

公 告 本

101年5月10日修正本

第 93/01775 號專利案101年5月修

發明專利說明書

※ 申請案號：93101775

※ 申請日期：93年1月27日

※IPC 分類：H04L12/58 (2006.01)

一、發明名稱：(中文/英文)

可適性垃圾訊息過濾系統與方法

ADAPTIVE JUNK MESSAGE FILTERING SYSTEM AND METHOD

二、申請人：(共1人)

姓名或名稱：(中文/英文)

美商・微軟公司

Microsoft Corporation

代表人：(中文/英文)

艾莘那諾爾 D 巴特萊

EPPENAUER, D. BARTLEY

住居所或營業所地址：(中文/英文)

美國華盛頓州列德蒙微軟路1號

One Microsoft Way, Building 8, Redmond, WA 98052-6399, USA

國籍：(中文/英文)

美國/USA

三、發明人：(共5人)

姓 名：(中文/英文)

1. 羅斯威特羅伯特 L/ROUNTHWAITE, ROBERT L.

2. 古德曼喬休爾 T/GODMAN, JOSHUA T.

3. 海克曼大衛 E/HECKERMAN, DAVID E.

4. 普拉特約翰 C/PLATT, JOHN C.

5. 凱迪卡爾 M/KADIE, CARL M.

住居所地址：(中文/英文)

1. 美國華盛頓州佛爾市東南 287 大街 4148 號

4148 287th Avenue SE, Fall City, Washington 98024, U.S.A.

2. 美國華盛頓州瑞蒙市東北 38 街 17424 號

17424 NE 38TH Street, Redmond, Washington 98052, U.S.A.

3. 美國華盛頓州貝拉弗市東北薩米希巷西湖 648 號

648 W. Lake Sammamish Lane NE, Bellevue, Washington 98008, U.S.A.

4. 美國華盛頓州瑞蒙市東北 166 庭園 4963 號

4963 166th Ct NE, Redmond, Washington 98052, U.S.A.

5. 美國華盛頓州貝拉弗市東北 1 街 15937 號

15937 NE 1st Street, Bellevue, Washington 98008, U.S.A.

國籍：(中文/英文)

1. 美國/USA

2. 美國/USA

3. 美國/USA

4. 美國/USA

5. 美國/USA

四、聲明事項：

◎本案申請前已向下列國家(地區)申請專利 V 主張國際優先權：

【格式請依：受理國家(地區)；申請日；申請案號數 順序註記】

美國；2003 年 2 月 25 日；10/374,005

五、中文發明摘要：

本發明係關於一種用於過濾訊息的系統，該系統包括一種子過濾器 (seed filter)，該種子過濾器具有與該種子過濾器相關聯之一誤正率 (false positive rate) 及一誤負率 (false negative rate)。該新的過濾器亦被提供來過濾訊息，係根據該種子過濾器的誤正率及誤負率來評估該新的過濾器，而用來決定該種子過濾器的誤正率及誤負率的資料亦被用來決定該新的過濾器之一新的誤正率及誤負率，以作為閾限的一函數。如果對於該新的過濾器而言存在一閾限而使得該新的誤正率及新的誤負率一起考量時優於該種子過濾器之誤正率及誤負率的話，則用該新的過濾器取代該種子過濾器。

六、英文發明摘要：

The invention relates to a system for filtering messages – the system includes a seed filter having associated therewith a false positive rate and a false negative rate. A new filter is also provided for filtering the messages, the new filter is evaluated according to the false positive rate and the false negative rate of the seed filter, the data used to determine the false positive rate and the false negative rate of the seed filter are utilized to determine a new false positive rate and a new false negative rate of the new filter as a function of the threshold. The new filter is employed in lieu of the seed filter if a threshold exists for the new filter such that the new false positive rate and new false negative rate are together considered better than the false positive and the false negative rate of the seed filter.

七、指定代表圖：

(一)、本案指定代表圖為：第 1 圖。

(二)、本代表圖之元件代表符號簡單說明：

100	垃圾訊息偵測系統	102	訊息
104	過濾器控制器構件	106	第一(種子)過濾器
108	第二(新的)過濾器	112	收件匣
114	使用者更正元件		

八、本案若有化學式時，請揭示最能顯示發明特徵的化學式：

九、發明說明：

【發明所屬之技術領域】

本發明係關於辨識出不想要的資訊(如，垃圾郵件)的系統及方法，及更特定地係關於可促進此一辨識功能之可適性過濾器。

【先前技術】

全球通信網路(如網際網路)的來臨提供了接觸大量潛在客戶的商業機會。電子訊息，特別是電子郵件(“e-mail”)變得已逐漸普遍成為將所不想要的廣告及推銷(亦被稱為“垃圾信(spam)”)散布給網路使用者的一種方法。

Radicati Group 公司(一家顧問及市場調查公司)在 2002 年八月所作的評估中指出每一天約有兩十億封垃圾郵件被寄發，而此數字預計每兩年成長三倍。個人及實體(如，公司，政府機關等等)逐漸變得不便且經常受到垃圾郵件的侵犯。因此，垃圾電子郵件現在已成為或即將成為可信計算的一項主要的威脅。

一種用來阻礙垃圾電子郵件的關鍵技術為使用過濾系統/方法。一種經過證實的過濾技術是根據一種機器學習方式，機器學習式過濾器將訊息是垃圾的或然率指定給一進來的訊息。在此方式中，特徵典型地從兩類別之示例訊息

(如，垃圾及非垃圾訊息)中被擷取出，且一學習過濾器被用來在這兩類訊息之間作或然率的區別。既然許多訊息特徵與內容(如，該訊息的主題及/或主體內的文字及文詞)相關，此類過濾器被通稱為”以內容為基礎的過濾器”。

某些垃圾/垃圾信過濾器是可適性的，對於需要一種可適合特殊之需求的多語使用者及說稀有語言的使用者而言是很重要的。更進一步地，並不是所有的使用者都同意何種是垃圾/垃圾信又何種不是垃圾/垃圾信。因此，藉由一種可隱含地訓練(如，經由觀察使用的行為)的過濾器，各別的過濾器可動態地整修以符合一使用者特殊的訊息辨識需求。

用於過濾的可適性的方法係要求使用者將訊息標示為垃圾或非垃圾。很不幸地，此種密集的手動式訓練技術對於許多使用者而言由於複雜性而讓該方式成為所不想要的，該複雜性係關聯於為了達到適當的訓練效果所需的時間量。另一種可適性過濾器訓練方式為使用隱含的訓練提示。例如，如果使用者回覆或轉寄一訊息，則該方法假設該訊息為非垃圾。然而，只使用此種訊息提示會將統計上的偏見導入此訓練過程中，而導致過濾器具有較低之精確性。

