財經新聞自動分類之研究

張凌翔 著

國立台灣大學

圖書館學研究所

碩士學位論文

經考試合格特此證明

指導教授

謝清俊

碩士學位論文考試委員

陳建光　謝清俊　董芳苑　謝清俊　吳明德

所主任

吳明德

中華民國　　年　　月　　日

誌　謝

論文寫作是每個研究生必經的考驗與磨練，藉由研究的過程，方能體會出治學的基本精神，窺得知識領域的浩瀚；在此由衷感謝謝清俊教授及陳克健教授的啓發與悉心指導，感謝他們在過去一年裏，給予我不斷的鼓勵與教誨；同時，李德竹教授、吳明德教授及盧秀菊教授等系上老師也給我許多賜正與關懷，在此一併致謝。

論文進行期間，最感謝的朋友，應該是中研院計算中心林晰和資訊所楊允言，從論文籌備到完成，他們一直盡力地幫助我撰寫相關程式，並和我一起解決各種層出不窮的問題，更經常無條件的漏夜趕工，使本論文得以順利進行；還要感謝中國時報楊明憲在資料等各方面的強力支援，沒有他們三人的鼎力相助，實沒有今日論文之完成，在此致上最深的謝意；此外，也要感謝永麒公司張健良、楊女華在論文後期所提供的統計程式支援，幫助我資料能夠順利的處理。

這份論文，更要感謝中國時報及中研院資訊所等單位的協助，感謝中國時報不吝提供資料供我研究，在原始新聞稿件的收集上助益良多；以及中研院資訊所提供的研究經費與各項軟、硬體設備，使本論文得到充份的後盾支援。

這段期間，感謝中國時報鄭副總家鐘、王副處長台鳳、羅副主任鴻進、洪明洲博士等長官的幫助與愛護，以及企劃處謝有智、劉碧宜等好友的精神支持；更要感謝台大總館編目組孫喜雲主任及張蓓蕾股長對我的照顧與體諒，使我能夠儘快完工，還有採錄股王梅玲主任、作標股連鍾蕙股長、西編股吳玉齡等同事及手下工讀生黃玉琴、謝愛香等人在生活、工作等各方面的支援與不斷鼓舞，也要謝謝研究所周利玲、施孟雅等同學；周曉雯、林彥君、林荷鵑等學妹，中研院張翠玲，以及關心我的大學同學們洪玉珠、張文熙等人，眞的非常感謝大家。

最後要感謝家人無時無刻的關懷，尤其是三哥在電腦軟、硬體及程式方面的援助與教導，以及辛苦的媽媽，對我所有的包容與無微不至的照顧，讓我在這三年早出晚歸的日子裏，每天都有熱熱的飯菜吃，也謝謝我家的小狗飛力，陪伴我熬夜趕論文。

摘要

本研究的目的，乃在於探究利用電腦將財經新聞自動分類之可行性。藉由國外發展的各種統計方法，來了解中文新聞資料中，詞彙與新聞及類別之間的關係，及各種統計方式在中文處理上的差異性，以工商時報2870篇財經新聞爲實例，分析出各小類專業詞彙的數量、出現頻率、分佈狀態等特性，以支持用專有詞彙作爲學習語料(training corpus)的理論基礎。並藉此研究歸納出人工分類與機械分類之間處理邏輯的異同，以提供日後資料庫更新維護的改進方向。

本論文是以工商時報民國80年7月至81年1月間2870篇見報之相關新聞作爲抽樣統計的主體，分類範圍限於產業、商業、電機、機械、資訊等五大類，共46小類，實驗方式是將抽取樣本全部予以人工分類，並劃分爲學習組(2583 篇) 和測試組(287 篇)兩大部份，分別進行各項統計分析，最後比較各種自動分類方法的效果；從實驗中發現，採用詞彙標準化頻率統計法的方式較爲簡易，就整體而言，電腦自動分類的正確率也較高。

根據研究結果可得到以下結論：

1.從本論文中可以發現，利用電腦儲存新聞，並作自動分類處理，確實是個可行的新作法，只要將學習語料的數量擴充，使詞彙總數趨於穩定，並修改統計缺失，當可使正確率提高，不過人工分類在實驗裏看來，仍有其價值存在，如何使機械和人工分類邏輯判斷方式更契合，則有待日後努力。
2.中文資料雖不似英文般有明顯的空白區隔，但藉助發展漸趨成熟的中文斷詞系統，仍可以解決中文詞彙選取的問題。
3.自動分類的方式雖無法百分之百和專家分類的結果相符，但如果有足夠的學習語料，還是能有不錯的成績，且人工分類有因人而有不同的分法，藉由自動分類的精神，也可以減少人爲分類的主觀偏差。

