探勘中文新聞文件

許中川　　陳景揆
雲林科技大學資訊管理系　雲林科技大學資訊管理系

摘　要

新聞報導每天發生的重要事件，大量的新聞文件中，往往蘊含重要的資訊。文件資料探勘技術用來發覺隱藏在大量文件中的特徵。然而，目前的文件探勘研究集中在歐美語系文件，且代表文件的關鍵詞彙的擷取，都是人工處理。本研究以中文新聞文件為探勘對象，試圖發覺其中隱含的知識。針對新聞文件的特殊結構，在收集關鍵詞彙方面，以混合式斷詞法進行中文斷詞，經過關鍵既有詞彙擷取與關鍵新生詞彙擷取步驟，獲得每篇新聞文件的關鍵詞彙，代表該文件重要概念，供後續探勘之用。在資料探勘方面，首先為切合新聞文件知識開採需求，使用概念階層樹建構背景知識與關鍵詞彙。然後以關聯法則為基礎，我們提出三個改良式關聯模式：第一個是新生詞彙關聯法則，第二個是結構化資料與高頻詞彙關聯，第三個是結構化資料與某同類詞彙關聯；另外，以線性迴歸及卡方分配技術，分別探勘關鍵詞彙的報導趨勢與分佈情況。最後並以實驗驗證此探勘架構的可行性。

關鍵詞：文件資料探勘、知識發覺、關鍵詞彙擷取、關聯法則、趨勢分析

Data Mining in Chinese News Articles

Chung-Chian Hsu Jing-Kuei Chen
Department of Information Management
National Yunlin University of Science and Technology

Abstract

News reports important daily events. Implicit information hides in huge collection of news articles. Text data mining technology aims at discovering knowledge hidden in large collection of texts. However, current reported research focus on English texts and keywords are given manually. This paper studied text data mining in Chinese news articles. Utilizing the special structure of news articles, existing keywords and new keywords, representing the content of a news article, are automatically extracted using hybrid segmentation technique. Then, the mining process guided by domain knowledge proceeds. We proposed three types of extended association rules: new keywords association rules, association rules of structured data and high frequency keywords, and association rules of structured data and homogeneous keywords. Further, linear regression technique and Chi-square test technique are used to analyzing the reporting trend of keywords and the distribution of important concepts. Experiments are conducted to verify the feasibility of the proposed architecture.

Keywords: text data mining, knowledge discovery, keyword extraction, association rules, trend analysis

壹、 緒論

1.1 動機

隨著數位化時代的來臨與網際網路的興起，許多文件由傳統紙上記錄方式，逐漸轉換為電子化文件。例如，電子新聞網站、醫學百科網站及最受矚目的大英百科全書的數位化，並提供網上使用等。在時代的推演下，以數位化呈現文件資料的方式，日益普遍，已成為一種重要的傳播及儲存媒體。

新聞文件具報導功能，蘊含大量資訊。特別是報導當時重要事件，反映社會運作現況。經由仔細分析大量的新聞文件，可以從中探勘出有用的特徵(pattern)或知識。如果長期觀察社會、經濟、政治等各領域的新聞文件，可以從新聞報導中隱含的逐漸變化，進而發現社會現象的消長。

傳統以人工處理及分析文件的方式，耗費人力和時間，僅能處理少量資料。而電腦化的資訊檢索(Information Retrieval)技術，無法有效探勘出隱含在大量文件中的知識。目前普遍使用的檢索系統為布林(Boolean)檢索系統，使用者輸入一個或多個關鍵詞彙，經由交集、聯集與差集的運算，配合模糊比對，尋找與檢索詞彙相關的文件。但只能檢索出個別文件中，具有符合或接近查詢條件的文件，無法進一步歸納及推衍隱含在大量文件中的知識 [Singh 1997]。有別於檢索技術，資料探勘技術可用來挖掘隱含在大量資料中的有用知識[Fayyad 1996b]。

大部份資料探勘技術運用於傳統資料庫中的大量資料。例如，應用於分析大量消費性交易資料，可從資料庫中找出產品間銷售的關聯情形，進一步幫助決策者訂定有利的行銷策略[Agrawal 1993]。傳統資料探勘技術主要針對結構化的表格資料，而忽略了非結構化或半結構化的文件資料[Singh 1997]。相對於關聯式資料庫中，定義明確的表格與欄位，所謂非結構化資料，其內容為一長串的文字或數字，通常無法直接取得關鍵資訊。半結構化資料介於結構化與非結構化資料之間，同時具備結構化資料與非結構化資料。例如，新聞文件就屬於半結構化資料，包含記者名字、報導日期、報導地點等結構化部分與新聞本文的非結構化部分。

目前文件資料探勘研究都針對歐美語系文件，而且代表文件內涵的關鍵詞彙都由人工擷取，轉成結構化資料，儲存在資料庫中，進行後續的探勘處理。人工處理速度慢且成本高。此外，中文文件的詞彙組成方式和歐美語系文件不同，需要特別的前置處理程序。歐美語系的文件以空白作為詞與詞之間的區隔，然而中文詞彙在詞與詞之間沒有明顯的斷詞點，電腦無法直接判斷出有意義的詞彙。因此使用中文文件進行資料探勘時，需要特別的處理程序，以獲得具有完整概念的中文詞彙。

1.2 目的

本研究主要目的，探討如何將資料探勘技術，應用於探勘蘊藏在大量中文新聞文件中的特徵或知識。依據主要目的，分別針對以下各項加以深入探討：

1. 中文新聞文件的探勘架構：提出一個適用於中文新聞文件的探勘程序，包括前置處理、自動化關鍵詞彙擷取及探勘模式(model)等相關問題。
2. 如何擷取中文新聞文件中的關鍵詞彙及新生詞彙：中文斷詞為中文文件處理的基本步驟，雖然已有一些研究成果[陳克健 1986][Sproat 1990]，然而有別於一般性的中文文件，新聞文件具有特殊的撰寫風格。我們探討如何利用此特性提高中文斷詞、關鍵詞彙及新生詞彙的擷取效率。
3. 傳統資料庫探勘模式如何應用於文件資料探勘：針對資料庫資料的探勘技術已經應用在一些商業用途[Brachman 1996][Fayyad 1996c]。但相較於資料庫探勘，文件資料具有其特殊的意義及結構，因此如何將探勘模式應用到文件資料上，

值得加以探討。

4. 如何由新聞文件探勘社會趨勢：新聞報導反映社會現象，越熱門的事物，被報導的機率越大；消逝中的事物，出現在新聞中機率自然逐漸降低。因此，社會現象的消長可以藉由長期觀察新聞報導中得到線索。

貳、 文獻探討

2.1 資料庫探勘

隨著時間的累積，多樣且大量的資料中，雖然蘊藏寶貴的知識，資料量卻造成人工分析上的困難。面對大量的資料，資料庫知識挖掘(Knowledge Discovery in Database, KDD)研究即針對此類問題，以電腦化的探勘流程，企圖從大量資料中找出令人感興趣的特徵，用以表示隱含的知識 [Brachman 1996][Fayyad 1996a, 1996b, 1996c]。

Fayyad 提出一個一般化的資料庫知識挖掘流程，包括選擇、轉換、資料探勘及解譯 [Fayyad 1996c]。然而，針對不同的應用領域，因不同的資料型態、領域知識及探勘需求，必須對該通用架構進行特殊化，以便能適用該領域，同時充分利用該領域的特性，提高探勘的效率及效果。通常該特殊化的處理並非一件瑣碎的事，需要投注相當的心力。例如，應用在文件資料的探勘，需要考慮資料是非結構化及半結構化的問題。

2.2 文件資料探勘

文件探勘有別於一般資料庫探勘。一般的資料探勘主要針對存放在資料庫中的結構化資料。資料屬性定義非常明確，可以直接從欄位中擷取資料。相對於資料庫的結構化資料，文件資料是半結構化或是非結構化 [Feldman 1995]。例如，新聞文件是屬於半結構化資料，一篇新聞包含記者名字、報導地點、時間等結構化資料及本文部分的非結構化資料。非結構化部份為一長串敘述性文字，無法直接取得關鍵特徵資料。[Dörre 1999]指出文件探勘具有兩個主要困難點：(1)人工進行多樣且大量的文件特徵選擇，缺乏效率且不敷成本。(2)文件資料的內容維度數量過多，即特徵的屬性不易清楚定義或界定。相較於資料庫探勘，文件探勘流程需要更複雜的步驟，以解決這兩個困難點。另外，需針對文件資料領域，提出探勘的模式。

