國立交通大學

資訊管理研究所

碩士論文

網際網路新聞文章心情偵測之研究

Research on Mood Detection of Internet News Articles

研究生：林揚書

指導教授：柯皓仁

林妙聰

中華民國 九十八 年 七 月

網際網路新聞文章心情偵測之研究

Research on Mood Detection of

Internet News Articles

研 究 生：林 揚 書　　Student: Yang-Shu Lin

指導教授：柯 皓 仁　　Advisor: Dr. Hao-Ren Ke

林 妙 聰　　Dr. Miao-Tsong Lin

國 立 交 通 大 學

資 訊 管 理 研 究 所

碩 士 論 文

A Thesis

Submitted to Institute of Information Management

College of Management

National Chiao Tung University

in Partial Fulfillment of the Requirements

for the Degree of

Master of Science

In

Information Management

July 2009

Hsinchu, Taiwan, the Republic of China

中華民國　九十八 年 七 月

網際網路新聞文章心情偵測之研究

指導教授：柯 皓 仁、林 妙 聰　　研究生：林 揚 書

國立交通大學資訊管理研究所

摘要

全世界每天有數以萬計的新聞被報導，在這些新聞裡，僅有少部份與自己有關的，大多數是毫不相關的。隨著網際網路使用者數量大幅度增加，網路已取代傳統媒體成為最受矚目的大眾媒體，如何從網路上眾多的新聞之中，以最短時間去篩選出自己需要的、喜愛的及完全不需閱讀的新聞乃是一個值得關注的議題。

本研究會以預測讀者閱讀新聞後的心情為目標，使用 Yahoo!奇摩新聞的心情投票資料，透過 CKIP 的斷詞切字處理，計算出每個詞彙的 Log Likelihood Ratio 值，與其心情比例分數結合之後排序篩選，找出優秀的特徵值作為分類依據，最後再放入 LibSVM 分類建構出模型，預測讀者閱讀新聞後可能呈現的心情狀況，並進一步設計出關鍵詞彙挑選系統，供讀者在選擇閱讀新聞時參考。

關鍵詞：文章心情分類、文件分類、支援向量機、特徵挑選、資訊檢索

Research on Mood Detection of Internet News Articles

Advisor: Dr. Hao-Ren Ke, Dr. Miao-Tsong Lin Student: Yang-Shu Lin

Institute of Information Management

National Chiao Tung University

Abstract

There's millions and thousands news coming out everyday. Only limited number of these news are relevant to a particular person. In the digital era, Internet has surpassed traditional media and become one of the most attractive media. How do we effective and efficiently filter through the huge amount of information on the Internet for finding those pieces of information which we need, like or we don't need to read?

The objective of this study is to predict the feelings of readers after reading the news from Yahoo! Kimo news. By using the CKIP system's tokenization and part-of-speech tagging processes, candidate phrases are indentified. Then the Log Likelihood Ratio values of candidate phrases are combined with the mood tendency to discover significant phrases for each mood. The discovered significant phrases are used as the input of a Support Vector Machine (SVM) software, LibSVM, to conduct the classification. Furthermore, a keyword screening system is designed to provide readers information for choosing what to read and predicting the readers' possible mood after reading a particular the news.

Keyword：Mood detection, Text categorization, Support Vector Machine, Feature Selection, Information Retrieval

誌　　謝

隨著畢業口試的結束，兩年的研究所生涯也到了尾聲，回顧兩年來的經歷，內心充滿了幸福與感謝。

首先要感謝指導教授柯皓仁老師與林妙聰老師諸多指導與協助，引導我一步步地將論文完成，柯皓仁老師更不遺餘力地點出論文中該注意的細節，心中的感激真是無法以筆墨形容。感謝口試委員林瑞盛老師與黃夙賢老師，在口試時所給予的建議，使得本論文能更加嚴謹。在此向上述老師致上學生最高的敬意與謝意。

同時也要感謝研究室同伴姿婷、筑婷、雅雯，所辦的淑惠，幫本論文蒐集新聞的婷瑱，以及MB208研究室裡陪伴我渡過每個夜晚的伙伴們，不論在研究上或生活上，大家總是一起努力打拚，共同分享喜怒哀樂，你們都是陪伴我一起成長的好夥伴，與大家相處的點點滴滴，都將是我人生中美好的回憶。

最後則是感謝家人的全力支持和包容，讓我能無後顧之憂的朝目標前進，有你們的支持才有現在的我！

林揚書　謹誌

2009 年 7 月

目錄

圖目錄

表目錄

第一章　緒論

1.1 研究背景與動機

在這資訊爆炸的時代，生活中有愈來愈多管道可以取得五花八門的資訊。1970 年 Alvin Toffler 於他的著作 Future Shock[1]中首先提到資訊超載的問題，當能取得的資訊已超越人所能負荷的程度，不但對決策無益，甚至進一步造成干擾現象。

每天一大早起床，許多人的第一件事情就是翻閱報紙、或是打開電視接收最新的新聞內容。一天有上萬件新聞被記者所報導，新聞對於許多人來說擁有極高的閱讀價值，2000 年政大新聞系的大學報「網路成學子閱讀新寵 閱讀率達二成八 直逼報紙 側重娛樂影視休閒資訊」[2]中提到大學生閱讀電子報的比例已跟傳統報紙差不多，但在那麼多的電子報中如何以最短時間去篩選出自己需要的、喜愛的及完全不需閱讀的呢？

1974 年 Katz, Blumler & Gurevitch 所提出的使用與滿足理論[3]中，說明了在三十年前讀者就會主動地去尋找自己所需的資訊。值此網際網路資訊爆炸的時代，每天在電子報網站出現的新聞不計其數，讀者都希望以最短的時間內找到自己所想要看的新聞，正好印證了使用與滿足理論。而電子報網站上預先設定好之「最多人閱讀的新聞」等格式並非完全符合所有人需求，讀者會主動地尋求自己想要的新聞資訊，並非將媒體所給的資訊照單全收，因此本研究試圖以心情類型為導向替使用者篩選、預測出他們所需要新聞內容，節省使用者在選擇有興趣的新聞時所浪費的時間。

2009 年四月南加大傳播學院數位未來中心(the Center for the Digital Future at USC's Annenberg School of Communications)的研究發現[4]，隨著線上新聞讀者人數的增加，每個讀者花在線上閱讀新聞的時間也與日俱增，從 2007 年的網路使

用者每週平均 41 分鐘至 2008 年提升至 53 分鐘。此份研究報告提出造成此現象的四個原因是：

1. 分類廣告逐漸轉移至線上；
2. 關心紙本報紙對環境造成的不良影響；
3. 經濟不景氣；
4. 缺乏對紙本報紙感興趣的新讀者。

數位未來中心的總監 Jeffery I. Cole 認為讀者改變閱讀習慣的速度超乎他的預期，由於新一代的年輕人處在資訊社會中，大多數都有使用電腦和網路的習慣，這些使用者可以在網路上輕易閱讀最新的新聞並且不需付任何費用，導致額外購買紙本報紙的意願低落，造成傳統紙本報紙的新客源逐漸減少，他更認為線上閱讀新聞是未來的趨勢。

2009 年四月中旬，美國建置了一個很特別的「新聞懶人包」網站(Newsy.com)，他們研究現代人「看電視新聞」與「上網看新聞」的習慣[5]，從以前只有少數幾間電視台到現在有上百台可供選擇，而在網路上有非常多的網站提供不同國家、不同地區各式各樣的新聞，進一步發現現代人都以「跳躍」的方式瀏覽資訊，只想要看自己所感興趣的部份、不斷地轉台、跳過一篇又一篇不感興趣的文章，在這選擇自己所需資訊的過程中浪費了不少時間。「新聞懶人包」網站因而採用一種主題式的包裝方式，雇用了一個編輯團隊，隨時觀看線上新聞、報章雜誌、電視新聞，然後將這些統整成一個「新聞短片」，當這網站的瀏覽者想要了解某個議題的完整內容時，只需要點選該議題對應的影片就能得到不同媒體、不同角度、不同觀點對於此議題的綜合整理報導。

每個人在閱讀一篇文章時，先看到人類所使用的文字符號，經由腦部思考後理解文字所敘述的主題，進一步由心理產生對這篇文章的感覺(心情)。近年來很流行在部落格發表文章，由於部落格是一個能讓每個人抒發自己情緒的平台，不少部落格提供在文章後面加註心情標記的服務，因此目前的文章心情偵測研究

中，已有不少將部落格文章做為資料來源，以研究作者心情為目標，分析作者心情是否與週遭環境或是季節、月份間的關係，造成在寫作時的心情差異；但卻僅有少數研究以讀者為目標，分析讀者讀完文章後的心情。

知名的入口網站 Yahoo!奇摩提供了線上閱讀新聞的服務，不同於其它提供新聞網站的在於它讓讀者閱讀後能依自己的心情為新聞選擇適當的標籤，並提供讀者可由之前使用者對新聞所加註的心情評價來做為選擇新聞的依據，而不同的新聞內容與讀者閱讀後的心情是否有特別的關係，如何有效率地由新聞內容偵測讀者可能出現的心情，乃是一值得關注的議題。畢竟在這資訊氾濫的時代，光是選擇電子報種類就會讓讀者頭痛，讀者在心情憂傷時通常不想看到會讓自己難過的新聞，在生氣的時候通常也不會想看到令人更加憤怒的新聞。

綜上所述，本研究之目的在於從讀者角度分析網路電子報文章所帶給讀者的感受(心情)，並將網路電子報文章以心情加以分類，讓讀者能根據自己閱讀新聞後想要感受的心情，過濾並找出適合自己心情的新聞文章。

1.2 研究目的

本研究希望在目前不斷成長的線上新聞閱讀人數以及閱讀者閱讀時間的環境下，以心情為導向為使用者過濾掉他們不期望看到之心情的新聞，節省他們在搜尋自己所需要新聞時花費的時間。本研究將利用特徵挑選(Feature Selection)和支援向量機(Support Vector Machine, SVM)，與Yahoo!奇摩新聞之心情相關資訊結合，期望在新的新聞文章出現時即可預測出讀者看完後的心情，並進一步研究心情與新聞文章分類、時間、詞彙等特徵是否有特殊關係。整體的研究方向如下：

1. 可以對大量的文章進行心情偵測；
2. 找出每個期間內各個心情分類具有代表性的詞彙，並提供給讀者相關新聞題材。

1.3 論文架構

本論文分成五章。第一章說明本研究的動機與目的；第二章介紹文章心情分析相關的研究與方法；第三章敘述新聞文章心情預測系統的設計方式，闡述如何將網路新聞文章處理，導入特徵值挑選並結合LibSVM分類器，將LibSVM分類結果進一步延伸應用；第四章說明實驗結果與分析；第五章為結論與未來改善方向。

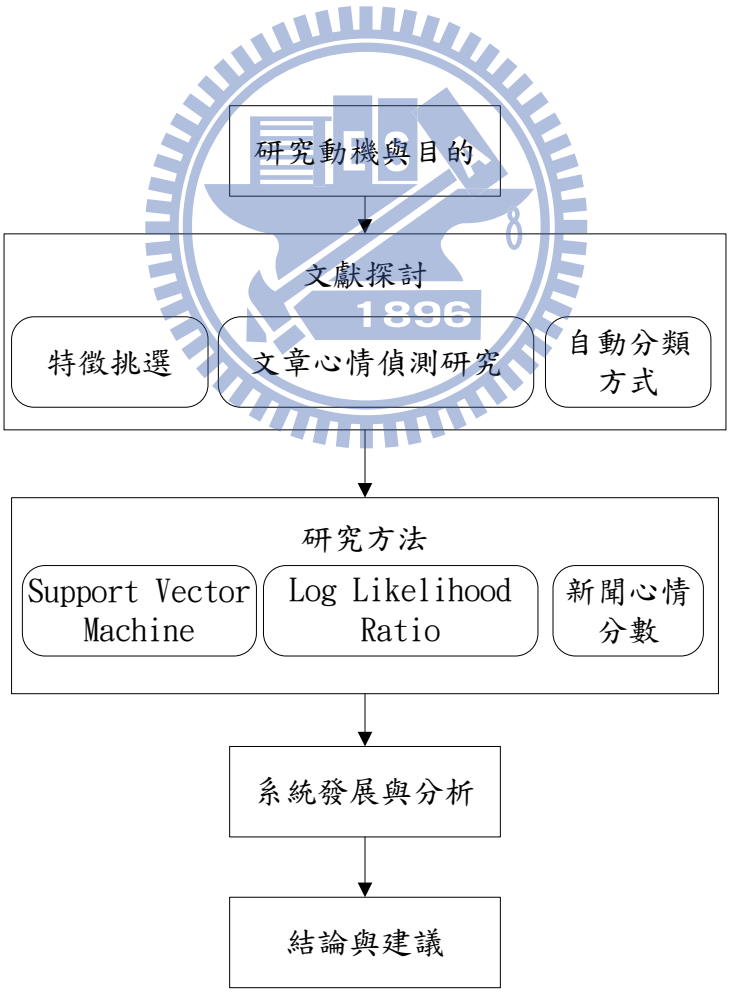

圖 1 論文架構

第二章 文獻探討

本章將介紹與本研究相關的理論與文獻：在2.1節描述文章心情偵測的相關研究，針對不同的文章類型(部落格文章、新聞文章、影評)採用不同的偵測方法；2.2節闡述自動分類的理論與常見的自動分類方法;2.3節中更進一步地介紹與本研究有關的特徵挑選方法與分類器。

2.1 文章心情偵測相關研究

「文章心情偵測」是指當人寫作或閱讀一篇文章，大腦思考後產生對文字敘述的認知，進而產生心理的反應，譬如喜、怒、哀、樂；有些研究著重於寫作者的心情如何反映在文字上、有些則是著重於閱讀者從文字中感受到的心情。本節將介紹與文章心情偵測的相關研究：在2.1.1節中概述以部落格文章為資料來源，以作者心情為偵測目標的研究；在2.1.2節介紹以新聞文章為資料來源，以讀者心情為偵測目標的研究；2.1.3節介紹採用影評為資料來源，以影迷對電影的評價為偵測目標的研究。

2.1.1 部落格文章

目前線上已有不少提供部落格服務的網站，而在部落格中常見的一項功能為功能為提供作者在發表文章後，能為自己的文章標記一個心情標籤，讓其它讀者或往後檢視自己部落格文章時，能知道當時作者的心情(如圖 2 所示)，也提供依心情標籤分類的文章列表供讀者選取(如圖 3 所示)，而要如何以文章的文字訊息就能去預測出作者在寫作時可能的心情狀況，為此類型研究的主要目標。

圖 2 部落格文章編輯的心情標籤

資料來源:部落格(blog.udn.com)

圖 3 心情標籤文章列表

資料來源:部落格(blog.udn.com)

本節將在部落格文章內依詞彙、文章特徵、時間序列三種不同方向的部落格文章心情偵測研究再作更詳細的說明。

1. 詞彙心情強度：

在Building Emotion Lexicon from Weblog Corpora[6]一文中，以Blog文章與其心情標籤作為資料，提出要以詞彙與心情計算出詞彙w在心情e中出現的機會，試圖找出常與某心情一起出現的關鍵字

$$co(e,w) = c(e,w) \times \log \frac{P(e,w)}{P(e)P(w)} \quad \textbf{(1)}$$

P(e,w)為詞彙w在e心情裡出現的機率，P(e)為詞彙在e心情裡出現的機率，P(w)為所有字彙裡出現w字彙的機率，c(e,w)為字彙w在心情e裡出現的次數，co(e,w)為字彙w在心情e中的心情強度分數。

接下來將所有句子拆解成許多詞彙，由詞彙的co心情強度選擇最強的作為此詞彙的心情，再將各詞彙心情結果以投票的方式決定此句的心情類型，最後由Blog的所有句子以投票方式預測出最有可能出現的作者心情。