目次

參考書目

附　件 (一)　日本經濟新聞分類表
(二)　新聞稿樣本
(三)　電腦無法分類的學習樣本群
(四)　各類詞彙頻率詳表

圖表說明

一、表說明

二、圖說明

三、公式說明

第一章 緒論

第一節 取材緣起

歷史是人類鑑往知來的重要根本，也是人類生生不息的重要軌跡，換言之，有人類便有歷史的存在；所謂歷史，其實就是一頁頁過去的新聞，只是時空情境的不同，而賦予不同的時代意義。西方新聞學學者卡爾・比察（Carl Biicher）曾說：「新聞爲心靈交通的手段」。（註1）是故，無論是人類的各項言行、社交活動或思想紀錄，均可謂爲新聞。由此可知，新聞乃是以社會心靈交通爲使命，從事現實之報導與批判，並在最短時間內作迅速傳播的一種方式。（註2）這些新聞傳播的綜合體，即爲大衆傳播。

新聞傳播的種類衆多，舉凡報紙、電視、廣播、雜誌等媒體均是，在這些傳播媒體當中，又以報紙出現最早，存在的年代最久遠，筆者認爲，報紙的取材範圍深入社會各層面，時效性又高，是提供社會動態的重要媒體之一，因而引發了筆者擬以報紙爲研究對象之主體的原始動機。

報紙爲社會文化的重要產物故必須隨著時代變化，才能適切地反應人生萬象；社會的發展是多元化的，不僅有政治的改革，也會有科技的發明、貿易的行爲等現象，爲滿足讀者各式各樣的閱讀需求，於是，各種不同訴求內容的報紙也就應運而生，而有所謂的綜合性報紙如：紐約時報（New York Times）、聯合報、中國時報；財經性報紙如：華爾街日報（The Wall Street）、工商時報、經濟日報；休閒性報紙如：民生報等。

上述各類型報紙中，財經性報紙因其報導內容以經濟、財政、金融及工商產業動態的新聞見長，最能及時反應全球或世界各國之

景氣循環、經濟現況以及貿易往來等現象；就我國而言，經濟發展之成就更是令人刮目相看的奇蹟，我國自民國三十八年國民政府播遷來台以後，由於國家發生劇烈的轉變，台灣又剛從日本的統治下收回，社會架構極待重建，於是政府遂致力推動各項經濟建設，以期改善人民的生活，故開始逐步施行經濟改革方案，如：三七五減租、耕者有其田、公地放領，使人民得以安居樂業，政治局勢也日趨穩定；六十年代以後，在當時行政院院長蔣經國先生的主導策劃之下，又陸續推動十項建設及十二項建設，使我國工商產業漸次起飛，國民所得亦大幅成長，經濟發展正式邁入另一個新紀元，創造了輝煌的「台灣經濟奇蹟」模式。然而這一切的成就，多要歸功於經濟政策的成功及財金制度的規劃得宜；由此可知，財經政策制定後實施的成功與否，的確和一國未來之發展有著密切的關連。而財經性報紙也正因爲此一因素，一發行就立刻受到政府決策當局、工商團體、以及各階層人士的重視，是故，其起步雖較綜合性報紙遲，但發展卻甚爲蓬勃，廣受好評。

我國最早的財經性專業報紙經濟日報於民國五十六年創刊（註5），其後工商時報亦於民國六十七年成立，爲我國重要的兩大財經性專業報紙。財經性新聞在內容上不只要求證消息來源的正確性，並爭取時效性，同時更要兼顧資料的累積、回溯及後續發展的預測與追蹤，這不僅可供讀者分析時局變化、金融波動、各產業動向的參考，掌握機先，同時更是執政當局制定決策的重要參考指標；由於這些源源不斷的需求，使此類型報紙能在報業發展歷程中獨樹一格，佔有前瞻經濟發展之領導地位。這也是筆者選擇以財經性新聞作爲研究範圍之考量因素。

