# 第一章 緒論

## 第一節 研究動機

### 電子郵件已是資訊生活中不可或缺的工具

從 1990 年代起 Internet (網際網路)的規模急速地成長，同時帶動了多元網路服務的發展，而這些網路服務廣泛的用途也直接或間接的幫助使用者創造更好、更愉快的生活方式。在眾多的網路服務當中，除了目前廣為大眾熟悉的全球資訊網 (World Wide Web, www)之外，電子郵件(Electronic Mail, E-Mail)在數位通訊上的貢獻尤為重要，特別是當年人們仍然停留在使用市內電話和傳統紙本書信的年代，電子郵件的出現確實為使用者節省不少郵務往返的時間和金錢，讓遠在國外的親友能夠透過電子郵件將訊息在彈指之間送達，而提供使用者編撰多媒體樣式和附加文件等便利的功能，更是改變現代人撰寫書信習慣的重要原因。

資策會(財團法人資訊工業策進會)於 2006 年 1 月 17 日所完成的「2005 年我國家庭寬頻、行動與無線應用現況與需求調查」[1]報告中指出，在國內上網人口中，有高達 77.6%的民眾將收發電子郵件列為主要的網路應用，僅次於瀏覽資訊的 89.2%；雖然近幾年 MSN、Yahoo!即時通等即時通訊軟體(Instant Messages)有迎頭趕上的趨勢，不過在未來行動通訊技

[1] 參考資策會，2006，http://www.find.org.tw/find/home.aspx?page=many&id=126

術成熟[2]和智慧型手機(Smart Phone)等手持通訊裝置的推波助瀾下，電子郵件仍可穩坐網路應用的前三強。

在相關行動通訊系統佈建完成後，收發電子郵件將更加地方便，使用者不再需要枯燥地坐在電腦前操作。對商務用戶而言，可以在乘坐交通工具時透過行動電子郵件隨時隨地收發客戶往來的文件；而一般使用者則可以在戶外活動時隨時將拍攝的照片分享給親友。相較於目前行動電話乏味的 SMS (Short Message Services, 短訊服務)或昂貴的 MMS[3] (Multimedia Message Services, 多媒體訊息服務)，電子郵件得以利用其多媒體技術及價格的優勢取而代之。

電子郵件不僅僅為個人用戶帶來極大的幫助，它也為企業提供了顯而易見的實質效益，主要包括三點：(1)節省郵務成本；(2)簡化文件遞送流程；(3)增加資訊傳遞的效率。因此多數企業都樂於使用電子郵件取代傳統公文。

然而在享受到電子郵件的便利之後，部份企業開始嘗試將電子郵件應用在行銷活動上，以便讓客戶能夠隨時清楚掌握最新的產

---

[2] 行動通訊技術除了 GPRS 及 3G 系統外，尚有由政府推動的「M 臺灣計畫」中所使用的 WiMAX 技術，尤其是 WiMAX 因為能夠帶來更大的頻寬及更低廉的使用成本，可望推動各項應用的發展，跳脫目前由於行動網路連接費用高昂、連線頻寬不足而導致市場使用意願不高的惡性循環。

[3] 以國內遠傳為例，MMS 的單次使用費率最低為 7 元，並且傳輸量每超出 30KB 需另外加收 0.18 至 0.24 不等的費用。

品資訊或服務動態，這種將活動宣傳單(Direct Mail, DM)數位化，並透過電子郵件傳遞的行銷方式，無論是在降低成本或維繫客戶關係上，都達到相當不錯的效果。

雖然自電子商務開始發展以來，電子郵件就一直扮演一個重要的角色，但是由於網路上的行為原本就缺乏有效的管理方法，再加上中、小型企業經營者普遍未建立電子行銷的正確觀念，濫發濫寄的結果，使得垃圾郵件(SPAM[4])反而成為電子郵件廣告的代名詞。此外，部份網路業者提供為廣告主大量發送廣告電子郵件的服務，也讓垃圾郵件的問題雪上加霜，這些大量發送的郵件也經常夾帶著許多色情資訊入侵到一般家庭，嚴重影響家庭中青少年的身心發展。

垃圾郵件不僅讓全球網路系統每年必須浪費龐大的成本，更直接地侵害了個人和社會的利益，因此如何有效地防制垃圾郵件，已經成為全球共同關注的焦點。

---

[4] SPAM 原為美國 Hormel 公司所生產的一種肉質罐頭的名稱，而後引申成為垃圾郵件代名詞的原因，一般較普遍的說法是指在一部描述一對夫妻在餐館用餐的戲劇中，妻子並不想吃 SPAM，但是餐館中其他人卻開始高聲讚頌 SPAM，甚至聲音大到整個劇場就只能聽到 SPAM 這個單字，後來就將強迫他人接受且泛濫的訊息稱為 SPAM，之後由於網際網路流行，因此又將不請自來、強迫收件者接受的的商業郵件稱為 SPAM。

## 第二節 研究目的

雖然在全球科技業界的努力下，網路資訊傳遞服務不斷地推陳出新，卻始終難以憾動電子郵件在網路通訊領域的地位，而且近年來包含微軟(Microsoft)在內的數家國際資訊大廠，也積極的合作研議下一代的電子郵件通訊協定，可想見電子郵件在未來仍然會是相當重要的網路應用，然而因為電子郵件之普遍性與便利性，造成垃圾郵件泛濫的問題，也讓全球網路付出愈來愈高的代價。

本研究針對上述議題分為兩個部份進行討論。

**1. 電子郵件通訊協定**

在此研究中將探討目前電子郵件通訊協定的架構和運作模型，並以 Microsoft .Net Framework 2.0 實作通訊元件，以簡化電子郵件軟體的開發程序，讓開發人員不必投注太多心力去理解郵件伺服器端(Server)與客戶端(Client)如何運作，能夠更專注於「功能」上的撰寫；另外，在未來新一代電子郵件通訊協定公佈後，開發人員也僅需要重新佈署通訊元件，不需要將軟體大幅翻新，間接延長軟體使用週期。

**2. 垃圾郵件防制策略**

本研究的另一個目的則是要針對氾濫的垃圾郵件研擬出有效

的防制策略。由於目前大部份垃圾郵件防制系統多半建置在郵件伺服器端，主要應用技術在於郵件發送端的驗證以及各種關鍵字演算法的分析，雖然技術上對垃圾郵件的攔截率已達一定水準，但是實際仍然無法滿足多數使用者的需求，其原因之一就在於伺服器端的防制系統不易進行訓練，而且為了符合多數使用者的標準，往往必須降低郵件過濾的靈敏程度，或是將已判定的垃圾郵件另置於其它資料庫中，以供使用者隨時回頭檢閱郵件是否遭系統誤判、誤刪，前者可能會導致系統對某一類型的垃圾郵件攔截率過低，而後者不僅讓使用者必須重新以人工方式檢查郵件清單，同時也降低使用者對郵件系統的信任。

另一項造成攔截率不佳的原因則是因為目前知名垃圾郵件過濾軟體大多來自歐美國家，這些軟體對於使用標準 ASCII 單位元組字集的英語系統有較高的辨識率，但是亞洲地區非英語系國家的文字系統多採 double bytes 雙位元組字集，造成過濾軟體辨識上的盲點，直接減低軟體過濾的成效。因此，本研究將以 RFC(The Request for Comments)的多用途網際網路郵件延伸標準 MIME(Multipurpose Internet Mail

Extensions)規格為藍本，建立 double bytes 辨識規則，並採用較廣泛的自然貝氏分類器(naive Bayesian classifier)，將郵件本文內容依關鍵字篩選後，計算其可能為垃圾郵件的聯合機率，以提供使用者自行訓練郵件過濾的模型，並且在客戶端自主建構第二道防線，不需要完全依賴郵件伺服器端的過濾機制。

## 第三節 研究範圍

本研究之主要研究項目有以下幾點：

**1. 分析 POP3 通訊協定**

首先分析 POP3(Post Office Protocol version 3)電子郵件通訊協定運作模型，以及伺服器端和客戶端之間進行溝通的方式，並進一步撰寫軟體通訊元件 Experiment Mail。

**2. 實驗 MIME 多用途多媒體格式**

依據 RFC 2045 公佈的 MIME 編碼格式，實驗 Base64 和 QP(Quoted-Printable)兩種電子郵件本文傳輸編碼格式的編/解碼程序。

**3. 驗證郵件標頭及本文特徵**

取得郵件本文並驗證郵件是否具備相關特徵，包含(1)郵件

是否經過正確編碼；(2)郵件標頭(head)所登載日期是否為合理的 UTC(Coordinated Universal Time)時間。

**4. 建構自然貝氏分類器辨識模組**

建構以自然貝氏分類器為基礎的辨識模組，並且透過觀察實際效果，逐一修正無效或辨識率過低的關鍵字，以期達到同時提高垃圾郵件攔截率、降低正常郵件誤判率的目標。

## 第四節 研究架構

本研究於第二章首先探討電子郵件的格式和垃圾郵件演進的歷史，以及各國對於垃圾郵件防制的相關法規，第三章則開始論述研究進行的方式和實驗的架構，第四章即進行系統的實作和執行結果的分析，最後在第五章敘述本研究成果之結論以及未來研究的方向。

# 第二章 文獻探討

本研究主要在發展一個可用的電子郵件存取元件，以及個人化的垃圾郵件防制策略，因此本章先探討電子郵件通訊協定和垃圾郵件相關文獻，其後再依目前業界使用中的垃圾郵件過濾技術，分為「郵件本文識別」、「寄件者身份識別」兩大類，並個別討論其相關的技術。

## 第一節 電子郵件

電子郵件在網路發展初期就已經出現，但是一直到 1980 年代中後期才開始具有比較完整的功能，並且隨著電腦系統的發展，也逐漸加入對多媒體格式的支援。

### (一) 電子郵件的格式

一封電子郵件的本文(Content) 以兩個 CrLf(歸位符號加換行符號)分為上、下兩個部份。上半部為「標頭」(head) ，用以描述郵件的基本資訊；下半部則為「主體」(body)，記錄寄件者所要傳遞的資訊內容。

郵件的標頭必須依照 RFC 4021 文件所規定的欄位編寫，雖然在文件中所記載的標頭欄位項目多達七十五項，但常用的欄位僅有十項左右，表 2.1 為常用的標頭欄位。

表 2.1 常用標頭欄位

| 欄位 | 說明 | 參考文件 |
|---|---|---|
| Date | 郵件的傳送日期 | RFC 2822 section 3.6.1 |
| From | 作者郵件地址 | RFC 2822 section 3.6.2 |
| Sender | 代寄者的郵件地址 | |
| Reply-To | 回覆的郵件地址 | |
| To | 收件人地址 | RFC 2822 section 3.6.3 |
| Cc | 副本的郵件地址 | |
| Bcc | 密件副本郵件地址 | |
| Message-ID | 郵件的識別 ID | RFC 2822 section 3.6.4 |
| In-Reply-To | 回覆郵件的識別 ID | |
| Subject | 郵件主旨 | RFC 2822 section 3.6.5 |
| Return-Path | 郵件回覆地址 | RFC 2822 section 3.6.7 |
| Received | 郵件傳遞記錄 | |

若郵件本身沒有額外的附加檔案(attach file)，主體部份除了必須對雙位元組字集進行編碼外，沒有其它特別規則，但是如果郵件當中夾帶附加檔案，就必須將附加檔案依據 MIME 規格加以編碼，並描述附件的文件類型、字元語系及傳輸時所用的編碼方式。

