以遞增式分群為基之分類方法過濾具偏斜類別及概念漂移之垃圾郵件

蕭文峰
屏東商業技術學院資訊管理系

張德民
中山大學資訊管理系

胡國信
財團法人金屬工業研究發展中心

摘要

電子郵件是現代人最常用以接收資訊的媒介之一，然而許多人利用它的方便、快速、及成本低廉等特性散佈大量電子郵件，以達到廣告宣傳的效果。如此造成電子郵件用戶的信箱中充斥著大量未經用戶許可的垃圾郵件。因此解決垃圾郵件問題是一個重要且急迫的議題。本研究的目的即在於提出適當的分類方法以提高過濾垃圾郵件的績效。所提方法「以遞增式分群為基之分類器(Incremental Clustering-Based Classifier; ICBC)」是以分群為基礎之分類演算法：ICBC會將文件分成數群，並找出每群等量具代表性的特徵，以解決垃圾郵件資料偏斜的問題；同時ICBC也具有遞增學習的能力，能以較低的成本及較快的速度來適應環境的改變，以解決電子郵件主題漂移的問題，並避免每次用所有資料重新學習的成本負擔。本研究共進行了四個實驗以瞭解所建構分類器的效能，實驗結果顯示ICBC可有效地同時處理中英文垃圾郵件之過濾；也能克服資料偏斜與主題漂移的問題。這些實驗結果驗證了ICBC的適用性。

關鍵字：垃圾郵件過濾、資料偏斜、主題漂移、遞增式學習

An Incremental Cluster-Based Classification Approach to Filtering Spam with Skewed Classes and Drifting Concepts

Wen-Feng Hsiao
Department of Information Management, National Pingtung Institute of Commerce

Te-Min Chang
Department of Information Management, National Sun Yat-sen University

Guo-Hsin Hu
Metal Industries Research & Development Centre

Abstract

E-mail has become one of the most popular communication channels for people to disseminate information nowadays. However, because of its convenience, speediness, and low cost, some people abuse this channel to spread information for advertisement and promotion purpose. This often causes users' troubles in managing their mailboxes. These unsolicited and undesired emails are referred to as spam (or junk emails). Spam filtering, therefore, is an essential issue to help users get rid of annoying emails. The purpose of this research is to propose an appropriate classification approach to filtering spam with skewed classes and drifting concepts. An Incremental Cluster-Based Classification method (ICBC) is proposed accordingly. ICBC first clusters documents into several groups, and an equal number of keywords are then extracted from each group to alleviate the problem of skewed class distributions. In addition, ICBC also possesses the ability of incremental learning that can adapt itself to the changing environment with drifting concepts and avoid the cost of re-training. Four experiments are conducted to evaluate ICBC. The results show that ICBC can effectively classify both Chinese and English spam. It can also deal with the issues of skewed class distributions and drifting concepts. The feasibility of ICBC is thus justified.

Key words: Spam filtering, skewed class distribution, concept drifting, incremental learning

壹、緒論

在現今網路發達的時代，電子郵件已經變成現代人最常用以接受資訊的媒介之一。它的方便性、不受地形的限制、傳送快速、成本低廉等優點，使得我們對電子郵件的依賴程度與日俱增。但由於電子郵件的傳送簡單，許多人便利用它散佈大量訊息，包括散佈商業廣告、致富訊息、色情廣告、特定思想、及政治宣傳等，使得用戶的信箱中充斥著大量未經他/她許可的垃圾郵件。因此解決垃圾郵件問題是許多公司、政府、甚或個人使用者的當務之急。

為解決垃圾郵件問題，許多學者提出不少的嘗試解法。早期的方法為簡單欄位比對或稱為法則式的過濾。這種方法是透過郵件欄位(比對特定寄件人、收件人、日期、主旨、附件檔等簡單人工法則式過濾)來進行郵件的過濾。雖然可以濾去一些垃圾郵件，但對較複雜的垃圾郵件幾乎很難發生效用，因為垃圾郵件的關鍵問題即在於傳送者能夠隱藏其真實身份及傳送郵件的主機，及變化郵件主旨，使反垃圾郵件程式(anti-spam program)無法區別其真實性。為了解決此類問題，更進階的過濾方法則倚賴郵件的內容分析，透過分析郵件，並萃取出垃圾郵件及正當郵件的特徵，以做為判定郵件類別的基礎。這些技術主要源自於文件探勘(text mining)。

但由於垃圾郵件的一些特性，大部份的程式並無法有效地過濾垃圾郵件 (Fawcett 2003)：首先是偏斜的類別分佈(skewed class distribution)，所謂的偏斜類別是指垃圾郵件與正當郵件之數目差異太大的情形，經常使用之帳戶所收的垃圾郵件數大多遠高於正當郵件。一般的分類演算法對偏斜類別的資料通常無法適切處理，易造成數目較少類別的召回率過低，而這些小類別的分類錯誤通常是決策成本的主要來源(Weiss & Provost 2001)；其次是主題漂移(concept drift)，其主題可能會隨著流行物(e.g., Viagra)或季節性(e.g., 耶誕商品)之變化而改變，如不能適切地調整學習架構，舊的分類架構將不足以判斷新的資訊。

本研究的目的即是針對上述問題，提出適當的分類方法以提高過濾垃圾郵件的績效。對於類別偏斜問題，本研究提議以分群為基礎之分類方式：利用階層式分群演算法(Jain 1999)當作前置處理，將文件切分成數個子集合，並找出每群等量具代表性的特徵，能有效彰顯出小類別之獨特性，而不致被忽視；對於主題漂移的現象，則以遞增式學習(微調分群架構)來調適，如此能以達較快的速度及較低的成本適應環境的變化。本研究所提之方法稱為「以分群為基礎之遞增式分類法(Incremental Clustering-Based Classification; ICBC)」，在後續的章節中，我們將展示此方法之運行及其有效性。

本文的章節安排如下：第二節探討偏斜類別分佈、概念漂移、及垃圾郵件過濾之相關研究。第三節闡述本研究所提之方法「具分群機制之遞增式分類法」。第四節以實驗來瞭解本研究所提方法之績效。第五節總結本研究的貢獻及未來可能的研究方向。

貳、文獻探討

本節回顧過去在偏斜類別分佈、概念漂移、分群分析及垃圾郵件等議題之相關研究。

一、偏斜類別分佈(skewed class distribution)

