圖書資訊學研究 13：1 (December 2018)：87-124

使用多數決策略之圖書自動分類的研究

The Study of Automatic Book Classification Using Majority Vote Strategy

郭俊桔

June-Jei Kuo

國立中興大學圖書資訊學研究所助理教授
Assistant Professor
Graduate Institute of Library and Information Science
National Chung Hsing University

【摘要 Abstract】

大多數的圖書館館員多半只懂得圖書資訊學領域的知識，卻必須負責所有到館圖書的分類工作。因此常常有因學科背景知識不足造成分類困難的情形。再加上，近年隨著科技的進步圖書出版量大幅度的增加，造成編目館員越來越重的負擔而使得無法提高分類速度，且容易受到主觀認知差異的影響，導致 inter-consistency 和 intra-consistency 等一致性低落的編目品質問題。本論文探討傳統圖書自動分類的課題並結合多種分類器的優點，提出使用多數決策略之多層次圖書自動分類。為了探討此研究的效能，首先使用大學碩博士論文及其對應之圖書分類號為訓練與測試語料。針對其文件內容，研究各種內容組合對文件表徵擷取的影響後，找出應用於圖書自動分類之最佳的內容組合，例如，摘要與目錄等。同時，針對各種分類器的組合，利用分類器間具有互補的特性探討應用於本論文的最佳分類器與階層之組合。針

通訊作者：郭俊桔 jjkuo@dragon.nchu.edu.tw
投稿日期：2017 年 1 月 23 日；接受日期：2018 年 6 月 21 日

對小量論文或少量多種類網路書店的語料得到令人滿意的實驗結果。進一步，使用大量少種類網路書店之書目資料執行自動分類。使用 10 折交叉驗證的方法驗證其效率，實驗結果顯示使用本論文提出之使用多數決策略之多層次圖書自動分類比傳統圖書分類具有更佳的分類效能。

Most librarians understand the knowledge of the library information sciences and a few other academic fields, but they are responsible for the bibliography section of all the academic fields. Due to a lack of background knowledge, the bibliography becomes more and more difficult for the librarians. Moreover, thanks to the recent rapid improvement of technology, the amount of publication in every academic field increases very quickly, and the bibliography load further increases. The quality of the bibliography, such as high inter-consistency and high intra consistency of library classification, is not easy to be maintained. Thus, this paper dealt with issues of traditional automatic book classification and employed the complementary attribute of various classifiers to propose a multiple layered automatic book classification using majority voting strategy. First, the collection of theses from a university library was utilized as the training and testing corpus. The classification codes of those theses were employed as the gold standard as well. Each thesis contained various components such as a title, author, table of contents, abstract or cited papers et al. To understand the classification effect of all the combinations of these components, various combinations were studied and the best combination (i.e., the combination of abstracts and a table of content) were recommended. On the other hand, to obtain the best classification performance, the layer allocation of classifiers was also studied and the best combination was recommended. Moreover, the thesis classification results were promising. Furthermore, to conduct the large amount and multiple categories automatic book classification experiment, the book content pages from online bookstore were collected. Under the principal of 10-fold cross-validation, experimental results showed that the performance of the proposed automatic book classification outperformed the traditional automatic book classification as well.

【關鍵字 Keywords】

圖書自動分類；多數決策略；階層式；分類器

automatic book classification; majority vote strategy; layer style; classifier

壹、緒論

圖書館除了保存圖書外，將圖書做有系統的整理進而提供讀者利用一直是圖書館重要的使命。文件分類則是所有圖書館和資訊中心的重要功能之一。為了達到此目的，圖書館專家與學者，例如杜威（Melvil Dewey），克特（Charles A. Cutter）和阮甘那桑（S.R. Ranganathan）等人發展了各種書目分類（Bibliographic classification）技術。編目（Cataloging）是指對資訊進行整理與組織，依據共同規範的規則將資訊記載著錄，使各類型與範圍的資訊都能以相同格式呈現，以方便讀者在檢查與利用時能有效率找到所需的資訊（何世文，2004；賴鼎銘等，2001）。編目[1]工作內容一般可分為記述編目（Descriptive Cataloging）和主題編目（Subject Cataloging），記述編目是對資料進行形體上的分析，並按照編目規則逐項著錄，把讀者能瞭解該資料形體上的特性（書名、作者、出版社、高廣、頁數等）資訊記錄下來，日後供讀者查詢圖書之檢索點。而主題編目是對一件資料作內容上的分析，用以決定該本書的分類號及主題標目 。其主要的工作有（1）給分類號（分類編目）： 依據分類法將圖書資料依其主題或性質不同，予以分別歸類，並以簡單符號組合代表，使同類書籍資料集中、並區別不同類書籍資料者稱之。（2）定標題：根據標題表，用幾個單一名詞或簡短詞句，標明圖書的主題內容（陳和琴、吳琉璃、江琇瑛，1996）。

分類編目指的是依據圖書的內容和採用的分類法（例如，中國圖書分類法[2]），選擇最適當的類目給予類號（陳光華、羅思嘉、林純如，2002；黃純敏，2002；黃嘉宏，2008；Kwok, 1975；Yang & Pederson, 1997；Yi, 2006）。其主要目的是給予書本在書架上有固定的位置，進一步將內容相近的圖書聚集在一起。另一方面，傳統的分類編目工作是由具有相關領域知識背景的專家先閱讀文件內容，並依其文意和內

1 詳細資料參見網站 http://techserviceslibrary.blogspot.tw/2011/04/cataloging.html

2 詳細資料參見網站 http://lib.fsvs.ks.edu.tw/ccl.htm

容主旨決定該文件的類別。此類方法不但耗費大量的人力、時間以及金錢，且人工分類的作業容易受到專家主觀認知差異性的影響，因而導致分類結果不一致。再者，圖書館館員多半只懂得圖書資訊領域的相關知識，卻必須負責圖書館內的所有圖書之分類工作，而常常有因學科背景知識不足造成分類困難的情形。例如，陳光華等人（2002）在『探討圖書資訊學學術期刊文獻主題編目的一致性』文中，運用一套主題分類體系，實際分析中國圖書館學會學報第一期到六十五期出版的 956 篇文獻，發現大類類號的一致性指數平均 82.00%，然而子類號的一致性指數平均為 59.20%~74.09%；完整類號（大類＋子類）的一致性指標平均為 60.82%。再加上，近年由於資訊科技的進步已在各個學科領域使得數位化檔案增加的速度越來越快，更造成編目館員負擔越來越重而無法提高速度外，容易受到主觀認知差異的影響，導致 inter-consistency（不同標引人員）和 intra-consistency （相同標引人員，不同時間）等一致性低落的編目品質問題（Tonta, 1991）。

另一方面，為解決上述不一致的課題，圖書資訊領域開始導入資料探勘之文件自動分類（曾憲雄等，2007；Chou, Han, & Chen, 2007；Wei, Lin, & Yang, 2011）的研究成果，文件自動分類的效能有些已達人工分類水準（郭俊桔、張瑞珊、張育蓉，2013；Chen & Wu, 1999；Kuo & Chen, 2005），並可以同時進行多類別多文件的即時分類．例如黃純敏（2002）探討自動分類技術於學術論文分類，使用民國 80 年及 88 年全國碩博士論文，以類神經網路訓練與測試其分類模型。然而，其實驗結果顯示於全碼自動分類平均精確率為五成五，逐層減少，前兩碼的精確率仍只有 76%。

然而，隨著網際網路和數位圖書館的電子資源的快速成長，圖書館使用傳統之人工編目或分類已捉襟見肘。（Mengle & Goharian, 2009；Wang et al., 2009）不論是人工或自動文件分類目前都無法達到令人滿意的分類效率（Efficiency）與效用（Effectiveness），使用者經常無法找到滿足個人資訊需求的圖書或期刊，造成對圖書館使用的滿意度大大降低。（Pong, Kwok, Lau, & Wong, 2008） 另一方面， Delveen and Shereen（2013）提出傳統分類器間具有互補的效果，同時使用多數個分類器和多數決策略，可以進一步提升文本分類效率。因此，如何導入多數個分類器和多數決策略以提供一個高品質自動化圖書分類，快速且有效地協助人工的分類作業，從而大幅改進圖書管理效率，已成為圖書館管理一個重要且熱門的研究課題。

貳、文獻探討

文件分類（Document Classification / Text Categorization）問題一直以來是圖書資訊、資訊科學和計算機科學等領域的熱門研究題目。文件分類係指依文件的資料內容與主旨給予該文件適當的類別標籤，進而對文件進行分門別類的加值處理，以達到分類管理與利用之目的。文件分類的相關研究可以分為人工分類與自動分類。文件人工分類的研究（Tonta, 1991；陳光華等，2002）主要存在於圖書資訊學領域，而自動文件分類的研究（王梵志、張俊聖，2001；陳信源等，2008；曾元顯，2002；黃嘉宏，2008；Antonie & Zaiane, 2002；Broko & Bernick, 1963；Desale & Kumbhar, 2013）則以資訊科學和計算機科學領域為主。因此。文件分類可以說是跨學科（interdisciplinary）的研究議題。自動文件分類的主要分成三種研究：基於規則式（Rule Based Methods）、基於統計式（Statistics Based Methods）和基於擷取式（Retrieval Based Methods）。基於規則式（Hamill & Zamora, 1980; Gopinath & Prasad, 1994）主要是使用引文內容（Citation Content）所建立各類模板（Pattern）進行比對後自動分類。基於統計式（Cheng & Wu, 1995; Tokkola, 2002）則是使用統計排序和權重的方法執行自動分類。再者，基於擷取式（Carpineto & Romano, 1994; Farbrizio, 2002）強調在資訊檢索系統中使用人工智慧的理論執行自動分類。多數的相關研究著重於辭彙分類（Term Classification）、文本與文件分類（Text and Document Classification），應用於改善網際網路上的資訊檢索效率。然而，早在電子計算機被發明以前大多數的圖書分類法已被提出，目前的自動分類技術無法有效對應。因此，對於圖書館和資訊科學而言，仍缺乏設計自動圖書分類的理論 / 典範（Desale & Kumbhar, 2013）。

