天主教輔仁大學圖書資訊學系碩士班碩士論文

指導老師：陳舜德

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

A Study of Book Title Feature Extraction Based on The Automatic Classification -An Example of Bibliography Automatically Classified System

研究生：黃嘉宏 撰

中華民國九十七年六月

謝辭

首先感謝我的指導教授陳舜德老師。總願意在學生碰到瓶頸而焦慮不堪之時，撥冗與學生溝通討論，解決問題的癥結點，學生才得以完成這篇論文。辛苦了!老師!

此外感謝曾元顯、吳政叡、張淳淳三位老師於學生大學至研究所這段時間的教導，讓學生在程式寫作、對資訊檢索領域的認識以及論文寫作技巧上皆打下了深厚的基礎。尤其是撰寫系統及論文寫作期間，學生更深切的感受到老師們所傳授的知識與技能，對學生完成論文的幫助有多大。

LE507 三朵花(靜宜、小童、瑩儒)，謝謝妳們總在我低潮時聽我說話、給我勇氣。對妳們的感激真的不是三言兩語就能道盡。 Anyway有妳們真好!!

MS493的同學們，真的很懷念那段大家都還在學校的日子。雖然大家的畢業時間不同，二年級後大家也多忙於自己的研究或工作。但還是時不時的可以感受到大家對我的關心(淑美、麗純、文樺，甘溫啊~)。謝謝你們!!

炳魁，最好的戰友，感謝有你與我在LE504裡面一起打拼，我想這段日子所堆砌出的酸甜苦辣、慘綠午後兩個研究僧從光華商場走到西門町時所留下的足跡。這些許多許多都將成為你我心中永難磨滅的印記。雖然，你即將離開台北回到故鄉打拼，但未來的日子...一起加油，好嗎?

最後感謝我的父母，感謝你們忍受我這段時間的壞脾氣，並支持我完成學業。

總之，要感謝的人太多了，就謝天罷!!

2008 6月 黃嘉宏 筆

摘要

有別於新聞標題或者網路書店所提供的書目資訊，圖書館館藏書目資料，其記載之資訊，多為描述圖書之外在表徵；與圖書內容較相關的紀錄欄位以圖書題名與作者欄位為主。其中圖書題名又有詞彙運用較為自由（題名不一定能反應圖書內容），以及資訊負載量不一的情況。倘若單用圖書題名進行文件自動分類，則容易導致成效不彰的結果。有鑑於此，本文提出以搜尋引擎進行擴展文件特徵並以作者欄位資訊進行輔助之策略。希望能就擴展文件特徵之觀點，改善以圖書館書目資料進行文件自動分類的成效。實驗證實，此策略確實能夠有效輔助藉由圖書書目分類的成效。

Abstract

In comparison to the bibliographic information from news headlines or Internet bookshops, most library bibliographies describe collections by their initial appearance. Among the record fields in library bibliographies, the title and the author can mainly reflect the content of the collections. The title can be described in a more flexible way; however it does not exactly reflect the content of the collection. Also, the information load is different. If collections are automatically classified only by titles, it may not reach a reasonable success in the automatic classification. In light of the potential problems, the research puts forward a strategy, which aims to utilise both search engine, such as Google, to expand document features and the author as a record field to assist classifying. The experiment proves that this proposed strategy is helpful for promoting the effect of automatic bibliography classification effectively and efficiently.

目錄

圖目錄

表目錄

圖表目錄

第一章 緒論

第一節 研究背景與動機

一直以來圖書館除了保存圖書的使命外，將圖書做有系統的整理進而提供讀者使用一直是圖書館最主要的工作。而有系統的整理指的便是分類編目。

所謂的分類編目指的是兩樣工作：1.圖書分類、2.圖書編目。圖書分類指的是根據圖書的內容，並依據所採用的分類法，選擇最適當的類目而給予類號，其目的在給予書本在書架上有個固定的位置，並將內容相近之圖書聚集在一起。而圖書編目指的是根據書籍的外部表徵如：書名、作者、出版社、高廣、頁數等等資訊記錄起來，為的是日後供讀者查詢圖書之用。

目前圖書分類的工作皆由編目館員根據：1.書名、2.目次、3.內容簡介、4.序跋凡例、5.導言或緒論、6.正文、7.參考其他書目或專家意見等資訊經過館員自己的認知以及經驗來進行類目的判別。(王省吾，1982)

然而我國圖書館教育的方式與英美有所不同。國內的圖書館系所所培養出來的人才，多半只懂得圖書資訊領域的知識。但編目館員必須負責所有到館圖書的分類編目工作，因此常有因學科背景知識不足造成給號困難的情形。再者我國自解除戒嚴之後出版業蓬勃發展，再加上近年來各學科領域皆有長足進步造成出版圖書的數量年年增加。編目館員的負擔可說是越來越沈重。

相似的情形也發生在網際網路上。以往的入口網站，為了要讓其使用者方便利用網際網路的資訊，除了提供搜尋引擎外也提供目錄的方式將相同主題的網站集中起來，減少讀者過濾過多資訊的時間。

以往這些網站的分類工作都是經由學科專家以人工的方式，根據網站的內容

依據其制定的類目進行分類的動作。但網際網路上的網頁、文件每日都以驚人的速度成長。若只憑著人工的方式來進行分類的工作顯得十分的不符合經濟效益。因此文件自動分類的技術便因應而生。

目前文件自動分類技術已經廣泛的運用如：網頁、新聞、電子郵件、專利文件等各種不同的文件形式。而利用圖書館書目資料進行自動分類以協助編目工作的研究則相當的稀少。因此，筆者希望能透過這次的機會，以圖書館書目資料中的圖書題名等資訊進行自動分類的實驗，找出適合於以圖書書目資訊的特徵擷取方法及書目中其他資訊用於輔助自動分類的搭配方式。希望不久的將來這項研究的成果能減輕圖書館編目館員在圖書分類工作上面的負擔，進而增進圖書館提供服務給讀者的效率。

第二節 研究問題與目的

在圖書書目資料中，通常最能反映圖書內容的欄位，即為圖書題名。因此本實驗的研究方向，基本上是以圖書書目中的圖書題名為主，而其它在書目中並有助於分類的項目，我們亦將透過實驗的方式，逐步將該因子加入分類器中，來改善分類的成效。

已有一些利用文件標題來進行文件自動分類的研究。從發表的文獻中顯示利用少量的文件標題就可以得到不錯的叢集分散效果（Kwok，1978），而國內也有研究者利用新聞標題進行自動分類的研究（ 杜海倫，1999；許雅芬，2002）。而文件或新聞標題在意義與訊息長度上與書目類似，但用字上書目資料更顯精簡，所以在分類技術上的困難度也較高。

在利用新聞標題自動分類的研究上。由於新聞標題必須讓讀者一眼就能了解新聞大致的內容，因此有用字精準、精簡並能反映內文的特性，可說是新聞主文的縮影，讓讀者在閱讀新聞標題時即可大略的了解新聞內文的大意，進而判斷是否要繼續閱讀新聞內文。因此雖然新聞標題的文字量不如新聞內文，但卻有足以代表本文的訊息可供分類器進行分類。目前在國內外已有許多的研究者投入新聞標題自動分類的研究。

反觀以圖書題名做為自動分類實驗資料的研究，能見度可說是相當的低。因此直接利用書目中的圖書題名來進行自動分類會遭遇到什麼樣的問題，筆者僅能以自己實際參與編目工作時的經驗來臆測。

根據筆者實際觀察書目時的經驗發現，與新聞標題相比，圖書題名之特性如下：

1. 題名長度不一：

新聞標題之長度雖然精簡，但往往能夠囊括與反應內文中的重要訊息。使人們在觀看新聞標題的同時，便能了解內文的大意。反觀圖書題名在

長度亦有相似的現象，部分圖書的題名是可以反應圖書內容意義，但也常見以一字詞作為題名的圖書，要對這些字數較少的圖書題名，進行自動編目分類，便會因特徵量不足而影響分類的成效。

2. 用字較為自由：

根據筆者實際觀察書目時的經驗發現，在以人工進行分類時，大部分的圖書題名是可以讓編目館員做分類的依據。王省吾(1982)曾在其著作中提到：「書名大都可以代表一書的內容，所以分析一書書名的涵義，可以幫助了解這本書的性質。」，黃淵泉先生也在其著作中提到類似的看法(黃淵泉，1986)。不過與新聞標題不同的是，圖書題名用字可說是相當的自由，特別是在文學、哲學、社會科學等類別的書，這些類別的書名常無法適切的表示圖書的內容與範圍(王省吾，1982；黃淵泉，1986)。

因此筆者認為在以圖書書目之圖書題名為主要分類素材的分類系統進行特徵擷取時若沒有適當的特徵擷取方法，可能會影響系統分類的成效。

在研究初期時筆者為了改善上述圖書題名之特徵，曾朝著「擴展文件特徵」、「分析語意」兩個方向構思，希望能改善圖書題名之「題名長度不一」、「用字較為自由」之特性所可能造成分類成效不彰的情況。在分頭進行可行性分析時發現，目前可見的語意分析工具多半只針對特定領域，通用的語意分析工具則往往有詞彙不夠齊全的情況。然而圖書館所典藏之藏書，包羅萬象且涵蓋各時期之圖書。利用現有的語意分析工具來進行圖書題名之特徵語意分析，往往有不敷使用之感。因此在擷取圖書題名特徵方面，本研究將朝著擴展文件特徵之方向進行實驗。

另外圖書館所利用的分類號也較新聞所用的分類要來的複雜許多。在國內目前主要使用的圖書分類法為「中國圖書分類法」，是一個階層式的十進分類法，分為三層，第三層後亦可根據書籍的性質進行複分。也就是不將複分列入計算的

話，中國圖書分類法理論上的類別數就有999類，類別的數目可說相當的多。另外由於學科發展不一致的緣故，造成在圖書館中有些類別有資訊超載的現象，而有些類別卻只有少數幾本書目。一般而言當分類系統的訓練文件越多，其分類成效也會越好。當大類小類的差距越大時，分類系統對小類的判斷能力可能會較差，而影響分類系統的成效。以上所述，可能只是中國圖書分類法可能會影響自動分類成效之因素的冰山一角。還有待透過實驗來挖掘其它隱藏在中國圖書分類法分類架構中的影響自動分類成效之因素。

綜合以上所述，本研究的目的在於以現有常用的特徵擷取方法為基礎，發展出適用於圖書書目資訊中圖書題名之特徵擷取方法，並希望能以書目中其他有利於圖書分類的資訊來輔助分類系統提升分類成效。此外有鑑於大多數台灣地區圖書館所採用的中國圖書分類法要比一般書店或新聞所採用的分類方式要複雜，因此本研究將透過實驗進行觀察，希望能找出因圖書分類法而影響圖書自動分類的因素，以提供日後研究者參考之用。希望本研究的成果，在不久的將來可以減輕圖書館編目館員的負擔，並進一步縮短圖書館新書上架服務讀者的時間。

第三節 研究範圍與限制

本研究之研究範圍與限制如下：

1. 字元集轉換問題：

由於我國圖書館書目資料庫中的字碼採用的是CCCII字元集，其優點在於其編碼空間足以容納所有的中文字。然而大五碼（Big5）在所收錄的中文字數不如CCCII但因為其在台灣較為普及的關係，因此線上公用目錄不得不採用大五碼（Big5）來顯示書目查詢結果。由於CCCII字元集裡的字數較Big5字元集多，若是在查詢時，書名內有Big5字元集中所沒有涵蓋的字，便會顯示為亂碼，成為日後字串處理的雜訊。另外為了日後研究字元集顯示上的方便，會先將辭彙從Big5碼轉換成萬國碼（Unicode）來存入資料庫，在字元集的轉換過程，未能對應的字元會產生雜訊的機會。

2. 資料收集：

為考量收集資料的便利性以及書目資料及其書目分類的權威性，本實驗的書目資料來源是從國家圖書館館藏查詢系統所收集而來。由於國家圖書館內的館藏並非都是採用中國圖書分類法做書目分類的依據。因此為了日後實驗在分類法上有統一的依據，收集書目時只收集使用中國圖書分類法做分類依據的書目資料。由於書目資料庫中的資訊是不斷的在更新的，因此日後若是有人要以本實驗的實驗方法重現實驗，其實驗數據可能會因為資料收集的時間點不同而有所改變。

3. 斷詞器所產生的研究限制：

在中文資料的處理上，通常首要的步驟就是斷詞，選用斷詞器的不同，實驗的結果亦會有些許的出入。實驗中我們採用中央研究院所開發的CKIP自動斷詞系統，其內建有中央研究院詞庫小組的中文詞庫，及從中央研究院平衡與料庫中所抽取約兩萬目詞的額外詞條，外加定量詞及重疊詞構詞律，做為其進行斷詞的基礎。

第四節 論文架構

本論文將分為五章，分別如下：

第一章 緒論：

分為四小節，主要闡述本論文的研究背景與研究動機、研究問題與目的與研究的範圍與限制以及簡略介紹本論文各章節之內容概要 。

第二章 文獻分析：

針對與本研究相關的文獻做有系統的回顧與分析。主要回顧的文獻主題分別有：1.特徵擷取、2.特徵權重計算、3.向量空間模型（Vector Space Model, VSM）、4.其他相關之研究、5. 文件自動分類常見的評估方式 。

第三章 研究方法：

本章將介紹本實驗所採用的研究方法與實驗的評估方式。

第四章 實驗與分析：

在本章節中，將針對不同的特徵擷取方法分別測試他們的分類成效，並觀察分類成效的結果與特徵擷取方法之間的關係。

第五章 實驗發現與未來展望：

針對第四章所得之研究結果做詳細的探討與解析，並針對實驗中發現之不足做一整理，供未來有興趣的研究者做參考之用。

第二章 文獻分析

第一節 特徵擷取（Feature Extraction）

由於目前的自然語言處理技術，在語意層次的理解上雖無很大的成果，但藉由一些語言學的知識，將文件內容分解成較小的語言單位，通常為文件的關鍵詞彙或稱「特徵辭彙」（曾元顯，2002）。將文件分解成較小的語意單位的動作則稱為特徵擷取或關鍵詞擷取（Keyword Extraction）。

目前所常用的特徵擷取方式大致上可分為：利用斷詞器來擷取特徵詞彙、N-Gram選詞法以及共現詞分析法（Co-occurrence Analysis）。

1. 利用斷詞器進行詞彙分析：

在眾多的中文斷詞器中，中央研究院所開發的CKIP中文斷詞系統為目前國內於資訊檢索領域及中文文件自動分類研究中經常被引用的中文斷詞器之一。此系統除藉由詞庫比對法外，亦針對詞庫中未存放的未知詞，透過統計式語料庫模型，(馬偉雲，2004)並輔助文法剖析規則來設計此中文斷詞系統，增加中文斷詞系統對於未知詞彙的辨識能力。

2. N-Gram 選詞法：

設定N個相連的字串作為特徵詞彙的特徵擷取方法。常用於擷取中文文件的特徵辭彙。由於中文詞彙以二字詞出現的頻率最多，因此許多的研究都以雙連字串（Bigram）作為關鍵辭彙擷取的單位（林政男，2004）。此種方法其優點在於可以藉由N-Gram(通常是Bigram)的組合來解決詞庫比對法或者文法剖析法無法擷取專有名詞的問題。但由於是利用N個相鄰字串作為擷取特徵詞的方式，因此有擷取出的特徵辭彙數量過於龐大的問題。系統不但需要更大的空間去儲存這些特徵辭彙，日後進行分類運算時的計算量也十分可觀。不過儘管有以上的缺點，大多數的研究顯示以N-Gram選詞法所擷取出的特徵辭彙對文件自動分類的效果要比利用斷詞器來擷取特徵辭彙來的好。如杜