當資料集的類別分佈是偏斜時，一般的分類器通常無法針對量少的類別進行正確的預測，然而這些量少的目標類別卻常是我們所關切的。例如，企業每月的流失客戶、信用卡詐欺偵測、罕見疾病判斷、保險保費詐領偵測等，這些領域中量少類別的錯誤成本都是相當高的。過去用以解決偏斜類別分佈的方法概分為四類，依次為增加少數法(over-sampling) (Honda et al. 1997)、減少多數法(under-sampling) (Hart 1968)、指定分類錯誤成本(assign misclassification costs) (Monard & Batista 2002)、及多分類器委員會法(multi-classifier committee approach) (Chan et al. 1999)。其中，增加少數法是想辦法增加量少類別的範例數，減少多數法則相反地是將量多的類別的範例數減少，使之與量少類別的範例數相當。此兩種方法都有其缺點，前者可能引入雜訊，後者則無法充份使用原有的資料。因此，指定分類錯誤成本則由指定較高的成本至小類別的分類錯誤，而多分類器委員會法利用資料切割方式來達到解決偏斜分佈的問題。

二、概念漂移(concept drift)

分類器是由歸納過去範例，找出區分各類別(概念)邊界的規則，並藉此預測未知範例所屬類別。當所學習的概念(類別邊界)不變，分類器通常可達到不錯的績效。但當概念會隨著時間而改變時，一般的分類演算法並無法適切處理。我們稱概念隨著時間改變的現象為「概念漂移」，例如人們對產品的偏好會隨時間而改變，垃圾郵件的主題會因當時的流行物品而改變等。過去對概念漂移的探討主要從兩方面，一是依漂移的速度(Stanley 2001)可分為：立即的概念轉變、中度的概念漂移、及緩慢的概念漂移。另一方面是依漂移的內容(Forman 2006)可分為：主要類別分佈改變、子類別分佈改變、類別分佈無常的改變。為應付概念漂移的現象，分類器本身必須具備遞增學習(或依環境調適)的能力。

三、遞增式學習

遞增式學習最主要的目標是為了在一個動態的環境中，使得概化(Generalization)的錯誤最小化(d'Alché-Buc & Ralaivola 2002)，換句話說也就是為了能以較快的速度及較少的成本來適應各種不同的環境，使得分類結果正確無誤。為了提昇分類器的正確率，Wu et al.(2002)提出一個漸進的方法改善資料與分類模式不配適的問題。這個方法並不改變分類方式及其模式，而是透過持續的分類修正來改善模式不配適。其方法如圖 1所示，設一個二元分類問題，給定一組事先分類好的訓練範例D，使用某一分類技術Cl訓練分類

器。一開始由全部的範例D中訓練出初始的分類器Cl_0，並分別將D分成正負兩種類別：DP及DN。但通常不會第一次就完全正確分類，所以再就D_P及D_N分別訓練子分類器：ClP及ClN，並分別將D_P及D_N再分成正負兩種類別。這個方法持續分割資料並訓練分類器，直到分割資料後的F-measure 值不再增加才停止。經由這樣的反覆進行可使得每一節點的資料越來越單純，資料混亂的程度就會降低，每一樹葉節點的資料所表達的意念也更為一致。

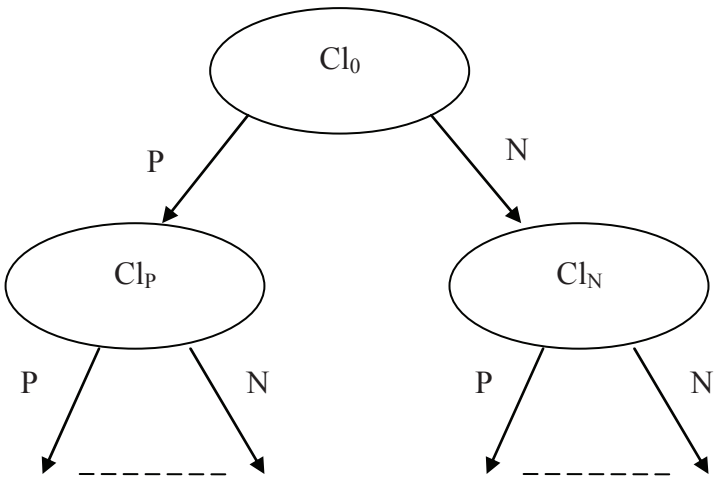

圖 1：文件分類遞增式學習的架構

四、分群演算法

分群方法一般可分為階層式(hierarchical)及分裂式(partitional)兩種(Jain et al. 1999)。分裂式分群法(e.g., k-means)是將原始資料直接拆除分成各群。此法會訂定一個評估分群好壞的準則，如平方誤(Squared Error)，並依此最佳化分群結構。亦即，此類分群法會去搜尋(通常是局部搜尋)能最佳化此準則的分群方式。

階層式分群演算法則根據群的形成方式又可分為聚合型(agglomerative)與分裂型(divisive)兩類。聚合型分群方式是藉由結合最相似的資料集成一群，使分群後的每一子群之郵件具有較高的相似度，異質性也因此而降低。依其相似度計算方式的不同又可分為單一連結法(Single Link)、完全連結法(Complete Link)及平均連結法(Average Link)：

單一連結法：目標點依次和每一群的最近點計算距離，此代表目標點與各群的距離，目標點會被結合到最近的群中。

完全連結法：目標點依次和每一群的最遠點計算距離，此代表目標點與各群的距離，目標點會被結合到最近的群中。

平均連結法：目標點依次和每一群的所有點計算平均距離，此代表目標點與各群的距離，目標點會被結合到最近的群中。

相反地，分裂型(divisive)分群法是一開始先將所有資料點視為同一群，並使用某一準則，例如均方誤(Mean Squared Error; MSE)，來當成分裂條件。此過程會一直遞迴進

行，直至子群的MSE值低於某一門檻值才停止。MSE的計算方式如下：

$$MSE = \frac{\sum_i (X_i - C)^2}{N}$$

其中，X_i 表示觀測值，C表示群心，N表示該群的資料個數。

五、垃圾郵件過濾相關研究

目前最常用的垃圾郵件過濾方法是採用文件探勘的分類技術(Kiritchenko & Matwin 2001; Lian 2002)。在應用這類方法時，使用者必須告訴過濾(分類)器那些是垃圾郵件、那些是正當郵件，過濾器依照使用者界定的類別，學習各類別中郵件內文的屬性特徵以當作分類的基準。

過去已有許多文件探勘技術被用於垃圾郵件的過濾，包括(但未窮舉[1])：法則式(CN2、Ripper)、決策樹(Boosting Decision Tree)、機率式(Naïve Bayes、Bayesian Network、Rocchio)、函數式(Support Vector Machine、Self Organizing Map)、及案例式(K Nearest Neighbor)。

