第二章　資料處理與資料探勘理論

就財經股市交易領域而言，影響股市交易活動會受到多個層面的影響，例如：基本面、籌碼面、消息面、政治面、技術面等[45]。由於有這些不確定的因素存在，進而使得預測股市的交易是較為不易的。其中消息面（財經新聞事件）往往蘊含著豐富的訊息[15]，甚至會影響股市的交易活動。數值面的資訊通常只是能夠知道當天股票交易情形與漲跌狀態，但並無法得知造成這些狀態的真正原因，而這些原因則是可以從新聞事件當中來發現的。若是從中探勘股市交易與新聞事件的關聯性，這將會是一個可探討的研究方向。歸納過去相關文獻中發現，僅有極少數的研究是在做有關新聞處理與交易策略整合的研究[15]。

2.1 股市投資概說

一般而言，影響股市股價的因素包括基本面、籌碼面、消息面、政治面、技術面等[45]，如圖 2-1 所示。若是以目前網際網路上的資訊而言，這些因素可以從不同的網站中獲得。例如：新聞事件的來源可以是各大入口網站或報社電台網站等，股市交易的來源可由各大證券商網站或臺灣證券交易所[44] 等，尤其是各大證券商網站可以同時獲得技術面與消息面的這兩種訊息。近年來由於網際網路的進步，消息面也可以即時的從網際網路上取得。將這些訊息有效的整合並且提供分析與判斷的結果，也是本研究所要探討的主要目的。

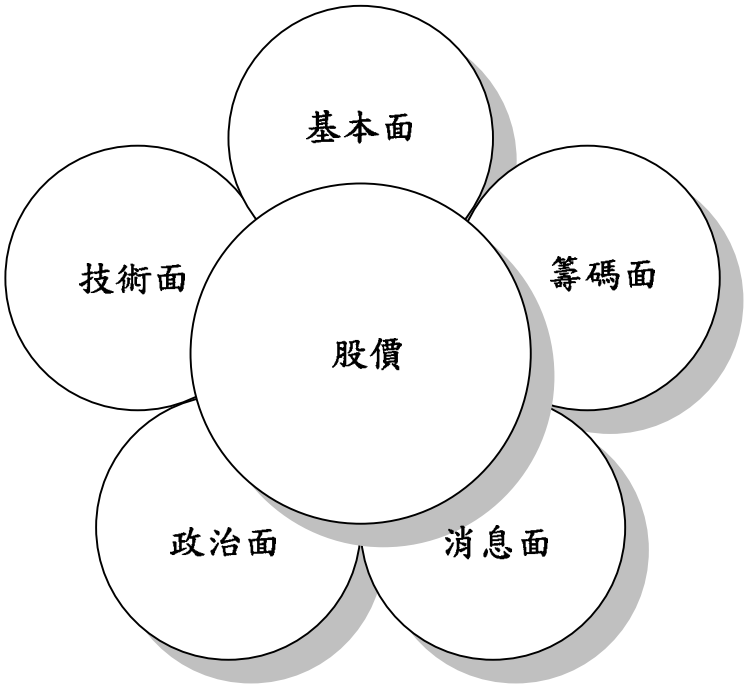


圖2-1 影響股價漲跌因素

以消息面而言，股市投資裡面有所謂的「年輪理論」[46]，在學理上年輪理論指的是資訊獲得的不對稱性，如圖 2-2 所示。位於最核心的人（大老闆與大股東）是最先得到消息面的訊息並且往外擴散出去，所以最外層的人（散戶投資人）永遠是最後得到訊息的。台灣證券市場的組成結構主要是由三大法人（外資、投信、自營商）和散戶所組成，其中又以散戶佔絕大多數。一般的投資散戶除了缺乏專業分析能力之外，並且在取得資訊方面較為不利，若是沒有一套完善的交易策略的話，往往成為最後的犧牲者。因此本研究擬提出「基於關聯式規則在影響個股漲跌之新聞事件探勘以臺灣股市為例之應用研究」的研究，結合資訊擷取、資訊處理以及資訊探勘等不同層面的技術，藉由資訊探勘之關聯式規則方法中的支持度（Support Level）與信賴度（Confidence Level）兩種判斷條件，找出股市消息面與股市交易活動的隱含關聯規則。