自動文件分類技術主要參考已分類的訓練資料，藉由機器學習（machine learning）演算法（Christopher et al., 2008），自動建立文件分類模型。日後如果有新的文件需要進行分類時，分類模型便能依據已學習的分類規則，自動將新文件歸類到一個或多個適當的類別。另一方面，以分類器的結構來區分，自動文件分類方法可區分為階層式結構與非階層式結構。其中，階層式結構指的是所建立的分類類別具有階層（先後執行）關係；非階層式結構則是分類類別間沒有特別的關係。

建立自動文件分類的流程，依序為如下所示：

一、文件表徵之擷取

從文字資料中擷取有意義（重要）的詞彙，以作為建立分類器之文件表徵。而所謂文件表徵，指的是每個文件對於分類比較有助益的關鍵詞向量。首先，文件表徵擷取主要是在決定分類所要根據的文件內容片段，主要可分為五種：（1）題名；（2）摘要；（3）關鍵詞；（4）本文；（5）參考文獻之題名。接著，由於本論文的訓練與測試文件為中文，將選取的文件內容使用中文斷詞器[3]進行斷詞（segmented）後，文件表徵的選取與相關權重計算方法有：詞頻（TF），詞頻 - 反轉文件頻率（TFIDF），文件頻率（DF），資訊增益，卡方檢定，互斥資訊等。（曾元顯，2002；Farbrizio, 2002；Tokkola, 2002）

二、建立文件表徵

利用表徵集合及所對應的表徵值建立文件表徵（Yang & Pederson, 1997），以作為計算機運算的依據。一般，文件會表示成以關鍵詞表徵所建立的高維度空間中的一個向量表徵，即文件向量（Document Vector）。

三、分類模型之建立

參考已分類的訓練資料，藉由機器學習演算法，自動學習並建立分類模型。常見的分類演算法（Wang, 1996；Delveen & Shereen, 2013），包括決策樹、Naïve Bayesian、K-Nearest Neighbor、支援向量機（Support Vector Machine）、Regressive Model、機率分類器和類神經網路等。

四、自動分類評估

針對訓練資料，將分類器的判斷類別與答案類別比較，計算此分類器的精確率（precision, P）和召回率（recall, R）後，用來比較分類器的效率。由於好的分類器同時是高精確率和高召回率，另一個常用的評估函數為 F-score，可以依需要分別調整精確率和召回率的權重，例如調和 F-score（Christopher et al., 2008）。另外，由於評估分類效率，有時必須數個類別一起考量，為避免各類別的個數不一致所造成評量

3　中央研究院 CKIP 斷詞系統：http://ckipsvr.iis.sinica.edu.tw/

結果的高估而導入 micro-average 與 macro-average 兩種計算方式（曾元顯，2002；曾綜源、吳俊儀，2008）。計算分數時，micro-average 針對每一個文件給予相同的權重，macro-average 則先計算每一類的分數後，才計算類別平均精確率；換句話說，對於比較小的類別會給予比較大的權重。

自 1960 年代開始，就有文獻提出利用計算機執行自動文件分類。大多數的參考文獻主要都是討論文件分類的技術，注重文件表徵擷取（Chen & Wu, 1999；Tokkola, 2002；Yang & Pederson, 1997）或分類模型建立（Farbrizio, 2002；Jaillet et al., 2006），鮮有以圖書分類為對象的研究文獻。其中有關圖書自動分類的參考文獻，英文文件分類主要有 Larson（1992）、Yi（2006）和 Avila-Arguelles 等人（2010）。Larson（1992）主要由書目資料中的題名和主題詞擷取表徵詞，針對 283 筆圖書資料，使用 60 種分類模型和三種表徵組合執行分類工作。實驗結果 inside test 平均精確率可以達到 86%，然而 outside test 平均精確率只有 46%。Yi（2006）是一篇調查報告，主要敘述圖書分類號的應用和機器學習的相關技術外，針對數位圖書館介紹圖書分類號的進行中計畫以及未來可能的挑戰。其中，作者提到由於自動圖書分類所需的訓練資料將是非常龐大，造成不易自動化的問題點，希望大家重視這個課題。再者，Avila-Arguelles 等人（2010）提出監督式圖書分類器，只以題名作為文本，混合兩種文本分類技術：Lesk 投票典範（Voting Scheme）和詞彙頻率（Term Frequency）擷取表徵與計算其表徵值後，訓練其美國國會圖書館分類號（Library of Congress Classification, LCC）的圖書分類模型。分別使用 489,726 訓練資料和 122,431 測試資料，其實驗結果 inside test 平均精確率達到 92.95%，然而 outside test 平均精確率只有 36.13%。

此外，對於中文文件分類，陳信源等（2008）於博客來書店取得「偵探／懸疑小說」、「科幻／奇幻小說」、「愛情文藝小說」三個類別各 900 本圖書書目資料。圖書書目資料內所包含資訊分別為描述部份與詮釋部份兩個部份。描述部份包含的資料有：書名、內容簡介和作者簡介；詮釋資料則包含作者與出版社資訊。使用支持向量機分類器進行實驗。實驗中，以描述部份作為主要分類對象，並搭配 Log-Likelihood（LLR）方式挑選類別表徵並加入詮釋資料輔助以提升效能。最後實驗結果以 F-measure 的評估下皆有 94% 以上的效能。然而，類別的數量會影響分類器的分類效能，該研究僅處理三個類別與實際書店或圖書館所需相距甚遠。因此其實驗結果是否可以代表實際應用

下的分類則值得商榷。而黃嘉宏（2008）鑑於從描述部份和詮釋資料中抽取文件表徵，仍不足以描述文件的向量表徵；所以利用搜尋引擎進行擴展文件表徵並以文件作者欄位資訊進行輔助的策略，以期能就擴展文件表徵之觀點，改善以圖書館書目資料進行文件自動分類之效能。其實驗的素材取自國家圖書館館藏目錄資料，總訓練文件數為19,499 筆（由 400~499 類號，800~899 類號篩選出 156 類），總測試文件為4911筆。採用三種表徵詞策略為:（1）雙連詞表徵和作者資訊，（2）所有 google 之擴展詞和作者資訊，（3）所有 google_nv_un 之擴展詞和作者資訊。其實驗結果的最佳精確率分別為 0.5960、0.6286 和 0.6261。

由上述國內外的參考文獻，自動圖書分類系統仍有如下所示課題，有待克服：

1. 由於圖書館館員無法精通各個領域的知識，無法正確指定圖書分類號而使得讀者可能無法找到架上的自己所需的圖書。
2. 針對自動圖書分類的研究，只有參考圖書之書目資料的部份內容，例如只使用題名，無法充分表達文件的表徵。
3. 使用傳統文件分類的學習模型（單層式圖書分類系統），例如類神經網路或支持向量機等，針對大量測試資料時的分類精確率大都只有 45%~63%，對於編目品質的提升助益不大。（Desale & Kumbhar, 2013）
4. 資訊爆炸的世紀，書籍大量且快速出版，如何即時進行書目分類，不須重新訓練分類模型等的議題，都是重要的研究課題。

本論文為解決課題 1 和課題 4，提出使用多數決策略之圖書自動分類，以解決人工分類的一致性以及分類速度與可調適性等問題。針對課題 2，透過不同文件內容片段組合，探討適用於自動分類的文件資訊表徵。至於課題 3 之單一分類器的精確率不佳，使用資料探勘之分類器的各種組合和投票策略探討編目品質的可否提升。

參、研究設計與實施

過去提出的圖書自動分類系統（陳信源等，2008；黃嘉宏，2008；Avila-Arguelles et al., 2010；Larson, 1992）僅使用一種分類器的方法，只可得到差強人意的分類效果。然而，為了進一步提高自動分類效果，導入多種分類器可以有互補效果的優點（Wang, 1996），本論文提出使用多數決策略之階層式圖書自動分類，試著結合多種分類演

算法，並以多數決（Majority Voting）的方式決定圖書分類。以兩層分類器為例，當輸入測試文件之文件表徵後，分別由第一層的分類器 A 和分類器 B 執行分類後，如果兩者的分類結果一致時，則接受該分類結果，如果分類結果不一致時，交由第二層的分類器 C 執行分類，再使用多數決策略判斷其分類結果。

換句話說，如果分類器 C 的分類結果和前述第一層的兩個分類器的任何一個相同時，則判定此分類為分類結果；否則判定為未知。再者，由於圖書的內容片段除書目資料（題名、作者、出版單位、出版時間、關鍵詞）外，為彌補書目資料的不足，進一步導入如摘要、全文和參考文獻題名等文件內容片段，本論文也研究何種文件內容片段的組合對於多層式圖書自動分類具有最好的執行效率和分類結果，例如，題名 + 作者、題名 + 作者 + 出版社、題名 + 參考文獻題名等各種組合。

本論文透過第一階段實作五種單一分類器的模型，篩選出具較佳分類效能的分類模型，並據以作為第二階段實作的多層式分類器之候選分類模型，最後針對分類結果使用多數決策略取得最大分類精確度的方式，實作出比單一分類器效能更好的多層式分類器，詳細之研究方法與步驟如下：

一、資料蒐集

第一部分博碩士語料蒐集包括 95 學年到 99 學年台灣中部某大學五個學系博碩士學位論文電子檔 79 冊，第二部分的網路書店[4]書目語料蒐集，則取民國 100 年 ~104 年出版的書籍資料，共五個類別，『商業理財』、『室內設計』、『醫療保健』、『電腦資訊』和『食譜』各取出 50 冊（共 250 冊），供多樣少量實驗使用。接著，為了少樣多量實驗，針對『商業理財』、『室內設計』和『醫療保健』等三個類別從網路書店再分別取出 300 冊（共 900 冊）。

