基於語義之多層式圖書自動分類實證研究
Empirical Study of Semantic-based Multi-layered Automatic Book Classification

陳雨沛

Yu-Pei Chen

國立中興大學圖書資訊學研究所碩士生

Master, Graduate Institute of Library and Information Science, National Chung-Hsing University

郭俊桔

June-Jei Kuo

通訊作者：jjkuo@dragon.nchu.edu.tw

國立中興大學圖書資訊學研究所副教授

Associate Professor, Graduate Institute of Library and Information Science, National Chung-Hsing University

【摘要 Abstract】

本研究旨在探討多層式圖書自動分類系統應用於圖書館分類編目的工作流程，並嘗試導入語義概念作為改進分類成效之策略，目的是為解決人工分類的一致性、分類效率等問題，改善分類編目的品質，並針對資料量或文件特徵量可能不足之課題，利用 Word2Vec 能夠保留目標詞與其上下文之間的語義關係之特性，將帶有語義的詞彙擴展特徵詞彙，藉此改善分類成效。本研究蒐集的資料取自大學圖書館 HyRead ebook 中文電子書，共 17,650 冊，分類類別為《中文圖書分類法》圖書分類號，從中隨機挑選的圖書分類號涵蓋十大類，共 581 個圖書分類號，並使用四種分類器應用於多層式圖書自動分類系統；於語義方面，

實驗使用 Word2Vec 訓練語料，並建構類索引典之同義詞詞典，再以擴充的方式擴展特徵詞彙，最後以正確率評估分類效能。實驗結果顯示將多層式圖書自動分類系統應用於圖書館分類編目具有更好的分類成效，並且所提出的策略確實能夠提升圖書分類的準確度，經詞彙擴充之後，正確率為 62.50%，有效減少多層式分類最終仍無法判斷的數量。然而，對圖書自動分類而言，存在多個主題的圖書內容可能是影響分類成效的因素之一，若將其他圖書館系統所給予的圖書分類號視為正確類別，得到的正確率接近 70%。因此，在未來圖書館實務上，尚須考慮到存在多個主題的圖書內容對圖書自動分類的影響。

This study aims to explore the application of the multi-layered automatic book classification using the majority voting strategy in a library environment, and proposes a semantic-based approach as a strategy to improve the classification performance. The proposed strategy uses Word2Vec, which can extract the deep semantic relationships between words, to expand words features for improving the classification performance. The collection of data from HyRead ebook, which was ordered by the university library, was utilized as the training and testing corpus, including book titles, table of contents, abstracts. In total, 17,650 books were collected. And the classification classes were the book classification numbers. A total of 581 classification numbers were selected randomly from New Classification Scheme for Chinese Libraries. Naïve Bayes, SVM, Decision Tree C4.5, and kNN were applied to the multi-layered automatic book classification. Regarding the proposed semantic-based approach, this study used Word2Vec as a tool for training word embedding. First, a thesaurus was built by the training results, and next the word features of the data set for classification were expanded. Under the principle of accuracy, experimental results showed that the performance of the multi-layered

automatic book classification outperformed the traditional automatic book classification in a library environment. And the strategy proposed in this study could indeed improve the accuracy of book classification, achieving 60%. After expanding the word features of the data set, the accuracy rate promoted 1.79%. The above results proved that Word2Vec tool was helpful for automatic book classification. It could effectively reduce the number of unable classified books, which were classified through the multi-layered automatic book classification. However, book content with multiple topics may be one of the factors that affect classification effectiveness for automatic book classification. If the book classification number, given by other library systems, was regarded as the correct class, the accuracy rate could be approximately 70% by the semantic-based multi-layered automatic book classification using the majority voting strategy. The accuracy rate could be further increased by 7.14%. Therefore, it is necessary to consider the effect of book content with multiple topics for automatic book classification in the future library practice.

【關鍵詞 Keywords】

分類號、階層式、圖書自動分類、Word2Vec
Classification number; Layer style; Automatic book classification; Word2Vec

壹、前言

編目（Cataloging）是圖書館為達成組織與整理資訊物件的方法，為圖書館建立目錄的過程，目的是將足以識別一件圖書資料的有關項目找出來，按照一定的規則加以排列，讓使用者能依據要項的描述，對該件圖書資料有所認識，而對其作有效利用（王梅玲、謝寶煖，2014）。編目分為記述編目（Descriptive Cataloging）及主題編目（Subject Cataloging），記述編目是對一件圖書資料作形體上的分析（Physical Analysis），以描述此一圖書資料形體上的特點，讓使用者透過各著錄項目的描述來瞭解該圖書資料的外在特性；而主題編目是對一件圖書資料內容的主題加以分析，以決定其標題（Subject Heading）與分類號（Classification Number），讓使用者掌握圖書資料內容的主題或學科性質。由於主題編目相較於記述編目較為主觀，在實務面難以達到完全一致性。

長期以來，主題編目的品質、準確性備受質疑與探討，尤其分類號的一致性。Subrahmanyam（2006）指出以抄編的方式可以有效改善分類號不一致的情形，另一相關研究趨勢便是將文件自動分類應用於圖書分類。隨著自動化技術的發展，圖書館專家認為自動分類將能使分類更有效率、快速且正確，加上處於資訊爆炸的時代，不僅標題表或索引典（Thesaurus）逐漸難以應付新興的詞彙，文件分類也愈來愈難有效率地管理、花費的成本也愈高，使得分類的品質出現降低的疑慮，因而開始投入圖書自動分類領域，藉由自動化技術將分類程序標準化，企圖改善圖書館館藏分類號的一致性，解決人工分類產生的問題，並提升圖書分類的準確度與分類編目的品質。

自 1990 年代開始，基於機器學習（Machine Learning）方法逐漸成為自動文件分類之主流，廣泛應用於資訊科學、圖書資訊學等學科領域，相關研究專注於機器學習方法的階段及步驟，嘗試各種方法或導入已經發展成熟的理論，以取得良好的分類結果。而在圖書自動分類之相關

研究，較專注在圖書特徵或敘述資料及詮釋資料（metadata）對自動分類成效的影響，以及嘗試在文件分類流程中找出適合圖書分類的方法。如在分類器方面，過去研究多以傳統單一分類器或多重分類器進行自動分類，少數研究則透過階層的方式結合多個分類器進行自動分類，如吳慧貞（2015）、林昕潔（2006）、郭俊桔（2018）。無論是自動文件分類或圖書自動分類之相關研究，時常受限於分類對象本身資訊量或特徵值不足等因素影響分類成效，曾元顯與莊大衛(2003)、黃嘉宏(2008)、Liang 等人（2006）、Shen 等人（2018）等相關研究進而嘗試使用其他輔助工具，用以擴展文件量或特徵值，作為改進分類成效之策略。

從圖書館分類編目工作的現況及圖書自動分類之研究的角度來看，仍有下列五項待克服的課題：

一、圖書分類號存在不一致性或準確度不高，影響使用者查找圖書資料的效率。

二、針對圖書自動分類之研究，用來作為訓練分類模型的資料內容，尤其圖書題名對自動分類的幫助不大。然而，圖書題名對館員、使用者而言，是圖書分類或辨識該圖書資料是否為需要或想要的資料之重要線索。

三、在訓練分類器方面，傳統單層式圖書分類於測試資料的正確率介於 36%至 62.86%之間，對於分類編目品質的提升助益仍有改善空間。

四、為解決資料量、文件特徵量不足的現象，相關研究進而提出改進策略，以改善分類成效，但所提出的策略方法仍存在限制。

五、在資訊爆炸的時代，圖書館分類編目的對象已經不僅僅是傳統實體館藏，如何有效率地分類管理不同類型的館藏，都是重要的研究課題。

本研究為解決課題一、三、五，本研究採用郭俊桔（2018）之多層式分類進行圖書分類，以圖書館館藏作為分類對象，探討多層式圖書自動分類系統對依據圖書分類法給予分類號之圖書分類是否有相同或更好的分類成效，解決人工分類的一致性、分類效率等問題。針對課題二於資料內容的選擇方面，本研究蒐集圖書題名連同其他圖書片段內容，透過不同片段內容的組合，探討適合用於自動分類的圖書特徵，同時觀察使用圖書館館藏提供的題名於自動分類的表現。關於改進分類成效之策略的課題四，過去研究利用外部詞典詞彙與詞彙之間的關係作為改進自動文件分類成效之策略，但使用現存詞典的劣勢為詞彙具領域性，詞彙的更新與維護無法及時趕上時代背景帶來的變化，因此，本研究嘗試使用由 Mikolov、Chen、Corrado 與 Dean（2013）提出的 Word2Vec 建構同義詞詞典，同樣能夠保留詞彙與詞彙之間的語義關係，具即時、動態等優勢，用以作為改進分類成效之策略的輔助工具。

貳、 文獻探討

一、 圖書自動分類之研究

文件分類是根據文件的內容主題給予類別的工作，傳統上由人工閱覽文件，根據其主題大意，給予適當的類別標示，若由機器自動執行，則稱為自動文件分類（曾元顯，2002）。Mirończuk 與 Protasiewicz（2018）將自動文件分類流程分為六個階段：1.資料蒐集，2.資料分析與給予標籤或類別，3.特徵建構及加權，4.特徵選擇，5.訓練分類模型，6.分類結果評估，特徵建構及加權、特徵選擇、訓練分類模型為自動文件分類最受關注的議題。

圖書自動分類研究可視為文件分類於圖書資訊學領域之實務應用，尤其是圖書館的分類編目。分類編目的品質將影響圖書館的服務、館藏管理的效率等層面。隨著網際網路的發展，處於資訊爆炸的時代背景，除了傳統圖書分類法是否能有效率地管理倍增的文件量而受質疑之外，

人工分類的準確度也開始動搖，接踵而至的挑戰使得圖書館工作愈來愈難以負荷。

從文件分類流程的角度來看圖書自動分類，Yi（2007）指出三項主要的挑戰：1.圖書分類法的框架、2.知識來源、3.分類方法／模型。首先，從圖書分類法的特性來看自動分類之可行性，需考量每個類別的資料量是否足夠作為訓練資料，以及所取得的圖書資料可能會有版本不一的疑慮。知識來源，指從書目紀錄或訓練資料中取得知識，以訓練學習系統制定出分類規則，Nonaka（1994）將知識分為內隱知識（tacit knowledge）與外顯知識（explicit knowledge），內隱知識強調可意會不可言傳、高度個人化的知識，外顯知識則是可取得、能透過文字清楚表達的知識，例如：書籍。站在圖書分類的角度，不易取得的內隱知識會是影響圖書分類正確性之潛在因素之一。然而，圖書資訊學領域所發展的控制詞彙，對於探勘內隱知識具有潛在能力，例如：美國《國會圖書館標題表》（Library of Congress Subject Headings，LCSH）。除此之外，文件分類相關研究使用的資料集所屬的學科主題較為集中，可能無法適用於學科廣泛的圖書分類，因此，若能建立出適合用於圖書分類的訓練資料，將為圖書自動分類研究領域帶來重要貢獻。

