國立交通大學

資訊科學與工程研究所

碩 士 論 文

以 SVM 與詮釋資料設計書籍分類系統

A Book Classification System
Using SVM and Meta-information

研 究 生： 林昕潔

指導教授： 柯皓仁 教授

楊維邦 教授

中 華 民 國 九 十 五 年 六 月

以 SVM 與 詮 釋 資 料 設 計 書 籍 分 類 系 統
A Book Classification System Using SVM and Meta-information

研 究 生：林昕潔　　Student：Hsin-Chieh Lin

指導教授：柯皓仁　　Advisor：Hao-Ren Ke

楊維邦　　Wei-Pang Yang

國 立 交 通 大 學

資 訊 科 學 與 工 程 研 究 所

碩 士 論 文

A Thesis

Submitted to Institute of Computer Science and Engineering

College of Computer Science

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of

Master

in

Computer Science

June 2006

Hsinchu, Taiwan, Republic of China

中華民國九十五年六月

以 SVM 與詮釋資料設計書籍分類系統

研究生：林昕潔　　指導教授：柯皓仁　博士
楊維邦　博士

國立交通大學資訊科學與工程研究所

摘要

本研究提出一套書籍自動分類系統的設計方法，用以有效地節省分類大批書籍時所需的人力，提升書籍分類的效率；透過學習的方式，本研究所提出的系統可以套用於不同的分類架構上，使書籍類別更具彈性、更切合資訊脈動與使用者的需求。

本研究以文件分類為基礎進行書籍分類，輔以專家的經驗挑選類別特徵，並且將書籍的詮釋資料加入，以提高分類成效。本研究將書籍資訊分為敘述資料(Description)與詮釋資料(Meta-information)兩部分，其中敘述資訊包含書名、書籍簡介與作者簡介，詮釋資料包含作者與出版社資訊。本研究所提出的方法分為三大步驟：1) 對敘述資料進行前置處理，透過特徵挑選公式過濾出具有類別代表性的特徵，接著借助專家的智慧增加或刪除特徵，並將專家所指定的特徵予以加權，再配合選取的特徵將敘述資料轉換為向量表示式後，透過 Support Vector Machines (SVM)分類器產生分類模型；2) 分析統計書籍詮釋資料，發掘有助於書籍分類的資訊，且單獨運用這些資訊進行書籍分類；3) 以線性組合將 SVM 與詮釋資料的分類結果加以合併，完成書籍分類的工作。

在實驗中，使用的是「博客來網路書店」的書籍資訊，以 9-fold cross validation 的方式進行實驗，同時並列 Accuracy 與 F-measure 兩項評估數據，以求一個客觀整體的比較。實驗結果顯示，加入專家智慧挑選特徵並給予適當權重，可以提升 SVM 分類成果約 5%，再將 SVM 分類結果融合詮釋資料中隱含的分類資訊，可再提高約 5%，整體正確率達 95%。

關鍵字：書籍自動分類、網路書店、圖書館、SVM、詮釋資料

A Book Classification System Using SVM and Meta-information

Student：Hsin-Chieh Lin

Advisor：Dr. Kao-Ren Ke
Dr. Wei-Pang Yang

Institute of Computer and Information Science
National Chiao-Tung University

ABSTRACT

This thesis proposes an automatic book classification system to reduce the labor and time in classifying a batch of books. By means of machine learning, the proposed system can be utilized in various class structures. Using this system, the cataloging task in librares or online book stores can be more efficient than ever.

The proposed system classifies books based on document classification. It uses experts' knowledge in feature selection. The kernel classification algorithm used is Support Vector Machines (SVM), the result of which is integrated with books' meta-information to improve the classification correctnesss. In the beginning, data of books are divided into description (book title and prospectus) and meta-information (author and publicher). Description of each book is preprocessesd and features of a book are selected according to term frequentcies and log likelihood ratio. In the next step, experts refine these selected features, and give a weight for those features selected by them. After feature selection, descriptions of books are transformed into vector forms and use the SVM classifier to learn and classifiy. On the other hand, the meta-information is statictically analysised and extracted some hidden information useful for book classification. The final step is to linearly combine the SVM classification and meta-information for obtaining the final classification.

To prove the feasibility of our method, we use the book data in ‘books.com.tw’, experiment through 9-fold cross validation, and evaluate accuracy and f-measure. The experimental results show that adding experts’ knowledge may improve SVM results by 5%, and combining SVM and meta-information may hava an additional 5% improvement. The overall performance may achieve 95%.

Keywords: Book Classification, SVM, Support Vector Machine, Meta-information

致謝

寫到了致謝，代表了這篇論文、兩年的 DB Lab 研究生活，以及六年的交大求學生涯來到了尾聲。人生中最重要的一個求學階段得以圓滿地話下句點，有這麼充實的收穫與回憶，我的心中滿是感謝。

在研究生活中，首先要感謝楊維邦老師在關鍵的時刻帶領我進入資料庫實驗室，得以和交大再續兩年的緣分、並且在實驗室的引領下拓展更新更寬廣的視野；感謝柯皓仁老師盡心的指導，柯老師談話時詼諧幽默、對研究嚴謹仔細，輕鬆卻絲毫不放鬆，亦師亦友的引導啟發我向前；亦感謝黃明居老師不厭其煩的指導與建議，使論文得以一步一步向前邁進。在三位老師領導下，啟發我獨立思考並積極發掘問題的學習態度，這是老師們給我人生中最重要的資產。

資料庫實驗室的生活溫馨而充實，鎮源學長值早班、忠億學長值大夜班，實驗室幾乎 24 小時不打烊。鎮源學長在研究的過程中，總是積極的給予我許多方針，並兼顧學妹們的論文進度與身材，在用餐時刻安排論文進度，使我們頓時胃口全消、食不下嚥；在每週固定的 Meeting 中，信源學長提供了很多寶貴的經驗與建議，平時也會照三餐威脅利誘，叮嚀我們時時把研究進度放在心頭，不可以鬆懈；忠億學長則有許多獨到的見解，引領我學習使用不同的角度、不同的觀點去看事情，透過和忠億學長討論(其實是爭辯)，引導我更深入去思考我的題目，清晰且具體地描繪出我想呈現的觀點與方向，除了耐心的指導與訂正論文，偶爾還得兼任心靈導師，鼓勵我們、替我們建立信心。此外還有親愛的同窗好友家寧，兩年來一同修課、一同學習成長，不時互相鼓勵，在研究的路上一點都不孤單。

最後感謝我的家人對我全力的支持與包容，也謝謝我的兩隻小兔子喵喵與啵啵陪我度過一個又一個研究、實驗、寫論文的夜晚。

August, 2006

目錄

表目錄

圖目錄

一、緒論

本章描述本研究的背景、動機，與以及希望解決的問題。1.1 節由圖書館與網路書店的發展背景與功能來看書籍的分類與呈現方式；1.2 節為研究動機，以文件自動分類為基礎，進一步拓展應用到書籍自動分類；1.3 節為本研究希望達成的目的；1.4 節則說明本論文各章節的內容架構。

1.1. 書籍分類系統

在今日資訊發達的社會中，知識激增，對於各種資訊的吸收與學習，是身為現代人的必備工作。加上電腦與網路的發明，縮短了人與人之間的距離，資訊的傳遞更是無遠弗屆，而在資訊的傳遞上，書籍仍具有不可替代的地位。傳統上圖書館是書籍與知識的殿堂，而隨著網路的迅速擴張，網路書店亦日漸壯大，成為書籍流通的重要管道。圖書館與網路書店，兩者皆需要統整分類大批書籍資訊，由於其背景、考量、功能的差異，對書籍的管理分類方式也大不相同。1.1.1 節與 1.1.2 節將分別介紹圖書館與網路書店。

1.1.1. 圖書館

圖書館在人類文化發展上，自有文字紀錄的需求以來，一直扮演著一定的角色。早期圖書館的功能僅是單純蒐集與保存文獻，到十八世紀學者諾德(Gabriel Naudé)[39]提出圖書館有系統地展示所有紀錄下來的知識，向所有學者開放，建立近代圖書館的管理概念。1833 年，第一座以州教育公債資金建立的圖書館在美國新罕布夏州(New Hampshire)成立，1850 年代鄰州也紛紛通過州法，允許運用稅收來建立圖書館，開啟圖書館公辦的風氣。到二十世紀，科學研究和工業研究的發展，以全球為範圍的專業

資訊出版物(大部份是期刊形式)大量增加，導致使用者產生對快捷地檢索到廣泛期刊文獻的要求和對特定題目提供資訊和參考書目的要求。而圖書館目前主要的功能，是以有系統的方式蒐集、整理書籍、期刊、多媒體……等知識媒介，並使公眾易於接近、取得這些資訊。

圖書館的核心在於資訊的分類編目，有好的分類架構，才便於讀者找到所需要的資訊，而圖書館編目館員除了要有專業素養外，也要有好的參考分類工具，才能提供良好的使用環境給讀者。在國內各大專院校傳統圖書編目中，大部分西文圖書都使用「美國國會分類法」(Library of Congress Classification, LCC)[22]或「杜威十進位分類法」(Dewey Decimal Classification, DDC)[4]，中文部分則使用「中國圖書分類法」(New Classification Scheme for Chinese Libraries, CCL)[1]。

圖書分類與編目，乃是圖書館的重要工作之一，是個複雜且繁鎖的專業領域，圖書館編目員通常需要經過專業訓練方能勝任。一般而言，編目員進行圖書分類與編目的過程，可歸納為如下程序[2]：

1) 編目員瀏覽書名、目次表、篇章名、前言、序跋、參考書目、附註、正文及書評等之書籍內容做初步主題分析；
2) 決定書籍內容的學科重點與層面關係；
3) 主客觀地給予適當大類的類號(類表)；
4) 經由更詳細的書籍內容大綱、性質、用途及重點，而給予更明確的複分號與子目。

因此，一本書籍要完成分類，必須依次查看「類表」、「簡表」、「綱要表」以決定適當的類目及類號，由類號再進而查其「詳表」找到最能代表圖書資料的類號。因此，在分類與編目的過程中，編目員除了需要具備圖書館學專業素養外，對於各領域學科知識也要有一定程度了解，以利於掌握更明確的學科領域，而給予適當的類號及書目資料。

1.1.2. 網路書店

自 1990 年以來網際網路的發展突飛猛進,與其相關的活動及產品在全球各地迅速擴展，生活周遭也無時無刻受到網際網路的影響。在快速變遷的資訊時代中，人類不斷地創新、顛覆傳統，隨著科技的發展，網路寬頻世界的到來，消費者和網路生活可說是密不可分，電子商務的興起，線上消費者不斷地增加，根據表 1-1 與表 1-2,「書籍及雜誌」是高居榜首的在家上網的購物商品。

表 1-1　在家上網購買的產品與服務資訊[1]

產品與服務資訊	比例
書籍或雜誌	6.7%
3C 資訊產品	4.3%
時尚精品與服飾	4.0%
電腦軟體與網路使用時數與空間	3.0%
美容保養	2.7%
訂票/預約	2.1%
玩具遊戲	1.6%
居家生活	1.3%
旅遊休閒資訊	1.2%
行動通訊	1.0%
音樂 CD、VCD、DVD	1.0%

[1] 資料來源：經濟部技術處「產業電子化指標與標準研究」科專計畫／資策會電子商務研究所 FIND (2003)。本表為複選題，因此百分比相加不等於 100%。

表 1-2　最近一個月在家網路應用行為[2]

最近一個月的網路應用行為	2004 年 百分比	2005 年 百分比	年成長率
瀏覽資訊	87.9	89.2	2%
收發 EMAIL	73.3	77.6	6%
傳送即時短訊	49.6	55.0	11%
上傳、下載檔案	62.3	53.5	-14%
玩線上遊戲	37.0	37.6	2%
從事線上影音視迅活動	21.1	22.7	8%
以商家標定的價格購買產品或服務(線上購物)	13.3	19.6	47%
聊天室	14.2	17.4	23%
管理使用網路日誌	N/A	15.0	N/A
拍賣物品或服務、或者有參與競標行為(網路拍賣)	13.8	14.8	7%
利用轉帳或信用卡繳交帳單	9.8	11.4	16%
使用網路電話	6.1	10.7	75%
使用電子化政府服務	11.1	9.7	-22%
付費線上學習	9.3	8.0	-14%
從事線上投資理財之交易行為	5.7	6.2	9%

電子商務的興起帶動了線上消費者，而「書籍及雜誌」更是在家上網購物的主要商品，原因除了此類商品的服務已趨向專業成熟外，最重要的是「書籍及雜誌」屬於易衡量、預測且價格適中的商品。由發展進程來看，美國知名的 Amazon 網路書店成立於 1995 年 7 月，1997 年 1 月正式在美國掛牌上市；而台灣的博客來網路書店在 1995 年 12 月成立，於 2000 年 6 月開始與 7-ELEVEN 合作到店取貨。與美國相比之下，台灣的網路書店發展顯得時間較漫長。但是台灣近年來，除了博客來網路書店，也有許多線上書店陸續發展，像是小知堂先讀網、遠流博識網。這些新興的網路書店大約可以分成三種類型，一為通路型網路書店，像是博客來網路書店、絲路網路書店；二為

[2] 資料來源：資策會 ACI-IDEA-FIND/經濟部工業局「電信平台應用發展推動計畫」。「2005 年我國家庭寬頻、行動與無線應用現況與需求」調查

出版型網路書店，像是時報悅讀網、天下書房，此種的數量最為龐大；三為書店型網路書店，如誠品、金石堂、新學友網路書店等[41]。

網路書店蒐集各式各樣的書籍資訊，建立豐富的資料庫，利用搜尋引擎及資料庫技術以提供使用者完整的書籍資訊。雖然其興起不過短短數年，但所受到的歡迎程度及擁有的顧客群相當廣泛。基於商業考量，網路書店必須精確掌握使用者對資訊的需求與市場脈動，市面上林立的網路書店也各有一套將書籍分門別類的標準。除此之外，網路書店與圖書館最大的不同之處，在於不需要考慮到書籍排架擺放的問題，因此書籍可以不受限制地同時出現於多個類別中，在使用者進行類別瀏覽時增加書籍的曝光率。

1.2. 研究動機

不論是使用何種類型的書籍分類方法，最終目的就是將書籍做有規律性、一致性、有效性的分類，以便圖書館館員或網路書店員工能有效率地排列、整理書籍，亦便於使用者查找資料，才能解決「資訊爆炸」、「資訊超載」的現象。

現今網際網路的時代裡，各種資料型態不斷增加，新的議題不斷產生，資訊持續不斷累積。在這樣快速變遷的環境下，書籍的分類編排方式能否跟上腳步，切合使用者的需求？更甚者，不論是圖書館或網路書店，書籍動輒萬冊以上，一旦分類架構有所更動，以人工重新對書籍進行分類將會是一個浩大的工程。

在另一方面，文件自動分類技術已經發展多時且廣泛運用，因此本研究嘗試將文件自動分類的方法應用於書籍的分類上。

1.2.1. 文件自動分類

「文件主題分類」或簡稱「文件分類」(Document Classification or Text Categorization)

是指依文件「內容主旨」給定「類別」(Class or Category)。例如，新聞文件可按其報導的內容，給予「政治」、「外交」、「運動」、「娛樂」等類別。

文件分類的目的，在於對文件進行分門別類的加值處理，使得文件易於管理與利用。分類後的文件，可提供使用者依主題查找文件而不受文件用詞的限制。在全球資訊網出現後，文件分類在協助使用者尋找網路資訊上，扮演非常重要的角色。例如，有些入口網站聘請大量的人員進行文件分類，以提供網站或網頁分類目錄的服務。以Yahoo!而言，其搜尋系統甚至委外建置，過去數年來從 Infoseek 換成 Altavista 再換成 Google，但是其分類目錄則由本身持續不斷地維護，足見分類對企業本身的價值與重要性。

近年來，拜資訊技術普及之賜，各企業與機構的數位文件不斷累積，數量大到難以有效地管理與利用，文件分類的需求也就因應而生。為此，如何利用自動化的技術，快速有效地協助人工分類，以應付大量暴增的分類需求，是現今資訊服務與知識管理的重要課題。

文件分類自動化後廣泛運用於各種應用，除了提供主題檢索(Topic-based Retrieval)、文件管理(歸檔、調閱、分享)外，還可應用在網頁過濾、電子郵件過濾、資訊選粹(SDI, Selected Dissemination of Information)、資訊配送(Information Filtering or Routing)、甚至是文字探勘(Text Mining)、新知發掘(Knowledge Discovery)、知識管理(Knowledge Management)等領域。

1.2.2. 書籍自動分類

本研究嘗試將文件自動分類技術應用於書籍分類上，表 1-3 比較文件資訊與書籍資訊欄位，一般文件分類對文件的本文(Text)進行分析；而在書籍方面，在書籍內文尚未數位化之前，可利用的的資訊有書名、簡介、作者與出版社等各項資訊。本論文將書名、書籍簡介與作者簡介對應到文件本文，稱為敘述資料(Description)，其餘的作者與出版社資訊則歸類為詮釋資料(Meta-information)。觀察表 1-3 可以發現書籍資料比

一般文件多了豐富的詮釋資料，若能處理分析當中所隱含的與類別相關的資訊，則將對書籍分類有所幫助。

綜上所述，本研究希望基於文件分類的技術來進行書籍分類，並運用書籍的詮釋資料提高分類成效；此外，不論是圖書館或網路書店，皆有專業的分類編目人員，如何運用這些專家的經驗提昇分類效能亦是本論文探討的議題之一。

表 1-3　文件資訊與書籍資訊之比較

	文件資訊	書籍資訊
敘述資料	本文	書名 簡介
詮釋資料	-	作者 出版社

1.3. 研究目的

本研究之目的為建立一套書籍分類系統，自動對書籍進行分類。建立此系統有兩個優點：

1) 自動分類大批書籍，提升書籍分類的效率。
2) 此系統可以很容易地套用在各種分類架構上，如此一來，若對現有的分類架構不滿意，或者是希望能將分類架構調整得更切合使用者需求，則更動分類架構、重新分類現有書籍的工程不再令人望之卻步。

1.4. 論文架構

本論文分成五章。第二章介紹與文件自動分類相關的研究；第三章闡述書籍分類系統的設計方法，說明如何導入專家的智慧於書籍特徵挑選(Feature Selection)的過程，並融合Support Vector Machines (SVM)分類器與詮釋資料建立書籍分類系統；第四章說明實驗與實驗結果的分析討論，以驗證本論文所提方法的可行性；最後一章是結論與未來可繼續發展的方向。

二、 相關研究工作

本章說明本論文「書籍分類系統」之相關研究工作，主要分為三方面：

- 常用的分類法
- 常用的特徵挑選法
- 評估方法

由於本論文所提出的「書籍分類系統」是以文件分類為基礎，因此首先在 2.1 節介紹各種常見的文件分類方法。文件分類大多以「字」作為分類的特徵屬性，常常會有特徵過多的問題，因此過濾挑選特徵的方法廣泛地應用在文件分類中，在 2.2 節中將介紹常見的特徵挑選法。 2.3 節則討論分類結果正確性的評量方式。

