一個以異常寄件行為來偵測郵件病毒的方法

林大為
中央大學資訊管理系
中壢市五權里 2 鄰 38 號
s4364007@cc.ncu.edu.tw

陳奕明
中央大學資訊管理系
中壢市五權里 2 鄰 38 號
cym@cc.ncu.edu.tw

摘要

電子郵件病毒問題進來變得更加嚴重，如最近的納妲(Nimda)病毒，利用多重管道的感染方式，其感染的速度擴散的廣度，都非以前所能比擬，非常短的時間內，多數的電腦已遭到感染，在防毒軟體尚未更新病毒碼的這段期間，損失與災害早已造成，因此我們亟需一種機制或方法，在病毒出現時就能發現其存在，採取因應對策，發現越早損失越低。

本文提出一簡單有效寄件者行為規範(User Behavior Profile)的建立方法，可用來偵測異常郵件行為，包含郵件病毒所產生寄件行為，不管病毒是已知或是未知，都能有效偵測。我們的方法是以正常的寄件行為為基礎，經過群組關係與依存關係分析，然後產生寄件者的行為規範。我們以郵件伺服器的郵件紀錄做了相關的分析與模擬檢驗，效果相當理想。

關鍵詞: 異常行為偵測、郵件病毒、郵件行為分析

一、簡介

以往病毒感染的速度與擴展的範圍，都是局部較為緩慢，通常需要好幾個星期或幾個月，才能產生較大的影響，對於企業而言不會構成什麼威脅，就算碰到了病毒，通常在其前造成危害或大量擴散之前，就已遭到控制，面對病毒處理反應的時間都相當充裕，防毒軟體的更新速度就不是那麼迫切。防毒公司在取得病毒樣本，分析與製作解毒劑等處理流程的時間，可以人工或半人工化來因應都來得及[1]。

隨著依賴網際網路的方便日深，全球連接的環境，變成病毒流行的新管道，病毒透過網際網路，可以傳遍世界，隨著病毒撰寫技巧的改進，多重管道的感染方式(IIS 伺服器、郵件附件、網路芳鄰)，大量而快速的感染，如最近的思坎(Sircam)與納妲(Nimda)病毒[15][12]，在病毒的出生率(birth rate)與死亡率(death rate)之間，明顯出現時間延遲上的關係，出生率快速成長，死亡率遠遠落後，傳染的速度與範圍是傳統病毒無法比擬的，等發現其存在時，感染的範圍已經相當廣，即使事後受到控制，但傷害已經造成，系統攤瘓，資料的毀損，機密資料的外洩等，以往防毒較為緩慢的因應方式，不再適用，我們需要更快速來因應這樣的的問題，才能將損失降至最低[6][18]。

目前的防毒軟體通常是以病毒的特徵碼(virus pattern)來找出該病毒，但這樣的前題是必須先取得病毒特徵碼，若沒有病毒特徵碼，對於這樣的病毒是沒有什麼防衛能力的。以特定的病毒特徵碼來防毒的缺點是

- 特定的病毒特徵碼解特定的病毒
  對於新病毒無能為力，況且同一類型的變種病毒，特徵碼可能會改變，必須更新病毒碼，並且已有所謂的病毒產生器[24]，各式各樣的病毒層出不窮，在防治病毒上相對處於被動地位，許多傳統病毒的技術如多型(polymorphism)，也會被用來使用在郵件病毒上，病毒的特徵碼將會變得更短更難以補捉甚至於造成誤判[5]。
- 時程較慢，失了先機，危害已造成
  當有新病毒產生，必須先取得病毒的特徵碼，在取得病毒的特徵碼前這段時間內，對

於新病毒豪無抵抗力，可能已造成損害。

安裝防火牆，但是對於郵件病毒而言，防火牆的幫助並不大，防火牆的主要功用是防範外部對內部的非法存取，電子郵件通常經由正常合法的管道通過防火牆，內部的使用者只要接收了含有郵件病毒的郵件，不小心開啟之後，在內部擴散開來，防火牆是無能為力的。面對這樣的問題，我們需要一種方法或機制，即時發現異常郵件，採取對策，將傷害降至最低，不管這樣的病毒是已知的或未知的。

我們提出一套方法可以即時有效的偵測異常郵件，雖然目前我們對於資料只做離線(Off-Line)的分析，但衡量行為規範的資料大小與處理方法，我們相信其可達到即時偵測的目標。

首先必須建立寄件者正常行為規範(behavior profile)，當寄件者的寄件行為不同於行為規範時就表示寄件行為異常，而這樣的異常行為可能由於寄件者行為改變所造成，或是由於郵件病毒所造成，這時可以更進一步檢查寄件行為是否含有可疑的夾帶檔，來決定這樣的行為是由於寄件者行為改變所造成或是郵件病毒作用所產生。

我們的方法在資料分析上可以分成兩個階段，第一階段稱為群組關係的建立，找出寄件行為資料間的群組關係，第二個階段更進一步找出同一群組關係內的依存關係。我們的方法有很大的彈性，當資料量不足時，可以只建立群組關係，不建立依存關係，避免高誤報率的產生，但通常也伴隨較高的 False Negative，當資料量足夠時可以建立到依存關係的階段，降低 False Negative 的情形。

