兩階層式垃圾郵件過濾機制之研究
A Study of Two-tier Filtering Schemes for Anti-spam

***葉生正[1]　蘇民揚[2]　張僩鈞[1]**
[1]銘傳大學資訊傳播工程學系所 [2]銘傳大學資訊工程學系所
***E-mail: peteryeh@mcu.edu.tw**

摘要

垃圾郵件氾濫至今，造就各種防堵機制群雄並起，且在內容過濾比對機制中又以機械學習理論的支援向量機(SVM)與貝氏演算法(Naïve Bayes)最為著名。故本論文主要擷取 SVM 以超平面快速分類的特點及貝氏演算法的彈性，研究設計一兩階層式之垃圾郵件過濾機制。本研究先將中、英文郵件訓練樣本於中文斷詞與英文斷字後，再以資訊增益(Information Gain)計算結果決定 SVM 所訓練之關鍵字。最後，將 SVM 對測試樣本之分類結果，以本論文提出之 4 種邊界距離挑選出落於模糊區間的郵件樣本，經由貝氏機率改良模型進行計分以判斷郵件類別。實驗結果呈現 4 種邊界距離擷取出資料再計算後的準確率皆有所提升，其中又以最大距離或平均距離的改善最顯著；且若加上在最佳化模式的預測下，中、英文郵件整體分類的精確度皆達97%以上，因此可驗證本研究提出之兩階層式過濾機制與貝氏演算法改良模型的可行性。

關鍵詞：支援向量機、貝氏演算法、資訊增益。

Abstract

The Support Vector Machine (SVM) and Naïve Bayes are well-known machine-learning algorithms for the application of content filtering against spam. On the basis of fast classification through the hyper-plane of SVM and flexible threshold setting of Bayes, this paper proposes a two-tier filtering scheme which combine SVM and new Naïve Bayes model for anti-spam. In the first tier, Information Gain is the way to decide keywords for training vector of SVM. The paper also provides four kinds of margin of the hyper-plane, and picks out the sampling data which locates on the scope for the second tier Bayesian probability calculation to decide the classification. The experimental results indicate that all kinds of the margin setting bring the improved accuracy about 1% to 4%, especially the Maximum Distance and Average Distance Margin. Additionally, the optimal model performs the total accuracy of Chinese and English sampling mails above 97%. However, the proposed two-tier filtering scheme and new Naïve Bayes model were verified with availability.

Keywords: SVM, Naïve Bayes, Information Gain.

1. 前言

透過電子郵件所衍生出的變相問題，不外乎是早期的「病毒郵件」以及現今吵得沸沸揚揚的「垃圾郵件」。且此二者又能相互結合，例如當病毒郵件感染了某一部電腦後，病毒作者再將中毒電腦名單販賣給垃圾郵件發送者，使其利用該電腦當作跳板來發送垃圾郵件形成更大的危害。這種病毒郵件與垃圾郵件交相賊的情況，令病毒為垃圾郵件創造更多攻擊機會，至少已有超過 30%的垃圾郵件都是透過此種中毒電腦的方式來發信[1]。

目前各家防毒廠商偵測病毒的技術日趨成熟，更新病毒碼之效率已臻於快速且穩定，從發現新病毒到釋出新病毒碼時間相差不會大於八小時，故對於病毒信件的攔截，防毒軟體已可做到幾近完善；但針對「垃圾郵件」的防堵，由於人人對信件的合法性與非法性的定義不同，因此沒有百分之百精確的方法。可是對於大多數人而言，只要郵件信息本身有其「目的」，有想要表達廣告、商業的意圖和立場，皆會被認定為「垃圾郵件」。然而在過濾垃圾郵件的方法中，以機器學習理論而言，由於訓練方式與演算方法的不同，造就出精確率的結果就不盡相同。本研究主要目的即在於提出結合支援向量機(Support Vector Machine, SVM)與貝氏(Naïve Bayes)兩種演算法的特性，先以 SVM 做初步分類，將落入模糊區的郵件交由 Bayes 進行機率判斷並給定分數，如此對郵件進行所謂兩階層式的過濾與分類，以期提高防堵垃圾郵件的精確率，並降低錯分正常郵件的誤判率[2][3]。

本論文共分 5 個章節，第 1 章簡述本研究之動機、背景與目的；第 2 章介紹相關的垃圾郵件判別過濾方法；第 3 章則詳述本研究論文之實驗方法與步驟；第 4 章將呈現所提出之機制效能分析的結果。第 5 章則歸納研究結果並提出未來方向。

