胡雅涵、黃正魁、楊承翰（2014），『以基因演算法為基礎建立自動化文件分類模式』，*資訊管理學報*，第二十一卷，第三期，頁 305-340。

以基因演算法為基礎建立自動化文件分類模式

胡雅涵
國立中正大學資訊管理學系
國立中正大學前瞻製造系統頂尖研究中心

黃正魁*
國立中正大學企業管理學系

楊承翰
國立中正大學資訊管理學系

摘要

數位資訊迅速地成長，增加人們在找尋資訊上的搜尋成本，如何有效地分類管理文件已是一項重要的研究議題。因此，文件分類研究的重要性與日俱增，在文件分類領域中存在文件特徵維度過高的問題，因此，我們以基因演算法（Genetic Algorithm, GA）為基礎選取文件中特徵字詞，透過對 GA 染色體於文件特徵向量設計和調整 GA 設定的參數，讓分類器（Classifier）從訓練資料中選取特徵字詞，並進行文件分類模式建構。本研究提出之 GA 特徵選取（GA-based Feature Selection, GAFS）方式，透過讓各單一分類器都能自我學習達到最佳化，進而提升各分類器的分類效能，以建構出分類效果最佳化的文件分類模式。實驗部分，本研究採用 WebKB 網頁文件資料集，評估 GAFS 所建立的文件分類模式，並與傳統將所有特徵集合進行訓練之方法（簡稱 TOTAL）做比較。本研究採用六種不同的分類器模式，包含貝氏分類器（Naïve Bayesian Classifier）、決策樹（Decision Tree）、分類迴歸樹（Classification and Regression Tree）、隨機森林（Random Forest）、支援向量機（Support Vector Machine），以及 k 最近鄰居法（k Nearest Neighbor）。實驗結果顯示，本研究提出之 GAFS 方法能夠有效地改善各分類模式的分類效能，證實以 GA 為基礎之 GAFS 自動化文件分類模式明顯優於 TOTAL，並且在特徵維度逐漸擴大的情況下，GAFS 仍能有效地改善分類效能，並且擁有穩定地分類準確率。

關鍵詞：文件分類、基因演算法、特徵選取、分類器

* 本文通訊作者。電子郵件信箱：bmahck@ccu.edu.tw
2013/10/18 投稿；2014/03/24 修訂；2014/05/18 接受

Hu, Y.H., Huang, T.C.K. and Yang, C.H. (2014), 'A Genetic Algorithm Based Approach for Text Categorization', *Journal of Information Management*, Vol. 21, No. 3, pp. 305-340.

A Genetic Algorithm Based Approach for Text Categorization

Ya-Han Hu
Department of Information Management, National Chung Cheng University
Advanced Institute of Manufacturing with High-tech Innovations, National Chung Cheng University

Tony Cheng-Kui Huang*
Department of Business Administration, National Chung Cheng University

Cheng-Han Yang
Department of Information Management, National Chung Cheng University

Abstract

Purpose: Digital data has been accumulated rapidly resulting in the significant increase in the cost of searching information from the data source. How to effectively manage documents (i.e., text categorization, TC) has become an important research issue. However, in TC, huge amount of index terms are selected for representing document vectors, resulting in poor prediction outcomes. This study proposes a genetic algorithm based feature selection (GAFS) method to optimize the selection of index terms.

Design/methodology/approach: Before training classifiers, GAFS selects a reduced set of index terms that can optimize the prediction accuracy of classifiers. In experimental study, the WebKB dataset was used to evaluate the performance of GAFS. A total of six well-known classification techniques were considered, including naïve Bayesian classifier (NB), decision tree (DT), classification and regression tree (CART), random forest (RF), support vector machine (SVM) and k-nearest neighbor (kNN). The baseline model, denoted as TOTAL, is to consider complete set of index terms in all

* Corresponding author. Email: bmahck@ccu.edu.tw
2013/10/18 received; 2014/03/24 revised; 2014/05/18 accepted

experiments.

Findings: The results show that the proposed GAFS method outperforms the TOTAL method. The performance of kNN and RF classifiers deteriorates as the number of features increases. Under different number of features, the SVM, NB, and DT classifiers perform stably but the CART classifier has relatively unstable performance.

Research limitations/implications: This study only considers the WebKB dataset. Future research is recommended to include other well-known datasets in the TC domain. Other feature selection methods can be also considered in the experimental evaluation.

Practical implications: Two practical implications are provided. First, this study reveals that different parameter settings in genetic algorithm (GA) can significantly affect the performance of feature selection in TC. Second, the proposed GAFS method allows users to systematically construct a robust classifier for TC.

Originality/value: This paper investigates the influence of the parameters used in GA for the feature selection in TC. It advances the literature in choosing GA parameters and classification techniques for optimizing the TC performance.

Keywords: document categorization, genetic algorithm, feature selection, classifier

壹、緒論

隨著資訊科技的進步，資訊成長的速度與累積的數量已不可同日而語，大量繁雜的資訊迅速地累積，增加尋找資訊的困難度，如能建立一個有效率的文件分類管理方法，將減少人們在尋找所需文件的搜尋成本。文件分類主要從預先定義好的類別集中，給予文件主題類別標籤的一項工作（Sebastiani 2002）。簡單來說，就是根據文件的內容將其分門別類，使每份文件都能歸屬至合適的分類之中，以便讀者查詢其所需之文件。例如，新聞文件可按其報導的內容，給予「政治」、「財經」、「娛樂」等類別，如此一來，讀者就能有效率地瀏覽自己感興趣的文件。

傳統文件分類的做法是以人工方式進行，也就是讓領域知識工作者檢閱文件，並依照他們各自的知識與經驗了解文件內容，再將文件歸屬至適當的類別之中。由於以人力方式進行文件分類是困難且耗時的，所以 Joachims（1998）認為透過分類技術去學習固有的文件範例，再進行文件分類是比較有效率的方法。而「自動化文件分類」，即是利用機器學習中監督式學習的方法，從已知類別文件的資料集歸納出分類模式，再依據分類模式給予新文件合適的類別標籤（Yang et al. 1999）。

自動化文件分類通常可分為兩個步驟－特徵選取及機器學習（Chou et al. 2007; Man et al. 2009; Wang et al. 2012; Wei et al. 2011）。首先，由於電腦不具備人類閱讀文件的能力，所以我們必須將文件整理成系統可分析的結構化格式，其通常是把文件中非結構化的句子分割成許多字詞，再用這些字詞代表文件做進一步的分析，此即為特徵選取的部分；接著，在機器學習的部分，我們使用已知類別的文件（訓練資料）進行監督式學習，使分類技術能夠自動學習文件分類的經驗與知識，經過訓練後，即會歸納出文件分類規則，就可以對未知類別的新文件（測試資料）讓文件分類系統自動分類。自動化文件分類已被廣泛地應用在不同的領域中，其涵蓋的範疇包括文件管理（Larkey 1999; Li et al. 2006; Zhang & Zhou 2006）、文件過濾（Drucker et al. 1999; Tauritz et al. 2000）、語意辨識（Escudero et al. 2000）、以及網頁文件分類（Denoyer & Gallinari 2004; Zhang & Zhou 2007; Pietramala et al. 2008）等。

在文件分類中，多數研究使用特徵字詞所形成的「特徵向量模式」表達文件，「字庫法」是最常被用以表示文件屬性的方法之一，然而這樣的文件表達方式，容易導致文件的特徵維度相當高。對於分類演算法來說，欲處理這類的資料，將是很大的阻礙，這種現象稱之為「維度詛咒（Curse of Dimensionality）」，其意指當資料的維度增加後，想要找到所有可能屬性組合的困難度將逐漸提高，整個資料空間就會成幾何倍數增加。再者，高維度的文件資料中，多數的特徵對於分類並

沒有幫助，甚至對於分類演算法而言會是雜訊，反而容易導致分類錯誤。因此，我們必須從原始文件中選取出部分關鍵字詞集合進行特徵維度的縮減，進而能夠有效地改善分類演算法的效能（Aghdam et al. 2009）。過去已有許多學者採用各種特徵選取方法進行文件分類之研究，多數是以統計方法為基礎，計算字詞出現頻率來進行篩選，其中較廣泛使用的方法，例如：文件頻率（Document Frequency, DF）（Xia et al. 2009）、資訊獲利率（Information Gain, IG）（Lee & Lee 2006）、互訊息（Mutual Information, MI）（Lu et al. 2009），以及卡方統計量（CHI）（Zheng et al. 2004）等。而在（Yang& Pedersen 1997）研究中，針對上述幾種特徵選取方法進行比較，發現各方法所適用的情況皆不盡相同。換言之，目前已有許多特徵選取方法被提出，但並沒有任一特徵選取方法能夠適用於多數情形中（Cheatham & Rizki 2006）。

目前基因演算法（Genetic Algorithm, GA）已被廣泛地應用於搜尋空間大的特徵選取研究，由於它具備解決困難問題最佳化的能力，所以在特徵選取研究中近來備受關注（ElAlami 2009; Sikora & Piramuthu 2007）。因此，本研究欲採用 GA 能夠求解搜尋空間大及計算複雜度高之問題的特性，來處理文件分類此一高維度的問題，進行分類演算法的訓練文件特徵選取之最佳化。而特徵選取簡單來說就是從原始特徵集合中，選擇出最佳化的特徵子集合給予演算法進行運算的一項工作，該如何進行最佳化通常是取決於應用問題之需求，因此，我們即可透過 GA 染色體於文件特徵字詞編碼方式的設計，讓基因不斷的演化、突變與交配不停地產生新基因，且淘汰不良的基因，有效地找尋出全域最佳解。然而，過去文獻已有類似的研究（Bai et al. 2007; Chen et al. 2013; Chou et al. 2007; Martin-Bautista&Vila1999; Özel2011; Qi & Sun 2004; Uğuz2011; Yang et al. 1998），但他們的採用方式，仍然設定固定的 GA 參數，忽略了當資料集呈現不同樣式時，應適性調整族群大小、演化世代數、交配率與突變率等參數於特徵選取的效能影響，因此，如何建構一個尋找最佳特徵集合且設定合適的 GA 參數，成為本研究的探討重點。

本研究的目的在於提出一個以 GA 為基礎進行特徵選取之文件分類方法（GA-based Feature Selection, GAFS），透過機器學習的方式讓電腦代替人工分類大量的文件資料，並且能夠解決文件高維度字詞特徵的問題，本方法主要是以 GA 為基礎建立分類模式，透過 GA 染色體於文件特徵向量編碼方式的設計，讓各分類器能夠達到較佳的分類效能，演化出問題的最佳解答，進而建立一套分類效果佳的自動化文件分類模式。

本文結構如下，論文第貳章回顧文件分類相關研究；第參章詳細說明本研究提出之文件分類方法 GAFS；第肆章討論相關實驗結果；最後綜合探討研究結果，並提出未來研究方向。

貳、文獻探討

一、文件分類流程

以字詞為基礎的文件分類主要在分析文件的內容，並找出具有代表性或類別區隔力的特徵字詞，做為文件分類的依據。一般來說，以字詞為基礎的文件分類流程主要由三個階段組成，包括(1)特徵文字萃取與選擇、(2)文件表達，以及(3)文件分類（Apté et al. 1994），以下將分別詳細敘述各階段之方法。

