使用 SVM 與詮釋資料之圖書自動分類
Automatic Book Classification Using Support Vector Machine and Meta-Information

林昕潔
國立交通大學
資訊工程學系
jesi.cis89@nctu.edu.tw

葉鎮源
國立交通大學
資訊工程學系
jyyeh@cis.nctu.edu.tw

陳信源
國立交通大學
資訊工程學系
sychen@cs.nctu.edu.tw

黃明居
國立交通大學
圖書館助理教授
mjhwang@lib.nctu.edu.tw

柯皓仁
國立交通大學
資訊管理研究所教授
claven@lib.nctu.edu.tw

楊維邦
國立東華大學
資訊管理學系教授
wpyang@ndhu.edu.tw

摘要

本文探討如何將文件自動分類技術應用於圖書分類。本研究將圖書資料劃分為敘述資料(書名、內容簡介與作者簡介)與詮釋資料(作者與出版社)。針對圖書資料的特性，本文提出一套完整的圖書自動分類法：(1) 對敘述資料進行處理，透過特徵挑選擷取出具有類別代表性的特徵。同時，針對專家事先建立的類別特徵，給予適當的加權後，透過 Support Vector Machine 產生分類模型；(2) 以統計方法分析詮釋資料，發掘有助於圖書分類的資訊；(3) 整合 SVM 的分類模型與詮釋資料的統計資訊建立可行的圖書分類模型。實驗使用博客來網路書店的圖書資料，以 *K*-fold Cross-Validation 的方法驗證可行性。實驗結果顯示，當整合 SVM 的分類模型與詮釋資料的統計資訊時，可使整體分類的正確率達到 95%。

關鍵詞：圖書自動分類、SVM、詮釋資料。

Abstract

This thesis discusses how to use document classification techniques to perform automatic book classification. The kernel classification algorithm in this thesis is Support Vector Machines (SVM). The SVM classification result can be integrated with books' meta-information to improve the classification correctness. The classification processes are as followings. In the beginning, data of books are divided into description (book title and prospectus) and meta-information (author and publisher). Description of each book is preprocessed, and features of a book are selected according to term frequencies and log likelihood ratio. In the next step, experts refine these selected features according to their domain knowledge, and give a proper weight for those features. After feature selection, descriptions of books are transformed into vector forms and use the SVM classifier to learn and classify. The final step is to linearly combine the SVM classification results and meta-information for obtaining the final classification. The experiments use the book data from "www.books.com.tw", and *K*-fold cross-validation is used to verify the feasibility. The experimental results show that combining SVM and meta-information can make the classification accuracy to 95%.

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

1. 前言

文件分類(Document Classification/Text Categorization)是指依文件的內容給予適當的類別標籤(Class Label)，以達到分類管理與利用的目的 [9]。舉例來說，新聞文件可依其報導的內容，歸類為政治、財經、科技、娛樂等類別。

傳統的分類工作藉由專家閱讀文件的內容，進而決定該文件的類別。此方法耗費大量的人力與時間，且人工分類會因主觀認知的差異性造成分類的不一致。現今，資訊科技的進步使得大量的數位化文件不斷累積，導致已無法單單藉由人工進行分類。因此，如何利用文件自動分類技術，快速且有效地協助人工的分

類作業，已然成為現今資訊服務與知識管理的重要研究課題。

文件自動分類技術參考已分類的訓練資料(Training Data)，藉由機器學習(Machine Learning)可以建立分類模型。當有新文件需要進行分類時，便可依據既有的分類模型自動將該文件歸類到一個或多個類別。有鑒於文件自動分類技術發展多時且已被廣泛地運用，本研究嘗試將文件自動分類的方法應用於圖書分類。研究成果可使用於圖書館或網路書店，有效地節省分類大批書籍時所需的人力及時間。分類過的書籍，亦可提供讀者依主題類別的方式查詢。

表 1: 文件與圖書資料的比較

	文件資訊	圖書資訊
敘述資料	本文	書名
		內容簡介
		作者簡介
詮釋資料	--	作者
		出版社

一般來說，文件自動化分類針對文件的文本(Text)進行分析。然而，就圖書分類而言，大部份的情狀下缺乏已數位化的內文，可以利用的資訊通常只有書名、內容簡介、作者與出版社等資料。如表 1 所示，本研究將書名、內容簡介與作者簡介等資料，稱為圖書的敘述資料(Description Data)。作者與出版社等資訊則歸類為圖書的詮釋資料(Meta-Information)。本研究的目的為建立一套圖書自動分類系統。導入文件自動分類技術分析圖書的敘述資料作為分類依據；同時，運用圖書的詮釋資料提高分類成效。本研究亦探討如何整合專家的知識提昇分類效能。

針對圖書資料的特性，本文提出一套完整的圖書分類方法。說明如下：(1) 對敘述資料進行處理，透過特徵挑選(Feature Selection)技術擷取出具有類別代表性的特徵。同時，針對專家事先所建立的類別特徵，給予適當的加權後，透過 Support Vector Machine (SVM)分類器產生分類模型。(2) 以統計方法分析詮釋資料，發掘有助於圖書分類的資訊。(3) 以線性組合方式(Linear Combination)整合 SVM 的分類模型與詮釋資料的統計資訊建立可行的圖書分類模型。