圖 2-1 依照年份及技術整理了相關分類法研究的發展，粗體字的部分是本論文主要參考的研究。

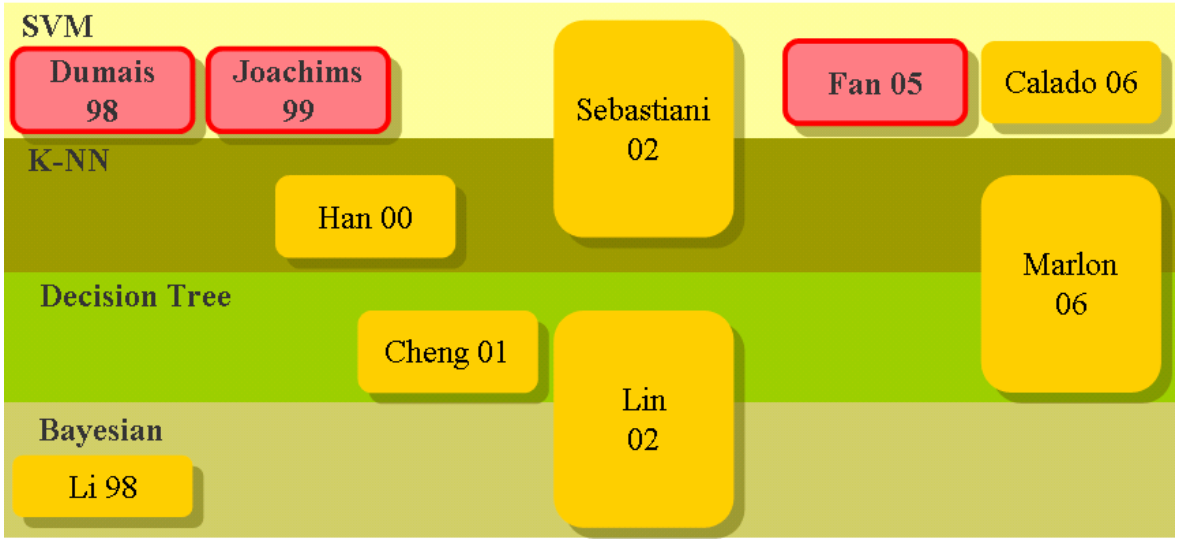

圖 2-1　相關研究發展

2.1. 分類法相關研究

一般分類工作之流程如圖 2-2 所示。蒐集文件資料後，進行前置處理過濾雜訊，再經由特徵挑選進一步地精簡特徵，盡量取出最有利於分類的特徵，而後將文件以特殊的形式表現特徵(多數是採用向量表示法)，經由分類演算法分類後轉為有用的資訊進行應用。

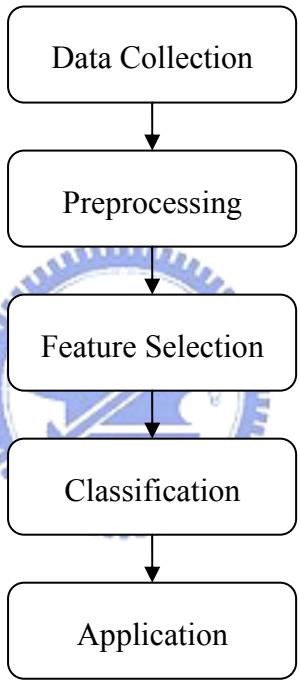

圖 2-2　一般分類流程圖

應用於文件分類常見的演算法有決策樹(Decision Tree)、Naïve Bayesian、K-NN，以及近期新興的 SVM。決策樹是一個類似流程圖的樹狀結構，由各種不同的條件於不同階段形成的節點及節點間的分支所構成；Naïve Bayesian 利用條件機率與獨立的假設預測分類結果；而 K-NN 則利用訓練資料與測試資料之間的距離，進行分類的動作；SVM 分類器使用核心函數，進行訓練學習。

2.1.1. 決策樹 (Decision Tree)

決策樹[28][30]是使用於資料探勘(Data Mining)與機器學習(Machine Learning)最受歡迎的分類演算法之一。決策樹是一個類似流程圖的樹狀結構。此樹狀結構之內部節點(Node) 表示某一種屬性的測試，如圖 2-3 所示，每個分枝(Branch)則代表一種測試結果，而每個樹葉節點(Leaf Node)則代表類別(Class)或類別分佈(Class Distributions)。

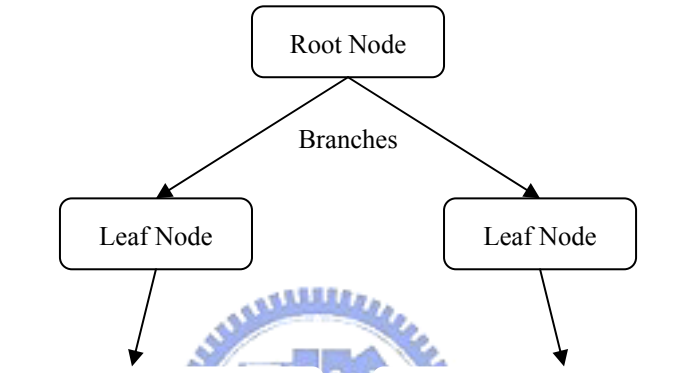

圖 2-3 決策樹示意圖

決策樹歸納法(Decision Tree Induction)的基本演算法是一貪婪演算法(Greedy Algorithm)，它是以從上到下遞迴的(Top-down Recursive)且採用 Divide-and-Conquer 的方式來建構決策樹，使得訓練資料 (Training Data)中屬於同一類別的資料最後歸在此決策樹的同一個樹葉節點(Leaf Node)中。由訓練資料建立好決策樹後，一筆新進的資料可經由決策樹被歸到某個樹葉節點，完成此資料的分類工作。

在決策樹的每個節點上使用信息增益(Information Gain, IG)，其是基於熵(Entropy)的度量作為啟發資訊(Heuristic Information)，並選擇某個能夠將樣本分類的最佳屬性。這種度量稱作屬性選擇度量(Attribute Selection Measure)或分裂的優良性度量(Measure of the Goodness of Split)。透過 2.2.3 節的方法分別計算每個屬性的 IG，選 IG 值最高的作為分裂的測試屬性。圖 2-4 為決策樹依測試屬性分枝的示意圖，若有 *m* 個屬性，則此決策樹高度不會超過 *m*。用此方法以 Top-down Recursive 的方式建立出決策樹。

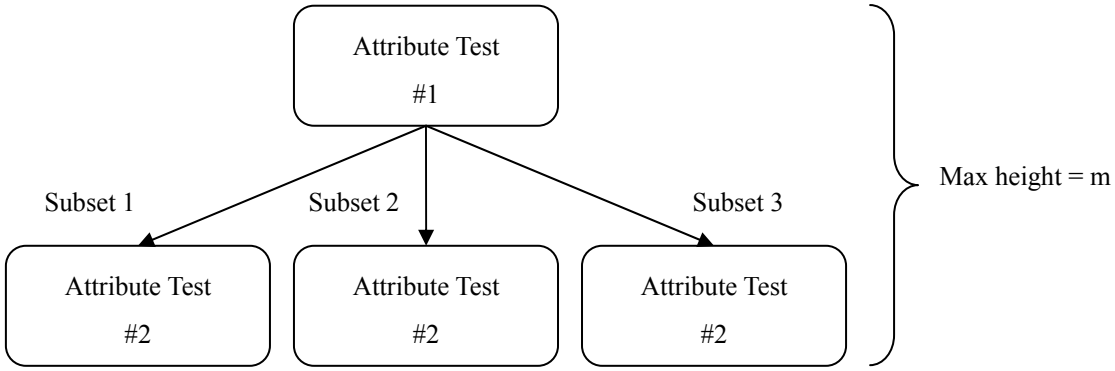

圖 2-4　決策樹測試屬性示意圖

圖 2-5 為著名的決策樹歸納演算法 ID3[26]的一種版本，使用 IG 以 Top-down Recursive 方式建立的過程：

1) 由根節點(Root Node)開始。
2) 選一個尚未被使用於上方節點(Ancestor Node)且具有最高 IG 的屬性。
3) 以 2) 之屬性為基礎，為此屬性每一個可能的數值結果新增一個子節點(Child Node)。
4) 把所有的訓練資料加入與其屬性數值相符的子節點中。
5) 若位於某子節點中的訓練資料皆屬於同一類別 C_i，將此節點標為樹葉節點，且歸類為 C_i。
6) 跳回 2)，直到所有的屬性都用完。此時若有樹葉節點尚未被標示類別，則以此節點中最多訓練資料所屬的類別為作為標示。

圖 2-5　ID3 演算法

2.1.2. Naïve Bayesian Classifier

貝氏(Bayesian)分類法[11][12]是基於統計學所發展的分類方法，可以預測分類成員關係的可能性，例如給定樣本屬於一個特定類別的機率。

現有一筆資料 d_i 和類別 c_k，依照貝氏定理(Bayes Theorem)，d_i 屬於 c_k 的機率為：

$$P(c_k \mid d_i) = \frac{P(d_i \mid c_k)P(c_k)}{P(d_i)} \tag{2.1}$$

d_i 將會被分類到計算出來機率最高類別。

然而計算 $P(d_i \mid c_k)$ 的運算成本可能非常大，為了降低運算成本，一般都做類別條件獨立(Class Conditional Independence)的假設。假設類別 c_k 和 d_i 的屬性值是條件獨立的，也就是在屬性之間並不存在相依關係，如此一來 $P(d_i \mid c_k)$ 可簡化為：

$$P(d_i \mid c_k) = \prod_{j=1}^{m} P(d_{ij} \mid c_k) \tag{2.2}$$

其中 d_{ij} 是 d_i 中的第 j 項屬性。

而如此化簡之後所建構之分類器，即為 Naïve Bayesian Classifier。儘管這個條件獨立的假設在現實生活中鮮少成立，許多研究中 Naïve Bayesian 卻有不錯的效果。

2.1.3. K-最鄰近分類法 (K-nearest Neighbor Classifier, K-NN)

2.1.3.1. 最鄰近分類 (Nearest Neighbor Rule, NN)

最鄰近分類法(Nearest Neighbor Rule, NN)在許多分類統計教科書中都有討論，如[12][20]，其所根據的基礎是「物以類聚」，換成數學語言來說：若以空間中的點來表示物件，則屬於同一類別的點之間距離會比較接近；因此對於一筆未知類別的資料，只要找出在訓練資料中和此資料最「接近」的點，就可以判定此筆資料應該和最接近的點屬於同一類別，在這裡所謂的「接近」通常是以歐基里德空間之距離為度量的方式。最鄰近分類法是一個最直覺的分類法，在測試各種分類器時常被當成是最基礎的分類器，以便和其他複雜的分類器進行效能比較。

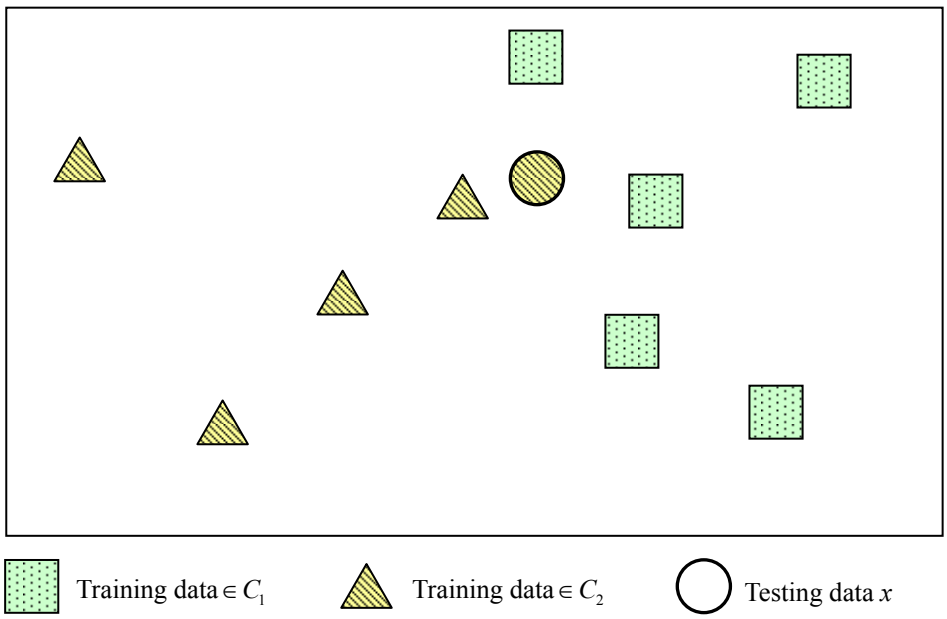

圖 2-6　NN 實例

圖 2-6 為 NN 的實例：訓練資料被分入兩個類別 C_1(三角形)與 C_2(正方形)，其中 C_1 的資料與未知資料 x(圓形)最接近，於是 x 被標示屬於 C_1。

2.1.3.2. K-最鄰近分類 (K-Nearest Neighbor Rule, K-NN)

在訓練資料雜訊很強的情況下，若只用最靠近的資料來決定類別，可能會失之武斷，因此另外有一個常見的做法，先求取最接近的 k 個資料點，再根據對應的 k 個類別資訊進行投票，以決定最後的類別。這種方法稱為 k-最鄰近分類法(K-nearest Neighbor Rule, K-NN)，也就是對未知類別之資料求出前 k 個最靠近的鄰居來投票決定此資料之類別。

圖 2-7 為 K-NN 的實例，取 k=3，同前例，有兩個類別 C_1(三角形)與 C_2(正方形)。與未知資料 x(圓形)最接近的三個鄰居中，因為有兩個來自 C_2，於是 x 被標示屬於 C_2。

決定 k 值大小有以下兩個原則：

1) k 值必須大到足以減少誤分 x 的機率。

2) 為了正確評估 x 真正屬於的類別，k 值必須夠小，讓選出來的參考資料盡量靠近 x。

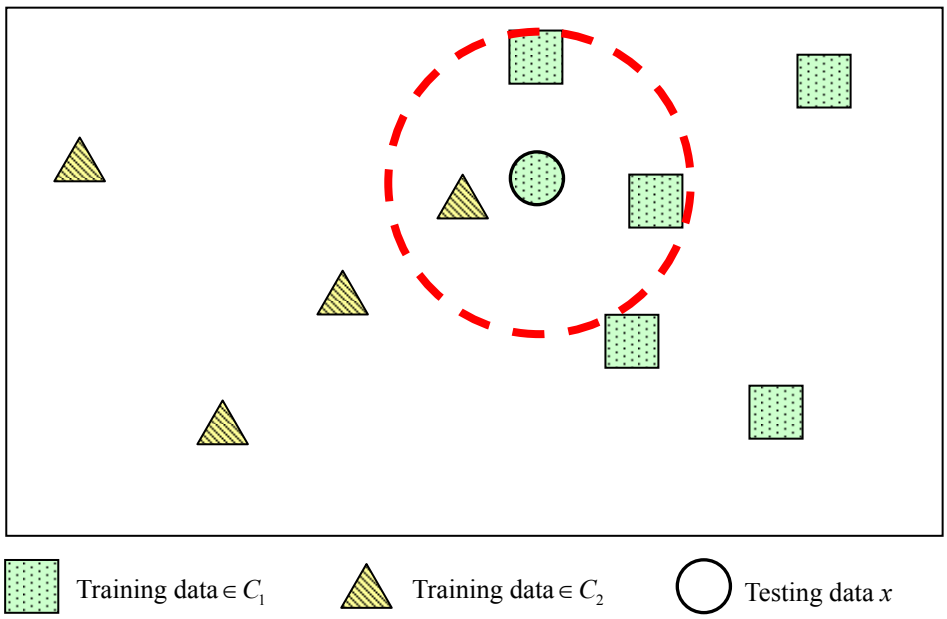

圖 2-7　K-NN 實例

2.1.4. Support Vector Machines (SVM)

支援向量機(Support Vector Machines, SVM)最早由 Vapnik 於 1979 年提出[36]，是一種以統計學習理論(Statistical Learning Theory)為基礎，而發展出來的機器學習系統。SVM 的應用領域相當廣，包含：生物資訊(Bioinformatics)、資訊探勘、影像識別(Image Recognition)、文字分類(Text Categorization)、手寫數字辨識(Hand-written Digit Recognition)等範圍。

當有一些資料要分成兩類，這些資料點不一定位在二維空間 $\boldsymbol{R}^2$，可能在多維的空間，SVM 的概念為找出一個超平面(Hyperplane)[3]，將分屬於兩個類別的資料點[4]分隔開。這個超平面與最靠近的資料點之間的距離稱為邊界(Margin)，其值越大越理想，因為能將這些訓練資料盡量明確地區隔開。當有新進未知類別的資料，期望可利用此超平面正確判定新資料所屬的類別，為此要盡量找出一個邊界值最大的超平面

[3] 超平面(Hyperplane)，由高維空間對應到三維空間上的平面。
[4] SVM 中的資料皆以向量(Vector)表示之，即每筆資料都是高維空間中的一個向量。

(Maximum-margin Hyperplane)或稱為最佳超平面(Optimal Hyperplane)，而空間中最靠近最佳超平面的資料點稱為支持向量(Support Vector)。

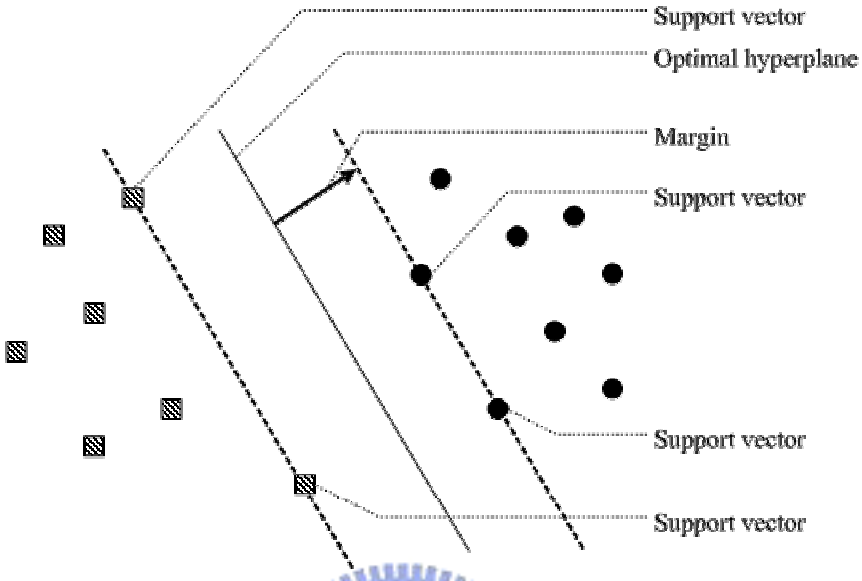

圖 2-8　SVM 概念示意圖

假設有 n 個資料點表示為：$\{(\vec{x_1},c_1),(\vec{x_2},c_2),...,(\vec{x_n},c_n)\,|\,\forall i=1...n, c_i\in\{+1,-1\}\}$ ，意即用±1 來標示資料點 $\vec{x_i}$ 所屬的類別。拿以上這些資料點作為訓練資料，找出最佳超平面並用以建構 SVM，表示為：

$$\vec{w}\cdot\vec{x}-b=0 \tag{2.3}$$

其中 $\vec{w}$ 代表 margin，b 為一常數。

假設 x_1 與 x_2 分別為+1 類別與−1 類別之 support vector，對於分別通過 x_1 與 x_2 且與最佳超平面平行的兩個超平面，我們將其定義為 p_1 與 p_2。經由(2.4)推導可得知要求得最大的 Margin 相當於求 $\|\vec{w}\|$ 的最小值：

$$
\begin{aligned}
& p_1 : \vec{w} \cdot \vec{x_1} - b = +1 \\
& p_2 : \vec{w} \cdot \vec{x_2} - b = -1 \\
& \Rightarrow \vec{w} \cdot (\vec{x_1} - \vec{x_2}) = 2 \\
& \Rightarrow \frac{\vec{w}}{\left\| \vec{w} \right\|} \cdot (\vec{x_1} - \vec{x_2}) = \frac{2}{\left\| \vec{w} \right\|}
\end{aligned}
\tag{2.4}
$$