在特徵文字萃取與選擇階段，首先會從文件中萃取出特徵文字（例如：名詞），再將其中無意義的字（Stopwords）排除，最後形成一組字詞集合（Term set）。接著，字詞集合中的每個字詞將被一一衡量其在此文件集合中的重要性或代表性。最後，前 k 個較具重要性或代表性的字詞將被選取做為下一階段用以重新表達文件的特徵文字。目前常見的字詞衡量方法多數以統計的方法來進行衡量，例如：文件頻率（Document Frequency, DF）（Yang & Pedersen 1997）、字詞頻率與反向文件頻率（Term Frequency Inverse Document Frequency, TF×IDF）（Salton & McGill 1983）、資訊獲利率（Information Gain, IG）（Quinlan 1986）、互訊息（Mutual Information, MI）（Church& Hanks 1990），以及卡方統計量（Chi-statistic）（Schütze et al. 1995）等。依據文件分類領域中過去研究顯示，多數研究皆採用 TF×IDF 作為特徵選取方法（Damerau et al. 2004; Lu et al. 2010; Luo et al. 2011; Sun et al. 2009; Teichert & Mittermayer 2002; Wei et al. 2011; Zhang et al. 2011），在文件前處理階段事先過濾特徵字詞，有效地區分出具代表性的特徵。其優點是運算簡單不複雜且效率佳，因而獲得許多研究者的青睞，故本研究也使用 TF×IDF 特徵選取方法。字詞頻率（Term Frequency, TF）為某一個文字在所有文件中出現的次數，出現次數愈多則視該文字愈具重要性。由於字詞頻率的方法只求文字出現的次數，因而可能會將字詞頻率高但分散在多數文件中，不具有區隔力的文字視為重要性字詞；或是將字詞頻率少但集中在少數文件中，而可能具有代表性的文字視為不具重要性字詞。因此 TF×IDF 的方法即以字詞頻率為基礎，結合文件頻率（Document Frequency）的特性，將出現在少數文件中的字詞，視為可能具有區隔能力的文字並納入考量，其定義如式(1)：

$$tfidf(t_j) = \sum_{1}^{n} tf(d_i, t_j) \times \log \frac{|D|}{df(t_j)} \tag{1}$$

$tf(d_i, t_j)$ 為文字 t_j 在文件 d_i 中出現的次數；$df(t_j)$ 為有出現文字 t_j 的文件數；$|D|$為所有文件總數；根據字詞頻率與文件頻率的特性及可知，字詞愈常出現表示它愈具代表性，字詞出現在愈多文件中則表示它的鑑別度愈差，若文字 t_j 的 TF×IDF

值愈高，則表示該文字不但在某文件中出現的次數高，且亦能具有類別代表性，因此該文字能視為具重要性的特徵文字。

接者，在文件表達階段，主要任務在於以前一階段選取的特徵文字將文件進行重新表達，並將文件轉換成 k 個維度的字詞向量空間。常見的文件表達方法有三種，包括二元法，即表達文件中有或沒有某特徵文字，並以 0 或 1 來表達該特徵文字的向量值；頻率法，以特徵文字在文件中出現的頻率來做為特徵文字的向量值；以及權重法，亦即以文字在文件集合中的重要程度做為特徵文字的向量值（例如 TF×IDF、卡方統計量）。

最後，在文件分類階段，則利用分類演算法將新進文件分門別類至適當的文件類別中。過去研究所提出之分類演算法從分類過程的角度大致可分為兩種，第一種方式主要利用人工智慧的自動學習方法，從訓練文件資料集合中分析出文件分類的法則，再進一步利用該分類法則來進行新進文件的類別預測，常見的方法包括「類神經網路」（Trappey et al. 2006; Yu et al. 2008; Zhang & Zhou 2006）、「決策樹」（Apté et al. 1994; Damerau et al. 2004; Johnson et al. 2002）、「貝氏網路」（Chen et al. 2009; Lu et al. 2010; Wei et al. 2011），以及「支援向量機」（Luo et al. 2011; Sun et al. 2009; Zhang et al. 2011）等。另一種方式則不透過分類法則的學習，直接比較新進文件與現有類別中文件的相似度來進行分類，此方法中以「k 個最鄰近法」（Ambert & Cohen 2012; Tan 2006）較具代表性。

二、文件分類研究概況

文件分類相關之過去研究彙整如表 1 所示。目前文件分類的研究已為數眾多，從表 1 我們首先可以得知，在特徵選取方法中，多數研究皆是採用以統計為基礎的 TF×IDF，它的優點是運算簡單不複雜且效率佳，同時能夠有效地區分出具代表性的特徵字詞，因此獲得許多研究的青睞。再者，在各研究使用的資料集中，以 Reuters 新聞文件資料集被最多研究採用，Newsgroups 次之，也有少數研究是針對中文文件、專利文件 WIPO 以及網頁 WebKB 做分類。在分類技術部分，從過去的研究中可以發現，各種分類技術無論是過去或是近期都還是都有研究者投入文件分類的研究，主要是因為大部分的學者認為各分類技術在不同的情況下會有不一樣的表現，例如：不同性質的文件資料。因此，為數眾多的研究因應而生，故我們將採用各類分類技術進行比較，並從中評估本研究是否能改善各分類技術之分類效能。

表 1：文件分類研究文獻彙整

Type	Research	Feature Selection	Dataset
Decision Tree, DT	(Apté et al. 1994)	Binary, TF	1. Reuters
	(Johnson et al. 2002)	TF	1. Reuters-21578
	(Damerau et al. 2004)	Binary, TF, TF×IDF	1. Reuters
Support Vector Machine, SVM	(Sun et al. 2009)	TF×IDF	1. Reuters-21578 2. 20 Newsgroups 3. WebKB
	(Zhang et al. 2011)	TF×IDF, Multi-words, LSI	1. Chinese documents 2. English documents
	(Luo et al. 2011)	TF, TF×IDF	1. Reuters-21578 2. 20 Newsgroups 3. WebKB
Artificial Neural Network, ANN	(Trappey et al. 2006)	Correlation analysis	1. Patent documents(WIPO)
	(Yu et al. 2008)	Latent Semantic analysis	2. 20 Newsgroups
Naïve Bayesian, NB	(Chen et al. 2009)	MOR, CDM	1. Reuters-21578 2. Chinese text corpus
	(Lu et al. 2010)	TF×IDF	1. Chinese text (Electronic 2. Theses and Dissertations)
	(Wei et al. 2011)	TF×IDF	1. Literature 2. News Press 3. three document corpora
k Nearest Neighbor, kNN	(Tan 2006)	IG	1. Reuter-21578 2. TDT-5
	(Ambert et al. 2012)	IG	1. BioCreative II.5 FEBS
Committees	(Fall et al. 2003)	IG	1. Patent documents(WIPO)
	(Teichert et al. 2002)	TF×IDF	1. Patent documents

三、GA 於文件分類研究概況

進行文件分類前，通常會把文件以特徵向量表示，透過字詞與文件構成特徵向量空間，而眾多字詞所構成的集合即可視為文件的特徵。由此可知，文件特徵

是一高維度的向量，由許多字詞所構成，然而傳統機器學習分類技術要對上述高維度特徵向量文件進行分類時，將會是一件高負荷且複雜困難的工作，所以透過特徵選取減少文件維度已是一件無可避免的前置作業。

過去文件分類的研究中已有許多特徵選取方法被提出，多數研究是以統計方法為基礎，計算文件字詞出現頻率來進行篩選的動作，最常使用的方法如：TF×IDF（Yang et al. 1998）、Information Gain（IG）（Lee et al. 2006; Uğuz 2011）、Mutual Information（MI）（Lu et al. 2009），以及卡方統計量（Chi-square statistics, CHI）（Zheng et al. 2004）等。由於以統計為基礎的特徵選取方法僅能計算出文件中各個字詞的重要程度（權重值），並無法直接從中計算得出最具代表性的特徵集合，也就是說，此類研究皆是由研究者自行挑選前 k 個權重最高的字詞當作文件特徵向量，是比較主觀的作法。

近年來有學者提出以 GA 為基礎之特徵選取方法於文件分類問題上。學者 Martin-Bautista et al.（1999）在 GA 特徵選取於文字探勘領域的研究調查中提及，過去 GA 尚未於文字探勘領域中被深入研究解決特徵選取問題。學者 Yang 等（1998）使用 TF×IDF 進行特徵選取，接著再利用 GA 針對 TF×IDF 選取後的特徵進行特徵子集合的篩選，但並未詳述 GA 運作方法，其參考文獻中亦無說明。學者 Qi 與 Sun（2004）結合 GA 與 *k*-means 分群演算法來發展自動化網頁文件分類器，此研究首先由網頁萃取出關鍵字詞，並給予初始權重，接著透過 GA 演化來進行權重最佳化的動作，因此權重較低的字詞可以被過濾掉，而達到屬性刪減的目的。學者 Bai 等（2007）首先使用約略集合（Rough Sets）來降低文件特徵個數，接著透過支援向量機進行文件分類並使用 GA 來調整參數，實驗結果顯示，其所設計的方法相較於傳統支援向量機能有更高的分類正確率。學者 Chou 等（2007）利用 GA 演化出特徵選取方法的最佳門檻值，具體來說，該研究首先透過四種統計指標來計算字詞的重要性，接著再利用 GA 來找出四種指標之最佳化權重，因此，每個字詞的重要性可透過四個統計指標加權計算求得，使用者可以設定一門檻值，若字詞重要性大於門檻值，則認為該字詞具代表性；反之，則過濾該字詞。學者 Özel（2011）利用 GA 找出網頁文件之特徵集合權重，該研究首先利用 HTML 標籤來識別出網頁標題、標頭、粗體字、斜體字等關鍵字詞，接著透過 GA 來進行字詞權重的最佳化，已建構出最佳的網頁文件分類器，該研究透過真陽率（True Positive Rate）與真陰率（True Negative Rate）的乘積做為適應函數值。學者 Uğuz（2011）亦提出與上述類似的作法來建立最佳化的網頁文件分類系統，但在其所提出的系統中，HTML 標籤與網頁內的文字內容同時用來萃取文件特徵，之後再透過 GA 找出文件特徵集合之權重。Chen 等（2013）結合了混沌最佳化演算法（Chaotic Optimization Algorithm, COA）和 GA，提出了一個 Chaos Genetic Feature Selection Optimization（CGFSO）的演算法，此演算法主要是從文件的特徵中，尋

找當中最佳化的子集合，而會結合 COA 的原因，是因為 GA 在實務上有提早收斂和演化過程中低搜尋效率的問題，但會結合 GA 的原因，是 COA 在搜尋過程中，容易受困於區域最佳化的空間。