## （二）通訊協定

電子郵件系統為一種 Client/Server 的架構，其傳遞分別依循 SMTP(Simple Mail Transfer Protocol)以及 POP 兩項通訊協定。其中 SMTP 主要負責郵件投遞、繞徑的工作，使用通訊連接埠 25 與其它伺服器連結，最近一次協定的更新是在 2001 年 4 月(RFC 2821)。

當初在制定時，為了讓 SMTP 協定簡單、容易使用，因此並未加入身份防偽機制，寄件者很輕易就能夠偽造身份，這是造成目前網路詐騙、垃圾郵件充斥的主要原因之一。

至於 POP 則負責透過連接埠 110 提供客戶端收取郵件的服務，最新的版本為 1996 年更新的第三版(RFC 1957)，故一般又稱為 POP3 協定。由於前兩代的 POP 協定在功能上漸漸無法滿足使用者的需求，因此 IETF(Internet Engineering Task Force)又公佈了 IMAP(Internet Mail Access Protocol)協定用以取代 POP，而後因為一方面 POP3 推出後已增加許多進階功能，另一方面則因為部份郵件伺服器仍然未支援 IMAP，因此許多郵件軟體開發人員仍然以 POP3 為主要的協定。

## 第二節 垃圾郵件

### (一) 垃圾郵件的定義

從過去許多研究垃圾郵件防制的文獻資料，和國外對於郵件管理的法令規定中，歸納整理出下列三點特徵。

(1) 未經收件人同意

合法的郵件廣告商必定以「選擇加入(Opt In)」當作名單的主要來源，意即經過使用者同意後，才將使用者的郵件地

址加入到名單當中，並且也提供「選擇退出(Opt Out)」的機制，讓使用者能夠自行決定是否退出發送名單；而垃圾郵件廣告商的名單來源多半是從網路搜集，或是直接與其他廣告商交換、購買，因此垃圾郵件都是未經使用者同意的情況下不請自來的。

(2) 內容以商業行銷為主要目的

依據消費者文教基金會在民國 95 年的統計，台灣地區所收到的垃圾郵件廣告中，推銷情色光碟、情趣用品和成人交友網站的垃圾郵件就佔了 90.48%，另外其它像是投資理財、瘦身美容等，也都是以行銷商品為主要目的。

(3) 偽造寄件人身份

為了逃避 ISP 的封鎖或追查，垃圾郵件會偽造不同的身份和郵件地址，而且因為過去在設計郵件通訊協定時，對郵件寄件人身份驗證並沒有詳加規範，使得這種手法多半可以有效的躲過郵件伺服器的稽核。

某些特殊的郵件未必會偽造寄件人身份，例如選舉期間候選人在網路上發送文宣，又或是某公司舉辦週年慶活動所投遞的網路廣告，因此垃圾郵件未必一定完全符合上述三點特徵。

## (二) 垃圾郵件的發送

隨著網路的演進，垃圾郵件發信的方式和內容也不斷地跟著轉變，依發送的方式和內容格式大約可以分為三個時期。

第一個時期是撥接式網路的時代(1990~1995)，由於當時撥接式網路的速率侷限在 33.6Kbps 到 56Kbps，廣告商為了提昇發信效率，會先從網路上搜尋允許「轉信(reply)」的郵件主機，再將郵件大量傳送給郵件主機處理，使得這些主機在效能和頻寬成本上的負擔相當沈重，另外因為當時撥接式網路的頻寬不高，所以郵件的內容僅能作簡單的文字描述。

在第二個時期(1996-2000)時，ADSL 和 Cable 等寬頻技術進入網路市場。因為部份國際反垃圾郵件聯盟逐漸建立起轉信郵件主機黑名單，並且無償提供給其它合法郵件主機參考，因此廣告商除了使用允許轉信的郵件主機外，也開始使用網路業者提供的免費信箱來發送廣告信。

這個時期因為頻寬的提昇，廣告內容以網頁設計的方式大量加入插圖和 HTML 語法以吸引消費者的注意，「郵件點閱率」乙詞也在這個時期流行，指的是廣告郵件被使用者開啟閱覽佔發信總數的比例。

在第三個時期中，寬頻網路技術掛帥，撥接式網路逐漸淡出市場。因為免費信箱太過於容易取得，ISP 為了防止垃圾郵件進入自己的網路系統，開始封鎖其他網路業者的郵件主機，於是從免費信箱發送的郵件紛紛被阻擋在外，連帶使得廣告商不得不轉而尋求其它郵件發送方式。

因為技術的普及，伺服器架設的門檻大幅降低，廣告商嘗試自建郵件伺服器以取代免費信箱，另一方面，網路上也出現專業的廣告信發送軟體，僅需要本機接上網路，完全不需要透過代理的郵件伺服器即可發送，而且在操作上也相當地簡單，即使是網路初學者也能夠快速上手，使得垃圾郵件的數量近年來持續暴增。

為了躲避各種「防制垃圾郵件(anti-spam)」軟體的過濾，廣告郵件的內容反而回到一般的文字訊息表現方式，內容也愈來愈簡短，有時候甚至只有附上簡單的網址或圖片連結[5]。

茲將各時期的發送方式及內容格式整理於表 2.2。

[5] 使用圖片兩個目的，一是為了規避軟體的過濾，另外就是為了測試收件地址是否為有效的電子郵件信箱。

表 2.2 垃圾郵件轉變的三個時期

| | 第一時期 | 第二時期 | 第三時期 |
|---|---|---|---|
| 發送方式 | 1.採用撥接式網路。<br>2.透過 reply 方式，經由其它合法的郵件主機代送。 | 1.窄頻與寬頻網路並行。<br>2.透過 reply 代送。<br>3.使用免費信箱。 | 1.寬頻網路為主。<br>2.自建郵件伺服器。<br>3.使用廣告信發信軟體。 |
| 內容格式 | 內容簡單，純文字。 | 內容較豐富，使用 HTML 網頁圖文方式。 | 內容簡單，可能為純文字或 HTML 網頁。 |

## (三) 各國對於垃圾郵件管理的現況

有鑑於垃圾郵件對個人和企業所造成的影響甚大，近幾年來，許多國家都已著手制訂相關法規來約束商業電子郵件的行為，以下分別介紹美國、歐盟、日本以及台灣等四個國家的立法現況。

### 1. 美國

2003 年 11 月，美國國會表決通過「**管制濫送色情及行銷之侵擾法**(Controlling the Assault of Non-Solicited Pornography and Marketing Act, CAN-SPAM Act of 2003)」，美國總統於同年 12 月簽署，並於 2004 年 1 月 1 日生效。

該法案規範商業電子郵件不得偽造寄件人郵件地址，以及郵件主旨、內容，並且在郵件中必須提供使用者退出名單的機制，違反規定者，使用者可向其索賠最高 200 萬美元的賠償。

美國為全球垃圾郵件發送數量最龐大的國家，該法案的施行曾引起各國的注目，但是因為 CAN-SPAM 採取較寬鬆的「選擇退出」原則，而不是較嚴格的「選擇加入」，因此即使是色情郵件，只要郵件當中提供「選擇退出」，仍然視為合法的電子郵件，可逕自寄送到使用者的電子郵件信箱中，部份團體認為這項法案反而有助於垃圾郵件「就地合法」，因此使得 CAN-SPAM 自推動以來就倍受各界的抨擊。

依據美國聯邦交易委員會(Federal Trade Commission；簡稱 FTC)2005 年年底對外公布的「垃圾電郵防治法執行成效報告」中指出，垃圾電郵占所有電郵的比例持平[6]。

**2. 歐盟**

2002 年 7 月，歐洲議會與歐盟理事會通過第 58 號有關「**電子通訊中個人資料處理與隱私保護指令**(Directive 2002/58/EC of the European Parliament and of the Council, of 12 July 2002, concerning the processing of personal data and the protection of privacy in the electronic communications sector)」，簡稱「隱私及電子通訊指令(Directive on privacy

[6] 垃圾郵件相對於電子郵件總數的比例維持在 CAN-SPAM 法案實施前的水準。雖然從比例上觀察垃圾郵件並沒有繼續增加的現象，但由於受到資訊化社會的影響，電子郵件的數量仍然持續以驚人的速度成長，因此也讓垃圾郵件的數量不斷增加。

and electronic communications)」，取代了 1997 年之「電信事業個人資料處理及隱私保護指令(Directive 97/66/EC)」。

該指令共計 21 條，其中第 13 條「未經請求之通訊(unsolicited communications)」就是特別針對垃圾郵件所制定，當中規範寄件人不得以虛偽或匿名的身份或未提供有效回覆地址方式發送以直銷為目的之電子訊息，並且採取較嚴格的「**選擇加入**」原則。至於因出售產品或服務而取得顧客電子郵件地址者，則被許可利用取得之地址進行其固有的類似產品或服務之直銷，但是仍然必須在郵件中提供「選擇退出」的機制。由於歐盟的體制較為特別，因此違反者的相關罰則是由各會員國自行訂定國內法。

**3. 日本**

2002 年 4 月(平成 17 年 4 月)，日本制訂「**特定電子郵件傳送標準化法**」，2003 年 7 月、2005 年 5 月分別再對當中的條文進行修訂。

日本在商業電子郵件傳送的名單也是採取「**選擇退出**」的原則，不過該法明訂商業電子郵件發送須經由主管機關核准，違反法令者可處以一年以下的有期徒刑並科罰金日幣一

百萬元，郵件寄件人若偽造不實身份則可處罰金三十萬日幣。

**4. 台灣**

我國立法院曾在民國 89 年 5 月及 91 年 6 月提出「電子廣告信件管理條例草案」，惟在一讀付委後均未能及時完成立法程序而遭擱置，後由行政院「通訊傳播委員會籌備處(National Communications Commission, NCC)」於 94 年 2 月草擬完成「**濫發商業電子郵件管理條例**」，並送交立法院完成一讀程序，目前正由科技及資訊委員會審查中。

該案主要參考美國 CAN-SPAM 採用「選擇退出」原則，對於商業電子郵件的定義和發送的規範大致相同，在罰則部份則明定商業電子郵件若導致收信人損害時，收信人得以向發信人請求每人每封郵件五百元以上二仟元以下的賠償，發信人最高賠償額為新台幣二千萬元，但若發信人所獲得的利益超過此一賠償上限時，則參考「電腦處理個人資料保護法」修正草案第二十八條第四項規定，以該所得利益為限。

同時，為了加強對於商業電子郵件的約束力，該案也參考英國「隱私與電子通訊條例」第二十二條及第二十三條規

定，明定廣告主或廣告代理商於明知或可得而知發信人有違法發送商業電子郵件情事時，應與發信人共同對外負連帶損害賠償責任，促使廣告主於委託發送廣告時，要求發信人不得違反相關規定。

## 第三節 垃圾郵件的過濾技術

從過去的研究中我們可以發現，垃圾郵件的過濾方法以分類器演算法為主，也就是從郵件本文中分析該郵件可能為垃圾郵件的機率，而後因垃圾郵件對全球網路所造成的損害日益增加，於是微軟等國際資訊大廠也開始著手積極制定新一代的電子郵件傳送機制，以下將其分為「郵件本文識別」、「身份識別技術」兩類，分別進行探討。

### (一) 郵件本文識別

#### 1. 貝氏分類法

貝氏分類法【5】是一種以統計學的貝氏分析理論為基礎所設計的分類方法，其底下又分為「自然貝氏分類器」(或稱簡易貝氏分類器)和「樹狀貝氏分類器」兩種。這兩種分類器的差別在於自然貝氏分類器假設每一項條件彼此之間的關係是獨立的，因為運算流程簡單，因此系統容易實作；而

