中文新聞文件的關聯法則探勘

許中川 陳俊男 胡勝傑 林彥成 邱宣諭 羅道祈 卓恆
國立雲林科技大學資訊管理系所
雲林縣斗六市大學路三段 123 號
hsucc@mis.yuntech.edu.tw

摘要

新聞文件記錄每天發生的重要事件，在這些大量的新聞文件中，往往蘊含重要的資訊。本研究提出一個自動化探勘架構，從大量的新聞文件中擷取出有用的關鍵詞彙，以關聯法則進一步萃取出潛藏的知識。在探勘過程中，針對中文新聞文件結構的特殊性，我們以結合詞庫式斷詞與統計式斷詞的混合式斷詞法進行中文斷詞；根據新聞撰寫經驗法則，提出四個處理程序，取得新聞文件中較具代表性的關鍵詞彙；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，搭配改良後的關聯法則，提出四個關聯模式：第一個是基本關聯法則，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同類詞彙關聯，第四個為非結構化資料的分佈差異。最後我們以實驗驗證此探勘架構的可行性。

關鍵詞：文件資料探勘 關聯法則 中文斷詞 關鍵詞擷取 分佈差異

1 緒論

隨著數位化時代來臨，文件的紀錄方式由傳統的紙上資料轉向數位化儲存。網際網路的風行，為利用數位傳輸的便利及快速，以數位化儲存的新聞媒體文件亦隨之快速成長。在這些大型文件資料庫中往往隱含有用的資訊[6]。傳統人工的分析方法耗費人力及時間，僅能處理少量的資料。電腦化的資料檢索系統只能檢索出滿足查詢條件的文件，而無法分析文件歸納出有用的潛藏知識[13]。有別於檢索技術，資料探勘技術可以用來挖掘大量文件中的知識[3,4]。

在傳統的資料庫中，資料探勘技術運用於大量消費性資料，可從資料庫中找出產品銷售間的關聯情形，並進一步幫助決策者訂立有利的行銷策略[1]。由於傳統上的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中，可能隱含的大量資訊[13]。相對於關聯資料庫中定義明確的表格與欄位，所謂非結構化資料，其內容並無一定的格式且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。[6,7,8,13]運用資料探勘技術於文件資料上，在文件的探勘過程中，單一詞彙構成個別的概念，藉由概念與概念間的關聯關係，可以進一步發現有意義的資訊，而這些隱含在文件中不易察覺的資訊，可以適切地應用在不同的用途上。[16]運用資料探勘技術於網際網路的財經新聞網頁中，經由找出關鍵詞彙配合領域知識的訓練資料，進而預測每日股票指數的漲跌。

目前文件資料探勘研究都針對印歐語系文件，且代表文件的關鍵詞彙都由人工擷取。中文文件的詞彙組成方式和印歐語系文件不同，需要特別的前置處理程序。本研究的目的針對中文新聞文件的資料探勘提出架構。探討架構中各組成的功能及技術，並透過實驗驗證架構以及探勘方法的可行性。本文的內容組織如下：第二節介紹文件探勘方面的相關研究，第三節說明本文所提出的中文新聞文件資料探勘架構，第四節為實驗結果，第五節討論，第六節結論。

