor, layer, channel, amorphous, effect	1:artifact, artefact:0.1390 2:structure, body structure:0.1225 **3:semiconductor:0.1029**	1:anatomical structure:0.500 2:artifact, artefact:0.040 3:validity, validness:0.000

第二節 路透社文件集

一、 類別特徵詞選取方法

路透社前十大類新聞共有 6018 篇文件。由於文件集特性是短文件、類別中文件數多，因此採用基本的 CC 方法取出類別特徵詞。由表 6 可以看到相當巧合的現象，有超過十個以上的機器選詞與人工給的標題詞相同，證實了 CC 的成效。與 Ron Bekkerman [2]的結果相比，CC 的成效要比 MI 產生的標題詞更好(見表 5)。MI 選出的類別特徵詞對於機器分類來說是具有代表性的關鍵特徵，然而對人來說卻難以解讀其意涵。

表 5：由 MI 選出的類別特徵詞[3]

Category	1st word	2nd word	3rd word	BEP
earn	vs+	cts+	loss+	93.5%
acq	shares+	vs-	inc+	76.3%
money-fx	dollar+	vs-	exchange+	53.8%
grain	wheat+	tonnes+	grain+	77.8%
crude	oil+	bpd+	OPEC+	73.2%
trade	trade+	vs-	cts-	67.1%
interest	rates+	rate+	vs-	57.0%
ship	ships+	vs-	strike+	64.1%
wheat	wheat+	tonnes+	WHEAT+	87.8%
corn	corn+	tonnes+	vs-	70.3%

二、 產生概念性標題及成效評估

由表 6 可以看出不論是 WordNet 或是 InfoMap，有 70%以上的概念性標題都被評定為合理的標題詞。由 Ido Dagan [4]的研究可知路透社的類別原本就是階層式分類。例如 grain, wheat, 以及 corn 可以被合併為大類：foodstuf。由表 6 中也可看出我們產生的概念詞正好描述這樣的階層關係：第四類、第九類以及第十類選出的上位詞都有 foodstuf。

表 6：路透社文件集概念性標題詞評估結果

ID	Category	Descriptors	WordNet	InfoMap
1	Earn	NET, QTR, Shr, cts Net, Revs	1:goal:0.2744 2:trap:0.2389 3:**income**:0.2389	1:trap:1.000 2:game equipment:1.000 3:fabric, cloth, textile:1.000
2	Acq	**acquire**, **acquisition**, stake, company, share	1:device:0.1514 2:stock certificate, stock:0.1386 3:wedge:0.1386	1:asset:0.500
3	Money-fx	currency, **money market**, central banks, TheBank, yen	1:backlog, stockpile:0.0850 2:airplane maneuver:0.0850 3:marketplace, mart:0.0770	1:medium of exchange, **monetary system**:0.750
4	Grain	wheat, **grain**, tonnes, agriculture, Corn	1:seed:0.1848 2:cereal, cereal grass:0.1848 3:**foodstuff**, food product:0.182	1:weight unit:1.250 2:grain, **food grain**:1.250 3:cereal, cereal grass:1.250
5	crude	**crude oil**, bpd, OPEC, mln barrels, petroleum	1:lipid, lipide, lipoid:0.2150 2:oil:0.1646 3:**fossil fuel**:0.1211	1:oil:0.750 2:**fossil fuel**:0.750 3:lipid, lipide, lipoid:0.500
6	Trade	**trade**, tariffs, Trading surplus, deficit, GATT	1:**business**:0.0823 2:UN agency:0.0823 3:prevailing wind:0.0823	1:liability, financial obligation, indebtedness, pecuniary obligation:0.062
7	interest	rate, money market, BANK, prime, discount	1:charge:0.1372 2:airplane maneuver:0.0850 3:allowance, adjustment:0.0850	1:charge:0.111
8	Ship	**ship**, vessels, port, Gulf, TANKERS	1:craft:0.2469 2:instrumentation:0.1959 3:**vessel**,	1:craft:0.812 2:physical object:0.456 3:**vessel**,