樹狀貝氏分類器則假設條件之間具有一定的關聯性，因為運算過程將條件的關聯納入計算，所以能夠獲得更高的正確性，但是由於運算規則稍較自然貝氏分類器複雜，系統建置的難度也相對提昇。

本研究即以自然貝氏分類器為基礎，將各項發生的條件視為獨立事件，並透過計算其聯合機率，以推導郵件本身是否為非法的垃圾郵件。在研究中我們將此問題描述為「在 A 條件下發生 B 事件的機率等於 AB 同時發生的機率除以 A 發生的機率」，其中 A 條件為所有符合的項目的合併機率，B 事件即為可能是垃圾郵件的機率，將以上描述再以數學式表示為：

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} = \frac{P(B) \times P(A \mid B)}{P(A)} \qquad \text{(式 2-1)}$$

**2. 基因演算法**

基因演算法(Genetic Algorithms, GAs)乃於 1975 年由 Holland 首先提出，它的精神在於摸擬生物界中「物競天擇，適者生存」的進化法則，並且實際透過複製(亦稱為選擇)、交配和突變三個步驟進行演化，以求出最佳的染色體配對。

在演化的過程中需要有對應的條件來評估產生的染色

體是否適當，這個條件就稱為適應函式(fitness function)，由於適應函式是由最佳化的參數(基因)所組成，因此參與最佳化的參數愈多，相對地，系統搜尋的範圍就會愈廣，所花費的時間也會增加，需要演化更多代之後，才有可能找出最佳的參數解。

## (二)身份識別技術

寄件者身份識別是近幾年才開始推行的垃圾郵件防制技術，它的特色是系統的過濾流程簡單，同時不需要複雜的演算就能夠判別郵件是否合法，而且對於郵件誤判(將正常郵件判定為垃圾郵件)的機率遠低於使用郵件本文過濾的方式。

雖然比較起來有較好的效能和準確率，但是這一項技術目前在推動上尚有困難，主要原因是因為寄件者身份識別並不像郵件本文過濾器是一個可以獨立建置的系統，它在運作的時候還必須與網域名稱伺服器(Domain Name Server, DNS)交換以查詢寄件人所公佈的網域資訊，因此需要對 DNS 現行的規格額外加以定義，同時郵件伺服器也必須能夠支援身份識別技術的應用，因此要實作這項技術必須大規模改變現行之郵件系統架構，在執行上的困難度較高。

此外，雖然美國線上、Yahoo!、微軟以及 EarthLink 共同建立了反垃圾郵件聯盟，但是聯盟旗下的會員卻又各自推出不同的識別技術，也是導致這項技術一直難以標準化的原因。

以下就針對 Yaoo!的 Domain Keys 和微軟的 Sender ID 兩家主流的身份識別技術加以介紹。

**1. Domain Keys**

Domain Kyes(網域認證鑰匙)【6】是由「Yahoo!反垃圾郵件小組」所研發的認證技術，主要是透過非對稱式加密的方法識別郵件來源是否為合法的寄件人，系統運作流程說明如下。

**寄送郵件**

(1) 支援 Domain Keys 技術的寄信伺服器會在郵件的傳送過程中產生兩組金鑰，一組為公開金鑰(public key)，另一組則為私密金鑰(private key)。

(2) 郵件傳送前，寄信伺服器會嘗試將公開金鑰存入網域名稱伺服器(Domain Name Server, DNS)中，並由 DNS 驗證網域名稱是否有效以及是否經過寄信端偽造。

(3) 通過 DNS 認證後，寄信伺服器會依據私密金鑰產生

一組數位認證簽名檔，並將簽名檔附加於郵件的標頭中一併傳送到收信伺服器。

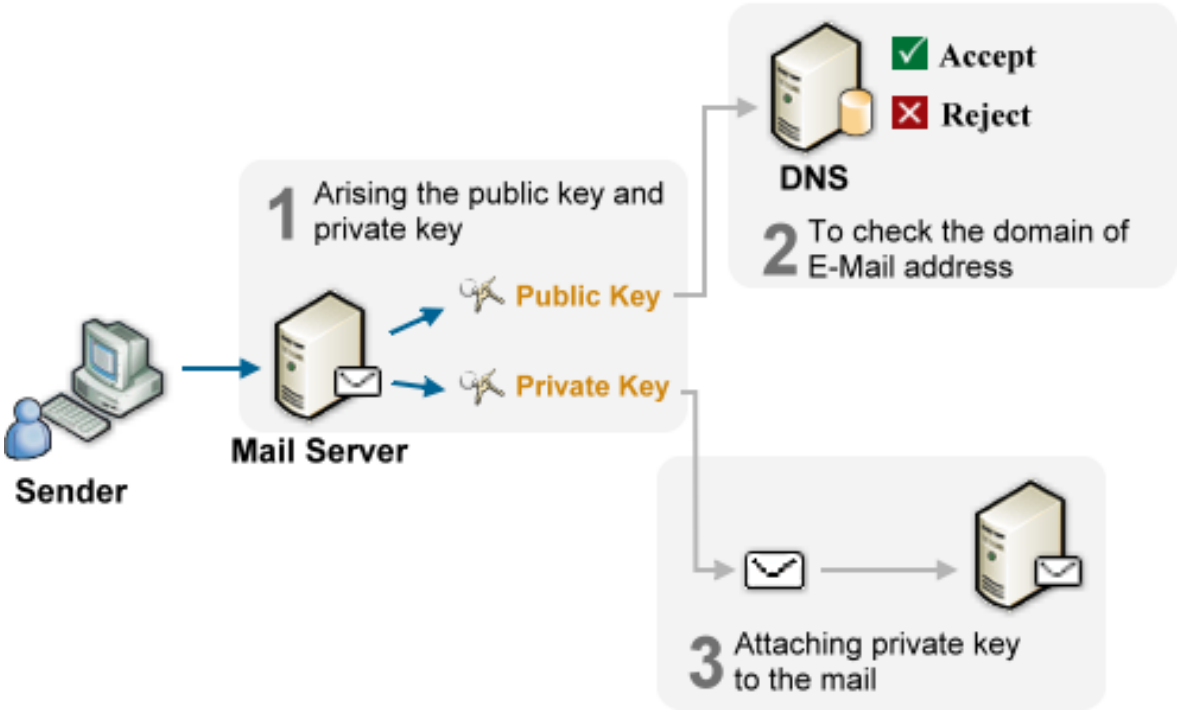


圖 2.1 網域認證鑰匙傳送模型

**接收郵件**

(1) 當收信伺服器接收到郵件時，會先判別郵件是否使用 Domain Keys 技術，並且從郵件和 DNS 系統分別取得私密金鑰及公開金鑰。

(2) 如果郵件內含私密金鑰，但是收信伺服器卻無法向 DNS 取得公開金鑰，則表示郵件異常或寄信來源不

合法，則可將該封郵件丟棄或隔離。

(3) 取得兩組金鑰後，比對郵件的寄件者名稱是否符合此網域，若兩組金鑰不相符，表示郵件偽造他人的網域名稱，因此收信伺服器可直接丟棄該封郵件。

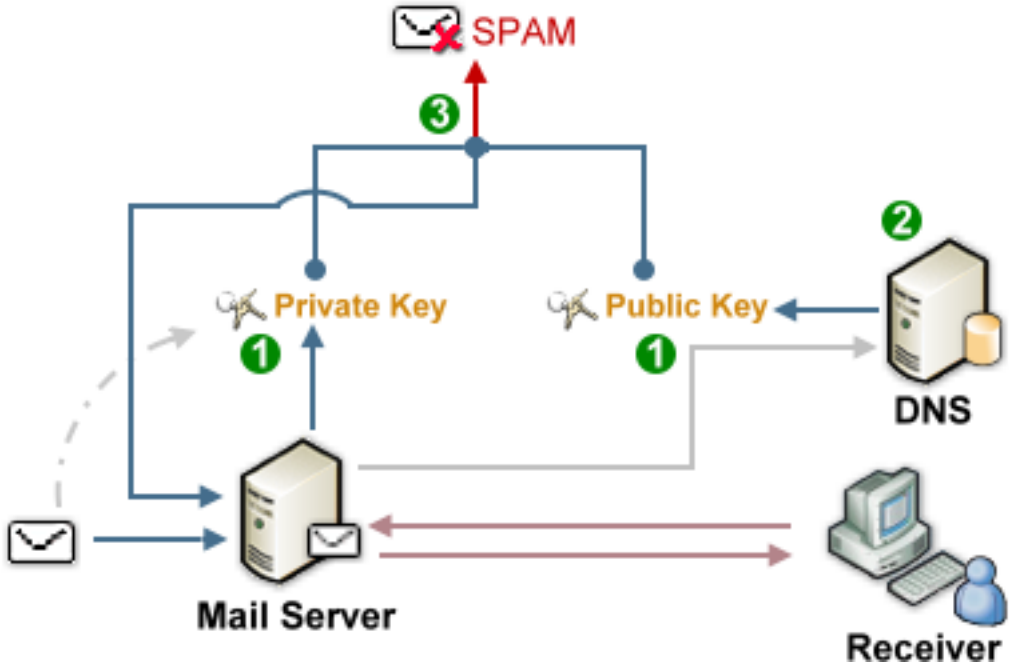


圖 2.2 網域認證鑰匙接收模型

**2. Sender ID**

Sender ID(寄件者身份識別)【7】最初稱為 Caller ID for E-Mail(電子郵件來電顯示技術)是由微軟(Microsoft)所推動的身份識別技術，在架構上和 Domain Keys 類似，兩者都是透過 DNS 去驗證寄件人是否偽造網域名稱，不同的是採用 SenderID 技術的郵件伺服器在收取信件後，會將信件交由

Sender ID Framework(Sender ID 的運作核心)從郵件中取得 PRA(Purported Responsible Address)，並查詢 DNS 的 SPF(Sender Policy Framework, 寄件者政策框架)，檢查 PRA 中所描述的網域名稱是否和 SPF 相符，流程如圖 2.3 所示。

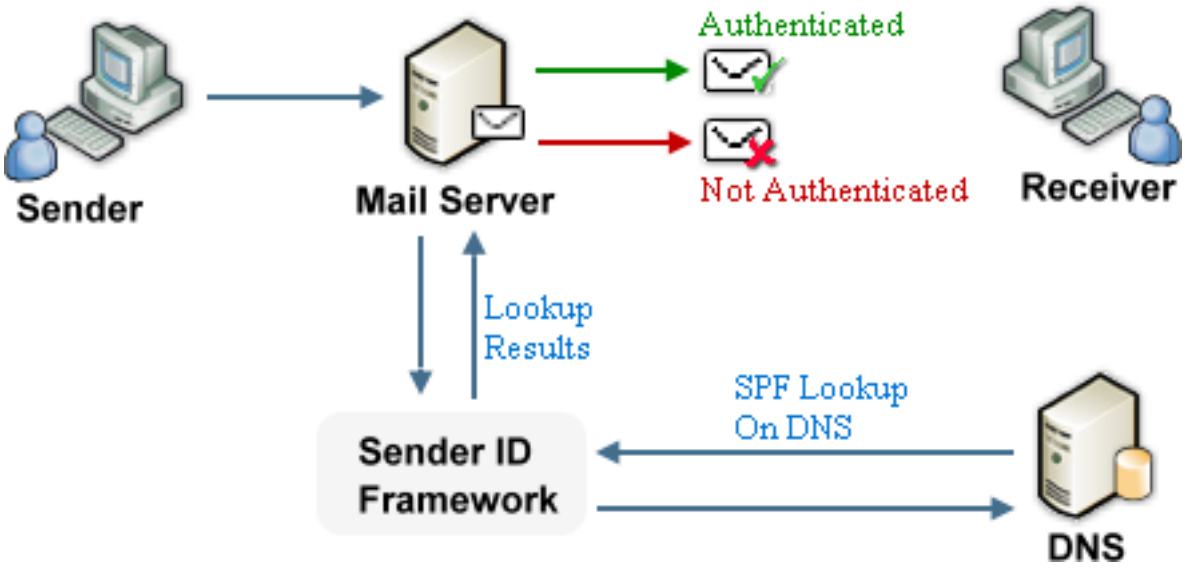


圖 2.3 Sender ID 運作模型

微軟在推動 Sender ID 的過程中遇到不少的問題，其中之一是因為微軟要求使用 Sender ID 技術的郵件服務商必須簽署一份技術授權合約，這項訊息在公佈之後就立即遭到網路業界的反彈。