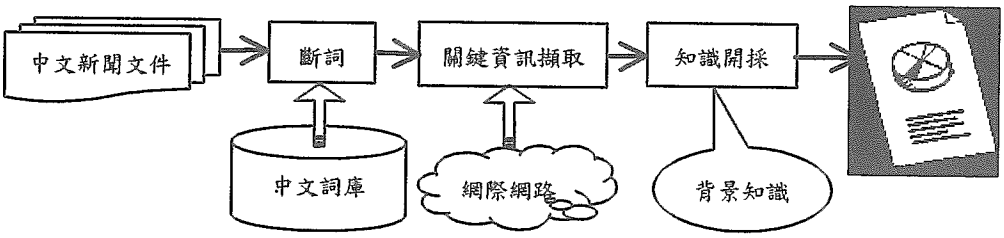
2 相關研究

Feldman 首先提出 KDT(Knowledge Discovery in Texts)的概念。作者以概念階層(concept hierarchy 表示相關的背景知識，利用背景知識將文件結構化並限制及導引後續資料探勘的進行[6,8]。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，以利進一步探勘進行。研究中以概念分佈(concept distributions 方式找出文件中的重要資訊。概念分佈為各個子概念對於其他子概念的分佈情形。其分佈比較方式可分三大類，包括一般分佈分析(uniform distributions) 、兄弟分佈分析(sibling distributions)及趨勢分析(past distributions)。利用上述的分析模式，分析人員可以得到對大量文件資料的概觀。[7]提出視覺化技術，用關聯圖表達文件庫中關鍵詞彙之間的關聯關係。[13]從一群半結構化的文件資料中，得到結構化資料與非結構化資料，並找出在某一結構化資料下，非結構化資料之間的關聯性。Singh 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(rule)——一般法則(general rules)、父法則(parent rules)、子法則(child rules)、兄弟法則(siblin rules)。[16]利用文件探勘方式預測股市漲跌。不同於傳統預測方式，該研究採用各大財金網站上的文件資料進行股市預測。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，推導股市漲跌走勢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。

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及關鍵詞彙。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，上述研究都是針對英文文件，文件的處理技術和中文不相同。因此本研究提出適合中文新聞文件探勘的架構，以及針對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。

3 探勘架構

中文新聞文件探勘需要特殊的探勘架構。目前大部分文



圖一　中文新聞文件資料探勘架構

件資料探勘研究主要應用於印歐語系文件探勘，比較於中文文件探勘需先斷出一個個有意義的詞彙，有其先天上的差異，而無法直接使用於中文文件探勘。因此，有必要針對中文文件提出特殊的探勘架構。因應此需求，本研究提出一個中文新聞文件的探勘架構如圖一。

一般資料庫探勘流程包括選擇、處理、轉換資料、探勘及解譯等[5]，其中前三個步驟可視為前置處理。本研究的斷詞與關鍵資訊擷取可視為前置處理。在斷詞步驟，我們提出一個特殊混合式斷詞方法，斷出一個個有意義的詞目。在關鍵資訊擷取步驟，關鍵詞擷取的品質會影響後來探勘的成果，因此在關鍵詞擷取步驟中，本研究使用了四個有次序性的程序，設法過濾掉不具代表性的一般詞彙，得到較具代表性的關鍵詞彙。並且在斷詞的同時，加入產生新生詞彙的方法，我們以詞類標記加上子集合統計的方式取得新生詞彙。並由於新生詞彙具有高度重要性的關鍵資訊，因此在新生詞彙取得之後，過濾掉專有名詞後加入關鍵詞彙裡面。

在斷詞與關鍵資訊擷取之後是知識開採的階段。在關鍵資訊擷取後，可以得到非結構化與結構化的關鍵詞彙，我們以概念階層表示背景知識，運用四種關聯模式探勘中文文件中潛在的特徵或知識，主要包括以條件機率為基礎的關聯法則與以卡方分配為基礎的分佈差異性分析，各步驟的詳細內容敘述如下。

3.1 斷詞

文件探勘的第一步驟為斷詞，而中文斷詞的處理有別於印歐語系文件的斷詞程序。目前的文件探勘研究都是針對印歐語系文件，印歐語系文件在詞(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。但中文文件字字相連，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一個個的詞目。