我們的偵測方法可以在初期階段找出異常郵件行為，不管這是已知或未知名的郵件病毒所造成，若其行為違背寄件者的行為規範，就會被找出，但關於病毒名稱、對系統的作用與影響及解毒方式，仍需靠專家或解毒軟體的幫助，找出該病毒的解決之道。

本文的安排如下，第 2 節介紹相關的研究，第 3 節說明我們的方法與原理，第 4 節資料處理與分析模擬，第 5 節結論。

二、相關的研究

對於病毒的研究或是異常行為的偵測的方法，學者專家採用下列數種方法來偵測病毒[1][5][9][10][20][22][23]：

字串搜尋法[1]就是病毒特徵碼比對的方式，可用來檢查已知的病毒，精確度高，對於未知的病毒則無能為力。演算法搜尋法、疫苗法與反盜法[1]較適合於會附加於檔案而修改檔案的病毒，而有些郵件病毒本身通常以獨立檔案的型式存在，不需附加於其它檔案中，因此演算法搜尋法與疫苗法也不完全適合於郵件病毒的偵測，而調查法[1]，也必須進一步驗證，才能確定其成效。

使用類神經網路，必需先決定有那些屬性，能夠有效的區分正常與異常的資料，然後經由訓練資料(training data)，調整類神經網路的權重值(weight)，使用測試資料(testing data)驗證效果，其中屬性的決定是非常重要的因素，關係整個類神經的好壞，類神經系統的結構如果太過複雜，運算速度可能會太慢，難以用再即時偵測，結構簡單可能精確度不夠[9]。

數位免疫系統(Digital Immune System)是一套由 IBM 與 Symantec 所建構的一套商業化自動化防衛系統，能夠發現病毒、分析病毒、產生解毒劑，這些流程可以完全自動化，無需人為因素介入，大大縮短從病毒的發現到最後解毒的流程[20][22]，”如果不能夠發現病毒，就沒有辦法治療”，病毒的偵測主要是在前端(客戶端)，其使用 Symantec 的一項啟發式技術 Bloodhound[23]，利用病毒程式的一些行為特徵，宣稱可以偵測出 80%新的未知的可執行檔型式的病毒，90%新的未知的巨集型的病毒，不過啟發式的方法，通常伴隨較高的誤報率[5][18]，仍需更進一步確認。

異常行為偵測法(abnormal behavior)是基

於正常行為為基礎，首先必須建立正常行為的規範(Behavior Profile)，異常行為的決定是由該行為是否背離正常行為或是預期的行為來判斷，因此可利用這樣的特性來補捉未知的病毒，一個正確有效的行為規範是異常行為偵測正確率的關鍵，關於正常行為規範的建立，有不同方法與理論，如：採用統計學的方法，類神經網路、資料採礦(data mining)等各種不同的方法，觀察的資訊也不同，如：系統資源(系統記憶體與 CPU)的使用狀況、使用者在不同時間的行為關係，執行碼間的執行順序關係，使用者的工作規範(User Work Profile)，很難界定這些方法間的好壞，通常比較客觀的衡量標準是由誤報率來決定。

Forrest 發現 Unix 系統中擁有 root 權限的正常程序的系統呼叫(system calls)，存在穩定的短執行碼序關係，可以利用這項特性來區分正常與異常的程式行為[14]，這些的執行碼序關係是觀察程式執行時的低階行為紀錄所得，Forrest 的方法是用來檢查系統本身程式行為的正常與否，而我們研究異常郵件的問題，並非在判定郵件伺服器本身的程式是否出了問題(此指 sendmail 程式本身)而是郵件記錄資料中的寄件行為是否異常，這些資料主要是由寄件者電腦所產生，因而是檢查寄件者端 mail client 程式是否出了問題，如遭受郵件病毒的作用等。我們的方法是以寄件者的行為來衡量，並未做高低階間格式的轉換，可在語意(semantic)層次來解釋，這樣的關係不會因為作業系統、版本與 mail client 程式而不同，我們使用郵件伺服器的紀錄分析資料，Forrest 的方法在非 Unix 的作業系統或不同的硬體平台是否能得到同樣的結果，仍需進一步驗證。

Lane 所處理的問題是在傳統文字模式的 UNIX 環境中[21]，經由使用者的所使用過的命令，來判定使用者是否異常，使用者所下的命令資料是以連續的資料串流(stream flow)方式存在，而 Lane 為了做相似度的計算，將連續的資料以等長的方式切割，而命令串流是屬於名目性資料，必需轉換成數值性的資料才能進行相似度的運算，後續為了節省空間，計算群組的中心，只保留中心與半徑的資料，這樣的轉換過程是否仍能有效保留使用者的特性。我們的問題，在資料的特性上，每一次的寄件行為都可以看成個別一次的行為事件，不同寄件行為間分隔得很清楚，寄件行為的資料長度也不等長，在資料的特性上有所不同，處理方法不見得能適用。Lane 所處理的問題源於傳統 Unix 環境的文字模式下，現今大都採用 Unix 環境使用者改採 X-window 或是使用 Microsoft windows 的圖形介面環境，資料特性上有很大的改變。

異常行為(abnormal behavior)的偵測與困難點[17]。一般而言，誤用(misuse)的偵測的誤判率低，而異常偵測的誤判率高。大家也知道誤用的方法對於未知的病毒無法防範，而異常行為偵測的方法雖然可以用來偵測未知的病毒或攻擊，建立的正常行為規範如果太寬鬆，將導致 False Negative 的錯誤，太嚴格將導致 False Positive。