以圖書分類為對象的文件分類研究，最早開始於 1990 年代。Larson（1992）、Frank 與 Paynter（2004）同樣以《國會圖書館分類法》（LCC）為編目依據的機讀編目格式（Machine-Readable Cataloging Format，MARC）書目紀錄作為分類訓練資料及測試資料，得到的分類正確率由 46.6%提升至 55.32%，以《國會圖書館分類法》（LCC）具以等級表示類目的從屬關係之特性來看各層正確率，最高可達到 80.27%（Frank & Paynter, 2004），Frank 與 Paynter（2004）亦藉由將訓練資料割分為不同大小的量，觀察分類的正確率及時間長度的變化，發現當資料量愈大，得到的正確率愈高，但相對需要的訓練時間愈長。

Á vila-Argüelles、Calvo、Gelbukh 與 Godoy-Calderón（2010）僅使

用圖書題名進行自動分類，在實驗中混合 Lesk 投票典範（Voting Scheme）、詞頻（Term Frequency，TF）兩種文件分類方法，並嘗試不同權重值及邏輯組合方法，以實證題名對於正確分類的貢獻。Á vila-Argüelles 等人（2010）以 489,726 冊圖書的題名作為訓練樣本，122,431 冊作為測試資料，僅使用詞頻進行分類時，約 28%的圖書被分類到正確的類別，而在不同權重值及邏輯組合方法中，得到最好的正確率為 36%，提升了 8%，表示所提出的方法有效改善僅使用圖書題名之分類成效，未來可再透過其他圖書特徵，比較不同圖書特徵對正確分類的貢獻程度。

林昕潔（2006）、陳信源等人（2009）以文件分類為基礎進行圖書分類，輔以專家的經驗進行特徵選擇及賦予權重值，並且加入圖書的敘述資料及詮釋資料，再透過支持向量機（Support Vector Machine，SVM）訓練分類模型，以完成圖書分類。研究所使用的圖書資料取自「博客來網路書店」偵探／懸疑小說、科幻／奇幻小說、愛情文藝小說等三個類別，每個類別各取 900 冊，共 2,700 冊，並使用九折交叉驗證法評估分類模型對新進未知資料進行預測分析時的表現。經由專家選擇過濾特徵並將這些特徵加權之後，針對圖書敘述資料的分類成效由 83%提高到約 90%，若再加入詮釋資料，則進一步提高至 95%，顯示加入詮釋資料能夠補強特徵值不足的問題。

針對可能因特徵值不足而導致分類成效不彰的情形，黃嘉宏（2008）於圖書分類導入資訊檢索及相關領域「擴展」概念，藉由 Google 搜尋引擎之查詢結果擴展特徵值，利用圖書題名作為查詢語句，再將查詢結果中的網頁標題及部分描述內容與原本的圖書題名特徵結合，並以圖書作者資訊為輔助策略，進行自動分類實驗。實驗資料取自國家圖書館館藏書目資料，包含訓練資料 19,949 筆、測試資料 4,911 筆，涵蓋《中文圖書分類法》兩大類，共 156 個分類號。實驗結果顯示所提出的策略確實能夠提升圖書分類的準確度，得到最佳的正確率為 62.86%。

在分類器方面，林昕潔（2006）、陳強強（2018）等相關研究嘗試使用單一分類器、多重分類器或層次分類器進行自動分類，或是透過多個分類器進行自動分類，再依各個分類器得到的結果進行比較，進而探討何種分類器較為適合用於圖書分類。陳強強（2018）提出的層次分類器比傳統單層分類器更加重視類目之間的層次關係，其在每一層類目均建構出分類器，如此可以降低計算複雜度，適用於層次類目較多的情況；在研究中，亦嘗試屬於人工類神經網路（Artificial Neural Network，ANN）的極限學習機（Extreme Learning Machine，ELM）訓練分類器，企圖改善傳統機器學習演算法存在訓練速度慢、學習能力差等問題，實驗結果顯示極限學習機的正確率高於支持向量機、反向傳播類神經網路，且訓練時間僅為其他兩種分類器的三分之一。

二、 多層式圖書自動分類系統

多層式圖書自動分類系統由郭俊桔（2018）提出，此研究嘗試結合多種分類演算法，並以多數決（Majority Voting）的方式決定圖書類別。多層式分類法，第一階層包含兩種分類器，而第二階層之後為一種分類器，各階層使用的分類器皆不相同。分類流程操作以二階層分類為例（參見圖 1），包含三種不同的分類器 A、B、C，當輸入分類實例之文件表徵之後，分別由第一階層的分類器 A、B 執行分類，而後針對兩者預測類別逐筆進行比對，若兩者的預測類別為一致，則接受該預測類別為最終類別；若存在預測類別為不一致的資料，則繼續交由第二階層的分類器 C 執行分類，再將分類器 C 得到的預測類別與第一階層分類器 A、B 不一致資料的預測類別進行比對，並使用多數決策略決定其最終類別，意即當分類器 C 的預測類別與第一階層兩種分類器中任何一種分類器的預測類別一致時，則判定為最終類別，最後再將最終類別與實際類別核對，計算分類正確率。

實驗結果顯示針對少量資料進行自動分類時，使用多數決策略之多層式圖書自動分類能夠有效地提高分類正確率，比傳統圖書分類具

有更佳的分類效能，得到最佳分類器與階層之組合依序為單純貝氏、支持向量機、決策樹（Decision Tree，DT），而最佳圖書片段內容的組合為摘要＋目次。

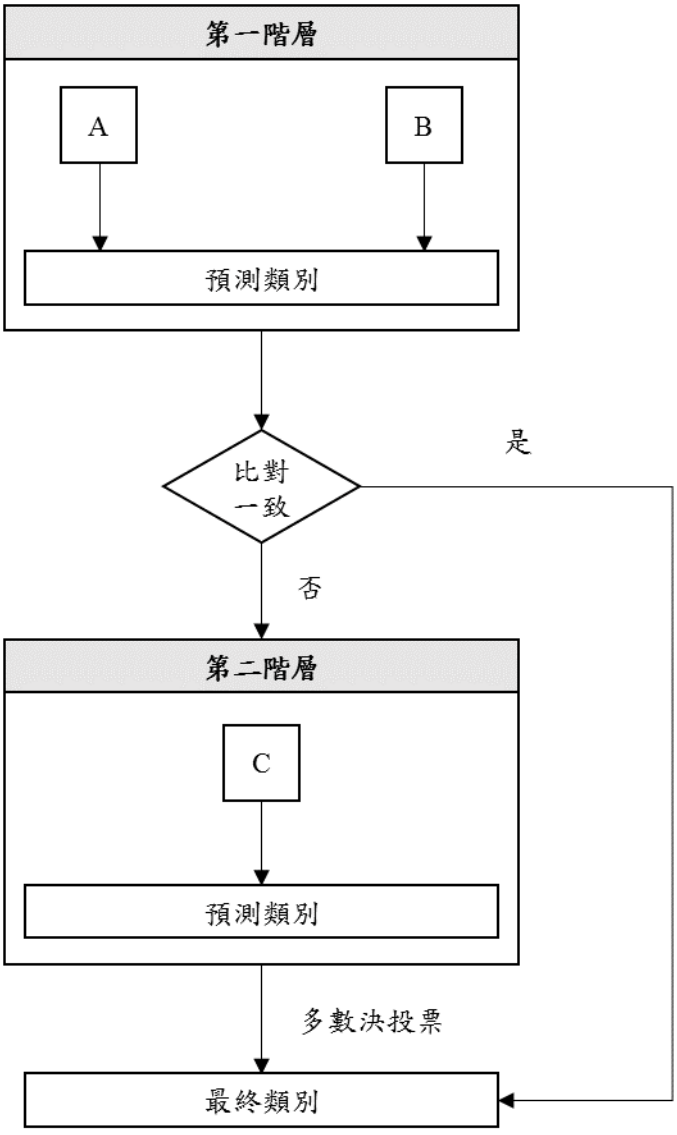

圖 1 多層式分類流程圖

三、 分類成效評估方式

常見的評估方式有交叉驗證（Cross-validation）和評估指標，交叉

驗證主要用於評估分類模型對新進未知資料進行預測分析時的表現，評估指標則應用於資訊檢索領域，常被用來比較不同的技術或系統之間的成效。

交叉驗證較為常見的方法為 k 折交叉驗證（k-fold Cross-validation），做法為將所選定的資料集分為 k 個子集，之後重複做 k 次保留交叉驗證法，每一次都從 k 個子集中選定一個作為測試集，其餘 k-1 個子集作為訓練集，最後再取這 k 次結果的平均值作為整體的評估結果。

常見的評估指標包含精確率（Precision）及召回率（Recall）。從文件分類的角度來看，精確率指在分類器分類類別的總數中為實際類別的比例，而召回率指實際類別的總數中被分類器正確分類的比例，每一個類別所有的文件將產生四種結果（表 1）。真陽性（True Positive，TP）和真陰性（True Negative，TN）為正確分類結果，即預測類別與實際類別相符合，而假陽性（False Positive，FP）指不屬於該類卻被系統分到該類者，假陰性（False Negative，FN）為不屬於該類且沒有被系統分到該類者。

表 1 文件數量分布表

系統分類結果 / 實際類別	系統分為該類	系統不分為該類
屬於該類別	TP	FN
不屬於該類別	FP	TN

四種結果統計之後，即可計算精確率（公式 1）、召回率（公式 2）、正確率（Accuracy）（公式 3）及錯誤率（Error）（公式 4）：

$$\text{Precision (P)} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad \text{（公式 1）}$$

$$\text{Recall (R)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad \text{（公式 2）}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad \text{（公式 3）}$$

$$\text{Error} = \frac{FP + FN}{TP + TN + FP + FN} \quad \text{（公式 4）}$$

此外，為了能同時兼顧精確率及召回率，而有 F 度量（F-measure）評估指標（公式 5、公式 6、公式 7）：

$$F - \text{measure} = \frac{2 \times P \times R}{P + R} \quad \text{（公式 5）}$$

$$\text{Micro} - F = \frac{2 \times \sum_{i=1}^{C} TP_i}{2 \times \sum_{i=1}^{C} TP_i + \sum_{i=1}^{C} FP_i + \sum_{i=1}^{C} FN_i} \quad \text{（公式 6）}$$

$$\text{Macro} - F = \frac{1}{C} \sum_{i=1}^{C} \frac{2 \times TP_i}{2 \times TP_i + FP_i + FN_i} \quad \text{（公式 7）}$$

其中，C 是類別總數，i 代表某一類別。由於 Micro-F 是全部文件一起累加統計且不區分類別，因而容易受到少量的大類別表現好壞的影響；而 Macro-F 考慮每個類別的成效後再取平均，因此容易受到大量的小類別影響。將兩種平均值列出並比較，可以瞭解大多數文件的分類成效（Micro-F），以及大多數類別的分類成效（Macro-F）（曾元顯，2002）。

最後，整體錯誤率（overall error）使用公式 8 計算：

$$\text{overall error} = \frac{\sum_{i=1}^{C} (FP_i + FN_i)}{\sum_{i=1}^{C} (TP_i + FP_i + FN_i + TN_i)} \quad \text{（公式 8）}$$

