

中華民國第四十六屆中小學科學展覽會  
作品說明書

---

高職組 電子、電機及資訊科

佳作

091005

越學越能擋的垃圾郵件守門員

學校名稱： 台北市立內湖高級工業職業學校

|        |       |
|--------|-------|
| 作者：    | 指導老師： |
| 職二 鄭博瑋 | 陳昭安   |
| 職二 盧珮穎 | 陳榮家   |

關 鍵 詞：垃圾郵件、分類權重演算法

## 摘要

垃圾郵件一直是十分令人苦惱的問題，平時都必須花費多於閱讀正常郵件的時間，來篩選及刪除像是一些廣告、色情郵件或藏有病毒的郵件；雖然現在的 ISP 大多已經設置可以濾除垃圾郵件問題的方法，效果也不錯，但卻只能服務到他們的會員，而對於學校、公司或政府單位的自設的郵件伺服器，都只能自求多福或者在垃圾郵件洪流中，過著水深火熱的生活。基於以上的理由，使我們決定要開發一個可安裝於個人電腦、自我學習並且獨立於 ISP 及收信軟體的垃圾郵件過濾器，以防止垃圾郵件的入侵。

本研究參考貝葉斯演算法，設計出「分類權重演算法」作為垃圾郵件過濾器分類核心，使用者只要花費幾天相等於清理垃圾郵件的時間，訓練郵件過濾器，便有很好的效果。經實驗結果，本垃圾郵件過濾器可以得到高達 97%~99%正確率的過濾效能，已媲美市面上宣稱 95%~99.5%效能的過濾器。

## 壹、研究動機

每天打開電子郵件信箱，又看到一大堆的垃圾郵件和廣告郵件，而且都要花上很多時間來檢查看看，哪些是包括廣告、病毒、匿名的垃圾郵件。據金山毒霸反垃圾郵件調查平均每人每天收到 1.85 封垃圾郵件，處理這些郵件每天則不得不花 3.65 分鐘。全國每年因此浪費了 15 億小時的時間，造成的損失高達 48 億元，而全球的損失竟高達 94 億美元。而花上無數的時間，就只為了怕錯殺一封重要的郵件，虛耗了多少寶貴的時間呀！因此常常會想到如果有一個垃圾郵件過濾器，可過濾含有垃圾郵件的元素或匿名的電子郵件並加以阻擋或刪除，這樣便可以省去檢查及刪除郵件的無謂時間。

目前有許多 ISP 如：雅虎奇摩、SeedNet、Hinet 都已經有提供不錯的垃圾郵件過濾方式，其中 RBL 就是最常用來提供給免費申請電子郵件者的一種過濾方式，RBL 是用 IP address 製作成一個黑名單，並以黑名單的電子郵件地址為阻擋的對象，但其中最大的缺點是很容易會造成誤攔的現象。

在「套裝軟體實習」的課程中，學過 Visual Basic 6.0 的程式語言。而在「資訊研習社」，曾經學過使用這一套語言寫過 TCP/IP 的網路連線遊戲程式。於是便聯想是否可以改用 POP3 的通訊協定，來開發 POP3 的收信程式，藉由收集並分析郵件的內容，來開發電子郵件過濾器。

## 貳、研究目的

本研究的目的是製作一個可以放置在個人電腦，並獨立於 ISP 之外並且可以學習的垃圾郵件過濾器。運用 Visual Basic 6.0 及 Winsock 控制元件製作 POP3 的收信程式收取郵件標題，透過統計標題中關鍵字的權重計算出符合垃圾及非垃圾郵件的權重，經過一段時間的訓練及學習，可產生出一個可靠的「分類權重表」作為分類演算法的核心。當達到可靠的程度垃圾郵件過濾器便可以在下載郵件前，事先阻擋垃圾郵件，並提示使用者予以刪除。

此外本研究將保留垃圾郵件做分析之用，主要是要蒐集足夠的樣本考驗過濾器的能力，並找出最佳的垃圾郵件過濾器的演算法。

## 參、研究設備及器材

表一 使用設備

| 設 備              | 規 格     | 數 量 |
|------------------|---------|-----|
| 電腦               | 586以上配備 | 2 部 |
| Visual Basic 6.0 |         | 1 套 |
| Windows XP 相關字詞庫 |         | 1 套 |

## 肆、相關知識

### 一、常用郵件編碼

原始的電子郵件協定只允許傳送 7 Bit 的ASCII 碼，這樣的設計在使用中文的台灣來說，是很大的一個問題，因為中文字採用的 Big5 或萬國碼 (Unicode)，都會用到 ASCII 128 ~ 255 間的高位元的文字，也就是會使用8個Bit。因此中文撰寫的電子郵件，SMTP 伺服器都無法順利的傳送這些電子郵件，此外如圖形檔或壓縮檔等二進位檔也有同樣的情況。為了把中文或二進位檔案之類的 8 Bit 資料，轉換成7Bit 或6Bit 的編碼，並調整在本文的文字編碼內，才能讓電子郵件順利的傳送這些資料。然而依照這些編碼規則所產生的檔案，雖然都是本文格式，但無法直接閱讀。因此必須瞭解並破解這些常見的電子郵件編碼，才可取得正確的郵件資料。

#### (一) Base 64編碼原理

1. Base 64編碼是將3個8位元位元組轉換為4組6位元， $(3 \times 8 = 4 \times 6 = 24)$  而這4組6位元前方再補上00，使其成為8位元，也就是一個位元組。
2. 當一個位元組只有6位元有效時，它的取值空間為0 ~ 26-1，也就是0~63，因此Base64編碼的每一個編碼的取值空間為0~63。
3. 以三個字節 (8 位) 用四個字節 (6 位) 表示，由於編碼後內容為六位，因此可避免截去，不過缺點是檔案的大小會被膨脹。

