廣告電子郵件的分流過濾及回覆訊息之萃取

[ab]王雅慈[b]曾黎明[ab]游象甫[c]陳奕明[a]國立中央大學電子計算機中心 [b]資訊工程學系 [c]資訊管理學系
center11@cc.ncu.edu.tw tsenglm@cc.ncu.edu.tw
center3@cc.ncu.edu.tw cym@mgt.ncu.edu.tw

摘要

近年來由於廣告信的問題日益嚴重，各方都在努力找出一個防制的機制，但由於事涉個人隱私問題，系統管理者、使用者、單位主管看法均不相同，使得過濾機制難以面面俱到，同時廣告信的認定隨使用者而不同，讓廣告信的防制機制很難達到百分百正確，本論文提出一個分流機制的架構，可以兼顧使用者的意願、並維持郵件伺服器的穩定及擴充性。同時由觀察廣告信中發現，其提供的回覆訊息(即 URI)並不會造假，因為發送廣告者希望使用者能點選，本論文借由資料的蒐集、統計及分析，探討如何從郵件提供的回覆訊息中找尋廣告信。

關鍵詞：廣告信、分流機制、黑名單

1 緒論

電子郵件已經成為現代人溝通及分享心情的一種管道，透過主動訂閱電子報，每天可自動且即時的收集有興趣的資訊；同時網路行銷也經由電子郵件寄送最新消息、促銷商品等資訊給客戶，不僅可以提供平面媒體做不到的動畫及互動性，也可以透過資料分析給予個人化行銷，但是因為 Simple Mail Transmission Protocol (SMTP) [10] 的協定容易冒充，例如冒充好友的來信，同時不需要註冊反查位址即可寄信，加上成本低廉，這些原因使得不請自來的廣告信在網際網路四處充斥，導致郵件伺服器停擺，使用者抱怨，針對此問題，本論文探討各式過濾架構的優缺點，提出一個分流機制希望可以符合使用者意願，具備人性化介面及自動化管理，並實際運用於中央大學郵件伺服器上 [1]。

從管理郵件的經驗中，我們發現廣告信為了逃避過濾系統，延伸出許多的多樣性，從過去的單一寄件者、統一內容，到目前利用不同寄件者、不同主旨、不同的內容，來逃避被分類成大量寄件的黑名單，而且從過去不合法的郵件伺服器，到目前在網域伺服器上可以找到完整正反查的主機，這些”進步”，讓過濾判斷機制面臨重大考驗，但是我們由觀察中發現，廣告信提供的回覆訊息並不會造假，因為發送廣告者希望使用者能點選。本文探討如何從郵件提供的回覆訊息中找尋廣告信，自動收集黑名單，並以網域伺服器技術存放，提供其他相關軟體取用及分享。

本文其餘內容如下，第二節介紹相關研究，探討廣告信氾濫的原因、各種過濾架構的優缺點及現有廣告信分類的技術；第三節提出信件分流機制架構，並藉由資料的蒐集，分析探討廣告信的發展及特性；第四節描述系統實作，量測運作後的成效；第五節總結研究成果及未來研究方向。

2 相關研究

本節分析及探討 SMTP 的冒充問題及防制廣告信的機制。

2.1 SMTP 的冒充問題

SMTP 是一種高度信任寄件者的協定，它並沒有設立嚴謹的認證機制，寄送電子郵件時無需驗證來者身份，就好像我們平常慣用的郵政系統，只要收件者的地址是對的，郵資是夠的，郵差都會把這封信送到，不管收件人是否住在這個地方，也不看寄件者是誰，我們以圖 1 的例子來說明冒充的實例，圖 2 是真正收到這封信的表頭，以 from: ytwang@msn.com 這行可以看到，我們假造自己的郵件地址為 ytwang@msn.com，這個訊息清楚的呈現出來，而由 Received: from *msn.com*(center11-3.dd.ncu.edu.tw