開放原始碼組織 OSI(Open Source Initiative)對微軟的行為表示無法理解，他們認為既然微軟宣稱這是一項免費的技術，卻又要求業界簽署授權合約的舉動是不合理的；同時開

放原始碼團體 Apache Foundation 也認為合約中的授權條件太過於嚴苛，將會阻礙這項技術成為開放原始碼領域的標準，因此未來 Apache 將不會支援 Sender ID 技術。

造成問題的另一項原因則是因為剛開始微軟與美國線上互相結盟，成為反垃圾郵件技術上的夥伴，並且在 Sender-ID 中採用美國線上所開發的 SPF 協定，但是在 Sender ID 公開發行後不久，美國線上卻發現微軟修改了這個協定，使得 Sender ID 無法完全回溯相容於原始的 SPF，讓人認為微軟自行擴展 SPF 並藉以取代原始版本的意味相當濃厚，因此美國線上一度相當惱火，曾公開表示將不再支持微軟的 Sender ID，負責審議反垃圾郵件協定的標準組織 IETF(Internet Engineering Task Force, 網際網路工程工作小組)也基於上述兩項原因退回微軟 Sender ID 的提案。

事後微軟不得不立即採取補救措施，將舊版 SPF 規格也納入 Sender ID 系統當中，並且修改授權合約的內容，以求重新爭取各界的支持，雖然後來與美國線上的合作未破局，但是已經比 Domain Keys 的應用慢了一步。

# 第三章 研究設計

## 第一節 研究架構

本研究之系統架構如圖 3.1 所示，共分為三個部份，第一個部份為系統的輸入，此次研究的資料來源以中文電子郵件為主；第二部份是系統的郵件處理程序，其中包含我們自行開發用來接收來源郵件的元件，以及以 Double Bytes、自然貝氏分類器為核心的郵件過濾器；第三部份則為系統的輸出，即研究的結果。

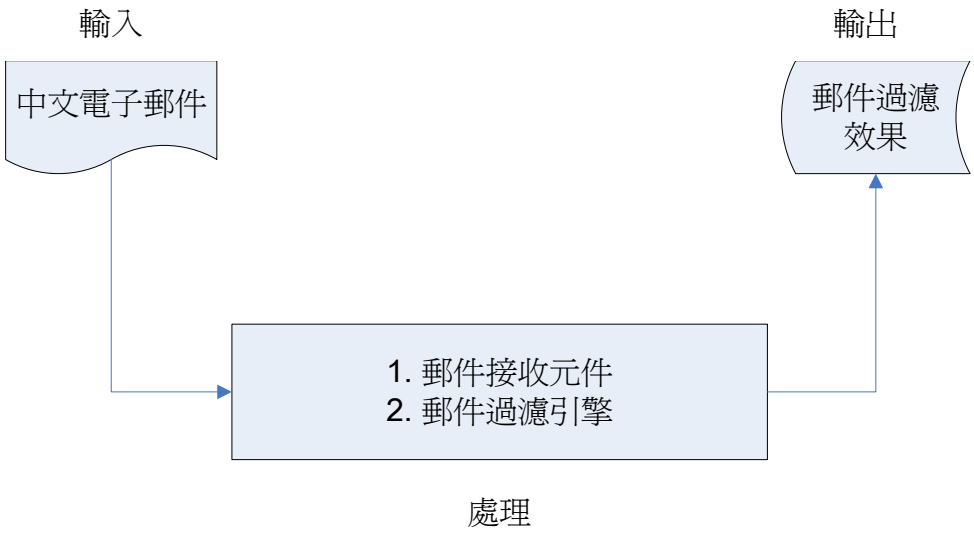


圖 3.1 研究架構圖

## 第二節　研究流程

本研究之系統流程分為兩個階段，第一階段為郵件接收元件的開發流程，如圖 3.2 所示。

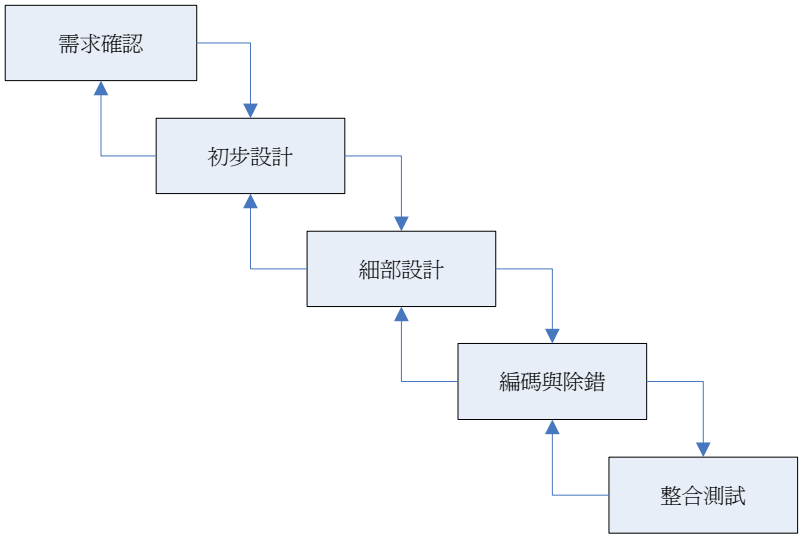


圖 3.2 郵件接收元件開發流程圖

針對系統需求，我們在開發初期訂定了三個明確的目標（1）具備連接郵件伺服器的能力；（2）具備非同步傳輸指令功能；（3）接收指定郵件的標頭。接著在初步設計階段，主要工作在於建立網路通訊功能，以及制定相關介面(Interface)。細部設計則開始實作前述介面功能，並增加錯誤處理函式，讓元件能夠處理各種例外狀況。各部功能完成即進行程式封裝及單元測試，並調校程式效能。如單元測試無誤，則將元件嵌入系統中進行整合測試，

第二階段則導入上一階段所開發之元件，實際進行郵件過濾效果的分析，如圖 3.3 所示。

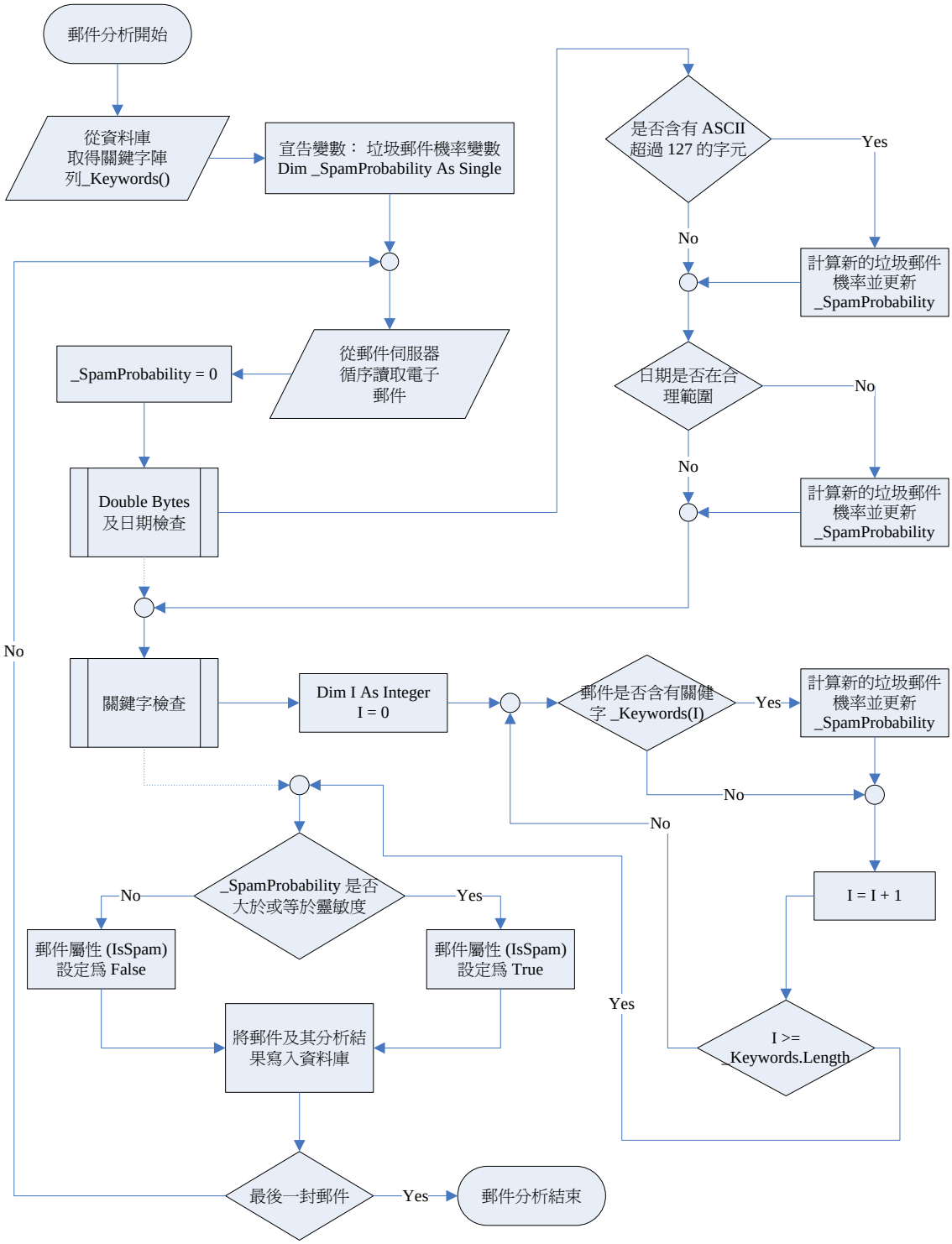


圖 3.3 郵件分析流程圖

圖 3.4 為系統完整的架構模型。模型左側表示系統元件關係，最底層開始為 Microsoft .Net Framework 2.0，其上再分為兩個類別，一是架構在 System.Net.Socket 類別之上的 Extend.Net.Mail 類別，主要負責與郵件伺服器之間的郵件存取；另一個類別則為系統過濾的核心類別 Extend.MyWeb.MyAntiSpam。模型右側則為系統的儲存單元，主要由 Microsoft SQL 2000 MSDE 組成。