2. 文章特徵：

Gilad Mishne[7]提出以Blog文章為資料，考慮了字彙頻率、文章長度、詞彙語意特徵(Kim&Hovy[8]及Turney&Littman[9]先前所做的研究結果)、心情的PMI-IR值、文章強調的詞彙、特別符號等特徵來進行分類預測。PMI(Pointwise Mutual Information)是用來計算兩個詞彙(t1、t2)間的結合程度

$$PMI(t_1,t_2) = \log \frac{p(t_1 \& t_2)}{p(t_1)p(t_2)} \quad \textbf{(2)}$$

P(t1)為詞彙t2 出現的機率，P(t2)為詞彙t2 出現的機率，P(t1&t2)為此兩詞彙共同出現的機率。算出兩個詞彙之間的PMI值後，試圖要找出是否有哪些詞彙傾

向在同一種心情內一起出現，並將此特徵拿來作為預測心情的依據，而結果可以提升預測的準確率。後續有Yuchul Jung, Yoonjung Choi和Sung-Hyon Myaeng學者更深入研究，在Determining Mood for a Blog by Combining Multiple Sources of Evidence一文中[10]提出藉由之前所提的特徵值與PMI-IR值一起放到SVM分類器進行分類，效果明顯優於其它分類方法。

3. **時間序列分析：**

Krisztian Balog ,Maarten de Rijke [11]提出了要以時間序列分析的方法檢視Blog使用者的心情資料，將大量附有心情標籤的部落格文章作為資料來源，將所有心情的數量變動趨勢依時間序列分析分成(季節性、有特定的變動趨勢、循環性、以及不規則性等四類)，並討論心情與四類間的關係，例如「酒醉」心情標籤通常出現在深夜、清晨發表的部落格文章；「寒冷」心情標籤就跟著時間一同變化，越接近冬天數量呈現成長的趨勢，夏天就很明顯的降低；「睏」心情標籤就隨著每天的早晨與傍晚呈現循環的變動方式。

Gilad Mishne與Maarten de Rijke[12]將部落格作者所選擇的心情以整體數量的方式來呈現，描繪出在所有時間點每種心情標籤篇數的曲線表，並且與當時所發生的新聞事件一起觀察其中是否有關聯。而作者發現當 2005 年七月時英國倫敦地鐵遭到恐怖份子惡意攻擊時，在生氣、難過等心情分類的曲線圖明顯暴增；作者認為部落格文章心情標籤篇數會受到社會事件影響。

2.1.2 新聞文章

在Emotion Classification of Online News Articles from the Reader's Perspective[13]中作者提出拿先前Blog所建出的情緒詞彙庫及Yahoo!奇摩新聞當作資料，以詞彙的bigram (BI)、詞彙的情緒分數(WE)、詮釋資料(MT)、字綴相似度(AS)、及字彙本身(WD)作為特徵值，將這些特徵值放入LibSVM軟體[14]並使用特定的參數做測試，得到0.7688的準確率，實驗結果也顯示了SVM優於Naïve

Bayes、及其它兩位學者所提出的Passive-aggressive(PA)[15]與Cui's n-gram features(CN)[16]分類法。

Hsin-Yih Lin在Ranking Reader Emotions Using Pairwise Loss Minimization and Emotional Distribution Regression[17]中認為一般心情偵測研究都只預測單一作者寫作文章的一個心情類別，但若從眾多讀者看完文章後可能產生的心情，應該要以排名的方式列出各種可能出現的心情類別名次，才能更符合眾多讀者的需求，並非像先前預測單一作者寫作心情僅用一種預測結果就可滿足。作者採用SVM以Pairwise Loss Minimization方式處理心情標籤名次上比對的問題，降低失誤率，再將Support Vector Regression與情緒分配迴歸結合，描繪出可能的各心情分類頻率比例，最終合併預測出結果。

2.1.3 影評

Sentiment classification of movie reviews using mutiple perspectives[18]的作者將影評依整體、導演、卡司等三部分做影迷可能對一部影片的評價預測，作者認為在預測影迷看完影片的心情是正面或是負面前，需要將影評內容更進一步的分析，才能得到更佳的預測結果。一部電影有人可能對整體評價是正面的，但討厭其中某個演員的表現，如果不細分很難得到一個客觀的預測結果。作者認為要很有效率地分析影評可能要藉由更進步的資訊擷取工具，才能更精確的辨識出影評內容，否則系統效能將受限於此。此篇研究將SVM與優秀的資訊擷取工具(GATE-ANNIE)[19]結合，將影評分三部分以不同角度預測皆達到不錯的準確率。

2.2 自動分類

人類有時會希望將資料進一步處理分類，將資料深入分析與應用、加值處理，依據資料的特性及內容來分到預先設定好的類別，讓資料能更容易管理、更充份地被利用。

現代科技越來越發達，網路蓬勃發展造成交換資訊越來越容易與普及，產生的資料數量大到人類難以有效管理與使用，大量的資料更需要有效的分類方法去妥善處理，許多自動分類的方法也因應而生。文件的自動分類需要預先定義好分類的類別，其目標是將具有相同特徵的文件歸屬到同一類別。與人工分類不同之處在於文件的自動分類處理是採用電腦演算法找出文件特徵，一般的自動分類會先將資料分為訓練資料與測試資料兩部分(如圖 4 所示)：訓練資料係用來找出各類別的特徵，將這些特徵用來當分類依據；測試資料則用來測試分類依據的分類準確率。

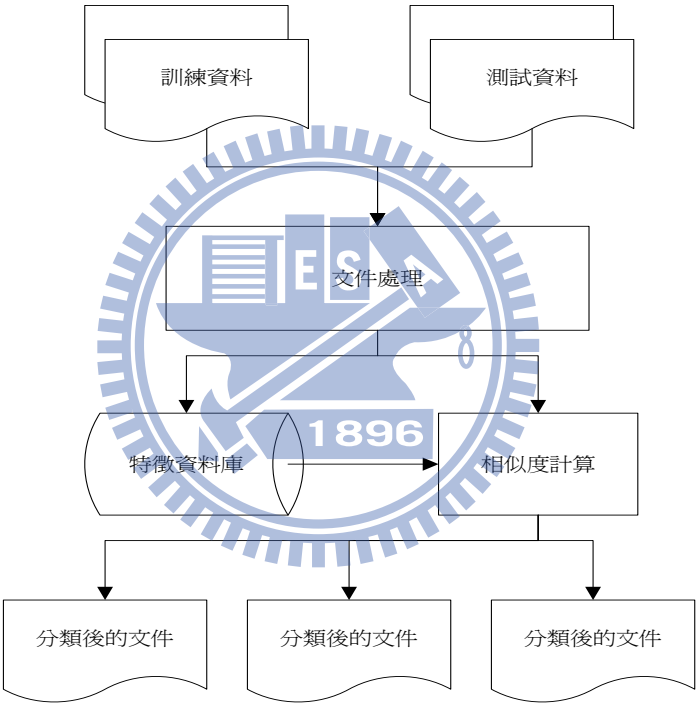

圖 4 自動分類

資料來源:[20]

在自動分類的過程中，需要分類的資料集可以由一群物件 0 所表示，$O=\{O_1,O_2,O_3,...,O_n\}$ 。每一個被分類的物件 O_i ，皆可以用一個集合 S_i 來表示，S_j 是由一連串的特徵值 F 所組成，可表示成 $S_j=\{F_1,F_2,F_3,...,F_n\}$

本節主要介紹幾種自動分類常見的方法，分述如下。

2.2.1 K-Nearest-Neighbor (KNN)

K-Nearest-Neighbor 是一種常見的自動分類方法，它的概念係基於物以類聚的想法，同類通常會聚集在一起，再由最靠近的幾個點(視 K 值而定)的類別去投票，哪一個類別得到最多票就分到哪一類。

舉例而言，若以頭髮長度和臉面積做為特徵值來分辨人的男女，則可將訓練資料標示在二維平面上(如圖 5 ，X 軸表示頭髮長度，Y 軸表示臉面積大小，黑點代表男生，白點代表女生)。

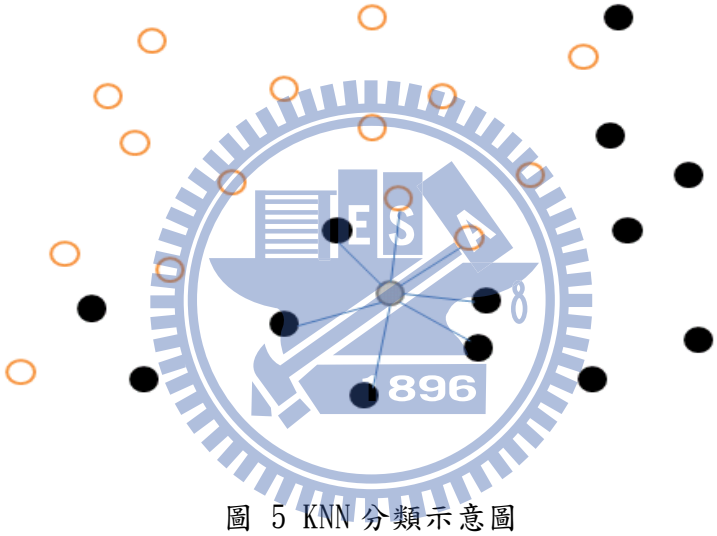

圖 5 KNN 分類示意圖

當有一筆新的資料要用 KNN 演算法做預測時，會先依它的頭髮長度以及臉面積大小，在此平面上標出此點(灰點)，假設是用 K=7 的 KNN 演算法，則由圖 5 所示，最靠近此點的 7 個點裡有 5 個黑點，2 個白點，黑色的點佔了大多數，故依 KNN 演算法的概念將此點分類到男生的類別。

但仍有幾種狀況是選擇使用 KNN 前需要考慮的。第一個是如何選擇適合的特徵去表示資料的空間，而前面所用的頭髮長度及臉的面積當特徵，常見的例外像是玩樂團留長髮的男生，女生臉的面積天生就較大，都會造成分類判斷上的錯誤。假若改採身高、體重來當作特徵，仍然會出現此問題。另外一個是距離的問題，若使用頭髮長度及臉部面積來做特徵，一個只是長度另一個卻是面積，在座

標圖上該以什麼比例取距離來呈現，距離函數該如何設定，這也會影響到分類的結果。還有許多的特徵無法以距離來量化，像是膚色、髮色這些特徵無法量化。語義的問題不適用於 KNN 處理，假設有暗紅色、桃紅色、粉紫色，依字面上來看暗紅色與桃紅色僅差一個字，應該距離會較近些，跟粉紫色會差很遠因為並無相同的字串出現，但是實際上依鮮豔程度其實桃紅色與粉紫色應該較近些。

2.2.2 決策樹(Decision Tree)

決策樹是一種模擬決策過程可能面臨到的抉擇流程圖，一般來說決策過程中考量到許多不同的狀況、組合、選擇，依此想法建立成抉擇流程樹狀圖即為決策樹。在 Kdnuggets[21]所做的票選中，決策樹是在資料探勘領域最常被使用的一項技術(圖 6)。其核心技術概念是很淺顯易懂，如圖 7 所示，可用來分析複雜的情況，透過很簡單的概念畫出流程，產生樹狀規則的結構圖，將相同特徵資料依樹狀圖規則分到同類。一般常見的 Decision Tree 演算法有 ID3，C4，C4.5，C5，CART，CHAID，QUEST。

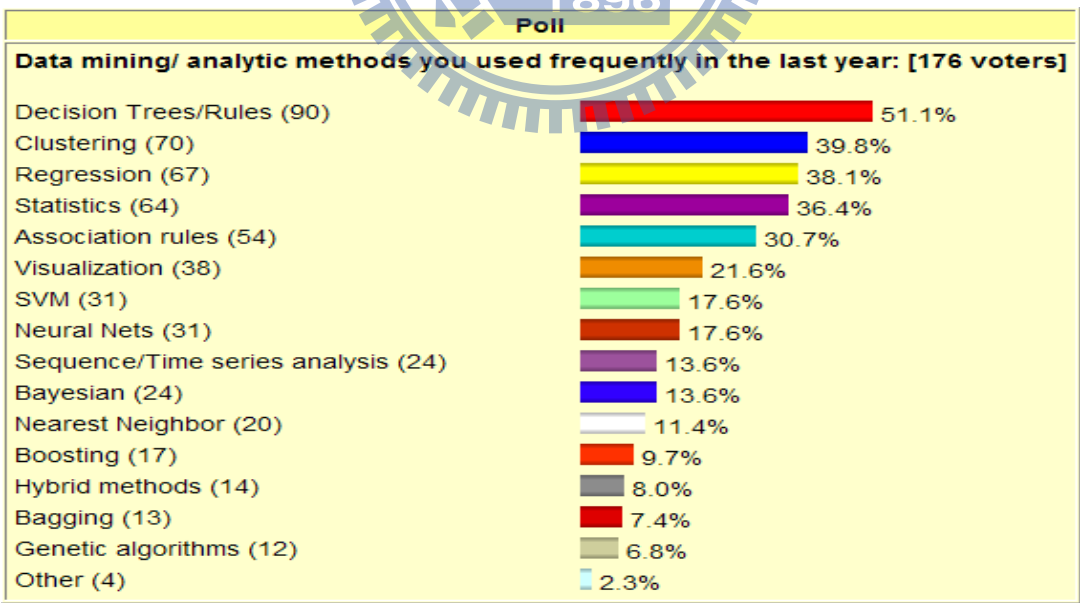

圖 6 最常用的資料探勘技術票選結果

資料來源: [21]

建立 Decision Tree 的流程如下[22]：

1. 定義在達到目標前可能會做出的所有屬性選擇；
2. 選擇屬性的增益比值或是資訊報酬較佳者做為分岔，產生子節點；
3. 根據每個子節點的案例分派狀況設定分類結果；
4. 讓決策樹不斷成長，並在最終結果上採用修剪技術移除不必要的規則；
5. 最後的決策樹結構描述了每種選擇組合的結果，並標示在樹枝的末端上。

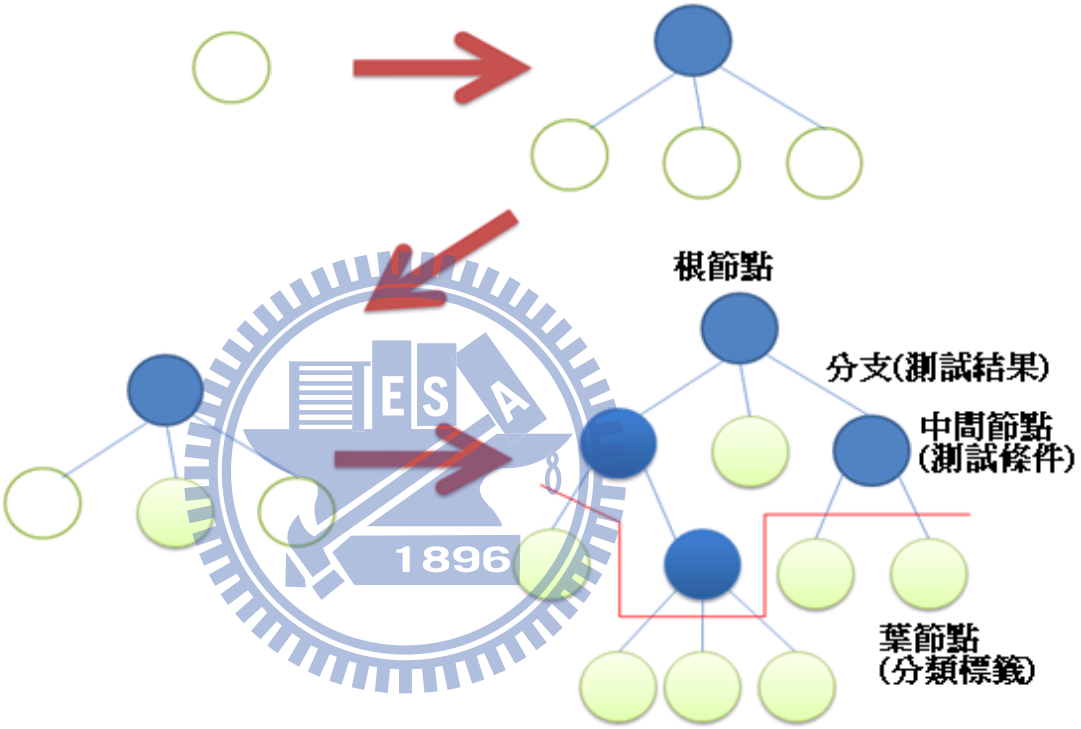

圖 7 決策樹建立過程

要建立優秀的決策樹必須盡可能地找出最佳分隔，最佳分隔指的是當決策樹分類出來的群組皆屬於同一個類別，這也是在分類過程後最期望達到的結果，要找出最佳分隔必須先採用最適合的分隔準則，常見的分隔方式有三類[22]：

