適應性文件分類系統

王稔志，張俊盛
清華大學資訊工程所
mr884366@cs.nthu.edu.tw ,jschang@cs.nthu.edu.tw

摘要

在網際網路蓬勃發展的今天，資訊的產生與傳播也越來越快速與多元化，為了讓使用者能在浩瀚無涯的網際網路上快速有效的尋找所需要的資訊，文件分類是解決的方法之一。然而耗費人力的人工分類已經無法滿足現實情況，因此一種良好分類機制是不可或缺。本文利用已有的分類架構與分類說明，進行統計式關鍵詞過濾，最後加上回授機制，期能使新聞分類更有效率。

在這次實驗中，為了設計出有效的自動分類系統，我們針對文件分類的重點與特性進行實驗與討論，發現：1.分類架構的重要性：需要符合分類資料的領域或類型。2.關鍵詞的取法：雙連字（Overlapping Bigram）比斷詞效果好。3.關鍵詞所在的位置越前面，與分類的相關強度越高。4.關鍵詞利用文件 Df 值與分類類別 Df 值進行過濾，得到好的分類關鍵詞。

最後我們討論回授機制，利用機率統計式的單字特徵關鍵詞進行回授關鍵詞的過濾，並加入原本的分類特徵詞中。使得召回率大幅上升。證明回授機制有助於分類效率的提升。

1. 簡介

面對網路上爆炸性的資訊量，讓使用者從不易取得資訊，慢慢的轉變成被大量沒有組織結構的資訊給淹沒。在網路資訊風起雲湧的這個關鍵時代，能有效且正確的利用網路資訊，就能確切掌握時代的脈動。

然而傳統搜尋引擎採用的技術在面對使用者的需求時，強調的是樣式比對（Pattern Matching），找到符合使用者查詢（Query）的文章。然而在資訊爆炸的這個時代，傳統樣式比對技術對於任何簡單的查詢字串（Query String）所傳回的資料量少則三、四十筆，多則上百上千筆資料。這些豐富的網路資訊雖然提供使用者眾多選擇但也會造成資訊超載（Information Overload）。文件分類可以把眾多的資訊有系統有條理的整合在特定的分類架構供使用者使用。

一般而言，文件分類系統可大致分為兩類：一、自動分類。二、專家分類。面對龐大的資料，需要大量人力的專家分類很顯然的已不符合現實狀況。自然而然地自動分類成為重要的研究方向。自動分類作法可大致歸為兩類（1）督導式(Supervised Training)（Duda and Hart 1973）和（2）非督導式（Unsupervised Training）(Yarowsky 1995)。所謂督導式分類作法為先有專家對訓練資料的文章進行分類，再由分類系統自行學習。雖然研究指出督導式分類會比非督導式分類有較好的效果 （Lewis 1996)。但是大量的訓練資料不易取得是督導式分類的最大問題所在。

本論文探討完全不用人為介入的自動分類作法，提出一種利用分類描述（Description Based）進行分類特徵選取的機制，並探討特徵選取的方法的好壞與如何過濾不重要的關鍵詞以提高正確率。最後並討論回授機制（Relevance Feedback）是否有幫助分類準確性的提升，進而可使分類達到可行的準確程度。

2. 相關研究

文件分類以往主要在討論如何把文件分到一組線性的分類架構。路透社新聞語料庫（Reuters-21578 corpus）的 135 個分類資料是典型的線性分類架構，也是大家最常研究的語料庫。不斷的有人針對路透社語料庫進行分類演算法的比較（Yang, 1997）。近年來已經開始有人對路透社語料庫進行『階層式分類』研究。(Chakrabarti et al.,1997)(Koller and Sahami, 1997)。並認為階層式文件分類可以得到較佳的正確率。

財經紀事的分類資料是現階段中文分類研究常用的語料庫，其資料為民國八十一年取自各大報紙的新聞標題資料，經過人工標示設定類別，共分九個大類，且細分為三十八個中類別（表 2.1），中類別下面又分小類別。總共有 131136 筆新聞標題資料。接下來我們會深入探討有關中文分類的相關研究。

大類別	中類別
公營事業篇	公營事業
服務業篇	交通運輸業
	觀光旅遊
	服務業（商業）
金融篇	金融
	銀行
	外匯
	股票
	租賃
國際篇	國際政經
產業篇	各項產業
貿易篇	貿易
農業篇	農業
	林、牧、漁、礦

表 2.1 財經紀事分類表（大類、中類）

總體篇	經濟
	消費
	稅賦
	關稅
	土地
	工業
	勞工
	商標、智慧財產
其他篇	人口
	公共建設
	大眾傳播
	郵政、電信
	企業管理
	財政
	社會
	人事動態
	人物檔案
	政府、政治
	教育
	醫療衛生
	科技
	環境
	公司檔案