#### (二) Unicode

- (1) 1991年，美國IBM、DEC、Sun、Apple、Xerox、Novell、Microsoft等廠商共同出資成立一個編碼組織 ( The Uniform Consortium)，制定出一套全球通用的文字編碼系統。
- (2) 又稱統一碼、萬國碼。

- (3) 以2 Bytes (16 bits)來表示一個字元。
- (4) 共可表示65536個字元或符號。
- (5) 前128個字元編碼和ASCII相同,其餘字元可涵蓋各國常用的文字、字母及符號。
- (6) 目前如Windows XP的作業系統,都已支援Unicode的編碼。

### **(三) EBCDIC**

- (1) EBCDIC是IBM所制定的一種編碼方式。 以8 bits來表示一個字元。
- (2) 最多可表示 $2^8=256$ 種字元或符號。
- (3) 用於IBM 廠牌的大型電腦。

### **(四) BIG-5碼**

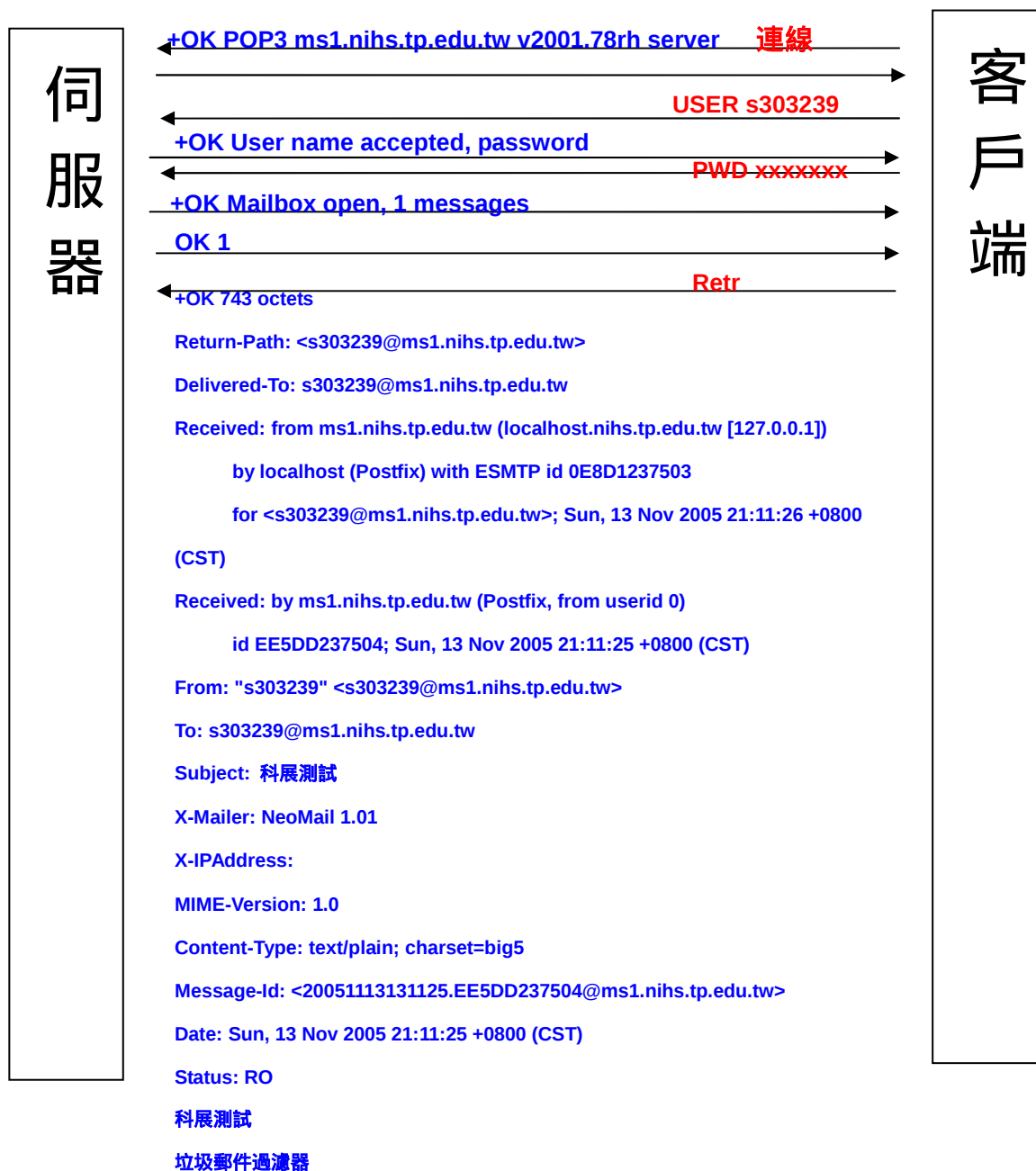
- (1) 1984年,由台灣資訊工業策進會(資策會)所制定的一套中文編碼系統,以兩個位元組一個中文字。第一個位元組稱為「高位位元組」,第二個位元組稱為「低位位元組」。
- (2) 最多可表示65536個字元。 目前已制定13053個中文字。
- (3) 其他常見中文碼 - 倚天碼、GB碼(中文簡體字)...等。

### **(五) QP 編碼**

內容都以 ASCII 碼表示的符號,因此可避免被截去,主要的特徵是有“=”和大量的英文字母,將一個字元用二個16進位法的數值表示,然後前面再加個”=”字元,例:「垃圾郵件」→「=A9U =A7=A3 =B6I =A5=F3 」