三、我們的方法(寄件行為規範的建立)

在介紹我們所採用的方法前，必須對於一些基本的名詞給於定義

- 寄件行為(Mailing Behavior)

指寄件者在某一個時間點上所產生的一次郵件寄送行為，一次的寄件行為包含下列資料，發生的時間，加上某些收件者

(寄件時間　收件者對象)

例子：

(Aug 20 13:30:30　A B C D)

這是指在該時間寄件者同時寄信給 A B C D 4 人，ABCD 這些符號是四個收件者的代稱。

- 寄件行為特徵(Behavior Pattern)

其中 (A B C D) 這些收件者在一次寄件

行中同時出現，這些收件者群就是一次行為特徵，對寄件者而言，累積一段時間之後將會有許多筆的寄件行為特徵。

例子：

(Aug 20 13:30:30 A B C D)

(Aug 20 13:40:30 E)

(Aug 20 13:45:30 A C D)

其中

(A B C D)

(E)

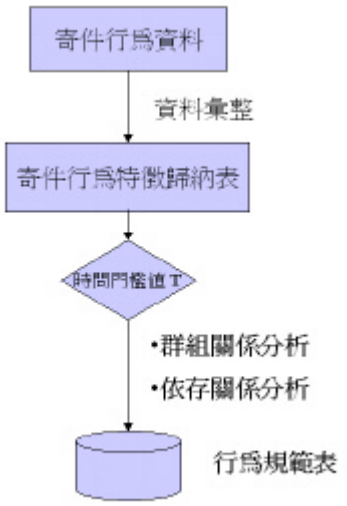
(A C D)

這些就是寄件者所累積的寄件行為特徵。

- 寄件行為規範(Behavior Profile)

由一堆寄件行為特徵經由推演而得到寄件者行為規範(Profile)，正常的寄件行為的行為都應該會涵蓋於行為規範中某一群組中。

寄件行為規範的建立包含兩個階段，第一個階段是群組關係的建立，第二個階段依存關係的建立，為了能夠快速有效建立行為規範，在中間階段加入寄件行為特徵歸納表與時間門檻值得設定，以後若要重新建立行為規範時，不須從寄件行為原始資料重新開始，只須從行為特徵歸納表推演即可，整個行為規範的建立流程如下所示



圖一 行為規範的建立流程

接下來我們先說明行為規範中的兩個重要關係的建立：群組關係與依存關係，然後再說明為什麼中間還需要寄件行為特徵歸納表與時間門檻值。

3.1 行為規範中群組關係的建立

3.1.1 郵件行為群組關係的概念

我們希望建立寄件行為特徵間所存在的群組行為關係，這樣的群組行為關係具有下面的特性。

> 寄件者所產生的一次寄件行為只會出現在行為規範(Behavior Profile)的某一個群組中且只有一次。

當上述條件成立時，如果寄件者的寄件行為涵蓋了行為規範中的兩個或兩個以上群組時，就是一個異常寄件行為。

我們希望建立的群組行為規範，具有互斥周延的特性

- 互斥:不同的群組間不會有交集，相同的群組元素不會出現在不同的群組中，也就是說群組間的收件者不會重覆。
- 周延:涵蓋完整的收件者，儘可能使得所有的收件者都出現在通訊錄中，否則在判斷寄件者的寄件行為，新的收件者將難以判斷。

3.1.2 群組關係的建立

對一位寄件者而言，累積一段時間將會有許多筆的寄件行為資料，這些寄件行為資料間的行為特徵，有些會相同或相似或不相同，下面我們以一個例子來說明這些行為特徵，欄位一表寄件行為代號，以寄件行為發生的時間順序排列，欄位二表行為特徵中括號內的數字表收件者，不同的數字代表不同的收件者，行為特徵就是一些收件者的集合，亦可稱為收件者群。

表一 寄件者的每次寄件行為特徵

寄件行為代號	行為特徵(收件者群)
0	[0]
1	[1]
2	[2]
3	[3]
4	[4]
5	[5]
6	[1 4]
7	[6]
8	[7]
9	[8]
10	[1 8]
11	[4 7 8]

我們可以看出來第 4 筆資料與第 6 筆資料彼此有交集，9、10、11 筆資料有交集，6 與 10 筆資料間也有交集，我們也看到有一些行為特徵間沒有交集，如：1、2、 3，只要是行為特徵有相關的(交集)，我們就將其行為特徵取聯集以產生這些行為特徵間的最大集合，結果如下表所示，原先的 12 筆寄件行為資料歸納之後產生了下面的行為規範表。

表二 寄件者的行為規範表

行為規範代號	群組行為特徵
0	[0]
1	[1 4 7 8]
2	[2]
3	[3]
4	[5]
5	[6]

我們可以清楚的發現在原表一的寄件行為只會隸屬於行為規範的某一個群組，參見表三，而且只出現在行為規範中的一個群組。

表三 寄件行為與行為規範間的對應關係

寄件行為代號	行為特徵	隸屬的行為規範代號
0	[0]	0
1	[1]	1
2	[2]	2
3	[3]	3
4	[4]	1
5	[5]	4
6	[1 4]	1
7	[6]	5
8	[7]	1
9	[8]	1
10	[1 8]	1
11	[4 7 8]	1