另一種方法為，使用所有使用者的電子郵件來訓練，

其中最初的標籤是由一既有的過濾器來指定，使用者有時可用顯性的(explicit)提示(如，“使用者更正”方法)來覆蓋 override)這些指定--例如，選取“當作垃圾刪除”及“非垃圾”的選項一及/或用隱含的提示來覆蓋這些指定。雖然此種方法比前述的技術要來得好一些，但仍比本案所揭示及所請求的發明來得差。

【發明內容】

下文所呈現的是本發明的一簡化的概要，係用以提供本發明的某些態樣的基本瞭解。此概要並非本發明的一廣泛性的綜覽。此概要並不是要指出本發明的關鍵/主要的元件或是要用來描繪本發明的範圍。此概要的唯一目的是要以一簡化的形式來呈現本發明的某些概念，詳細的說明將在稍後參照附圖加以說明。

本發明提供一種系統及方法來促進一可用的過濾器(如，種子過濾器或新的過濾器)的使用，該可用的過濾器最適合辨識出垃圾/垃圾信訊息。本發明利用一種子過濾器來過濾訊息，該種子過濾器具有與該種子過濾器相關聯的一誤正率(如，非垃圾郵件被錯誤地歸類為垃圾)及一誤負率(如，垃圾郵件被錯誤地歸類為非垃圾郵件)。一新的過濾器亦被用來過濾訊息—該新的過濾器係根據與該種子過

濾器相關聯的誤正率及誤負率來加以評估的。用來決定該種子過濾器的誤正率及誤負率的資料用來決定該新的過濾器之一新的誤正率及誤負率為閾限的一函數關係。

對於該新的過濾器而言，如果存在一閾限使得該新的誤正率及新的誤負率一起被考量時優於該種子過濾器之誤正率及誤負率，則用該新的過濾器取代該種子過濾器。根據被使用者標示為垃圾及非垃圾的訊息(如，經由使用者更正處理)來決定該新的誤正率及新的誤負率。該使用者更正處理包括覆蓋該訊息之最初的分類，當使用者接收到該訊息時，藉由該種子過濾器自動實施該最初的分類。該閾限可以是一單一閾限值，或是從複數個產生的閾限值中選取。如果使用複數個值，則可藉由選取有效的閾限值的範圍(如，具有最低誤正率之閾限值，或可將根據一 p^* 實用函數之使用者預期效用最大化的閾限值)之內的一中位閾限值來決定選取的閾限值。或者，只有在該新的過濾器的誤正率及誤負率至少與該種子過濾器在選取的閾限值的誤正率及誤負率一樣好時才會選取該閾限值，且該新的過濾器的誤正率及誤負率之中的一個是較佳的。此外，可提供選擇準則以使得只有在新的過濾器的過濾率不只在該選取的閾限值處，更在附近的閾限值處都優於種子過濾器的過濾率時，才會選取該新的過濾器。

本發明的其它態樣提供一種圖形式的使用者介面，該圖形式的使用者介面可促進資料的過濾。該界面提供一可與一配置系統相溝通的過濾器界面，該配置系統與配置一過濾器有關。該介面提供複數個使用者可選擇的過濾器階層，該等使用者可選擇的過濾器階層包括有內定，加強，及獨家中之至少一者。該介面提供許多可用來實施本發明之上述系統及方法的工具。

為了要完成上述及相關的目的，本發明之某些舉例性態樣在本文中藉由參照附圖來加以說明。然而，這些態樣為可運用本發明的原理來實施的一些例子，且本發明包含這些態樣及等效物。本發明之其它的優點及新穎的特徵從以下參照附圖之本發明的詳細說明中將會變得很明顯。

【實施方式】

本發明現將參照附圖來加以說明，其中在所有圖中相同的標號被用來表示相同的元件。在下面的說明中，為了說明的目的，有許多特定的細節被敘述用以提供對本發明之徹底的瞭解。然而，可在沒有這些特定的細節下操作本發明。在其它的例子中，習知的結構及裝置係以方塊圖的形式來表示以便於描述本發明。

在本文中所使用之「構件」及「系統」等詞係指一與

電腦相關的實體，為硬體，硬體與軟體的組合，軟體，或執行中的軟體。例如，一構件可以是，但並不侷限於，在一處理器上運行的處理，一處理器，一物件，一可執行的程式，一執行緒，一程式，及/或一電腦。舉例而言，在一伺服器上運行的應用程式及該伺服器兩者都可以是一構件。一或多個構件可位在一處理及/或執行緒內，且一構件可位在一構件上及/或分配在兩個或更多個構件之間。

本發明可將不同的干擾方案及/或與垃圾訊息過濾相關的技術相併合。在本文中所用之”干擾”一詞係指該系統，環境，及/或使用者從一組由事件及/或資料所攔截到的觀察而實施的推論處理或干擾狀態。舉例而言，干擾可被用來辨識一特定的上下文或動作，或可產生狀態的或然率分布。該干擾可以是或然性的，亦即，根據資料或事件的考量所進行之重要性狀態的或然率計算。干擾亦可指用來從一組事件及/或資料構成更高階事件的技術。此干擾的結果為從一組被觀察到的事件及/或被儲存的事件資料來建構新的事件或動作，不論事件是否是以短暫親近的關係相關連，及不論事件與資料是來自一或數個事件及資料源。

應被瞭解的是，雖然訊息一詞在整個說明書中被經常使用到，此用詞並不侷限於電子郵件本身，而是包括了可透過任何適當的通信架構散布之任何形式的電子訊息。例

如，因為不想要的文字會在使用者交換訊息時被電子地散置於正常的聊天訊息中，及/或如一前導訊息，一結束訊息般地被插入，促進兩個或多個人之間的會議進行之會議應用程式(如，互動式聊天程式，及立即傳訊程式)亦可運用本文所揭示之過濾的好處。在此特殊的應用中，一過濾器可配置成能夠自動地過濾特定的訊息內容(文字及影像)用以攔截不想要的內容(如廣告，推銷，或宣傳)並貼上垃圾標籤。

現參照第1圖，圖中顯示出依據本發明之一垃圾訊息偵測系統100。該系統100接收一串進來的訊息102，該串進來的訊息102被加以過濾用以促進垃圾訊息的偵測與移除。訊息102被收入到一過濾器控制器104，根據本發明的可適性態樣所決定的過濾規範，該過濾器控制器104，可將訊息102繞經一第一過濾器106(如，種子過濾器)與一第二過濾器108(如，新的過濾器)之間。因此，如果第一過濾器106在偵測垃圾訊息上是充分有效的，則第二過濾器108將不會被使用，且該過濾器控制器104將會持續將訊息102繞經該第一過濾器106。然而，如果第二過濾器108被認定為至少與第一過濾器106一樣有效的話，則該過濾器控制器104可決定將訊息102繞經該第二過濾器108。用來作出此一認定的規範將在下文中詳細說明。當最

初被使用時，該過濾器系統 100 可配置為一預設內定過濾器設定，使得訊息 102 會被繞經第一過濾器 106 以進行過濾(典型如，第一過濾器 106 為被明確地訓練的種子過濾器，並與與一特定產品一起交貨)。