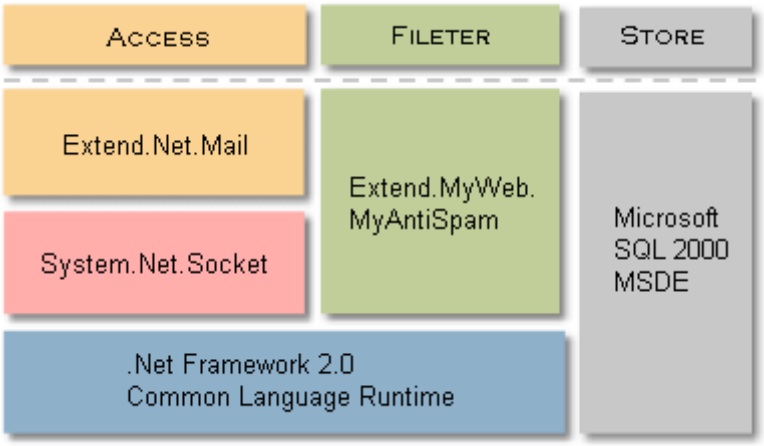


圖 3.4 系統架構模型

## 第三節 研究步驟

本研究初期先行建置郵件接收元件，再進行郵件過濾系統的分析，完整實驗分為九個步驟，分別為：

1. 電子郵件元件之開發
2. 驗證元件是否能夠正常運作
3. 資料集（郵件）之取得
4. 郵件標頭分析
5. 解讀郵件主題、寄件人及收件人名稱
6. 使用 Double Bytes 進行郵件過濾
7. 檢查郵件日期是否正確
8. 使用自然貝氏分類器進行郵件過濾
9. 分析實驗結果

### (一) 電子郵件元件之開發

研究首先以 Microsoft .Net Framework 建置類別庫專案(Class Library Project)，並封裝命名為 Extend Mail，同時實作與郵件伺服器的連接(Connection)、中斷連接(Disconnection)、登入(Login)、登出(Logout)、取得指定郵件(GetMail)等方法，以方便系統後續的開發及研究測試。

## (二) 驗證元件是否可以正常運作

網路程式可能發生的狀況遠多於單機程式，為了確保在研究中能夠順利取得正確的資料，Extend Mail 元件需要具備處理例外狀況的能力，並且通過「內部封閉測試」和「外部開放測試」兩階段的驗證。

內部封閉測試由本研究的成員進行，重點在於元件效能的改進，以縮短未來研究所需要的時間。外部開放測試則將 Extend Mail 當中負責資料傳輸的程式碼抽離，另外封裝為 TCP Socket 元件，測試重點為網路連接及資料傳輸的正確性，我們將該元件公佈於程式設計社群「藍色小舖」中，提供給其他程式設計人員測試使用，說明暨下載網址為 http://www.blueshop.com.tw/download/show.asp?pgmcde=PGM20060203095636KHZ

## (三) 資料集之取得

本研究的分析對象為電子郵件，為避免實驗結果受到人為控制，而導致產生的數據有所偏剖，資料來源將實際自 talk.idv.tw 及 yahoo.com.tw 兩個網域共三個郵件帳號中取得，由於研究過程中 yahoo.com.tw 信箱於 2006 年 9 月曾進行系統變更，導致該信箱無法再參與實驗，因此實驗後期的資料收集方式也配合稍加改變，以下將資料收集分為兩階段說明。

第一階段，因為 yahoo.com.tw 信箱免費會員僅能透過網站的使用者介面進行郵件的操作，不允許直接使用 POP3 方式讀取郵件，所以需要另外透過免費軟體 FreePOPs[7]進行。實驗過程三個信箱皆可正常收發電子郵件,每日約可取得不重覆的郵件樣本約五百封，此階段合計郵件樣本數為 6363 封。

第二階段，yahoo.com.tw 信箱介面變更，FreePOPs 無法正常讀取該信箱郵件，因此放棄該信箱，並將收集頻率從一至三日收集一次，變更為一個月收集一次，此階段累積郵件樣本數為 4211 封。

Extend Mail 元件最後版本為 1.2.0.3，表 3.1 列示其成員。

表 3.1 Extend Mail 元件之成員

| 成員 | 說明 |
|---|---|
| 公開方法 | |
| New | 多載, 建構子 |
| Connect | 連接郵件伺服器 |
| Disconnect | 中斷郵件伺服器的連接 |
| Login | 登入郵件伺服器 |
| Logout | 登出郵件伺服器 |
| Receive | 多載, 接收自郵件伺服器的訊息 |
| Send | 送出訊息到郵件伺服器 |
| Delete | 刪除伺服器上指定的郵件 |

[7] FreePOPs 為一免費信箱輔助軟體，能夠幫助使用者透過 POP3 的方式取得免費信箱的郵件，並且支援包含 Yahoo!、Hotmail、AOL、Gmail …等十幾種 web mail 格式。

## (四) 郵件標頭分析

取得郵件後，可自標頭資訊中再分析取得寄件人(From)、收件人(To)、主題(Subject)、日期(Date)四項實驗所需的主要資訊，郵件標頭樣本範例如下所示。

*Received: from mail.talk.idv.tw ([221.169.52.181]) by company.mail ([127.0.0.1])*
*with MultiPOP (MDaemon.PRO.v6.0.3.R)*
*for ; Mon, 30 Oct 2006 08:50:56 +0800*
*Received: from enews.cph.com.tw ([59.120.174.134])*
*by talk.idv.tw ([221.169.52.181])*
*with SMTP (MDaemon.PRO.v6.0.3.R)*
*for ; Sun, 29 Oct 2006 05:39:37 +0800*
*Date: Fri, 27 Oct 2006 19:17:31 +0800*
*Reply-To: =?big5?B?oW1Eb3dubG9hZCEguvS49L7Hst+7eKFu?=*
*From: =?big5?B?oW1Eb3dubG9hZCEguvS49L7Hst+7eKFu?=*
*To: "Johnny123"*
*thread-index: MTE3N18xNjcxXzQ3NzMwNw==*
*Subject:*
*=?big5?B?p0u2T6/BqPqhdaVkpNq0tbDyqL6scrNuxemhdrjVpc6qqaFFwvi7eLnPrtGmQ*
*Q==?=*
*=?big5?B?pfS/76FJ?=*
*MIME-Version: 1.0*
*Content-Type: multipart/alternative;*
*boundary="----=_NextPart_000_3EC60_01C6F9FC.8D011110"*
*Content-Class: urn:content-classes:message*
*Importance: normal*
*Priority: normal*
*X-MimeOLE: Produced By Microsoft MimeOLE V6.00.3790.2757*
*X-MDMultiPOP: johnnyfang@mail.talk.idv.tw*
*X-MDRemoteIP: 221.169.52.181*
*X-MDRcpt-To: JohnnyFang@company.mail*
*X-MDaemon-Deliver-To: JohnnyFang@company.mail*

## (五) 解讀郵件主題、寄件人及收件人名稱

由於 RFC 822 在制定郵件規格時，規範郵件標頭及本文僅能使用 127 個 ASCII 字元，後來到了 RFC 2822 才加入對多媒體格式的支援，因此在郵件中非 ACSII 的字元都需經過編碼。

經由檢查主題及寄件人(收件人)名稱格式可判別字串是否經過編碼。編碼可採用 Base64、QP 或 UU[8](unix-to-unix encoding)編碼格式，經過編碼後的字串格式應為：

=?語系字集?編碼格式縮寫?編碼後字串?=

茲舉前一小節標頭樣本中的寄件人欄位說明。在該欄位中以「?」字符分隔即可取得語系字集、編碼格式、編碼後字串三項資訊，分別為「big5」、「B」及「oW1Eb3dubG9hZCEguvS49L7Hst+7eKFu」。

**big5**：表示郵件原始語系為 Big5 中文大五碼。

**B**：表示郵件為使用 Base64 編碼，若為 QP 編碼，則顯示為 Q。

依前兩項資訊將編碼後的字串以 Base64 解碼還原，可得到編碼前原始字串為「《Download! 網路學習誌》」。

## (六) 使用 Double Bytes 進行郵件過濾

因為中文字字數遠多於 ASCII 所能表達的範圍，因此 Big5[9]內碼使用兩個位元組來組成一個中文字，並將其區分為高位組與低元組，其中高位組的最高位元必定為 1。

由於 ASCII code 大於 127 的字元[10]就應該進行 MIME 編碼，但

---

[8] UU 為早期 UNIX 所使用的編碼格式，目前多數系統已轉為使用 MIME 標準的 Base64 及 QP。

[9] 中文內碼除 Big5 外，尚有「通用碼」、「倚天碼」、「公會碼」、「王安碼」、「IBM5550」、「電信碼」、「漢英碼」等，但以 Big5 較為普遍使用。

[10] ISO646 規範 ASCII 內碼為 7bits，使用 8bits 來表示時，其最高位元為 0。

是分析目前網路上的中文垃圾郵件多數未經編碼程序，因此可對郵件的標頭進行內碼的檢測，如果發現任一位元組大於 127，就假定該郵件為垃圾郵件，並計算其垃圾郵件機率，計算公式如下。

$$SpamP = \frac{DBSpamMails}{TotalSpamMails} \quad \text{(式 3.1)}$$

$$NormalP = \frac{DBNormalMails}{TotalNormalMails} \quad \text{(式 3.2)}$$

**TotalSpamMails**：所有的垃圾郵件總數。

**TotalNormalMails**：所有的正常郵件總數。

**DBSpamMails**：符合 Double Bytes 規則的垃圾郵件總數。

**DBNormalMails**：符合 Double Bytes 規則的正常郵件總數。

**SpamP**：郵件為垃圾郵件的獨立機率。

**NormalP**：郵件為正常郵件的獨立機率。

Double Bytes 雖然是獨立事件，但是在下一階段會將其納入貝氏分類器中與其它關鍵字共同計算聯合機率。

## (七) 檢查郵件日期是否正確

因為使用者操作郵件軟體多半習慣依收件日期排序，在實際研究垃圾郵件發送情況時可發現，垃圾郵件的發送者會將日期變造為尚未到達的日期，讓郵件在排序的時候能夠始終在郵件列表的最前頭，以達到增加郵件的曝光的機會。

因此研究中也以系統日期為標準進行假設，若郵件發送日期在系統日期之後，則將該郵件視為變造日期的垃圾郵件，並計算其垃圾郵件機率，機率計算公式同式 3.1、式 3.2。

## (八) 使用自然貝氏分類器進行郵件過濾

將資料庫中的關鍵字逐一與自步驟三中取出的郵件主題及寄件人名稱比對，檢查是否有符合的項目，若符合則計算該項目為垃圾郵件的機率，計算公式仍然與式 3.1、式 3.2 相同。

在完成所有關鍵字檢查後，依據式 3.3 及式 3.4 將所有符合項目的個別機率合併計算，求得垃圾郵件及正常郵件的合併機率分別為 SpamProbability 及 NormalProbability。

$$Spam\Pr obability = \left(SpamP_1 \times SpamP_2 \times ... \times SpamP_n\right) \quad \text{(式 3.3)}$$

$$Normal\ \Pr obability = \left(NormalP_1 \times NormalP_2 \times ... \times NormalP_n\right) \quad \text{(式 3.4)}$$

最後將兩項機率代入式 3.5，即可求得垃圾郵件的聯合機率，其公式如下所示，其中 $SpamP_n$ 為每一個獨立項目的個別垃圾郵件機率，$NormalP_n$ 則為個別的正常郵件機率。

$$\text{Probability} = \frac{\left(\frac{\text{TotalSpamMails}}{\text{TotalMails}}\right) \times \text{SpamProbability}}{\left[\left(\frac{\text{TotalSpamMails}}{\text{TotalMails}}\right) \times \text{SpamProbability}\right] + \left[\left(\frac{\text{TotalNormalMails}}{\text{TotalMails}}\right) \times \text{N}ormal\text{Probability}\right]}$$