其中，K個最近鄰居法(以下簡稱KNN)屬於記憶式(或範例式)的學習法。記憶式法會儲存訓練範例在一個記憶體結構，並使用他們來做分類判斷。他們並未對所要學習的概念建構任何的模式，只單純的儲存訓練範例。記憶式學習的最簡單形式是以所選用的屬性來定義一個多維空間，並且將每個訓練範例視為是這空間中的一個點。當欲分類測試範例時，則透過計算此測試範例與空間中的那個訓練範例最接近來獲得預測值。為避免雜訊的可能影響，相對於只取最接近的範例，KNN是藉由K個最近的訓練範例進行投票以求得測試範例的類別。

Payne與Edward (1997)利用法則式歸納法(CN2)來自動學習如何分類郵件，並開發出一套軟體命名為Magi。Magi 透過觀察使用者與郵件間的互動(如轉寄或刪除)的方式來決定新進郵件的處理方式，屬於一種代理人程式。研究中提出兩種門檻值來當作執行的準則，分別為預測性門檻(predictive threshold)及信賴門檻值(confidence Threshold)。若預測效果滿足預測性門檻值，則預測該郵件可能的類別並提供給使用者參考；若滿足信賴門檻值，則可直接依據該類別的處理方式執行(轉寄或移除)，毋需詢問使用者。

此外，貝氏機率模式(Bayes Probabilistic Model)也被常應用於文件分類中。Sahami et al. (1998)使用天真貝氏法(Naïve Bayes)來分析垃圾郵件，Androutsopoulos et al. (2000)也使用天真貝氏法學習郵件的內文之特徵來建構垃圾郵件分類器。由於其效果不差，所以許多研究亦使用天真貝氏法來當作比較的基準(baseline)。但在天真貝氏法存在著兩個基本的假設：亦即假設類別與文件間有一對一的關係(one-to-one correspondence)及字詞出現的頻率是相互獨立(Nigam et al. 2000)。此兩點假設與實際狀況不一定相符。所以後來的貝

[1] 此分類方式是參考Weka (version 3.5.3)(http://www.cs.waikato.ac.nz/ml/weka/)系統之分類方式。

氏網路(Bayesian Network)研究朝向有限放鬆獨立的假設。

Drucker et al. (1999)使用支援向量機(Support Vector Machine; SVM)來過濾垃圾郵件，並與其他三種分類方法比較：Ripper、Rocchio及Boosting Decision Tree。他們發現使用二元(binary)的方式來表示郵件特徵向量時，SVM分類的效果為最佳。就整體而言，SVM及boosting decision tree 的預測效果為最佳，速度也較快。Luo與Zincir-Heywood(2005)提出以SOM (Self Organized Map)為基的次序分析系統(SOM Based Sequence Analysis; SBSA)用以過濾垃圾郵件。此系統結合以SOM為基之次序性資料表示法與利用字序資訊之KNN分類器。其結果顯示SBSA的表現比天真貝氏法較佳。

Delany et al. (2005) 提出一垃圾郵件過濾系統叫ECUE。此系統使用範例式(instance-based)的學習技術來追蹤概念的漂移。藉由學習錯誤分類的郵件(不管是垃圾郵件或正當郵件)，ECUE可以依使用者偏好的特定性進行個人化設定，並能適應垃圾郵件改變的本質。但由於採用的是KNN演算法(Delany & Cunningham 2006)，一旦其錯誤分類觸發了學習機制，ECUE就必須重建其範例庫並重新進行屬性選取的程序。本研究所提出的方法ICBC在概念漂移問題上的處理類似ECUE，但最主要差別在於ICBC是遞增式地修改分群的結構而非重建範例庫來重新學習。

參、 遞增式分群之分類法

本研究所提方法為「以分群為基礎之遞增式分類法(ICBC)」，主要在解決電子垃圾郵件之資料偏斜問題，以及增加過濾器之遞增學習能力。因此，本研究方法包含兩階段：分類器訓練階段以及遞增學習階段。各階段之內容詳述如下：

一、 初始分類器建立階段

本階段的目的在於以所蒐集之資料訓練電子郵件分類器。我們採用以分群為基礎之分類方式，以分群當作前置處理。傳統文件探勘若由訓練文件中挑選特徵字(如以TFIDF作判斷)，通常便以這組特徵字作為分類的依據。但若文件類別大小不一，小類別特徵字的權重較低(字頻較低)，通常無法被挑出。因此用於分類時，屬於小類別的新文件會被分到有較多特徵字的大類別中，而造成錯誤。

ICBC將所蒐集之訓練郵件，分別對正當郵件及垃圾郵件作分群分析，將概念差異大的郵件分離，概念相近的郵件集結一起而形成子群。再對每一子群抽取同樣數量的特徵字，如此小類別之特徵字便容易浮現。用於分類時，屬於小類別的新文件就較不易分錯。此一作法呼應部份研究者(Sebastiani 2002)主張文件分類應採局部特徵字選取之論點[2]。圖 2顯示傳統方法與本研究方法比較的示意圖。

[2] 相對於全域特徵字選取，局部選取法是在各類別內選取特徵字。

整個分群過程都是以TFIDF來選取特徵字及群心，但一旦分群結束後，再針對每一群重新抽取特徵字，此時抽取的指標僅採用TF，因為此步驟欲求得群內共同概念，所以不需要以IDF來找出同群內彼此的不同點。最後由計算群中所有文件之TF平均，可求得每一群群心之特徵向量。

