國立政治大學資訊管理學系

碩士學位論文

指導教授：楊建民博士

整合文件探勘與類神經網路預測模型之研究
-以財經事件線索預測台灣股市為例

研究生：歐智民

中華民國一百年七月

誌謝

在政大的兩年研究過程中，首先要感謝楊建民老師的諄諄教誨，讓我們從興趣中探索研究方向，言談中更是學習到待人處事的道理，在在讓我對老師產生敬佩及感恩的心意；也感謝邱光輝老師、季延平老師及林我聰老師在提報論文與口試期間給予的指導及建議，使得架構及內容能夠更加嚴謹、詳細，有助於提升論文整體品質。

此外，感謝在政大資管所認識的朋友們，同儕間互相切磋，讓我能在忙碌中快速成長，而研究室的夥伴們，柏均、振和、漢瑞及章威，互相鼓勵、討論，度過同甘共苦的過程，能一同完成論文更是兩年來最感動的時刻，也感謝取向這兩年來一路的幫忙。

生命中的好朋友們，孟婷、依玟、瑋之和謹仔，我們在不同領域努力生活著，但因為有你們的陪伴，讓我不斷地提醒自己要一起為了夢想努力，成為一個很棒的人。而家人這兩年的鼓勵及支持，讓我能毫無顧慮的專心研究，感謝你們，讓我在過程疲憊的時候總是能夠得到前進的力量。

最後，謹將這份榮耀獻給所有關心我及我關心的人。

摘要

隨著全球化與資訊科技之進步，大幅加快媒體傳播訊息之速度，使得與股票市場相關之新聞事件，無論在產量、產出頻率上，都較以往增加，進而對股票市場造成影響。現今投資者多已具備傳統的投資概念、觀察總體經濟之趨勢與指標、分析漲跌之圖表用以預測股票收盤價；除此之外，從大量新聞資料中，找出關鍵輔助投資之新聞事件更是需要培養的能力，而此正是投資者較為不熟悉的部分，故希望透過本文加以探討之。

本研究使用 2009 年自由時報電子報之財經新聞（共 5767 篇）為資料來源，以文件距離為基礎之 kNN 技術分群，並採用時間區間之概念，用以增進分群之時效性；而分群之結果，再透過類別詞庫分類為正向、持平及負向新聞事件，與股票市場之量化資料，包括成交量、收盤價及 3 日收盤價，一併輸入於倒傳遞類神經網路之預測模型。自台灣經濟新報中取得半導體類股之交易資訊，將其分成訓練及測試資料，各包含 168 個及 83 個交易日，經由網路之迭代學習過程建立預測模型，並與原預測模型進行比較。

由研究結果中，首先，類別詞庫可透過股票收盤價報酬率及篩選字詞出現頻率的方式建立，使投資者能透藉由分群與分類降低新聞文件的資訊量；其次，於倒傳遞類神經網路預測模型中加入分類後的新聞事件，依統計顯著性檢定，在顯著水準為 95%及 99%下，皆顯著改善隔日股票收盤價之預測方向正確性與準確率，換言之，於預測模型中加入新聞事件，有助於預測隔日收盤價。最後，本研究並指出一些未來研究方向。

關鍵字：事件偵測與追蹤、kNN 分群、倒傳遞類神經網路預測模型

Abstract

With the development of the Internet rapidly, retrieving correct information from lots of sources for investors is very important today. In this study, we want to explore whether unstructured information like news' events will affect Taiwan stock market. Therefore, we used cluster technology to group finance news as news' events, and further classify as "positive, flat or negative" events; then combine transaction information in TEJ's database to build a BPN prediction model.

From the results, in the one hand, we found that investors could use clustering and classification technologies to abstract information. On the other hand, the predict model with news' events will improve the accuracy of the direction and the precision of the prediction. In other words, the new model is more significant than one without news' events.

Keywords: event detection, event tracking, kNN clustering, BPN prediction model

目錄

圖目錄

表目錄

公式目錄

第一章 緒論

第一節 研究背景與動機

當前投資市場所流通的企業併購與轉投資訊息，皆可能對投資決定造成重大影響，足見資訊已成為評估投資風險的重要管道。因此，對於金融業者與投資者而言，擴充蒐集資訊的數量，並從快速更新、種類繁多的各種資訊中，分辨正確的投資脈動，已成為投資者作為審慎判斷進入市場的要件，以上在在說明了由於台灣市場規模較小，對於資訊的接收往往使股票市場隨之造成影響，這也是典型「淺碟性市場」的特性之一（李春淋，2010；吳真蕙，2000；林章德，2000），總的來說，掌握資訊對於台灣投資者而言是一重要的投資利器。

面對股票市場的波動起伏，投資者除了透過總體經濟趨勢及對公司營運狀況做分析之外，仍需要累積敏銳的觀察力，運用正確的投資策略，及選擇適當時機進入或退出市場。而隨著資訊科技的進步，無論在空間或時間層次，訊息傳播的速度都較以往快速，人們可隨時隨地取得四面八方的消息。然而，繁雜的訊息卻可能阻礙投資者判斷正確的資訊來源，更添增找尋其所需資料的時間。因此，在各個領域中，與股票市場相關之研究日漸受到關注，學術著作也有可觀產出，其所發展的方向有三，一為結構化資訊，由總體經濟的趨勢建構具有經濟意義的模型，如時間序列模型（楊踐為、李家豪、類惠貞，2007）等，以觀察未來趨勢；二多從非結構角度觀察，在資訊領域中透過人工智慧（Artificial Intelligence, AI）技術結合財金重要指標以發掘隱藏的資訊（Nygren, 2004；陳稼興、楊孟龍，2000；黃馨瑩、楊建民、李耀中，2009）；三則將上述兩者合併之研究（Armano, 2005）。針對後兩者之方向，常用的工具如類神經網路預測模型，原因在於其具有學習、聯想、歸納推演等能力，使得非結構化之資訊能透過訓練得到適當的預測模型，

雖然訓練過程易於落入區域解問題，然而透過演算法（如基因演算法等）之全域搜尋法可避免此問題（林聖哲，2002）。

由於新聞具有時效性、區域性的特性，能適時揭露重要資訊給地區民眾，而目前網路新聞事件的分類多以人工判斷為主，投資者能直覺地從財經新聞中選擇相關的新聞，當作觀察股票市場變化的工具之一（Ahmad et al., 2002）。事實上，並非所有相關的新聞皆受到投資者的關注，而是部分與投資市場相關之事件，投資者會從中發掘出隱含的訊息以幫助決策。Khurshid（2002）的研究認為，影響財務市場的資訊，不論這些資訊的來源形式為何，新聞內所隱含的資訊對股價的影響扮演重要的中介因素。

新聞事件偵測與追蹤的領域，以卡內基美隆大學（Carnegie Mellon University, CMU）與麻州大學（University of Massachusetts, UMass）最為著名（古倫維，2000；戴尚學，2003），而 CMU 以 kNN（k-Nearest Neighbor）分類器之概念實現新聞事件偵測與追蹤的技術，並加入時間區間（Time Window）的概念，以符合新聞具備的時效性，使分群效果更完整。

近年來，從新聞文件預測股票趨勢之研究相當熱門，部分學者透過文字探勘將新聞文件分類，進而對股票趨勢做對應，用以幫助預測（喻欣凱，2008），亦有學者嘗試將文字探勘與人工智慧結合，從文件中擷取新聞文件特徵字後，以訓練的方式建構預測模型，所建立的模型則以類神經網路為大宗（黃馨瑩、楊建民、李耀中，2009；Nygren, 2004；Kim & Han, 2001；Kim & Han, 2000），然而新聞往往因事件大小造成重複發佈的機會，使重要訊息分散於各新聞文件中，降低了「新聞事件」之焦點。

故本研究期望能以「新聞事件偵測與追蹤」的技術為出發點，透過類別詞庫的建立，探討新聞事件對於市場股價之影響，期望能將新聞文件以新聞事件分類的方式，與類神經網路預測模型做結合，進而提供投資者有效的投資線索及資訊。

第二節 研究目的

依上述之背景與動機，本研究將針對以下二點做為研究目的：

1. 以每日報酬率為基礎，透過詞彙於一篇文章中所出現次數（Term Frequency, TF，以下簡稱為詞頻）及權重值之篩選建立新聞類別詞庫，使新聞事件能根據類別詞庫判斷屬於哪一分類，透過類別詞庫的建立，也能給予投資者在預測股票市場上的觀察指標。
2. 於倒傳遞類神經網路預測模型中加入非結構化之資訊，使預測模型顯著提升預測方向正確性與預測準確率。

第二章 文獻探討

第一節 新聞與股價之相關性研究

Khurshid 等學者（2002）認為無論文字消息的形式為何，皆可能為影響金融市場的波動，換言之，文字消息為傳遞事件的一種方式，而投資者可透過文字消息如新聞文件觀察並加以評估投資的最佳時機。以新聞文件為例，媒體可根據事件大小評估該新聞事件之重要程度發佈大大小小的新聞文件，研究亦指出，新聞量與股價趨勢具有正向影響（黃馨瑩、楊建民、李耀中，2009），因此資訊的傳遞上實際是由事件所揭露，而新聞文件僅為傳遞資訊的方式之一（黃馨瑩、楊建民、李耀中，2009；Khurshid et al., 2002）。

Lavrenko 等學者（2000）採用分段線性配對（Piecewise Linear Fitting）觀察股價漲跌趨勢，並將高度相關的文件做連結，訓練出 Language Model，模型也證實財經新聞與股價趨勢具有相關性，且能用來有效預測股價。

Chen 等學者（2003）採用 PNN 訓練歷史資料，將模型用來預測指數報酬率的方向，引導投資者的交易策略。研究顯示以 PNN 為基礎投資策略能比其他投資策略獲得更高的報酬率。

Mittermayer（2004）應用支援向量機（Support Vector Machine, SVM）將新聞分成正向新聞、無影響新聞及負向新聞三類，並實現於其所提出的系統 NewsCATS（News Categorization and Trading System），用來預測新聞發布後 60 分鐘之 NMS（National Mittermayer System）股票指數趨勢。結果顯示以此系統交易的平均獲利大於隨機投資策略，因此認為新聞分類能幫助提供更多資訊以進行股價趨勢的預測（Mittermayer, 2004；吳昀錚，2008）。

吳昀錚(2008)以線上財經新聞分類為基礎,決定投資者的短期之投資策略，並以台灣股票加權指數評估系統之績效。實驗結果顯示，由投資策略所獲得之報酬率可勝過銀行定存利率，具有參考價值。

第二節 探勘技術

2.1. 資料探勘

隨著科技的進步，資料產生的速度亦突飛猛進，資料量隨著時間大量成長，從大量資料中萃取有意義的特徵或規則，並集結成有用的知識，成了各領域期望能獲得的分析能力，而資料探勘即為萃取資訊的技術之一。許多人認為「資料探勘」和「從資料中發掘知識」(Knowledge Discovery in Database, KDD)是同義的，然而，資料探勘僅為 KDD 的重要程序之一，研究亦顯示兩者之間有著相輔相成的關係（Fayyed, 1996；Han & Kamble, 2001；黃孝文，2010）。