(式 3.5)

## (九) 分析實驗結果

將郵件過濾前的真實情形與過濾結果比較，可得到垃圾郵件及正常郵件的召回率(Recall)及精確率(Precision)兩項數據，並計算兩數的調和平均值(即 F1-measure)，以為評估系統的依據。

由於一般在統計學上只會考慮事件單一面向發生的情形，像是車廠在召回車輛檢修時，多半僅會注意實際有問題的車輛數目，而乎略或不考慮沒有問題的車輛數目，也因為如此，通常召回率愈高，精確率就會愈低；反之，精確率愈高，召回率就會愈低。但是因為在郵件系統中必須同時兼顧過濾的效能和正確性，因此研究的目標是如何同時提昇垃圾郵件及正常郵件的召回率、精確率。

實驗結果產生後，可能發生的情形有以下四種：

1. 對垃圾郵件的分析結果正確。

2. 對正常郵件的分析結果正確。

3. 未擊中(Miss, 系統未能將垃圾郵件挑出)。

4. 誤判(Mistrial, 系統誤將正常郵件判斷為垃圾郵件)。

# 第四章 系統實作與評估

## 第一節 系統環境設定

以下先就使用之硬體設備、軟體環境及開發工具分別說明。

### （一）系統硬體

**1. 實驗主機一**

中央處理器：Intel PentiumM 2.4G

記憶體：2GB

硬碟：60GB

網路卡：10/100 MB

**2. 實驗主機二**

中央處理器：Intel Pentium 4 1.8G

記憶體：1GB

硬碟：80GB

網路卡：10/100 MB

### （二）軟體環境

作業系統：Microsoft Windows XP Service Pack 2

資料庫系統：Microsoft SQL 2000 MSDE

網站伺服器：IIS 5.x + Microsoft .Net Framework 2.0

## （三）開發工具

網站開發：Microsoft Visual Web Developer 2005 Express

程式開發：Microsoft Visual Basic 2005 Express

Microsoft .Net Framework 2.0 SDK

文件撰寫：OpenOffice 2.4.x

Microsoft Word 2003

## （四）系統畫面

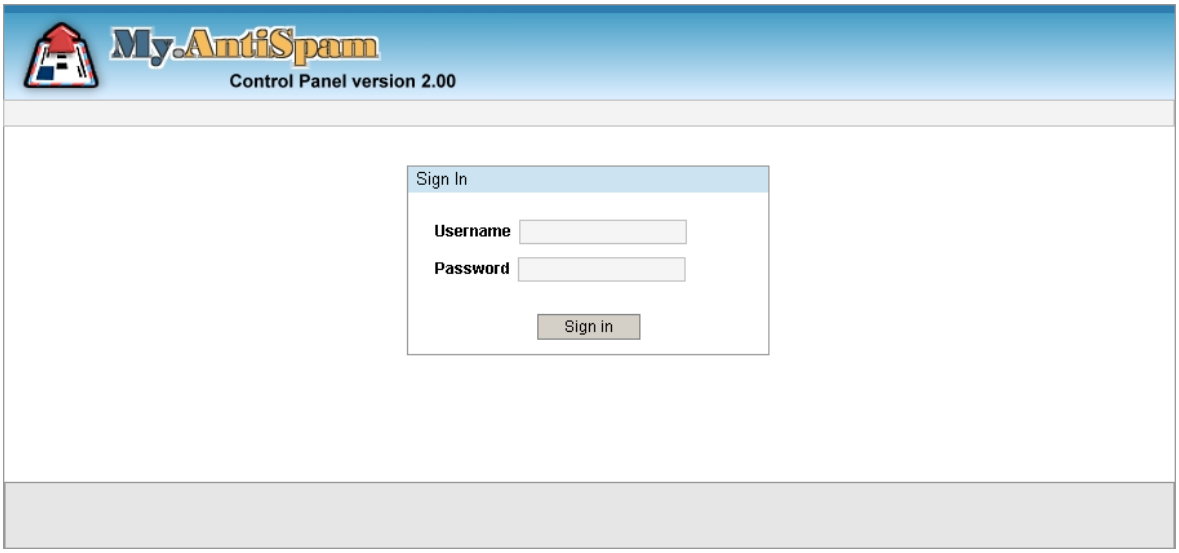


圖 4.1 系統登入畫面

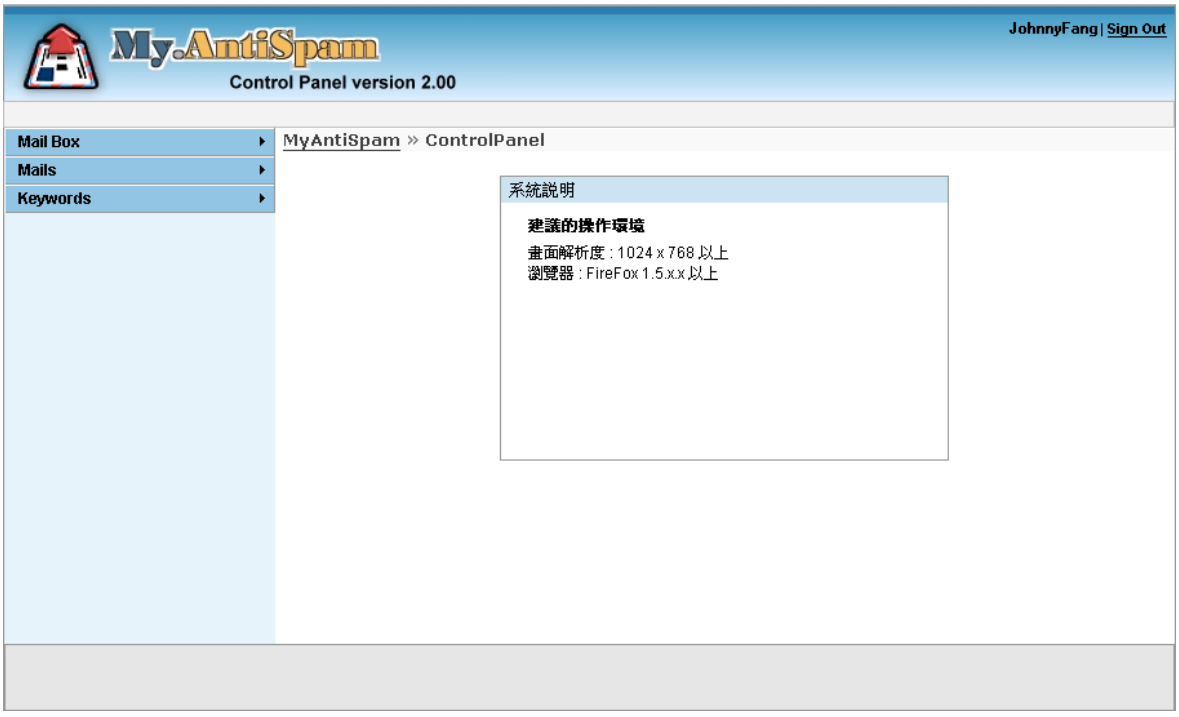


圖 4.2 系統主畫面

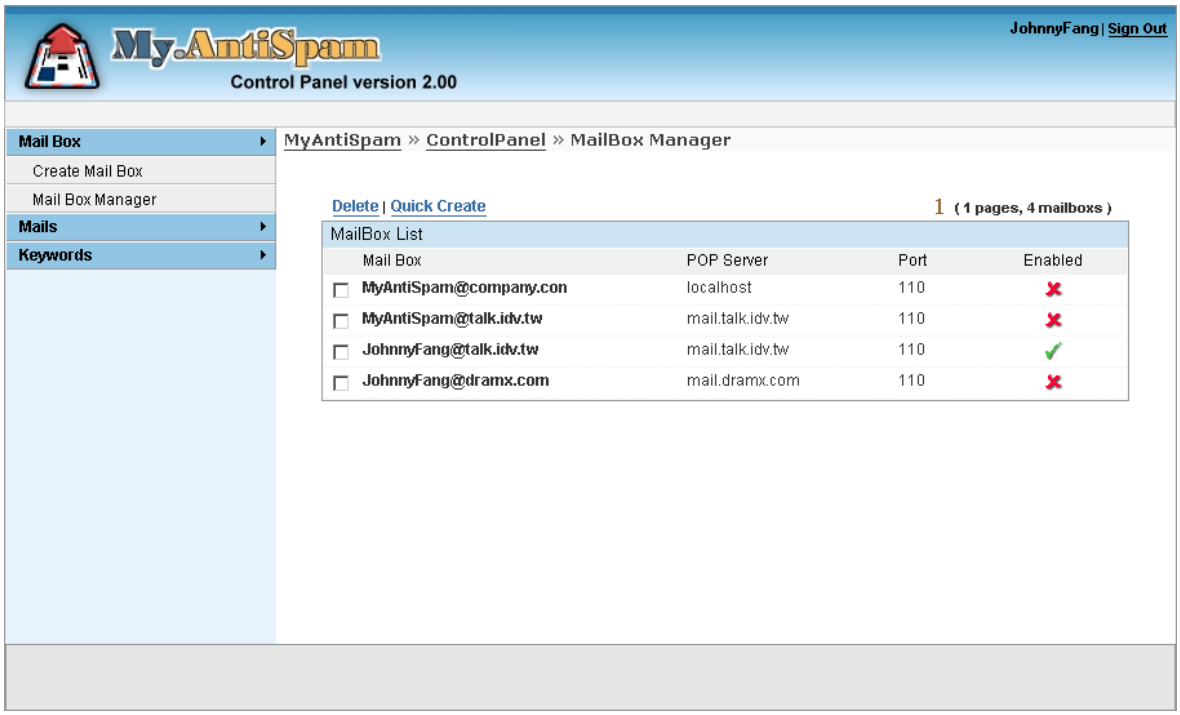


圖 4.3 信箱管理主畫面

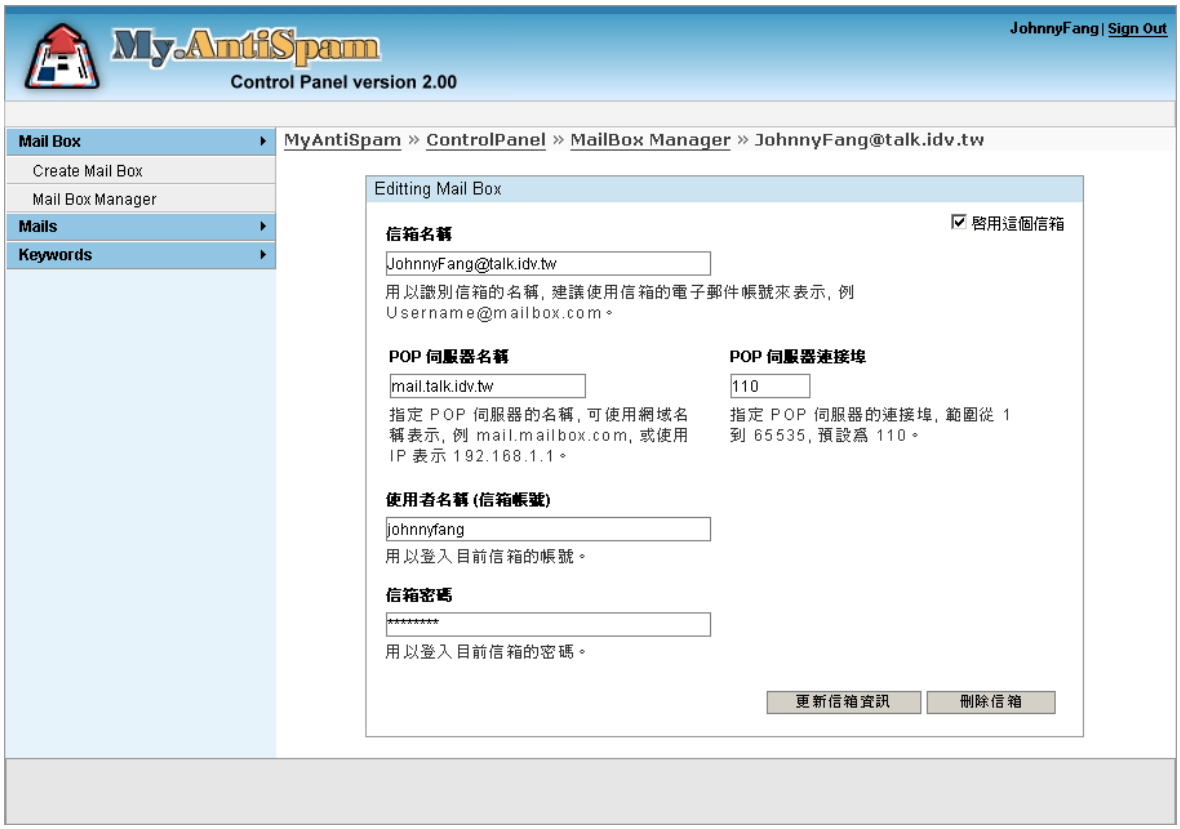


圖 4.4 信箱資料維護畫面

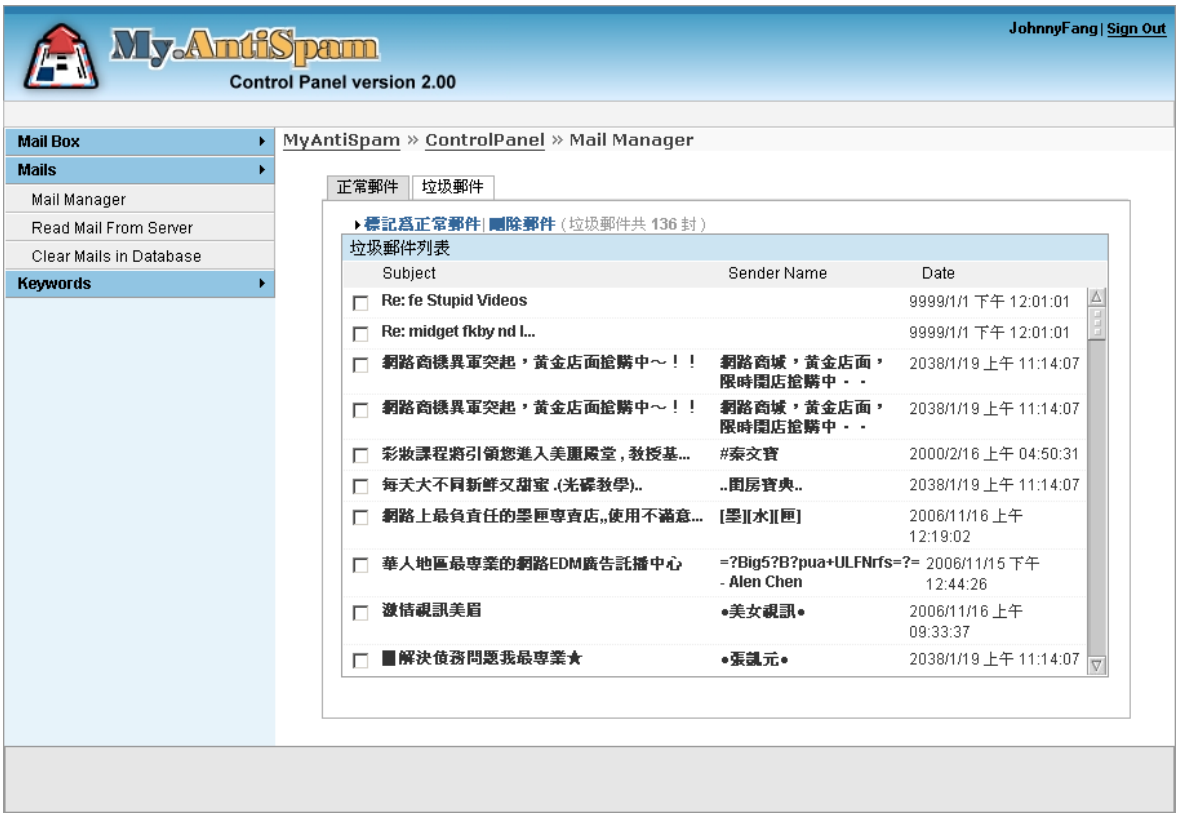


圖 4.5 郵件管理主畫面

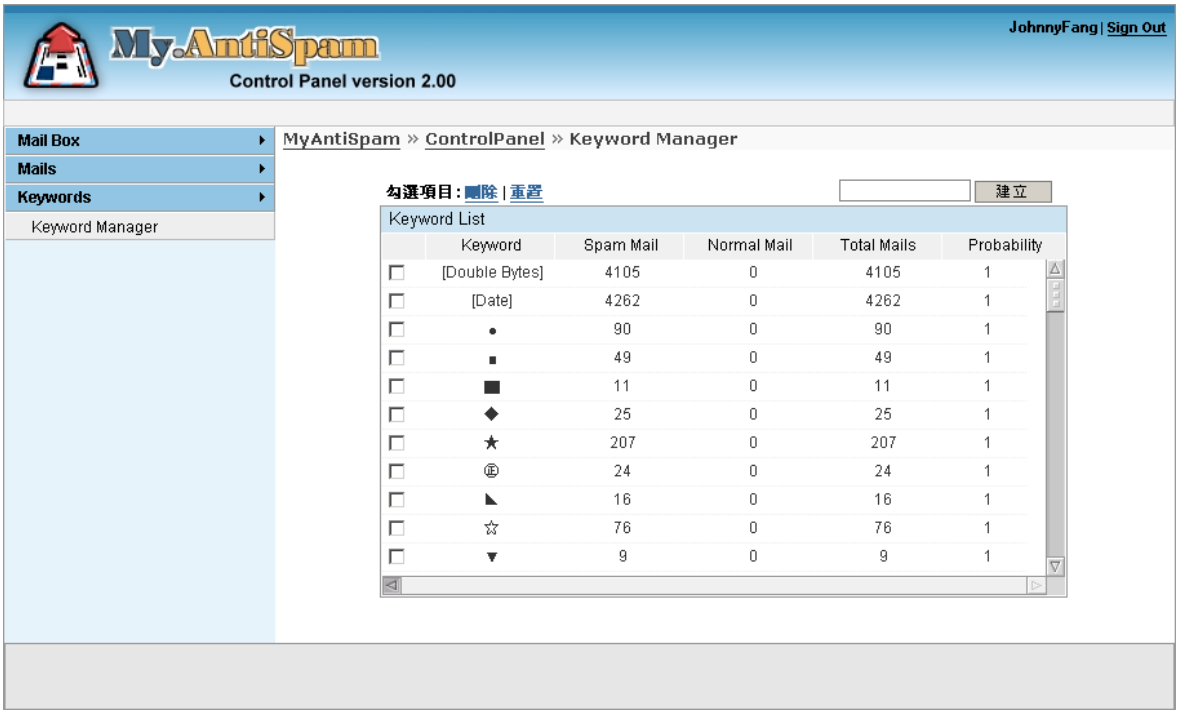


圖 4.6 關鍵字維護主畫面

## 第二節　資料集結果分析及系統調整

表 4.1 為第一階段實驗的分析結果，第一至十五回為系統建置初期，尚未進行任何調整前的實驗數據，第十六回以後，參考前十五回測試的情形，系統有相當程度的改良。

表 4.1 第一階段分析結果

| 導入系統的第一個版本進行實驗 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 計次 | Total | SR | SP | SF1 | NR | NP | NF1 | F1 |
| 1 | 146 | 0.82 | 1 | 0.9 | 1 | 0.26 | 0.42 | 0.57 |
| 2 | 267 | 0.83 | 1 | 0.9 | 1 | 0.32 | 0.48 | 0.63 |