二、資料前置處理

此模組將電子文件轉換成分類器能閱讀之格式，也就是將整個文

4　博客來網路書店網址為 http://www.books.com.tw

件集表示成為表徵向量（feature vector）集合。首先，使用包含斷詞切字（tokenization）及刪除停用詞（stopwords）等步驟（陳信源等，2008），分析上述圖書的書目資料內容，並擷取重要詞彙作為分類學習的文件表徵。本論文由中央研究院平衡語料庫中抽出出現頻率最高的 100 個詞作為中文停用詞，另外也採用 Onix Text Retrieval Tookit[5] 中的 429 個英文停用詞。再者，本論文使用 TF-IDF 方法，透過詞彙出現之頻率和反轉檔案頻率的乘積來計算每一個詞彙的重要性後，挑選權重高的詞彙作為每一份文件之文件表徵。

三、分類器訓練

將所有文件表示成表徵向量後，本論文將以機器學習方法進行文件分類模型之建構。此分類模型將能依據各種內容片段組合的文件表徵，預測出不同之分類結果。文件的主要內容片段可分為（1）題名，（2）摘要，（3）本文，（4）參考文獻之題名，（5）目錄和（6）關鍵詞等六大部分；然而觀察小量論文之前測結果發現（1）題名和（6）關鍵詞具有高度重複性以及含有雜訊會導致分類結果的錯誤；例如，基因（gene）一詞常常出現在園藝學系、昆蟲學系和動物科學系等之論文關鍵詞中。因此，本論文以台灣中部某大學碩博士論文之為文件來源，探討（2）、（3）、（4）、（5）、（2）+（3）、（2）+（4）、（2）+（5）、（2）+（3）+（4）、（2）+（3）+（5）、（2）+（4）+（5）、（2）+（3）+（4）+（5）、（3）+（4）、（3）+（5）、（3）+（4）+（5）、（4）+（5）等 15 種內容組合，了解哪一種組合可以得到最佳的分類效能。

本論文先依其類號將文件分別儲存至不同的類別資料庫中，將各類別資料庫中的文件，擷取上述不同的 15 種內容片段組合的文件表徵，分別訓練五種分類器。參考分類器之效能比較結果（Delveen & Shereen, 2013；Wu et al., 2008），採用效能最好的三種分類器演算法 NaiveBayes、SVM、決策樹 C 4.5 外，也導入 KNN（1）、投票法（使用多數決策略）。

四、實例驗證

模型建構完成後，即可使用分類器來進行自動化文件分類之作業。

5 http://www.lextek.com/onix/

在本論文中，以不同的文件內容片段組合作為分類器之文件表徵訓練資料。在訓練階段結束後，即可得知針對類似博碩士論文此類文件，最具代表性的文件內容片段組合為何。此外，使用分類效能較佳之文件片段組合作為測試資料，透過多種多層式分類器架構得到不同的分類結果，且據以進行多層式分類器架構之效能評估。

五、分類器效能評估

使用文件資料之圖書分類號作為分類的正確答案。傳統的分類效能評估方式，類似資訊檢索效能評估。曾元顯（2002）對每一個類別，所有的文件將被劃分為四種狀況，（1）屬於該類的文件，且被系統正確分為該類的有 a 篇、（2）屬於該類的文件，沒被系統正確分為該類的有 b 篇；（3）不屬於該類的文件，卻被系統分為該類的有 c 篇、（4）不屬於該類的文件、且沒被系統分為該類的有 d 篇。對每個類別都做這樣的統計後，即可計算「正確率」（A, accuracy）、「精確率」（P, precision）、「召回率」（R, recall），如下：accuracy =（a+d）/（a+b+c+d），precision=a/（a+c），recall=a/（a+b）。同時考慮 precision 和 recall 的評估函數，則有 F-score =2P*R/（P+R）。為了和其他系統比較，本論文採用「精確率」為評估函數，來評估不同分類器模式之分類效用。

肆、研究結果與分析

一、使用學位論文資料集（Data Sets）

在訓練語料和測試語料蒐集方面，採用 95 學年到 99 學年中部某大學碩博士學位論文，含園藝學系 12 冊，昆蟲學系 12 冊，動物科學系 17 冊，水土保持學系 18 冊，中國文學系 20 冊，全部合計 79 冊。由於該校內學位論文編目以學院和系所作為區分，而國家圖書館為該篇論文指定的分類號亦為相同方式，故研究中各篇論文的分類方式遵循上述方法，亦即以學院和系所作為分類結果之表徵值。換句話說，此表徵值可以作為本論文的評估標準答案。

由於本論文的論文格式屬於 PDF，經過轉換成為 word 格式並刪除亂碼和錯誤後，記錄筆數與表徵數量為表徵（記錄筆數，表徵筆數）：題名（61, 54），摘要（76, 734），本文（73, 784），參考文獻（30,

483），目錄（74, 971），關鍵詞（48, 213），指導教授（17, 2）。例如，題名有 61 筆，經過表徵值自動抽取後，得到 54 個表徵。然而，由於指導教授的名字在論文分類上是一個很有效的分類表徵，為了解其他屬表徵值組合的效能，本實驗暫不導入。

進行訓練分類器的實驗使用紐西蘭 Waikato 大學所開發的 WEKA（Waikato Environment for Knowledge Analysis）版本 3.7.8，進行資料前置處理（原始文件資料檔案轉換成字詞矩陣格式）、資料維度精簡、表徵值選取。歷年來應用於文件自動分類的機器學習法很多，包括「類神經網路」（Artificial Neural Network, ANN）、「案例式推理」（Case-Based Reasoning, CBR）、「k 個最鄰近法」（k-Nearest Neighbor Algorithm, kNN）、「貝氏分類」（Naïve Bayes, NB）、「支援向量機」（Support Vector Machine, SVM），以及「決策樹」（Decision Tree）等。本論文第一階段依據 Delveen and Shereen （2013）的分類演算法的評比結果，挑選分類效率好的單純貝氏分類、支援向量機、決策樹、kNN，以及結合上述四種方法的投票法等五種方法進行實驗。在資料探勘軟體 WEKA 中，所使用的分類器之名稱對應如下：

單純貝氏分類：weka.classifiers.bayes.NaiveBayes；

支援向量機：weka.classifiers.functions.SMO；

決策樹 C4.5：weka.classifiers.trees.J48；

kNN（分 1 群）：weka.classifiers.lazy.IBk；

投票法（採多數決）：weka.classifiers.meta.MultiSchem。

再者，本論文之評估方法則採 10 折交叉驗證（10-fold Cross Validation），將這些資料切割為 10 份，進行 10 次之驗證。當進行其中一個 fold 之訓練與測試時，即利用該 fold 之文件作為測試資料，其餘 9 個 fold 之文件為訓練資料。

（一）文件表徵及其組合之效能

為挑選最能代表碩博士論文資料集的表徵值，由前述 79 冊論文中，先抽取其（1）題名、（2）摘要、（3）本文、（4）參考文獻、（5）目錄、（6）關鍵詞等六大部分，並採取上述五種機器學習法，分別訓練出五種分類器，透過效能比較，藉以初步篩選出具分類精確率在 80% 的表徵值項目。因此，本階段訓練的資料集共有：（1），（2），（3），（4），（5），（6）共六項。分別執行文件分類後，其結果如表 1 所示。其中，（1）題名的部份沒有包含無題名的論文之測試結果。由精確率的平均值來看，可以發現摘要、本文、參考文獻題名和

目錄都是 70% 以上。其中，摘要和本文之表徵值的分類效能更是明顯。另外，由於各系所之論文的題名皆很相似，例如，「基於…. 之研究」或「…一致性之探討」等，造成精確率下降的主因。

表 1
使用各種文件內容片段之分類精確率（單位 : 百分比）

資料集	Naïve Bayes Classifier	SVM (SMO)	C4.5 (J48)	kNN(1)	MultiScheme（投票法，採 majority vote）	平均
（1）題名	51.60	53.62	36.74	47.50	51.62	48.22
（2）摘要	**97.91**	**95.96**	51.21	63.46	**97.79**	**79.33**
（3）本文	**89.77**	**90.80**	76.75	75.18	**88.57**	**81.75**
（4）參考文獻	**66.33**	**83.33**	69.00	70.00	**81.33**	**74.00**
（5）目錄	**90.89**	**84.11**	73.64	31.11	**90.89**	**74.13**
（6）關鍵字	33.50	45.35	33.50	18.15	45.95	35.29

（二）單層分類器架構之實驗

由上述六項文件內容片段中，篩選出具有平均分類效能較佳的前四種文件內容，再據以做資料集的組合，以持續提供後續分類器訓練使用。假設如果選擇的文件內容為（2），（3），（4），（5），則資料集中的組合為：（2）+（3）、（2）+（4）、（2）+（5）、（2）+（3）+（4）、（2）+（3）+（5）、（2）+（4）+（5）、（2）+（3）+（4）+（5）、（3）+（4）、（3）+（5）、（3）+（4）+（5）、（4）+（5）等 11 種內容。此外，本實驗僅留下在表徵抽取的文件內容片段之實驗中分類精確率在 85% 以上的分類法持續進行實驗，以聚焦找出對碩博士論文此類文件分類有最佳表現的機器學習法及內容片段組合，研究哪一種組合可以得到最佳的分類效能。故經過組合擷取，本實驗的資料集其記錄筆數與表徵數量如表 2。

接續表徵抽取的文件內容之實驗，分類器則採 NaiveBayes、SVM 與 WEKA 的 MultiScheme（多數決，採 majority vote）此三種方法，資料集則以（2）摘要、（3）本文、（4）參考文獻、（5）目錄此四種，分別再以排列組合方式產生 15 種不同組合面貌，故本階段分類結果如表 2。