2. 相關研究與技術

本章將介紹目前常見之垃圾郵件判別與過濾的方法和理論，分別是即時性黑名單(Real Time Black-hole Lists, RBL)、DCC、Razor、Pyzor、支援向量機(SVM)和貝氏演算法(Naïve Bayes)等[2][5][6]。

2.1 即時性黑名單

面對全球 Spammer 每日更迭的 IP 位置、主機名稱，網路上已有專門收集資訊，建立完整黑名單資料庫，稱作「即時性黑名單」(Real Time Black-hole Lists, RBL)，各郵件伺服器可透過即時查詢 RBL 的資料以判斷郵件是否為垃圾郵件來源，而後決定是否拒收相應的郵件。然而目前線上即時黑名單之管理制度寬鬆不一，較寬者是將對方加入黑名單前會主動發信通知 SPAM 的郵件主機管理者，要求改善；若主機管理者遲遲未處理，才會將該主機列入黑名單；較嚴的作法則為一旦使用者舉報或經主動偵測認定為可疑對象，一律列入黑名單。這種方法雖然可有效檔下非常多數的廣告信件，但由於過度制式，誤攔率必定提高，因此衍生出許多問題。現今 RBL 服務商大部分是非營利性機構或民間組織，基本上皆屬免費服務，然而在使用這個技術之前，還是必須慎選一些聲譽良好且值得信任的服務網站，如目前公認較準確且免費的 ORDB(www.ordb.org)。因為目前各種黑名單資料庫皆難以保證其正確性和及時性，若使用不完善的資料庫，如北美的某些 RBL 網站包含了我國大量的主機名字和 IP 位址，其中有些是早期 Open Relay 所造成，有些則是由於誤報所造成。這些遲遲不糾正的資料庫，在一定程度上必定阻礙了我國與北美地區的郵件聯繫。

2.2 DCC、Razor 與 Pyzor 法

DCC、Vipul's Razor 與 Pyzor 法是將信件內容取樣，並計算其檢查碼（Checksum），透過網路向某些集中式資料庫查詢，以辨別此封信件是否已經遭到其他人提報為廣告信件。此三種技術透過許多下游裝設收集程式的電子郵件伺服器，取得回報已知垃圾郵件訊息的 checksum value 或密碼雜湊（Cryptographic Hash），然後製作成該郵件訊息的特徵或類似指紋的數據供其餘下游的電子郵件伺服器查詢。在 Spammer 短時間內大量寄發了上千或上百萬封相同郵件訊息之後，該郵件訊息很容易會被資料庫端的判斷主機發現，並依特徵被辨識出來。其中 Razor 是一種線上的垃圾郵件比對資料庫。如果有一封垃圾信同時寄給上百位使用者，它會記錄所有的垃圾郵件，計算其指紋碼 (使用 SHA 雜湊演算法)，然後存至 Razor 之線上資料庫中。而事後郵件主機可以經由查詢 Razor 之線上資料庫來判斷該封郵件是否為垃圾郵件，若是，則會自動的將這封垃圾信封鎖。DCC 分散式檢查碼交換為一大宗郵件辨識技術，為免費原始碼。其作法為讓啟動 DCC 的 Mail Server 在接收到信件時產生信件相關的檢查碼(Checksum)，再將這些 Checksum 通報至 DCC 伺服器，DCC 伺服器會自動更新並告知郵件伺服器此檢查碼出現的次數。其中不同內容產生的 Checksum 會完全不同，其數值代表著此郵件曾在其它郵件伺服器上被傳送的次數。當伺服器發現此一次數超過了管理者所設定的門檻時，就可認定此一為大宗寄發的垃圾郵件。而由於前述之“Razor”線上垃圾郵件比對資料庫所使用的伺服器並非免費原始碼，所以“Pyzor”的目的即是要用來取代 Razor。但上述三種方法最大好處是準確性高以及節省 CPU 計算資源，因為它僅需針對信件內容的某幾段取樣，無須逐字比對，對於大量寄送重複內容的廣告信件（包括電子報），若郵件的特徵符合了資料庫中的特徵，皆能夠有效大量攔截。此外由於並不是針對完整信件內容做 checksum，因此即便廣告信利用一些小技巧更動信件內容，例如：信件首行的收信者姓名、內容夾雜空白行及無意義單字等，都不容易影響其算出的 checksum 結果。