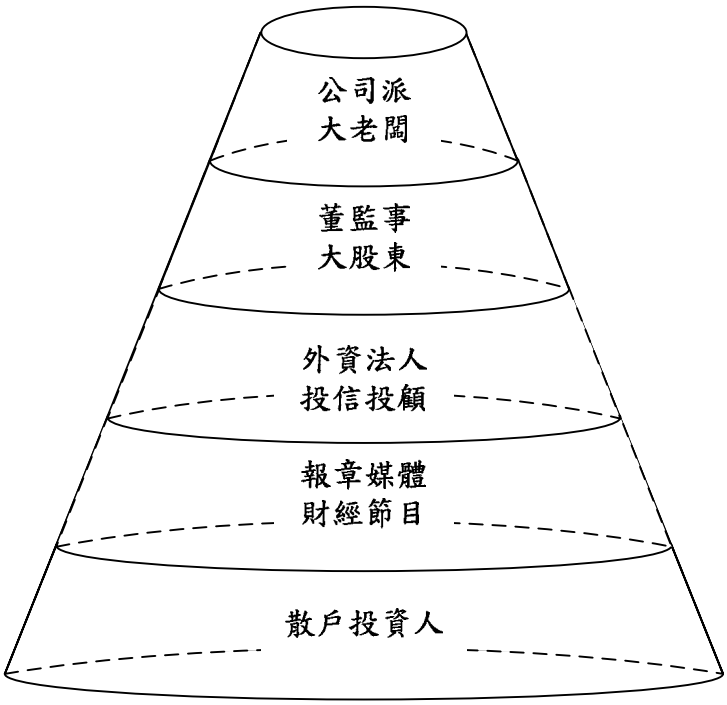


圖2-2 年輪理論

2.2 資料擷取

由於今時今日是一個資訊爆炸的年代，不管是報紙、廣播、電視或是網際網路，人們可以隨時隨地獲得所需的訊息。因此，一些學者針對股市交易的這方面已有一些相關的論文研究被提出以及發表，像是 Wuthrich 及 Cho 等人[21]-[23]利用文件探勘的方式來預測股市漲跌，該研究經由蒐集各大金融報紙網站上的文件資料，透過統計與金融有關的關鍵詞彙並給予權值，以權值和收盤價的關係推導股市漲跌與關鍵詞彙之間的關聯性，進一步預測股市收盤的指數。Kuo 及 Chen 等人[24]的研究則提出利用基因演算法為基礎的模糊類神經網路，整合技術指標與新聞消息面因素於股市投資的決策。而 Fung 及 Yu 等人[25]的研究是提出利用當時的新聞文件探勘來預測股票未來的走勢。Mittermayer[26]則發展一個新聞分類及交易系統，利用即時的新聞事件作股票價格漲跌的預測。Peramunetilleke 及 Wong[27]提出利用政治及財經新聞標題預測匯率的走勢。Cheung 及 Huang 等人[28]發展一個財經知識管理系統，可以同時處理中文及英文新聞，並

且擷取出其新聞的詮釋資料及新聞本文的關鍵字詞，再利用非監督式學習法則將新聞事件作事件歸類。綜合這些研究可知，在[15][29]所提出關於新聞事件影響股市投資交易的研究方面，利用探勘新聞事件與股市交易行為之間的關聯研究，從財經觀點上已經是一個相當重要的研究議題。然而這兩者之間的關聯是難以用簡單量化的方法來探勘出來的，並且沒有充分的股市交易資訊以及新聞事件的數量，這將會是難以進行更深入的資訊探勘與分析的。由此可知，股市交易資訊與個股新聞事件的蒐集是格外重要的，也是本研究所要努力的方向之一。

在新聞事件與交易資訊蒐集方面，必須透過一套有效的擷取系統來獲得所需的相關資訊。目前國立花蓮師範學院的學習科技研究所已有這方面的相關研究，其中智慧型網路學習系統實驗室[48]已經成功發展出一套〝股市新聞擷取與典藏系統〞，如圖 2-3 所示。