海倫在1997年的研究利用財經紀事新聞語料的新聞標題作為實驗資料，其研究結果顯示利用雙連字串的效果較利用字典斷詞要來的好（杜海倫，1997）；而王稔志與張俊盛在2001年的研究以第二屆NTCIR資訊檢索比賽中文檢索方面的資料進行分類實驗，其分類結果證實利用雙連字串的分類效果要比斷詞處理要來的好（王稔志、張俊盛，2001）。

3. 共現詞分析法(Co-occurrence Analysis)：

兩個詞彙同時出現在同一份文件中稱為共現詞或共現語詞（Co-occurrence）。一般而言此種特徵擷取方式需要先利用斷詞方式或者N-Gram的方式擷取出單一語詞，再將文件中所擷取出的單一語詞做兩兩配對而成為共現詞。許雅芬於2002年的研究發現若以雙連字串作為特徵擷取的單位進而產生共現詞來作為文件的特徵來進行文件分類其效能會比只使用雙連字串要來的好（許雅芬，2002）。此種方法的缺點與N-Gram選詞法一樣，會擷取出大量的特徵詞彙。系統分類時需要更多的時間來進行計算。為了改善這樣的情況，Tseng在2001年的研究中將共現詞限制為合併同一個句子所出現的不同詞彙。這樣的做法可以減少文件中所擷取出共現詞的數量大大的減少計算詞彙關連性的時間，還可以擷取出一般辭典中尚未納入的特殊新聞詞彙（Tseng，2001）。而林政男於2004年的研究則認為利用斷詞處理所產生出的共現詞，相較於利用雙連字串所產生的共現詞更能代表文件的特徵。實驗中利用VSM、KNN、SVM三種分類器進行實驗。實驗結果發現利用KNN、SVM兩種分類方法時，利用雙連字串產生的共現詞在分類的精確度會比利用斷詞處理產生的共現詞要來的好。而利用VSM作為分類方法時結果則相反(林政男，2004)。

第二節　特徵權重計算

利用特徵擷取方法擷取出文章的特徵後，亦須剖析各特徵於該文件或所屬文件集合的重要性，並根據該特徵於所屬文件或文件集合的重要性賦予一權重值。由於目前電腦的技術仍無法讓電腦了解人類語言，因此為了要讓電腦了解文件或文件集合的組成，需利用「特徵-特徵權重」的搭配，讓電腦能夠根據詞彙權重的大小來辨識特徵在所屬文件或文件集合的重要性。以下便針對一些常見的特徵權重計算方式做一介紹。

1. 詞彙頻率（Term Frequency：TF）：

此種方式是最常用的幾種計算特徵權重的方法之一。計算權重的方式即為計算文件中各詞彙出現的頻率。當詞彙出現在文件中的頻率越高代表該詞彙在文件中的重要性越高。其公式如下：

$$W_{ij} = tf_{ij} \qquad \text{（公式 1）}$$

W_{ij}：代表文件 i 中詞彙 j 的權重

tf_{ij}：文件 i 中出現詞彙 j 的次數。

雖然詞彙頻率可以反應各詞彙在文件中的重要性。不過若只利用詞彙頻率來作為特徵選取的依據可能會有下列所述的缺點：

（1）當某些詞彙頻繁的出現於測試集中的文件中時，由於普遍出現於各個文件中，該詞彙對日後分類的影響力可能會比只出現在少數文件的詞彙要來的小。

（2）未考慮文件長度對於詞彙出現頻率的影響，當文件長度較長時，詞彙出現的頻率也越大。倘若測試集中所收錄的文件長度不一致的情況嚴重的話。則可能會影響日後分類處理的成效。

2. 文件頻率倒數（Inverse Document Frequency：IDF）：

1988年Salton等人認為除了詞彙出現在文件中的頻率外，詞彙出現的文件數量也應該是評估詞彙重要性的指標之一（Salton and Buckley，1988）。

其概念如下：若一詞彙出現在少數文件時，則該詞彙對文件的代表性則越大；相反的若是一詞彙普遍出現在各文件中，如：「我們」、「因為」這類的詞彙。由於大多數的文件都可能出現該詞彙，我們可以假設該詞彙對文件類別的代表性越低。文件頻率倒數的計算公式如下：

$$IDF_j = \log \frac{N}{df_j} \qquad \text{（公式 2）}$$

N：整個文件集合中的總文件數。

df_j：出現詞彙 j 的文件數量。

3. TF×IDF：

詞彙頻率可以反應該詞彙在單一文件的重要性，但若只考慮詞彙頻率來作為特徵選取的依據，則很可能會選出在單一文件中詞頻很高，也頻繁的出現在各文件中的詞彙，這樣的詞彙對分類的效用不大。而考慮文件頻率倒數，則可以彌補這樣的缺點。於是有研究者同時考慮這兩項因素利用詞彙頻率與文件頻率倒數的乘積作為特徵選取的依據(杜海倫，1997)。這種特徵權重計算方式在碰到單一文件中出現頻率高且普遍出現於各文件的詞彙時，可以有效的抑制該詞彙的權重。目前TF×IDF的特徵選取方法已廣泛的應用於資訊檢索（Information Retrieval）相關研究中。TF×IDF計算方式如下：

$$W_{i,j} = TF_{i,j} \times IDF_j \qquad \text{（公式 3）}$$

$TF_{i,j}$：關鍵字 j於文件 i中出現的次數。

IDF_j：關鍵字 j出現於文件集合中的文件數倒數。

$W_{i,j}$：關鍵字 j於文件 i中之重要性。

第三節 向量空間模型（Vector Space Model, VSM）：

向量空間模型是由Salton所提出，本是資訊檢索中利用關鍵字所產生的二維向量來表示查詢與文件，進而計算查詢與文件相似程度的方法。經過修改而用於文件自動分類的向量空間模型則是將關鍵字索產生的二維向量分別用來表示類別與測試文件(Salton，1988)。

進行分類時分為兩個階段：其一為訓練階段、其二為測試階段。在訓練階段中，首先系統須建立一「類別-特徵詞彙」的空間模型，其中每個類別會根據其所包含的特徵之權重建構成一個空間上的資料點。之後將各類別的特徵權重做正規化的動作以平衡每個類別的資料長度。正規化的計算方法如下：

若 $\overrightarrow{C_j} = (W_{1,j}, W_{2,j}, ..., W_{t,j})$，則正規化的公式為

$$\left\| \overrightarrow{C_j} \right\| = \sqrt{\sum_{i-1}^{t} W_{i,j}^2} \qquad \text{(公式 4)}$$

$$\overrightarrow{C_j'} = \left(\frac{W_{1,j}}{\left\| \overrightarrow{C_j} \right\|}, ..., \frac{W_{t,j}}{\left\| \overrightarrow{C_j} \right\|} \right) = (W_{1,j}', W_{2,j}', ..., W_{t,j}') \qquad \text{（公式 5）}$$

其中，$\overrightarrow{C_j'}$ 為正規化後的類別向量。當訓練階段建構起「類別-特徵詞彙」之空間模型後，即可進行測試文件的階段。

測試文件階段的步驟如下：當新文件要進行分類時，也需要進行前置作業將文件中的特徵詞取出並計算文件中個特徵詞之權重形成「文件-特徵詞彙」之空間模型。建立起待測文件之空間向量後，即可與測試階段所建立的「類別-特徵詞彙」模型進行相似度的比對，所找出相似度最高的類別即可決定新文件的類別為何。

在相似度計算方面一般常見的計算方式如表1：

Similarity Measure Sim(X,Y)	Evaluation for Binary Term Vectors	Evaluation for Weighted Term Vectors
Inner Product	$\lvert X \cap Y \rvert$	$\sum_{i=1}^{t} x_i \cdot y_i$
Dice Coefficient	$2 \cdot \frac{X \cap Y}{\lvert X \rvert + \lvert Y \rvert}$	$\frac{2\sum_{i=1}^{t} x_i \cdot y_i}{\sum_{i=1}^{t} x_i^2 + \sum_{i=1}^{t} y_i^2}$
Cosine Coefficient	$\frac{\lvert X \cap Y \rvert}{\lvert X \rvert^{1/2} \cdot \lvert Y \rvert^{1/2}}$	$\frac{\sum_{i=1}^{t} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{t} x_i^2 \cdot \sum_{i=1}^{t} y_i^2}}$
Jaccard Coefficient	$\frac{X \cap Y}{\lvert X \rvert + \lvert Y \rvert}$	$\frac{\sum_{i=1}^{t} x_i \cdot y_i}{\sum_{i=1}^{t} x_i^2 + \sum_{i=1}^{t} y_i^2 - \sum_{i=1}^{t} x_i \cdot y_i}$

表 1：相似度計算方式，資料來源：Salton, G., Automatic Text Processing, Addison Wesley, 1989, pp.318.

此種自動分類方法有著以下限制：

1. 不適合處理過長的文件。

   由於此種分類方法，須將查詢文件與被查詢文件轉換成「文件-特徵詞彙」與「類別-特徵詞彙」之向量空間模型。當文件的字數越長，所涵蓋的特徵詞彙越多，所佔的向量空間也會越大。當向量空間越大時，系統在進行相似度計算所耗費的時間便會越多。為此，利用向量空間模式進行文件自動分類時，多半會將資訊覆載量低或者較無意義的虛詞濾除掉。這樣的步驟稱為特徵詞彙刪減（Feature Reduction）。

2. 無法辨別語彙間之關聯性。

由於現在的電腦技術還沒達到可以判別語意的階段。因此在進行運算時，語彙字串必須相符才會被電腦判別為關聯；當出現語意上有關聯之語彙時卻會因為字串不相符而影響相似度計算的結果。在英文處理上，最簡單的就是將詞尾變化去除，處此之外還有潛在語意縮減法（Latent Semantic Indexing ，LSI）（Sullivan，2001），在國內亦有利用關聯規則技術用於尋找語彙間的關聯來提昇利用向量空間模式的文件群集計算（李維平、吳澤民、王美淳，2007）。

第四節 其他相關之研究

上述文獻分析部份，主要著重在一般文件自動分類之流程。以下便就近年來利用圖書書目資訊進行文件自動分類研究或本研究將運用之技術進行介紹。

1. 以圖書書目資料進行文件自動分類之相關研究：

過去數年，國內外有關文件自動分類的研究，已相當豐富。可能是因為實驗資料取得不易，國內外以圖書書目資料進行文件自動分類的研究則顯得相對的稀少。相關的研究有林昕潔(2005)所提出的『以SVM與詮釋資料設計書籍分類系統』。

該研究於博客來書店取得「偵探/懸疑小說」、「科幻/奇幻小說」、「愛情文藝小說」三個類別各900本圖書資料。圖書資料內所包含的資訊分別為描述部份（Description）與Meta-information部份；描述部份所包含的資料有：書名、關於書籍內容、作者與譯者的簡介而Meta-information所包含的資訊有：作者與出版社資訊。並以SVM分類器進行實驗。

實驗中，以描述部份作為主要分類對象，並利用專家挑選類別特徵並加入Meta-iformation輔助以提昇成效。實驗結果證實，利用專家挑選類別特徵並加入Meta-information確實能提昇分類成效。最後的實驗結果以F-measure的評估下皆有94%以上的成效。

一般來說，類別的數量會影響分類器的分類成效。該研究僅處理三個類別，與一般書店或圖書館所需處理的類別相差甚遠，因此得出的實驗數據是否能代表實際應用下的分類成效則值得商確。

2. 擴展概念運用於文件自動分類之相關研究：

在第一章時筆者曾提到，欲利用「擴展」之概念改善圖書題名之「題名長度不一」所可能造成自動分類成效不彰的情況。以下便就文件自動分類領域利

用「擴展」的概念之研究加以介紹。

目前「擴展」的概念已廣泛的運用於資訊檢索及其相關領域。如資訊檢索系統所常用的查訊擴展（Query Expansion）技術。

在文件自動分類領域中，亦有運用擴展概念的相關研究。不過針對文件或分類對象本身資訊量不足的情況，來提昇文件自動分類成效的研究，能見度並不高。近年來國內較為接近的研究有曾元顯，莊大衛（2003）所提出『文件自我擴展於自動分類之應用』。

該研究之目的在於解決因訓練文件不足而影響自動分類系統分類成效之情況。文中分別利用摘要擴展法與詞彙擴展法，從現有的訓練文件中擷取每篇文件的片段，組成新的文件。其目的在於增加訓練文件數量，以提昇分類成效。其實驗中顯示，研究中所提出的摘要擴展法與詞彙擴展法進行訓練文件的自我擴展，有助於分類成效的提昇，特別是在於訓練文件較少的情況下，改進成效的幅度越明顯。

第五節 文件自動分類常見的評估方式

在資訊檢索的領域中，召回率（R, recall）與精確率（P, precision）被視為評鑑資訊檢索系統的基本工具。所謂的召回率指的是所有符合需求的文件被系統檢索出的比例；而精確率所指的是所有被檢索出的文件與被檢索出而符合檢索需求之文件的比例。

這種評估方式經修正後亦可用來評估文件自動分類系統。以下則為利用於文件分類的精確率與召回率之公式：

	系統分為該類	系統不分為該類
屬於該類別	a	b
不屬於該類別	c	d

表2：文件數量分布表，資料來源：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， 第 68 期，2002。

$$\text{Recall}=\frac{a}{a+b} \qquad \text{（公式 6）}$$

$$\text{Precision}=\frac{a}{a+c} \qquad \text{（公式 7）}$$

此種方式運用在文件分類時其主要的問題在於精確率與召回率兩者呈反比，即若精確率高召回率勢必降低，反之亦然而無法兩者兼顧。至於何者較為重要，須視系統使用者的使用需求而定，並無定論。因此有人以精確率與召回率的乘積（Recall×Precision）作為系統成效的依據，當值越大時代表系統的成效越好。目前亦有同時參考精確率與召回率的評估指標：F1測試值，其公式如下頁公式8：

$$F_1 = \frac{P \times R \times 2}{P + R} \qquad \text{（公式 8）}$$

P：精確率(見上述公式8)

R：召回率(見上述公式7)

倘若同時有數個類別要一起考量的則有micro-average與macro-average兩種計算方式(曾元顯，2002)。其定義如下：

$$micro\,\mathrm{Pr}ecision = \frac{\sum_i a_i}{\sum_i a_i + \sum_i c_i} \qquad \text{（公式 9）}$$

$$micro\,\mathrm{Re}call = \frac{\sum_i a_i}{\sum_i a_i + \sum_i b_i} \qquad \text{（公式 10）}$$

$$macro\,\mathrm{Pr}ecision = \frac{1}{m}\sum_{i=1}^{m} \frac{a_i}{a_i + c_i} \qquad \text{（公式 11）}$$

$$macro\,\mathrm{Re}call = \frac{1}{m}\sum_{i=1}^{m} \frac{a_i}{a_i + b_i} \qquad \text{（公式 12）}$$

i ：指第 i 個類別。

m ：是類別的總數。

此外F1測試值也可以根據原公式利用micro-average與macro-average的定義帶入相關的精確率與召回率來計算出micro-F與macro-F值。我們從上述定義可以發現由於micro-average是將所有文件一起累加統計，並不區分類別，因而容易受到文件數量較多的類別所影響。而相反的macro-average則是考慮每個類別的成效再取平均，因此容易受到大量文件數較少的類別所影響。因此在利用micro-average與macro-average做評估系統的依據時，常會將兩種數值並列，以便比較(曾元顯，2002)。