表 2.1 財經紀事分類表（大類、中類）(續)

2.1督導式訓練的分類研究

督導式訓練是抽出部分資料經過人為標示每篇文章的分類，再經由訓練的步驟得到每個類別的特徵詞。大部分的分類研究都是利用訓練資料找出特徵詞，在經由特徵選取演算法的篩選留下重要的分類特徵詞。而督導式中文分類研究中主要分線性分類與階層式分類：

1. 線性分類

過去在中文分類的研究中，楊允言等利用雙連字並針對財經紀事的新聞資料進行線性分類；共處理 24 類。在文件特徵的選取上考慮關鍵詞的出現頻率、集中度及廣度，並以機率模式與向量模式進行分類，所得的分類正確率為 67%。

而利用特徵關鍵詞在朗文多功能字典（McArthur, 1992）中所佔主題的重要性或以 chi-square 進行關鍵詞過濾，找出每類中特中關鍵詞所形成的特徵向量，再用 K-nearest neighbor 進行分類的研究所得到的正確率將近 50%（Chen 1998）

根據許多結果顯示，利用雙連字串作為特徵的效果最好（楊允言, 1992)。關鍵詞過濾則是 chi-square 比傳統 tfxidf 來得有效。而分類演算法則是以向量式的結果最佳（楊允言, 1993)。

2. 階層式分類

線性分類由於類別間是完全獨立的，而階層式分類中，若有兩個類別的父節點相同，則代表這兩類別內容關係相近。因此柯淑津與陳振南(1999)利用階層式關係來對關鍵詞進行權重調整，把關鍵詞分成正項特徵與負項特徵。對於財經紀事的資料進行線性與階層式的分類。其中發現階層式分類（正確率 66%）比線性分類（正確率 61%）效果來的好。

2.2 非督導式訓練的分類研究

非督導式訓練的分類則是訓練資料並無標示分類別，進而找出相似的文章，性質上比較近似分群（Clustering)。利用文章中關鍵詞所形成的向量空間進行比對，並把相似的文章分在一起。通常

督導式訓練的分類效果比非督導式的分類效果效果來得好(Lewis, 1996)。督導式的分類訓練雖可得到較佳的分類效果，然而需耗費大量的時間與人力。非督導式的中文分類研究較為少見，在這本文中我們將會提出一種 description-based 的非督導式分類作法，並藉由回授機制加強分類之效果。

3. 文件分類作法

文件自動分類涉及的問題很廣。除了依照資料選取適當的分類架構外，如何選取關鍵詞，並利用關鍵詞與分類間的關係，過濾選取合適的分類特徵等，都是自動分類所要面對的課題。接下來我們會對如何決定文件分類架構、文件分類特徵的選取與文件自動分類三方面進行討論。

3.1 文件分類架構的決定

分類架構通常因為特定的領域或不同的需求而有所不同，比如說之前提過的新聞網站和網路服務公司（搜尋引擎）等等。一個好的分類架構，往往需要經過專業的圖書資訊學專家或語言學家詳細的規劃與討論。其決定分類架構所需的時間、資源與人力是超乎想像的。

然而當耗費大量資源完成的分類架構完成之後，往往又因社會變遷、產業結構改變或社會趨勢等外在因素所影響，而需持續不斷的修改與調整[註一]。這種資源與人力的不斷支出是無法避免的。

[註一]、圖書分類是階層式架構是最好的例子。現今圖書館分類架構大多採用民國七十八年賴永祥編定『中國圖書分類法』（增定七版）爲藍本。那時還沒有把電腦科技與資訊方面視爲主要的類別。然而十幾年後的今天，電腦資訊科技已成爲今天生活與產業結構上重要的一部份，而圖書館分類限於分類架構而把電腦資訊分到數學下面，這就是因爲產業結構改變而勢必要修正分類結構的例子。

在這次實驗中，直接採用兩個現成的行業分類架構來進行實驗，一、主計處行業分類結構，二、根據天下雜誌分類，縮減的主計處分類。

一、主計處行業分類結構：

我們採用的階層式分類架構是我國的行業標準主分類架構。是由行政院主計處第三局於民國五十六年訂定，其間經過六十年、六十四年、七十二年、七十六年與八十年數次修改而趨於完善。民國八十四年初因社會環境變遷、產業結構改變，而廣泛徵詢各界意見，經過六次專案修改而於八十五年五月完成最新版修訂。而未來為了使行業標準分類應用日趨廣泛，各界行業分類判定、解釋及修正建議日多，因此主計處自八十六年起採逐年部分修正，以符合供商業及學術應用發展之需，符合台灣產業結構的最新狀況。共計分為十一大類、七十中類、二二九小類及六一五細類[註二]。其說明文件中對每個類別都有詳細的描述(說明如下)。對於階層式最底層的細類更提供豐富的分類關鍵詞來加強分類標準的明確性。