由上可知，過去透過 GA 進行文件特徵選取的研究上大致分為兩類：二元編碼型 GA 與權重型 GA。首先，二元編碼型 GA 的研究多數是將 GA 染色體中的基因作為文件候選關鍵字詞的編碼，染色體中的基因值代表候選關鍵字詞有無被選擇作為特徵字詞（Cheatham & Rizki 2006; Zhang et al. 2005）；而權重型 GA 的做法則是賦予每個字詞不同的權重，並透過遺傳運算進行權重最佳化的動作。從上述 GA 於文件分類領域的研究中，我們可以得知其主要都是用於解決文件特徵選取的問題，透過 GA 染色體基因的編碼設計不斷地進行演化，達成文件維度縮減的目的，同時也期望能夠求解出文件特徵向量集合的最佳解。然而，過去絕大部份的研究對於 GA 運算，均採用固定參數，意即忽略了 GA 運算中族群大小、演化世代數、交配率與突變率等參數對於特徵選取的效能影響，由於遺傳運算需耗費大量的運算資源，如何在找出最佳字詞特徵集合與設定適當的 GA 參數以建構最佳分類器，是值得進一步探討的研究議題。

參、研究方法

本章將介紹本研究所提出的自動化文件分類方法，以基因演算法為基礎進行特徵選取（GA-based Feature Selection, GAFS）。GAFS 流程如圖 1 所示，主要可以分為三個部份，首先為文件前處理階段，包含字詞擷取、過濾與還原、以及文件表達等步驟；其次為基因演算法於分類器特徵選取最佳化；最後則是分類模式之建構。以下將詳細說明各階段中的流程與任務，依序為「文件前處理」、「基因演算法之最佳化」、以及「分類模式」。

一、文件前處理

在建立分類模式前，必須先將資料集做前處理的動作後，才能作為文件分類模式的輸入資料。文章的結構通常是數個段落組合而成，而文章各段落是由一些字詞所構成，描述一篇文章最基本單位就是字詞。因此，我們從文章中把一些有意義的字詞聚集起來，形成足以代表文章的特徵向量，使文件資料集從人類可閱讀的原始型態，表示成分類器程式能解讀的特徵向量。資料前處理的過程包括停用詞處理、字詞還原及文件表達等三個步驟，以下將依序分三部份說明，表 2 為本研究相關變數符號定義。

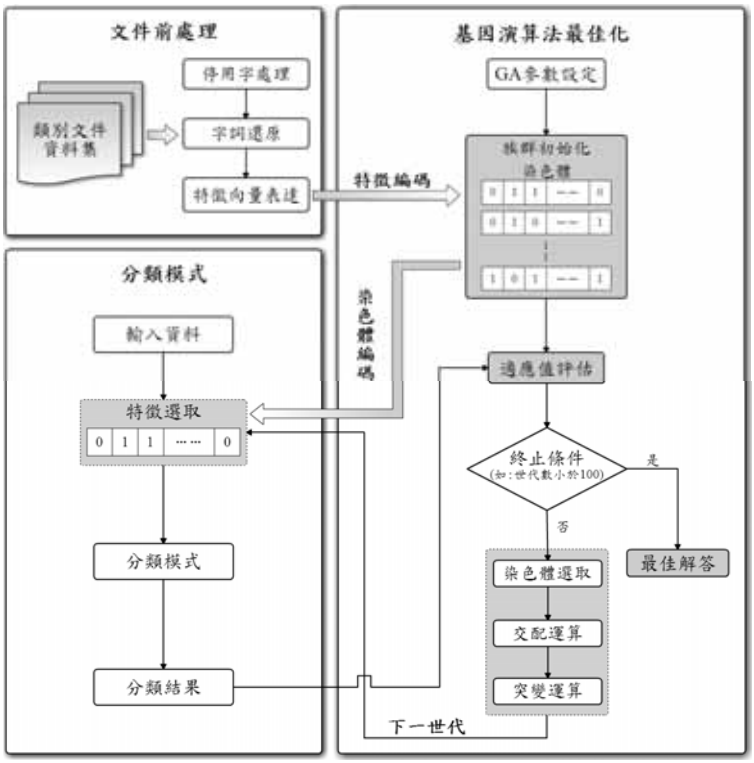

圖 1：GAFS 流程圖

表 2：變數符號定義表

符號	定義與說明
T	所有字詞之集合。
t_i	所有字詞中，第 i 個字詞。
D	所有文件之集合。
d_j	所有文件中，第 j 份文件。
w_{ij}	字詞 t_i 在文件 d_j 的重要程度（權重值）。
C	所有文件類別之集合。
c_k	所有文件類別中，第 k 個類別。

（一）停用詞處理（Stopwords processing）

將原始文件拆解成字詞是取得文章語意的一種典型方法，但在英文文章中，存在著許多定冠詞、介係詞，以及連接詞等輔助性的字詞，這些字詞由於出現頻繁，本身又沒有甚麼意義，因此我們會把這些不重要的字詞過濾掉，這些被忽略的字詞，就叫「停用詞」(Stop words)，例如”the”、”a”、”an”、”to”、”of”，以及”and”

等。它們在每篇文章中，都大量且重複地出現，不僅不具文章代表性，也沒有鑑別不同文章的能力。因此，本研究採用 SMART 資訊檢索系統（Buckley 1985; Salton 1971）524 個停用字詞（資料源自於 ftp://ftp.cs.cornell.edu/pub/smart/english.stop），將資料集中的這些字詞過濾掉，確保索引原始文件資料集的特徵字詞庫是良好且有效用的。

（二）字詞還原（Stemming）

英文都是由某個基本型態的詞語，例如：act，經由某個文法規則的轉換，而變成 acting、acted、action、active 等其他動詞、動詞過去式、名詞、形容詞等類型。本研究採用（Porter 1980）字詞還原演算法，經由「Stemming」的處理，把這類文字的各種型態，依照文法的規則，轉換成文法上最原始的型態。

（三）文件表達（Document representation）

本研究採用普遍使用的特徵選取方式 TF×IDF 來計算各個特徵字詞之權重。假設文件資料集 $D=\{d_1,d_2,\ldots,d_i,\ldots,d_{|D|}\}$ 為整個原始文件資料集，經過上述步驟後，即可得到所有特徵字詞集合 $T=\{t_1,t_2,\ldots,t_j,\ldots,t_{|T|}\}$，這些字詞在資料集中至少都出現過一次。透過 TF×IDF 計算每篇文件中各特徵字詞的權重，每份文件皆可用一特徵向量來表示，文件 d_i 即可表示為 $\vec{d}_i=\{w_{i1},wi_2,\ldots,w_{ij},\ldots,w_{i|T|}\}$， w_{ij} 代表字詞 t_j 在文件 d_i 中的權重大小。另外 $C=\{c_1,c_2,\ldots,c_k,\ldots,c_{|C|}\}$ 是所有類別集合，每份文件皆會存在一既有已知的類別，如文件 d_i 之類別標籤為 $c_{k_i}(1\le k_i\le|C|)$，其代表的意義是文件 d_i 所屬的類別為 c_{k_i}。經過上述所有前處理的步驟後，可以得到一所有文件（D）與字詞（T）及類別標籤（C）的矩陣，如表 3。

表 3：訓練文件資料集之字詞特徵向量權重與類別

文件	字詞						文件類別
	t_1	t_2	…	t_j	…	$t_{\|T\|}$	
d_1	w_{11}	w_{12}	…	w_{1j}	…	$w_{1\|T\|}$	c_{k_1}
d_2	w_{21}	w_{22}	…	w_{2j}	…	$w_{2\|T\|}$	c_{k_2}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
d_i	w_{i1}	w_{i2}	…	w_{ij}	…	$w_{i\|T\|}$	c_{k_i}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
$d_{\|D\|}$	$w_{\|D\|1}$	$w_{\|D\|2}$	…	$w_{\|D\|j}$	…	$w_{\|D\|\|T\|}$	$c_{k_{\|D\|}}$

註：c_{k_i} 代表 d_i 的類別，$k_i=\{1,2,...,|C|\}$，$|C|$為所有文件類別總數。

二、基因演算法之最佳化

（一）基因演算法流程

本研究 GA 之運作流程如圖 1 的右半部所示，首先針對族群中的染色體進行初始化（Initialization）的運算，此階段每個染色體可以定義出被挑選的字詞集合，所有文件可依染色體建立相對應的文件向量。接著，我們針對初始族群進行染色體適應值（Fitness value）的計算，此階段將依不同的分類技術來建構分類器，並進行正確率評估，有關適應函數的設計將於第（四）小節中說明。再取得染色體的適應值後，我們可依所設定的中止條件決定結束演化或是繼續進行演化，若為中止演化則結束學習程序，若為繼續進行演化則開始進行遺傳演算，包括挑選（Selection）、交配（Crossover）及突變（Mutation），其皆是依據所設定之機率值（交配機率與突變機率）來決定是否需要執行，在交配與突變結束後，再回到適應函數值的計算，依此類推，直到達成終止條件，挑選出最佳解。

（二）染色體編碼

本研究提出之 GAFS 方法其染色體編碼方式採用二元編碼方式，將染色體中的基因以二元型態進行編碼，文件經前處理後之各特徵字詞皆要被編碼至染色體基因中，染色體中的基因即代表各特徵字詞，而基因值編碼內容僅有兩種即為 $\{0,1\}$。將被選取進行訓練的特徵，其所對應的基因將被編碼為 1；反之，不被選取進行訓練的特徵，其基因編碼為 0。如圖 2 所示，假設文件有十個特徵字詞 $(|T|=10)$，那麼染色體長度即為 10，圖中第一個基因(t_1)編碼為 0，代表t_1特徵字詞將被捨棄不進行訓練；而第九個基因(t_9)編碼為 1，代表t_9特徵字詞將被選取進行訓練。

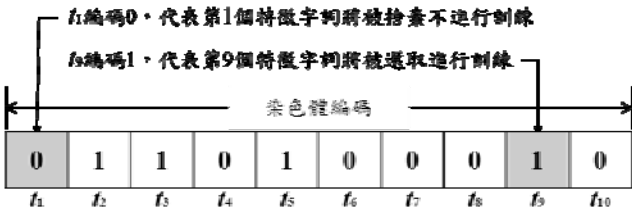

圖 2：GAFS 之染色體編碼

（三）族群初始化（Initialize population）

進行族群初始化的動作是為了產生染色體族群，在 GA 演化的過程中所產生的多對染色體稱為母體，每個演化世代都是由母體中成對染色體依基因演化產生下個世代的母體。一般而言，第一代母體稱之為初始母體，首先須決定族群大小，

接著制定初始化策略，再來依據染色體編碼規則及初始化策略來產生初始化的族群染色體。本研究初始化策略採用隨機的方式，事先定義基因值的範圍，接著透過隨機函數決定族群中染色體初始值，其優點就是可以增加初始母體的多樣化，使得之後的演化運算可以在較大的搜尋空間中搜尋，隨機產生染色體族群公式如式(2)：

$$\begin{aligned} &Population[\,i\,][\,j\,] \;=\; Rand[0 \text{ or } 1] \\ &i \;=\; 1,\ 2,\ \ldots,\ P_{size}\ ;\ j \;=\; 1,\ 2,\ \ldots,\ Chromosome_{size} \end{aligned} \tag{2}$$

Population 為族群，*i* 與 *j* 代表族群中第 *i* 條染色體的第 *j* 個基因；*Rand* 是隨機函數；P_{size} 為族群大小，$Chromosome_{size}$ 為染色體大小。