1. 分隔候選值，當此 X 值(分隔所採用的變數)比常數 N 小時，就會被分配到一個子節點，但當 X 值等於或是大於 N 時，會分配到另一個子節點
2. 類別分隔：依所採用分隔變數的特性、種類建立類別分支，例如，可以是性別即可分為男、女兩類別。
3. 面臨遺漏值分隔:將零視為一個合理且具分支的值或將該筆資料取代、丟棄

在以數值方式分隔中常用的方式有以下兩種[22]：

I. 變異減化(Reduction in variance)：

以測量母節點與子節點資料值與平均值的變異情況計算，將結果視為變異數，以此數值來作為調整的依據。

II. F-test：

計算兩組樣本是從同一母體產生的機率(平均數、變異數、樣本數不一定相同)，當樣本數夠大且為隨機樣本時，樣本的變異數是母體變異數很好的估計值。

以類別方式分隔常見的方式有以下三種

吉尼索引值(Gini index):

I. 當資料為連續屬性資料時經常被採用，Gini 又稱 Population diversity，計算從同一個母體，隨機選取兩筆資料，此兩筆資料屬於同一種類的機率。樣本集合 D 中包含 n 類樣本，且每一種預測值在該節點中的出現頻率為 Pj，則吉尼索引法將該樣本集合 D 的 Gini score 定義為

$$Gini(D) = 1 - \sum_{j=1}^{n} p_j^{\,2} \quad \textbf{(3)}$$

熵亂度(Entropy)：

II. 源自於 Information Theory ，在 1949 年由 Shannon 與 Weaver 所提出，在決策樹中，用來定義一個節點的同質性。同質性高者包含較少資訊，因此熵亂度比較小。若一節點包含 n 種預測值，且每一種預測值在該節點中的出現頻率為 pi ，則該節點的熵亂度為

$$-\sum_{i=1}^{n} p_i \times \log(p_i) \quad \textbf{(4)}$$

資訊獲利率(Information Gain Ratio)：

III. 以某一屬性作為決策樹節點，由其所產生的子決策樹熵亂度與物件集合的熵亂度所決定，假設訓練資料中的集合 S 有 n 個類別 Ci, i=1, 2, 3, …, n，每個類別的資料個數皆可以 freq=(Ci, S)表示，|S|代表 S 中所有資料個數，則各類別其資料出現的機率為

$$\frac{freq(C_i,S)}{|S|} \quad (5)$$

，根據資訊增益理論，各類別資訊為

$$-\log_2(\frac{freq(C_i,S)}{|S|}) \quad (6)$$

由各類別的資訊量可以計算出訓練集合的平均資訊量(亂度)，將所有類別的資訊量乘上各類別資料的出現機率總合可表示為

$$\inf o(S) = -\sum_{i=1}^{n} \frac{freq=(C_i,S)}{|S|} \log_2(\frac{freq=(C_i,S)}{|S|}) \quad \textbf{(7)}$$

據 info(S)的計算方式，當集合的某屬性 A 切割成多個集合時，其分割後所佔的資訊量與各子集合的資訊量乘上各子集所佔的比例總合兩者相等，如以下所示：

$$\inf o_A(S) = -\sum_{i=1}^{n} \frac{|S_i|}{|S|} \times \inf o(S_i) \quad \textbf{(8)}$$

因此集合 S 由屬性 A 分割後所獲得的資訊量則為分割前的資訊量減去分割後的資訊量，表示為

$$gain(A) = \inf o(S) - \inf o_A(S) \quad \textbf{(9)}$$

而決策樹用在統計分析上的優點有[22]：

1. 能清晰地表示出重要欄位；
2. 能產生容易理解的規則；

3. 計算次數較少；

4. 輸出即有包含屬性排序的結果。

缺點則有：

1. 不易分析具有連續值特性的屬性；

2. 對於有順序特性的資料，前置處理的負荷(Pre-processing loading)較重；

3. 當類別愈多，誤差愈高；

4. 無法處理含有空值的資料；

5. 不易於分類多個屬性資料。

2.2.3 Support Vector Machine

支援向量機(Support Vector Machine,SVM)是在 1970 年代，由學者 Vapnik[23]提出，是一種由統計理論發展出來的機器學習技術，1998 年時 Joachims[24]、Taira 與 Haruno[25]等學者在文件分類以及 Kudo 和 Matsumoto 在名詞組標[26]示(利用 SVM 辨識中文基底名詞組的初步研究)讓它結果更優於其它方法的準確性，近年來 SVM 也常應用於資料探勘、影像辨識、文字分類等領域，而在與本研究相關的自然語言處理領域中更涉及了語意分析、詞性標記、未知詞辨識等，都有相當不錯的準確率。

SVM 通常用來處理兩類別的問題，但亦可處理多類別的問題。在處理兩類別的問題時先將訓練資料以+1 或是-1 加以標註(代表不同類別)，以數學式表示為

$$\{x_i, y_i\}, i = 1 \ldots l, y_i \in \{1, -1\}, x_i \in R^d \qquad (10)$$

假設有一個超平面可以將+1 及-1 的資料加以區分，則此超平面就可稱為區分平面(Separating Hyperplans)，若在此超平面上的 x 必須滿足：

$$w \cdot x + b = 0 \qquad (11)$$

w 為超平面的法向量。而 SVM 的目標是要在高維度的特徵空間，找出一個具有最

大邊界(margin)的區分平面來隔開兩類資料，如圖 8。

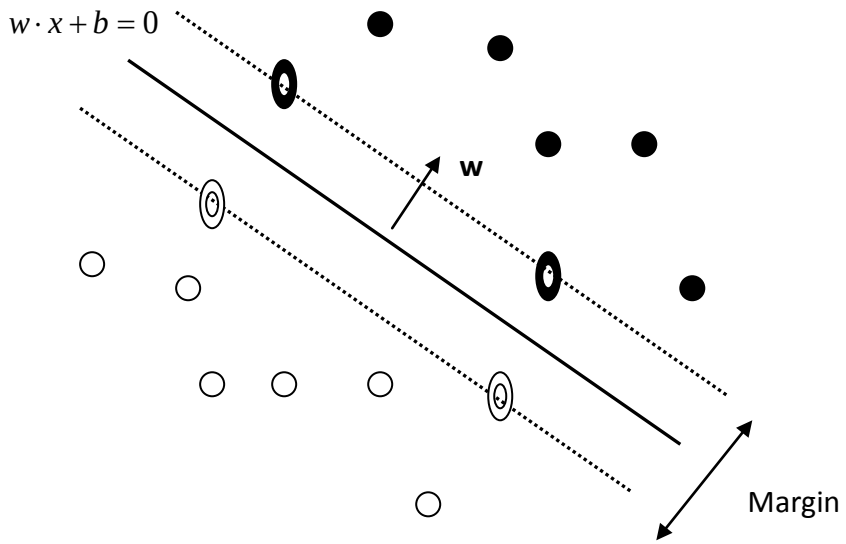

圖 8 支援向量機

SVM 的優點如下[27]

1. 有較強的泛化能力以及低複雜度
2. 它可使用少量樣本得到最佳解
3. 優越的分類與預測能力
4. 非線性資料的效能優越

缺點則有：

1. 測試的資料數目過大時，需要花比較多的時間；
2. 隨著訓練資料集的數量線性成長，一方面可能造成過度調適（overfitting）的問題，另一方面則浪費計算時間。
3. 無法得到機率式的預測。一般人會比較偏好機率式的預測，因為機率式的預測能夠給人確定程度的資訊。
4. SVM 的使用者必須給定一個誤差參數 C，這個參數對結果有很大的影響。不幸的是，大部分的情況下，使用者都必須猜過各種可能值，才能找最好的結果

目前 SVM 的工具很多，本研究係採用 LibSVM，因此以下略為介紹 LibSVM:

LibSVM[14]:

是台灣大學林智仁(Chih-Jen Lin)博士等開發設計的一個操作簡單、易於使用、快速有效的通用 SVM 套裝軟件，可以解決分類問題（包括 C- SVC、n - SVC ）、回歸問題（包括 e - SVR、n - SVR ）以及分佈估計（one-class-SVM ）等問題，提供了線性、多項式、徑向基和 S 形函數四種常用的核函數供選擇，可以有效地解決多類問題、交叉驗証選擇參數、對不平衡樣本加權、多類問題的機率估計等。

LibSVM 結合了 SVM-light 和 Keerthy 改良過的 Sequential Minimal Optimization[28]的算法，在很多地方考慮的比 Joachims[29]還要仔細。LibSVM 是以 SVM-light 為基礎來發展卻沒有完全承襲 SVM-light 的想法。在不同資料的分類種兩者分類結果各有優劣。

本論文採用支援多類別分類的 LibSVM，通常 SVM 設定的參數都是以嘗試錯誤的方式(try and error)去慢慢嘗試盡可能找出適合分類的最佳參數，通常會調整 SVM 的在分類錯誤時的懲罰參數 C，去加重分類誤差的權重，或是選擇不同的核心函數，如表 1 所示，而 LibSVM 已有開發不少相關的程式(model selection tool-grid.py)可降低使用者在嘗試參數上所花費的時間。

表 1 LibSVM 核心函數

核心函數	公式
Linear	u'*v
Polynomial	(gamma*u'*v + coef0)^degree
Radial Basis function	exp(-gamma*|u-v|^2)
sigmoid	tanh(gamma*u'*v + coef0)

參考來源: [14]

使用 LibSVM 的流程如下圖 9 所示：

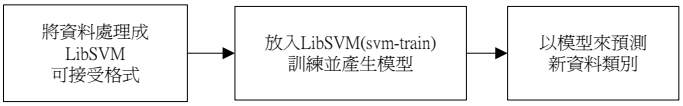

圖 9 LibSVM 分類流程圖

2.3 特徵值挑選(Feature Selection)

在處理文件分類時，通常會將整篇文章斷詞切字，處理成以詞彙的方式呈現，再將一些不值得考慮的停用字刪去，但是當資料量大時，會出現數十萬個詞彙，容易造成分類器在分類過程中採納的特徵數量過於龐大，可能導致分類效果不佳與執行速度緩慢等問題，因此需要倚靠特徵值挑選過濾出值得考慮的詞彙，降低資料維度、提升系統效能，接下來將介紹本研究有採用的特徵挑選方法。

2.3.1 TF-IDF

在資訊檢索、資訊探勘的領域中，TF-IDF 常被用來對每個 term 依其重要性做額外加權的動作。採用統計的概念，用來評估一個 term 在這份文件中與整體文件間的重要性，TF(Term Frequency)： 概念起源於 Luhn[30]在自動索引的實驗裡發現，term 在經由分析之後可分為高、中、低頻三類，而高頻及低頻不具任何意義，但中頻(middle-frequency)的詞彙大多數都較有意義，Luhn 再進一步提出 Resolving Power of Significant words 的概念[30]，如圖 10 所示，在中頻的部分是作者認為比較具有參考價值的。

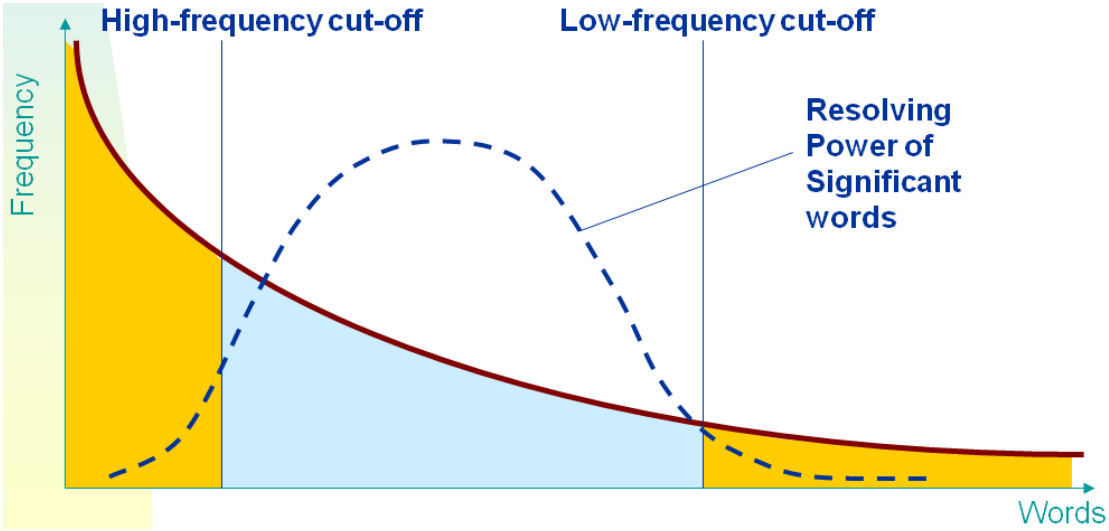

圖 10 詞頻與 Resolving Power of Significant words 關係

參考來源：[31]

IDF(Inverted Document Frequency)：由 Spark Jones 於 1972 年所提出來的[32]，是一個詞彙普遍重要性的度量。某一特定詞彙的 IDF，可以由總文件數目除以包含該詞彙之文件的數目，再將得到的商取對數計算出來。

2.3.2 Log Likelihood Ratio

Likelihood Ratio 指的是在兩個不同的假設下找出它最大機率的比例，通常用一個希臘字母 Λ 來表示，而 Likelihood Ratio 檢定是一種以兩個假設為基礎數值做決定的比例統計方法。也可以說是比較兩種模型的方法，原理是將兩個模型拿來比較，在同樣的資料集下找出最適合的模型。主要是以 Log likelihood ratio 檢定統計量在樣本很大時、近似於卡方分配的性質的概念來處理，計算方式為在虛無假設下找出最大的 likelihood ratio 值

在虛無假設下參數 θ 屬於參數空間 θ 的特定子集合 θ_0 中，likelihood 公式可進一步表示成一個參數 θ 與一個實際觀察後擁有固定值的參數 χ [33]

$$\Lambda(x)=\frac{\sup\{L(\theta\mid x):\theta\in\Theta_0\}}{\sup\{L(\theta\mid x):\theta\in\Theta\}} \qquad (12)$$

這是一個表示資料 χ 在 Likelihood Ratio 函數中的統計量。當這個統計量太小時 Likelihood Ratio Test 會否定虛無假設。如何衡量這個統計量的多寡標準，取決於 Type I error 機率[34]的容忍程度。

在 Pearson 檢定拋擲硬幣的例子中[34]，他嘗試著比較兩個硬幣同時出現「正面」的機率。他把硬幣出現的機率情況以表 2 表示，行代表硬幣出現正面及反面出現的次數，列代表的是硬幣 1 及硬幣 2 出現的次數：