分類編號：A0111

分類名稱：稻作栽培業

分類說明：凡從事以稻米栽培為主之行業均屬之。　　陸稻栽培；水稻栽培；

分類編號：A0112

分類名稱：雜糧栽培業

分類說明：凡從事以雜糧作物，如麥類、玉米、小米、高粱、豆類、甘藷等栽培為主之行業均屬之。　　小麥栽培；大麥栽培；甘藷栽培；黑麥栽培；蕎麥栽培；花豆栽培；紅豆栽培；綠豆栽培；大豆栽培；薏仁栽培；玉米栽培；落花生栽培；粟(小米)栽培；蜀黍(高粱)栽培；

分類編號：A0114

分類名稱：蔬菜栽培業

分類說明：凡從事食用根菜類、莖菜類、葉菜類、花菜類、果菜類、芽苗類等蔬菜之栽培為主之行業均屬之。食用竹筍、馬鈴薯、豆薯等，以及生鮮用香、辛料如薑、蒜、蔥、辣椒等之栽培亦歸入本細類。 竹筍栽培；胡瓜栽培；甘藍栽培；冬瓜栽培；南瓜栽培；越瓜栽培；韭類栽培；甕菜栽培；豆芽栽培；牛蒡栽培；苦瓜栽培；菱角栽培；菜豆栽培；茄子栽培；番茄栽培；草莓栽培；洋蔥栽培；蔥類栽培；大蒜栽培；蘆筍栽培；荸薺栽培；豆薯栽培；蘿蔔栽培；西瓜栽培；芋類栽培；甜瓜栽培；

分類編號：A0115

分類名稱：果樹栽培業

分類說明：凡從事各種水果如柑橘、荔枝、龍眼、桃、李、杏、梨、木瓜、芒果、棗等，乾果如胡桃、栗、榛等，蔓生鮮果如葡萄，其他水果如鳳梨等種植、栽培而以收穫其產品為目的之行業均屬之。果園之兼營種苗供應者亦歸入本細類， 李栽培；梅栽培；杏栽培；桃栽培；柿栽培；棗栽培；栗栽培；梨栽培；枇杷栽培；芒果栽培；橄欖栽培；胡桃栽培；木瓜栽培；蘋果栽培；榛子栽培；楊桃栽培；龍眼栽培；鳳梨栽培；香蕉栽培；檳榔栽培；葡萄栽培；蓮霧栽培；

[註二]、詳細的主計處行業分類表請參考網頁（第六次修訂）

http://www.dgbas.gov.tw/dgbas03/bs1/text/indu/indu.htm

主計處已於民國九十年一月份第七次修訂新的行業分類表，並已經正式放在網頁上供人下載。在第七次修訂時，已由十一大類擴充爲十六大類，其中包括了最新的公共行政業、文化、運動及休閒服務業等最新的產業結構，請參考網頁

http://www.dgbas.gov.tw/dgbas03/bs1/text/indu89/indu.htm （第七次修訂）

二、縮減版的主計處分類

主計處的行業分類別是非常詳細的類別說明，然而對於新聞分類而言，七百多類的分類架構太詳細。另外就分類的實施而言，部份類別的類別說明（description）較短，無法找出足夠的特徵關鍵詞，反而會影響新聞分類的正確率。因此，我們利用天下雜誌的分類架構來縮減主計處行業分類結構，找出主計處行業分類架構和天下雜誌分類架構之間的相關類別對照表加以縮減分類結構，因此產生了一個 89 類的階層式分類架構（表 3. 1）。

大類	中類	小類	對照代碼（主計處行業分類）
農漁礦	農漁礦	農林漁牧	A（包含 B 下面的所有類別）
農漁礦	農漁礦	礦業	B（包含 B 下面的所有類別）
製造	民生工業	食品飲料飼料	C11(包含 C11 以下除 C1181、C1182 之外的所有類別）
製造	民生工業	菸酒	C1181、C1182、C12（包含 C12 以下的所有類別）
製造	民生工業	紡織服裝	C13、C14、C15 等類
製造	民生工業	家具用材	C16、C17 等類

表 3.1 部分分類架構與對照表

3.2 分類特徵的選取

在分類架構下文件分類最重要的課題乃在於各類別特徵（features）的選取。特徵的好壞決定分類的正確與否。所以如何取得真正可以代表某一特定分類的特徵，以及特徵評分機制（weighting）是文件分類中重要的課題。接下來我們分別說明擷取關鍵詞與評分步驟。

一、特徵選取（Feature Extract）

然而中文和英文最大的不同就在於特徵字或特徵詞組的選取。英文因為本身結構不同於中文，不需要斷詞這個步驟。英文利用一些分隔符號（最常見的空白字元或標點符號）就可將可能的特徵詞分割開來。中文在詞或詞組的選取方面則面臨到斷詞的難題，尤其語言會不斷的成長，對於未知詞與專有名詞方面的辨識更是中文處理方面比較困難的問題。（表 3.2）