Feldman 首先提出文件知識發覺(Knowledge Discovery in Texts or KDT)概念[Feldman 1995, 1997b, 1998a]。作者以概念階層(concept hierarchy)表示相關的背景知識，利用背景知識限制及導引後續資料探勘的進行。作者以人工方式給予每篇文件數個詞彙標籤(tag)代表該文件內涵，以利進一步探勘。研究中以概念分佈(concept distributions)方式分析文件集合中各個子概念對於其他子概念的分佈情形。FACT 系統是以查詢為導向的文件探勘系統[Feldman 1996, 1997c]，不同於傳統資料檢索，查詢條件為個別文件必須滿足的條件；FACT 系統的查詢條件是跨文件的文件集合必須滿足的條件。

[Singh 1997]也是以人工方式給予半結構化文件數個詞彙標籤，並以擴充式概念階層(extended concept hierarchy, ECH)建構背景知識。擴充式概念階層擴充了概念與概念之間的兄弟關係，可探勘出四種法則：一般法則(general rules)、父法則(parent rules)、子法則(child rules)及兄弟法則(sibling rules)。[Singh 1999]將非結構化或半結構化的文章對應到結構化的資料表格，一篇文章對應(mapping)為一筆交易(transaction)資料，並運用概念相關性(concept-relatives)從稀疏矩陣得到重要項目集合(large itemset)，這些重要項目集合即為符合使用者特殊需求，並具高度相關性的知識。IBM 的 Intelligent Miner for Text 利用群聚(clustering)及分類(categorization)技術，對文件集合進行群聚處理及分類處理。可以應用到顧客關係管理，處理顧客抱怨電話或郵件[Dorre 1999]。

在文件資料探勘相關的視覺化技術方面，[Feldman 1997a]將文件關鍵詞的關聯以關聯

圖的方式表示，視覺化的呈現文件庫中，重要詞彙的關聯以及關聯強度。[Feldman 1998c] 面對大量文件資料，從主題圖(context graphs)的建構，發展出趨勢圖(trend graphs)，趨勢分析的方法為關聯性標界同時發生的頻率，透過趨勢圖的圖形化介面，可探勘文件中概念的發展趨勢。[Aumann 1999]提出圓圈圖(circle graph)藉以觀察大型文件庫中，概念和概念之間的關係。WEBSOM 利用自組圖(self-organizing maps)神經網路將內容性質相近的文件群聚在一起 [Lagus 1996]。

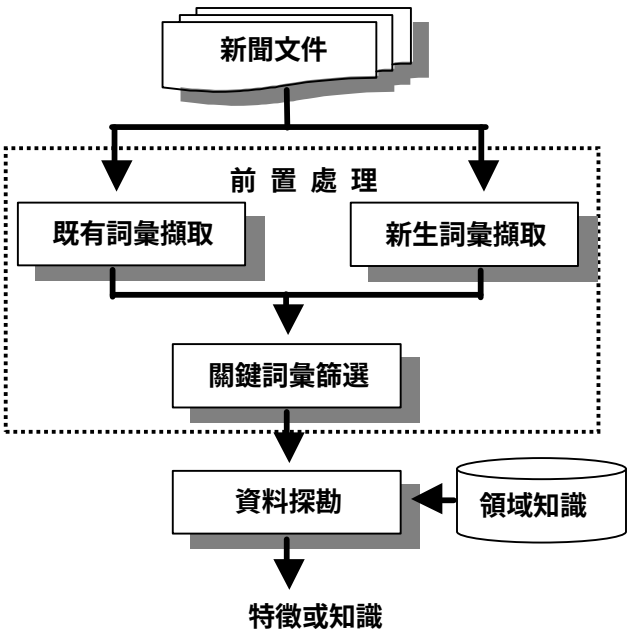
在文件探勘應用方面，Wuthrich 利用文件探勘方式預測股市漲跌。不同於傳統預測方式，該研究經由蒐集各大金融報紙網站上的文件資料，統計文件內與金融有關的關鍵詞彙，並予以權值，再以權值與收盤價的關係，推導股市漲跌與關鍵詞彙間的關聯。利用這些探勘出的關聯，進一步預測每日歐美亞的股市收盤指數 [Wuthrich 1998] [Cho 1999]。Lent 利用一個形狀定義語言 SDL 為基礎[Agrawal 1995]，以查詢導向的方式，探索文件庫內的趨勢。例如，查詢連續五年都大幅增加的專利申請 [Lent 1997]。[Shewhart 1999]利用事先儲存的關鍵詞彙，透過比對及計算，監測新聞文件中的熱門話題。

從以上文獻可發現文件探勘漸漸受到重視。然而，上述研究都探討歐美語系文件，應用於中文文件的探勘技術則付之闕如。再者，上述文件資料探勘方法大多以人工方式給予各文件相關的概念或關鍵詞彙。總而言之，我們認為目前的文件探勘方法有兩項不足之處。首先，利用人工給予關鍵詞彙不夠客觀，每個人所認定的重要詞彙可能會有所不同，而且花費大量的人力成本。其次，上述研究都是針對歐美語系文件，文件的處理技術和中文不相同。中文句子字字相連，詞與詞之間沒有明顯中斷符號，需要斷詞處理。因此本研究提出適合中文新聞文件探勘的架構，以及探討新聞的中文文件處理技術，以便利進行中文新聞文件資料探勘。

參、中文新聞文件探勘架構

3.1 探勘架構

Fayyad 提出一個一般化的資料庫知識發覺流程，包括選擇、轉換、資料探勘及解譯 [Fayyad 1996c]。以此流程為基礎，我們針對中文文件的特性，提出一個中文文件資料探勘架構(如圖一)。



圖一 中文新聞文件探勘架構

此探勘架構包括幾個主要子系統：

1. 既有詞彙擷取：新聞文件經由詞庫式斷詞，得到詞庫中既有的詞彙。
2. 新生詞彙擷取：新生詞彙不在詞庫中，無法直接擷取。本研究以統計式斷詞的字元相鄰頻率的基本概念為基礎，發展新生詞彙的斷詞法則。
3. 關鍵詞彙擷取：並非所有詞彙都是重要的詞彙，經由此步驟篩選既有詞彙與新生詞彙中的重要詞彙。
4. 領域知識庫：存放領域知識，引導資料探勘，以提昇探勘效率及效果。
5. 資料探勘：設定探勘模式，配合領域知識對擷取出的關鍵詞彙，進行資料探勘。本研究探討的探勘模式包括新生詞彙關聯、關鍵詞彙報導趨勢及詞彙分佈差異。

3.2 前置處理

前置處理主要目的是，從半結構化的新聞文件中，擷取出關鍵詞彙，代表文件的內涵，以利後續的資料探勘。主要的處理包括中文斷詞、擷取既有詞彙、擷取新生詞彙及篩選關鍵詞彙。

3.2.1 中文斷詞

電腦化文件分析的第一步驟為斷詞。中文文件斷詞處理有別於歐美語系文件的斷詞程序。歐美語系文件在詞(word)與詞間以空白隔開，只需以空白為中斷點，即可斷出獨立的詞彙。但中文句子字字相連，有意義的詞與詞間並無明顯區隔，因此探勘程序需要包含中文斷詞，將中文新聞文件斷成個別的詞目。

中文文件斷詞大致上分為三種：詞庫式斷詞法[Chen 1992]、統計式斷詞法[Fan 1988、Sproat 1990]及混合式斷詞法[Nie 1996]：

- 詞庫式斷詞法
  為目前普遍使用的斷詞方式，其演算法相當直覺且實作容易。基本上將文件和詞庫中收集的詞彙比對，進行斷詞。斷詞的品質和詞庫的詞彙多寡有關，必須時常對詞庫的內容加以更新。另外，有學者將詞庫斷詞法，輔以一些詞性的結構，發展出規則式斷詞法 [陳克健 1986]，提昇斷詞的品質。
- 統計式斷詞法
  統計式斷詞法乃參考一大型語料庫(corpus)的詞彙統計資訊，以鄰近字元同時出現頻率高低作為斷詞的依據。由於語料庫和應用領域有關，不同語料庫間的統計資訊不適合互用[Nie 1996]。再者，因處理的時間複雜度，統計式斷詞常受限於一階馬可夫模式(first-order Markov models) [Li 1991]，大多只針對二字詞進行處理，進一步擴充此模式會提高演算法的時間複雜度[Nie 1996]。所以兩字以上的詞彙如「大賣場」、「小額投信」等就無法斷出。
- 混合式斷詞法
  將詞庫斷詞法及統計斷詞法整合。[Nie 1996]利用詞庫斷出不同組合的詞彙，然後利用詞彙的統計資訊，找出最佳的斷詞組合。此法仍需要大型的語料庫提供統計資訊。