根據第一過濾器 106 的設定，被接收至第一過濾器 106 中的訊息將會被詢問是否有與垃圾資料相關的垃圾資訊。該垃圾資訊可包括，但不侷限於，以下所列各者：傳送者資訊(來自於已知寄送垃圾郵件的寄送者)，像是 IP 地址，傳送者名稱，傳送者 e-mail 地址，傳送者網域名稱，及在識別子欄位中之難理解的文字與數字串；經常使用在垃圾郵件中之訊息本文用字與用詞，像是”貸款”，”性愛”，”利率”，”限量供應”，”立即購買”，等等；訊息本文特徵，像是字型大小，字型顏色，特殊字母使用；及彈出式廣告之嵌入式鏈結。垃圾資料可至少部分地根據預設且動態決定垃圾之規範來加以決定。該訊息亦被詢問是否為”好的”資訊，像是典型地不會出現在垃圾郵件中的字眼，如”天氣”及”隊伍”，或來自於已被認知為只會傳送好的郵件的傳送者或傳送者 IP。應瞭解，如果該產品最初沒有與一種子過濾器一起交貨，且沒有任何建立的過濾規範，則所有的訊息都會未加標記地通過第一過濾器 106 而進入使用者的收件匣 112(亦被稱為第一過濾器輸出)。應被瞭解的是，收件

匣 112 可以只是位在許多位置(如，伺服器，大量儲存單元，客戶端電腦，分散式網路，...)的資料儲存。又，應被瞭解的是，第一過濾器 106 及/或第二過濾器 108 可被複數個使用者/構件所使用，且該收件匣 112 可被分割用以將各別使用者/構件的訊息分開來儲存。另，該系統 100 可使用複數個第二過濾器 108，以使得一個最適當的第二過濾器被使用在一特定的工作中。本發明的這些態樣將於下文中詳細說明。

當使用者檢閱過信箱中的訊息之後，某些訊息將會被決定為垃圾，其它的則不會。這一部分是根據使用者顯性地標記(如，按下按鈕)為垃圾郵件或非垃圾郵件，及經由使用者對特定的訊息所作的動作來隱含地標記該訊息。一訊息可根據下列的使用者動作或訊息處理而被隱含地決定為非垃圾：該訊息已被閱讀且仍被留在收件匣中；該信件已被閱讀且被轉寄；該信件已被閱讀且被放在除了垃圾檔案夾之外的其它檔案夾中；該訊息被回覆；或使用者開啟並編輯該訊息。使用者的其它動作亦可被定義為與非垃圾訊息相關聯。一訊息可根據以下的現象而被隱含地決定為垃圾，如該訊息一個星期沒有被閱讀，或沒閱讀該訊息即刪除該訊息。因此，系統 100 經由一使用者更正元件 114 來監視這些使用者動作(或訊息處理)。這些使用者動作或

訊息處理可被預先配置在使用者更正元件 114 中，使得使用者最初在檢閱並對訊息實施動作時，系統 100 即可開始發展出第一過濾器 106 的誤正率及誤負率。沒有被實質上預先配置至該使用者更正元件 114 中的任何使用者動作（或訊息處理）都將自動地允許該”未知的”之未標記訊息通過以到達過濾器輸出 112，直到系統 100 被調適來應付此訊息類型為止。應被瞭解的是，本文中所用的”使用者”一詞是要包括：人類，一群人類，一構件以及人類與構件的組合。

當在使用者收信匣 112 中的一訊息接收以作為未標記訊息，但該訊息實際上是一垃圾訊息時，系統 100 會將該訊息當作一誤負資料值來處理。然後，該使用者更正元件 114 將此誤負資訊回授給該過濾器控制器 104，以當作用來查明該第一過濾器 106 的有效性的資料值。在另一方面，如果該第一過濾器 106 將一訊息標記為垃圾郵件，而該訊息事實上並非一垃圾訊息時，系統 100 會將該訊息當作一誤正資料值來處理。然後，該使用者更正元件 114 將此誤正資訊回授給該過濾器控制 104，以當作用來確定該第一過濾器 106 的有效性的資料值。因此，隨著使用者更正在使用者收信匣 112 中被接收的訊息，亦發展該第一過濾器 106 的誤正及誤負資料。

系統 100 決定第二過濾器 108 是否存在一閾限，以使得該第二過濾器的誤正及誤負率都比第一過濾器 106 的誤正及誤負率低(如，在一可接受的或然率的範圍內)。如果是如此的話，則系統 100 會選取一個可接受的閾限。該系統亦可在誤負率一樣好，但誤正率較佳時，或在誤負率一樣好但誤正率較佳時選取第二過濾器。因此，本發明決定該第二過濾器 108 是否存在閾限(及該閾限應為何)，以保證在一可接受的或然率的範圍內，該第二過濾器可提供有關垃圾偵測之相等或更佳的實用性，無論一特定使用者的實用功能及該使用者是否已無誤地更正第一過濾器 106 的錯誤。

有鑑於誤正及誤負識別的使用者確認，系統 100 根據對於新的訓練的需求來訓練新的(或第二)過濾器 108。詳言之，系統 100 使用經由使用者更正方法決定以標記垃圾及非垃圾的資料。藉由使用此資料，決定該第一(既有或種子)過濾器 106 的誤正(如，被錯誤地標示為垃圾之非垃圾訊息)率及誤負(如，被錯誤地標示為非垃圾之垃圾訊息)率。相同的資料被用來學習(或”訓練”)該新的(如，第二)過濾器 108，該資料亦被使用在與該第二過濾器的誤正及誤負率為一閾限的函數有關的決定上。既然評估資料與訓練第二過濾器所用的資料相同，所以最好是使用一交叉確

認方法（交叉確認是一種熟習此技藝者所習知的技術），這將於下文中被詳細說明。如果第二組資料被決定為至少與第一組一樣好的話，則啟用第二過濾器 108。該控制構件 104 接著將所有進來的訊息繞經該第二過濾器 108，直到比率比較處理因為第一過濾器具有較佳的過濾實用性而決定過濾應被移回到第一過濾器 106 為止。

本發明的一特殊的態樣依賴兩項保證。第一項保證為第一確認(如，使用者更正)不會有錯誤(如，使用者不會將非垃圾訊息當作垃圾訊息加以刪除)。在此保證之下，資料標籤雖然不是永遠正確，但”至少”與該第一過濾器 106 所指定的標籤一樣正確。因此，根據此等標籤，如果第二過濾器 108 的實用性不低於既有過濾器的話，該第二過濾器 108 之真實的預期實用性不會比第一過濾器 106 的實性性差。第二項保證為，降低誤正及誤負率是所想要的。有關此項保證，如果第二過濾器 108 的兩個錯誤率都不會比第一過濾器 106 的兩個錯誤率大的話，則第二過濾器 108 在垃圾偵測上至少與第一過濾器 106 一樣好，而與使用者的特定實用函數無關。