3.2 行為規範中群組內的依存關係 (Dependence Relationship)

同一群組內的關係仍然可能太過於粗糙，因其是聯集產生的，可以更進一步找出其間所存在的關係，更精確的掌握住同一群組內的行為關係，這有利於降低異常行為被誤認為正常行為。

依存關係所考慮的是同一群組內，兩兩元素間所存在的關係，以表二為例，其中行為規範中代號二的群組由[1 4 7 8]所組成，若我們能找出若有 7 出現則 8 一定會出現的關係，也就是 7→8，這就是所謂的依存關係，這個觀念與資料探勘(data mining) 所談 confidence 概念相同，這裡我們不強調 support 的概念，以 confidence 為主[17]，而且所找出的依存關係是以 confidence=100% 為準，建立同一行為規範內元素間的依存關係，如在[1 4 7 8]的行為規範中，若有 7→8 與 8→1 的關係存在，我們就可以將同一群組內元素間的關係描述的更清楚，這個階段所建立出來稱為依存關係。

接下來我們以例子說明一個完整的寄件者的行為規範的內容

表四 寄件者的完整行為規範

群組代號	群組關係	群組內的依存關係
1	[0]	
2	[1 4 7 8]	7→8 8→1
3	[2]	
4	[3]	
5	[5]	
6	[6]	

在上表的行為規範中總共包含 6 個群組，其中代號為 2 的群組包含有 4 個收件者 1 4 7 8，而這個群組內存在的依存關係有 7→8 與 8→1，其他的群組代號都是由單一收件者所組成的群組，因此只有群組關係的資料，無依存關係的資料。

3.3 郵件行為異常與否的判斷

依據上面方式所建立的行為歸範，包含有兩種資料，群組關係與依存關係，寄件者的寄件行為將會只出現在行為規範中的一個群組中，而且應該只有一個，也就是寄件行為不會涵蓋行為規範中兩個(含)以上的群組，這是群組關係的檢查標準，若寄件者的行為滿組足了群組關係，可進一步檢查是否滿足群組內的依存關係，這兩種關係的檢查標準整理如下：

群組關係：

寄件行為僅屬於一個群組，且只有一個，否則為假。…….. (1)

依存關係：

若有 A→B 的依存關係存在，若 A 出現則 B 一定出現，否則為假。…… (2)

行為規範中的群組關係是採取聯集的方式產生，除了涵蓋歷史資料所有可能之外，有些不屬於寄件者的寄件行為，也會涵蓋在聯集之後的群組中，聯集是屬於寬鬆的標準，此點可能導致將異常行為當做正常行為來判斷，而依存關係是在同一群組內再找出存在其中的元素相對關係，可以將群組關係檢驗中異常行為誤當做正常行為的情形降低。

3.4 快速有效的產生群組行為規範

隨著時間的前進，寄件者的行為也會隨著改變，新的寄件行為產生，行為規範必須重新訓練，而早期的行為已經不再出現，也必須加以排除，否則將會模糊群組規範，降低判斷寄件行為的有效性。基於上述的原因，我們提出以下"行為特徵歸納表"與"時間門檻值"的設定，此總表的建立可以幫助我們在短時間快速的建立新的行為規範，而時間門檻值的決定可以幫助我們濾除早期不再發生的行為。

首先我們說明行為特徵歸納表的建立方法，將寄件者的每一次寄件行為特徵資料歸納出如下表的資料，稱為寄件行為特徵歸納表，這個歸納表包含四個欄位，欄位一行為特徵代號，不同的行為特徵就會賦予一個代號，欄位二紀錄這個行為特徵總共出現的次數，欄位三紀錄這個行為特徵最近一次出現的時間，欄位四紀錄行為特徵的內容，這個歸納表累計了寄件行為特徵的次數與最近發生的時間，這些寄件行為特徵若是不完全相同就視為不同兩筆資料，在歸納表就會存在兩筆資料，因此歸納表中所紀錄的資料與行為規範中的資料是不同的，行為規範中的群組資料是聯集後的結果，歸納表記錄了所有曾經出現過的行為特徵，不做聯集的動作，因此歸納表的資料筆數會比行為規範表的筆數多，其行為特徵歸納總表處理邏輯如下所描述：

每當有新的寄件行為產生都會跟表中以往出現過的行為特徵做比對，若完全相同則在該對應的行為特徵代號上，累計次數欄上加 1，並更新最近出現的時間欄的時間，若是以往未曾出現過的行為特徵，則新增一筆行為特徵的資料，累計次數欄位設為 1，並將時間填入最近出現的時間欄中。

表五 寄件行為特徵歸納表

行為特徵代號(p_no)	累計次數(count)	最近出現時間(last)	行為特徵(Pattern)
0	3	620	[0]
1	122	579	[1]
2	1	11	[2]
3	1	12	[3]
4	22	544	[4]
5	1	18	[5]
6	1	24	[1 4]
7	5	564	[6]
8	14	90	[7]
9	14	559	[8]
10	6	182	[1 8]

時間門檻值的作用，主要是幫助我們濾除早期不再發生的行為。透過此歸納表與時間門檻值，可以快速有效產生行為規範，不必再從原始寄件行為資料重新推導，大幅縮短建立行為規範所需的時間，以我們所分析的寄件行為