未知詞	未知詞說明
兩國論	是李前總統於 1999 年接受訪問後所引起的話題。
E 世代	網路蓬勃發展後對於現今生活的形容。

表 3.2 中文未知詞範例說明

在這篇論文中，對於關鍵詞選取我們採用兩種方法：斷詞、雙連字（Bigram）。並比較兩種方法之正確性。

三、特徵評分（Weighting）

在早期資訊檢索時代就有如何對具特徵的關鍵詞進行評分的研究，當時認為若關鍵詞在文件中的詞頻很高（TF，term frequency），代表這個關鍵詞足以代表這篇文章。後續更有學者認為詞在各文件中的分佈也要考慮（IDF：Inverse document frequency）。若這個詞出現在很多文章中，則重要性應該下降。因此合併詞頻與詞的重要性便成為文件分類常用的關鍵詞評分方法 tf x idf。

在這次實驗中我們會針對 tf x idf 以及其他的評分公式來進

行討論，並利用資訊檢索的方法來進行文章分類。

3.3 文件自動分類

在傳統資訊檢索（Information Retrieval）中最常利用精準率（Precision）和召回率（Recall）來進行分類結果的評分(圖 3.1)。因此我們也沿用召回率和精準率評分。

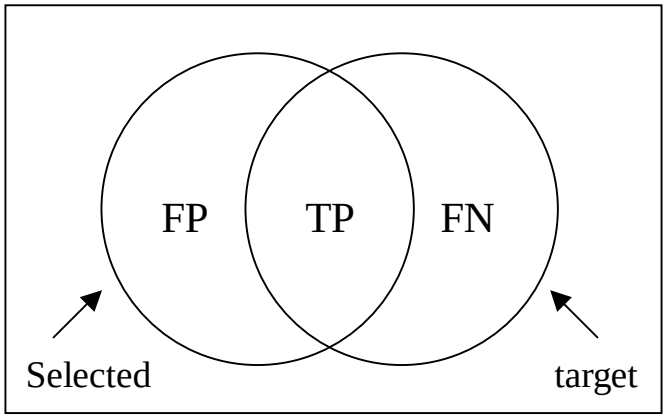


圖 3.1 精準率與召回率之圖示

FP：已分類文件中錯誤的部分

TP：已分類文件中正確的部分

FN：未分類的文件

公式： Precision = TP/(TP+FP)

Recall = TP/(TP+FN)

Applicability = (TP+FP)/(TP+FN)

精準率是所分類的文件中，正確的比率。召回率是所有文件的分類正確率。應用率為分類系統產生分類結果，不論對錯的比率。

4. 實驗設計說明

本實驗不同於以往的研究，在於應用資訊檢索原有的技術來進行分類研究。接下我們設計了一連串的實驗，並探討一些分類中值得研究的問題，如特徵關鍵詞的取得與過濾和相關性回授機制(Relevance Feedback）的效果。接下來我們會針對 1.實驗方法與實驗資料：2.實驗評估：3.實驗步驟三方向加以說明。

4.1 實驗方法

傳統 IR 的方法主要是針對使用者的查詢字串進行處理（如：Query expansion)，和許多的文件進行查詢比對，再把查最符合的文件傳回給使用者（圖 4.1)。我們把分類的特徵轉成是一種查詢的格式。每個類別都有一個由特徵詞構成的 Query。而比較不同的是傳統 IR 是一個 query 針對多篇文章下去查詢，並找出符合 query 的文章，然而我們的方法是所有的分類主題針對一篇文章進行比對評分，並按照每個類別對這篇文章的分數排序，取出分數最高的分類主題做為這篇文件的類別。(圖 4.2)