[140.115.11.137])也可以看出來造假的地方，雖然括弧內仍然呈現真正連線的機器，但這已足以說明 SMTP 的冒充問題，也是造成廣告信氾濫的原因。

C:\>telnet smtp.cc.ncu.edu.tw 25 ← 連結 smtp server
220 smtp2.cc.ncu.edu.tw ESMTP Sendmail 8.12.9/8.12.9/dove/0.0.4; Sun, 3 Jul 2005 01:50:01 +0800 (CST)
helo msn.com ← 宣稱自己來自 msn.com
250 smtp2.cc.ncu.edu.tw Hello center11-3.dd.ncu.edu.tw [140/115.11.137], pleased to meet you
mail from: ytwang@msn.com ← 宣稱自己的email 爲ytwang@msn.com
250 2.1.0 ytwang@msn.com… Sender ok
rcpt to: center11@cc.ncu.edu.tw ← 宣稱要送達的 email 爲 center11@cc.ncu.edu.tw
250 2.1.5 center11@cc.ncu.edu.tw … Recipient ok
data
354 Enter mail, end with "." On a line by itself
from : ytwang@msn.com
.
250 2.0.0 j62Ho180020974 Message accepted for delivery
quit

圖 1 SMTP 冒充問題(1)

Return-Path: <ytwang@msn.com>
Received: from smtp2.cc.ncu.edu.tw (smtp2 [140.115.17.128]) by aries.cc.ncu.edu.tw (8.13.0/8.13.0) with ESMTP id j62Hu1qE007373 for<center11@cc.ncu.edu.tw>; Sun, 3 Jul 2005 01:56:01 +0800 (CST)
Received: from *msn.com* (center11-3.dd.ncu.edu.tw [140.115.11.137]) by smtp2.cc.ncu.edu.tw (8.12.9/8.12.9/dove/0.0.4) with SMTP id j62Ho180020974 for center11@cc.ncu.edu.tw; Sun, 3 Jul 2005 01:54:58 +0800 (CST)
Date: Sun, 3 Jul 2005 01:50:01 +0800 (CST)
Message-Id: <200507021754.j62Ho180020974@smtp2.cc.ncu.edu.tw>
from: ytwang@msn.com
subject: this is a test mail

圖 2 SMTP 冒充問題(2)

2.2 郵件分類的架構

由於 SMTP 協定容易冒充送信者，再加上發送廣告信高利潤而低成本，導致大量的廣告信充斥於網際網路，為了解決這個問題，有下列不同的作法。

2.2.1 MUA 上的策略

可以在使用者端的電腦上安裝過濾軟體，例如 POPFile [17]、K9 [13]，這一類軟體大部份都是使用貝氏分類法 [9]，經由使用者不斷的訓練後通常可以有 95%以上的正確率，而且符合個人的喜好，其缺點在於需額外安裝設定且無法分享彼此的資訊，另一項缺點是必須在信件主旨加上辨識的字串，才能提供後端讀信軟體的分類，此外這樣的作法無法減少信件量，無法保護郵件伺服器。

2.2.2 網路流量預測及防制

以網路管理的角度來看，可以利用 SMTP 的流量分析及單位時間所引發的連線數，從網路傳輸特性、流量統計發現異常提早通知管理者處理 [2]；被埋入後門而形成開放性郵件主機的電腦也有文章提供自動檢測的機制 [4]，這些方法可以及早發現不當的流量，尤其是中毒、被埋入後門及惡意攻擊，但是無法對郵件分類，使用者仍然必須面對大量的垃圾郵件。

2.2.3 郵件伺服器端的過濾

在郵件伺服器上對郵件分類，可以幫助使用者避免面對大量的廣告信，同時可以降低伺服器負荷，保護郵件伺服器。郵件伺服器上防治廣告信的方法，可以分成下列幾類：

第一類是提供 webmail 的讀信軟體，因為信件集中在伺服器上，可由系統分類出不同的資料匣，但本策略須考慮系統負荷及提供大量磁碟空間（例如 yahoo.com，相關軟體如 Openwebmail [14]）。