			watercraft:0.1848	watercraft:0.361
9	wheat	**wheat**, tonnes, grain agriculture, USDA	1:weight unit:0.1792 2:**foodstuff**, food product:0.151 3:executive department:0.1386	1:weight unit:1.250 2:**foodstuff**, food product:0.500
10	Corn	**corn**, maize, tonnes, soybean, grain	1:seed:0.1848 2:cereal, cereal grass:0.1848 3:**foodstuff**, food product:0.182	1:cereal, cereal grass:1.250 2:weight unit:1.250 3:**foodstuff**, food product:0.790

第三節 FJU-CTC 文件集

一、類別特徵詞選取方法

與路透社文件集特性相同，FJU-CTC 是短文件、類別中文件數多，因此採用基本的 CC 方法取出類別特徵詞。由表 7 的第三欄可以看到選出的辭彙，粗體字代表被評定為適當的詞。由表中可以看出選出的辭彙有許多是三字以上的長詞，並且都是基於時間與空間的因素下，特有的辭彙。自動選詞已經可達到近八成的解釋度，因此可以看出本研究的斷詞以及取詞方法應用在領域特殊的中文文件一樣具有穩定的成效。

表 7：FJU-CTC 文件集標題詞與概念性標題詞評估結果

CID	Category	Descriptors	Sinica BOW
P52	文化大革命： 1 最重要 2.次要或一般性 3.各地情況(分區：A.華北、B.東北、C.西北、D.華東、E.中南區、F.西南區)	**資產階級幫派體系：0.0920**	None
		冤案：0.0797	None
		清查工作：0.0791	None
		反革命政治綱領：0.0766	None
		權陰謀活動：0.0758	None
P10	軍事： 1.軍事綜合資料 2.兵役法、兵役制度、徵兵、民兵 3.蘇聯(反對原子戰爭)各國 4.各兵種 5.志願軍 6.抗美援朝 7.韓戰始末 8.對台軍事 9.擁政愛民 10.優撫工作(安置復員軍人) 11.軍中文化 12.軍事幹部訓練 13.軍人生產	**部隊：0.4718**	組織, 群體
		指戰員：0.3351	None
		幹部戰士：0.3007	None
		我軍：0.2931	None
		連隊：0.2625	None
Cu4	1.青年團的工作 2.少先隊 3.兒童 4.兒童讀物(影劇) 5.青聯 6.青年思想生活問題	**共青團：0.4428**	None
		團員：0.3365	成員
		團組織：0.3294	None
		名知識青年：0.2944	None
		家長：0.2939	社會角色
E34	1.農村行政 2.農民生活	**公社：0.2582**	None
		生產隊：0.2126	None
		社員：0.1906	成員
		貧下中農代表大會：0.1751	None
		農村：0.1708	None

E3	農業 1.綜合資料 2.地方資料 3.稻米與小麥 4.肥料 6.農業技術 7.雜糧(玉米等)、農業技術	**作物: 0.4137**	None
		糧食: 0.3503	食物
		農業生產: 0.3090	None
		省農業: 0.3040	None
		早稻: 0.2954	None
P5	黨 1.黨代表大會、會議、中全會、對黨的評論 2.中央要文 3.地方代表大會 4.分局、省市直屬機關 5.縣、區、支部 6.建黨及新黨員入黨 7.黨工礦企業工作 8.黨農村工作 9.黨員作風(貪污、整風、反黨) 10.黨史 11.外國共產黨 12.黨中央機關及人員	**黨章: 0.2759**	None
		新黨: 0.2661	None
		黨代表大會: 0.2453	None
		廣大黨員: 0.2431	None
		中國共產黨: 0.2200	None
E6	勞動 1.綜合資料 2.工會工作 3.生產(競賽、先進) 4.教育(文娛、體育) 5.工資 6.工人福利 7.勞動保險 8.勞動保護 9.失業就業	**職工: 0.2533**	None
		工人階級: 0.2249	None
		省總工會: 0.2135	None
		車間: 0.2130	None
		工會組織: 0.2102	None
P4	1.公安 2.戶口(外僑) 3.對反革命份子方法、報告 4.反革命份子 5.游民、勞改、娼妓	**民兵工作: 0.3646**	None
		民兵建設: 0.3637	None
		社會治安: 0.3188	None
		公安機關: 0.3090	None
		好民兵: 0.2978	None