本研究結合詞庫式及統計式斷詞的優點，進行既有詞彙擷取及新生詞彙擷取。先採用詞庫式斷詞擷取既有詞彙，剩下的文字，再利用統計式斷詞進行新生詞彙擷取。

3.2.2 既有詞彙擷取

詞庫式斷詞擷取既有詞彙。詞庫式斷詞作法為，對照詞庫內收集的詞目，以比對句子中可能隱含的詞目，找出可能的中斷點，以中斷點斷出個別詞目。這些詞目皆為原來詞庫所收集的詞彙，其演算法相當直覺，為目前普遍使用的方法。

3.2.3 新生詞彙擷取

隨著潮流不斷演進，新生詞彙陸續被創造。因此，新生詞彙很有可能代表一種新興社會現象或新發明事物，在資料探勘中扮演重要角色。詞庫的更新速度，通常無法趕上新生詞彙的創造。所以，文件中不在詞庫中的詞彙，可以視為新生詞彙。

新生詞彙的擷取，我們使用統計式斷詞法的基本概念，配合新聞文件結構的特性擷取新生詞彙。觀察中文新聞文件可以得知，報導性的新聞寫作通常會在第一段作整篇新聞的概要性描述，第二段之後才作更詳盡的報導。事實上，許多報導性或報告性的文件，都有類似的文件格式。例如，調查報告、研究報告及學術論文等。我們稱此撰寫風格為報告性文件撰寫特性。因此，新聞文件的重要詞彙會在概要性描述的第一段中出現，同時也會重覆出現在後續的段落。利用此特性，我們可以新聞文件的第一段為基礎，擷取至少在後續段落重覆出現 N 次的連續字彙，作為新生詞彙。N 為可設定的系統參數。

本方法不同於傳統統計式斷詞法是不需要使用大量的語料庫，而是利用報告性文件撰寫特性，提供統計資訊。此外，詞庫式斷詞已將已知詞彙斷出，從剩下的文字中擷取新生詞彙，大幅減少需要統計式斷詞的詞彙。

3.3 關鍵詞彙篩選

關鍵詞彙篩選步驟在於取得少數關鍵詞彙，以這些關鍵詞彙表達文件內涵。文件經過斷詞處理後雖然可以鑑別出各個詞彙，但各個詞彙在新聞文件中重要性不同。例如，在犯罪新聞中的「槍枝」就比「終於」等一般性詞彙重要，因此有必要從文件中過濾掉一般性的詞彙，留下重要的關鍵詞彙以代表該文件的關鍵資訊。這些關鍵資訊代表文件中一個個的重要概念，可供往後探勘使用。

經過之前的中文斷詞步驟，可以得到既有詞彙及新生詞彙，本節分別以四個步驟過濾這些詞彙，擷取關鍵既有詞彙及關鍵新生詞彙，代表關鍵詞彙。

3.3.1 關鍵既有詞彙

新聞文件通常包含類似的內容結構。在新聞報導中，新聞文件通常具有下列的結構——「標題」、「日期」、「類別」、「內容」及「記者ｘｘｘ報導」等結構化資訊，接著為新聞的本文。如果將一篇新聞分成具有結構性的前半部份及不具結構性的本文部份，新聞文件可視為半結構化資料。針對此特性，關鍵既有詞彙擷取可分為結構化部分與非結構化部分。由於結構化資料擷取相當直覺且明確，在此不予贅述，以下則說明本研究在非結構化部分的截取程序。

在非結構化本文的關鍵既有詞彙擷取上，本研究針對新聞文件的特性，提出四個主要過濾方法：

1. 剔除單一字元的詞目：從斷詞結果觀察可以發現，單一字元的詞目通常無法表達一個完整的概念，很少具備成為關鍵詞彙的特質。例如：到、及、與、時、酒、水、肉….等。雖然酒、水、肉等名詞表達完整概念，但大部份還是以多字元詞目情況下，成為文件中的關鍵詞彙。例如葡萄酒、自來水、豬肉。
2. 擷取名詞與動詞：經觀察發現重要關鍵詞彙詞性主要有名詞與動詞兩種，或由這兩種詞性複合而成的複合詞彙。因此第二步驟經由剔除掉一些不重要的詞性類別，可以提升關鍵詞彙擷取的品質。
3. 首段詞彙及頻率規則：依據新聞撰寫特性通常重要的詞彙會出現在概要性描述的第一段，然後再重複地出現在其他段落。在這個步驟只留下出現在第一段的詞彙，而且出現頻率超過門檻值 N，也就是在第一段以後至少出現 N-1 次。
4. 過濾一般性詞彙：此步驟為過濾一般性詞彙留下關鍵詞彙。在資料檢索領域，最常使用的方法是逆向文件頻率(Inverse Document Frequency, IDF)，反映詞彙在文件集合中的分佈情形[Spark Jones 1972]：

$$IDF(w)=\log_2(n)-\log_2(O(w))+1$$

其中 n 是文件集合的文件總數，O(w)是包含詞彙 w 的文件總數。當 w 出現在一半以上的文件，則其 IDF 小於等於 0，我們可以認為這個詞彙出現在大部分文件中，因而對於文件集合中的文件較不具有鑑別性。例如有一文件集內含 1000 個文件數，有兩詞 A 和 B 在文件集都分別出現了 2000 次，詞頻皆為 2000，無法由此區分兩詞彙的重要性。以文件數的角度來看，若 A 與 B 分別出現在 100 篇及 900 篇文件之中，則 A 及 B 兩詞的逆向文件頻率分別為 4.3

及 1.1，由此可發覺 A 的逆向文件頻率 4.3 明顯高於 B 的逆向文件頻率 1.1，表示 A 比 B 更具區別文件的能力，是為較關鍵的詞彙。

3.3.2 關鍵新生詞彙

在新生詞彙擷取步驟中，以統計式斷詞斷出不在詞庫中的重複性詞目組合。但有些為無意義的詞彙，必須刪除。另外，有意義的詞彙不一定是關鍵性詞彙，例如「用來」、「但是」等一般性詞彙。根據新聞撰寫特性及實驗，本研究歸納出以下四個過濾規則，篩選關鍵新生詞彙。

1. 去除含功能性詞性字詞：根據實驗觀察，發現重複性字詞中，如果包含某些詞性的字時，大都無法成為關鍵性詞彙，可以剔除。例如：含有「喔」、「嗎」。表一為可考慮剔除的詞性類別。
2. 首段詞彙及頻率規則：如同篩選關鍵既有詞彙，關鍵的新生詞彙必須出現在首段及其他段落，至少共 N 次以上。
3. 出現持續性：為確認新詞彙代表的是一個新的社會現象或新事物，而不是一個突發個案事件所產生的臨時性詞彙，新生詞彙必須具有出現持續性。我們以 x 天為一個觀察區間，將文件庫的新聞文件集合依據報導的時間，切割成連續的區間，詞彙必須出現在至少 y 個觀察區間，並且於觀察區間中出現至少 z 次，表示新生詞彙出現的持續性。x, y 及 z 都是系統參數。
4. 領域相關經驗法則：從實驗觀察得知，關鍵新生詞彙的形成方式，和應用領域有關。很難全部用通用的規則過濾，需要輔以領域相關的經驗法則。例如在社會新聞方面，常會斷出「長張溫鷹」、「長廖正豪」及「長李大維」等之類的重複性字詞。針對此類的重複字詞，可以考慮利用領域相關的停字清單(stop list)方式過濾。