四、 詞向量模型

詞向量（word vector 或 word embedding）指使用一個向量來表示每一個詞，即將詞轉換成實數向量的表示，藉此利用向量空間計算詞與

詞之間的相似度，若轉換對象為文件，則稱文件向量（document vector），同樣利用向量空間計算文件與文件之間的相似度，此種表示法廣泛應用於資訊檢索領域。詞袋模型（Bag of Words model，BOW）為文件表示法的一種，此表示法基於獨立假設，每一個詞彙皆是獨立的單位，並不考慮其相依性，優勢在於此表示法簡化了許多文件自動處理的計算，因而廣被採用，例如：空間向量模型（Vector Space Model，VSM）（Salton, Wong & Yang, 1975）。

然而，傳統以類神經網路語言模型訓練詞向量的方法，存在時間過長的問題，為了改進訓練詞向量的方法、解決時間過長等問題，Mikolov 等人於 2013 年提出 Word2Vec，Word2Vec 是一個產生詞向量的開放原始碼工具，使用前饋式類神經網路語言模型（Feed-forward Neural Net Language Model，NNLM），並以非監督式學習方法建構這些隱含前後文特徵的單詞向量。Word2Vec 基於以下假設：計算兩個單詞在語義上為基礎的相似度，決定於目標詞的前後字或詞分布是否相似。因此，若要看出單詞向量空間中的規律性，其強弱程度取決於是否具有足夠向量維度的大型語料來訓練 Word2Vec 模型。首先會依據所輸入的純文字訓練語料建立一個詞彙表（vocabulary），之後依據上下文隱含的特徵計算出單詞的詞向量，透過多單詞視窗的方式，來保留目標詞與其上下文之間的語義關係，以建立特徵。

Word2Vec 包含 CBOW（Continuous Bag-of-Words）、Skip-gram 兩種訓練詞向量的模型，兩者皆包含輸入層（input layer）、投影層（projection layer）、輸出層（output layer），模型架構見圖 2。

（一） CBOW

CBOW 模型架構與前饋式類神經網路模型相似，兩者不同之處在於 CBOW 去除非線性隱藏層（Non-Linear Hidden Layer），並且在輸入層的所有單詞皆共享投影層。此模型的輸入為上下文範圍內的詞（w_{t-2}，

…, w_{t+2}）之詞向量，輸出則是目標詞（w_t）的詞向量，即利用上下文範圍內的詞來預測目標詞。

（二） Skip-gram

Skip-gram 模型是利用每個目標詞（w_t）之詞向量作為具有連續投影層的對數線性分類器之輸入，輸出為上下文範圍內所包含的詞（w_{t-2}, …, w_{t+2}）之詞向量，即透過目標詞來預測其上下文範圍內的詞。

相較之下，Skip-gram 模型的訓練時間較長，但其對罕見字詞的處理具優勢，且適合作為文件分類之特徵，因此，本研究嘗試使用 Word2Vec 之 Skip-gram 模型訓練詞向量，用以建構類索引典的同義詞詞典，並以訓練分類模型所挑選出來的特徵詞彙從中取出相似度高的語義相似詞，建立後續實驗使用的詞表，進而探討導入語義概念對圖書自動分類的影響。

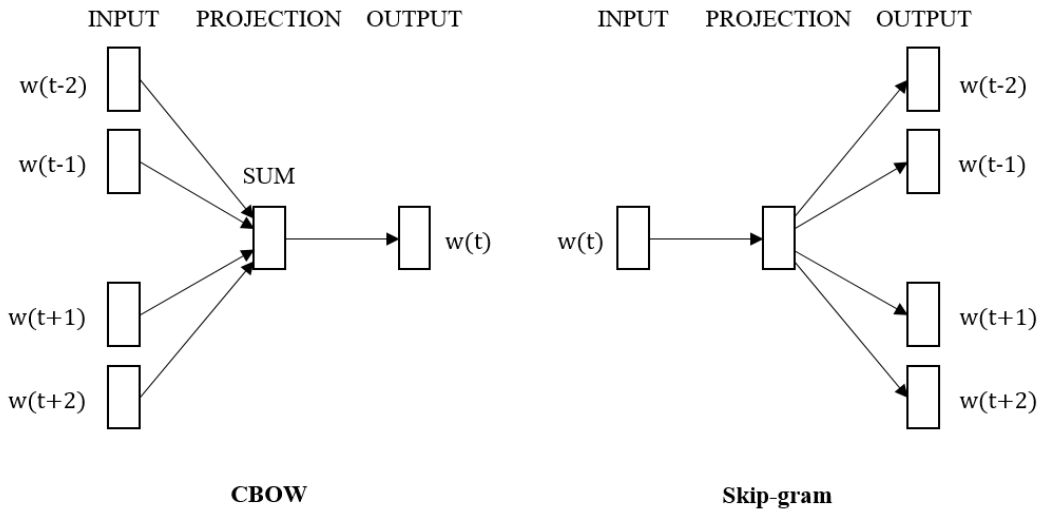

圖 2 Word2Vec 模型架構圖

資料來源：“Efficient estimation of word representations in vector space,” by T. Mikolov, K. Chen, G. Corrado, & J. Dean, 2013. Retrieved from https://arxiv.org/abs/1301.3781v3.

參、研究方法與設計

一、 研究架構

本研究的研究架構分為四個模組：1.資料前置處理、2.語義詞彙處理、3.圖書自動分類、4.分類結果評估。首先，將蒐集的資料進行資料前置處理，資料蒐集的片段內容包含圖書題名、目次、摘要，接著將資料前置處理完成的資料作為訓練分類器使用的訓練資料，以及作為Word2Vec 詞向量工具使用的訓練語料。語義詞彙處理模組的主要目的，是為建構基於 Word2Vec 之同義詞詞典，並建立轉換或擴充詞彙使用的詞表，用於將 Word2Vec 訓練得到的相似詞加進分類資料中。之後，將原始分類資料及經過語義詞彙處理的分類資料進行圖書自動分類，並以十折交叉驗證法、正確率評估分類成效與分析分類結果。

實驗內容共分為四個階段，前兩個階段使用少量資料進行圖書分類，後兩個階段接續前兩個階段得到的初步結果，再以增加資料量的方式進行大量語料之圖書自動分類實驗。

（一） 第一階段

第一階段之目的為測試 WEKA wordsToKeep 參數，以及篩選何種片段內容之單項或組合能得到較佳的分類成效，同時觀察多層式分類器之間的搭配組合，以及階層數對分類正確率的影響。WEKA wordsToKeep 指嘗試保持的詞彙數量，在本研究用於挑選對於類別具代表性的詞彙。

此階段以不同片段內容之單項及組合作為分類資料，片段內容包含題名、目次及摘要，三者代號依序為（1）、（2）、（3），三者之組合共四種組合，各為：（1）＋（2）、（1）＋（3）、（2）＋（3）、（1）＋（2）＋（3）。

依據 AL-Nabi 與 Ahmed（2013）的分類演算法之評比結果，以及

過去常被應用於文件分類的分類演算法，挑選出分類成效佳的單純貝氏、支持向量機、決策樹、k 個最近鄰（k-Nearest Neighbor Algorithm，kNN）等四種分類演算法進行實驗，觀察多種分類器在階層之間的搭配組合，測試的階層數為兩階層及三階層。

（二） 第二階段

第二階段之目的為測試 Word2Vec 詞向量工具對圖書分類是否有幫助，藉由實作語義詞彙模組及圖書自動分類，評估方法的可行性。測試項目為語義相似詞的個數，以餘弦相似度（cosine similarity）由高到低依序選出作為特徵詞彙的語義相似詞，以找出得到較好的分類表現之最佳抽取個數。

（三） 第三階段

此階段之目的為找出適合電子書書目分類的類別資料量，並觀察每個小類資料量與分類成效之間的關係。透過類別資料量的測試，進一步觀察每個小類包含的資料量與分類成效之間的關係。所實作的分類資料分訓練資料、測試資料，以訓練資料訓練分類器，再以測試資料評估分類模型的效能。測試資料為從每個類別中挑選兩筆，其餘作為訓練資料。

（四） 第四階段

最後階段使用第一階段針對多層式分類器篩選出的最佳階層數、片段內容進行，使用第二階段篩選出的語義相似詞最佳抽取個數建立詞表。在詞彙擴充方面，針對多層式分類各階層不一致的資料加進語義相似詞，進而觀察何階層的不一致資料詞彙擴充之後得到的分類正確率最佳。

二、 研究對象

本研究使用的實驗資料取自中部某大學圖書館所訂購的電子資源

－HyRead ebook 中文電子書，蒐集的書目內容包含圖書題名（含副題名）、目次及摘要，目次、摘要於 HyRead ebook 中文電子書平台對應的名稱依序為「章節」、「簡介」。

蒐集的中文電子書書目分為兩部分，第一部分為實驗第一階段與第二階段使用，為少量資料，第二部分為實驗第三階段與第四階段使用，為增量資料。第一部分依《中文圖書分類法》從中隨機挑選 10 個分類號，再於 HyRead ebook 中文電子書平台蒐集該分類號 100 冊，合計 1,000 冊中文電子書。第二部分隨機挑選的分類號涵蓋《中文圖書分類法》十大類，共 581 個分類號，合計 17,650 冊。

為了使實驗貼近當前圖書館實際處理待編目的圖書情形，於第二部分蒐集的資料中，每個分類號所蒐集的資料量屬於不平均分配，增量資料之類別分布比例與國立中興大學圖書館 HyRead ebook 中文電子書之類別分布比例的差距介於 0%至 5%之間（統計期間：2019 年 1 月至 2019 年 12 月），差距最多者為「社會科學類」（5%）（圖 3）。

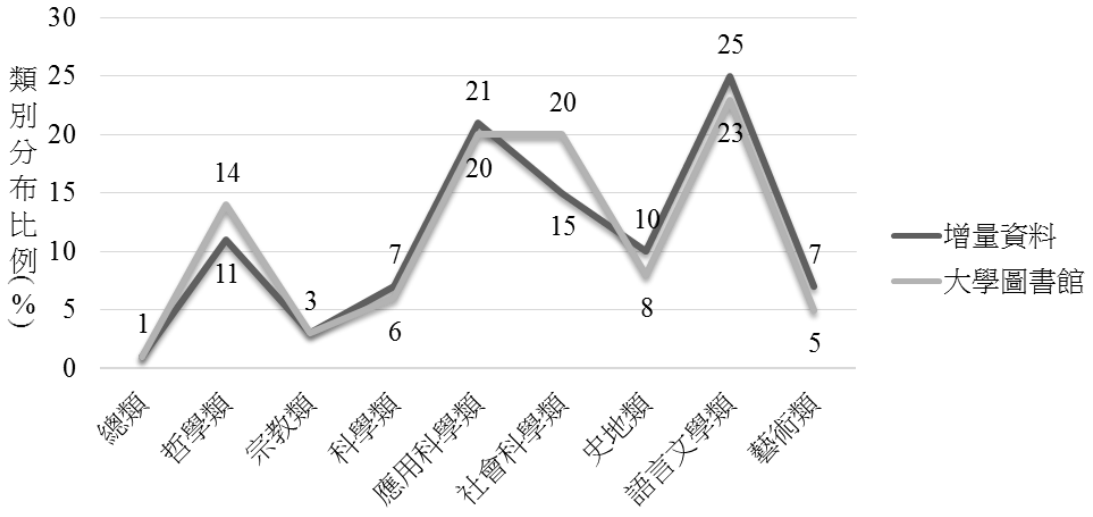

圖 3 增量資料與大學圖書館館藏主題之類別分布－HyRead ebook

三、 研究工具

（一） CKIP 中文斷詞系統

CKIP 中文斷詞系統由中央研究院詞庫小組所開發維護，此系統為一具有新詞辨識能力並附加詞類標記的選擇性功能之中文斷詞系統，包含一個約十萬詞的詞彙庫及附加詞類、詞頻、詞類頻率、雙連詞類頻率等資料，分詞依據為此一詞彙庫及定量詞、重疊詞等構詞規律及線上辨識的新詞，並解決分詞歧義問題。