	6.罪犯 7.交通、槍械、無線電...則例 8.街道工作 9.民兵		
E5	工業 1.綜合資料 2.鋼鐵工業 3.機械工業 4.水泥 5.化學工業 6.其他工業	企業: 0.3014	公司／法人, 意向性歷程, 金融交易
		產品: 0.2826	物體,主觀評價屬性,結果
		省工業學大慶: 0.2817	None
		省工交戰線: 0.2805	None
		工業生產: 0.2723	None
E36	人民公社 1.綜合資料 2.地方資料 3.人民公社(鄉村)	社員: 0.3049	成員
		生產隊: 0.2788	None
		生產責任制: 0.2687	None
		自留地: 0.2356	None
		農業: 0.2270	維持

二、產生概念性標題及成效

將取出的辭彙以人工方式透過 Sinica Bow 查詢，結果很明顯的幾乎大部分都沒有相對應的資料。

我們可以歸納原因如下：

1. 基於文件集本身時空環境的特殊性，文件用詞屬於「文化大革命」時期大陸的用詞，因此，許多概念在一般性語料庫中查不到是很正常的結果。
2. 中文長詞原本就容易因為字串不匹配而造成查詢的低召回率。

由 FJU-CTC 實驗結果可推斷，當其他變因都在控制的情況

下，成較不佳是基於語料庫領域不匹配。因此可以看出整個流程的六大變因中，語料庫的選擇雖然與技術方法無關，卻是直接影響成效的關鍵。

注釋

[1] Yuen-Hsien Tseng, Dai-Wei Juang, Yeong-Ming Wang, and Chi-Jen Lin, "Text Mining for Patent Map Analysis," Proceedings of IACIS Pacific 2005 Conference, May 19-21, 2005, Taipei, Taiwan, pp.1109-1116.

[2] Ron Bekkerman, Ran El-Yaniv, Yoad Winter, Naftali Tishby, “On Feature Distributional Clustering for Text Categorization,” Proceedings of the 24th ACM-SIGIR Conference, 2001, pp.146-153.

[3] 同註 2

[4] Ido Dagan and Ronen Feldman, "Keyword-based browsing and analysis of large document sets," Proceedings of the Symposium on Document Analysis and Information Retrieval (SDAIR-96), Las Vegas, Nevada, 1996.

第五章 結論與後續研究方向建議

第一節 結論

對歸類結果給訂類別標籤是協助人工解讀歸類結果的重要工作。為了產生概念性標題詞，本研究提出了一種以階層式語料庫為基礎的上位詞演算法。因為這個演算法僅考慮了上位詞在階層架構中的深度與出現次數，因此可以很容易應用在其他階層式的語料庫。

這個方法應用在專利文件集得到 50%的成效。然而這是源於 WordNet 本身的限制所達到的成績。WordNet 並未涵蓋所有由專利文件萃取出來的辭彙。同時 WordNet 的上位詞結構也未能反應專利分析需要的領域知識。一方面可以說，假如我們欲分析的文件與 WordNet 涵蓋的領域相同，那麼成效就會提升；另一方面來說，我們可以預期假如有更適合的語料庫供我們使用，那麼我們提出的上位詞搜尋方法就能產生更好的標題詞，Reuters 以及 FJU-CTC 的實驗結果分別以正面及反面的數據支持這個結論。我們也嘗試過將專利文件集萃取出的辭彙透過 Google 目錄查詢，然而得到的結果並不符合分析專利的需求。截至目前尚未能找到更佳的詞庫改進專利文件集的概念性標題擷取成效。