第二類是單層簡易型的過濾架構，直接在郵件伺服器上安裝判斷過濾的軟體，並在主旨上加上{Spam?}，以便使用者後端的讀信軟體，可判別及分類，使用者不需要改變已有的習慣，可以快速的學會使用，而缺點是改變信件主旨，如果判斷錯誤時，容易造成回信或轉寄此封信時的困擾，對使用者而言，信件並沒有減少，只是可以比較快速的處理重要的信件。

第三類是串接型過濾架構，利用串接的觀念，額外安裝一台過濾器，同時更改網域伺服器 MX 的紀錄，先行過濾判斷後，再將信件導向真正的郵件伺服器，一方面不受限於帳號的管理，可以彈性的使用不同的作業平台，增加處理的選擇性，另一方面獨立於使用者的郵件空間，可以彈性調整增加過濾系統，缺點是增加硬體成本，

且廣告信仍然必須加上主旨才能提供後端的使用者辨識，使用者的信件也沒有減少。

2.3 郵件過濾的技術

目前已存在許多種方法，其成功率均相當高，不論是裝在郵件伺服器或個人電腦上，大致上可分為信件標頭及信件內容過濾兩種方式。

2.3.1 header test

利用 Received Header 中 HELO 的 IP 與 sender IP 不相同 [16]，或者是 MSGID 不符合 mail server 格式，例如全部都是大寫英文或包含本機端的伺服器名稱，如圖 3。

```
Received: from 140.115.17.208 ([61.95.185.146]) by
aries.cc.ncu.edu.tw (8.13.0/8.13.0) with SMTP id
j5O3Avoa0 Fri, 24 Jun 2005 11:11:06 +0800
宣稱從 140.115.17.208 來，但實際上為 61.95.185.146
Message-Id: <HAWIVFJJFYVSFCGZTSHZYB@yahoo.com>
Message-Id:
<200506240510.j505AY4o029981@aries.cc.ncu.edu.tw>
```

圖 3 信件表頭

2.3.2 Naïve Bayes 分類法

Naïve Bayes 分類法是基於機率理論的分類方法，在特徵選取後，由已知文件計算出該特徵與該類別之間的條件機率關係，分類時藉由此機率關係計算文件屬於各類的機率，由其中選出機率最高的類別作為該文件的類別。這是一個常見的判斷廣告郵件的方法 [6][7][8]，運用在許多知名的軟體上例如 Spamassassin [19]，POPFile [17], K9 [13]。

2.4 黑名單拒絕

利用黑名單直接拒絕來自不受歡迎的 IP 連線，例如不斷寄送 User Unknown 的 IP 及 open relay 的機器，利用 DNS-based Blackhole List (DNSBL) 的觀念，可以達到聯防的功能，例如國外的網站 relays.ordb.org [15]。

黑名單不只用於送件者端的阻擋，也可用於郵件內容的過濾中，本論文希望找到在信件內容中的不當回覆訊息，最常見的就是網址，因為寄件者可能來自許多不同的電子信箱，連線的 IP 也可以不斷的更換，但是廣告希望點選的網址應該是不能造假，也不應該是不存在的，這個觀念在網路上已經有人實行了[20]，但是在參考了這個國外的名單後，我們發現仍然有許多網站是無法阻檔的，這應該是因為廣告信本身具有文化及語言的相關性，因此除了參考國外的黑名單網站，自行維護一個屬於本地的黑名單網站有其必要性。