第三章 研究方法

第一節 實驗資料

本研究實驗的素材取自國家圖書館館藏書目資料。其特性與篩選原則如下：

1. 採用中國圖書分類法所分類之書目資料。
2. 分類號前三碼（不包括特藏符號）為400至499以及800至899。
3. 所取材之書目資料限於中文書目資料。

	bid	class	book_name	author	publisher	t
	1	400	考工記導讀圖譯	聞人軍著	明文 初版民79	0
	2	400	進入汽電共生的世界	涂寬編著	全華 初版民84	0
	3	400	節約能源短文及繪畫比賽優勝作品專?...	經濟部能源委員會編	經濟部能源委員會 民79	0
	4	400	多元能源時代能源間競爭分析研究報?...	日本能源綜合推進委員會著	經濟部能源委員會 初版民79	0
	5	400	能源產業之轉機：日、美、歐各國比?...	柴田益男著	經濟部能源委員會 民78	1

圖 1：實驗原始資料（部份）

實驗原始資料如上圖1所示，其中bid為系統號、class代表該書所屬之分類號、book_name代表圖書題名、author代表作者欄位、publisher代表出版項，而t則代表該書是否為測試文件，t欄位為1者則為測試文件。

考慮資料的代表性對於所蒐集的書目資料中，少於25篇書目類別的資料予以刪除。再將各類別取其20%篇的書目為測試文件，其餘80%則為訓練文件。經刪減實驗資料後。具測試價值的類別數為156類；總訓練文件為19949筆書目；總測試文件為4911筆。由於類別數目較多，因此各類別做為訓練與測試文件的數目，可參見附錄A。

第二節 實驗設計

在實驗的設計上，針對圖書題名的擷取採用下列策略：1.以不同特徵擷取方法擷取圖書題名之特徵，進行圖書自動分類；2.導入擴展文件特徵策略，進行圖書自動分類；3.只利用圖書題名與導入擴展文件特徵策略，分別導入作者資訊進行圖書自動分類。

即時(vi),0.00136912295131455
臺北縣(n),0.00106194466053807
前來(vt),0.000299646190550155
證照(n),0.000517231996689708
依據(n),0.0021089977189778
守法(nv),0.00176750853377106
灰燼(n),0.000950863064526688
冷卻水(n),0.0023042470682173
緩和期(n),0.00176750853377106
賽車(n),0.00184506708367384
停車(vi),0.000872784153282355
章句(n),0.000793062219298086
九十二年度(n),0.00750631638272751

圖 2：類別與特徵向量（分類號為400）

論文中採用向量空間模式（VSM，Vector Space Mode）做為實驗的文件自動分類方法。其中「類別-特徵」向量（參見圖2）與「文件-特徵」向量的特徵權重計算採用TF×IDF法計算（參見第二章之公式3），並在取得特徵權重後進行正規化（參見第二章之公式4、公式5）以調整各類別的「類別-特徵」向量或各文件之「文件-特徵」向量的維度。在「類別-特徵」向量和「文件-特徵」向量的相似度計算方面，則採用內積法（參見第二章之表1）進行相似度運算。詳細的實驗設計如下頁表3：

實驗階段	各階段實驗目的
以不同特徵擷取方法擷取圖書題名之特徵。	只利用圖書題名時，不同特徵擷取方法對分類成效之影響
導入擴展文件特徵策略。	利用Google搜尋引擎擴展文件特徵對分類成效之影響
並加入作者資訊進行圖書自動分類。	比較只利用圖書題名以及利用Google搜尋引擎擴充文件特徵於導入作者欄位後的分類成效

表 3：各實驗階段及其實驗目的

1. 以不同特徵擷取方法擷取圖書題名之特徵 ：

首先，我們將利用斷詞器、雙連字串（Bigram）以及上述兩種方法分別搭配共現詞分析法（Co-occurrence Analysis）進行圖書題名的特徵擷取。共設計了十組實驗，各特徵擷取方法及實驗代號如下表4所示：

實驗代號	特徵擷取方法
seg	斷詞處理不含詞性標記。
seg_nv	斷詞處理後只留下具有動詞及名詞屬性的詞彙且不含詞性標記。
seg_with_word_class	斷詞處理包含詞性標記。
seg_nv_wc	斷詞處理後只留下具有動詞及名詞屬性的詞彙並包含詞性標記。
bigram	以雙連字串擷取特徵詞彙。
co_seg	斷詞處理不含詞性標記，並進行共現詞分析。
co_seg_nv	斷詞處理後只留下具有動詞及名詞屬性的詞彙且不含詞性標記。 並進行共現詞分析。
co_seg_with_word_class	斷詞處理包含詞性標記，並進行共現詞分析。
co_seg_nv_wc	斷詞處理後只留下具有動詞及名詞屬性的詞彙且包含詞性標記，並 進行共現詞分析。
co_bigram	以雙連字串擷取特徵詞彙，並進行共現詞分析。

表 4：實驗代號與其所代表之特徵擷取方法

在這個階段中我們採用中央研究院所研發的CKIP中文斷詞系統做為實驗的斷詞器。至於雙連字串的特徵擷取以及共現詞分析部份，則是利用Perl語言自

行開發處理程式。對於共現詞的分析部份，我們將共現詞分析的範圍限定為一筆圖書題名內，來進行共現詞彙分析。

2. 導入擴展文件特徵策略，進行圖書自動分類：

研究中設計了一擴展文件特徵之特徵擷取方法，在前一階段的實驗設計中我們發現僅就圖書題名進行相關分析在特徵的取樣上確實不足，針對此部分我們利用圖書題名做為搜尋引擎的查詢語句，搜尋引擎會將與圖書題名相關的查詢結果做為回饋(Feedback)，因此我們將擷取搜尋引擎所回饋的網頁標題以及搜尋結果描述部份來擴展文件特徵（參見圖3所示）。

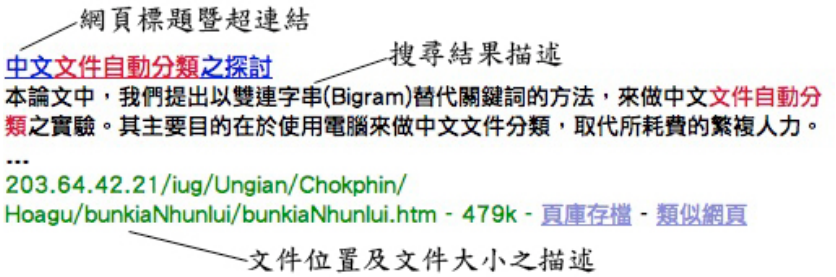


圖 3： Google搜尋引擎之搜尋結果截圖

在實驗中我們藉由Google搜尋引擎來擴展文件特徵，實驗步驟如下：

(1) 首先利用圖書題名做為查詢語句，藉由Google搜尋引擎進行查詢，擷取搜尋引擎所回饋的搜尋結果網頁(如圖3)。

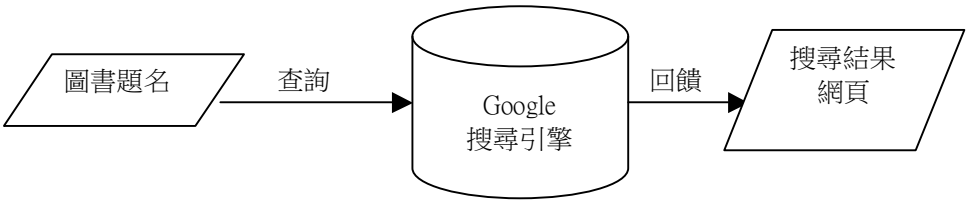


圖 4：利用Google搜尋引擎進行擴展文件特徵示意圖—步驟一

(2) 為了避免較不相關的查詢結果被納入擴展範圍，實驗只採搜尋引擎回饋結果中關聯性較高(第一頁)的網頁標題以及搜尋結果描述部份進行圖書題名擴展特徵的處理，示意圖如下圖5。

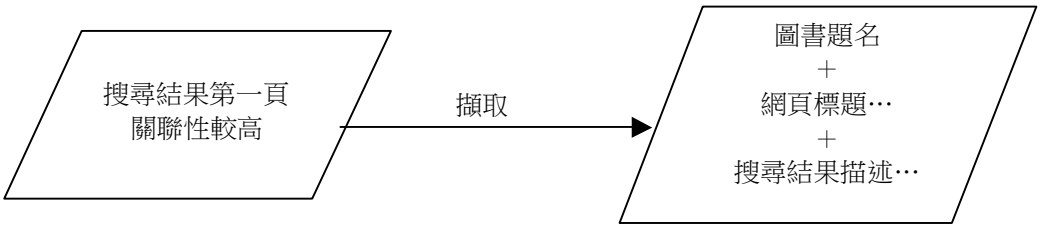


圖 5： 利用Google搜尋引擎進行擴展文件特徵示意圖—步驟二

(3) 將圖書題名與其利用搜尋引擎擴展文件特徵的部份，當作一訓練文件或測試文件，送入分類器進行分類器訓練或文件分類（見圖6）。

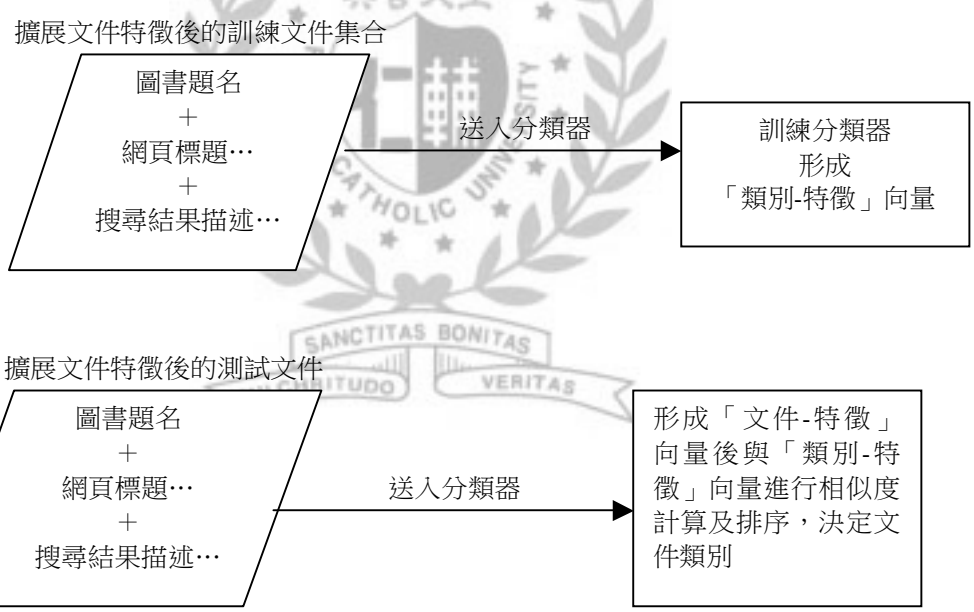


圖 6： 利用Google搜尋引擎進行擴展文件特徵示意圖—步驟三

由於利用Google搜尋引擎擴充文件特徵，勢必會造成文件特徵的增加，為了避免特徵的過度增加，影響自動分類時系統的運算時間。我們將參考第一階段的實驗結果，選擇適當的特徵擷取方法， 並考慮進行詞彙過濾。

上述擴展文件特徵策略的分類成效將與第一階段（只利用圖書題名時，不同特徵擷取方法對分類成效之影響）成效最好的組別比較。並藉此觀察此擴展文件特徵策略，是否能增進以圖書書目為分類素材的分類成效。

3. 只利用圖書題名與導入擴展文件特徵策略，分別導入作者資訊進行圖書自動分類 ：

透過觀察書目，我們發現到大多數的圖書作者，其著作只出現在少數類別或單一類別。因為這樣的特性我們推測，雖然作者欄位中所蘊含的作者資訊，以特徵數量的角度來看，可能比圖書題名要來的少，但作者欄位中所負載的資訊是可以有效提昇以圖書書目進行自動分類的成效。此外加入作者欄位的因素在 林昕潔（2005）的研究中以獲得證實確實能提高圖書分類之成效。

圖 7：作者姓名部份訓練後所產生的類別-特徵向量class400

因此在這個階段，我們將觀察只利用圖書題名以及利用Google搜尋引擎進行擴展文件特徵策略，兩者分別導入作者欄位資訊後的分類成效。因此將以第一階段（只利用圖書題名時，不同特徵擷取方法對分類成效之影響）成效最好的組別和利用Google搜尋引擎擴展文件特徵策略，各別導入作者欄位的因素進行分類實驗，作者欄位的資訊再特徵擷取的方式上採用Bi-gram方式取詞亦利用TF×IDF法賦予特徵權重，圖7表作者資訊部份之「類別-特徵向量」。導入作者因素的相似度計算公式如下公式13、公式14：

$$S(C_i,Book_j) = S(C_i^{\{BN\}},Book_j^{\{BN\}}) + \alpha \times S(C_i^{\{AN\}} + Book_j^{\{AN\}}) \quad \textbf{(公式 13)}$$

$$S(C_i,Book_j) = S(C_i^{\{BNex\}},Book_j^{\{BNex\}}) + \alpha \times S(C_i^{\{AN\}} + Book_j^{\{AN\}}) \quad \textbf{(公式 14)}$$

$S(C_i,Book_j)$：類別i與書籍j的相似度。

$S(C_i^{\{BN\}},Book_j^{\{BN\}})$： 類別i所形成之「類別-圖書題名特徵」空間向量與由書籍j的圖書題名所產生「書籍-圖書題名特徵」空間向量的相似度。

$S(C_i^{\{BNex\}},Book_j^{\{BNex\}})$：經過擴展文件特徵處理後類別i所形成的「類別-圖書題名文件特徵」空間向量與書籍j的圖書題名經過擴展文件特徵處理後所產生的「書籍-圖書題名文件特徵」之相似度。

$S(C_i^{\{AN\}} + Book_j^{\{AN\}})$：類別i所形成的「類別-作者特徵」空間向量與由書籍j的作者名稱所產生的「書籍-作者名稱特徵」空間向量之相似度。

α： $S(C_i^{\{AN\}} + Book_j^{\{AN\}})$之權重；用以調整作者部份相似度計算的權重。以經驗法則及實驗中逐步調整，以達到最佳的實驗成效。其中：$S(C_i^{\{BN\}},Book_j^{\{BN\}})$、$S(C_i^{\{BNex\}},Book_j^{\{BNex\}})$、$S(C_i^{\{AN\}} + Book_j^{\{AN\}})$的計算公式如下：

$$S(C_i^{\{BN\}},Book_j^{\{BN\}}) = \sum_{x=1}^{t} C_{ix}^{\{BN\}} \cdot Book_{jx}^{\{BN\}} \quad \text{(公式 15)}$$

$$S(C_i^{\{BNex\}},Book_j^{\{BNex\}}) = \sum_{x=1}^{t} C_{ix}^{\{BNex\}} \cdot Book_{jx}^{\{BNex\}} \quad \text{(公式 16)}$$

$$S(C_i^{\{AN\}} + Book_j^{\{AN\}}) = \sum_{x=1}^{t} C_{ix}^{\{AN\}} \cdot Book_{jx}^{\{AN\}} \quad \text{(公式 17)}$$