表 2 兩個硬幣投擲結果

	正面	反面
硬幣 1	k_{1H}	k_{1T}
硬幣 2	k_{2H}	k_{2T}

這裡的 Θ 包含了下列幾個參數： p_{1H} 、 p_{1T} 、 p_{2H} 、 p_{2T} ，這四個參數為表 2-2 所列情況出現的機率。將這個假設空間定義成一個符合一般分佈限制的 H 空間：$0\le p_{ij}\le 1$，且 $p_{iH}+p_{iT}=1$。當虛無假設 H_0 的 $p_{1j}=p_{2j}$ 時為其子空間。在以上的條件中 i=1，2 及 j=H，T

在假設 H 與 p_{ij} 的情況下最大的可能性 n_{ij} 可表示成

$$n_{ij}=\frac{k_{ij}}{k_{iH}+k_{iT}} \qquad (13)$$

在假設 H_0 與 p_{ij} 的情況下最大的可能性 m_{ij} 可表示成

$$m_{ij} = \frac{k_{1j} + k_{2j}}{k_{1H} + k_{2H} + k_{1T} + k_{2T}} \tag{14}$$

由表 2.2.4 的情況可再將 Log Likelihood ratio 表示成

$$-2\log\Lambda = 2\sum_{i,j} k_{ij} \log \frac{n_{ij}}{m_{ij}} \tag{15}$$

許多特徵挑選都會有存在一個問題，出現在某分類裡頻率高的特徵內容裡，同時包含了在其它分類裡也很常出現的、在其它分類裡很少出現的等情況，而使用這種方法概念可以篩選出僅在此分類裡經常出現，在其它分類裡鮮少出現的特徵內容。將此概念運用至本研究的前置處理過程裡，試圖找出僅在每項心情分類中出現的特徵詞彙，進一步降低資料維度來增加分類系統效能。本研究考慮前面章節所提的 LLR($-2\log\Lambda$)[33]，由於此方法在處理圖書資訊等文件分類情況正確率高達八、九成，故本研究採用此方法來作為前置處理的特徵挑選。

第三章 系統設計

本章將闡述本研究中所提出之新聞文章心情分類系統的設計結構與使用的方法，3.1 節介紹本系統所使用的資料；3.2 節介紹此系統整體結構；3.3 節介紹系統前置處理；3.4 描述特徵值之挑選方法，用以篩選出特徵詞彙並放入 SVM 分類系統進行訓練與分類；3.5 節將 SVM 分類系統與特徵挑選出的詞彙結合，延伸開發出關鍵字詞彙系統。

3.1 資料收集

本研究收集台灣 Yahoo!奇摩新聞中的部分新聞以及該新聞所附帶的心情標籤與投票數據為資料集。

在本系統所用來訓練的每一篇新聞文章中，皆附帶一個心情的標籤[35](包含新奇、溫馨、誇張、難過、實用、高興、無聊、生氣等八類心情)，該分類取決於所有閱讀某篇新聞的讀者依其閱讀感受的心情投票，票數最高的分類，即為該篇文章的心情分類。當針對一篇文章的投票人數愈多，則該文章所附帶的心情分類愈能反映大多數讀者閱讀後所呈現之心情寫照。除了心情標籤，每篇新聞文章依其文章內容也具有一個新聞分類，Yahoo!奇摩新聞有政治、社會、地方、國際、財經、科技、運動、健康、教育、藝文、影劇、旅遊、生活等十三種新聞分類。

圖 11 為 Yahoo!奇摩新聞網頁的其中一篇新聞，本研究係以新聞分類、心情分類標籤、新聞標題、新聞文章內容、新聞日期、八項心情分類投票比例做為資料的來源，每收集一篇新聞資料，資料庫將記錄所使用的新聞分類(news_cate)、心情標籤(mood_cate)、新聞標題(news_title)、新聞文章(news_text)、新奇心情投票比例(mood1)、溫馨心情投票比例(mood2)、誇張心情投票比例(mood3)、難過心情投票比例(mood4)、實用心情投票比例(mood5)、高興心

情投票比例(mood6)、無聊心情投票比例(mood7)、生氣心情投票比例(mood8)、以及投票人數(voters)、新聞日期(news_date)等內容。

圖 11 新聞文章資料

資料來源：[35]

以圖 11 為例，假設要收集一篇以「84 年次出生 確定免當兵」為標題的新聞，則資料庫會記錄為如表 3。

表 3 蒐集新聞實際範例

news_cate	mood_cate	news_title	news_text
11(影劇)	8(生氣)	84 年次出生 確定免當兵	…國防部已經確定民國八十四年元旦以後出生的免當兵，也就是今年國二升國三的役男不用當兵了…

mood1	mood2	mood3	mood4	mood5	mood6	mood7	mood8	news_date	voters
3	0	10	13	2	23	9	40	2009-7-15	125

3.2 系統整體結構

本研究所提出之分類系統採用多元分類器，每篇新聞文章移除停用字與不經採用的詞性詞彙後，僅採用部分的特徵詞彙做為分類依據，並送到分類器進行處理。延伸應用部分將特徵詞挑選、分類預測、新聞文章心情分數、時間特性結合，採納多種特性將結果做不同的呈現與進一步推薦。

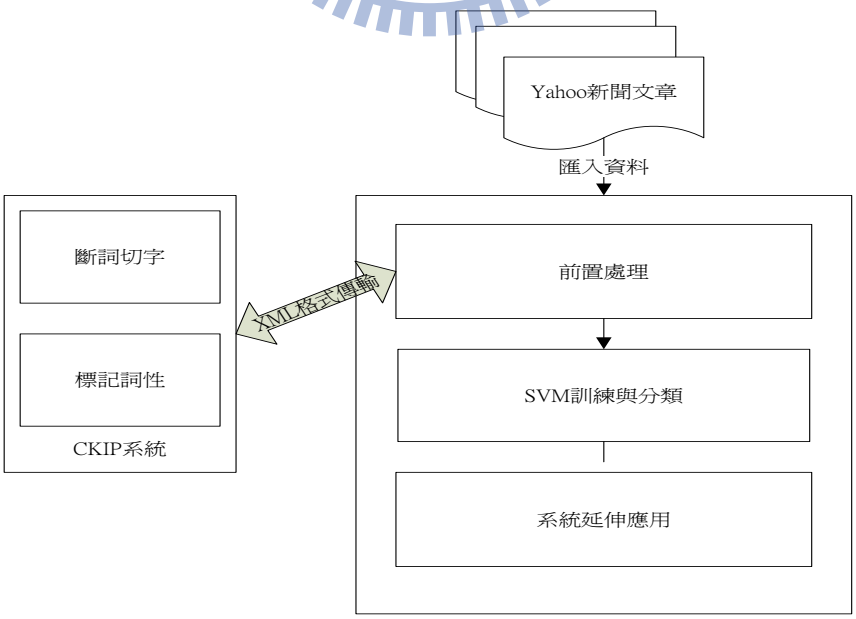

圖 12 系統整體架構圖

3.3 前置處理

當一則新聞要經過分類器預測出其代表的心情前，將經過斷詞切字，讓文章改以每個「詞」的方式呈現;但並不是每個詞對分類過程皆有幫助，過多不重要的詞反而造成分類過程中的誤差，容易誤判結果，所以資料的前置處理是非常重要的，良好的前置處理可以提升系統效能、降低誤差。本研究將前置處理分成以下四部分進一步討論：斷詞切字、詞性標記、移除停用字、特徵挑選。(如圖 13 所示)。

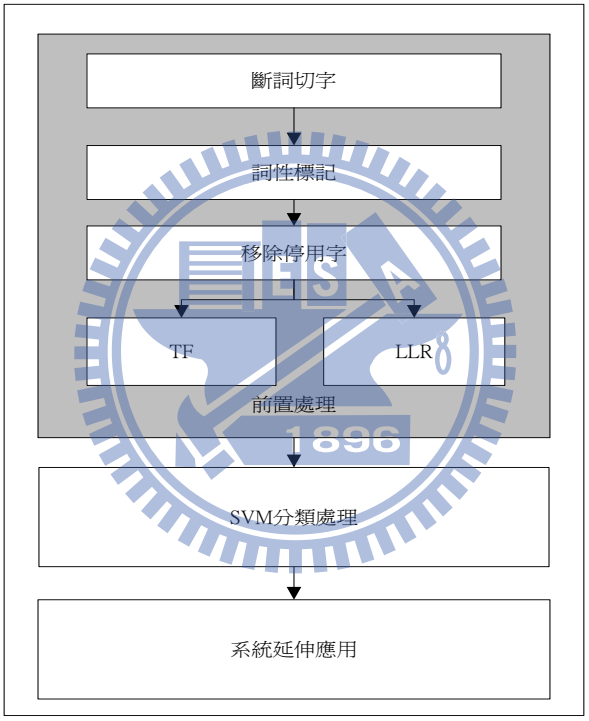

圖 13 系統架構-前置處理

3.3.1 斷詞切字

在文章處理過程中，最常用的基本單位是「詞」，而「詞」所代表的是由語言學家所定義的「能夠獨立運用，具有完整語意的最小語言成份」[36]。但在不同語言中，對「詞」的處理方式也不同，例如在英文中每個單字都是具有意義的，可直接由字與字間的空白將全部的詞輕易斷出，僅有少數的片語需要額外處理;

但在中文的表達方式中，無法使用空白符號將每個字準確地切割出來，可能一個字串裡面包含了許多「詞」，不像英文一個字串就代表一個「詞」，所以要處理中文必須使用一套更精準的斷詞切字方法。

本研究採用中研院中文詞知識庫小組所開發的 CKIP(Chinese Knowledge Information Processing Group)[37]系統來處理斷詞切字的工作。由於中文詞集是一個隨時都在增加的開放集合，並不存在任何一個辭典能涵蓋所有詞彙，所以在針對特定領域的研究中，必須進一步地加入該領域特有的詞彙，才能減少錯分的情況。中研院的 CKIP 系統也有處理此種狀況的機制，可以自動搜集新詞建立該領域用詞，具有自動擴增新詞與辨別詞彙詞性的功能，本研究使用此系統的線上斷詞切字服務 API，只需要將檔案以 XML 格式型態建立 SOCKET 傳遞即可迅速獲得結果。

3.3.2 詞性標記

本研究所使用的 CKIP 斷詞切字系統亦可同時標記詞性，CKIP 採用中央研究院詞庫小組八萬目辭典中的 178 個詞類經簡化後所得的 43 個標記，另外針對特殊符號增加三個標記。而本研究只採用其中較具代表性的動詞與名詞，也就是精簡標記 Vi、Vt、N、Nv 等四類精簡。以下舉一個實際操作的例子。新聞標題為：「台鐵端午車票熱賣 高鐵宣布開加 58 班列車」，經過中研院 CKIP 斷詞切字處理系統處理過後，可得到如圖 14。

台鐵(Nc)　端午(Nd)　車票(Na)　熱賣(VD)　高鐵(Na)　宣布(VE)　開(VC)　加(VC)　５８(Neu)　班(Nf)　列車(Na)

圖 14 前置處理第一階段－斷詞切字與詞性標記

但由於詞類細分為太多種類，CKIP 線上服務系統會進一步地處理成精簡的詞彙標記如圖 15：

台鐵(N)　端午(N)　車票(N)　熱賣(Vi)　高鐵(N)　宣布(Vt)　開加(N)　58(DET)　班(M)　列車(N)

圖 15 前置處理第二階段－精簡詞彙標記

而本研究只採用其中的四種詞性，將 CKIP 系統所回傳的詞彙進一步挑選後僅餘如圖 16 的結果:

台鐵(N)　端午(N)　車票(N)　熱賣(Vi)　高鐵(N)　宣布(Vt)　開加(N)　列車(N)

圖 16 前置處理第三階段－僅挑選名詞和動詞

3.3.3 刪除停用字

停用字指的是，有些字在文章中會出現很多次，但字的本身卻不任何意思，往往僅用來修飾語句，平順語意。而在分類的過程中，由於部分停用字出現頻率很高，容易造成在分類時的雜訊進而影響到分類結果。例如：你、我、他等代名詞，以及較不重要的介系詞、語助詞。

除了一般常見被許多人所認定的停用字外，在特定領域裡有些字的出現也是不太具有意義。當新聞文章中一再出現：記者、報導、新聞等字，這些字也屬於本篇研究所使用的動、名詞類型，但是在此領域中出現這些字是不具有任何意義的。故本研究整合了中央研究院平衡語料庫詞集及詞頻統計[38]使用頻率最高的前 100 個詞彙作為停用字，以及 Oracle Text Reference[39]的停用字表，再額外加入新聞文章類的停用字，期能 提升系統準確率。本研究所用的停用字列表如表 4。

表 4 停用字列表

你(N)，它(N)，他(N)，我(N)，我們(N)，你們(N)，阿(N)，妳(N)，妳們(N)，他們(N)，自己(N)，她(N)，人(N)，是(Vt)，上(N)，後(POST)，到(Vt)，無(Vt)，小(Vi)，們(N)，今(N)，好(Vi)，後(POST)，者(N)，大(Vi)，那(ADV)，年(N)，時(POST)，說(Vt)，有(Vt)，個(M)，這(DET)，種(Vt)，中(N)，讓(Vt)，此(DET)，做(Vt)，沒有(Vt)，位(Vt)，想(Vt)，其(DET)，高(Vi)，沒(Vt)，何(N)，不同(Vi)，一(DET)，兩(DET)，各(DET)，每(DET)，次(M)，三(DET)，目前(N)，希望(Vt)，有關(Vt)，包括(Vt)，最近(N)，是(Vt)，引起(Vt)，最後(N)，加強(Vt)，繼續(Vt)，有(Vt)，了解(Vt)，過去(N)，任(Vt)，左右(N)，經過(Vt)，使得(Vt)，相關(Vi)，當時(N)，進入(Vt)，現在(N)，需要(Vt)，原因(N)，如此(Vi)，什麼(DET)，問題(N)，學生(N)，表示(Vt)，公司(N)，大家(N)，記者(N)，新聞(N)，媒體(N)，採訪(Vt)，報導(N)，發現(Vt)，台灣(N)

3.3.4 特徵挑選

本研究嘗試運用詞頻(Term Frequency)與 Log Likelihood Ratio Test 等二種方式，來選擇對新聞文章心情分類有所助益的特徵，分述如后。

1. 詞頻(Term Frequency):

在高中學英文的過程中，許多英文老師都會要求學生去背教育部所統計最常使用的英文 7000 字、8000 字、甚至到 10000 字。這麼做的用意主要是因為這些字經常出現，先學會這些常出現的單字對於學習英文的幫助很大，顯見衡量詞的重要性時經常會考慮到詞頻。在文件分類的過程裡，詞頻高的詞通常會包含一些人類常用詞以及該類別裡重要的詞，無法光倚靠詞頻就去斷定此詞與該分類間是否有關係，本研究中以 Yahoo!奇摩新聞的八個心情分類依詞頻最高的前二十名

如下表 5 與表 6 所示

表 5 詞頻最高的前二十名詞彙(全詞彙)

詞彙	出現次數	詞彙	出現次數
的	7349	都	4362
在	6488	個	4274
是	6170	人	4138
一	6055	還	4096
有	5634	說	4072
也	5343	上	3711
不	5107	但	3653
這	4696	會	3569
了	4523	到	3560
就	4457	後	3493

由這八個分類所統計的詞頻表 5 與表 6 中可發現,在詞頻前二十名的詞中仍夾雜許多人類慣用字,必須再藉助其它特徵挑選方法才有辦法找出每個分類中的特徵詞。

表 6 詞頻最高的前二十名詞彙(動詞與名詞)

詞彙	出現次數	詞彙	出現次數
是	6170	大	2632
有	5634	前	2507
人	4138	沒有	2401
說	4072	為	2313
到	3560	報導	2147
要	3439	發現	1942
讓	3360	自己	1941
他	3344	沒	1882
表示	2849	她	1849
記者	2668	得	1740