（四）適應函數設計（Fitnessfunction）

適應函數設計主要作用於搜尋最佳解，它決定每個族群中染色體適應環境的能力，合適的適應度函數可以將染色體的優劣有效地比較出來，故適應度函數通常會依照設計者對問題求解的要求而有所不同。在監督式分類問題中，多半最佳解具有特定的目標可供計算，例如以準確度作為適應函數。本研究中，我們首先計算依每個染色體所建立分類器之分類準確度（Accuracy），計算方式如圖 3 所示，其中，*predict*(d_i)代表文件 d_i 被預測的類別；*actual*(d_i)代表文件 d_i 實際的類別。

```
For each document di in D
   If predict(di)=actual(di) then
AccurateSum=AccurateSum + 1
End loop
Accuracy = AccurateSum / |D|
```

圖 3：計算分類正確率之虛擬碼

此外，為了避免染色體間準確度過於相近，而無法有效地區分出適應程度，所以本研究進而依據各染色體之分類準確度（Accuracy）由高至低進行排序，並依序給予其排序值（AccuracyRank）（由 1 至 P_{size}），最後，該染色體之適應函數值可以透過其本身的 AccuracyRank 計算求得，如式(3)所示：

$$FitnessValue = P_{size} - AccuracyRank + 1 \tag{3}$$

式(3)中，*FitnessValue* 代表染色體適應值。

（五）複製與選擇運算（Reproduction and Selection operator）

適應函數設計完後，我們即可計算出各染色體的適應值，接著就是針對族群中的染色體進行複製的步驟，複製的步驟主要是希望族群中表現較佳的染色體可以被複製較多的數量，透過複製的運算，新的子代族群可以保留母代中表現較好的染色體，藉以加速族群收斂的速度。複製流程如圖 4 所示，主要可分為兩部分，首先是複製階段，為了確保子代中表現較佳的染色體不會因為世代演化而被淘汰，所以新子代族群中前 *i* 條染色體是直接由母代複製而來，如此便能把母代中表現較佳的染色體保留至下個子代中；而另一部分則是選擇階段，此階段會使用輪盤式選擇法，依染色體的適應程度分配其被挑選的機率，適應程度越佳的染色體被挑選的機率越高，而挑選出合適的染色體後，將會再進行交配與突變演化運算。由於複製可以加速族群的收斂，所以複製的策略需要審慎的選擇，若複製過多的母代族中表現較佳的染色體，可能會造成染色體族群過早收斂，而複製太少的母代染色體可能會造成母代優良基因過少而收斂過慢。

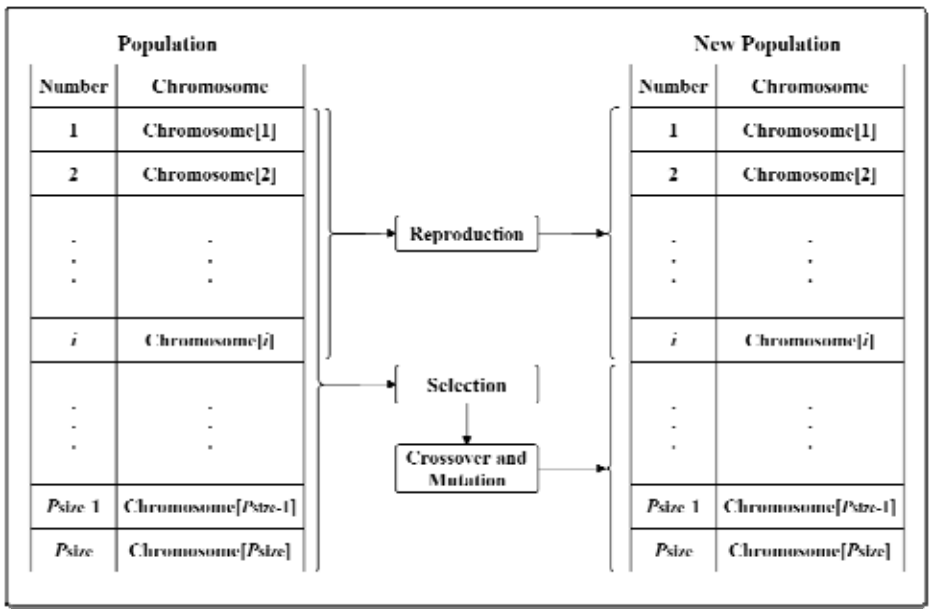

圖 4：基因演算法學習流程圖

本研究採用輪盤複製與選擇法，此方法是透過適應函數判斷各組染色體的優劣，若適應函數值較高者，則其被保留的機率將隨之較大；反之，若染色體適應度函數值較低者，則其被保留的機率也較小。當確定所有染色體的優劣之後，便透過複製將較好的染色體予以保留。其步驟如下：

步驟一：將排序好的母代族群中的染色體依下式(計算適應函數比例值：

$$TotalFitness = \sum_{i=1}^{P_{size}} Population[i].FitnessValue,$$
$$Population[j].FitnessRate = \sum_{i=1}^{j} \frac{Population[i].FitnessValue}{TotalFitness} \quad (4)$$
$$where\ j = 1, 2, ..., P_{size}$$

其中，*TotalFitness* 代表族群中所有染色體適應值的總和；*Population*[*i*].*FitnessValue* 代表族群中第 *i* 條染色體的適應函數值；*Population*[*j*].*FitnessRate* 代表族群中第 *j* 條染色體的適應函數值占總適應函數值的累進比例值；P_{size} 為族群大小，即一族群中全部染色體的數量。由式(4)計算後，即可得出一族群中各染色體適應值所佔的比重。

步驟二：以式(5)隨機產生一隨機值，其介於 0 到 1 之間。

$$RandValue = Rand[0, 1] \quad (5)$$

步驟三：依式(判斷應該選擇複製到子代族群的染色體。

$$SelectedPopulation_i = Population[k],$$
$$if\ fitnessRate_{k-1} < RandValue < fitnessRate_k \quad (6)$$
$$where\ i = 1, 2, ..., P_{size}$$

其中，$SelectedPopulation_i$ 代表子代族群中第 i 條染色體；Population[k]代表母代族群中第 k 條染色體。

步驟四：重複步驟二與三，直到子代染色體數達到族群大小為止。

（六）交配與突變運算（Crossover and Mutation operators）

過去研究中提出的交配運算類型有單點交配、雙點交配、及定點交配等方法，而許多研究顯示雙點交配較優於單點交配（Beasley et al. 1993），只有當族群中染色體的適應函數值接近收斂狀態時，雙點交配的效能才會比單點交配的結果差（Beasley et al. 1993; Spears et al. 1993）。因此，本研究採用雙點交配策略，運算步驟如下：

步驟一：以式(7)產生兩個交配點

$$CrossoverSite_i = Rand[1, Chromosome_{size}],$$
$$where\ i = 1\ and\ 2. \quad (7)$$

步驟二：接著將所挑出來的母代染色體中第一條染色體兩交配點間基因，與第二條染色體兩交配點間基因互換，即完成雙點交配運算。

經由複製與交配的運算後，子代族群中的染色體可以經過互相交換母代而具有較佳的適應程度，但卻沒有新的基因值加入染色體中，容易陷入局部最佳解的情況，因此，突變運算可讓 GA 在演化過程中跳脫區域最佳解，使搜尋空間不被侷限住，因而能逼近全域最佳解。本研究之突變運算主要是從選取的染色體中，根據突變機率決定基因是否要進行突變，運算如式(8)：

$$\begin{aligned} &If\ \ MutationRate > RandValue_j[0,1] \\ &\quad Population[i][j] = 1 \quad if\ \ Population[i][j] = 0 \\ &\qquad\qquad\qquad = 0 \quad otherwise. \\ &where\ \ j = 1, 2, \ldots, Chromosome_{size} \end{aligned} \tag{8}$$

MutationRate 為突變率；$RandValue_j$ 代表染色體中第 *j* 個基因位置所產生的隨機值，其介於 0 到 1 之間；$Chromosome_{size}$ 染色體大小。各基因位置皆會產生一隨機值，若小於突變率，則該位置基因值就會進行突變，基因值為 0 則突變為 1；反之，則突變為 0。

三、分類模式之建構

本研究採用六種不同的分類器模式，其中包含貝氏分類器（Naïve Bayesian Classifier, NB）（Rish 2001）、決策樹（Decision Tree, DT）（Quinlan 1993）、分類迴歸樹（Classification And Regression Tree, CART）（Breiman et al. 1984）、隨機森林（Random Forest, RF）（Breiman 2001）、支援向量機（Support Vector Machine, SVM）（Cortes & Vapnik 1995），以及 *k* 最近鄰居法（*k* Nearest Neighbor, *k*NN）。

肆、實驗設計與結果

本章將詳細說明本研究之實驗設計與結果，首節敘述實驗資料來源，第二節介紹實驗環境與參數調整，第三節解釋本研究之評估指標，第四節說明實驗程序與結果；最後一節探討實驗結果。

一、實驗資料

本研究使用之資料集為 WebKB 網頁文件，其是由 Carnegie Mellon University（CMU）文字探勘學習小組於全球資訊知識庫（World Wide Knowledge Base）專案中收集而成。網頁文件內容源自於 1997 年四所大學（Cornell, Texas, Washington, Wisconsin）計算機科學系，其中各網頁已透過人工手動方式共分為七個不同的類

別：課程（course）、學生（student），教師（faculty），職員（staff），部門（department），專案（project），及其他（other）。在本研究中，由於「部門」與「職員」兩類別的資料較少，因此我們捨棄上述兩個類別的網頁文件，另外在「其他」類別中的網頁文件內容差異性較大，所以我們也不採納。

實驗中，我們使用 Cardoso-Cachopo 與 Oliveira（2007）資料集切分方式，隨機將資料集依文件類別以二比一比例分切分為訓練與測試資料集，分別有訓練文件 2803 篇與測試文件 1396 篇，總共 4199 篇文件（資料源自於 http://web.ist.utl.pt/~acardoso/datasets/），各類別訓練與測試文件數量分佈如表 4 所示。

表 4：WebKB 資料集各類別文件數量

Class	Train docs	Test docs	Total docs
project	336	168	504
course	620	310	930
faculty	750	374	1124
student	1097	544	1641
Total	2803	1396	4199

二、實驗環境與參數設定

在實驗環境上，本研究採用 Intel Xeon E5620 2.4G 處理器，搭配 32.0 GB 系統記憶體，作業系統為 Microsoft Windows Server 2008 R2。演算法部分主要可分為各分類模式與基因演算法兩部分，各單一分類模式採用資料探勘軟體 WEKA 3.6.2 版開放 API 之演算法 NaïveBayes（NB）、J48（DT）、SimpleCart（CART）、RandomForest（RF）、SMOreg（SVM），以及 IBk（kNN），並以 JAVA 程式語言實作 GA 與上述分類模式結合，各分類模式參數請參閱附錄一。