此外，由於 support vector 為最佳超平面之訓練資料，因此不可能有其他訓練資料存在於 p_1 與 p_2 之間，則可得條件式(2.5)：

$$
\begin{aligned}
& \vec{w} \cdot \vec{x_i} - b \geq 1 \ \text{ or } \ \vec{w} \cdot \vec{x_i} - b \leq 1 \\
& \Rightarrow \exists k_i \ni k_i (\vec{w} \cdot \vec{x_i} - b) \geq 1, \forall i \ 1 \leq i \leq n
\end{aligned}
\tag{2.5}
$$

因此，要解決的問題為 minimize $\left\| \vec{w} \right\|$，並同時符合(2.5)之條件，這是一個二次規劃最佳化問題(Quadratic Programming Optimization Problem, QP)。

SVM 經過學習之後，對於未知類別的新資料，可以依照規則(2.6)分類之：

$$
\hat{c} = \begin{cases} +1, & \text{if } \vec{w} \cdot \vec{x} + b \geq 0 \\ -1, & \text{if } \vec{w} \cdot \vec{x} + b \leq 0 \end{cases}
\tag{2.6}
$$

由(2.6)可以發現，SVM 在對未知類別之資料進行分類時，僅是進行簡單的向量運算，並不會佔用太多運算時間，這也是 SVM 的優點之一。

2.2. 特徵挑選法(Feature Selection)

分類文件時，通常以文件內容的詞彙(Term)作為分類的特徵。然而一般文件進行前置處理(Preprocessing)過濾雜訊後，仍會留下大量的詞彙，此時需要進一步做特徵挑選的動作，取出對分類最有利的特徵。配合不同的演算法，常見的特徵挑選法有最基本以計算頻率的 TF-IDF，以機率為基礎的互見訊息(Mutual Information)、信息增益(Information Gain)，以及配合統計的 χ^2- test 與 likelihood-ratio test。

2.2.1. Tf-IDF

TF(Term Frequency)[31][32][33]衡量一個詞彙對其所存在特定文件的重要性，計算方式如(2.7)所示，其中 n_i 為被考慮的詞彙的出現次數，而分母的部分則代表所有詞彙出現次數的總和。而 IDF (Inverse Document Frequency)則用來評斷詞彙對整體資料(文件集中的所有文件)的重要性；如(2.8)所示，其中$|D|$是文件集(Corpus)中的文件總數，$\left|(d_j \supset t_i)\right|$則代表內容中出現過詞彙 t_i 的文件總數。以 TF-IDF 作為權重的方法常見於資訊擷取(Information Retrieval)與文字探勘(Text Mining)，也時常應用在搜尋引擎(Search Engine)上。

$$tf = \frac{n_i}{\sum_k n_k} \tag{2.7}$$

$$idf = \log_2\left(\frac{|D|}{\left|(d_j \supset t_i)\right|}\right) \tag{2.8}$$

從(2.7)與(2.8)可得 TF-IDF 權重如(2.9)：

$$tf - idf = \frac{n_i}{\sum_k n_k} \times \log_2\left(\frac{|D|}{\left|(d_j \supset t_i)\right|}\right) \tag{2.9}$$

TF-IDF 是以統計學的方法評估詞彙對文件的重要性。一個詞彙的重要性與其在特定文件中的出現次數成正比；但若在文件集中，這個詞彙過於常見於各篇文件，則被視為普遍而不重要。因此，被 TF-IDF 權重法評斷為重要的詞彙，其在給定文件中會具有較高的 TF 值，並且僅在文件集的少數文件中出現。透過這種加權方式可以過濾掉過於普遍的詞彙。

2.2.2. Mutual Information (MI)

在機率理論(Probability Theory)中的資訊理論(Information Theory)領域，互見訊息(Mutual Information, MI) [9][29]可用以決定兩個隨機變數(Random Variable)之間互相依

賴(Mutual Dependence)的程度。假設兩個隨機變數 X 與 Y，直覺來說，MI 測量的資訊是屬於 X 並且與 Y 共享。如果 X 與 Y 互相獨立(Independent)，則 X 中沒有任何與 Y 相關的訊息，反之亦然，Y 亦沒有任何關於 X 的訊息，此時 MI 的值將為零。如果 X 與 Y 完全相同(Identical)，則 MI 等同於單獨由 X 或 Y 獲得的資訊，又稱為 X 的熵(Entropy)。

對於兩個離散(Discrete)的隨機變數 X 與 Y，MI 定義為：

$$I(X;Y)=\sum_{y\in Y}\sum_{x\in X}p(x,y)\log_2\frac{p(x,y)}{p(x)p(y)} \tag{2.10}$$

其中 $p(x,y)$是 X 與 Y 的聯合機率分佈函數(Joint Probability Distribution Function)，$p(x)$ 與 $p(y)$則分別對應到 X 與 Y 的邊界機率分佈函數(Marginal Probability Distribution Function)。

當 X 與 Y 互相獨立，則 $p(x,y)=p(x)\,p(y)$，因此可經由推導得知：

$$\log\frac{p(x,y)}{p(x)p(y)}=\log 1=0 \tag{2.11}$$

進而獲得(2.12)之結論：

$$I(X;Y)=0\leftrightarrow X \text{ and } Y \text{ are indepdent} \tag{2.12}$$

除此之外，MI 是對稱(Symmetric)的，意即 $I(X;Y)=I(Y,X)$。

應用在分類過程中的特徵挑選上，term t_i 與類別 C_k 之間的 MI 可寫為：

$$I(t_i;C_k)=\log\frac{p(t_i,C_k)}{p(t_i)p(C_k)}=\log\frac{\frac{freq(t_i,C_k)}{N}}{\frac{freq(t_i)}{N}\times\frac{1}{CN}} \tag{2.13}$$

N 為文件集中的文件總數，CN 為類別總數，$freq(t_i)$為 term 在文件集中出現過的總次數，$freq(t_i,C_k)$為 term 在 C_k 中的出現次數。當 MI 大於零，表示 t_i 和 C_k 傾向一起出現，值越大，「共現率」越高。如果 MI 小於零，表示 t_i 和 C_k 傾向「不」一起出現，負值越大，「互斥率」越高。

2.2.3. Information Gain (IG)

選擇一個屬性來作為分枝的測試屬性是決策樹演算法中最重要的步驟之一。選擇的方法有很多種，其中最廣為人知的方法為信息增益(Information Gain, IG) [27][30]。

進行分類時，定義訓練資料的集合為 S，假設類別的數量為 m，則以 $C_1, C_2, \ldots, C_m$ 代表這 m 個類別。假設 s_i 是屬於類別 C_i 的訓練資料總數，p_i 是任一訓練資料屬於 C_i 的機率，則 $p_i = \frac{s_i}{|S|}$，對一個給定的資料，可以求出它的期望資訊：

$$I(s_1, s_2, ..., s_m) = \sum_{i=1}^{m} p_i \log_2(p_i) \tag{2.14}$$

假設屬性 A 具有 v 個不同的值：$\{a_1, a_2, ..., a_v\}$。可用 A 將 S 劃分為 v 個子集合 $\{S_1, S_2, \ldots, S_v\}$，其中 S_j 所包含的樣本屬於 S 且在 A 上的值為 a_j。假設 s_{ij} 是子集合 S_j 中類別 C_i 的樣本數，根據由 A 劃分成之子集合的熵值為：

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + \ldots + s_{mj}}{s} I(s_{1j}, \ldots, s_{mj}) \tag{2.15}$$

其中項目 $\frac{s_{1j} + \ldots + s_{mj}}{s}$ 充當第 j 個子集合的權重，並且等於子集合(即 A 值為 a_j)中的樣本個數除以 S 中的樣本個數。熵值越小，則子集合劃分的純度越高。

在 A 上分枝將獲得的 IG 為：

$$Gain(A) = I(s_1, s_2, ..., s_m) - E(A) \tag{2.16}$$

儘管 IG 在決定相關性時通常是一個不錯的測量方法，但它並非完美。當 IG 使用在計算擁有大量不同數值的屬性上，會有一個值得注意的問題。舉例來說，假設用關於顧客的商業資料建立決策樹，IG 通常被使用來決定哪一個屬性是最有關聯的，則此屬性的測試位置會接近根節點。這些資料中可能有一項屬性是顧客的信用卡卡號，這個屬性具有很高的 IG，因為每位顧客的信用卡卡號都是獨一無二的，但我們並不希望把這個屬性放到決策樹中，因為對一個從未見過的新客戶而言，用信用卡卡號來決定如何對待他很可能會失敗。因為有這種狀況的存在，決策樹選擇屬性時，對於擁有大量不同數值的屬性偏向不予採用。

2.2.4. χ^2-test

2.2.4.1. χ^2- distribution

在機率理論與統計學的領域裡，χ^2-分佈(χ^2-distribution)[5][10][37]是被廣泛運用在推論統計(Inferential Statistics)上的機率分佈假設。在合理的假設下，如果 null hypothesis[6] 為真, 那麼在這些前提成立的時候, 則可以用方便的方法計算趨近 χ^2- distribution 的值，圖 2-9 為 χ^2- distribution 之機率分佈圖。

現有 k 個互相獨立且正規分佈(Normally Distributed)的隨機變數 $X_1, X_2,..., X_k$,。對於每個 Xi，平均值為 μ_i，變異數(Variance)為 σ_i^2，則可得依照 χ^2- distribution 分佈的隨機變數 Z：

$$Z = \sum_{i=1}^{k} \left(\frac{X_i - \mu_i}{\sigma_i} \right)^2 \tag{2.17}$$

標記為：

$$Z \sim \chi_k^2 \tag{2.18}$$

χ^2- distribution 有一參數 k，k 為正整數，代表自由度(Degrees of Freedom, DF)[7]，也就是 X 的個數。χ^2- distribution 是 gamma distribution 的一個特例。

[5] χ^2 distribution 亦可寫作 chi-square distribution 或 chi-squared distribution。

[6] 在統計學中，null hypothesis 是一種建立來被證明無效或錯誤的假說，用以支持另外一個替換的假說(Alternative Hypothesis)。使用上，null hypothesis 一開始會被假設為真，然後再提出統計上的證據反證。使用 null hypothesis 有時會受到爭議。

[7] 評估參數這項工作可以基於不同的資訊量來處理。評估一個參數時，其中所用到的獨立資訊數量稱為自由度(Degrees of Freedom, df)。

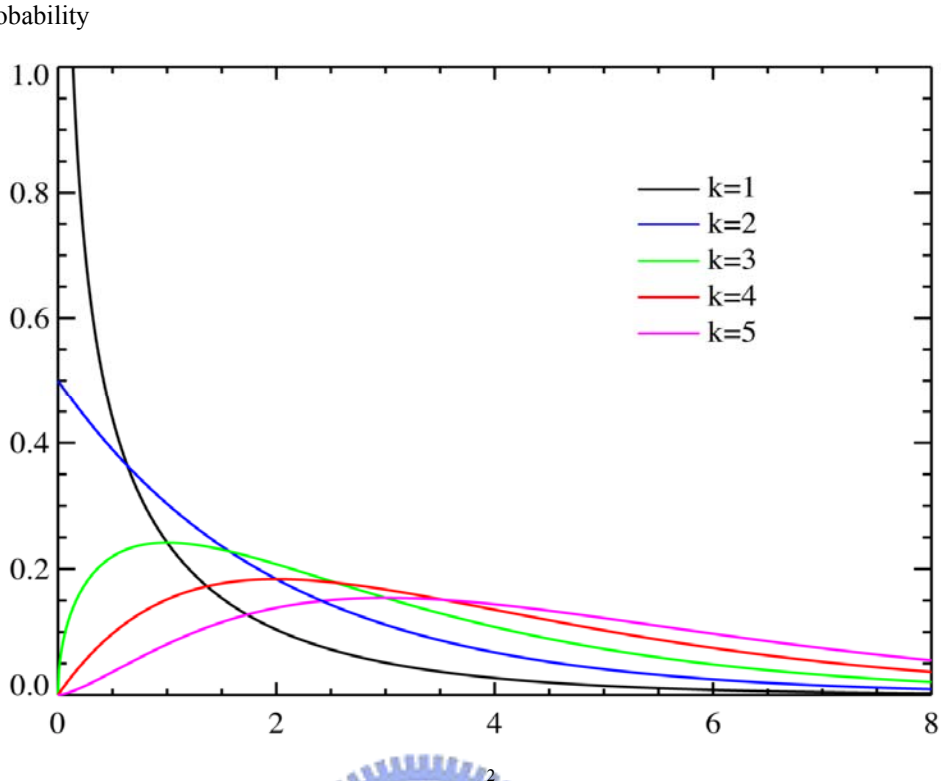

圖 2-9 χ^2- distribution 之機率分佈圖[8]

2.2.4.2. χ^2-test

任何統計的 hypothesis test，在 null hypothesis 為真的情況下，若結果呈現 χ^2-distribution 則可以稱為 χ^2-test[10][37]。χ^2-test 是一種統計工具，用來區別隨機變化(Random Variation)所造成真正的影響，適用於符合下列條件的資料：

1) 由大量資料中隨機抽取的樣本。
2) 呈現方式為原始計算出來的頻率，而非百分比或比率。
3) 所測量的變數必須是互相獨立的。
4) 在獨立變數和相依(Dependent)變數上的數值必須互斥(Mutually Exclusive)。
5) 所觀察的頻率不可過小。

χ^2-test，不論是檢定兩組資料為互相獨立的效果或兩者適合，效果都不錯。

[8] 圖片來源：http://en.wikipedia.org/wiki/Chi-square_distribution (2006/5/4)。

2.2.5. Likelihood-ratio Test

2.2.5.1. Likelihood Function

Likelihood[21][28]是一種假設機率(Hypothetical Probability)：某個已經發生的事件(Event)將會產生一種特定的結果。不同於以往：「機率不受過去結果的影響，只與未來的事件有關」[9]的概念，likelihood 會考慮到過去事件的結果。

Likelihood function 是一個考慮二個參數的條件機率函數，它的第一個參數是固定的。令 A 與 B 為兩個事件，b 為 B 的一種結果，可得：

$$b \mapsto p(A \mid B = b) \tag{2.19}$$

也就是說，B 的 likelihood function 即為函數(2.20)的等價類別(Equivalence Class)：

$$L(b \mid A) = \alpha p(A \mid B = b) \tag{2.20}$$

其中 α 為大於零之比例常數。因此 $L(b|A)$的實際數值並不是重點，我們所關注的是(2.21)這個比例式：

$$\frac{L(b_2 \mid A)}{L(b_1 \mid A)} \tag{2.21}$$

這是關於 α 的不變量(Invariant)，即不論 α 值為何，(2.21)值不變。

2.2.5.2. Likelihood-ratio Test

Likelihood-ratio test[21][28]為一種用來評斷兩個模型(Model)之間合適程度的統計檢定，使用的方法是在 null hypothesis 的限制下求 likelihood function 比率的最大值。令此比率為 Λ，當 null hypothesis 成立，則就一般常見的機率分佈而言，$-2\log\Lambda$ 的特殊漸進分佈(Asymptotic Distribution)簡單易用。許多常見的統計檢定，如 Pearson's χ^2 test，可以代換為 log-likelihood ratio 或是其近似值。

---

[9] 例如：投擲一個公正的銅板，不論先前結果是什麼，下一次擲出正面的機率仍為 1/2。

通常假設 null hypothesis 的參數 θ 位在參數空間 Θ 的某個特定子集合中 Θ_0，則經由推導 $L(\theta) = L(\theta|x) = p(x|\theta) = f_\theta(x)$，likelihood function 可表示為一個有參數 θ 與變數 x 的函數。則 likelihood ratio 可表示為：

$$\Lambda(x) = \frac{\sup\{L(\theta|x) : \theta \in \Theta_0\}}{\sup\{L(\theta|x) : \theta \in \Theta\}} \tag{2.22}$$

這是一個表示資料 x 的函數，因此是個統計量。若此統計量過小，則 likelihood-ratio test 否定 null hypothesis，同時此性質可由 Neyman-Pearson lemma[10]印證。究竟「多小」才算是「太小」呢？這要取決於這個測試的重要程度，亦即所能容忍的 Type I error[11]機率("Type I error" consist of rejection of a null hypothesis that is true)。

若 null hypothesis 為真，且觀察結果為 n 個一連串的獨立獨特分佈(Independent Identically Distribute, i.i.d)的隨機變數，則樣本大小 n 將會趨近於∞，$-2\log\Lambda$ 則會趨近於 χ^2- distribution，其中 χ^2- distribution 的 DF 等同於 Θ 與 Θ_0 之間的維度差距。

以 Pearson's test[9]其中一個例子來說明：投擲兩枚硬幣，想要比較其正面朝上的機率是否相等。將所有狀況之觀察結果以表 2-1 表示，其中各項皆為投擲結果的累計次數。

表 2-1　Pearson's test 的一個例子

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

[10] Neyman-Pearson lemma：若存在一個 critical region C，其大小為 α，且存在一個非負的常數 k 使得在 C 之中的點符合：$\frac{\prod_{i=1}^{n} f(x_i|\theta_1)}{\prod_{i=1}^{n} f(x_i|\theta_0)} \geq k$，且不在 C 中的點符合：$\frac{\prod_{i=1}^{n} f(x_i|\theta_1)}{\prod_{i=1}^{n} f(x_i|\theta_0)} \leq k$，則以大小為 α 言 C 為最佳的 critical region。

[11] Type I Error，亦即 False positive。發生在測試結果錯誤，且錯誤情況是出現正向結果，但是不應該存在正向結果。換句話說，可以想成 Type 1 error 否決 null hypothesis 其實並不正確，即接受這個替換的假說(Alternative Hypothesis) 及使 null hypothesis 為真。

在此 ω 由參數 p_{1H}、p_{1T}、p_{2H} 及 p_{2T} 組成，分別代表硬幣 1、硬幣 2 所擲出正反面結果的機率。令這整個 hypothesis space H 符合一般分佈的限制：$p_{ij} \geq 0, p_{ij} \leq 1$，且 $p_{iH} + p_{iT} = 1$。而是 null hypothesis H_0 符合條件 $p_{1j} = p_{2j}$ 之子空間。在以上所有的限制中 $i = 1,2$ 且 $j = H,T$。此 hypothesis 與 null hypothesis 可以稍加修改使其符合 log-likelihood ratio，以獲得更理想的分佈狀況。因為這項限制使得原本為二維的 H 減少成一維的 H_0，這個測試的漸進分佈為 $\chi^2(1)$，其中 χ^2 distribution 的 DF=1。