表一　可剔除的詞性類別

詞性	說明
感嘆詞(I)	一般出現句子前，例如：喔，瞭解
語助詞(T)	幾乎都出現在句尾，如：你來嗎？
連接詞(C)	包括 1.對等連接詞，如：張三和李四，2.列舉連接詞，如：身分證、戶口名簿等證件，3.句尾關聯連接詞，如：你不來的話，我也不來，4.關聯連接詞，如：雖然他聰明，但是不用功
副詞(D)	在動詞與主詞之間，如：他可能回去了
介詞(P)	一種帶論元的功能詞，如：他從家裡來
代名詞(Nh)	如：他、她、它….
定詞(Ne)	包括 1.數量定詞，如：兩輛、年約三十，2.特指定詞，如：某名人，3.後置數量定詞，如：五十歲開外，4.指代定詞，如：這代表什麼
後置詞(Ng)	出現在詞組尾的帶論元功能詞，如：三年來、理論上
V_2	此詞性標記只有一個字：有

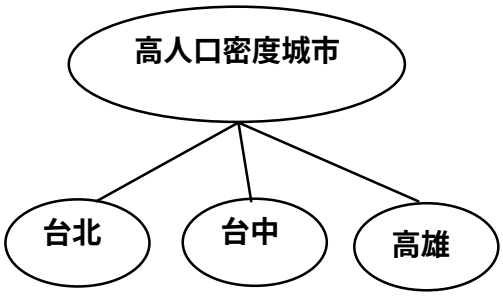
3.4 資料探勘

在取得關鍵資訊之後，配合領域知識庫中的領域知識，進行資料探勘。在本研究，我們提出關聯法則探勘、趨勢分析及分佈差異探勘。

3.4.1 領域知識庫

領域知識可以引導資料探勘，適當的運用領域知識可大幅提升探勘效率與品質。本研究的資

料探勘過程亦加入領域知識的輔助，利用概念階層建構背景知識 [Han 1993][Feldman 1995]。概念階層可視為一種階層樹，越接近根節點存放越高層次的一般化概念，越接近葉節點則為越明確化的概念，如圖二所示。本研究的概念階層結構中，葉節點存放的資訊是新聞文件中的關鍵詞彙，每個葉節點有指標指向包含該節點詞彙的文件，以加速探勘處理。例如：概念「台北」代表所有與「台北」相關聯的文件集合。利用概念階層可以有效地進行一般化的運算，找出更概觀或簡潔的資訊，如「70%的犯罪案件發生在台北及高雄市」，這樣的資訊在導入概念階層後，可得到「犯罪事件大都發生在高人口密度的地方」。



圖二 「高人口密度城市」概念階層樹範例

3.4.2 詞彙關聯探勘

Agrawal 於百貨業的銷售產品分析研究中，提出關聯法則概念 [Agrawal 1993]。關聯法則「A」→「B」表示產品 A 與產品 B 於銷售上的關聯性。本研究以關聯法則觀念為基礎，針對文件特性，加以修改及擴充，進行關鍵詞彙關聯探勘。經由文件前置處理取得新聞文件中的結構化與非結構化資訊，輔以階層結構的概念階層樹，我們提出三種擴充式關聯探勘模式：新生詞彙關聯法則、結構化資料與高頻詞彙關聯及結構化資料與某同類詞彙關聯。

新生詞彙關聯法則

一個新生的詞彙可以視為一個新的特徵，可能代表新事物或新興的現象。再者，一個詞彙的實質意義和該詞彙出現的場合或文章有很大的關聯。當我們不了解一個新生詞彙時，由伴隨該詞彙的一些句子或常一起出現的詞彙，多少可以增進我們對該詞彙的了解。因此，我們以新生詞彙的關聯法則，提供發覺新事物或新興現象的線索。

給予一文件庫及一個詞庫，視不在詞庫中的詞彙為新生詞彙。進行混合式斷詞及擷取關鍵新生詞彙。然後，進行新生詞彙關聯法則探勘，探勘結果為一個新生詞彙關聯法則集合。新生詞彙關聯法則的格式如下：

$$X/S,D \rightarrow <W_1/c_1, W_2/c_2, W_3/c_3, \ldots>$$

其中 S 為含有新生詞彙 X 為關鍵詞彙的文件篇數比率。D 為關鍵新生詞彙 X 在文件資料庫中，最早出現的日期。兩個參數反映新生詞彙 X 受重視的程度：D 值相近的兩個新生詞彙，S 值愈大表示被報導的越多，越受重視。式子的箭頭右邊為一串列，按 c_i 由大至小排列。W_i 為 X 之外的其他關鍵詞彙，c_i 表示 W_i 和 X 的關聯強度；也就是 X 出現下，W_i 也會出現的機率。其值為含有 X 及 W_i 詞彙的文件數除以含有 X 詞彙的文件數。例如一條新生詞彙關聯法則如下，其中的 D 值 142 表示距離文件庫中的最早日期的天數：

- 紅籌股/0.09%, 142→<香港/0.83, 中國/0.61, 概念/0.55, 大陸/0.38, 表現/0.33, 企業/0.33, 指數/0.22, 股價/0.22, 交易/0.16, 證券/0.16, 預估/0.16, 經濟 0.16, 基金/0.16, 股票/0.16 >

結構化資料與高頻詞彙關聯

觀察在不同結構化資料節點的高頻率關鍵詞彙，可以瞭解對應於不同結構化資料，新聞報導的特性。此模式的關聯法則如下：

$sw_{ch,d,i}/s \rightarrow <uw_1/c_1, uw_2/c_2, \cdots, uw_n/c_n>$

其中 $sw_{ch,d,i}$ 代表某結構化資料概念階層樹 ch、深度 d 節點 i 的概念。s 為支持度，亦即為 $sw_{ch,d,i}$ 概念在總文件中出現的比率。式子右邊的 uw_i 為與 $sw_{ch,d,i}$ 節點概念關聯的非結構化關鍵詞彙，c_i 為此關鍵詞彙的信心度值，亦即文章中出現 $sw_{ch,d,i}$ 概念，也出現 uw_i 詞彙的比例，uw_i 按關聯信心度 c_i 由大到小排序。例如以「報導地點」建構結構化資料的概念階層樹，並將新聞文件中的報導地點一般化到區域層次「北部」、「中部」、「南部」等。其中北部地區的報導中，前十個高頻率詞彙的關聯法則如下：

- 北部/55.2%→<逮捕/6%, 死亡/3.8%, 收押/3.7%, 命案/3.6%, 在逃/3.4%, 綁架/2.1%, 自殺/2.1%, 綁匪/2.1%, 逃亡/2.1%, 血案/2.1%>

結構化資料與某同類詞彙關聯

針對某一結構化資料的概念階層的節點，和其它某一同類關鍵詞彙的關聯比較。此模式的關聯法則如下：

$sw_{ch,d,i}/s \rightarrow <uw_1/c_1, uw_2/c_2, \cdots, uw_n/c_n>$

式子左邊的 $sw_{ch,d,i}$ 代表某結構化資料概念階層樹 ch、深度 d 節點 i 的概念，s 為 $sw_{ch,d,i}$ 在總文件中出現的比率，亦即支持度；式子右邊的 uw_i 為其它某一同類的非結構化關鍵詞彙，c_i 為此關鍵詞彙的信心度值。例如以「報導地點」建構結構化資料的概念階層樹，並一般化到「北部」、「中部」、「南部」，收集有關犯罪工具的關鍵詞彙，建構犯罪工具的概念階層樹，並進行適當的一般化，可以獲得如下形式的關聯法則：

- 北部/40.0%→<手槍/56.5%, 持刀/30.2%, 衝鋒槍/13.3%>

3.4.3 探勘新聞文件趨勢

新聞事件報導的多寡，可以反應社會現象的趨勢。例如，飆車事件報導的增加，很可能表示飆車現象有增加的趨勢。關鍵詞彙代表該篇新聞報導的重要內涵，因此關鍵詞彙出現的頻率及關鍵詞彙頻率的變化情形，可以當成社會現象趨勢的一個重要指標。

針對擷取出的關鍵詞彙，新聞文件的趨勢探勘，我們是以線性方程式模式，表示詞頻的變化趨勢：利用線性回歸計算線性方程式，取得方程式的參數：斜率、截距及母體判斷係數。這些參數提供一些趨勢的線索，可引導分析人員進行深入分析。