由表 2 可得知，透過不同文件內容片段組合，所訓練出的分類器

精確率互有高下。為了更精確擷取出最具碩博士論文語料庫表徵值代表性之組合，在後續之二階層分類器架構實驗，擬選擇在分類法 NaiveBayes、SVM 兩者精確率皆大於 85% 的資料集。因此由表 2 數據，共可篩選出（2）、（3）、（2）+（5）、（3）+（5）、（2）+（3）+（4）、（2）+（3）+（5）、（3）+（4）+（5）、（2）+（3）+（4）+（5）共八種組合。

表 2
單層分類器架構之精確率（單位 : 百分比）

資料集 (2) 摘要	(3) 本文	(4) 參考文獻	(5) 目錄	記錄筆數	表徵數量	Naïve Bayes Classifier	SVM（SMO）	MultiScheme（投票法，採 majority vote）
*				**76**	**734**	**97.91**	**95.96**	**97.79**
	*			**73**	**784**	**89.77**	**90.80**	**88.57**
		*		30	483	66.33	83.33	82.33
			*	74	971	90.89	84.11	90.89
*	*			149	770	57.50	91.14	91.14
*		*		106	770	79.92	83.77	83.39
*			*	**150**	**666**	**97.60**	**92.60**	**97.60**
	*	*		103	471	82.15	85.78	85.12
	*		*	**147**	**492**	**90.45**	**96.19**	**96.19**
		*	*	104	424	83.28	83.07	83.49
*	*	*		**179**	**666**	**89.95**	**95.10**	**94.99**
*	*		*	**223**	**650**	**93.36**	**97.81**	**97.81**
*		*	*	180	666	84.14	89.27	90.34
	*	*	*	**177**	**660**	**89.10**	**97.07**	**97.07**
*	*	*	*	**253**	**653**	**91.35**	**97.79**	**97.79**

（三）多層分類器架構之實驗

主要依據表 2 實驗結果，測試多層式分類器之效能而將分類器分為二階層。二階層分類器架構為第一階層先以 Naïve Bayes Classifier 與 SVM Classifier 進行投票，若結果一致，則輸出分類結果。若不一致，則再將分類不同的資料紀錄轉至決策樹 C4.5（J48）再度進行分類，再將其分類結果與一階層分類器的兩個分類結果進行多數決投票，以判斷類別。另外，資料集有（2）、（3）、（2）+（5）、（3）+（5）、（2）+（3）+（4）、（2）+（3）+（5）、（3）+（4）+（5）、（2）+（3）+（4）+（5）共八種組合。本實驗結果如表 3 所示。不論哪一種方法的執行時間都低於 10 分鐘。

表 3
二階層分類器架構之精確率（單位 : 百分比）

資料集				方法 1: MultiScheme (NB/SVM 投票法，採 majority vote)	方法 2: MultiScheme (NB/SVM/J48 投票法，採 majority vote)	方法 3: MultiScheme (NB/SVM→ J48) Train data: 全部 instances	方法 4: MultiScheme (NB/SVM→ kNN) Train data: 全部 instances
(2) 摘要	(3) 本文	(4) 參考文獻	(5) 目錄	全部 instances 皆執行訓練	全部 instances 皆執行訓練	Test Data: 僅留下當 Naïve Bayes 與 SVM 不一致時的 instances	Test Data: 僅留下當 Naïve Bayes 與 SVM 不一致時的 instances
*				97.79	97.79	100.00	99.33
	*			88.57	87.84	100.00	100.00
*			*	97.60	97.07	100.00	99.45
	*		*	96.19	96.61	99.55	99.55
*	*	*		94.99	97.49	98.81	99.60
*	*		*	97.81	97.62	100.00	100.00
	*	*	*	97.07	96.92	100.00	100.00
*	*	*	*	97.79	97.87	99.03	99.03
平均				**95.97**	**96.16**	**99.67**	**99.62**

（四）討論

如表 3 所示，方法 1 和方法 2 為使用多數決策略的單層分類器，而方法 3 和方法 4 則為使用多數決策略之二階層式分類器。由表 3 數據得知，由於方法 1 與方法 2 之精確率互有優劣，顯示多數決策略在單層時作用不明顯。然而，方法 3 及方法 4 的精確率皆優於方法 1 和方法 2 的精確率。

傳統的圖書自動分類方法的執行結果如表 1 所示，最高平均精確率只有 81.75%。而本論文使用投票策略之多層式圖書自動分類的精確率，皆優於傳統的圖書自動分類方法，可證明多層式圖書自動分類之優越性。再者，由實驗數據得知，又以二階層分類器架構之實驗的方法 3 及方法 4 得到的最佳的分類精確率。在上述數據中，方法 3 及 4 分別可達到 99.67% 和 99.62% 的精確率，顯示研究中所採用兩層式分類架構具分類實務上之可行性。

針對碩士論文的：（1）題名、（2）摘要、（3）本文、（4）參考文獻、（5）目錄、（6）關鍵詞等六方面資料，本論文發現碩士生使用的論文題名和關鍵詞相似度很高，造成自動分類效率的下降，然而其他四個方面的資料則是有效的文件表徵。另一方面，多層式圖書分類系統之分類器組合，由表 3 數據中可得知，kNN 的整體表現略遜於 J48，故可合理推測，若未來以大量數據進行實驗，則雙階層分類法中的第二階層採 J48，會比採 kNN 能獲得較佳效率。因此，由於二階層式分類器的第二層只須執行第一層兩個分類器分類不一致的檔案，需要分類的檔案數大幅減少，即使使用須要較久訓練和處理時間的分類器，例如，關連規則分類器，也可以利用多層式分類器的優點，有效執行圖書自動分類。

二、使用網路書店書目資料

由於前述實驗（採用博碩士論文為實驗語料）的使用多數決策略之階層式文件自動分類可得到令人滿意的精確率。為進一步驗證多層式分類器架構對於圖書分類的效能，本節語料來源是博客來網路書店[6]，任意挑選本數多於 50 本的五類，取「商業理財」、「室內設計」、「醫療保健」、「電腦資訊」與「食譜」於民國 100 年～ 103 年出版的書籍資料，共五個類別、各 50 本，總共 250 本。書目資料如附錄一所示，包括書名，目錄和摘要等。再者，各類別書籍書目資料經過前

6　博客來網路書店網址為 http://www.books.com.tw

置處理後擷取之文件表徵數量（括弧內數字）分別為摘要（495），目錄（500）以及摘要和目錄（500）。

資料集建立完成後，分兩階段訓練第一和第二階層分類器架構。在第一階層分類器實驗中，精確率測試方法採 10 折交叉驗證。第二階段作法則接續第一階層的分類結果，挑出該階段分類不一致的資料，匯入第二階層分類器中，進行再次分類判斷，並計算其精確率。

（一）第一階層分類器架構之實驗

採取較常用且效能較佳的機器學習法 Naïve Bayes Classifier 與 SVM，並抽取語料中（1）摘要、（2）目錄和（3）摘要 + 目錄三部分，各別訓練出三種分類器。由表 4 中可得知，實驗資料集不論是採用摘要、目錄，或是摘要加上目錄，所得到的分類精確率皆有 98% 以上的水準。接著，使用多數決投票（majority vote）策略，其分類精確率分別為 98.0%（5），97.6%（6）和 97.2%（7）。括號內的數字為分類不一致的筆數。

「摘要」語料不管使用何種分類器或多數決投票策略皆得到較好的分類結果，可得知在書籍分類上的表徵值選用上，摘要具重要代表性。另一方面，雖然「摘要 + 目錄」語料比「目錄」語料具有更多資料內容，同時也引進雜訊而使得精確率些許下降。

表 4
第一階層分類器架構之精確率

資料集		Naïve Bayes Classifier	SVM（SMO）	Majority Vote
（1）摘要	精確率	98.8%	98.4%	98.0%
	資料筆數（正確 / 錯誤）	247/3	246/4	245/5
（2）目錄	精確率	98.0%	97.6%	97.6%
	資料筆數（正確 / 錯誤）	245/5	244/6	244/6
（3）摘要 + 目錄	精確率	98.4%	98.0%	97.2%
	資料筆數（正確 / 錯誤）	246/4	245/5	243/7

（二）第二階層分類器架構之實驗

分類器訓練為接續第一階層的分類結果，挑出該於第一階層分類器架構分類不一致的資料，繼續匯入第二階層分類器中進行類別判斷，並再將第二階層分類結果與第一階層的兩種分類器進行投票（多數決）。本階段採行 kNN 與決策樹 C4.5（J48）平行實驗，實驗數據如表 5，在第二階層分類器架構中，決策樹 C4.5 的整體精確率仍優於 kNN。其中，使用目錄或摘要 + 目錄語料，C4.5 皆可得到 100% 之精確率，顯示出多數決策略之階層式文件自動分類運用於圖書自動分類的可行性。至於使用摘要語料，仍有一個錯誤殘留的原因是第一階層的兩分類器皆將其分類至相同的錯誤類別而被認為正確，不會交由第二階層進一步處理。

表 5
第二階層分類器架構之精確率

	資料集	kNN	C4.5（J48）
（1）摘要	精確率	98.8%	99.6.%
	資料筆數（正確 / 錯誤）	247/3	249/1
（2）目錄	精確率	98.0%	100%
	資料筆數（正確 / 錯誤）	245/5	250/0
（3）摘要 + 目錄	精確率	98.8%	100%
	資料筆數（正確 / 錯誤）	245/3	250/0

三、增加各類筆數之網路書店書目資料

由於前述實驗（採用博碩士論文、網路書店書目資料）為實驗語料的多層式文件自動分類法精確率達 97% 以上。然而，上述語料皆具有類別多和每類資料少的特性。然而，圖書資訊分類下，每類之下皆有大量圖書。因此，想承續上述分類器訓練架構，進一步了解減少類別個數但增加各類筆數的情形下，是否會影響多數決策略之階層式文件自動分類的執行效率。