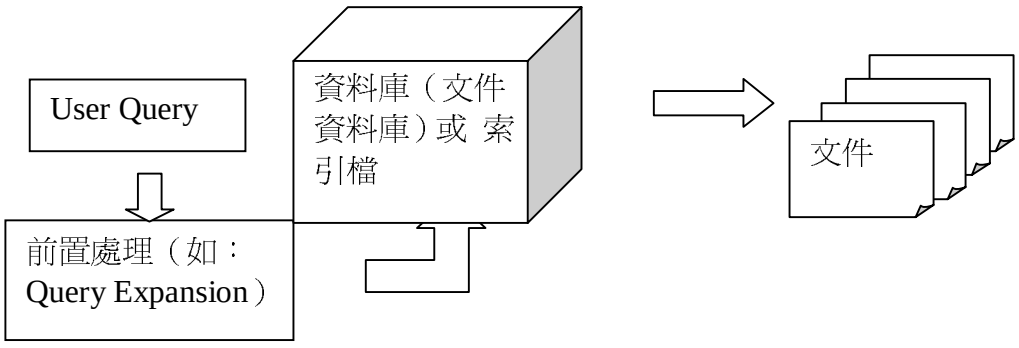


圖 4.1 傳統資訊檢索

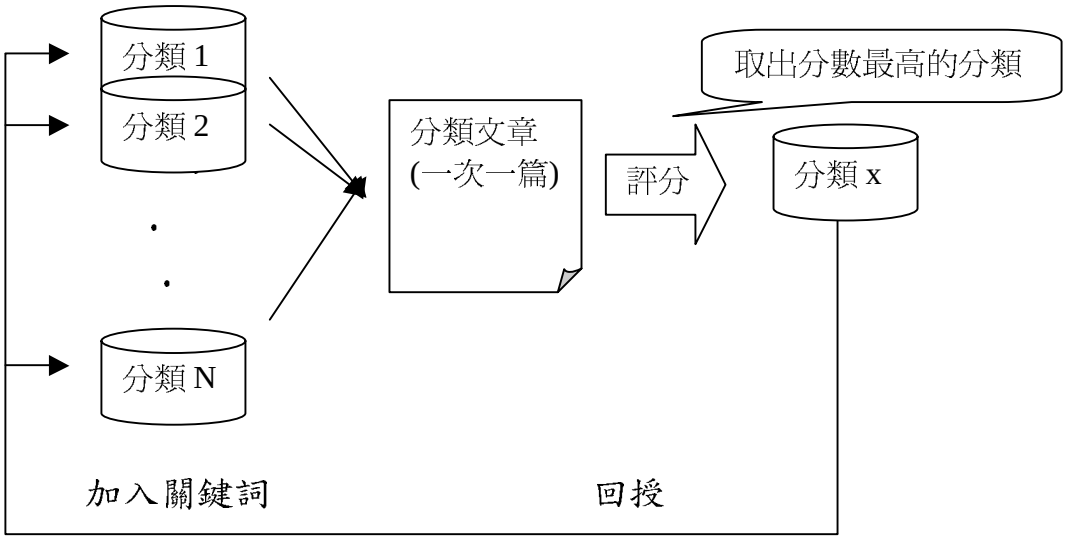


圖 4.2 文件分類系統

柏克萊大學(Cooper et al., 1994; Gey et al., 1996)認為文章和查詢相關與不相關的機率模型跟四個參數有相關，分別是字詞在查詢中出現頻率、查詢所包含的字詞總數與字詞在所有文件中出現的頻率、所有文件所包含的字詞總數呈現正相關，而與字詞在單一文件中出現的頻率、單一文件所包含的字詞總數呈現負相關，並利用文件和查詢中相符的字詞數當作重要參數對 NTCIR-2 的中文檢索資料（Kando et al. 2001）進行訓練，得到公式如下。

$$R_i = \log O(R \mid S_i, Q)$$
$$\approx \log \frac{P(R \mid S_i, Q)}{P(\overline{R} \mid S_i, Q)}$$
$$\approx -3.51 + 37.4 * X_1 + 0.330 * X_2$$
$$+ (-0.1937) * X_3 + 0.929 * X_4$$

$P(R \mid S_i, Q)$：S_i和查詢（*Query*）相關的機率

$P(\overline{R} \mid S_i, Q)$：$S_i$和查詢（*Query*）不相關的機率

X_1, X_2, X_3, X_4：此公式的四個參數，計算方法如下

$$X_1 = \frac{1}{\sqrt{N}+1}\sum_{i=1}^{N}\frac{qtf_i}{ql+35}$$
$$X_2 = \frac{1}{\sqrt{N}+1}\sum_{i=1}^{N}\log\frac{dtf_i}{dl+80}$$
$$X_3 = \frac{1}{\sqrt{N}+1}\sum_{i=1}^{N}\frac{ctf_i}{cl}$$
$$X_4 = N$$

N :< 文件 > 和 < 查詢 > 中相符的字詞數
qtf : 字詞在 < 查詢 > 中出現的頻率
ql :< 查詢 > 所包含的字詞總數
dtf : 字詞在 < 單一文件 > 中出現的頻率
dl :< 單一文件 > 所包含的字詞總數
ctf : 字詞在 < 所有文件 > 中出現的頻率
cl :< 所有文件 > 包含的字詞總數

分類完成後，針對三千篇新聞文件進行關鍵詞回授，利用單字特徵詞進行關鍵詞過濾比對，即可得到不錯的特徵關鍵詞。

4.2 實驗評估

傳統資訊檢索評估正確性是利用召回率與精準率（3.3 節）來進行評估，我們也採用這種方法。先請專家對於訓練資料前五十篇新聞依照第三章第 1 節所提到的兩種分類結構進行人工分類，找出標準答案。在比對系統所提供的答案，並算出正確率。這樣一來比較客觀，且評分結果也具有參考價值。