3 系統架構及資料分析

第一小節討論如何設計一個適合使用者及可擴充的分流機制系統；第二小節介紹增強郵件伺服器的保護機制；最後討論透過資料萃取，分析廣告信的特性。

3.1 分流機制系統架構設計

由於學校大部份使用者對廣告信的看法是寧可放行 100 封也不能錯殺一封信，另外考慮 23,000 個使用者及未來每年繼續以 2,000 人增加的使用者，使得廣告信的過濾，不只是正不正確，必須建立一個分流架構，兼顧使用者意願、人性化介面及自動化管理，才能順利的運作及經營。

我們提出的分流機制系統，讓使用者可以選擇是否啟動分流，分流後的信件(即可能的廣告信)會放在另一個空間，如此可讓使用者減少閱讀許多信件，同時系統會每天提供廣告信清單，使用者隨時可以去將被分流的信件取回。同時為避免磁碟空間不足，用來存放可疑的廣告信系統也會自動刪除過期的信件 [12]。

本文提出的分流機制可以解決第二節所提到的單機或串聯式架構所帶來的缺點，提供使用者選擇是否攔截廣告信，可以避免抱怨，其次分流可以讓信件減少，達到保護郵件伺服器的目的；同時本系統又不改變郵件內文，使用者不需要更動任何設定，可直接導入系統；另外主動通知被過濾的信件，提供使用者可以隨時取回的機制，可對錯誤判斷進行補救；最後，定期刪除垃圾信件，可以避免資源的浪費。

本系統中因為增加郵件轉信站，所以必須解決其所帶來的相關安全性問題，同時為了使用者的郵件空間伺服器可應使用量的需求增加而擴充，利用 pop3 proxy 的觀念，在不改變使用者設定的前題下，自動將使用者的讀信需求，導向

正確的郵件空間伺服器，因此分流過濾機制必須要能將信件送到正確的郵件空間伺服器上，我們利用虛擬地址(virtusertable)的觀念，在第一層的郵件轉信站上將所有的帳號列表，同時依不同郵件空間伺服器的位置，給予不同的轉向。在圖 4 中 user1 不啟動分流，分流伺服器上直接導向 pop3 伺服器，user3、user4 啟動分流，所以導向 spam-1spam-2 伺服器，實際上 spam-1 spam-2 可以是同一台主機，當過濾主機負荷加重時或基於備援的問題，可以增加一台過濾主機，利用網域伺服器的 round robin 機制或直接指明那些使用者由 spam-1 過濾、那些使用者由 spam-2 過濾，這個架構同時也可以運用在將 spam-1 及 spam-2 的過濾條件設定為不同的程度，提供不同條件的過濾，讓使用者可以選擇不同的偏好。

3.2 郵件伺服器保護機制

大量廣告信不僅造成使用者困擾，對郵件伺服器的管理也是一大考驗，分流機制可有效保護郵件伺服器，除了讓使用者容易取信外，還必須注意郵件伺服器的流量是否異常，以中央大學而言，每天平均有 10 萬到 20 萬封信件，且不斷的