2.Log Likelihood Ratio Test：

令 S_i 為訓練資料中屬於類別 M_i 的新聞資料，$\overline{S_i}$ 為訓資料中不屬於類別 M_i 的新聞資料，並針對詞彙 t_j 死類別 M_i 提出下列兩個假設

假設 1 $P(S_i \mid t_j) = p = P(S_i \mid \overline{t_j})$，文件間是否有關聯與 t_i 無關;

假設 2 $P(S_i \mid t_j) = p_1 \neq p_2 = P(S_i \mid \overline{t_j})$，文件間是否有關聯與 t_i 是否存在，有很大的關係，由此可推論 p_1 遠大於 p_2

仿表 2 列出本研究各種可能的情況表，表 7 中，n_1指的是詞彙 i 在心情分類 j 中的出現次數；n_2指的是詞彙 i 在非心情分類 j 出現的次數；n_3指的是除了詞彙 i 以外其它詞彙在心情分類 j 中的出現次數；n_4除了詞彙 i 以外其它詞彙在非心情分類 j 中的出現次數(採用 Document Frequency 計算)。

表 7 詞彙與心情類別關係

	S_j	$\overline{S_j}$
t_i	n_1	n_2
$\overline{t_i}$	n_3	n_4

假設機率分佈情況為二項式分佈

$$b(k;n,x)=\binom{n}{k}x^k(1-x)^{(n-k)}$$

可將H_1與H_2改寫成

$$L(H_1)=b(n_1;n_1+n_2,p)b(n_3;n_3+n_4,p)$$

$$L(H_2)=b(n_1;n_1+n_2,p_1)b(n_3;n_3+n_4,p_2)$$

並代入$-2\log\Lambda$

$$=-2\log\frac{L(H_1)}{L(H_2)}$$

$$=-2\log\frac{b(n_1;n_1+n_2,p)b(n_3;n_3+n_4,p)}{b(n_1;n_1+n_2,p_1)b(n_3;n_3+n_4,p_2)}$$

$$=-2((n_1+n_3)\log p+(n_2+n_4)\log(1-p)-(n_1\log p_1+n_2\log(1-p_1)+n_3\log p_2+n_4\log(1-p_2)))$$

其中機率 $p=\frac{n_1+n_2}{n_1+n_2+n_3+n_4}$, $p_1=\frac{n_1}{n_1+n_2}$, $p_2=\frac{n_3}{n_3+n_4}$

而本研究所計算出的八種新聞心情 LLR 排名前十的詞彙如表 8、表 9、表 10 所示。在表 8、表 9、表 10 中，可以看出一些很明顯的特徵：實用類別的詞彙大多跟醫學報導、常識有關係，這些詞彙容易讓人覺得這類型新聞很實用；高興類別的詞彙跟運動勝利有關係，感覺運動類型報導容易讓人很高興有意願去投票，但是詞彙比較偏向勝利的運動新聞，也有可能是因為通常落敗的標題讓人點閱意願或是投票意願低落所造成的；無聊與生氣兩個類別的詞彙都與政治有關係，不分藍綠政黨都易使人在閱讀時產生負面的情緒。

表 8 新奇、溫馨、誇張三類新聞心情分類 LLR 前十名

新奇		溫馨		誇張	
詞彙	LLR 值	詞彙	LLR 值	詞彙	LLR 值
科學家	360.1088	感動	583.7297	警方	447.7907
研究	260.3479	愛心	479.8703	男子	364.9363
人類	210.5044	感謝	288.7076	女子	149.8070
嘖嘖稱奇	196.8792	照顧	273.1356	台灣	147.6521
生物	182.8508	孩子	250.9976	發現	142.1874
牠	161.9498	捐出	227.3476	烏龍	124.7958
研發	160.8925	志工	224.8288	報案	115.2022
廟	158.0893	幫助	195.0384	調查	114.1598
神奇	157.9359	弱勢	194.5427	逮捕	110.8561
地球	151.7349	捐	186.9286	報警	106.8442

表 9 難過、實用、高興三類新聞心情分類 LLR 前十名

難過		實用		高興	
詞彙	LLR 值	詞彙	LLR 值	詞彙	LLR 值
不幸	300.1170	醫師	974.2429	投手	546.5510
死亡	282.6576	患者	771.5972	先發	463.3743
不治	256.4122	症狀	747.3640	勝	383.3057
家屬	239.0486	飲食	727.2914	拿下	382.5888
身亡	235.8778	建議	702.2004	比賽	370.3191
自殺	229.5752	容易	628.6636	球員	370.1879
難過	193.8598	治療	598.0166	安打	316.2670
送醫	185.0590	疾病	594.8826	球	308.3993
急救	184.6250	預防	531.6418	擊敗	307.6599
死者	178.7346	避免	520.5768	冠軍	298.8540

表 10 無聊、生氣二類新聞心情分類 LLR 前十名

無聊		生氣	
詞彙	LLR 值	詞彙	LLR 值
民進黨	401.7913	立委	389.7735
前總統	346.4427	質疑	281.9694
陳水扁	346.0811	立法院	253.6003
蔡英文	274.0449	政府	242.6274
緋聞	266.3180	國民黨	217.6709
馬英九	253.3650	要求	189.2973
總統	214.8017	公務員	167.5531
主席	212.0275	扁家	153.2738
看守所	208.1930	爭議	152.9293
馬	206.7946	痛批	152.0251

3.4 SVM 分類處理

本節將詳細介紹如何將前置處理後的資料，轉換成適合 SVM 處理的向量，放入 LibSVM 分類器再進一步分類出結果茲將 SVM 向量、利用 LibSVM 分類器的步驟分述如下(如圖 17 所示)。

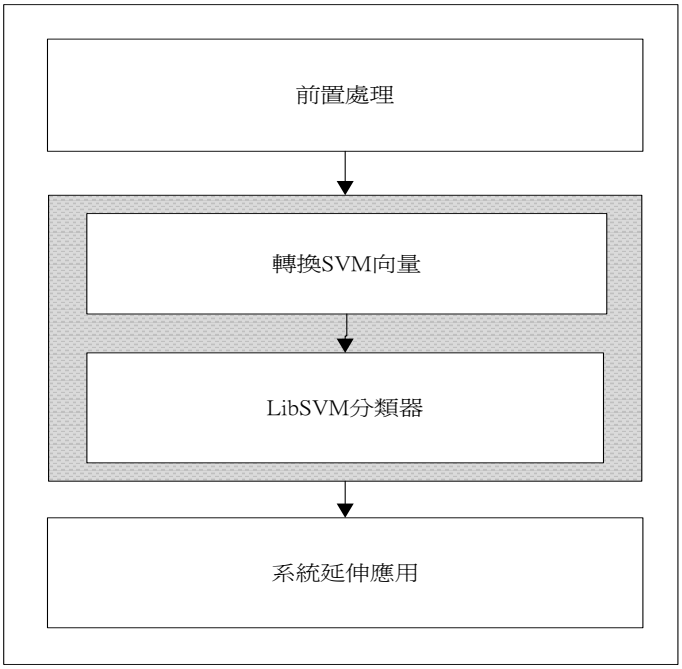

圖 17 系統架構-SVM 分類處理

3.4.1 轉換 SVM 向量

本研究將前置處理過後的詞，進一步地以二元值(Binary)方式轉換成 LibSVM 可以接受的檔案格式，而二元值所呈現的方式為依詞編號將有出現的詞標記為 1，沒出現的詞標記為 0，將此二元值依單純的存在與否概念轉換成 SVM 向量處理，實際例子如下。圖 18 為經前置處理後的新聞文章標題範例。

台鐵(N)　端午(N)　車票(N)　熱賣(Vi)　高鐵(N)　宣布(Vt)　開加(N)　列車(N)

圖 18 經前述處理後的詞彙列

而依照上述詞轉換成詞的編號與二元值，依詞編號升冪排列，如圖 19 所示：

10005:1　19982:1　21121:1　39445:1　40124:1　46219:1　51224:1　60231:1

圖 19 LibSVM 向量格式

由於 LibSVM 會將所有沒輸入的詞彙編號預設成 0，所以只需要列出此筆資料有出現的詞編號即可。

3.4.2 LibSVM 分類

在嘗試 SVM 最佳參數時參考[13]一文的方法，使用 SVM　Linear Kernel 代入(2^{-4}、2^{-3}、2^{-2}、2^{-1}、2^{1}、2^{2}、2^{3}、2^{4})作為 Cost 與 SVM 的三個 Kernel(linear、radial、sigmoid)交叉測試，其正確率普遍不佳(正確率平均約 60%)。因此本實驗收集自 2008 年 10 月 15 日自 2009 年 2 月 28 日的資料共四千八百篇新聞文章，每個心情分類各六百篇的資料集，採用 9-fold cross validation 進行參數測試，其原因是當資料集愈來愈大時，每次測試參數的時間將會愈來愈長，無法在短期間內找出最佳參數，先以較小的資料集找出較優秀的參數供往後實驗所使用，所得結果如表 11 所示，可以看出在核心函數為 radial 與 sigmoid、懲罰參數為 500 時有較佳的表現，下一章將只討論這兩個核心函數的正確率。

表 11 SVM 參數設定與結果

資料集	懲罰參數 C	核心函數	正確率
所有詞彙	100	linear	87.6566%
所有詞彙	100	polynomial	12.8232%
所有詞彙	100	radial	87.8432%
所有詞彙	100	sigmoid	86.5103%
所有詞彙	250	linear	87.6566%
所有詞彙	250	polynomial	12.8232%
所有詞彙	250	radial	88.1898%
所有詞彙	250	sigmoid	87.8966%
所有詞彙	400	linear	87.6566%
所有詞彙	400	polynomial	12.8232%

資料集	懲罰參數 C	核心函數	正確率
所有詞彙	400	radial	88.1632%
所有詞彙	400	sigmoid	88.1898%
所有詞彙	500	linear	87.6566%
所有詞彙	500	polynomial	12.8232%
所有詞彙	500	radial	88.1632%
所有詞彙	500	sigmoid	88.2165%
所有詞彙	750	linear	87.6566%
所有詞彙	750	polynomial	12.8232%
所有詞彙	750	radial	87.8699%
所有詞彙	750	sigmoid	88.1365%
所有詞彙	1000	linear	87.6566%
所有詞彙	1000	polynomial	12.8232%
所有詞彙	1000	radial	87.7899%
所有詞彙	1000	sigmoid	88.0832%
所有詞彙	5000	linear	87.6566%
所有詞彙	5000	polynomial	21.6206%
所有詞彙	5000	radial	87.55%
所有詞彙	5000	sigmoid	87.63%
所有詞彙	10000	linear	87.6566%
所有詞彙	10000	polynomial	29.3788%
所有詞彙	10000	radial	87.55%
所有詞彙	10000	sigmoid	87.6566%

此外，linear kernel 雖然並不是在所有 cost 參數配合下都有最佳表現，但變動起伏較小，表現雖然不是最好但也與最佳表現差距不大，而在表 11 中表現最佳的

是當懲罰參數設定為 500 與核心函數選擇 sigmoid 時的 88.2165%。

3.5 關鍵詞彙挑選系統

本研究將前面所提的特徵挑選與 LibSVM 分類器進一步應用，構思出關鍵詞彙挑選系統。

此系統考慮了前面所提的 LLR 與 TF 方法[33]，考量到有不少詞彙僅在某一篇新聞文章僅出現一次，而此篇新聞心情比例卻高達八九成，造成部份詞彙新聞心情分數偏高但出現次數過少產生誤差。但由於這兩種統計方法只能依各詞彙的出現分佈情況計算出 LLR 值與 TF 值後進一步評估，依此兩值的權重去挑選出各類別較具代表性的詞彙，忽略了人類可能對於不同詞彙中有不同的情感差異。本系統額外考慮奇摩新聞心情投票人數比例，希望能進一步在心情的呈現上表現出強度強弱差異。例如兩個人名的詞彙，「陳水扁」與「王永慶」，雖然說都是同一類型的詞彙，而在本系統所使用的資料集出現頻率也不低，但兩個詞彙造成讀者的心情分佈狀況截然不同。本系統的流程如圖 20。

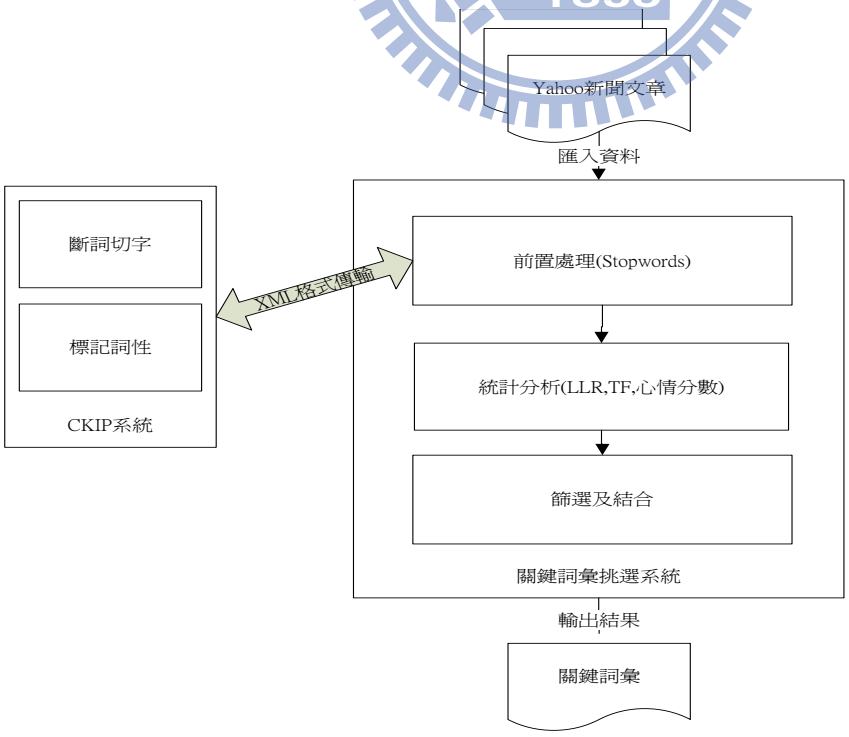

圖 20 系統架構圖

考量到有不少詞彙僅在某一篇新聞文章出現一次，而此篇新聞心情比例卻高達八九成，造成部份詞彙新聞心情分數偏高但出現次數過少產生誤差，統計分析的處理流程如圖 21。

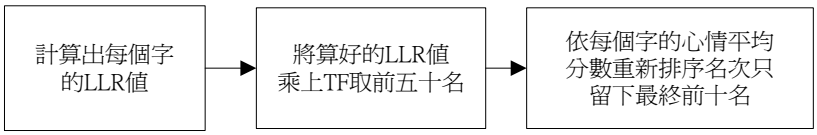

圖 21 統計分析方式

而其中在心情 e 中詞彙 i 平均分數計算方式為:

$$S_{e,詞彙i} = \frac{1}{F_{e,詞彙i}} \times \sum_{n} (S_{e,詞彙i,n} \times F_{e,詞彙i,n})$$

其中 $F_{e,詞彙i}$ 指的是在心情 e 中詞彙 i 所有新聞裡出現的總次數;n 為新聞編號; $S_{e,詞彙i,n}$ 在心情 e 中詞彙 u 於編號 n 的新聞中的心情分數，此值為 3.1 節所提的八項心情投票比例; $F_{e,詞彙i,n}$ 在心情 e 中詞彙 i 於編號 n 的新聞中出現的總次數。

3.5.1 系統介面

系統在中間呈現每個月的關鍵詞彙，如圖 22 所示，並以 3.5 節所提之計算方式找出最具重要性的詞彙，以其為中心，右上方為其相關的新聞標題，下方為時間捲軸，本系統採用資料自 2008 年 10 月至 2009 年 4 月，將此系統依月份與心情分類為區間來呈現：