4.3 實驗設計

為驗證本研究提出的適應性分類系統，設計了一連串的實驗，並以第二屆 NTCIR 資訊檢索比賽中文檢索方面的資料進行分類實驗。主

要探討的方向為：1.關鍵詞位置 2.關鍵詞取法 3.關鍵詞過濾 三大方向。

4.3.1 關鍵詞位置：

關鍵詞在文章出現的位置是很重要的討論項目，尤其新聞所強調的重點通常在文章前半部，而後半部有時轉而討論其他議題（表4.3），對於分類判斷並沒有太大的幫助。在實驗中，我們加入了處理文件長度的考慮，觀察特徵的位置對分類的影響。

文章內容
因應匯率變局釋金保守 受到印尼暴動影響，新臺幣貶勢不止，臺北外匯市場波動加劇，為避免炒匯之風再起，中央銀行貨幣政策已明顯由鬆轉緊，昨日有郵匯局轉存款存單質借一百五十億元到期，央行全數收回，而透過公開市場操作釋金一百零六億元，兩者相抵，央行淨收回四十四億元，短期平均利率也跳升至百分之六點九四五，瀕臨百分之七。昨日有一百五十億元郵政儲金轉存款存單質借到期，加上全體銀行體系累計超額準備不足部位達新臺幣四百一十六億元，央行為調節金融市場，昨日透過公開市場操作採固定利率，以附買回方式買進票券一百零六億元，期限二十天，買進利率百分之六點七。**由於目前國際外匯市場局勢混沌，新臺幣貶值壓力大，央行官員表示，在匯市較不穩定時，適度拉高市場短期利率，可避免投機客利用易取得的便宜資金炒匯，以維持匯率的穩定。**
芬蘭電梯公司來臺尋求商機 經過一連串針對亞洲市場所作的測試與研究後，全球排名第三大的電梯公司—通力公司正式宣布在臺灣推出「通力３０００型無機房式電梯」，**將以高級公寓、別墅、辦公大樓、飯店及高層建築區域運輸為主要銷售目標，年銷售量初期訂為三百台至五百台。**

表 4.3 由上面兩篇文件，我們可以發現文章後半段（底線部分）比較屬於討論部分，對於分類無太大幫助。

觀察 NTCIR 的資料我們設計一個參數就是關鍵詞只比對文章前 N 個字(N:100、200、300 與全部比對)。接下來的實驗會有詳細的討論。

4.3.2 特徵取法：

本篇論文利用第三章所提主計處分類架構中的分類描述對每個類別進行特徵詞抽取。我們觀察每個細類的分類說明，發現到最後面都會加上這個類別的性質說明，甚至是產品名稱(請參閱 3.1 節分類架構)。利用這些詳細的類別說明我們進行兩種關鍵詞的取法：

1. 雙連字 ：把所有細類的類別說明斷成雙連詞

例如：陸稻栽培; 水稻栽培;

結果：陸稻　稻栽　栽培　水稻　稻栽　栽培

2. 斷詞 ：把所有的分類說明用斷詞程式加以斷詞（中研院語料庫訓練出來的斷詞程式）

例如：陸稻栽培; 水稻栽培;

結果：陸 | 稻 | 栽培 | 水稻 | 栽培 |

Nb | bb| VC | Na | VC|

4.3.3 關鍵詞過濾：

在這邊我們設計兩種關鍵詞過濾方法，主要針對關鍵詞 Df 值進行過濾，方法如下：

1.Df 值過濾：(Document Df)

我們的訓練資料有一萬筆新聞文件，把所有關鍵詞對這一萬篇文章進行查詢，找出每個關鍵詞出現的文章數目。

若這個關鍵詞出現的文章數目過多，則代表這個關鍵詞沒有分類

的作用，因此我們決定了一個文章篇數的限制。當關鍵詞出現的文章篇數大於 N/C（N：訓練文章篇數 C：類別數目），則把這個關鍵詞過濾掉。

縮減版的分類架構的類別總數為八十九類，而總共我們的訓練資料有一萬筆新聞文件，若這一萬筆新聞文件很均勻的分佈在這八十九類中，則每個分類的新聞文件數目將會是 10000/89 = 112.35，若這個關鍵詞所在的文章數目大於 112.35，代表這個關鍵詞可能分佈在許多類的文章，會導致分類錯誤。

2.類別 QDf 值過濾：關鍵詞出現的類別數目

天下分類架構為三層的階層式分類架構，最小類別總共有八十九類，也就是每個中類下面平均擁有 4.46(89 開三次根號)個小類別。若關鍵詞出現在不同類別的數目大於 4.46，代表這個關鍵詞有可能出現在其他中類別下面的小類別（相同中類別下面的小類別內容、特性相同，表 4.4 為一些不好例子）所以可能導致分類錯誤。