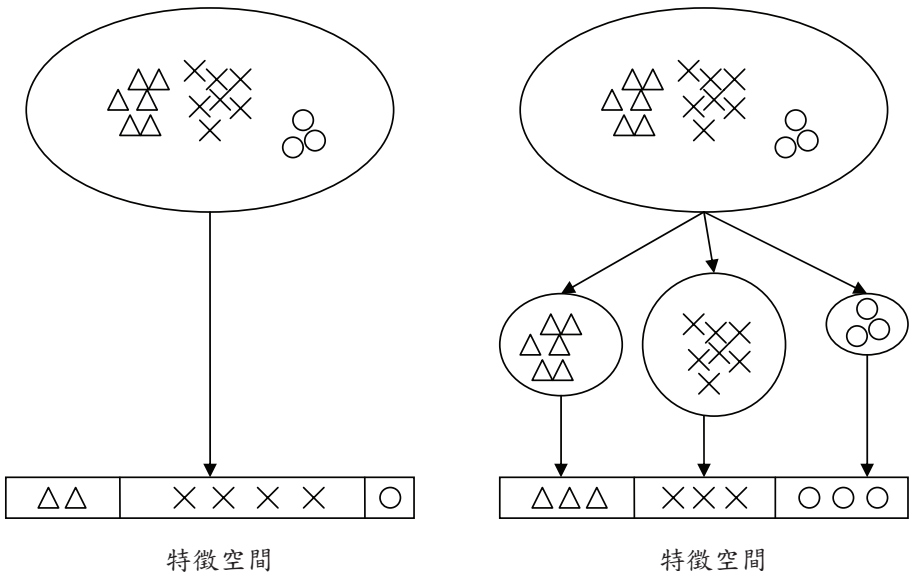

圖 2：傳統特徵字選取與本研究特徵字選取比較之示意圖

本階段的處理流程如圖 3所示。首先將停用字移除，並將字根還原，可消除雜訊及減少特徵空間的維度[3]。本研究採用McCallum(2002)所發展的bow，裡面共列出524個停用字；還原字根的部分則使用Porter stemming algorithm(Porter 1980)來處理。接著再針對各子群[4]內之樣本以TFIDF抽取特徵字，如此每篇文件之特徵向量就可以其所屬子群之特徵字集來表示，例如，文件Di 可表示成(wi1, wi2, ...win)。每一群的群心則可由該群內所有文件的特徵向量平均而得。有了群心則可計算群內各文件對群心之均方誤(MSE)值，以判斷是否要再細分子群。本研究的分群是採用分裂型(divisive)階層分群法，以MSE值當成分裂之準則；當某群之MSE值高於門檻值則進行分裂，系統會於對應群的郵件資料夾產生子群(子資料夾)。不採分類式(partitional)分群的原因是其分群容易受初始值(群心)及群數的影響。階層式分群法除了較不受到這些因素的影響外，也有助於後續階段的遞增式學習(詳見下階段的說明)。

[3] 移除停用字及字根還原部份僅針對英文資料集處理，中文資料則是進行中文斷詞(採長詞優先法，Wong & Chan 1996)，再抽取特徵字。

[4] 最一開始只有正當郵件及垃圾郵件兩群，當郵件異質性高時正當郵件及垃圾郵件都有可能產生自己的子群。

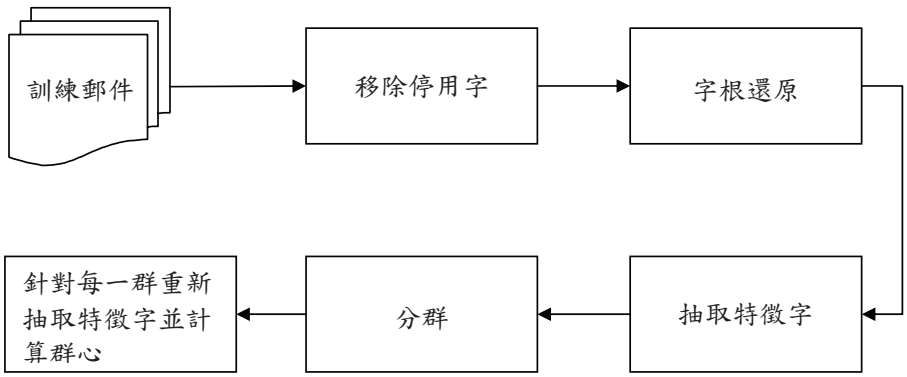

圖 3：訓練分類器之步驟

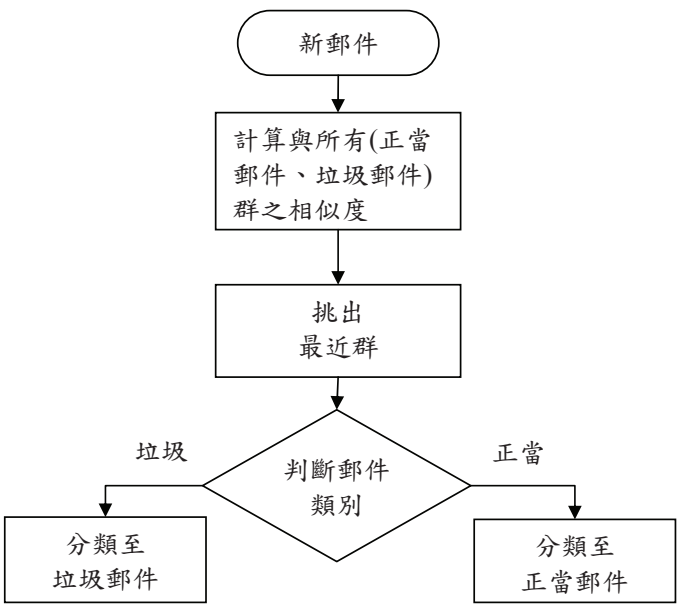

圖 4：郵件分類之過程

當上述的分群分析結束後，就可以對新(測試)郵件作類別的判定。其步驟是(參考圖4)：挑出與新郵件最近似的群，並將新郵件分類至該群所屬的類別。我們所用的方法是比對新郵件與每群群心特徵字相符的數量[5]。當相符的數量越多，也就表示該群與新郵件的相似度越高。最後我們取特徵字相符數量最多的那一群為新郵件的最近群，並據以判斷新郵件的類別。

[5] 此處相似度的計算方法，本研究由實驗不同的資料集發現採布林法(而非由特徵字之權重向量計算Cosine夾角相似度)來決定最近群可得到較佳的效果。

二、遞增式學習階段

本階段的目的在於使分類器具有遞增學習的能力。分類器遞增學習的目的有二：其一為當有分類錯誤的情形發生時，能找出錯誤原因並加以修正(重新萃取特徵字)，使下次不會有類似的錯誤；另一則是當有新的觀測資料被分類時，能學習該新觀測值的特性，並修改、微調現有的分類架構。

如前所述，由於電子郵件通常具有主題漂移的特性，垃圾郵件的主題經常隨著環境的變化而變化，所以分類器必須能遞增學習，適切地調整分類架構，才足以判斷新的資訊。上階段所提之分群分類法可以隨著新郵件的加入而改變其架構，因此適用於遞增式學習。本階段的流程如圖 5所示，其中判定新郵件類別的步驟與上階段相同，以粗體包含的部分則是遞增學習的過程。