第二過濾器 108 不會永遠與第一過濾器 106 一樣有效的一個原因為，第二過濾器所依據的資料比第一過濾器 106 少。第一過濾器 106 可以是一”種子”過濾器，種子過

濾器具有由其它使用者的資料所產生的種子資料。如果不是全部也是絕大部分的可適性過濾器都是與一種子過濾器一起交貨，因而提供一過濾器配置給使用者，該過濾器配置可辨識出典型的垃圾電子郵件訊息且無需使用者來配置該過濾器，此舉提供了沒有經驗的電腦使用者一良好的“即開即用 (out-of-the-box)”的經驗。第二過濾器 108 不會永遠與第一過濾器 106 一樣有效的另一個原因為，第二過濾器 108 更為敏感。這與兩個因子有關：過濾器是不完美的，及無法被校準。這兩個因子都將依次被討論，然後將再回到決定第二過濾器 108 是否較佳的議題上。

現參照第 2 圖，第 2 圖圖示性能取捨 (tradeoff) 與攔截率 (被正確標示的垃圾信的百分比，等於 1 減掉誤負率) 及誤正率 (標示為垃圾的非垃圾訊息的百分比) 之間的關係的圖表。如在本文中所顯示的，且將為熟習此技藝者所瞭解的，沒有過濾器是完美的。因此，在辨識並攔截更多的垃圾訊息與意外地將非垃圾訊息誤標示為垃圾之間存在著取捨。此性能取捨 (在本文中亦被稱為正確率) 係描述為習知之接收者 - 操作者曲線 (ROC) 200。在該曲線上的每一點對應於一不同的取捨。一使用者藉由調整一或然率閾限來為一過濾器選取一“操作點”，或該或然率閾限可被預設。當一訊息為垃圾 (被過濾器所認定) 的或然率 p 超過此閾限

時，該訊息即被標示為垃圾。因此，如果使用者決定要在一高正確率的體系下操作的話(如，誤正數目與正確標示的訊息數比較起來很低)，則在曲線 200 上的操作點會靠近原點。例如，如果使用者選取 ROC 曲線 200 上的操作點 A 的話，則誤正率約為 0.0007 且代表正確標示的訊息數之相應的 y 軸值約為 0.45。該使用者將會約有一 $0.45/0.0007=643$ 的過濾器正確率，亦即，大約每六百四十三個正確標示的訊息會有一個誤正訊息。在另一方面，如果操作點為點 B 的話，計算出之較低的正確率約為 $0.72/0.01=72$ ，亦即，大約每七十二個正確標示的訊息會有一個誤正訊息。

因為語言的決定理論，不同的人具有不同的實用函數來過濾垃圾訊息，不同的使用者將會依據他們各自獨特的偏好來作出取捨。例如，有一類別的使用者可能對於不正確地標示一非垃圾訊息及無法攔截 N 個垃圾訊息毫不在意。對於此類別使用者而言，用於垃圾之最佳的或然率閾限(p^*)可用以下的關係式來定義：

$$p^* = \frac{N}{(N+1)}$$

其中 N 為訊息數，且 N 可隨著不同類別的使用者而改變。

因此，此類別的使用者被說成是具有” p^* 實用函數”。在此一瞭解之下，如果一使用者具有一 p^* 實用函數且第二

過濾器被校準，則一最佳閾限可被自動地選取，亦即，該閾限應被設定為 p^* 。另一類別的使用者可能想要讓他或她的非垃圾訊息只有不大於 $X\%$ 的比例被標示為垃圾。對於該等使用者而言，最佳閾限係依附於第二過濾器 108 指定給訊息的或然率分布。

第二註記係為過濾器可以或不可以被校準。經過校準的過濾器具有的特性係為，當決定一組電子郵件訊息是垃圾的或然率為 p 時，則這些訊息中的 p 個即為垃圾。許多機器學習方法會產生經過校準的過濾器，並使使用者教條式地更正既有過濾器的錯誤。如果使用者只有在有些時候(如，少於 80%)才更正錯誤的話，則過濾器將不再是經過校準的過濾器，亦即，關於不正確的標籤，這些過濾器將是經過校準的，但關於正確的標籤，則是未校準的。另一方面，本發明提供一種決定第二過濾器 108 是否存在閾限(及該閾限應為何)的方式，並可保證(在某些或然率內)第二過濾器 108 供應與第一過濾器 106 相同或更佳的實用性，無論該使用者的實用函數為何，及該使用者是否已無誤地更正第一過濾器 106 的錯誤。

現參照第 3 圖，第 3 圖圖示依據本發明的一態樣的處理的流程圖。為了簡化說明的目的，該方法被顯示及描述為一連串的動作，應被瞭解及認知的是，本發明並不侷限

於動作的順序，依據本發明，有些動作可以不同的順序發生及/或與其它的動作同時發生。例如，熟習此技藝者將可瞭解及認知的是，一方法可用一連串互相關聯的狀態或事件（如在一狀態圖中者）來代表。又，根據本發明，並非需要所有圖示的動作以實施該方法。

基本的方法依賴兩個假設。一個假設為，使用者更正不包含錯誤（一個錯誤例為使用者將一非垃圾的訊息當作垃圾刪除掉）。在此一假設下，該資料上的標籤雖然不是永遠都是正確的，但”至少”是跟第一/種子過濾器所指定的標籤一樣正確。因此，根據這些標籤，如果第二過濾器所具有的實用性不會比第一過濾器的實用性低的話，第二過濾器的真正被預期的實用性並不會比第一過濾器差。第二個假設為，所有使用者都喜好低的誤正率及低的誤負率。在此假設下，如果第二過濾器的兩個錯誤率都不會比第一過濾器的兩個錯誤率高的話，則第二過濾器即不會比第一過濾器差，而不論使用者的特定實用函數為何。

在 300，提供一介面構件至第一及第二過濾器（如，用來改變設定，及控制過濾器的安排及配置）。在 302，第一過濾器被配置成可根據一或多個過濾器設定自動地過濾進來的訊息。這些設定可包括由製造商所提供之內定的設定。一旦接收經過濾的訊息（如，進入到收件匣中），在 304

該等訊息被檢閱並作出(如，藉由使用者更正方法)哪些非垃圾的訊息被錯誤地標記為垃圾(如，誤正訊息)及哪些垃圾訊息被錯誤地標記為非垃圾(如，誤負訊息)的決定。在 306，可藉由顯性地或隱含地將誤負訊息標記為垃圾郵件，及去除誤正訊息之標籤以標記為非垃圾來實施使用者更正功能。此使用者更正功能藉由決定第一過濾器之誤正及誤負率資料來提供一正確率給第一過濾器。在 308，第二過濾器依據該第一過濾器 106 之使用者更正過的資料來加以訓練。在 310，相同的資料被用來決定第二過濾器的誤正及誤負率為一閾限的函數關係。在 312，該閾限被決定。作出關於第二過濾器是否存在一閾限的決定，以使得與第二過濾器相關的誤正率及誤負率低於第一過濾器的誤正率及誤負率(在某些合理的或然率範圍內)。亦即，在 314 會決定第二過濾器的正確率($Accuracy_{SF}$)是否優於第一過濾器的正確率($Accuracy_{FF}$)。如果是的話，則該適當的閾限被選取且該第二過濾器被用來過濾進來的訊息，如 316 所示。如果不是的話，則該處理前進至 318，其中該第一過濾器被保留以實施訊息的過濾。在必要時，該處理可動態地循環前述的動作。

