致謝

首先，我要向柯淑津老師致最深的謝意，感謝老師這些年來辛苦的教導，讓我改掉許多壞習慣，學會正確的學習方法和態度。也感謝陳培敏老師、陳振南老師以及李御璽老師於口試期間給予的指導與建議，讓我得以完成此論文。

接著我要感謝我最親愛的家人，感謝他們對我求學過程不斷的支持與鼓勵，讓我有勇氣面對學業以及生活上種種的挑戰。對於研究所的同學，祥賓、紘正、皆興、喬毅、鴻儒、介文、耀升、一志以及宣尹，我也很感激大家對我情感上的支持及學業上的協助。最後，我要謝謝可愛的學弟妹們，謝謝你們陪我走過這一、兩年的歲月，也謝謝高中以及大學時代認識的同學和學長們的支持與幫助，讓我得以完成學業。

中文摘要

現今的人們隨時渴望著有用的資訊。但是，面對龐大的資料量，人們往往無法快速有效地找到想要的資訊，為了解決這樣的問題，許多資訊科技應運而生，其中，文件分類的領域是學者們的重要研究議題。當我們先將文件分類後，讀者就可以依需求文件是屬於那個類別，再到那個類別裡頭尋找，即可讓使用者節省搜尋時間。傳統文件分類的做法是由專家對文件內容加以分析了解後之後，再指派類別給文件。時至今日，在面對龐大資料量的情形下，使用人工進行文件分類需耗費極大的人力及物力資源，顯得沒有經濟效益，所以我們需要文件自動分類器，由電腦協助執行這項任務。而所謂文件自動分類的工作，就是針對給定的一篇文件，透過自動化處理來指派適當的類別給予該文件。

文件自動分類通常分為兩個重要的步驟，第一個步驟是特徵的選取，第二個步驟是選擇適合的相關函數，本論文提出以選取共現語詞作為特徵，以及對雙連字串考慮位置因素的方法，希望能夠提高自動分類的精確度和效率。本文主要以這兩種方法來對財經紀事新聞中的標題資料作研究。另外，本研究也提出其他不同的特徵選取方法來作為實驗的對照研究，有選取單一語詞當特徵、雙連字串當特徵、斷詞當特徵、以及考慮斷詞的位置因素等等。

經實驗證明，本文所提出的方法，採取以共現語詞當作特徵選取的策略比起單純使用雙連字串或單純使用斷詞作為特徵的方法，平均約有 15%的提昇，在效能上改善許多。另外，本文的實驗也驗證了我們對文件的觀察，亦即雙字詞比單字詞具有代表性，其次，新聞文件關鍵詞的位置和重要性具有相當正向的關係。

關鍵詞：文件分類、共現語詞、統計式處理、新聞語料。

英文摘要

Nowadays, people are eager to get new information. People can't easily and efficiently find out the wanted information among such huge data. So, we have to classify the documents and then users can efficiently search these documents in the category they belong. Traditionally, by understanding the document experts assign specific categories to that document. However, it costs a lot of resources and has no economic benefits. So, we need an automatic text classifier to heap classification process. Automatic text categorization is the task of assigning predefined categories to free text documents.

In text classification, there are always two important steps. The first step is features selection, and the second one is relevance function selection. Here we propose two techniques to improve the precision of classification by using co-occurrence terms and by considering the positions which bigram occurs. Moreover, this research also provides some other different features selection methods as the contrast for the experiment, including single terms features, bigram features, segmentation features and the position which segmentation occurs.

The experimental result shows that the strategy which uses the co-occurrences as features did perform relatively well. Comparing with using pure bigram, there is about 15% improvement of the performance in average. Besides, the experiment also proves our observation of the texts, that is, bigram is more representative than single terms. In the next place, the positions of the key words have quite positive relation to importance.

Keywords: text classification, text categorization, co-occurrence, statistical processing, news corpus.

目錄

表目錄

圖目錄

1. 簡介

現在是一個資訊爆炸的時代，到處充斥著各式各樣的資訊，尤其網際網路的盛行更加速資訊累積的腳步。當使用者想要尋找需求的資訊時，往往無法有效率地在這茫茫大海中找到想要的資料。而在這大量的多樣化資訊中，「新聞」資訊在人們的生活裡佔了一個相當重要的角色。每日新聞記錄著人們周邊所發生的最新訊息，每個人幾乎每天都需要從新聞報導中吸取新知。但是，由於大家對新聞的選擇有不同的喜好，若讓讀者自己在眾多新聞文件中找尋所需要的資訊，恐怕會造成讀者的負擔。因此，電子新聞通常都會以分類方式呈現，給予讀者使用上的方便。傳統上，為了能夠得到高正確率，文件分類都是由專業領域中的專家學者，以他們的專業知識對文件進行分類。不過，這樣的處理步驟，會浪費龐大的人力、物力資源，且面對現今網際網路上大量湧入的新資訊，人工式文件分類的作法亦無法應付即時性的需求。

單純以人工進行分類除了花費許多的人力資源成本外，人工的判斷也可能不夠客觀、產生不一致性的現象，或者造成誤判的情形。杜海倫的論文中也提到，原始新聞資料的錯誤分類比率相當高，而自動分類系統卻能夠指定正確的類別給予這些文件（杜海倫，1999）。因此，我們急需一個良好的文件自動分類系統，先將資訊分門別類，協助使用者可以快速又正確地擷取所需資訊。另外，文件自動分類除了能提供使用者更方便的檢索處理之外，對於文件摘要、查詢句擴充、查詢句更正、及資訊過濾等多項應用也有很大的助益（林頌堅，1998）。Joachims 等人更提出自動文件分類現在已經被應用在許多不同的領域上，從平常的自動化或半自動化文件索引，到個人的資訊傳遞、過濾不適當的內容、階層目錄下進行網頁分類、語音文件分類、文件類別偵測，甚至辯論出文件的來源偵測（Joachims & Sebasiani, 2002）。Sebastiani 在其論文中也提到，文件自動分類已經由於下述原因，而變得日益重要：1. 與日俱增的數位文件讓自動分類的應用極多且重要；2. 目前有一定數量的文件等待分類，而且需要極短的分類反應時間，這非常符

合自動分類的效益；3. 自動分類可以增加人工分類的生產率；4. 目前的自動分類技術漸趨成熟，可以和人工專家分類並駕齊驅（Sebastiani，2002）。

所謂文件自動分類（TC, Text Classification）的工作，就是針對給定的一篇文件，透過自動化處理來指派適當的類別給予該文件。通常，建構一個分類系統，必須包含特徵的選取以及相關函數設計兩個步驟。首先，必須從訓練語料中選取和文件類別關係密切的特徵出來，再設計適當的相關函數，以計算文件與類別的關聯強度。最後，將與待測文件關聯強度最大的類別指派給此文件。

過去的研究在特徵選取部分，常用單一語詞（single term）作為特徵。本文建議以共現語詞（co-occurrence）作為文件特徵，以雙連字（bigram）作為語詞的單位。一組共現語詞便是由兩個共同出現的雙連字所組成。經實驗證明，共現語詞較之於雙連字串有更好的效能表現。

接下來，本文的第二節將介紹過去有關文件自動分類處理的文獻；第三節描述我們對於財經紀事新聞標題資料（卓越出版社，1992）的觀察；第四節說明我們提出的分類方法；為驗證本研究提出方法之效能，我們設計了四組實驗，在第五節的內容即為我們的實驗規劃、實驗結果，以及相關討論分析；最後，是我們的討論及未來研究方向。

2. 文獻探討

文件自動分類處理，通常分為兩個階段，第一階段先自文件中選取有用的特徵來代表文件，第二階段則是採用適合的相關函數，來量度文件特徵與類別間的關聯度，再將最大關聯度類別指定為文件所歸屬的類別。以下將分別就特徵選取、相關函數以及其他相關方面的文獻作探討。

2.1 特徵選取

一般來說，要建構一套文件自動分類系統，必須先由文件本身自動擷取出足以代表該文件的特徵，再給予各個特徵不同的權重值。特徵的選取可以從文件表面的語詞（字、詞或片語）資訊獲得，例如語詞出現的頻率、語詞所在的位置和語詞的詞性等等。之前的研究多以文件表面的語詞資訊來代表文件的特徵，認為文件中的高頻語詞與文件主題有較高的關聯度，可以取之為特徵（Frakes and Baeza-Yates, 1992），來代表整個文件。1988 年，Salton 等人認為除了語詞在文件出現的頻率（*tf*, term frequency）外，語詞出現的文件數多寡（*df*, document frequency）也應該是很重要的指標。若一語詞只出現在少數文件時，則此語詞對文件的代表性，應具有較高的關聯度；相反地，若一語詞幾乎出現在每篇文件中，就可以假設此語詞對文件類別不具代表性（Salton and Buckley, 1988）。另外，有些研究同時採用 *tf* 與 *df* 來決定語詞對文件的代表性強度，即使用 $tf \times idf$ 來計算語詞與類別之間的相關強度（杜海倫，1999）。

在選取特徵的語詞單位方面，英文的文件分類通常只考慮單一語詞，而不管片語或諺語。中文研究則有斷詞、雙連字串（bigram）及三連字串（trigram）等等不同的單位考量。杜海倫在其論文中表示，二字詞佔的比例高出其他語詞甚多，而且三字詞的重要語意常呈現在其所含的二字詞中，如「研究生」的「研究」是一個二字詞，但已可以表現出「研究生」的特徵，再加上四字詞以上通常是成語或形容詞，與文件主題沒有太大關係，因此她認為將語詞長度限制在二字詞

內，似乎是較好的選擇（杜海倫，1999）。而王稔志和張俊盛在 2001 年的研究，也提出實驗證明，認為在關鍵詞的取法上，雙連字串比斷詞具有較好的效果（王稔志和張俊盛，2001）。在杜海倫的論文中，使用財經紀事新聞語料作實驗，採用 $tf \times idf$ 的方法計算語詞與類別的相關度，在選取雙連字串當特徵的情況下，得到平均 39.96%的精確率；在選取斷詞當特徵的情況下，得到平均 30.7%的精確率。雖然兩者的精確率都不高，我們還是可以由此看出選取雙連字串當特徵的自動分類效能比選取斷詞當特徵的效能好。

另一方面，語詞所帶來的表面資訊雖然可以作為文件分類的依據，但是光靠語詞的字面層次來作分類，可能會有失誤的情形，例如，語詞的同義詞以及歧義問題。以「費用」和「成本」這兩個語詞為例，假若我們只看語詞表面，那就會認為它們是兩個不同的語詞。但是，事實上它們都是指同一個意思。同樣地，「做作」和「忸怩作態」雖然是不同的語詞，但卻是表達相同的意思。假若我們可以分辨出哪些語詞是同義詞，應該對自動分類有很大的幫助。

因為這些語詞可能無法提供完整的資訊以獲得更精確的分類，有學者開始研究以語詞所蘊含的語意來代表文件特徵，例如，Liddy 等人先利用朗文機讀字典轉換語詞成為主題碼（SFC, Subject Field Code）後，再以向量模式表示文件特徵（Liddy, Paik and Yu, 1994）。Hull 以及 Schutze 等人利用隱含語意索引（LSI, Latent Semantic Indexing）的概念，將所有可能的文件都視為在特徵向量空間中的一個向量，再利用奇異值分析（SVD, Singular Value Decomposition）技術，將向量空間中高維的特徵向量轉換成一組較低維的特徵向量（Hull, 1994; Schutze, Hearst and Saund, 1995）。古倫維利用同義詞擴展系統（Thesaurus Expansion System）找到選取出來特徵的同義詞，以增加相似度比對率（古倫維，2000）。

2.2 相關函數

在文件自動分類的研究範疇中，通常我們使用相關函數來計算文件特徵與類

別之間的相關程度，相關程度愈高，表示文件愈有可能屬於此類別。一般來說，相關函數有兩種不同的方法：中心向量法（Centroid）與 k-最鄰近法（k-Nearest Neighbor）。中心向量法是先使用統計方法對訓練資料作一個整體分析，找出各個類別的中心向量，之後再拿個別的測試資料與這些中心向量作比對，取最接近的類別中心之類別當作測試資料的歸屬類別。K-最鄰近方法則是個別找出每筆訓練文件的特徵向量後，直接拿測試文件與訓練資料中所有文件的特徵向量作比較，再選擇與測試資料向量最相像的 k 個訓練資料文件，然後再將該文件的歸屬類別指派給測試文件。可想而知，k-最鄰近方法因為要與整個訓練語料作全面性比對，所以需要較大的計算量。相對的，中心向量法則需要大量的訓練資料，才能使研究結果較客觀。杜海倫在其論文中，使用了中心向量法的 Dice 係數法與 k-最鄰近法的 Centroid Dice 法作實驗，分別得到 24.56%和 34.53%的精確率，證明在新聞標題文件的語料之下，k-最鄰近法的效能優於中心向量法（杜海倫，1999）。

在文件自動分類領域中，有幾個常用的模式，例如向量模式（vector space model）（Borko and Bernick, 1963; 陳淑美，1992; Liddy et al., 1994; Yang and Chute, 1994; Joachims, 1998; Joachims, 2001）與機率模式（probabilistic model）（Maron, 1961; Weiss, Kasif and Brill, 1996; 王稔志等人，2001; Burstein, Marcu, Andreyev and Chodorow, 2001）。向量模式將文件表達成一組向量，而每個類別都由一個中心向量表示，類別中心向量與文件向量在幾何空間的夾角越小，就表示相關度越高；機率模式則使用各種機率的計算來求得文件與類別的關聯度。之後，有人以類神經網路（陳俊凱，1995; Dasigi and Mann, 1996; Ng, Goh and Low, 1997; 黃政偉，1998; 曾祥泰，1998; 陳彥呈和蔣榮先，2001）、模糊理論（楊雪花，1997）以及專家系統（Hayes and Weinstein, 1990; Blosseville, Hebrail, Monteil and Penot, 1992）為基礎來作文件自動分類之研究。其中，Hayes 等人發展的 CONSTRUE 專家系統（Hayes et al., 1990）從路透社新聞語料中選擇性地擷取 3%的科技新聞語料，因此在少量的語料上可以得到相當不錯的成績，平均有 90%的召回率和精