給予一新聞文件庫 W 及觀察區間長度 n 天。|W|表示整個新聞文件庫涵蓋的時間區間天數，將文件庫切割成 |W|/n 個長度為 n 的連續觀察區間。累計每個觀察區間中，各個關鍵詞彙出現的文件篇數，然後對每一個關鍵詞彙進行線性回歸計算，方程式模式如下：

$$Y=SX+O$$

其中 X 表示觀察區間，Y 表示詞彙的文件篇數。S 為斜率，O 為截距。利用最小平方差的線性回歸技術 [Keller 1994]，斜率 S、截距 O 及母體判定係數 R^2 可以分別計算如下：

$$S=\frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(X-\bar{X})^2}$$

$$O=\bar{Y}-S\bar{X}$$

$$R^2=1-\frac{SSE}{SST},\quad 0\le R^2\le 1$$

$$\left\langle SSE=\sum(Y_j-\hat{Y}_j)^2;\ \ SST=(\sum Y_j^{\,2})-\frac{(\sum Y_j)^2}{n}\right\rangle$$

其中

$\bar{X}$、$\bar{Y}$：為 X 與 Y 的平均值

$\hat{Y}$：利用迴歸方程式計算出的預測值

R^2：母體判定係數

我們以一個六維的值組表示詞彙 K 的趨勢探勘結果：

<K, Q, P, S, O, R^2>

Q 為在文件庫中包含有詞彙 K 為關鍵詞彙的比率；亦即含有關鍵詞彙 K 的文件篇數除以總文件篇數。P 為指定的觀察區間長度。斜率 S 表示關鍵詞彙文件篇數的變化情形，正斜率表示增加的趨勢，負斜率表示減少的趨勢，S 值愈大表示變化越大，趨近於零表示篇數幾乎維持穩定。截距 O 的大小和第一個觀察區間的詞彙頻率有關，篇數越多，截距越大。負的截距表示第一個區間或前幾個區間的詞頻可能為零或比較小，而後面幾個區間的詞頻較大。母體判斷係數 R^2 為預測的可靠性，估算單一關鍵詞彙實際詞頻與預測詞頻之影響變異，R^2 越大，表示 X 對 Y 之直線型影響密切度越高，趨近效果越好。可視為趨勢的穩定性，母體判斷係數愈小表示趨勢的變異性較大，不是穩定的增加、減少或持平。

如果觀察區間由零開始編號，每次增加一，則斜率代表平均每個觀察區間增減的詞彙文件篇數。例如 <教育, 1.3%, quarter, 4.0, -1.5, 0.74>表示在文件庫中包含有「教育」為關鍵詞彙的文件篇數占 1.3%，平均每季約增加四篇。回歸結果的可靠性為 0.74。

3.4.4 分佈差異

新聞事件的分佈可能有差異。例如針對區域分佈或發生時間分佈的差異性。一種分佈分析的方法是利用文件自動分類技術，先將新聞分類，再進一步統計及比較分佈情形。文件自動分類技術頗複雜，通常需先人工分類，再利用分類好的文件訓練分類系統。本研究提出另一種分佈分析方法：透過關鍵詞彙分佈情形，提供分析人員新聞事件的概略分佈狀況。

本研究使用無母數統計方法的卡方分配檢驗分佈情形。非結構化詞彙針對某同類結構化資料的分佈情形，可用下列式子表示：

$$uw/s \rightarrow \langle sw_1/c_1, sw_2/c_2, \cdots, sw_n/c_n \rangle \quad (\chi^2)$$

其中

$$x^2 = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i}$$

O_i ：uw 詞彙與 sw_i 概念關聯的新聞文件篇數

e_i ：uw 詞彙依含 sw_i 概念文件比例分配的預期篇數

其中，關聯法則左邊的 uw 表示某一非結構化關鍵詞彙，s 為 uw 在總文件中出現的比率。關聯法則右邊的 sw_i 表示某結構化資料概念階層中同一深度的節點，c_i 為此關鍵詞彙的信心度值，也就是文章中出現 uw，亦出現 sw_i 概念或詞彙的比例。χ^2 愈大，則表示 uw 詞彙相對於 $sw_1, \cdots, sw_n$ 的預期分佈相差越大。

對於卡方分配的預期分佈，以新聞文件在概念階層中同一深度節點的比例為預期分佈。例如：所有新聞文件報導地點在「北部」、「中部」、「南部」所佔比例為 10:5:1，而文章中包含「掃黃」此一關鍵詞彙，出現在此三地的比例卻為 18:5:1，此時出現明顯分佈上的差異，也就是「掃黃」詞彙出現在報導地點為「北部」的頻率相對於中部及南部，比預期高許多，據此應可推論掃黃事件的分佈很可能北部高於其他地區。

肆、 實驗

本研究以 86 及 87 年社會及財經新聞作為實驗的文件庫，社會新聞共 9,282 篇，財經共有 20,236 篇。由於網路上取得的文章為 HTML 格式，此格式檔案內容除了新聞報導外，還包含額外的顯示控制資料，雛形系統自動篩選過濾出網頁中所需之新聞內容，其中主要包括結構化與非結構化資料部分。

本實驗以中研院的中文詞庫當作斷詞詞庫，進行詞庫式斷詞與詞類標記。沒有收錄在該詞庫的詞彙，就當作是新生詞彙。中央研究院詞庫小組建構的詞庫中，所使用的詞彙是由中研院所蒐集的平衡語料庫中擷取出來的詞彙，語料

庫的內容總計有 9,529,233 個詞彙[中文詞知識庫小組 1993, 1995]，經過合併處理後，目前在該詞庫中收錄 78,410 個詞彙，

4.1 前置處理

關鍵既有詞彙擷取方面。首段及詞頻規則的詞頻門檻值設定為 2，也就是除了在首段外，至少還需出現在其他段落一次。在過濾一般性詞彙及低頻詞彙中，扣除 IDF 值小於 6.5 及大於 9 的詞彙，亦即出現的篇數超過 447 或小於 79 篇的詞彙。表二為經過各個步驟過濾後，社會與財經新聞剩餘的詞目數，詞彙數目分別由 4,563,904 及 7,527,102 降至 1,182 及 1,018。

關鍵新生詞彙擷取方面。首段及詞頻規則的門檻值設為 2，也就是除了首段外，至少還需出現在其他段落一次以上。在持續性步驟的詞彙觀察區間 x 與出現的觀察區間數 y 及出現次數 z，分別設定一個月為一個觀察區間與至少於兩個觀察區間都各出現至少一次。表三為各步驟過濾後，剩餘的詞彙數。

表二 關鍵既有詞彙擷取各步驟及剩餘詞彙數

進行步驟	社會新聞剩餘詞彙數	財經新聞剩餘詞彙數
1.經詞庫式斷詞詞目數	4,563,904	7,527,102
2.剔除單一字元的詞目	1,952,560	4,273,004
3.擷取動詞與名詞	1,604,705	3,738,878
4.首段及詞頻規則	509,149	1,826,467
5.過濾一般及低頻詞彙	90,054	186,198
6.合併及累計相同詞彙	1,182	1,018

表三 關鍵新生詞彙擷取各步驟及剩餘詞彙數

進行步驟	社會新聞剩餘詞彙數	財經新聞剩餘詞彙數
1.經統計式斷詞詞目數	56,792	174,563
2.去除含功能性詞性詞彙	32,847	94,379
3.首段及詞頻規則	5,244	13,147
4.出現持續性	3,412	4,715
5.合併重複性新生詞彙	401	1,225
6.領域相關經驗法則	347	918

4.2 關鍵詞彙擷取評估

本研究以 8 位資管所研究生及 2 位大四學生，進行文件關鍵詞彙擷取實驗。實驗人員都不清楚本研究系統中，所使用的關鍵詞彙擷取方法。文件方面，依時間順序將新聞文件庫切割成二十區段，每個區段各取出一篇社會與財經新聞，形成社會與財經新聞各 20 篇。每一篇新聞文件給予五人進行人工關鍵詞彙擷取。紀錄每個關鍵詞彙同時被幾個人擷取為關鍵詞彙，至少為一，至多為五。

表四為系統及人工擷取關鍵詞彙平均個數比較。針對每一篇新聞，人工擷取重疊四次或五次的詞彙，個數相當少，約只有一個到兩個左右，表示評估人員對於關鍵詞的認知也存在相當大的差異性。驗證了人工擷取關鍵詞彙會有客觀性問題的說法。

4.2.1 首段及詞頻規則法評估

首段及詞頻規則法的實驗主要在於，觀察重要詞彙出現在首段的比例，以及出現在首段的重要詞彙亦重複出現在後面段落的比例。首先，

以「出現在首段的詞彙數」除以「每篇新聞中評估者圈選的詞彙總數」，評估重要詞彙出現在首段的比例(以 F 值表示)。另外，出現在首段的重要詞彙亦重複出現在後面段落的比例(以 S 值表示)，則以「出現在首段及後面段落的詞彙數」除以「出現在首段的詞彙數」獲得其值。