圖2-3 股市新聞擷取與典藏系統

此一系統能夠透過〝系統排程〞與〝擷取代理人系統〞的機制，完整

地透過網際網路向特定的網站蒐集所需的相關資料，以提供相關的資料探勘研究使用。然而這些典藏的資訊來源可分為兩個部分，一為個股新聞事件，另一為股市交易資訊。在個股新聞事件方面，目前系統已完整典藏在鉅亨網[41]、Google News[42]及公開資訊觀測站[43]上的個股新聞事件，如圖 2-4 所示。然而本研究所需的新聞資料來源必需要具備有〝深入性〞與〝全面性〞的特質，因此選定 Google News 與鉅亨網這兩個網站作為本研究不可或缺的資料來源。

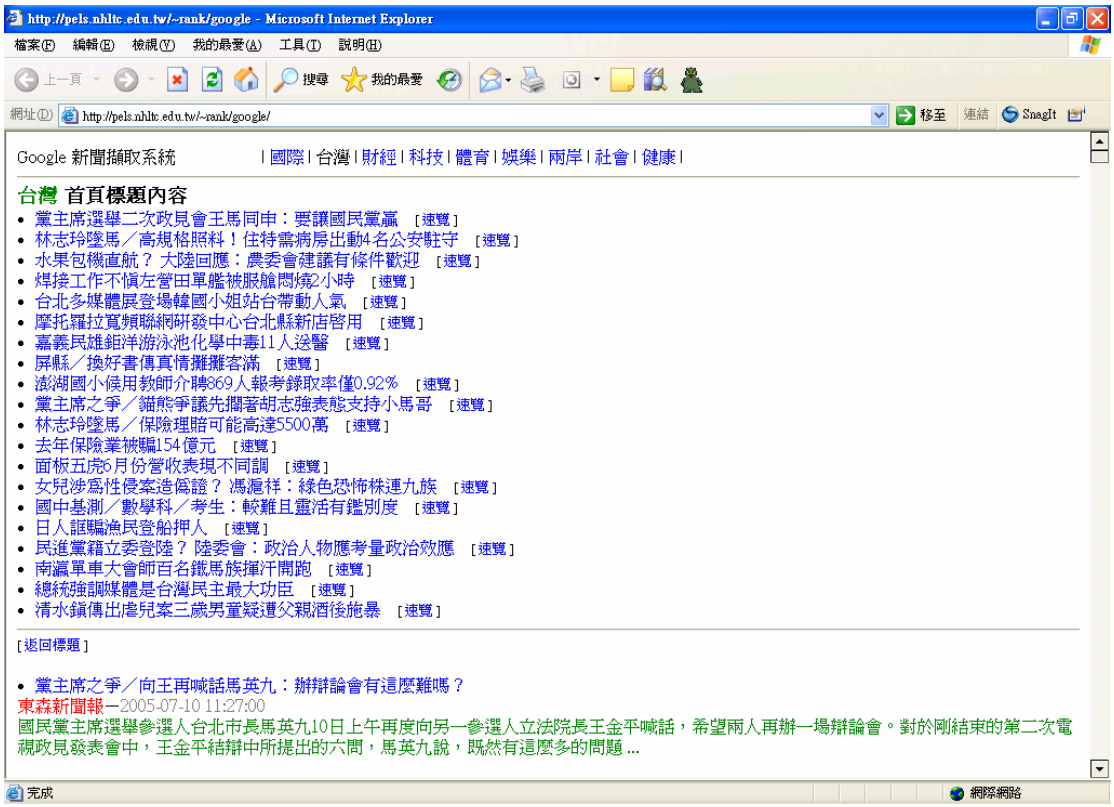
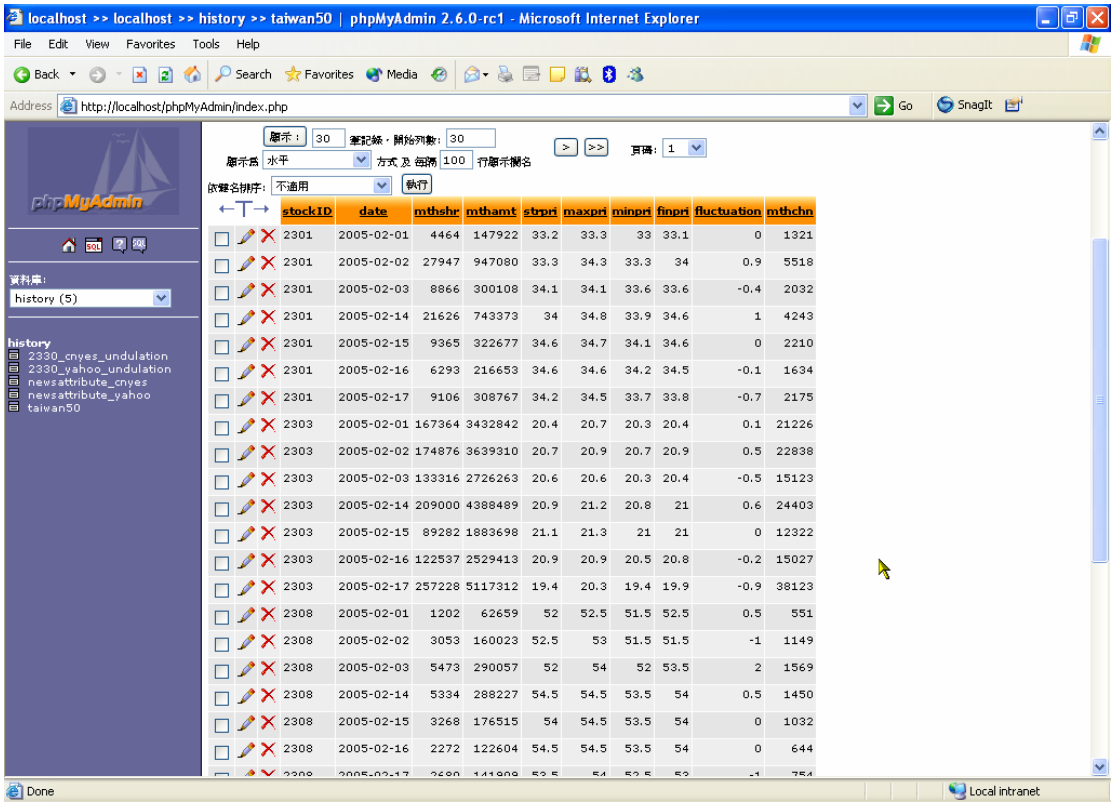