（一）語料與資料前置處理

語料來源仍為前述之博客來網路書店，取民國 100 年～ 104 年出

版的書籍資料，共「商業理財」、「室內設計」、「醫療保健」三個類別、各 300 本，總共 900 本。接著，抽出其書目資料（1）摘要、（2）目錄、（3）摘要 + 目錄三類進行分類實驗。900 筆書目資料經過前置處理後之文件表徵數量（括號內的數字）為摘要（1786），目錄（1879）以及摘要和目錄（1846）。

資料集建立完成後，分兩階層訓練分類器。實驗的做法是在第一階層分類器實驗中，精確度測試方法仍採 10 折交叉驗證，第二階段作法則接續第一階層的分類結果，挑出該階段分類不一致的資料，匯入第二階層分類器中，進行再次分類判斷，並計算其精確率。

（二）實驗結果

1. 第一階層分類器架構之實驗

在第一階層，採取了文件自動分類中較常用的機器學習法 Naïve Bayes Classifier、SVM，分別訓練出三種分類器，即（1）「摘要」、（2）「目錄」、（3）「摘要 + 目錄」。所得之數據如由表 6 中可得知，實驗資料集不論是採用摘要、目錄，或是摘要加上目錄，所得到的分類精確率皆有 93% 以上的水準；其中，又以書籍的「摘要 + 目錄」在分類上表現最佳，以本次實驗蒐集的 900 筆之網路書店書目資料集，透過 Naïve Bayes Classifier 與 SVM 的演算法進行分類，精確率達到 94% 以上，可得知在書籍分類上的表徵值選用上，「摘要」及「目錄」語料皆具備重要代表性。

與表 4 結果比較，「摘要 + 目錄」語料不管使用何種分類器或多數決投票策略皆得到較好的分類結果，而不是只有使用「摘要」語料。分析實驗結果，發現其原因是由於類別減少，文本數量如果不足，機器學習將無法訓練出具有類別或文件的代表性表徵值而容易造成分類錯誤。另一個有趣的發現是 SVM 分類器的分類結果不論在哪一種訓練語料皆優於 Naïve Bayes Classifier。因此，綜合表 4 和表 6 的實驗結果，當每類資料很多，若欲使用傳統單層分類器時，推薦使用 SVM 分類器；相對的，當每類資料較少時，則推薦使用 Naïve Bayes Classifier。

表 6
第一階層分類器實驗之精確率

資料集		Naïve Bayes Classifier	SVM（SMO）	Majority Vote
（1）摘要	精確率	94.0%	98..6%	93.7%
	資料筆數（正確 / 錯誤）	846/54	888/12	844/56
（2）目錄	精確率	97.8%	98.4%	97.2%
	資料筆數（正確 / 錯誤）	880/20	244/6	875/25
（3）摘要 + 目錄	精確率	98.7%	99.1%	98.2%
	資料筆數（正確 / 錯誤）	892/11	889/8	884/16

2. 第二階層分類器架構之實驗

基於欲提高分類精確率之目標，多層式文件分類的第二階層分類器訓練。為接續第一階層的分類結果，挑出該於第一階層分類不一致的資料，繼續匯入第二階層分類器中進行類別判斷，不一致的資料筆數（括號內的數字）分別為摘要（56），目錄（25）以及摘要和目錄（16）；並且再將第二階層分類結果與第一階層的兩種分類器進行投票（多數決）。經過第一階層分類篩選後，第二階層再採用表 5 分類器架構實驗中效果較佳之決策樹 C4.5（J48）分類演算法，經過加入決策樹的分類法，其投票後之精確率如表 7。此外，為了方便比較效能，表 8 為傳統單階層分類器與多層式分類器之精確率比較，相較於單一階層，三種資料集在兩階層分類器中精確率均有提升，由此可見，多了一層分類器的篩選，資料分類精確率必可以再次提升。

表 7
第二階層分類器實驗之精確率

資料集		第二階層加入 C 4.5（J48）後，並採投票法決定分類結果
（1）摘要	精確率	99.3%
	資料筆數（正確 / 錯誤）	894/6
（2）目錄	精確率	99.4%
	資料筆數（正確 / 錯誤）	895/5

（3）摘要 + 目錄	精確率	99.6%
	資料筆數（正確 / 錯誤）	896/4

表 8
傳統單階層分類器與多層式分類器之精確率比較

資料集（900 筆）	傳統單層式自動分類器 NB 的精確率	傳統單層式自動分類器 SVM 精確率	多層式自動分類器之精確率
（1）摘要	94.0%	98.7%	99.3%
（2）目錄	97.8%	98.4%	99.4%
（3）摘要 + 目錄	99.6%	99.4%	99.6%

四、效能評估與討論

以 900 筆測試資料為例，如表 8 所示其精確率皆達 99% 以上，本論文的多層式圖書自動分類具有實用價值。為了進一步了解錯誤發生的原因，本節討論以網路書店書目為資料集的多層式分類器實驗下，造成分類錯誤之原因，以及效用評估。

（一）分類錯誤原因探討

從表 6 的實驗結果，第一層使用 Naïve Bayes 分類器之效果明顯比使用 SVM 分類器差。因此，為進一步探究造成 Naïve Bayes 分類器分類錯誤的原因，經針對分類錯誤之紀錄逐筆取出檢視，可歸納出如下所兩個原因如下；

1. 書目資料內容片段太短，無法擷取足夠數量之文件表徵。

書目資料（摘要或目錄）文句如果太短，經斷詞、去除停用詞後，留下的文件表徵數量就太少，以致分類器無法正確判斷。例如：有一筆屬於 " 摘要 " 的書目資料，其內容如下，經查此筆資料，標準答案是 " 財務管理 "。

1（DET）亞洲（N）\n　2（DET） 中東（N）\n　3（DET） 歐洲（N）\n　4（DET） 非洲（N）\n　5（DET）　美洲（N）\n

但是，就人類專家用字面上來看，根本看不出來此筆紀錄屬於「財務管理」領域，故三種分類法中，Naïve Bayes Classifier 分類為「財務管理」領域，SVM 與決策樹 C4.5 都分類成是「室內設計」領域。看

起來 Naïve Bayes Classifier 似乎比較正確（它分結果的跟標準答案是一樣的），但是經人類專家檢視資料集內容，判斷後反而覺得分成室內設計比較正確。因為在室內設計的書目資料（摘要）中，的確有好幾筆資料是在探討 " 異國風室內設計風格 "，裡面的字詞含有亞洲、歐洲風 ... 等風格。故 SVM 與 C4.5 將此筆紀錄當成室內設計，實屬合理。由於本論文使用多數決策略之多層式自動分類，故 2:1 的情況下，分類的答案為「室內設計」，但實際上原來標準答案給的是「財務管理」，故此筆就被認為是分類錯誤。事實上，造成謬誤的因素，就是因為受限於該筆資料（摘要）文句太短，造成文件表徵數量太少，以至於導致辨識度不夠。

然而，相同的一本書，在「目錄」的書目資料集中，就沒有發生分類錯誤，因為此本書在「目錄」資料集中，留下的文件表徵數量比較多，故可以明確讓三種分類法都分類正確。並且，在「目錄」+「摘要」中，也被兩階層分類器分類正確，故可推斷，這要歸功表於此筆書目在「目錄」資料集中的所留下的文件表徵具備足夠的代表性，換言之，也就是該筆書目資料的文件表徵數量夠多。

因此，為朝提高分類器精確率方向邁進，且必須避免上述分類錯誤，資料集中各文件內容片段應具備足夠的內容與品質。

2. 書目資料經斷詞、去除停用詞後，留下來的文件表徵數量不足，會產生文件表徵不具備代表性之狀況。例如：下面此筆書目資料，屬於 " 摘要 "，內容如下：

KB（FW） 1909（DET） 總部（N）\n 商業（Nv） 1909（DET） HEADQUARTERS（FW）\n 讓（Vt） 工作（N） 變得（Vt） 有趣（Vi） THQ（FW） 工作室（N）\n THQ（FW） WORK（FW） AT（FW） PLAY（FW）\n CONNECT（FW） Sydney（FW） 工作室（N）\n CONNECT（FW） SYDNEY（FW）\n 媒體（N） 工作室（N）\n ATTRACTION（FW） MEDIA（FW）\n K（FW） 辦公室（N）\n OFFICE（FW） K（FW）\n NFOE（FW） 公司（N） 辦公（Nv） 空間（N）\n BUREAU（FW） 100（DET） A（FW）\n NEW（FW） OFFICE（FW） FOR（FW） THE（FW） CENTURY-OLD（FW） FIRM（FW） NFOE（FW）\n BPGM（FW） 律師（N） 事務

所（N）\n BPGM（FW） LAW（FW） OFFICE（FW）\n 荷蘭（N） 紅牛（N） 總部（N）\n RED（FW） BULL（FW）\n SC（FW） 工作室（N）\n STUDIO（FW） SC（FW）\n CUMLUS（FW） 辦公室（N）\n CUMULUS（FW） OFFICE（FW）\n 微軟 （N） 維也納（N） 總部（N）\n HEADQUARTER（FW） MICROSOFT（FW）\n VIENNA（FW）\n “（FW） 巴拉巴拉（N）”（FW） 辦公室（N）\n BARRA & BARRA（FW） OFFICE（FW）\n BCC（FW） BOVES（FW） 銀行（N）\n BCC（FW） BOVES（FW） BANK（FW）\n 日航（N） 房地產（N） 開發（Vt） JAL（FW） REAL（FW） ESTATE（FW） DEVELOPMENT（FW） OFFICES（FW）\n BOUKLIS（FW） 集團（N） 農產品（N） 進出口（Nv） 公司（N） 牙科（N） 診所（N） 金融（N） 機構（N）\n ARTIS（FW） CAPITAL（FW） MANAGEMENT（FW）\n 倫敦（N） 考文（N）特花（N） 園（N） GOOGLE（FW）\n PENSON（FW） GOOGLE（FW） CSG（FW） COVENT（FW） GARDEN（FW） LONDON（FW）\n 豐田（N） 通商（Nv） 辦公（Nv） TOYOTA（FW） TSUSHO（FW）\n