就一般的狀況表而言，可以把 log-likelihood ratio 的計算式表示為：

$$-2\log\Lambda = \sum_{i,j} k_{ij} \log \frac{p_{ij}}{m_{ij}} \tag{2.23}$$

本論文即採用 log-likelihood ratio 作為特徵挑選的方法之一。

2.3. 評估方法

由於學習演算法有可能對資料過份最佳化(Overfitting)，使用訓練資料建立模型，在評估分類結果導致太過於樂觀的估計。保持交叉檢定(Holdout Cross-validation)與 k-折交叉檢定(K-fold Cross-validation)基於給定的資料作隨機抽樣劃分，常用於評估分類法的準確性。[17]

2.3.1. 交叉檢定 (Cross-validation)

模型評估方法中，交叉檢定法優於其他的方法，因為其他的評估法並無法指出這個學習演算模型對一筆新進的未知資料進行預測分析時會有怎樣的表現。克服這個問題的方法之一是：不要把全部的已知資料都拿去進行訓練，在一開始就先將它們保留下來。然後當訓練完成，再把當初保留下來的資料當作「新的資料」來測試這個模型。這就是交叉檢定進行評估模型的基本概念。

交叉檢定的方法由 Seymour Geisser(1929-2004) 提出。當樣本資料較少，且進一步的樣本資料難以取得的情況下(無法蒐集、花費昂貴、或是有危險性等狀況) ，交叉檢定更顯得顯得重要。

在交叉檢定中，用於初始分析的樣本子集資料稱為訓練集(Training Set)。例如資料探勘、人工智慧等領域，系統利用這些訓練資料建立模型，這個步驟就稱為「訓練(Training)」。而剩下用來對初始分析進行檢定的樣本子集(當作「未知的新資料」)則稱為檢定集(Validation Set)，或測試集(Testing Set)。

2.3.1.1. 保持交叉檢定法 (Holdout Cross-validation)

保持交叉檢定法(Holdout Cross-validation)是最簡單的交叉檢定法。原始的資料集被隨機劃分成兩個獨立的集合，也就是所謂的訓練集與測試集。只使用訓練集的資料來產生模型，接著對測試集預測輸出值，再利用原本已知的答案來評估正確性，如圖2-10。這個方法的優點是提供模型實地應用的表現成效，然而資料的分割方法可能會影響到評估的結果，如果將訓練集與測試集資料重新劃分，評估效能有可能會大不相同。一般來說，作為測試的資料量不會超過原始樣本的三分之一。

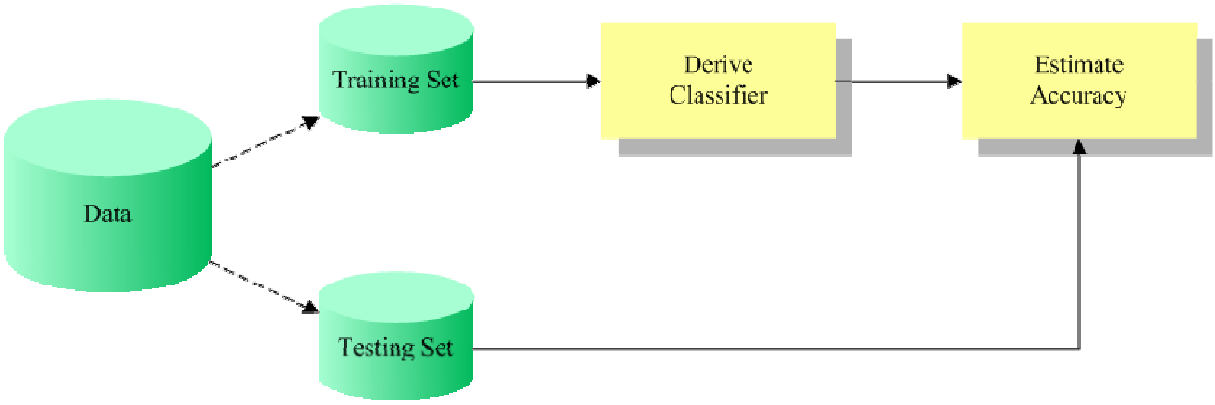

圖 2-10 用 Holdout Cross-validation 評估分類法的準確性

2.3.1.2. K-折交叉檢定(*K*-fold Cross-validation)

由於使用 holdout cross-validation 方法建立模型時只用了一部份的初始資料，因此

評估的結果較為保守，*k*-fold cross validation 可以用來改進 holdout cross-validation。將樣本資料分為 *k* 個子集，然後重複 *k* 次 holdout cross-validation，在每一次進行中，選這 *k* 個子集的其中一個作為測試集，其餘的 *k*-1 個子集作為訓練集，取這 *k* 次結果的平均作為整體評估。這個方法的優點是比較不會受到資料分割方式的影響，每一筆資料都當了一次的測試資料以及 *k*-1 次的訓練資料，當 *k* 值越大則結果變異度越小。而這個方法的缺點就是演算法必須重複執行 *k* 次，也就得花上 *k* 倍的計算量。此外這個方法還有一個變形：隨機將資料分為訓練與測試資料 *k* 次，這個方法的優點是使用者可以自由的將訓練集大小與測試次數分開來考慮。

2.3.2. 評估方法(Evaluation Metric)

靈敏性(Sensitivity)與明確性(Specificity)是常用於醫學檢測的評估方法，也可用來評量二元分類器(Binary Classifier)的效能。假設我們要對一群人進行某種疾病的檢驗，當有些人患病且檢驗結果呈陽性反應，則稱為真-正例(True Positive)；有些人患病，但檢驗為陰性反應，則稱為偽-反例(False Negative)；有人並未患病，且檢驗結果亦為未患病，稱為真-反例(True Negative)；最後偽-正例就是指那些健康卻被檢驗成有患病的人。因此，真-正例、偽-反例、真-反例與偽-正例(False Positive)的總和為樣本資料。

靈敏性的計算公式如(2.24)，由上述例子來看，靈敏性就是「所有患病的人中被檢驗出來的比例」。靈敏性為 100%則代表所有病患都被檢驗出來了，或是由工廠品質控制的角度來看，所有不良品都被找出來，避免外流至市場中。然而單靠靈敏性並不能獲得完整的檢測資訊，若將樣本全都標示為正(陽性)亦可獲得百分之百的靈敏性，因此必須與明確性配合。

明確性的計算公式如(2.25)，也就是「所有健康的人中檢測結果呈陰性反應的比例」。明確性越高，則越少健康的人被誤判為病患。由工廠品管的角度來看，就是避免把可以出售獲利的良品誤判為瑕疵品。同樣的，也不能排除靈敏性來單獨看明確性，若要獲得 100%的明確性，只需要將所有測試資料都標示為反(陰性)即可達成。

$$Sensitivity = \frac{number\ of\ true\ positives}{number\ of\ true\ positives + number\ of\ flase\ negatives} \tag{2.24}$$

$$Specificity = \frac{number\ of\ true\ negatives}{number\ of\ true\ negatives + number\ of\ flase\ positives} \tag{2.25}$$

除了靈敏性與明確性之外，亦可以用正負預測值(Positive and Negative Predictive Values)來計算二元分類法的效能。正預測值可以回答以下問題;「當我的檢定結果是陽性反樣，那我已經患病的可能性有多大？」計算方式如(2.26)，所有標示為 positive 的結果中，真-正例的機率。負預測值亦同，只是替換為 negative。

$$Positive\ predictive\ value = \frac{number\ of\ true\ positives}{number\ of\ true\ positives + number\ of\ flase\ positives} \tag{2.26}$$

然而值得注意的是，以上兩種概念之間有一個很重要的差異：因為靈敏性和明確性不會因為正反例的比例而改變，靈敏性和明確性和總體(Population)是互相獨立的。更確切的說，只需要測試正例便可以計算出靈敏性，反之預測值就需要考慮總體。

舉例來說明，假設某個對疾病的檢測，其靈敏性及明確性皆為 99 %。對 2000 個人進行檢測，且其中有 1000 個人患病，則可能的檢驗結果為 990 個真-正例、990 個真-反例，且偽-正例與偽-反例皆為 10 個。而正預測值與負預測值亦皆為 99%，則民眾可以信任這個檢測結果。

然而在同樣的條件下，若 2000 個受檢者中，只有 100 個人患病，則代表真-正例s 為 99 個、偽-反例只有一個，且真-反例為 1881 個、偽-正例為 19 個。因此在 19+99 個檢測結果為陽性反應的人中，只有 99 個真的患病—這意味著當某個人進行這項檢測，若結果為陽性反應，則此人有 84%的機會是真的患病了。

同時考慮正例與反例，可計算正確率(Accuarcy)如(2.27)。此外，正預測值又稱為

準確率(Precision)，而靈敏性稱為召回率(Recall)。

$$Accuracy = sensitivity \times \frac{number\ of\ all\ postivites}{total} + specificity \times \frac{number\ of\ all\ negatives}{total}$$
$$= \frac{true\ positives + true\ negatives}{total} \tag{2.27}$$

而準確率與召回率不能單獨使用，理由是系統很容易做出高準確率、低召回率，或低準確率、高召回率的結果。為同時兼顧這兩個數據，經常再定義 F-Measure(2.28)，來比較不同系統的成效。

$$F - Measure = \frac{2 \times precision \times recall}{precision + recall} \tag{2.28}$$

三、系統設計

本章描述書籍分類系統的架構設計與各步驟所採用的方法，3.1 節介紹系統整體架構與各部分功能；3.2 節針對書籍相關資訊做前置處理(Preprocessing)；3.3 節進而將適合用於分類器的資訊進行特徵挑選的工作；3.4 節使用 SVM 分類器進行學習；3.5 節分析統計書籍詮釋資料，從中獲取有助於分類書籍的訊息；3.6 節合併 SVM 與詮釋資料的分類結果，完成書籍分類的工作。

3.1. 系統設計

比起一般文件，書籍資料具有完整且豐富的詮釋資料(Meta-Information)，清楚地介紹了書籍的來源、作者、出版社、出版日期等資訊。本論文的概念為：書籍的詮釋資料中可能隱藏了有利於進行書籍分類的特性，例如：大部分作者通常擅長撰寫某些特定類別的書籍。系統輸入書籍資料後，首先進行前置處理，將書籍資料分為詮釋資料與敘述資料(Description)。將敘述資料視為文件(Document)，進行自動分類；另方面統計分析詮釋資料。最後透過分析詮釋資料所獲得的資訊配合分類敘述資料的結果，提高分類的正確性。

本論文將書籍分類系統設定為二元分類器(Binary Classifier)，對每個類別個別產生專屬的分類器。對於新進的書籍，分別送到各個分類器進行運算，判斷是否適合歸屬於該類別。因此，本系統不限制一本書只能分類到單一類別，亦即一本書可以同時標示兩種以上的類別。

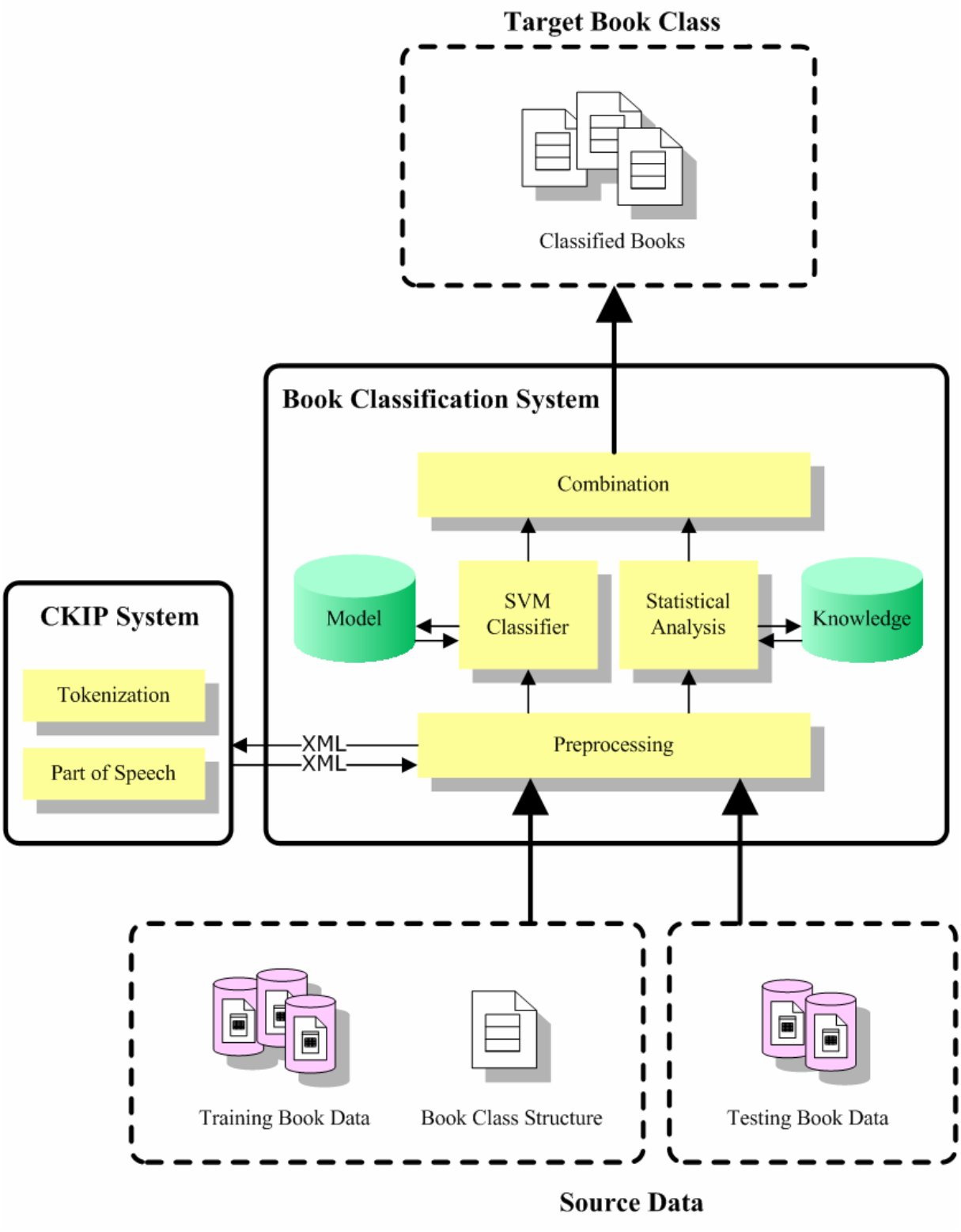

圖 3-1　圖書分類系統架構

圖 3-1 為系統架構示意圖。書籍分類系統透過 XML 與中研院 CKIP 斷詞系統溝通，對敘述資料進行斷詞切字與詞性分析，之後過濾詞性並刪除停用字，再進行特徵挑選並加入專家的經驗與知識，最後將書籍敘述資料以向量的形式呈現，經由 SVM 學習產生各類別的分類模型；另一方面，對詮釋資料進行統計分析，由現有的歷史資訊整理出作者與出版社的類別特性。完成以上準備工作後，輸入未知類別的書籍資訊，經由 SVM 模型分類，並配合詮釋資料的分類特性，決定書籍類別。

3.2. 前置處理 (Preprocessing)

分類之首要工作在於將雜亂無序的敘述資料，經過整理，成為實驗所需要的資料，因為若無此項程序，則系統較不易處理資料，且易造成分類成效不彰之結果，因此，資料前置處理便顯得格外重要。本系統所進行之前置處理包含斷詞切字、詞性標示、去除標點符號、去除非動詞名詞的單字詞、去除停用字等步驟，如圖 3-2。

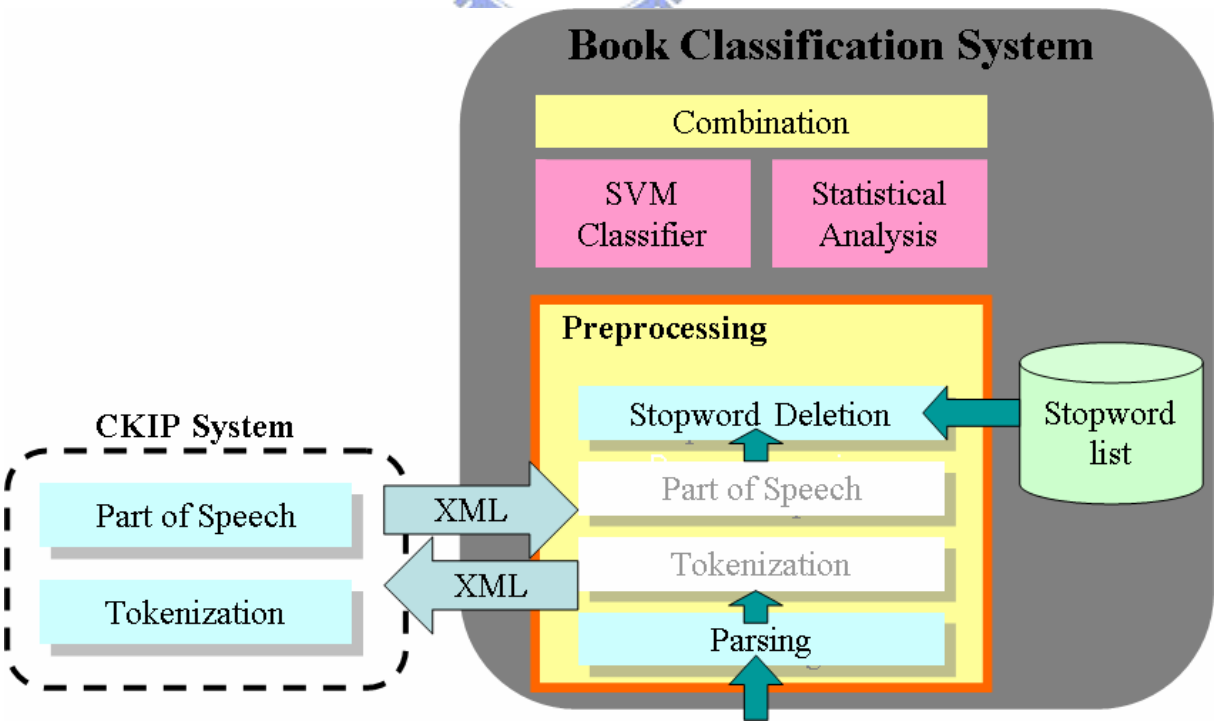

圖 3-2 前置處理流程圖

3.2.1. 斷詞切字 (Tokenization)

斷詞切字的目的在於從文字資料中擷取出含有語意的詞鍵。中文書寫系統的一大特色，就是書寫單位「字(Character)」，對應至語言單位的「詞素(Morpheme)」--比「詞(Word)」還小的單位。然而「詞彙」又是自然語言處理的一個基本單位，因此需要將文件作正確的斷詞，取得含有語意的詞彙，往後的工作才得以進行。然而中文詞的結構，有單字詞、多字詞等多種不同的型態，且中文文件中只有字的界線，詞與詞的界線不明；不像英文，英文詞通常除了極少數的片語(Phrase)外，極大多數都是一個詞(Word)，就是一個意義單位(Meaning Unit)，因此中文處理起來較英文困難。

自動斷詞大多利用詞典中收錄的詞和文本做比對，找出可能包含的詞，由於存在歧義的切分結果，因此多數的中文分詞程式多著重討論如何解決分詞歧義的問題，而較少討論如何處理詞典中未收錄的詞出現的問題(新詞如何辨認)。