確率。不過，若把 CONSTRUE 用在其他的應用領域上面時，會花費較多成本以及人力資源（Yang，1999）。另外，曾祥泰於 1998 年以並聯式的倒傳遞網路來進行文件的分類學習，實驗結果證實他們提出的分類模組比起單一式倒傳遞網路分類模組有較高的精確度（曾祥泰，1998）。此外，陳智偉認為文件分類與圖形辨識有異曲同工之妙，所以結合數種圖形辨識方法應用在文件分類上，對已量化的文件群作分類（陳智偉，1998），實驗證實 k-最鄰近法雖然需要較大的計算量，效果卻最好。

2.3 其他

除了特徵的選取以及相關函數的適當採用之外，文件自動分類的處理尚有其他的考量點。例如，文件分類的對象、實驗語料的選擇以及分類系統類別間組織的架構。

對於文件分類的對象，有些以全文內容為研究對象（Apte', Damerau and Weiss, 1994; Ng et al., 1997）；有些是取文件的摘要（Maron, 1961; Kar and White, 1978; Yang, 1996）；有些以文件的題目為實驗對象（Hamill and Zamora, 1980; 陳淑美，1992; 蔡憲文，1998）；杜海倫利用統計方法，針對新聞文件標題的特殊用詞特點，找出語詞與文件歸屬類別之間的關係，進而讓新聞分類達到自動化目的（杜海倫，1999）；Kwok 在對醫藥文章進行自動分類時，則是同時考慮文件的摘要以及題目（Kwok, 1975），並且證明僅考慮少量的標題文件，就可以得到不錯的叢集分散效果。

實驗的語料選擇方面，新聞文件因為有其不同於一般文件的特點，例如，為了讓民眾容易閱讀以及瞭解，所以通常用字較簡單、標題長度簡短易懂、每篇新聞要表達的主旨明顯，以及人、事、時、地、物出現在文件偏高的比率，故常被用來作為文件分類的研究資料。英文文件分類研究最常使用的是路透社（Reuters）的新聞資料（Frakes and Baezay-Yates, 1992; Lewis, 1992; Koller and Sahami, 1997;

Ng et al., 1997; Yang, 1997; Alessio, Murray, Schiaffino and Kershenbaum, 1998; Lodhi et al., 2002)，其他則有期刊科技性論文（Maron, 1961; Hamill et al., 1980）以及 OHSUMED 語料庫（Yang, 1997）等。中文文件方面則有取自民國八十一年各大報紙新聞標題資料的財經紀事新聞（陳淑美，1992; 陳智偉，1998; 杜海倫，1999)、醫藥新聞（黃仲瑋，1999)、中央通訊社八十四年出版的通訊稿（楊雪花，1997）以及中華民國科技期刊論文摘要（蔣俊霞，1994）等。

另外，分類系統類別間的組織架構也是文件分類的議題之一。以往的研究在類別組織上大都採用單階層式架構，亦即線性分類，而不考慮類別與類別之間的關聯情形（Frakes and Baeza-Yates, 1992; Yang, 1997)。除此之外，還有學者提出階層式文件自動分類系統，考慮類別之間的關係而形成階層式的組織架構（ Yang 1996; Chakrabarti et al., 1997; Koller and Sahami, 1997; Ng et al., 1997; D'Alessio et al., 1998; 陳彥呈等人，2001)。1999 年有學者先初步選完特徵集後，再依各特徵與相近類別間所具的分類意義做適當的調度，實驗結果顯示階層式分類的效能優於線性分類（柯淑津，陳振南，1999)。

3. 文件的觀察

我們觀察新聞語料後，發現存在著以下幾個現象：雙字語詞比起單一語詞在語意上往往更為明確，共現語詞與類別的關係比起單一語詞為強，還有語詞出現的位置也常扮演著重要的角色。這些現象我們將分別在以下各小節說明。

3.1 雙字詞較單字詞具代表性

我們從語料中發現雙字詞比單字詞在語意上更具明確性，因為許多的單字詞往往具一字多義的情形，而雙字詞的意義通常較確定，以下面的三則文件標題為例：

例 1 「公共建設」類：執政黨台北市黨與中國時報合辦「基層建設座談會」專題報導之(6)/文山區:合併景美、木柵、文山向繁榮邁進

例 2 「金融」類：央行無意引導國內利率下跌,謝森中說,美國調低重貼現率是為刺激景氣,我國貨幣政策則以穩定物價為優先

例 3 「農業」類：宣揚中華美食我農產品向美出擊,因應加入 GATT 後進口貨衝擊,農委會未雨綢繆先尋出路

例 1 到例 3 中的「美」分別代表「地名」、「國名」以及「美好」，假若只看單字詞，那麼它們都是「美」，如此將混淆它們真正代表的意義，但是，若以雙字詞作為處理單位，意義就會被區分開，「景美」、「美國」與「美食」很清楚地代表「地名」、「國名」以及「美好」。

另外，例 4 到例 6 的「長」字也有同樣的情形，它們個別代表「人名」、「職

稱」及「時間量度」，若只看單字詞，則都是「長」，無法看出它們真正的意義，若以雙字詞作為處理單位，就能明確知道其表示的意義。

例 4 「經濟」類：落實六輕核四,81 年經濟重點,蕭萬長新年表達新希望,將視大陸善意回應情形,調整兩岸經貿步調

例 5 「政府、政治」類：正副院長改選此其時:立院應即時完成權力與經驗傳承

例 6 「稅賦」類：鼓勵長期投資、抑制短線投機,財部正式函告立院千分之 6 證交稅不宜降低 ,顧及券商經營,同意可出租閒置土地,購買定額開放型受益憑證

除此之外，我們也從字典的觀察中，發現同樣的情形。以「下」字為例，當單字詞時，有幾種不同的解釋，分別為「當兒」、「日後」、「底端」、「記」、「鄙」、「副」、「投」、「脫」、「拆」、「攻佔」、「發出」、「使用」、「出來」及「生產」等。雙字詞時，則個別有其明確的解釋，例如：「下人」指的是「奴僕」，而「下凡」、「下工」、「下午」、「下任」以及「下毒」分別指「降謫人間」、「下班」、「午後」、「卸任」及「放毒」。

在同義詞詞林[註1]（梅家駒等，1993）中，單字詞共計有 3,921 個，每個單字詞平均擁有 2.19 個不同的字義。而雙字詞共計有 33,488 個，每一個雙字詞平均有 1.17 個不同的詞義。也就是說，單字詞歧義的現象比雙字詞來的普遍。

由以上的說明可以知道，我們常常無法從單字詞看出清楚的意思，而雙字詞

[註1] 同義詞詞林一書收錄詞語近七萬，全部按意義進行編排，為一本類義詞典。分類原則：以詞義為主，兼顧詞類，並充分注意題材的集中。全書將詞語分大、中、小類三級，共分 12 個大類，1428 個小類，小類下再以同義原則劃分詞群，每一個詞群以一個標題詞立目，共 3925 個標題詞。除此之外，本書除了收錄同義詞，尚收了同類詞。同類詞指詞義上屬同一範疇的詞語。

比起單字詞可以提供較明顯的意義。

3.2 共現語詞的特性

一般的文件自動分類研究皆使用單一語詞為選取的特徵，由上一個章節的說明，我們可以知道新聞語料中的雙字詞比單字詞在語意上更確定，應該有助益於文件的自動分類。假若將單一語詞合併起來變成共現語詞，則自動分類使用共現語詞當特徵之效能應該比單一語詞當特徵之效能來的好。經由我們對文件的觀察，也發現共現語詞比單一語詞更具文件的代表性，而且我們發現某些語詞的配對經常被使用在某些特定類別。我們以下面的幾個例子來說明共現語詞與單一語詞在語詞與類別間的關聯度差異：

例 7 「國際政經」類：國際景氣回顧與前瞻系列報導(5):取代舊蘇聯崩潰邊緣經濟;資源配置不當矛盾更為凸顯,新國協政策若不協調,經濟將更惡化

例 8 「人物檔案」類：跟著景氣趨勢調整經營觸角,陳盛沖要抓緊年輕人的心(聲寶公司董事長)

例 9「關稅」類：修正海關進口稅則公告取消退稅並停止按內銷比率課稅貨品項目清表

例 10「外匯」類：如果未來 1 年出超與入超無法縮小差距,新台幣匯率有持續升值壓力

以例 7 來看，假若我們只看單一語詞，就很容易因為「景氣」、「經濟」以及「配置」等語詞，而將這則新聞文件指派給「經濟」類別。假若我們考慮共現語詞的話，可以發現「國際」與「蘇聯」、「國際」與「國協」以及「蘇聯」與「國協」等共現語詞都和「國際政經」類別有相當高的關係，這樣就可以將此則文件正確地分類。

我們再以例 8 來看，只考慮單一語詞的時候，將因為「趨勢」、「經營」、「聲寶」等單一語詞，而將這則新聞文件分到「股票」類別，而正確類別應該為「人物檔案」類別。要是我們能夠考慮共現語詞，就會很明顯地看到「陳盛沖」與「聲寶」、「陳盛沖」與「公司」、「陳盛沖」與「董事長」、「董事長」與「聲寶」以及「公司」與「董事長」這些共現語詞與「人物檔案」類別有相當高的關係，可以幫助我們的自動分類。

例 9 中的「進口」、「課稅」、「修正」、「內銷」、「項目」、「公告」等關鍵詞與各類別的關係強度都不太一樣，例如：語詞「進口」在「各項產業」、「經濟」、「貿易」等類別都有滿高的關聯度；語詞「課稅」與「稅賦」類別有高關聯度；語詞「內銷」在「經濟」和「各項產業」類別有高關聯度。因此，假若我們只採取單一語詞當作特徵選取的單位，那麼這篇文件就會被系統錯分至不對的類別。假若我們採取共現語詞當作選取的特徵時，上述的問題就不會發生，自動系統也就能夠正確地分類，因為「海關」與「進口」、「進口」與「稅則」、「海關」與「稅則」以及「海關」與「課稅」等共現語詞都與「關稅」類別有很高的關聯度，其餘的共現語詞對於自動分類則沒有什麼影響力。由以上的例子說明，我們可以發現，共現語詞比起單一語詞在新聞文件更具有代表性。

最後，以例 10 為例，單一語詞「台幣」與「金融」類別有相當強的關聯度，而且語詞「壓力」和「政府,政治」類別有很高的關係，這樣會導致自動分類的錯誤。若以共現語詞來看，則「台幣」與「出超」、「台幣」與「入超」、「台幣」與「匯率」、「台幣」與「升值」、「匯率」與「升值」以及「出超」與「入超」等等共現語詞都和「外匯」關係密切，可以選取出來當特徵以代表這則文件，這樣應該會對自動分類有所幫助。

由以上的例子可以知道，共現語詞常常比單一語詞更能決定一則文件的類別特性，因此，比較可以幫助新聞文件自動分類的進行。

3.3 位置的重要性

新聞文件關鍵詞的位置與重要性有很大的關係，較重點的關鍵詞常常出現在文件中較前面的位置，而在後面位置的語詞則其重點性較低，甚至有可能會誤導分類，以下面的幾個例子來看：

例 11 「人物檔案」類：章孝慈人生曲徑堅定行:宦海波濤漸遠,學海航向無涯

例 12「人物檔案」類：細數風雲人物系列(5):環保執行者施展魄力抓污染,趙少康取締污染不論公民營或地下工廠均一視同仁力求落實執法

例 13「股票」類：上市公司剖析系列報導:業外投資獲利估算逾 3 億,石膏板廠將量產,環球水泥紮底,跨足金融闢財源

例 14「股票」類：上市公司下半年景氣仍是幾家歡樂幾家愁,台幣匯率、建國、兩岸互動等因素,對個別產業影響仍深,營運因此展現不同風貌

例 15「勞工」類：勞委會提出「勞工福利促進法」草案,有關提撥職工福利金規定,對大企業採強制方式,對中小企業採鼓勵性質

其中，例 11 是屬於「人物檔案」檔案類別，關鍵詞是「章孝慈」，然而後面位置的語詞「波濤」可能將此篇文件誤導至「各項產業」類別；語詞「學海」可能將之誤導至「其他」下的「教育」類別。例 12 原屬於「人物檔案」類別，有利於正確分類的關鍵詞「人物」和「趙少康」都位於標題文件的較前面位置，後面位置出現的「環保」、「污染」等語詞的影響，容易將分類系統誤導至「環境」類別。例 13 和例 14 的正確類別為「股票」類別，當我們只以純粹的語詞當特

徵，沒有考慮位置因素時，發現分類系統可能會因為「板廠」和「水泥」等語詞，而將例 13 錯分至「經濟」類別，或者因為「金融」而將之錯誤分類至「金融」類別；考慮位置因素時，由於「上市」及「公司」等出現於較前面位置的語詞，經過位置因素的權重值加權之後，在「股票」類別會擁有較高的權重值，所以能夠將之正確分類。