## 二、 POP3伺服器的運作說明

POP3伺服器主要是負責電子郵件的收信，使用TCP/IP的110 Port的協定。透過客戶端與伺服器端來回交談，來達到連線、登入帳號密碼、取得郵件等動作。POP3伺服器的運作流程如下圖所示：



圖一 POP3協定的通訊流程

### 三、垃圾郵件的篩選方式及原理

目前使用的垃圾郵件過濾方式，各有其優點及尚待克服的缺陷，以下為幾種最常用垃圾郵件過濾方式：

#### 1. 關鍵字過濾

這是一種簡單的內容過濾方式來處理垃圾郵件，它的基礎是必須創建一個龐大的過濾關鍵字列表。這種技術有很大的缺點，過濾的能力同關鍵字有明顯聯繫，關鍵字列表也會造成錯報可能比較大，當然使用這種方法來處理郵件所消耗的資源也相對比較多。而現在使用關鍵字過濾常常無法刪除像是拆詞，或組詞。

#### 2. 黑名單過濾 (BL, Block List)

現在有很多組織都在做 BL (Block List)，將經常發送垃圾郵件的 IP 地址（甚至 IP 地址範圍）收集在一起，做成 Block List。BL 技術也有明顯的缺點，因為不能在 Block list 中包含所有的（即便是大量）的 IP 位址，而且垃圾郵件發送者很容易通過不同的 IP 位址來製造垃圾。

#### 3. HASH技術

HASH 技術是郵件系統通過創建 HASH 來描述郵件內容，比如將郵件的內容、發件人等作為參數，最後計算得出這個郵件的 HASH 來描述這個郵件。如果 HASH 相同，那麼說明郵件內容、發件人等相同。這在一些 ISP 上在採用，如果出現重複的 HASH 值，那麼就可以懷疑是大批量發送郵件了。

#### 4. 智慧和概率系統：Bayesian 貝葉斯演算法

其實大部份人已經使用類似貝氏定理的概念來分析垃圾信件。如信件中出現「未滿十八歲」字眼有可能為垃圾信件，但如果這封信件同時出現「極品」、「熟女」、「偷拍」等字眼，便幾乎可以斷定是垃圾信件。也就是利用過去的收信經驗，判定新的信件是否為垃圾信。而貝氏定理應用在內容過濾領域方面有強大效能，貝氏過濾法即是採用類似，但更客觀的統計方式來偵測其為垃圾信件之機率，比對訓練過的「貝氏過濾法資料庫」，分析過往的經驗，以精確評判是否為垃圾信件的機率。透過訓練「貝氏過濾法資料庫」的步驟，貝氏過濾法的精準率必能達到 95%~99.95% 之準值。貝氏過濾法具有極高的準確度、學習方式容易、可針對不同產業調整、垃圾信發送者不容易將垃圾信穿透。

#### 四、文件分類

1. **詞庫式斷詞法**：為目前普遍使用的斷詞方法，然而品質與詞庫的大小有相對應之關係，因此常需要刪除或擴充。
2. **混合式斷詞法**：將詞庫斷詞法及統計斷詞法整合。利用詞庫斷出不同組合的詞彙，然後利用詞彙的統計資訊，找出最佳的斷詞組合。

#### 五、字詞權重函數

1. **字詞出現頻率(Term Frequency, TF)**: TF是指某一關鍵詞在某類文件中出現的次數，文件分類用d來表示、而關鍵字用(t)來表示則權重(W)