資料為例，寄件行為資料量約 500 筆的資料，歸納表的資料量平均可以縮減成 5 分之 1 約 100 筆的資料量，這個行為特徵歸納表是持續累積的，因此只需針對新發生的寄件行為，這更有助於縮短處理的時間。

接下來舉一例說明，表五是行為特徵歸納表的內容，從累計次數欄我們可以看出來縮減了大量原始的寄件行為資料，接著選定適當的時間門檻值 T，我們以 T 必須在 100 個時間單位之後為例(數值越大越接近現在)，行為特徵代號 2、3、5、6、8 這些早期行為特徵便不會加入行為規範之中，下表中以時間門檻值 T 100 為例，其有效行為特徵的數量將從 11 降至 6。

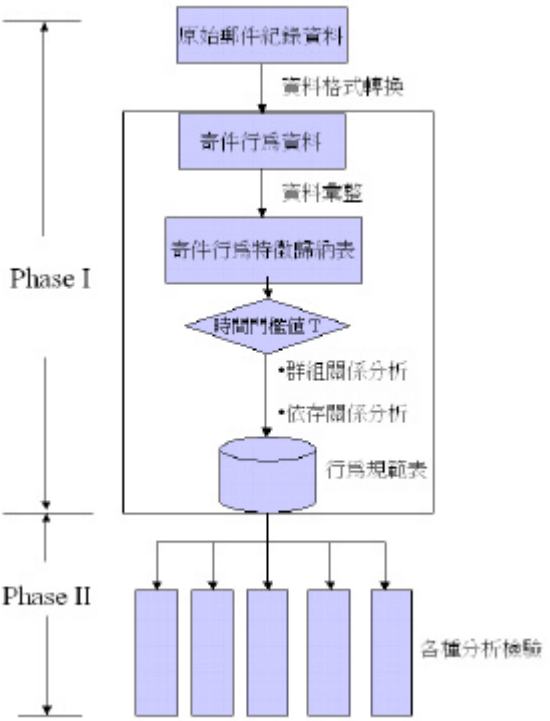
表六 寄件行為特徵歸納表(last 欄 >=100)

行為特徵代號(p_no)	累計次數(count)	最近出現時間(last)	行為特徵(Pattern)
0	3	620	[0]
1	122	579	[1]
2	22	544	[4]
3	5	564	[6]
4	14	559	[8]
5	6	182	[1 8]

從表六我們也可以彈性決定參數的條件(累計出現次數(count)與最近出現時間(last))，比如說累計出現的次數越多，時間門檻值可以越小，保留越久，不同的行為特徵可以有不同的時間門檻值等。

四、資料處理與分析模擬檢驗

資料處理流程與分析可以參見圖二，原始資料來源是學校計算機中心的 SMTP(sendmail) 的郵件紀錄(mail log)，包含有 6 千多個寄件者的資料，涵蓋期間從 2001.02~2001.09 月，資料量約為 1.5GB，而這樣的龐大的資料也只涵蓋了 mail client 與 SMTP 伺服器溝通所需的基本資料而已，並沒有郵件的內容(指信件的主題、內容…等)，為了確保寄件者的隱私權，在我們取得郵件紀錄前，已將寄件者與收件者資料使用 MD5 演算法處理過。



圖二 資料處理與分析流程

在資料格式轉換的階段，所做的工作是將原始的 SMTP 郵件紀錄的格式轉換成以個別寄件者為主的寄件行為資料，以利後續行為規範的建立，為了簡化處理與表示，我們將已經過 MD5 編碼過的寄件者的 Email 與收件者的 Email 以數字表示。

寄件者行為規範的建立，是以正常的寄件行為資料為基礎，因此為了確認原始寄件行為資料中沒有包含郵件病毒行為資料，我們以寄件者所有的寄件行為最大寄件量都不超過 25 的寄件者為挑選對象，因為從以往至今所發現的郵件病毒來看，都是以大量寄送為主。

4.1 行為規範的完整性檢驗

要建立寄件者的行為規範，訓練資料是否涵蓋足夠的寄件者行為特性是很重要的，否則建立出來的行為規範，將會有較高的誤報率，我們以下列的方式挑選出較能完整涵蓋寄件

者行為特性的寄件者資料，以做為後續相關分析使用，將寄件者所有的資料，以餘數運算(mod 6)的方式分成兩個集合，其中一個集合當做訓練資料，一個當做測試資料，彼此沒有交集，用訓練資料來產生寄件者行為規範，然後用測試資料來檢驗行為規範的完整性，若誤報率高(False Positive)表示訓練資料不足於涵蓋完整的寄件行為特性，誤報率低表訓練資料能夠完整涵蓋寄件者的行為特性，訓練資料與測試資料的數量比是 5:1，採用這樣的方式可以將訓練資料與測試資料平均分佈於相同的時間區段中，排除時間因素造成行為改變所帶來的影響，結果如下表所示。

表七 行為規範的完整性檢驗

寄件者代號	通訊錄人數/寄件行為總數	群組關係檢驗*	依存關係檢驗*
User130	51/770	0.02	0.62
User1409	33/645	0.00	0.01
User1688	22/684	0.02	0.03
User2242	39/734	0.02	0.02
User3101	34/801	0.00	0.00
User4676	18/1698	0.00	0.00
User744	22/531	0.03	0.03