另外，例 14 可能因為位於文件後面位置的「兩岸」和「產業」等語詞，而被錯分到「經濟」類別。例 15 的正確類別為「勞工」類別，假若不考慮位置因素，沒有特別加權給位於前面位置的「勞委會」以及「勞工」等關鍵詞，則文件可能會因為位於文件後面位置的語詞「企業」而被錯分至「經濟」類別。而且此篇文件一共包含了兩個「企業」，故使得整篇文件落在「經濟」類別。當我們考慮語詞的位置時，語詞「企業」因為位於文件較後面的位置，所以位置加權後的權重值被拉低，相反地，位於文件較前面位置的「勞委會」以及「勞工」等語詞，待位置加權後，皆拉高此篇文件與「勞工」類別的關聯度，所以就正確分類至「勞工」類別。

4. 研究方法

文件分類研究通常分成訓練與測試兩個階段進行。訓練階段(training stage)主要的任務是擷取語料中有用的知識；測試階段(testing stage)則利用這些知識，給予待分類的測試文件一個應屬的類別。在訓練階段，首先面對的問題是如何辨識出最能突顯文件所屬類別特色的內容作為特徵。之後，於測試階段擷取這些特徵以代表待測文件，再利用相關函數計算文件與類別之間的關聯度，達成文件自動分類的目標。

由第三章的文件觀察，我們可以知道，雙連字串比單字詞在語意的表達上更為明確、共現語詞與類別的關係比單一語詞與類別的關係強，而且，語詞在新聞文件中出現的位置常常與其重要性有關聯。所以本文以共現語詞當作選取的特徵以及使用雙連字串的位置變數當作研究的對象。以下，我們先分別就本論文所使用的特徵選取以及相關函數作說明，接著，我們分別在第三、四節中提出本研究在訓練階段與測試階段所作的處理。

4.1 特徵選取

本論文在特徵選取步驟中，考量了共現語詞的影響。通常，文件分類相關的研究都是對單一語詞或者片語（phrase）作為語詞特徵的選取。本論文的研究，不只是挑選單一語詞當特徵，還探討共現語詞對自動分類的幫助。

本文採用中心向量法中的 $tf \times idf$ 來計算文件與類別之間的相關程度，即使用 $tf \times idf$ 來評估每個語詞的重要程度。假設給予一個文件 d，以及出現在 d 中的特徵語詞 a。我們決定 a 在 d 中所具的重要性，為 a 在 d 中的出現次數 tf 乘以特徵語詞 a 本身的重要性 idf。

首先，我們先對訓練資料的文件一一拆成雙連字，再將每篇文件中的語詞作兩兩配對成一組，這些配對的語詞就是所謂的共現語詞。之後，我們使用統計方法，計算每組共現語詞出現在各類別的頻率，頻率高者，我們就覺得此組共現語

詞與這個類別相關程度較高。例如，「大陸」與「投資」這組共現語詞在「經濟」類別擁有高頻率值，我們可以說「大陸」與「投資」這組共現語詞在「經濟」類別是相當具有代表性；同樣的情形也發生在「政黨」與「執政」這組共現語詞，「政黨」與「執政」這組共現語詞在「政府、政治」類別有高頻率的出現，表示「政黨」與「執政」這組共現語詞與「政府、政治」類別有相當高的關聯度。而「理報」與「權威」這組共現語詞在「國際政經」類別就只有出現一次，很明顯的，這組共現語詞和此類別的關聯度非常小。另外，我們也考慮每組共現語詞在整個訓練語料的出現類別數，類別數多者表示這組共現語詞比較不具代表性，亦即不能突顯文件與類別的關係。例如：在所有的 38 個類別當中，「系列」與「報導」這組共現語詞就曾出現在 36 個類別中，表示「系列」與「報導」這組共現語詞與類別之間並沒有特殊的關聯性；而「外勞」與「核定」這組共現語詞只有出現在「勞工」及「各項產業」兩個類別中，可以看得出來，「外勞」與「核定」這組共現語詞和「勞工」以及「各項產業」這兩個類別有相當高的關聯度。

本研究計算訓練資料中共現語詞的頻率，為每一個類別選取適當的特徵集，且依據每個特徵語詞 a 與類別 c 間的相關程度，給予權重值 $\mathrm{W}(a, c)$。本文分別選取單字詞、雙連字串、雙連字串組成的共現語詞、斷詞、及以斷詞所組成的共現語詞為特徵，它們的權重計算法分別於下面各小節說明。

4.1.1 單字詞的權重計算

單字詞 s 在類別 c 的權重值為 $\mathrm{WS}(s, c)$，是由單字詞 s 出現在類別 c 中的頻率值 tf 乘以單字詞 s 本身的重要性 idf 得到。如公式（1）和公式（2）所示。

$$\mathrm{WS}(s,c) = tf_{s,c} \times idf_s \text{，} \quad (1)$$

$$idf_s = \log\left(\frac{T}{df_s} + 1\right)\text{，} \quad (2)$$

其中，　$tf_{s,c}$　：單字詞 s 出現在類別 c 中的頻率值，

df_s　：單字詞 s 出現的類別數，

$\mathrm{WS}(s,c)$　：單字詞 s 在類別 c 的權重值，

T　：所有類別總數。

4.1.2 雙連字串的權重計算

由字詞 term_1 與字詞 term_2 所組成的雙連字串 b 在類別 c 的權重值為 $\mathrm{WB}(b, c)$，由雙連字串 b 出現在類別 c 中的頻率值 tf 乘以雙連字串 b 本身的重要性 idf 得到。如公式（3）和公式（4）所示。

$$\mathrm{WB}(b,c) = tf_{b,c} \times idf_b \quad , \tag{3}$$

$$idf_b = \log\left(\frac{T}{df_b} + 1\right) \quad , \tag{4}$$

其中，　$tf_{b,c}$　：雙連字 b 出現在類別 c 中的頻率值，

df_b　：雙連字 b 出現的類別數，

$\mathrm{WB}(b,c)$　：雙連字 b 在類別 c 的權重值，

T　：所有類別總數。

4.1.3 共現語詞的權重計算

由雙連字串 b_1 與雙連字串 b_2 所組合而成的共現語詞 t 在類別 c 的權重值為 $\mathrm{WC}(t, c)$，由共現語詞 t 出現在類別 c 中的頻率值 tf 乘以共現語詞 t 本身的重要性 idf 得到。如公式（5）和公式（6）所示。

$$\mathrm{WC}(t,c) = tf_{t,c} \times idf_t \text{，} \quad (5)$$

$$idf_t = \log\left(\frac{T}{df_t} + 1\right)\text{，} \quad (6)$$

其中， $tf_{t,c}$ **：共現語詞** t **出現在類別** c **中的頻率值，**

df_t **：共現語詞** t **出現的類別數，**

$\mathrm{WC}(t,c)$ **：共現語詞** t **在類別** c **的權重值，**

T **：所有類別總數。**

4.2 相關函數計算

相關函數的計算方面，我們先從待測文件 d **中擷取能夠代表該文件的特徵** a**，再到訓練語料中找尋這些特徵** a **與類別** c **的關聯度** $\mathrm{W}(a, c)$**。並針對各個類別，加總所有的** $\mathrm{W}(a, c)$**，求得文件** d **與類別** c **的相關強度** $\mathrm{R}(c, d)$**。接著，取最大相關強度的類別為指派類別。本文針對位置因素的考量與否而個別使用不同的兩個相關函數計算公式，分別說明於** 4.2.1 **小節及** 4.2.2 **小節。**

4.2.1 不考慮特徵出現位置的相關函數計算

在不考慮語詞出現在文件位置的情況下，我們採用公式（7**）來計算待測文件與類別之間的相關函數。**

$$\mathrm{R}_1(c,d) = \sum_{a \in d} w(a,c) \quad \text{，} \quad (7)$$

其中， $\mathrm{w}(a, c)$ **：特徵語詞** a **在類別** c **的權重值，**

$\mathrm{R}_1(c,d)$ **：文件** d **與類別** c **的相關強度。**

4.2.2 考慮特徵出現位置的相關函數計算

在考慮語詞出現在文件的位置因素時，因為我們認為較前面位置的語詞較為重要，所以將測試文件中語詞位置的序數 *order* 開根號後取倒數當作位置權重，即公式（8）。其文件與類別的相關強度公式如公式（9）所示。

$$p_b = \frac{1}{\sqrt{order_b}} \quad (8)$$

$$R_2(c,d) = \sum_{b \in d} w(b,c) \times p_b, \quad (9)$$

其中， $order_b$ ：特徵 b 在文件內的位置序數，

p_b ：特徵 b 在文件內的位置權重，

$w(b, c)$ ：特徵語詞 b 在類別 c 的權重值，

$R_2(c,d)$ ：文件 d 與類別 c 的相關強度。

4.3 訓練階段

在訓練階段，我們將訓練語料中的每則新聞標題，依照其實驗所選取的特徵單位，分解成一個個的語詞。在我們的研究中，語詞代表的單位，可能是單字詞、雙連字串、斷詞、雙連字串合併成的共現語詞或者斷詞合併成的共現語詞。擷取代表各新聞標題文件的語詞之後，我們計算這些語詞在訓練語料中的頻率值以及語詞本身的重要性。之後，利用公式（1）至公式（6）等權重公式來計算各語詞與類別的關聯權重值。詳細的訓練階段部分之演算法如下圖所示。

訓練階段的演算法

步驟一：將訓練語料中的新聞標題文件斷成一個個的語詞。

步驟二：計算各個語詞出現在各個類別的頻率（tf)。

步驟三：計算各個語詞出現的類別數（df)。

步驟四：利用權重公式（1）至（6）來計算各語詞與類別的關聯權重值。依語詞單位為單字詞、雙連字串、斷詞或共現語詞，而分別使用公式（1）（2)、(3)　（4）或（5）　（6)。

圖 1　訓練階段的演算法

以下面的例 16 訓練文件為例，在步驟一的時候，我們將此新聞標題文件利用斷詞系統斷成一個個的語詞，變成「憲改 與 議會 民主 的 迷思 從 煽動 者 到 政治 的 新 現實」等十四個語詞。接著執行步驟二，即計算這十四個語詞出現在各個類別的頻率值以及計算這些語詞出現的類別數。最後，執行步驟四，利用權重公式（1)、(3）或（5）來得到這些語詞與每個類別的關聯度權重值。

例 16「政府,政治」類：憲改與議會民主的迷思:從煽動者到政治的新現實

例 17「政府,政治」類：梁肅戎:資深立委功過,留給歷史裁斷,臨別提出 3 點期許:籲立委認同國家、致力調和國家利益與地方利益、促朝野政黨加速關係正常化

4.4 測試階段

在測試階段，我們將測試資料中的每則測試新聞標題文件，依照其實驗所選

取的特徵單位，分解成一個個的語詞。再利用公式（7）或公式（9）等關聯強度公式來計算測試文件與各個類別的關聯強度：假若實驗不考慮語詞在文件的位置關係，則直接使用公式（7）來計算關聯強度；假若實驗有考慮語詞在文件的位置關係，則必須在使用公式（9）之前，先利用公式（8）計算出各語詞在測試文件中的位置序數，再將此位置序數開根號後取倒數當作位置權重。之後，我們將擁有與測試文件最高關聯強度的類別指派給測試文件，當作其所屬類別。詳細的測試階段演算法如下圖所示。

測試階段的演算法

步驟一：將測試文件分解成一個個的語詞。

步驟二：假若實驗有作位置考量的話，就先利用公式（8）計算每個語詞在文件中的位置序數，若不考慮語詞出現的位置，則以 1 當所有語詞的位置序數。之後，再使用公式（7）或（9）計算文件與各類別的關聯強度。

步驟三：將擁有最高關聯強度的類別指派給測試文件。

圖 2　測試階段的演算法

以上面的例 17 測試文件為例，我們依步驟一將之分解成「梁肅戎　資深　立委　功過　留給　歷史　裁斷　臨別　提出　點　期許　籲　立委　認同　國家　致力　調和　國家　利益　與　地方　利益　促　朝野　政黨　加速　關係　正常化」。再來則利用公式（12）來計算這則測試文件與各個類別的關聯強度。最後在步驟三時，將擁有最高關聯強度的類別指派給測試文件。

接著，我們以雙連字串為特徵單位，並且考慮語詞在文件中的位置因素，以本篇測試文件為例，在第一步驟時，我們將此文件分解成如表 1 所描述的第一個欄位。再來，利用公式（13）計算每個語詞在文件中的位置序數，例如「資深」這個語詞位於此分解後的文件之第三個位置，則它的位置序數 $order_w$ 就等於 3，

然後使用公式（12）計算文件與各類別的關聯強度。最後在步驟三時，將擁有最高關聯強度的類別指派給測試文件。

表 1　雙連字串、其位置序數及其在文件內的位置權重

雙連字串	位置序數	p_b	雙連字串	位置序數	p_b	雙連字串	位置序數	p_b
梁肅	1	1	點期	17	0.2	與地	33	0.2
肅戎	2	0.7	期許	18	0.2	地方	34	0.2
資深	3	0.6	籲立	19	0.2	方利	35	0.2
深立	4	0.5	立委	20	0.2	利益	36	0.2
立委	5	0.4	委認	21	0.2	促朝	37	0.2
委功	6	0.4	認同	22	0.2	朝野	38	0.2
功過	7	0.4	同國	23	0.2	野政	39	0.2
過留	8	0.4	國家	24	0.2	政黨	40	0.2
留給	9	0.3	致力	25	0.2	黨加	41	0.2
給歷	10	0.3	力調	26	0.2	加速	42	0.2
歷史	11	0.3	調和	27	0.2	速關	43	0.2
史裁	12	0.3	和國	28	0..2	關係	44	0.2
裁斷	13	0.3	國家	29	0.2	係正	45	0.1
臨別	14	0.3	家利	30	0.2	正常	46	0.1
別提	15	0.3	利益	31	0.2	常化	47	0.1
提出	16	0.3	益與	32	0.2			