由於本研究對象僅針對中文，因此選用中央研究院詞庫小組開發維護的 CKIP 中文斷詞系統。本研究實驗進行的資料前置處理及作為 Word2Vec 訓練語料的資料前置處理皆使用 CKIP 中文斷詞系統，進行中文的斷詞。

（二） WEKA

WEKA（Waikato Environment for Knowledge Analysis）是一種使用 Java 語言編寫的資料探勘機器學習軟體，是 GNU 協定下分發的開放原始碼軟體。WEKA 主要用於科研、教育和應用領域，是一套完整的資料處理工具、學習演算法及評估方法，其中包含資料視覺化的圖形化使用者介面，同時提供比較及評估不同學習演算法之性能的功能，除了提供大量學習演算法之外，亦提供資料前置處理工具，讓使用者能夠快速靈活地將已有的處理方法應用於新的資料集。在學習演算法訓練使用的資料集割分方面，提供多個選項供使用者選擇，例如：全部使用訓練資料、部分使用測試資料、交叉驗證法、自行設定百分比值將資料集分為訓練資料與測試資料等。

本研究使用的版本為 WEKA 3.8.2，於 2017 年 12 月發布。選用 WEKA 作為分類器訓練之工具的主要原因在於 WEKA 為開放原始碼軟體，在資料探勘機器學習方面具備強大功能且趨近成熟，在文件分類領域的成效表現上，相較於其他套裝軟體，WEKA 軟體具較佳的表現（AL-Nabi & Ahmed, 2013; Wahbeh, Al-Radaideh, Al-Kabi, & Al-Shawakfa, 2011）。

（三） Word2Vec

Word2Vec（Mikolov et al., 2013）的優勢在於模型的架構較為簡單且效能高，特別適合從大量語料中取得高精度的詞向量表示。不同於僅針對所要處理的文件之傳統詞向量技術，為文件中每個單詞建構出相對應的向量，於 Word2Vec，則是針對輸入的大量語料中出現過的每一個單詞，給予一個獨特的向量，每一個獨特的向量，都帶有大量隱含於單詞前後文的特徵（或用法），再運用餘弦相似度來找出語義相似的單詞。上述所描述的特性，為本研究選用 Word2Vec 的原因，用於建構類索引典的同義詞詞典。

（四） LibSVM

LibSVM(Library for Support Vector Machines)(Chang & Lin, 2011) 為一套支持向量機的函數庫，此函數庫的運算速度快且開放原始碼，支援 C#、Java、MATLAB、Net、Python、R 等多種程式語言。

相較於以序列最小最佳化（Sequential Minimal Optimization，SMO）演算法訓練的支持向量機分類器，使用 LibSVM 建構的支持向量機分類器速度更快，也能解決當資料量增大時，WEKA 軟體無法正常執行的問題。

肆、 研究結果與分析

一、 實驗資料概況

實驗資料經過資料前置處理之後，增量資料中的題名、目次、摘要之空值數量各為 15、147、15 筆，分布於「哲學類」、「應用科學類」、「社會科學類」、「世界史地」、「語言文學類」、「藝術類」。

Word2Vec 訓練語料使用題名＋目次＋摘要之組合，輸出檔案以二進制格式儲存。於實驗第一階段及第二階段使用的訓練語料詞彙量共 439,884 個，訓練得到的詞彙共 11,633 個，於實驗第三階段及第四階段

使用的訓練語料詞彙量共 5,831,881 個，訓練得到的詞彙共 63,268 個。

二、 少量少類別之多層式圖書自動分類

為實驗的前兩個階段，以少量資料、少類別進行圖書分類，所使用的四種分類器於 WEKA 對應的名稱如下：

- 單純貝氏（NB）：weka.classifiers.bayes.NaiveBayes
- 支持向量機（SVM）：weka.classifiers.functions.SMO（SMO）
- 決策樹（DT）：weka.classifiers.trees.J48（C4.5）
- k 個最近鄰（kNN）：weka.classifiers.lazy.IBk

（一） 決定類別詞彙量

在開始進行圖書分類實驗之前，先行測試 WEKA wordsToKeep 參數，測試的數量為 25、50、100，以四種單一分類器進行圖書分類，分類資料使用題名＋目次＋摘要之組合，再以十折交叉驗證法評估分類成效。

從圖 4 可觀察出類別嘗試保持詞彙量（wordsToKeep）對分類的影響，單純貝氏分類器的正確率呈現些微下滑的現象，但仍維持在 90% 以上，支持向量機分類器的正確率則是持續上升，決策樹分類器的正確率是四種分類器中最高者，並且無論 wordsToKeep 值為 25、50 或 100，正確率皆為 99.10%，然而，k 個最近鄰分類器的正確率在過了 wordsToKeep 值為 50 之後，明顯下滑，若從平均分類效能來看，wordsToKeep 值為 25 和 50 兩者僅相差 0.025%。後續圖書自動分類實驗 wordsToKeep 值以 50 進行。

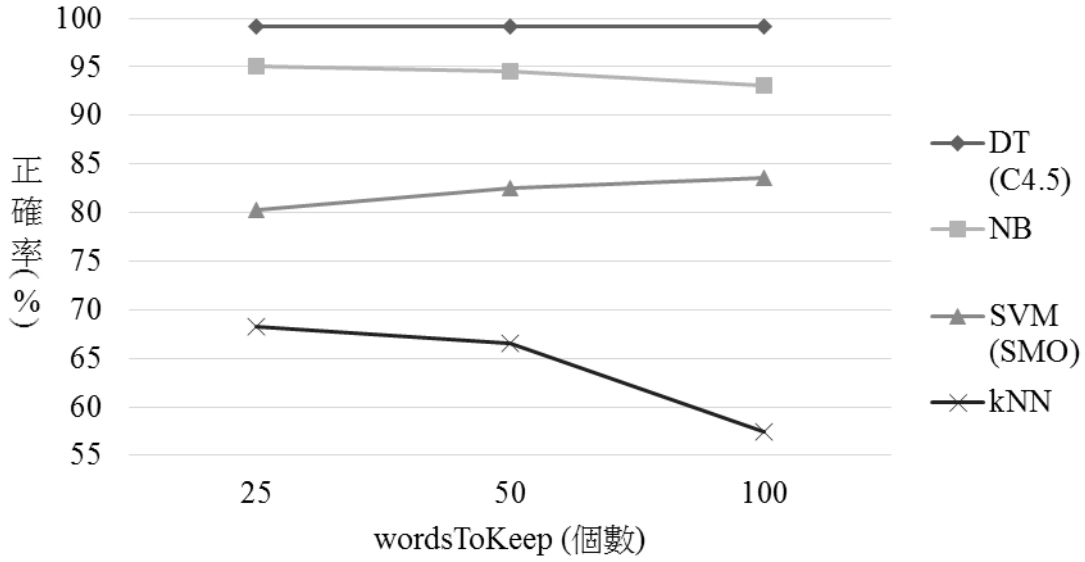

圖 4 類別詞彙量對分類的影響

（二） 片段內容於圖書自動分類

四種分類器於多層式分類法中對應的代號如下所示：

- 單純貝氏：A
- 支持向量機（SMO）：B
- 決策樹（C4.5）：C
- k 個最近鄰：D

以下階層組合代號中的「1」代表第一階層，「2」代表第二階層，「3」代表第三階層。

表 2 顯示題名、目次、摘要及其四種組合於多層式分類第一階層、第二階層與第三階層的表現，得到最佳分類成效的片段內容皆為題名，其次為題名＋目次＋摘要之組合，而目次在三個階層中皆為最差者。

雖然題名＋目次＋摘要之組合在多層式分類得到的正確率僅次於

題名，但題名＋目次＋摘要之組合在三階層分類 12 組中得到的最佳分類正確率與最差正確率相差 5.7%，而題名在三階層分類 12 組中得到的最佳分類正確率與最差正確率卻相差了 10.9%，相對來說，題名＋目次＋摘要之組合對於分類的準確度較為穩定，除此之外，利用片段內容的組合作為分類資料進行圖書分類，可以避免發生因資料為空值而無法分類的情形。

表 2 少量資料於多層式分類之正確率－最佳組合

片段內容	第一階層		第二階層		第三階層		無法判斷
	階層組合	正確率	階層組合	正確率	階層組合	正確率	
(1) 題名	AC	96.60%	1AB2C 1AC2B 1AC2D 1AD2C 1BC2A 1CD2A	97.70%	1AB2C3D 1AC2B3D 1AC2D3B 1AD2C3B 1BC2A3D 1CD2A3B	97.80%	0
(2) 目次	AC	90.40%	1AB2C 1AC2B 1BC2A	93.60%	1AC2D3B 1AD2C3B 1CD2A3B	94.00%	10
(3) 摘要	AC	92.60%	1AB2C 1AC2B 1BC2A	94.60%	1AC2D3B 1AD2C3B 1CD2A3B	95.40%	4
(1)＋(2)	AC	91.10%	1AB2C 1AC2B 1BC2A	93.70%	1AC2D3B 1AD2C3B 1CD2A3B	94.40%	11
(1)＋(3)	AC	93.00%	1AB2C 1AC2B 1BC2A	94.80%	1AC2D3B 1AD2C3B 1CD2A3B	95.30%	4
(2)＋(3)	AC	93.60%	1AC2D 1AD2C 1CD2A	95.80%	1AC2D3B 1AD2C3B 1CD2A3B	96.00%	3

片段內容	第一階層		第二階層		第三階層		無法判斷
	階層組合	正確率	階層組合	正確率	階層組合	正確率	
(1) + (2) +(3)	AC	93.70%	1AB2C 1AC2B 1BC2A	95.90%	1AC2D3B 1AD2C3B 1CD2A3B	96.10%	3

（三） 多層式分類階層組合

在分類器於階層的搭配組合方面，當（A）單純貝氏分類器與（C）決策樹分類器一起放在第一階層進行分類時，得到的正確率最高，但當（B）支持向量機分類器與（D）k 個最近鄰分類器一起放在第一階層進行分類時，得到的正確率最低。於三階層，得到最佳正確率之階層組合為第三階層放支持向量機分類器之組合（1AC2D3B、1AD2C3B、1CD2A3B）。詳見表 2。

由於第一階層及第二階層之預測類別比對結果，仍然有不一致的資料，因此，仍需要再繼續到下一階層進行預測類別的比對，直到第三階層為止，從表 2 也可發現，正確率隨著階層數的增加而提升，除了逐層減少不一致的資料之外，亦能降低分類的錯誤率。表 2 中的欄位「無法判斷」，指經過三個階層比對之後，仍有預測類別不一致的資料數量，將被視為無法分類（分類錯誤）。