$C_{ix}^{\{BN\}}$：類別i所形成之「類別-圖書題名特徵」空間向量。其中圖書題名特徵是將類別i所包含的圖書題名，以適當的特徵擷取方法，取得圖書題名特徵。

$C_{ix}^{\{BN\}}$為類別i所形成之「類別-圖書題名特徵」空間向量，其index為x的特徵權重。

$C_{ix}^{\{BNex\}}$：類別i所形成之「類別-圖書題名文件特徵」空間向量。其中圖書題名文件特徵是將類別i中所包含的圖書題名經過擴展文件特徵策略處理後，以適當特徵擷取方法取得。$C_{ix}^{\{BNex\}}$為類別i所形成之「類別-圖書題名文件特徵」空間向量，其index為x的特徵權重。

$C_{ix}^{\{AN\}}$：類別i所形成的「類別-作者特徵」空間向量。其中作者特徵，是將類別i中所包含的作者名稱以雙連字串進行特徵擷取而得。$C_{ix}^{\{AN\}}$為類別i所形成的「類別-作者特徵」空間向量，其index為x的特徵權重。

$Book_{jx}^{\{BN\}}$：書籍j的圖書題名以適當之特徵擷取方法取得其特徵所形成的「書籍-圖書題名特徵」空間向量。$Book_{jx}^{\{BN\}}$為書籍j所形成「書籍-圖書題名特徵」空間向量，其index為x的特徵權重。

$Book_{jx}^{\{BNex\}}$：書籍j的圖書題名經過擴展文件特徵策略處理後，以適當的特徵擷取方法取得特徵而形成的「書籍-圖書題名文件特徵」空間向量。

$Book_{jx}^{\{BNex\}}$為書籍j所形成的「書籍-圖書題名文件特徵」空間向量，其index為x的特徵權重。

$Book_{jx}^{\{AN\}}$：書籍j的作者名稱，經適當的特徵擷取方法，取得作者名稱之特徵而形成的「書籍-作者名稱特徵」的空間向量。$Book_{jx}^{\{AN\}}$為書籍j所形成的「書籍-作者名稱特徵」的空間向量，其index為x的特徵權重。

t：$C_{ix}^{\{BN\}}$ 或 $C_{ix}^{\{BNex\}}$ 或 $C_{ix}^{\{AN\}}$ 的向量空間大小。

其中 $C_{ix}^{\{BN\}}$、$C_{ix}^{\{BNex\}}$、$C_{ix}^{\{AN\}}$、$Book_{jx}^{\{BN\}}$、$Book_{jx}^{\{BNex\}}$、$Book_{jx}^{\{AN\}}$ 之特徵權重皆以TF×IDF法計算(參見第二章之公式3)，並對上述各空間向量進行正規化(參見第二章之公式4、公式5)。

第三節 成效評估

論文所取得的實驗資料，若以訓練文件總數除以類別數，則各類別平均訓練文件數量約為128筆(四捨五入)。以訓練文件的平均數量作為分界，類別中訓練文件大於128者為大類別（訓練文件較多的類別），反之為小類別。其中實驗資料大類別與小類別的比例約為41:57。由於大類別與小類別的比例相近，而每種評估方式都有其強調的對象，只看單一的評估方法難以通盤的了解本研究所提出的分類方法對大、小類別在分類成效上的影響。於第二章得文獻分析中得知，正確率代表測試文件被正確分類的比例而 Micro-F與Macro-F可了解大多數測試文件的分類成效（Micro-F）以及測試文件較少之類別的分類成效（Macro-F）。因此本實驗採用正確率(Accuracy)、Micro-F以及Macro-F三種評估方式，同時呈現分類成效。三種評估方式的計算方法如下：

$$Accuracy = \frac{Test^{\{acc\}}}{Test} \quad \text{(公式 18)}$$

$Test^{\{acc\}}$：分類正確的測試文件數

$Test$：所有測試文件數

$$Micro-F = \frac{2 \times \sum_{i=1}^{c} TP_i}{2 \times \sum_{i=1}^{c} TP_i + \sum_{i=1}^{c} FP_i + \sum_{i=1}^{c} FN_i} \quad \text{(公式 19)}$$

$$Macro-F = \frac{1}{C} \sum_{i=1}^{c} \frac{2 \times TP_i}{2 \times TP_i + FP_i + FN_i} \quad \text{(公式 20)}$$

C：類別總數。

i：某一類別。

TP_i：測試文件為類別i並且正確的被分入類別i的篇數。

FP_i：測試文件不是i類卻被分入類別i的篇數。

FN_i：是類別i卻沒有被分入i類的篇數。

第四章 實驗與分析

如第三章實驗設計所述，本研究所進行的實驗共分三階段為：1. 以不同特徵擷取方法擷取圖書題名之特徵，進行圖書自動分類 ；2. 導入擴展文件特徵策略，進行圖書自動分類 ；3. 只利用圖書題名與導入擴展文件特徵策略，分別導入作者資訊進行圖書自動分類。以下便以各階段的實驗結果進行說明及分析。

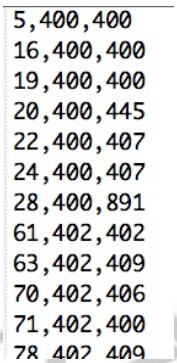


圖 8： 實驗分類結果的截圖，分三個欄位第一個欄位表系統號，第二個欄位表原始分類號，第三個欄位表系統所給予之分類號

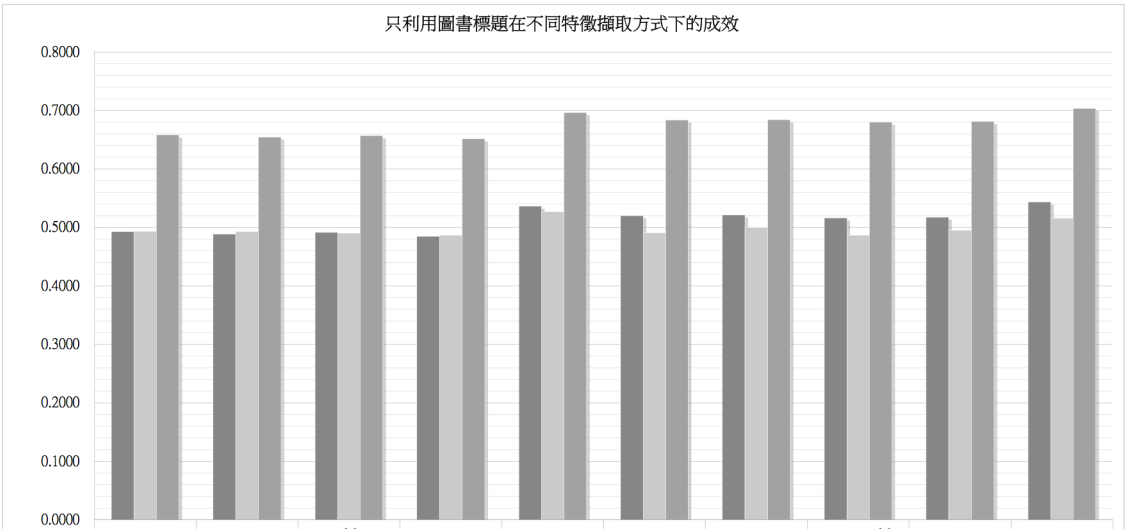
第一節 以不同特徵擷取方法擷取圖書題名之特徵

在這個階段我們利用了10種特徵擷取方法，進行實驗，其中分類方法採用向量空間模式(VSM，Vector Space Mode)，而其中的特徵權重則採用TF×IDF法，其各特徵擷取方法的實驗代號如下表5：

實驗代號	特徵擷取方法
seg	斷詞處理不含詞性標記。
seg_nv	斷詞處理後只留下具有動詞及名詞屬性的詞彙且不含詞性標記。
seg_with_word_class	斷詞處理包含詞性標記。
seg_nv_wc	斷詞處理後只留下具有動詞及名詞屬性的詞彙並包含詞性標記。
bigram	以雙連字串擷取特徵詞彙。
co_seg	斷詞處理不含詞性標記，並進行共現詞分析。
co_seg_nv	斷詞處理後只留下具有動詞及名詞屬性的詞彙且不含詞性標記。並進行共現詞分析。
co_seg_with_word_class	斷詞處理包含詞性標記，並進行共現詞分析。
co_seg_nv_wc	斷詞處理後只留下具有動詞及名詞屬性的詞彙且包含詞性標記，並進行共現詞分析。
co_bigram	以雙連字串擷取特徵詞彙，並進行共現詞分析。

表 5：實驗代號與其所代表之特徵擷取方法

實驗結果如下頁圖表1：



	seg	seg_nv	seg_with_word_class	seg_nv_wc	bigram	co_seg	co_seg_nv	co_seg_with_word_class	co_seg_nv_wc	co_bigram
Accuracy	0.4926	0.4885	0.4913	0.4850	0.5361	0.5203	0.5213	0.5162	0.5174	0.5439
Macro-F	0.4932	0.4927	0.4901	0.4868	0.5264	0.4908	0.4995	0.4867	0.4952	0.5160
Micro-F	0.6583	0.6546	0.6571	0.6514	0.6962	0.6839	0.6844	0.6802	0.6810	0.7035

圖表 1：只利用圖書標題在不同特徵擷取方式下的成效

由上表可知，在只利用圖書題名進行自動分類時，在特徵擷取方法上，以雙連字串擷取文件特徵並進行共現詞分析這組實驗（co_bigram）所得到的正確率（Accuracy）數值最高為0.5439，其次為只利用雙連字串擷取文件特徵（bigram）其正確率為0.5361。

就分類結果而言，以雙連字串擷取文件特徵並進行共現詞分析這組實驗（co_bigram）的確優於其他特徵擷取方法。但由於向量空間模式的計算上，當特徵數量越多時所需要的計算時間也就越多（訓練文件+實際分類），除此之外電腦儲存這些自訓練文件所擷取的文件特徵所需要的空間也會隨之增加。

因此觀察了各特徵擷取方法擷取訓練文件後所統計出來的特徵數量以及特徵種類數量後(圖9、圖10)，我們不難發現 以雙連字串擷取文件特徵並進行共現詞分析這組實驗（co_bigram）所擷取出的特徵數量以及特徵種類數量遠大於其他的特徵擷取方法。但在整體成效卻只比利用雙連字串擷取文件特徵（bigram）這組高出0.0078。因此若將電腦所需耗費的時間空間納入考量，利用雙連字串擷取文件特徵（bigram）這組特徵擷取策略或許是較好的選擇。

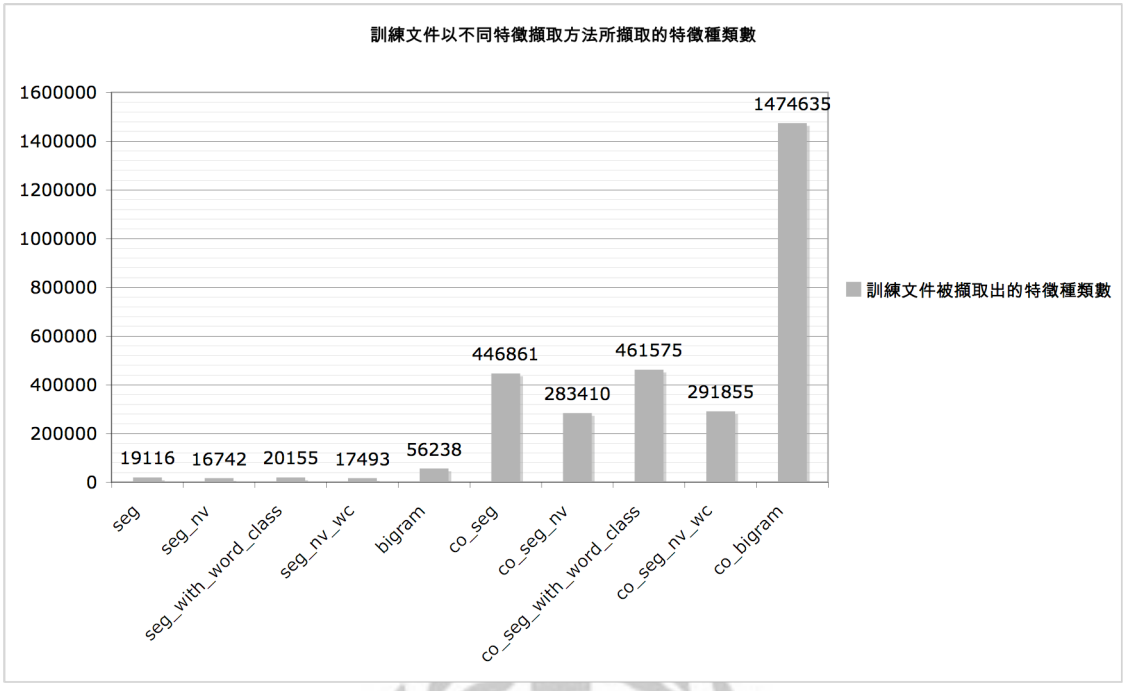


圖 9：訓練文件以不同特徵擷取方式所擷取的特徵種類數量

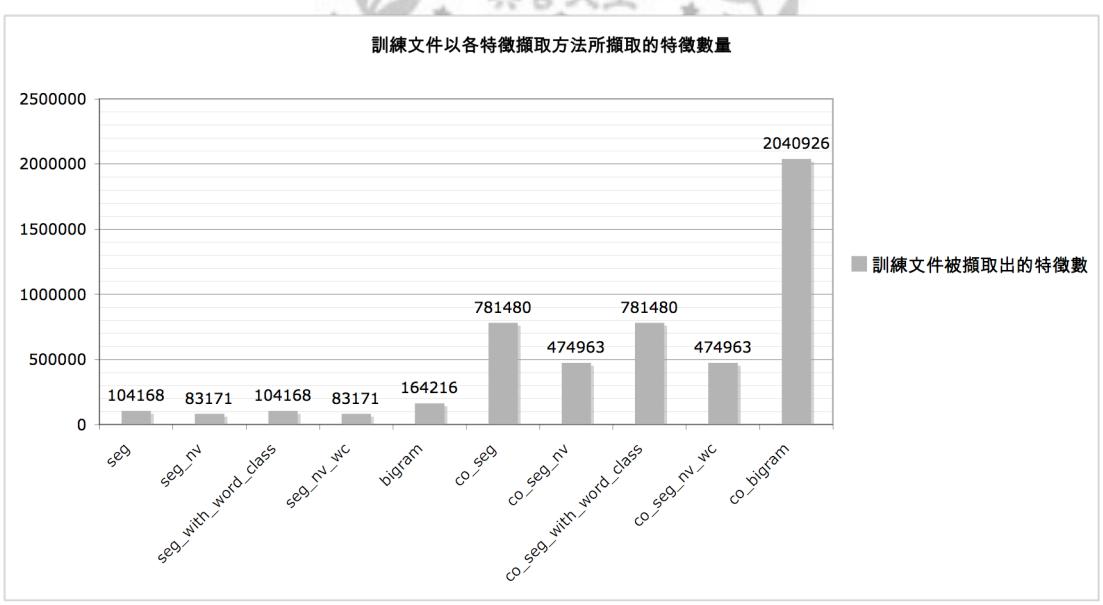


圖 10：訓練文件以不同特徵擷取方式所擷取的特徵數量

第二節 導入擴展文件特徵策略

從前階段實驗中，發現圖書題名可擷取的特徵相當有限，所以如何從短短幾個字的題名去延伸擴展，取得與該題名相關的特徵，對於提昇圖書書目自動分類的成效是相當重要的關鍵之一。