Han & Kamber（2001）提出 KDD 的程序可分成四個步驟，如圖 2.2.1：

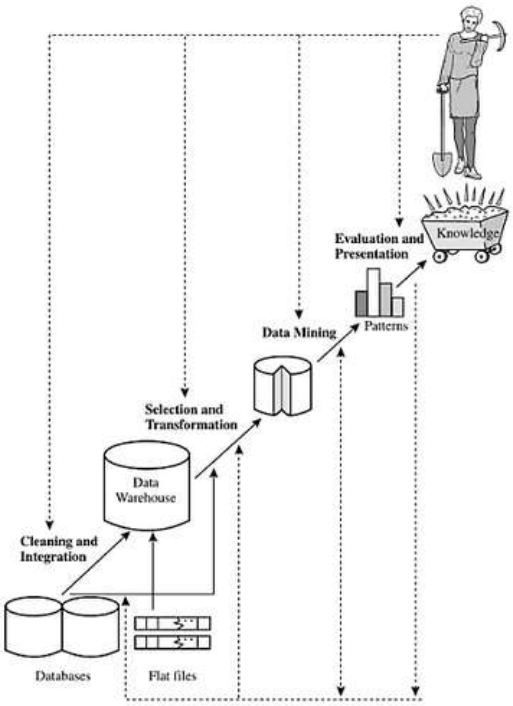

圖 2.2.1 資料探勘為 KDD 的程序之一

資料來源：Han & Kamber，2001

1. 過濾與整合：欲對龐大的資料庫進行知識萃取前，由於資料可能有錯誤、不完整、遺失或重複的狀況，因此必須先過濾資料中的雜訊或整合同義資料，使資料成為有意義的訊息，並放置資料倉儲中。
2. 選擇與轉換:從資料倉儲中選擇特定的知識領域後,資料量仍可能相當龐大,因此需要適當的將資料簡化、轉換成適當的格式，使後續的工作能順利進行。
3. 資料探勘：此為 KDD 之重要過程，透過關聯規則、分類預測、分群分析等演算法，分析並挖掘資料中隱藏的規則。
4. 評估與解釋：為了檢驗前一步驟所發現的規則是合理的，需要對其作出合理的評估或解釋，此結果可透過簡易的圖表呈現，讓使用者能評估是否可成為決策分析的依據。

資料探勘的過程中，使用者可以根據資料的類型與範圍，選擇適當的演算法做相關分析，因此資料探勘所採用的演算法成為知識挖掘的關鍵因素。常見的演算法包括關連規則分析、分類分析及分群分析（Han & Kamber，2001；黃孝文，2010）。

1. 關連規則分析（Association Analysis）

此演算法以統計機率為基礎，從大量的資料中發掘出，在某一規則下兩種不同類型的項目經常共同出現之現象。商業上透過市場決策分析，了解客戶購買商品的隱性規則，經過有效的推測後，有助於主管執行有效的策略決定。

2. 分類分析（Classification Analysis）

資料分類包括兩步驟，首先要先透過分類演算法訓練資料，得到分類的規則，通常規則形式為「IF …, THEN …」，然後分類器經過訓練後，才能透過分類規則預測測試資料所屬的分類。分類的技術包含簡單貝氏分類（Naïve Bayes Classification）、kNN（k-Nearest Neighbor）、支援向量機（Support Vector Machine, SVM）等。而 Joachims（1998）將此三種分類器與最小平方誤差法（LLSF）及類神經分類（ANN）以統計方法比較效率與分類結果，優異程度為：{kNN、SVM} ＞LLSF＞ANN＞NB。

3. 分群分析（Clustering Analysis）

透過觀察將大量資料分割、分群，使群集內資料的相似度提高，而群集間的相似度降低，分群以統計的基礎對資料做分析，由於分群時目標值並不存在，屬於非監督式學習。Han & Kamber（2006）將分群法分成五大類：分隔式分群（Partitioned）、階層式分群（Hierarchical）、密度基礎分群（Density-based）、網格式分群（Grid-based）與類神經網路分群（Neural network），尤以分隔式分群之 K-means 最為常見（胡舜禹，2009），而 kNN 分類器之概念亦被用於分群技術上（戴尚學，2003）。

2.2. 文字探勘

文字探勘屬於資料探勘中的一重要分支，透過觀察文件中文字、段落、主題等關聯，期待能從中尋找文件趨勢，甚至進一步進行預測（Han & Kamber, 2001）。

袁立安（2007）將文字探勘分成三個步驟：文件準備、文件處理與文件分析。文件準備階段需對文件做前處理，並萃取有用的關鍵字詞；文件處理步驟使用資料探勘技術發掘文件中有意義、有趣的型態；最後，文件分析進行輸出結果的驗證以確定所擷取之知識是否有用，下列將整理各步驟之使用方式。

2.2.1. 文件準備

1. 斷詞

中文文件是由字與標點符號以非結構化的方式所組成，然而字卻未必能成為有意義的單位，因此，在處理中文文件前必須採取斷詞的動作，使字能以有意義的詞彙之方式呈現。研究顯示，斷詞的方式大致分成詞庫斷詞法、N-Gram 選詞法及混合斷詞法三種（顧皓光，1996；戴尚學，2003），其優缺點將列於表 2.2.1。

表 2.2.1 斷詞方式之比較

名稱	方式
詞庫斷詞法	以既有詞庫做為標準，將文件以比對的方式找出斷詞。比對的效果與詞庫完整程度成正比。
N-Gram 選詞法	依照所選取的字數長短，計算字所出現的數量，以統計分析判斷是否為有意義的詞，再做出適當斷詞。優點是不需要知道詞彙的意義，亦不被詞庫限制，但斷詞過程中不易發現不合適的詞彙。
混合斷詞法	先對文件做詞庫比對，再對無法比對的字詞做統計斷詞，找出可能的斷詞位置，此方法擷取詞庫與統計斷詞的優點，然而仍需對詞庫做維護才能有效提升斷詞結果的品質。

資料來源：顧皓光，1996；戴尚學，2003

2. 特徵值選擇

文件處理時，為了使效率增加，減少計算複雜度，往往會先移除文件中不具代表性的詞彙，找出特徵值（Liu & Motoda, 1998）。常見的特徵值的選擇方式包括：文件頻率（Document Frequency）挑選出現於文件數量較高的字詞，將其當成類別的特徵值；資訊增益量（Information Gain）以字詞在各類別中出現及不出現的機率判斷字詞重要性；交互資訊量（Mutual Information）著重於詞彙間共同出現的程度；卡方統計量（χ^2-Statistic）以卡方統計量檢定字詞與類別間的相關程度；詞彙強度（Term Strength）在訓練集內利用餘弦找出相似度到達門檻的文件，再以條件機率計算兩詞彙的關聯強度。其中，以資訊增益量及卡方統計量成效較佳（Yang & Pedersen, 1997；Aas & Eikvil, 1999）。

3. 權重值計算

縱然文件準備後能得到斷詞結果，卻難以評斷哪些字詞具有文件代表性，而權重值計算則是為了此目的產生。根據 Popescu（2001）研究整理，權重值可由三個部分組成，包括區域權重（L_{ij}）、全域權重（G_i）及文件之正規化因子（N_j），區域權重以字詞於「特定文件」中出現的頻率為基礎，而全域權重則以字詞於「所有文件」中出現的頻率為基礎，正規化因子則是為了讓不同字詞的權重得以比較而產生。

舉例來說，Jing 等學者（2002）研究中權重值之區域函數為字詞於一篇文章中出現的次數，全域函數為字詞於所有文章出現次數之倒數值，並給予對數以做調整，最後則採用餘弦正規化。此方法即為 TFIDF（Term Frequency–Inverse Document Frequency），也是常被使用的權重計算方式之一。

2.2.2. 文件處理

1. 向量空間模型

向量空間模型是目前資訊檢索中效果較好的方式（Salton, 1988），也是目前最廣泛使用的資訊檢索模型（戴尚學，2003）。每篇文件以一組向量表示，維度代表的是關鍵字詞，而維度的數值則代表該字詞的權重，如圖 2.2.2 所示。

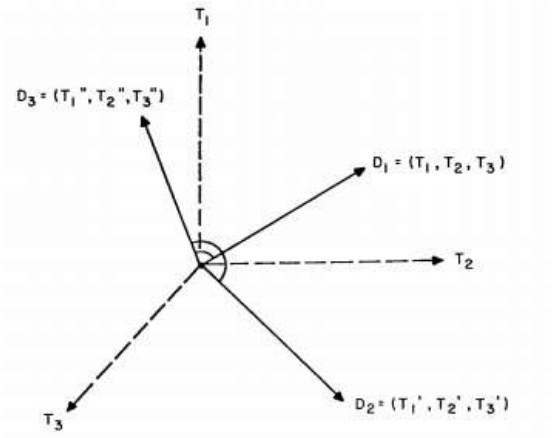

圖 2.2.2 向量空間模型

資料來源：Salton et al., 1975

此外，為了使文件間能互相比較，使用向量空間模型時必須轉化為單位向量，以避免文件長短不一所造成的誤差。當文件數量增加時，可利用「詞彙—文件矩陣」表達詞彙與文件間的關係。以圖 2.2.3 為例，文件集選出 i 個特徵字，而每一列則代表一篇文章中各個特徵字的權重值。

$$\begin{bmatrix} & Term_1 & Term_2 & \cdots & \cdots & \cdots & Term_i \\ Doc_1 & W_{11} & W_{12} & \cdots & \cdots & \cdots & W_{1i} \\ Doc_2 & W_{21} & W_{22} & \cdots & \cdots & \cdots & W_{2i} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ Doc_k & W_{k1} & W_{k2} & \cdots & \cdots & \cdots & W_{ik} \end{bmatrix}$$

圖 2.2.3 詞彙—文件矩陣

資料來源：Salton & Gill, 1983

2. 相似度計算

將文件以向量空間模型表達後，可藉由相似度計算實現分群或分類技術。常用的相似度計算方式有 Jaccard 係數及 Cosine 係數（Salton, 1988），前者計算文件字詞出現於交集之機率意即$\frac{|x\cap y|}{|x|\cup|y|}$，而後者則計算兩向量間之餘弦值。

3. 分類、分群技術

(1) 貝氏分類器

此分類器以貝氏定理為基礎，且假設屬性間彼此獨立下，以事前機率計算事後機率，再判斷資料屬於哪個類別（黃孝文，2010；章秉純、許清琦，2001）。透過大量的學習，能有效處理欲分類的資料。

(2) kNN分類器

此演算法搜尋與新文件最相似的 k 份文件，並比較兩者之相似度，選擇各分類中相似度最高的類別，因此演算過程中，最重要的即為 k 值大小之決定。一般採用 M-way kNN，演算步驟如下（戴尚學，2003）：

Step1、將文件以向量空間模型表示。

Step2、取出前 k 份與新文件相似度最高之文件，此 k 份文件之類別則為候選類別。

Step3、將文件與新文件之相似度以類別為基礎做加總，分數最高之類別則為新文件之所屬類別。

(3) 支援向量機

支援向量機（Support Vector Machines, SVM）由 Vapnik 於 1979 年提出，分類方式是在多維度空間中，以超平面（Hyperplane）對資料作分割，使分類邊界最大（章秉純、許清琦，2001）。