表四 系統及人工擷取關鍵詞彙平均個數統計

	系統擷取個數	人工擷取關鍵詞彙個數				
		受一人認同個數	受二人認同個數	受三人認同個數	受四人認同個數	受五人認同個數
財經新聞	11.5	18.1	8.45	4.2	2.0	0.7
社會新聞	10.3	16.7	8.5	4.8	2.4	1.0

實驗結果如表五，可發現本實驗之新聞文件，具有首段摘要及重要資訊出現在後面段落的特性。二十篇社會新聞中，使用者評選之關鍵詞彙中出現在首段的比例，在兩人、三人、四人及五人認同的關鍵詞彙集合下，分別有平均 78.1%、95.1%、100%、及 100%的比例，而出現在首段的關鍵詞彙中，亦出現在後面段落的比例，分別有 95.2%、96.9%、98.3%、及 100%等。財經新聞的實驗結果亦有相當高的比率。由此也顯示，社會與財經兩類新聞文件格式沒有明顯差異，即表示此兩類的新聞寫作皆具有高程度的首段摘要寫作特性。

表五 評估關鍵詞彙分佈在首段及後續段落假設

	關鍵詞彙受認同人數									
	一人		二人		三人		四人		五人	
	F	S	F	S	F	S	F	S	F	S
社會新聞	59.1%	89.1%	78.1%	95.2%	95.3%	96.9%	100.0%	98.3%	100.0%	100.0%
財經新聞	57.9%	86.7%	80.0%	94.6%	86.9%	97.2%	100.0%	100.0%	100.0%	100.0%

4.2.2 關鍵詞彙評估

以人工擷取的關鍵詞彙為標準，比較系統擷取的成效。比較的方法為資訊檢索領域常使用的正確率(precision)與召回率(recall)。正確率為同時被人工及系統擷取之詞彙個數與系統擷取之詞彙個數的比率。召回率為系統所擷取之個數和人工所擷取個數之比率。表六為 20 篇社會新聞與 20 篇財經新聞關鍵詞彙擷取平均正確率、召回率。

$$\text{正確率} = \frac{\text{評估者與系統同時擷取的詞彙數}}{\text{系統擷取詞彙數}}$$

$$\text{召回率} = \frac{\text{評估者與系統同時擷取的詞彙數}}{\text{評估者擷取的詞彙數}}$$

表六 關鍵詞彙擷取「正確率」與「召回率」統計

	關鍵詞彙受認同人數									
	一人		兩人		三人		四人		五人	
	正確率	召回率	正確率	召回率	正確率	召回率	正確率	召回率	正確率	召回率
社會新聞	39.7%	23.3%	30.5%	34.6%	20.8%	44.3%	12.6%	52.3%	8.6%	87.9%
財經新聞	53.1%	34.4%	38.9%	52.0%	25.8%	67.1%	16.5%	81.5%	7.7%	90.9%

本研究的結果和 Chien 的研究做比較。[Chien 1997]的中文關鍵字擷取評估方法中的「書目索引」方式為：人工方式從包含 20 萬中文字元的一本書中，擷取出 190 個關鍵字，然後比對系統擷取結果。當考慮關鍵字完全相符的情形，約有三成正確率，四成三的召回率；如果

只考慮關鍵字的意義，不須完全字面相符的話，約有三成八正確率與五成六的召回率。而本研究在財經新聞有五成三的正確率與三成四的召回率，正確率優於 Chien 的研究，召回率遜於其研究；社會新聞在有三成九的正確率與兩成三的召回率，正確率略優於 Chien 的研究，召回率遜於其研究。但 Chien 研究中的評估人員只有一位，較容易受個人主觀影響。如觀察本研究在多人同時認同的關鍵詞彙的召回率，則明顯提高。顯示本研究對於新聞文件關鍵詞彙擷取上，對多人同時認同的關鍵詞彙大部分可以擷取出。

4.2 資料探勘

新聞文件經由關鍵資訊擷取之後，在知識探勘階段，分別以關聯法則探勘、趨勢探勘、與分佈差異探勘，進一步進行知識擷取工作。

4.2.1 詞彙關聯探勘

關聯法則探勘的結果分為三個小節，分別為新生詞彙關聯法則探勘、結構化詞彙與高頻詞彙關聯、結構化詞彙與同類詞彙關聯。

新生詞彙關聯法則

探勘基本關聯法則方面，以關鍵既有詞彙與新生詞彙為基礎的關聯法則探勘模式，部份財經新聞的基本關聯法則探勘結果如下，法則右邊的關鍵詞彙只顯示關聯度超過 15%或前十五個詞彙：

- 適足率, 0.12%, 229→<資本/1, 自有/0.51, 財政部/0.47, 增資/0.40, 規定/0.40, 新版/0.36, 風險/0.36, 國際/0.31, 比率/0.23, 資產/0.23, 董事長/0.23 清算/0.20, 轉投資/0.20>
- 晶圓廠, 0.07%, 101→<半導體/0.71, 園區/0.61, 代工/0.53, 產業/0.53, 預估/0.53, 合作/0.38, 生產/0.38, 全球/0.38, 廠商/0.38, 興建/0.38, 成長/0.38>
- 詢價圈購/0.07%, 93→<增資/0.92, 現金/0.92, 承銷/0.69, 上市/0.53, 制度/0.53, 證管會/0.53, 市價/0.53, 股東/0.46, 發行/0.46, 價格/0.38, 證券/0.30, 辦理/0.30, 股票/0.30, 行情/0.23, 調整/0.23, 規定/0.23, 財政部/0.23, 問題/0.23, 股價/0.15>
- 換匯點/0.07%, 248→<國外/0.61, 新台幣/0.53, 法人/0.53, 報價/0.46, 交易/0.38, 外匯/0.38, 投機/0.38, 央行/0.30, 套利/0.30, 亞洲/0.23, 大型/0.23, 匯率/0.23, 貶值/0.23, 台幣/0.23, 利率/0.15>

從關聯關係中可發現，適足率與資本的信心度為 100%，可推演得知，兩詞彙有非常密切的關係，事實上「資本適足率」是一完整財經重要詞彙，與財政部規定企業增資有關。晶圓廠的關聯關係可推測與半導體代工有關。在社會新聞方面的一些關聯特徵如下：

- 簽賭案, 0.24%, 142→<法官/0.36, 職棒/0.31, 專案/0.31, 偵辦/0.31, 檢察官/0.31, 表示/0.31, 英豪/0.27, 球員/0.27, 嘉義/0.27, 嘉義市/0.22, 放水/0.22, 地方法院/0.22, 省議員/0.18>
- 土石流, 0.24%, 213→<昨天/0.45, 發生/0.40, 活埋/0.40, 調查/0.31, 造成/0.31, 人員/0.27, 進行/0.27, 士林/0.27, 檢察官/0.22, 災情/0.22, 前往/0.22, 地檢署/0.22, 災害/0.22, 災變/0.22, 豪雨/0.22, 上午/0.22, 三芝/0.22, 政府/0.18, 要求/0.18, 檢方/0.18, 事件/0.18, 疏散/0.18>
- 躁鬱症, 0.11%, 116→<警方/ 0.5, 自殺/ 0.5, 精神/ 0.5, 身亡/ 0.40, 台北/ 0.40, 現場/ 0.40, 發現/ 0.40, 年級/ 0.30, 表示/ 0.30, 驗屍/ 0.30, 神情/ 0.30, 母親/ 0.30, 高雄/ 0.30, 清晨/ 0.30, 父親/ 0.30, 殺死/ 0.30, 跳樓/ 0.30, 學生/ 0.30, 臥室/ 0.20>

結構化詞彙與高頻詞彙關聯

本文選用社會新聞，結構化資料選用報導地點，並一般化到北中南東離島等區域。實驗結果發現針對報導地點而言，北中南東等不同區域的高頻率詞彙的類別，沒有顯著的不同。離島的高頻詞彙則跟軍方較有關聯。

本研究另外用 86 年度的焦點新聞實驗，總共 11359 篇，其中北、中、南、東、離島各 6955、271、247、29、16 篇，及 3841(33.8%)篇沒有註明報導地點。實驗報導地區和各區前 20 個高頻詞彙的關聯，結果可以發現不同區域的高頻率詞彙類別有明顯的差異。從下列的關聯法則，可以發現發生在北部的焦點新聞偏重在政