（四） 語義相似詞之最佳抽取個數

以得到最佳正確率之階層組合接續進行導入語義概念之圖書自動分類，分類結果見表 3，使用三階層、第三階層放支持向量機分類器之組合。經過詞彙轉換之正確率皆高於未經過詞彙轉換之正確率，並且當語義相似詞抽取個數為 3 時，對分類成效較具明顯的變化，語義相似詞的加入具有語義加強的作用。題名、目次、摘要及其四種組合於多層式分類，得到最佳的片段內容為題名（98.00%），其次為題名＋目次＋摘要之組合（96.80%），而目次為最差者（93.50%）。

雖然題名、目次、摘要及其四種組合在多層式分類得到的正確率，仍然僅次於單以決策樹分類器進行圖書分類得到的正確率，但與未經過詞彙轉換之正確率相比，提升了 0.2%至 0.9%，由此說明，對以少量資料進行的圖書分類而言，作為輔助工具的 Word2Vec 有助於圖書分類準確度的提升，尤其以多層式分類進行的圖書分類。

表 3 經詞彙轉換分類資料於多層式分類之正確率（單位：百分比）

片段內容	少量資料	Word2Vec 語義相似詞抽取個數		
		2	3	4
(1) 題名	97.80	98.00	97.60	97.40
(2) 目次	94.00	93.50	94.00	94.50
(3) 摘要	95.40	95.00	95.80	95.80
(1) ＋ (2)	94.40	94.50	94.50	94.20
(1) ＋ (3)	95.30	95.50	96.20	95.80
(2) ＋ (3)	96.00	96.10	96.70	96.70
(1)＋(2)＋(3)	96.10	96.30	96.80	96.70

三、 多量多類別之多層式圖書自動分類

為實驗的後兩個階段，以大量語料、多類別進行圖書分類，所使用的四種分類器於 WEKA 對應的名稱如下：

- 單純貝氏（NB）：weka.classifiers.bayes.NaiveBayes
- 支持向量機（SVM）：weka.classifiers.functions.LibSVM（LibSVM）
- 決策樹（DT）：weka.classifiers.trees.J48（C4.5）
- k 個最近鄰（kNN）：weka.classifiers.lazy.IBk

（一） 類別資料量

為了確保大量語料的分類成效，先行實作排除類別資料量少於 5、

25、30、35、40、45、50、55、60 筆的類別，藉以找出適合電子書書目分類的類別資料量，並觀察類別資料量與分類成效之間的關係。各小類的資料概況詳見表 4，包含訓練資料及測試資料的資料量、涵蓋的大類與小類數量等，訓練資料量皆至少一萬筆以上。

表 4 增量資料之類別概況

類別資料量	資料量	訓練資料	測試資料	小類個數	涵蓋大類個數
>= 5	17,626	15,382	1,122	561	10
>= 25	14,088	13,840	248	124	10
>= 30	13,729	13,507	222	111	10
>= 35	13,082	12,900	182	91	10
>= 40	12,717	12,555	162	81	9
>= 45	12,428	12,280	148	74	9
>= 50	11,997	11,867	130	65	9
>= 55	11,536	11,424	112	56	9
>= 60	11,307	11,203	104	52	9

透過圖 5 呈現的曲線，可觀察到測試資料的分類正確率隨著類別資料量的增加而提升，最高可達 60.71%（多層式分類器，類別資料量至少 55 筆）。在多層式分類器方面，當類別資料量至少 55 筆以上時，原本正確率僅次於單純貝氏分類器的多層式分類器開始位居第一，相較於單純貝氏分類器，正確率提升了 3.57%至 5.77%，並且多層式分類器的最佳組合亦產生變化，從原本第三階層放支持向量機分類器得到的正確率是 12 個組合中最高，到了類別資料量至少 55 筆時，最佳組合的第三階層為決策樹分類器，詳見表 5。

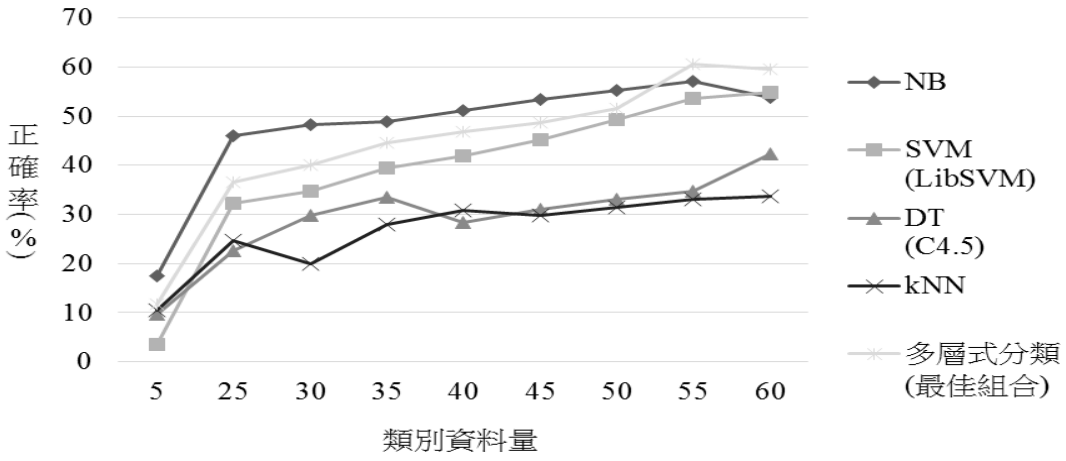

圖 5 測試資料之正確率

表 5 測試資料於多層式分類之正確率

類別資料量	正確率	三階層最佳組合	無法判斷
>= 5	11.68%	1AC2D3B, 1AD2C3B, 1CD2A3B	647
>= 25	36.69%	1AC2D3B, 1AD2C3B, 1CD2A3B	67
>= 30	40.09%	1AC2D3B, 1AD2C3B, 1CD2A3B	54
>= 35	44.51%	1AC2D3B, 1AD2C3B, 1CD2A3B	41
>= 40	46.91%	1AC2D3B, 1AD2C3B, 1CD2A3B	27
>= 45	48.65%	1AC2D3B, 1AD2C3B, 1CD2A3B	29
>= 50	51.54%	1BC2D3A, 1BD2C3A, 1CD2B3A, 1AC2D3B, 1AD2C3B, 1CD2A3B, 1AB2D3C, 1AD2B3C, 1BD2A3C	22
>= 55	60.71%	1AB2D3C, 1AD2B3C, 1BD2A3C	10
>= 60	59.62%	1BC2D3A, 1BD2C3A, 1CD2B3A, 1AC2D3B, 1AD2C3B, 1CD2A3B, 1AB2D3C, 1AD2B3C, 1BD2A3C, 1AB2C3D, 1AC2B3D, 1BC2A3D	10

（二） 於多層式分類導入語義概念

實驗結果顯示，得到最好的分類成效為在第三階層不一致的資料

加進語義相似詞，正確率增加 1.79%，無法判斷的數量從 10 筆減至 5 筆，說明在多層式分類第三階層進行詞彙擴充，能有效提升圖書分類的準確度，亦實證由 Word2Vec 詞向量工具訓練出來的相似詞，有助於改善圖書自動分類的成效，適合作為圖書自動分類的輔助工具之一。結果詳見表 6。

表 6 階層詞彙擴充之正確率

多層式階層	擴充次數	正確率	三階層最佳組合	無法判斷
—	0	60.71%	1AB2D3C, 1AD2B3C, 1BD2A3C	10
第一階層	1	58.93%	1CD2B3A	12
第二階層	1	59.82%	1AB2D3C, 1AD2B3C, 1BD2A3C	16
第三階層	1	62.50%	1AB2D3C, 1AD2B3C, 1BD2A3C	5
第一、二階層	2	58.93%	1CD2B3A	14

四、 圖書分類號前三層次之評估

以實驗四個階段得到最佳的結果，評估《中文圖書分類法》分類號前三層次的正確率，進一步觀察錯誤分類資料之前兩層次為正確的情形。分類號前三層次指圖書分類法的第一級類目至第三級類目，以下依序以 X00、XX0、XXX 表示。

增量資料之最佳結果為多層式分類第三階層放決策樹分類器，片段內容為題名＋目次＋摘要之組合。表 7 顯示無論是在階層詞彙擴充之前或之後，分類號前兩層次 X00、XX0 之分類成效皆有所提升，正確率至少提升 5.36%。

表 7 分類號層次之評估

分類號層次	階層詞彙擴充之前	無法判斷	階層詞彙擴充之後	無法判斷	三階層最佳組合
X00	78.57%	3	80.36%	0	1AB2D3C, 1AD2B3C, 1BD2A3C
XX0	66.07%	5	67.86%	2	1AB2D3C, 1AD2B3C, 1BD2A3C
XXX	60.71%	10	62.50%	5	1AB2D3C, 1AD2B3C, 1BD2A3C

伍、 討論

（一） 分類錯誤原因探討

於本研究中觀察分類錯誤原因主要有三：特徵詞彙量多寡及代表性、訓練資料量不足、存在多個主題的圖書內容。

在片段內容組合方面，雖然題名或許已經能取得很好的分類結果，假若再加進其他片段內容，得以擷取更多具代表性的特徵詞彙，有助於分類器的訓練，以及提升圖書自動分類的成效，達到輔助判斷合適的分類號之目的。透過片段內容組合的方式，除了解決圖書題名太短而無法擷取足夠的特徵詞彙（文件特徵）的問題，也使多種片段內容互相補足因資料前置處理而出現的空值。以電子書《時之一》為例，經過資料前置處理之後，題名變為空值，不為空值的目次與摘要便能補上題名為空值的部分，避免因資料為空值而無法分類，導致分類錯誤率增加。

經過資料前置處理留下來的特徵詞彙多寡及代表性與正確分類的影響，以電子書《這樣開會最聰明!》為例，此例在僅題名分類時，單純貝氏分類器將之分類到 528「各種教育」，支持向量機、k 個最近鄰分類器將之分類到 413「中國醫學」，而決策樹分類器將之正確分類到 494「企業管理」，若將此資料加上目次及摘要進行分類，只有 k 個最近鄰分類器仍分類錯誤，其餘三種分類器皆將之分類到正確的類別，此筆資料在採用多數決策略之多層式分類中，最終類別也會是 494「企業管理」，分類正確。由此例說明，將圖書片段內容組合分類，有效克服特徵詞彙

量的不足及代表性的疑慮。

在訓練資料量方面，過去研究考慮到資料的代表性，排除少於一定數量的資料之類別，再以保留的資料進行分類。然而，由於本研究蒐集的增量資料，類別範圍至少涵蓋九大類，小類的個數至多 581 個，因此，除了排除類別資料量少於 25 筆的類別，亦排除類別資料量少於 30、35、40、45、50、55、60 筆的類別，藉以找出適合大量語料進行分類的類別資料量，確保分類結果。訓練資料量愈多，分類成效愈好，另一方面，分類類別數量愈多，每個類別需要準備或蒐集的資料量也愈多，才能夠確保最後的分類結果達到一定的成效。