調整 GA 合適的世代數（Generation Size, GS）、族群大小（Population Size, PS）、交配率（Crossover Rate, CR）及突變率（Mutation Rate, MR），我們將進行族群演化效能之評估，本階段實驗將評估各分類模式合適之參數。首先為了有效地觀察出 GA 演化效能之趨勢，我們將 GS 參數擴大至 50，使 GA 有足夠世代進行演化，同時改變 PS 參數，由 20 開始以 10 為單位遞增至 50，進一步觀察一世代中存在多少染色體能使 GA 有較佳的演化效能。再者，我們將 GS 與 PS 調整為上述實驗中得出之結果，同時改變 CR 與 MR 參數，CR 由 0.7 以 0.1 為單位遞增至 0.9，而 MR 由 0.05%開始以 0.1%為單位遞增至 0.25%，從中觀察 CR 與 MR 參數對於 GA 演化效能之影響。

GS 與 PS 參數調整結果如圖 5，橫軸為演化過程中的世代數，如上述共有 50

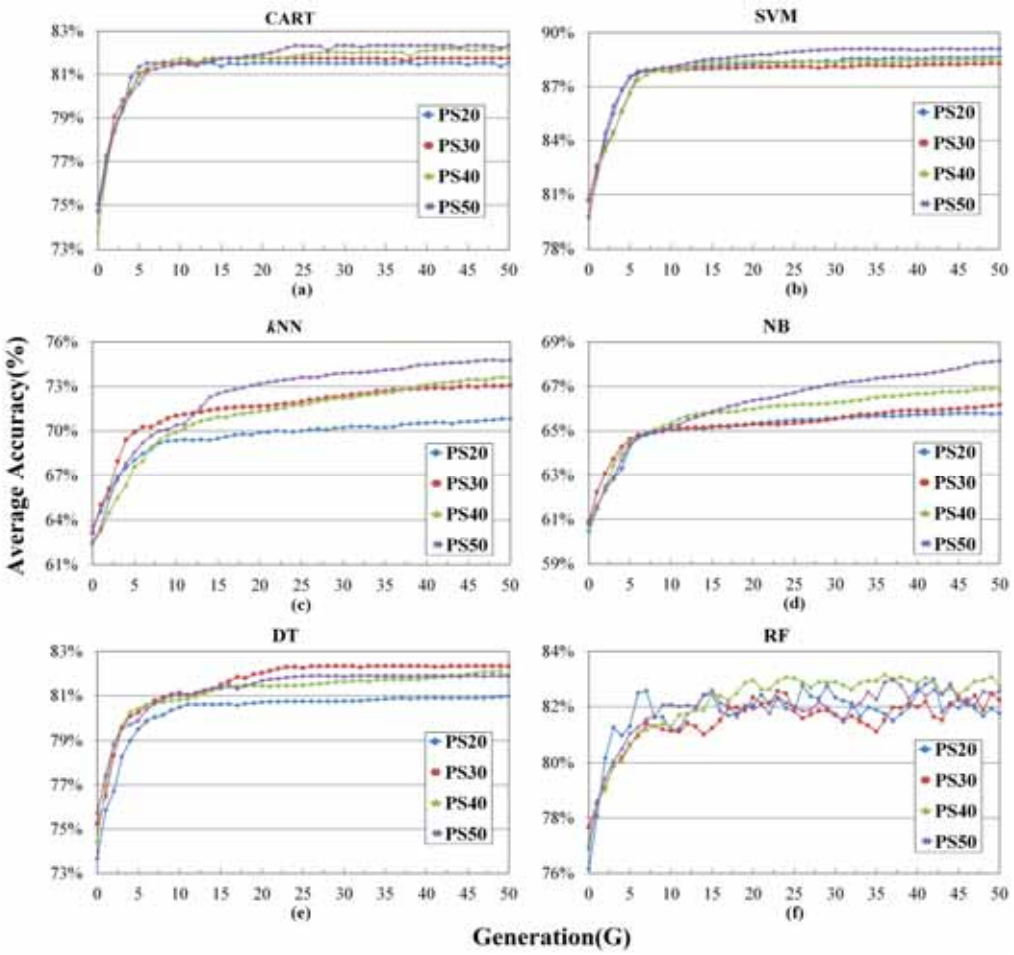

圖 5：演化世代數（GS）與族群大小（PS）對於演化效能之影響

個世代；縱軸則為一世代中所有染色體的平均準確率。圖中編號(a)至(f)分別是 CART、SVM、*k*NN、NB、DT，以及 RF 分類模式之演化結果，從圖中我們可以觀察出各分類模式演化的結果皆不盡相同，以下將依序說明圖 5 中各分類模式之參數實驗結果：在(a)與(b)中，PS50 的演化效能最佳，GS 約為 25 時各組 PS 的演化已趨向平緩；在(c)與(d)中，PS50 的演化效能最佳，GS 約為 50 時各組 PS 的演化才逐漸趨向平緩；在(e)中，PS30 的演化效能最佳，而 GS 約為 25 時各組 PS 的演化已趨向平緩；而在(f)中，雖然演化效能較不穩定，但我們還是可以觀察出 PS40 的演化效能較佳，GS 約為 35 時演化漸漸趨向平緩。最終各分類模式設置之 GS 與 PS 如表 5。

表 5：GS 與 PS 參數設定

Classifier Model	Generation Size(GS)	Population Size(PS)
CART	25	50
SVM	25	50
*k*NN	50	50
NB	50	50
DT	25	30
RF	35	40

透過上述實驗調整結果，本階段我們將分類模式之 GS 與 PS 值依表 5 設定。CR 與 MR 參數調整結果如圖 6，以下將說明圖 6 中各分類模式之參數實驗結果：

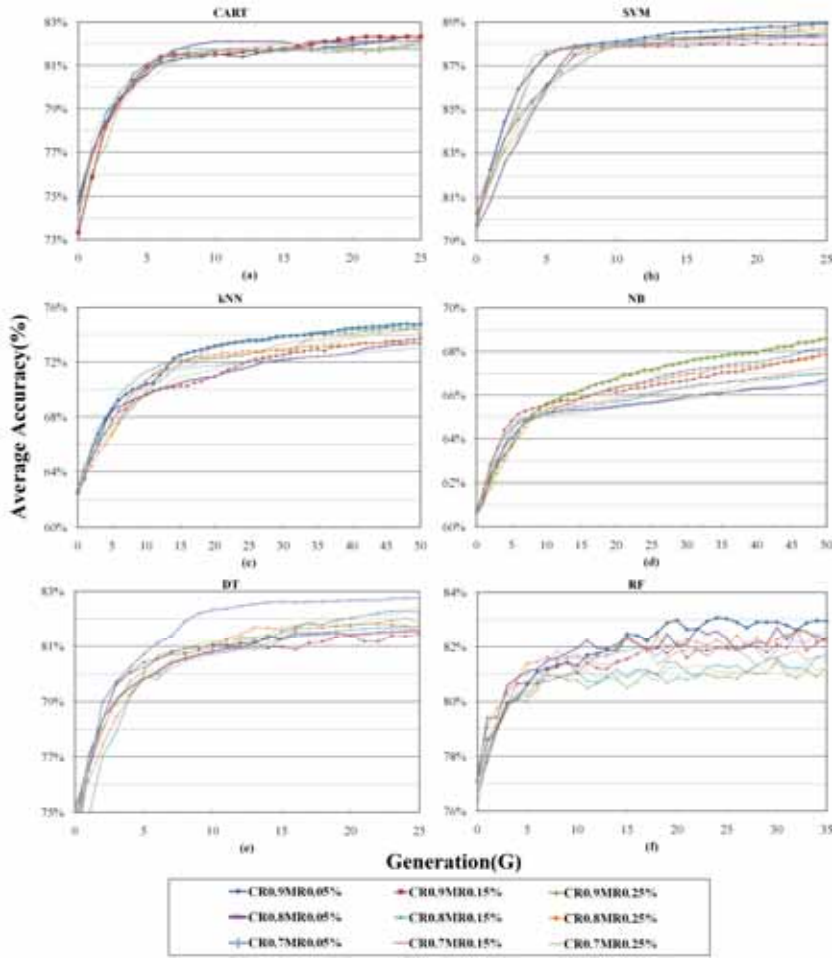

圖 6：交配率（CR）與突變率（MR）對於演化效能之影響

在(a)中，雖然各組參數演化效能差異不大，但仔細觀察可發現 CR0.9MR0.15%該組演化效能最佳；在(b)、(c)與(f)中，皆是 CR0.9MR0.05%該組的演化效能較佳；在(d)中，各組參數之演化效能有明顯地差異，其中以 CR0.9MR0.25%該組最佳；在(e)中，除了 CR0.7MR0.05%該組演化效能明顯優於其他組外，其餘參數之演化效能皆無太大差異。另外，仔細觀察各分類模式的差異，我們發現 CART、SVM 與 kNN 較不易受參數變動影響，各組數據差異性不大；而 NB、DT 與 RF 則較容易受參數變動影響演化效能。最終得出結論：雖然各分類模式演化結果不盡相同，但多數分類模式在 CR 較高時（0.9），而與 MR 較低時（0.05%），演化效能會有比較穩定地表現。最終各分類模式設置之 CR 與 MR 如表 6。

表 6：GA 參數值設定

Classifier Model	Crossover Rate(CR)	Mutation Rate(MR)
CART	0.9	0.15%
SVM	0.9	0.05%
*k*NN	0.9	0.05%
NB	0.9	0.25%
DT	0.7	0.05%
RF	0.9	0.05%

三、評估準則

建構完成之分類器必須透過客觀的評估方法來計算其效用（Effectiveness），且當文件分類之型態改變時，相對應之評估指標也會有所不同。根據 Sebastiani（2002）學者對於自動化文件分類研究一系列的整理，在單一標籤文件分類中，通常會使用 Accuracy、Precision、Recall，以及 F-measure 等評估方法，針對模式的分類效用來進行評估。其中，Accuracy 為測試資料之分類預測結果的正確率；Precision、Recall，以及 F-measure 則為資訊擷取領域中常用之評估指標。這些評估指標之計算方法如下（式(9)、式(10)、式(11)和式(12)）：

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} = \frac{\sum_{i=1}^{|C|}(TP_i + TN_i)}{\sum_{i=1}^{|C|}(TP_i + FP_i + FN_i + TN_i)} \tag{9}$$

$$Precision = \frac{TP}{TP + FP} = \frac{\sum_{i=1}^{|C|} TP_i}{\sum_{i=1}^{|C|}(TP_i + FP_i)} \tag{10}$$

$$Recall = \frac{TP}{TP + FN} = \frac{\sum_{i=1}^{|C|} TP_i}{\sum_{i=1}^{|C|} (TP_i + FN_i)} \tag{11}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{12}$$

其中，*TP*、*FP*、*TN*，以及 *FN* 之定義如表 7 與表 8 所示。表 7 代表文件類別 c_i 之混亂矩陣，橫列代表分類器預測文件是否屬於類別 c_i；縱行代表文件的真實類別是否屬於類別 c_i，因此對於 c_i 類別可能產生上述四種預測評估情形。TP_i（True Positives）代表分類器預測正確與文件的真實類別都是屬於類別 c_i；FP_i（False Positives）代表分類器預測錯誤，文件的真實類別不屬於 c_i 而分類器預測屬於類別 c_i；FN_i（False Negatives）代表分類器預測錯誤，文件的真實類別屬於 c_i 而分類器預測不屬於類別 c_i；TN_i（True Negatives）代表分類器預測正確與文件的真實類別都是不屬於類別 c_i。表 8 是根據表 7 特定文件類別的混亂矩陣所彙整之所有類別集合的混亂矩陣。