(4) K-means分群

此演算法由 J. B. MacQueen 於 1967 年所提出，分群之前須先設定群集數量 k，以質心的概念對群集做迭代，直到質心趨於穩定，群集收斂為止（A Tutorial on Clustering Algorithms, 2011）。K-means 雖然能得到較佳的分群結果，質心的概念卻容易受到資料的離散程度影響，分析者在事前未必能正確決定群集數量，若資料量龐大將造成整體效率降低。圖 2.2.4 為群集及群集質心之示意圖。

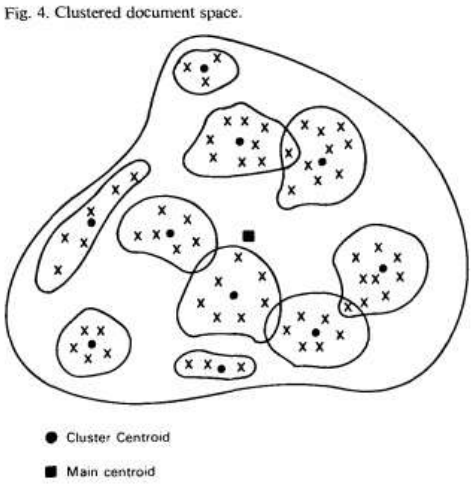

圖 2.2.4 分群與群集質心示意圖

資料來源：Salton et al., 1975

K-means 分群步驟(A Tutorial on Clustering Algorithms, 2011)如下：

Step1、隨機選取 k 個資料當作質心。

Step2、將資料與此 k 個質心計算相似度，選擇較近的分為同一群，最後再計算 k 群資料所產生的新質心。

Step3、若質心產生變動，意即尚未收斂，則重複 Step2，直到質心收斂為止。

2.2.3. 文件分析

通常，文字處理後需透過客觀的方法評估其效用，根據 Sebastiani(2002)的整理，評估方法較常使用的有 Accuracy、Precision、Recall，以及 F-measure 等，其中，Accuracy 評估預測結果中分類預測結果的機率；Precision 評估預測結果正確中分類預測結果亦為正確的機率；Recall 評估預測結果與分類預測結果相同中，預測結果正確之機率；F-measure 綜合 Precision 及 Recall 之評估方式而成。此四種為資訊擷取領域中常用之評估指標（Sebastiani, 2002）。

第三節 事件偵測與追蹤

根據學者定義：「在特定的時間、地點所發生的事情，即為事件」，其目的在於能利用資料間的關係描述此事件。事件偵測是為了辨識已存在事件的特徵，或尚未發生的事件；而事件追蹤則是從事件既有的樣本中尋找子事件（Allan, 1998）。

3.1 事件偵測

事件偵測分成「回顧偵測（Retrospective Detection）」與「線上偵測（On-line Detection）」兩種（Allan, 1998）。兩者差異在於前者是從固定範圍的文件中偵測事件是否存在，而後者則具有範圍不定，且具有時間性，此種方式較適用於新聞事件。

CMU 在新聞事件偵測上，採用向量空間模型表示一篇文件，而偵測方式則是採用單一連結法（戴尚學，2003；Yang, 1997；Kurt, 2001），新聞事件偵測可分成字詞權重值計算及群聚方式。前者採用 TFIDF 計算，而後者採用向量空間模型表示，而事件將以事件質心（Centroid）表示，新文件僅需與事件質心作相似度比對，偵測新文件所屬之事件是否存在即可。CMU 實現新聞事件偵測時，加入了時間區間（Time Window），避免文件與事件時間間隔太遠仍被考慮的情況（戴尚學，2003；Yang, 1998）。

3.2 事件追蹤

事件偵測僅用來找尋相似度高的事件，如何正確歸類則需依賴事件追蹤。CMU 採用 2-way kNN 分類法對每個事件進行事件獨立追蹤(戴尚學，2003；Yang, 2000)。此分類法將事件分成「目標（Positive）事件」及「非目標（Negative）事件」，計算結果即為新文件與目標事件的相關分數（Relevance Score）。為避免 k 值大小造成分類不正確的問題，亦提出平均加以修正。

第四節 類神經網路

4.1 概論

類神經網路（Artificial Neural Network, ANN）最早由 McCulloch 和 Pitts 在 1943 年共同提出，是人工智慧領域中的重要領域之一。它使用電腦模擬神經細胞接收外界刺激後傳遞訊息的動作，是一種用來表達生物神經網路的數學模型，屬於平行運算的模式。由於其模型適合用來解決較複雜的非線性模型，擁有學習、聯想、歸納推演等能力，因此常被用來解決最佳化、分類、預測等問題，而這些問題通常具有下列特徵（張斐章、張麗秋，2005）：

(1) 問題及相關條件難以完整定義。

(2) 需要快速得到問題解答且解答不用完全精確。

(3) 問題非常複雜或是非線性的問題，無法由一連串已知的數學方程式來描述。

4.2 網路結構

類神經網路模擬生物神經元傳遞訊息的結構，由許多運算單元組成，而運算單元之間透過非線性關係連結組成，而運算單元則分成輸入層、隱藏層及輸出層。每一個神經元之輸入值與輸出值對應方式，需要先將輸入值與對應權重之乘積相加，並透過活化函數轉化，才能得到輸出值。活化函數的目的是為了使輸入值做非線性轉換，可用來模擬門檻值、轉化輸出值範圍（張斐章、張麗秋，2005）。

類神經網路可分成監督式及非監督式學習（張斐章、張麗秋，2005），監督式學習必須不斷修正神經元的權重，因此每一次的訓練都會包含輸入項及目標值，而非監督式學習則是依照輸入資料的特性去學習、調整權重，因此，學習的方式通常應用於聚類的問題。兩者的差異性可由圖 2.3.1 表示：

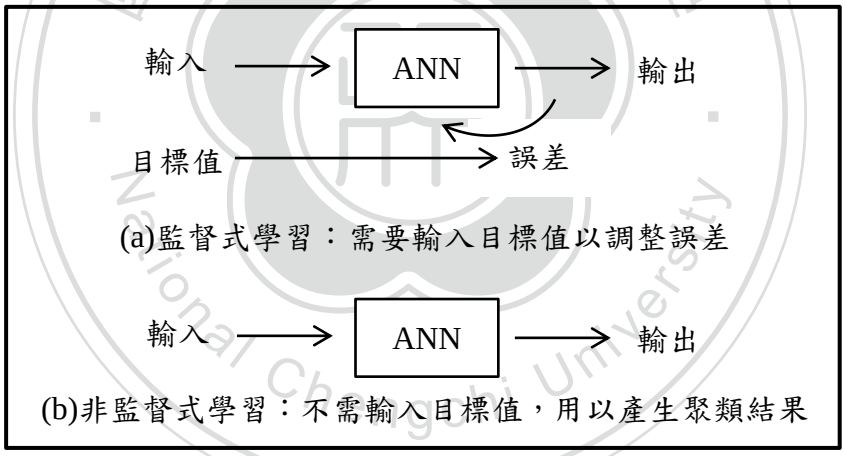

圖 2.3.1 監督式學習與非監督式學習

資料來源：張斐章、張麗秋，2005

類神經網路的運作模式可分成兩部分，即正向傳播及負向傳播。正向傳播能將每一層的神經元經過加權運算及活化函數處理後，輸出至下一層神經元；負向傳播則出現於監督式學習之類神經網路，目的在於修正輸出值與目標值的誤差，透過誤差值的回傳，各層神經元能夠適當調整、修正其權重值及偏權值，值到誤差值達容忍範圍內為止。

4.3 股票市場之應用

由於時間序列受到統計及經濟意義上的限制，使類神經網路從輸入、輸出之對應關係中更能提供股票市場之隱藏訊息（周宗南、劉瑞鑫，2005），因此許多學者嘗試採用不同方法於類神經網路上實現，使股票市場之應用更為廣泛，以下為本研究之整理：

Kim & Han（2000）採用特徵離散的概念，並使用基因演算法決定類神經網路之連結權重以預測股市價格指數，實驗證明此方式優於傳統的選擇部分特徵字和拓墣優化的概念。

陳稼興、楊孟龍（2000）應用類神經網路於預測個股股價在波段漲跌走勢，並搭配資金配置、交易策略用於選股，實驗結果證實，在操作方式相同的情況下，此選股方式之報酬率較佳。

Nygren（2004）提出以誤差修正為基礎之 ECNN（Error Correction Neural Network）網絡結構，預測結果是成功的，但僅適用於每週預測，說服力有限。

周宗南、劉瑞鑫（2005）比較時間序列與類神經網路模型於台股指數報酬率之預測，結果顯示時間序列模型由於統計上的假設與限制，相異於人工智慧模型在資料間所尋找的對應關係，而結合時間序列與類神經網路之模型，能有效改善類神經網路過適化問題，且能捕捉到其他模型所沒有的訊息，使預測效果更好。

黃馨瑩、楊建民、李耀中（2009）採用類神經網路技術，訓練期間學習股價與隔日股價的關係，探討股市新聞量的資訊對於面板股價趨勢的影響，實驗結果顯示新聞量與股市指標之間具有顯著關聯，而類神經網路加入新聞量因素後，亦能提高預測股價趨勢之正確率。

第三章 研究設計

根據文獻探討之綜合分析，本研究採用分群技術使新聞文件聚集為新聞事件以降低資訊量，爾後透過新聞事件與類別詞庫做相似度比對，使其分成正向、持平及負向新聞事件。另一方面，將各類別之新聞事件數量與量化資料（成交量、收盤價與 3 日平均價）一併輸入於倒傳遞類神經網路預測模型，將此預測模型與未含新聞事件之預測模型進行預測隔日收盤價之比較，進一步發掘新聞事件對於股市漲跌是否呈現顯著影響。