在多階層分類器研究法中， Naïve Bayes Classifier 分類為「財務管理」領域，SVM 分類為「室內設計」領域，決策樹分類成「醫療」領域；但因為三種分類器答案都不同，最後只好判斷該筆分類錯誤。事實上該筆書目的標準答案是「室內設計」。以人類專家的再次人工確認，經檢視三類型的書目資料內容，發現其實不論分成哪一種都似乎都對。例如：在「財務管理」領域中，常常出現 " 商業 "、"WORK"、" 工作 "、" 銀行 "、" 金融 "、" 房地產 " 等之文件表徵。而「室內設計」領域中，也常會出現 " 工作室 "、" 花園 "、" 空間 " 等文件表徵。在「醫療」領域中，也曾出現過 " 工作 "、"PLAY"、" 牙科 "、" 診所 "、" 農產品 " 等文件表徵字眼，故三種分類法分類成三種結果，經人類專家的判斷，認為應歸咎於留下來的文件表徵不夠具代表性。

因此，為朝提高分類器精確率方向邁進，使用大量資料分類時，為避免上述分類上的錯誤，應再導入「目錄」或其他書目資料以使抽

取之文件表徵更具代表性。

（二）實驗結果討論

1. 分類精確率之探討

由表 8 中之傳統單階層式自動分類器與多層式自動分類器之精確率比較數據得知，第一階層的精確度測試方法，不論是以 Naïve Bayes Classifier 與 SVM 分類器，三種文件內容片段皆可得到的 94% 以上的分類精確率。若再加入本論文提出之多層式自動分類，則在表中的精確率欄位中的數據可看出皆有提升至 99% 以上，故採多層式分類法比僅採單層式之傳統分類法，確實可以效提升分類精確率。

2. 標準檔案格式之制定

若希望提高日後自動分類精確率，所須匯入系統之電子檔格式應具備一套標準存檔規則，並且文件內容及書目資料文件之撰寫也應統一格式，如此除了可以在前置處理階段時，就降低無效紀錄的比率外，亦可提高分類績效。

3. 運用於圖書館館藏編目實務的可行性

本論文藉由以機器學習為基礎的多層次分類法，建立書目資料自動分類知識架構，除了可輔佐圖書館編目館員從事館藏編目工作、降低編目工作錯誤率、提高工作時效外，更可將機器學習演算法適用於現實生活中的文件自動分類之範圍，在應用層面亦具有實務創意研究成果。並且，由於分類精確率相較於過往單多層式自動分類，亦有較高的精確率，因此可以應用於圖書自動分類系統上。再者，以多層式自動分類進行博碩士論文及網路書目資料進行文件自動分類，惟該類檔案涉及個別不同領域之專業知識，專有名詞甚多，加上多數皆以電子文件 PDF 檔案格式、或網頁檔案 HTML 格式呈現，無純文字檔，因此加添前置處理之困難。經剔除無效檔案後，將所有檔案請圖書資訊專業人員進行快速分類，並將分類結果作為多層式分類法之學習與訓練機制及分類標準答案。研究結果分類精確率可高達 99% 左右。因此，可以預期的是若將本論文中的網路書店書目分類對應到圖書館現行的分類規則，亦是具體可行作法。

由此合理推斷，採用多層式文件分類模式在博碩士論文資料與書籍資料上都具備相同程度的實務應用性；且網路書店書籍書目資料，目前尚僅採用摘要與目錄的組合，因此，若再納入其他書目資料，相信一定可以達到更令人滿意的分類效能。

伍、結論與未來研究方向

文件分類方法的出現為大量圖書歸類問題帶來許多解決方案，然而，在傳統自動圖書分類模式常僅針對單種分類模式進行參數調整與效能評估，實務上模式應用範疇也就有所限制。本論文基於此動機進行了使用多數決策略之階層式圖書分類之建構與實例驗證。

一、結論

根據研究結果顯示，在網路書店書目資料集中，本論文提出之使用多數決策略之多層式自動分類，相較於傳統只採用單種分類器做分類，精確率可達 99%，因此證實能夠有效地提高分類之精確率。並且因為在多層式分類法中的第二階段，已經大幅度減少欲執行分類之資料量，故也可改善傳統單一分類器在執行大量資料時會發生的分類時間過久、效能降低之困境。另外，從以多數決投票法作為第二階段實驗主體方式中可以發現，分類模式之效能，易隨著資料紀錄之原始文句字詞量影響，亦即若所蒐集到的書目資料集原始文句字詞若太過簡短（例如：只有正常中文文章的 1~2 句話），則在資料維度精簡後，所蒐集到的文件表徵數量將會不夠具備代表性，以至於經過多層式分類法的分類後，使得分類效能隨之遞減或浮動。但若是該筆書目資料文句字詞為 5 句以上，則採用多層式分類方法後，分類結果便能夠提升精確率，如此亦可使分類模式表現得更加穩定。

二、未來研究方向

本論文多數實驗的平均精確率皆超過 95%，然而受限於資料取得困難或清理費時的限制，僅有少量類別與圖書數量可以用於實驗，跟真實圖書館分類任務，動輒上百個類別、數十萬本書，差距過大。未來重要的研究課題與方向是如何讓分類器成效數據進一步接近真實情境的情況下做實驗，呈現自動分類在圖書館館藏分類的可行性。

此外，在未來研究上，仍有下列方向可繼續進行：

1. 以機器學習為基礎的階層式分類法，雖具延展性與擴充性特質，仍無法有效全面適用於圖書館各類館藏資料的編目作業。因此，未來可考慮針對個別館藏特性，尋求合適之分類類目表作為分類學習與編目參考，或利用非監督式類神經網路相似表徵群聚特性，依館藏別進一步改善目前所呈現之分類架構與細目，亦

是提高分類效能的做法。

2. 由於圖書分類號是由三個層次的號碼所組成，為探討多層式圖書自動分類系統對兩層或三層的分類號是否具有相同或更好的分類結果，實務上做法亦可將分類結果對應到圖書館現行採用的分類號。亦即，若能將本論文中的網路書店書目分類結果，對應到圖書館現行採用之的圖書分類規則，針對分類精確率進行後續研究，如此能更切合館員的需求，也能提高館方的使用意願。
3. 可以採用不同的文件表徵選取方法進行比較與測試。在文件表徵選取與表 示方法中，本論文僅採用較常被使用的 TF 與 TFIDF 方法進行初步的文件表徵選取與表示。由於是使用 bag of word 的觀念，詞和詞之間沒有順序關係，是否能真正反應文件語義的表徵向量，仍有待商榷。在未來研究上，建議可以採用不同的文件表徵選取方法進行比較與測試，例如，導入本體論（Ontology）等。
4. 由於網路資源普及，世界上各國語系的資料，日日夜夜以爆量方式劇增，可針對不同語言資料集進行後續擴展實驗。例如，參照網路書店目前區分之類別進行網路新聞的實驗，更可提升實驗結果的參考性。

誌謝

國科會計畫 NSC 101-2221-E-005 -090 對於本論文的部份經費補助和支持， 以及吳慧貞研究助理的熱心協助，在此致上由衷感謝之意。

參考文獻

王棯志、張俊盛（2001）。適應性文件分類系統。*在第十四屆計算語言學研討會論文集*（頁 99-121）。台南市：成功大學。【Wang, R. J., & Jhang, J. S,. (2001). Shih ying sing wun jian fen lei si tong. In *Di Shih Sih Jieh Ji Suan Yu Yan Syueh Yan Tao Huei Lun Wun Ji* (pp.99-121). Tainan: National Cheng Kung University. (in Chinese)】

何世文（2004）。分類法與主題法之比較。*國立中央圖書館臺灣分館館刊*，*10*(1)，25-37。【He, S.W. (2004). Fen lei fa yu jhu ti fa jhih bi jiao. *Bulletin of NCL Taiwan Branch, 10*(1), 25-37. (in Chinese)】

郭俊桔、張瑞珊、張育蓉（2013）。導入矩陣分群之視覺化圖書推

薦系統。教育資料與圖書館學，*51*(1)，5-35。【Kuo, J. J., Chang, J. S., & Zhang, Y. J. (2013). Visualized Book Recommender System Using Matrix Clustering. *Journal of Educational Media & Library Sciences, 51*(1), 5-35. doi:10.6120/JoEMLS.2013.511/0560.RS.AM. (in Chinese)】

陳光華、羅思嘉、林純如（2002）。圖書資訊學學術期刊主題編目一致性之探討。在資訊傳播學術研討會論文集（頁 125-142）。臺北：世新大學。【Chen, K. H., Lo, S. C., & Lin, C. J. (2002). The Investigation of the Consistency of Subject Cataloging for Academic Journal Articles of Library and Information Science .In *Zih Syun Chuan Bo Syueh Shu Yan Tao Huei Lun Wun Ji* (pp. 125-142). Taipei: Shih Hsin University. (in Chinese)】

陳和琴、吳琉璃、江琇瑛（1996）。圖書分類編目。臺北縣：空中大學。【Chen, H.Q., Wu, L. L., & Jiang, S. Y. (1996). *Tu shu fen lei bian mu*. Taipei: National Open University (in Chinese)】