治方面。中部偏弊案，南部則偏社會新聞。東部則偏環境及交通。離島則和軍方及交通較有關聯。

- 北部/61.2%→<修正案/0.9%，營建/0.9%，約見/0.9%，處分/0.9%，總統制/0.9%，比率/0.9%，項目/0.9%，盈餘/0.8%，納入/0.8%，法人/0.8%，變更/0.8%，黨務/0.8%，中央政府/0.8%，衛生署/0.8%，刑事/0.8%，委員/0.8%，不足/0.8%，主計處/0.8%，規範/0.8%，副院長/0.7%>
- 中部/2.4%→<賄選/6.2%，農會/5.9%，質詢/4.7%，偵訊/4.4%，暴力/4.4%，到案/3.6%，收押/3.6%，省政/3.6%，芳苑/3.6%，總幹事/3.3%，黨員/3.3%，總部/2.9%，綁票/2.5%，南下/2.5%，設廠/2.5%，搜索/2.5%，道路/2.5%，理事長/2.5%，老人/2.2%，財團/2.2%>
- 南部/2.2%→<園區/6.4%，瓦斯/.2%，球員/4.8%，爆炸/4.0%，醫院/4.0%，南下/4.0%，設廠/3.6%，急救/3.6%，巡視/3.6%，關係企業/3.2%，工業區/3.2%，警局/2.8%，生命/2.4%，進駐/2.4%，危險/2.4%，開槍/2.4%，國中/2.4%，恐嚇/2.4%，事故/2.4%，醫生/2.0%>
- 東部/0.3%→<公園/13.7%，面積/13.7%，登陸/13.7%，災害/13.7%，房屋/10.3%，公路/10.3%，中斷/10.3%，有期徒刑/10.3%,合法/6.8%，公務員/6.8%，正常/6.8%，交保/6.8%，圖利/6.8%，審核/6.8%，儀式/6.8%，農業/6.8%，許可/6.8%，偵訊/6.8%，動工/6.8%，鐵路/6.8%>
- 離島/0.1%→<班機/18.7%，軍方/12.5%，國華/12.5%，監督//6.2%，豬肉/6.2%，妻子/6.2%，任期/6.2%，演習/6.2%，功能/6.2%，法律/6.2%，長官/6.2%，查獲/6.2%，質詢/6.2%，管制/6.2%，爆炸/6.2%，空難/6.2%，老人/6.2%，輔選/6.25，戰機/6.2%，文化/6.2%>

結構化詞彙與某同類詞彙關聯

以「報導地點」建構概念階層樹並一般化到「北部」、「中部」、「東部」、「南部」和「離島」。另外，建一「犯罪型態」概念階層，子節點包括「綁架」、「搶劫」、「強暴」、「賭博」及「偷竊」關鍵詞彙。運用結構化資料與同類詞彙關聯模式，得到下列關聯法則。從中可觀察北中南部對指定的犯罪型態詞彙的關聯沒有明顯的不同。但東部及離島的犯罪型態詞彙的關聯則明顯不同於北中南部。

- 北部/56.1%→<綁架/33.0%，搶劫/24.4%，強暴/15.9%，賭博/15.6%，偷竊/11.1%>
- 中部/13.3%→<搶劫/29.2%，綁架/23.3%，賭博/20.0%，強暴/17.5，偷竊/10.0%>
- 南部/11.7%→<綁架/25.0%，搶劫/25.0%，賭博/20.6%，強暴/14.7%　偷竊/14.7%>
- 東部/1.9%→<賭博/100.0%，　綁架/0.0%，搶劫/0.0%，強暴/0.0%，偷竊/0.0%>
- 離島/0.7%→<偷竊/100%，綁架/0.0%，搶劫/0.0%，　賭博/0.0%，強暴/0.0%>

4.2.2 趨勢分析

在探勘發展趨勢方面，以關鍵既有詞彙與關鍵新生詞彙為基礎的趨勢探勘模式，部分社會新聞趨勢結果如下：

- <環境, 3.6%, quarter, 3.2, -1.7, 0.86>
- <教育, 1.3%, quarter, 4.0, -1.5, 0.74>
- <省議員, 0.7%, quarter, -1.4, 14.6, 0.66>

環境與教育方面的報導平均每季頻率分別增加 3.2 及 4 篇，可推估此議題漸受重視。省議員相關的報導佔文件庫比率並不高，而且有減少的趨勢。部分財經新聞趨勢結果如下：

- <升值, 1.4%, quarter, 6.6, 6.4, 0.72>
- <重整, 0.3%, quarter, 1.5, 0.7, 0.69>
- <裁員, 0.25%, quarter, 1.8, -1.7, 0.67>

4.2.3 分佈差異

以非結構化資料對結構化資料報導地點一般化到區域為例。以 86 年為例，3819 篇社會新聞文章中，扣除沒有報導地點 624 篇，其餘分佈「北部」、「中部」、「南部」、「東部」和「離島」的篇數各為 2141(67%)、

508(15.8%)、448(14.2%)、72(2.2%)和 26(0.8%)篇。探勘出的法則如下所示。其中出現「手槍」詞彙的篇數共 79 篇，分佈情形在「北部」、「中部」、「南部」、「東部」和「離島」的篇數各為 42、16、21、0 和 0 篇，經運算卡方分配值為 14.5。由此值可知含「手槍」詞彙新聞文件對報導地區的分佈情形和所有文件對報導地區的分佈情形有差異：基本上是「手槍」詞彙出現在南部及中部的報導，比預期的比例還多。「軍方」詞彙的分佈差異更大，在離島的分佈比預期多出許多。「抗議」及「贓物」詞彙的則和預期分佈叫接近，顯示這些詞彙的分佈較沒區域性的區別。

- 手槍/2.5%→<北部/53.2%, 南部/26.6%, 中部/20.3% 東部/0 離島/0> (14.5)
- 軍方/0.9%→<北部/43.3%, 中部/16.7%, 南部/23.3% 東部/0% 離島/16.7%> (97.7)
- 抗議/0.9%→<北部/69%, 南部/13.8%, 中部/13.8% 東部/3.5 離島/0> (0.5)
- 贓物/0.5%→<北部/64.7%, 南部/17.7%, 中部/17.7% 東部/0 離島/0> (0.7)

伍、討論

本研究探勘的結果，基本上是巨觀的特徵，而非微觀的知識。也就是，提供大量資料中隱含的特徵，作為進一步深入探討特徵形成的原因以及該特徵所代表的精確意義。例如，趨勢分析中的「省議員」詞彙有下降趨勢，只提醒分析人員省議員的相關報導減少，至於為何減少及這些關於省議員新聞的議題，則需分析人員進一步探索。例如，搭配關鍵詞彙關聯分析，觀察常和省議員詞彙一起出現的其他關鍵詞彙。

受限於實驗資料取得不易，如果能分析更長時間區段的新聞文件集合，例如十年的新聞文件。相信在趨勢探勘方面可以探勘出更有顯著的特徵。另外，從實驗中發現，在同一篇或同一類文件中，會因寫作習慣或需要，以不同詞彙表達相同概念，例如「自殺」、「自盡」與「自裁」。對此同義詞彙的問題，如果可以引進一同義詞詞典，在關鍵資訊擷取步驟處理中，將同意詞彙頻率合併，可讓詞彙頻率統計更為準確。

從我們較熟悉的社會新聞判斷，探勘所得的結果，頗能反映社會的現象。因此，本探勘架構的探勘出的特徵有相當程度的參考性。特別是應用到分析人員所不孰悉的專業領域，有助於了解該陌生領域；或應用於分析早期年代的新聞資料，例如 20 年前的新聞文件，本探勘方法有助於不同時代的人，了解當時的社會概況。

本研究提出的文件探勘方法，其適用性事實上不只侷限於中文新聞的文件類別。雖然本研究以中文新聞集合為例，但文中所提出的方法，亦適用於其他的中文文件，只要該文件撰寫格式具有全文摘要及全文詳述等兩個部分的格式架構，而且文件集合具有跨時間紀錄的特性，即可適用於本研究探勘流程中的處理方法。