$$W(d,t)=TF(d,t)$$

此缺點為特徵不夠明顯，若關鍵詞過度密集出現則分類的特徵就不明顯，因此頻率高的關鍵詞必須移除以提高正確度。

2. **逆文件頻率(Inverse Document Frequency, IDF)**: IDF是指某一關鍵詞在各文件分類中出現的普遍程度。IDF定義如下：

$$IDF(t) = N/df(t)$$

其中

N：代表文件總數

df(t)：代表出現過這一個關鍵詞 t 的文件總數

例：有一關鍵詞在垃圾郵件出現為 1，在非垃圾郵件出現也為 1 則

$$2/(1+1)=2(\text{此為出現的類別總數})$$

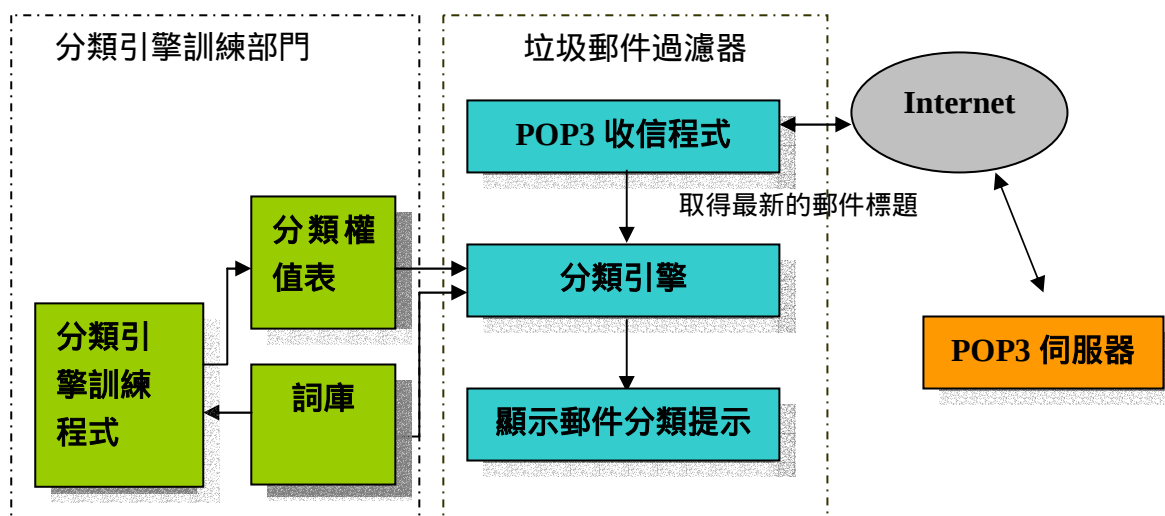
$$2/(1+1)=1(\text{此為普遍性})$$

以上例子普遍性太高，因此沒有參考價值

所以如果普遍出現的比例越高，則越無法突顯分類的特徵；相對的越低，最好是集中出現在同一個分類之中，則越容易突顯。

## 伍、研究過程及方法

下圖為系統架構圖，左邊的方框代表分類引擎訓練部門，分類引擎是利用分類權重表和詞庫訓練出來的，而在中間的方框是垃圾郵件過濾器，此部份首先利用 POP3 的收信程式，透過網際網路連到 POP3 伺服器上收取新的信件，並取得最新的郵件標題，再使用訓練完成後的分類引擎分析是否為垃圾郵件，最後可得到郵件分類的提示訊息。



圖二 系統架構

分類引擎訓練部門透過使用者檢視郵件標題並判定其分類，然後用詞庫斷詞法取得郵件標題的關鍵字詞，並更新關鍵字詞在其分類的出現次數，而形成分類權重表，以提供給分類引擎作為分類判斷的依據。

在系統設計完成之後，也將透過回歸測試、關鍵字詞數量與正確率關係比較測試及使用者差異測試，了解系統的可行性及效率，並且從過程中了解演算法的盲點，進而修正方法以求得最佳的效能。

### 一、 分類權重演算法說明

分類引擎的分類演算法，在本研究設計將採用字詞出現頻率(TF)，轉化成字詞出現機率的構想，主要是轉成概率後，同時可以表達出現頻率及普遍性的概念。此外方法上也參考部份貝葉斯演算法的概念，產生分類權重表。分類權重表將紀錄關鍵字詞出現在「垃圾」及「非垃圾」的比例，如：字詞 A 在「垃圾」及「非垃圾」出現的次數分別為 10、30 次，則出現的比例就是 0.25 及 0.75，在本研究將其命名

為分類權重演算法。

單一字詞的出現比例就判定是否為垃圾郵件，可能會發生極大的誤判。如：「本校的同學請勿瀏覽色情網站」當中隱含著「色情」這個垃圾郵件的字詞，憑這個字就斷定為垃圾郵件，結果是誤判。但結合其它關鍵字詞在各分類的出現比例，就可以得到更客觀的結果。

Q 為某郵件標題所分析出的關鍵字詞及出現的次數，依據 Q 所列的關鍵字詞，從分類權重表中，找出這些關鍵字詞及其垃圾及非垃圾的出現機率，而形成 M 的分類權重表。

$$Q = [f_1 \ f_2 \ f_3 \ \dots \ f_m]$$

$$M = \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \\ w_{31} & w_{32} \\ \vdots & \vdots \\ w_{m1} & w_{m2} \end{bmatrix}$$

將Q及M的矩陣相乘，得到M<sub>Q</sub>的矩陣：

$$M_Q = Q \cdot M = [f_1 \ f_2 \ f_3 \ \dots \ f_m] \cdot \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \\ w_{31} & w_{32} \\ \vdots & \vdots \\ w_{m1} & w_{m2} \end{bmatrix} \quad (\text{公式1})$$

從M<sub>Q</sub>矩陣找出最大的元素

$$D = \text{Max}(M_Q) \quad (\text{公式2})$$

例：以下列這兩個矩陣為例，在M裡面，由上到下是關鍵字詞，而由左而右分別是垃圾郵件及非垃圾郵件的權重，經過（公式1）的計算後，再以（公式

2) 計算後，取權值總合較大的那一邊為郵件的分類。假設左邊為各關鍵字詞與垃圾郵件相關的權值，右邊為各關鍵字詞與非垃圾郵件相關的權值。

計算如下：

$$Q = [6 \quad 4 \quad 2 \quad 8] \quad M = \begin{bmatrix} 0.3 & 0.7 \\ 0.1 & 0.9 \\ 0.8 & 0.2 \\ 0.5 & 0.5 \end{bmatrix}$$

$$M_Q = [1.8+0.4+1.6+4.0 \quad 4.2+3.6+0.8+4.0] = [7.8 \quad 12.6]$$

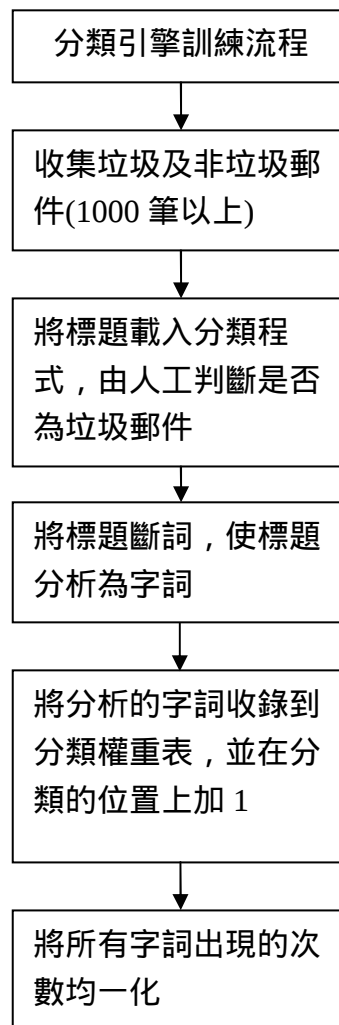
$$D = \text{Max}([7.8 \quad 12.6]) = 12.6$$