遞增學習的精神在於每次判定新郵件類別後，會將該郵件加入所屬之最近群，並將該群之特徵字重新萃取以適應新郵件的加入。另外，為偵測新主題的產生，當新郵件加入某一群後，會同時檢查該群是否滿足分裂條件，以確保群體的內聚力及特徵字的代表性。換句話說，透過這樣的機制判斷是否有新的主題產生，若有，則該群會分裂產生若干子群，我們隨即重新萃取分裂群之特徵字。每次僅針對一個群體來進行微調以達到遞增式學習的目地。圖 6 顯示偵測新主題之示意圖。

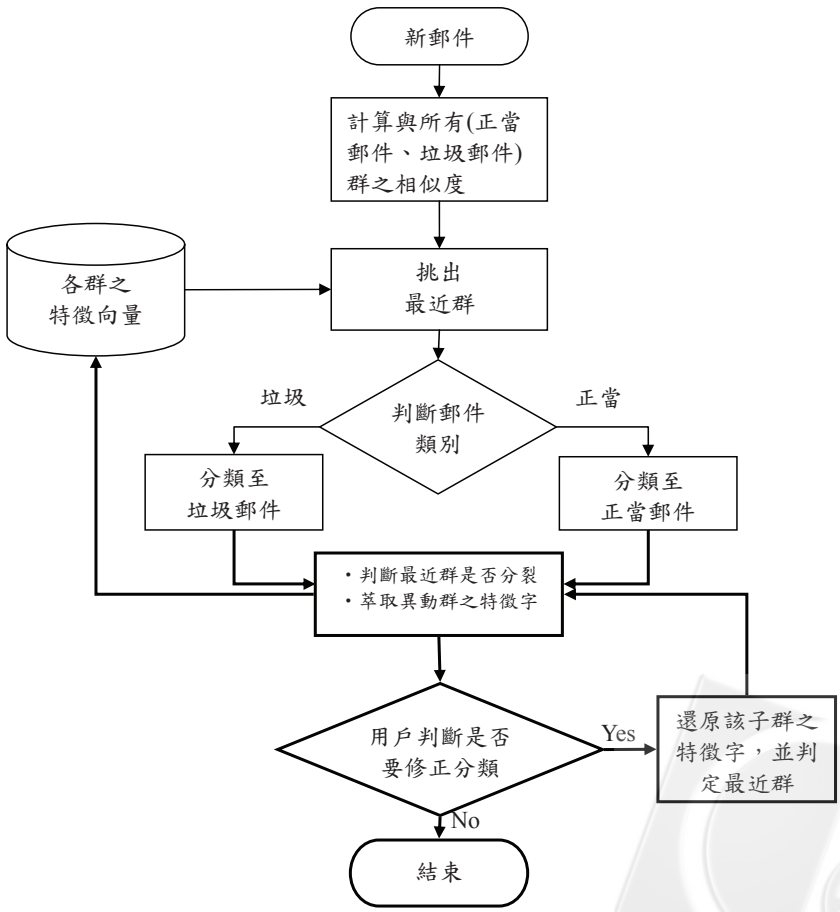

圖 5：分類器遞增學習流程

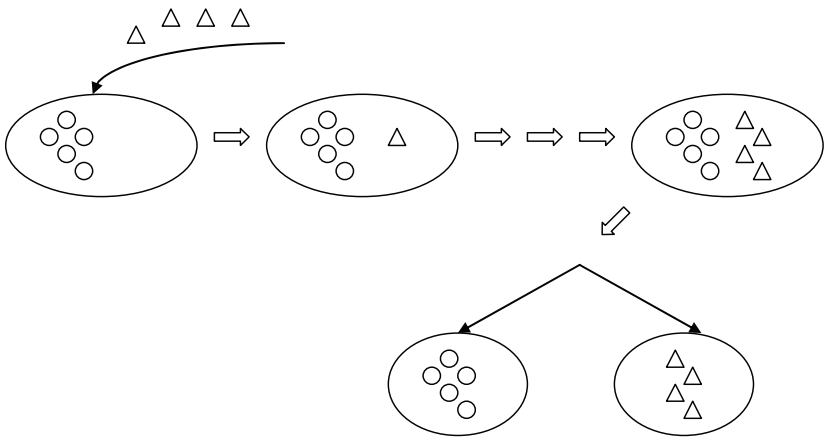

圖 6：偵測新主題之示意圖

由於此分類結果不一定正確，因此此分類結果將以圖形介面呈現給使用者，若使用者認為分類無誤則結束；若認為有誤，可透過介面將錯誤分類的郵件搬移至正確類別，此舉會同時觸發修正程序，對原本郵件所在群重新萃取其特徵字，再針對該郵件找出所屬類別(使用者指定的類別)之最近群，加入該群，判斷該群是否分裂，萃取其特徵字，完成修正程序。此遞增學習步驟先將分群結構改變，再詢問使用者是否分類正確的作法，主要是因為使用者並非即時處理新郵件，大部分以批次處理；此作法可以加快分類器之遞增學習，可避免因使用者之因素而延遲對垃圾郵件主題漂移之敏感性。

肆、實驗與討論

本研究共進行了四個實驗，實驗一及三為純英文資料集測試，實驗二及四則採中文測試，中文部份會夾雜著部份英文；這符合一般華人的使用習慣，通常文字中會夾雜一些專有名詞或是無法以適當中文解釋的原文。實驗一與實驗二之設置是在沒有資料偏斜及沒有遞增學習的情境下測試，主要是用來佐證分類器的效能(無論是對中文或英文電子郵件應皆能有不錯的效能)；實驗三則在資料偏斜的情況下進行測試；實驗四是針對遞增學習的部份來加以測試。由於KNN目前應用於垃圾郵件分類上相當廣泛 (Delany et al. 2005; Delany & Cunningham 2006; Fdez-Riverola et al. 2007; Blanzieri & Bryl 2007)，故本研究之前三個實驗皆以KNN作為比較基準。

一、實驗一

本實驗採用的英文資料集是以Spamassassin[6]中擷取的非垃圾郵件(檔名以ham結尾)

[6] Spamassassin 郵件資料集，http://spamassassin.apache.org/publiccorpus；抓取時間94年2月，當時共8624封ham，現已刪減。

當成正類別(共8624封)，而以E. M. Canada[7]中2002年1月至2003年12月的垃圾郵件當負類別。其中E. M. Canada所提供的郵件比較雜亂，有多種語言及編碼方式夾雜其中，例如有繁體中文、簡體中文、英文及義大利文，而部份郵件中沒有text part，故本研究以程式進行簡單的資料篩選，只留下MIME中附有英文text part的郵件(共15213封)，若有無法正確解碼之郵件，則一律剔除[8]。本實驗中僅使用初始的知識來分類，亦即，在分群並抽取特徵字後即馬上進行分類，並無任何遞增學習的過程。