針對增量資料之測試資料分類錯誤的資料，經查詢全國圖書書目資訊網（National Bibliographic Information Network，NBINet），發現至少 23 筆圖書資料擁有一個以上的圖書分類號。可能原因如下列四種情形（Subrahmanyam, 2006）：1.給予的分類號具替換性，2.因圖書分類法之彈性造成的差異，3.與圖書館本身給予分類號的規則有關，4.給予的分類號同時具替換性及差異性，前述四種情形使得同一圖書資料之分類號在不同的圖書館系統之間不一致，也導致同一圖書資料會有一個以上的圖書分類號。

在這至少 23 筆圖書資料中，由分類結果亦發現當中 3 筆於多層式分類得到的最終類別，雖然為分類錯誤，但該最終類別事實上為其他間圖書館給予的分類號。以電子書《看花猶是去年人》為例，此筆圖書資料經查詢 NBINet，不同間圖書館給予的分類號有 078「現代論叢」、573「中國政治制度」，依序屬於「總類」、「社會科學類」大類，在實驗中的實際類別為 078，以題名＋目次＋摘要之組合進行分類時，於多層式分類最佳組合得到的最終類別為 573，雖然為分類錯誤，但得到的最終類別 573 事實上為其他間圖書館給予的分類號之一（以下稱「其他類別」）。因此，若將多層式分類得到的最終類別為其他類別的資料視為分類正確，正確率可再提升 1.79%，經階層詞彙擴充之後，正確率提升

7.14%，其中，進行階層詞彙擴充之前，此種情形共 2 筆，經階層詞彙擴充之後，再新增 1 筆。圖書分類號前三層次之評估得到的正確率見表 8，對照表 7。

表 8 分類號層次之評估－其他類別

分類號層次	階層詞彙擴充之前	無法判斷	階層詞彙擴充之後	無法判斷	三階層最佳組合
X00	83.93%	3	85.71%	0	1AB2D3C, 1AD2B3C, 1BD2A3C
XX0	71.43%	5	75.00%	2	1AB2D3C, 1AD2B3C, 1BD2A3C
XXX	62.50%	10	69.64%	5	1AB2D3C, 1AD2B3C, 1BD2A3C

（二） 多層次分類效能分析

實驗結果顯示，在經過第二階層仍有預測類別不一致的前提之下，無論是以何種片段內容及其組合進行分類，三階層的分類成效比二階層更好，無法判斷的數量亦逐層減少，降低因最終類別無法判斷而增加的錯誤率。於使用增量資料進行的實驗中，得到最佳的組合為前兩階層放單純貝氏、支持向量機、k 個最近鄰分類器，第三階層放決策樹分類器。

分類器於階層的搭配組合亦會影響分類結果，以電子書《超有趣的英文單字故事書》為例，此筆資料以題名＋目次＋摘要之組合進行分類時，發生四種分類器得到兩種分類號（預測類別）的情形，不同的組合會使此筆資料被判斷為分類正確或分類錯誤，如多層式分類最佳組合為第三階層放決策樹分類器（1AB2D3C、1AD2B3C、1BD2A3C），得到的最終類別會是 800，分類錯誤，但階層組合為第三階層放單純貝氏或支持向量機分類器時（1BC2D3A、1BD2C3A、1CD2B3A、1AC2D3B、1AD2C3B、1CD2A3B），得到的最終類別則是 500 或 520，分類正確。因此，分類錯誤的部分原因為分類器於多層式分類階層的搭配組合，但此種情形於增量資料的測試資料中，僅發生在分類號的前兩層次。

另一方面，當類別資料量達到一定的數量，多層式分類效能愈好，當每個小類資料量少於 55 筆時，多層式分類的效能皆排名第二，但當每個小類資料量至少 55 筆以上時，多層式分類的效能比其他四種分類器的效能更好。藉此實證多層式圖書自動分類系統有助於圖書館分類編目實務工作，使分類器之間達到互補效用，減低分類演算法特性對分類的影響，提升整體分類效能。

（三） Word2Vec 詞向量工具作為改進分類成效之輔助工具探討

相較於傳統單層式單一／多重分類器的分類表現，導入語義概念於多層式圖書自動分類系統能得到更好的分類表現。以電子書《精神醫療的美麗境界:大溫哥華精神衛生照護模式》為例，此筆資料以題名＋目次＋摘要之組合進行分類時，單純貝氏分類器將之分類到 538「民俗學；各國風俗」，支持向量機分類器將之分類到 857「小說」，決策樹分類器將之分類到 177「應用心理學」，而 k 個最近鄰分類器將之分類到 782「中國傳記」，經過詞彙擴充之後，單純貝氏及支持向量機分類器皆將此筆資料分類到正確的小類 415「西醫學」，決策樹分類器改分類至 418「藥學；藥理學；治療學」，而 k 個最近鄰分類器仍分類至 782，於多層式分類 1AB（第一階層為單純貝氏、支持向量機）組合中，此筆資料的最終類別也會是 415，分類正確。其中，雖然經過詞彙擴充之後，決策樹分類器將此筆資料分類到錯誤的類別，但卻與實際類別為相同的大類，從「哲學類」改分類到「應用科學類」，接近於實際類別，此情形亦顯示出詞彙擴充的方式可能有助於大類的判斷。

另一方面，階層詞彙擴充的次數不宜過多，擴充兩次得到的正確率反而下降，無法判斷的數量增加，導致錯誤率上升。原因在於詞彙擴充的過程中，可能加進了其他代表性不足或語義相似度較低的相似詞，反而影響分類成效，雖然使用 Word2Vec 詞向量工具建構的同義詞詞典具即時、動態等優勢，並且建立的詞彙不易引進其他不在實驗語料範圍的雜訊詞彙，得以確保詞典內容的品質，但是卻有可能因加進的語義相似

詞過多或語義模糊等因素，反而為分類資料帶進內部可能存在的雜訊詞彙，因而影響分類準確度，不過就整體而言，Word2Vec 詞向量工具所建構的同義詞詞典及基於 Word2Vec 同義詞詞典建立的詞表，確實有助於圖書分類準確度的提升。

（四） 圖書分類號前三層次之分析

針對圖書分類號前三層次進行的評估，分類的正確率隨著分類號層次遞減而提升，如分類號層次 XXX 到 X00 於階層詞彙擴充之後的正確率，從 62.50%提升至 80.36%，與陳強強（2018）使用層次分類器所得到的結果相符合。此結果實證本研究提出的基於語義之多層式圖書自動分類系統整體在大類（X00）的分類表現，與分類號前三碼（XXX）的分類表現相比較，為有效提升。因此，對僅需要分類到大類的圖書分類應用而言，使用本研究提出的基於語義之多層式圖書自動分類系統，將可得到良好的分類表現。

（五） 與相關研究結果的比較

本研究與其他同樣以圖書分類號作為分類類別之相關研究的不同，整理至表 9。在圖書特徵選擇方面，過去相關研究皆選擇了較能簡潔表達書籍內容的圖書題名，其他如作者資訊、摘要等圖書特徵，而本研究除了選擇圖書題名之外，亦蒐集目次與摘要，並在實驗中嘗試以組合的方式進行分類，例如：題名＋目次、題名＋摘要、目次＋摘要、題名＋目次＋摘要，實驗結果顯示組合的方式能夠避免發生因資料為空值而無法分類的情形，亦能解決題名太短而無法擷取足夠數量之文件特徵等問題。

面臨可能由於每個類別的資料量非平均分配，將影響最後的分類結果，於此，相關研究提出不同策略以改進分類成效，如黃嘉宏（2008）藉由 Google 搜尋引擎之查詢結果擴展特徵值，用以克服特徵值可能不足的課題，本研究則是利用 Word2Vec 作為改進分類成效之輔助工具，

嘗試在分類資料中加進語義相似詞，尤其是所加進的詞彙皆取自於內部資料本身，非來自於外部資源內容，如此避免引進其他不在實驗語料範圍的雜訊詞彙。

在分類器方面，陳強強（2018）提出層次分類器，針對圖書分類號前三層次建構分類器，以此方法降低計算複雜度，同時顧及分類法類目之間的層次關係及類目繁多的情況。針對分類號前三層次建構的分類器，皆使用相同的分類演算法訓練分類器，而本研究使用郭俊桔（2018）提出的多層式分類器，則是三個階層皆使用不同的分類演算法訓練分類器，但分類號層次不在本研究的研究範圍之內，僅依據分類結果作分類號層次的評估。相較之下，本研究較為不足的地方在於應用的分類器，陳強強（2018）使用了三種分類演算法訓練層次分類器，其中一種分類演算法為極限學習機，是一種求單隱層前饋式類神經網路的學習演算法，因而得到較佳的分類效能。在未來研究方向，亦可嘗試使用類神經網路訓練分類器，並應用於多層式分類，或是將多層式分類器應用於分類號層次方面，以求得更精確的分類結果。

除了上述比較內容，在正確率方面，本研究得到的正確率略低於黃嘉宏（2008）得到的正確率，主要的差別為分類類別涵蓋的大類數量及小類資料量，黃嘉宏（2008）蒐集的資料涵蓋兩大類，每個小類資料量至少 25 筆，本研究蒐集的資料則涵蓋九大類，每個小類資料量至少 55 筆時得到的正確率最佳。另一方面，本研究在蒐集的資料量方面尚有不足，受限於大學圖書館訂購的 HyRead ebook 中文電子書數量，因此，未來或許可擴大電子書蒐集的平台，一方面增加電子書書目數量，另一方面也能蒐集更多種不同類別的電子書，如此更貼近圖書館館員分類編目的工作流程。

表 9 相關研究之比較

比較項目 \ 相關研究	黃嘉宏 (2008)	Á vila-Argüelles et al. (2010)	陳強強 (2018)	本研究
圖書題名	√	√	√	√
作者資訊	√			
目次				√
摘要			√	√
中文圖書分類法	√			√
中國圖書館分類法			√	
國會圖書館分類法(LCC)		√		
Google 搜尋引擎	√			
Lesk 投票典範、詞頻		√		
LDA + TF×IDF			√	
Word2Vec				√
單層單一分類器	√	√		
層次分類器			√	
多層式分類器				√
大類數量	2	1	8	9
訓練資料	19,949	489,726	205,500	11,424
測試資料	4,911	122,431	—	112
正確率	62.86%	36.00%	72.39%	62.50%

陸、 結論與建議

一、 結論

實驗結果顯示將多層式圖書自動分類系統應用於圖書分類具有更佳的分類成效，正確率為 60.71%，加進 Word2Vec 語義相似詞之後的正確率為 62.50%，提升了 1.79%，若將其他圖書館給予的分類號視為分類正確，加進 Word2Vec 語義相似詞之後的正確率則為 69.64%，接

近 70%。說明多層式圖書自動分類系統有助於圖書分類準確度的提升，並且 Word2Vec 適合作為圖書分類的輔助工具，其訓練出來的語義相似詞能使圖書分類得更準確。

研究發現可歸納下列三項：

（一） 多層式圖書自動分類系統應用於圖書館館藏的分類表現

題名或許已經能取得很好的分類結果，但若是透過片段內容組合的方式，除了可以解決圖書題名太短而無法擷取足夠數量之文件特徵的問題，也能互相補足因資料前置處理而出現的空值，避免發生因資料為空值而無法分類的情形。因此，對基於機器學習方法而言，具代表性且適合用於圖書自動分類的圖書特徵應為圖書題名＋目次＋摘要之組合。