第二階段的實驗所要探討的是本研究所設計之擴展文件特徵策略，能否有效的提升圖書書目自動分類的成效。

在本文中我們嘗試的擴展文件特徵策略，是將圖書題名做為查詢語句，藉由網際網路上所提供的搜尋引擎進行查詢，搜尋引擎所回饋的條項內容中具有分類價值的部分可分為網頁標題及搜尋結果描述部分。我們將擷取搜尋引擎回饋之查詢結果中關連性較高的條項（實驗中取回饋的第一頁）中網頁標題及搜尋結果描述部分與圖書題名結合成一新文件，再對此一新文件進行特徵擷取的步驟，爾後進行之後的自動分類作業。

由於圖書題名經過上述擴展文件特徵處理後，文件特徵便隨之增加。倘若如第一階段選擇以雙連字串擷取文件特徵並進行共現詞分析這個特徵選取方法。勢必擷取出大量的文件特徵，使得向量空間的維度大幅增加，讓系統需要花更多的時間來進行計算。因此，在使用擴展文件特徵策略的特徵擷取部分，本實驗參考了第一階段各特徵擷取方法的分類成效，選擇適合的特徵擷取方法來擷取經擴展文件特徵處理後的文件。

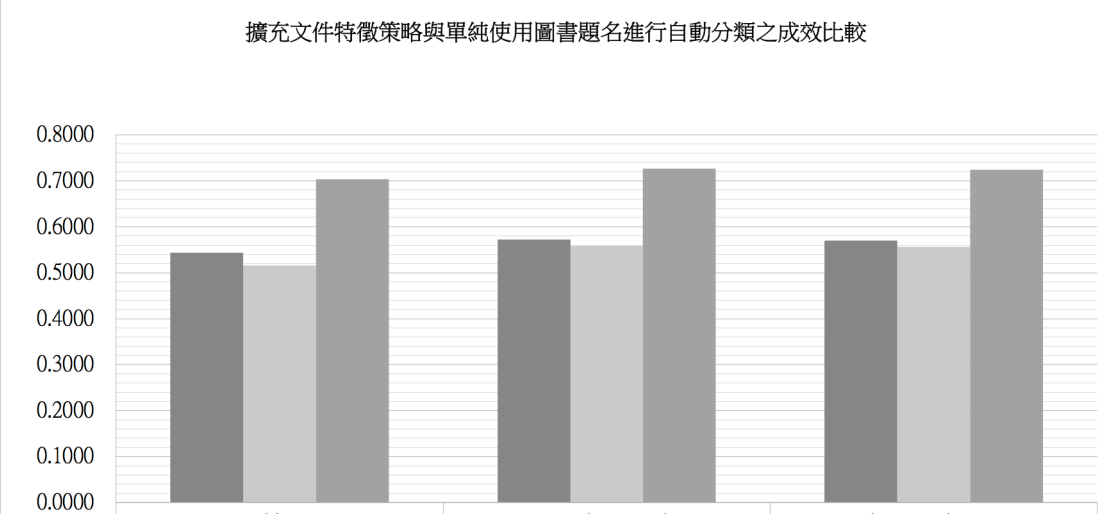
經過第一階段的實驗觀察，筆者選擇了斷詞處理後只留下具有動詞及名詞屬性的詞彙且不含詞性標記(seg_nv)。以及斷詞處理後只留下具有動詞及名詞屬性的詞彙並包含詞性標記(seg_nv_wc)。這兩種方法分別做為擴展文件特徵處理後所形成之新文件的特徵擷取方法。在擴展文件特徵策略中選擇以斷詞處理後只留下具有動詞及名詞屬性的詞彙且不含詞性標記(seg_nv)其實驗代號為「google_extend_nv_unwc」；另外採用斷詞處理後只留下具有動詞及名詞屬性的

詞彙並包含詞性標記(seg_nv_wc)做為擴展文件特徵策略之特徵擷取方法，其實驗代號為「google_extend」。

之所以選擇這兩種特徵擷取方法，主要原因為這兩種特徵選取方法，所擷取出的特徵數量最少。一般認為動詞、名詞兩種詞性與其他詞性相比，其資訊負載量比其他詞性要來的大。雖然在第一階段的實驗中這兩種特徵擷取方法的實驗組別的成效是最差的。但在這兩種特徵擷取方法進行共現詞分析後所得到的分類成效卻只略遜於以雙連字串擷取特徵並進行共現詞分析(co_bigram)這組實驗。觀察了各特徵擷取方法擷取的特徵詞彙種數以及數量，我們推測可能是因為「seg_nv」與「seg_nv_wc」所擷取出的特徵種類數量以及特徵數量不足以讓分類器辨別156的類別，而造成分類成效的低落。

由於經過擴展文件特徵處理後所產生的文件，在文件長度上要比單純使用圖書題名進行自動分類時要大的多，因此採用特徵擷取數量較少且所擷取出之特徵的資訊負載量較大的特徵擷取方法，可能較適合做為擴展文件特徵策略的特徵擷取方法。此外由於在第一階段的實驗中詞性標記對分類成效的影響不大，因此在擴充文件特徵策略的特徵擷取方法上我們分別選取了「seg_nv」(無詞性標記)、「seg_nv_wc」(有詞性標記)，分別進行自動分類的實驗。

此外為了要與未使用擴展文件特徵策略的自動分類成效進行比較。在此筆者選擇了第一階段分類成效最好的組別(co_bigram)做為對照組與使用擴展文件特徵策略後的分類成效進行比較。本階段之各組別的分類成效如下頁圖表2所示。



	co_bigram	google_extend	google_extend_nv_unwc
Accuracy	0.5439	0.5716	0.5693
Macro-F	0.5160	0.5586	0.5564
Micro-F	0.7035	0.7260	0.7242

圖表 2：擴展文件特徵策略與單純使用圖書題名進行自動分類之成效比較

由上圖表2我們可以發現，使用擴展文件特徵策略的兩個實驗組在各評估方法上的成績，均比只使用圖書題名進行自動分類的成效要來的好。特別是「google_extend」這組表現得最好，在各項評估指數均高於其他兩組實驗組別。

在這個階段的實驗結果中，沒有詞性標記的組別(google_extend_nv_unwc)其分類成效略低於有詞性標記的組別(google_extend)，這個部份正巧與第一階段的實驗結果相反。單看利用斷詞器擷取特徵且沒有做共現詞分析的組別來看，沒有詞性標記的組別(seg,seg_nv)在分類的成效上均比有詞性標記的組別（seg_with_word_class,seg_nv_wc）要來的高。由於第二階段兩組利用擴展文件特徵策略的組別，所擷取出的特徵種類與數量，皆高於第一階段利用斷詞器擷取且沒有做共現詞分析的組別（seg,seg_nv, seg_with_word_class,seg_nv_wc）要多。因此筆者推測在文件長度大於某臨界值時，利用斷詞器擷取特徵並附有詞性標記，才能有助於自動分類的成效。而這個文件長度的臨界值則可能會受如類別數量、等等的因素而改變。但整體來說，文件長度越大，使用利用斷詞器擷取特徵並附有詞性標記，才能有助於自動分類的成效，而文件長度越短（如書名），利用斷詞器擷取特徵並附有詞性標記，則可能降低自動分類的成效。

由於google_extend與google_extend_nv_unwc兩組組別在正確率的指標上僅相差0.0023，差距可說是相當的小。因此這兩組經過擴展文件特徵策略處理的組別，將一起進入下個階段進行加入作者欄位因素後的實驗。

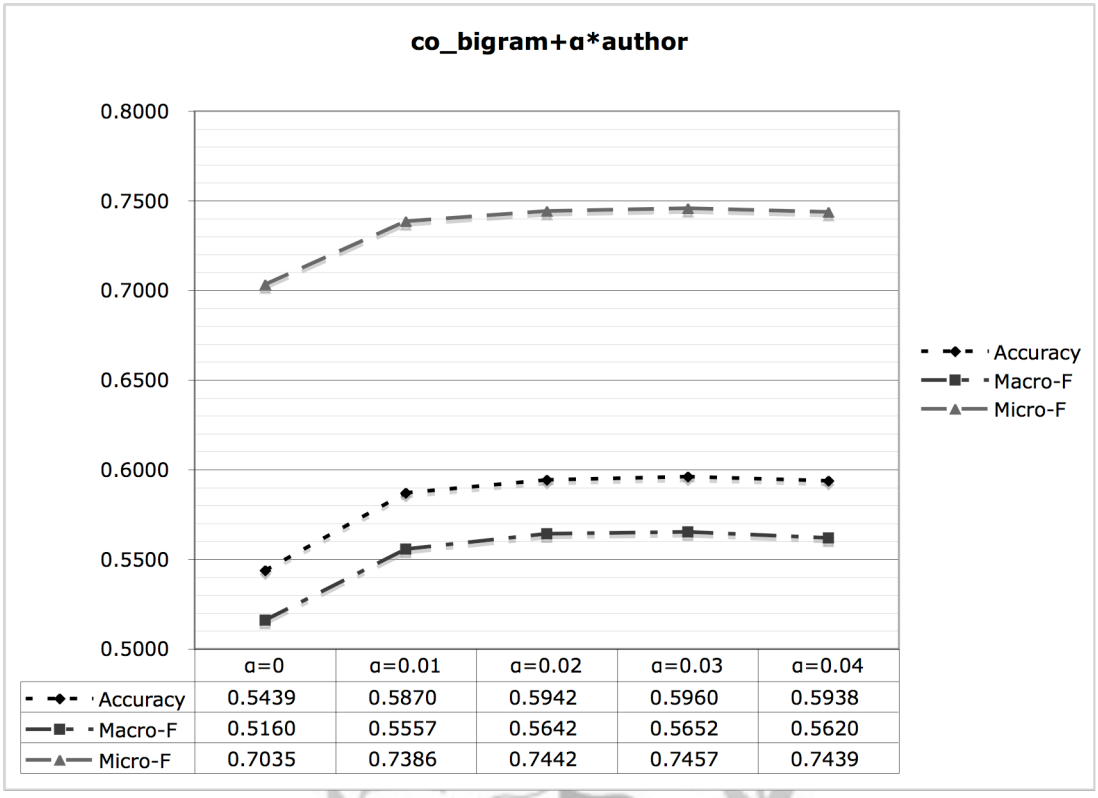
第三節 只利用圖書題名與導入擴展文件特徵策略，分別導入作者資訊

在第一階段的實驗中發現，只利用圖書題名進行文件自動分類時，一些書名較短的書籍常因為其特徵不足而導致系統分類錯誤。因此我們在構思第二階段實驗之擴展文件特徵方法外，亦觀察了書目中其他欄位資訊，其中我們發現了作者資訊欄位與類別具有相當程度的關聯性。此外國內已有研究證實，加入作者欄位資訊能有效的提昇圖書自動分類的成效（林昕潔，2005）。由於第二階段之擴展文件特徵方法是利用搜尋引擎進行檢索，擷取其回饋之內容作為書名之延展內容。這樣的方法難免會擴展出與圖書內容較不相關的特徵，進而影響分類成效。因此希望能藉由加入作者資訊欄位的因素來減輕上述情況所帶來的影響。

這個階段的實驗，我們將在上兩個階段的實驗中取得成效最好的組別再加入作者欄位的因素來進行自動分類實驗。因此我們選擇了第一階段中成效最好的實驗代號為co_bigram組；由於第二階段導入擴展文件特徵策略的兩個實驗組別成效相當接近，因此這兩個組別也被納入這個階段的實驗。

在計算公式上第三章已有說明，在此便就未說明的部分進行解說。作者欄位與類別之相似度計算這個部分，考量到中央研究院所研發的CKIP中文斷詞系統並未對人名進行最佳化，實際運用上常無法正確的取出人名。因此在權衡之下，本實驗在作者欄位與類別之相似度計算上，其特徵擷取方式採用雙連字串擷取作者欄位的特徵。

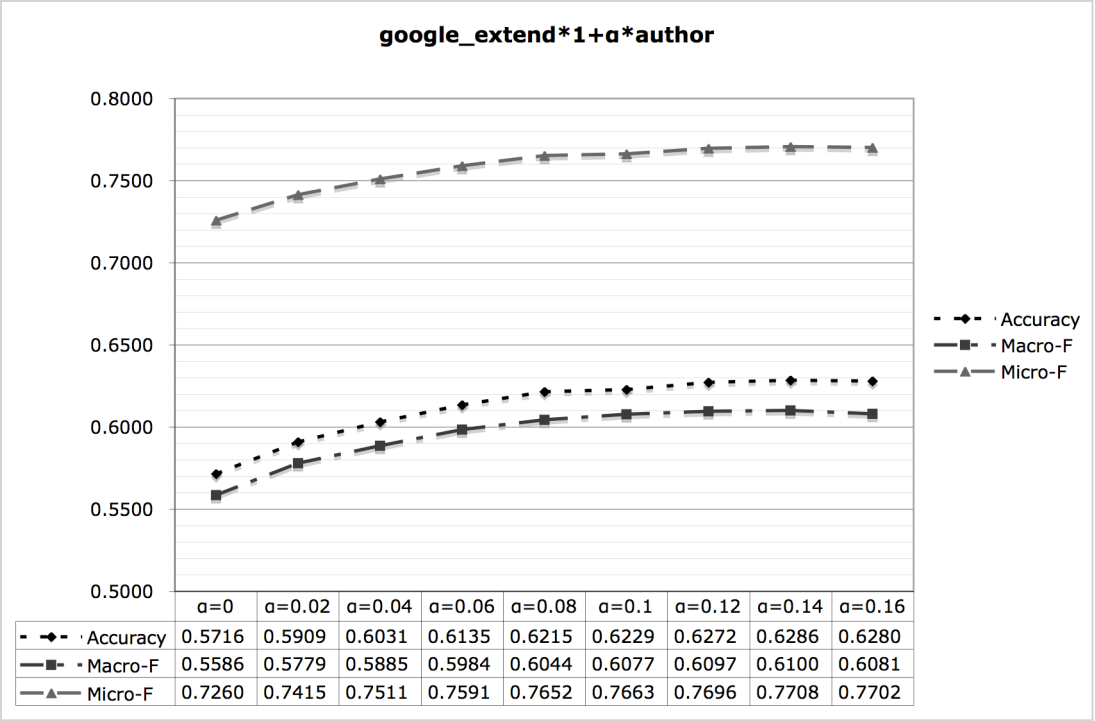
在第三章的公式13與公式14中表作者欄位與類別的相似度所占權重的α值的取法，是先觀察「類別-作者欄位特徵」向量中各作者欄位特徵的權重，根據經驗法則先以大略的數值進行測試，並根據實驗的結果增加或減少α的值，直到該組實驗的分類成效成長趨緩或衰退為止。以下為各實驗組分別加入作者欄位與類別相似度在不同權重下的分類成效。



	α=0	α=0.01	α=0.02	α=0.03	α=0.04
Accuracy	0.5439	0.5870	0.5942	0.5960	0.5938
Macro-F	0.5160	0.5557	0.5642	0.5652	0.5620
Micro-F	0.7035	0.7386	0.7442	0.7457	0.7439