中文文件斷詞的方法可分為詞庫式斷詞[2]、統計式斷詞[15]與混合式斷詞[12]。詞庫式斷詞的作法，對照詞庫內收集的詞目比對句子中可能隱含的詞目，找出可能的中斷點，以中斷點斷出一個個詞目。其演算法相當直覺，為目前普遍使用的方法。此法必須時常對詞庫內容加以更新，才能避免因為詞庫中未收錄新生詞彙而影響斷詞品質。統計式斷詞則參考一大型語料庫(corpus)上的統計資訊，單純以鄰近字元同時出現頻率高低作為斷詞的依據，由於語料庫屬於領域相關(domain dependent)，不同語料庫間的統計資訊不適合互用[12]。再者，統計式斷詞常受限於一階馬可夫模式(first-order Markov models)[11]，進一步擴充此模式會提高演算法的時間複雜度[12]，所以大多只針對二字詞進行處理，三字詞如：「登革熱」就無法有效擷取，而且只針對鄰近字元出現的頻率決定斷詞點，沒有考慮語意的正確性，常會產生無意義的斷詞結果。[12]結合詞庫式斷詞與統計式斷詞發展出混合式斷詞，利用詞庫找出可能詞彙輔以語料庫的統計資訊，決定最可能的斷詞點，降低無意義詞彙的機率，此法亦需要詞庫的維護與收集適用的語料庫。我們針對中文新聞文件特性建構一混合式斷詞法，以切合中文新聞文件資料探勘的需求。

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，我們稱此經驗法則為新聞撰寫特性。因此，重要的詞彙會在概要性描述的第一段中出現，如果以第一段字元緊鄰出現的頻率為基礎，可加強同一篇文章統計式斷詞的統計資訊。另外，重要信息通常出現在概要性描述的第一段並在後面段落再詳述一次，因此重要詞彙通常會在一篇文章中同時出現兩次以上。

本研究結合詞庫式及統計式斷詞兩種方式進行斷詞。先採用詞庫式斷詞將所有在詞庫中的詞彙斷出，得到有意義的詞彙。由於詞庫式斷詞已將大部份已知詞彙斷出，大幅減少需要統計式斷詞的詞彙，如此可降低之後統計式斷詞的複雜度。接著使用統計式斷詞。我們根據新聞撰寫特性，認為重要的新生詞彙會出現在第一段及重複地出現在其他段落。因此，本研究從未知詞彙的字詞集合，找出未知詞彙可能的組合情形，再從當中取出詞頻大於門檻的詞彙[19]。

根據實驗觀察，在這些新組合成的詞彙中，若含有介系詞或連接詞，例如「在台中港」，通常不符合後續探勘所需。因此，我們剔除這些組合，以避免無意義新詞彙。

3.2 關鍵資訊擷取

關鍵資訊擷取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識出各個詞彙，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。

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(semi-structured)資料。針對此特性，關鍵資訊擷取可分為結構化資訊擷取與非結構化關鍵詞彙擷取。

3.2.1 結構化資訊擷取

結構化資訊為固定在新聞文件特定位置的提示文字，可以得到切確的資訊。例如：「日期」、「報導地點」，由結構化資訊推導出來的知識，將會具有明確的語意。此外，網路上亦存在許多各式各樣的資料，如股價資料、天候

資料及經濟成長率等，若擷取及適當處理後，可以當成結構化資訊，進一步和新聞本文的非結構化詞彙進行關聯分析。

3.2.2 非結構化關鍵詞彙

在非結構化本文的關鍵詞彙擷取上，我們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，(2)擷取名詞與動詞，(3)首段及詞頻規則法，(4)過濾一般性詞彙。

第一步驟為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、....等。我們首先將單一詞目的詞彙剔除。

在第二步驟中，配合探勘目標的不同需求，以詞彙的詞性，過濾掉一些不重要詞性類別的詞彙。[17,18]以詞類標記來標示一個詞在句子中的語法功能，由於本研究目的在找出具代表性的關鍵詞彙，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。例如表一為一些在大部分的應用中，可考慮剔除的詞性類別。