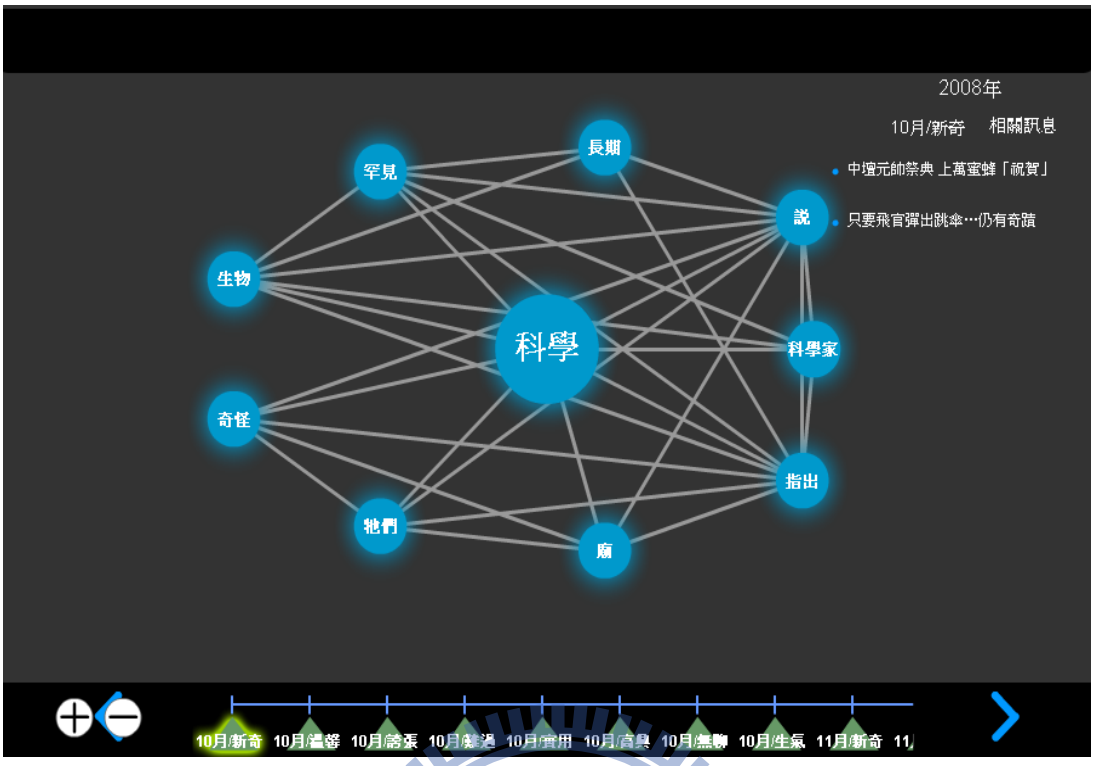

圖 22 系統介面圖

在詞彙間以圖 22 的計算方式計算出最高分數的詞彙為中心，而球與球之間的連線為是否有共同出現在同一篇新聞，如圖 23 所示。

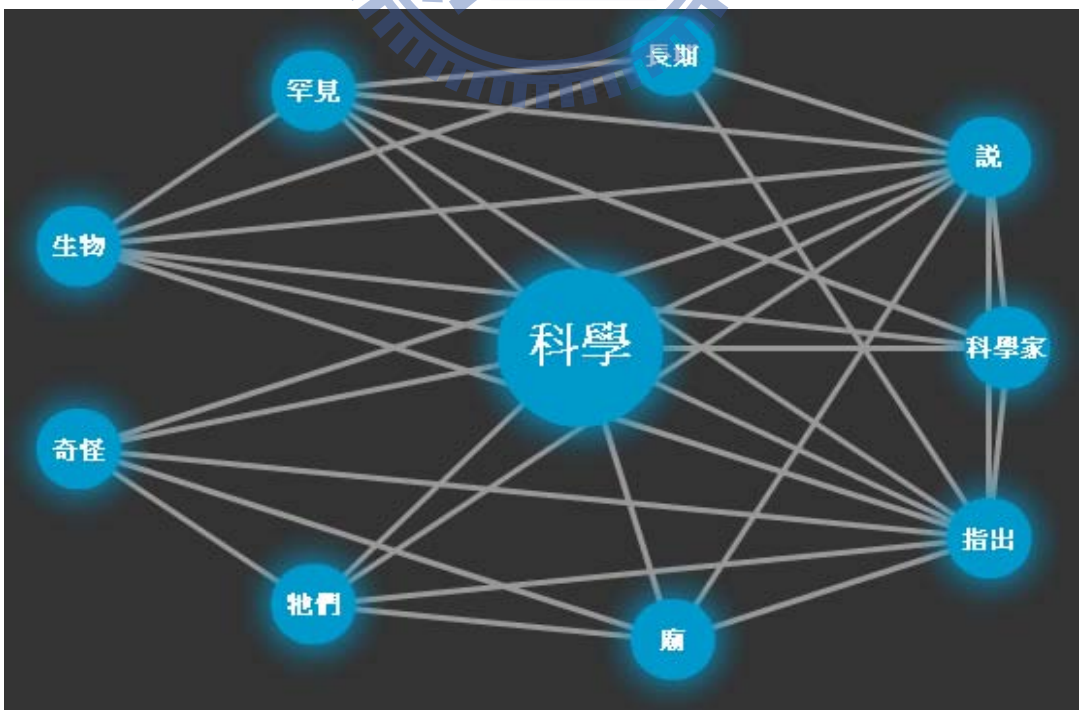

圖 23 詞彙關係圖

在網頁底下可以選擇使用者感興趣的月份與心情，選項為本系統所採用的時間 2008 年 10 月、11 月、12 月，2009 年 1 月、2 月、3 月、4 月與新奇、溫馨、誇張、難過、實用、高興、無聊、生氣等八類心情。如圖 24 所示。

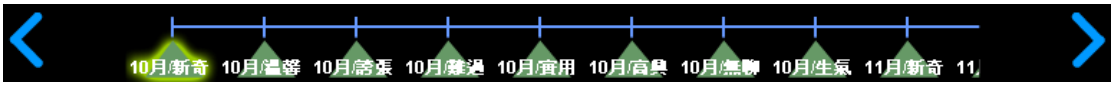

圖 24 詞彙心情月份選擇工具列

而右半部將與詞彙曾出現過的新聞標題列出，能更清楚了解該詞彙曾出現在哪些類型的新聞中，如圖 25 所示。

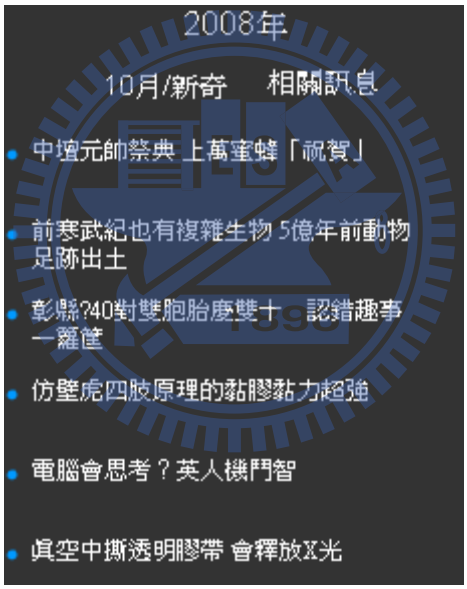

圖 25 新聞標題列表

選擇有興趣的標題後進一步呈現該篇新聞完整內容，如圖 26 所示

中壇元帥祭典 上萬蜜蜂「祝賀」

〔記者郭靜慧?潮州報導〕縣級古蹟泗林里朝林宮日前舉辦中壇元帥祭典，突然飛來上萬隻蜜蜂在廟內飛舞聚集，停留一天一夜直到祭典結束後，泗林里長李正順祝禱「請」蜂群離開，蜜蜂才集體飛出廟外，令地方稱奇。 傾巢而出 可能煙薰所致「嘸曾看過這麼多的蜜蜂」，朝林宮廟祝陳水勝指出，大約4、5年前有蜜蜂飛進廟裡，在中壇元帥神像上方屋頂築巢，每年農曆9月9日中壇元帥祭典都可以看見蜜蜂飛舞，但沒有像這次數量這麼多；泗林社區總幹事李晉榮認爲，從科學角度來看，可能祭典當天的進香人潮香火太旺，蜜蜂受不了煙薰傾巢而出。 祭典從農曆9月8日展開，之前就有民眾在廟內看到些許蜜蜂，李正順說，當天下午信眾到村落週邊祭拜五營兵馬後，蜜蜂不知道從哪起冒出來，一下子整個廟內都是蜜蜂，有些分據在幾個角落，入夜後全部聚集在地板上，黑壓壓一片，大夥都說是五營兵馬來祝壽了。 雖然蜜蜂沒有螫人，但數量太多，信眾看了怕怕不敢進廟上香，隔天一早李正順虔敬向蜂群說，祭典已經結束，煩請牠們回歸原位，說也奇怪，不一會兒蜂群即飛起，在廟內盤旋後飛出廟外，再繞行燒金紙的金爐一圈飛走，好像在向信眾道別般。

繼續其他查詢

圖 26 新聞內容

四、實驗與分析

本章介紹實驗所使用的資料集特色，以及在不同的資料集與特徵集下 LibSVM 的分類結果，了解改變特徵集種類與數量是否能提升 LibSVM 分類正確率。本章各節內容如下：4.1 節介紹實驗環境、資料特色、實驗步驟以及評估系統效能所使用的方法；4.2 節說明三種特徵的實驗結果；4.3 節詳細分析 Cross Validation 的結果；4.4 節介紹實驗產生的分類模型用於預測新資料的結果；4.5 節則是檢視 LLR 值與結合新聞心情分數之後選取的特徵對系統效能的影響；4.6 節討論關鍵詞彙挑選系統的結果。

4.1 實驗環境、資料、步驟與評估方法

4.1.1 實驗環境

本實驗在計算 Log Likelihood Ratio 值、應用系統、SVM 訓練與預測、新聞資料搜集等動作使用電腦來輔助，詳細資料如表 12。

表 12 實驗環境

工作	環境說明
Log Likelihood Ratio 值計算、應用系統	OS:Microsoft Windows XP Professional Version 2002 Service Pack3 Hardware：Intel(R) Core(TM)2 CPU 6300@1.86GHz 512ram
SVM 資料建立、訓練模型、新聞資料庫	OS： Microsoft Windows XP Professional Version 2002 Service Pack3 Hardware:Intel(R) Pentium(R) 4 CPU 3.0GHz 2.99GHz 512+256 ram

SVM cross validation	OS:Microsoft Windows XP Professional Version 2002 Service Pack2 Hardware： Intel(R) Core(TM) 2 CPU T5600 @ 1.83GHz 1.81GHz 3.0GB ram

4.1.2 實驗資料

本研究之實驗資料使用 2008/10/15~2009/4/29 的 Yahoo!奇摩新聞的部分資料，如 3.1 節所述。心情分類是由讀者在閱讀一篇文章後主動對於該篇新聞給予符合自己心情的標籤，經由 Yahoo!奇摩新聞系統統計之後，以最高比例的心情分類訂為該篇新聞的心情分類，心情標籤的投票結果會不斷更新，讓讀者能在 Yahoo!奇摩網站上取得最新的統計結果。本研究每天以各心情分類收錄五篇投票超過 50 人的新聞為原則，表 13、表 14、表 15 為資料集的統計資料。表中詞彙總數為不重複採計的詞彙數量(不除去標點符號與停用字)。而本研究所使用的資料中新聞分類與心情分類的關係如圖 27。

表 13 月份與新聞篇數的統計

年/月份	新聞篇數	詞彙總數
2008/10	583	17286
2008/11	1134	28536
2008/12	1295	32822
2009/1	1165	30833
2009/2	1012	28346
2009/3	1217	32085
2009/4	1209	32625
總計	7615	76314

表 14 心情分類與新聞篇數統計

心情分類	新聞篇數	詞彙總數
新奇	944	25446
溫馨	942	24992
誇張	944	23129
難過	944	21963
實用	944	23676
高興	942	22542
無聊	943	20501
生氣	942	22904

表 15 新聞分類與新聞篇數的統計

新聞分類	新聞篇數	詞彙總數
政治	483	14186
社會	772	19471
地方	891	22975
國際	575	16956
財經	272	11761
科技	438	15289
運動	448	13112
健康	955	22652
教育	446	16094
藝文	183	10313
影劇	1007	22369
旅遊	101	5900
生活	970	22564

圖 27 新聞分類與心情分類關係

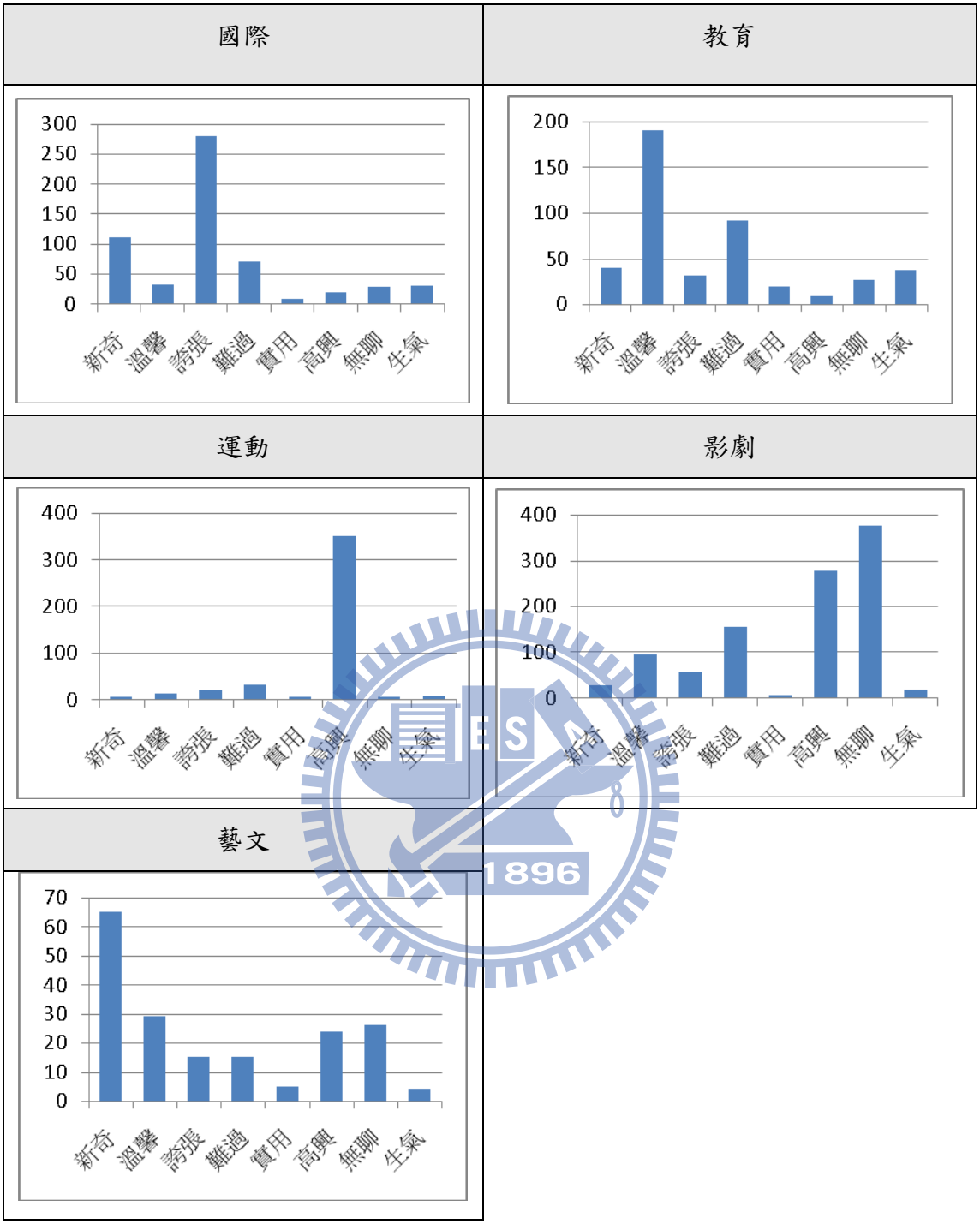

圖 27 新聞分類與心情分類關係(續)

