主動式管理垃圾郵件與病毒郵件擴散架構

莊斯凱
元智大學電機研究所／桃園縣政府教育局網路中心
eric@tyc.edu.tw

摘要

在眾多的網路應用中，電子郵件扮演著聯絡溝通的重要角色，在商業上也是行銷方式中的絕佳方案，不僅可以低成本的方式進行網路行銷，更能以無國界的特性進行資訊傳播，然而病毒也利用此特性感染通訊錄清單裡的聯絡人，進而產生大量的病毒郵件與進行商業行銷的垃圾郵件。

在廣義的垃圾郵件可以定義為：先前不存在商業關係的一方，在未經同意的狀況下所發送的商業電子郵件，且收件者不想閱讀的信件。在此定義下的郵件均不受歡迎。本文將提出主動式管理控制架構，在網路提供商或是網路管理中心的角色上，優先管控電子郵件的發送者，避免成為網路上垃圾郵件與病毒郵件的製造者。

關鍵詞：Spam、Virus、Mail、Initiative。

Abstract

Among many Internet applications, e-mail plays an especially important part in communication. It is also an effective solution in promotion and marketing. E-mail is not only an inexpensive way of Internet marketing, also a borderless tool to circulate information. However, it is also this attribute that computer viruses can take advantage of to influence all on the users' contact lists, and produce huge volumes of virus, spam and junk mail.

The generalized definition of junk mail could be this: unwanted, or uncalled-for commercial e-mails sent by a non-business-associate party, which the receiver often does not wish to read. The e-mails under the definition are unwelcome. This essay is going to present a control system that manages the e-mail senders for the ISP or network management centers, and prevent them from being the producers of junk and virus mail.

Keywords: Spam, Virus, Mail, Initiative.

1. 前言

隨著頻寬網路的快速普及，電子郵件系統已成為網路使用者日常運作必不可少的一個部分。根據多份的研究報告[13][14][15]指出，在所有的 Internet 服務中，電子郵件服務是最基本的一項，也是使用率僅次於網頁服務的一項功能。透過電子郵件進行交流具有方便快捷的特點。可以說，幾乎沒有一個 Internet 的使用者不在收發電子郵件，甚至擁有兩個以上的電子郵件地址。

然而電子郵件就像兩面刃。一方面，電子郵件對企業的資訊化建設、生產效率和服務品質的提高帶來了極大的促進作用，個人使用電子郵件方便、快捷、高效、經濟；但是另一方面，由於網路的開放性和電腦資訊易於擴散的特性，它在處理、存儲、傳輸和使用上有著嚴重的脆弱性，即很容易被干擾、濫用，大量的垃圾信、電子炸彈、郵件病毒使用戶受到極大的困擾。隨著企業業務的擴展，收發電子郵件的數量逐漸增加，受到郵件病毒、垃圾信攻擊的可能性也在逐漸增大。[13][14]

更進一步分析發送來源的 IP，大多分佈在學術研究網路上，而在學術研究網路上，基礎建設較為健全，頻寬也比一般家用網路來得高，用戶的網路安全保護也較為鬆散，無法與企業網路安全政策比擬，十分容易遭受駭客或不當資訊的網頁植入後門程式，在後門程式上種類五花八門，電子郵件領域上常見的手法就是植入垃圾郵件寄送（轉送 Relay）程式。

有鑑於此，研究者工作於教育局網路中心，一般縣市網路中心均可以比擬為小型的網際網路提供商（ISP），研究者為避免學術研究網路成為大量垃圾郵件或病毒郵件的發送者，著手研究防治方案，於本文提出「主動式管理垃圾郵件與病毒郵件的防治擴散架構」，以期能避免學術研究網路成為網路上的問題製造者。