本文共分成五個章節。首先，第二章介紹文件自動分類的相關研究。第三章說明我們所提出之圖書分類系統的設計；詳敘如何導入專家的知識於圖書特徵挑選，並整合 SVM 分類器與詮釋資料統計資訊建立圖書分類系統。第四章說明實驗方法，並評估分類成效。最後，第五章總結本研究，並提供未來可繼續發展的研究方向。

2. 相關研究

過去的研究已提出多種不同運用機器學習方法的文件分類器，如[2][8]。以文件中所擷取的詞彙(Keyword)作為分類特徵或屬性，其特徵值定義為詞彙出現於該文件中的頻率，並套用分類演算法進行文件分類。常見的分類演算法包含 Decision Tree [5]、Naive Bayesian [1]、*K*-Nearest Neighbor [4]與 Support Vector Machine [7]等。

2.1 Decision Tree

如圖 1 所示，決策樹 (Decision Tree) [5]是一個樹狀結構，其每個內部節點(Node)代表一種分類條件的測試，每個分枝(Branch)代表一種測試的結果，而每個樹葉節點(Leaf Node)則代表類別(Class)或類別分佈(Class Distributions)。決策樹採用 Divide-and-Conquer 的方法，以一種從上而下遞迴(Top-down Recursive)的方式依序建構樹狀結構。在每個節點上根據 Information Gain (IG)，計算每個資料屬性的 IG，並選取具有最高 IG 值的屬性，作為分裂的依據。

決策樹可以使訓練資料中屬於同類別的資料最後歸於同一個樹葉節點。由訓練資料建立好決策樹後，新進的測試資料經由一連串的條件測試後，最後可歸類於某個樹葉節點完成該筆資料的分類。

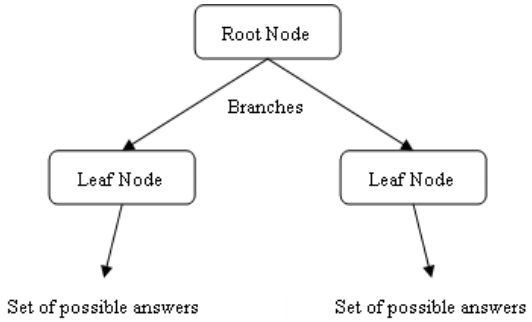

圖 1: Decision Tree 示意圖

2.2 Naive Bayesian

Naive Bayesian 分類器[1]是一種基於統計機率所發展的分類法。給定測試資料 d_i 和類別 C_k，藉由貝氏定理(Bayes' Theorem)可定義 d_i 屬於 C_k 的機率：

$$P(d_i \in C_k) = \frac{P(d_i \mid C_k)P(C_k)}{P(d_i)} \quad \text{Eq. (1)}$$

其中，$P(d_i)$與 $P(C_k)$為常數。

藉由計算 $P(d_i \mid C_k)$可決定 d_i 的類別。假設 d_i 具有 m 個屬性且屬性間的關係是獨立，則可推導得到 Eq. (2)：

$$P(d_i \mid C_k) = \prod_{j=1}^{m} P(d_{ij} \mid C_k) \quad \text{Eq. (2)}$$

其中，d_{ij} 是 d_i 的第 j 項屬性，$P(d_{ij} \mid C_k)$可經於訓練資料事先運算得到。

2.3 *K*-Nearest Neighbor

最鄰近分類法 (Nearest Neighbor Algorithm, NN) [4]以空間中的點表示資料，並假設屬於同類別的點之間的距離會比較接近。依此，對於一筆未知類別的測試資料，只要尋找在訓練資料中和該資料距離最接近的點，便可決定該筆資料和此最接近的點屬於同類別。圖 2 為一 NN 的實例。圖中訓練資料分屬兩個類別 C_1 (三角形)與 C_2 (正方形)，其中 C_2 的資料 y 與測試資料 x (圓形)距離最近，因此 x 被標示為類別 C_2。

通常，在訓練資料具有很多雜訊的情況下，若只用最靠近的資料來決定測試資料的類別時，可能會導致分類的效果不彰。常見的做法為先取與該筆測試資料最接近的 K 個訓練資料點，再依據其對應的 K 個類別進行投票(Voting)或加權投票(Weighted Voting)以決定最後的類別。此方法稱為 K-最鄰近分類法(K-nearest Neighbor Algorithm, K-NN)。

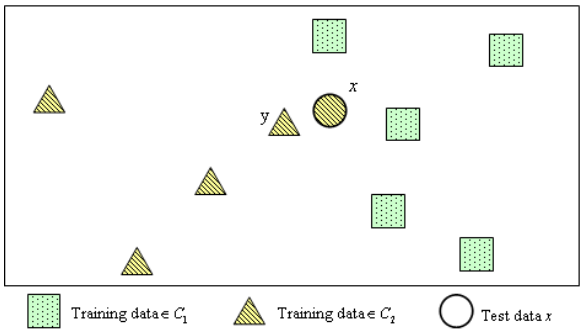

圖 2: NN 實例

2.4 Support Vector Machine

支援向量機(Support Vector Machine, SVM) [7]，是一種以統計學習理論(Statistical Learning Theory)為基礎所發展出來的方法。如圖 3 所示，當多維空間中的資料要分成兩類時，SVM 找出一個超平面(Hyperplane)，將屬於兩個類別的資料點分開。超平面與其最靠近的資料點間的距離稱為邊界(Margin)。SVM 的目的在於尋找一個具最大邊界的超平面(Maximum-Margin Hyperplane)，使得訓練資料點明確地分開，如圖中所標示之 Optimal Hyperplane。空間中最靠近最佳超平面的資料點稱為支持向量(Support Vector)。當有新進未分類的測試資料時，利用此超平面可正確判定該資料所屬的類別。