表一　可剔除的詞性類別

詞性	說明
感嘆詞(I)	一般出現於句前，例：喔，我知道了
介詞或前置詞(P)	帶論元的功能詞，例：他從家裡來
語助詞(T)	大都出現於句尾，如：你來嗎
連接詞(C)	例：張三和李四，大又圓，雖然他很聰明，但是不用功，妳不來的話，我也不來
副詞(D)	緊接在狀態動詞前，如：萬分難過。緊接在動詞後，如：難過萬分。通常出現在句首的句副詞，如：總而言之，你不對
代名詞(Nh)	如：這、那、哪、什麼、其....
位置詞(Ncd)	如:南北、側面、頂端、對面....
量詞(Nf)	如：千萬、小時、公分、世紀、克拉、輩子、學期....
後置詞(Ng)	如：之前、其間、為止、當中....
時間詞(Nd)	如：一月、七夕、下午、今天、次月....
狀態謂賓動詞(VL)	如：不惜、招致、長於、故意、急於、負責、開始、輪流....
狀態類及物動詞(VI)	如:不遺餘力、心軟、生疏、求助無門、受氣、受教、知情、相符、偏心、深信不疑、纏身....
分類動詞(VG)	如: 充當, 出任, 轉任, 類似, 當選, 做為, 晉升, 命名
雙賓動詞(VD1)	如: 分配, 出租, 捐贈, 傳送, 轉交, 讓與
地方名詞(Nc)	包括 1.名方式地方名詞，例：海外、身上、腳下，2.表示物相對位置的地方詞，及 3.雙音節位置詞，如：上頭、中間、右方、西北....

第三步驟的首段及詞頻規則是依據新聞撰寫特性，認為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θ，也就是在第一段以後至少出現 θ-1 次。

第四步驟為過濾一般性詞彙留下關鍵詞彙。我們目前使用反轉文件頻率(Inverse Document Frequency, IDF)方法[14]。反轉文件頻率反映詞彙在文件集合中的分佈情形。反轉文件頻率計算公式如下：

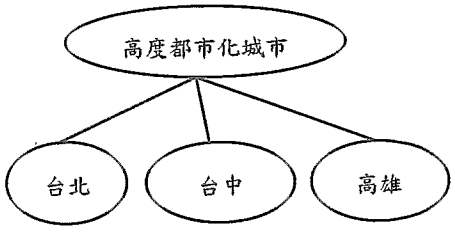
$$IDF(w)=\log_2(n)-\log_2(O(w))+1 \quad (\text{式一})$$

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。

3.3 背景知識

探勘巨量的資料是一項浩瀚的工程，適當的運用背景知識可以大幅提升探勘效率與品質。知識的表示有許多種方法，在資料探勘領域中，常用的背景知識表示方法包括階層樹、語意網路、機率網路、一階邏輯等。

本研究的資料探勘過程亦加入背景知識的輔助，利用概念階層(concept hierarchy)[9]的方式建構背景知識。概念階層是一種階層樹，在本研究的概念階層結構中，葉節點存放的資訊是文件中的原始詞彙或概念，每個葉節點有指標指向包含該節點資訊的文件，以加速探勘處理。例如概念「台北」代表著所有與「台北」相關聯的文件集合。另外，各階層的概念間存在一般化(generalization)及明確化(specification)的關係，愈接近根節點為愈一般化的高層次概念，如圖二所示。有效地利用概念階層可以容易地作出一般化與明確化的運算，找出更概觀或簡潔的資訊，如「70%的犯罪案件發生在台北市、高雄市」，這樣的資訊在導入概念階層後，將得到「犯罪事件大都發生在高度都市化的地方」。



圖二　概念階層範例

3.4 知識開採

在取得關鍵資訊之後，本研究的知識開採階段分為關聯法則探勘與分佈差異探勘。在關聯法則探勘中，運用三種關聯模式探勘關鍵資訊的關聯關係。在分佈差異探勘中，運用卡方分配(ki-square)檢驗異常分佈。

3.4.1 關聯法則

本研究以 Agrawal 提出的關聯法則觀念為基礎，針對文件特性，加以修改及擴充。 [1]提出關聯法則的觀念並應用於銷售產品的分析。關聯法則「A」→「B」表示產品 A 與產品 B 於銷售上的關聯性。我們提出配合新聞文