關鍵詞	QDf	類別名
包裝	11	鋼鐵金屬、機械設備、機械及設備租賃、廣告公關、農林漁牧、塑膠、家具用材、食品飲料飼料、法律會計工商顧問、印刷、人力影印保全服務
合成	8	製藥、塑膠、清潔用品及化妝品、紡織服裝、家具用材、食品飲料飼料、化學材料、化工製品
製品	7	鋼鐵金屬、機械設備、紡織服裝、家具用材、食品飲料飼料、其他工業製品、水泥玻璃陶瓷、化工製品

表 4.4 高 QDf 值的關鍵詞

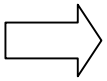
4.3.4 關鍵詞回授

回授機制可以分為兩種：正項回授、負項回授。所謂正項回授是從分類文章中找出關鍵詞並加入原本的類別關鍵詞中。而負項回授則是從分類文章中比對類別關鍵詞，針對不好的特徵使其分數（Weighting）降低。而我們這邊提出一種統計式的正項回授機制並進行實驗與討論。

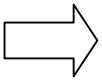
利用已分類之文件，取分數高的進行斷詞，使用原本分類特徵詞之單字特徵字並利用 QDf 值過濾（4.3.3 節中 QDf 值過濾）（表 4.5）留下良好的單字特徵字，再利用單字特徵字取得不錯的回授關鍵詞進行回授（表 4.6）。

分類名稱	關鍵詞
礦業	礦業
礦業	鐵礦
礦業	長石
礦業	河砂
礦業	砂採
礦業	然氣
礦業	煙煤

原本分類特徵詞



分類名稱	特徵單字
礦業	礦
礦業	業
礦業	鐵
礦業	長
礦業	石
礦業	河
礦業	砂
礦業	採
礦業	然
礦業	氣
礦業	煙
礦業	煤



分類名稱	特徵單字
礦業	礦
礦業	業
礦業	鐵
礦業	石
礦業	河
礦業	砂
礦業	採
礦業	然
礦業	煤

表 4.5 由原本分類特徵詞取得單字特徵字並進行 QDf 值過濾後留下的單字特徵詞

分類名稱	關鍵詞
礦業	砂石
礦業	砂石
礦業	紅燈
礦業	院會
礦業	對策
礦業	行政院長
礦業	蕭萬長

分類名稱	關鍵詞
礦業	砂石
礦業	砂石
礦業	行政院長
礦業	蕭萬長

分類名稱	關鍵詞
礦業	砂石
礦業	砂石

表 4.6 關鍵詞過濾說明

4.4 實驗結果

實驗結果發現，天下雜誌縮減主計處分類架構的效果比較好，且雙連字比斷詞效果來的好（表 4.7）。原因我們探討過後認為雙連字串比較能詳盡的分析類別特徵。國際貿易就是一個好的例子。在我們斷詞系統中，國際貿易是可以被斷出來的，然而這個關鍵詞內部還有包含一些不錯的特徵資訊例如：國際、貿易這兩個特徵關鍵詞，而雙連字能夠清楚的表達出國際和貿易的重要性，更能利用”際貿”這個雙連字串表達出國際和貿易之間的關連性，進而有助於分類準確性的提高。

兩種關鍵詞過濾方法過濾不具分類特徵的關鍵詞，有效的提升分類的正確性，證明關鍵詞 Df 值有助於過濾關鍵詞與提升文件分類準確率。

由於本次實驗的資料（共一萬篇）平均長度約為 618 個字，實驗結果發現關鍵詞比對文章前兩百個字就可達到最佳分類結果，因此我們提出關鍵詞只要比對新聞文件前三分之一的內容就可進行文件分類。

天下雜誌分類架構之分類結果

雙連字

比對文章字數	應用率	召回率	精準率
0100(一百字)	29/50	21/50=42%	21/29=73%
0200	35/50	26/50=52%	26/35=75%

斷詞

比對文章字數	應用率	召回率	精準率
W100(一百字)	16/50	9/50=18%	9/16=56%
W200	26/50	15/50=30%	15/26=58%

表 4.7 主計處分類結果

針對已分類文件前三分之一進行斷詞並經由單字特徵關鍵詞過濾取得回授關鍵詞，發現可以取得有助於分類的特徵關鍵詞（表 4.8)。經由回授加入原本分類特徵，再進行分類實驗，發現精準率與召回率都可達到 77% and 68%（表 4.9)，證明單字特徵關鍵詞過濾方法有助於取得良好特徵詞並可提升分類效果。