本論文採用中央研究院中文詞知識庫小組(Chinese Knowledge Information Processing Group, CKIP)[12]所研發之中文斷詞系統(包含未知詞擷取與標記)[13]。由於中文詞集是一個開放集合，不存在任何一個詞典或方法可以盡列所有的中文詞，當處理不同領域的文件時，領域相關的特殊詞彙或專有名詞，常常造成斷詞系統因為參考詞彙的不足而產生錯誤的切分。為了解決這個問題，最有效的方法是補充領域詞典，加強詞彙的搜集。因此新的詞彙或關鍵詞的自動抽取成為斷詞的先期準備步驟。領域關鍵詞彙多出現在該領域的文件中而少出現在其它領域，因此抽取關鍵詞時多利用此特性。高頻的關鍵詞比較容易抽取，少數低頻的新詞不容易事先搜集，必須線上辨識。構詞律(Morphological Rule)、詞素、詞彙及詞彙共現訊息，為線上新詞辨識依據。中研院的中文斷詞系統系統提供了一個解決方案，可以自動抽取新詞建立領域用詞或線上即時斷詞功能，為一具有新詞辨識能力並附加詞類標記功能之中文斷詞系統。此一

[12] 中文詞知識庫小組(CKIP) http://rocling.iis.sinica.edu.tw/CKIP/
[13] 中文斷詞系統 http://ckipsvr.iis.sinica.edu.tw/

系統包含一個約拾萬詞的詞彙庫及附加詞類、詞頻、詞類頻率、雙連詞(Bigram)類頻率等資料。分詞依據為此一詞彙庫及定量詞(Quantifier)、重疊詞(Reiterative)等構詞規律及線上辨識的新詞，並解決分詞歧義問題。

本論文採用此系統之線上斷詞服務，使用一 API 呼叫，資料的交換方式採用 XML，經由 TCP Socket 連線傳送驗證資訊及文本至伺服器，伺服器經過處理後經由原連線傳回結果。

3.2.2. 詞性判斷 (Part of Speech, POS)

本論文所採用之斷詞系統除斷詞功能外，亦可指定輸出簡化之詞類標記。此系統的線上斷詞處理採用中央研究院資訊科學所詞庫小組所編列的中研院平衡語料庫詞類標記集之「簡化詞類」進而對照精簡成「精簡詞類」列表，詳見附錄一。

經由斷詞系統處理過的文件以 XML 格式傳回結果，對每個擷取出來的詞以括弧標示詞性，每個詞之間以全形空白隔開。

哈佛大學的宗教符號學教授羅柏・蘭登到巴黎出差的深夜， 突然接到一通緊急電話，通知他羅浮宮年高德邵的館長遭人謀殺，就在博物館內，屍體旁邊留下了一個令人困惑的密碼。蘭登與法國美女密碼專家 Sophie Neveu 在整理分析謎團的過程中，驚訝地發現在達文西的作品中藏有一連串的線索。這些線索人人可見，卻被畫家巧妙地偽裝，加以隱藏。

圖 3-3 前置處理實例 — 原文[14]

將圖 3-3 的原文輸入 CKIP 中文斷詞系統之後，進行斷詞處理並標示詞類，輸出結果如圖 3-4 所示。

[14] 摘錄自 博客來網路書店《達文西密碼》內容簡介

哈佛(N) 大學(N) 的(T) 宗教(N) 符號(N) 學(Vt) 教授(N) 羅柏・蘭登(N) 到(Vt) 巴黎(N) 出差(Vi) 的(T) 深夜(N) ，(COMMACATEGORY)

突然(ADV) 接到(Vt) 一通(N) 緊急(Vi) 電話(N) ，(COMMACATEGORY)

通知(Vt) 他(N) 羅浮宮(N) 年高德邵(Vi) 的(T) 館長(N) 遭(P) 人(N) 謀殺(Vt) ，(COMMACATEGORY)

就(ADV) 在(Vt) 博物館(N) 內(N) ，(COMMACATEGORY)

屍體(N) 旁邊(N) 留下(Vt) 了(ASP) 一個(N) 令(Vt) 人(N) 困惑(Vi) 的(T) 密碼(N) 。(PERIODCATEGORY)

蘭登(N) 與(C) 法國(N) 美女(N) 密碼(N) 專家(N) Sophie(FW) Neveu(FW) 在(ADV) 整理(Vt) 分析(Vt) 謎團(N) 的(T) 過程(N) 中(POST) ，(COMMACATEGORY)

驚訝(Vt) 地(N) 發現(Vt) 在(Vt) 達文西(N) 的(T) 作品(N) 中(POST) 藏有(Vt) 一連串(A) 的(T) 線索(N) 。(PERIODCATEGORY)

這些(DET) 線索(N) 人人(N) 可見(ADV) ，(COMMACATEGORY)

卻(ADV) 被(P) 畫家(N) 巧妙(Vi) 地(N) 偽裝(Vt) ，(COMMACATEGORY)

加以(ADV) 隱藏(Vt) 。(PERIODCATEGORY)

圖 3-4 前置處理實例 — 斷詞切字與標示詞性結果

中文的意義單位，在文言文中很多都是單字詞，所以在當時，字(Character)確實是意義單位；但在現代的中國語文當中，單字詞的型態已經很少，現代的中國語文慢慢發展以雙字詞為主的一種語言型態，因此針對單字詞的部分，本論文只保留詞性為名詞(Nouns)與動詞(Verbs)的單字詞，其餘的單字詞予以刪除。此外，亦將標點符號等不具有語意的詞刪除。圖 3-5 為圖 3-4 的內容去除標點符號、外文及非名詞、動詞的單字詞之後的結果。

哈佛(N)　大學(N)　宗教(N)　符號(N)　學(Vt)　教授(N)　羅柏・蘭登(N)　到(Vt)　巴黎(N)　出差(Vi)　深夜(N)

突然(ADV)　接到(Vt)　一通(N)　緊急(Vi)　電話(N)

通知(Vt)　他(N)　羅浮宮(N)　年高德邵(Vi)　館長(N)　人(N)　謀殺(Vt)

在(Vt)　博物館(N)　內(N)

屍體(N)　旁邊(N)　留下(Vt)　一個(N)　令(Vt)　人(N)　困惑(Vi)　密碼(N)

蘭登(N)　法國(N)　美女(N)　密碼(N)　專家(N)整理(Vt)　分析(Vt)　謎團(N)　過程(N)

驚訝(Vt)　地(N)　發現(Vt)　在(Vt)　達文西(N)　作品(N)　藏有(Vt)　一連串(A)　線索(N)

這些(DET)　線索(N)　人人(N)　可見(ADV)

畫家(N)　巧妙(Vi)　地(N)　偽裝(Vt)

加以(ADV)　隱藏(Vt)

圖 3-5　前置處理實例 — 初步篩選

3.2.3. 停用字 (Stopword)

停用字指的是在文章中沒有語意但是可以用來平順語意的詞，通常包括介系詞、指示代名詞、連詞、助詞等，如：的、是、之、我們。某些停用字在文件資料中出現的頻率極高，若以頻率計算字彙重要程度的話，有些停用字會因此突顯出來，但是這與語意的豐富與否並沒有關係，因此將它們歸納於停用字一覽表(Stopword List)中，在前置處理中需要先過濾掉，以達到清理雜訊的目的。本系統參考 Oracle Text Reference[15]

[15] Orcal Text Reference
http://www.utexas.edu/its/unix/reference/oracledocs/v92/B10501_01/text.920/a96518/astopsup.htm#45728

並加以補強，設置繁體中文的停用字一覽表，共 90 個停用字。停用字的擇定一來不可太寬鬆，以免降低分類的成效，但又不能太少，以免遺漏重要的資訊，亦會影響分類結果。圖 3-6 為部分停用字的範例。

目前	由於	因此	他們	可能	沒有	希望
有關	不過	可以	如果	對於	因為	是否
但是	相當	其中	其他	雖然	我們	包括
必須	以上	之後	所以	以及	許多	最近
至於	一般	不是	不能	而且	引起	如何
除了	不少	最後	就是	分別	加強	甚至
繼續	另外	共同	只有	了解	根據	已經
過去	所有	不會	以來	任何	一直	不同
進入	並不	據了解	現在	只是	需要	原因
只要	否則	並未	什麼	如此	不要	…

圖 3-6　停用字範例

將圖 3-5 初步去除標點符號與非名詞、動詞之單字詞的結果，配合停用字一覽表可進一步將圖 3-3 之原文精簡為圖 3-7。

哈佛(N) 大學(N) 宗教(N) 符號(N) 學(Vt) 教授(N) 羅柏・蘭登(N) 到(Vt) 巴黎(N) 出差(Vi) 深夜(N)

突然(ADV) 接到(Vt) 一通(N) 緊急(Vi) 電話(N)

通知(Vt) 他(N) 羅浮宮(N) 年高德邵(Vi) 館長(N) 人(N) 謀殺(Vt)

博物館(N) 內(N)

屍體(N) 旁邊(N) 留下(Vt) 一個(N) 令(Vt) 人(N) 困惑(Vi) 密碼(N)

蘭登(N) 法國(N) 美女(N) 密碼(N) 專家(N)整理(Vt) 分析(Vt) 謎團(N) 過程(N)

驚訝(Vt) 發現(Vt) 達文西(N) 作品(N) 藏有(Vt) 一連串(A) 線索(N)

這些(DET) 線索(N) 人人(N) 可見(ADV)

畫家(N) 巧妙(Vi) 偽裝(Vt)

加以(ADV) 隱藏(Vt)

圖 3-7 前置處理實例 — 刪除停用字

3.3. 特徵挑選 (Feature Selection)

特徵挑選一直是影響著分類效率的一項重要環節，由於構成敘述資料的詞彙數量相當龐大，換句話說，即表現敘述資料的向量之維度(Dimension)相當大，但真正重要的詞彙卻是少數，因此，若無經過特徵選取的過程，則原始的敘述向量會產生許多的冗餘資料(Redundancy)。特徵選取對系統最主要的目的有兩個，第一，節省運算時間、增進系統效率。第二，使敘述向量能更具體地代表該文件之意義。由於敘述資料中，普遍存在許多詞彙對於敘述資料整體的意義並無太大的影響，若剔除這些對敘述資料分類影響力小的詞彙，對敘述資料表現其意義時，並無太大影響，但卻可省去大部分的運算量與使用空間。因此，良好的特徵選取程序，為鞏固精確分類準確率的第一要

務。

特徵挑選方法的種類繁多，參考比較之後，本論文選用最簡便的 TF 配合 Log Likelihood Ratio (LLR)作為選取特徵的方法，再加入專家的經驗作最後把關的工作，使特徵的選取更為精準。

3.3.1. TF

如 2.2.1 節所介紹，TF 很直接地表達詞彙的出現頻率。若統計樣本已經過適當的前置處理，將不具語意的詞彙排除，則透過 TF，每個類別常用的詞彙一目了然。

然而前置處理的步驟只能大略篩選較不具語意的詞彙，為了避免誤刪有用的資訊，停用字列表僅列出明顯不具代表性的詞彙，因此光靠 TF 稍嫌不足，仍需要其他挑選法補強。

表 3-1 為實驗中「偵探/懸疑小說」、「科幻/奇幻小說」與「愛情文藝小說」類別以 TF 作為排序標準之前 10 名的詞彙，可以明顯觀察出仍有數個詞彙語意不明顯，無法代表這三個類別。

表 3-1　特徵選取實例 — TF

偵探/推理小說		科幻/奇幻小說		愛情文藝小說	
詞彙	***TF***	**詞彙**	***TF***	**詞彙**	***TF***
人	941	人	700	愛情	910
小說	525	世界	511	人	811
上	520	上	432	愛	550
推理	424	自己	307	自己	495
書	353	小說	295	故事	427
發現	340	故事	291	上	376
地	301	奇幻	256	地	201
一個	291	一個	254	說	292
自己	282	人類	252	心	286
故事	280	書	244	想	286

3.3.2. Log Likelihood Ratio (LLR)

根據[24]的概念：可以從一個事先分類好的文件集中選出某個目標概念高度相關的一些詞彙。基於「相關的詞彙傾向同時出現」的假設，可以利用 χ^2-test 或其他統計測試及資訊理論的測量方法，從事先分類好的文稿(Text)中建立主題特徵(Topic Signature)。根據[14]，LLR Λ 比 χ^2-test 更適用於稀疏資料且-2logΛ 的分佈與 χ^2 相近，因此本研究使用 LLR (-2logΛ)作為挑選特徵的參考。

令 S_i 為訓練資料中屬於類別 C_i 的書籍資料，$\overline{S_i}$ 為訓練資料中不屬於類別 C_i 的書籍資料。對於類別 C_i 與詞彙 t_j，有以下兩個 hypothesis：

Hypothesis 1 (H_1)：$P(S_j \mid t_i) = p = P(S_j \mid \overline{t_i})$，文件之間是否有相關性，與 t_i 沒有關係；

Hypothesis 2 (H_2)：$P(S_j \mid t_i) = p_1 \neq p_2 = P(S_j \mid \overline{t_i})$，$t_i$ 的存在對文件之間的相關性有很大的影響，因此 $p_1 \gg p_2$；

以下為各種可能情況的列表：

表 3-2 詞彙與類別關係狀況列表

	S_j	$\overline{S_j}$
t_i	O_{11}	O_{12}
$\overline{t_i}$	O_{21}	O_{22}

其中 O_{11} 是 t_i 在 S_j 中的出現頻率(次數)；O_{12} 是 t_i 在 S_j 以外，其他訓練資料中的出現頻率；O_{21} 是 S_j 中所有非 t_i 的詞彙的出現頻率；O_{22} 是所有非 t_i 的詞彙在 S_j 以外，其他訓練資料中的出現頻率。

舉例來看，假設考慮的詞彙是「睡覺」，類別是「偵探小說」，則 O_{11} 是偵探類裡的「睡覺」出現過的次數；O_{12} 是「睡覺」在其他的類別裡出現過的次數；O_{21} 是偵

探類裡「睡覺」除外的其他的詞彙所出現過的總次數；O_{22} 是其他類別裡，「睡覺」除外，其他詞彙所出現過的總次數。

令 tf_{t_i} 為 t_i 在整個文件集中的出現次數，tf_{t_i,S_j} 為 t_i 在 S_j 中的出現次數，N 為整個文件集裡所有詞彙出現的總次數，N_{S_j} 為 S_j 中所有詞彙出現的總次數。則四種狀況可以代換為(3.1)：

$$\begin{aligned} O_{11} &: tf_{t_i,S_j} \\ O_{12} &: tf_{t_i} - tf_{t_i,S_j} \\ O_{21} &: N_{S_j} - tf_{t_i,S_j} \\ O_{22} &: N - N_{S_j} - (tf_{t_i} - tf_{t_i,S_j}) \end{aligned} \tag{3.1}$$

假設機率分布是二項式分佈(Binomial Distribution)：

$$b(k;n,x) = \binom{n}{k} x^k (1-x)^{(n-k)} \tag{3.2}$$

可得 H_1 與 H_2 之 likelihood 為(3.3)：

$$\begin{aligned} L(H_1) &= b(O_{11};O_{11}+O_{12},p)\,b(O_{21};O_{21}+O_{22},p) \\ L(H_2) &= b(O_{11};O_{11}+O_{12},p_1)\,b(O_{21};O_{21}+O_{22},p_2) \end{aligned} \tag{3.3}$$

則-2logΛ 可以用以(3.4)計算：

$$\begin{aligned} &-2\log\Lambda \\ &= -2\log\frac{L(H_1)}{L(H_2)} \\ &= -2\log\frac{b(O_{11};O_{11}+O_{12},p)\,b(O_{21};O_{21}+O_{22},p)}{b(O_{11};O_{11}+O_{12},p_1)\,b(O_{21};O_{21}+O_{22},p_2)} \\ &= -2((O_{11}+O_{21})\log p + (O_{12}+O_{22})\log(1-p) \\ &\quad -(O_{11}\log p_1 + O_{12}\log(1-p_1) + O_{21}\log p_2 + O_{22}\log(1-p_2))) \end{aligned} \tag{3.4}$$

配合(3.1)與(3.4)對每個詞彙計算，可以求得每個詞彙的-2logΛ，高低排序找出與類別相關性較高的詞彙。

表 3-3 為實驗中三個類別以 LLR(-2logΛ)作為排序標準之前 10 名詞彙。

表 3-3　特徵選取實例 —LLR

偵探/推理小說		科幻/奇幻小說		愛情文藝小說	
詞彙	$-2\log\Lambda$	詞彙	$-2\log\Lambda$	詞彙	$-2\log\Lambda$
推理	1071.35	奇幻	707.49	愛情	2188.00
愛情	922.05	愛情	541.41	愛	1184.70
偵探	659.12	推理	440.76	男人	429.39
福爾摩斯	590.64	人類	401.09	幸福	377.12
愛	478.47	世界	381.08	推理	357.01
羅蘋	475.59	愛	293.67	女人	353.46
兇手	455.42	魔法	273.10	心	312.75
犯罪	322.28	吸血鬼	263.03	偵探	281.21
殺人	320.58	科幻	255.83	感情	236.00
奇幻	284.09	地球	241.64	戀愛	230.73

接著配合 TF 與-2logΛ，以(3.5)計算每個詞彙對類別的代表性，值越高則代表性越高，實驗中前十名的結果列於表 3-4。

$$-2\log\Lambda \times TF \tag{3.5}$$

表 3-4　特徵選取實例 —LLR×TF

偵探/推理小說		科幻/奇幻小說		愛情文藝小說	
詞彙	$-2\log\Lambda \times TF$	詞彙	$-2\log\Lambda \times TF$	詞彙	$-2\log\Lambda \times TF$
推理	454254.30	世界	194732.45	愛情	1991069.11
偵探	181257.91	奇幻	181118.69	愛	651585.24
福爾摩斯	128168.04	人類	101075.83	男人	104342.86
兇手	95636.47	愛情	42771.39	女人	91192.02
羅蘋	81802.03	魔法	36868.11	心	89445.34
殺人	58986.36	冒險	33157.36	自己	81973.05
小說	45250.50	傳說	31994.54	幸福	69768.03
犯罪	42541.36	地球	25855.45	想	52614.18
愛情	41492.35	吸血鬼	24987.97	故事	43547.01
事件	69616.51	科幻	23024.74	說	29635.23

3.3.3. 專家挑選

透過計算(3.5)選出對類別代表性較高的詞彙之後製成「候選特徵列表」，再加入專家的智慧，讓特徵選取更精準。

專家的工作有兩項，首先是刪除不必要的詞彙，雖然前置處理已經做了很多篩選過濾的動作，但語言文字變化性極大，光靠停用字列表難免有遺漏疏失，使得候選特徵列表中出現一些語意不足但在計算上佔有優勢的詞彙，因此在這裡做最後把關的動作，讓專家將對候選特徵進行篩選。

接著是依專家的經驗，加入類別相關性非常高、非常具有代表性的詞彙，表 3-5 專家對「偵探/懸疑小說」、「科幻/奇幻小說」及「愛情文藝小說」三個類別加入重要詞彙的例子，完整列表詳見附錄二。

表 3-5　特徵選取實例 — 專家加入之類別相關詞彙

偵探/推理小說	科幻/奇幻小說	愛情文藝小說
偵探	哈利波特	分手
動機	女巫	邂逅
密室	科幻	寂寞
推理	精靈	浪漫
殺人	召喚	心愛
懸疑	衛斯理	廝守
陰謀	外星人	愛情
意外	地球人	失戀
綁架	法術	交往
福爾摩斯	托爾金	深情