第二節　報紙儲存的新趨勢

早期世界各國報社均成立資料室，並派有專人負責資料的剪貼、整理及保存等工作，以提供社內記者查閱資料，但是報紙每天所刊載的新聞量非常龐大，以中國時報爲例，現在每日正常的出刊數量爲八大張，共三十二個全開版面，扣除廣告之外的報導性新聞稿約佔了四百篇左右（註７），若以每篇新聞稿長度五百字計算，則一天的新聞總字數就在二十萬字以上，日積月累下來，資料量的成長速率是極爲快速的；換句話說，如果要採用人工方式作地毯式地搜尋某類新聞，勢必要耗費龐大的人力、時間和體力，投注在資料的瀏覽、過濾之上，於是，有人開始構想把報紙當作一批批的文件資料，儲存在電腦裏；這方面起步較早者爲美國的紐約時報，該報在西元一九五一至一九六〇年代，每個月均花費衆多的人力與金錢，收集該報每篇新聞所出現的關鍵詞（ keyword ），經過統計、篩選之後，定期出版索引典；直到西元一九六一年末期，才開始利用電腦發展財經新聞資料庫，這不僅可以節省許多人力和資料整理的工夫，同時也使得該索引典取樣報紙的種類擴增，更增進了研究參考的價值。（註８）

就我國而言，目前各報社的資料室或資料中心，仍以人工剪報的方式來儲存報紙，這種傳統的資料儲存方式主要的缺點有：

1. 空間日益不敷需求。
2. 紙本日久會變黃、變脆、受潮變形，維護與保存不易。
3. 同樣的資料如無複本，則無法同時提供多人使用。
4. 紙本資料容易產生遺失或缺漏的困擾。
5. 合訂本使用不便，但如不裝訂，翻閱時又不甚便利。
6. 統計資料需要倚賴人工方式做地毯式的收集工作。
7. 報紙剪報資料的蒐集和整理必須花費可觀的人力。

是故，如果能夠經由其他報業先進國家發展新聞資料庫的前車之鑑，把中文報紙之新聞資料庫架構建立起來，透過電腦處理資料大量、快速、準確的優點，便能夠達到符合經濟效益的需求，同時，也可以同步掌握世界各國即時性的消息動態，促進我國新聞傳播事業的發展。

第三節　研究動機

論及新聞資料庫架構的建立，首先須考慮到資料的處理方式；由於報紙所刊載之新聞稿的篇幅並不大，描述內容的重點也較明確，因此，如果能夠將各篇新聞預先作分類處理，再從中抓取重要的相關性詞彙，就能迅速地找到所需要的資料，由此可見，若上述之假設成立，則如何在一篇新聞中挑選有用的詞彙，便是個關鍵性的步驟；然而，這種挑選關鍵詞彙的工作若由人工來處理是十分可觀的，以每種報紙每天的新聞量在四百篇左右計算，要將每篇新聞稿各入其類的話，至少得花上半天的時間，但是若改用電腦處理，就可以節省許多；關於關鍵字的研究，西方各國早已開始著手；從英語本身的結構來看，英語是屬於拼音式語言，整體來說都是以一個個詞（ word ）爲最小單位，詞和詞之間也有空白（ space ）加以區隔，對電腦而言，辨識方式較爲容易，所以發展過程穩定，已有較完整且一致性的處理方向，然而中文的處理方式就棘手許多，就中文本身來說，中國文字是屬於方塊性文字，每個字（ character ）本身除虛字外，都有獨立的意義，如：東、西、南、北、中等，但組合之後卻可能會有不同的含義，如：作東（表請客）、西藥（表藥品）、南瓜（表食物）、敗北（表示輸給對方）、中和（表示某地區的地名），其間就牽涉到語言學方面的知識，所以處理起來問題便複雜許多，目前國內已有中央研究院、國立交通大學、財團法人工業技術研究院等機構從事中文斷詞的研究，正確率可達到九十

九％以上（註 9），中國大陸於西元一九七〇年代開始，也投注了二千多名人力，進行中文詞彙的整理（註10），對於詞彙頻率的研究，也已將成果集結成册出版（註11），但對於新聞分類的架構和各類之下關鍵字（類目）的安排，尚未見有研究之先趨，因此引發筆者研究的動機，筆者假設藉助現有利用電腦從事中文斷詞可行的成就，把研究的範圍加以界定，將財經類目相關的詞彙整理出來之後，於各類之下的關鍵詞依據出現頻率的高低給予不同的比重，用這些具有類別辨識意義的關鍵詞去測試不同的新聞稿，以了解電腦將新聞稿自動分類的可行性。