*群組關係的檢驗標準參見 3.3 ..(1)的說明

*依存關係的檢驗標準參見 3.3 ..(2)的說明

上表所呈現的結果，整體而言都相當不錯，只有在 User130 依存關係檢驗的誤報率(False Positive)較高 0.62，這表示 User130 的訓練資料尚不足於完整涵蓋寄件行為特性，群組關係檢驗屬於較為寬鬆的標準，而依存關係檢驗屬於較嚴格的檢查標準，是在滿足群組關係檢驗的前提下再進行依存關係檢驗，因此依存關係檢驗的誤報率會比群組關係檢驗來得高。

4.2 時間因素下對行為規範的檢驗

這一部份的分析主要是想檢驗時間因素對於行為規範的影響[13]，資料來源與前一分析一樣，訓練資料與測試資料的產生方式不同，將寄件者的寄件行為資料以時間先後關係，前 5/6 的資料當做訓練資料，後 1/6 的資料當做測試資料，訓練資料與與測試資料涵蓋不同的時間區段，資料量比 5:1，結果如下表所示。

表八行為規範的檢驗(時間)

使用者代號	群組關係檢驗	依存關係檢驗
User130	0.04	0.04
User1409	0.65	0.94
User1688	0.01	0.01
User2242	0.05	0.05
User3101	0.22	0.24
User4676	0.00	0.00
User744	0.06	0.06

從上表的結果來看，User1409 在表七中，雖然誤報率很低，但在表八中加入時間因素後，卻有很高的誤報率，這表示在近期中行為有較大的改變，導致於誤報率變大，User3101 的行為也有一些改變，整體而言本表會比表七有較高的誤報率，此因為寄件者的行為會隨時間改變所致，這也表示寄件者的行為規範在一段時間後必須重新訓練，才能抓住寄件者的行為。

4.3 測試資料期間分佈長短對於誤報率的影響

接下來我們想更進一步分析測試期間長短對於誤報率的影響，取訓練資料 500 筆，緊隨其後的 50 筆資料當測試資料 Test1，100 筆資料當 Test2 及 150 資料當 Test3，如下表所示，結果參見表十。

表九 訓練資料與測試資料時間上的分佈關係

訓練資料	Test1	Test2	Test3
第 1~500 筆	第 501~550 筆	第 501~600 筆	第 501~650 筆

結果如下表所示

表十 測試資料時間長短對誤報率的影響

寄件者代號	Test1 誤報率	Test2 誤報率	Test3 誤報率
User130	0.08	0.11	0.12
User1409	0.05	0.75	0.82
User1688	0	0.01	0.01
User2242	0	0	0.03
User3101	0.04	0.06	0.04
User4676	0	0	0
User744	0.07	0.07	0.07

上表中我們列出依存關係的檢驗結果(已包涵蓋群組關係檢驗)，群組關係檢驗的情形

類似，隨著測試資料分佈的時間越長，通常誤報率也會隨著增加，其中 User3101 的 Test3 的誤報率比 Test2 小，是因為 Test3 的資料中出現歷史的行為，比率雖然比 Test2 小，但誤報的筆數仍較多，從上面的結果也可以得知，經過一段時間之後，寄件者的行為規範必須重新訓練，重新訓練的時機對不同的寄件者而言並不相同。

4.4 時間門檻值 T 的設定

隨時間前進寄件者的行為會變化，老舊的歷史行為可能不再發生，若不加以濾除，對於行為規範的建立會有負面的影響，進而影響檢驗的正確性，因此設定適當的時間的門檻值，以濾除不再發生的行為特徵，避免這些資料影響寄件者的行為規範的建立，接下來我們以時間檻值當條件，第一組的訓練資料包含第 1~500 筆資料(時間門檻值 T=0)，第二組訓練資料從第 51~500 筆(時間門檻值 T=50)，測試資料同為第 501~600 筆，用這兩組訓練資料分別建立寄件者的行為規範，用測試資料來驗證，時間門檻值間的差異，表十二分別列出兩組訓練資料所產生行為規範的群組數與誤報率，訓練資料與測試資料分佈如下表所示。

表十一 訓練資料與測試資料間的關係

	訓練資料	測試資料
第一組	第 1~500 筆(共 500 筆)	第 501~600 筆
第二組	第 51~500 筆(共 450 筆)	第 501~600 筆

表十二 時間門檻值對於行為規範與誤報率的影響

使用者代號	第一組訓練資料所產生的行為規範群組數(T=0)	誤報率	第二組訓練資料所產生的行為規範群組數(T=50)	誤報率
User130	32	0.11	30	0.11
User1409	3	0.75	3	0.76
User1688	9	0.01	9	0.01
User2242	10	0	8	0
User3101	5	0.06	4	0.06
User4676	18	0	7	0
User744	17	0.07	12	0.07

從分析的結果可以看出來，第二組訓練資料，加入時間門檻值的設定後，有一些寄件者行為規範中的群組數明顯降低了，但對於誤報率卻沒什麼影響，可以有效縮減行為規範大小。

4.5 不同寄件者間的比較(False Negative 分析)

由於我們所選擇的資料都是正常的並沒有不正常的測試資料，先前的分析也都是針對 False Positive，因此我們以其他的寄件者的寄件行為當測試資料用來檢驗行為規範，當做對於異常資料的偵測能力(False Negative 的檢驗)，是否行為規範能夠有效區別本身與其他的寄件者的寄件行為，為了避免通訊錄大小不同的差異干擾分析，我們挑選寄件者的通訊錄大小在 32 正負 2 的範圍，下表左邊欄位表示以該寄件者的資料為訓練資料產生行為規範，水平欄位表示以該寄件者的資料為測試資料，測試資料的資料筆數為 100 筆，結果如下表所示。