3.4. SVM 分類法

將書籍敘述資料轉為特徵向量表示式，再經由 SVM 學習產生分類模型，利用模型對測試集進行分類。

3.4.1. 向量表示式

決定特徵之後，要將書籍的敘述資料轉換為適合 SVM 的向量表示式。首先以二元(Binary)的方式來呈現，若敘述資料中有出現這個特徵就標示為 1，否則為 0。接著進一步對經由專家指定加入的特徵給予適當加權。

表 3-6、圖 3-8 及圖 3-9 為向量表示式轉換的實例。表 3-6 為特徵列表，包含八個由專家指定的特徵，及八個提系統自動挑選並經由專家過濾的特徵；圖 3-8 為前置處理實例的延伸，將經由前置處理後保留下來的詞彙與特徵列表進行比對，以外框標示出符合的詞彙，並且以灰底標示為專家特選的特徵；假設專家挑選的特徵權重為 3，則圖 3-8 內容文字配合表 3-6 特徵類表，可以轉換為圖 3-9 向量表示式。

表 3-6　向量表示式實例 — 特徵列表

專家指定的特徵							
屍體	動機	迷團	密碼	偵探	密室	推理	殺人
系統自動挑選的特徵							
線索	教授	政府	追查	一連串	謀殺	刑警	發現

哈佛　大學　宗教　符號　學　教授　羅柏・蘭登　到　巴黎　出差　深夜
突然　接到　一通　緊急　電話
通知　他　羅浮宮　年高德邵　館長　人　謀殺
博物館　內
屍體　旁邊　留下　一個　令　人　困惑　密碼
蘭登　法國　美女　密碼　專家整理　分析　謎團　過程
驚訝　發現　達文西　作品　藏有　一連串(A)　線索
這些　線索　人人　可見
畫家　巧妙　偽裝
加以　隱藏

圖 3-8　向量表示式實例 — 比對敘述資料

<3, 0, 3, 3, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1, 0, 1>

圖 3-9　向量表示式實例 — 轉換結果

3.4.2. SVM 分類器

本研究採用的分類方法為 Support Vector Machine，使用的工具是 SVM*light*[16][40]。首先將訓練資料經由前置處理及特徵挑選，轉換為向量表示式，送入 SVM 分類器學習，產生分類模型；再配合測試資料的向量表示式與分類模型進行分類。圖 3-10 為 SVM 學習及分類之流程示意圖。

SVM 以計算空間中最大邊界(Margin)的做法確實同時盡可能避免了過份最佳化(Overfitting)訓練資料的問題，且同時確實利用了所有的統計資料，儘管它的成效確實較之前所有統計方法都來得更好，但面對新資料時，特徵可能不足的問題依然存在，因此本論文嘗試以加入書籍詮釋資料的方式提升分類準確性。

[16] SVM*light* http://svmlight.joachims.org/

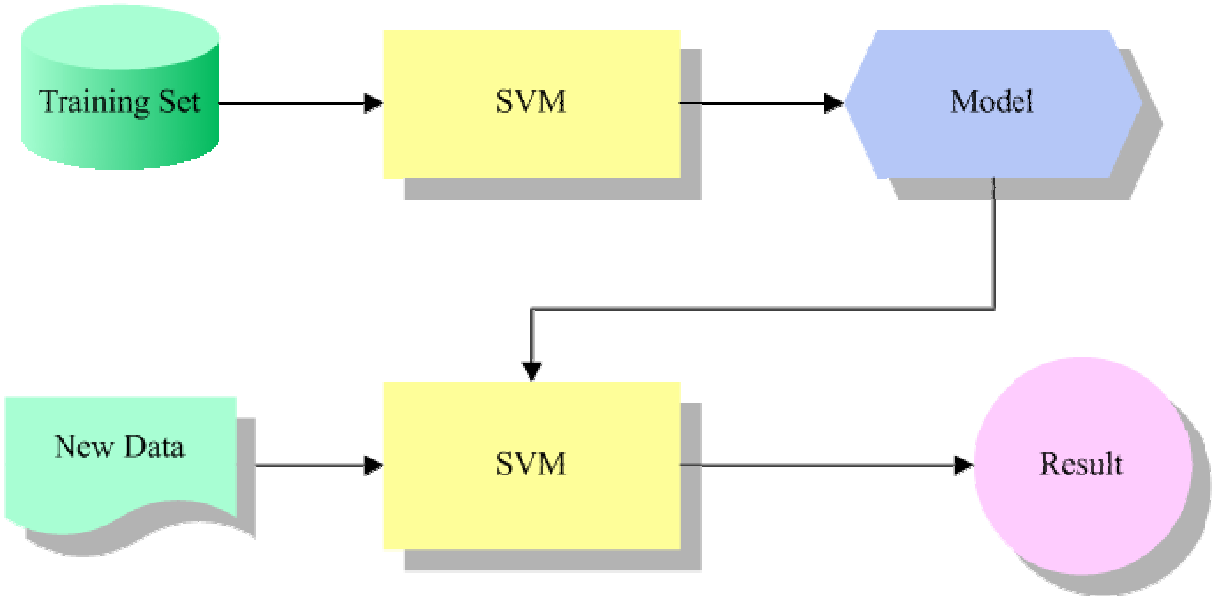

圖 3-10　SVM 分類流程圖

3.5. 詮釋資料

本論文嘗試從現有的詮釋資料進行統計分析，希望藉由這些出版的歷史資訊，獲得有利於書籍分類的情報。本系統所採用的詮釋資料有「作者」與「出版社」兩項。

在作者部分，其中主要的概念是：「一般作家不會橫跨太多的領域」，若有一位作家在某個類別有很多著作，可判斷他對這個類別的代表很高，當這位作家有新的著作，便能合理推敲這一本新書屬於同類別的機會很大；若一位作者在此類別沒有著作，而在其他類別有相當多的出書紀錄，則他出的新書在這個類別的機會也不高。

圖 3-11 為計算作者與類別之間相關資訊的流程圖，其中 $\left|A_{j,C_i}\right|$ 代表出版社 A_j 在類別 C_i 的著作數量；$\left|A_{j,\overline{C_i}}\right|$ 為 A_j 在 C_i 的以外的所有著作數量；$\left|A_j\right|$ 是 A_j 的所有著作數量；$W(A_j,C_i)$ 即為 A_j 在 C_i 的類別相關度；而 θ_A 是一個常數。

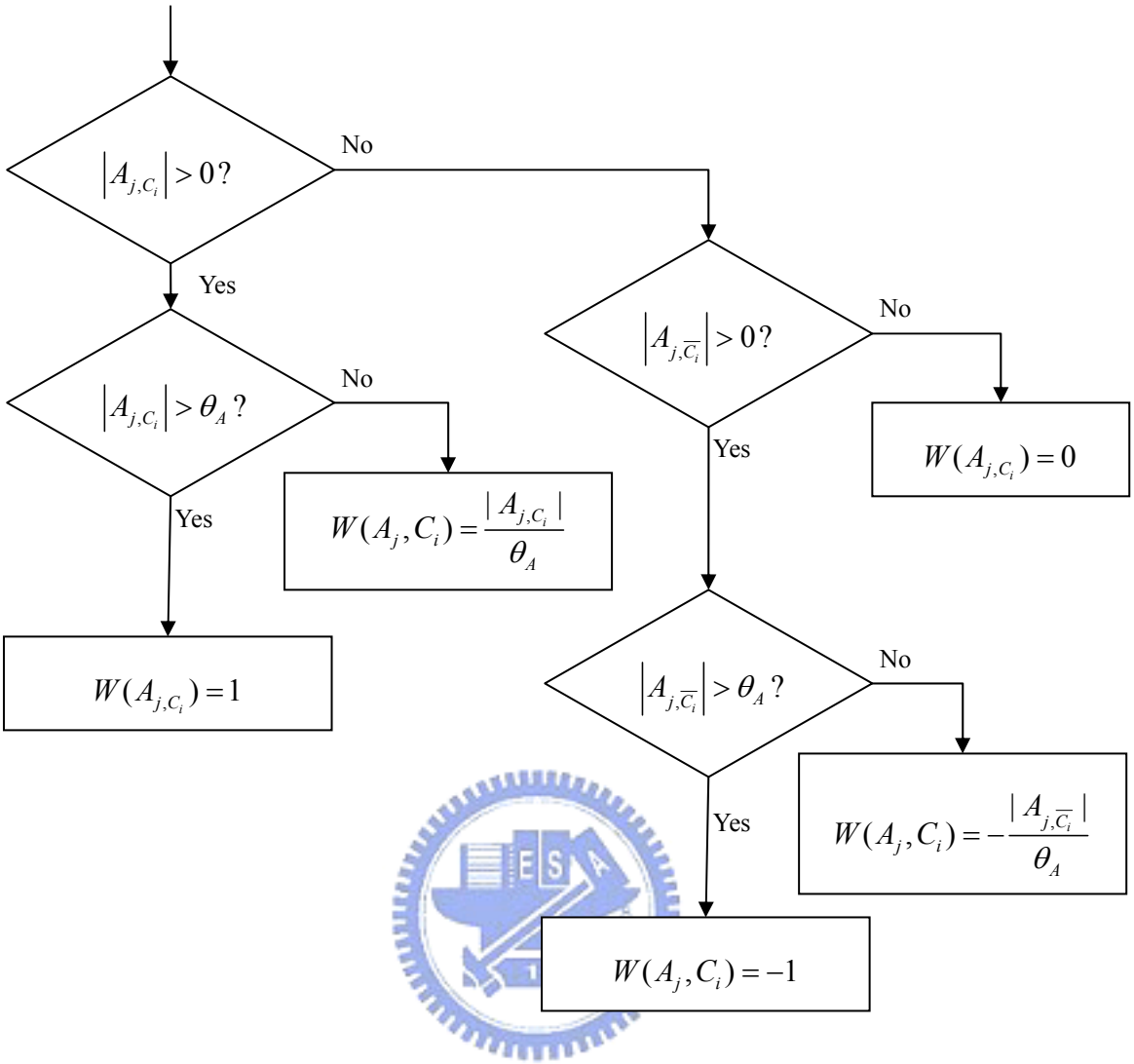

圖 3-11　作者-類別相關資訊計算流程圖

出版社資訊也是類似的流程。但出版社有規模大小之分，某些大出版社出版項目橫跨各個類別，也有一些出版社雖然規模不大，但卻專精於某個類別的出版品。例如某出版社 A 的總出版品有 5,000 冊，其中有 200 冊為類別 C，另外有一個小出版社 B，其 150 本的出版品全部都是屬於 C，則雖然 A 在 C 的出版品數量較多，本系統仍認為 B 對 C 的類別相關性較高。因此本研究以出版品的類別比例來判斷相關程度。

圖 3-12 為計算出版社與類別之間相關資訊的流程圖，其中 $|P_{k,C_i}|$ 代表出版社 P_k 在類別 C_i 的出版品數量；$|P_{j,\overline{C_i}}|$ 是 P_j 在 C_i 的以外的所有出版品數量；$|P_j|$ 為 P_j 的所有出版品數量；$W(P_j, C_i)$ 則代表 P_j 在 C_i 的的類別相關度；而 θ_P 是一個常數。

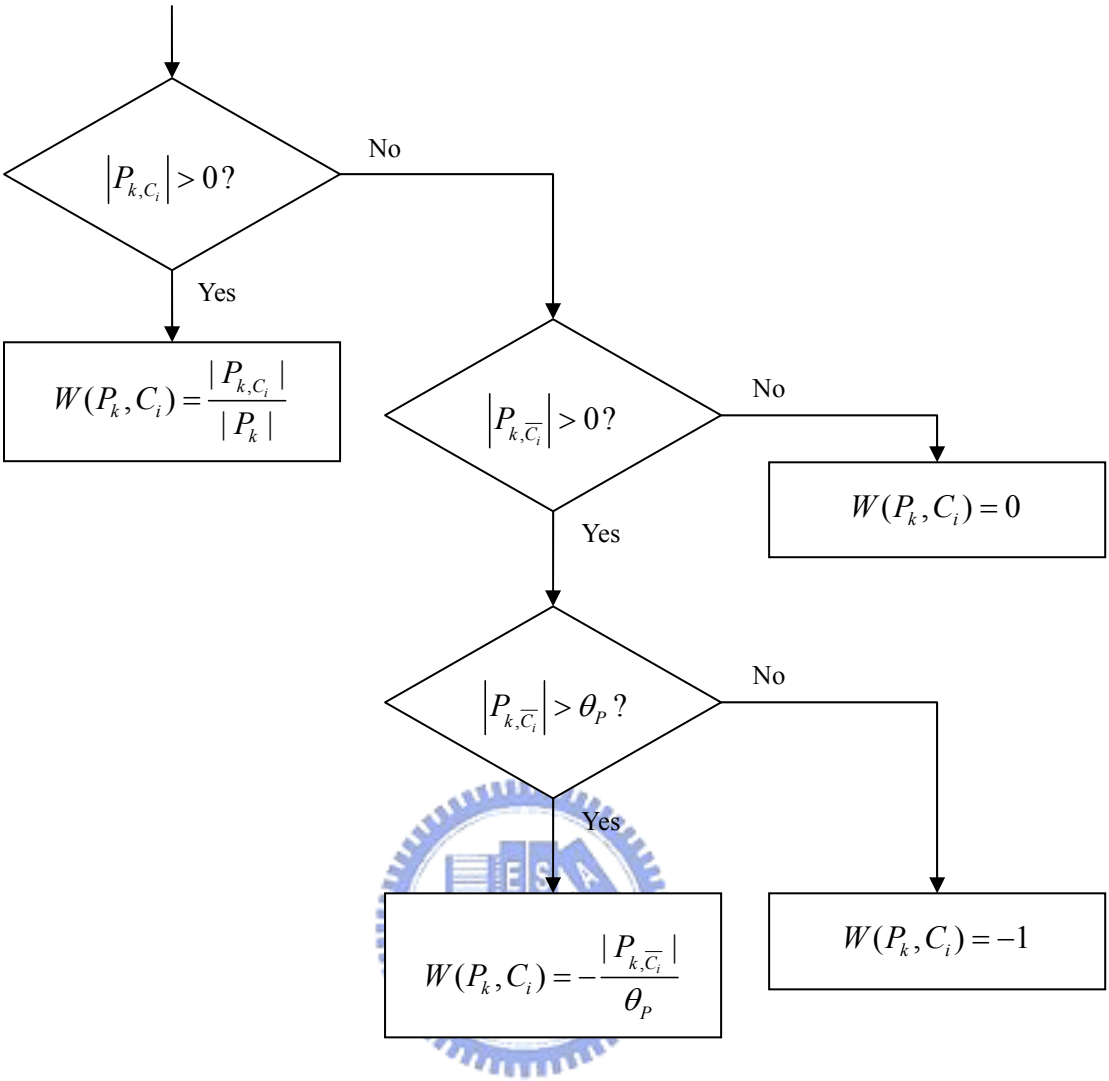

圖 3-12　出版社-類別相關資訊計算流程圖

3.6. 整合 SVM 與詮釋資料

最後將 SVM 的分類結果與出版社、作者資訊作線性組合(Linear Combination)(3.6)，其中 $W(B,A,P,C)$代表一本書 B 要分到類別 C 裡面的分數，而且這本書的作者為 A、出版社為 P；$WS(B,C)$則為 SVM 分類法將 B 代入類別 C 的學習模型所得的結果，配合作者資訊 $WA(A,C)$與出版社資訊 $WP(P,C)$，找出一組最合適的 α、β 及 γ。

$$W(B,A,P,C)=\alpha\times WS(B,C)+\beta\times WA(A,C)+\gamma\times WP(P,C) \qquad (3.6)$$

$W(B,A,P,C)>0$則將書籍 B 標示為屬於類別 C；否則$W(B,A,P,C)\le 0$，則 B 不屬於 C。

四 、 實驗與分析

本章的內容主要在評量並分析 SVM 與加入專家篩選特徵的分類效果，以及配合詮釋資料之後的表現。4.1 節說明實驗環境、實驗資料、實驗步驟，以及評估方法；4.2 節則從多個面向描述實驗結果。

4.1. 實驗環境、資料、步驟與評估方法

4.1.1. 實驗環境

進行實驗的軟硬體環境如表 4-1，PC1 用於開發程式、進行實驗與分析結果；PC2 則為資料庫伺服器。

表 4-1 實驗環境

		PC 1	PC 2
Hardware	CPU	Intel Pentium(R) 4 3.20GHz	Intel Pentium(R) 4 2.80GHz
	Memory	1.50 GB	1.00 GB
Software	OS	Microsoft Windows XP Professional	Microsoft Windows Server 2003
	Database	SQL Server	SQL Server
	Tool	Borlend C++ Builder 6	

4.1.2. 實驗資料

本論文實驗資料的來源是博客來網路書店(http://www.books.com.tw)，取「偵探/懸疑小說」、「科幻/奇幻小說」與「愛情文藝小說」三個類別的書籍資料各 900 本。表 4-2 列出各項統計資料，包含將敘述資料進行前置處理後，每個類別中平均每本書所包含的相異詞彙個數、每本書的詞彙總數，以及整個類別的相異詞彙個數、整個類別

的詞彙總數;除此之外,詮釋資料的部分,對每個類別的作者與出版社進行統計。表 4-3 為敘述資料與詮釋資料分別使用到的書籍資料欄位。

表 4-2　實驗資料

類別	數量	前置處理後的敘述資料				詮釋資料	
		冊		類別		作者	出版社
		相異詞彙	詞彙總數	相異詞彙	詞彙總數		
偵探/懸疑小說	900	84	100.05	19,571	90,046	274	82
科幻/奇幻小說	900	69	80.02	17,269	79,214	319	93
愛情文藝小說	900	75	81.39	17,517	73,249	644	122
整體統計	2,700	76	89.81	36,064	242,509	1,206	210

表 4-3　資料欄位

敘述資料	詮釋資料
書名	作者
內容簡介	出版社
作者簡介	

4.1.3. 實驗步驟

本研究分別將三個類別的書籍資料隨機分成九份,編號為 1 至 9,每一份包含 300 本書(三個類別各 100 本),進行 9-fold cross validation。實驗共進行九次,依序取其中一份作為測試集,其餘 800*3 本書則用來訓練產生分類器,再將九次實驗的數據平均作為評估結果。

實驗分為以下四部分進行:

1) 選擇適當的特徵個數進行 SVM 分類;
2) 比較加入專家智慧挑選特徵的 SVM 分類成效,並為專家指定的特徵選擇適當的權重;
3) 比較 SVM、ID3 與 Naïve Bayesian 這三種分類演算法之分類成效;

4) 加入詮釋資料，調整 SVM 分類結果與詮釋資料之適當比重。

4.1.4. 評估方法 (Evaluation Metric)

由於 Accuracy 與 F-measure 各有其盲點存在，本系統評估時採用兩者並列的方式，同時觀察兩項數值以期得到整體表現的客觀趨勢。

表 4-4 為分類結果正確性之列聯表，將作為測試集的 300 本書的分類結果與博客來網路書店對這 300 本書的分類進行比對，針對單一類別考慮「屬於」、「不屬於」兩種情況。S_C 為測試集之中由博客來網路書店斷定屬於類別 C 的書籍，$S_{\overline{C}}$ 代表不屬於 C 的書籍；R_C 與 $R_{\overline{C}}$ 分別代表測試集之中被本系統標示為屬於與不屬於 C 的書籍，O_{11}、O_{12}、O_{21} 與 O_{22} 則分別代表真-正例(True Positive)、偽-正例(False Positive)、偽-反例(False Negative)與真-反例(True Negative)。