然而即使只做到現在的程度，這個方法與自文件叢集中選詞一樣，對於詮釋歸類結果依然相當有幫助。

舉例來說，圖 11 顯示的是分析 612 篇專利文件後所得的主題地圖。圖中每個圈代表一個文件叢集，圓圈的大小代表叢集的文件數量多寡，而圓圈中的數字代表主題編號。我們採用 MDS 方法[1]將叢集主題對應到二維圖形，主題間的距離代

表主題間關聯度的強弱，但是主題間的相對位置並沒有意義。主題標題語類別 ID 細節在表 8 可以看到。這個知識地圖非常清楚的呈現了國科會擁有的專利技術部分部情形，標題詞更協助人工可以快速解讀這 612 篇深度技術性文件的分群是基於什麼樣的原則。

總體而論，歸類技術、標題詞擷取技術以及概念性標題對於協助人工作主題分析來說可以達成相當程度的幫助，是值得更加深入研究的技術。

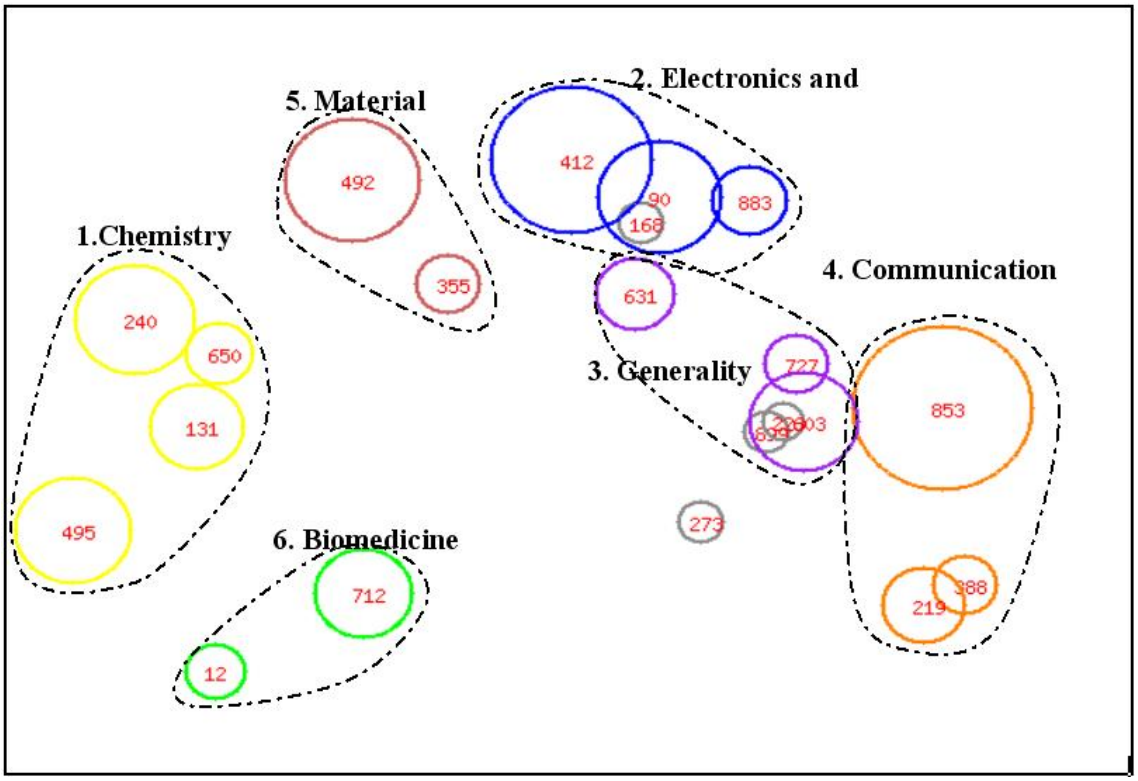

圖 11：NSC 專利文件歸類結果視覺化

表 8：NSC 專利文件最後歸類結果

1: 122 docs. : 0.2013 (acid:174.2, polymer:166.8, catalyst:155.5, ether:142.0, formula:135.9) * 108 docs. : 0.4203 (polymer:226.9, acid:135.7, alkyl:125.2, ether:115.2, formula:110.7) o 69 docs. : 0.5116 (resin:221.0, polymer:177.0, epoxy:175.3, epoxy resin:162.9, acid: 96.7) + ID=131 :26 docs.:0.2211(polymer: 86.1, polyimide: 81.1, aromatic: 45.9, bis: 45.1, ether …) + ID=240 : 43 docs.:0.1896(resin:329.8, acid: 69.9, group: 57.5, polymer: 55.8, monomer: 44.0) o ID=495:39 docs.:0.1385(compound: 38.1, alkyl: 37.5, agent: 36.9, derivative: 33.6, formula …) * ID=650 : 14 docs. : 0.1230(catalyst: 88.3, sulfide: 53.6, iron: 21.2, magnesium: 13.7, selective: 13.1)
2: 140 docs. : 0.4068 (silicon:521.4, layer:452.1, transistor:301.2, gate:250.1, substrate:248.5) * 123 docs. : 0.5970 (silicon:402.8, layer:343.4, transistor:224.6, gate:194.8, schottky:186.0) o ID=412 : 77 docs. : 0.1503(layer:327.6, silicon:271.5, substrate:178.8, oxide:164.5, gate:153.1) o ID=90 : 46 docs. : 0.2556(layer:147.1, schottky:125.7, barrier: 89.6, heterojunction: 89.0, … * ID=883 : 17 docs. : 0.1035(film: 73.1, ferroelectric: 69.3, thin film: 48.5, sensor: 27.0, capacitor …)