本文其餘內容如下，第二章介紹相關研究，以及過濾技術；第三章提出主動式管理的架構；第四章實際的驗證與量測運作後的成效；第五章總結研究成果。

2. 介紹相關研究資料與過濾技術

2.1 垃圾郵件之定義

近年來最令人困擾的兩大類電子郵件，分別是病毒與垃圾信件。病毒信件具備非常明確之定義：凡是含有會傷害電腦、盜取資料之有害程式，包括後門與木馬等，我們都可以將之歸納為病毒信件。檢測方式也非常容易，任何一套防毒軟體，都可以輕鬆勝任此項工作，簡單來說，判斷是否為病毒信件，可以說是定義明確之零和問題。[14]

相較於病毒信件，垃圾信件常常處於模糊的灰色地帶，不易認定。以條列式整理可以歸納如下列

四點[12]：

- 收件人事先沒有提出要求，或者同意接收的廣告、電子刊物、各種形式的宣傳品等宣傳性的電子郵件。
- 收件人無法拒收的電子郵件。
- 隱藏發件人身份、位址、標題等資訊的電子郵件。
- 含有虛假的資訊源、發件人、路由等資訊的電子郵件。

2.2 發送垃圾郵件之行為定義

目前國際上不僅僅對直接發送垃圾郵件的個人或者組織採用懲罰措施，並且還對發送垃圾郵件提供支援的個人或組織採用懲罰措施。在網際網路上的交換中心與管理中心所定義垃圾郵件相關行為如下[12]：

A.直接通過主機發送垃圾郵件。

B.對發送垃圾郵件提供支援的行為包括：

- 為垃圾郵件組織的網站提供平臺。
- 使用 DNS 伺服器替該網站做 DNS 解析。
- 為垃圾郵件的發送提供 openrelay 郵件服務器或者代理伺服器等。

2.3 垃圾郵件的行為特點[16]

- 發件人位址隨機變化。
- 郵件主題隨機變化。
- 偽造郵件頭干擾資訊。
- 信體內容隨機變化內容。
- 正文以圖片方式顯示，難以識別。
- 對垃圾郵件的定義和分類因人而異。
- 垃圾郵件在不同時段內的傳播內容不一。
- 垃圾郵件在不同範圍內的傳播內容不一。

2.4 過濾分類技術

2.4.1 kNN 方法

kNN 是最常用的基於 Memory 的方法。kNN 沒有訓練過程，分類時直接將待分類文本與訓練集合中的每個文本進行比較，然後根據前 k 篇相似的文本得到新文本的類別（最簡單的情況可以根據 k 篇文本所屬的類別數的多少來確定最後類別）。kNN 的原理非常直觀，也很容易理解。在文本分類中，kNN 也常常能夠取得好的結果。[3]

2.4.2 Rocchio 方法

Rocchio方法由是資訊檢索領域常常用於相關回饋的方法。它用於分類的基本思路很簡單：將所有訓練文本向量化，類別向量等於所有正例向量和反例向量的加權差。形式地：

$$\vec{C} = \frac{\alpha}{|D^+|}\sum_{\vec{x}_i \in D^+} \vec{x}_i - \frac{\beta}{|D^-|}\sum_{\vec{x}_i \in D^-} \vec{x}_i \text{。}$$

其中 D^+、D^-分別表示正例和反例集合，$|D^+|$、$|D^-|$分別表示正例集合和反例集合的大小。α、β為加權係數。計算得到的結果$\vec{C}$ 表示該類的類別向量。用於垃圾郵件過濾時，通過上式可以得到垃圾郵件類的類別向量。新的郵件與類別向量計算距離，距離小於某個閾值θ則判定該郵件屬於垃圾郵件類，否則為合法郵件。實際應用中，α可以設定為 1，β和θ可以通過訓練得到(使得訓練集合的分類錯誤率最低)。[6]