正確性分析處理可在每一次使用者更正功能發生時，使得第二過濾器可在任何時間根據閾限的決定而被使用或

被解除作用。因為第一過濾器的評估資料與用來訓練第二過濾器的資料相同，所以一交叉確認方法被使用。亦即，資料被分段成 k 個區段 (k 為整數) 以用於每一使用者更正處理，且對於每一區段而言，第二過濾器係使用其它 $k-1$ 個區段內的資料來訓練。第二過濾器的性能(或正確性)係針對從該 $k-1$ 個區段中被選取的區段來加以評估。另一種可能為等待，直到具有垃圾及非垃圾標籤的訊息分別被累積到 N_1 及 N_2 的數目(如， $N_1=N_2=1000$)，並在每次額外的垃圾及非垃圾訊息被累積到 N_3 及 N_4 個訊息時(如， $N_3=N_4=100$)再實施一遍。另一種方式為根據行事曆時間來安排此一處理。

如果有多於一個的閾限值可讓第二過濾器不會比第一過濾器差的話，則選取哪一個閾限值來使用存在多種可能。其中一種可能係為選取在使用者具有一 P^* 實用函數的假設下，可讓使用者的預期實用性最大化的閾限。另一種可能係為選取具有最低的誤正率的閾限。再另一種可能係為選取合格閾限值範圍內的中間點。

關於在經測量的錯誤率中的不確定性，設 k_1 及 k_2 為分別來自於第一及第二過濾器之誤標示為非垃圾(或垃圾)的錯誤數目。顯示一簡單的統計分析，如果：

$$k_1 - k_2 \geq f\sqrt{(k_1 + k_2)},$$

則可假定約可 $x\%$ 確定第二過濾器的錯誤率不會比第一過濾器差 (如，當 $f=2$ ， $x=97.5$ ；當 $f=0$ ， $x=50$)。保守一些，如果 k_1 或 k_2 等於 0 的話，則 1 值應被使用在該平方根 (sqrt) 項中。應注意的是， x 為一保守性的調整，當 x 接近 100 時，第二過濾器的確定性必定優於第一過濾器在使用第二過濾器之前的確定性。此確定性(或不確定性)計算包括假設介於第一過濾器與第二過濾器之間的錯誤是獨立的。避免此一假設的一種方法為評估共同錯誤的數目，亦即，在該獨立的假設下應有的錯誤數目。如果發現比此數目多出 k 個錯誤的話，則用在上述計算中的 (k_1-k) 及 (k_2-k) 來取代 k_1 及 k_2 。此外，隨著在訓練資料內的訊息數增加，則第二過濾器(在任何的閾限都)將更可能比第一過濾器正確。上述的不確定性評估忽略此“先前知識”。熟悉 Bayesian 機率學/統計學之熟習此技藝者將會瞭解到，存在著將此先前知識結合至不確定性評估中的方法。

在此基本方法的一個態樣中，想像一垃圾訊息被第一過濾器標示為非垃圾訊息。又，假設使用者並沒有更正此一錯誤，所以該系統內定將此訊息決定為非垃圾。具有更精確的訓練資料的第二過濾器可能將此訊息標示為垃圾。因此，第一過濾器的誤正率將被低估，而第二過濾器的誤正率則被高估。在一閾限處，許多垃圾訊息標示為非垃圾，

以保持低的誤正率，而因為大多數的垃圾電子郵件過濾器的操作是在該閾限處實施，而由此事實會擴大此效應。

有數種方法可被組合使用，以應付此基本方法的態樣。第一種方法係為假設使用者具有 p^* 實用函數（如， $N=20$ ），並在可找到讓第二過濾器不比第一過濾器差的閾限的任何時候佈署該第二過濾器。在此處，在第二過濾器的誤正率大於第一過濾器的誤正率時，可佈署第二過濾器。亦即，在此方法中，第二過濾器更可能被佈署。

第二種方法係為，限制測試組，使得標示為非垃圾的訊息以一極高的確定程度被確定為不是垃圾。例如，該測試組包括被使用者按下”非垃圾”按鈕而加以標記的訊息，被閱讀且沒有被刪除的訊息，被轉寄的訊息，及使用者已回覆的訊息。

第三種方法係為，該系統可使用由一經過校準的過濾器（如，第一過濾器）所產生的或然率來產生第二過濾器之誤正率的較佳預估。亦即，該系統可將每一正常（非垃圾）訊息的或然率（根據一經過校準的過濾器）加總起來，而不是簡單地計算在該資料中具有一非垃圾標籤的訊息數及具有來自第一過濾器的垃圾標籤的訊息數。此加總將小於該計數，且將會是比讓使用者更正所有的訊息的計數還要好的預估。

在一相較簡單的第四種方法中，使用者使用”非垃圾”及”垃圾”按鈕來更正標籤的預期次數被加以監視。在此處，預期係與一經過校準的過濾器有關(如，第一/種子過濾器)。如果實際的更正次數落在(絕對數字或百分比)預期的次數之下的話，則該系統並沒有訓練第二過濾器。

在使用時，使用者介面可提供數個閾限來讓使用者選擇。在此情況中，只有當在使用者所選定的閾限下，新的過濾器的性能表現優於種子過濾器時，才會佈署新的過濾器。此外，想要該新的過濾器在其它的閾限設定下，特別是在使用者目前的選擇值附近的設定亦優於該種子過濾器。下面的演算法則為可促進此法的方式。輸入一被稱為 SliderHalfLife(SHL) 的參數，該參數為具有一內定值為 0.25 的實數。針對每一閾限值來決定該新的過濾器是否優於第一過濾器或與第一過濾器一樣好。然後使用目前選取的閾限值。然而，如果新的過濾器在目前的閾限設定上優於第一/種子過濾器且如下文所說明的總權重值 (TotalWeight, w) 大於或等於 0，則切換新的過濾器。最初，總權重值 = 0。對於每一非目前的閾限設定而言：

根據每一非目前的閾限設定與目前的設定的距離來指定一權值

$$d = \text{abs} \left[\frac{(IS - ICS)}{(IMAX - IMIN)} \right]$$

d = 距 離

IS = 設 定 的 指 標

ICS = 目 前 設 定 的 指 標

$IMAX$ = 最 大 設 定 的 指 標

$IMIN$ = 最 小 設 定 的 指 標

$w = 0.5^{(d/SHL)}$

如 果 新 的 過 濾 器 在 此 設 定 下 比 較 好 的 話 ， 則 將 新 的 過 濾 器 之 權 值 加 至 總 權 重 值 ； 否 則 的 話 將 新 的 過 濾 器 之 權 值 從 總 權 重 值 中 減 掉 。