圖2-4 新聞擷取系統

股市交易資訊方面，則是透過臺灣證券交易所[44]提供的盤後資訊下載系統，經由系統排程與擷取代理人的步驟，每日擷取當天的各項交易資訊，例如：成交價、買賣成交張數與漲跌幅度等等，並且將這些資訊予以儲存至資料庫之中，如圖 2-5 所示。台灣加權股價指數漲跌之消息面因素與新聞事件探勘的研究，必須整合財經消息面及數值面資訊，這些財經新聞資訊來源，有助於我們探勘出影響股市漲跌之消息面因素。



stockID	date	mthshr	mthamt	strpri	maxpri	minpri	finpri	fluctuation	mthchn
2301	2005-02-01	4464	147922	33.2	33.3	33	33.1	0	1321
2301	2005-02-02	27947	947080	33.3	34.3	33.3	34	0.9	5518
2301	2005-02-03	8866	300108	34.1	34.1	33.6	33.6	-0.4	2032
2301	2005-02-14	21626	743373	34	34.8	33.9	34.6	1	4243
2301	2005-02-15	9365	322677	34.6	34.7	34.1	34.6	0	2210
2301	2005-02-16	6293	216653	34.6	34.6	34.2	34.5	-0.1	1634
2301	2005-02-17	9106	308767	34.2	34.5	33.7	33.8	-0.7	2175
2303	2005-02-01	167364	3432842	20.4	20.7	20.3	20.4	0.1	21226
2303	2005-02-02	174876	3639310	20.7	20.9	20.7	20.9	0.5	22838
2303	2005-02-03	133316	2726263	20.6	20.6	20.3	20.4	-0.5	15123
2303	2005-02-14	209000	4388489	20.9	21.2	20.8	21	0.6	24403
2303	2005-02-15	89282	1883698	21.1	21.3	21	21	0	12322
2303	2005-02-16	122537	2529413	20.9	20.9	20.5	20.8	-0.2	15027
2303	2005-02-17	257228	5117312	19.4	20.3	19.4	19.9	-0.9	38123
2308	2005-02-01	1202	62659	52	52.5	51.5	52.5	0.5	551
2308	2005-02-02	3053	160023	52.5	53	51.5	51.5	-1	1149
2308	2005-02-03	5473	290057	52	54	52	53.5	2	1569
2308	2005-02-14	5334	288227	54.5	54.5	53.5	54	0.5	1450
2308	2005-02-15	3268	176515	54	54.5	53.5	54	0	1032
2308	2005-02-16	2272	122604	54.5	54.5	53.5	54	0	644