2.4.3 Naïve Bayes 方法

Bayes 方法是通過計算文本 d 屬於每個類別 C_i（$i=1,2,\ldots,M$，M 為類別個數）的或然率 $P(C_i|d)$並將它們排序取其最大值來得到 d 所屬的類別。根據 Bayes 公式，最後歸結於求每個類別的或然率 $P(C_i)$和從類別 C_i 生成文本 d 的或然率 $P(d|C_i)$。這兩個或然率都可以通過訓練語料得到。Naïve Bayes 是 Bayes 方法中使用最廣泛的一種。在這種方法中，假設 d 由互相獨立的多個特徵 w_j(j=1,2,…,N，N 是 d 中不同特徵數)生成，於是 $P(d|C_i)$由可以歸結成求 $P(w_j|C_i)$。Naïve Bayes 方法被廣泛用於文本分類中，取得了不錯的效果。[1][4]

2.4.4 Winnow 方法

Winnow 是一種線性分類器。它訓練的目的就是為了找到某個類所有特徵的權重向量 $w=<w_1,w_2,\ldots,w_N>$(N 是特徵數)和閾值θ，對於新文本 $x=<x_1,x_2,\ldots,x_N>$，如果 $w^T x>\theta$，則判定屬於該類。否則，不屬於該類。Winnow 在學習 w 時採用的是一種錯誤驅動的方法。在訓練時，一旦發生錯誤，將根據需要降低或者升高 w 裡相應特徵的權重值。[2]

2.4.5 決策樹方法(Decision Tree)

決策樹是著名的規則方法之一。通過按照某種屬性的順序自頂向下地生成一棵樹，樹的每個節點是屬性名，而每條邊是屬性值。從樹根到樹葉的一條路徑便對應一條規則。基於資訊增益進行屬性順序選擇是決策樹中常用的方法之一。著名的決策樹演算法有 ID3、C4.5 等。Carreras[9]使用決策樹來過濾垃圾郵件，得到的垃圾郵件過濾的正確率和召回率都在 88%左右。[8][9][10]

目前市場上提供垃圾郵件與防毒過濾掃瞄的產品眾多，經過與廠商洽談與借測後，發現能夠提

供給分析的資訊眾多，也能以充足的分析資訊將發送垃圾郵件或病毒郵件的單位給予警告，於是研究者提出了新的使用管理架構，希望能夠阻斷病毒郵件的對外發送，同時將發送垃圾郵件的使用者列入管理黑名單，也將其使用的 IP 與學術網路連接線路隔離。

規劃建置此架構時，也發現這類產品經過市場競爭與今年度中信局提供此採購項次，已比往年較為低價。依照病毒的變化度，必須倚靠商業的病毒碼或是過濾引擎，才能有效的防治，所以我們以商業產品並配合負載平衡設備，設計出新架構。新架構由下一章節接續說明。

3. 主動式管理防治架構

一般過濾郵件架構計設著重於寄入伺服器時的防護，所以幾乎是以設定 DNS 的 MX 紀錄，將郵件先導入過濾設備，再由過濾設備或是伺服器利用 SMTP-route 的記錄，把過濾後的郵件送至郵件伺服器，如圖 1。但是卻無法防止使用者發送出垃圾郵件或病毒郵件，本文提出新概念，使用者不用更改任何設定即可引入過濾機制，防止寄發擾人的垃圾郵件與病毒郵件。