陳信源、葉鎮源、林昕潔、黃明居、柯皓仁、楊維邦（2008）。結合支援向量機與詮釋資料之圖書自動分類方法。資訊科技國際期刊，*3*(1)，2-21。【Chen, S. Y., Yeh, J. Y., Lin, S. J., Huang, M. J., Ke, H. R., & Yang, W. B. (2008). Jieh he jhih yuan siang liang ji yu chyuan shih zih liao jhih tu shu zih dong fen lei fang fa. *International Journal of Advanced Information Technologies (IJAIT), 3*(1), 2-21. (in Chinese)】

曾元顯（2002）。文件主題自動分類成效因素檢討。中國圖書館學會會報，*68*，62-83。【Tseng, Y. H.(2002). Effectiveness Issues in Automatic Text Categorization. *Bulletin of the Library Association of China, 68*, 62-83. (in Chinese)】

曾綜源、吳俊儀（2008）。文件內容來源對文件分類之績效評估。「2008 數位科技與創新管理研討會」發表之論文，新北市。【Tseng, Z. Y., & Wu, J. Y. (2008). Performance Evaluation of Various Document Content Sources on Document Classification. Paper session presented at *the 2008 Conference of Digital Technology and Innovation Management.* New Taipei, Taiwan. (in Chinese)】

曾憲雄等（2007）。資料探勘。台北市：旗標。【Tseng, S. S. (2007). *Data Mining*. Taipei: Qi Biao (in Chinese)】

黃純敏（2002）。學術論文自動分類技術研究。行政院國家科學委員會補助專題研究計畫成果報告（NSC 902416-H-224-016）。

【Huang, C. M. (2002). *Syueh Shu Lun Wun Zih Dong Fen Lei Ji Shu Yan Jiou*. Sing Jheng Yuan Guo Jia Ke Syueh Wei Yuan Huei Bu Jhu Jhuan Ti Yan Jiou Ji Hua Cheng Guo Bao Gao (in Chinese)】

黃嘉宏（2008）。基於自動分類爲基礎圖書題名特徵擷取之研究 - 以輔助圖書分類系統爲例（未出版之碩士論文）。輔仁大學圖書資訊學研究所，新北市。【Huang, J. H. (2008). *A Study of Book Title Feature Extraction Based on The Automatic Classification -An Example of BibliographyAutomatically Classified System* (Unpublished master's thesis). Fu Jen Catholic University, Department of Library and Information Science, Taiwan. (in Chinese)】

賴鼎銘等（2001）。圖書資訊學概論。台北縣：空大。【Lai, T. M. (2001). *Tu shu zih syun syueh gai lun*. Taipei: National Open University (in Chinese)】

Antonie, M. L., & Zaiane, O. (2002). Text Document Categorization by Term Association. In *Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM'02)* (pp. 19-26). Maebashi City, Japan.

Avila-Arguelle, R., Calvo, H., Gelbukh, A., & Godoy-Calderon, S. (2010). Assigning Library of Congress Classification codes to books based only on their titles, *Informatica, 34*.77-84.

Broko, H., & Bernick, M. (1963). Automatic Document Classification. *Journal of the ACM, 10*(2), 151-162.

Carpineto, C., & Romano, G. (1994). Dynamically bounding browsable retrieval spaces: an application to Galois Lattices. *In proceedings of RLAO 94: intelligent multimedia information retrieval systems and management (pp. 520-533)*. New York.

Chen, K.H., & Wu, C.T. (1999). Automatically Controlled-Vocabulary Indexing for Text Retrieval. *In Proceedings of Research on Computational Linguistics Conference XII (pp.171-185)*. Hsin-Chu, Taiwan.

Cheng, P. T.K., & Wu, A. K.W. (1995). ACS: an automatic classification system. *Journal of information science, 21*. 289-299.

Chou, C. H., Han, C. C., & Chen, Y. H. (2007). *GA based optimal keyword extraction in an automatic Chinese web document classification system*. In Proceedings of the Frontiers of High Performance Computing and

Networking ISPA 2007 Workshops (pp. 224-234). Niagara Falls, Canada.

Christopher, D. M., Prabhakar, R., & Hinrich, S. (2008). *Introduction to Information Retrieval*. Cambridge University Press, 2008.

Delveen Luqman Abd AL-Nabi, & Shereen Shukri Ahmed. (2013). Survey on Classification Algorithms for Data Mining:(Comparison and Evaluation). *Computer Engineering and Intelligent Systems, 4*(8), 18-24.

Desale, S., & Kumbhar, R. (2013). Research on Automatic Classification of Documents in Library Environment: A Literature Review. *Knowledge Organization. 40*. 295-304.

Farbrizio, S. (2002). Machine Learning in Automated Text Categorization. *ACM Computing Surveys, 34*(1), 1-47.

Gopinath, M. A., & Prasad, A.R.D. (1994). A knowledge representation model for analytico-synthetic classification. *In Proceedings of the third international ISKO conference* (pp.320-327). Copenbangen, Denmark.

Hamill, K. A., & Zamora, A. (1980).The Use of Titles for Automatic Document Classification. *Journal of the American Society for Information Science, 31*(6), 396-402.

Jaillet, S., Laurent, A., & Teisseire, M. (2006). Sequential Patterns for Text Categorization. *Intelligent Data Analysis, 9*, 1-16.

Kuo, J. J., & Chen, H. H. (2005). Cross Document Event Clustering Using Knowledge Mining from Co-Reference Chains. In *Proceedings of Second Asia Information Retrieval Symposium (AIRS 2005)* (pp. 121-134). Jeju Island, Korea.

Kwok, K. L. (1975). The Use of Title and Cited Titles as Document Representation for Automatic Classification. *Journal of Information and Management, 11*, 201-206.

Larson, R. R. (1992). Experiments in Automatic Library of Congress Classification. *Journal of the American Society for Information Science, 43*(2), 130-148.

Mengle, S. S. R., & Gohnraian, N. (2009). Ambiguity measure feature-selection algorithm. *Journal of the American Society for Information Science and Technology. 60*. 1037-1050.

Pong, J. Y. H., Kwok, R. C. W., Lau, R. Y. K., Hao, J. X., & Wong, P. C. C.

(2008). A comparative study of two automatic document classification methods in a library settings. *Journal of information Science. 34*. 213-230.

Tokkola, K. (2002). Discriminative Features for Document Classification, In *Proceedings of the 16th International Conference on Pattern Recognition (ICPR02)* (pp. 472-475). Quebec City, QC, Canada.

Tonta, Y. (1991). A Study of Indexing Consistency between Library of Congress and British Library Catalogers. *Library Resources & Technical Services, 2*, 177-185.

Wang, J. T. L. (1996). Complementary classification approaches for protein sequences. *Protein Engineering, 9*(5), 381-386.

Wang, J. T. L., Marr, T. G., Shasha, D., Shapiro, B. A., Chirn, G. W., & Lee, T. Y. (2009). An extensive study on automated Dewey Decimal Classification. *Journal of the American Society for Information Science and Technology, 60*. 2269-2286.

Wei, C. P., Lin, Y. T., & Yang, C. C. (2011). Cross-lingual text categorization: Conquering language boundaries in globalized environments. *Information Processing & Management, 47*(5), 786-804.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., … Steinberg, D. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems, 14*(1), 1-37.

Yang, Y., & Pederson, J. O. (1997). A Comparative Study on Features Selection in Text Categorization, In *Proceedings of the 14th International Conference on Machine Learning* (pp. 412-420). Nashville, TN, USA.

Yi, K. (2006). Challenges in Automatic Classification using Library Classification Schemes. In *Proceedings of World Library and Information Congress: 72nd IFLA General Conference and Council*, Seoul.

The Study of Automatic Book Classification Using Majority Vote Strategy

June-Jei Kuo

Assistant Professor
Graduate Institute of Library and Information Science
National Chung Hsing University

Cataloging refers to assigning a publication the most appropriate category and classification number according to its content and the adopted classification scheme (such as New Classification Scheme for Chinese Libraries). The main purpose of classification is to assign the publication to a specific location on the shelf and group publications of similar contents together. The traditional classification follows this procedure below. Experts with relevant knowledge background read the content first, and decide the categories according to the publication's themes and contents. This procedure requires a large amount of manpower, time and money, and the manual classification is easily influenced by experts' differences of subjective cognitions, thus resulting in inconsistencies of classification. In addition, most librarians have relevant knowledge only in the library information field, but are required to be responsible for all the cataloging and classification in libraries, thus encountering classification difficulties resulted from insufficiency of background knowledge. Furthermore, because of the advancements of information technologies, the digitalized documents in all subject fields have increased in a large number, creating an even heavier load for cataloging librarians to enhance the speed of cataloging. Cataloging is also easily influenced by differences of catalogers' subjective cognitions, resulting in inconsistencies of cataloging, including inter-consistency (among different catalogers) and intra-consistency (among different cataloging times of the same cataloger).

To solve the problems mentioned above, automatic book classification systems of data mining have been introduced into the library and information science field. Some of the functions of automatic book classification have achieved the quality of manual classification and can do real-time classification

of multi-categories and multi-documents simultaneously. The results of experiments reveal that the accuracy rate of all digits of classification numbers through automatic book classification is 55%, decreasing with more digits counted, and the accuracy rate of the first two digits is only 76%. However, no matter it is manual or automatic book classification, the overall efficiency and effectiveness of classification are still not satisfying. Library users often fail to find books or journals that meet individual information needs, thus the satisfaction degree of library uses is largely decreased.

Methodology

Automatic book classification systems of the past have adopted only one classifier, generating barely satisfying results of classification. For enhancing the performance of automatic book classification systems, multiple classifiers have been incorporated. In this paper, the hierarchical automatic book classification system with a majority-ruled decision making strategy has been investigated, in which multiple classification algorithms are integrated, and the classification results are decided by a majority voting approach. Taking the two-level classification system as an example, after the features of testing document are keyed in, the classifier A and classifier B of the first level conduct the classifications. If the results of classifications are in consistency, then the result is accepted. If the results are not consistent, then the classifier C of the second level does another round of classification, and the result is decided by a majority voting strategy. In other words, if the classification result of the classifier C is the same with any one of the results of the two classifiers of the first level, then the result is accepted; otherwise, the result is decided as unknown.