陸、結論與未來方向

面對文件量暴增的時代，需要電腦化的文件處理工具，幫助使用者進行文件分析與知識擷取。查詢導向的資料檢索技術，無法挖掘隱藏在大量文件中的特徵。目前的文件探勘技術都針對歐美語系文件，而非中文文件，且關鍵詞彙擷取的前置處理都是人工作業。本研究的貢獻在於，提出一個包含自動化前置處理的中文新聞文件資料探勘架構。並針對報告性文件的撰寫格式，提出文件的關鍵詞彙擷取方法。在探勘模式方面，利用新生詞彙提供發覺新社會現象或新事物的線索；以關聯法則為基礎，我們提出三類擴充式關聯法則；以及分析關鍵詞彙分佈差異等方法，挖掘大量文件中隱含的特徵。

下列兩點是值得未來繼續努力的方向：

1. 找尋其他適用於文件探勘的興趣指標：本研究在關聯法則探勘模式中以信心度及支持度作為過濾法則數量的興趣指標，可獲得一些探勘成果。針對不同的需求，可考慮其他的興趣指標。資料探勘領域中，其它的興趣指標，包括功效 (strength) [Dhar 1993]、收益量 (gain) [Fukuda 1996]、興趣 (interest) [Brin

1997]、定獻(conviction) [Brin 1997]、拉普拉斯[Webb 1995]等，分別具有不同的特性和用途。如何將其它興趣指標應用到文件資料探勘，或配合文件特性，尋找其他特殊指標，是值得未來研究的方向。

2. 中英夾雜文件處理：現在許多中文文件都夾雜英文，特別是科技類文件或財經類文件，部分英文詞彙隱含有用的訊息。但對於夾雜的英文字彙部分，需要特別的處理與判斷，是值得進一步探討的問題。

誌謝

本研究承蒙國科會的贊助(計畫編號：NSC89-2416-H-224-018)，特此致謝。另外，感謝胡勝傑的初期參與及林彥成協助開發雛型系統。

參考文獻

1. 中文詞知識庫小組，1993，"新聞語料詞頻統計表"，技術報告，TR-93-02，中央研究院，南港
2. 中文詞知識庫小組，1995，"中央研究院平衡語料庫"，技術報告，TR-95-02，中央研究院，南港
3. 陳克健、陳正佳、林隆基，1986 ，"中文語句的研究－斷詞與構詞"，技術報告，TR-86-006，中央研究院，南港
4. Agrawal, R., T. Imielinski, and A. Swami, 1993, "Mining Association Rules between Sets of Items in Large Databases", Proceedings of the 1993 ACM SIGMOD Conference, pp. 207-216.
5. Aumann, Y., et al.,1999, "Circle Graphs: New Visualization Tools for Text-Mining", Proceedings of Third European Conference on KDD(PKDD-99), pp. 165-173.
6. Brachman, R. J., et al., 1996 November, "Mining Business Database", Communication of ACM, 39(11), pp.42-48.
7. Brin, S., et. al., 1997, "Dynamic Itemset Counting and Implication Rules for Market Basket Data", Proceedings Of the ACM-SIGMOD international Conference On the Management of Data, pp.255-264.
8. Chen, K. J. and S. H. Kiu, 1992, "Word identification for Mandarin Chinese sentences", Fifth International Conference on Computational Linguistics, pp. 101-107.
9. Chien, L.-F., 1997, "PAT-Tree-Based Keyword Extraction for Chinese Information Retrieval", Proceedings of the 20th Annual ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 50-58.
10. Cho, V. and B. Wüthrich, 1999, "Combining Forecasts from Multiple Textual Data Sources", Proceedings of 3'rd Pacific-Asia Conference on KDD (PAKDD-99), April 1999, pp.174-178.
11. Dhar, V. and A. Tuzhilin, 1993, "Abstract-driven discovery in databases", IEEE Transactions on Knowledge and Data Engineering, 5(6), pp.926-938.
12. Dörre, J., P. Gerstl and R. Seiffert, 1999, "Text Mining: Finding Nuggets in Mountains of Textual Data", Proceedings of the 5's ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.398-401.
13. Fan, C. K. and W. H. Tsai, 1998, "Automatic Word Identification in Chinese Sentences by the Relaxation Technique", Computer Proceeding of Chinese and Oriental Languages, pp.33-56.
14. Fayyad, U. and R. Uthurusamy, 1996a, "Data mining and knowledge discovery in databases", Communications of the ACM, 39(11), pp. 24-26
15. Fayyad, U., G. Piatetsky-Shapiro and P. Smyth, 1996b, "The KDD Process for Extracting Useful Knowledge from Volumes of Data", Communication of the ACM, 39, pp. 27-34.
16. Fayyad, U., G. Piatetsky-Shapiro and P. Smyth, 1996c, "From Data Mining to Knowledge Discovery: An Overview", Advances in Knowledge Discovery and Data Mining, pp.1-36.

17. Feldman, R., and I. Dagan, 1995, “Knowledge Discovery in Textual Database(KDT)”, Proceedings of the first ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.112-117.
18. Feldman, R., and H. Hirsh, 1996, “Mining Association in Text in the Presence of Background Knowledge”, Proceedings of 2’nd international Conference on Knowledge Discovery and Data Mining, pp.343-346.
19. Feldman, R., W. Klosgen and A. Zilberstein, 1997a, "Visualization Techniques to explore Data Mining Results for Document Collections", Proceedings of the Third International Conference on Knowledge Discovery & Data Mining, pp.16-23.
20. Feldman, R., et al., 1997b, “Pattern Based Browsing in Document Collections”, Proceedings of First European Symposium on Principles of Data Mining and Knowledge Discovery, pp.112-122.
21. Feldman, R. and H. Hirsh, 1997c, “Exploiting Background Information in Knowledge Discovery from Text”, Journal of Information System, 9, pp.83-97.
22. Feldman, R. and I. Dagan, 1998a, “Mining Text Using Keyword Distribution”, Journal of Information System, 10, pp. 281 - 300.
23. Feldman, R., et al., 1998b, “Text Mining at the Term Level”, Journal of Intelligent Information System, pp.65-73.
24. Feldman,R., et al., 1998c, “Trend Graphs: Visualizing the Evolution of Concept Relationships in Large Document Collections”, 2’nd European Conference on KDD, pp.38-47.
25. Han, J., Y. Gai and N. Cercone, 1993, “Data-driven discovery of quantitative rules in relation databases”, IEEE Transactions On Knowledge and Data Engineering, pp.29-40.
26. Keller, G., B. Warrack, and H. Bartel, Statistics for Management and Economics, 3rd ed., Duxbura Press, Belmont California, 1994.
27. Lagus, K, T. Honkela, S. Kaski, and T. Kohonen, “Self-Organizing Maps of Document Collections: A New Approach to Interactive Exploration”, Proceedings of Conf. on Knowledge Discovery and Data Mining, 1996, pp.238-249.
28. Lent, B., R. Agrawal, and R. Srikant, 1997, “Discovering trends in text databases”, Proceedings of Conference On Knowledge Discovery and Data Mining, pp.227-230.
29. Li, B.-I., et al., 1991, “A maximal matching automatic Chinese word Segmentation algorithm using corpus tagging for ambiguity resolution”, R.O.C. Computational Linguistics Conference, Taiwan, pp.135-146.
30. Nie, J., M. Briscbois and X. Ren, 1996, “On Chinese Text Retrieval”, Conference Proceedings of SIGIR, pp.225-233.
31. Shewhart, M. and M. Wasson, 1999, “Monitoring a newsfeed for hot topics”, Proceedings of the 5th Int’l Conf. On Knowledge Discovery and Data Mining, pp. 402-404.
32. Singh, L., P. Scheuermann and B. Chen, 1997, “Generating Association Rules from Semi-Structured Documents Using an Extended Concept Hierarchy”, ACM IKM, pp.193-200.
33. Singh, L., et al., 1999, “An Algorithm for Constrained Association Rule Mining in Semi-structured Data”, PAKDD-99, April, pp.148-158.
34. Sproat, R. and C. Shih, 1990, “A Statistical Method for Finding Word Boundaries in Chinese Text”, Computer Processing of Chinese and Oriental Languages, pp. 336-351.
35. Webb, G.. I., 1995, “OPUS: An Efficient Admissible Algorithm for Unordered Search”, Journal of Artificial Intelligence Research, 3, pp.431-465.
36. Wuthrich, B., et al., 1998, “Daily Stock

Market Forecast from Textual Web Data", IEEE International Conference on SMC, pp.1-6.