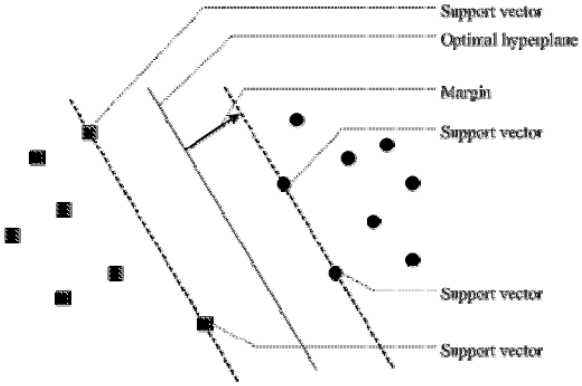

圖 3: SVM 概念示意圖

假設空間中有 n 個資料點集合為 $\{(\vec{x}_1, c_1), (\vec{x}_2, c_2), ..., (\vec{x}_n, c_n)\}$，其中 $c_i \in \{+1, -1\}$ 代表資料點 $\vec{x}_i$ 所屬的類別。利用這些資料點作為訓練資料，所找出的最佳超平面可表示為：

$$\vec{w} \cdot \vec{x} - b = 0 \quad \text{Eq. (3)}$$

其中 $\vec{w}$ 代表 margin，b 為一常數。

$\vec{w}$ 可能有很多解，但是只有具最大邊界的 $\vec{w}$ 才是最好的解。此求解問題相當於求解 Eq. (4)的最佳化問題。

$$\text{minimize} \frac{1}{2}\|\vec{w}\|^2 \quad \text{Eq. (4)}$$

$$\text{subject to } c_i(\vec{w} \cdot \vec{x}_i - b) \geq 1, 1 \leq i \leq n$$

SVM 經過學習之後，對於未知類別的測試資料，可依照 Eq. (5)計算得到其分類 $\hat{c}$。

$$\hat{c} = \begin{cases} +1, & \text{if } \vec{w} \cdot \vec{x} - b \geq +1 \\ -1, & \text{if } \vec{w} \cdot \vec{x} - b \leq -1 \end{cases} \quad \text{Eq. (5)}$$

3. 圖書自動分類方法

本研究針對每個圖書類別建立專屬的二元分類器(Binary Classifier)。對於新進的書籍資料，依序輸入各個分類器進行運算，判斷是否屬於該類別。因此，本研究所提之分類方法並不限制一本書只能分類到單一個類別。換句話說，一本書可以同時標示兩種以上的類別。

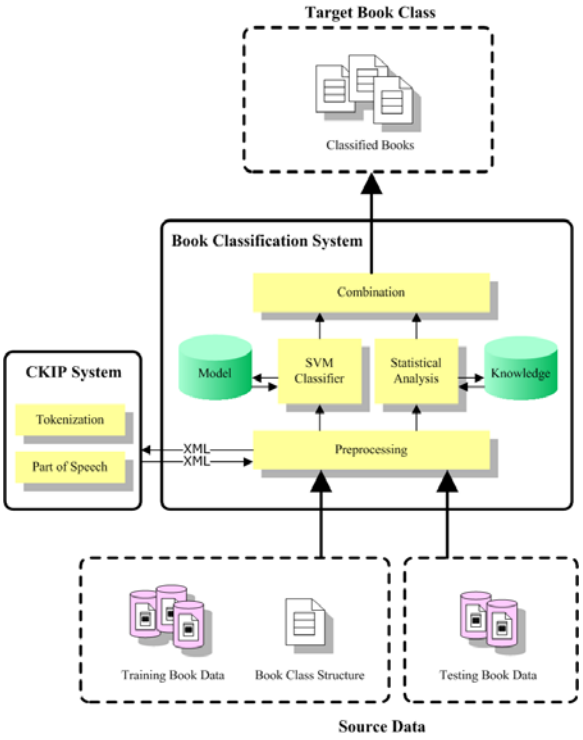

圖 4: 書籍分類系統架構

圖 4 為圖書分類系統架構圖。首先利用斷詞系統對敘述資料進行斷詞切字與詞性分析，之後過濾詞性並刪除停用字，再進行特徵挑選並加入專家的所建立之特徵詞，最後將圖書資料以向量的形式表示，經由 SVM 學習產生各類別的分類模型。此外，針對詮釋資料進行統計分析，由現有的歷史資訊整理出作者與出版社的類別特性。完成以上分類器的學習工作後，輸入未知類別的書籍資料，經由 SVM 模型分類，並整合詮釋資料的統計資訊決定書籍的所屬類別。