表十三 不同寄件者間行為的比較檢驗

測試 / 訓練	User88	User2279	User2776	User3101	User3543
User88	0	63	100	32	100
User2279	44	18	96	47	40
User2776	91	63	4	60	36
User3101	53	59	98	6	47
User3543	93	96	98	62	12

對角線的部分表示以該寄件者的資料訓練也以該寄件者的資料測試，理論上，如果訓練資料完整涵蓋寄件者的行為，這一部份的資料誤報率應該很低(False Positive 的觀點)，其他的欄位(相對於 False Negative 的觀點)應該越高越好，表示越能分辨出不同寄件者間的行為特性，整體而言相當不錯，當然這會因為不同寄件者行為特性而有高低。

4.6 模擬檢驗

從以前到現在所出現過的郵件病毒 Melissa、Love letter、Sircam、Nimda

[3][11][12][15]，大部分都以通訊錄為主，大量寄送，因此這一部份模擬檢驗的目的是想要是以隨機選擇通訊錄名單的方式來模擬郵件病毒的寄送行為，檢驗這樣的行為能符合寄件者的行為規範可能性有多少，而我們除了想檢驗大量寄送的行為外，也希望能量測行為規範對於小量寄送的行為的檢驗能力，況且未來難保沒有小量寄送的郵件病毒存在。

我們並沒有寄件者的真正通訊錄，在此做了一個假設，以累計所得的收件者清單當做寄件者的通訊錄，模擬資料的產生是以通訊錄名單的 10%、15%、20%、25%、50%、75%、100% 的收件者數量為條件，分別產生 100 筆測試資料，使用行為規範加以檢驗，訓練資料以寄件者的所有寄件資料當訓練資料來訓練[18]，下表是我們的模擬檢驗的結果，小括號中表示在信心水準 95%的區間界限[2]。

表十四 隨機模擬檢驗的結果

寄件者代號	通訊錄人數	10%	15%	20%	25%	50%	75%	100%
User130	51	1	1	1	1	1	略	略
User1688	22	0.94(±0.035)	0.97(±0.032)	0.99(±0.009)	1	1	1	1
User1921	70	1	1	1	1	略	略	略
User2776	29	0.91(±0.039)	0.99(±0.019)	1	1	1	1	略
User3101	34	0.54(±0.096)	0.82(±0.074)	0.88(±0.062)	0.93(±0.049)	1	1	略
User3543	29	0.86(±0.059)	1	1	1	1	1	略
User744	22	1	1	1	1	1	1	1
User88	31	0.77(±0.081)	0.85(±0.07)	0.94(±0.046)	0.95(±)0.041	1	1	略

註:小括號中表在 95%信心水準下的區間值

模擬檢驗的結果相當理想，在中大量寄送的情形下，都能完全偵測出來，在小量與微量的情形下效果也相當理想，其中有些欄位的資料略，是因為我們為了挑選的原始資料不會含有病毒行為資料，設定最大寄件量不大於 25。

郵件病毒若想要達到快速擴張，其出生率不能太低，寄件量就必須大，如現在大部份的郵件病毒，行為相對容易曝露，若為了提高隱敝性，降低出生率，其擴張速度相對變慢低，威脅較弱，可以有較長的因應時間，也就不那麼危急。整體而言，以我們的方法不管是低中高量的寄件量，都能得到相當好的結果。

4.7 綜合討論

群組關係分析的階段所取得是從歷史的寄件行為的最大集合，這樣的關係仍然可能太過模糊(fuzzy)，包容許多未曾實際發生的行為，當寄件者的寄件行為資料總數不足時(訓練資料)，這樣的方法可以包容可能發生的寄件行為但目前尚未發生，當寄件行為的資料總數足夠的情形下，為了更精確的把握其間所存在的關係，我們可以將其提昇至依存關係的層次，更精確掌握寄件者的行為特性，但當資料量不足時提昇至依存關係的層次，反而會因為資料量不足推導出不正確的依存關係，誤將正常的行為當做異常行為，在處理上必須小心。

當寄件者的寄件行為是屬於單一收件者類型，也就是收件對象只有一位時，只需做到群組關係的建立即可，因為依存關係的建立是針對多收件者才需要，在找出同一群組內收件者間所存在的關係，因此依存關係的建立主要是針對多收件者的群組進行的。

我們處理的資料是 mail Client 與郵件伺服器交換的資訊，這些資訊是維持郵件系統能夠運作的必要資訊，否則無法完成郵件行為，難以造假，且資料量相當少，能保有快速的處理，對於郵件內容的資訊並不涉入，兼顧寄件者的隱私權。

五.結論

寄件者行為規範的建立，包含兩個階段群組關係與依存關係的建立，由資料的完整性程度，我們可以彈性決定要建立至群組關係的階段就好或是要到依存關係的階段。為了能夠快速有效重新建立行為規範，我們導入了行為特徵歸納表與時間門檻值的設定，濾除老舊不再發生的行為，縮減行為規範的大小，這些特性都有助於用來做為異常郵件的即時偵測。