實驗結果如表1所示，不論是F值或是正確率，KNN皆略優於ICBC。這是由於在資料偏斜不明顯的情況下，以一般方法(如TFIDF)所擷取的特徵字尚具各類別代表性。KNN在決定測試資料的類別時，是比對所有訓練資料，找出二者特徵字向量最大的相關程度；而ICBC只跟群心的特徵字向量作比對，找出測試資料與某群最大的相關程度。在F值或是正確率上，自然不若KNN以窮盡比對方式而得的結果。

表 1：田英文資料集實驗數據

		Precision	recall	F	accuracy
ICBC	正當郵件	0.93177	0.92302	0.92738	0.94730
	垃圾郵件	0.95762	0.96100	0.95931	
KNN	正當郵件	0.97538	0.92461	0.94931	0.96416
	垃圾郵件	0.95927	0.98657	0.97273	

但是就因為KNN是採窮盡比對的處理方式，其所花費的執行時間就相當冗長。為了比較KNN與ICBC的執行時間，我們以相同的英文資料集合，分別以10封、20封、…、100封當成訓練資料，測試資料則每次皆固定200封，分別針對ICBC及KNN來計算分類所需時間。其結果如圖 7所示。

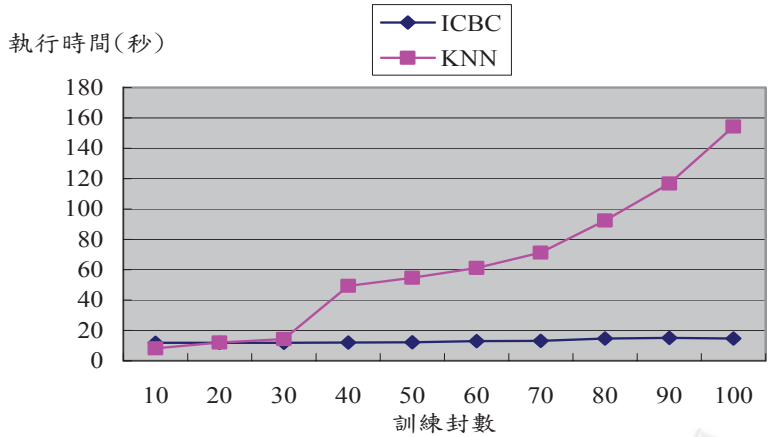

圖 7：分類器執行時間比較

[7] E. M. Canada垃圾郵件資料集，http://www.em.ca/~bruceg/spam

[8] 這裡所謂「無法正確解碼」指的是郵件的編碼無法由JavaMail API解出。即該email不是由一般的Mail Client寄出。

從圖 7中可以發現ICBC的執行速度明顯優於KNN，且較穩定。這是由於ICBC是以群為單位作比對，而KNN是以文件為單位作比對。當訓練封數愈多，KNN比對的時間愈長，而ICBC則較不受訓練封數多寡之影響(群數的增加較不顯著)。

二、實驗二

實驗資料集是取自中國復旦大學李榮陸博士建立的「文本分類語料庫」[9]，原始資料為簡體中文，故利用簡繁轉換程式將資料轉換成繁體中文。從中隨機抽取6個類別來進行訓練及測試，每個類別皆隨機抽取50封當訓練，200封當測試。主題分佈情形如表 2。實驗結果如表 3所示。

表 2：中文訓練主題

類別	主題	訓練個數	測試個數
正當郵件	Space	50	200
	Computer	50	200
	Sport	50	200
垃圾郵件	Agriculture	50	200
	Politics	50	200
	Environment	50	200

表 3：中文資料集實驗數據

		precision	recall	F	accuracy
ICBC	正當郵件	0.92797	0.92333	0.92565	0.92583
	垃圾郵件	0.92371	0.92833	0.92602	
KNN	正當郵件	0.96623	0.93000	0.94777	0.94875
	垃圾郵件	0.93253	0.96750	0.94969	

由表 3可看出，在中英文夾雜的情況下，其績效皆較純英文垃圾郵件過濾之表現低。其原因可能是：(1) 未使用停用詞表，在中文停用詞的相關研究較少，故並無移除停用詞；(2)中文斷詞不易，以及中英文夾雜，故特徵空間較為混亂、資料複雜度較高，分類更不易，不過整體而言仍能達到滿意的效果(各項指標皆達90%以上)。而ICBC與KNN之結果比較，則與實驗一中所述相同。

[9] 中文自然語言處理開放平台，http://www.nlp.org.cn/docs/doclist.php?cat_id=16&type=15。

三、實驗三

本實驗是在測試ICBC處理資料偏斜的能力。傳統的分類方法(如決策樹及決策法則)是以資料屬性來當作節點或項目集合(item sets)，當類別間訓練範例數的數量極度不對等時(不平衡的類別分佈，或簡稱資料偏斜)，會導致其建構出的模式中，節點或項目集合皆出自於範例較多的類別，而範例較少的類別因為本身的文件數不多，故總字數有限，再經由特徵字挑選的階段後，剩下用來當節點或項目集合的個數寥寥無幾。那麼分類結果就大多會指向範例較多的類別。

由於本研究的分群機制是透過TFIDF來挑選特徵字，所以資料中若有偏斜的類別分佈就會影響到特徵字的挑選。故本研究在分群後會再針對每一群進行二次挑選。其作法是先移除停用字，並以字頻(TF)當成挑選的基準。由於是針對每一群分別挑選，故資料的純度會較高，並且對每群都抓取等量的特徵字，故資料偏斜影響的程度會變小。

在緒論中提到，偏斜的類別分佈會造成範例較少的類別的召回率過低，故本節的重點將放在召回率的衡量上。實驗設計如下：由實驗一(英文資料集)的正當郵件中抽取50封(分別以10封、20封、30封、40封及50封當訓練集進行五次試驗) ，而由垃圾郵件中抽取200封。亦即，在訓練資料集中正當與垃圾郵件的比例依次為10:200、20: 200、…以此類推，測試資料則固定為100封。