2.3 支援向量機

SVM是一種用在機器學習的演算法，其主要的概念就是針對訓練資料集，利用定義的特徵值，以訓練函式計算出一個最理想的超平(Hyper-plane)，此後透過此超平面分類測試資料判斷其準確率，當準確率超過一標準值且具意義時，即可分類新的未知資料，將所有欲分類的資料快速分類至正確的類別。SVM 的基本方法論是在建構一個線性分割超平面(The linear separating hyper-plane)，以線性的模式去執行非線性的分類範圍，而此超平面則是最大化正集(Positive class)與負集(Negative class)之間的距離所得出。由於分類垃圾郵件時會有大量的資料與特徵值，其訓練過程將耗費不少時間，但使用支援向量機透過數學的方式可以讓事後整體分類速度大幅提升，達到最佳的分類效果[4][5][8]。

2.4 貝氏演算法

貝氏演算法運用於垃圾郵件的過濾效果卓越，Spam Conference 的創辦人 Paul Graham 曾表示過，使用此方法的過濾軟體 CRM114 一個月內偵測的精準度達 99.75%，同時 7000 封信件中，只有 8 封有誤判的情形。而 Bayesian Filtering 即是應用機率學中的貝氏定理來對郵件進行分類，其涵義主要是假設字與字之間存在著獨立的關係，最早的第一篇應用在垃圾郵件的論文是在 1998 年由 Sahami 學者提出，該分類的演算法則說明如下[1]。

給定 N 個文件的種類（在垃圾郵件問題上，N=2，一是正常信件，一是垃圾信件），$C_1, C_2, \cdots C_N$ 表示不同文件的種類，以英文字母 d 來表示一個文件，其中 d 文件包含了 $W_1, W_2, \cdots W_m$ 個不同的單字（Keywords），這裡的目的是在已知 d 文件由不同單字組成的前提下，該 d 文件屬於 C_j ，$j = 1, 2, \ldots, N$，去求得一個機率值 $P(C_j \mid W_1, W_2, \cdots W_m)$ ，其計算公式如式(1)：

$$P(C_j \mid W_1, W_2, \cdots, W_m) = \frac{P(C_j) * P(W_1 \mid C_j) * P(W_2 \mid C_j) * \cdots * P(W_m \mid C_j)}{P(W_1, W_2, \cdots, W_m)}$$
$$= \frac{\prod_i P(W_i \mid C_j) * P(C_j)}{P(W_1, W_2 \ldots W_m)}$$
$$= \frac{P(C_j)}{P(W_1, W_2 \ldots W_m)} \prod_i \frac{P(C_j \mid W_i)}{P(C_j)} \quad (1)$$
$$= P(C_j) \frac{\prod P(C_j)}{P(W_1, W_2 \ldots W_m)} \prod_i \frac{P(C_j \mid W_i)}{P(C_j)}$$

未知文件的類別則為式(1)中所求得機率值最大的C_j（意指最可能）。式(1)中$P(C_j)$代表所有文件中，C_j類別的可能性，計算上是以「C_j類別文件的總數」除以「全部的總文件數目」。$P(W_k \mid C_j)$則代表在給定C_j類別的前提下，W_k出現在C_j類別的比率（「出現關鍵字W_k的信件數」除以「該類別的總信件數」所得的條件機率）。而$P(W_1, W_2, \cdots, W_m)$代表關鍵字W_k出現在全部總文件的機率。

分類器依下式將文件 d 歸類為C_k類（設C_k為某一已知類別），則從 d 在各類別得到機率的最大值，來判斷其所屬類別，如式(2)所示：

$$P(C_k \mid W_1, W_2, \cdots W_m) = \underset{j=1\ldots N}{Max} P(C_j \mid W_1, W_2, \cdots W_m) \quad (2)$$

貝氏演算法在郵件系統的應用十分廣泛，許多著名軟體如 SpamAssassin 已內嵌經此方法作垃圾郵件的過濾。其最大優點在於經由貝氏演算法計算後，會針對郵件產生一組易於識別機率分數，屆時與使用者於郵件伺服器所設定之門檻值比對，若分數超過門檻值，則判定此文件為垃圾郵件，而門檻值又可依經驗設定，若發覺過濾器誤刪太多正常信件，可以將門檻值訂得較為寬鬆，對個人化而言，貝氏分類法算是較彈性的一種規則[3]。