5. 實驗

本文提出以共現語詞當作選取的特徵，以及使用雙連字串的位置變數當作研究的對象，為驗證本論文所提出方法之效能，我們設計了六組實驗，用來比較不同特徵以及位置的考量之下，對文件自動分類效果的影響。接下來，我們就本論文的實驗資料、實驗設計以及實驗結果進行說明與討論。

5.1 實驗資料

本論文使用的中文新聞語料來自『財經紀事』（卓越出版社，1992）中所含的新聞標題，其內容取自民國八十一年間中國時報、工商日報、聯合報、民生報等各報社之新聞標題，共含有 124,940 則新聞標題，每則新聞標題皆經過人工標示所屬類別。這些標題採三層式分類：大分類、中分類以及小分類。其中，大分類共分為「公營事業篇」、「服務業篇」以及「金融業篇」等九大類別；大類別下細分為 38 個中類別，例如「金融業篇」類別之下含有「金融」、「銀行」、「外匯」、「股票」以及「租賃」等中類別；中類別下又細分成小類別。附錄一為這些新聞資料的介紹。本研究為避免類別過於細分，僅就中類別層進行實驗，並以隨機方式取樣其中的百分之十為測試資料，其餘的百分之九十為訓練資料，即訓練資料共含 112,446 則標題，測試資料含 12,494 則標題。

5.2 實驗設計

本文為了探討共現語詞以及位置因素對新聞文件自動分類的影響，我們一共設計了六組實驗，分別取單字詞、單純的雙連字串、以雙連字串配對而組成的共現語詞、單純的斷詞、還有以斷詞配對而組成的共現語詞當作特徵，並且在雙連字串方面加入位置的考量。

本論文為了得到實驗要用的斷詞特徵，所以使用中研院詞庫小組所發展的斷

詞系統[註 2] 對文件作斷詞的處理。斷句的原則是以「句」為單位，而斷句是以標點符號為參考點，所以文章的句子之間一定要有標點符號分隔。空白鍵及換行符號不被視為分隔符號，在執行斷詞過程中會被刪除。本文所使用的斷詞系統認得的標點符號有”，”、”。”、”；”、”：”、”！”、以及”？”等。接下來，我們詳述本文的實驗如表 2 以及其下的說明。

表 2　實驗介紹

實驗	特徵選取	位置的考量與否
實驗一	單字詞	無
實驗二	雙連字串	無
實驗三	雙連字串配對而成的共現語詞	無
實驗四	雙連字串	有
實驗五	斷詞	無
實驗六	斷詞配對而成的共現語詞	無

在實驗一，我們將訓練語料分解成一個個的單字詞，再計算這些單字詞的 $tf \times idf$ 作為語詞與類別相關程度的權重值。接著，將待測文件以單字詞表示，由待測文件中的單字詞去找尋訓練語料中相同的語詞，並得到相對應的權重值，再針對每個類別作加總，以得到待測文件與各個類別的關聯度分數，我們取關聯度分數最高的類別為指派給待測文件的類別。而實驗二的做法和實驗一相似，只是實驗一使用單字詞當作選取的特徵，實驗二使用雙連字串當作選取的特徵。實驗三的做法和實驗二相似，不過我們先將雙連字串一一配對組成共現語詞，之後就

[註 2] 斷詞系統由中央研究院中文詞庫小組所發展而成。本文採用的版本為「中文自動斷詞系統 1.0 版」，出版日期為西元 1999 年 11 月 1 日。中文自動斷詞系統具有中文自動斷詞以及中文詞類自動標記功能，使用者可以根據自己的需求，而選擇不同的詞典，作為斷詞及標記系統的參考。在這個版本中，內建有中央研究院詞庫小組的中文詞庫，以及從中央研究院平衡語料庫中所抽取的額外詞條（約二萬目詞），外加定量詞及重疊詞構詞律，以供使用者斷詞時的使用。

取這些共現語詞當作選取的特徵。至於實驗四，其做法和實驗二類似，但是實驗二沒有考慮位置因素，實驗四則將特徵關鍵詞在文件的位置因素考慮進來。實驗五的做法和實驗二相似，不過實驗二使用的特徵是雙連字串，實驗五使用的特徵是中研院斷詞系統處理後的斷詞。最後，實驗六的做法和實驗三類似，實驗三使用雙連字串配對組成的共現語詞，實驗六使用斷詞配對組成的共現語詞。

5.3 門檻值的設定

本研究在未對共現語詞設定任何門檻值的情況之下，共現語詞約有兩千六百萬組。為了解選取共現語詞的特徵數目對精確率的影響，我們另外作了 17 組實驗，對每種類別各自取擁有最高權重值的前 n 組共現語詞出來當作特徵，n 分別為 20、40、60、…、300、320 及 340。我們可以從圖 3 看到整個精確率變化的情形，？ 軸的數值 10000 表示沒有設定門檻值的情況，為了方便閱讀，我們只取出前五個數列，即涵蓋率從 10％到 50％這五組數列。整體來看，此圖在門檻值為 280 組的時候趨近穩定，加上特徵數的考量，所以本文建議採用門檻值為 280 組，也就是總共 280×38 組，即共現語詞特徵數為 10,640 組。

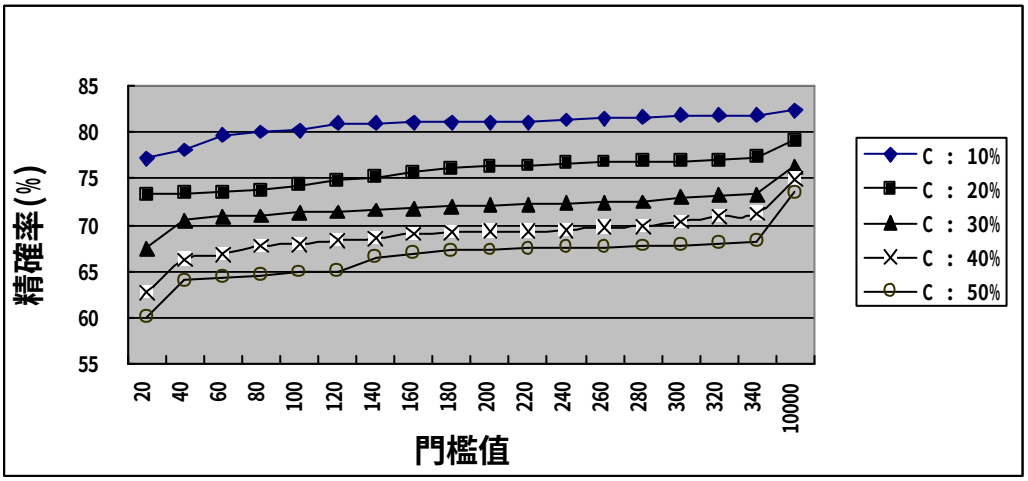

圖 3　不同的特徵選取門檻值設定對系統精確率的影響

5.4 實驗評估與討論

我們將實驗的結果以涵蓋率 C（Coverage）、精確率 P（Precision）來進行評估。兩者的公式分別如公式（14）及公式（15）所示。涵蓋率顯示系統能分類的文件數比率；精確率表示找回來的文件中，被系統正確分類的文件數比率。表 9 是本文所有實驗在不同涵蓋率下的精確率值，其中，由於本分類系統的涵蓋率只達到 99 左右，所以本表只呈現到涵蓋率為 99 的精確率。

$$\text{涵蓋率} = \frac{\text{可以標示分類的文件數}}{\text{所有測試文件的總數}} \times 100\%, \quad (14)$$

$$\text{精確率} = \frac{\text{正確分類的文件數}}{\text{可以標示分類的文件數}} \times 100\%, \quad (15)$$

表 3　不同涵蓋率下的精確率

涵蓋率（%）	實驗一	實驗二	實驗三	實驗四	實驗五	實驗六
10	42.99	62.85	82.39	70.54	61.01	76.46
20	37.39	59.01	79.06	64.45	58.73	74.46
30	35.87	58.31	76.30	61.20	56.61	72.51
40	34.21	56.65	74.88	59.71	54.02	71.00
50	32.95	55.58	73.53	58.67	52.36	69.61
60	31.69	54.20	72.48	57.85	51.17	68.55
70	30.93	53.75	71.20	56.81	50.28	67.61
80	30.20	53.18	70.03	55.92	49.60	66.87
90	29.71	52.45	68.74	54.79	48.84	65.82
99	28.65	50.91	65.58	52.99	48.30	63.91
平均	33.46	55.69	73.42	59.29	53.09	69.68

首先，我們看實驗一、實驗二和實驗五的實驗結果，由圖 4 可以清楚看出，當我們使用單字詞當作選取的特徵時，獲得最差的結果，不管是使用雙連字串的實驗二或者使用斷詞的實驗五，實驗結果效能都比實驗一來的好。因此，我們可以將實驗一當作本文的基準實驗。另外，我們也可以看出，實驗二雙連字串當特徵的表現比實驗五斷詞當特徵的表現好。

接著，我們比較單純使用雙連字串當特徵的實驗二以及加入位置考量的實驗四，我們由圖 5 可以看到有加入位置考量的精確率高於沒有加入位置考量的實驗二。證明我們提出的特徵在文件中位置的變數因素的確有助於自動分類。

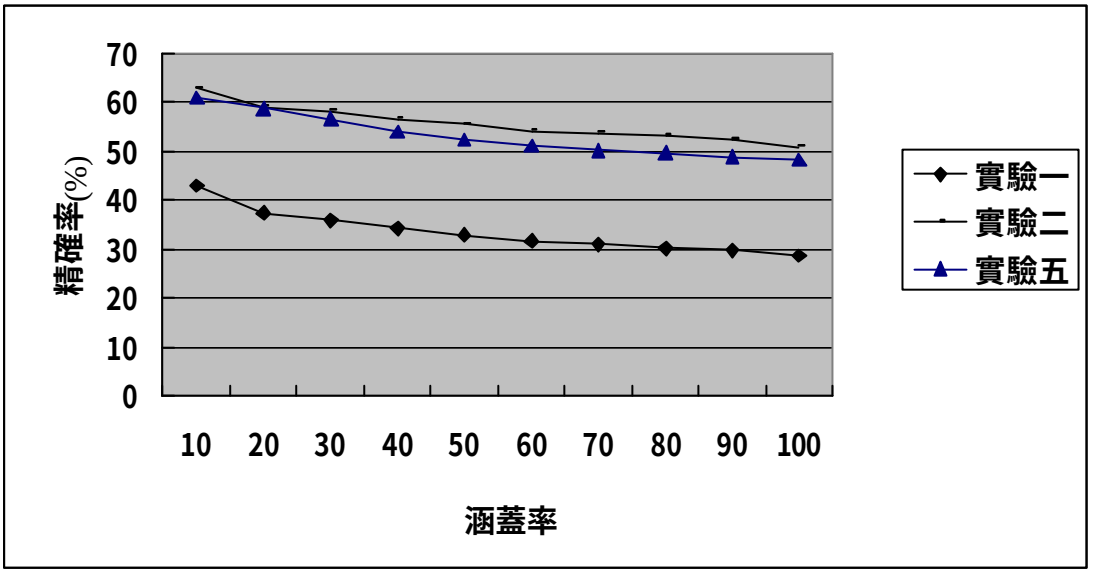

圖 4　　實驗一、實驗二和實驗五的結果數據

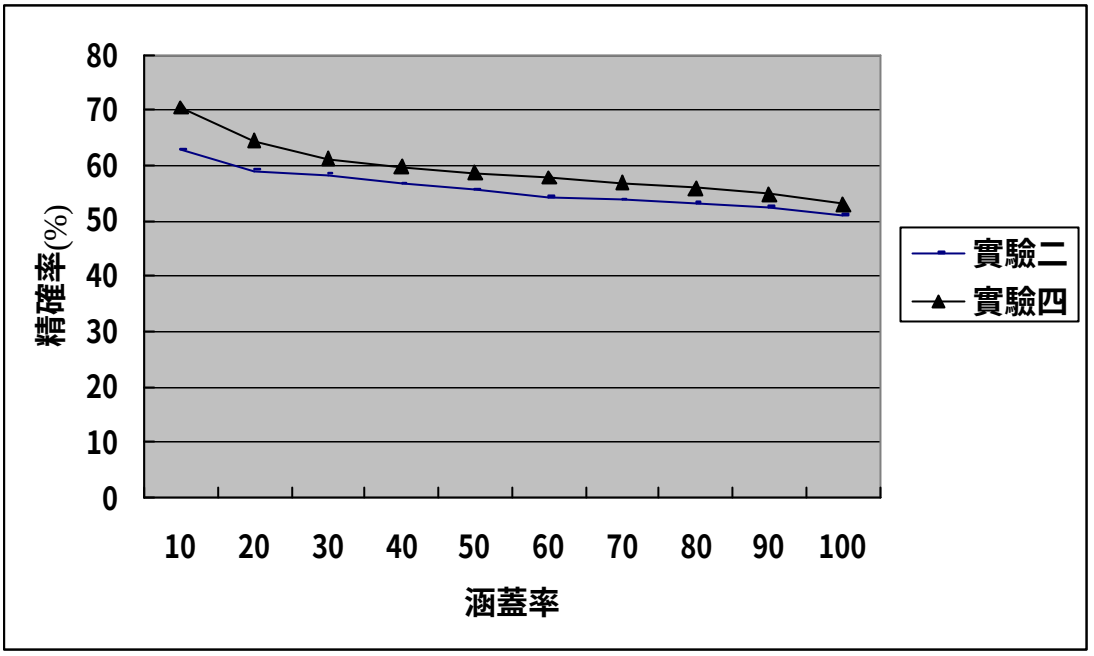

圖 5　　實驗二和實驗四的結果數據

由圖 6 可知實驗三的精確率明顯地比實驗二高，亦即使用由雙連字串組成的共現語詞當特徵的方法，比起只單純使用雙連字串當特徵的方法有效。這證明了本文提出的共現語詞當特徵的想法的確有助益於文件自動分類的工作。