件特性的關聯探勘方式，經由文件前置處理可取得結構化與非結構化資訊，如果輔以階層結構的概念階層樹，可以得到以下三種探勘模式：

基本關聯法則

針對文件集合 D，經斷詞及關鍵詞彙擷取後，每篇新聞文件以擷取出的詞彙代表。|D|表示文件數，N(w)表示包含關鍵詞彙 w 的文件數。文件集合內隱含的詞彙關聯法則以下列式子表示。

$$w/s \rightarrow w'/c \quad (\text{式二})$$

其中

w, w' 為關鍵詞彙

$s = N(w)/|D|$

$c = N(w \cap w')/N(w)$

關聯法則左邊的 w 表示某一關鍵詞彙，s 為 w 在總文件中出現的比率，稱為支持度。關聯關係右邊的 w'為另一關鍵詞彙。c 為 w 出現的文章中，亦出現 w'的比率，稱為信心度。例如：販毒/1.13%→海洛英/60%，代表「販毒」關鍵詞彙出現在總文件的比率為 1.13%，而含有「販毒」詞彙文章中出現「海洛英」詞彙的比率為 60%。

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。例如結構化資料報導地點一般化成北部、中部、南部及東部等的新聞，各地區的高頻詞彙可以反應該區新聞事件特性。此模式的關聯法則如下：

$$sw_{ch,d,i}/s \rightarrow <uw_1/c_1, uw_2/c_2, \ldots, uw_n/c_n> \quad (\text{式三})$$

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 概念的比例，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%>

結構化資料與同類詞彙關聯

針對某一結構化資料的概念階層的節點，和同一類關鍵詞彙的關聯比較。例如針對報導地點「北部」、「中部」及「南部」，觀察犯罪工具「手槍」、「持刀」及「衝鋒槍」的關聯情形。此模式的關聯法則如下：

$$sw_{ch,d,i}/s \rightarrow <uw_1/c_1, uw_2/c_2, \ldots, uw_n/c_n> \quad (\text{式四})$$

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 為同一類別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%>

網際網路盛行，網路上多樣且豐富的資料來源亦可作為建構概念階層之用。除了使用新聞文件作為資料探勘的來源之外，在背景知識建構的同時可以參考其他相關的有用資訊，輔以概念階層的模式建構適用的概念階層樹。例如擷取網路上的股市盤價，適當處理後當成結構化資料。以每日股市股價漲跌點數建構收盤價概念階層樹，或許可以得到「護盤基金」詞彙與「股市上漲」概念間的關聯性。

3.4.2 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。

我們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：

$$uw/s \rightarrow <sw_1/c_1, sw_2/c_2, \ldots, sw_n/c_n> \ (\chi^2) \quad (\text{式五})$$

其中

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i} \quad (\text{式六})$$

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，式五關聯法則左邊的 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 $sw_1, \ldots, sw_n$ 的預期分佈相差越大。

對於卡方分配的預期分佈，我們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。

4 實驗

本實驗雛形系統以 Borland C++ 4.0 版開發，作業系統為 Windows NT 4.0，使用民國 86 年份的網路電子報共 3819 篇社會新聞作為探勘資料的來源，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，本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。

在本研究的實驗中，比對中研院的詞庫作為判斷某一詞彙是否為新生詞彙的依據。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，語料庫的內容總計有 9,529,233 個詞彙[17,18]，經合併處理後，目前在該詞庫中收錄 78,410 個詞彙，先經由與此詞庫中詞彙的比對斷詞。然後，再用統計方式，斷出不在詞庫中的詞彙，視為新生詞彙。

表二　新聞文件處理範例

結構化資料
標題：狂犬病若入侵 立委要求邱茂英下台負責 日期：860429 類別：社會傳真 報導地點：台北
非結構化資料
記者崔慈悌台北報導 因口蹄疫事件受到立委要求下台的農委會主委邱茂英，二十八日在立法院經濟委員會中，被多位立委要求保證狂犬病一旦侵入台灣後不致造成民眾死亡，否則應立即辭職。不過邱茂英表示....
非結構化關鍵詞彙
農委會，防疫，下台，卸責
新生詞彙
口蹄疫、邱茂英