3. 兩階層式的過濾機制

本研究論文主要是設計一兩階層式垃圾郵件過濾機制，取 SVM 分類演算法所能形成之模糊區間，及其快速分類的特性；搭配貝氏演算法透過樣本的大量建立，提高機率分數給予之準確性，期望二者的結合優於單一演算法，展現更佳的過濾效能。

在郵件樣本的選取上，分為訓練樣本與測試樣本。訓練樣本旨在做為 SVM 關鍵字之挑選以及貝氏關鍵字資料庫的建立。挑選上依目前一般使用者收到信件類別的比例，垃圾郵件比正常郵件為 4：1，故在英文郵件的訓練樣本中垃圾郵件為 800 封，正常郵件 200 封；中文的訓練樣本亦以垃圾郵件 800 封，正常郵件 200 封進行訓練。測試樣本的數量，中英文郵件皆為垃圾郵件與正常郵件各 500 封，共 1,000 封進行測試。樣本來源方面，英文郵件使用 Ling-spam、TREC Spam Corpus 以及個人收集之混合型樣本；中文郵件則為收集多位使用者信件之混合型樣本，其中測試樣本與訓練樣本為個別收集，刻意避免文件之重複，最後進行效能分析評估。

本機制之流程主要為程式訓練、挑選 SVM 關鍵字、SVM 分類、決定 SVM 分類後模糊區間範圍，以及貝氏演算法計算機率分數等，以下將詳述其步驟程序。

3.1 訓練關鍵字

由於辭典式斷詞法須具備一龐大資料庫，且須人工建立，定期維護、更新，加上中研院 CKIP 斷詞軟體價格高昂，故本研究訓練關鍵字的方式主以統計式斷詞法的概念，於程式紀錄字詞出現次數，包括在垃圾郵件中出現次數、正常郵件出現次數、垃圾郵件中出現封數，與正常郵件出現封數等資訊。其中「出現封數」為關鍵字在不同郵件類別出現過的封數紀錄，可適用於決定 SVM 分類的 Information Gain 關鍵字挑選條件，以及貝氏資料庫中關鍵字機率分數的紀錄[7]。

在關鍵字的擷取方面，中英文樣本方法不一。英文主要以空白字元或標點符號來決定欲擷取單字之位置，而後挑選出來做次數、封數的紀錄。由於語文的特性，英文在資訊索引上較容易識別，且單字通常即可代表完整意義，故關鍵字的決定則以其在文件中出現之次數、封數的頻率為主。然而有些單字出現頻率甚高，但其具備「關鍵」字特性的地位並不高，例如人稱代名詞、連接詞、定冠詞等，可濾過不需比對。

至於中文字詞的擷取，由於一個全形中文字大小為 2bytes，且文句中字與字，或詞與詞間並沒有明顯的空白或標點符號隔開，加上目前的中文郵件事實上多屬中英文夾雜信件(英文字母大小為 1byte)，故在斷詞方法上本研究先以 ASCII 碼比對，分離出英文或中文字，而後再進行英文關鍵字的訓練，以及中文字串的斷詞，紀錄其出現次數與封數。針對中文斷字本研究於程式中並無使用類似 CKIP 的文字資料庫，主要原因在於其所佔空間與資源量大、申請價格高昂、對詞庫判斷依賴性高(若詞庫沒有相同的詞句，則全部斷成一個個單獨的中文字)；本研究以每兩個中文字做斷詞，蓋因中文多以二字組合即具詞意，超過二個字以上的詞句，實際上也以二個字為基本單位。另外，以每兩個字做斷詞或許有些詞不具意義，但在樣本數量足夠的訓練情況下，具備意義的關鍵字排名亦會超越前序。

3.2 關鍵字轉換 SVM 特徵向量

經由 Information Gain 計算關鍵字分數後，本研究分別擷取「垃圾信關鍵字」與「正常信關鍵字」，而擷取的方向分為二：

1. Overlapping Keywords: 垃圾信關鍵字的條件為其出現在垃圾信的封數大於在正常信出現之封數，且依 Information Gain 值由大至小排序，

取前 100 名；同樣的正常信關鍵字為其出現在正常信之封數大於在垃圾信出現的封數，依 Information Gain 值排序取前 100 名，然而在訓練樣本中，正常信為 200 封，加上一般正常信的性質為字數較少、無特定類別，因此最後訓練出來符合上述條件的正常信關鍵字自然比垃圾信關鍵字數量少。故在此中文信件所使用的關鍵字共 132 個，英文信共 136 個。