因此，本研究將研究設計分為兩階段做相關討論——「新聞文件分群與分類」及「倒傳遞類神經網路預測模型」，而研究架構如圖 3.1.1 所示：

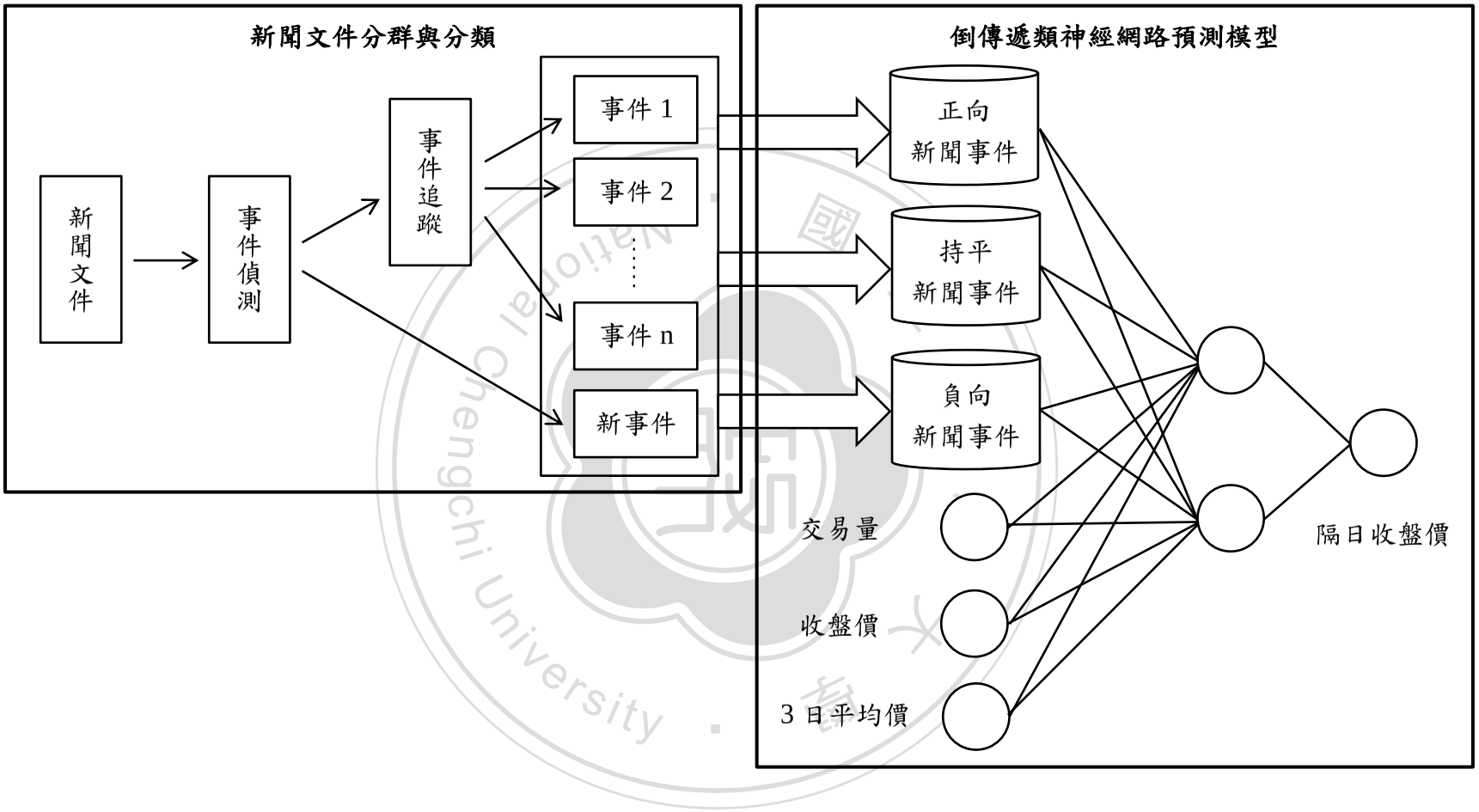

圖 3.1.1 研究架構

資料來源：本研究整理

第一節 新聞文件分群與分類

此階段將新聞文件透過分群技術聚集為新聞事件，使其與類別詞庫做相似度比對，分成正向、持平及負向新聞事件。此階段希望以分群降低資訊量，並利用分類使新聞事件能在倒傳遞類神經網路預測模型訓練過程中，挖掘各類別於預測模型之影響程度，進一步反應於模型預測能力。下述將針對研究方法加以介紹：

1.1 斷詞工具

由於本研究採用中文新聞文件，進行文字探勘前需先對文件進行斷詞之前處理作業，在此介紹目前常見的工具：中研院提供的 CKIP 及 Yahoo API，兩者的比較整理如表 3.2.1：

表 3.2.1 CKIP 與 Yahoo API 之比較

	CKIP	Yahoo API
輸入編碼	BIG5	UTF-8
特色	使用剖析樹、結構樹等技術斷詞，並具有辨識新詞的能力	可分析多國語言
服務限制	一次處理一句，因此不能超過 80 字	每天只能處理 2000 字

資料來源：本研究整理

新聞為提供新訊息的工具之一，使得出現新詞之機率相對於其他中文文件更為頻繁，使得文句需要自我判斷詞意之重要性相對增加，此外，本研究所採用之新聞（資料來源）又以中文為主，因此本研究將採用 CKIP 之斷詞工具以進行後續研究流程。

1.2 新聞文件分群——事件偵測與追蹤

1.2.1. 以 kNN 分群技術實現

新聞的事件具有時效性，且每天不斷更新，事件大小也隨著媒體追蹤程度有差異，因此新聞的分群並無法事前決定群數，本研究將採用 CMU 的方式，應用 kNN 分群法於新聞事件追蹤上。而執行事件偵測前，需先將文件轉換為向量空間模型表示，因此需要計算文件字詞之權重值，權重值計算方式採用 TFIDF，如公式(1)，並採用正規化調整權重值大小：

$$w_{ij} = tf_{ij} \times \log\left(\frac{N}{df_i}\right) \quad \text{公式(1)}$$

根據 CMU 所提出之事件偵測與追蹤之流程如下：

Step1、計算新進文件之權重值，並與既有事件計算相似度，CMU 在新聞偵測上，將相似度加入了時間區間（Time Window），避免文件與事件時間間隔太遠仍被考慮的情況，而計算方式以公式(2)表示（戴尚學，2003；Yang, 1998），若超過門檻值則表示新進新聞文件可能屬於該事件，將其列入候選事件，並進入 Step2，否則新進新聞文件不屬於該事件，結束此演算法。

$$score(x) = 1 - \max_{c_i \in window}\left\{(1 - \frac{k}{m}) \times sim(\vec{x}, \vec{c_i})\right\} \quad \text{公式(2)}$$

$\max_{c_i \in window}\{sim(\vec{x}, \vec{c_i})\}$表示當新文件與已存在的事件作相似度比較的最大值，因此$score(x)$可視為事件存在的門檻，當$score(x)$大於所設定的門檻，則表示新文件屬於一新事件，反之，表示新文件存在於目前的事件。

時間區間之m為時間區間中的文件數，k值為該群集中最新一篇所收錄的時間至x所收錄的時間所增加的文件數；相似度則使用餘弦相似度公式，計算方式以公式(3)表示：

$$\mathrm{sim}(x,c)=\frac{\sum_{j=1}^{M} w_{jx}\times w_{jc}}{\sqrt{\left(\sum_{j=1}^{M} w_{jx}^{2}\right)\times\left(\sum_{j=1}^{M} w_{jc}^{2}\right)}} \quad \text{公式(3)}$$

Step2、若候選事件採用 2-way kNN 分類法，將候選事件列為目標事件，而非候選事件則為非目標事件，並以平均的概念避免 k 值大小所造成分類不正確的問題，以公式(4)表示。

$$r(\vec{x},kp,kn,D)=\frac{1}{|U_{kp}|}\sum_{\vec{y}\in U_{kp}}\cos(\vec{x},\vec{y})-\frac{1}{|V_{kn}|}\sum_{\vec{z}\in V_{kn}}\cos(\vec{x},\vec{z}) \quad \text{公式(4)}$$

公式(4)表示新文件需分別與目標函數及非目標函數之最近鄰kp、kn個數比較，意即$|U_{kp}|\le kp$，$|V_{kn}|\le kn$

Step3、若計算結果大於 0 則表示新進新聞文件屬於該事件，此時將新進新聞文件標記為該事件，並調整此事件之質心，質心調整方式如公式(5)；反之，表示新進新聞文件不屬於該事件，結束此演算法。

$$w_{jc}=\frac{N\times w_{jc}+w_{jx}}{N+1}\ \text{for all j} \quad \text{公式(5)}$$

其中w_{jc}為原來的質心權重，w_{jx}為新進新聞文件之權重，

N為原來事件涵蓋的文件數

若新進新聞文件已對所有既有事件做偵測與追蹤，仍未被分類至適當的事件，表示其不屬於既有事件，則建立新事件，並訂定新事件之質心，即為該新進新聞文件，結束此演算法。

1.2.2. 以 RTD-based（Relative Text Distance-based，相對文件距離）kNN 分群技術實現

基於新聞文件無法事前得知群集個數，事件偵測與追蹤將採取 kNN 分群技術，雖然 kNN 採用非監督式學習，然而 kNN 分群於迭代的過程需要不斷計算不同文件的相似度，卻難以記錄其演算過程，使複雜度隨著資料量得增長而提高，效率相對降低。因此本研究欲以特徵字代表文件之單位向量為基礎，嘗試以文件間的歐氏幾何距離取代相似度的方式進行新聞事件分群。研究中將使用 RTD-based kNN 取代 kNN 之分群技術提高系統整體效率。下列將敘述如何以文件距離取代相似度的軌跡。

文件距離取代相似度的方式，本研究採取歐智民、陳柏均（2011）之研究，取代方法如下：

1. 取第一篇文件做為基準點，因此每篇文件都能儲存與第一篇文件計算後之距離。
2. 有了基準點後，與新文件之相似度範圍便能改以距離取代，如圖 3.2.1。

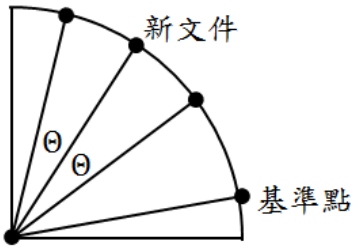

圖 3.2.1 文件相似度與距離關係

資料來源：本研究整理

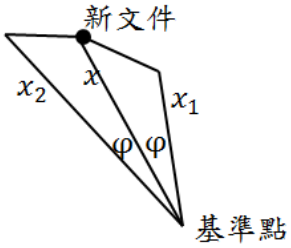

圖 3.2.2 文件相似度轉換為距離關係

資料來源：本研究整理

3. 由於三角形的角度φ恰為弧度θ之圓周角，因此$\varphi = \frac{\theta}{2}$（參考圖 3.2.2），而距離範圍則可以餘弦公式求之，證明如下：

證明：$\cos\frac{\theta}{2} = \frac{x^2+x_1{}^2-a^2}{2\times x\times x1}$，

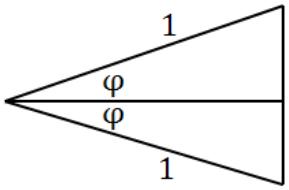

圖 3.2.3 公式轉換圖

資料來源：本研究整理

其中 $a = 2\sin\varphi = 2\sin\frac{\theta}{2}$（圖 3.2.3）

由餘弦定理知，$\cos\frac{\theta}{2} = \frac{x^2+x_1{}^2-a^2}{2\times x\times x_1}$

得$x_1 = \left|\frac{2\times x\times\cos\frac{\theta}{2}-\sqrt{D}}{2}\right|$

其中$D = 4\times x^2\times\cos\frac{\theta}{2} - 4\times x^2 - 16\times\cos^2\frac{\theta}{2} + 16$

同理，$x_2 = \left|\frac{2\times x\times\cos\frac{\theta}{2}+\sqrt{D}}{2}\right|$，而D與上式相同。

1.3 新聞事件分類

本研究假設每天所發生的新聞事件將整體影響到隔日收盤價之漲跌，為了找出各個新聞事件之影響，需要先將新聞事件分類成正向、持平、負向新聞事件，如此倒傳遞類神經網路預測模型才能夠根據各類別所具有的特徵加以學習、訓練。

以往關鍵詞可透過權重值（如 TFIDF、TFC 等）做為評估字詞之重要程度，然而各類別之代表關鍵詞在該類別中所出現的頻率亦成了相當重要的因素。本研究採用之權重值為為 TFC，其考慮區域權重及全域權重之影響，因此當權重值高時卻可能因為字頻過低造成，依照權重值高低可能擷取到較不具類別意義之關鍵字，另一方面，字詞出現次數若過於頻繁，關鍵字之重要性又會降低。根據以上分析，類別詞庫之選擇需考量字頻及權重值之因素，本研究將選取 DF 介於 0.075~0.3 之範圍做為篩選，爾後再從權重值的大小挑選類別詞庫之關鍵字代表。