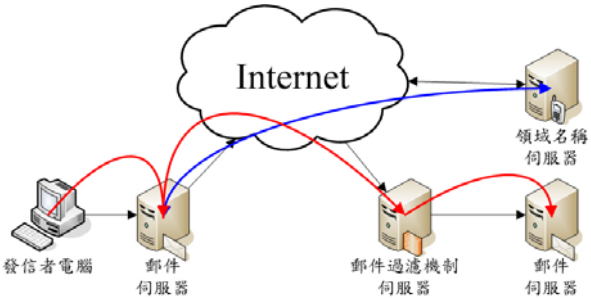

圖 1　一般郵件過濾架構圖

3.1　系統架構

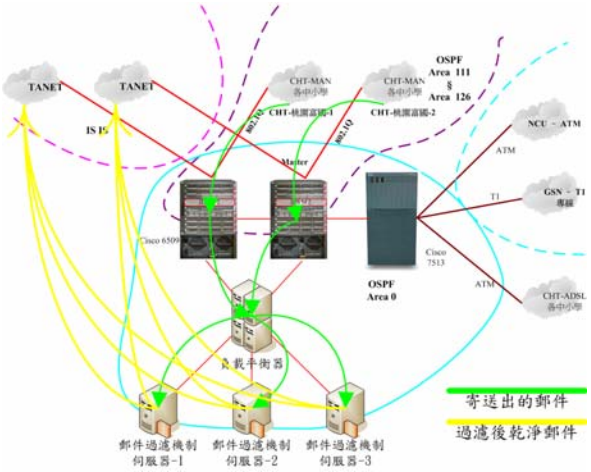

圖 2　系統架構圖

主動式管理架構，如圖 2；由使用者或電子郵件伺服器寄送出信件時，均會通過教育局網路中心第三層骨幹交換器，使用交換器的功能，將郵件轉向送至負載平衡器設備，再有負載平衡器設備分配至過濾機制，經由過濾機制分析後，寄出正常的郵件，可以有效地避免下游使用單位成為問題製造者，或是電子郵件廣告商的跳板。

3.2　第三層骨幹交換器路由政策

郵件送出的路徑特性，為寄送來源 IP 為下游使用單位，目的地的服務埠為 SMTP Service（TCP 25），所以必須在骨幹交換器上設定將此路徑導向負載平衡器，如表 1、表 2、表 3。

表1　ACL 控制表

```
access-list 113 permit tcp 163.30.0.0 0.0.255.255
any eq smtp
access-list 113 deny      tcp any any
```

表 1 設定規則，來源使用單位 IP，目的地埠為 TCP 25，其餘均予以 deny。

表 2　IP Based 的導向表

```
Route-map SPAM-Rule permit 10
  match ip address 113
  set ip next-hop 163.30.255.10
!
```

表 2 使用 Route-map 的功能，將符合規則 113 的流量導向 163.30.255.10（負載平衡器的 IP）。

表 3　啟動 IP Policy 的功能

```
interface Vlan111
  description N-RING-1
  ip address 10.12.16.225 255.255.255.0
  ip access-group School_ACL in
  ip access-group School_ACL out
  ip accounting output-packets
  ip accounting mac-address input
  ip route-cache flow
  ip policy route-map SPAM-Rule
!
```

表 3 在對下游使用單位的連接介面上，啟動 IP Policy 的功能，實際攔截寄送郵件流量至負載平衡器。再由負載平衡器修改目的地 IP 後，送至過濾系統，此時過濾系統僅需設定允許使用單位的 IP 可以 Relay Mail。

3.3　Persistent Connections

此架構必須承受大量的郵件流量，所以採分散

式處理的模式，將郵件分送至不同的過濾系統，以執行過濾機制，但因為 SMTP 為交談式協定，如圖 3 表示；所以每一個 Session 必須能保持在同一台過濾設備，否則將造成連線中斷的狀態，可以稱為 Persistent Connections，如圖 5。

```
Trying 163.30.4.4...
Connected to mqueue.tyc.edu.tw.
Escape character is '^]'.
220 relay1.tyc.edu.tw ESMTP
HELO tyc.edu.tw
250 relay1.tyc.edu.tw
mail from:eric@tyc.edu.tw
250 sender <eric@tyc.edu.tw> ok
rcpt to:eric2q@yahoo.com.tw
250 recipient <eric2q@yahoo.com.tw> ok
data
354 go ahead
subject:TEST CLI to KIMO
$:$e
TEST TEST  ETS
TEST TEST kkk
.
250 ok:  Message 4468854 accepted
quit
221 relay1.tyc.edu.tw
Connection closed by foreign host.
```