3.1 前置處理

前置處理包含斷詞切字(Tokenization)、詞性標記(POS Tagging)及刪除停用字(Stopword Removal)等步驟。此模組分析圖書的敘述資料，並擷取詞彙作為分類學習的特徵。

斷詞切字的目的在於從文字資料中擷取具有語意的詞彙。中文詞的結構，有單字詞、多字詞等型態，且中文文件僅有字與字的界線，詞彙與詞彙間並無明顯界線，因此必須先經由斷詞的程序將詞彙互相區隔。本研究利用中央研究院所研發的中文斷詞系統[1]，將圖書的敘述資料進行斷詞。該系統包含一個約 10 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，且具有新詞辨識及附加詞類標記等功能。

斷詞的結果，首先將標點符號等不具有語意的符號刪除。接著，針對單字詞的部分，本研究僅保留名詞(Nouns)與動詞(Verbs)的單字詞，其餘的單字詞均予以忽略。剩餘的多字詞則透過刪除停用字的步驟過濾。

停用字(Stopword)為出現頻率高但不具分類區別價值的詞彙，包含介系詞、指示代名詞、連詞、助詞等。這類詞彙會混淆分類學習的效果，因此必須先行剔除。我們建立中文的停用字表，共 90 個停用字，如圖 5 所示。

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

圖 5: 停用字範例

3.2 特徵挑選

在不影響分類效果的條件下，特徵挑選(Feature Selection)過濾對分類具較小影響力的詞彙或干擾分類的雜訊詞，以降低特徵詞的數目，提升系統效率。本研究採用詞頻(Term Frequency, TF)，搭配 Log Likelihood Ratio (LLR)作為特徵選取的方法。同時，亦加入專家的所建立的專業詞彙庫篩選，可使特徵的選取更為精準。

(1) 結合 TF 與 LLR 的特徵挑選:

觀察發現，詞彙與類別的關係與詞彙在特定類別中出現的次數相關。依此，本研究定義類別 C_i 與詞彙 t_j 的關聯強度 $R_{TF}(C_i, t_j)$ 為 t_j 出現於 C_i 的頻率，如 Eq. (6)所示。

$$R_{TF}(C_i, t_j) = \frac{n_{ij}}{\sum_k n_{kj}} \qquad \text{Eq. (6)}$$

其中，n_{ij} 代表 t_j 出現於 C_i 的次數。

表 2: 詞彙 t_j 與類別 C_i 關係表

	C_i	$\neg C_i$
t_j	O_{11}	O_{12}
$\neg t_j$	O_{21}	O_{22}

表 2 列出類別 C_i 與詞彙 t_j 的關係表，其中，O_{11} 為 t_j 出現於 C_i 類別的總數，O_{12} 為 t_j 出現於非 C_i 類別的總數，O_{21} 為非 t_j 的詞彙出

[1] CKIP 斷詞系統: <http://ckipsvr.iis.sinica.edu.tw/>

表 3: 特徵自動挑選實例 (前 10 名)

偵探/推理類		科幻/奇幻類		愛情文藝類	
詞彙	$R(C_i, t_j)$	詞彙	$R(C_i, t_j)$	詞彙	$R(C_i, t_j)$
推理	454254.30	世界	194732.45	愛情	1991069.11
偵探	181257.91	奇幻	181118.69	愛	651585.24
福爾摩斯	128168.04	人類	101075.83	男人	104342.86
兇手	95636.47	愛情	42771.39	女人	91192.02
羅蘋	81802.03	魔法	36868.11	心	89445.34
殺人	58986.36	冒險	33157.36	自己	81973.05
小說	45250.50	傳說	31994.54	幸福	69768.03
犯罪	42541.36	地球	25855.45	想	52614.18
愛情	41492.35	吸血鬼	24987.97	故事	43547.01
事件	69616.51	科幻	23024.74	說	29635.23

表 4: 專家建構的專業特徵詞 (部分)

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

現於 C_i 類別的總數， O_{22} 為非 t_j 的詞彙出現於非 C_i 類別的總數。本研究定義類別 C_i 與詞彙 t_j 的關聯強度 $R_{LLR}(C_i, t_j)$為 Eq. (7)。

$$R_{LLR}(C_i, t_j) = -2.0 \times \log \frac{b(O_{11}; O_{11}+O_{12}, p) b(O_{21}; O_{21}+O_{22}, p)}{b(O_{11}; O_{11}+O_{12}, p_1) b(O_{21}; O_{21}+O_{22}, p_2)} \quad \text{Eq. (7)}$$

其中:

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

$$p_1 = \frac{O_{11}}{O_{11}+O_{12}}, p_2 = \frac{O_{21}}{O_{21}+O_{22}}$$

$$p = \frac{O_{11}+O_{21}}{O_{11}+O_{12}+O_{21}+O_{22}}$$

最後，如 Eq. (8)所示，結合 $R_{TF}(C_i, t_j)$與 $R_{LLR}(C_i, t_j)$，可計算類別 C_i 與詞彙 t_j 的關聯強度。

$$R(C_i, t_j) = R_{TF}(C_i, t_j) \times R_{LLR}(C_i, t_j) \quad \text{Eq. (8)}$$

當 $R(C_i, t_j)$的值越高時，則表示詞彙 t_j 可用來區別類別 C_i 的代表性越高。表 3 列出幾個類別中所自動挑選的 10 個詞彙以供參考。

(2) 專家建立專業特徵詞彙:

透過特徵挑選選出對類別代表性較高的詞彙製成「候選特徵列表」，再經由專家新增或刪除詞彙，可建立專業詞彙列表。

專家的工作項目有兩項。首先，從候選特徵列表刪除不必要的詞彙。接著，依專家的經驗及知識，加入與類別相關性高且具代表性的特徵詞彙。表 4 列出專家對「偵探/懸疑」、「科幻/奇幻」及「愛情文藝」三種類別圖書所加入之專業詞彙的實例。

3.3 SVM 分類器

本研究採用 Support Vector Machine 做為核心分類演算法，使用的工具是 SVM^{light}[2]。首先，將圖書敘述資料轉換為以特徵為維度的向量表示式(Vector Representation)，再輸入 SVM 學習產生分類模型。

將敘述資料轉換為向量表示式的步驟如下：首先，以二元(Binary)的方式來表示。若敘述資料中有出現某個特徵，則其相對應的維度向量值設為 1，否則為 0。接著，針對由專家所指定加入的特徵，額外給予適當權重。

以下舉例說明如何將敘述資料轉換為向量式。表 5 為一特徵列表，包含 8 個由專家指定的特徵及 8 個系統自動挑選所得的特徵。圖 6 為一圖書敘述資料。其中，以外框標示所有有效的特徵詞彙，以灰底標示為專家特選的特徵。假設專家挑選的特徵權重設定為 3，則圖 6 的範例可表示為下列的的向量表示式。

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

[2] SVM^{light}: <http://svmlight.joachims.org/>

表 5: 圖書向量表示式實例: 使用特徵列表

專家指定的特徵							
屍體	動機	迷團	密碼	偵探	密室	推理	殺人

系統自動挑選的特徵							
線索	教授	政府	追查	一連串	謀殺	刑警	發現

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

圖 6: 向量表示式實例: 圖書敘述資料

3.4 詮釋資料

本研究嘗試從圖書的詮釋資料進行統計分析，藉由過去出版的歷史紀錄，探勘有利於圖書分類的資訊。所採用的詮釋資料包含「作者」與「出版社」兩項。

在作者的部分，假若作者 A_j 在類別 C_i 出版很多著作，則可判斷 A_j 對於 C_i 具有高度代表性。因此，當 A_j 出版新著作時，便可合理推敲該著作屬於類別 C_i 的可能性很大。反之，假若 A_j 沒有出版過任何屬於類別 C_i 的著作，則可推論 A_j 所出版的新書屬於類別 C_i 的可能性亦不高。圖 7 為計算作者 A_j 與類別 C_i 間相關資訊的流程圖。其中，$|A_{j,C_i}|$ 代表 A_j 在 C_i 類別中的著作數量，$|A_{j,\overline{C_i}}|$ 代表 A_j 在非 C_i 的類別中所有的著作數量，$|A_j|$ 是 A_j 的所有著作數量，$W(A_j,C_i)$ 為 A_j 在 C_i 的類別相關度，而 θ_A 是一個常數。

在出版社的部分，考量到出版社有規模大小之分，例如某些大出版社出版項目可能橫跨各種類別，也有些出版社雖然規模不大，但卻專精於某個類別的出版品。因此，本研究以出版品的類別比例來判斷出版社與類別的相關程度。舉例來說，假設出版社 P_1 的總出版品有 5,000 冊，其中有 200 冊為類別 C。另外，出版社 P_2 其總出版品為 150 本，卻全部屬於 C。雖然 P_1 在 C 的出版品數量較多，但是其類別出版比例為 1 (150/150=1.0)，我們仍認為 P_2 對 C 的類別相關性較高。

圖 8 為計算出版社與類別之間相關資訊的流程圖。其中 $|P_{k,C_i}|$ 代表出版社 P_k 在類別 C_i 的出版品數量，$|P_{k,\overline{C_i}}|$ 是 P_k 在非 C_i 類別的所有出版品數量，$|P_k|$ 為 P_k 的所有出版品數量，$W(P_k,C_i)$ 則代表 P_k 在 C_i 的的類別相關度，而 θ_P 是一個常數。

3.5 整合 SVM 與詮釋資料統計分析

3.1 節至 3.4 說明如何從訓練資料集中建構圖書分類器，本節介紹當輸入未分類的圖書時，如何利用前幾節所建立的 SVM 分類模型，並配合詮釋資料的分類特性，以決定該書所屬的類別。

依 Eq. (9)將 SVM 的分類結果與出版社、作者資訊作線性組合，可得到最後的分類準則。若 $W(B,A,P,C)>0$ 則將書籍 B 標示為屬於類別 C。反之，若 $W(B,A,P,C)\le 0$，則 B 不屬於 C。

$$W(B,A,P,C)=\alpha\times WS(B,C)+\beta\times WA(A,C)+\gamma\times WP(P,C) \quad \text{Eq. (9)}$$