圖表 3：利用雙連字串擷取圖書題名特徵並做共現詞分析加入作者欄位後，作者欄位於不同權重下的成效比較。

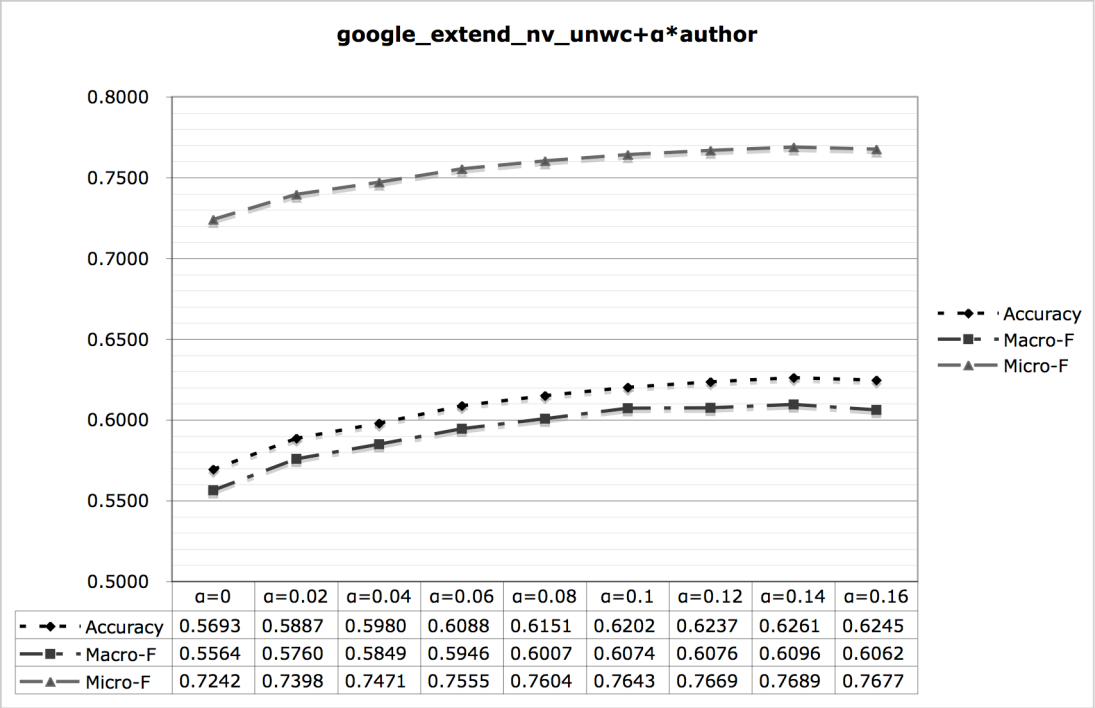
上圖表3為co_bigram實驗組加入作者欄位元素後於不同作者欄位元素之權重下的分類成效，權重以0.01增加，α為0時表示沒有加入作者欄位元素。由上圖可以觀察到，α在0與0.01這個區間，其分類成效的成長最為明顯，在0.01至0.03之間，分類成效呈現緩慢成長，α為0.04時各分類成效評估指數開始下降。因此在這co_bigram加入作者欄位元素的實驗中，當α=0.03時成效最佳。



	α=0	α=0.02	α=0.04	α=0.06	α=0.08	α=0.1	α=0.12	α=0.14	α=0.16
Accuracy	0.5716	0.5909	0.6031	0.6135	0.6215	0.6229	0.6272	0.6286	0.6280
Macro-F	0.5586	0.5779	0.5885	0.5984	0.6044	0.6077	0.6097	0.6100	0.6081
Micro-F	0.7260	0.7415	0.7511	0.7591	0.7652	0.7663	0.7696	0.7708	0.7702

圖表 4：圖書題名利用擴展文件特徵（含詞性標記）並加入作者欄位。作者欄位於不同權重下的成效比較。

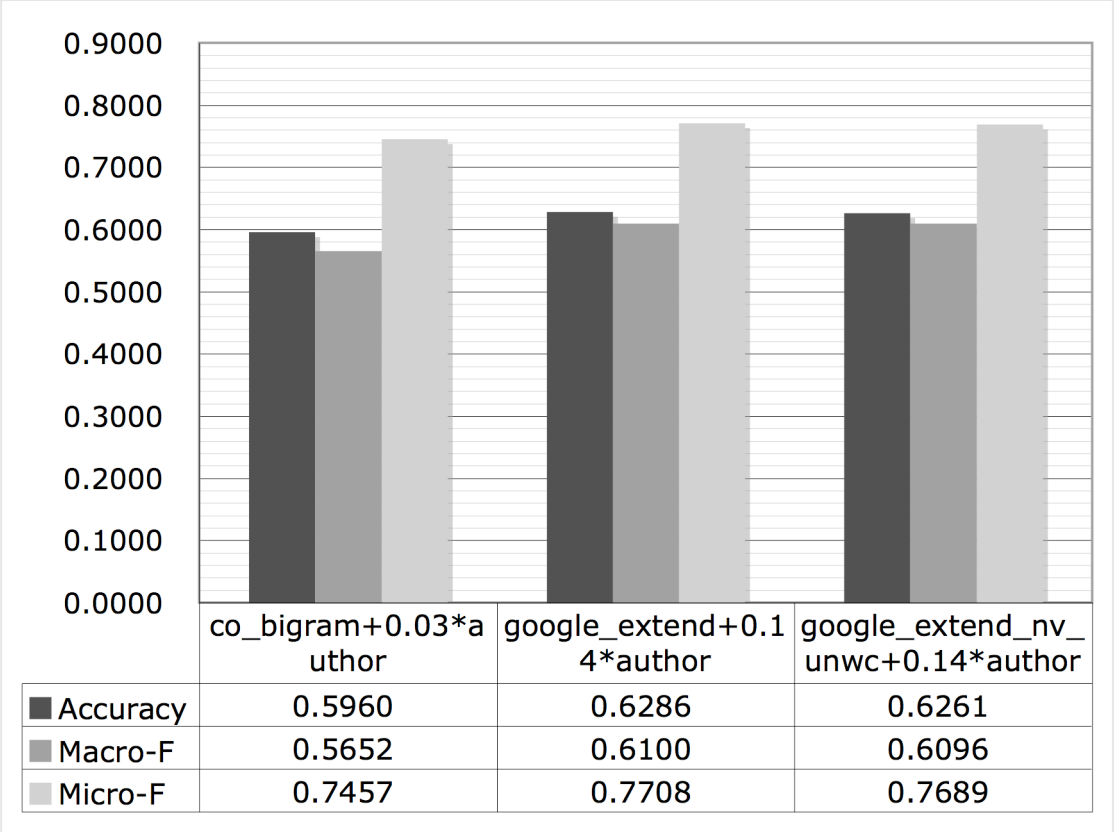
上圖表4為google_extend加入作者欄位因素後，不同作者欄位元素之權重下的分類成效，權重α以0.02增加，α為0時表示沒有加入作者欄位元素。上圖所示，α在0到0.02這個區間的分類成效成長最快，α於0.08至0.14這幾組其分類成效的曲線則呈現緩慢成長的趨勢，α為0.16時三個分類成效之評估指數皆開始下降。在google_extend加入作者欄位因素的實驗中，作者欄位與類別相似度之權重為0.14時，可得到最佳的分類成效。



	α=0	α=0.02	α=0.04	α=0.06	α=0.08	α=0.1	α=0.12	α=0.14	α=0.16
Accuracy	0.5693	0.5887	0.5980	0.6088	0.6151	0.6202	0.6237	0.6261	0.6245
Macro-F	0.5564	0.5760	0.5849	0.5946	0.6007	0.6074	0.6076	0.6096	0.6062
Micro-F	0.7242	0.7398	0.7471	0.7555	0.7604	0.7643	0.7669	0.7689	0.7677

圖表 5： 圖書題名利用擴展文件特徵（無詞性標記）並加入作者欄位。作者欄位於不同權重下的成效比較。

上圖表5為google_extend_nv_unwc加入作者欄位因素後，不同作者欄位元素之權重下的分類成效，權重α以0.02增加，α為0時表示沒有加入作者欄位元素。這個部分的曲線圖，與google_extend加入作者欄位因素後，不同作者欄位元素之權重下的分類成效相當類似。其權重α的最佳解亦為0.14。



	co_bigram+0.03*author	google_extend+0.14*author	google_extend_nv_unwc+0.14*author
Accuracy	0.5960	0.6286	0.6261
Macro-F	0.5652	0.6100	0.6096
Micro-F	0.7457	0.7708	0.7689

圖表 6：本階段各實驗組別之最佳成效比較

圖表六為第三階段各組實驗中，各組成效最好的成效比較。綜合第三階段各組之實驗圖表來看，無論是只使用圖書題名進行分類的組別或導入擴展文件特徵策略的組別，在導入作者欄位進行自動分類時，都有助於提升圖書自動分類的成效，再度驗證了林昕潔(2005)研究中所指出圖書資訊中的Meta-information(包含作者及出版社資訊)是有助於增進圖書自動分類的成效。在第三階段的各組實驗中又以google_extend+0.14*author這組的分類成效最好。

與未加入作者欄位元素的分類成效相比，在正確率成長了0.057；Macro-F成長了0.0514；Micro-F成長了0.0448。其中以正確率成長幅度最大，並且該組的Macro-F指數亦高於其他兩組。在第二章文獻分析時有介紹到，Macro-F此評估方式因容易受到大量文件較少的類別所影響，可藉此種評估方式了解大多數文件較少之類別其分類情況。可見導入擴展文件策略並加入作者欄位元素進行圖書自動分類，除了可以提昇整體的分類成效外亦能有效的幫助訓練文件數較少的類別提升辨識度。

在進行自動分類的計算上google_extend+0.14*author這組實驗，在導入擴展文件特徵策略這部分的相似度計算中其訓練文件所擷取出的特徵總共有5261774個、251879種，相對於co_bigram實驗組別的訓練文件所擷取出的特徵總數為2040926個、1474635種。相較之下，雖然特徵數量比co_bigram所擷取出的數量要高出兩倍，但種類數卻遠小於co_bigram所擷取出的種類數。由於在進行相似度計算時當訓練文件所擷取出的特徵種類越多時，所形成的「類別-特徵」向量也就越大，矩陣相乘時所需耗費的計算時間也隨之增加。因此，google_extend+0.14*author這組實驗組別，除了能有效的提升圖書自動分類的成效外，在自動分類時所需要的時間上亦能維持一定的水準。

第五章 實驗發現與未來展望

本章節將就實驗過程中所發現的現象做一歸納整理。由於在第四章的研究結果中，我們發現利用擴展文件特徵策略並加入作者欄位因素的實驗組別，其成效要比單純使用圖書題名進行自動分類的成效要好，並將測試文件完全分類錯誤的類別減少剩一類(見表一)。由於導入作者欄位來輔助分類，已在林昕潔(2005)的研究中證實在權重設置得宜的情況下可以增進圖書書目自動分類的成效，因此在本章節中關於加入作者欄位來輔助分類的部分便不多做探討。我們將從google_extend+0.14*author這組實驗中，測試文件完全分類錯誤的類別做為切入點，觀察本研究所提出的擴展文件特徵策略的缺失，並進一步剖析擴展文件特徵策略，能有助於增進圖書書目自動分類成效的原因。

實驗代號	測試文件完全分類錯誤的類別數
Seg	3
seg_nv	2
seg_with_word_class	3
seg_nv_wc	2
Bigram	3
co_seg	4
co_seg_nv	4
co_seg_with_word_class	4
co_seg_nv_wc	4
co_bigram	4
google_extend	4
google_extend_nv_unwc	4
co_bigram+0.03*author	1
google_extend+0.14*author	1
google_extend_nv_unwc+0.14*author	1

表 6：實驗代號及其對應的測試文件完全分類錯誤之類別數

此外由於本文所採用的訓練文件資料，其原本的分類是由專業圖書館員根據圖書內容以中國圖書分類法為分類依據進行分類而得。然而訓練文件的分類品質，可能會影響分類的成效。因此在本章中，筆者將以圖書資訊學的角度來觀察本研究的實驗資料，希望可以找出一些在圖書館員分類圖書的過程中，對於將來

進行圖書書目自動分類時可能會影響自動分類成效的因素。

最後，筆者將整理本研究中的實驗結果與發現，並進一步提出在圖書書目自動分類於未來的研究方向。以提供之後研究相關領域的研究者做為參考之用。

第一節 由分類結果檢討擴展文件特徵策略之缺失

在第四章的研究結果中，我們發現利用擴展文件特徵策略並加入作者欄位因素的實驗組別，並將測試文件完全分類錯誤的類別減少剩一類。特別是google_extend+0.14*author這組實驗組，其成效是本研究中所有研究組別裡分類成效最好的組別。儘管如此，我們還是希望知道在該實驗組別中，完全分類錯誤的類別中各測試文件分類錯誤的原因。因此我們將針對google_extend+0.14*author這組實驗組中測試文件完全分類錯誤的類別進行分析，試圖找出其分類錯誤的原因。

此一完全分類錯誤的類別為編號864，在中國圖書分類法中代表的是「中東文學」。由於在未加入作者欄位因素時(實驗組別：google_extend)，原本屬於該類別的測試文件便完全分類錯誤，且分類系統所給予的類號也與實驗組別：google_extend+0.14*author相同。可見加入作者欄位因素時，亦無法有效的使原本屬於該類的測試文件被系統正確的分類。

因此我們進一步觀察擴展文件特徵部分將其測試文件特徵與類別特徵相關聯的特徵提出進行觀察。並歸納出幾個造成分類錯誤的原因：

1. 影響分類結果的特徵詞彙在其他類別的相關度較高：

   測試文件代號，17276、17281、17289其分類原因屬於此類。見下表7-9：

<table>
<tr><td colspan="4">測試文件代號：17276</td></tr>
<tr><td colspan="4">書名：白色城堡</td></tr>
<tr><td colspan="2">系統分類之類別：875(德國文學)</td><td colspan="2">正確類別：864(中東文學)</td></tr>
<tr><td colspan="2">測試文件與類別875相關聯之特徵及其經過內積法計算過後的權重(由大至小排列，取前十表列)</td><td colspan="2">測試文件與類別864相關聯之特徵及其經過內積法計算過後的權重(由大至小排列，取前十表列)</td></tr>
<tr><td>城堡(n)</td><td>0.031678</td><td>土耳其(n)</td><td>0.012351</td></tr>
<tr><td>有(vt)</td><td>0.003614</td><td>白色(n)</td><td>0.006125</td></tr>
<tr><td>威尼斯(n)</td><td>0.003548</td><td>城堡(n)</td><td>0.004135</td></tr>
<tr><td>作者(n)</td><td>0.002712</td><td>土耳其人(n)</td><td>0.003467</td></tr>
<tr><td>故事(n)</td><td>0.002091</td><td>有(vt)</td><td>0.002525</td></tr>
<tr><td>出版社(n)</td><td>0.00207</td><td>故事(n)</td><td>0.002257</td></tr>
<tr><td>小說(n)</td><td>0.001871</td><td>機智(n)</td><td>0.001967</td></tr>
<tr><td>作品(n)</td><td>0.001799</td><td>作品(n)</td><td>0.001924</td></tr>
<tr><td>世界(n)</td><td>0.001551</td><td>世界(n)</td><td>0.001102</td></tr>
<tr><td>文學(n)</td><td>0.001431</td><td>小說(n)</td><td>0.001033</td></tr>
<tr><td colspan="4">錯誤分析：測試文件中特徵詞彙：「城堡(n)」與類別875較相關，且其相關程度大於類別864相關程度最高的詞彙：「土耳其(n)」</td></tr>
</table>