第二節 倒傳遞類神經網路預測模型

本研究之預測模型所採用之對照組中，輸入項目以量化資料為基礎，包括交易量、收盤價及 3 日平均價，而輸出項目則為隔日收盤價，模型以天為單位的方式訓練模型；而實驗組於輸入項目中加入量化資料，即為正向、持平及負向之新聞事件數量。透過訓練進行權重值修正，以預測結果進行分析與討論。

倒傳遞類神經網路由 Rumelhart、McClelland 等人於 1986 年提出，屬於多層前饋式網路，並以監督式學習修正輸入與輸出之間的關係(張斐章、張麗秋，2005)。由於其擁有學習精度高、回想速度快、輸出值可為連續值的優勢，因此能處理複雜度高與高度非線性函數之問題。

一般來說，倒傳遞類神經網路的運作是一迭代的過程，透過不斷的權重修正以獲得最佳權重值，流程如圖 3.2.4：

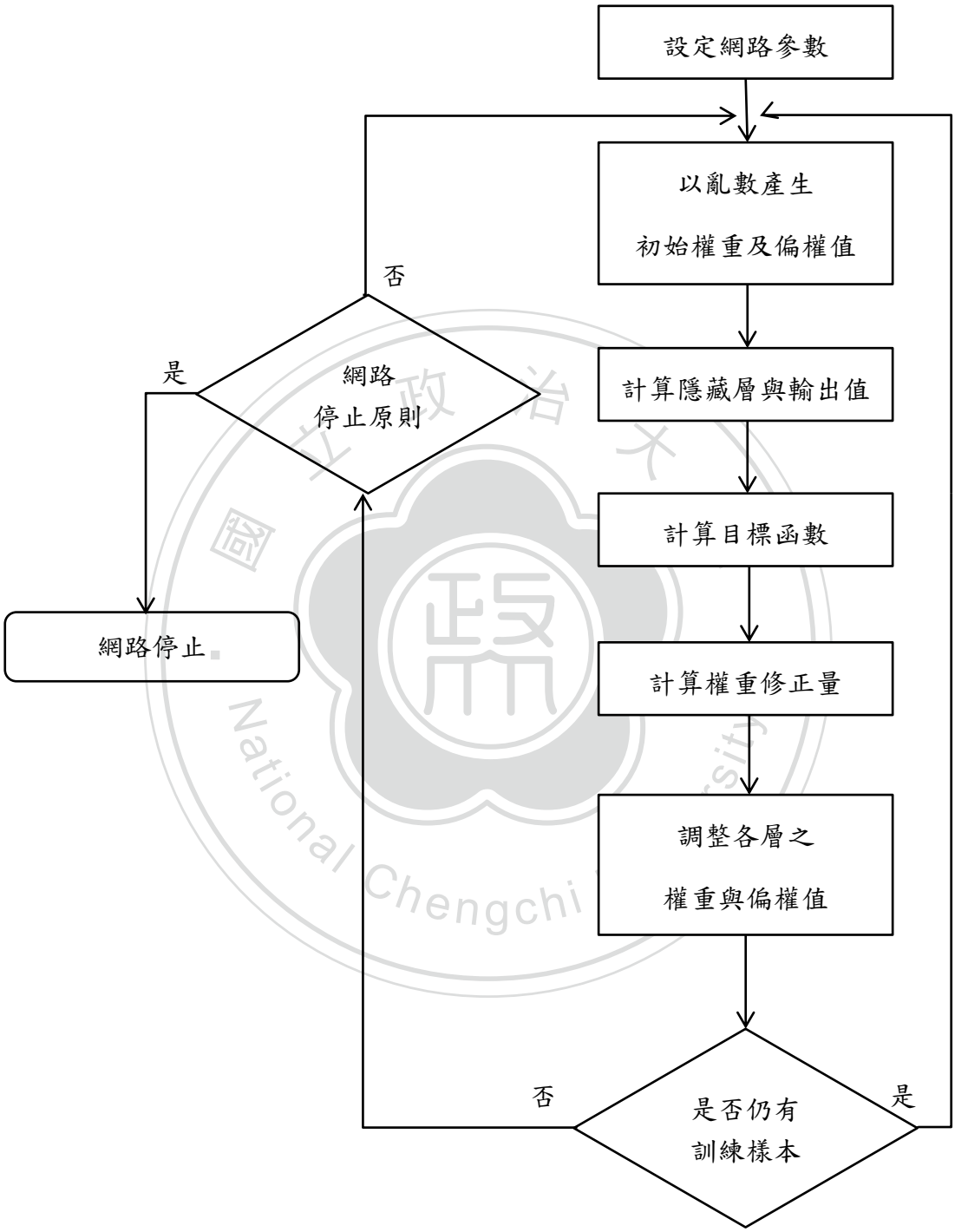

圖 3.2.4 倒傳遞類神經網路流程

資料來源：張斐章、張麗秋，2005

建構倒傳遞類神經網路預測模型時，訓練階段將以天為單位，使模型能在有限的迭代次數（2000）及容忍誤差（0.01）下產生預測模型，然而模型建構過程中仍需考慮以下問題：

1. 權重初始化

類神經網路模型的訓練過程是為了尋找適當的權重值，根據研究，其權重初始值可採用下列兩種方法：

(1) Ham & Kostanic（2001）提出權重初始值應落在$\left[\frac{0.5}{N}, \frac{-0.5}{N}\right]$之間，$N$為該層神經元之個數。本研究將採取此方法設定權重初始值。

(2) Nguyen & Widrow（1990）表示，可先取得$[0.5, -0.5]$之隨機亂數，再透過公式(6)或公式(7)進行修正：

$$\gamma = 0.7\sqrt[n_0]{n_1} \quad \text{公式(6)}$$

$$w_{ji} = \gamma \frac{w_{ji}}{\sqrt[n_0]{\sum_{j=1}^{n_1} w_{ji}^2}} \quad \text{公式(7)}$$

2. 活化函數、誤差函數

倒傳遞類神經網路模型較常使用的活化函數為 purelin 函數、tansig 函數及 logsig 函數，原因在於此三種函數皆為可微分之連續函數，整理如圖 3，故權重值或偏權值之修正可採用最陡坡降法及迭代過程達到容忍誤差，實現模型學習目的（羅華強，2005）。本研究於預測模型之隱藏層所使用之活化函數為 logsig 函數，原因在於模型中欲加入新聞事件數量之比例，其值將介於 0 到 1 之間，因此採用 logsig 函數較為適合，而輸出層為隔日收盤價，此數值在股票市場上並沒有範圍，因此所採取之活化函數為 purelin 函數。

當倒傳遞類神經網路預測模型之輸出層無法得到目標值，則需將誤差函數透過最陡坡降法修正其權重值與偏權值，使誤差值達到容忍範圍內或迭代次數達到上限為止。（張斐章、張麗秋，2005）

3. 隱藏層層數

研究顯示，隱藏層不需超過兩層以上，而一或兩層則沒有定論（Chester, 1990；Hayashi et al., 1990；Kurkova, 1992；Hush & Horne, 1993；張斐章、張麗秋，2005），其中 Hush & Horne（1993）指出，某些問題中使用兩層隱藏層的網路，各隱藏層只需有少量神經元即可以取代「使用一層，但需要數量龐大神經元」隱藏層的網路。

4. 隱藏層神經元個數

可由兩種方式達成（Dawson & Wilby, 2001；張斐章、張麗秋，2005）：網路修剪法（Pruning algorithm）（Abrahart et al., 1998）及網路增長法（Constructive algorithm）（Kwok & Yeung, 1997）。前者先將隱藏層個數設為極大，並逐一修剪神經元個數，直到超過誤差容忍範圍為止，然而此方式需耗費大量時間，效率較低；後者則將隱藏層個數設為最小，再逐一增加神經元個數，直到誤差達到容忍範圍為止，相較之下，此方式的效率較佳，能以較經濟的方式達成，因此本研究將採用網路增長法選取適合之神經元個數。

整體而言，倒傳遞類神經網路預測模型之結構整理如圖：

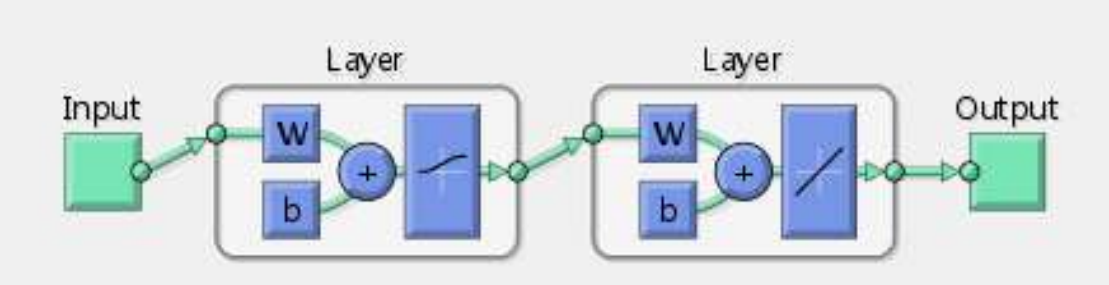

圖 倒傳遞類神經網路預測模型之結構

資料來源：本研究整理

其中，輸入層之質化資料以各類別之新聞事件數量輸入，量化資料以交易量、收盤價及 3 日平均價為指標；輸出層則為隔日收盤價；活化函數之隱藏層及輸出層分別以 logsig 及 purelin 函數為主。而層數及各層之神經元個數之決定則將由研究結果進一步探討之。

第三節 研究樣本與統計檢定

3.1 研究樣本

為了避免新聞經過大量報社的整合（如 Yahoo 新聞或 Google 新聞等），造成對權重計算或選擇特徵值的影響，本研究資料來源將選擇以報社出身之「自由時報電子報」為新聞文件樣本，採取 2009 年 1 月 1 日至 2009 年 12 月 31 日之新聞資料，共 5767 筆新聞文件，平均一天有 15 篇新聞文件；而股市基本面資訊如交易量、收盤價、3 日平均價等資訊，則選擇「台灣經濟新報資料庫」所提供之半導體類股做為研究對象，而 2009 全年包含 251 個交易日。圖 3.1.1 及圖 3.1.2 分別為自由時報電子報及台灣經濟新報資料庫畫面。

圖 3.3.1 自由時報電子報

資料來源：本研究整理

O2324 OTC 半導體 | 查詢日期 20090101-20091231 | 日期排序

未調整股價(日)-均價 (個 股 總 覽)←說明網頁

	日期	當日成交量	收盤價(元)	3日均價(元)
1	2009/01/05	276,942	23.31	22.81
2	2009/01/06	367,887	23.72	23.25
3	2009/01/07	299,386	23.76	23.60
4	2009/01/08	228,599	22.38	23.29
5	2009/01/09	174,962	22.54	22.89
6	2009/01/10	148,380	22.11	22.34
7	2009/01/12	130,215	21.98	22.21
8	2009/01/13	116,496	21.95	22.01
9	2009/01/14	225,651	22.28	22.07
10	2009/01/15	122,792	21.16	21.80
11	2009/01/16	200,728	21.35	21.60
12	2009/01/17	110,547	21.18	21.23
13	2009/01/19	148,602	20.79	21.11
14	2009/01/20	125,203	19.94	20.64