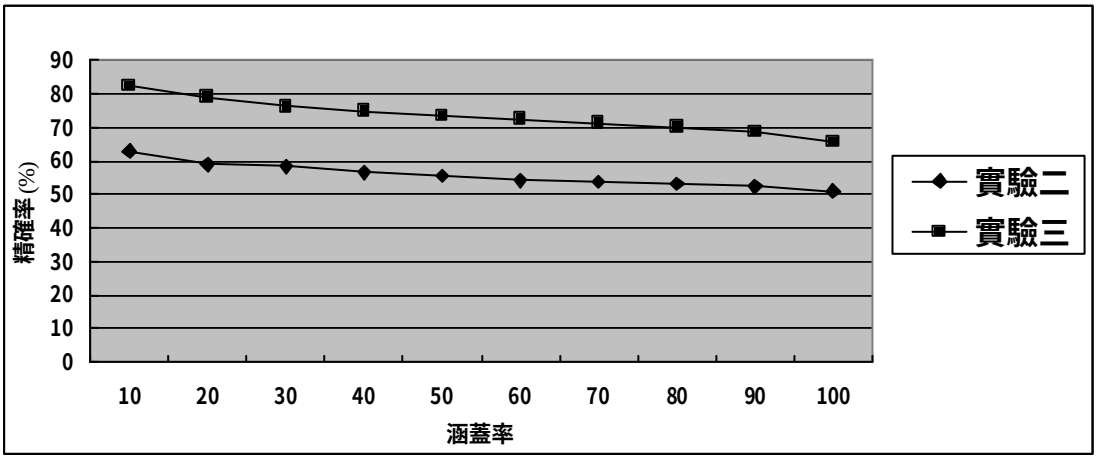

圖 6　　實驗二和實驗三的結果數據

至於為何會導致共現語詞的效能比起雙連字串好的原因，我們可以從測試文件「隨著財團銀行陸續登陸,3 邊金融交流將更熱絡」為例來看看。當我們的特徵語詞使用雙連字串時（表 4），本篇文件會因為雙連字串「銀行」在「銀行」類別擁有極高的權重值，導致系統傾向於將本文件歸屬到「銀行」類別，與其它類別相關的雙連字串，權重值都比較小，加總起來都敵不過擁有最高權重值得「銀行」類別。反觀共現語詞的情形（表 5），雖然共現語詞「金融」與「銀行」在「銀行」類別也是擁有最高的權重值，但是後面的「金融」與「銀行」、「金融」與「登陸」、「登陸」與「銀行」、「交流」與「金融」…等共現語詞皆在「經濟」類別有滿高的權重值，所以整體對各類別作加總之後，待測文件就會歸屬於正確的「經濟」類別。也就是說，共現語詞可以解決權重值過於極端的現象，可以平衡權重值的表現，使自動分類任務可以正確進行。另外，本文實驗的共現語詞與類別關係見附錄二，因為總共有 38 個類別，數目龐大，所以我們只取「經濟」類別以及「醫療衛生」類別當代表，。

表 4　測試文件「隨著財團銀行陸續登陸,3 邊金融交流將更熱絡」的雙連字串與類別的關聯度權重值

雙連字串	類別	權重	雙連字串	類別	權重
銀行	銀行	1219.5	金融	銀行	216.8
交流	經濟	515.3	金融	國際政經	187.1
金融	金融	485.3	登陸	經濟	139.6
銀行	金融	394.5	交流	政府、政治	125.6
金融	經濟	286.5	登陸	各項產業	85.6
銀行	國際政經	243.4	財團	政府、政治	84.0
銀行	經濟	225.4	銀行	股票	56.4

表 5　測試文件「隨著財團銀行陸續登陸,3 邊金融交流將更熱絡」的共現語詞與類別的關聯度權重值

共現語詞		類別	權重	共現語詞		類別	權重
語詞一	語詞二			語詞一	語詞二		
金融	銀行	銀行	147.9	登陸	銀行	銀行	21.5
金融	銀行	金融	122.1	金融	融交	金融	16.5
金融	銀行	經濟	56.4	交流	金融	經濟	15.1
金融	登陸	經濟	36.6	金融	融交	經濟	14.1
金融	銀行	國際政經	32.9	交流	融交	經濟	13.1
登陸	銀行	經濟	28.0	交流	登陸	經濟	12.9
陸續	銀行	銀行	23.7	交流	流將	經濟	11.8

此外，由圖 7 的實驗數據也可以清楚的看出，實驗六比起實驗五有較高的精確率，亦即，使用斷詞組成的共現語詞當特徵比起單純使用斷詞當特徵精確率高。這也證明了本文所提出的共現語詞在新聞文件自動分類上面的有效性。

再來我們看圖 8 的實驗三和實驗六的結果比較。實驗三是選取雙連字串組成的共現語詞當作特徵，實驗六是選取斷詞組成的共現語詞當作特徵，由圖 8 可以看出，由雙連字串所組成的共現語詞當特徵的效果比起斷詞組成的共現語詞當特徵的效果來的好，這樣的結果跟單純使用雙連字串以及單純使用斷詞當特徵的實驗結果是一樣的，也就是，對新聞標題文件的自動分類而言，雙連字串當特徵的效果比斷詞當特徵的效果好。

最後，我們從圖 9 探討所有實驗的結果，使用共現語詞當特徵的實驗三和實驗六擁有較高的精確率，其中，使用雙連字串合併而成的共現語詞效能比使用斷詞合併而成的共現語詞效能來得好。再來是考慮位置因素的實驗四，然後是單純使用雙連字串當特徵的實驗二以及單純使用斷詞當特徵的實驗五，最後，效能最低的是使用單字詞當特徵的實驗一。整體而言，精確率高低的比較為：共現語詞_雙連字串>共現語詞_斷詞>雙連字串加上位置考量>雙連字串>斷詞>單字詞。

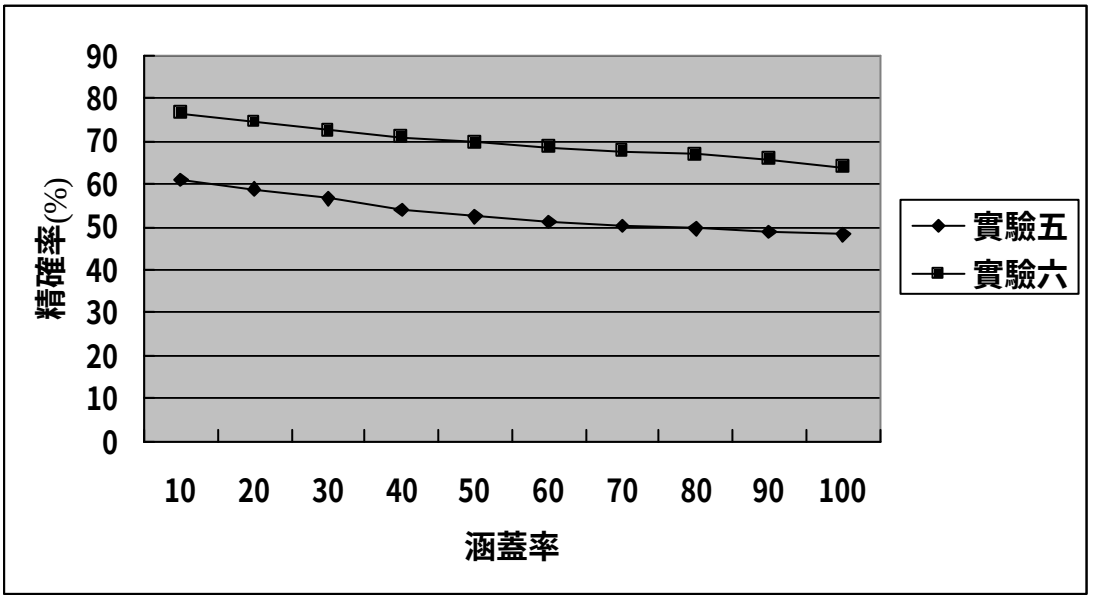

圖 7　　實驗五和實驗六的結果數據

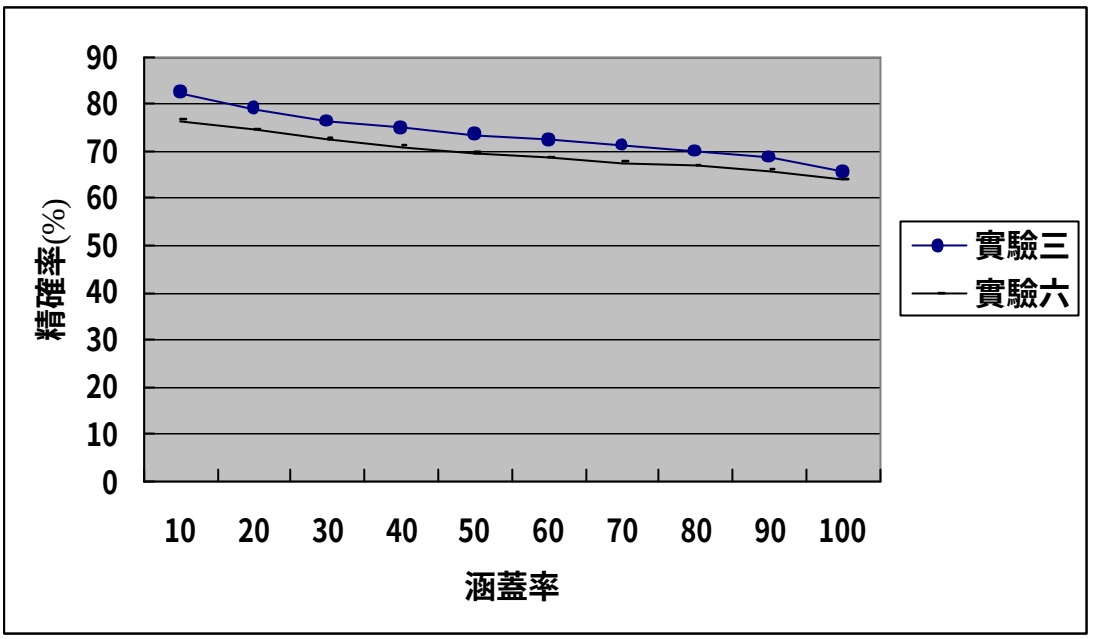

圖 8　　實驗三和實驗六的結果數據

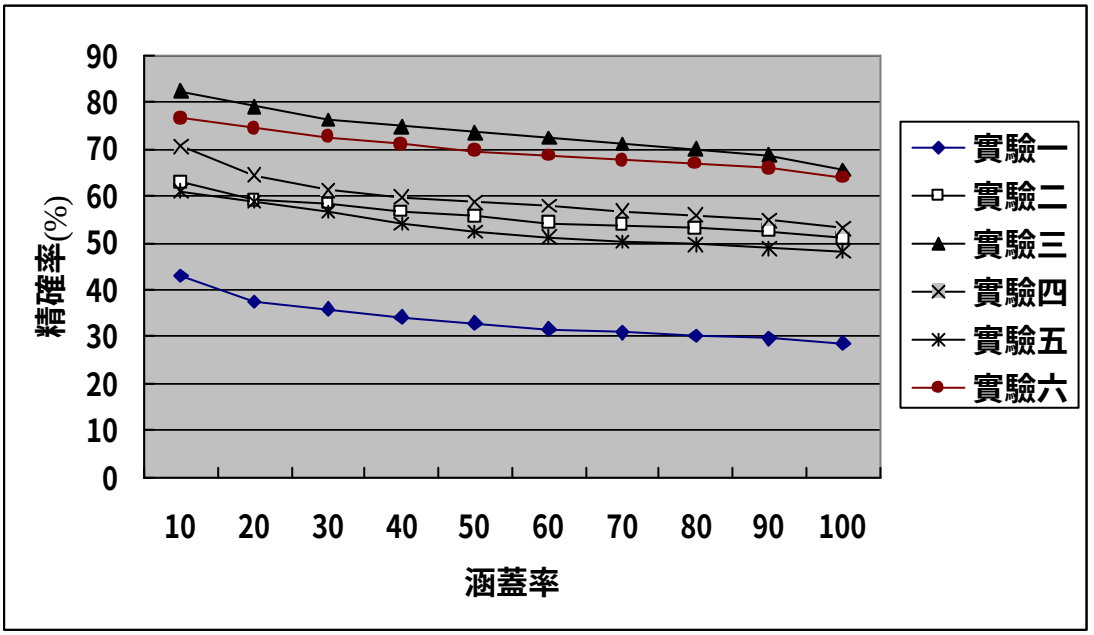

圖 9　　所有實驗的結果數據

5.5 錯誤分析

雖然本文使用共現語詞當特徵在分類任務上的效能已經提昇不少，但是我們依舊想知道在這樣的方法之下，有哪些情況是本系統無法正確分類的原因。因此，我們從實驗三中隨機取出 100 筆系統錯分類別的文件紀錄出來做錯誤分析。

實驗的錯誤分析如表 6 所示。系統錯分的部分，我們發現這些文件多含有其他類別的關鍵詞，導致系統的錯分。如例 18，應屬於「政府, 政治」類別，但是因為「中共」與「兩岸」這組共現語詞與「經濟」類別有相當高的關聯度，所以便被系統錯分至「經濟」類別；例 19 應屬於「國際政經」類別，卻因為「經濟」與「投資」以及「市場」與「經濟」這兩組共現語詞在「經濟」類別有很高的關聯度，所以就被系統錯分到「經濟」類別。