表 7：文件類別 c_i 之混亂矩陣

Text Category c_i (i=1, 2, …, $\|c\|$)		Text Category	
		YES	NO
Classifier Prediction	YES	TP_i	FP_i
	NO	FN_i	TN_i

表 8：所有文件類別集合之混亂矩陣

Category set $C = \{c_1, c_2, \ldots, c_{\|c\|}\}$		Text Category	
		YES	NO
Classifier Prediction	YES	$TP = \sum_{i=1}^{\|C\|} TP_i$	$FP = \sum_{i=1}^{\|C\|} FP_i$
	NO	$FN = \sum_{i=1}^{\|C\|} FN_i$	$TN = \sum_{i=1}^{\|C\|} TN_i$

四、實驗程序與結果

由於文件字詞數量龐大，若採用所有字詞當作分類模式之特徵進行訓練，則會耗費相當多的訓練時間。因此，在進行文件分類模式訓練前，我們首先選用字詞頻率（Term Frequency, TF）先篩選出最頻繁出現的前 500、1000、1500、2000 與 2500 等字詞作為五組不同的特徵集合，依序稱為 TF500、TF1000、TF1500、TF2000 與 TF2500。接著，評估本研究提出之 GAFS 方法與直接採用上述各特徵集合之方法（簡稱 TOTAL）。透過上述之實驗程序，我們希望從中觀察出本研究提出之 GAFS 在各單一分類器上的分類效能，並與 TOTAL 方法做比較。同時，觀察隨著特徵字詞的遞增是否對於文件分類效能有所影響。

以下將依序進行 TF500（圖 7）、TF1000（圖 8）、TF1500（圖 9）、TF2000（圖 10）與 TF2500（圖 11）實驗結果分析，各圖中橫軸代表各分類模式，由左至右分別是 CART、SVM、kNN、NB、DT，以及 RF，其中各分類模式依特徵選取方法不同又可分為 TOTAL 與 GAFS 兩組不同的結果。而縱軸部分則為分類準確度，此部分的分類準確度與先前所取的平均準確度有所不同，差別在於平均準確度是指一世代族群中各染色體的準確度加總平均；而此部分的分類準確度是指最終演化世代族群中表現最佳之染色體的準確度。

（一）TF500

TF500 特徵集之分類結果如圖 7, 由圖中可清楚地觀察出本研究所提出之 GAFS 在各分類模式中皆優於 TOTAL 方法。其中，kNN 與 NB 改善幅度較大，準確度分別由 TOTAL71.61%和 63.22%提升至 GAFS76.47%和 68.21%；而 CART 與 RF 也有不錯的表現，準確度分別由 TOTAL 80.45%和 83.06%提升至 GAFS82.77%和 85.01%；至於 SVM 雖然提升幅度較小，但其為單一分類器中分類效能最佳的分類模式。由此可證實本研究提出之 GAFS 方法確實能改善各分類模式之分類效能。

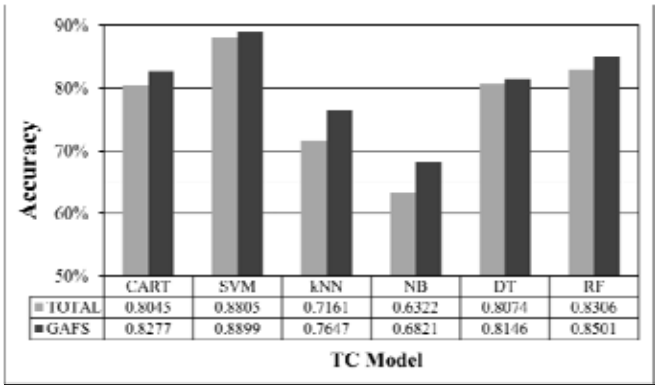

圖 7：TF500 文件分類效能之比較

（二）TF1000

TF1000 特徵集之分類結果如圖 8，GAFS 在各分類模式中皆優於 TOTAL 方法。其中，kNN 與 NB 改善幅度較大，準確度分別由 TOTAL66.55%和 64.81%提升至 GAFS74.95%和 69.08%；而 DT 與 RF 也有不錯的表現，準確度分別由 TOTAL79.87%和 80.45%提升至 GAFS82.77%和 83.64%；SVM 雖然提升幅度較小，但其為單一分類器中分類效能最佳的分類模式。

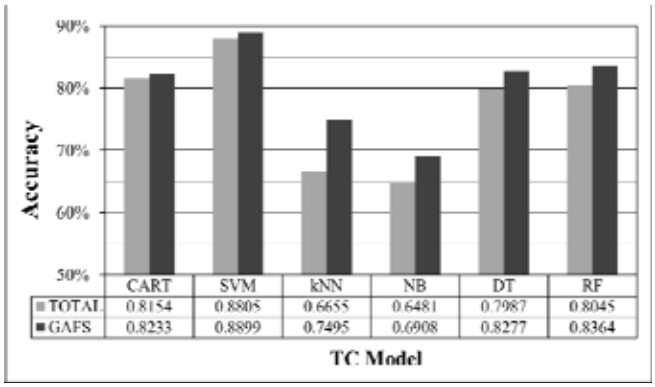

圖 8：TF1000 文件分類效能之比較

（三）TF1500

TF1500 特徵集之分類結果如圖 9，GAFS 在各分類模式中皆優於 TOTAL 方法。其中，kNN 改善幅度最為明顯，準確度由 TOTAL57.93%提升至 GAFS73.50%；而 NB、DT 與 RF 也有不錯的表現，準確度分別由 TOTAL64.74%、80.38 和 79.44%提升至 GAFS68.86%、82.33%和 83.35%；SVM 雖然提升幅度較小，但其為單一分類器中分類效能最佳的分類模式。

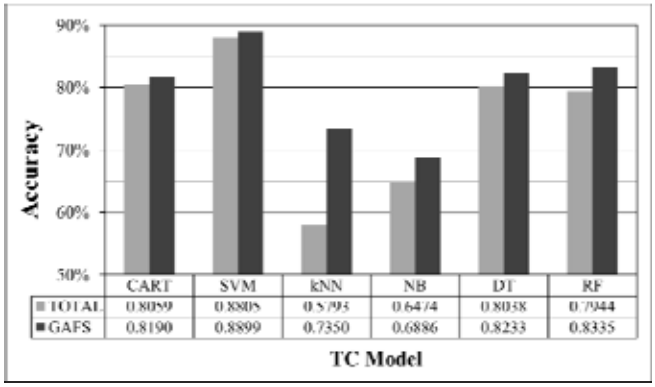

圖 9：TF1500 文件分類效能之比較

（四）TF2000

TF2000 特徵集之分類結果如圖 10，GAFS 在各分類模式中皆優於 TOTAL 方法。其中，kNN 模式分類效能有明顯地改善，準確度由 TOTAL57.93%提升至 GAFS73.50%；而 NB、DT 與 RF 也有不錯的表現，準確度分別由 TOTAL 64.74%、80.38 和 79.44%提升至 GAFS68.86%、82.33%和 83.35%；SVM 雖然提升幅度較小，但其為單一分類器中分類效能最佳的分類模式。

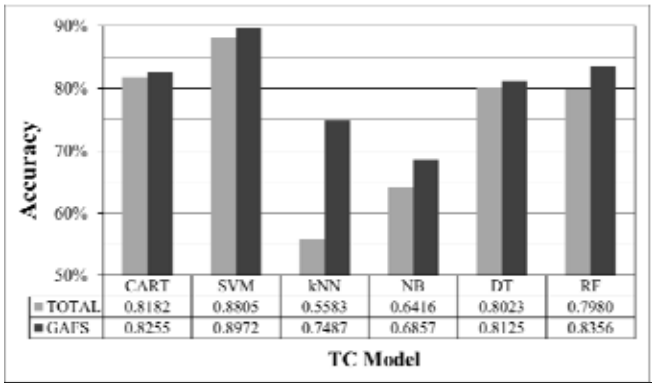

圖 10：TF2000 文件分類效能之比較

（五）TF2500

TF2500 特徵集之分類結果如圖 11，GAFS 在各分類模式中皆優於 TOTAL 方法。其中，kNN 模式分類效能有明顯地改善，準確度由 TOTAL55.83%提升至 GAFS73.71%；而 NB、DT 與 RF 也有不錯的表現，準確度分別由 TOTAL 64.81%、80.38 和 76.61%提升至 GAFS67.99%、82.11%和 82.69%；SVM 雖然提升幅度較小，但其為單一分類器中分類效能最佳的分類模式。

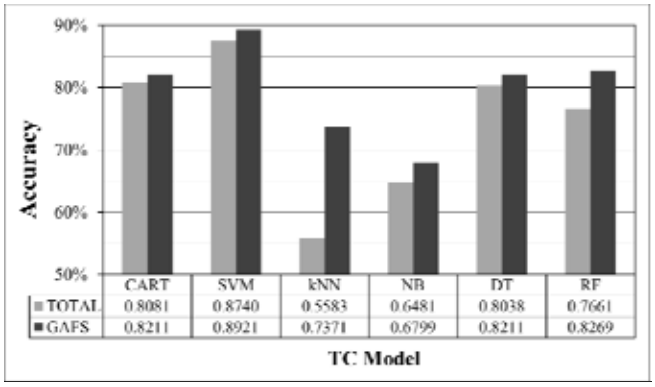

圖 11：TF2500 文件分類效能之比較

（六）綜合討論

圖 12 為特徵字詞遞增於 GAFS 與 TOTAL 方法之比較，首先，從中可以明顯地觀察出，無論是在任何分類方法或是特徵集合大小下，本研究提出之 GAFS 方法皆明顯優於 TOTAL，由此可證實 GAFS 確實能夠改善各分類模式之分類效能。另外，分析圖中各分類模式：圖 12(c)中 kNN 採用 TOTAL 方法深受特徵字詞數量影響，隨著特徵增加分類效能明顯地下降，而採用本研究提出之 GAFS 方法雖也有微幅降低，但相對而言穩定許多，在分類效能上也提升相當多，圖 12(f)中 RF 也是同理；圖 12(a)中 CART 採用 TOTAL 方法分類效能與特徵數量並無一致性，