圖 3.3.1 半導體類股交易資訊

資料來源：本研究整理

3.2 統計檢定

一般類神經網路預測模型以 MSE（Mean Square Error）或 RMSE（Root Mean Square Error）為評估標準，分別為公式(8)、公式(9)所示，其中x_i表示預測值，而t_i表示實際值：

$$MSE = \frac{(x_i - t_i)^2}{n} \quad \text{公式(8)}$$

$$RMSE = \sqrt{\frac{(x_i - t_i)^2}{n}} \quad \text{公式(9)}$$

本研究採取對照的實驗模型，因此將使用 RMSE 做成對樣本之顯著性檢定加以檢驗兩模型之差異。值得一提的是，無論是 MSE 或 RMSE 之評估標準，類神經網路在建置過程中容易產生過適化，亦即實驗模型之訓練過程中，誤差值越低將使學習能力過於合適，卻相對造成驗證過程之誤差結果不佳的情形。

第四章 研究結果

本研究將研究架構分為兩部分——「新聞事件分群與分類」及「倒傳遞類神經網路預測模型」，本章將針對研究架構依序進行三個實驗：首先針對類別詞庫之建立做相關討論，接著對於倒傳遞類神經網路預測模型之參數建構進行分析；最後則以統計檢定評估預測模型之預測方向正確性及預測準確率是否呈現顯著差異。

第一節 類別詞庫之建立

本研究使用 RTD-based kNN 分群技術將新聞文件轉化成新聞事件，期待降低資訊量之非結構化資訊亦能對於股票市場造成影響。然而為使新聞事件能在倒傳遞類神經網路預測模型中獲得有效的學習能力，本研究將建立事件之類別詞庫，使新聞事件能夠依照相似度高低將其分類為正向、持平或負向新聞事件。

為了建立新聞事件之類別詞庫，本研究對於 2009 年半導體類股報酬率之天數分佈初步觀察，統計結果繪製折線圖如圖 4.1.1，圖中可看出，半導體類股報酬率之漲跌幅度介於-7%至 7%之間，而其中 68%則集中於-1%至 3%，推論半導體類股大致呈現左偏之趨勢，屬於正向緩和成長。本研究欲藉由報酬率區間之界定，從新聞事件中挑選適當之關鍵字代表。

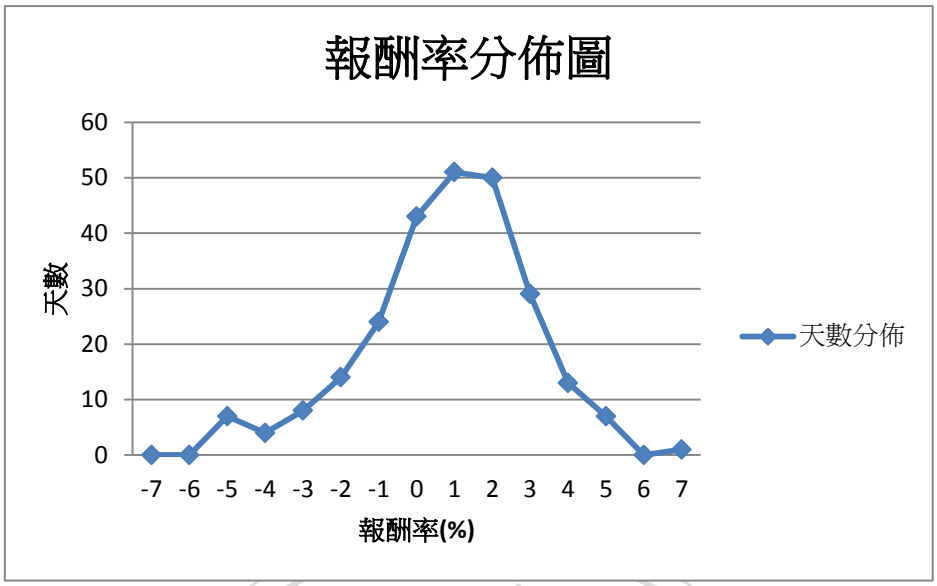

圖 4.1.1 報酬率分佈-2009 年

資料來源：本研究整理

因此研究將以報酬率範圍做為初步篩選，將新聞事件以日期為基礎分成正向、持平及負向類別，爾後再對詞彙做篩選。鍾任明等人（2007）之研究指出，關鍵詞彙可依據詞彙於一篇文章中所出現次數（Term Frequency, TF，以下簡稱為詞頻），及關鍵詞權重作為篩選考量，其中詞頻太低可能造成太多雜訊，而詞頻過高則難以表現該類別之特徵，使經過挑選之詞彙無法具有代表性。而本研究所挑選之詞彙為類別代表，因此將依此概念衍伸，從全域性觀點修正，即以 DF（詞彙於所有文件中出現的次數）替代 TF，範圍則設定為 0.075 至 0.3，再選擇前百分之二之權重值為關鍵詞代表。下列將根據報酬率範圍分析類別詞庫代表之關鍵詞，整裡如表 4.1.1。

表 4.1.1 類別詞庫數量比較

±4%	類別	正向	持平	負向
	天數	11	232	8
	詞庫數量	39	17	26
±3%	類別	正向	持平	負向
	天數	19	211	21
	詞庫數量	33	18	31
±2%	類別	正向	持平	負向
	天數	33	168	50
	詞庫數量	30	19	33

資料來源：本研究整理

表 4.1.1 分別以報酬率作為類別詞庫之實驗：當報酬率介於±4%時，雖然正向類別及負向類別所佔天數較少（8 天、11 天），詞庫數量卻明顯集中於此二類別，顯示正向類別及負向類別擁有較明顯特徵之關鍵字。然而當報酬率介於±3%及±2%時，正向與負向之關鍵詞已有趨於平均的現象。為了呈現報酬率介於±4%之正負向類別詞庫數量比例，本研究將選取報酬率±3%範圍做為類別詞庫之分界，並進一步對新聞事件之分類結果討論。

此處以 2009 年 10 月 29 日一正向新聞事件及其關鍵字做類別詞庫為例加以討論，其標題如下：

第三季獲利亮眼 楠梓電、尖點抗跌
Q3 台達電 EPS1.21 元 光寶科 1.13 元
Q3 營收與獲利 緯創、仁寶創新高
景氣復甦 矽品明年資本支出達百億
Q3 聯詠 EPS1.91 元 凌陽 0.59 元

此五篇新聞文件敘述第三季於太陽能個股呈現景氣復甦之現象，包括產能率提升、營收與獲利增加，使得投資成本提高等，觀察該新聞事件所挑選出來的關鍵字，其中「盈餘」、「營收」、「獲利」、「成長」、「盈餘」、「提升」、「增加」出現次數最為頻繁，其他如「改善」、「復甦」、「受惠」、「抗跌」、「搶眼」等皆有助於用以描述該新聞事件之關鍵字，整體而言，可由類別詞彙看出該事件所表達的為一正向新聞事件。又，對應至半導體類股報酬率所造成之正向影響可推測，太陽能產業與半導體產業具有密切關係，因而當太陽能相關之正向新聞事件產生時，造成半導體類股之報酬率亦呈現正向影響。

第二節 倒傳遞類神經網路預測模型之參數建構

在倒傳遞類神經網率預測模型中，為了選擇適當參數，本研究將迭代參數設為 2000，而學習參數為 0.7，又研究顯示隱藏層層數不需超過二層（Chester, 1990；Hayashi et al., 1990；Kurkova, 1992；Hush & Horne, 1993；張斐章、張麗秋，2005），因此本研究將以此為基礎實驗神經元個數，採用網路增長法觀察其 RMSE，預測模型之架構如圖 4.2.1，依照此網路模型進行實驗，而 251 筆資料來源中，本研究隨機挑選 168 筆做為訓練資料，其餘 83 筆則做為測試資料，實驗結果列於表 4.2.1，並依此探討神經元個數之選擇方式。

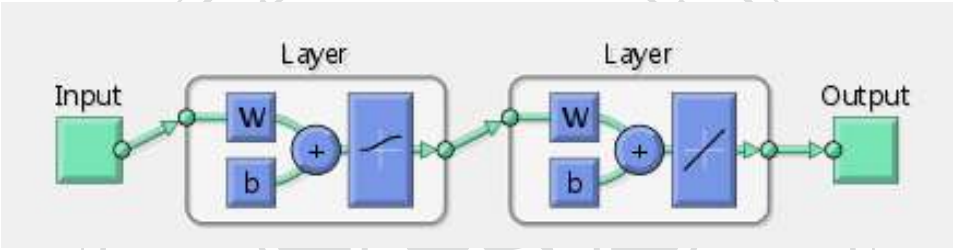

圖 4.2.1 預測模型之架構

資料來源：本研究整理

表 4.2.1 不同參數之模型 RMSE 比較（*含新聞事件，**不含新聞事件）

層數	個數		訓練*	測試*	訓練**	測試**
1	1		0.0279	0.0448	0.0282	0.0469
	2		0.0274	0.0446	0.0281	0.0469
	3		0.0268	0.0446	0.0278	0.0466
	4		0.0263	0.0440	0.0272	0.0456
	5		0.0264	0.0442	0.0274	0.0461
平均			**0.0270**	**0.0444**	**0.0277**	**0.0464**
2	1	1	0.0334	0.0497	0.0350	0.0562
		2	0.0303	0.0507	0.0294	0.0504
		3	0.0368	0.0571	0.0429	0.0695
平均			**0.0335**	**0.0525**	**0.0358**	**0.0587**
2	2	1	0.0287	0.0458	0.0296	0.0484
		2	0.0496	0.0700	0.0516	0.0742
		3	0.0300	0.0474	0.0303	0.0497
平均			**0.0361**	**0.0544**	**0.0372**	**0.0574**
2	3	1	0.0282	0.0461	0.0289	0.0476
		2	0.0294	0.0464	0.0303	0.0499
		3	0.0298	0.047	0.0304	0.0496
平均			**0.0291**	**0.0465**	**0.0299**	**0.0490**
2	4	1	0.0288	0.0461	0.0299	0.0495
		2	0.0289	0.0465	0.0302	0.0497
		3	0.0280	0.0460	0.0288	0.0484
平均			**0.0286**	**0.0462**	**0.0296**	**0.0492**

資料來源：本研究整理

由表 4.2.1 可看出，一層預測模型之 RMSE 範圍介於 0.0270 至 0.0277 之間，而二層預測模型介於 0.0286 至 0.0501 之間，初步推論，層數的增加雖然使權重值之修正方向增加，卻因為受限於相同迭代次數，難以有效改善預測模型之誤差值。另一方面，若訓練階段所得 RMSE 越低，將有效降低測試階段之 RMSE，亦即兩者具有正向關係，因此預測模型可根據訓練期間所得之 RMSE 決定其參數。因此本研究將設定訓練過程中 RMSE 最低之參數作為下階段實驗的基準，即層數為 1，神經元個數為 4 之預測模型。