圖 3　SMTP 的交談過程

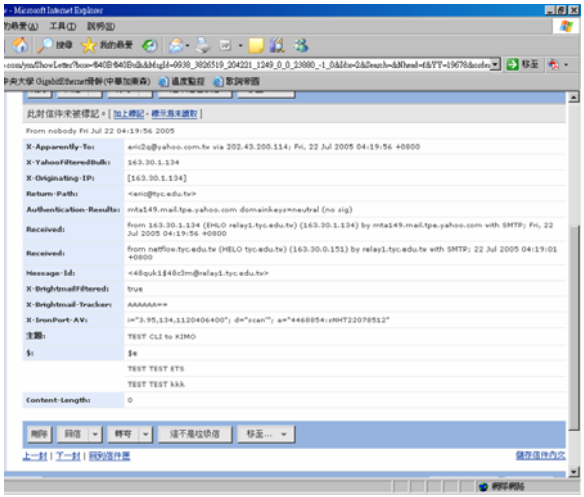

圖 4　信件的標頭檔

圖 5　Persistent Connections

4. 郵件發送結果分析

4.1　整併入現有的網管系統

可以即時顯示目前垃圾信件與病毒郵件的發送狀態，如圖 6。

圖 6　即時狀態

4.2　單一 IP 發送垃圾郵件

以單一 IP 主針對垃圾郵件進行監控分析，如圖 7。

4206773	163.30.94.2	有卡就可變現金30-60萬給你用喔 ！	垃圾信	2005-07-19 14:46:37
4206774	163.30.94.2	兩岸升學就業移民最佳情報	垃圾信	2005-07-19 14:46:37
4206778	163.30.94.2	如何面對憂鬱，一起來探討?	垃圾信	2005-07-19 14:46:38
4206780	163.30.94.2	◢◢◢貸款直接找銀行,您會更有保障(限新竹以北地區)◣◣◣	垃圾信	2005-07-19 14:46:38
4206781	163.30.94.2	您好!! 合作提案	垃圾信	2005-07-19 14:46:39
4206784	163.30.94.2	現代孫子兵法EDM	垃圾信	2005-07-19 14:46:40
4206789	163.30.94.2	專業理財顧問為你量身定做省息專案!	垃圾信	2005-07-19 14:46:40
4206792	163.30.94.2	限量班尼頓手錶送給你	垃圾信	2005-07-19 14:46:41
4206796	163.30.94.2	周年慶！狂賣中！三件式800喇叭，只要399元，再送　姘甏I	垃圾信	2005-07-19 14:46:41
4206801	163.30.94.2	憑勞保卡可貸10~300萬!	垃圾信	2005-07-19 14:46:42
4206802	163.30.94.2	不影響正職工作 增加收入	垃圾信	2005-07-19 14:46:42
4206804	163.30.94.2	信貸1.88%、每月10萬月繳1280	垃圾信	2005-07-19 14:46:43