應 注意 的 是 ， 此 演 算 法 只 決 定 該 新 的 過 濾 器 在 每 一 閾 限 設 定 是 否 比 較 好 。 而 並 未 考 慮 第 二 過 濾 器 比 第 一 / 種 子 過 濾 器 好 多 少 或 差 多 少 。 該 演 算 法 可 使 用 以 下 的 功 能 來 加 以 修 改 用 以 考 慮 改 進 或 惡 化 的 程 度 : 新 及 舊 的 誤 負 率 , 誤 正 率 , 誤 負 數 及 / 或 誤 正 數 。

現 參 照 第 4a 圖 ， 第 4a 圖 圖 示 一 舉 例 性 的 使 用 者 介 面 400, 使 用 者 介 面 400 可 呈 現 给 使 用 者 進 行 本 文 中 所 揭 示 之 可 適 性 垃 圾 過 濾 器 系 統 及 使 用 者 電 子 信 箱 的 基 本 配 置 。 介 面 400 包 括 一 帶 有 選 單 列 402 之 垃 圾 郵 件 頁 面 (或 視 窗) 401, 該 標 題 列 包 括 但 不 僅 限 於 下 列 下 拉 式 選 單 標 頭 : 檔 案 , 編 輯 , 檢 視 , 登 出 , 及 說 明 & 設 定 。 視 窗 401 亦 包 括 一 鏈 結 列 404, 該 鏈 結 列 404 可 促 進 向 前 及 向 後 導 航 ，

以允許使用者導航至其它的頁面、工具、及介面 400 的能力，介面 400 的能力包括首頁、我的最愛、搜尋、郵件及其它、即時通訊、娛樂、理財、購物、人物&聊天、學習及相片。一選單列 406 可促進選取該垃圾電子郵件配置視窗 401 的一或多個配置視窗。如所圖示的，一設定子視窗 408 允許使用者選擇數個過濾垃圾電子郵件的基本配置選項。第一個選項 410 允許使用者能夠啟動垃圾電子郵件過濾。使用者亦可選擇不同等級的電子郵件保護。例如，第二選項 412 允許使用者選取一內定的過濾器設定，而該內定的過濾器設定只會攔截最明顯的垃圾郵件。第三選項 414 允許使用者選擇更多進階的過濾功能，使得更多的垃圾郵件被攔截並丟棄。第四選項 416 允許使用者可選取只從被信賴的一方，例如從列在使用者的通訊錄上及安全名單上的一方，收取郵件。一相關的設定區 418 提供一導航至這些表列區的構件，包括垃圾郵件過濾器，安全名單，郵寄名單，及封鎖寄件人名單。

現參照第 4b 圖，第 4b 圖圖示呈現使用者信箱外貌的該使用者介面 400 的一使用者信箱視窗 420。該信箱視窗 420 包括該選單列 402，該選單列 402 包括但不侷限於下列下拉式選單標頭：檔案，編輯，檢視，登出，及說明&設定。該信箱視窗 420 亦包括該鏈結列 404，鏈結列 404 可

促進向前及向後導航，以允許使用者導航至其它的頁面、工具、及介面 400 的能力，介面 400 的能力包括首頁、我的最愛、搜尋、郵件及其它、即時通訊、娛樂、理財、購物、人物&聊天、學習及相片。視窗 420 亦包括電子郵件控制工具列 422，電子郵件控制工具列 422 包括以下所列：允許使用者建立新的訊息的一寫訊息選項；用以刪除一訊息的一刪除選項；用以將一訊息標記為垃圾的一垃圾選項；用以回覆一訊息的一回覆選項；用以將一訊息移動至其它檔案夾的一放入檔案夾選項；及用以轉寄一訊息的一轉寄圖像。

視窗 420 亦包括一檔案夾選項子視窗 424，該檔案夾選項子視窗 424 提供顯示收信匣、垃圾筒、及垃圾郵件檔案夾的內容的選項給使用者。使用者亦可存取不同檔案夾內的內容，不同檔案夾內的內容包括被儲存的訊息、寄件匣、送出的訊息、垃圾筒、草稿匣、一示範程式及一舊的垃圾郵件檔案夾。在垃圾郵件及舊的垃圾郵件檔案夾內每一者的訊息數目亦被列在各自的檔案夾名稱旁邊。根據在檔案夾選項子視窗 424 內的檔案夾選項，在一訊息列子視窗 426 中呈現一列接收到的訊息。在一訊息預覽子視窗 428 中，呈現選取的訊息的部分內容給使用者看以進行預覽。視窗 420 可被修改用以包括使用者偏好的資訊，使用者偏

好的資訊係呈現在一使用者偏好子視窗(未示出)中。該偏好子視窗可被包括在所圖示的視窗 420 的右側的一部分內，如第 4a 圖所示。這包括但不侷限於天氣資訊、股票市場資訊、喜愛的網路連結等等。

圖示的介面 400 並不侷限在圖中所示者，而是可包括其它傳統的圖形、影像、說明性文字、選單選項等等，介面 400 可被實施用以進一步幫助使用者來進行過濾器選擇及導航至該介面的其它頁面。

現參照第 5 圖，第 5 圖圖示一利用所揭示的過濾技術的架構的示意方塊圖。一網路 500 被提供來促進電子郵件來回於一或多個客戶端 502, 504 及 506(亦被標記為客戶端 1, 客戶端 2, ..., 客戶端 N)之間的通信。網路 500 可以是一全球通信網路(GCN)(如網際網路)、或一 WAN(廣域網路)、LAN(地區網路)、或其它的網路架構。在此特定的應用例子中，一 SMTP(簡單郵件傳送通信協定)閘道伺服器 508 界接至該網路 500，以提供 SMTP 服務給一 LAN510。一可操作地設置在該 LAN510 上的電子郵件伺服器 512 界接至該閘道 508，以控制並處理客戶端 502, 504 及 506 的進來及出去的電子郵件，其中客戶端 502, 504 及 506 亦被設置在該 LAN510 上，以至少存取被提供於該 LAN510 上的郵件服務。

客戶端 502 包括一中央處理單元(CPU)514，CPU514 控制著客戶端處理，應被瞭解的是，CPU514 可包含多個處理器。CPU514 執行指令，這些指令與提供上述的一或多個過濾功能有關。該等指令包括，但不侷限於：至少執行上述的基本過濾方法之經過解碼的指令、讓使用者實施使用者更正以應付錯誤而被組合地使用之任何或所有方法、不確定性的決定、閾限的決定、使用誤正及誤負率資料的正確性計算、及使用者互動性選擇。一使用者介面 518 被提供來促進 CPU514 與客戶端作業系統之間的溝通，使得使用者可互動地配置過濾器設定並存取電子郵件。