2) Independent Keywords: 垃圾信關鍵字條件為其在正常信件中出現次數為 0 者，再依 Information Gain 值排序挑選前 100 名；正常信關鍵字的挑選以此字詞在垃圾信件中出現次數為 0，再依 Information Gain 值排序挑選。如同上述提及正常信件的性質，符合此條件而產生的中文郵件關鍵字共 111 個，英文郵件共 109 個。

之後依決定的關鍵字，轉換成SVM向量，由於本研究以LIBSVM作為第一階段分類工具，故轉出向量的格式須符合程式要求為：

<label> <index1>:<value1> <index2>:<value2>…

其中 label 為信件所屬類別標籤，若欲分類的資料集類別為二，則 label 可標示為+1、-1；index 則為向量編號，視當初所選特徵數目多少而定；而 value 即為特徵向量之值。以本研究將一封信件轉為向量，垃圾信的 label 值為+1，關鍵字若有 136 個，index 值則為 1~136，而 value 為每個關鍵字在此信件中出現的次數。

3.3 SVM 設定與訓練

在SVM的使用上，本研究主要以LIBSVM程式執行，此程式能因應不同資料類型、使用者對於分類結果之需求，藉由參數設定來達成期望的效果。在初始環境設定上，首先主要在決定SVM型態與SVM核心函數。SVM型態主要分為Classification與Regression。二者間的差別，依定義Classification主要意圖在產生使未來測試資料錯誤可能性最低的最佳邊界範圍，而Regression則為計算出未來預測資料之最佳迴歸數值。本研究「二階層式分類」之主要目的，即是在SVM首先訓練出最佳超平面，之後將遺落在正負邊界範圍內的資料挑選出，交由Bayes進行第二層的過濾，故在SVM型態上選擇Classification [4][7]。

初始環境設定的第二部份 kernel_type，在現實空間中資料並非完全能以線性分類，當遇到無法以線性分割的資料集時，則須借助核心函數將資料從 Input Space 對應到 Feature Space，核心函數定義為 $K(x,t)=\phi(x)\bullet\phi(t)$。

在 LIBSVM 的核心函數中有以下四種類型，其中 γ、r、d 為核心參數：

1) Linear: $K(x_i,x_j)=x_i^{\ T}x_j$

2) Polynomial:

$$K(x_i,x_j)=(\gamma * x_i^{\ T}x_j+r)^d,\gamma>0$$

3) Radial basis function (RBF):

$$K(x_i,x_j)=exp(-\gamma * \left|x_i-x_j\right|^2)$$

4) Sigmoid: $K(x_i,x_j)=tanh(\gamma * x_i^{\ T}x_j+r)$

本研究使用的SVM型態為Classification中的C-SVC，而核心函數則選擇RBF。主要目的在於C-SVC能藉由設定的參數-c與-g，訓練出將錯誤最小化的超平面，而參數-c和-g的訓練，LIBSVM內所附的grid.py程式能經由反覆測試，找出最佳的c值與g值，之後在執行訓練程式時將參數鍵入。

其次RBF為目前較普遍使用的核心函數，能將極複雜非線性分布的資料轉換至特徵空間，另外學者亦建議，選擇核心函數時應優先考慮放射型(RBF)核心函數，因為它具有能分類非線性且高維度的資料，並僅需調整兩個參數(c和g)，不但減少操作上的複雜性與運算時間，亦能達到較高的預測能力。故本研究在初始環境設定上以此二者為主。

3.4 設定邊界範圍

經由SVM分類後的資料，主要以正、負極區分，意即 $y=w\bullet x+b=0$ 計算後大於0者屬於垃圾郵件，小於0者則判為正常郵件。在實際情形中，待預測信件通常與資料庫訓練樣本不重複，又很有可能一封正常信，因出現少許偏向垃圾信的關鍵字，經計算後結果為+0.002，依SVM分類的精神，將毫不考慮地將此封郵件歸類為垃圾信，但其所存在於的平面空間，距離超平面其實是很近的。故為了避免類似的偶發情形，需要設定一範圍，將落在接近 $y=w\bullet x+b=0$ 兩側模糊區的資料挑選出，交由第二層貝氏過濾法計算。故本研究採用四種範圍，框架出需進行第二道過濾的資料：