由計算的結果可以得知，權值總和的最大值為12.6，屬於垃圾郵件的可能性比較大，因此判定分類結果為垃圾郵件。

## 二、系統架構介紹

### (一) 分類引擎部門

訓練分類引擎的過程，首先需要收集約 1000 多筆的郵件標題，接著採用人工判斷是否為垃圾郵件再，經由斷詞程式分析出郵件標題的關鍵字詞，將關鍵字詞的出現頻率 (TF)，累加分類權重表該關鍵字詞在相關分類的出現頻率，為了不要因為偏重於經常出現的關鍵字，而干擾分類引擎的準確性，因此必須進行均一化 (Uniform) 的動作，均一化的目地在於了解某關鍵字詞出現於「垃圾」及「非垃圾」這兩個分類的概率，最大值為 1、最小值為 0，而「垃圾」及「非垃圾」這兩個分類的比例加總為 1，也就是說只要其中一個分類為 1 的話，另一個必為 0。而 1 則代表 100%出現於該分類，0 則反之，此兩者都極具參考價值。而在 0~1 之間的數值，越大者或越小者代表參考越高，而 0.5 的數值則完全沒有參考價值。



圖三 分類引擎訓練流程圖

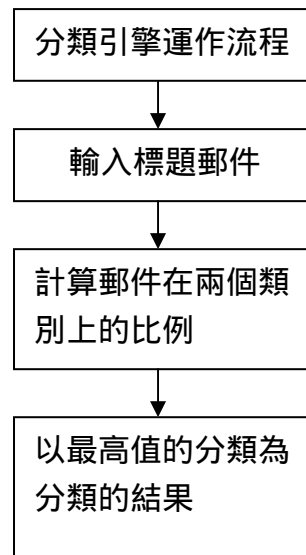
| 字詞 | 垃圾 | 非垃圾 | 均一化 | 字詞 | 垃圾    | 非垃圾   |
|----|----|-----|-----|----|-------|-------|
| A  | 5  | 10  |     | A  | 0.333 | 0.666 |
| B  | 3  | 6   | ➡   | B  | 0.333 | 0.666 |
| C  | 4  | 9   |     | C  | 0.307 | 0.692 |
| D  | 11 | 13  |     | D  | 0.458 | 0.541 |
| E  | 6  | 9   |     | E  | 0.4   | 0.6   |
| F  | 2  | 8   |     | F  | 0.2   | 0.8   |
| G  | 4  | 5   |     | G  | 0.444 | 0.555 |

圖四 分類權重表及均一化的過程

如圖四，以 D 字詞為例在垃圾的權重為 11，在非垃圾的權重為 13，則均一化就是將 11 除以 11+13 所以我們可以得到 D 字詞在均一化後的垃圾權重得到機率為 0.458，而在均一化後的非垃圾權重得到平均值為 0.541，而兩者算出來的結果都接近 0.5，所以 D 字詞的參考價值並不高，在以 F 字詞為例在垃圾的權重為 2，在非垃圾的權重為 8，則均一化就是將 2 除以 2+8 所以我們可以得到 F 字詞在均一化後的垃圾權重得到平均值為 0.2，而在均一化後的非垃圾權重得到平均值為 0.8，而非垃圾權重較高，所以 F 字詞應該是屬於非垃圾郵件關鍵字的參考價值很高。

## (二) 分類引擎運作流程

分類演算法，採用圖四已經均一化的分類權重表做為核心。首先輸入郵件的標題，然後使用分類權重表及關鍵字詞出現的次數，在依照分類權重表演算法計算垃圾及非垃圾郵件上的比例，並以比例值總和最高的分類為最後的類別。



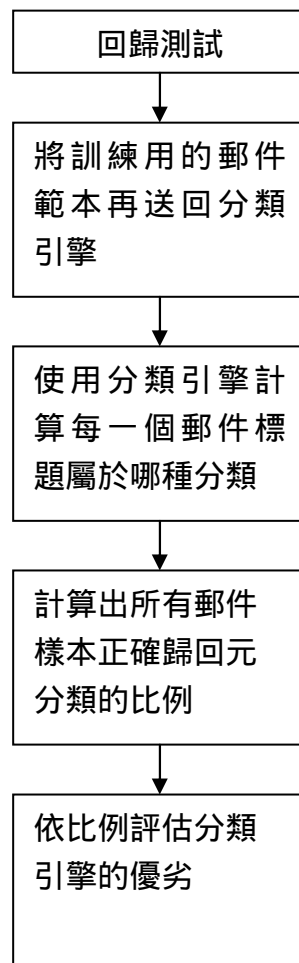
圖五 分類引擎運作流程圖

### 三、 分類效能評鑑方式

本研究分類效能評鑑方法，依郵件樣本、郵件數量及使用對象，設計回歸測試、關鍵字詞數量與正確率關係比較測試及使用者差異測試三個重點：

#### (一) 回歸測試

回歸測試主要是將用來學習的郵件標題樣本，再送入過濾器分類，並檢視是否正確的回到原本的分類，並計算出正確回歸的比例，以判定分類引擎的優劣與否。



圖六 回歸測試的流程圖

#### (二) 關鍵字詞數量與正確率關係比較測試

假設垃圾郵件過濾器的效能，與分類權重表的關鍵字詞數量成正比，也就是說從第一封郵件開始累積關鍵字詞的樣本，舉例來說，假設 120 筆的關鍵字詞時，經過垃圾郵件過濾器的分類，可得到 40% 的正確率，但當累積到 1000 筆的郵件標題而得到 2500 筆的關鍵字詞時，可能達到 95% 正確率。計劃透過關鍵字數量及正確率的曲線來驗證。此外將委由數個不同的使用者測試，以瞭解曲線是否相似或相近。