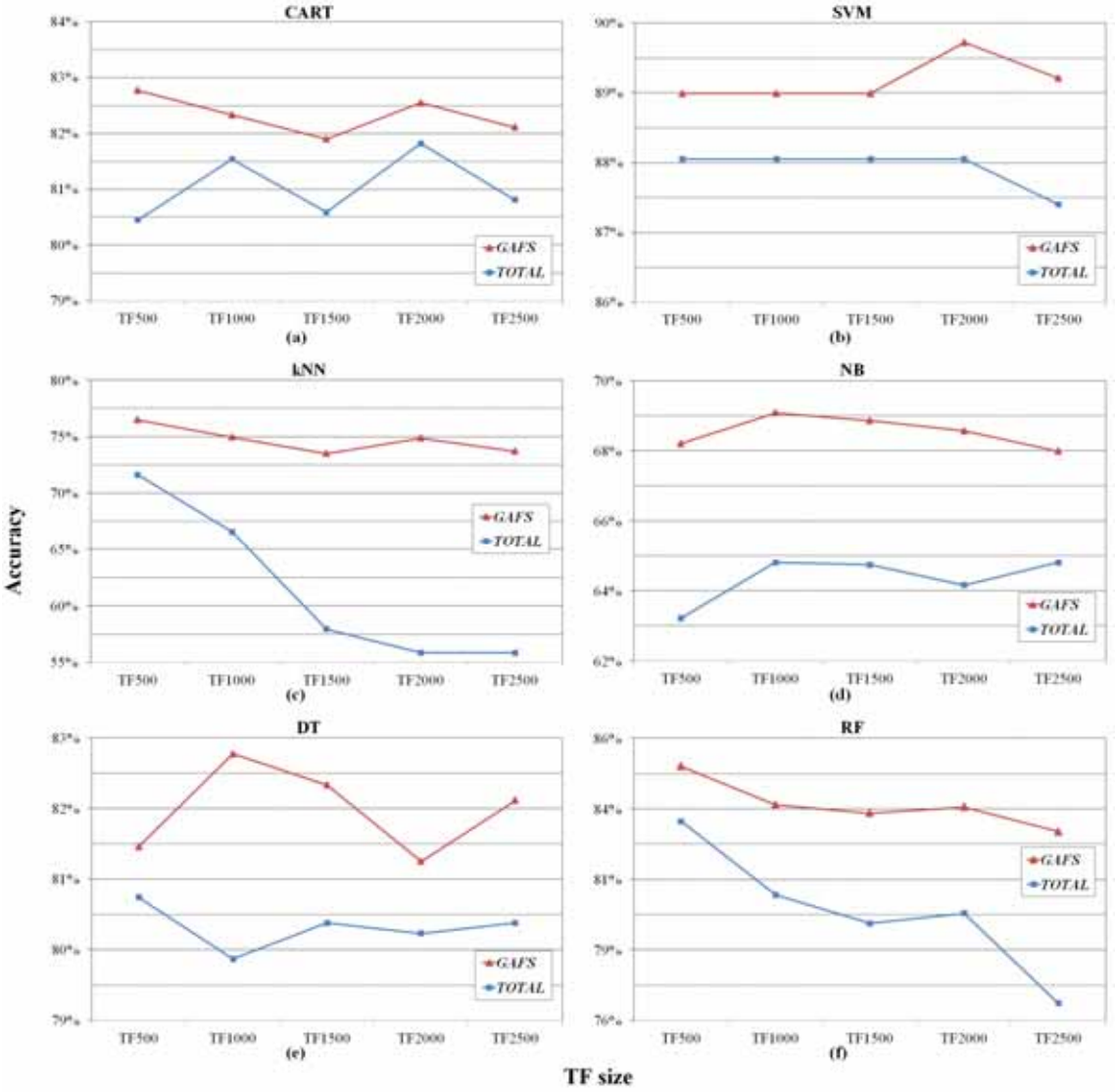

圖 12：遞增特徵集合之分類結果比較

呈現不穩定的狀況，而 GAFS 方法不僅穩定許多，同時對於分類準確性也有所提升；而圖 12(b)、圖 12(d)與圖 12(e)中各分類模式皆是較不受特徵字詞數量影響的分類方法，僅有圖 12(e)中 DT 採用 GAFS 方法有稍微不穩定，但分類效能也並無隨著特徵增加而持續下降，浮動的情形也僅有 1%至 2%間，相對而言不算太差，而其餘分類模式不論是在 TOTAL 或是 GAFS 上都有穩定地表現。經由上述分析與探討，我們可以發現在 TOTAL 方法中，隨著特徵字詞數量增加各分類模式被影響的程度皆不盡相同，大致可分為以下三類：(1)分類模式易隨著受特徵數量增加反而降低分類效能，例如：kNN 與 RF；(2)分類模式不易受特徵數量影響，例如：SVM、NB 與 DT；(3)分類模式則是呈現不穩定的狀況，例如：CART。而從實證研究中得證，採用本研究提出之 GAFS 方法，不僅能有效地改善分類效能，亦能使分類模式較不易受特徵數量影響而表現更加穩定。

伍、結論與建議

本章將探討研究結果，共分為三部分，首節歸納出研究結論與貢獻；次節說明研究限制；最後一節則給予未來研究方向與建議。

一、研究結論與貢獻

現代數位資訊爆炸性地成長，資料量與日俱增，同時也增加人們在找尋資訊上的搜尋成本。因此，對組織而言，如何有效地分類管理文件將是目前一項重要的研究議題。傳統人力進行分類管理的方式缺乏效率，而自動化文件分類管理能夠透過電腦快速地分析文件內容，並自動將文件迅速地進行分門別類，有效地節省使用者的搜尋時間，使其在短時間內就能瀏覽所需資訊，由此可見自動化文件分類研究的重要性。

本研究的目的在於提出一套完善的自動化文件分類方法，透過機器學習的方式讓電腦代替人工分類大量的文件資料，並以 GA 為基礎調整 GA 設定的參數，來建立文件分類模式，期望能夠解決文件高維度字詞特徵的問題，改善各分類模式的分類效能，建立一套分類效果佳且穩定的自動化文件分類模式。

根據第四章實證研究結果顯示，在 WebKB 網頁文件資料集中，本研究提出之 GAFS 方法能夠有效地改善各分類模式的分類效能。另外，從特徵數量遞增的實驗中可以發現，在未採用本研究方法的情況下，分類模式之效能易隨著特徵字詞數量遞增而受到影響，使得分類效能隨之遞減或浮動。而採用 GAFS 方法後，各分類模式不僅能夠提升分類效能，同時也較不易受到特徵字詞數量影響分類效能，各分類模式表現得更加穩定。

本研究主要貢獻在於提升各自動化文件分類模式之分類效能，實驗結果證實

以 GA 為基礎之 GAFS 自動化文件分類模式明顯優於 TOTAL，在特徵維度逐漸擴大的情況下，GAFS 皆能有效地改善分類效能，並且擁有穩定地分類準確度。

二、研究限制

本研究提出之文件分類模式是以 GA 為基礎進行特徵選擇之演化，進行特徵挑選的過程需要許多世代的演化，而各世代中又必須透過許多組染色體進行分類準確度之評估，不斷地依據染色體中所挑選之特徵重複建立不同的分類模式，所以若是採用訓練時間較長的分類模式，則整體訓練時間將大幅增加。因此，本研究受限於分類模式的訓練時間，而沒有選擇較耗時的分類模式進行文件分類之評估（例如：類神經網路）。另外值得注意的是，本研究在 GA 參數調整階段的實驗同樣是礙於效率上的考量，所以僅採用特徵集合中較小的 TF1000 作為評估準則，因此歸納出的參數可能僅較適合 TF1000 特徵集合進行演化，而 TF1000 以上的特徵集中字詞較多，所以可能會需要更長時間的演化世代或是更多樣化的族群，而必須增加 GA 演化世代數或是族群數量，才能有更佳的分類結果。

三、未來研究方向

本研究雖已採用具備公證性的 WebKB 網頁文件資料集進行實證研究，並以文件分類準確度做為評估準則，透過各階段的實驗證實本研究方法確實能夠改善分類準確度，但在未來研究上，希望能夠採用多組類型不同的文件資料集進行實驗，例如：Reuters-21578、20Newsgroups，以及 Cade12 等資料集，進一步驗證本研究方法是否在多數的資料集中都能有穩定的分類效能。另外，在特徵選取與文件表達方法中，本研究僅採用較常被使用的 TF 與 TFIDF 方法進行初步的特徵選取與文件表達，而未考量其他方法，在未來建議可以採用不同的特徵選取方法進行比較與測試，分析何種特徵選取與文件表達方法較適合本研究。

誌謝

感謝審查委員提供許多的寶貴建議，使本論文之內容更臻完美；本研究承蒙國科會專題研究計畫經費補助，計畫編號為 NSC 102-2410-H-194-104-MY2 和 NSC 102-2410-H-194-100-MY2，謹致謝忱。

參考文獻

Aghdam, M.H., Ghasem-Aghaee, N. and Basiri, M.E. (2009), 'Text feature selection using ant colony optimization', *Expert systems with applications*, Vol. 36, No. 3, pp. 6843-6853.

Ambert, K.H. and Cohen, A.M. (2012), 'K-information gain scaled nearest neighbors: a novel approach to classifying protein-protein interaction-related documents', *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 9, No. 1, pp. 305-310.

Apté, C., Damerau, F. and Weiss, S.M. (1994), 'Automated learning of decision rules for text categorization', *ACM Transactions on Information Systems*, Vol. 12, No. 3, pp. 233-251.

Bai, R., Wang, X. and Liao, J. (2007), 'Combination of rough sets and genetic algorithms for text classification', *Lecture Notes in Artificial Intelligence*, Vol. 4476, pp. 256-268.

Beasley, D., Bull, D.R. and Martin, R.R. (1993), 'An overview of genetic algorithms: Part 1. Fundamentals', *University computing*, Vol. 15, pp. 58-58.

Breiman, L. (2001), 'Random forests', *Machine learning*, Vol. 45, No. 1, pp. 5-32.

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984), 'Classification and Regression Trees', *Wadsworth*, Belmont, CA.

Buckley, C. (1985), *Implementation of the SMART information retrieval system*, Cornell University.

Chen, H., Jiang, W., Li, C. and Li, R. (2013), 'A heuristic feature selection approach for text categorization by using chaos optimization and genetic algorithm', *Mathematical Problems in Engineering*, Vol. 2013, pp.1-6.

Cardoso-Cachopo, A. and Oliveira, A. (2007), 'Combining LSI with other classifiers to improve accuracy of single-label text categorization', *First European Workshop on Latent Semantic Analysis in Technology Enhanced Learning-EWLSATEL, (EWLSATEL2007)*, Heerlen, The Netherlands, March 29-30.

Cheatham, M. and Rizki, M. (2006), 'Feature and prototype evolution for nearest neighbor classification of web documents', *IEEE Third International Conference on Information Technology*, Las Vegas, NV, April 10-12, pp. 364-369.

Chen, J., Huang, H., Tian, S. and Qu, Y. (2009), 'Feature selection for text classification with Naïve Bayes', *Expert Systems with Applications*, Vol. 36, No. 3, pp. 5432-5435.

Chou, C.H., Han, C.C. and Chen, Y.H. (2007), 'GA based optimal keyword extraction in an automatic Chinese web document classification system', *Frontiers of High Performance Computing and Networking ISPA 2007 Workshops*, Niagara Falls, Canada, August 28-September, pp. 224-234.

Church, K.W. and Hanks, P. (1990), 'Word association norms, mutual information, and lexicography', *Computational Linguistics*, Vol. 16, No. 1, pp. 22-29.

Cortes, C. and Vapnik, V. (1995), 'Support-vector networks', *Machine learning*, Vol. 20, No. 3, pp. 273-297.