表 4-4 分類結果列聯表

	S_C	$S_{\overline{C}}$
R_C	O_{11}	O_{12}
$R_{\overline{C}}$	O_{21}	O_{22}

配合表 4-4，可以依照(4.1)計算 Accuracy；再利用 Precision(4.2)與 Recall(4.3)求得 F-measure(4.4)。

$$Accuracy = \frac{true\ positives + true\ negatives}{total} = \frac{O_{11} + O_{22}}{O_{11} + O_{12} + O_{21} + O_{22}} \tag{4.1}$$

$$Precision = \frac{O_{11}}{O_{11} + O_{12}} \tag{4.2}$$

$$Recall = \frac{O_{11}}{O_{11} + O_{21}} \tag{4.3}$$

$$F - Measure = \frac{2 \times precision \times recall}{precision + recall} \tag{4.4}$$

4.2. 實驗結果與分析

本節從多個面向說明與分析實驗結果。4.2.1 節中先決定進行 SVM 分類時所需的特徵個數；4.2.2 節裡利用專家的智慧與經驗進行特徵挑選；4.2.3 節則比較 SVM 與其他分類演算法；4.2.4 節則說明並分析結合 SVM 與詮釋資料進行分類的實驗結果。

4.2.1. 特徵個數

為比較選取的特徵多寡對 SVM 分類結果的影響，本實驗以自動選取特徵的方式，比較從每個類別選取 50 個、100 個與 150 個特徵進行 SVM 分類對分類結果正確性的影響，由實驗結果顯示(表 4-5)，選取 150 個特徵有利於 SVM 分類。

表 4-5　選擇特徵個數

Features	偵探/懸疑小說		科幻/奇幻小說		愛情文藝小說	
	Accuracy	F-measure	Accuracy	F-measure	Accuracy	F-measure
50	89.3%	82.2%	85.3%	75.1%	89.1%	82.5%
100	90.7%	84.9%	86.3%	77.3%	89.6%	83.6%
150	91.1%	85.6%	88.1%	80.8%	89.8%	83.9%

4.2.2. 加入專家智慧挑選特徵

本實驗探討由專家進行特徵挑選對 SVM 分類結果的影響。每個類別皆由專家決定 50 個必要的特徵，並且對其他自動選取的特徵進行過濾。

由於實驗採用 9-fold cross validation，為了維持過濾特徵的一致性，先找出每個類別 $-2\log\Lambda\times TF$ 由大到小排序前 1000 名的詞彙，由專家進行刪除標記，製成「停用特徵列表」，進行實驗時統一比對此列表進行過濾刪除的動作。表 4-6 為實驗結果，「Without Expert」列代表由系統自動選取 150 個特徵且專家完全不參與特徵挑選的分類結果；其餘各列以 W_Exp 代表專家指定特徵之權重。

表 4-6　專家權重

	偵探/懸疑小說		科幻/奇幻小說		愛情文藝小說	
	Accuracy	F-measure	Accuracy	F-measure	Accuracy	F-measure
Without Expert	91.1%	85.6%	88.1%	80.8%	89.8%	83.9%
W_Exp=1	91.0%	85.5%	87.7%	79.7%	90.9%	85.7%
W_Exp=2	92.3%	84.5%	89.0%	82.1%	92.5%	87.9%
W_Exp=3	93.5%	89.5%	90.2%	83.9%	92.9%	88.4%
W_Exp=4	93.9%	90.1%	90.3%	84.0%	93.2%	88.8%
W_Exp=5	93.9%	90.1%	90.9%	85.0%	93.3%	89.0%

實驗結果顯示，若僅是讓專家介入特徵挑選，卻不將專家指定的特徵加權(W_Exp=1)，則在「偵探/懸疑小說」及「科幻/奇幻小說」類別上，SVM 分類結果較完全自動挑選的成效稍差。這是因為由專家「空降」的特徵，在以頻率為基礎的計算上並不一定佔有優勢，若不給予加權則 SVM 分類時無法凸顯其重要性；且專家刪除特徵的動作，會將原本由系統自動選取、以頻率為基礎的計算較具有代表性的特徵去除，由較不具代表性的特徵遞補，因而造成 SVM 分類的正確性下降。相對地，若將專家指定的特徵加權，則可發現對 SVM 分類的結果有正面的提升。將表 4-6 之實驗結果分別依照 Accuracy 與 F-measure 製成折線圖，如圖 4-1 與圖 4-2，可以更明顯觀察出專家指定特徵的權重與分類成果之間的關係。

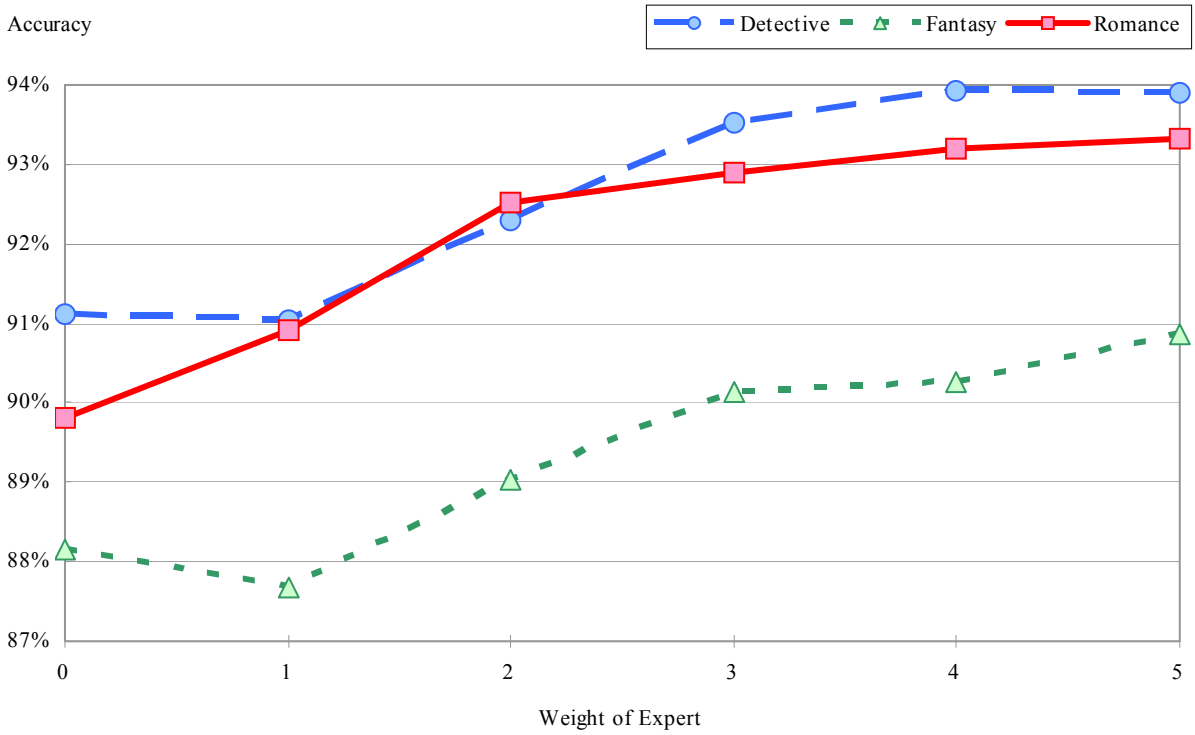

圖 4-1　專家權重 —Accuracy

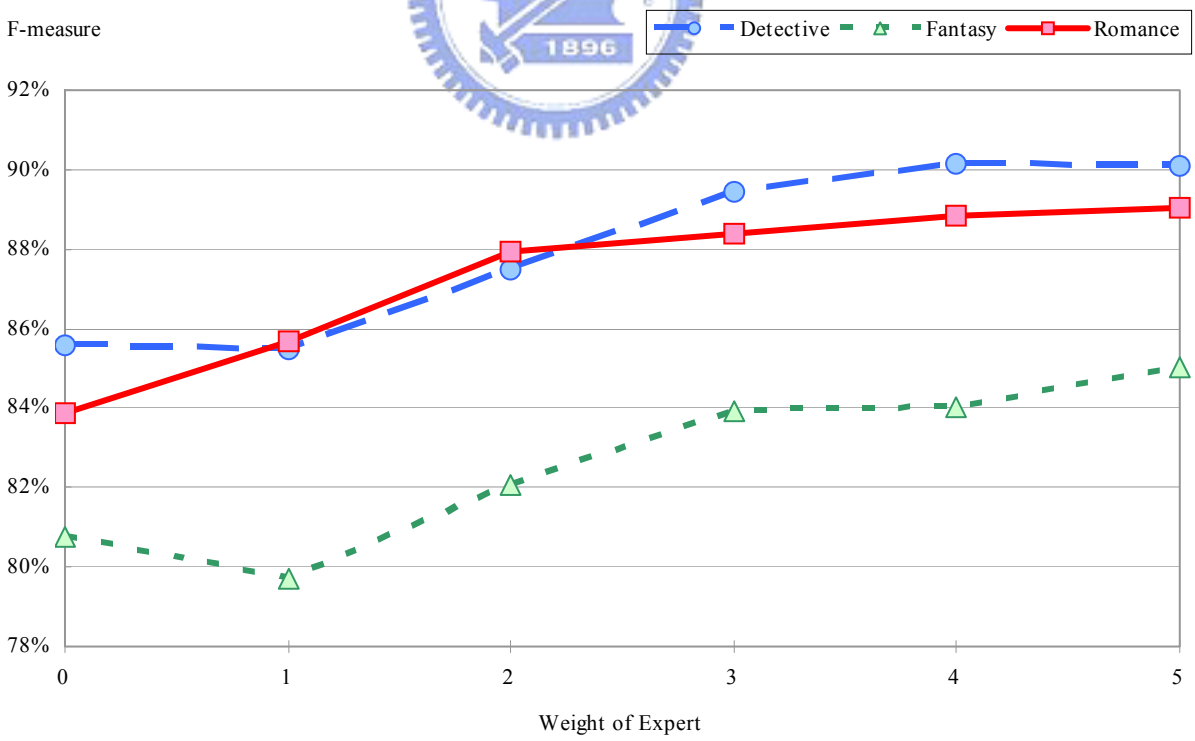

圖 4-2　專家權重 — F-measure

4.2.3. 比較分類演算法

表 4-7 比較 SVM、ID3 與 Naïve Bayesian 這三種分類演算法之分類成效，在特徵條件皆為 150 個自動挑選的特徵時，Naïve Bayesian 的分類結果最好，SVM 略差一點，ID3 明顯遜色許多。然而若與加入專家挑選結果並加權的 SVM 分類結果相比，則後者不論 Accuracy 或 F-Measure 都比較高。

表 4-7　比較分類演算法

Classification	偵探/懸疑小說		科幻/奇幻小說		愛情文藝小說	
Algorithm	Accuracy	F-measure	Accuracy	F-measure	Accuracy	F-measure
SVM	91.1%	85.6%	88.1%	80.8%	89.8%	83.9%
Naïve Bayesian	91.7%	87.0%	88.3%	82.3%	91.3%	87.5%
ID3	81.7%	74.0%	81.0%	72.0%	80.5%	71.7%
SVM (W_Exp=5)	93.9%	90.1%	90.9%	85.0%	93.3%	89.0%

4.2.4. 配合 SVM 與詮釋資料進行分類

本實驗結合 SVM 與詮釋資料進行分類，進而配合各方法實驗取得表現最好的組合方式。表 4-8 中 WS 為 SVM 之分類結果(已加入專家挑選的特徵，並將權重值設定為 5)，WA 代表作者資訊，WP 為出版社資訊。

首先分別單獨以此三種方法進行分類，藉此可觀察出各類別的資料特性，實驗結果列於表 4-8。「科幻/奇幻小說」類的 SVM 分類結果比其他兩類差，可以藉此推斷此類別的敘述資料較為分散，較不容易找出具有類別代表性的詞彙；在單以作者資訊進行分類的情況下，「愛情文藝小說」類的分類結果最差，顯示此類別有許多著作量較少的作家，或是一直有新作家；而「偵探/懸疑小說」類別在依照出版社資訊分類的結果明顯優於其他兩類，顯示此類別的出版品較具有獨佔性，有較多專門出版此類書籍的出版社。前述的討論和表 4-2 的實驗資料大致吻合：「愛情文藝小說」類的作家數量是其他類別的兩倍以上；「偵探/懸疑小說」類的出版社數量最少；而「科幻/奇幻小說」

敘述資料量最少，以致無法取得較佳的特徵。

表 4-8　比較 SVM 與詮釋資料分類結果

WS	WA	WP	偵探/懸疑小說		科幻/奇幻小說		愛情文藝小說	
			Accuracy	F-measure	Accuracy	F-measure	Accuracy	F-measure
1	0	0	93.9%	90.1%	90.9%	85.0%	93.3%	89.0%
0	1	0	90.3%	84.8%	89.5%	83.6%	77.9%	55.3%
0	0	1	70.3%	68.6%	55.9%	59.5%	56.7%	59.5%

表 4-9　SVM 配合詮釋資料分類結果

WS	WA	WP	偵探/懸疑小說		科幻/奇幻小說		愛情文藝小說	
			Accuracy	F-measure	Accuracy	F-measure	Accuracy	F-measure
1	1	0	96.9%	95.3%	95.6%	93.4%	94.6%	91.2%
1	0	1	94.5%	91.1%	91.2%	85.6%	93.4%	89.1%
0	1	1	84.4%	80.6%	80.3%	76.6%	89.1%	85.3%
1	1	1	97.0%	95.4%	95.7%	93.6%	94.6%	91.2%

表 4-9 列出搭配一種以上的方式進行分類，若僅從 SVM 分類結果、作者資訊，以及出版社資訊中組合兩種方式進行分類，則經由實驗後發現 SVM 配合作者資訊的分類結果優於其他兩種組合，因此將(3.6)之合併計算式修改為(4.5)，先調整 SVM 與作者資訊兩者之比例求得最佳 α、β，再加入出版社資訊計算 γ 與 δ 之最佳組合。

$$W(B,A,P,C)=\delta(\alpha\times WS(B,C)+\beta\times WA(A,C))+\gamma\times WP(P,C) \qquad (4.5)$$

表 4-10　SVM 配合詮釋資料組合分類結果

WS	WA	WP	偵探/懸疑小說		科幻/奇幻小說		愛情文藝小說	
			Accuracy	F-measure	Accuracy	F-measure	Accuracy	F-measure
1	0	0	93.9%	90.1%	90.9%	85.0%	93.3%	89.0%
α	β	0	97.1%	95.6%	94.7%	91.6%	95.1%	92.8%
δ×α	δ×β	γ	97.4%	96.1%	96.2%	94.5%	96.5%	94.6%

表 4-10 為配合(4.5)之實驗結果，其中 α 為 0.625，β 為 0.375，γ 為 0.05，δ 為 0.95。原本單純用 SVM 分類的正確率約為 90%上下，實驗證明，加入詮釋資料後，可將成效提高至 95%。

為了解加入詮釋資料後對 SVM 分類結果所產生的影響，本研究進一步對詮釋資料所提升的分類成果進行分析。實驗中僅針對 SVM 分類結果落於指定區間的書籍加入詮釋資料做進一步分類，其餘部分書籍直接以 SVM 輸出作為分類結果。表 4-11 為實驗結果，其中 SVM 列代表僅採用 SVM 分類結果；其餘各列為取部分書籍配合詮釋資料進行分類，Range=±0.25 代表僅取 SVM 輸出值為±0.25 之書籍套用(4.5)計算分類結果，其餘書籍不加入詮釋資料，直接以 SVM 之輸出作為分類結果；Total 列為將全部書籍配合 SVM 與詮釋資料進行分類。將表 4-11 之實驗結果分別依照 Accuracy 與 F-measure 製成折線圖，如圖 4-3 與圖 4-4 所示，觀察圖形可以發現 0~±1 區間加入詮釋資料對分類有很大的幫助，此區間之資料即為落於圖 2-8 之 Margin 內的測試資料。

表 4-11　取部分區間配合詮釋資料分類

Range	偵探/懸疑小說		科幻/奇幻小說		愛情文藝小說	
	Accuracy	F-measure	Accuracy	F-measure	Accuracy	F-measure
SVM	91.1%	85.6%	88.1%	80.8%	89.8%	83.9%
±0.25	96.2%	94.0%	92.3%	88.0%	94.1%	90.4%
±0.5	96.4%	94.4%	93.7%	90.3%	94.1%	90.5%
±0.75	96.6%	94.7%	94.3%	91.4%	94.3%	90.8%
±1	96.8%	95.0%	94.8%	92.2%	94.4%	91.0%
±1.5	97.2%	95.7%	95.3%	93.1%	94.7%	91.5%
±2	97.2%	95.8%	95.5%	93.3%	95.0%	92.0%
±3	97.2%	95.8%	95.8%	93.9%	95.4%	92.8%
±4	97.4%	96.1%	95.9%	94.1%	96.0%	93.8%
Total[17]	97.4%	96.1%	96.2%	94.5%	96.5%	94.6%