表 7：測試文件代號：17276之錯誤分析表

<table>
<tr><td colspan="4">測試文件代號：17281</td></tr>
<tr><td colspan="4">書名：巴格達部落格</td></tr>
<tr><td colspan="2">系統分類之類別：875(德國文學)</td><td colspan="2">正確類別：864(中東文學)</td></tr>
<tr><td colspan="2">測試文件與類別875相關聯之特徵及其經過內積法計算過後的權重(由大至小排列，取前十表列)</td><td colspan="2">測試文件與類別864相關聯之特徵及其經過內積法計算過後的權重(由大至小排列，取前十表列)</td></tr>
<tr><td>部落(n)</td><td>0.00649</td><td>部落(n)</td><td>0.004589</td></tr>
<tr><td>格(n)</td><td>0.00424</td><td>格(n)</td><td>0.00243</td></tr>
<tr><td>書(n)</td><td>0.004023</td><td>生活(n)</td><td>0.00143</td></tr>
<tr><td>上(n)</td><td>0.001981</td><td>書(n)</td><td>0.00102</td></tr>
<tr><td>網路(n)</td><td>0.001785</td><td>上(n)</td><td>0.000934</td></tr>
<tr><td>世界(n)</td><td>0.001088</td><td>戰爭(n)</td><td>0.000798</td></tr>
<tr><td>人(n)</td><td>0.001058</td><td>世界(n)</td><td>0.000773</td></tr>
<tr><td>作者(n)</td><td>0.000951</td><td>有(vt)</td><td>0.00059</td></tr>
<tr><td>有(vt)</td><td>0.000845</td><td>人(n)</td><td>0.000537</td></tr>
<tr><td>譯者(n)</td><td>0.000835</td><td>幽默(vi)</td><td>0.000406</td></tr>
<tr><td colspan="4">錯誤分析：由於近年來網際網路上部落格的興起，利用google查詢時，許多網頁標題都帶有「部落格」一詞。再加上與中東較具關聯的詞彙「巴格達」可能因測試文件數不足而未被收入至類別864的特徵向量中。造成因「部落(n)」與「格(n)」兩個詞彙與類別875的相關度較高使本測試文件被分類至類別875。</td></tr>
</table>

表 8：測試文件代號：17281之錯誤分析表

<table>
<tr><td colspan="4">測試文件代號：17289</td></tr>
<tr><td colspan="4">書名：近東開闢史詩</td></tr>
<tr><td colspan="2">系統分類之類別：850(中國各類文學)</td><td colspan="2">正確類別：864(中東文學)</td></tr>
<tr><td colspan="2">測試文件與類別850相關聯之特徵及其經過內積法計算過後的權重(由大至小排列，取前十表列)</td><td colspan="2">測試文件與類別864相關聯之特徵及其經過內積法計算過後的權重(由大至小排列，取前十表列)</td></tr>
<tr><td>史詩(n)</td><td>0.033271</td><td>圖書(n)</td><td>0.0045717</td></tr>
<tr><td>文學(n)</td><td>0.009152</td><td>出版社(n)</td><td>0.0009459</td></tr>
<tr><td>圖書(n)</td><td>0.008425</td><td>書(n)</td><td>0.0008558</td></tr>
<tr><td>出版社(n)</td><td>0.002872</td><td>作者(n)</td><td>0.0007432</td></tr>
<tr><td>書(n)</td><td>0.001805</td><td>文學(n)</td><td>0.0006306</td></tr>
<tr><td>信息(n)</td><td>0.000906</td><td>新(vi)</td><td>0.0006306</td></tr>
<tr><td>作者(n)</td><td>0.000839</td><td>網(n)</td><td>0.0005743</td></tr>
<tr><td>新(vi)</td><td>0.000739</td><td>相關(vi)</td><td>0.0004504</td></tr>
<tr><td>價格(n)</td><td>0.000689</td><td>我們(n)</td><td>0.0004015</td></tr>
<tr><td>定價(n)</td><td>0.000672</td><td>價格(n)</td><td>0.0003716</td></tr>
<tr><td colspan="4">錯誤分析：比對書名後可發現，上面兩個欄位的特徵詞彙裡與書名相關的特徵只有「史詩(n)」，該特徵相較於其他特徵，對類別850具有高度的相關。此外觀察了其餘的特徵，大多數的特徵似乎無助於了解該書之內容。</td></tr>
</table>

表 9：測試文件代號：17289之錯誤分析表

除了以上三個測試文件的錯誤分析外，將以上三個測試文件的錯誤分析並列觀察時，筆者亦觀察到在測試文件中與「中東」這個地域概念相關的特徵，在進行相似度計算時，不是沒比對到就是強度不足以使測試文件被分類至類別864中。

其改善方法，除了可增加類別864的訓練文件外。亦可利用專家選詞的方式針對描述地域性的語彙，根據類別所涵蓋的地域觀念增減其權重。

2. 因擴展特徵詞彙策略而造成類別或測試文件向量納入不相關的特徵：

測試文件代號：17280其分類原因屬於此類。見下表10：

<table>
<tr><td colspan="4">測試文件代號：17280</td></tr>
<tr><td colspan="4">書名：男味</td></tr>
<tr><td colspan="2">系統分類之類別：875(德國文學)</td><td colspan="2">正確類別：864(中東文學)</td></tr>
<tr><td colspan="2">測試文件與類別875相關聯之特徵及其經過內積法計算過後的權重(由大至小排列，取前十表列)</td><td colspan="2">測試文件與類別864相關聯之特徵及其經過內積法計算過後的權重(由大至小排列，取前十表列)</td></tr>
<tr><td>男味(n)</td><td>0.002488</td><td>小知堂(n)</td><td>0.00291</td></tr>
<tr><td>價格(n)</td><td>0.00197</td><td>小說(n)</td><td>0.000957</td></tr>
<tr><td>人(n)</td><td>0.001861</td><td>人(n)</td><td>0.000944</td></tr>
<tr><td>小說(n)</td><td>0.001732</td><td>價格(n)</td><td>0.000519</td></tr>
<tr><td>商品(n)</td><td>0.001483</td><td>有(vt)</td><td>0.000519</td></tr>
<tr><td>文化(n)</td><td>0.000917</td><td>商品(n)</td><td>0.00051</td></tr>
<tr><td>比較(vt)</td><td>0.000877</td><td>故事(n)</td><td>0.000464</td></tr>
<tr><td>作者(n)</td><td>0.000837</td><td>文化(n)</td><td>0.000429</td></tr>
<tr><td>有(vt)</td><td>0.000744</td><td>世界(n)</td><td>0.00034</td></tr>
<tr><td>書(n)</td><td>0.000708</td><td>部落(n)</td><td>0.000336</td></tr>
<tr><td colspan="4">錯誤分析：比較測試文件與兩個類別相關之相似度前五名的特徵，可發現測試文件特徵中與類別875相關的特徵，其相似度大多大於測試文件特徵與類別864相關的特徵。特別的是書名「男味」，卻出現在類別875中，進行反向分析時筆者發現會造成該詞彙出現在類別875的原因是因為，類別875的訓練文件(系統號：18164，書名：菠菜吸血鬼)進行擴展特徵詞彙時，該書籍的譯者亦有翻譯「男味」這本書，並出現在擴充詞彙策略蒐集的範圍內被系統蒐集為該類的特徵詞彙。</td></tr>
</table>

表 10：測試文件代號：17280之錯誤分析表

為了減低上述因擴展特徵詞彙策略而造成類別或測試文件向量納入不相關的特徵。可在要送入搜尋引擎查詢的查詢句中加入作者等改進檢索策略的手段，進而提高檢索結果與書籍的相關性。來避免因擴展特徵詞彙策略而造成類別或測試文件向量納入不相關的特徵進而影響分類結果。

3. 該圖書亦可被分入系統所賦予之類別

測試文件代號：17282其分類原因屬於此類。見下表11：

<table>
<tr><td colspan="4">測試文件代號：17282</td></tr>
<tr><td colspan="4">書名：紀伯倫散文詩全集 : the complete prose poems</td></tr>
<tr><td colspan="2">系統分類之類別：865(阿拉伯文學)</td><td colspan="2">正確類別：864(中東文學)</td></tr>
<tr><td colspan="2">測試文件與類別865相關聯之特徵及其經過內積法計算過後的權重(由大至小排列，取前十表列)</td><td colspan="2">測試文件與類別864相關聯之特徵及其經過內積法計算過後的權重(由大至小排列，取前十表列)</td></tr>
<tr><td>紀伯倫(n)</td><td>0.237851</td><td>出版社(n)</td><td>0.001965</td></tr>
<tr><td>散文詩(n)</td><td>0.009476</td><td>相關(vi)</td><td>0.001871</td></tr>
<tr><td>全集(n)</td><td>0.004708</td><td>文學(n)</td><td>0.00131</td></tr>
<tr><td>出版社(n)</td><td>0.002657</td><td>出版(vt)</td><td>0.001123</td></tr>
<tr><td>文學(n)</td><td>0.001832</td><td>資訊(n)</td><td>0.001123</td></tr>
<tr><td>相關(vi)</td><td>0.001721</td><td>著(vt)</td><td>0.000755</td></tr>
<tr><td>卡里．紀伯倫(n)</td><td>0.001606</td><td>文化(n)</td><td>0.000708</td></tr>
<tr><td>出版(vt)</td><td>0.001313</td><td>譯(vt)</td><td>0.000695</td></tr>
<tr><td>資訊(n)</td><td>0.001238</td><td>中國(n)</td><td>0.000561</td></tr>
<tr><td>詩(n)</td><td>0.001082</td><td>詩(n)</td><td>0.000419</td></tr>
<tr><td colspan="4">錯誤分析：這個案例，經筆者查證該圖書內容後認為系統沒有分類錯誤。紀伯倫是阿拉伯文學著名作者之一，其利用阿拉伯文所著作之散文詩，本該分類至阿拉伯文學中。不過正確類別將該本圖書分類至中東文學，也不算分類錯誤。因為阿拉伯文學本屬於中東文學中的一支。實務上碰到這樣可以分入兩個類號的圖書都會根據該圖書館的政策來選擇適當的類號。但當館員對此種狀況有分類不一致的情況發生時，以這些人工分類的書目進行分類器的訓練，便容易造成本案例的情況發生。</td></tr>
</table>

表 11：測試文件代號：17282之錯誤分析表

第二節 以圖書資訊學角度探討影響圖書自動分類成效之因素

在實驗的過程以及進行錯誤分析時，筆者也發現了一些可能影響圖書自動分類成效的因素隱藏在中國圖書分類法的結構以及圖書館進行圖書分類的實務中。本節希望就實驗結果中的實際例子，以圖書資訊學的角度探討並提出可能的改善方法，希望在實際的應用上能進一步的改善圖書自動分類的成效。

由於分類器需要經過機器學習後，才能進行自動分類作業。因此訓練文件的品質也是影響自動分類成效的因素之一。但由於在圖書館中的圖書分類是圖書館員根據圖書內容來判斷歸類，不可避免的會有分類不一致的情況存在，進而影響到往後進行圖書自動分類的成效。特別是中國圖書分類法的類目結構又更增加了圖書館員分類不一致的機率。

我們在第一章中提到中國圖書分類法是一個三層的十進階層式分類法。每一層結尾為零的類目（例：100、010）基本上為該大類的總論，圖書內容為該大類的概論或者擁有這大類下數種子類目內容的圖書皆可歸類至總論。因為具有總論屬性之類目有這樣的特性，當館員缺乏某一學科的領域知識時，便容易將本該繼續往子類目細分的圖書歸類至概念涵蓋範圍較廣的總論屬性類目中。

此外在進行實驗時，筆者也發現中國圖書分類法在某些類目下的子類目，也不一定按照概念的階層排列。如賴永祥先生（2001）於中國圖書分類法第八版的例言中提到：「***其標記採用數字，以層累原則而編成，但不拘於十進***」。如類號為860的東方文學總論下的子類目變有這樣的情況。

類號	類目概念
860	東方文學總論
861	日本文學
862	韓國文學
863	遠東各地文學
864	中東文學
865	阿拉伯文學
866	伊朗文學
867	印度文學
868	東南亞各國文學
869	南洋文學

表 12：類目及其對應之類目概念對照表

如上表12所示，若依照階層原則，日本文學及韓國文學理應隸屬於遠東各地文學的子類目。可能因為該國文學於台灣地區為顯學，圖書出版的總類較多，需要向下細分的空間，因此才與遠東各地文學分屬同層類號。不過這樣的情況，亦有可能增加圖書館員分類不一致的情況產生。如本章第一節錯誤分析的案例：測試文件代號17282，書名為：紀伯倫散文詩全集：the complete prose poems。圖書館員給予這本書的類別為864(中東文學)，但經筆者了解作者的生平以及其於阿拉伯現代文學的地位後，則認為該本圖書應歸類至865（阿拉伯文學）中。但若是館員有依據864(中東文學)的指示利用世界分國表進行複分，則館員的歸類方法亦無不當。每個人進行歸類的觀點本來就會有相異之處，若以檢索的立場來看，這樣的例子應該給予兩個分類號以方便讀者檢索。但分類類號在圖書館內同時代表著該書在館內的排架位置。在一本書只能有一個固定位置的原則下，館員只能選擇一個類號作為該書的類號，進而增加分類不一致的機會。

在編目實際作業上，要減低上述因圖書內容或分類架構而造成一本書需要在一個以上的類號中選擇一個類號，而造成分類不一致的情形。常見的方法有下列

兩種：

1. 制定該圖書館的分類策略。

   較大型的圖書館為求其服務品質，多半都會根據該館的館藏發展政策，制定適當的分類策略，使館員在碰到一書多主題或可分入一個以上類號的情形時能有個固定的依據，減輕館員分類不一致的情況。

2. 鼓勵館員進修增進了解其他領域的知識。

   有時候分類不一致的情況也可能是因為分類館員對該本書的學科領域不熟悉，導致給號的困難。由於我國圖書館學教育的關係，國內圖書館中的專業館員多半只擁有圖書館學的知識，又更加重了分類不一致的可能性。因此鼓勵館員進修增進了解其他領域的知識，亦為減輕圖書館員分類不一致的方法之一。

除上述策略外，倘若圖書館有引進自動分類技術來協助館員分類，盡可能落實上述兩個既有方法來減輕館員分類不一致的情況，以維持圖書館書目的分類品質。相信對於館員能更準確且有效率進行編目工作。

此外，在實驗的過程中，我們也有發現因為中國圖書分類法改版，因類號更動而造成分類錯誤的問題。舉例而言，筆者在檢視分類結果時發現，類號424以及類號425的圖書相當相似，其中分類錯誤的測試文件也多被系統分入類號424及類號425(例：原本類號應歸入424的書籍，系統卻歸入類號425；而原本類號應歸入類號425的書籍卻被歸入類號424)，進而影響書目自動分類的成效。

經查證發現，這兩個類號的類目皆為「美容」，只是第七版中國圖書分類法中類號為424的美容類在第八版時改入類號425，原本類號424則被空而不用。(國家圖書館增訂類目表，2001)一般來說圖書館內的圖書一旦給號，基於更改類號牽動到更動書標、排架等等作業的考量下，即使分類法的類目更動，也不會進行類號更新的工作(從實驗資料中即可證實此一說法)。這樣的情況更加重了圖書館書目自動分類的困難度。要避免上述因分類法的類號變動而影響自動分類成效的問