客戶端 502 亦包括至少一第一過濾器 520(與第一過濾器 106 相似)及一第二過濾器 522(與第二過濾器 108 類似)，第一過濾器 520 與第二過濾器 522 可依據上文所述的過濾器描述來操作。客戶端 502 亦包括一電子郵件收件匣儲存位置(或檔案夾)524，電子郵件收件匣儲存位置(或檔案夾)524 用來接收來自於第一過濾器 520 及第二過濾器 522 中至少一者之經過過濾的電子郵件及預期已適當地標示的電子郵件訊息。一第二電子郵件儲存位置(或檔案夾)526 可被提供來容納垃圾郵件，該垃圾郵件已被使用者決定為垃圾郵件，且被使用者選擇儲存在第二電子郵件儲存位置(或檔案夾)526 內，此第二電子郵件儲存位置(或檔

案夾)526 亦可以是一垃圾筒檔案夾。如上文提及的，收件匣檔案夾 524 可包括已被第一過濾器 520 或第二過濾器 522 過濾的電子郵件，而電子郵件經過何者過濾是依據是否使用第二過濾器 522 代替第一過濾器 520 提供相同或更佳的電子郵件過濾。

一旦使用者接收來自電子郵件伺服器 512 的電子郵件，使用者會瀏覽收件匣檔案夾 524 的電子郵件，以閱讀並決定該等經過濾的收件匣電子郵件訊息之實際狀態。如果一垃圾電子郵件通過了第一過濾器 520 的話，則使用者將實施一顯性或隱含的使用者更正功能，以對該系統表明該訊息實際上應為垃圾電子郵件。然後根據此使用者更正資料訓練第一及第二過濾器(520 及 522)。如果決定該第二過濾器 522 具有比第一過濾器 520 更佳的正確率的話，第二過濾器 522 將被用來取代第一過濾器 520 用以提供相同或更佳的過濾。如在上文中提及的，如果第二過濾器 522 具有一實質上相等於第一過濾器 520 的正確率的話，則第二過濾器 522 可被使用，也可以不被使用。根據上述的數個預定的規範，過濾器訓練可由使用者選擇而發生。

現參照第 6 圖，第 6 圖圖示具有一或多個客戶端電腦 602 的系統 600，該一或多個客戶端電腦 602 可供多個使用者登入，並根據本發明的過濾技術來過濾進來的訊息。客

戶端 602 包括多人登入的能力，而使得一第一過濾器 604 及一第二過濾器 606 分別提供訊息過濾給每一登入到該電腦 602 上之不同的使用者。因此，提供一使用者介面 608，該使用者介面 608 呈現一登入畫面作為該電腦作業系統開機處理的一部分，或當有需要時，在使用者可存取他或她的進來的訊息之前接觸有關使用者的簡介。因此，當一第一使用者 610(亦被標示為使用者 1)選擇存取訊息，第一使用者 610 藉由使用者介面 608 的登入畫面 612 輸入存取資訊(通常為使用者名稱及密碼的形式)，以登入到該客戶端電腦 602 上。CPU514 處理此存取資訊以允許第一使用者透過一訊息通信應用程式(如，一郵件客戶端)而只能存取第一使用者收件匣位置 614(亦被標記為使用者 1 收件匣)及第一使用者垃圾訊息位置 616(亦被標記為使用者 1 垃圾訊息)。

當 CPU514 接收到使用者登入存取資訊時，CPU514 存取第一使用者這過濾器偏好資訊，以使用第一過濾器 604 及第二過濾器 606 來過濾下載至該客戶端電腦 602 的進來訊息。允許登入到該電腦的所有使用者(使用者 1，使用者 2，...，使用者 N)的過濾器偏好可被儲存在本地的一過濾器偏好表中。當第一使用者登入到電腦 602 上或接觸與第一使用者有關的簡介時，CPU514 可存取該過濾器偏好資

訊。因此，處理第一使用者 610 之第一及第二過濾器(604 及 606)的誤負及誤正率資料，以決定使用第一過濾器 604 或第二過濾器 606 來過濾下載的訊息。如在上文中所揭示說明的，誤正率及誤負率資料至少是從使用者更正處理中衍生出來的。一旦第一使用者 610 下載訊息，即可根據錯誤地標記之訊息來更新誤負及誤正率資料。在另一使用者登入到電腦 602 上之前的某一時間點，該第一使用者之經更新的資料會被存回到過濾器偏好表中以供未來參考之用。

當一第二使用者 618 登入時，該誤負及誤正率資料會根據與該第二使用者 618 相關的過濾偏好而改變。在第二使用者 618 進入他或她的登入資訊之後，CPU514 存取第二使用者的過濾器偏好資訊，並據以接觸第一過濾器 604 或第二過濾器 606。結合該電腦訊息應用程式的電腦作業系統限制訊息服務，以讓第二使用者 618 只存取第二使用者收件匣 620(亦被標記為使用者 2 收件匣)及第二使用者垃圾訊息位置 622(亦被標記為使用者 2 垃圾訊息)。處理第二使用者 618 之第一過濾器及第二過濾器(604 及 606)的誤負及誤正率資料，以決定使用第一過濾器 604 或第二過濾器 606 來過濾第二使用者 618 下載的訊息。如在上文中所揭示說明的，誤負及誤正率資料至少是從使用者更正

處理中衍生出來的。一旦第二使用者 618 下載訊息時，即可根據錯誤地標記之訊息來更新誤負及誤正率資料。

第 N 個使用者 624(被標記為使用者 N)的操作是以與第一及第二使用者(610 及 618)相類似的方式提供。與所有其他的使用者一樣，第 N 個使用者 624 被限制只能存取與第 N 個使用者 624 相關的使用者資訊，因此第 N 個使用者 624 只被允許可存取第 N 個使用者收件匣 626 及第 N 個使用者垃圾訊息位置 628，且在使用該訊息應用程式時，不能存取其它的收件匣(614 及 620)及垃圾訊息位置(616 及 622)。

電腦 602 被適當地配置以與該 LAN510 上的其它客戶端通信，並藉由利用一客戶端網路介面 630 來存取位於 LAN510 上的網路服務。因此，提供該訊息伺服器 512 從 SMTP(或訊息)閘道 508 接收訊息，以控制及處理客戶端(602 及 632(亦被標記為使用者 N))，及其它有線或無線裝置的進來和出去的訊息，該有線或無線裝置可透過 LAN510 經操作來與訊息伺服器 512 通信訊息。客戶端(602 及 632)被設置以可經操作與 LAN510 通信的方式，以至少存取提供於 LAN510 上的訊息服務。該 SMTP 閘道 508 界接至該 GCN500，以在 GCN500 的網路裝置與 LAN510 上的訊息實體之間提供相容的 SMTP 訊息服務。

應被瞭解的是，如上所述之比率資料平均值可被用來決定使用過濾器 604 及 606 的最佳平均設定。相似地，允許登入到電腦 602 上的使用者的最佳比率資料亦可用來配置可登入的所有使用者的過濾器。