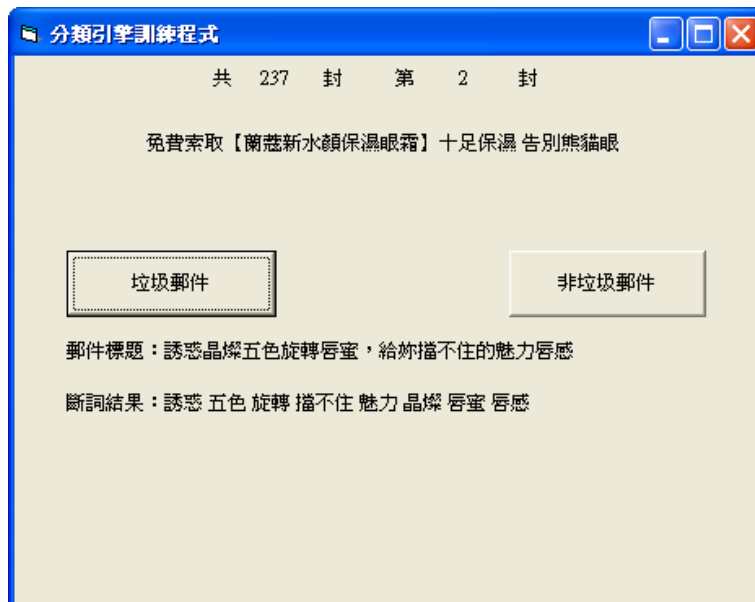
### (三) 使用者差異測試

假設垃圾郵件過濾器所訓練出來的分類權重表交給不同的使用者，是否會因為收信習慣的不同而有不同的正確率。計劃將訓練後的分類權重表交給 A、B、C、D 四個使用者，透過記錄郵件數量及正確率，了解其差異的程度。

## 陸、研究結果

依照上述的研究方法設計以下4個部份的設計及實驗，實驗過程及結果說明如下：

### 一、 分類權重表訓練結果



圖七 分類引擎訓練程式

當此程式統計完時，會出現以下 2 種統計表，左邊的是出現的次數表而右邊的則是出現的比例表，以此做區分。出現次數表是在所有的郵件中所出現的次數。

表二 關鍵字詞出現次數表及出現機率表比較

|     |      |      |                                                                   |
|-----|------|------|-------------------------------------------------------------------|
| 投資  | 1    | 0    | <p>出現次數表：</p> <p>左邊為垃圾郵件中字詞所出現的次數</p> <p>右邊為非垃圾郵件中字詞所出現的次數</p>    |
| 風險  | 1    | 0    |                                                                   |
| 小額  | 1    | 0    |                                                                   |
| 時間  | 1    | 1    |                                                                   |
| 準備  | 2    | 1    |                                                                   |
| 好幾  | 1    | 0    |                                                                   |
| 範本  | 1    | 1    |                                                                   |
| 學網  | 1    | 0    |                                                                   |
| 學網頁 | 1    | 0    |                                                                   |
| 千個  | 1    | 0    |                                                                   |
| 改一  | 1    | 0    |                                                                   |
| 就好  | 1    | 0    |                                                                   |
| 就好囉 | 1    | 0    |                                                                   |
| 好囉  | 1    | 0    |                                                                   |
| 知道  | 2    | 1    |                                                                   |
| 創意  | 0    | 1    |                                                                   |
| 還是  | 3    | 1    |                                                                   |
| 投資  | 1    | 0    | <p>出現機率表：</p> <p>左邊為垃圾郵件中字詞所出現的機率數值</p> <p>右邊為垃圾郵件中字詞所出現的機率數值</p> |
| 風險  | 1    | 0    |                                                                   |
| 小額  | 1    | 0    |                                                                   |
| 時間  | 0.5  | 0.5  |                                                                   |
| 準備  | 0.67 | 0.33 |                                                                   |
| 好幾  | 1    | 0    |                                                                   |
| 範本  | 0.5  | 0.5  |                                                                   |
| 學網  | 1    | 0    |                                                                   |
| 學網頁 | 1    | 0    |                                                                   |
| 千個  | 1    | 0    |                                                                   |
| 改一  | 1    | 0    |                                                                   |
| 就好  | 1    | 0    |                                                                   |
| 就好囉 | 1    | 0    |                                                                   |
| 好囉  | 1    | 0    |                                                                   |
| 知道  | 0.67 | 0.33 |                                                                   |
| 創意  | 0    | 1    |                                                                   |
| 還是  | 0.75 | 0.25 |                                                                   |

## 二、 回歸測試

本實驗將用來訓練分類權重表的郵件樣本，送回垃圾郵件過濾器並使用此一分類權重表作為分類的核心，測試垃圾郵件過濾器是否可以將原來的郵件樣本回歸到原來的分類，並以正確率作為評估效能的依據。

### (一) 實驗設計

1. 將 1000 筆送入分類引擎訓練程式，以訓練出分類權重表。
2. 使用垃圾郵件過濾器分別針對 1000 筆的樣本郵件做分類。

3. 計算出總正確率、垃圾郵件的正確率及非垃圾郵件正確率。

## (二) 實驗結果

1. 經過分類引擎訓練程式訓練之後取得 5003 的關鍵字詞。
2. 經計算後取得總正確率=97.20%
3. 無法判定的狀況：在回歸測試的過程中逐一檢視發生辨識錯誤的情況，並分析結論如下：