| 3 | 510 | 0.79 | 1 | 0.88 | 1 | 0.08 | 0.15 | 0.25 |
| 4 | 174 | 0.55 | 1 | 0.71 | 1 | 0.73 | 0.85 | 0.77 |
| 5 | 174 | 0.79 | 1 | 0.88 | 1 | 0.15 | 0.26 | 0.4 |
| 6 | 99 | 0.96 | 0.91 | 0.94 | 0.43 | 0.67 | 0.52 | 0.67 |
| 7 | 331 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 |
| 8 | 42 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 9 | 196 | 0.94 | 0.98 | 0.96 | 0.75 | 0.45 | 0.56 | 0.71 |
| 10 | 140 | 0.95 | 0.97 | 0.96 | 0.8 | 0.73 | 0.76 | 0.85 |
| 11 | 100 | 0.95 | 0.96 | 0.95 | 0.88 | 0.84 | 0.86 | 0.9 |
| 12 | 427 | 0.94 | 0.98 | 0.96 | 0.78 | 0.54 | 0.64 | 0.77 |
| 13 | 552 | 0.99 | 0.99 | 0.99 | 0.96 | 0.96 | 0.96 | 0.97 |
| 14 | 171 | 0.77 | 0.98 | 0.86 | 0.8 | 0.18 | 0.29 | 0.44 |
| 15 | 474 | 0.93 | 0.99 | 0.96 | 0.54 | 0.17 | 0.26 | 0.41 |
| 系統重新規劃、開發 | | | | | | | | |
| 16 | **786** | **0.77** | **1** | **0.87** | **1** | **0.19** | **0.33** | **0.47** |
| 17 | **1375** | **0.83** | **0.99** | **0.91** | **0.91** | **0.22** | **0.35** | **0.5** |
| 18 | **121** | **0.8** | **0.99** | **0.88** | **0.96** | **0.52** | **0.68** | **0.77** |
| 19 | **278** | **0.71** | **1** | **0.83** | **1** | **0.46** | **0.63** | **0.72** |
| 合計 | 6363 | 0.85 | 0.99 | 0.91 | 0.94 | 0.46 | 0.62 | 0.74 |

**Total**：郵件的樣本數。**SR**：垃圾郵件召回率。**SP**：垃圾郵件精確率。**SF1**：垃圾郵件召回率與精確率的調和平均值。**NR**：正常郵件召回率。**NP**：正常郵件精確率。**NF1**：正常郵件召回率與精確率的調和平均值。**F1**：SF1 與 NF1 的調和平均值。

從表 4.1 中可以觀察到系統在多次的測試中，垃圾郵件的召回率 SR 及精確率 SP 大多都可以維持在八成以上的水準，由於精確率 SP 反映出正常郵件被誤判的情形[11]，實驗結果說明了系統將郵件誤判的情形並不多，但是因為誤判對使用者而言嚴重性遠大於未擊中，因此仍然必須儘可能將 SP 提昇到 100%。

回過頭再看正常郵件的精確率 NP，在十五回的測試中就有八回的 NP 不到六成，這表示被系統歸類為正常郵件的郵件當中，經常夾帶了大量的垃圾郵件。

再將 SF1、NF1 以及 F1 製成關係圖(圖 4.7)後，可以發現實驗的結果其實並不穩定，特別是 NF1 的結果，任一點相臨的兩次測試幾乎都會發生大幅度的波動。