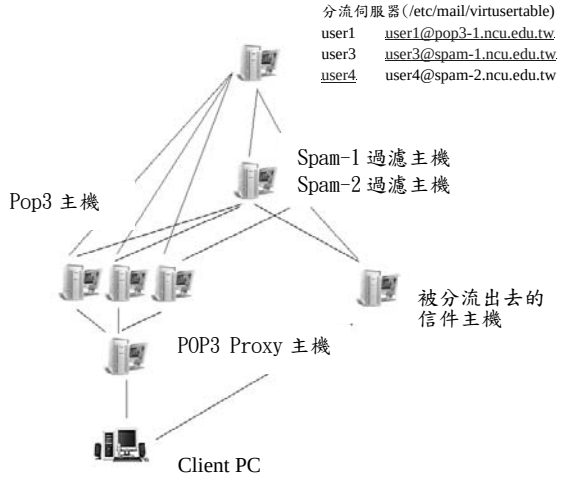

圖 4 分流機制擴充示意圖

成長，若遇到惡意攻擊則可能暴增多倍，所以郵件伺服器應該要具有偵測異常、自我保護及容錯的機制 [3][5]。

以中央大學為例，我們設置了幾個檢查點，用以預防並即時告知服務是否異常或已中斷：

(1) 異常流量監測：

a. 每小時個人信件超過一定封數的檢測－避免系統被不當使用或惡意攻擊造成停擺

b. 針對不特定使用者的送信主機，拒絕連線

c. 根據郵件伺服器的紀錄，每天平均會有 30 萬筆的紀錄，因數量太大無法以人工觀察，所以自動過濾正常紀錄，只留下異常訊息，提供管理者快速了解狀態

(2) 郵件掃毒伺服器除了過濾病毒外，並記錄及統計中毒 IP，校內使用者隔離，校外使用者透過 Rwhois [18] 資料庫線上查詢通知各校、區網中心及各大 ISP 管理者。

(3) 為了只提供校內使用者寄信，最簡單的作法就是在郵件伺服器中(relay-domains) 只放入校內的 IP，只接受校內轉信功能，另一方面我們也啟動 SMTP AUTH 的認證機制，並推廣使用中。

3.3 資料萃取

由於廣告信不斷的在變化，萃取郵件內容可以幫助了解廣告信的發展及特徵，資料萃取的過程，為尊重個人隱私權，只萃取來源端資料，包含 Msgid、From address、Subject、URI list。

郵件內容的資料屬於非結構化的資料，每筆資料沒有共通的結構性可言，經常為長短不一，剛好符合文字探勘(Text Mining, TM)的精神，因此我們透過知識探勘的步驟對資料做蒐集、清理、轉換、Text Mining 後將結果呈現與解讀

廣告信的內容開始有變小的趨勢，用以避免被貝氏分類法或 keyword 重點加分法發現。由萃取的資料我們發現廣告信可能有下列特徵：

(1). uri address 以 random 的方式，同一個 IP ，同一個 domain，產生兩層到三層不等的非常不規則性的 URI 名稱。例如：

04Ab51Y.ENiCG8vX.fffgggbgeess.cn
05B1Za.Y4Uv248E.fffgggbgeess.cn

(2). uri address 以 random 的方式，同一個 IP，同一個 domain，產生兩層到三層不等但是非常有意義的字。例如：

backfill.gopamania.com
backorder.gopamania.com

(3).uri address 以大小寫來混亂相同的名稱，例如 Com.tw　cOm.Tw

(4).uri address 以 16 進位的方式表示，例如

http://%77%57%77.%54888.%43%6f%4d.%54%77 ➔ wWw.T888.CoM.Tw

http://%6e%65%77%5398.%6f%52%67 ➔ newS98.oRg

(5).參考國碼網域（country code TLDs, ccTLDs）[11] 的資料，經過合併計算後，發現幾個現象，在每天 13000 個不同的網址中，只有約 5000 個不同的網域，以每個網域所擁有的不同網址個數劃出曲線圖(圖 5)，其中 10 個以下約佔所有網域的 98.21%，也就是大部份的網域不會有太多不同網址，表 1 是 6 月 28 日同一網域有 100 個不同網址

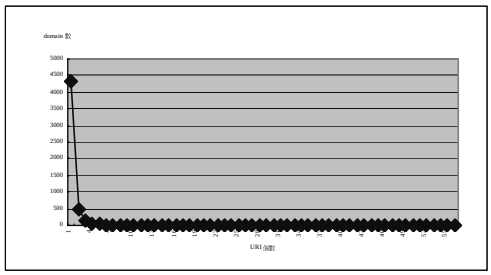

圖 5 同一網域含有不同網址的個數分

的列表，扣除少數知名的網域如 yahoo.com 及 ncu.edu.tw 等幾個大學的網域後，這些高達 500 個不同網址的網域，存在著相當可疑的訊息，我們將在系統實作時透過一些參數的不同權重，試著找出黑名單。