3: 66 docs. : 0.2203 (plastic:107.1, mechanism: 83.5, plate: 79.4, rotate: 74.9, force: 73.0) * 54 docs. : 0.3086 (plastic:142.0, rotate:104.7, rod: 91.0, screw: 85.0, roller: 80.8) o ID=631 : 19 docs.:0.1253(electromagnetic: 32.0, inclin: 20.0, fuel: 17.0, molten: 14.8, side: 14.8) o ID=603 : 35 docs. : 0.1275(rotate:100.0, gear: 95.1, bear: 80.0, member: 77.4, shaft: 75.4) * ID=727 : 12 docs. : 0.1155(plasma: 26.6, wave: 22.3, measur: 13.3, pid: 13.0, frequency: 11.8)
4: 126 docs. : 0.4572 (output:438.7, signal:415.5, circuit:357.9, input:336.0, frequency:277.0) * 113 docs. : 0.4886 (signal:314.0, output:286.8, circuit:259.7, input:225.5, frequency:187.9) o ID=853 : 92 docs. : 0.1052(signal:386.8, output:290.8, circuit:249.8, input:224.7, light:209.7) o ID=219 : 21 docs. : 0.1934(finite: 41.3, data: 40.7, architecture: 38.8, comput: 37.9, algorithm: … * ID=388 : 13 docs. : 0.1531(register: 38.9, output: 37.1, logic: 32.2, addres: 28.4, input: 26.2)
5: 64 docs. : 0.3131 (powder:152.3, nickel: 78.7, electrolyte: 74.7, steel: 68.6, composite: 64.7) * ID=355 : 12 docs. : 0.1586(polymeric electrolyte: 41.5, electroconductive: 36.5, battery: 36.1, ... * ID=492 : 52 docs. : 0.1388(powder:233.3, ceramic:137.8, sinter: 98.8, aluminum: 88.7, alloy: 63.2)
6: 40 docs. : 0.2501 (gene:134.9, protein: 77.0, cell: 70.3, acid: 65.1, expression: 60.9) * ID=12 : 11 docs. : 0.3919(vessel: 30.0, blood: 25.8, platelet: 25.4, dicentrine: 17.6, inhibit: 16.1) * ID=712 : 29 docs. : 0.1163(gene:148.3, dna: 66.5, cell: 65.5, sequence: 65.1, acid: 62.5)