In addition to basic bibliographic information (title, author, publisher, publication date, and keywords), abstracts, full text and bibliographies tend to be added to the bibliography fields of each document for a more comprehensive document description. In this paper, the researcher tried to investigate which combination of document content fragments is the best for achieving the highest efficiency and generating satisfying classification results in a multi-level automatic book classification system. Fifteen combinations of document fragments mentioned above were generated and used to train five classifiers. Based on the comparison of efficiencies of the commonly used classifiers, the three most efficient classifier algorithms,

Naïve Bayes, SVM and Decision Tree C4.5, were adopted, and KNN and the majority voting strategy were incorporated.

Results

Three datasets including thesis and dissertations, online bookstore catalogs, and online bookstore catalogues with category decreases and quantity increases were tested, and the experiment results of various classifiers are discussed as below.

Dataset of Thesis

When collecting training and testing corpora, the thesis of an university in central Taiwan submitted during the period of 2006-2010, were collected, including 12 in Horticulture, 12 in Entomology, 17 in Animal Science, 18 in Soil and Water Conservation, and 20 in Chinese Literature, with a total of 79 degree papers. Three classification algorithms including Naïve Bayes, SVM and MultiScheme of WEKA (with the majority voting strategy) were adopted for the classifiers. Four datasets including abstracts, texts, bibliographies and catalogues were included, and 15 permutation combinations of different fragments were generated.

At the first level, Naïve Bayes Classifier and SVM Classifier conducted classifications. If the results were consistent, it was accepted. Among the combinations of various data fragments, the best accuracy rate of the first level was 97.79%. At the second level, the inconsistent data records were transferred to Decision Tree C4.5 (J48) or KNN for another round of classification. The results were compared with the classification results of the two classifiers of the first level, and the final result was decided with a majority voting strategy for the attribution of categories. Among the combinations of data fragments, Decision Tree C4.5 and KNN reached the accuracy rate of 100.00%.

Dataset of Online Bookstore Catalogues

Since the hierarchical automatic classification systems with a majority voting strategy reached a satisfying accuracy rate on cataloging the experimental data of thesis and dissertations, for further verifying the efficiency of this multi-level classification system on cataloging books, more than fifty different books in five categories including Business and Finance, Interior Design, Medical and Health, Computer Science and Recipes published

during 2011 and 2014 were chosen as the dataset, with a total of 250 books. The data was tested and classified for three combinations of data fragments, including (1) abstracts, (2) catalogues, and (3) abstracts and catalogues.

In the experiment of the first-level classifiers, the commonly used machine learning approaches, Naïve Bayes Classifier and SVM, were adopted. In the experiment datasets, no matter it is abstracts, catalogues or abstracts plus catalogues, the accuracy rates of classification were all over 98%. In the parallel experiment of the second-level classifiers KNN and Decision Tree C4.5 (J48), the overall accuracy rate of Decision Tree C4.5 was higher than kNN, especially for the "Catalogues" or "Abstracts plus Catalogues" parts, the accuracy rate of C4.5 could reach an accuracy rate of 100%.

Online Bookstore Catalogues with category decreases and quantity increases

To further investigate whether the efficiency of this hierarchical automatic classification system with the majority voting strategy would be influenced under the circumstance of a fewer number of categories but a larger amount of books, the researcher of this study added 300 books published between 2011 and 2015 in each of the three categories of Business and Finance, Interior Design, and Medical and Health, with a total of 900 books. A classification experiment was conducted on the bibliographic data including (1) abstracts, (2) catalogues, and (3) abstracts plus catalogues.

At the first level, the commonly used and efficient machine learning approaches, Naïve Bayes Classifier and SVM, were adopted to train the three classifiers, for generating classification results on (1) abstracts, (2) catalogues, and (3) abstracts plus catalogues. The accuracy rates of classification through the algorithms of Naive Bayes Classifier and SVM were over 94%. It suggested that the corpora of Abstracts and Catalogues were of importance and representativeness. After the first-level classification, the efficient algorithm of Decision Tree C4.5 (J48) was used for the second-level classification on (1) abstracts, (2) catalogues, and (3) abstracts plus catalogues. The accuracy rates were 99.3%, 99.4% and 99.6%, respectively.

Conclusion and suggestions for future studies

The research results suggested that, in the dataset of online bookstore

catalogues, compared with the traditional classification systems with one single classifier, a multi-level automatic classification system with the majority voting strategy could reach an accuracy rate of 99%, thus enhancing effectively the accuracy rates of classification. Furthermore, at the second level of multi-level classification systems, since the quantity of data to be classified had decreased largely, it could improve the problem of taking too much time when the traditional classification systems with one single classifier were processing a large quantity of data. From the experiment of the second-level classification with a majority voting strategy, we can see that the efficiencies of classification models were influenced by the numbers of original sentences and phrases of data records. That is, if the original sentences and phrases were too short (for example, just one or two sentences in Chinese), then after data dimension deduction, the quantity of collected data representation would not be representative, resulting in decreasing or fluctuation of classification efficiencies through a multi-level classification system. However, if the number of original sentences and phrases of a bibliographic record was more than five, then through the multi-level classification system, the accuracy rates of classification would be enhanced, resulting in a more stable performance of the classification model.

It is suggested that researchers of future studies should conduct experiments in real situations of libraries and examine the outcome numbers, for exploring the feasibility of using automatic classification systems to do classification for library collections. In addition, several suggestions were made for future studies.

1. To investigate whether a multi-level automatic classification system has similar or better classification outcomes when processing two or three-level classification numbers, in addition to the practical approach of enlarging datasets, it is also suggested to test the classification results on the classification number systems now used in libraries.
2. In addition to the approaches of TF and TFIDF, different choices of document representation should be compared and tested.
3. With the prevalence of online resources, data in various languages in different countries around the world has been increasing dramatically every day, and it is suggested that subsequent experiments should expand to test on datasets in different languages.

附錄一　網路書店之書目資料範例

主要分成書名、目錄及摘要。

書名：

30 歲後你會站在哪裡？ 2：想脫離「貧性循環」，就必須打造自己的「印鈔機」

目錄：

【作者序】吃了一百天泡麵省下的錢，一天內就貶掉？

PART1「貧性循環」像高血壓，無聲無息卻要人命

01 窮人永遠不知什麼是「貧性循環」？
02 你再繼續「踩油門」，早晚會翻車
03 你口袋裡的一百元，只是一張紙
04 每天買地攤貨和吃泡麵，是賠本生意
05 需求和欲望，便當和名牌包
06 窮人沒有資格享受「欲望」
07 害怕沒錢的人，才是窮人
08 生老病也是「貧性循環」的推力
09 寫手和作家，合音和歌手的不同
10 小心！貧窮是會遺傳的病毒
11 父母的支援，是看不見的負債
12「貧性循環」像高血壓，悄悄地吃掉你的人生

PART2 高 CPI 和負利率時代，是個不斷蔓延的沙漠

13 窮人都是埋頭苦幹的傻子
14 富人的財富，都來自非勞務收入
15 太在意薪水的人，都是不愛錢的人
16 你付出的只是價格，得到的卻不是價值
17 薪水制，是窮人的母親
18 加薪只是一種安慰劑
19 月薪超過九萬，不是窮人，但也只算是「貧民」
20 擁有房子，不見得就是「有產階級」
21 這個時代，每個人都口渴卻又渴不死
22 別想用薪水趕上 CPI 這輛高鐵列車

摘要：

繼暢銷書《三十歲後，你會站在哪裡？》出版後，
超人氣職場勵志作家「狄驤」，
為了協助年輕人打造未來，再次提筆上陣，強勢推出第二集！
脫貧致富的關鍵，不在節儉省錢，
而在於增加「非勞務收入」。
如果三十歲前，你仍看不清這事實，
即使你每天吃泡麵省錢，你也只會「愈省愈窮」。
相信我，當你出國度假或蹺腿看電視時，
你的印鈔機仍日夜不停地為你印鈔，
這才是脫離「貧性循環」的最佳方案。
資本主義的壞處一堆，但好處就是：
不管你是誰，人人都有機會打造自己的印鈔機。
如你不想三十歲後，仍站在提款機前領死薪吃泡麵，
那麼，你最好三十歲前就開始打造印鈔機，
否則，你一輩子都只能活在「錢不夠用」的狗籠裡。
如果你沒有祖產，也沒有富爸爸，更不可能和老闆有一腿
與其在資本主義的巨輪下墮落

不如及早規畫自己的印鈔機藍圖，靠著源源不絕的系統性收入
在三十歲之後，成功晉升有錢人的行列

財經雜誌常報導，許多三十幾歲的年輕人，靠著玩股票和投資理財，就能提早享受財務自由的賺錢故事。也告訴年輕人：只要你努力工作，手邊存點資金，再一股腦地把錢投入股票市場，就能開啟另一個錢滾錢的帳戶。

然而，如果你真的以為，照著這些「勵志故事」理財就能致富，你只是把自己推入貧性循環的陷阱中，永世不得超生罷了。

當你還在被薪水制奴役時，有錢人早已搶先一步，根據自身的專業和發展優勢，替自己量身打造一台，全天候運轉、不需要休息的印鈔機。

本書將破解各種理財迷思，點醒你對金錢的錯誤認知，又如迎頭痛擊般，要你好好認清現實，規畫屬於自己的印鈔機藍圖，才能真正扭轉你的貧窮命運。

切記，不論你是藍領或白領，甚至是速食店或飲料店的工讀生，三十歲前，你們就該擁有，即使失敗也不會放棄的意志力，並靠著耐心持續打造你的印鈔機，在三十歲後，它將能為你賺進大把鈔票，順利脫離貧性循環的枷鎖。