類別名稱	原本關鍵詞	回授關鍵詞
國際貿易	貿易、入貿、出入、出貿、輸入	國際貿易、出超、出口、進出口
食品飲料飼料	牛奶、牛肉	奶粉、肉鬆

表 4.8 具分類性的回授特徵關鍵詞

回授字數/比對文章字數	應用率	召回率	精準率
100/100	80%	58%	73%
100/200	88%	68%	77%

表 4.9 回授關鍵詞之分類結果

5. 結論與未來研究方向

本文提出一種只需分類架構與分類描述的適應性分類系統，並證實其可行性。另外我們也得到下列幾個結論：(1)關鍵詞取法以雙連詞效果最好，(2)特徵比對的範圍只需文章的前前兩百字或前三分之一就足以標示文件的類別，(3)適當的回授機制可提高文件分類的準確率。

未來我們將針對專有名詞選取。回授機制可以增加新的關鍵詞，而這些關鍵詞確實有助於分類正確性與召回性的提高。然而斷詞系統對於未知詞會有判斷上的問題。如：『陸砂』會斷成『陸｜砂』、『進口砂』斷成『進口｜砂』、『有機感光鼓』:『有機｜感光｜鼓』，而這些未知詞及詞組會降低關鍵詞回授的效果，若能有良好的機制能夠加強為知詞與關鍵詞組的分析，會對分類系統正確率的提升將有很大的幫助。

Reference

1. Chakrabarti, S., Dom B., Agarawal R., and Raghava P. 1997. “Using Taxonomy, Discriminants and Signatures for Navigating in Text Databases.” Proceedings of the 23rd VLDB Conference; Athens, Greece
2. Chen, Aitao, Hailing Jiang and Fredric C. Gey Chinese, 2001. Japanese, and English IR Experiments at NTCIR-2. Chinese Information Retrieval Task, pp 32-40.
3. Chen, J.W.(陳 智 偉), 1998. Methodologies and Analysis for Document Classification. National Tsing-Hua University. master Thesis.
4. Chen, Yu-Jin(陳鈺瑾). 2000. Scalable Summarization for Chinese Text. National Tsing-Hua University, master thesis.
5. Church, K.W. and P. Hanks, 1990. “Word Association Norms, Mutual Information, and Lexicography,” Computational Linguistics, Vol. 16, pp. 22-29.
6. Cooper, W. S., A. Chen, and F. C. Gey. 1994. Full text retrieval based on probabilistic equations with coefficients fitted by logistic regression. In D. K. Harman, editor, *The Second Text REtrieval Conference (TREC-2)*, pp. 57-66.
7. Duda, Richard O., and Peter E. Hart. 1973. Pattern classification and scene analysis. New York: Wiley
8. Gey, F. C., A. Chen, J.He, L. Xu, and J. Meggs. 1996 Term importance, Boolean conjunct training, negative terms, and foreign language retrieval: Probabilistic algorithms at TREC-5. In D. K. Harman, editor, *Text REtrieval Conference (TREC-5)*.
9. Kando, Noriko, Kenro Aihara, Koji Eguchi and Hiroyuki Kato. 2001. Proceedings of the Second NTCIR Workshop Meeting on Evaluation of Chinese and Japanese Text Retrieval and Text Summarization, National Institute of Informatics, Japan.
10. Koller, D. and Sahami M. (1997) “Hierarchically Classifying Documents using Very Few Words”. International Conference on Machine Learning, Volume 14, Morgan-Kauffman.
11. Lewis, D. D., 1996 “Challenges in Machine Learning for Text Classification,” In Processing of the Ninth Annual Conference on Computational Learning Theory, pp. 1.

12. Salton, G. and C. Buckley, 1988. “Term Weighting Approaches in Automatic Information Retrieval,” Information Proceeding and Management, Vol.24, No. 3, pp. 513-524.
13. Salton, G. and M. J. McGill, 1983. Introduction to Modern Information Retrieval, McGraw-Hill, NY, USA.
14. Yang, C.C., J.Yen and H. C. Chen, 2000,“Intelligent Internet searching agent based on hybrid simulated annealing,“ Decision Support System, Vol.28, No.3, pp.269-277
15. Yang, Y., 1997. “An Evaluation of Statistical Approaches to Text Categorization,” Technical Report CMU-CS-97-127, Computer Science Department, Carnegie Mellon University.
16. Yarowsky, 1995.“Unsupervised word sense disambiguation rivaling supervised methods.” In ACL 33,pp.189-196
17. 杜海倫, 1999. “以標題進行新聞自動分類,” 清華大學資訊工程研究所碩士論文, 新竹.
18. 柯淑津, 陳振南, 1999. “階層式文件自動分類之特徵選取研究,” 中華民國八十八年第十二屆計算語言學研討會論文集, pp.137-146.
19. 楊允言, 張俊盛, 陳克健, 1993. “文件自動分類及其相似性排序,” 清華大學資訊科學研究所碩士論文, 新竹.
20. 楊允言, 謝清俊, 陳淑美, 陳克健, 1992. “中文文件自動分類之研究,” 中華民國八十二年第六屆計算語言學研討會論文集, pp.217-233.