4.1.3 實驗步驟

本研究將 2008 年 10 月中旬至 2009 年 4 月底所搜集之資料分成兩部分來進行實驗：

1. 將資料分成九份，進行 9-fold cross validation，重複執行九次，每次以八份資料做為訓練集，用以建立訓練模型；餘下一份做為測試集，用以驗証分類模型的效能，最後將九次執行的效能平均。實驗的目的是為了驗証不同特徵挑選方式(以所有詞彙、動詞與名詞、動詞與名詞中高 Log Likelihood Ratio 值者)對 SVM 效能的影響。
2. 以前述資料做為訓練模型，將 2009 年 5 月之新聞資料做為測試資料。

4.1.4 評估方法

本論文採用 Accuracy 以及 F-measure。計算方式源自資訊擷取理論，以錯差矩陣(Confusion matrix)四個準則(如表 16)來評估分類效果。

計算前必須定義出何為 relevant 以及 irrelevant，本實驗將系統自動分類的類別與實際新聞資料心情分類相符為 relevant;不符則為 irrelevant。在 relevant 與 irrelevant 定義完後可以做出具有四元素的錯差矩陣。

表 16 錯差矩陣

		實際的類別	
自動分類的類別		p	n
	Y	True Positive	False Positive
	N	False Negative	True Negative
Column Totals		P	N

$$precision = \frac{TP}{TP + FP}，recall = \frac{TP}{P}$$

$$accuracy = \frac{TP + TN}{P + N}$$

$$F - measure = \frac{2}{\frac{1}{precision} + \frac{1}{recall}}$$

4.2 三種特徵的實驗結果

本節分別說明以全部詞彙、動詞與名詞、LLR 值高的名詞與動詞等三種特徵的分類結果。由於當資料量愈大會使 LibSVM 的執行時間愈長，無法在短期間內測出最佳的參數，而在第三章的表 11 中可以看出表現普遍較好的為採用核心函數 radial 與 sigmoid 以及懲罰參數設為 500，故後面實驗將僅使用 radial 與 sigmoid 兩種核心參數，懲罰參數設定 500 並額外加入 100、1000 兩種用來對照。

1. **全部詞彙：**

在 Yahoo!奇摩新聞首頁所擷取下來的新聞直接透過 CKIP 處理，不去除掉任何停用字與標點符號的情況下將所有內容轉換成對應的詞彙編號，以最原始的新聞文章面貌放入 LibSVM 進行分類處理，結果如表 17(詞彙數量為不重複採計的詞彙總數)：

表 17 使用全部特徵的分類結果

詞彙數量	懲罰參數 C	核心函數	正確率
76314	100	radial	72.32%
76314	100	sigmoid	75.68%
76314	500	radial	73.05%
76314	500	sigmoid	76.77%
76314	1000	radial	71.24%
76314	1000	sigmoid	73.22%

2. **動詞與名詞：**

參考[33]一文中的方法，在圖書資訊分類過程中僅考慮動詞與名詞，可以達到提升準確率的目標，在奇摩心情新聞首頁所擷取下來的新聞透過 CKIP 處理，僅使用 CKIP 所傳回來資料中標記為 N、Nv、Vi、Vt 詞性的詞彙，並除去停用字，將這些內容轉換成對應的詞彙編號放入 LibSVM 處理，所得的結果如表 18：

表 18 使用動詞與名詞的分類結果

詞彙數量	懲罰參數 C	核心函數	正確率
68964	100	radial	71.79%
68964	100	sigmoid	68.31%
68964	500	radial	76.92%
68964	500	sigmoid	79.67%
68964	1000	radial	77.36%
68964	1000	sigmoid	78.25%

3. LLR 值高的動詞與名詞：

此實驗將說明在使用不同特徵數量時，是否會改變到分類預測正確率，當使用過多的特徵數量時將會導致系統處理速度緩慢，效能低落，而降低特徵數量又可能會造成分類預測正確率受到影響，在本實驗所使用的新聞資料集下，使用多少特徵數才能對系統效能與分類效率都達到最佳表現，在表 19 可以看出特徵數量與正確率的關係，核心函數僅選擇前面實驗結果普遍較佳的 sigmoid 與 radial 兩種：

表 19 SVM 參數設定與 LLR 特徵數量結果

資料集	懲罰參數 C	核心函數	正確率
LLR 前 800	100	radial	72.32%
LLR 前 800	100	sigmoid	73.35%
LLR 前 800	500	radial	73.05%
LLR 前 800	500	sigmoid	73.63%
LLR 前 800	1000	radial	71.81%
LLR 前 800	1000	sigmoid	73.22%
LLR 前 1000	100	radial	74.21%

資料集	懲罰參數 C	核心函數	正確率
LLR 前 1000	100	sigmoid	73.58%
LLR 前 1000	500	radial	72.24%
LLR 前 1000	500	sigmoid	75.13%
LLR 前 1000	1000	radial	75.68%
LLR 前 1000	1000	sigmoid	75.45%
LLR 前 1500	100	radial	76.27%
LLR 前 1500	100	sigmoid	73.79%
LLR 前 1500	500	radial	78.31%
LLR 前 1500	500	sigmoid	77.94%
LLR 前 1500	1000	radial	78.29%
LLR 前 1500	1000	sigmoid	77.82%
LLR 前 2500	100	radial	75.26%
LLR 前 2500	100	sigmoid	79.16%
LLR 前 2500	500	radial	82.34%
LLR 前 2500	500	sigmoid	83.77%
LLR 前 2500	1000	radial	84.91%
LLR 前 2500	1000	sigmoid	82.39%
LLR 前 3000	100	radial	83.69%
LLR 前 3000	100	sigmoid	82.81%
LLR 前 3000	500	radial	85.69%
LLR 前 3000	500	sigmoid	84.91%
LLR 前 3000	1000	radial	85.19%
LLR 前 3000	1000	sigmoid	84.81%

表 19 中 LLR 前 800 為八個心情類別(新奇、溫馨、誇張、難過、實用、高興、無聊、生氣)裡分別取出 Log Likelihood Ratio 值最高的前八百名;LLR 前 1000 為八個心情類別裡分別取出 Log Likelihood Ratio 值最高的前一千名,依此類推。而此五種資料集的不重複詞彙數量如表 20 所示。

表 20 資料集的不重複詞彙數量

資料集	不重複詞彙數量
LLR 前 800	4589
LLR 前 1000	5606
LLR 前 1500	7967
LLR 前 2500	12552
LLR 前 3000	15117

在表 19 中可以明顯看出選擇的特徵值越多,正確率也越高,但在取前 3000 名的 Log Likelihood Ratio 值時接近極限,其原因是在第三千名左右的詞彙 Log Likelihood Ratio 值已經接近於 0,再往後取下去的值(小於或等於零)對於分類已幫助不大,如表 21 所示。

表 21 LLR 前三千名詞彙最低採用值

心情類型	Log Likelihood Ratio 值
新奇	3.67202506889589
溫馨	4.39227412338369
誇張	3.54172265983652
難過	3.46731747558806
實用	6.70959477871656
高興	4.31943313125521
無聊	3.76446460781153
生氣	4.65359306475148

4.3 分析 Cross Validation 內容

本節將詳細分析將資料切成九份做 Cross Validation，在九份資料中是否有特定資料預測正確率偏低，更進一步剖析其可能造成正確率降低的原因，只以前面實驗正確率最高的，本次實驗採用 7615 篇新聞，將 846 篇分成一份，共九份資料，依新聞輸入先後順序切割並依序編號，再以一份為預測資料，八份為訓練資料的方式透過 LibSVM 以前面實驗最佳的 Log likelihood Ratio 值前三千名特徵詞彙、懲罰參數 C 設定為 500、核心函數選擇 radial 詳細內容如表 22。

表 22 Cross Validation 分析結果

心情類別	新奇	溫馨	誇張	難過	實用	高興	無聊	生氣
正確率	62%	71%	77%	74%	86%	75%	83%	83%
F-measure	56%	68%	72%	70%	80%	71%	78%	76%

由此份資料平均投票人數與預測錯誤的資料屬於哪一個心情類別，如表 22 與表 23，可以看出在預測新奇的錯誤篇數最高，實用的錯誤篇數最低：

表 23 預測錯誤的心情分類篇數

心情分類	新奇	溫馨	誇張	難過	實用	高興	無聊	生氣
錯誤篇數	362	276	219	247	133	238	161	162
總篇數	953	952	952	952	951	952	951	952

4.4 預測新資料

前面所做的資料分析都是以所有詞彙在前置處理前就已預先放入資料庫給予其編號，預先計算出全部詞彙的 LLR 值並取出前幾名作為特徵，但在新聞類型的資料中可以發現每個月所常出現的重要詞彙變動差異很大，如前面第三章所做的各月重要詞彙表。例如在去年年底因為大陸三鹿毒奶粉事件而出現的新詞彙「三聚氫氨」，在近幾個月常出現的「H1N1」，這些詞彙都是因為在某時期中發生的偶發事件而突然出現在新聞報導，也說明了新聞類型的資料很容易出現以往不曾出現的詞彙，無任何的歷史資料可供參考，造成分類預測上的困難，而表 24 也可以清楚看出五月份的新資料預測正確率比舊資料差。

表 24 不同模型預測新資料的結果

模型	懲罰參數 C	核心函數	舊資料正確率	新資料正確率
所有詞彙	500	sigmoid	76.77%	28.73%
動、名詞	500	sigmoid	79.67%	30.19%
LLR 前 1000	500	sigmoid	75.13%	42.34%
LLR 前 3000	500	sigmoid	84.91%	45.12%

4.5 結合 LLR 值與心情分數

在 3.1 節提過每篇新聞皆有一個心情分數的數據，依八種心情分類各有一個以百分比表示的分數，而此數據的計算方式為所有投票人數中投給此項心情分類的百分比。而本實驗將原先計算好的 Log Likelihood ratio 值與各詞彙的百分比分數總平均相乘會造成名次上不同於先前實驗 LLR 值的變化，並將之放入 SVM 進行分類看是否能提升系統在處理新資料的效能，由表 25 的結果可以看出正確率略有提升。而其中詞彙 i 在心情 e 中的新分數為：

$$LLR_{心情e,詞彙i} \times S_{心情e,詞彙i}$$

表 25 LLR 值與心情分數結合的預測結果

模型	懲罰參數 C	核心函數	正確率
結合前 LLR 前 1000	500	sigmoid	42.34%
結合前 LLR 前 3000	500	sigmoid	45.12%
結合後 LLRxS 前 1000	500	sigmoid	48.67%
結合後 LLRxS 前 3000	500	sigmoid	50.03%

4.6 關鍵詞彙挑選系統實驗結果

本節將討論 3.4 節所設計的關鍵詞彙挑選系統及其結果，將以七個月份綜合比較的方式呈現各個月份前十五名重要詞彙變化，如表 26、表 27、表 28、表 29、表 30、表 31、表 32、表 33 所示：

表 26「新奇」七個月份重要詞彙比較

十月	十一月	十二月	一月	二月	三月	四月
指出	指出	隕石	信徒	信眾	指出	信徒
廟	廟	信徒	內	廟	可能	信眾
說	分	內	廟	結束	像	才
牠們	認為	廟	牠們	科學	下午	分
奇怪	牠們	科學	奇特	奇特	科學	像
科學	科學	神像	地球	生物	隻	整
生物	上香	屋頂	科學家	科學家	法新社	可以
科學家	地球	地球	細胞	爆炸	地球	大約
罕見	生物	生物	喜歡	罕見	生物	隻
長期	科學家	科學家	牠	東京	科學家	地球
期刊	日本	細胞	案發	可愛	罕見	生物
牠	期刊	古老	動物	牠	日本	科學家
客人	海洋	日本	主人	掛	牠	赫然
動物	動物	動物	解剖	主人	海洋	長期
白	主人	猴子	提醒	張沛元	綜合	日本

表 27「溫馨」七個月份重要詞彙比較

十月	十一月	十二月	一月	二月	三月	四月
醫院	醫院	醫院	醫院	醫院	醫院	醫院
悲傷	開心	經濟	經濟	不久	不久	罹患
一輩子	好友	住院	罹患	罹患	罹患	老人
兒女	關懷	關懷	開心	南投	南投	關懷
得到	生命	探望	白米	關懷	關懷	生命
老人	小朋友	生命	老人	生命	生命	長大
辭世	長大	感謝	繞	進	進	感謝
關懷	感謝	認養	生命	年度	年度	感恩
父	認養	捐出	機車	累	累	傷心
南投縣	物價	登記	熱心	打破	打破	退休
小吃店	表達	表達	小朋友	春聯	春聯	員警
生命	接受	回饋	感謝	長大	長大	付出
接下	傷心	領	感恩	感謝	感謝	畢業
育有	退休	幫助	捐出	捐出	捐出	容易
擔任	植入	容易	回饋	桃園	桃園	感激

表 28「誇張」七個月份重要詞彙比較

十月	十一月	十二月	一月	二月	三月	四月
長	毒蛇	長	清點	早上	早上	長
監控	中午	撞上	高院	警方	警方	竟然
軍方	手槍	寬	警方	警察	現場	輛
當地	警方	警方	逃逸	現場	叫	處
警方	消防隊員	隨後	驚訝	叫	研判	穿
手機	槍	空中	落網	載	殺害	當地
提告	反映	技術	附近	研判	查獲	警方
空中	大樓	油門	研判	登上	行為	馬桶
消防隊員	專線	登上	嚇	殺害	花錢	挖
市	房間	外海	報案	查獲	口袋	欠
附近	離譜	火	罪嫌	行徑	攝影	下車
上下	通知	被害人	轎車	行為	編譯	來賓
礙於	查看	嚇	英國	花錢	派出所	槍
2樓	男童	濃煙	車子	被害人	市價	乘客
按鈕	賠償	名字	屋	口袋	要價	叫

表 29「難過」七個月份重要詞彙比較

十月	十一月	十二月	一月	二月	三月	四月
傍晚	正常	大礙	醫生	醫生	人生	正常
正常	拍	人生	孩子	人生	台幣	費用
香港	氣	正常	家	機會	進食	請假
臨近	嚇壞	香港	家屬	台幣	孩子	孩子
新婚	翻覆	景象	自殺	進食	媽媽	媽媽
聽到	孩子	孩子	身亡	醫療	自殺	家
孩子	媽媽	媽媽	送	孩子	表現	家屬
家	家	家	出席	媽媽	導演	身亡
家屬	家屬	家屬	起火	家屬	懂事	通告
自殺	自殺	自殺	演藝圈	自殺	送醫	表現
身亡	身亡	身亡	男星	表現	後方	出席
表現	表現	通告	送醫	導演	不治	看起來
出席	損失	表現	痕跡	演藝圈	死者	母親
骨折	生前	出席	不治	懂事	聖嚴	送醫
母親	不治	母親	談	送醫	低收入戶	不治