題，在不更改圖書分類法的情況下，在分類器進行訓練的過程中，加上檢查並更新訓練文件類號的處理，可能是較可行的處理方法之一。

以上主要是根據本研究中實驗之過程與結果所觀察到的情況進行查證與分析而得。由於實驗資料的限制，本研究只觀察到中國圖書分類法中應用科學及語文兩大類的自動分類結果，相對於擁有十大類圖書的圖書館而言。此處所觀察到因圖書分類法或館員而影響圖書書目自動分類的因素，或許只是冰山一角。還有賴更大規模的研究來發掘其他隱性之圖書書目自動分類的因素。進而增進圖書書目自動分類的實用性。

第三節 結論與未來展望

一直以來，圖書館引進資訊技術多半運用於協助讀者探索圖書館。而利用資訊技術來減輕編目館員工作負擔者，在國內的研究中能見度並不高。由於圖書館所採用的館藏分類遠比其他文件類型所需歸類的類別要多，結構上也較為複雜，而圖書館館藏書目所涵蓋的紀錄項目多半是用來描述書籍的外部表徵，與圖書內容相關的欄位並不多。圖書題名在用字上以及長度上也不如新聞標題來的穩定（新聞標題需反應內文，讓人了解新聞內容的大意）。

在文件自動分類處理程序中，文件特徵擷取是影響分類結果好壞的重要因素之一，所以特徵擷取方法的選擇與設計相當重要（曾元顯，2002）。如前所述：圖書題名在用字上以及長度上不如新聞標題來的穩定。要以這樣的文件類型進行文件自動分類，筆者認為找出適合於圖書題名的特徵擷取方法，使擷取出的文件特徵足以代表文件內容的意涵，乃為增進圖書館自動分類系統成效之上策。為此本研究設計了三個階段的實驗，希望能找出適合於圖書題名的特徵擷取方法。

在第一階段的實驗中，我們利用數個常見的特徵擷取方法，來觀察不同特徵擷取方法的分類成效。並藉由各特徵擷取方法所擷取出的特徵種類數及特徵數量分析其對分類成效的影響。實驗中我們發現利用雙連字串擷取圖書題名之特徵並進行共現詞分析的實驗組(實驗代號：co_bigram)成效最好，其次為只利用雙連字串擷取文件特徵的實驗組(實驗代號：bigram)。這兩組的分類成效其實相差不多，正確率的差距僅差0.0078，若將分類的運算時間納入考量，利用雙連字串擷取圖書題名所擷取出的特徵總類數較利用雙連字串擷取圖書題名之特徵並進行共現詞分析的實驗組少。利用雙連字串擷取圖書題名要優於利用雙連字串擷取圖書題名之特徵並進行共現詞分析。

第二階段的實驗則企圖針對圖書題名的特性，並以第一階段的實驗為基礎，希望能改良現有的特徵擷取方法，進而增進圖書書目自動分類的成效。由於圖書題名在用字上以及長度上不如新聞標題來的穩定特性。原本的改良方向有兩種，

一為將擷取後的特徵進行語彙的隱性分析；其二為利用擴展的概念，增加與該圖書題名內容相關的特徵。由於在前期實驗時發現，筆者所能取得的語意剖析工具，難以運用在圖書題名上，便將目光擺在擴展文件特徵上。

在文件自動分類的領域，擴展的概念多用於解決訓練文件不足的問題（ 曾元顯、莊大衛，2003）。在此我們則希望利用特徵擴展的概念解決因圖書題名較短的文件所產生特徵不足難以判斷類別的問題。由於網際網路上的搜尋引擎最主要的功能在於尋找與查詢語句相關的網頁，因此研究中假設當利用圖書題名做為查詢語句送入搜尋引擎進行查詢，搜尋引擎所回饋的網頁中搜尋結果描述與網頁標題亦與圖書題名相關。因此本實驗便根據上述的假設，以圖書題名為查詢語句利用Google搜尋引擎進行檢索，並擷取第一頁的回饋網頁中搜尋結果描述與網頁標題與圖書題名結合成一新文件。由於經過擴展後的文件長度增加，為了避免因擷取的特徵太多而造成計算時間過度增加的情況，因此我們採用了利用斷詞器進行斷詞只擷取具有動詞名詞屬性的詞彙，並過濾掉較不具功能的虛詞語彙，且分為有詞性標記、無詞性標記兩組實驗組進行實驗。與第一階段co_bigram實驗組相比，兩組經過擴展文件特徵的實驗組皆比只利用圖書題名進行自動分類在三種評估指標的成效上都要來的好。並且雖然利用擴展文件特徵之方法，增加了文件的長度，但因選擇了適當的特徵擷取方法，限制了特徵的種類數量，因此利用擴展文件特徵的實驗組別，在運算速度上，亦比第一階段的co_bigram實驗組要快。

在第三階段的實驗中，其目標在於改善第二階段的擴展特徵方法。由於利用第二階段的擴展文件特徵方法，因擴展的方法是利用搜尋引擎進行查詢，並擷取搜尋引擎回饋後的訊息。很難避免會有擷取到不相關資訊的情形，進而影響到自動分類的成效。前人的研究已有利用圖書書目中作者及出版社的資訊，來增進圖書自動分類的成效。本研究便採用作者資訊來改善擴展特徵方法的缺失，希望能進一步增進擴展特徵方法的成效。實驗中亦將第一階段的co_bigram實驗組加上作者資訊做為實驗對照組。實驗結果顯示，利用擴展特徵方法的兩組實驗組分別加入作者資訊來輔助分類，於適當的調整權重後在分類成效的各項指標上皆比實驗對照組的表現要好。特別是實驗組google_extend+0.14*author於各項分類成效指

標上表現最為優異，不但提升了整體的分類成效，在訓練文件數較少的類別之分類成效亦有長足的進步。

在錯誤分析方面，我們分析了三個階段中實驗最好的實驗組別google_extend+0.14*author，針對完全分類錯誤的類號之測試文件進行分析。分析結果發現，會造成此類別完全錯誤，最主要的原因為該類別的測試文件數不足。其次為擴展文件特徵方法所產生的雜訊與中國圖書分類法的結構、圖書館員分類不一致等等因素交互影響而造成該類的測試文件完全分類錯誤。

由於圖書館所採用之分類法較為複雜。因此在實驗的過程中，筆者亦針對因圖書分類法的結構而造成的分類錯誤情形進行觀察。發現，因圖書分類法結構而造成分類錯誤的因素主要分為兩種，其一為因圖書分類法改版造成的分類錯誤，以及因圖書分類法的結構造成館員分類不一致進而影響分類成效的情形。有意引進自動分類技術輔助圖書分類的圖書館，應制訂有效的分類規範減低館員分類不一致的情況。並在系統設計時，留意圖書分類法的改版所造成的類目更動。

以下便針對本研究的實驗結果及研究過程中所發現到的現象，提出幾點關於圖書館圖書自動分類未來的研究方向與實際建置輔助圖書分類系統的建議：

1. 提升擴展文件特徵方法的品質。

   本研究所提出的擴展文件特徵方法，仍未盡完善，當書名較短且所用詞彙為高頻詞時，往往會納入與該圖書內容較不相關的語彙，進而造成因雜訊影響導致分類困難的情況。可見，只利用圖書題名做為查詢語句擷取搜尋引擎所回饋的資訊，在分類成效上仍有進步的空間。以查詢出與圖書內容相關的資訊為前提，未來的研究可利用圖書中的作者、出版社等等欄位來搭配圖書題名進行查詢且縮小擷取搜尋引擎所回饋資訊的範圍，或以其他方式增加所擴展的特徵與圖書內容的相關性。此外因本文中所提出之擴展文件特徵方法，並沒有對擴展處理後的特徵進行特徵刪減的動作，因此如何過濾經擴展處理後的特徵，使留下的特徵更具有類

別辨識度（抑或更符合圖書內容），進而增進分類的效率與成效，是之後相關研究可以嘗試的方向之一。

2. 導入專家選詞。

在觀察中國圖書分類法時亦發現許多的類別是具有地域性的概念。但實際進行自動分類時，可能會因為類別中的訓練文件數不夠或其他因素，使得這些具有地域性質的詞彙在經過機器訓練後在該類別的權重仍不高或根本沒收入到這一類的詞彙，而因此即使測試文件中出現具有該類別的地域性詞彙，卻可能因該詞彙在此類別的權重不高而被分類至其他的類別。因此在此建議未來的實驗可以朝著針對地域性的詞彙進行專家選詞來改善上述所提到的問題。

3. 導入分散式系統。

由於本研究所提出的自動分類系統，係輔助圖書館員進行圖書分類之用，因此系統的反應時間亦相當重要。有鑑於圖書館的藏書、分類類目眾多，在實際的運用上，系統進行分類的時間，必定比本研究中所需要的運算時間要多。因此利用分散式系統，藉由多台主機分擔進行自動分類時的運算，進而縮短系統進行自動分類的運算時間亦有其必要性。

4. 嘗試不同的分類器。

由於研究時間的限制，在實驗的設計上，本研究於實驗設計上將分類器做為固定變項。因此未來有興趣的研究者亦可利用不同分類器進行圖書書目之自動分類實驗，觀察不同分類器對於圖書書目自動分類於成效上的影響。

5. 進行類目涵蓋範圍更大的實驗。

限於研究資源及研究時間，論文中所蒐集的資料，僅限於應用科學及語文兩大類的書目資料。雖然在類別的數量上較其他利用書店的分類架構進行圖書自動分類之研究相比已高出很多，但在卻與圖書館實際上所採

用的類別數量仍有段差距。再加上中國圖書分類法是一部具有彈性的分類法。因此本研究所發現的影響圖書自動分類成效的因素或許只是冰山一角。還有賴更大規模的實驗來挖掘隱藏在中國圖書分類法中影響圖書書目自動分類成效的因素。

6. 減少圖書分類系統所需處理的類別數量。

有鑑於類別文件的數量對自動分類成效的影響，在實際的圖書書目自動分類系統的建置上，我們可將原本細分至三層的分類號縮減成兩層(例：110-119皆做為11X類)，如此一來除了類別數量縮減外，亦可以避免三層類號的類別下容易造成訓練文件數不足的情況發生。最後一層的類號則可由系統提示館員。此法不但有助於提升系統的分類成效，亦能有效的輔助編目館員縮減圖書分類的時間。

希望以上的建議能有助於未來圖書書目自動分類之研究。進而提升自動分類技術運用於圖書館圖書分類的可行性。

中文參考書目：

1. 王省吾（1982）。圖書分類法導論。台北：私立中國文化大學。

2. 黃淵泉（1986）。中文圖書分類編目學。台北：臺灣學生。

3. 陳光華（1996）。資訊檢索查詢之自然語言處理。中國圖書館學會會報，第57期，p141-153。

4. 曾元顯（1997）。關鍵詞自動擷取技術之探討。中國圖書館學會會訊，第106期，pp. 26-29。

5. 杜海倫（1999）以標題進行新聞自動分類。 碩士論文，國立清華大學。

6. 王稔志、張俊盛（2001）。適應性文件分類系統。，第十四屆計算語言學研討會論文集，pp. 99-121。

7. 賴永祥（2001）中國圖書分類法2001增訂八版，台北：文華。

8. 許雅芬（2002）。新聞文件自動分類之研究。碩士論文，私立東吳大學。

9. 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，第 68 期，pp. 62-83。

10. 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（頁 129-141）。

11. 林政男（2004）。以共現語詞為基礎的特徵選取在文件自動分類上之研究。碩士論文， 銘傳大學。

12. 馬偉雲（2004）。未知詞擷取作法。網址：http://ckipsvr.iis.sinica.edu.tw/uwe.htm，上網日期：2008/04/08。

13. 林昕潔（2005）。以SVM與詮釋資料設計書籍分類系統。碩士論文，國立交通大學資訊科學與工程研究所。

14. 國家圖書館增訂類目表，民國89年版。網址：http://catweb.ncl.edu.tw/2-1-10a.htm ，上網日期 2008/04/08

英文參考書目：

1. Kwok, K. l. (1975), "The Use of Title and Cited Titles as Document Representation for Automatic Classification, Journal of Information and Management, Vol. 11 pp.201-206, 1975.

2. Tseng, Y-H. (2001), "Fast co-occurrence thesaurus construction for Chinese news, IEEE International Conference on System, Man, and Cybernetics, Vol. 2, pp. 853-858.

3. Salton, G. (1988)., Automatic text processing : the Transformation, Analysis, and Retrieval of Information by Computer, Mass. : Addison-Wesley.

4. Salton, G. (1989). Automatic Text Processing, Mass. : Addison-Wesley, 1989.

5. Salton, G. & C. Buckley, (1988) "Term Weighting Approaches in Automatic Information Retrieval," Journal of Information Proceeding and Management, Vol.24:3, pp. 513-524.

6. Sullivan, D. (2001). Document Warehousing and Text Mining, 326. Wiley Computer Publishing, New York, NY.

附錄A 各類別訓練文件與測試文件數目表

類號	訓練文件數量	測試文件數量
400	312	78
402	28	7
403	68	17
406	34	8
407	37	9
408	154	38
409	42	10
410	220	55
411	231	57
412	92	22
413	226	56
414	200	50
415	161	40
416	248	61
417	224	55
418	172	42
419	109	27
420	50	12
421	94	23
422	153	38
423	335	83
424	181	45
425	86	21
426	275	68
427	147	36
428	280	69
429	291	72
430	62	15
431	175	43
432	177	44
433	197	49
434	305	76
435	341	85
436	108	27
437	115	28
438	60	15
439	83	20
440	37	9
442	128	31
443	206	51
444	72	17
445	204	51
447	174	43
449	66	16
450	49	12
452	46	11
454	60	15
456	42	10

類號	訓練文件數量	測試文件數量
457	32	7
461	23	5
462	24	6
463	114	28
464	128	32
465	71	17
466	95	23
467	245	61
468	60	14
469	60	15
470	44	10
471	228	56
472	59	14
474	81	20
475	24	6
476	48	11
477	218	54
478	195	48
479	159	39
480	57	14
481	126	31
483	341	85
485	25	6
486	84	20
487	86	21
488	118	29
489	360	90
490	192	47
492	60	15
493	229	57
495	29	7
496	56	14
497	66	16
498	270	67
499	60	14
800	165	41
801	107	26
802	160	39
803	77	19
804	173	43
805	37	9
806	42	10
810	63	15
811	105	26
812	100	24
813	102	25
815	152	37
819	54	13
820	84	21
821	62	15
822	30	7
823	130	32

類號	訓練文件數量	測試文件數量
824	116	28
825	77	19
827	38	9
829	163	40
830	151	37
831	257	64
832	151	37
833	180	44
834	236	59
835	212	53
836	80	19
839	188	47
842	24	5
843	32	8
844	103	25
845	154	38
846	151	37
847	197	49
848	238	59
850	128	32
851	191	47
852	32	8
853	257	64
854	91	22
856	448	112
858	121	30
859	43	10
862	167	41
863	72	17
864	24	5
865	74	18
867	61	15
868	207	51
870	76	19
871	49	12
872	131	32
873	116	28
875	175	43
876	272	68
877	121	30
878	61	15
879	60	15
880	47	11
881	163	40
882	125	31
883	24	5
885	173	43
886	44	10
887	51	12
889	60	15
890	236	58
891	36	9

類號	訓練文件數量	測試文件數量
893	64	15
895	164	41
898	117	29
899	53	13