從網路上抓取原始 HTML 格式的新聞文件資料檔案並過濾掉無意義的內容之後，可獲得結構化與非結構化資料，再經過前置處理之後，可擷取得到非結構化關鍵詞彙與新生詞彙，如表二所示。經斷詞及關鍵資訊擷取後，詞彙數目由 1,975,671 降至 1171；另外，還有 8,705 個新生詞彙。在首段與詞頻規則的詞頻門檻值，我們設定 2，也就是除了在首段外，至少還需出現在其他段落一次。在過濾一般性詞彙步驟中，我們扣除 IDF 值小於 5.26 及大於 8.11 的詞彙，亦即出現的篇數超過 100 或小於 14 篇的詞彙。表三則為各步驟過濾後，剩餘的詞目數。

表三　關鍵詞擷取各步驟及剩餘詞彙數

進行步驟	剩餘詞彙數
1.經斷詞詞目數	1,975,671
2.剔除單一字元的詞目	759,347
3.擷取動詞與名詞	560,643
4.首段優先法	87,436
5.過濾一般性詞彙	1,171

初步的實驗結果顯示各種模式的探勘，都能找到一些令人感興趣的資訊。

基本關聯法則

探勘出的部分基本關聯法則如下：

- 販毒/1.13%→海洛因/60%
- 販毒/1.13%→安非他命/41%
- 飆車族/0.31%→青少年/75%
- 愛滋病/0.29%→雞尾酒療法/54%
- 按摩/0.24%→色情/66%
- 炸彈/0.45%→恐嚇/35%

結構化資料與高頻詞彙關聯

我們先使用社會新聞，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。

我們另外用 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/6.2%, 戰機/6.2%, 文化/6.2%>

結構化資料與同類詞彙關聯

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，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/100.0%, 綁架/0.0%, 搶劫/0.0%, 賭博/0.0%, 強暴/0.0%>

非結構化資料的分佈差異

我們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。在 3819 篇總文章中，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、508、448、72 和 26 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」詞彙則和預期分佈較接近，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/69.0%, 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)

5 討論

初步的實驗結果顯示，雖然可以從探勘的法則中發現一些有用的特徵，但我們也發現幾個問題。首先，由於記者的寫作習慣或為了增加文章的可讀性，在同一篇或同一類文件中會以不同詞彙表達相同概念，如自殺案件的描述就常常會以「自殺」、「自盡」、「自裁」或「自縊」等同義詞描述，這樣往往會造成「自殺」一詞的頻率被分散了，而可能錯失掉這些詞彙或頻率被稀釋。對此，我們認為可以引進一同義詞詞典，在關鍵資訊擷取步驟處理中，將同義詞彙合併。

另外一個問題是探勘出的關聯法則數目過多，且其中會因關鍵詞品質而產生出無意義的法則。從我們實驗結果發現，關鍵詞彙的數目平均每篇約有十個，在 3819 篇的新聞資料中共擷取關鍵詞彙 1,171 個(不同兩篇可能會有重複關鍵詞彙)。如果以基本關聯法則的詞彙兩兩關聯為例，會有 1,171*1,170 個關聯法則。雖然可以透過支持度及信心度門檻值減低法則數目，但有研究指出支持度與信心度指標過於粗糙[10]，設定低門檻值會有法則過多及參雜無用法則；然而設定高門檻值，卻容易漏掉有用法則，如何使用適當的興趣指標(interestingness measure)過濾關聯法則，值得未來進一步研究。

6 結論

在本論文中，我們針對中文新聞文件提出一個資料探勘流程，以此探勘流程挖掘出隱含在大量文件中的有用知識。雖然資料檢索提出許多處理文件資料的技術，然而無法自動從大量文件中擷取出隱含的知識，幫助決策分析。本研究在中文新聞文件利用新聞撰寫特性提出斷詞技術，混合式中文斷詞法可以斷出不包含在詞庫內的詞彙且不需要語料庫的支援；在關鍵詞彙擷取上，運用四個關鍵詞擷取步驟得到較具代表性的詞彙；最後以關聯法則為基礎，進行資訊的關聯探勘及分佈分析。實驗結果顯示本探勘方法，確能找出一些令人感興趣的特徵。