Damerau, F.J., Zhang, T., Weiss, S.M. and Indurkhya, N. (2004), 'Text categorization for a comprehensive time-dependent benchmark', *Information Processing & Management*, Vol. 40, No. 2, pp. 209-221.

Denoyer, L. and Gallinari, P. (2004), 'Bayesian network model for semi-structured document classification', *Information processing & management*, Vol. 40, No. 5, pp. 807-827.

Drucker, H., Donghui, W. and Vapnik, V.N. (1999), 'Support vector machines for spam categorization', *IEEE Transactions on Neural Networks*, Vol. 10, No. 5, pp. 1048-1054.

ElAlami, M.E. (2009), 'A filter model for feature subset selection based on genetic algorithm', *Knowledge-Based Systems*, Vol. 22, No. 5, pp. 356-362.

Escudero, G., Màrquez, L. and Rigau, G. (2000), *Boosting applied to word sense disambiguation*, Springer Berlin Heidelberg, pp. 129-141.

Fall, C.J., Torcsvari, A., Benzineb, K. and Karetka, G. (2003), 'Automated categorization in the international patent classification', *ACM SIGIR Forum*, Vol. 37, No. 1, Toronto, Canada, pp. 10-25.

Joachims, T. (1998), 'Text categorization with support vector machines: Learning with many relevant features', *Machine Learning*, Springer Berlin Heidelberg, pp. 137-142.

Johnson, D.E., Oles, F.J., Zhang, T. and Goetz, T. (2002), 'A decision-tree-based symbolic rule induction system for text categorization', *IBM Systems Journal*, Vol. 41, No. 3, pp. 428-437.

Larkey, L.S. (1999), 'A patent search and classification system', *Proceedings of the fourth ACM conference on Digital libraries*, Berkeley, CA, August 11-14, pp. 179-187.

Lee, C. and Lee, G.G. (2006), 'Information gain and divergence-based feature selection for machine learning-based text categorization', *Information processing &*

management, Vol. 42, No. 1, pp. 155-165.

Li, Y., Shiu, S.C.K., Pal, S.K. and Liu, J.N.K. (2006), 'A rough set-based case-based reasoner for text categorization', *International Journal of Approximate Reasoning*, Vol. 41, No. 2, pp. 229-255.

Lu, S.H., Chiang, D.A., Keh, H.C. and Huang, H.H. (2010), 'Chinese text classification by the Naïve Bayes Classifier and the associative classifier with multiple confidence threshold values', *Knowledge-Based Systems*, Vol. 23, No. 6, pp. 598-604.

Lu, Z., Shi, H., Zhang, Q. and Yuan, C. (2009), 'Automatic chinese text categorization system based on mutual information', *IEEE International Conference on Mechatronics and Automation (ICMA 2009)*, August 9-12, Changchun, China, pp. 4986-4990.

Luo, Q., Chen, E. and Xiong, H. (2011), 'A semantic term weighting scheme for text categorization', *Expert Systems with Applications*, Vol. 38, No. 10, pp. 12708-12716.

Man, L., Tan, C.L., Jian, S. and Yue, L. (2009), 'Supervised and traditional term weighting methods for automatic text categorization', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 31, No. 4, pp. 721-735.

Martin-Bautista, M.J. and Vila, M.A. (1999), 'A survey of genetic feature selection in mining issues', *Proceedings of the 1999 Congress on Evolutionary Computation (CEC 1999)*.Washington, DC, July 6-9, pp. 1314-1321.

Özel, S.A. (2011), 'A Web page classification system based on a genetic algorithm usingtagged-terms as features', *Expert Systems with Applications*, Vol. 38, No. 4, pp.3407-3415.

Pietramala, A., Policicchio, V.L., Rullo, P. and Sidhu, I. (2008), 'A genetic algorithm for text classification rule induction', *Lecture Notes in Artificial Intelligence*, Vol. 5212, pp. 188-203.

Porter, M.F. (1980), 'An algorithm for suffix stripping', *Program: electronic library and information systems*, Vol. 14, No. 3, pp. 130-137.

Qi, D. and Sun, B. (2004), 'A genetic k-means approaches for automated web page classification', *Proceedings of the 2004 IEEE International Conference on Information Reuse and Integration(IRI 2004)*, Las Vegas, NV, November 8-10, pp. 241-246.

Quinlan, J.R. (1986), 'Induction of decision trees', *Machine Learning*, Vol. 1, No. 1, pp. 81-106.

Quinlan, J. R. (1993), *C4.5: programs for machine learning*, Morgan kaufmann, San Mateo, CA.

Rish, I. (2001), 'An empirical study of the naive Bayes classifier', Proceedings ofIJCAI 2001 Workshop on Empirical Methods in Atificial Intelligence(IJCAI 2001), Seattle, WA, August 4-10, pp. 41-46.

Salton, G. (1971), *The SMART Retrieval System - Experiments in Automatic Document Processing*, Prentice-Hall, Inc. Upper Saddle River, NJ.

Salton, G. and McGill, M.J. (1983), *Introduction to Moderm Information Retrieval* McGraw-Hill, London.

Schütze, H., Hull, D.A. and Pedersen, J.O. (1995), 'A comparison of classifiers and document representations for the routing problem', *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, Seattle, WA, July 9-13, pp. 229-237.

Sebastiani, F. (2002), 'Machine learning in automated text categorization', *ACM Computing Surveys (CSUR),* Vol. 34, No. 1, pp. 1-47.

Sikora, R. and Piramuthu, S. (2007), 'Framework for efficient feature selection in genetic algorithm based data mining', *European Journal of Operational Research*, Vol. 180, No. 2, pp. 723-737.

Spears, W.M., Jong, K.A.D., Bäck, T., Fogel, D.B. and Garis, H.D. (1993), 'An overview of evolutionary computation', *Machine Learning: ECML-93*, Springer, pp. 442-459.

Sun, A., Lim, E.P. and Liu, Y. (2009), 'On strategies for imbalanced text classification using SVM: A comparative study', *Decision Support Systems,* Vol. 48, No. 1, pp. 191-201.

Tan, S. (2006), 'An effective refinement strategy for KNN text classifier', *Expert Systems with Applications,* Vol. 30, No. 2, pp. 290-298.

Tauritz, D.R., Kok, J.N. and Sprinkhuizen-Kuyper, I.G. (2000), 'Adaptive Information Filtering using evolutionary computation', *Information Sciences*, Vol.122, No. 2-4, pp. 121-140.

Teichert, T. and Mittermayer, M.A. (2002), 'Text mining for technology monitoring', *Proceedings of the 2002 IEEE International Engineering Management Conference,* Cambridge, UK, August18-20, pp. 596-601.

Trappey, A.J.C., Hsu, F.C., Trappey, C.V. and Lin, C.I. (2006), 'Development of a patent document classification and search platform using a back-propagation network', *Expert Systems with Applications,* Vol. 31, No. 4, pp. 755-765.

Uğuz, H. (2011), 'A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm', *Knowledge-Based Systems*, Vol. 24, No. 7, pp. 1024-1032.

Wang, D., Zhang, H., Liu, R. and Lv, W. (2012), 'Feature selection based on term frequency and T-test for text categorization', in: *Proceedings of the 21st ACM international conference on Information and knowledge management*, Maui, HI, October 29 - November 2, pp. 1482-1486.

Wei, C.P., Lin, Y.T. and Yang, C.C. (2011), 'Cross-lingual text categorization: Conquering language boundaries in globalized environments', *Information Processing & Management*, Vol. 47, No. 5, pp. 786-804.

Xia, F., Jicun, T. and Zhihui, L. (2009), 'A text categorization method based on local document frequency', IEEE Sixth International Conference on*Fuzzy Systems and Knowledge Discovery(FSKD 2009), Tianjin, China, August 14-16*, pp. 468-471.

Yang, J., Pai, P., Honavar, V. and Miller, L. (1998), 'Mobile intelligent agents for document classification and retrieval: A machine learning approach',*14th European Meeting on Cybernetics and Systems Research. Symposium on Agent Theory to Agent Implementation (EMCSR 1998)*, Vienna, Austria, April 14-17.

Yang, Y., Carbonell, J.G., Brown, R.D., Pierce, T., Archibald, B.T. and Liu, X. (1999), 'Learning approaches for detecting and tracking news events', *IEEE Intelligent Systems*, Vol. 14, No. 4, pp. 32-43.

Yang, Y. and Pedersen, J.O. (1997), 'A comparative study on feature selection in text categorization', ICML, pp. 412-420.

Yu, B., Xu, Z. and Li, C. (2008), 'Latent semantic analysis for text categorization using neural network', *Knowledge-Based Systems*, Vol. 21, No. 8, pp. 900-904.

Zhang, M.L. and Zhou, Z.H. (2006), 'Multilabel neural networks with applications to functional genomics and text categorization', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 18, No. 10, pp. 1338-1351.

Zhang, M.L. and Zhou, Z.H. (2007), 'ML-KNN: A lazy learning approach to multi-label learning', *Pattern Recognition*, Vol. 40, No. 7, pp. 2038-2048.

Zhang, W., Xu, B. and Cui, Z. (2005), 'A document classification approach by GA feature extraction based corner classification neural network', *Proceedings of the 2005 international conference on cyberworlds*, Singapore, November23-25, pp. 499-504.

Zhang, W., Yoshida, T. and Tang, X. (2011), 'A comparative study of TF-IDF, LSI and multi-words for text classification', *Expert Systems with Applications*, Vol. 38, No.

3, pp. 2758-2765.

Zheng, Z., Wu, X. and Srihari, R. (2004), 'Feature selection for text categorization on imbalanced data', *ACM SIGKDD Explorations Newsletter,* Vol. 6, No. 1, pp. 80-89.

附錄一：分類模式參數

Classifier	Option	Accuracy	Precision	Recall	F1
*k*NN	-k 2	0.6495	0.6387	0.6495	0.6363
	-k 5	0.6503	0.6467	0.6503	0.6265
	-k 10	0.6531	0.671	0.6531	0.6272
	-k 15	0.6445	0.6619	0.6445	0.6092
	-k 20	0.6655	0.6904	0.6655	0.6310
	-k 25	0.6481	0.704	0.6481	0.6119
	-k 30	0.6191	0.7052	0.6191	0.5818
	-k 50	0.5554	0.6788	0.5554	0.4999
CART	-M 2	0.8088	0.8089	0.8088	0.8076
	-M 5	0.8154	0.8169	0.8154	0.8155
	-M 10	0.7922	0.8018	0.7922	0.7957
	-M 20	0.7791	0.7824	0.7791	0.7804
	-M 50	0.7748	0.7721	0.7748	0.7695
	-M 100	0.7321	0.7317	0.7321	0.7191
DT	-M 2	0.7886	0.7878	0.7886	0.7881
	-M 5	0.7835	0.7877	0.7835	0.7844
	-M 10	0.7987	0.7949	0.7987	0.7954
	-M 20	0.7900	0.7884	0.7900	0.7865
	-M 50	0.7857	0.7853	0.7857	0.7799
	-M 100	0.7466	0.7617	0.7466	0.7496