目前資料的分析雖然是以離線(off line)的方式進行，衡量所產生的寄件者行為規範的方法與資料量的大小，所產生的群組數目與依存法則，我們相信本方法可以達到即時偵測的目標，如果與郵件伺服器整合在一起，可以在初期偵測到異常郵件的行為，以明尼蘇達大學為例，其郵件伺服器有安裝一套電子郵件病毒檢查系統，採用病毒特徵碼的比對方式，檢查郵件的內容是否包含病毒特徵碼，若是則會在郵件前加一些訊息告知收件者這是一封疑似包含"某某"病毒的郵件請小心，並且也會回一電子郵件告知寄件者已經中毒的訊息，這樣系統就算誤判也不會真的把郵件刪除，對於真的病毒，對寄件者與收件者而言，也已經達到警告的作用[4]。

從病毒的作用速度來看，傳統防毒軟體的防毒方式已經緩不濟急，我們需要一套整合的自動化機制，自動在第一時間發現問題即時處理，這樣的系統，前端可以結合我們的方法，偵測寄件者的異常寄件行為，發現問題，後端可以建立一套病毒的確認與清除的機制[7][22]，確認病毒的類型、作用方法與解決方式等，也許這樣的方式要完全自動化，仍然存在困難，但至少能縮短防治的時間，降低損失與傷害。

參考文獻

[1] "多層次的防毒專案", http://www.savetime.com.tw/web/symantec/submain.htm

[2] 黃文隆,抽樣方法,滄海書局,民國 88 年

[3] "ILOVEYOU", Wormhttp://www.sans.org/infosecFAQ/malicious/iloveyou_worm2.htm

[4] Automatic Virus Checks, http://www1.umn.edu/oit/newsletter/01/0601_itn/virus.html

[5] B. Le Charlier, A. Mounji and Morton Swimmer, "Dynamic detection and Classification of Computer viruses general behaviour patterns", Proceedings of Fifth International Virus Bulletin Conference., Sep 1995

[6] David Chess, "The future of viruses on the Internet", Virus Bulletin Intl. Conference in San Francisco, California, October 1-3, 1997 , access from http://www.research.ibm.com/antivirus/SciPapers/Chess/Future.html

[7] David M. Chess, Virus Verification and Removal Tools and Techniques, http://www.research.ibm.com/antivirus/SciPapers/Chess/CHESS3/chess3.html

[8] Emilie Lundin and Erland Jonsson, "Privacy vs Intrusion Detection Analysis", Proceedings of the 2nd International Workshop on Recent Advances in Intrusion Detection (RAID'99), September 7-9, 1999.

[9] G. Tesauro, J. O. Kephart and G. B. Sorkin, Neural Network for Computer Virus Recognition , IEEE expert, Vol 11, No 4, Aug 1996, pp5-6

[10] Intrusion Detection FAQ, http://www.sans.org/newlook/resources/IDFAQ

[11] Melissa, http://securityresponse.symantec.com/avcen

ter/venc/data/w97.melissa.a.html

[12] Nimda, http://www.symantec.com/avcenter/venc/data/w32.nimda.a@mm.html

[13] Ruby L. Kennedy, B.Van Roy, C. D. Reed and R. P. Lippmann, Solving Data Mining Problems through Pattern Recognition, Prentice Hall,1998

[14] S. Forrest, S. A. Hofmeyr, A. Somayaji and Thomas A. Longstaff, "A Sense of Self for Unix Processes", Proceedinges of the 1996 IEEE Symposium on Research in Security and Privacy, pp120-128

[15] Sircam, http://securityresponse.symantec.com/avcenter/venc/data/w32.sircam.worm@mm.html

[16] Srikant R. and Agrawal R., Mining Generalized Association Rules, Proceedings of the 21st international Conference on Very Large Data Bases, 1995, pp 407-419

[17] Stefan Axelsson, "On a difficulty of Intrusion Detection", 2nd Intl. Workshop on Recent Advances in Intrusion Detection (RAID'99), September 7-9, 1999,

[18] Steve R. White, "Open Problems in Computer Virus Research", Virus Bulletin Conference, Munich, Germany, October 1998, access from http://www.research.ibm.com/antivirus/SciPapers/White/Problems/Problems.html

[19] Steve R. White, Jeffrey O. Kephart and David M. Chess ,"Computer Virus: A Global Perspective", proceeding of the 5th Virus Bulletin International Conference, September 20-22, 1995 http://www.research.ibm.com/antivirus/SciPapers/White/VB95/vb95.distrib.html

[20] Steve R. White, Morton Swimmer, Edward J. Pring et al, "Anatomy of a Commercial-Grade Immune System", http://www.research.ibm.com/antivirus/SciPapers/White/Anatomy.html

[21] Terran Lane and Carla E. Brodley, "Temporal sequence learning and data reduction for anomaly detection", Proceedings of the 5th Conference on computer & Communications Security, pp 150~158, San Francisco, CA, USA, Nov 2-5, 1998, ACM

[22] The Digital Immune System, http://www.symantec.com/avcenter/reference/dis.tech.brief.pdf

[23] Understanding Heuristics: Symantec's Bloodhound Technology, http://www.symantec.com/avcenter/reference/heuristc.pdf

[24] VBS Worm Generator, http://www.sans.org/infosecFAQ/malicious/VBS_worms.htm