表 1　同一個網域含 100 個以上不同網址

domain	Uri 個數	Uri random 變化的平均數	Gethostbyname()
fffgggbgeess.cn	938	9.432836	5 個以上同一個 IP
hotblogz.biz	502	1.125498	5 個以上同一個 IP
gaizini.com	216	1	5 個以上同一個 IP (4 個 IP)
bhjhbv.biz	168	1.458333	5 個以上同一個 IP
brillianc3.net	165	5.381818	查不到
dhdas.biz	133	1.451128	5 個以上同一個 IP
yahoo.com	128	1.25	不同 IP
cheeked.net	120	1	5 個以上同一個 IP

第四章 系統實作

本章介紹分流機制實作及測試。

4.1 分流機制實作

目前本分流機制已經應用在中央大學郵件伺服器系統，每天過濾超過 10 萬筆信件，主要由分流伺服器、廣告信判斷主機、被分流信件暫存主機所組成，使用者不需要改變任何的設定，如果有被系統分流出去的信件，系統會每天主動通知，使用者可隨時進入暫存信件區取回被錯誤分類的信件。

4.2 系統評估與討論

由於廣告信偵測系統無法達到百分之百正確，因此我們以使用者的滿意度為本系統的主要評估依據，中央大學電算中心的電子郵件系統使用者為樣本做了一個意見回饋調查表，統計結果可參考 [1]。

4.3 黑名單更新的方法實作與評估

郵件內容的回覆訊息(即 URI)萃取流程及處理功能說明如下：

功能模組一：記錄郵件的 msgid、subject、from address、uri list

功能模組二：啟用 Spamassassin 的 URIDNSBL plugin 的功能，並參考本系統的黑名單

功能模組三：建立網域伺服器

本系統所建立的黑名單是以網域伺服器的一筆記錄方式存放，除可自己使用，也可提供分享。

功能模組四：每天統計可疑名單：

我們考慮多個參數給予不同的權重。第一個參數是每個網域其不同網址，以超過 100 個的例子而言，除了 yahoo.com，其行為都有假造及逃避過濾的意圖。第二個參數是 URI address 本身的名字變化，本文中只以簡單的變化量來看，讀入第一個字元與第二個字元比較，如果其 ascii 落在十進位的 65-90(A-Z)為同一群組，97-122(a-z)為同一群組，48-57(0-9)為同一群組，每一個字元之間如果從一個群組變化到另一個群組則 count 加 1，例如 t3eAf.a8VrE 計算得到為 9 而 01LpDr.i9a4THIBa 則為 10，一般正常的網址不會有變化，其值為 0，但是這樣的算法，對於以字典字來變化網址的部份，計算不出其差異性。

第三個參數是是利用 gethostbyname 的方式，取得每個網址的真正 IP 位置，但是此法最大問題在於查詢網域名稱需時相當久，且有時會被認為是攻擊遭到阻檔，所以改為抽樣，僅取其

中 5 個網址查詢，如果均為同一個 IP，即給予權重。

第四個參數是針對色情網站篩選，在主旨的部份，加上一些字串比對，例如：”情色”、”辣妹”等。

4.4 測試地域性黑名單更新後的成效

記錄 6/20-6/30 的統計量如表 2，本統計量的單位是連線數，並不是實際送給幾個人的信件數，為避免隱私權的問題，萃取資料沒有收件人(即 SMTP 中的 to 欄位)資料，所以無法計算實際上的信件數。

6/22-6/25 的結果並沒有顯示單獨因為我們的黑名單而被判定為廣告信的數量，因為這段時間實驗時並沒有把實際篩選出來的黑名單加入抵制行列，這個資料初步可以顯示，單獨參考國外的黑名單，可以過濾到一些不恰當的網址，但是沒有考慮到地域性的因素，因此維護一個地域性的黑名單仍然有其必要的存在性。