圖 4.2.2 為該預測模型之預測趨勢圖，上下圖分別為訓練及測試階段，顯現預測模型之訓練階段有助於幫助學習預測之能力，然而預測能力則依賴參數之調整。此外，圖 4.2.2 亦呈現半導體產業於 2009 年之股價呈現正向成長，呼應上一節報酬率分布圖中，多達 68%的交易日資料集中於-1%至 3%之間，顯示該產業確時有穩定成長之趨勢。

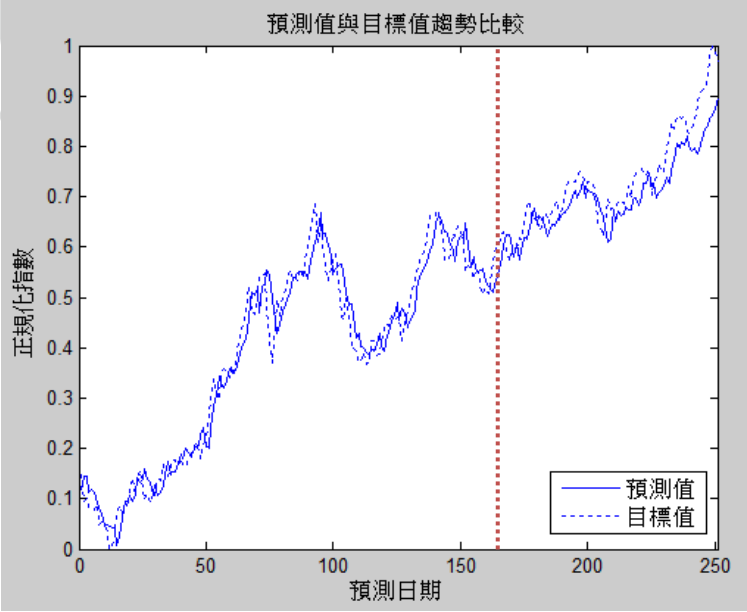

圖 4.2.2 含新聞事件之預測模型（層數為 1，神經元個數為 4）

資料來源：本研究整理

第三節 預測模型之顯著性檢定

透過神經元個數之之觀察結果，本研究採取層數為 1，神經元個數為 4 之預測模型，決定參數後即可將含新聞事件（實驗組）與不含新聞事件（對照組）之預測模型加以比較。由於兩預測模型之差異僅再於新聞事件之輸入與否，因此本節將從統計角度出發，以成對樣本評估新聞事件能否改善預測模型之學習能力，而評估方式則針對預測方向正確性及預測準確率做顯著性檢定。

3.1 預測方向正確性

兩預測模型的差異在於輸入項是否加入新聞事件，而預測模型經過學習若能使預測方向正確性提高，則代表新聞事件有助於預估隔日收盤價。又本研究使用±3%之報酬率作為正向、持平、負向新聞事件的分界，因此評估預測方向正確性時亦以±3%之報酬率將其分成上漲、持平及下跌三類，進行預測方向正確性之顯著檢定，本研究提出假說如公式(10)，H_0表示兩模型不具顯著性差異，反之H_1則否：

$$H_0 : \mu_2 - \mu_1 = 0$$
$$H_1 : \mu_2 - \mu_1 \neq 0 \quad \text{公式(10)}$$

本研究中包含 40 組樣本（n），資料列於表 4.2.3，依據中央極限定理，樣本大於 30 時樣本趨近於常態分配，因此檢定統計量及其分配如公式(11)：

$$\frac{D_{Acc}-(\mu_2-\mu_1)}{S_D/\sqrt{n}} \sim N(0,1) \quad \text{公式(11)}$$

其中，$D_{Acc} = Acc_2 - Acc_1 = 79.47\% - 82.04\% = -2.57\%$，

$$S_D = \sqrt{\frac{\sum_{i=1}^{n}(D_i - D_{Acc})^2}{n-1}} = 0.0074$$

而拒絕域$RR = \{D_{Acc} < CV_1 \text{，} D_{Acc} > CV_2\}$

表 4.2.4 預測方向正確性之顯著檢定（*含新聞事件，**不含新聞事件）

模型編號	Acc_1*	Acc_2**	D_{Acc}
1	82.00%	79.60%	-2.40%
2	82.00%	80.00%	-2.00%
3	81.60%	78.80%	-2.80%
4	81.60%	79.60%	-2.00%
5	81.60%	79.60%	-2.00%
6	82.80%	79.20%	-3.60%
7	81.60%	79.20%	-2.40%
8	81.60%	80.00%	-1.60%
9	81.60%	80.40%	-1.20%
10	82.40%	79.20%	-3.20%
11	82.00%	78.80%	-3.20%
12	82.80%	78.80%	-4.00%
13	81.60%	79.60%	-2.00%
14	81.60%	79.20%	-2.40%
15	81.60%	79.60%	-2.00%
16	82.00%	79.60%	-2.40%
17	81.60%	79.60%	-2.00%
18	82.40%	79.20%	-3.20%
19	82.00%	80.80%	-1.20%
20	82.00%	80.00%	-2.00%
21	83.20%	79.20%	-4.00%
22	82.40%	79.20%	-3.20%
23	82.00%	79.60%	-2.40%

24	82.80%	78.40%	-4.40%
25	82.00%	79.60%	-2.40%
26	82.00%	79.60%	-2.40%
27	82.00%	79.60%	-2.40%
28	82.40%	78.80%	-3.60%
29	82.00%	79.60%	-2.40%
30	82.80%	78.80%	-4.00%
31	82.00%	79.60%	-2.40%
32	82.00%	79.60%	-2.40%
33	82.00%	79.60%	-2.40%
34	82.00%	79.20%	-2.80%
35	82.00%	79.60%	-2.40%
36	82.00%	80.00%	-2.00%
37	82.00%	79.60%	-2.40%
38	82.00%	79.60%	-2.40%
39	81.60%	79.20%	-2.40%
40	82.00%	79.60%	-2.40%
平均	**82.04%**	**79.47%**	**-2.57%**
變異數	**$1.6200*E^{-5}$**	**$2.0831E^{-5}$**	**$5.4100*E^{-5}$**
標準差	**0.0040**	**0.0046**	**0.0074**

資料來源：本研究整理

由臨界值檢定法計算臨界值，如公式(12)：

$$CV_1, CV_2 = \mu \pm z_{1-\frac{\alpha}{2}} \frac{\sigma_D}{\sqrt{n}} = 0 \pm z_{1-\frac{\alpha}{2}} \frac{0.0074}{\sqrt{40}} \quad \text{公式(12)}$$

由式 3 可得，當顯著水準為 95%，$CV_1, CV_2 = \pm 0.002308$；當顯著水準為 99%時，$CV_1, CV_2 = \pm 0.003033$，然而不論顯著水準為何，$D = |-2.57\%| > |CV_1|$，表示$D \in RR$，亦即拒絕$H_0$，以統計觀點而言，此成對樣本具有顯著差異，該結果可說明於預測模型中加入新聞事件將有助於預測方向之正確率。

從表 4.2.2 中可初步看出，預測方向正確性皆以含新聞事件之預測模型較為優異，而含新聞事件之實驗組的平均正確率 82.04%亦高於不含新聞事件之對照組的平均準確率 79.47%，初步評估兩預測模型之比較，含新聞事件之預測模型結果顯著優於不含新聞事件之預測模型。

3.2 預測準確率

另一方面則檢驗測試階段之 RMSE 值，由於兩預測模型彼此相依，因此同樣以成對樣本之雙尾檢定觀察兩模型之 RMSE 是否顯著差異，提出假說如公式(13)：

$H_0：\mu_2 - \mu_1 = 0$

$H_1：\mu_2 - \mu_1 \neq 0$ ..... 公式(13)

同樣採取 40 組樣本（n），資料列於表 4.2.4，依據中央極限定理，樣本大於 30 時樣本趨近於常態分配，因此檢定統計量及其分配如公式(14)：

$\frac{\overline{D_{RMSE}}-(\mu_2-\mu_1)}{S_D/\sqrt{n}} \sim N(0,1)$ ..... 公式(14)

其中，$\overline{D_{RMSE}} = \overline{RMSE_2} - \overline{RMSE_1} = 21.2791 - 21.1841 = 0.0949$，

$S_D = \sqrt{\frac{\sum_{i=1}^{n}(D_i-\overline{D_{RMSE}})^2}{n-1}} = 0.0224$

而拒絕域$RR = \{\overline{D_{RMSE}} < CV_1，\overline{D_{RMSE}} > CV_2\}$

表 4.2.4 預測準確率之顯著性檢定（*含新聞事件，**不含新聞事件）

模型編號	$\overline{RMSE_1}$*	$\overline{RMSE_2}$**	$\overline{D_{RMSE}}$
1	21.2421	21.3695	0.1274
2	21.2759	21.3695	0.0936
3	21.1459	21.2525	0.1066
4	21.1511	21.2369	0.0858
5	21.1251	21.2213	0.0962
6	20.9094	20.9796	0.0702
7	21.2187	21.3305	0.1118
8	21.3876	21.5358	0.1481
9	21.4864	21.5774	0.0910
10	21.0628	21.1537	0.0910
11	20.9640	21.0472	0.0832
12	20.9276	21.0160	0.0884
13	21.2499	21.3435	0.0936
14	21.1719	21.2317	0.0598
15	21.1381	21.2395	0.1014
16	21.3097	21.3798	0.0702
17	21.3746	21.4786	0.1040
18	21.0862	21.1693	0.0832
19	21.6917	21.8165	0.1248
20	21.3954	21.5254	0.1300
21	20.9354	21.0212	0.0858
22	21.1433	21.2213	0.0780
23	21.2083	21.3149	0.1066

24	20.9068	20.9848	0.0780
25	21.1797	21.2889	0.1092
26	21.3435	21.4786	0.1351
27	21.1563	21.2525	0.0962
28	21.0550	21.1199	0.0650
29	21.1615	21.2473	0.0858
30	21.0628	21.1018	0.0390
31	21.1849	21.3019	0.1170
32	21.1381	21.2447	0.1066
33	21.0966	21.1771	0.0806
34	21.1641	21.2473	0.0832
35	21.1719	21.2837	0.1118
36	21.4812	21.5696	0.0884
37	21.0654	21.1225	0.0572
38	21.1563	21.2681	0.1118
39	21.1173	21.2135	0.0962
40	21.3227	21.4292	0.1066
平均	**21.1841**	**21.2791**	**0.0949**
變異數	**0.0273**	**0.0316**	**0.0001**
標準差	**0.1651**	**0.1778**	**0.0224**

資料來源：本研究整理

由臨界值檢定法計算臨界值，如公式(15)：

$$CV_1, CV_2 = \mu \pm z_{1-\frac{\alpha}{2}} \frac{\sigma_D}{\sqrt{n}} = 0 \pm z_{1-\frac{\alpha}{2}} \frac{0.0224}{\sqrt{40}} \quad \text{公式(15)}$$

由式 3 可得，當顯著水準為 95%，$CV_1, CV_2 = \pm 0.006945$；當顯著水準為 99%時，$CV_1, CV_2 = \pm 0.009128$，然而不論顯著水準為何，$\overline{D_{RMSE}} = 0.0949 > |CV_1|$，表示$\overline{D_{RMSE}} \in RR$，亦即拒絕$H_0$，以統計觀點而言，此成對樣本具有顯著差異，此研究結果可說明，除了成交量、收盤價及三日平均價外，新聞事件亦為隔日收盤價的重要因素之一。