- (1) 關鍵字詞權值總合相等：由於垃圾郵件權值和非垃圾郵件權值相等，所以無法判定其郵件類別。

共 291 封 第 114 封  
下一則分類的標題：月華劍士二代

垃圾

分析

分類結果

垃圾郵件：1 非垃圾郵件：1

郵件標題：綁鞋帶的藝術

分類：非垃圾 實際分類：垃圾

正確率：98% 已判斷筆數：114筆 正確筆數：112筆

|    |     |     |
|----|-----|-----|
| 鞋帶 | 0.5 | 0.5 |
| 藝術 | 0.5 | 0.5 |

- (2) 含有中性的文字：關鍵字詞在垃圾及非垃圾郵件都很常用，因此系統往往由於一個小小的差異而誤判。如下圖，玩具在垃圾郵件的比重0.75大於非垃圾郵件的0.25，而其他如茶杯、貴賓、尋找、主人和泰迪紅等字詞在兩種分類的比分相同，因此玩具這個字只要稍微偏向其中一方，就可能造成誤判。

共 291 封 第 211 封  
下一則分類的標題：成人影片一片 100元 買五送一 大優待回饋網友  
seventieth

垃圾

分析

分類結果

垃圾郵件：3.25 非垃圾郵件：2.75

郵件標題：玩具型茶杯型的泰迪紅貴賓在尋找主人囉!

分類：垃圾 實際分類：非垃圾

正確率：98% 已判斷筆數：211筆 正確筆數：207筆

|     |      |      |
|-----|------|------|
| 玩具  | 0.75 | 0.25 |
| 茶杯  | 0.5  | 0.5  |
| 貴賓  | 0.5  | 0.5  |
| 尋找  | 0.5  | 0.5  |
| 主人  | 0.5  | 0.5  |
| 泰迪紅 | 0.5  | 0.5  |

### 三、 關鍵字詞與正確率比較實驗結果

本實驗的目的，是要考驗及瞭解關鍵字詞的數量與垃圾郵件過濾器的效能是否為正比的關係，是否郵件標題數越多，分類引擎的效能也越好，最後是到多少的數量可以得到良好的分類品質。

#### (一) 實驗設計

1. 將垃圾郵件過濾器分別交給 A、B、C、D 四個人使用，每收到一封郵件便由使用者辨別是否為垃圾郵件，以訓練分類權重表。

越學越耐擋的垃圾郵件守門員

POP3伺服器: msl.nihs.tp.edu.tw 帳號: s303238 密碼: \*\*\*\*\* 連線

+OK POP3 msl.nihs.tp.edu.tw v2001.78rh server ready  
+OK User name accepted, password please  
+OK Mailbox open, 1000 messages  
+OK 1000 349273  
+OK Top of message follows  
Return-Path: <s303238@msl.nihs.tp.edu.tw>  
Delivered-To: s303238@msl.nihs.tp.edu.tw  
Received: from msl.nihs.tp.edu.tw (localhost.nihs.tp.edu.tw [127.0.0.1])  
by localhost (Postfix) with ESMTP id 394B02374EA

郵件標題: 轉寄: 給我的朋友, 一定要幸福囉

分類結果: 非垃圾郵件

正確 錯誤

正確率: 97.2 正確數: 892 郵件數量: 900

關鍵字詞總數: 5003

關鍵字運算結果  
計算結果 (5 10)

朋友 1 2  
一定 1 2  
幸福 3 6

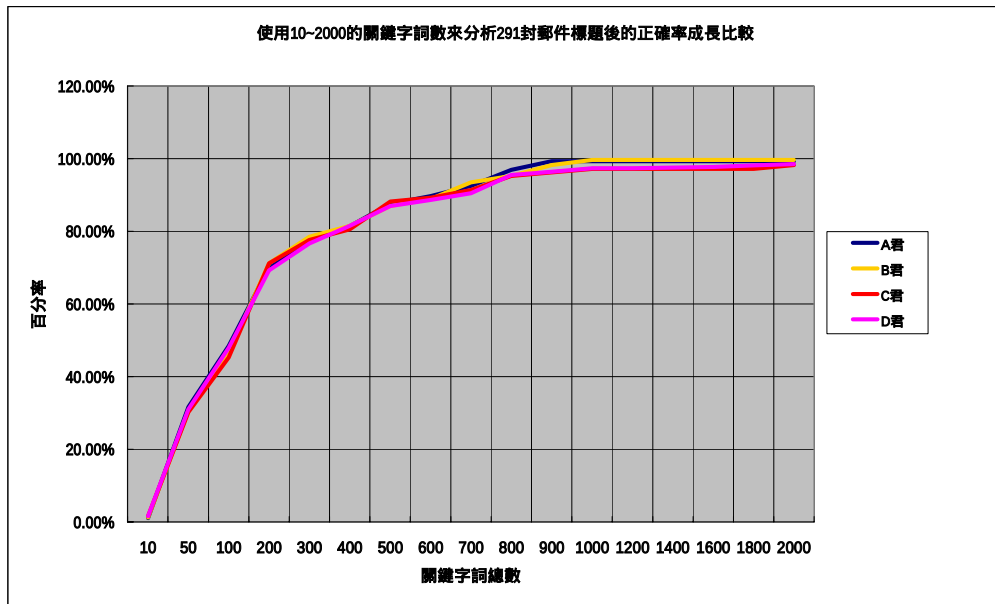
☒ 停止學習  
清除所有統計數據

圖七 本研究所開發的越學越耐擋 - 「垃圾郵件守門員」

2. 當收到下一封郵件時立即利用已經訓練的分類權重表判斷是否為垃圾郵件，在訓練及使用的過程中紀錄關鍵字詞的數量及正確率。
3. 測試到關鍵字詞數量為 2000 筆時將四個人最後的結果整理並以關鍵數量為 X 軸，正確率為 Y 軸繪製成圖表。
4. 比較圖表的曲線關係，並加以分析。