圖2-5 每日盤後股市交易資訊

2.3 資料處理

對於所收集的新聞事件必須針對個別事件做歸類，然而有關新聞事件處理的方式多偏向以詞彙的判斷來進行，而且只有為數極少的研究是關於將財經數值面與消息面整合作投資決策。因此，必須要針對新聞事件來進行分類，並且選擇適當的處理機制來完成此項目的。目前有關新聞處理方面的研究，有 Shah 及 Elbahesh[16]提出利用相似測度將新聞文件依據主題作聚類的研究；Maria 及 Silva[17]提出利用 Support Vector Machine 作新聞主題的分類研究；Kurtz 及 Mostafa[18]提出能依據個人日誌檔，利用詞彙的比對及聚類方法來追蹤使用者有興趣的新聞主題；Lam 及 Cheung 等人[19]利用歸納及探勘網路線上新聞，提出一套 Matching Algorithm，能夠將未看過的英文詞彙自動翻譯成對應的中文；Allan 及 Papka 等人[20]則提出利用相似測度來追蹤特定新聞事件的發展歷程。

本研究欲採用中文斷詞的方式來產生出各項新聞事件標題的關鍵字詞，由於中文句子是由一連串的字元所組合而成，不同的字元組合會有不同的表示意義。中文句子並無英文句子有明顯的字元區隔符號（空白）符號，而且句子中相同的字元也會代表不同的詞性，所以對中文句子（文章）作自動斷詞是極為困難的事。幸而海峽兩岸目前都有一群人正在進行研究，期望解決、降低中文自動斷詞的困難度。除了這方面的研究本來就並不多，而且鮮少有針對中文介面的新聞事件作斷詞處理。

新聞事件的處理多偏向於出現詞彙方式來判斷事件發生的趨勢，因此分析詞彙則極為重要。目前中央研究院資訊科學研究所發展的中文斷詞及詞性標註系統 CKIP[47]作為自然語言的處理工具，目前該系統為開放式介面 CGI（Common Gateway Interface）提供中文斷詞及詞性標註的服務，如圖 2-6 所示。一般斷詞系統是利用詞典中收錄的字詞來比對，找出句子中可能的字詞。以往這些詞典是由人工的方式來定義、產生出來的，所以詞典的產生極為耗時且困難，因此開發出一個有效率且精確的自動字詞判斷系統也是他們研究努力的目標。斷詞系統可分為兩個部份，一為已知詞，另一為未知詞。已知詞是指已經過判斷的字詞，而未知詞是指為尚未分析判斷的部份，而中文自動斷詞的真正困難處即是判別未知詞的部分。

當一篇財經新聞經過 CKIP 中文斷詞處理過後，除了可以將中文正確斷詞之外，還可以分析出中文字詞的詞性。因此，本研究將所擷取的股市新聞事件之標題利用網際網路程式，經由系統排程定時將資料庫之中的資料送至 CKIP 中文斷詞系統進行斷詞分析，並且將所獲得的斷詞結果儲存回到資料庫之中，以提供新聞事件歸類與資料探勘的需求。例如圖 2-6 中已經輸入一段文字到 CKIP 中文斷詞系統，並且經由系統運算後之結果如圖 2-7 所示，可以發現該中文斷詞系統已經可以將股市新聞事件有效的區分每一個字詞與詞性，而所代表的詞性可由表 2-1 得知其代表的意義。因此，將新聞事件的標題送至 CKIP 中文斷詞系統所產生的字詞，是有助於