[17] 分析 SVM 分類結果，超過 80%的輸出值落於±5 區間內，因此令 SVM 輸出值大於+5 者為+5，小於−5 者為−5。

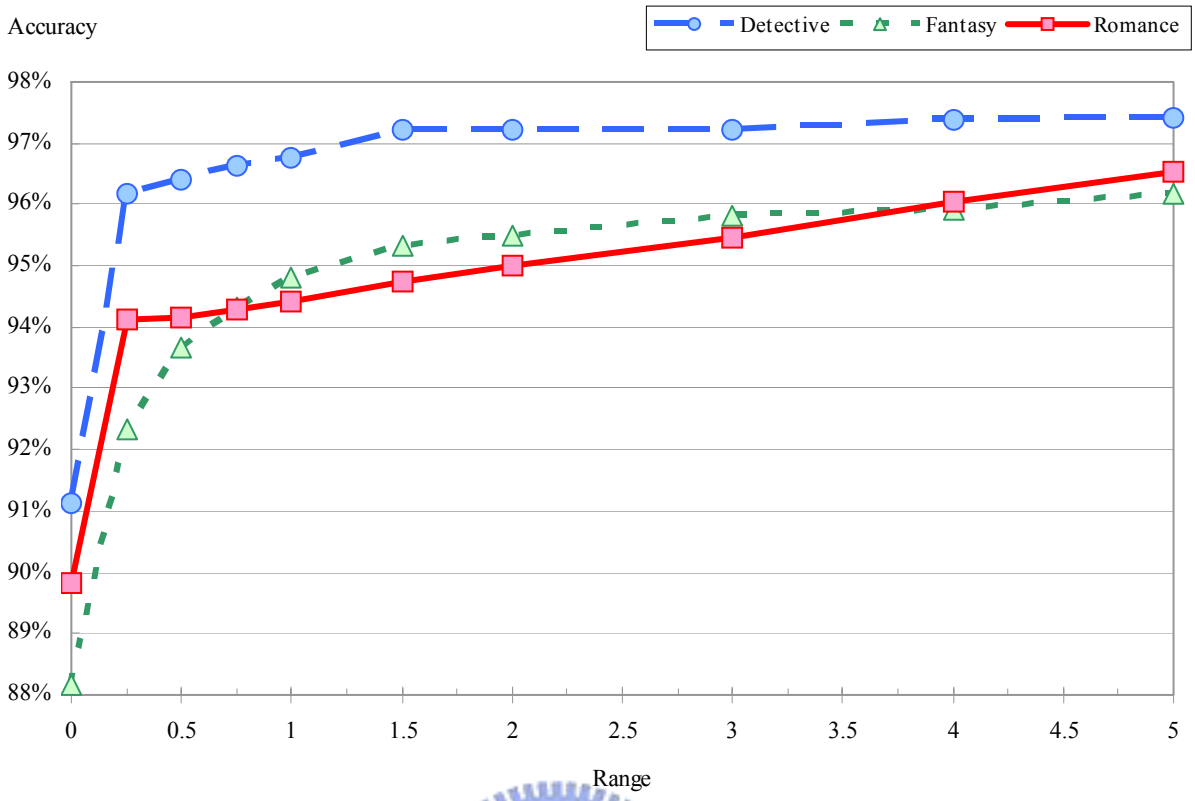

圖 4-3　取部分區間配合詮釋資料分類 — Accuracy

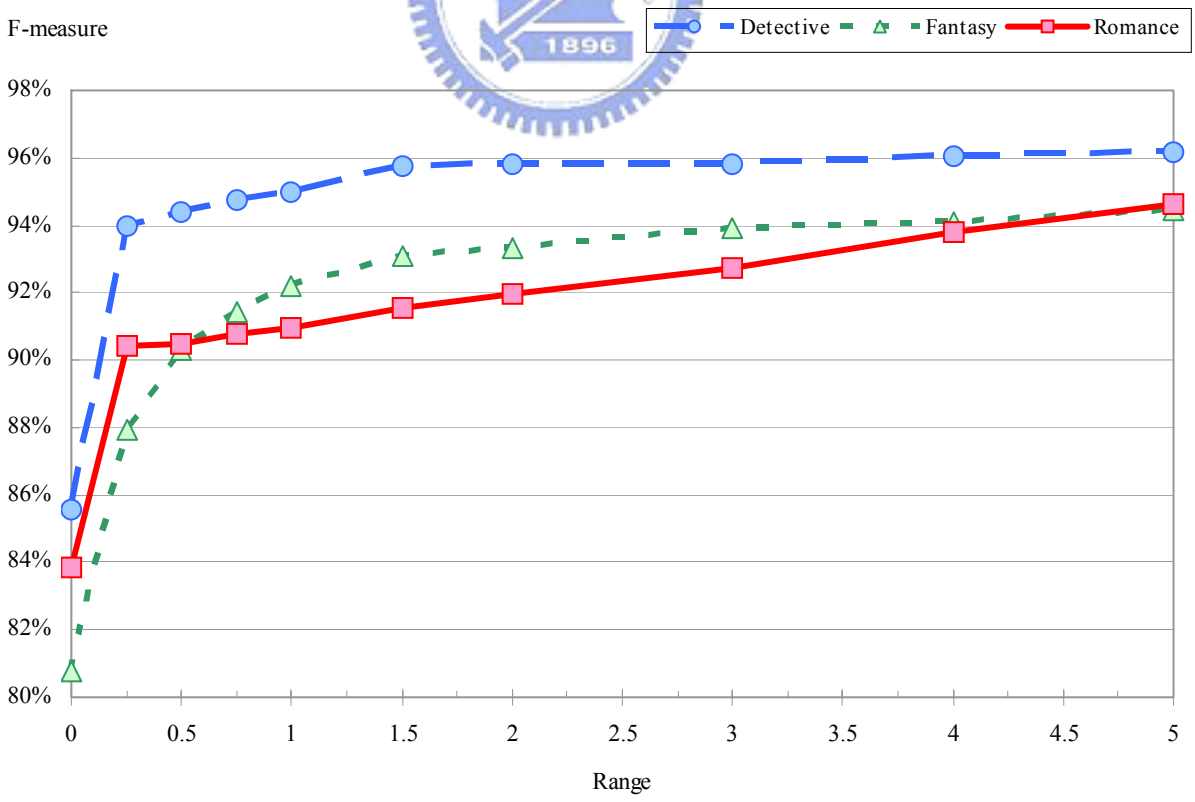

圖 4-4　取部分區間配合詮釋資料分類 — F-measure

五 、 結論與未來研究方向

本章總結整篇論文並提出未來可能的研究方向。5.1 節根據實作書籍分類系統的經驗，說明書籍詮釋資料與專家的智慧經驗如何用來提升分類成效；5.2 節由書籍分類的角度，提出不足或是可擴充之處做為未來的研究方向。

5.1. 結論

本論文將文件分類的概念應用在書籍分類上，建立了一個針對書籍的自動分類系統。本論文將書籍資訊分為敘述資料與詮釋資料兩部分，敘述資訊包含書名、書籍簡介與作者簡介，詮釋資料包含作者與出版社資訊。書籍分類系統首先以文件分類為基礎，配合專家的智慧經驗對敘述資料進行特徵挑選，以 SVM 分類法進行學習與分類；另一方面統計分析詮釋資料中有利於書籍分類的歷史資訊；最後將 SVM 分類結果融合詮釋資料中隱含的分類資訊，提升書籍分類系統的準確率。以下針對「專家介入特徵挑選」與「融合 SVM 分類法與詮釋資料」兩方面進行說明：

1) 在特徵挑選方面

專家刪除特徵的標準不宜過於嚴厲，這是因為去除掉原本以頻率計算較具有優勢的特徵，則不足的特徵缺額將由較不具代表性的特徵遞補，若刪除過多自動選取的特徵，反而有可能造成 SVM 分類的正確性下降；另一方面，經由專家指定加入的特徵必須給予適當的權重，否則在以頻率為基礎的計算上並不一定佔有競爭優勢，若無加權則輸入 SVM 分類器進行學習時無法凸顯其「經由專家指定」的重要性。經由專家挑選過濾並對重要的特徵加權後，SVM 分類書籍敘述資料的成效可由 83%提升到約 90%。

2) 在進行書籍分類方面

首先可藉由觀察 SVM 與單獨使用各項詮釋資料的分類結果了解各類別的資

料特性，進而以貪婪法則(Greedy)找出 SVM 分類法與詮釋資料之最佳的線性組合，將整體成效更進一步提高至 95%。除此之外，加入詮釋資料所改正分類結果之測試資料大多集中於 SVM 概念圖之 Margin 內，顯示加入詮釋資料能夠補強 SVM 面對新資料時，特徵可能不足的問題。

5.2. 未來研究方向

本論文結合文件分類、詮釋資料的分類特性，並且於特徵挑選過程加入專家智慧，在實際的測試中能有效地提升分類效果，正確率達 95%，但是仍有以下幾點可以改進，做為未來研究方向之參考。

1) 分類演算法

本論文嘗試以 SVM 分類法作為核心分類演算法，然而實驗過程中發現 Naïve Bayesian 分類法應用於分類敘述資料亦有不錯的成果，將來可嘗試以 Naïve Bayesian 或其他分類演算法作為分類核心；甚至可以更進一步，結合多種分類演算法，以投票表決的方式(Voting)決定敘述資料之分類結果，再融合詮釋資料進行書籍分類。

2) 詮釋資料的擴充、前置處理與分析

本論文初步探討詮釋資料對分類成果的影響，採用的詮釋資料為作者與出版社兩項，而書籍資料還有許多其他的資訊，例如主題標目(Subject Heading)與分類號(Classification Number)，其中亦可能隱藏許多與類別相關的訊息可供利用。

除此之外，本系統於前置處理作業時，僅將敘述資料進行轉換為向量表示式，未對詮釋資料進一步處理，若能對詮釋資料依照各分項的特性加以個別處理，使資料更為精準，必定對分類有所幫助。以作者資訊為例，作者名稱有許多為外文譯名，若音譯的不同、選字不同，在系統中便被視為不同的作者，

會造成作者代表性被分散的情況，若能統整建置作者的對照資訊，實現權威控制(Authority Control)，則能減輕譯名不同所帶來的影響。

最後，本論文以統計的方式分析詮釋資料，未來可以更進一步，以資料探勘或機器學習的方式，更精準地萃取詮釋資料中所隱藏的分類資訊，做更精確的分類預測。

參考文獻

[1] 賴永祥, "中國圖書分類法增訂七版," *現代圖書館學從刊第一種,* 1989.

[2] 陳和琴, 吳鎏璃 and 江琇瑛, *圖書分類編目*. 台北縣蘆洲市: 國立空中大學, 1996.

[3] J. Abrahams, "Probability, Random Variables, and Stochastic Processes," *Journal of the American Statistical Association,* vol. 79, pp. 957, 1984.

[4] M. C. Bauer, *Dewey Decimal Classification: 200 Schedules Expanded for use.* Catholic Library Association, 1988.

[5] R. S. Bot, Y. B. Wu, X. Chen and Q. Li, "A hybrid classifier approach for web retrieved documents classification," in *Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'04),* 2004, pp. 326.

[6] C. J. C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Min. Knowl. Discov.,* vol. 2, pp. 121-167, 1998.

[7] P. Calado, M. Cristo, M. A. Gonçalves, E. S. de Moura, B. Ribeiro-Neto and N. Ziviani, "Link-Based Similarity Measures for the Classification of Web Documents," *J. Am. Soc. Inf. Sci. Technol.,* vol. 57, pp. 208-221, 2006.

[8] C. H. Cheng, J. Tang, A. W. Fu and I. King, "Hierarchical classification of documents with error control," in *Proceedings of the 5th Pacific-Asia Conference on Knowledge Discovery and Data Mining(PAKDD '01),* 2001, pp. 433-443.

[9] K. W. Church and P. Hanks, "Word association norms, mutual information, and lexicography," in *Proceedings of the 27th Annual Meeting on Association for Computational Linguistics,* 1989, pp. 76-83.

[10] W. G. Cochran, "Some Methods for Strengthening the Common χ^2 Tests," *Biometrics,* vol. 10, pp. 417-451, 1954.

[11] P. Domingos and M. Pazzani, "Beyond independence: Conditions for the optimality of the simple bayesian classifier," in *13th International Conference on Machine Learning(ICML'96),* 1996, pp. 105-112.

[12] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis.* Wiley New York, 1973.

[13] S. Dumais, J. Platt, D. Heckerman and M. Sahami, "Inductive learning algorithms and representations for text categorization," in *Proceedings of the Seventh International Conference on Information and Knowledge Management(CIKM '98),* 1998, pp. 148-155.

[14] T. Dunning, "Accurate methods for the statistics of surprise and coincidence," *Comput. Linguist.,* vol. 19, pp. 61-74, 1993.

[15] M. Fuketa, S. Lee, T. Tsuji, M. Okada and J. Aoe, "A document classification method by using field association words," *Inf. Sci. Inf. Comput. Sci.,* vol. 126, pp. 57-70, 2000.

[16] M. Gregory, R. Scata and E. Brown, "Web document classification using machine learning clustering algorithms," *J. Comput. Small Coll.,* vol. 21, pp. 276-277, 2006.

[17] J. Han and M. Kamber, *Data Mining: Concepts and Techniques.* Morgan Kaufmann, 2000.

[18] E. Han and G. Karypis, "Centroid-based document classification: Analysis and experimental results," in *Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery*(*PKDD '00),* 2000, pp. 424-431.

[19] Jiawei Han and Micheline Kamber, *Data Mining: Concepts and Techniques.* Morgan Kaufmann, August 2000, pp. 500.

[20] M.James, *Classification algorithms.* John Wiley & Sons, Inc. 1985.

[21] E. Lehmann and J. P. Romano, *Testing Statistical Hypotheses.* Springer, 2005.

[22] Library of Congress. Cataloging Policy and Support Office, *Library of Congress classification. H. Social sciences* , Washington, D.C. : Library of Congress, Cataloging Distribution Service, 1994.

[23] C. Lin and E. Hovy, "The automated acquisition of topic signatures for text summarization," in *Proceedings of the 18th Conference on Computational Linguistics,* 2000, pp. 495-501.

[24] S. -. Lin, M. C. Chen, J. -. Ho and Y. -. Huang, "ACIRD: Intelligent Internet Document Organization and Retrieval," *IEEE Trans. Knowled. Data Eng.,* vol. 14, pp. 599-614, 2002.

[25] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing.* MIT Press, 1999.

[26] R. Michalski, J. Carbonell and T. Mitchell, *Machine Learning, an AI Approach,* 1983.

[27] S. K. G. Murthy, "Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey," *Data Mining and Knowledge Discovery,* vol. 2, pp. 345-389, 1998.

[28] J. Neyman, *Joint Statistical Papers.* University of California Press, 1967.

[29] A. Papoulis, "Probability, random variables and stochastic processes," *New York: McGraw-Hill, 1984, 2nd Ed.,* 1984.

[30] J. R. Quinlan, *C 4. 5: Programs for Machine Learning.* Morgan Kaufmann, 1992.

[31] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval.* McGraw-Hill, Inc. New York, NY, USA, 1986.

[32] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manage.,* vol. 24, pp. 513-523, 1988.

[33] G. Salton, E. A. Fox and H. Wu, "Extended Boolean Information Retrieval," 1982.

[34] A. Schenker, M. Last, H. Bunke and A. Kandel, "Classification of web documents using a graph model," in *ICDAR '03: Proceedings of the Seventh International Conference on Document Analysis and Recognition,* 2003, pp. 240-244 vol.1.

[35] F. Sebastiani, "Machine learning in automated text categorization," *ACM Comput. Surv.,* vol. 34, pp. 1-47, 2002.

[36] V. Vapnik, "The Nature of Statistical Learning Theory. 1995," *NY Springer,*

[37] F. Yates, "Contingency Tables Involving Small Numbers and the χ^2 Test," *Supplement to the Journal of the Royal Statistical Society,* vol. 1, pp. 217-235, 1934.

[38] Y. Li and A. Jain, "Classification of Text Documents," *The Computer Journal,* vol. 41, pp. 537-546, 1998.

[39] Encyclopaedia Britannica Online - http://www.britannica.com/

[40] SVM*light* support vector machine - http://svmlight.joachims.org/

[41] 元智電子報 - http://www.yzu.edu.tw/E_news/304/student/01.htm

附錄一　中研院平衡語料庫詞類標記集

精簡詞類	簡化詞類	對應的 CKIP 詞類標記[18]	
A	A	A	/*非謂形容詞*/
C	Caa	Caa	/*對等連接詞，如：和、跟*/
POST	Cab	Cab	/*連接詞，如：等等*/
POST	Cba	Cbab	/*連接詞，如：的話*/
C	Cbb	Cbaa, Cbba, Cbbb, Cbca, Cbcb	/*關聯連接詞*/
ADV	D	Dab, Dbaa, Dbab, Dbb, Dbc, Dc, Dd, Dg, Dh, Dj	/*副詞*/
T	DE		/*的，之，得，地*/
ADV	Da	Daa	/*數量副詞*/
ADV	Dfa	Dfa	/*動詞前程度副詞*/
ADV	Dfb	Dfb	/*動詞後程度副詞*/
ASP	Di	Di	/*時態標記*/
ADV	Dk	Dk	/*句副詞*/
FW	FW		/*外文標記*/
T	I	I	/*感嘆詞*/
N	Na	Naa, Nab, Nac, Nad, Naea, Naeb	/*普通名詞*/
N	Nb	Nba, Nbc	/*專有名稱*/
N	Nc	Nca, Ncb, Ncc, Nce	/*地方詞*/
N	Ncd	Ncda, Ncdb	/*位置詞*/
N	Nd	Ndaa, Ndab, Ndc, Ndd	/*時間詞*/
DET	Nep	Nep	/*指代定詞*/
DET	Neqa	Neqa	/*數量定詞*/
POST	Neqb	Neqb	/*後置數量定詞*/
DET	Nes	Nes	/*特指定詞*/
DET	Neu	Neu	/*數詞定詞*/.
M	Nf	Nfa, Nfb, Nfc, Nfd, Nfe, Nfg, Nfh, Nfi	/*量詞*/
POST	Ng	Ng	/*後置詞*/
N	Nh	Nhaa, Nhab, Nhac, Nhb, Nhc	/*代名詞*/
Vt	SHI		/*是*/
T	T	Ta, Tb, Tc, Td	/*語助詞*/
Vi	VA	VA11,12,13,VA3,VA4	/*動作不及物動詞*/

[18] 斜體詞類，表示在技術報告#93-05中沒有定義，即後來增列的。

精簡詞類	簡化詞類	對應的 CKIP 詞類標記[18]	
Vt	VAC	VA2	/*動作使動動詞*/
Vi	VB	VB11,12,VB2	/*動作類及物動詞*/
Vt	VC	VC2, VC31,32,33	/*動作及物動詞*/
Vt	VCL	VC1	/*動作接地方賓語動詞*/
Vt	VD	VD1, VD2	/*雙賓動詞*/
Vt	VE	VE11, VE12, VE2	/*動作句賓動詞*/
Vt	VF	VF1, VF2	/*動作謂賓動詞*/
Vt	VG	VG1, VG2	/*分類動詞*/
Vi	VH	VH11,12,13,14,15,17,VH21	/*狀態不及物動詞*/
Vt	VHC	VH16, VH22	/*狀態使動動詞/
Vi	VI	VI1,2,3	/*狀態類及物動詞*/
Vt	VJ	VJ1,2,3	/*狀態及物動詞*/
Vt	VK	VK1,2	/*狀態句賓動詞*/
Vt	VL	VL1,2,3,4	/*狀態謂賓動詞*/
Vt	V_2	V_2	/*有*/

附錄二　專家加入的類別相關詞彙

偵探/懸疑小說類							
神秘	凶手	分屍案	手法	失蹤	犯罪	兇手	兇殺案
死亡	死法	死者	自殺	身亡	事件	亞森	命案
奇案	怪盜	案件	真凶	真相	追蹤	偵探	動機
密室	推理	殺人	殺人案	殺害	殺機	陳屍	陰謀
意外	綁架	罪犯	慘遭	槍殺	福爾摩斯	謀殺	謀殺案
謎團	離奇	羅蘋	證據	懸案	懸疑	可疑	目擊
密碼	屍體						

科幻/奇幻小說類							
人魚	女巫	中土	元素	召喚	外星	外星人	地球人
地獄	托爾金	吸血鬼	妖怪	妖精	巫師	奇幻	武俠
法師	法術	冒險	哈利波特	封印	星際	星戰	科幻
科技	科學	神秘	神祕	基地	惡魔	詛咒	傳奇
預言	種族	精靈	銀河	衛斯理	機器人	龍	獵人
騎士	羅琳	魔王	魔戒	魔咒	魔法	魔界	魔族
靈異	靈魂						

愛情文藝小說類							
分手	心	心情	心痛	心愛	心酸	外遇	失戀
交往	同居	吻	吸引	村上春樹	孤單	幸福	恨
相愛	相遇	苦澀	浪漫	真心	真情	真愛	祝福
追求	寂寞	情	情人	情感	情愛	情話	深情
深愛	甜蜜	結婚	感動	感情	愛	愛上	愛情
愛慕	愛戀	暗戀	廝守	邂逅	癡情	戀	戀人
戀情	戀愛						