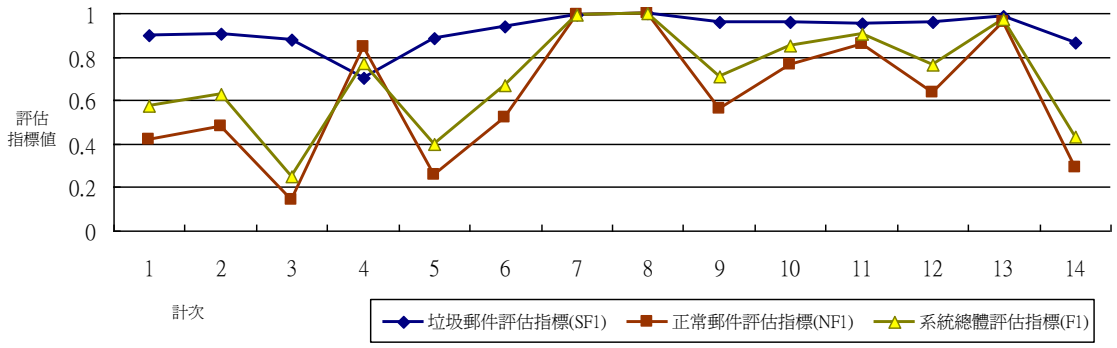


圖 4.7 第一階段 SF1、NF1 與 F1 關係圖

[11] 正常郵件被誤判的比例愈高，SP 就會愈低；相反地，SP 愈高則表示系統誤判的情形愈少。

線段波動情形在 SF1 就較為平緩，推敲可能的原因是因為系統僅過濾出特定類型的垃圾郵件，而其它類型的垃圾郵件則被乎略，因此研究假定目前的過濾系統對於某部份的垃圾郵件較為靈敏而且準確，並以此為基礎重新開發第二代系統。二代系統經過四次測試，實驗結果逐漸平穩。

到了第二階段由於雅虎信箱無法再使用的緣故[12]，需要重新調整郵件樣本的收集步驟和實驗方法，結果列於表 4.2。

表 4.2 第二階段分析結果

| 二代系統配合第二階段郵件樣本 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 計次 | Total | SR | SP | SF1 | NR | NP | NF1 | F1 |
| 1 | **4211** | **0** | **0** | **0** | **1** | **0.06** | **0.11** | **0** |
| 2 | **4211** | **0.84** | **0.99** | **0.91** | **0.9** | **0.26** | **0.4** | **0.56** |
| 3 | **4211** | **0.9** | **0.99** | **0.94** | **0.86** | **0.34** | **0.49** | **0.64** |
| 4 | **4211** | **0.93** | **1** | **0.96** | **0.96** | **0.45** | **0.61** | **0.75** |
| 5 | 2850 | 0.93 | 1 | 0.96 | 0.93 | 0.49 | 0.64 | 0.77 |
| 6 | 82 | 0.9 | 0.98 | 0.94 | 0.95 | 0.77 | 0.85 | 0.89 |
| 7 | 108 | 0.86 | 0.97 | 0.92 | 0.9 | 0.6 | 0.72 | 0.81 |
| 8 | 163 | 0.95 | 1 | 0.97 | 1 | 0.79 | 0.89 | 0.93 |

表 4.2 第一至四回測試使用相同的郵件樣本，並逐一加入過濾器及新的關鍵字以重新訓練系統。四次測試分別為：(1)關閉所有過濾器；(2)加入 Double Bytes 過濾器；(3)加入貝氏關鍵字過濾器；(4)加入新的關鍵字。

[12] 請參考第三章第三節實驗步驟三之說明。

同樣將結果製成關係圖(圖 4.8)，從圖中可以看出經過調整，新系統確實比舊系統有較穩定的表現，而且在幾次的分析後，系統 F1 就能夠達到並維持在 70%以上的水準。

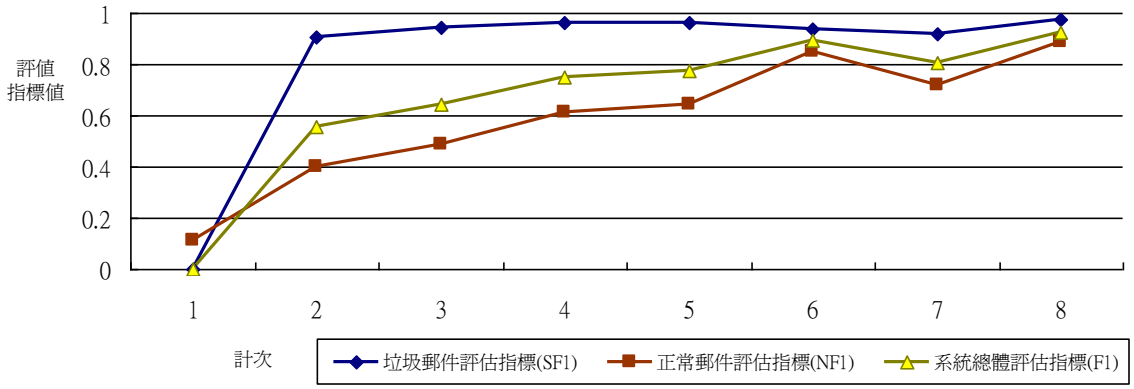


圖 4.8 第二階段 SF1、NF1 與 F1 關係圖

# 第五章 研究結論與建議

## 第一節 研究結論

### 1. 貝氏分類器關鍵字的選取

系統經過多次的訓練之後，對於垃圾郵件的過濾的確有相當顯著的效果，不過在此次的研究過程中也發現了一些問題，例如不適合的關鍵字造成判斷不正確。

不適合的關鍵字在研究中可以分為兩個類別。一是關鍵字達成的結果不符合使用者的期待，會造成這個現象主要是因為使用者在選取關鍵字的時候，僅注意到當下選取的那一封郵件是否屬於垃圾郵件(或正常郵件)，而乎略這個關鍵字可能在正常郵件(或垃圾郵件)裡頭才是多數，因此即使使用者訓練次數再多，都還是可能讓系統發生 Miss(或 Mistrial)。

另一種關鍵字不適合的情形對系統並不會造成過濾結果的失誤，但是卻會減低系統執行時的效率，這個情形是發生在符合該關鍵字的正常郵件與垃圾郵件數量幾乎相等的時候，雖然關鍵字本身並不會影響最後過濾的結果，但是系統在計算過程還是必須將其加入參考，造成中央處理器資源的浪費。

因此適時的整理關鍵字列表對於系統是有幫助的，雖然在這次的研究中沒有較正式的統計，不過在實驗操作的時候發現，將關鍵字數量控制在三百至五百之間較能同時兼顧過濾的效能與效率。

**2. 合法郵件的例外狀況**

實驗當中一直無法正確的判斷某一些正常郵件，分析結果發現部份郵件雖然來源合法，但是其標頭並不符合正確的格式，這類郵件以來自雅虎奇摩為最多，不過也因為雅虎的郵件中有不少是交友、拍賣及購物的電子報，因此系統並未打算特別針對雅虎郵件的例外狀況進行處理。

除了上述情形外，還有一種合法郵件誤判的情況發生在國外往來的郵件上，檢查郵件標頭後發現郵件的日期字串已經損毀無法辨示，由於其它欄位編碼都正常，所以目前無法確卻了解發生損毀的原因。

如果排除上述兩種狀況，合法郵件例外的情形並不多，而且郵件例外狀況的處置並不是這次研究的重點，因此這類型的郵件僅以人工方式修正。

### 3. 個人化的過濾器

因為系統參考使用者所建立的關鍵字列表，並且學習使用者個人對於郵件主觀的認知，使得系統分析結果能夠接近使用者的期待，也因為如此，研究結果也驗證使用自然貝氏分類器確實能夠達到在客戶端建立個人化郵件過濾系統的目標。

## 第二節 未來發展方向

雖然在研究過程當中，我們努力讓實驗盡可能更周延，但畢竟技術所能探討的問題是無止盡的，甚難一次考慮到所有可能的變數。評估此次研究成果，可發現仍然有許多問題存在於系統中等待一一克服，其中特別受到期待和關注的有以下三點。

### 1. 靈敏度的分析

靈敏度決定郵件被歸類於正常郵件或是垃圾郵件，對於系統過濾的正確性有一定的影響，但是因為時間及人力因素，這次研究中無法深入探討。

一個良好的靈敏度可幫助系統提昇召回率。因此日後在系統改進時，可再針對這個主題進行更詳細的統計分析，評估垃圾郵件機率的分佈情形，並依分佈情形決定出最佳的數值，以提供使用者作為設定系統參數時的參考。

### 2. 加入其它分析條件

貝氏分類器猶如一個不斷累積經驗的決策者，提供給它的訊息愈多、愈正確，它也愈能做出準確的判斷，因此我們可以將貝氏分類器當作包容其它過濾技術的容器，將不同技術分析後所得到的數據交由貝氏分類器進行最後的決策，這次 Double Bytes 與關鍵字的合作就是很好的例子。

### 3. 增強電子郵件通訊元件功能

此次研究中使用的 Extend Mail 元件，因為在時間上為了配合實驗進行，因此功能僅侷限於提供郵件標頭樣本及欄位的分析，真正要達到讓開發人員足以撰寫郵件軟體的目標仍然有很大的距離。

在日後的研究中，或許能夠計劃投入更多的時間以補齊功能不足之處，甚至能針對郵件存取的安全性予以加強，例如導入個人數位憑證、Sender ID 等身份試別技術。

# 參考文獻


1. 蔡瓊輝，2004，使用倒傳遞類神經網路學習垃圾郵件行為的類型，樹德科技大學資訊管理研究所碩士論文。
2. 劉鼎康，2005，使用類神經網路進行垃圾郵件過濾之研究，中原大學資訊管理學系碩士論文。
3. 行政院，2005，立法院第六屆第一會期第二次會議議案關係文書院臺經字第 0940081634 號。
4. 財團法人資訊工業策進會，2006，2005 年我國家庭寬頻、行動與無線應用現況與需求調查。
5. 尹相志，2006，Microsoft SQL Server 2005 資料採礦聖經。
6. Yahoo!奇摩，2004，網域認證鑰匙 DomainKeys http://tw.promo.yahoo.com/antispam/domainkeys.html
7. Microsoft，2004，寄件者使用名稱 Sender ID，http://www.microsoft.com/mscorp/safety/technologies/senderid/default.mspx
8. G. Klyne. 2005. Reqistration of Mail and MIME Header Fields (Request for Comments: 4021).
9. P. Resnick. 2001. Internet Message Format (Request for Comments: 2822).
10. Carnegie Mellon M.. 1996. Post Office Protocol – Version 3 (Request for Comments: 1939).

11. David H. Crocker. 1982. STANDARD FOR THE FORMAT OF ARPA INTERNET TEXT MESSAGES (Request for Comments: 822).
12. N. Freed, N. Borenstein. 1996. Multipurpose Internet Mail Extensions(MIME) Part One: Format of Internet Message Bodies (Requst for Comments: 2045).
13. J. Lyon Microsoft Corp., M. Wong. 2006. Sender ID: Authenticating E-Mail (Request for Comments: 4406).
14. J. Lyon Microsoft Corp., M. Wong. 2006. Sender ID: Authenticating E-Mail (Request for Comments: 4406).
15. G. Klyne, J. Palme. 2005. Registration of Mail and MIME Header Fields (Request for Comments: 4021).
16. M. Crispin. 2003. INTERNET MESSAGE ACCESS PROTOCOL – VERSION 4revl (Request for Comments: 3501).
17. J. Postel. 1975. On the junk mail problem (Request for Comments: 706).