後續資料探勘的重要運算過程，也符合本研究進行實驗前的條件之一。

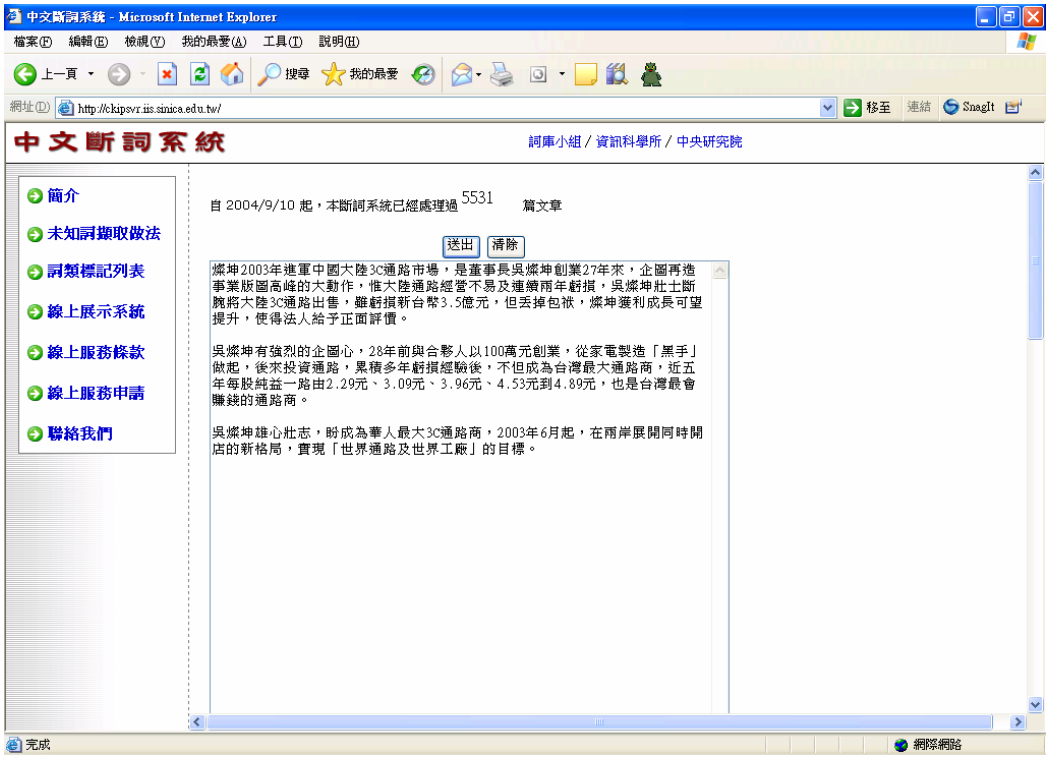


圖2-6 CKIP 中文斷詞系統

圖2-7 中文斷詞系統運算結果

表2-1 CKIP 標記詞性對照表

CKIP 標記	對應之詞類標記	CKIP 標記	對應之詞類標記
A	非謂形容詞	Ng	後置詞
Caa	對等連接詞(如：和、跟)	Nh	代名詞
Cab	連接詞(如：等等)	I	感嘆詞
Cba	連接詞(如：的話)	P	介詞
Cbb	關聯連接詞	T	語助詞
Da	數量副詞	VA	動作不及物動詞
Dfa	動詞前程度副詞	VAC	動作使動動詞
Dfb	動詞後程度副詞	VB	動作類及物動詞
Di	時態標記	VC	動作及物動詞
Dk	句副詞	VCL	動作接地方賓語動詞
D	副詞	VD	雙賓動詞
Na	普通名詞	VE	動作句賓動詞
Nb	專有名稱	VF	動作謂賓動詞
Nc	地方詞	VG	分類動詞
Ncd	位置詞	VH	狀態不及物動詞
Nd	時間詞	VHC	狀態使動動詞/
Neu	數詞定詞.	VI	狀態類及物動詞
Nes	特指定詞	VJ	狀態及物動詞
Nep	指代定詞	VK	狀態句賓動詞
Neqa	數量定詞	VL	狀態謂賓動詞
Neqb	後置數量定詞	V_2	有
Nf	量詞		

2.4 資料探勘

本研究除了資訊擷取與資料處理外，最重要的是資料探勘部分。在資料探勘方面目前已有許多研究被發表，但是運用於中文新聞探勘或股市交易資訊探勘上的研究卻是少數。目前有關資料探勘的研究有灰色關聯分析方法[33][34]、加權關聯式規則法[34][35]、決策樹或是類神經網路方法[30]-[32]。本研究採用資料探勘之中的關聯式規則（Association Rules）作為資料探勘的核心方法，利用中文斷詞系統 CKIP 對各則新聞事件標題來產生相對的關鍵字詞，將這些關鍵字詞由相似度鑑別可使新聞事件歸類並設定事件標籤。藉由關聯式規則將事件標籤作運算並產生相對映新聞事件標籤的支持度（Support Level）與信賴度（Confidence Level），依據極大項目集合（Large Itemsets）與這兩個參考判斷條件即可以找到當中的隱含關聯規則。