圖 7　單一 IP 發送垃圾郵件狀態

4.3　單一 IP 發送病毒郵件

以單一 IP 主針對病毒郵件進行監控分析，如圖 8。

2005-07-20　選擇開始日期　2005-07-20　選擇結束日期　送出查詢　上一頁 下一頁

MID	IP address	主旨	事件	日期時間
4276190	163.30.129.1	You~ve got 1 VoiceMessage!(W32/Zafi-B)	病毒信	2005-07-20 08:53:42
4276193	163.30.129.1	You~ve got 1 VoiceMessage!(W32/Zafi-B)	病毒信	2005-07-20 08:53:42
4276195	163.30.129.1	Check this out kid!!!(W32/Zafi-B)	病毒信	2005-07-20 08:53:42
4276198	163.30.129.1	xb9qxb8xa3xb1xd0xabxc7(W32/Zafi-B)	病毒信	2005-07-20 08:53:42
4276200	163.30.129.1	Don~t worry, be happy!(W32/Zafi-B)	病毒信	2005-07-20 08:53:42
4276203	163.30.129.1	You~ve got 1 VoiceMessage!(W32/Zafi-B)	病毒信	2005-07-20 08:53:42
4276208	163.30.129.1	You~ve got 1 VoiceMessage!(W32/Zafi-B)	病毒信	2005-07-20 08:53:43
4276211	163.30.129.1	Don~t worry, be happy!(W32/Zafi-B)	病毒信	2005-07-20 08:53:43
4276217	163.30.129.1	xb9qxb8xa3xb1xd0xabxc7(W32/Zafi-B)	病毒信	2005-07-20 08:53:45
4276219	163.30.129.1	Check this out kid!!!(W32/Zafi-B)	病毒信	2005-07-20 08:53:45
4276221	163.30.129.1	xb9qxb8xa3xb1xd0xabxc7(W32/Zafi-B)	病毒信	2005-07-20 08:53:46
4276223	163.30.129.1	Check this out kid!!!(W32/Zafi-B)	病毒信	2005-07-20 08:53:46
4276227	163.30.129.1	Check this out kid!!!(W32/Zafi-B)	病毒信	2005-07-20 08:53:46
4276228	163.30.129.1	Don~t worry, be happy!(W32/Zafi-B)	病毒信	2005-07-20 08:53:46

圖 8　單一 IP 發送病毒郵件狀態

4.4　Message-ID 標記

加以記錄每封信件的 Message-ID 標記，如圖 9。

圖 9　紀錄 Message-ID 標記

4.5　郵件發送統計圖與分佈圖

計算單一 IP 發送次數，並加以排名，如圖 10。

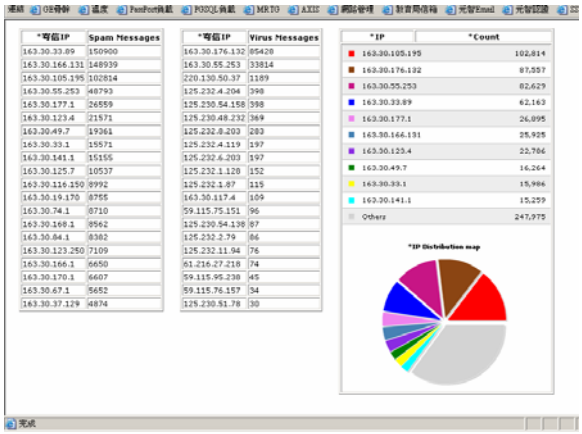

圖 10　郵件發送統計

4.6　統計紀錄

將一週內發送狀況加以記錄統計，如表 4。

表 4　統計紀錄

日期	垃圾郵件數量(中文標題)	垃圾郵件數量(含重寄)	病毒郵件數量
2005/7/19	333	28476	95
2005/7/20	246	41586	26004
2005/7/21	219	9537	67
2005/7/22	244	12056	100
2005/7/23	398	12646	113
2005/7/24	317	10775	101
2005/7/25	307	31735	100
2005/7/26	35	1455	7

4.7　學校發送排名

計算單一學校發送次數，並加以排名，如圖 11。

圖 11　學校發送排名

4.8　通知使用者協助排除障礙

通知寄發垃圾與病毒的下游使用單位，若是 Mail Server 利用 message_id 尋找發信者，若是一般單機使用者，可利用 netstat -n-o 等工具尋找發信程式，或利用軟體來移除發信程式，以避免學術網路成為垃圾郵件與病毒郵件的製造者。

5. 結論

本研究經長時間的驗證，確認可以主動優先的管理郵件，以保持對外寄送郵件時的清潔度，以能確保不發送出病毒郵件，而在垃圾郵件是可以達到有效的先期防止，在發送的初期就能夠通知單位的管理者，由現場單位的管理者有效的控制郵件發送的狀態。希望本研究可以給管理者不同的看法與概念，使垃圾郵件的防治上能更向前一步，達到確實防堵廣告信件的目的。

參考文獻

[1] H. Katirai, "Filtering Junk E-Mail: A Performance Comparison between Genetic Programming & Naive Bayes", available online at: http://members.rogers.com/hoomank/katirai99filtering.pdf, Sep. 1999[D]