1) -1 ~ +1: 測試資料經由SVM計算後，其值落在此範圍者，挑選出予以貝氏計算。

2) -0.5 ~ +0.5: 測試資料經由SVM計算後，其值落在此範圍者，挑選出予以貝氏計算。

3) Maximum Distance: 額外訓練一組資料集，挑選其被判錯中正極最大值與負極最小值，以此二者間距離作為下一組資料在SVM分類後而被挑選出的範圍。

4) Average Distance: 額外訓練一組資料集，挑選其被判錯中正極所有值的平均與判錯中負極所有值的平均值，以此二者間距離作為下一組資料在SVM分類後而被挑選出的範圍。

上述*1)*和*2)*屬於固定且對稱之範圍，*3)*和*4)*屬不對稱且會隨訓練資料改變而更動的邊界範圍，若範圍距離越大，被挑選出的資料量會越多，造成原先SVM已正確分類的資料，還會有再次經歷貝氏計算之機會。決定邊界範圍後，將待測資料交由SVM測試(predict.exe)，分數產生後再將落在範圍內的資料選取出，不論其當初SVM判別的準確度，最後交給下一層貝氏法進行雙重過濾。

3.5 修改貝氏過濾機制

不同於支援向量機演算法僅需少許特徵值轉化為向量，並搭配參數設定與核心函數即能找出最佳分割平面，貝氏過濾垃圾郵件之機制需要龐大的關鍵字資料庫，紀錄關鍵字屬於郵件類別的機率分數，提供往後測試信件的機率判定準則。本研究在實作貝氏資料庫方面，主要將信件斷字、斷詞後的關鍵字存於資料庫，另外紀錄該關鍵字屬於垃圾郵件之機率分數，之後針對從SVM模糊區間挑選出的資料，以貝氏資料庫關鍵字的機率分數作運算，並預設一門檻值分數，判斷其為垃圾抑或正常郵件。

根據貝氏演算法，任一關鍵字經由計算後所得之機率分數，可視其為一獨立事件，可對其作數學運算。以一關鍵字「專業」為例，其在垃圾郵件樣本中出現的機率為 0.8889，在文件類別只有二種的情況下，「專業」相對於出現在正常郵件樣本中的機率即為 0.1111。在此很明顯可看出此關鍵字出現在垃圾郵件的機率較高，故此關鍵字可獲得一機率分數 0.8889-0.1111=0.7778，且隸屬於垃圾信的關鍵字，同樣的方法亦可建立一套正常信的關鍵字機率分數。

獲得上述貝氏關鍵字機率分數資料庫後，針對待測郵件的記分，本研究提出四種記分方式，即貝氏機率演算方式之改良，其中 $P(W_i)$ 為垃圾信關鍵字之機率分數，$P(W_i')$ 為正常信關鍵字之機率分數，N_i 為關鍵字在單一文件中出現之次數：

1) $\prod_i P(W_i) - \prod_i P(W_i')$

2) $\prod_i P(W_i) * N_i - \prod_i P(W_i') * N_i'$

3) $\sum_i P(W_i) - \sum_i P(W_i')$

4) $\sum_i P(W_i) * N_i - \sum_i P(W_i') * N_i'$

4. 研究結果分析

首先對英文郵件之SVM訓練過程為將訓練樣本800封垃圾信與200封正常信轉化為特徵向量後，交由grid.py程式訓練，找出最佳c值與g值，以產生最適之model檔案提供未來測試工作。

第一階段在垃圾郵件關鍵字與正常郵件關鍵字之間相互獨立情況下(Independent Keywords)，以Information Gain挑選109個特徵值，轉換向量後得出最佳c值為2048，g值為0.0004882，對負極資料被錯分的懲罰度為5，之後會產生一個model檔，再將原本訓練檔案當作測試檔案，以svmpredict.exe程式進行前測。接著以垃圾信關鍵字與正常信關鍵字間不相互獨立之關鍵字(Overlapping Keywords)所產生的136組特徵向量，經由grid.py訓練得出最佳c值為32，g值為0.001953，同樣w值為5，接續以之產生model檔，並再次進行前測。分析結果以Overlapping Keywords作為特徵向量的分類平均準確性較高，如圖1所示。其中正確率(Spam Precision, SP)反應了過濾系統「找對」垃圾郵件的能力，正確率越大，將非垃圾郵件誤判爲垃圾郵件的數量越少。召回率(Spam Recall, SR)則反映過濾系統發現垃圾郵件的能力，召回率越高，漏網的垃圾郵件就越少。準確率(Accuracy)表示對所有郵件（包括垃圾郵件和正常郵件）的判別正準率。錯誤率(Error)則是對所有郵件的誤判率。