本研究的探勘目標是中文新聞文件，在中文文件的領域裡，新聞文件有其特殊寫作技巧，探勘流程的部分技術是屬於領域相關。如果將此探勘架構擴充範圍至一般性文件，將需要進一步對文件斷詞處理及關鍵詞擷取步驟作修改，以符合探勘一般性文件的需求。如何擴充此兩項技術，以便有效率的斷詞及擷取出高品質的關鍵詞彙，值得未來進一步研究。

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。

參考文獻

[1] Agrawal, R. and T. Imielinski, A. Swami, 1993, "Mining Association Rules between Sets of Items in Large Databases", Proceedings of the 1993 ACM SIGMOD Conference, pp. 207-216

[2] Chen, K. J. and S. H. Kiu, 1992, "Word identification for Mandarin Chinese sentences", Fifth International Conference on Computational Linguistics, pp. 101-107

[3] Fayyad, U. and R. Uthurusamy, 1996a, "Data minin and knowledge discovery n databases", Communications of the ACM, Vol. 39, No. 11, pp. 24-26

[4] Fayyad, U., G. Piatetsky-Shapiro and P. Smyth, 1996b, "The KDD Process for Extracting Useful Knowledge from Volumes of Data", Communication of the ACM, Vol. 39, pp. 27-34

[5] Fayyad, U., G. P atetsky-Shapiro, and P. Smyth, 1996c, "From Data Mining to Knowledge Discovery: An Overview", Advances in Knowledge Discovery and Data Mining, pp. 1-36

[6] Feldman, R. and I. Dagan, 1995, "Knowledge Discovery in Textual Databases (KDT)", Proceedings of the First International Conference on Knowledge Discovery & Data Mining, pp. 112-117

[7] Feldman, R., W. Klosgen, and A. Zilberstein, 1997a, "Visualization Techniques to explore Data Minin Results for Document Collections", Proceedings of the Third International Conference on Knowledge D s-covery & Data Mining, pp.16-23.

[8] Feldman, R., W. Klösgen, Y. Ben-Yehuda, G. Kedar and V. Reznikov, 1997b, "Pattern Based Browsing in Document Collections", Proceedings of First European Symposium on Principles of Data Mining and Knowledge Discovery, pp.112-122

[9] Han,J., Y.Gai and N.Cercone, 1993, " Data-driven discovery of quantitative rules in relation databases", IEEE Tran. On Knowledge and Data Engineering ", pp.29-40

[10] Kryszkiewicz, M, 1998, "Representative association

rules and minimum condition maximum consequence association rules", Proceedings of the Second European Symposium on Principles of Data Mining and Knowledge Discovery,pp.361 -369.

[11] Li, B.-I., et. al., 1991, "A maximal matching automati Chinese word Segmentation algorithm us ing corpus tagging for ambiguity resolution", R.O.C. Computational Linguistics Conference, Taiwan, pp. 135-146

[12] Nie, J., M. Briscbois and X. Ren, 1996, "On Chinese Text Retrieval", Conf. Proc. of SIGI , pp. 225-233

[13] Singh, L.,P. Scheuermann and B. Chen, 199 7, "Generating Association Rules from Semi-Structured Documents Using an Extended Concept Hierarchy", ACM IKM, pp.193-200

[14] Sparck Jones, K., 1972, "A static interpretation of term specificity and its application in retrieval", Journal of Document, Vol. 28, No. 1, pp. 11-20

[15] 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

[16] Wuthrich, B., V.Cho, S. Leung, D. Permunetilleke, K Sankaran, J. Zhang and W. Lam, 1998, "Daily Stock Market Forecast from Textual Web Data", IEEE Int'l Conf. on SMC

[17] 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

[18] 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

[19] 胡勝傑、許中川，1999，「中文新聞文件斷詞」，第十屆國際資訊管理學術研討會論文集，pp.968-974.