其中，$W(B,A,P,C)$代表一本書 B 要分到類別 C 裡面的分數，而且這本書的作者為 A、出版社為 P。$WS(B,C)$則為 SVM 分類法將 B 代入類別 C 的學習模型所得的結果，$WA(A,C)$為作者與類別的統計分數，$WP(P,C)$為出版社與類別的統計分數。

4. 實驗與結果分析

4.1 實驗資料

實驗資料取自於博客來網路書店[3]，分別從「偵探/懸疑」、「科幻/奇幻」與「愛情文藝」三個類別各下載 900 本圖書資料，總共 2,700 本書作為實驗的資料集。

4.2 實驗方法

實驗採用 K-fold Cross-Validation [3]。隨機將三個類別的圖書資料分成 9 份，每一份包含 300 本書(註：三個類別各 100 本)，進行 9-fold Cross-Validation。實驗時依序取其中一份作為測試集，其餘 2400 (800×3=2400)本書作為訓練集以產生分類器，再將九次實驗的數據平均作為評估結果。

實驗步驟分以下四個部分進行：

[3] http://www.books.com.tw/

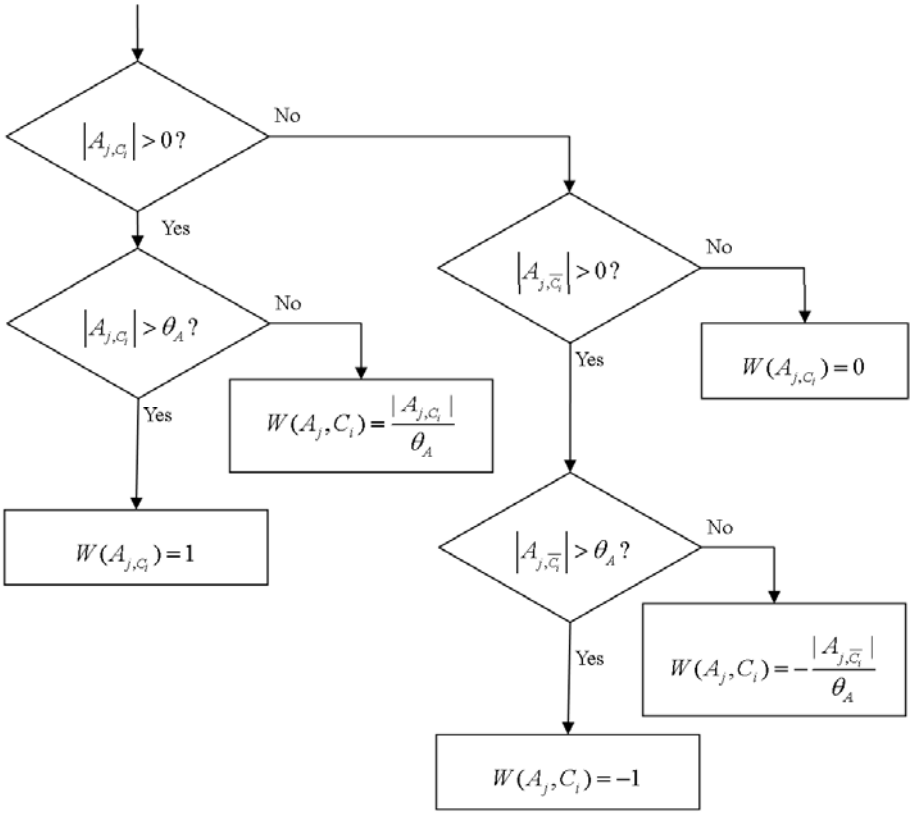

圖 7: 作者-類別相關資訊計算流程圖

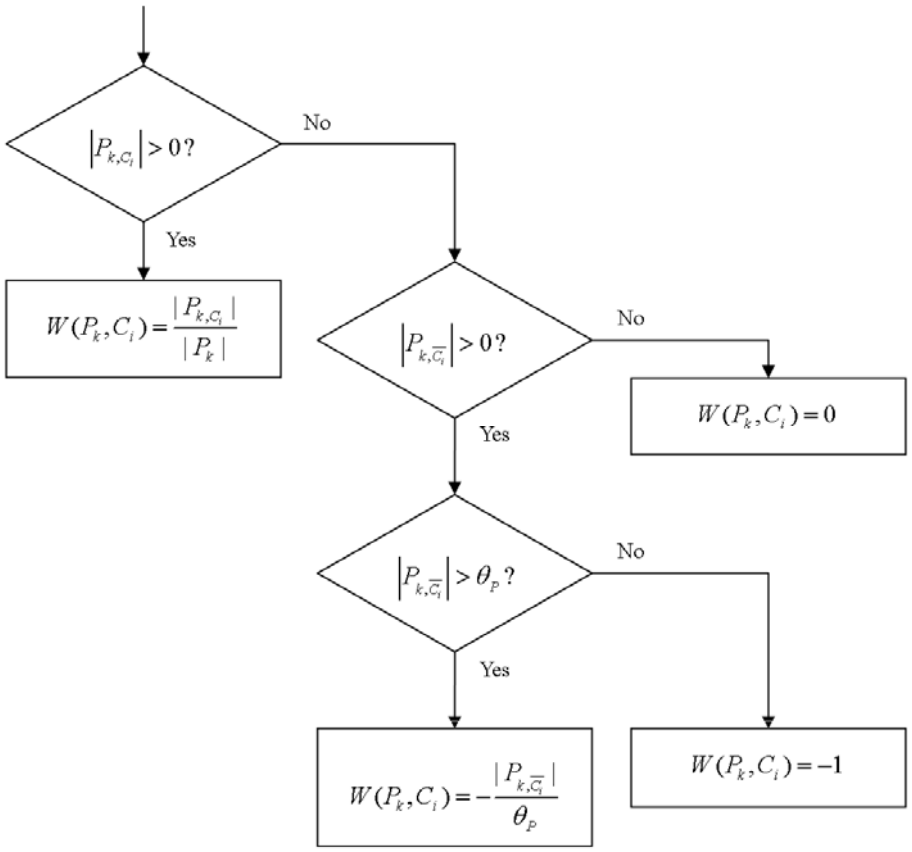

圖 8: 出版社-類別相關資訊計算流程圖

1. 選擇適當的特徵個數進行 SVM 分類；
2. 比較加入專家所建立特徵詞會的 SVM 分類成效，並為專家指定的特徵詞選擇適當的權重；
3. 比較 SVM [7]、ID3 [6]與 Naive Bayesian [1]三種分類演算法的成效；
4. 加入詮釋資料，調整 SVM 分類結果與詮釋資料之適當比重。

4.3 評估標準

首先，將測試集的圖書分類結果與博客來網路書店的分類進行比對，針對單一類別考慮「屬於」或「不屬於」兩種情況。如表 6 所示，S_c 為測試集中屬於類別 C 的書籍集合，$S_{\bar{c}}$ 代表不屬於 C 的書籍集合，R_c 與 $R_{\bar{c}}$ 分別代表測試集中被分類系統標示為屬於與不屬於 C 的書籍集合。O_{11}、O_{12}、O_{21} 與 O_{22} 則分別代表真-正例(True Positive)、偽-正例(False Positive)、偽-反例(False Negative)與真-反例(True Negative)的數目。舉例來說，O_{11} 代表屬於類別 C 且被系統標示為 C 類別的書籍數目。

表 6: 分類結果列聯表

	S_c	$S_{\bar{c}}$
R_c	O_{11}	O_{12}
$R_{\bar{c}}$	O_{21}	O_{22}

$$Accuracy = \frac{O_{11} + O_{22}}{O_{11} + O_{12} + O_{21} + O_{22}} \quad \text{Eq. (10)}$$

$$F - measure = \frac{2PR}{P + R} \quad \text{Eq. (11)}$$

$$P = \frac{O_{11}}{O_{11} + O_{12}}, \quad R = \frac{O_{11}}{O_{11} + O_{21}}$$

本研究採用正確率(Accuracy)與 F-measure 作為評估的分類結果好壞的標準。如 Eq. (10)與 Eq. (11)所示，可定義正確率與 F-measure。