另外，我們發現有些部分的文件並不限制於某種單一類別，也就是說，雖然系統作的結果和原始分類不同類別，但是經由人工專家的判斷後，認為系統的歸屬類別也可以算是正確，符合文件的分類。根據我們的觀察，發現原始資料的分類與我們系統實驗的分類結果皆符合文件的分類，這樣的文件佔有 11%，如例 20，原始分類為「金融」類別，系統的分類為「經濟」類別，我們認為這兩個類別皆可以當作此文件的所屬類別。例 21 也是相同的情形。

表 6　實驗三的錯誤分析

錯誤原因		比例（%）
測試文件中含有其他類別的關鍵詞		54
原始資料錯誤分類	系統正確分類	19
	系統錯誤分類	7
由於同一篇文件可能屬於兩個以上的類別，原始資料與系統的分類結果皆歸屬到正確類別，但系統的歸屬類別與原始資料類別不同，因此被指為錯誤分類		11
文件內容提供的語詞資訊不足或其他因素		8
共現語詞特徵未出現在訓練語料中		1

例 18「政府, 政治」類：中共擴大漁區與兩岸關係（系統分類類別：「經濟」類）

例 19「國際政經」類：越南走向市場經濟,外人投資興趣大增（系統分類類別：「經濟」類）

例 20「金融」類：如何抑制通膨才是最大難題（系統分類類別：「經濟」類）

例 21「銀行」類：花旗決在台北設亞洲地區信用卡處理中心, 外國來台設立的第 1 個區域性金融業務中心, 財政部表歡迎（系統分類類別：「金融」類）

還有，我們可以發現，在實驗系統錯分的文件中，原始資料有 26%的錯誤率，其中有 19%是我們的系統能正確分類，如下面的例 22 至例 25 所示，例 22 應屬於「政府, 政治」類別，原始資料卻將之分類到「經濟」類別；例 23 應屬於

「稅賦」類別，原始資料將之分至「企業管理」類別；例 24 應屬「銀行」類別，原始資料卻錯分至「股票」類別；同樣地，例 25 應屬「股票」類別，原始資料將之錯分至「國際政經」類別。

例 22「經濟」類：結束"對峙"的年代(兩岸關係) （系統分類類別：「政府,政治」類）

例 23「企業管理」類：夫妻如何報稅才是最省,4 種判斷原則提供參考(附選擇夫妻所得申報方式判別表) （系統分類類別：「稅賦」類）

例 24「股票」類：北市銀更名為台北銀行（系統分類類別：「銀行」類）

例 25「國際政經」類：海外基金 91 年績效評估專題（系統分類類別：「股票」類）

而原始資料錯誤比率中的 7%文件，本系統也分類錯誤，如例 26 與例 27，例 26 的原始分類為「各項產業」類別，系統分類為「稅賦」類別，很明顯地，我們可以看出此文件應屬於「租賃」類別；例 27 的原始分類為「農業」類別，系統分類為「各項產業」類別，其實此文件應該屬於「林, 牧, 漁, 礦」類別。另外還有一種情況使得系統錯分，如例 28 和例 29 就是因為新聞文件本身所提供的語詞資訊不足，導致我們無法從標題本身得知其應屬類別。

由表 18 可以看出，這一百篇錯誤分類的文件中，有 30%的文件其實我們的系統是正確地將之分至應屬類別，也就是說，假若原始文件的歸屬類別都是正確無誤的，那麼我們應該會得到比實驗數據更高的精確度。

除此之外，我們在觀察類別時發現，這些類別的分類有重疊的現象，以「經濟」類別和「金融」類別為例，「金融」類別的文件常常可以歸屬在「經濟」類

別之中，造成我們在區分類別上的困難。因此，或許使用階層式的類別架構可以改善這個問題。

例 26「各項產業」類：房地租金標準北高 2 市看漲,售屋 3 戶以上多巢氏財交所得標準則略降(附 79 年及 80 年房屋及土地當地一般租金標準比較表及 79 年及 80 年財產交易所得標準比較表) （系統分類類別：「稅賦」類）。

例 27「農業」類：減少魚塭損失,連戰宣佈補助措施(1)漁會加價收購成魚(2)以每公斤 12 元收購凍死魚苗(3)魚價跌至成本價業者可憑交易單申請補貼（系統分類類別：「各項產業」類）。

例 28「各項產業」類：締造 PC 王國,光男豪氣淩雲（系統分類類別：「人物檔案」類）。

例 29「企業管理」類：現代廣告人應具備的特質（系統分類類別：「服務業(商業)」類）。

6. 結論與未來展望

本研究提出使用共現語詞當作選取的特徵及考慮語詞出現在文件中位置的方法。文件自動分類一般的做法都是將文件以多個單一語詞來表示，再從這些語詞中選取有用的特徵。這樣的做法並沒有考慮到同一篇文件中語詞間彼此的關係，或許某些語詞的配對是經常地被使用在某些類別。因此，我們希望藉由考慮共現語詞的出現情形，來幫助增進文件分類的效能。由實驗的結果證明共現語詞確實可協助中文新聞文件的自動分類。另外，考慮語詞在標題中出現的位置對自動分類也有幫助。

由實驗的結果可以知道，使用斷詞當特徵的實驗效能比起使用雙連字串當特徵的效能差，這是因為新聞文件包含著許多的專有名詞，例如人名、地名或組織名等，而斷詞系統不見得能即時包含這些專有名詞，偏偏這些專有名詞對於新聞標題文件來說，大多是重要的關鍵詞。因此，在未來的研究方向，我們將加入專有名詞的考量，再配合雙連字串組成的共現語詞特徵來作研究，探討專有名詞對新聞文件自動分類的效能影響。另外，我們知道，在考慮單一語詞當特徵時，通常名詞對文件分類的影響最大，動詞與形容詞次之，其他詞性幾乎沒有貢獻。但是，當我們將這些語詞配對後，或許可以發現某些一起出現的詞性在某特定類別中被廣泛使用，我們可以經此觀察得到分類的規則，以期能求得更高的精確率。

另外，本文的分類研究，主要是以新聞標題為研究對象，進行文件的分類處理。然而，我們知道新聞的本文內容比簡短的標題含有較多的資訊可以供擷取，因此，或許我們可以試著將本系統應用於中文文本的自動分類。只是，由於文本的長度較長，將產生大量的共現語詞特徵數，所以我們必須設定適當的特徵數門檻值，以掌握特徵數目。

最後，在未來的研究中，為了測試本系統的強健性，我們擬將本文提出的方法應用到英文新聞文件上面，例如路透社的新聞語料。也就是將英文的關鍵詞也合併成共現語詞，再利用這些共現語詞當特徵，以進行文件的自動分類任務。

參考文獻

1. 王稔志和張俊盛，「適應性文件分類系統」，第十四屆計算語言學研討會論文集，pp. 99-121，2001。
2. 古倫維，「中英文新聞文件主題偵測方法之研究」，國立臺灣大學，碩士論文，2000。
3. 杜海倫，「以標題進行新聞自動分類」，國立清華大學，碩士論文，1999。
4. 林頌堅，「自動化文件分類在資訊服務上的應用」，資訊傳播與圖書館學季刊，五卷二期，pp. 87-102，1998。
5. 柯淑津和陳振南，「階層式文件自動分類之特徵選取研究」，第十二屆計算語言學研討會論文集，pp. 137-149，1999。
6. 財經紀事語料，卓越出版社，1992。
7. 梅家駒、竺一鳴、高蘊琦和殷鴻翔，同義詞詞林，台北：東華書局，1993。
8. 陳彥呈和蔣榮先，「基於階層式類神經網路之自動新聞文件分類方法」，第十四屆計算語言學研討會論文集，pp. 89-97，2001。
9. 陳俊凱，「利用類神經網路作文件自動分類之研究」，私立淡江大學，碩士論文，1995。
10. 陳淑美，「財經新聞自動分類之研究」，國立台灣大學，碩士論文，1992。
11. 陳智偉，「文件分類的方法與分析」，國立清華大學，碩士論文，1998。
12. 曾祥泰，「以類神經網路為基礎的中文文件分類研究」，國立交通大學，碩士論文，1998。
13. 黃仲瑋，「醫藥新聞的自動分類」，國立交通大學，碩士論文，1999。
14. 黃政偉，「具語句特徵選取能力的類神經網路文件分類器」，國立台灣科技大學，碩士論文，1998。
15. 楊雪花，「模糊理論結合遺傳演算法應用於中文自動化分類之研究」，國立中央大學，碩士論文，1997。

16. **蔣俊霞，「中文文件自動分類之探討」，私立淡江大學，碩士論文，**1994。
17. **蔡憲文，「利用基因演算法來做文件自動分類之研究」，私立淡江大學，碩士論文，**1998。
18. Alessio, S. D. , K. Murray, R. Schiaffino & A. Kershenbaum, "The Effect of Topological Structure on Hierarchical Text Categorization," In Proceedings of the Sixth Workshop on Very Large Corpora, pp. 66-75, 1998.
19. Apte', C., F. Damerau & S. M. Weiss, "Automatic Learning of Decision Rules for Text Categorization," Journal of ACM Translation Information System, 12(3), pp. 233-251, July 1994.
20. Blosseville, M., G. Hebrail, M. Monteil & N. Penot, "Automatic Document Classification: Nature Language Processing, Statistical Analysis, and Expert System Techniques Used Together," In Proceedings of SIGIR'92, pp. 51-58, 1992.
21. Borko, H. & M. Bernick, "Automatic Document Classification, "Journal of the ACM, Vol. 10, No. 1, pp. 151-162, 1963.
22. Burstein, J., Marcu, D., Andreyev, S. & Chodorow M., "Towards Automatic Classification of Discourse Elements in Essays," In Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics, pp. 1-8, 2001.
23. Chakrabarti, S., B. Dom, R. Agarawal & P. Raghavan, "Using Taxonomy, Discriminate and Signatures for Navigating in Text Databases," In Proceedings of the 23rd VLDB Conference, Athens, Greece, pp. 446-455, 1997.
24. Chen, A., J. He, L. Xu, F. C. Gey & J. Meggs, "Chinese Text Retrieval Without Using a Dictionary," In Proceedings of ACM International Conference on Research and Development in Information Retrieval, Philadelphia PA, USA, pp. 42-49, 1997.
25. D'Alessio, S., K. Murray, R. Schiaffino & A. Kershenbaum, "The Effect of Topological Structure on Hierarchical Text Categorization," In Proceedings of the

Sixth Workshop on Very Large Corpora, pp. 66-75, 1998.

26. Dasigi, V. & R. C. Mann, "Nerual Net Learning Issues in Classification of Free Text Documents," In Proceedings of AAAI 1996 Sprint Symposium on Machine Learning in Information Access, pp. 101-103, March, 1996.
27. Frakes, W. B. & R. Baezay, Information Retrieval: Data Structures and Algorithms, Prentice-Hall, 1992.
28. Hamill, K. A. & A. Zamora, "The Use of Titles for Automatic Document Classification," Journal of American Soc. for Information Science, Vol. 31, No. 6, pp. 396-402, Nov., 1980.
29. Hayes P. J. and Weinstein S. P, "Construe/Tis: A System for Content-based Indexing of A Database of New Stories," In Proceedings of Second Annual Conference on Innovative Applications of Artificial Intelligence, pp. 49-64, 1990.
30. Hull, D., "Improving Text Retrieval for the Routing Problem using Latent Semantic Indexing," In Proceedings of SIGIR'94, pp. 282-291, 1994.
31. Joachims T., "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," In Proceedings of 10th European Conference on Machine Learning (ECML), pp. 137-142, 1998.
32. Joachims T., "A Statistical Learning Model of Text Classification with Support Vector Machines," In Proceedings of the Conference on Research and Development in Information Retrieval (SIGIR), pp. 128-136, 2001.
33. Joachims T. & F. Sebastiani, "Guest Editors' Introduction to the Special Issue on Automated Text Categorization," Journal of Intelligent Infromaiton Systems, Vol. 18, pp. 1-3, 2002.
34. Kar, G. & L. J. White, "A Distance Measure for Automatic Document Classification by Sequential Analysis," Journal of Information Processing and Management, Vol. 14, pp. 57-69, 1978.

35. Koller, D. & M. Sahami, "Towards Optimal Feature Selection," In Proceedings of International Conference on Machine Learning, Morgan-Kauffman, Vol. 13, pp. 1-14, 1997.

36. Kwok, K. L., "The Use of Title and Cited Titles as Document Representation for Automatic Classification," Journal of Information and Management, Vol. 11, pp. 201-206, 1975.

37. Lewis, D. D., "Feature Selection and Feature Extraction for Text Categorization," In Proceedings of Speech and Natural Language, pp. 212-217, 1992.

38. Liddy, E. D., W. Paik & E. S. Yu, "Text Categorization for Multiple Users Based on Semantic Features from a Machine-Readable Dictionary," ACM Transactions on Information Systems, Vol. 12, No. 3, pp. 278-295, 1994.

39. Lodhi, H., C. Saunders, J. Shawe-Taylor, N. Cristianini & C. Watkins, "Text Classification using String Kernels," Journal of Machine Learning Research, Vol. 2, pp. 419-444, 2002.

40. Maron, M. E., "Automatic Indexing: An Experimental Inquiry, " Journal of the ACM, Vol. 8, No. 3, pp. 404-417, July 1961.

41. Ng, H. T., W. B. Goh & K. L. Low, "Feature Selection, Perceptron Learning and a Usability Case Study for Text Categorization," In Proceedings of 20th ACM International Conference on Research and Development in Information Retrieval, Philadelphia PA, USA, pp. 67-73, 1997.

42. Salton, G. & C. Buckley, "Term Weighting Approaches in Automatic Information Retrieval," Journal of Information Proceeding and Management, Vol.24:3, pp. 513-524, 1988.