實驗結果如表 4及圖 8所示，整體而言，ICBC面對資料偏斜情況下的召回率皆高於KNN，隨著資料偏斜程度越小，召回率當然也越來越高。這是由於KNN是將新範例(測試範例)與舊有的範例一一比對，並找出最相似的範例。因此KNN一般應較不受到資料偏斜的影響。但是在文件分類的情形下，由於採用全面性特徵字的擷取，所以所擷取的特徵字無法代表範例較少的類別，分類結果仍會指向範例較多的類別；相對地，ICBC特徵字的挑選是以群為單位，且對每群都抓取等量的特徵字，故其分類結果受資料偏斜的影響較小，較有機會避免指向範例較多的類別。

表 4：資料偏斜情況下之召回率比較

偏斜比率	ICBC	KNN
10:200	0.46	0.22
20:200	0.74	0.59
30:200	0.86	0.67
40:200	0.92	0.69
50:200	0.90	0.71

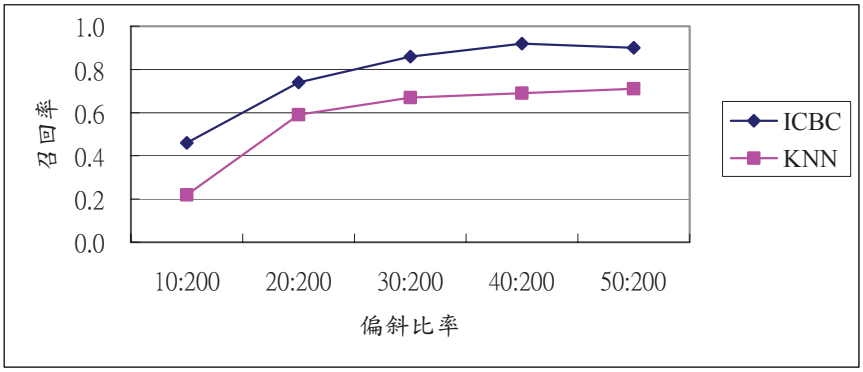

圖 8：ICBC資料偏斜實驗結果

四、 實驗四

為了偵測主題的變化，分類學習法必須具有遞增學習的能力，故設計具有主題變化的測試環境以進行實驗。首先從中文資料集(與實驗二相同)中挑選四個主題各20封當正當郵件及垃圾郵件，再使用另兩個主題當遞增學習訓練資料及測試資料，遞增學習各20封及測試各30封，詳細之資料運用如表 5所示。

本實驗欲得知當有新主題進入分類器後，ICBC能否以遞增學習的方式有效偵測出主題的變化，並將測試範例予以正確歸類。其遞增學習資料分別以5封、10封、15封及20封輸入分類器，而以相同的30封測試資料進行測試，最後計算新主題的測試範例被分類至對應類別的比例(召回率)，並判斷是否有越來越好的趨勢來獲知遞增學習是否有其成效。

表 5：遞增學習之資料集

類　別	主　題	郵件個數
正當郵件	Energy	20
	Electronics	20
垃圾郵件	Literature	20
	Education	20
遞增學習資料	Law	20
	Military	20
測試資料	Law	30
	Military	30

實驗結果如表 6及圖 9所示。當正當郵件及垃圾郵件分類完成後，初始知識架構訓練完成後並未包含Law及Military的資訊。此時先測試沒有遞增學習前的召回率，由於目前知識架構中並不具Law及Military的知識，故召回率分別只有0.467及0.433(表 6)。隨後

Law及Military的遞增學習資料加入訓練範例，利用ICBC所得之召回率也越來越高(0.9及0.767)，這即表示ICBC隨著新主題的文件越多時，已偵測到新主題，並對新主題加以學習，以降低錯誤率。同時，由於ICBC進行遞增學習，可以顯著減少每次使用所有資料重新學習的成本負擔。此結果說明了本研究方法確能有效地偵測並處理垃圾郵件概念偏移的問題。

表 6：ICBC遞增學習召回率

訓練個數	Law	Military	Average
0	0.46667	0.43333	0.45000
5	0.43333	0.63333	0.53333
10	0.60000	0.66667	0.63333
15	0.90000	0.66667	0.78333
20	0.90000	0.76667	0.83333

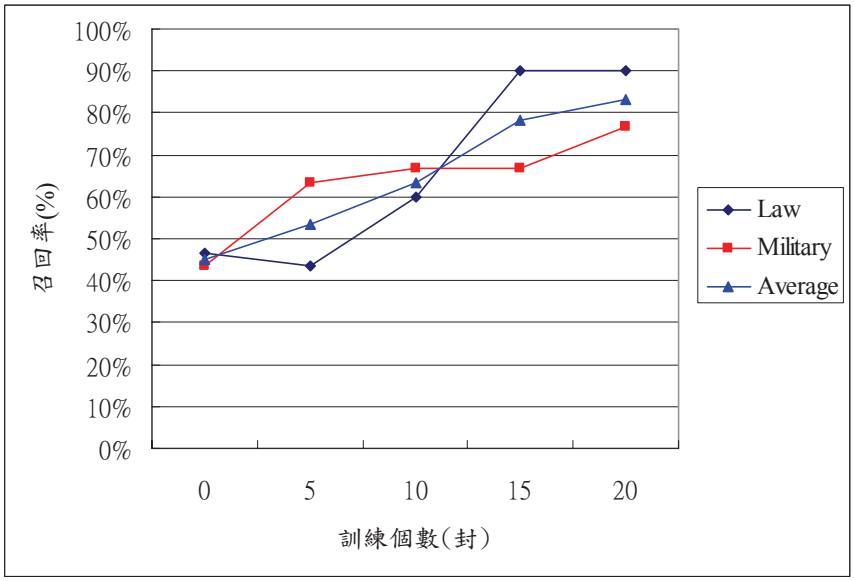

圖 9：ICBC遞增學習實驗結果

伍、結論

隨著網際網路的普及與電子郵件的廣泛使用，電子郵件在我們的生活中已是不可或缺，然而隨著垃圾郵件的數量日益增多，因此如何事先偵測並過濾大量垃圾郵件問題便成為了現今研究電子郵件處理的重要議題之一。

本研究提出ICBC—以分群為基礎之遞增式分類法來提高過濾垃圾郵件的績效。ICBC是以分群為基礎之分類方式，利用分群演算法將大概念分成數個子概念，並找出每群等量具代表性的特徵。藉由這樣的處理，小類別與大類別均有等量的特徵字代表，可以有效解決垃圾郵件資料偏斜的現象；同時該分群結構可以隨著新郵件的加入而改變，使得ICBC有遞增學習的能力，針對電子郵件主題漂移的特性，能以最快的速度及最少的成本來適應環境的變化，減少分類的錯誤。