第四節　研究目的

報紙是傳動訊息的重要媒體，更是人類生存發展與交流的資產，然而，由於報紙的資料量龐大，紙質易脆的保存障礙，在過去一直是不易解決的困擾；近年來，各種新科技產品相繼問世，人們也逐漸將其技術應用在報紙儲存上，如：製作縮影、光碟、影像處理等，這些方法均解決了空間儲存的問題，同時也延長了保存的期限；此外，爲了使報紙的應用更多元化，國外也開始朝資料庫的方向發展，如；紐約時報、日本經濟新聞，不僅將報紙儲存在電腦中，建立新聞資料庫，更積極發展線上即時系統，使人們能夠更方便、更快速地獲知各地消息；筆者此次研究，主要也是希望能把電腦大量處理資料的功能，應用在中文報紙上，同時考量分類的原理，建立一套適合中文資料的自動分類方式；就文獻得知，目前在海峽兩岸已有許多電腦自動分析文章、擷取關鍵字的實例，因此可將這些研究的精神引用到本論文所探討的主題上，以了解電腦自動分類之可行程度；及做爲新聞線上檢索的重要工具，並希望藉由此次研究，依據分類之基本原則，建議一種較爲合理，且適於電腦作業的方法，進而從研究過程中發現利用電腦處理新聞分類工作的價值，詞彙與文章之間的關係，以及可能遭遇到的問題，做爲日後處

理新聞分類的參考。

綜合上述所言，本論文進行研究之目的可歸納如下：

1. 藉由國外發展實例的經驗，應用於中文資料的整理，從統計中了解國外所提之各種統計法在中文資料處理效果上的差異，以建議較適合中文資料的自動分類方式。

2. 找出新聞自動分類方法的優缺點，並探討和人工分類處理流程上的差異，以便了解新聞自動分類的作法，是否能夠確實減少資料處理的人力資源和成本，提供節省人工分類工作的新方向。

3. 探討詞彙、新聞內容、類別三者之間的關係，以瞭解專有詞彙群是否可以作爲學習語料學習依據的理論基礎。

4. 研究人工分類與機械分類之間，對於詞彙處理與統計邏輯上的異同，以提供全文資料庫後續發展與改進的參考。

5. 實現報紙全文資料庫應用的理想，提供一種儲存與資料分類的新觀念。

第五節　研究方法與步驟

本論文所採行之研究方法和步驟敘述如下：

1. 選定分類系統及實驗類目範圍。
2. 以內容分析法分析研究範圍內的新聞，將同類的新聞整理在一起。
3. 用計劃抽樣法對所選定之半年期新聞作隔週之抽樣分析，每隔八天抽取一天的相關新聞作爲樣本基礎。
4. 用實驗法分析詞彙出現的頻率，以研究出詞彙和類別之間的向量關係。
5. 分別用dBASE Ⅲ 、C 語言撰寫驗所需之各項程式，

用UNIX系統執行統計結果並進行分析。
6. 根據向量模式所分析出來的詞彙與類別的關係，做爲電腦自動分類理論之實證。

換言之，本論文計畫以工商時報做爲抽樣統計的主體，將抽出來的樣本全部予以人工初步分類，並分爲學習組和測試組兩大部份，在學習組部份採用兩種方式進行詞彙的收集：一是藉助中央研究院中文全文資料庫系統的輔助，將每篇新聞用人工方式作標誌（mark-up）、選詞、統計的處理；另外則是由中央研究院資訊所提供系統詞彙判斷的功能，完全由電腦自行學習，建立學習語料（training corpus），判斷出屬於各類應有的關鍵詞，採兩法並行對照的方式，同時統計各類財經性新聞中各關鍵詞出現的數量和頻率，這些詞彙收集整理之後，分別建立詞彙檔，再進一步比較各關鍵詞和分類系統的關係，以期未來如有財經新聞存入時，能夠利用現有的詞彙檔進行比對分析，經由電腦分析判斷之後，建議較合適的類目，以實際掌握財經新聞自動分類之原始動機的可行程度。

第六節　名詞解釋

1. 經濟新聞

一般而言，經濟新聞的內容包括：財稅、金融、經濟政策、證券、貿易、工業、農礦業、市場及工商報導等。（註12）

2. 內容分析法（Content Analysis）

根據傳播媒體的內容，作客觀的、系統的、量化的分析研究，希望能解釋內容眞正的意圖。（註13）本論文中關於新聞之分類、標誌處理（mark-up editor）、裁文、關鍵字分析等，即屬於內容分析的範圍。