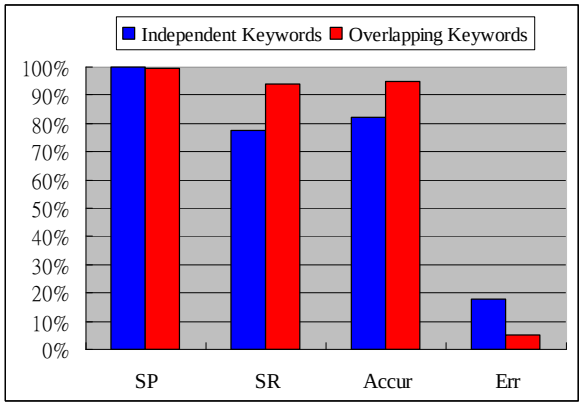


圖 1 英文關鍵字於 SVM 前測結果比較

至於中文信件則以訓練樣本中垃圾郵件關鍵字與正常郵件關鍵字相互獨立的 Independent Keywords 特徵值，轉換為每筆資料 111 向量，以訓練樣本 1000 筆資料作前測，最佳 c 值為 2048，g 值為 0.0004882，對負極資料被錯分的懲罰度 w 為 5。接著以訓練樣本中關鍵字不相互獨立的 Overlapping Keywords，經 Information Gain 計算找出 132 個關鍵字，轉化成向量對原本訓練樣本再進行前測，最佳 c 值為 32768，g 值為 0.000122，w 值亦為 5。同樣地在中文信件的前測結果中，以 Overlapping Keywords 作為特徵值分類之準確度較高，如圖 2 所示。

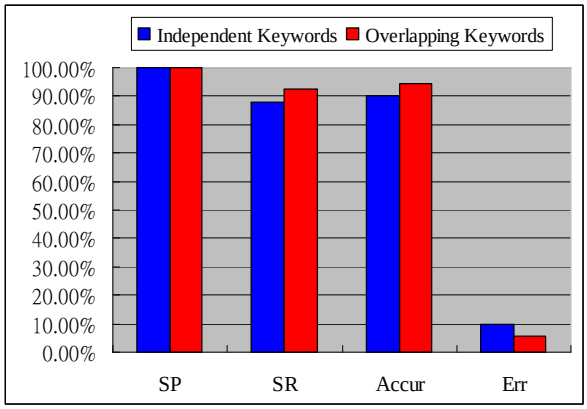
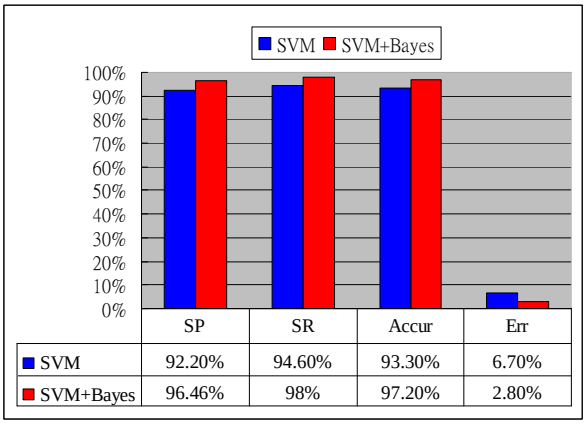