43. Schutze, M., M. Hearst & E. Saund, "Applying the Multiple Cause Mixture Model to Text Categorization," In Proceedings of the 13th International Conference in Machine Learning, pp. 435-443, 1996.

44. Sebastiani, F., "Machine Learning in Automated Text Categorization," ACM Computing Sruveys, Vol. 34, No. 1, pp. 1-47, March 2002.
45. Weiss, S. A., S. Kasif & E. Brill, "Text Classification in USENET Newsgroups: A Progress Report," In Proceedings of AAAI 1996 Sprint Symposium on Machine Learning in Information Access, pp. 125-127, 1996.
46. Yang, Y. & C. G. Chute, "An Example-Based Mapping Method for Text Categorization and Retrieval," ACM Transactions on Information Systems, Vol. 12, No. 3, pp. 252-277, July 1994.
47. Yang, Y., "An Evaluation of Statistical Approaches to MEDLINE Indexing," In Proceedings of the AMIA, pp. 358-362, 1996.
48. Yang, Y., "An Evaluation of Statistical Approaches to Text Categorization," Technical Report CMU-CS-97-127, Computer Science Department of Carnegie Mellon University, pp. 1-28, 1998.
49. Yang, Y., "An Evaluation of Statistical Approaches to Text Categorization," Information Retrieval, Vol. 1, pp. 69-90, 1999.

附錄一　財經紀事新聞介紹

『財經紀事』新聞由卓越出版社於西元 1992 **年提供。其中所含的新聞標題，內容取自民國八十一年間中國時報、工商日報、聯合報、民生報等各報社之新聞標題，共含有** 131,136 **則新聞標題，原始文字檔如附表一所示。**

附表一　原始文字檔範例

[1]J.01;810101;**中晚**;03(M);1;

[2]**社評**

[3]**憲改與議會民主的迷思**:**從煽動者到政治的新現實**

[1]HB.01;810101;**中晚**;04(M);1;

[2]**秦德山**

[3]APEC **首度擺脫亞銀情結**,**採「中華台北」名稱模式帶來外交大豐收**

[1]HDn.1501;810101;**華**;03(S);1;

[2]**編輯部**

[3]**國內核電廠運轉達世界水準**,**外國專家評估證明** 3 **座核電營不差**,**台電將針對** 95 **項建議在** 1 **月底前提出改善方案**

[1]J.01;810101;**華**;02(M);1;

[2]**社論**

[3]**再創政治奇蹟**,**建立尊嚴繁榮**:**中華民國** 81 **年元旦獻詞**

這些標題採三層式分類：大分類、中分類以及小分類。其中，大分類共分為「公營事業篇」、「服務業篇」、「金融業篇」…**等九大類別；大類別下細分為** 38 **個中類別，例如「金融業篇」類別之下含有「金融」、「銀行」、「外匯」、「股票」以及「租賃」等中類別；中類別下又細分成小類別。以附表一中的** J.01 **為例，** 01 **便是一個小類別，**J **為其中類別。另外，附表二為九大類別及其下的中類別以及它們的類別代碼，附表三為測試資料與訓練資料在各類別的文件數。**

附表二　類別代碼對照表

大類別	中類別	類別代碼
公營事業篇	公營事業	HDn
服務業篇	交通運輸業	HE
	觀光旅遊	HEt
	服務業（商業）	HF
金融篇	金融	HG
	銀行	HGb
	外匯	HGe
	股票	HGs
	租賃	HGt
國際篇	國際政經	HX
產業篇	各項產業	HDo
貿易篇	貿易	HFt
農業篇	農業	S
	林，牧，漁，礦	SD
總體篇	經濟	HB
	消費	HBc

附表二　類別代碼對照表（續）

大類別	中類別	類別代碼
總體篇	土地	HD
	工業	HDi
	勞工	HDl
	商標，智慧財產	HFc
	稅賦	HJd
	關稅	HJt
其他	人口	HBp
	公共建設	HDm
	大眾傳播	HEm
	郵政，電信	HEs
	企業管理	HFm
	財政	HJ
	社會	HN
	人事動態	HPt
	人物檔案	Hp
	政府，政治	J
	教育	L
	醫療衛生	RA
	科技	T
	環境	TD
	公司檔案	W
	集團企業	WG

附表三　測試資料與訓練資料在各類別的文件數

類別代碼	測試資料文件則數	訓練資料文件則數
HDn	193	1,703
HE	253	2,425
Het	167	1,521
HF	774	6,714
HG	316	2,686
HGb	256	2,429
HGe	58	455
HGs	486	3,921
HGt	88	886
HX	1,495	13,060
HDo	1,532	14,404
HFt	328	2,804
S	116	1,109
SD	54	460
HB	1,571	14,811
HBc	118	1,138
HD	150	1,346
HDi	168	1,635
HDl	257	2,452
HFc	87	799
HJd	263	2,505
HJt	38	315
HBp	18	194

附表三　測試資料與訓練資料在各類別的文件數（續）

類別代碼	測試資料文件則數	訓練資料文件則數
HDm	308	2,776
Hem	5	62
HEs	44	394
HFm	299	2,635
HJ	95	789
HN	205	1,982
HPt	81	630
Hp	243	2,075
J	1,465	12,935
L	224	2,085
RA	269	2,076
T	71	627
TD	287	2,564
W	89	839
WG	23	208

附錄二　共現語詞與各類別關聯度的權重值

以下的附表四與附表五是各類別中的共現語詞（由雙連字串組成）與類別的關係，權重值由大到小排列，取得前五十名排行的共現語詞。

附表四　「經濟」類別的前五十名排行的共現語詞

類別	語詞一	語詞二	權重值	類別	語詞一	語詞二	權重值
HB	大陸	投資	831.54	HB	中共	企業	320.78
HB	台商	投資	605.35	HB	陸經	經濟	320.05
HB	中共	經濟	576.83	HB	改革	革開	314.16
HB	大陸	台商	576.74	HB	中共	貿易	312.81
HB	小平	鄧小	550.85	HB	革開	開放	312.01
HB	大陸	經濟	521.71	HB	中共	市場	311.36
HB	交流	兩岸	482.46	HB	兩岸	直航	310.59
HB	中共	改革	460.25	HB	中共	政治	305.16
HB	大陸	兩岸	451.27	HB	投資	經濟	285.94
HB	大陸	市場	444.99	HB	改革	經濟	282.83
HB	大陸	中共	428.7	HB	十四	四大	282.42
HB	投資	陸投	428.19	HB	大陸	赴大	280.04
HB	大陸	陸投	400.51	HB	中共	北京	275.11
HB	兩岸	岸經	393.66	HB	企業	投資	273.58
HB	兩岸	經貿	385.6	HB	場經	經濟	267.95
HB	大陸	企業	366.21	HB	成長	經濟	266.69
HB	大陸	台灣	350.91	HB	中共	台商	265.44
HB	岸經	經貿	349.92	HB	大陸	中國	264.05
HB	中共	開放	345.46	HB	中共	共決	263.78
HB	大陸	開放	340.72	HB	兩岸	經濟	261.96
HB	大陸	改革	334.73	HB	上海	股市	260.11
HB	中共	兩岸	334.61	HB	兩岸	關係	257.13
HB	大陸	陸經	331.98	HB	中共	發展	254.92
HB	改革	開放	323.6	HB	發展	經濟	252.84
HB	投資	商投	322.33	HB	市場	經濟	251.26

附表五 「醫療衛生」類別的前五十名排行的共現語詞

類別	語詞一	語詞二	權重值	類別	語詞一	語詞二	權重值
RA	日咳	百日	208.82	RA	尿病	糖尿	61.14
RA	立醫	醫院	117.2	RA	民健	全民	60.59
RA	醫院	醫療	111.65	RA	衛生	醫院	60.25
RA	病患	醫院	101.98	RA	療網	醫療	59.91
RA	生署	衛生	98.6	RA	正確	看病	59.91
RA	市立	醫院	96.83	RA	正確	族如	59.91
RA	滋病	愛滋	96.83	RA	正確	確看	59.91
RA	市立	立醫	95.64	RA	如何	確看	59.91
RA	台大	醫院	87.67	RA	何正	看病	59.91
RA	日咳	病例	84.26	RA	何正	族如	59.91
RA	疫苗	接種	84.26	RA	何正	確看	59.91
RA	百日	病例	84.26	RA	看病	族如	59.91
RA	肝炎	型肝	83.68	RA	看病	確看	59.91
RA	醫師	醫院	81.06	RA	中心	醫學	59.91
RA	百日	疫情	76.93	RA	族如	確看	59.91
RA	病床	醫院	76.93	RA	日咳	疑似	58.62
RA	日咳	疫情	76.93	RA	百日	疑似	58.62
RA	大醫	醫院	76.03	RA	民健	健保	57.74
RA	大醫	台大	75.22	RA	全民	健保	57.74
RA	接種	預防	73.27	RA	正確	班族	56.92
RA	防接	接種	73.27	RA	班族	確看	56.92
RA	防接	預防	73.27	RA	病患	醫師	56.92
RA	疫苗	衛署	65.94	RA	病患	醫療	56.92
RA	日咳	流行	65.94	RA	上班	正確	56.92
RA	百日	流行	65.94	RA	上班	何正	56.92

附錄三　100 則測試文件

以下是我們擷取出 100 則測試文件，並且標註它們在語料中的原始歸屬類別以及本研究的自動分類系統在實驗三中指派的類別。附表六顯示出有哪些文件是我們的分類系統可以正確地給予指派類別，有哪些文件是我們的系統分類錯誤。

附表六 100 則測試文件的原始類別以及系統指派的類別

編號	新聞標題文件	原始的類別	系統指派的類別
1	修正海關進口稅則公告取消退稅並停止按內銷比率課稅貨品項目清表	HJt	HJt
2	廠商如何因應公平交易法系列之(2):外商公司、同業公會均適用	HF	HF
3	廠商如何因應公平交易法系列之(7):百貨業聯合折扣型態有待釐清	HF	HF
4	上市公司剖析系列報導:獨霸商用車市場,中華連年豐收,80 年營收 155 億,估計每股稅後盈餘 3.5 元	HGs	HGs
5	上市公司剖析系列報導:瓦斯售價調高在望,2 辦公大樓完全入帳可獲利 2.5 億元,營運左右逢源,大台北漸入佳境	HGs	HGs
6	修正海關進口稅則公告取消退稅並停止按內銷比率課稅貨品項目清表	HJt	HJt
7	廠商如何因應公平交易法系列之(16):媒體與出版業均受規範	HF	HF
8	廠商如何因應公平交易法系列之(21):加盟店可調整為「結合」關係	HF	HF
9	80 年度營利事業所得稅結算申報及查審應注意事項法令輯要(2)	HJd	HJd
10	全國經濟會議工商業分組提出建議:全面翻修公司法,廢除不合宜限制,經部決朝允許公司持有一定庫存股、刪除公司擴充生產設備限制,及對不具違法行為改採行政罰等方向修正	HB	HB
11	「傑出操盤人致勝秘招大公開」系列報導(1)/林一銘:早人一步才有附加價值	Hp	Hp
12	上市公司剖析系列報導:外有新銀行競逐壓力,內有派系之爭尚未平息,高企業績亮麗表現,將受考驗	HGs	HGs

編號	新聞標題文件	原始的類別	系統指派的類別
13	1 月僑外來台投資台商赴外投資,均較 80 年同期減少,中外合資案則增加 26.6%受注目(附經濟部核准中外投資和技術合作案比較表)	HB	HB
14	「傑出操盤人致勝秘招大公開」系列報導(6)/吳永祥:加強風險管理,不留套牢股票	Hp	Hp
15	執政黨中常會通過政策會升級案及相關人事案,謝隆盛:國大黨政協調工作會主任兼國大黨團書記長,王金平:立院黨政協調工作會主任兼立院黨團書記長,洪玉欽:政黨關係工作會主任,監院黨政協調工作會主任尚未決定;另派饒穎奇升任中委會副秘書長	HPt	HPt
16	新銀行搶攤前進北縣新莊戰雲密佈,插旗搶地盤,台新、聯邦、萬通湧向中港社區,另一波高潮,萬泰、中興另一分行將在三重宣戰(附地方金融業及新銀行保管箱出租行一覽表)	HGb	HGb
17	小川忠夫:改善中日貿易逆差,台灣業者也應盡心力,強化先端科技接收力,致力培養商業習慣,並呼籲日本 10 大家電 9 大綜合商社代表加強對台採購、投資及技術移轉	HFt	HFt
18	政院 22 日公布官方委託學者調查二二八事件研究報告,首度檢討蔣公、陳儀、柯遠芬等關係人的責任歸屬,研究小組成員建議對受難者的賠償、平反採取善後對應措施	J	J
19	上市公司剖析系列報導:業外投資獲利估算逾 3 億,石膏板廠將量產,環球水泥紮底,跨足金融闢財源	HGs	HGs
20	上市公司剖析系列報導:新纖致力多樣化生產,路遙知馬力,計劃赴印尼設廠,將發行國內及海外可轉換公司債支應	HGs	HGs
21	「傑出操盤人致勝秘招大公開」系列報導(16)/王文慧:早人一步始能海闊天空	Hp	Hp
22	福壽公司 26 日董事會決提撥 1 億 2000 多萬元資本公積配股、盈餘轉增資 3600 多萬元,80 年每股股票股利 1 元	HGs	HGs
23	81 年政治熱戲專題(1):執政黨三中全會(附 81 年重要政治進程分析表)(2)國大臨時會(3)執政黨十四全(4)81 年底立委選舉	J	J
24	土地增值稅將改按交易價課徵,王建　宣示決採漸進方式並考慮調低稅率	HJd	HJd