表 2 6/20-6/30 的統計量

日期	同時符合國外及本系統參考名單的廣告數量	國外名單中沒有，只符合本系統參考名單的廣告數量
6/20	1001	161
6/21	1694	218
6/22	466	0
6/23	769	0
6/24	304	0
6/25	397	0
6/26	3067	700
6/27	3711	639
6/28	4243	711
6/29	4897	863
6/30	4634	324

5 結論

本研究主要以中央大學電算中心所提供之郵件伺服器為主要研究對象，提出一個分流機制可以供使用者依意願啟動過濾機制，並且搭配郵件管理的檢查點及具備擴充性的架構，使得廣告信不會影響正常信件的流量及運作。同時，我們提出一個方法可從信件中萃取出含有不恰當回覆網址，以此發現發送廣告信者，並成功的予以抵制。本研究所搜集到的廣告信，雖然難以涵蓋所有廣告信的特性，但適合本校使用者且反映特殊群體與地域性，可彌補國外黑名單網址不足。同時所搜集之黑名單以網域伺服器技術存放，可提供其他相關軟體取用及分享。

目前本研究只找出回覆訊息中的可疑名單，並不能減緩廣告信的信件總量，未來應該可以從這些信件的 received IP 再得到確實送信的 IP ，直接抵制其寄送信件進入。其次，近來釣魚信件所帶來的損害遠比單純廣告信來得嚴重，如果可以從郵件的回覆訊息中歸納出其意圖不軌的可疑性，提供使用者警示，應該可以大量減少網路安全的問題。

參考文獻

[1]. 王雅慈、張慈敏、劉劍青、曾黎明、陳奕明，「Spam/Virus Mail 分流機制探討—以中央大學為例」，TANET 2004 論文集，2004.

[2]. 楊素秋、曾黎明，「IP 管理資訊查詢與攻擊事件自動通告系統的實現」，TANET 2003 論文集，2003.

[3]. 賴守全、謝木政、郭文曲，「具自我保護機制之多網域叢集式電子郵件系統」，TANET 2004 論文集，2004.

[4]. 賴守全、謝木政、藍松月，「開放性郵件中繼主機防治系統之設計與實作」，TANET 2003 論文集，2003.

[5]. 謝仁瀚、廖享進，「具備容錯機制及多連外線路之叢集式郵件伺服系統」，TANET 2003 論文集，2003.

[6]. Cormac O' Brien, Carl Vogel, "Spam Filters: Bayes vs. Chi-squared; Letters vs. Words", Proceedings of the 1st international symposium on Information and communication technologies, September 2003.

[7]. Karl-Michael Schneider, "A Comparison of Event Models for Naïve Bayes Anti-Spam E-Mail Filtering", Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics - Volume1, April 2003.

[8]. Le Zhang, Jingbo Zhu, Tianshun Yao, "An Evaluation of Statistical Spam Filtering Techniques", ACM Transactions on Asian Language Information Processing, December 2004.

[9]. R. Grier and S. Schappel, "Junk email filter using naïve Bayesian classification," http://www.ryangrier.com/news/archives/BayesianEmailFilter.pdf

[10]. SMTP,RFC821,http://www.faqs.org/rfcs/rfc821.html

[11]. Country-Code Top-Level Domains, http://www.iana.org/cctld/cctld.htm

[12]. Expire mail http://sunsite.cc.ncu.edu.tw/expire_mail/

[13]. K9, http://www.keir.net/k9.html

[14]. Openwebmail,http://www.openwebmail.org/

[15]. Ordb.org, http://ordb.org/

[16]. http://people.apache.org/~felicity/AC2004/Slides%20Export.pdf

[17]. POPFile,http://popfile.sourceforge.net/

[18]. Rwhois Server Operation Guide, http://www.rwhois.net/

[19]. SpamAssassin, http://spamassassin.apache.org/

[20]. SURBL - Spam URI Realtime Blocklists, http://www.surbl.org/