本研究由四個實驗來瞭解所提分類器的效能。實驗一及實驗二結果顯示，在資料偏斜不明顯的狀況下，ICBC的表現與KNN相當，但執行速度上則明顯較KNN有效率。另外，在資料偏斜的情況下(實驗三)，ICBC對小類別資料的召回率高於KNN，可知ICBC可以有效地處理資料偏斜的情況。最後，實驗四的結果顯示ICBC因有遞增學習的能力，可以適應郵件主題變化而減少錯誤。

本研究所提ICBC法，雖能初步解決電子郵件資料偏斜與主題漂移的問題，但仍有改善的空間。首先是若初始分群的結果不佳，可能會影響到一開始的分類的效果(需等到一定量的遞增學習來修正)。因此在未來研究中，可以透過詞網(WordNet)或本體論(Ontology)的輔助來強化各主題或概念間的關係，改善起始的分群。此外，目前ICBC法僅使用內容分析的技術來過濾郵件，若能搭配其他過濾技術，如法則(filtering rules)及線上自動更新黑名單(black list)等，其過濾效果將會更佳。最後，ICBC在處理中文郵件時必須先進行斷詞，此部份為速度上之瓶頸，後續研究將由選擇合適之斷詞演算法來提昇整體的速度。

致謝

本研究受國科會計畫補助(計畫編號：NSC 93-2218-E-251 -001)，特此致謝；作者並感謝兩位匿名審查委員的寶貴意見及建議。

參考文獻

1. Androutsopoulos, I., Koutsias, J., Chandrinos, K.V., and Spyropoulos, D. “Learning to filter spam e-mail: A comparison of a naive bayesian and a memory-based approach,” *the 4th PKDD's Workshop on Machine Learning and Textual Information Access*, 2000.
2. Blanzieri, E. and Bryl, A. “Instance-Based Spam Filtering Using SVM Nearest Neighbor Classifier,” *Proceedings of the 20th International FLAIRS Conference*, 2007, pp. 441-442.
3. Chan, P.K., Fan, W., Prodromidis, A.L., and Stolfo, S.J. "Distributed Data Mining in Credit Card Fraud Detection," *IEEE Intelligent Systems* (14:6), 1999, pp. 67-74.
4. d'Alché-Buc, F. and Ralaivola, L. “Incremental Learning Algorithms for Classification and Regression: Local Strategies,” *American Institute of Physics Conference Proceedings, 2002*.

5. Delany, S.J., Cunningham, P., Tsymbal, A., Coyle, L. "A case-based technique for tracking concept drift in spam filtering," *Knowledge-Based Systems* (18: 4-5), 2005, pp. 187-195.
6. Delany, S.J. and Cunningham, P. "ECUE: A Spam Filter that Uses Machine Learning to Track Concept Drift," *Technical Report*, Computer Science Department, The University of Dublin. Available online at https://www.cs.tcd.ie/publications/tech-reports/reports.06/TCD-CS-2006-05.pdf.
7. Drucker, H., Wu, D. and Vapnik, V.N. "Support Vector Machines for Spam Categorization," *IEEE Transactions on Neural Networks* (10:5), 1999, pp. 1048-1054.
8. Fawcett, T. ""In vivo" spam filtering: a challenge problem for KDD", ACM *SIGKDD Explorations Newsletter* (5:2), 2003, pp. 140 – 148.
9. Forman, G. "Tackling Concept Drift by Temporal Inductive Transfer," *SIGIR '06 ACM, 2006*.
10. Hart, P.E. "The Condensed Nearest Neighbor Rule," *IEEE Transactions on Information Theory*, IT-14, 1968, pp. 515-516.
11. Honda, T., Motizuki, H., Ho, T.B., and Okumura, M. "Generating Decision Trees from an Unbalanced Data SEt," Poster papers presented at the *9th European Conference on Machine Learning* (ECML), edited by Maarten vanSomeren and Gerhard Widmer, 1997, pp. 68-77.
12. Jain, A.K., Murty, M.N., and Flynn, P.J. "Data Clustering: A Review," *ACM Computing Surveys* (31:3), 1999, pp. 264-323.
13. Kiritchenko, S. and Matwin, S. "Email Classification with Co-Training," *Proceeding of CASCON 2001*.
14. Lian, Y. "E-mail Filtering," *Master thesis*, University of Sheffield, Department of Advanced Software Engineering, 2002.
15. Luo, X. and Zincir-Heywood, N. "Comparison of a SOM based sequence analysis system and naive Bayesian classifier for spam filtering," *Proceedings of IEEE International Joint Conference on Neural Networks -- IJCNN 2005*, pp. 2571 – 2576.
16. McCallum, A. "Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering," *http://www.cs.cmu.edu/~mccallum/bow/*, 2002.
17. Monard, M.C. and Batista, G.E.A.P.A. "Learning with skewed class distributions," *Advances in Logic, Artificial Intelligence and Robotics (LAPTEC'02)*, 2002, pp. 173-180.
18. Nigam, K., Mccallum, A.K., Thrun, S., and Mitchell, T. "Text Classification from Labeled and Unlabeled Documents using EM," *Machine Learning* (39), 2000, pp. 103-134.
19. Payne, T.R. and Edward, P. "Interface Agents that Learn: An Investigation of Learning Issues in a Mail Agent Interface," *Applied Artificial Intelligence*, 1997, pp. 1-32.
20. Porter, M.F. "An algorithm for suffix stripping," *Program (Automated Library and Information Systems)* (14: 3), 1980, pp. 130-137.
21. Sahami, M., Dumais, S., Heckerman D., and Horvitz, E. "A Bayesian approach to filtering junk e-mail," *AAAI-98 Workshop on Learning for Text Categorization*, 1998.

22. Sebastiani, F. "Machine learning in automated text categorization," *ACM Computing Surveys* (34:1), 2002, pp. 1-47.
23. Stanley, K.O. "Learning concept drift with a committee of decision trees," *http://citeseer.ist.psu.edu/ stanley01learning.html*, Computer Science Department, University of Texas-Austin. 2001.
24. Weiss, G.M. and Provost, F. "The effect of class distribution on classifier learning," *Technical Report ML-TR-44*, Department of Computer Science, Rutgers University, 2001.
25. Wong, P.K. and Chan, C. "Chinese Word Segmentation based on Maximum Matching and Word Binding Force," *Proceedings of 16th International Conference on Computational Linguistics*, 1996, pp. 200-203.
26. Wu, H., Pang, T.H., Liu, B. and Li., X. "A Refinement Approach to Handling Model Misfit in Text Categorization," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002, pp. 23-26.