第五章 結論

第一節 結論與建議

本研究以「新聞事件偵測與追蹤」的分群技術為出發點，透過類別詞庫的建立，探討新聞事件對於市場股價之影響，期望能將新聞文件以新聞事件分類的方式與「倒傳遞類神經網路預測模型」做結合，進而提供投資者有效的投資線索及資訊。此架構將新聞文件以事件為基礎做分類，希望將資訊量降低之餘，亦能提供預測隔日收盤價之訊息。

從研究結果分析可看出，透過分群使新聞文件降低資訊量，將顯著改善預測方向正確性及預測準確率，此現象表示新聞事件之質化資料將有效提供市場股價之訊息，換言之，新聞事件中所隱藏之市場資訊可透過各種文字消息傳遞，本研究所探討之新聞文件則為其中一種，而過去研究所探討之財報資料及重大訊息與股票市場之影響亦息息相關，可知兩者皆為市場所傳遞資訊的一種方式。此亦證實 Khurshid 等學者（2002）之論述：「無論文字消息的形式為何，皆可能為影響金融市場的波動」

此外，本研究於建立類別詞庫時使用 DF 篩選，再以權重值作為選擇，透過新聞事件之分類後的分析，也確實能使新聞事件做出分類。由於現今網路上資訊量龐大，大至入口網站、小至個人所使用的網誌、留言板等，都存在著分類的需求，其目的即為了簡化資料量，未來將可採用此方法幫助管理者建置類別詞庫，使分類得以自動化，成為管理資訊之工具之一。

第二節 未來研究方向

本研究整合文字探勘領域於類神經網路預測模型，透過分群分類技術使預測模型能達到學習能力，研究成果亦顯示能顯著提高預測能力。針對未來之研究方向，本節提出以下三點建議：

1. 本研究以詞頻及權重值高低篩選、選擇關鍵字，用以建立類別詞庫，此處提出類別關鍵字選擇之概念。而目前網站上往往具有分類功能，類別選擇卻需要透過人工決定，若能將此概念應用於網站上，將新的文章或網頁與類別詞庫做相似度比對，使其能夠完成自動分類，將能帶來便利之處。然而類別詞庫之正確程度將直接影響分類之準確度，因此若類別詞庫可隨著時間動態新增或修正，將使其可信度大幅提高。
2. 本研究僅以時間作為切割，用縱斷面角度針對每天的新聞之分群分類結果預測隔日收盤價，然而以橫斷面的角度來看，大型新聞事件通常會持續發展一段時間，若能將縱斷面及橫斷面之訊息加以結合，將能提供更完整的新聞事件資訊，相信也能得到較準確之預測能力。
3. 本研究針對半導體類股進行分析討論，若將範圍縮小至個股，則質化資料除了新聞文件外，亦可涵蓋該公司所提供之財報資料或重大訊息等資訊，此方式可提高非結構化資訊之完整度，而各產業所建立的類別詞庫，亦可用於討論不同產業之依賴關係。

參考文獻

英文文獻

1. K. Aas and L. Eikvil(1999), Text Categorization: a Survey, Technical Report, no.941, Norwegian Computing Center.
2. Abrahart R. J., See L. & Kneale P. E. (1998). New Tools for Neurobydrologists: Using Network Pruning and Nodel Breeding Algorithms to Discover Optimum Inuts and Architectures. In Proceedings of the 3rd International Conference on Geocomputation. University of Bristol.
3. Ahmad, K., Oliveira, P., Manomaisupat, P., Casey, M. & Taskay, T. (2002). Description of Events: An Analysis of Keywords and Indexical Names, Third International Conference on Language Resources and Evaluation, LREC 2002: Workshop on Event Modelling for Multilingual Document Linking, p29-35.
4. Allan, J., Papka, R. & Lavrenko. V. (1998). On-line New Event Detection and Tracking, Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, p37-45.
5. Annexstein, F. (2002). Indexing and Representation: The Vector Space Model Retrieved, December 25, 2003, from the World Wide Web: http://www.ececs.uc.edu/~annexste/Courses/cs690/Indexing%20and%20Representation.ppt.
6. Armano G., Marchesi M., & Murru A. (2005). A hybrid genetic-neuralarchitecture forstock indexes forecasting," Information Sciences, Vol.170, Issue 1, p3-33.

7. Chen, A. S., Leung, M. T., Daouk, H. (2003). Application of neural networks to an emerging financial market: forecasting and trading the Taiwan Stock Index. Computers & Operations Research. 30, 6.
8. Dawson C. W. and Wilby R. L. (2001). Hydrological Modeling Using Artificial Neural Networks. Progress in Physical Geography. 25(1), p80-108.
9. Fayyed, U., Piatetsky-Shapiro, G. & Smyth, P. (1996). The KDD Process of Extracting Useful Knowledge from Volumes of Data, Communication of the ACM, Vol.39, p27-34.
10. Ham F. M. & Kostanic I. (2001). Principles of Neurocomputing for Science & Engineering. McGraw-Hill: New York, NY.
11. Hsu, C. W., Chang, C. C., and Lin, C. J. (2010). A Practical Guide to Support Vector Classification http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf
12. Hush, D. R. & Horne, B. G. (1993). Progress in supervised neural networks. IEEE Signal Process. Mag. (January 1993), p8-39.
13. Jing, L. P. , Huang, H. K., Shi, H. B. (2002). Improved Feature Selection Approach TFIDF in Text Mining. 1st International Conference on Machine Learning and Cybernetics, Beijing.
14. Joachims, T. (1998). Text Categorization with Support Vector Machines: Learning with Many Relevant Features, Proceedings of the European Conference on Machine Learning Springer.
15. Han, J. & Kamber, M. (2001). Data Mining: Concepts and Techniuqes, Morgan Kaufmann Publishers, San Francisco, CA.
16. Kim, K. J., Han, I. (2000). Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index.

17. Kurt, H. (2001). On-line new event detection and tracking in a multi-resource environment, MS Thesis, The Institute of Engineering and Science of Bilkent University.
18. Kwok, T.Y. & Yeung, D. Y. (1997). Constructive Algorithms for Struture Learning in Feedforward Neural Networks for Regression Problems. IEEE Transactions on Neural Networks. 3: 630-645.
19. Kwok, T. Y. and Yeung, D. Y. (1997). Constructive Algorithms for Structure Learning in Feedforward Neural Networks for Regression Problems. IEEE Transcations on Neural Networks. 3: 630-645.
20. Lavrenko, V., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D., and Allan, J. (2000). Language models for financial news recommendation. In Proceedings of CIKM 2000, p389-396, New York, N.Y., ACM Press.
21. Liu, H. & Motoda, H. (1998). Feature Selection for Knowledge Discovery and Data Mining. Kluwer Academic, Norwell, MA, USA.
22. MacQueen, J. (1967). Some Methods for Classification and analysis of multivariate observations. Proc. 5th Berkeley Symp. Math. Statist, Prob., 1:281-297.
23. Nguyen, D. & Widrow, B. (1990). Improving the Learning Speed of the 2-Layer Neural Networks by Choosing Initial Values of Adaptive Weights. In Proceedings of the International Joint Conference on Neural Networks. 3. San Diego, CA.
24. Nygren, K. (2004), Stock Prediction – A Neural Network Approach. Master Thesis, Royal Institute of Technology, KTH.
25. Popescu, A. (2001). Implementation of term weighting in a simple IR system, Personal course project, University of Helsinki.
26. Salton, G. (1989). Automatic Text Processing. Addison-Wesley, Reading, Mass.

27. Salton, G. & Gill, M. (1983). Introduction to Modern Information Retrieval, McGraw-Hill.
28. Salton, G., Wong, A. & Yang, C. S. (1975). A Vector Space Model for Automatic Indexing.
29. Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys, 34(1), p1-47.
30. Yang, Y., Ault, T., and Pierce, T. (2000). Improving text categorization methods for event tracking , Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.
31. Yang, Y. & Pedersen, J. O. (1997). A Comparative Study on Feature Selection in TextCategorization. Proceedings of the Fourteenth International Conference on Machine Learning, p412-420, Nashville, TN, USA.
32. Yang, Y., Pierce, T. & Carbonell, J. (1998). A Study on Retrospective And On-Line Event Detection , Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, p28-36.
33. Wu, Y. C. (2008). Predicting the Trend of Taiwan Weighted Stock Index with Text Mining Techniques, NCU IM.
34. Mittermayer, M.-A.(2004). Forecasting Intraday Stock Price Trends with Text Mining Techniques. In: Proceedings 37th Annual Hawaii Int. Conference on System Sciences (HICSS). Big Island, p64.

中文文獻

[1]林章德，2000，上市公司重大投資宣告對股價影響之研究，東海大學管理研究所碩士論文。

[2]林聖哲，2001，針對認購權證建構不同之人工智慧評價，實踐大學企業管理學系研究所碩士論文

[3]李春淋，2010，個股新聞對股價影響之研究-以台股為例，輔仁大學應用統計學系碩士論文。

[4]吳真蕙，2000，專業性報紙頭版新聞對股票價量的影響，中原大學會計系碩士論文。

[5]周宗南、劉瑞鑫，2005，演化式類神經網路應用於台股指數報酬率之預測，財經論文叢刊，第三期，p77-94

[6]胡舜禹，2009，結合 PSO 及 K-Means 聚類分析演算法的圖像分割，中山通訊工程研究所碩士在職專班

[7]袁立安，2007，混合式自動文件摘要方法，國立中山大學資訊管理研究所碩士論文

[8]陳稼興、楊孟龍，2000，類神經網路於股市波段預測及選股之應用

[9]章秉純、許清琦, Combining Unsupervised Feature Selection Strategy for Automatic Text Categorization, In Proceedings of the 6th Conference on Artificial Intelligence and Applications, November 9, 2001.

[10]張斐章、張麗秋，2005，類神經網路，台北市：東華書局

[11]黃孝文，2010，雲端運算服務環境下運用文字探勘於語意註解網頁文件分析之研究，國立政治大學資訊管理研究所碩士論文

[12]黃馨瑩、楊建民、李耀中，2009，財經新聞探勘影響股價趨勢之探討-以跨兩岸面板產業為例，

[13]楊踐為、李家豪、類惠貞，2007。應用時間序列分析法建構台灣證券市場之預測交易模型。中華管理評論國際學報，10，3

[14]鍾任明、李維平、吳澤民，2007。運用文字探勘於日內股價漲跌趨勢預測之研究。中華管理評論國際學報，10，1

[15]戴尚學，2003，運用事件偵測與追蹤技術於中文多文件摘要之研究，國立雲林科技大學資訊管理研究所碩士論文

[16]顧皓光，1996，網路文件自動分類，國立台灣大學資訊管理研究所碩士論文

[17]羅華強，2001，類神經網路，台北市：清蔚科技

網站資料

[1]A Tutorial on Clustering Algorithms (2011), 2011 年 2 月 3 日取自 http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html

[2]自由時報電子報。2011 年 2 月 1 日取自 http://www.libertytimes.com.tw/index.htm

[3]中研院 CKIP。2011 年 1 月 17 日取自 http://ckipsvr.iis.sinica.edu.tw

[4]Yahoo API (2011)。2011 年 1 月 22 日取自 http://tw.developer.yahoo.com/cas