## (二) 實驗結果

1. A、B、C、D 的正確率的成長曲線都十分的相近。
2. 關鍵字詞數在 700 以後，開始穩定在 98~99%之間，甚至在 900 以後穩定在 99%的正確率。



圖八 關鍵字詞與正確率的成長比較圖

## 四、 同一分類權重表運用於不同使用者的個別差異

假設同一個分類權重表，讓不同的使用者使用，會得到不同的正確率，並且在不同的族群的使用者的正確率也會有所差異。

### (一) 實驗設計

先找到一位老師當作原使用者，使其訓練出符合個人的分類權重表。接著將這份分類權重表分別交給A、B、C、D君，分別取代其垃圾郵件過濾器的分類權重表。

### (二) 實驗結果

1. 與原使用者比較：如表三，經實驗後發現其A、B、C、D君的總正確率遠不及原使用者的高。
2. 使用族群的比較：如表三，經實驗發現老師組 (A、B君) 與學生組 (C、D君) 的總正確率相差約10~20%之多，教師組的A、B君相差13.27%，而學生組的C、D君相差僅有7.89%，可見不同族群使用本系統也有明顯的差異。

表三 使用者個別差異

| 使用者  | 角色 | 總郵件數 | 總正確率   |
|------|----|------|--------|
| 原使用者 | 老師 | 1000 | 97.20% |
| A 君  | 老師 | 548  | 83.39% |
| B 君  |    | 540  | 70.12% |
| C 君  | 學生 | 555  | 66.67% |
| D 君  |    | 537  | 58.78% |

## 柒、討論

### 一、為什麼會出現郵件亂碼？

- (一) 伺服器系統不同：由於某些郵件伺服器不支援8 位元的非ASCII碼的格式，尤其是歐洲的系統，因此在傳輸過程中，會造成8位元的編碼無法呈現的問題
- (二) 郵件編碼不同：各種電子郵件軟體的預設值及收件者及發件人的選項不一定相同。系統如果不能自動識別編碼的方法，就會出現所謂的亂碼了。

### 二、為何過濾器分析非垃圾郵件的正確率高過垃圾郵件？

非垃圾郵件的特徵遠較垃圾郵件明顯。

### 三、分類過程中如果遇到垃圾及非垃圾郵件比分相同如何判別？

應該是無法判別，此時仍列入非垃圾郵件，因為誤砍比不砍嚴重。

### 四、隨著分類權重表的增大，是否系統效率會變差？

會，由圖八的結果可以知道，關鍵字詞數量達到500以上，就已經可以很可靠的過濾，如希望能提升正確率，可將關鍵字增到1000字左右，不會影響到系統效能。

## 捌、結論

一個可靠的垃圾郵件過濾器，可以有效的阻擋垃圾郵件，節省網路頻寬並省下不少清理郵件的困擾。本研究所開發的垃圾郵件過濾器，除了過濾的演算法則簡單，並且可以得到正確率 97.20%的分類效率，與市面上 95%~99.95%的分類效率可以相媲美，只要使用者使用 POP3 收信伺服器的通訊協定，就可以在個人的本機電腦上使用本程式，沒有 ISP 的限制。經研究的結果，如果使用者可以持續的訓練垃圾郵件過濾器，則分類的效率便可以持續提升 97%~99%的正確率，而因人而異的特性，可以讓每一個使用者都可以訓練出一個符合個人使用特性的垃圾郵件過濾器。

另外在使用同一個分類權重表分析個別差異的實驗結果，委託 2 位老師及同學分別測試所得到的結果，很驚訝的發現當中隱含著使用族群的差異，是我們在研究設計的過程當中沒有意料到的。是否可以運用這個特性，經由使用過程的統計分析，劃分出網路上的使用族群呢？這點或許可以成為後續研究的課題。

## 玖、參考資料及其他

李春雄 (民 92)。 Visual Basic6.0 學習實務。文京圖書。

陳永佳 (民 79)。 離散數學與計算機科學之應用。全華科技圖書。

杜海倫 (民 88)。 以標題進行新聞自動分類。清華大學資訊工程學系碩士論文。

林頌華 (民 88)。 新聞標題自動分類。清華大學資訊工程學系碩士論文。

陳昭安 (民 91)。 建構試題自動分類系統之研究 - 以 MOCC 術科試題為例。  
師範大學工教系碩士論文。

電子郵件編碼的技術 。 <http://www.cdchen.idv.tw/?p=56>。

向垃圾郵件說不。

<http://www.digitimes.com.tw/n/article.asp?id=90AC874FFABA930A48256F71000DBA18>。

學習法過濾。 <http://mail.sfilc.com/plan/spam/7.htm>。

用智慧戰勝垃圾郵件。 <http://www.jituo.net/wangluo/5/2686.shtml>。

ShareTech 協助瑞耘科技阻擋郵件病毒及垃圾信。

<http://www.computerdiy.com.tw/modules/news/article.php?storyid=958>。

有效預防常見垃圾信手法。

<http://www.maiciao.com.tw/WebMail/%A6%B3%AE%C4%B9w%A8%BE%B1`%A8%A3%A9U%A7%A3%ABH%A4%E2%AAk.htm>。

毒霸信息安全。金山毒霸反垃圾郵件調查。 <http://db.kingsoft.com/product/inquiry/315/>。

## 091005 越學越能擋的垃圾郵件守門員

以分類權重演算法來區分垃圾信件及非垃圾信件是一個很好的想法，但一開始必須經過一段時間的學習才可以開始區別信件種類是主要缺點。其他使用者之經驗法則或共通的垃圾信件判別法則應可設為基本法則。