表 30「實用」七個月份重要詞彙比較

十月	十一月	十二月	一月	二月	三月	四月
治療	發生	治療	發生	發生	飲食	發生
造成	治療	意外	治療	治療	降低	治療
檢查	造成	造成	意外	相關	常見	意外
美國	檢查	檢查	相關	飲食	疾病	造成
現象	飲食	飲食	造成	美國	產生	嚴重
緊急	發表	降低	檢查	降低	增加	檢查
降低	降低	含	飲食	常見	患者	飲食
常見	含	常見	作者	疾病	肌肉	緊急
疾病	常見	疾病	降低	產生	醫生	降低
產生	疾病	產生	常見	增加	改善	疾病
增加	增加	疼痛	疾病	患者	刺激	產生
習慣	習慣	增加	增加	肌肉	症狀	增加
患者	患者	患者	習慣	減少	面	習慣
減少	肌肉	減少	患者	醫生	急救	患者
水果	減少	肥胖	減少	改善	食物	減少

表 31「高興」七個月份重要詞彙比較

十月	十一月	十二月	一月	二月	三月	四月
家人	成為	展出	家人	所長	偵辦	開球
笑	家人	家人	一樣	娘家	投手	家人
停	今年	東森	透漏	家人	球隊	一樣
唱	亞洲	笑	偵辦	一樣	球員	車
兒子	傳喚	出門	歌迷	偵辦	大賽	歌迷
下樓	扁	車	扁	扁	出色	沒想到
異常	出來	歌迷	唱	投手	聯盟	位於
費城人	態度	亞洲	董事長	球隊	季	扁
季後賽	董事長	扁	兒子	球員	勝	投手
出賽	投手	兒子	異常	大賽	冠軍	打者
國聯	球隊	投手	投手	出色	拿下	球隊
大聯盟	球員	出賽	球隊	聯盟	展現	球員
球員	球團	球隊	大聯盟	季	觀眾	穿著
球團	大賽	球員	球員	勝	敗	安打
檢查官	國務	球團	洞	冠軍	對手	夫人

表 32「無聊」七個月份重要詞彙比較

十月	十一月	十二月	一月	二月	三月	四月
女友	難過	專輯	張碩文	新人	新人	難過
大方	緋聞	難過	緋聞	難過	獻金	徵
懷疑	機要費	戲	大方	戲	追查	緋聞
今天	懷疑	手法	追查	獻金	問	大方
追查	民進黨	傻眼	民進黨	黑色	剛好	利菁
民進黨	嗆	緋聞	離去	追查	感情	今天
還好	星光	台視	出書	民進黨	父親	追查
海角	買單	懷疑	帶回	問	專訪	民進黨
露出	雲林	音樂	比賽	剛好	黨	來不及
隆起	遊行法	民進黨	新書	法鼓山	狗狗	相處
比賽	比賽	海角	陳前總統	帶回	站立	比賽
女星	圍城	攤位	攝	芝加哥	領先	新書
覺得	歐	鄭文龍	湯	比賽	氣勢	職棒
長達	新社	雲林	幫忙	幫忙	不滿	幫忙
想	措施	比賽	感情	感情	王牌	歲

表 33「生氣」七個月份重要詞彙比較

十月	十一月	十二月	一月	二月	三月	四月
5 月	漲	看守所	看守所	看守所	案	看守所
高	麵粉	案	案	案	買	棟
國小	馬	巡迴	漲	漲	監察院	違法
學校	國小	績效	監察院	買	馬	民進黨團
大學	學校	馬	民進黨團	金控	批評	馬
吸引	讀	下台	批評	監察院	搖頭	批評
特別	持	觀感	國小	中選會	國民黨	權益
強調	大學	國小	強調	馬	研究	給
政府	強調	學校	政府	批評	國中	國小
立委	政府	進行	立委	搖頭	成績	生
要求	立委	台	要求	權益	政策	政府
立法院	要求	強調	立法院	高	扁案	立委
開放	立法院	政府	國民黨	生	李慶安	要求
同意	研究	立委	研究	吸引	總統	立法院
研究	教育	要求	說明	立委	總統府	國民黨

將以上八項心情分類表(表 26、表 27、表 28、表 29、表 30、表 31、表 32、表 33)觀察後並與圖 27 的新聞類型比例比較，歸納下列幾種狀況：

1. 在新奇(表 26)詞彙表中可以看出大多是跟科技、環境有關的詞彙，而在新聞分類科技(圖 27)裡最容易令讀者產生新奇的心情；

2. 在溫馨(表 27)的詞彙表中，詞彙大多數與社會關懷用字有關，在新聞分類教育(圖 27)中最容易使讀者產生溫馨的心情，這跟本研究新聞所收錄的關懷弱勢家庭孩子教育問題，不謀而合；

3. 在誇張(表 28)的詞彙表中，詞彙大多數跟描述社會事件的用字相關，而 Yahoo!奇摩裡的新聞分類國際，經常收錄國外誇張的社會事件相關新聞，與圖 27 中新聞分類國際也最易使人感到誇張相符；

4. 在難過(表 29)的詞彙表中，詞彙用字不少都跟地方新聞遭受到不幸事件有關係，Yahoo!奇摩在地方新聞經常收錄台灣各地，弱勢家庭孩子教育與生活等新聞，與圖 27 中新聞分類地方心情比例中，難過位居第二的高比例情況相符；

5. 在實用(表 30)的詞彙表中，詞彙與醫療保健領域常用的字有關，每個月詞彙間差異不大，與圖 27 中新聞分類健康心情比例中實用分類佔了絕大多數情況相符；

6. 在高興(表 31)的詞彙表中，詞彙大多跟運動有關係，與圖 27 中新聞分類運動心情比例中高興分類佔了絕大多數情況相符；

7. 在無聊(表 32)的詞彙表中，詞彙跟政治、演藝圈兩類有關，與圖 27 中新聞分類政治、影劇心情比例中無聊分類佔了絕大多數情況相符；

8. 在生氣(表 33)的詞彙表中，詞彙跟政治、政策類型有關。與圖 27 中新聞分類政治、社會心情比例中生氣分類佔了絕大多數情況相符；

綜上所述，關鍵詞彙挑選系統可以挑出與該心情分類相關性高的關鍵詞彙。

第五章 結論與建議

5.1 結論

本論文將文件分類的概念應用於讀者閱讀新聞後的心情偵測，藉由特徵挑選方法建立了一個關鍵詞彙挑選系統。本研究首先將由 Yahoo!奇摩收錄的新聞文章轉換成詞彙的形式，計算出每個詞彙的 Log Likelihood Ratio 值，以 Log Likelihood Ratio 值挑選出重要的詞彙作為 SVM 分類的特徵集，放入 SVM 進行訓練與預測，提升分類正確率；另一方面將 Log Likelihood Ratio 值與新聞文章附帶的心情比例分數結合，找出每個月的關鍵詞彙。以下針對「結合特徵挑選與 LibSVM 分類法」與「關鍵詞彙挑選系統」兩方面進行說明：

1. 結合特徵挑選與 LibSVM 分類法：

   在兩次不同資料集大小的實驗中(2008 年 10 月至 2009 年 2 月、2008 年 10 月至 2009 年 4 月)，可以發現當資料集變大時正確率明顯下降，但藉由特徵挑選出重要的特徵值可以提升正確率，當使用特徵數量愈多時，正確率也隨之上升，假若使用過多不重要的特徵也會使正確率下降；另一方面，可以看出在「新奇」的分類預測正確率與 F-measure 值表現不佳，其可能原因是新聞的詞彙差異較大，容易出現新的詞彙，以特徵挑選的方式從舊資料中找出重要的特徵值無法明顯改善其正確率；「實用」的類別正確率與 F-measure 值表現最佳，由其挑選出的特徵內容可以看出與醫療保健相關的內容經常讓人會去選擇「實用」的心情選項，是可以藉由舊資料去找出此分類的重要特徵。

2. 關鍵詞彙挑選系統：

   本系統將每個月分的詞彙分開計算 Log Likelihood Ratio 值，並與各心情分類中每個詞彙所計算出的心情比例分數結合，可以找出每個月重要的關鍵字

在各心情分類裡的變化情況，提供使用者簡單明瞭地了解變化的趨勢。

5.2 後續研究建議

本研究將文件分類的技術應用於新聞文章心情偵測研究，期望能在讀者閱讀新聞前，可以透過分類預測就能達到初步過濾新聞的效果，避免看到自己不期望甚至厭惡看到的新聞，並可以透過關鍵詞彙挑選系統找出每天、每月、每年重要的詞彙。經過實驗結果的分析，本研究上仍有值得改善之處，茲闡述如下：

1. 分類正確率：

   在本研究中，明顯看出有少數分類因為新聞內容變化較大的關係，無法掌握新詞彙，故而造成在分類預測上正確率不佳。但本研究中所使用的技術僅限於資訊檢索領域，將來可以考慮由心理學領域方面尋找人類心情與詞彙的相關資料。

2. 詞彙挑選系統：

   本研究僅採用動、名詞兩種詞性來建構系統，但是在使用過程中發現，當月關鍵字是名詞時比動詞更容易想起相關新聞事件，而形容詞似乎跟人類情緒起伏也有關係，進而影響到人類的情緒，未來可以考慮僅考慮名詞或者形容詞來實作系統。

3. 預測其它類型資料：

   本研究是以新聞文章資料作為訓練資料，預測新加入的新聞資料效果普遍不佳，可能是新聞資料詞彙與新聞長度變動較大的關係，假若以此方法運用於部落格文章或是其他類型文章之心情偵測，效果跟預測新聞文章是否有明顯差異，亦是一可探究的課題。

參考資料

[1] A. Toffler, *Future Shock*, Random House, 1970.

[2] “網路成學子閱讀新寵 閱讀率達二成八 直逼報紙 側重娛樂影視休閒資訊”，*大學報*，政治大學新聞系, June 5, 2000. (Access Date: 2009/4/1)

[3] E. Katz, J. G. Blumler & M. Gurevitch , "On the Use of the Mass Media for Important Things,”in *American Sociological Review*, vol.38 , No.2, pp. 164-181, 1973.

[4] The Center for the Digital Future at USC’s Annenberg School of Communications, "2009 Digital Future Report," http://www.digitalcenter.org/ , 2009. (Access Date: 2009/5/1)

[5] 「Newsy打造每天５則「新聞懶人包」影片，給聰明人觀賞」, http://mr6.cc/?p=3040 , (Access Date: 2009/4/15)

[6] C.H. Yang, H.H. Chen, and H.Y. Lin, "Building Emotion Lexicon from Weblog Corpora," in *Proceedings of 45th Annual Meeting of Association for Computational Linguistics*, June 23rd-30th, pp. 133-136, 2007.

[7] G. Mishne, "Experiments with mood classification in blog posts," in *Style2005 – 1st Workshop on Stylistic Analysis of Text for Information Access, at SIGIR*, 2005.

[8] E. Hovy and S.M. Kim, "Determining the Sentiment of Opinions," in *Proceedings of the 20th international conference on Computational Linguistics*, pp. 1367–1373, 2004.

[9] M.L. Littman, P.D. Turney, "Unsupervised learning of semantic orientation from a hundred-billion-word corpus," *Technical Report*, National Research Council Canada, Jul 15, 2002.

[10] S.H. Myaeng, Y. Choi, Y. Jung, "Determining Mood for a Blog by Combining

Multiple Sources of Evidence,"in *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence table of contents*, pp.271-274, 2007.

[11] K.Balog, M.D. Rijke, "Decomposing Bloggers' Moods: Towards a Time Series Analysis of Moods in the Blogosphere, " in *the Third annual Workshop on the Weblogging Ecosystem*, 2006.

[12] G. Mishne, M. Rijke, "Capturing global mood levels using blog posts, " in *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs*, pp.145-152, 2006.

[13] C.H. Yang, H.H. Chen, K.H. Lin, "Emotion Classification of Online News Articles from the Reader's Perspective," in *Web Intelligence and Intelligent Agent Technology*, pp. 220-226, 2008.

[14] C.C. Chang, C.J. Lin, "LibSVM," - http://www.csie.ntu.edu.tw/~cjlin/libsvm/ , 2008.

[15] S.S. Shai, "passive-aggressive (PA) classifier," http://ttic.uchicago.edu/~shai/code/ , 2006. (Access Date: 2009/4/10)

[16] H. Cui, M. Datar, and V. Mittal, "Comparative Experiments on Sentiment Classification for Online Product Reviews, " in *Proc. of 21st Conference of the American Association for Artificial Intelligence, AAAI*, 2006.

[17] H.H. Chen ,and H.Y. Lin, "Ranking Reader Emotions Using Pairwise Loss Minimization and Emotional Distribution Regression,"in *Proceedings of EMNLP 2008: Conference on Empirical Methods in Natural Language Processing*, October 25-27,pp. 136-144, 2008.

[18] C.S. Khoo, J.C. Na, T.T. Thet , "Sentiment Classification of Movie Reviews Using Multiple Perspectives," in *Proceedings of the 11th International Conference on Asian Digital Libraries: Universal and Ubiquitous Access to Information* , pp. 184 - 193 , 2008.

[19] The University of Sheffield 2001-2009, "GATE's ANNIE System, " http://gate.ac.uk/ie/annie.html , 2009.

[20] A. Gelbukh, *Computational Linguistics and Intelligent Text Processing*, Springer, 2004.

[21] P.S. Gregory, "KDnuggets," http:// www.kdnuggets.com , 2007. (Access Date: 2009/5/15)

[22] 張云濤、龔玲，「資料探勘原理與技術」，五南出版社，2007 年。

[23] V. Vapnik, "Support vector machine, " http://en.wikipedia.org/wiki/Support_vector_machine , 1995. (Access Date: 2009/3/10)

[24] Joachims, Thorsten, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features, "in *Proceedings of the European Conference on Machine Learning (ECML)*, Springer, 1998.

[25] Haruno, Hirotoshi, Masahiko, Taira, "Feature Selection in SVM Text Categorization, " in *AAAI/IAAI 1999*, pp. 480-486, 1999.

[26] Kudo, Matsumoto, Taku, and Yuji, " Use of Support Vector Learning for Chunk Identification.," in *Proceedings of CoNLL-2000*, pp. 142-144, 2000.

[27] C.C. Chang, C.J. Lin, and C.W. Hsu, "A Practical Guide to Support Vector Classification," 2003.

[28] Keerthy, "Improvements to Platt's SMO algorithm for SVM classifier design," 1999.

[29] T. Joachims, "SVMlight," http://svmlight.joachims.org/ , 2008.

[30] H.P. Luhn, "A Statistical Approach to the Mechanized Encoding and Searching of Literary Information," in *IBM Journal of Research and Development, vol.11, no.4,* pp.309-307, Oct 1957.

[31] A.Singhal and G.Salton, "Automatic Text Browsing Using Vector

Space Model," in *Technical Report, Department of Computer Science*, Cornell University, pp. 145-151, 1993.

[32] K. Spark-Jones, "A statistical interpretation of term specificity and its application in retrieval, " in *Journal of Documentation, Vol. 28, No. 5*, pp. 111-121, 1972.

[33] 林昕潔，「以 SVM 與詮釋資料設計書籍分類系統」國立交通大學，碩士論文，2006 年。

[34] "Likelihood-ratio test, " http://en.wikipedia.org/wiki/Likelihood-ratio_test , 2009. (Access Date: 2009/3/18)

[35] 「Yahoo!奇摩新聞」， http://tw.news.yahoo.com/attitudelist.html，2008 年。

[36] 許世瑛，「中國文法講話」，中華開明書局，1984 年。

[37] 中央研究院，「CKIP中文詞知識庫小組」http://godel.iis.sinica.edu.tw/CKIP/ ，2009 年。

[38] 中央研究院，*中央研究院平衡語料庫詞集及詞頻統計*，http://www.aclclp.org.tw/doc/wlawf_abstract.pdf ，2008 年。

[39] "Oracle Text Reference," http://download.oracle.com/docs/cd/B19306_01/text.102/b14218/astopsup.htm#sthref2545, 2009.