探勘買賣交易與新聞事件項目之間的關係而形成關聯式規則（Association Rules），可以表達出項目之間的關聯性。關聯式規則以 $X \rightarrow Y$ 表示，其中 X 、Y 為項目組（Itemsets），且 $X \cap Y = \phi$。關聯規則是否成立，必須依據是否滿足最小支持度（Minimum Support Level）與最小信賴度（Minimum Confidence Level）而定。支持度定義為：在所有交易中出現 $X \cup Y$ 的機率；而信賴度定義為：在所有出現 X 的交易中也出現 $X \cup Y$ 的機率。對關聯式規則設定一個門檻值（Threshold）來作為判斷之條件，對於符合門檻值的項目集合，即可稱為該類別之中具有代表象徵的類別事件，也就是所要尋找的資料規則，如圖 2-8 所示。

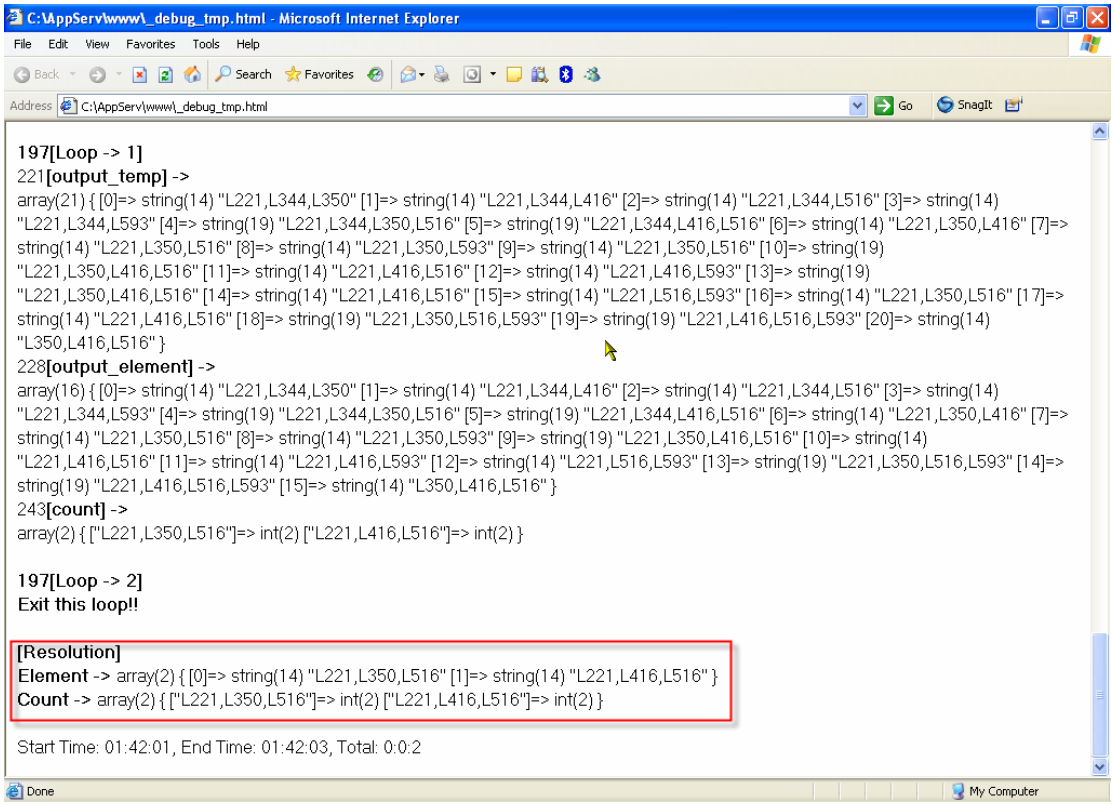


圖2-8 關聯式規則運算結果

由圖 2-8 可發現，當關聯式規則運算執行後，會對每次運算所獲得的關聯規則進行條件判斷。如果條件判斷的結果是符合條件門檻值時，則會將此次所得之關聯規則作為資料探勘的結果(如 L221, L350, L516 與 L221, L416, L516)，反之則將會繼續往下執行直到找出符合條件的結果。