在多層式圖書自動分類系統實作方面，當第二階層仍有預測類別不一致的資料時，接續進行的三階層分類得到的正確率較高。研究亦發現分類器於階層的搭配會因資料量的多寡而有不同的結果，於少量資料（1,000 筆），最佳階層組合為前兩個階層放單純貝氏、決策樹、k 個最近鄰分類器，第三階層放支持向量機分類器，於增量資料（至少 10,000 筆），最佳階層組合則是前兩個階層放單純貝氏、支持向量機、k 個最近鄰分類器，第三階層放決策樹分類器。再者，分類正確或錯誤亦會受分類器於階層的搭配組合影響，如某筆圖書資料於四種分類器得到兩種分類號，其中一種分類號為實際類別，分類器階層不同的組合會使此筆圖書資料被判斷為分類正確或錯誤。

（二） 導入語義概念對圖書自動分類的影響

透過自 Word2Vec 訓練結果抽取語義相似詞抽取個數之實驗，發現個數的多寡會影響分類結果，於本研究中，抽取個數為 3 時，得到的分類成效最佳。於使用大量語料進行的圖書自動分類，針對多層式分類各階層不一致的資料進行詞彙擴充，研究發現，階層擴充的層次和擴充詞

彙的次數皆會影響分類結果，針對第三階層不一致資料（擴充次數為 1 次）進行詞彙擴充之後得到的正確率為最佳，提升了 1.79%，無法判斷的數量從 10 筆減至 5 筆。前兩項 Word2Vec 語義相似詞之個數、於多層式圖書自動分類系統導入語義概念之分類表現，皆實證 Word2Vec 詞向量工具有助於改進圖書自動分類的成效，適合作為圖書自動分類的輔助工具之一，所提出的改進分類成效之策略確實能夠提升圖書分類的準確度。

（三） 訓練資料、類別資料量與分類成效之間的關係

當類別資料量至少 55 筆時，得到的正確率最佳，另一方面也發現，當類別資料量至少 55 筆以上時，多層式分類器得到的正確率高於其他四種分類器，說明多層式圖書自動分類比傳統圖書分類具有更佳的分類效能。

二、 建議

（一） 資料蒐集的範圍及分類類別的數量

資料蒐集範圍受限於 HyRead ebook 中文電子書，在實驗中為了確保分類結果，排除了資料量少於 55 筆的類別，從 581 個類別減至 56 個類別，雖然訓練資料量至少一萬筆以上，但分類類別的個數與圖書館館藏實際類別數量仍有差距，在未來研究可以擴大資料蒐集的範圍，不限於一種中文電子書資料庫，其他電子書資料庫如 iRead eBooks 華藝電子書、udn 數位閱讀館電子書庫等，以增加類別資料量及類別的數量，如此更貼近圖書館館藏的實際情形，再進一步觀察可能影響圖書自動分類成效的因素，尤其以《中文圖書分類法》圖書分類號作為分類類別。

（二） 作品內容包含多個主題之分類原則

本研究於分類錯誤原因之探討，發現部分錯誤原因在於經由多層式分類器得到的最終類別事實上為其他圖書館給予的分類號，但由於

與予以核對的實際類別不同，因而判斷為分類錯誤，此種情形說明作品內容可能包含多個主題。於基於機器學習方法之圖書自動分類，在進行分類之前，亦應針對分類資料內容制定原則，以因應遇到內容包含兩個以上主題的作品時，給予最適合該作品的分類號，並且該分類號能夠協助使用者掌握圖書資料內容的主題或學科性質，如若遇到包含之主題或學科太廣的作品，則歸入總類。

（三） 應用於多層式圖書自動分類系統之分類器

本研究使用四種過去常被應用於文件分類的分類演算法個別訓練分類器，並應用於多層式圖書自動分類系統，從階段性實驗發現，k 個最近鄰分類器無論訓練資料量多寡，分類表現皆不盡理想，雖然藉由階層的方式使每個分類器的特性達到互補的效果為多層式分類的優勢，但單個分類器若是分類表現不佳，仍然會影響到多層式分類的效能，因此，在選擇分類器方面，未來研究或許可以考慮使用其他常被應用於文件分類且分類表現尚佳的分類器，用以加強整體的分類表現。

（四） Word2Vec 語義關係

在語義相似詞的加入方面，多層式分類階層擴充的層次和擴充詞彙的次數對分類結果皆產生影響，語義相似詞的數量與擴充的對象似乎有所限制，未來研究可以結合外部現存詞典，進一步分析語義關係，如詞彙的上下位關係，輔助解決 Word2Vec 語義相似詞可能存在的語義問題，進而達到語義加強的效果。

（五） 中文圖書自動分類系統分析與設計

以圖書館館藏中文電子書書目為分類對象，建構基於 Word2Vec 之同義詞詞典，並使用大量語料進行多層式圖書自動分類，正確率為 62.50%，尤其大類的正確率可達到 80%以上，未來可以進一步將文件分類流程整合至分類系統，包含資料前置處理、特徵建構及加權、特徵選擇、訓練分類器、多層式分類比對、Word2Vec 詞向量訓練等子模組，

並設計人機介面，供圖書館或其他相關領域人員使用。在實務方面，將圖書自動分類系統實際應用於分類編目的工作流程，透過質性方法進行相關研究，如從編目人員的角度來看圖書自動分類系統的分類結果，所得到的圖書分類號達到協助使用者掌握圖書資料內容的主題或學科性質，以及協助館員掌握每個主題的圖書資料並管理等目的之程度如何。另外，為進一步縮小大學圖書館與公共圖書館的差異，未來將取用公圖的實際文本，以期更能貼近公圖實務。

【參考書目】

王梅玲、謝寶煖（2014）。*圖書資訊學導論（二版）*。臺北市：五南。

吳慧貞（2015）。*使用資料探勘探討多層式圖書自動分類系統之研究*。未出版之碩士論文，國立中興大學圖書資訊學研究所，臺中市。

林昕潔（2006）。*以 SVM 與詮釋資料設計書籍分類系統*。未出版之碩士論文，國立交通大學資訊科學與工程研究所，新竹市。

陳信源、葉鎮源、林昕潔、黃明居、柯皓仁、楊維邦（2009）。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*，2-21。

陳強強（2018）。*基於機器學習的中文書目自動分類的研究*。未出版之碩士論文，北方工業大學，北京市。

郭俊桔（2018）。使用多數決策略之圖書自動分類的研究。*圖書資訊學研究，13*，87-124。

曾元顯（2002）。文件主題自動分類成效因素探討。*中國圖書館學會會報，68*，62-83。

曾元顯、莊大衛（2003）。文件自我擴展於自動分類之應用。*第十五*

屆計算機語言學研討會論文集（129-141 頁）。新竹市：清華大學。

黃嘉宏（2008）。*基於自動分類為基礎的圖書題名特徵擷取之研究－以輔助圖書分類系統為例*。未出版之碩士論文，輔仁大學圖書資訊學系碩士班，臺北市。

AL-Nabi, D. A., & Ahmed, S. S. (2013). Survey on classification algorithms for data mining:(comparison and evaluation). *Computer Engineering and Intelligent Systems*, 4(8), 18-24.

Á vila-Argüelles, R., Calvo, H., Gelbukh, A., & Godoy-Calderón, S. (2010). Assigning Library of Congress Classification codes to books based only on their titles. *Informatica*, 34(1), 77-84.

Chang, C. C., & Lin, C. J. (2011). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3), 1-27. doi: 10.1145/1961189.1961199. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Frank, E., & Paynter, G. W. (2004). Predicting Library of Congress classifications from Library of Congress subject headings. *Journal of the American Society for Information Science and Technology*, 55(3), 214-227. doi: 10.1002/asi.10360.

Kim, J. H., & Lee, K. H. (2002). Designing a knowledge base for automatic book classification. *The Electronic Library*, *20*(6), 488-495. doi: 10.1108/02640470210454010.

Larson, R. R. (1992). Experiments in automatic Library of Congress Classification. *Journal of the American Society for Information Science*, 43(2), 130-148. doi: 10.1002/(SICI)1097-4571(199203)43:2%3C130::AID-ASI3%3E3.0.CO;2-S.

Liang, C. Y., Guo, L., Xia, Z. J., Nie, F. G., Li, X. X., Su, L., & Yang, Z. Y. (2006). Dictionary-based text categorization of chemical web

pages. *Information Processing & Management*, 42(4), 1017-1029. doi: 10.1016/j.ipm.2005.09.001.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. Retrieved from https://arxiv.org/abs/1301.3781v3.

Mirończuk, M. M., & Protasiewicz, J. (2018). A recent overview of the state-of-the-art elements of text classification. *Expert Systems with Applications*, 106, 36-54. doi: 10.1016/j.eswa.2018.03.058.

Nonaka, I. (1994). A dynamic theory of organizational knowledge creation. *Organization Science, 5*(1), 14-37.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM, 18*(11), 613-620. doi: 10.1145/361219.361220.

Shen, Y., Zhang, Q., Zhang, J., Huang, J., Lu, Y., & Lei, K. (2018, June). Improving medical short text classification with semantic expansion using word-cluster embedding. *In International Conference on Information Science and Applications* (pp. 401-411). Springer, Singapore. doi: 10.1007/978-981-13-1056-0_41.

Subrahmanyam, B. (2006). Library of Congress Classification numbers: issues of consistency and their implications for union catalogs. *Library Resources & Technical Services*, 50(2), 110-119.

Wahbeh, A. H., Al-Radaideh, Q. A., Al-Kabi, M. N., & Al-Shawakfa, E. M. (2011). A comparison study between data mining tools over some classification methods. *International Journal of Advanced Computer Science and Applications*, 8(2), 18-26.

Yi, K. (2007). Automated text classification using library classification schemes: trends, issues, and challenges. *International Cataloguing and Bibliographic Control*, 36(4), 78-82.

Empirical Study of Semantic-based Multi-layered Automatic Book Classification

Yu-Pei Chen

Master, Graduate Institute of Library and Information Science,

National Chung-Hsing University

June-Jei Kuo

Corresponding author: jjkuo@dragon.nchu.edu.tw

Associate Professor, Graduate Institute of Library and Information Science, National Chung-Hsing University

Abstract

This study aims to explore the application of the multi-layered automatic book classification using the majority voting strategy in a library environment, and proposes a semantic-based approach as a strategy to improve the classification performance. The proposed strategy uses Word2Vec, which can extract the deep semantic relationships between words, to expand words features for improving the classification performance. The collection of data from HyRead ebook, which was ordered by the university library, was utilized as the training and testing corpus, including book titles, table of contents, abstracts. In total, 17,650 books were collected. And the classification classes were the book classification numbers. A total of 581 classification numbers were selected randomly from New Classification Scheme for Chinese Libraries. Naïve Bayes, SVM, Decision Tree C4.5, and kNN were applied to the multi-layered automatic book classification. Regarding the proposed semantic-based approach, this study used Word2Vec as a tool for training word embedding. First, a thesaurus was built by the training results, and next the word features of the data set for classification were expanded. Under the principle of accuracy, experimental results showed that the

performance of the multi-layered automatic book classification outperformed the traditional automatic book classification in a library environment. And the strategy proposed in this study could indeed improve the accuracy of book classification, achieving 60%. After expanding the word features of the data set, the accuracy rate promoted 1.79%. The above results proved that Word2Vec tool was helpful for automatic book classification. It could effectively reduce the number of unable classified books, which were classified through the multi-layered automatic book classification. However, book content with multiple topics may be one of the factors that affect classification effectiveness for automatic book classification. If the book classification number, given by other library systems, was regarded as the correct class, the accuracy rate could be approximately 70% by the semantic-based multi-layered automatic book classification using the majority voting strategy. The accuracy rate could be further increased by 7.14%. Therefore, it is necessary to consider the effect of book content with multiple topics for automatic book classification in the future library practice.