圖 2 中文關鍵字於 SVM 前測結果比較

最後，呈現經過 SVM 分類出之後，把所有錯分的信件完全地挑選出交由貝氏法進行機率演

算，視其能否驗證兩階層式過濾方法在最佳化的情況下，準確率能有效提升。

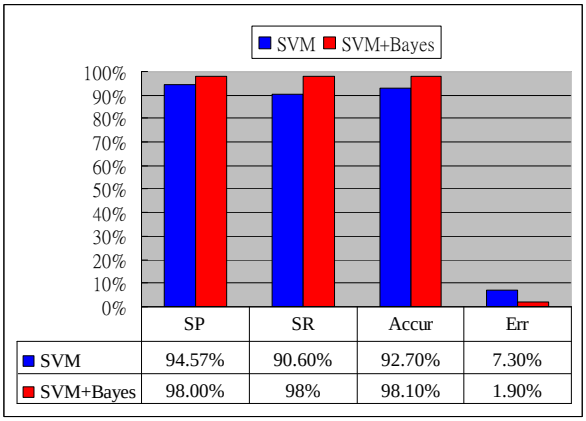
原先SVM對1,000封測試英文郵件的分類結果已於圖 2 呈現，在最佳化的前提下，將所有錯分信件挑選出來，交由第二層貝氏運算，英文郵件即是以 $\sum_i P(W_i)*N_i - \sum_i P(W_i')*N_i'$ 演算法來進行貝氏機率分數之判定，其得到效能之提升結果如圖 3 所示，可明顯看出運用兩階層式過濾法的結果與原本單一 SVM 相比，各評估指標包括正確率、召回率與準確率皆有效提升。



	SP	SR	Accur	Err
SVM	92.20%	94.60%	93.30%	6.70%
SVM+Bayes	96.46%	98%	97.20%	2.80%

圖 3 英文郵件之兩階層式過濾最佳化結果

SVM 對 1,000 封中文測試郵件的分類結果亦如圖 2 呈現，今將所有由 SVM 錯分的信件挑選出來，交由第二層貝氏機率分數判定，依中文郵件最佳貝氏運算結果 $\sum_i P(W_i) - \sum_i P(W_i')$ 演算法，得到最佳化效能提升結果如圖 4 所示，此結果亦充分表現出在最佳化之情況下，透過兩階層式過濾機制，對中文郵件分類的準確性能有顯著的提升，亦證明此方法針對中文郵件過濾分類之可行性。



	SP	SR	Accur	Err
SVM	94.57%	90.60%	92.70%	7.30%
SVM+Bayes	98.00%	98%	98.10%	1.90%

圖 4 中文郵件之兩階層式過濾最佳化結果

5. 結論與未來展望

以內容比對的垃圾郵件過濾方法而言，特徵值、關鍵字的選擇與擷取一直是相當重要的一環。本研究最大的特點在於作為測試之資料集與起始訓練關鍵字、特徵值的資料集分開收集，使之更符合「預測」的精神。在關鍵字的訓練與決定方面，針對不同演算法則有不同的方式，目的亦是希望結合二者發揮一加一大於二的功效。在樣本獨立、實驗環境固定的情況下，加上本研究提出貝氏機率演算的改良模式，驗證了二階層式垃圾郵件分類過濾的可行性與準確性，未來若有更多特徵明顯、類別固定之信件樣本，經由Information Gain計算得出更多屬於正常郵件和垃圾郵件的關鍵字，輸入SVM訓練向量權重後，相信能提供更佳的準確率。

誌謝

本研究感謝國科會專題研究計畫(計畫編號：NSC 94-2622-E-130 -003 -CC3)的經費支持。

參考文獻

[1] 吳昭逸，”具垃圾信過濾與安全機制之電子郵件收發系統”，國立台灣科技大學資訊工程系碩士論文，民國九十二年。

[2] 謝居呈，”應用機器學習理論改良分類竄改過之中英文垃圾電子郵件”，國立台灣科技大學電機工程系碩士論文，民國九十三年。

[3] 蘇士能，”具個人化中文垃圾郵件之過濾設計與實作”，國立東華大學資訊工程學系碩士論文，民國九十四年。

[4] Chih-Wei Hsu, Chih-Chung Chang, Chih-Jen Lin, “A Practical Guide to Support Vector Classification,” from http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf, 2003.

[5] Harris Drucker, Donghui Wu, Vladimir N. Vapnik, “Support Vector Machines for Spam Categorization,” IEEE Transactions on Neural Networks, Vol. 10, No. 5, 1999.

[6] Jenq-Haur Wang, Lee-Feng Chien, “Toward Automated E-mail Filtering – An Investigation of Commercial and Academic Approaches,” TANET 2003, p687-692, 2003.

[7] Kun-Lun Li, Kai Li, Hou-Kuan Huang, Sheng-Feng Tian, “Active Learning With Simplified SVMs for Spam Categorization,” First International Conference on Machine Learning and Cybernetics, Beijing, 2002.

[8] Pelossof, R. Miller, A. Allen, P. Jebara, T., “An SVM Learning Approach to Robotic Grasping,” ICRA 2004 IEEE International Conference, Vol. 4, 3512-3518, 2004.