第二節 後續研究

在中文的部分，由實驗可以看出由於語料庫領域不匹配，加上文件集本身時間空間的因素，使得文件用詞難以對應到適當的上位詞。本研究曾試圖以中國圖書分類法查詢，然而有兩個主要的問題：

1. 圖書分類的目的與文件主題詮釋並不相同，因此，即使有匹配的詞彙，其上位概念通常依然是以學科領域的角度來描述，並不是以主題的角度描述。
2. 中國圖書分類法涵蓋的領域從總類到哲學、宗教、自然科學…等，一開始制定的目的就是希望可以涵蓋人類所有知識範疇，因此階層架構本身就廣而不深。

因此實驗的結果並不理想。

在英文的部分，如前所述嘗試用 Google Directory，然而依然因為階層架構制定時的需求不同，成效也不佳。

本研究受限於時間與人力等因素，僅以三個文件集以及三個語料庫做實驗，未來可以再擴展實驗範圍以驗證成效或提出更佳的解決方案。

注釋

[1] Joseph B. Kruskal, "Multidimensional Scaling and Other Methods for Discovering Structure," pp. 296-339 in "Statistical Methods for Digital Computers" edited by Kurt Enslein, Anthony Ralston, and Herbert S. Wilf, Wiley: New York, 1977.

參考書目

一、 中文部分

(一) 圖書

林傑斌、劉明德、陳湘，「資料採掘與 OLAP 理論與實務」，台北：文魁書局，2002。

(二) 期刊

陳光華，"資訊檢索查詢之自然語言處理"，中國圖書館學會會報，第 57 期， 85 年 12 月，頁 141 - 153 。

曾元顯，"分類不一致對文件自動分類效果的影響"，大學圖書館，9 卷 1 期，2005 年 3 月，頁 11-13

簡立峰，"尋易系統（Csmart）與中文智慧型資訊檢索"，資訊傳播與圖書館學，3 卷 2 期， 85 年 12 月，頁 28-37。

二、 英文部分

(一) 圖書

Wordnet: An Electronic Lexical Database , pp. xviii-xix

(二) 期刊

Aggrawal, C. C., & Yu, P. S., "Finding Generalized Projected Clusters in High Dimensional Spaces," Proceedings of the 2000 ACM SIGMOD international conference on Management of data, Dallas, New York: ACM Press, 2000, 70-81.

Cutting, D. R. Karger, J. O. Pedersen, and J. W. Tukey. “Scatter/gather: A cluster-based approach to browsing large document collections,” Proceedings of the 15th ACM-SIGIR Conference, 1992, pp. 318-329.

Dubes, R. C. & Jain, A. K., "Algorithms for Clustering Data," New Jersey: Prentice Hall, 1988.

Frakes, W. B. & Baezay, R., "Information Retrieval: Data Structures and Algorithms," New Jersey: Prentice-Hall, 1992.

Franca Debole and Fabrizio Sebastiani, “An Analysis of the Relative Hardness of Reuters-21578 Subsets” to appear in Journal of the American Society for Information Science and Technology.

Griffith, A., Luckhurst, H. C. & Willet, P., "Using Inter-Document Similarity Information in Document Retrieval Systems," Journal of the American Society for Information Sciences, Vol. 37, No. 1, 1986, pp. 3-11.

Ido Dagan and Ronen Feldman, "Keyword-based browsing and analysis of large document ets," Proceedings of the Symposium on Document Analysis and Information Retrieval (SDAIR-96), Las Vegas, Nevada, 1996.

Joseph B. Kruskal, "Multidimensional Scaling and Other Methods for Discovering Structure," pp. 296-339 in "Statistical Methods for Digital Computers" edited by Kurt Enslein, Anthony Ralston, and Herbert S. Wilf, Wiley: New York, 1977.