4.4 實驗結果與分析

本節從多個面向說明與分析實驗的結果。4.4.1 節決定進行 SVM 分類時所需的特徵個數；4.4.2 節利用專家的智慧與經驗進行特徵挑選；0 節比較 SVM 與其他分類演算法；0 節說明並分析結合 SVM 與詮釋資料進行分類的實驗結果。

4.4.1 特徵個數

本實驗以自動選取特徵的方式，比較從每個類別選取 50 個、100 個與 150 個特徵進行 SVM 分類對分類結果的影響。由表 7 可知，選取 150 個特徵有利於 SVM 的分類。

4.4.2 加入專家建立的特徵詞

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

表 8 為實驗結果，其中 Without Expert 代表由系統自動選取 150 個特徵且專家完全不參與特徵挑選的分類結果，其餘各列以 W_Exp 代表專家指定特徵之權重。

實驗結果顯示，若僅是讓專家介入特徵挑選，卻不將專家指定的特徵加權(W_Exp=1)，則在「偵探/懸疑」及「科幻/奇幻」兩類別上，SVM 分類結果較完全自動挑選的成效稍差。這是因為由專家指定的特徵，在以二元的向量

表 7: 特徵個數對 SVM 分類成效的影響

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

表 8: 專家指定特徵詞設定不同權重對 SVM 分類成效的影響

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

表 9: 比較分類演算法

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

表 10: 比較 SVM 與詮釋資料的分類結果

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

表 11: SVM 配合詮釋資料分類結果

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

表 12: SVM 配合詮釋資料組合的最佳分類結果 (α=0.625*0.05;β=0.375*0.05;γ=0.95)

WS	WA	WP	偵探/懸疑小說		科幻/奇幻小說		愛情文藝小說	
			Accuracy	F-measure	Accuracy	F-measure	Accuracy	F-measure
A	B	γ	97.1%	95.6%	94.7%	91.6%	95.1%	92.8%

表 13: 取部分區間配合詮釋資料分類

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

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

表示式中無法與其他特徵詞區別，若不給予加權則 SVM 分類時無法凸顯其重要性。相反地，若將專家指定的特徵加權，則可發現對 SVM 分類的結果有正面的提升。

4.4.3 比較分類演算法

表 9 比較 SVM、ID3 與 Naive Bayesian 三種分類演算法的分類效果。當使用 150 個自動挑選的特徵時，Naive Bayesian 的分類結果最好，SVM 略差一點，而 ID3 則明顯遜色許多。然而，若考慮專家挑選的特徵與加權的話，SVM(W_Exp=5)則具有最好的分類結果。