編號	新聞標題文件	原始的類別	系統指派的類別
25	「傑出操盤人致勝秘招大公開」系列報導(3):亞東錢聲遠隨波逐「利」拿捏神準	Hp	Hp
26	上市公司剖析系列報導:歐洲線預估可成長 1 成以上,美洲線潛藏轉機,長榮鼓浪前行,81 年榮面居多	HGs	HGs
27	「傑出操盤人致勝秘招大公開」系列報導(5)/孫天山:掌握主升段,搶搭順風車	Hp	Hp
28	多層次傳銷事業有法可管了,公交會通過管理辦法草案,業者需在公布後 2 個月內完成報備	HF	HF
29	中共將制訂兩岸直航記者訪台法規,因應我方通過兩岸關係條例後形勢,希望以積極做法促進三通擴大交流	HB	HB
30	中共政改響前奏,縣級機構將「轉型」,為減輕財政負擔,縣級政府的管理職能將逐步變為服務職能;除精簡人事外,並將部分機關轉變為經濟實體獨立經營,既能服務也有收入	HB	HB
31	中共公開呼籲發展資本主義,人民日報認為大陸應學習西方經濟發展經驗,以持續開放政策	HB	HB
32	「傑出操盤人致勝秘招大公開」系列報導(13)/謝克興:掌握供需變化,擇優適時切入	Hp	Hp
33	中共計委會批准大陸 10 大經濟區計劃,東北、華北、南方沿海、長江黃河流域及新疆西藏分為 10 大區域,並依各區域特殊發展條件規劃建設重點	HB	HB
34	80 年度營利事業所得稅結算申報及查審應注意事項法令輯要(4)	HJd	HJd
35	新興航運 28 日董事會通過 80 年度配股配息方案,每股股利 2 元,每千股配發 100 股;並擬辦理現金增資 1 億 8300 萬元,用來購置新船	HGs	HGs
36	「傑出操盤人致勝秘招大公開」系列報導(26)/張慶隆:重視基本面,賺錢不投機	Hp	Hp
37	「上市公司剖析」系列報導:尚德調整產銷結構,將可揮別陰霾,以冰品業為經營重心加上業外收益,81 年應可扭轉頹勢	HGs	HGs
38	「傑出操盤人致勝秘招大公開」系列報導(32)/馮偉奇:賺大賠小,押寶憑真本事	Hp	Hp
39	執政黨台北市黨部與中國時報合辦「基層建設座談會」專題/松山區:期待重大工程完工,再創松山發展契機	HDm	HDm

編號	新聞標題文件	原始的類別	系統指派的類別
40	國泰人壽20日董事會中通過80年度盈餘分配,每股配發現金股利 1.5 元、股票股利 3 元	HGs	HGs
41	落實一定面積或一定金額以上土地買賣按實際交易價格課徵土地增值稅政策,內政部將研究建立土地正常買賣計價方式	HJd	HD
42	上市公司剖析系列報導:流通籌碼淹腳目,中國信託經營天蠶變,80 年度分派股利 2 元,增資後資本額將逾 115 億元,成為金融股大哥大	HGs	HGs
43	「上市公司剖析」系列報導 :積極跨足海外,台紙色彩絢爛,土地資產亦雄厚,尤其大肚廠開發計畫更受矚目	HGs	HGs
44	中共 85 計畫有利台灣產業發展,我應強化大陸台商協會功能,爭取更多投資保障	HB	HB
45	非自用住宅,一律按交易價課增值稅,自用住宅則可按公告現值課稅,財政部裁定原則,將與內政部協商	HJd	HJd
46	我將建立兩岸經貿全面預警指標,兩岸條例立法後,已赴大陸投資台商須補辦登記,逾期將從嚴處理	HB	HB
47	歐洲共同體和歐洲自由貿易協會簽署協定,全球最大單一市場歐洲經濟區誕生加速整合	HX	HX
48	中共擬進一步開放內銷市場,配合台商優惠貸款,可能掀起另 1 波大陸投資熱	HB	HB
49	保險代理人經紀人公證人資格測驗系列報導(6):電腦閱卷採選擇題	HGt	HGt
50	上市公司剖析系列報導:高雄廠採礦權可望延長,短線進出股市獲利豐,台泥 81 年業績,將比 80 年更輝煌	HGs	HGs
51	保險代理人經紀人公證人資格測驗系列報導(15):保險概論偏重學術性	HGt	HGt
52	保險代理人經紀人公證人資格測驗系列報導(7):兩種保險組織是重點	HGt	HGt
53	擴大吸引外資中共優惠開放掛帥,預計 6 月「全國外商投資工作會議」討論吸引外資辦法;包括內陸省市享有沿海開放區外商優惠政策,准許外商在沿海投資商業、交通、金融等試點,將可使大陸對外更開放	HB	HB
54	海峽兩岸經貿發展回顧及合作前景研討會論文系列之(6)/兩岸產業合作的可能方式:以機電工業為例	HB	HB

編號	新聞標題文件	原始的類別	系統指派的類別
55	「上市公司剖析」系列報導:致力品保制度,華夏為銷歐舖路	HGs	HGs
56	高縣高手雲集,有人忙佈樁已花千萬元,選情熱烈可期,執政黨:現任 3 立委,可望再獲提名;民進黨:傾向不辦初選直接提名,余政憲可望過關	J	J
57	政院研擬多項提振民間投資意願優惠措施,減輕土地承租負擔提供低利貸款等,期挽留企業界赴大陸投資腳步	HB	HB
58	金額不超過國內半數且在國內投資達一定規模,赴大陸間接投資經部擬從寬調整,對國內產業發展造成不利影響者,陸委會可能暫停開放相關投資項目	HB	J
59	赴大陸投資,水泥業上建言書,呼籲政府出面與大陸簽訂投資保障協議,爭取台商外銷比例在 50%以下,配合措施貨物從量課征以示公平加速開發和平專區並建港	HB	HB
60	執政黨通過極具彈性,參選立委提名辦法,①部份地區可視狀不辦初選,②特殊需要可辦徵召,③在不提名或不足額提名地區,經書面申請得報准參選,④任立委現優良者可優先考慮	J	J
61	中共擬議由海基會、海協會簽訂兩岸投資協定,北京以 5 號文件確定發展第三產業,列 4 個重點歡迎台商參與經貿合作	HB	HB
62	陸委會 22 日將通過 155 項服務業開放赴大陸投資,經部配合一措施將簡化報備和申請手續並考慮採負面表列方式	HB	HB
63	投資意願低落已成經濟發展隱憂,欲改善現況,有賴明確的兩岸經貿政策及各部會間的相互合作(存 HB.10)	HB	HB
64	立委參選大爆炸,國民黨內登記完畢,競爭態勢空前激烈,各方菁英齊出馬(附執政黨辦理 2 屆立委選舉黨內提名登記名單)	J	J
65	兩岸經貿交流,短期有利,長期潛伏危機,將造成過份依賴大陸市場,減少我談判籌碼,進而降低產業升級能力(附大陸間接投資對雙方長短期利弊分析摘要表)	HB	HB
66	以提名 60 人為最高限額:2 屆區域立委初選,民進黨 7 個選區 7 月 5 日舉行:不分區部分初步設定提名 9 人,7 月 6 日起受理參選登記	J	J

編號	新聞標題文件	原始的類別	系統指派的類別
67	大陸台商的生存與競爭,東南沿海專題報導系列之(1):地理位置最近,關稅簽證優待,平潭想做對台貿易重點市場	HB	HB
68	違反規定大陸投資,將處罰巨款,馬英九:兩岸關係條例規定課稅與罰款不會影響台商到大陸投資的意願	HB	HB
69	大陸投資案件,不准先斬後奏,投審會擬訂「對大陸地區投資或技術合作許可辦法」,嚴格規定,赴大陸投資者,須獲經濟部核准,不再允許「先投資後報備」(存 HB.1001)	HB	HB
70	「曾任監委」可能列入 2 屆監委提名資格,監院組織法修正草案初步完成,執政黨傾向「資格從寬,提名從嚴」	J	J
71	大陸民眾消費潛力,值得重視,中華經濟研究院指出:城鄉居民孝款累積高達 1 萬億人民幣,結餘購買力大幅提升,內銷市場規模不可小看(大陸取民及社會消費額成長率)(大陸 30 個地區消費水準分類)(大陸居民結餘購買力及商品零售額)	HB	HB
72	國民黨北市立委選舉黨內初選南區首場說明會,參選人暢所欲言,高潮迭起,公共政策成話題,立院亂象被抨擊	J	J
73	土地增值稅若改按實際交易價格課徵,銀行核貸成數勢將縮水	HJd	HGb
74	執政黨立委黨內初選揭曉,投票率 3 成 42 競爭激烈,提名名單預計 9 月 16 日報中常會通過	J	J
75	郝揆:加強兩岸文化交流應把人員交流擺在前面,指中共不可能接受兩岸報紙對等發行,但我仍歡迎更多的大陸記者來台灣看看(附海基會邀請來台訪問大陸記者個人資料)	HB	HB
76	對大陸經貿政策,央行內部口徑一致,在產業方面,主張政府「不能鼓勵」業者赴彼岸投資;在兩岸金融開放方面,主張「慢慢來」	HB	HB
77	據經濟部投資審議委員會統計:台商大陸投資近 10 億美元,附(主要台商進軍大陸市場一覽表)	HB	HB
78	民進黨不分區僑選立委 8 月 30 日初選,提名名單確定,不分區\葉菊蘭高居榜首,黃爾璇意外排第七,江鵬堅落在安全名單外,僑選\張旭成、陳唐山列為安全外單	J	J
79	匈牙利公車採購案驚傳弊端?約談官商 6 人依凱公司總經理田訓民被收押,檢調追查重點:①廢氣排放 2 期標準不符國內採購契約規定,是否涉嫌偽造	J	HE

編號	新聞標題文件	原始的類別	系統指派的類別
80	華隆企業股份有限公司公告(附本公司 81 年 8 月份董事、監察人、經理人及持股百分之 10 以上大股東股權變動情形)	HGs	HGs
81	大陸工作會議召開前夕,兩岸經貿問題是重頭戲,俞國華:我們應開放企業直接登陸投資	HB	HB
82	國民黨不分區立委提名業 11、12 展開,以專業性、功能性、均衡性為原則,現任立委部分著重國會資歷	J	J
83	台商投資熱「吃緊弄破碗」?;<大陸不動產投資指南系列報導之②>邱彰:賺錢的人守口如瓶,賠錢的人一出口不就更難「解套」?	HB	HDo
84	政院 22 日續審查土地增值稅課徵案;刪除營利事業納入按實價課稅對象,抑制土地炒作雷大雨小	HJd	HJd
85	採購匈牙利公車案情急轉往上升,急付公車尾款,證物不利處長,得標的台灣依凱和依凱原廠,聯合記者會保證競標文件真實(附台北市匈牙利公車採購案過程紀要)(存 HE.1007)	HE	J
86	匈車採購驗收過程諸多疑點仍待釐清①依凱是否以偽造資料投標②合約規定定過於寬鬆③初檢多項不合格仍堅持先上路④2 期排放標準問題⑤相關官員對不實資料事先是否知情(附台北市公車處 560 輛公車採購事件紀要)(存 HE.1007)	HE	J
87	企業人士:台商赴大陸投資應透視大陸「春風後母面」政策,規範廠商兩岸對等投資避免兄弟鬩牆	HB	HB
88	土地增值稅課微方式,內政部再提折衷方案,3 年內曾多移轉,變更用途後道次移轉,改按實際交易價格課稅	HJd	HJd
89	海外投資熱漸為大陸間接投資所取代,經部投審會議 5 日通過申請赴大陸投資案多達 19 件,同期對國外投資者僅有 4 件	HB	HB
90	龍邦建設股份有限公司公告:本公司 81 年 9 月份董事、監察人、經理人、及百分之 10 以上大股東股權變動情形	HGs	HGs
91	大陸外商投資 500 大製造業台商排名顯著,根據中共發布排行榜顯示,台商企業中,偶鞋業佔多數,其中海豐鞋業局第 60 位,協豐鞋業 88 位排名較高(附 1991 年大陸 20 家最大外資製造業排行變動表)	HB	HB

編號	新聞標題文件	原始的類別	系統指派的類別
92	158 項服務業可間接赴大陸投資,陸委會通過首批開放項目,最遲 12 月由 4 單位分別公布實施	HB	HB
93	經部闡述兩岸經貿「教戰手冊」,①加強整合輔導台商②規劃兩岸投資障制度③擴大開叉大陸制物品間接進口④爭取大陸內銷市場	HB	HB
94	民進黨公布「兩岸關係」及「中國政策」綱領,主張台灣主權獨立不屬於中華人民共和國,並強調以「一中一台」策略國際生存空間	J	J
95	立法院第 90 會期總質詢朝野議題焦點,,一中一台論調就是「台獨」主張,郝揆:堅持一個中國政策才能保障民眾福祉,18 標案及王建烜辭職案波濤洶湧,民進黨立院黨團表現了無新意	J	J
96	服務業赴大陸,兩岸暗中較勁,陸委會認藉推銷「台灣經驗」改變大陸人民價值觀,中共則將以「市場磁性」吸收台資	HB	HB
97	台商赴大陸投資技合辦法,限制閘門大開,抑制兩岸貿易過熱,辦法中賦予對違規廠商課處罰鍰權力	HB	HB
98	執政黨通過第 2 屆立委不分區及僑選提名名單,不分區應選 30 人提 27 名,僑選足額提 6 名	J	J
99	二屆立委選舉的觀察評估及未來衝擊座談會(3):選民投票取向,影響政黨政治發展,執政黨黨內有派,仍須拉大得票差距勝選,才能確保政權穩定	J	J
100	2000 萬人的抉擇,李登輝總統的政治改革軌跡	J	J