Krista Lagus, Samuel Kaski, and Teuvo Kohonen, “Mining Massive Document Collections by the WEBSOM Method,” Information Sciences, Vol 163/1-3, pp. 135-156, 2004.

Lee-Feng Chien, "PAT-Tree Based Keyword Extraction for Chinese Information Retrieval" CM SIGIR 1997.

Liu, T., Liu, S. & Chen, Z., 2003, "An Evaluation on Feature Selection for Text lustering," Proceedings of the Twentieth International Conference on Machine Learning, Washington, CA: AAAI Press, pp. 488-495.

Marti A. Hearst and Jan O. Pedersen, "Reexamining the Cluster Hypothesis: Scatter/ Gather n Retrieval Results," Proceedings of the 19th ACM-SIGIR Conference, 1996, pp. 76-84.

Mehran Sahami, Salim Yusufali, and Michelle Q. W. Baldonaldo, “SONIA: A Service for Organizing Networked Information Autonomously,” Proceedings of the 3rd ACM Conference on Digital Libraries, 1998, pp. 200-209.

Michele Banko, Vibhu O. Mittal, and Michael J. Witbrock, “Headline Generation Based on Statistical Translation,” ACL 2000.

Oren Zamir and Oren Etzioni, “Web document clustering: a feasibility demonstration,” Proceedings of the 21st ACM-SIGIR Conference, 1998, pp. 46-54.

Paul E. Kennedy, Alexander G. Hauptmann, "Automatic title generation for EM," Proceedings of the 5th ACM Conference on Digital Libraries, 2000, pp.

Ron Bekkerman, Ran El-Yaniv, Yoad Winter, Naftali Tishby, “On Feature Distributional Clustering for Text Categorization,” Proceedings of the 24th ACM-SIGIR Conference, 2001, pp.146-153.

Russell Swan and James Allan, "Automatic Generation of Overview Timelines," Proceedings of the 23rd ACM-SIGIR Conference, 2000, pp. 49-56.

Salton, G. & Buckley, C., "Term Weighting Approaches in Automatic Information Retrieval," Journal of Information Proceeding and Management, Vol. 24, No. 5, 1988, pp. 513-524.

Salton, G., "Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer," New York: Addison-Wesley, 1989.

Steinbach, M., Karypis, G. & Kumar, V., "A Comparison of Document Clustering Techniques," Technical Report 00-034, Computer Science and Engineering, University of Minnesota, 2000.

Wei,C.P.,Hu P.J.&Dong,Y.X.,"Managing Document Categories in E-Commerce Environments:an Evolution-Based Approach,,"European Journal of Information Systems,Vol.11,No.3,2002,pp.208-222

William B. Frakes and Ricardo Baeza-Yates, Information Retrieval: Data Structure and Algorithms, Prentice Hall, 1992.

(三) 網路資源

Antti Arppe, "Term Extraction from Unrestricted Text,",1995 <http://www.lingsoft.fi/doc/nptool/term-extraction.html>

David D. Lewis, “Reuters-21578 text categorization test collection, Distribution 1.0” README file (v 1.2), 1997 <http://www.research.att.com/~lewis/>

Document Understanding Conferences<http://www-nlpir.nist.gov/projects/duc/.>

Jean Godby, "Two Techniques for the Identification of Phrases in Full Text," <http://www.oclc.org/oclc/research/publications/review94/part1/twotech.htm>

Jen-Nan Chen, Jyun-Sheng, Chang and Huey-Chyun Chen, "Using Word Segmentation Model for Compression of Chinese Text" <http:// nlplab.cs.nhtu.edu.tw/~mathis/own/html/PAPER/JNL/95/cpcol/CPCOL95.htm>

Mathis H. C. Chen, Tsong-Yi Tseng, Jason J. S. Chang, "Automatic Generation of Indices or Chinese Books," <http://nlplab.cs.nthu.edu.tw/~mathis/own/html/PAPER/JNL/96/cpcol/BookIdx.htm>