Keywords: Classification number ; Layer style; Automatic book classification ; Word2Vec

SUMMARY

Introduction

Cataloging is a subset of a larger field that is called *information organization* (sometimes referred to as *bibliographic control* or as *organization of information*), and it is helpful to view it within that context. Cataloging is divided into description cataloging and subject cataloging. The quality and accuracy of subject cataloging have been questioned and discussed, especially the consistency of classification number. In addition to improve the consistency of classification number by copy cataloging, another related research trend is the application of automatic document classification to book classification in a library environment.

The study aims to explore the application of the multi-layered automatic book classification using the majority voting strategy in a library environment, and proposes a semantic-based approach as a strategy to improve the classification performance.

Regarding the selection of data content, the study collected the titles together with other book features, and explored the features of books suitable for automatic classification through the combination of different fragment contents. On the strategy of improving the classification effect, past researches have used the relationship between words and words in external dictionaries as a strategy. However, the disadvantage of using existing dictionaries is that words are territorial, and the updating and maintenance of words cannot be timely. The study attempted to use Word2Vec, which can extract the deep semantic relationships between words, to construct a thesaurus with real-time, dynamic.

Methods

The research structure of the study was divided into four parts: (1) data preprocessing, (2) semantic vocabulary processing, (3) automatic book classification, and (4) classification result evaluation. First, pre-process the collected data which included the title, table of contents, and abstract of the book. Then, the pre-processed data was used as the training corpus for training the classifiers and as the Word2Vec word vector tool. Next, the original classification data and the classification data processed by semantic vocabulary were automatically classified into books. The ten-fold cross-validation method and the correct rate were used to evaluate the classification effect and analyze the classification results.

The collection of data from HyRead ebook, which was ordered by the university library, was utilized as the training and testing corpus, including book titles, table of contents, abstracts. In total, 17,650 books were

collected, and the classification classes were the book classification numbers. A total of 581 classification numbers were selected randomly from New Classification Scheme for Chinese Libraries. Naïve Bayes, SVM, Decision Tree C4.5, and kNN were applied to the multi-layered automatic book classification.

Regarding the proposed semantic-based approach, the study used Word2Vec as a tool for training word embedding. First, a thesaurus was built by the training results, and next the word features of the data set for classification were expanded.

Results

The strategy proposed in the study could indeed improve the accuracy of book classification, achieving 60%. After expanding the word features of the data set, the accuracy rate promoted 1.79%. The above results proved that Word2Vec tool was helpful for automatic book classification. It could effectively reduce the number of unable classified books, which were classified through the multi-layered automatic book classification. However, book content with several topics may be one of the factors that affect classification effectiveness for automatic book classification. If the book classification number, given by other library systems, was regarded as the correct class, the accuracy rate could be approximately 70% by the semantic-based multi-layered automatic book classification using the majority voting strategy. The accuracy rate could be further increased by 7.14%. Therefore, it is necessary to consider the effect of book content with several topics for automatic book classification in the future library practice.

Conclusions and Suggestions

Under the principle of accuracy, experimental results showed that the

performance of the multi-layered automatic book classification outperformed the traditional automatic book classification in a library environment.

For the method based on machine learning, the book features that are representative and suitable for automatic book classification should be the combination of the title, table of contents, and abstract of the book. In future research, the scope of data collection can be expanded, in order to increase the amount of data and the number of categories. So as to be closer to the actual situation of library collections, and further observe the factors that may affect the effect of automatic book classification.

In terms of selecting classifiers, it may consider using other classifiers that are often used in document classification and have better performance to enhance the overall classification performance in future research. In terms of adding semantically similar words, the number of semantically similar words and the objects to be expanded seem to be limited. In future research, it can combine external existing dictionaries to further analyze the semantic relationship and assist in solving the possible semantic problems of Word2Vec semantically similar words, so as to achieve semantic enhanced effect.

ROMANIZED AND TRANSLATED REFERENCE FOR ORIGINAL TEXT

王梅玲、謝寶煖（2014）。*圖書資訊學導論（二版）*。臺北市：五南。【Wang, Mei-Ling & Hsieh, Pao-Nuan (2018). *Introduction to Library and Information Science, Second Edition.* Wu-Nun Book Inc., Taipei (in Chinese)】

吳慧貞（2015）。*使用資料探勘探討多層式圖書自動分類系統之研究*。未出版之碩士論文，國立中興大學圖書資訊學研究所，臺中市。【Wu, Huei-Chen (2015). *A Study on Multi-layered Automatic Book Classification System Using Data Mining.* (Unpublished master's thesis). Graduate Institute of Library & Information Science National Chung Hsing University, Taichung (in Chinese)】

林昕潔（2006）。*以 SVM 與詮釋資料設計書籍分類系統*。未出版之碩士論文，國立交通大學資訊科學與工程研究所，新竹市。【Lin, Hsin-Chieh (2006). *A Book Classification System Using SVM and Meta-information.* (Unpublished master's thesis). Institute of Computer Science and Engineering College of Computer Science National Chiao Tung University, Hsinchu (in Chinese)】

陳信源、葉鎮源、林昕潔、黃明居、柯皓仁、楊維邦（2009）。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*，2-21。【Chen, Shinn-Yuan, Yeh, Jen-Yuan, Lin, Hsin-Chieh , Hwang, Ming-Jiu, Ke, Hao-Ren & Yang, Wei-Pang (2009). Jiehhe jhihyuan siangliangji yu chyuanshih zihliao jhih tushu zihdong fenlei fangfa. *International Journal of Advanced Information Technologies (IJAIT), 3*, 2-21 (in Chinese)】

陳強強（2018）。*基於機器學習的中文書目自動分類的研究*。未出版之碩士論文，北方工業大學，北京市。【Chen, Qiang-Qiang (2018). *Jiyu jichi syuehsi de jhongwun shumu zihdong fenlei de yanjiou.* (Unpublished master's thesis). North China University of

Technology, Beijing (in Chinese)】

郭俊桔（2018）。使用多數決策略之圖書自動分類的研究。*圖書資訊學研究，13*，87-124。【Kuo, June-jei (2018). The Study of Automatic Book Classification Using Majority Vote Strategy. *Journal of Library and Information Science Research*, *13*, 87-124 (in Chinese)】

曾元顯（2002）。文件主題自動分類成效因素探討。*中國圖書館學會會報，68*，62-83。【Tseng, Yuen-Hsien (2002). Effectiveness Issues in Automatic Text Categorization. *Journal of Library and Information Science Research*, *68*, 62-83 (in Chinese)】

曾元顯、莊大衛（2003）。文件自我擴展於自動分類之應用。*第十五屆計算機語言學研討會論文集*（129-141 頁）。新竹市：清華大學。【Tseng, Yuen-Hsien & Juang, Da-Wei (2003). Application of Document Self-Expansion to Text Categorization. *Proceeding of ROCLING 2003*, 129-141. National Tsing Hua University, Hsinchu (in Chinese)】

黃嘉宏（2008）。*基於自動分類為基礎的圖書題名特徵擷取之研究－以輔助圖書分類系統為例*。未出版之碩士論文，輔仁大學圖書資訊學系碩士班，臺北市。【Huang, Jia-Hon (2008). *A Study of Book Title Feature Extraction Based on The Automatic Classification -An Example of BibliographyAutomatically Classified System.* (Unpublished master's thesis). Graduate Institute of Library and Information Science, Fu Jen Catholic University, Taipei (in Chinese)】

AL-Nabi, D. A., & Ahmed, S. S. (2013). Survey on classification algorithms for data mining: (comparison and evaluation). *Computer Engineering and Intelligent Systems*, 4(8), 18-24.

Á vila-Argüelles, R., Calvo, H., Gelbukh, A., & Godoy-Calderón, S. (2010). Assigning Library of Congress Classification codes to books

based only on their titles. *Informatica*, 34(1), 77-84.

Chang, C. C., & Lin, C. J. (2011). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3), 1-27. doi: 10.1145/1961189.1961199. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Frank, E., & Paynter, G. W. (2004). Predicting Library of Congress classifications from Library of Congress subject headings. *Journal of the American Society for Information Science and Technology*, 55(3), 214-227. doi: 10.1002/asi.10360.

Kim, J. H., & Lee, K. H. (2002). Designing a knowledge base for automatic book classification. *The Electronic Library*, *20*(6), 488-495. doi: 10.1108/02640470210454010.

Larson, R. R. (1992). Experiments in automatic Library of Congress Classification. *Journal of the American Society for Information Science*, 43(2), 130-148. doi: 10.1002/(SICI)1097-4571(199203)43:2%3C130::AID-ASI3%3E3.0.CO;2-S.

Liang, C. Y., Guo, L., Xia, Z. J., Nie, F. G., Li, X. X., Su, L., & Yang, Z. Y. (2006). Dictionary-based text categorization of chemical web pages. *Information Processing & Management*, 42(4), 1017-1029. doi: 10.1016/j.ipm.2005.09.001.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. Retrieved from https://arxiv.org/abs/1301.3781v3.

Mirończuk, M. M., & Protasiewicz, J. (2018). A recent overview of the state-of-the-art elements of text classification. *Expert Systems with Applications*, 106, 36-54. doi: 10.1016/j.eswa.2018.03.058.

Nonaka, I. (1994). A dynamic theory of organizational knowledge creation. *Organization Science*, *5*(1), 14-37.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for

automatic indexing. *Communications of the ACM*, *18*(11), 613-620. doi: 10.1145/361219.361220.

Shen, Y., Zhang, Q., Zhang, J., Huang, J., Lu, Y., & Lei, K. (2018, June). Improving medical short text classification with semantic expansion using word-cluster embedding. *In International Conference on Information Science and Applications* (pp. 401-411). Springer, Singapore. doi: 10.1007/978-981-13-1056-0_41.

Subrahmanyam, B. (2006). Library of Congress Classification numbers: issues of consistency and their implications for union catalogs. *Library Resources & Technical Services*, 50(2), 110-119.

Wahbeh, A. H., Al-Radaideh, Q. A., Al-Kabi, M. N., & Al-Shawakfa, E. M. (2011). A comparison study between data mining tools over some classification methods. *International Journal of Advanced Computer Science and Applications*, 8(2), 18-26.

Yi, K. (2007). Automated text classification using library classification schemes: trends, issues, and challenges. *International Cataloguing and Bibliographic Control*, 36(4), 78-82.