[2] H. Drucker, D. Wu, and V. N. Vapnik, "Support Vector Machines for Spam Categorization", IEEE Transactions on Neural Networks, Vol. 20, No. 5, pp. 1048-1054, Sep. 1999[J]

[3] I. Androutsopoulos, G. Paliouras, V. Karkaletsis, G. Sakkis, C.D. Spyropoulos and P. Stamatopoulos, "Learning to Filter Spam E-Mail: A Comparison of a Naive Bayesian and a Memory-Based Approach", in Proc. 4th European Conference on Principles and Practice of Knowledge Discovery in

Databases (PKDD 2000), pp. 1-13, Sep. 2000[C]

[4] I. Androutsopoulos, J. Koutsias, K.V. Chandrinos, G. Paliouras and C.D. Spyropoulos, "An Evaluation of Naive Bayesian Anti-Spam Filtering," in Proc. of the Workshop on Machine Learning in the New Information Age, 11th European Conference on Machine Learning (ECML 2000), pp. 9-17, May 2000[C]

[5] J. M. G. Hidalgo, "Evaluating Cost-Sensitive Unsolicited Bulk Email Categorization", in Proceedings of ACM Symposium on Applied Computing (SAC 2002), pp. 615-620, Mar. 2002[C]

[6] J. Rocchio, "Relevance feedback in information retrieval", in the SMART Retrival System: Experments in Automatic Document Processing, pp. 313-323, PrenticeHall Inc., 1971[M]

[7] K. Schneider, "A Comparison of Event Models for Naive Bayes Anti-Spam E-Mail Filtering", in Proc. 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003), Budapest, Hungary, pp. 307-314, Apr. 2003[C]

[8] M. DeSouza, J. Fitzgerald, C. Kemp and G. Truong, "A Decision Tree based Spam Filtering Agent", from http://www.cs.mu.oz.au/481/2001_projects/gntr/index.html, 2001[EB]

[9] X. Carreras and L. Marquez, "Boosting Trees for Anti-Spam Email Filtering", in Proceedings of Euro Conference Recent Advances in NLP (RANLP-2001), pp. 58-64, Sep. 2001[C]

[10] T. Nicholas, "Using AdaBoost and Decision Stumps to Identify Spam E-mail", Stanford University Course Project (Spring 2002/2003) Report, from http: //nlp.stanford.edu/ courses/cs224n/ 2003/fp/[EB]

[11] Y. Diao, H. Lu and D. Wu, "A Comparative Study of Classification Based Personal E-mail Filtering", in Proceedings of PAKDD-2000, pp. 408-419, Apr. 2000[C]

[12] 中國互聯網協會。2003。反垃圾郵件規範。

[13] 王雅慈，曾黎明，游象甫，陳奕明。2005。廣告電子郵件的分流過濾及回覆訊息之萃取。TANET 2005 台灣網際網路研討會

[14] 張傑生，唐瑤瑤，許凱平，陳啟煌，李秀惠，賴飛羆。2005。使用 Open Source 軟體進行 SPAM Mail 防制處理——以台灣大學電子郵件系統為例。TANET 2005 台灣網際網路研討會

[15] 陳志明，余興華，陳朝欽。2005。一個防廣告信件及內容過濾之整合系統架構。TANET 2005 台灣網際網路研討會

[16] 孫東紅，陳光英。2006。反垃圾郵件技術分析與中文垃圾郵件過濾規則研究。中國教育和科研電腦網緊急回應組。清華大學資訊網路工程研究中心（中國）。

[17] 科學人雜誌。2006。終結垃圾信。

[18] 古德曼（Joshua Goodman），赫克曼（David Heckerman），倫斯威特（Robert Rounthwaite）。2006。對於防堵垃圾郵件，我們能夠做些什麼？ http://www.sciam.com.tw/read/readshow.asp?FDocNo=672&CL=19