4.4.4 配合 SVM 與詮釋資料進行分類

本實驗結合 SVM 與詮釋資料進行分類，且配合先前各實驗中所取得表現最好的組合方式。表 10 至表 12 中 WS 代表 Eq. (9)中 α 的值，WA 代表 β 的值，WP 則代表 γ 的值。

由表 10 可知，「科幻/奇幻」類的 SVM 分類結果比其他兩類差。在單以作者統計資訊進行分類的情況下，「愛情文藝」類的分類結果最差，推測此類別有許多著作量較少的作家，或是一直有新作家。而「偵探/懸疑」類別在依照出版社統計資訊分類的結果明顯優於其他兩類，顯示此類別的出版品較具有獨佔性，有較多專門出版此類書籍的出版社。

表 11 列出搭配一種以上的方式進行分類。若僅從 SVM 分類結果、作者統計資訊，及出版社統計資訊中組合兩種方式進行分類，則經由實驗發現 SVM 配合作者統計資訊的分類結果較優於其他兩種組合。因此，我們利用 Greedy 的方式，先調整 SVM 與作者統計資訊兩者的比例，再加入出版社統計資訊以求得最後的參數值。由表 12 可知，相較於原本單純用 SVM 分類的正確率約為 90%上下，加入詮釋資料後，可將成效提高至 95%。

為更進一步了解加入詮釋資料後對 SVM 分類結果所產生的影響，本研究針對詮釋資料所提升的分類成果進行分析。實驗僅對 SVM 分類結果落於指定區間的書籍(亦即，落於圖 3 之 Margin 內的測試資料)，加入詮釋資料做進一步分類，其餘部分書籍直接以 SVM 輸出作為分類結果。

表 13 中，SVM 列代表僅採用 SVM 分類結果，其餘各列為取部分書籍配合詮釋資料進行分類。舉例來說，Range=±0.25 代表僅取 SVM 輸出值為±0.25 之書籍，再套用 Eq. (9) 計算分類結果，其餘書籍不加入詮釋資料，直接以 SVM 之輸出作為分類結果。Total 列為將全部書籍配合 SVM 與詮釋資料進行分類。由此表可發現考慮 0~±1 區間時，加入詮釋資料對分類有很大的幫助。

5. 結論

本研究嘗試將文件自動分類的方法應用於圖書分類。研究成果可應用於圖書館或網路書店，有效地節省分類大批書籍時所需的人力及時間。分類過的書籍，亦可提供讀者依主題類別的方式查詢。

本研究將圖書資訊分為敘述資料與詮釋資料兩部分。敘述資訊包含書名、內容簡介與作者簡介；詮釋資料則包含作者與出版社資訊。針對圖書資料的特性，本文提出一套完整的圖書分類方法。說明如下：(1) 對敘述資料進行處理，透過特徵挑選(Feature Selection)技術擷取出具有類別代表性的特徵。同時，針對專家事先所建立的類別特徵，給予特別的加權。最後，透過 Support Vector Machine (SVM) 分類器產生分類模型。(2) 以統計方法分析詮釋資料，發掘有助於圖書分類的資訊。(3) 以線性組合方式(Linear Combination)整合 SVM 的分類模型與詮釋資料的統計資訊建立最終的圖書分類模型。

實驗使用「博客來網路書店」的圖書資料做驗證，並以 9-fold Cross Validation 的方式進行實驗。實驗結果顯示，針對專家事先所建立的類別特徵，給予特別的加權，可以提升 SVM 分類成果約 5%。若再將 SVM 分類結果整合詮釋資料中隱含的分類資訊，可再提高約 5%，使得整體正確率達 95%。

5.1 未來研究方向

以下幾點可做為未來研究方向之參考。

(1) 分類演算法

本研究嘗試以 SVM 分類法作為核心分類演算法，然而實驗過程中發現 Naive Bayesian 分類法應用於分類敘述資料亦有不錯的成果，將來可嘗試以 Naive Bayesian 或其他分類演算法作為分類核心或是結合多種分類演算法，以投票表決的方式(Voting)決定敘述資料之分類結果，最後再整合詮釋資料進行圖書分類。

(2) 詮釋資料的擴充

本研究初步探討詮釋資料對分類成果的影響，採用的詮釋資料為作者與出版社兩

項。然而圖書資料還有許多其他的資訊，例如主題標目(Subject Heading)與分類號(Classification Number)，未來可針對這些資訊進一步探討其與分類的關係。

參考文獻

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

[2] Farbrizio, S., "Machine Learning in Automated Text Categorization," ***ACM Computing Surveys***, Vol. 34, No. 1, pp. 1-47.

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

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

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

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

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

[8] Yang, Y., and X. Liu, "An Examination of Text Categorization Methods," In ***Proceedings of SIGIR-99***. 1999.

[9] 曾元顯, "文件主題自動分類成效因素探討," ***中國圖書館學會會報***, 2002 年 6 月, 第 68 期, 頁 62-83.