現參照第 7 圖，第 7 圖圖示一系統 700，其中最初的過濾係實施在一訊息伺服器 702 上及第二次過濾是實施在一或多個客戶端上。提供該 GCN500 以促進來回於一或多個客戶端(704、706 及 708)(亦被標記為客戶端 1，客戶端 2，...，客戶端 N)之間的訊息溝通。該 SMTP 閘道 508 界接至該 GCN500，以在該 GCN500 上的網路裝置與 LAN510 上的訊息實體之間提供與 SMTP 相容的訊息服務。

訊息伺服器 702 可操作地設置在該 LAN510 上，並與閘道 508 相界接以控制並處理客戶端 704、706 及 708，及其它有線或無線裝置的進來與出去的訊息，該無線或有線裝置可操作以透過 LAN510 來與訊息伺服器 702 通信訊息。客戶端(704、706 及 708)(如，有線或無線裝置)設置成可操作地與 LAN510 通信，以至少存取提供於 LAN510 上的訊息服務。

依據本發明的一個態樣，訊息伺服器 702 藉由使用第一過濾器 710(與第一過濾器 106 類似)來實施最初的過濾，及客戶端使用第二過濾器 712(與第二過濾器 108 類似)

來實施第二過濾。因此，隨著第一過濾器 710 處理訊息以決定進來的訊息是垃圾或非垃圾訊息，進來的訊息從開道 508 被接收到該訊息伺服器 702 的一進來訊息的緩衝器 714 中作為暫時儲存處。緩衝器 714 可以是一單純的 FIFO(先進先出)架構，使得所有訊息都以先到先服務的方式被處理。然而，應被瞭解的是，訊息伺服器 702 可根據一已標記的優先權來過濾處理緩衝訊息。因此，緩衝器 714 被適當地配置以提供訊息優先順序，使得被傳送者標記為較高優先權的訊息從緩衝器 714 轉寄，以在其他標記為較低優先權的訊息之前過濾。優先權標記可根據與傳送者優先權標記無關的其它規範來實施，與傳送者優先權標記無關的其它規範包括但不限於：訊息的大小、該訊息傳送的日期、該訊息是否有附件、附件的大小、該訊息在緩衝器 714 中的時間有多長等等。

為了要發展出第一過濾器 710 的誤正及誤負率資料，一管理者可對第一過濾器 710 的輸出取樣用以決定有多少正常的訊息被錯誤地標示為垃圾及有多少垃圾訊息被錯誤地標示為正常。如在上文中參照本發明的一個態樣所作的說明，第一過濾器 710 的此一比率資料接著被用來當作第二過濾器 712 之新的誤正及誤負率資料的基礎。

無論如何，一旦第一過濾器 710 已將訊息過濾，根據

客戶目的地 IP 地址該訊息繞從伺服器 702 通過一伺服器網路介面 716 跨越網路 510 至適當的客戶端(如，第一客戶端 704)。第一客戶端 704 包括控制所有客戶端處理的 CPU514。CPU514 與訊息伺服器 702 相通信用以獲得第一過濾器 710 的誤正及誤負率資料，並實施與第二過濾器 712 的誤正及誤負率資料的比較，以決定何時應使用第二過濾器 712。如果比較結果為第二過濾器的比率資料不比第一過濾器的比率資料差的話，則第二過濾器 712 會被使用，且 CPU514 會與訊息伺服器 702 通信以允許預定到第一過濾器 710 的訊息未過濾地通過伺服器 702。

當第一客戶端 704 的使用者檢閱接收到的訊息並實施使用者更正時，第二過濾器 712 之新的誤正及誤負率資料會被更新。如果新的比率資料比第一比率資料差的話，則第一過濾器 710 將重新被使用，以提供過濾功能給第一客戶端 704。CPU514 持續作比率資料比較，以決定何時為該特定的客戶端 704 切換第一及第二過濾器(710 及 712)。

CPU514 根據提供上文中所述的任何一或多個過濾功能之指令來執行一可操作的演算法。該演算法包括但不侷限於：可至少執行上述的基本過濾方法之經過解碼的指令、讓使用者實施使用者更正以應付錯誤而被組合地使用之任何或所有方法、不確定性的決定、閾限的決定、使用

誤正及誤負率資料的正確性計算、及使用者互動性選擇。使用者介面 518 被提供來促進 CPUS14 與客戶端作業系統之間的溝通，使得使用者可互動地配置過濾器設定並存取電子郵件。

客戶端 502 亦包括至少該第二過濾器 712，該第二過濾器 712 可根據上文所描述的過濾器說明來操作。客戶端 502 亦包括訊息收件匣儲存位置(或檔案夾)524，該訊息收件匣儲存位置(或檔案夾)524 從第一過濾器 710 及第二過濾器 712 中至少一者街收經過濾的訊息及預期已被適當地標記的訊息。該第二訊息儲存位置(或檔案夾)526 可提供來容納垃圾郵件，該垃圾郵件已被使用者決定為垃圾郵件，且被使用者選擇儲存在第二電子郵件儲存位置(或檔案夾)526 內，此亦可以是一垃圾筒檔案夾。如上文提及的，收件匣檔案夾 524 可包括已被第一過濾器 710 或第二過濾器 712 過濾的訊息，而電子郵件經過何者過濾是依據是否使用第二過濾器 712 代替第一過濾器 710 來提供相同或更佳的進來的訊息的過濾。

如上文中提及的，一旦使用者從訊息伺服器 702 下載訊息，使用者會瀏覽收件匣檔案夾 524 的訊息，以閱讀並決定該等經過濾的收件匣訊息之實際狀態。如果一垃圾訊息通過了第一過濾器 710 的話，則使用者將實施一顯性或

隱含的使用者更正功能，以對該系統表明該訊息實際上應為垃圾訊息。然後根據此使用者更正資料訓練。如果決定該第二過濾器 712 具有比第一過濾器 710 更佳的正確率的話，第二過濾器 712 將被用來取代第一過濾器 710 用以提供相同或更佳的過濾。如在上文中提及的，如果第二過濾器 712 具有一實質上相等於第一過濾器 710 的正確率的話，則第二過濾器 712 可被使用，也可以不被使用。根據上述的數個預定的規範，過濾器訓練可由使用者選擇而發生。

應被瞭解的是，因為其它客戶端(706 及 708)使用訊息伺服器 702 來過濾訊息，所以各別客戶端(706 及 708)的新的誤正及誤負率資料將會影響到第一過濾器 710 的過濾操作。因此，各別客戶端(706 及 708)亦與訊息伺服器 702 溝通，以依據這些客戶端(706 及 708)各別的新的誤正及誤負率資料來使用或不使用第一過濾器 710。訊息伺服器 702 可包括與各別客戶過濾器要求相關的客戶端偏好的過濾器偏好表。因此，每一緩衝訊息被詢問目的地 IP 位址，且依據與儲存在過濾器表內的該目的地 IP 位址相關之過濾器偏好來加以處理。因此，雖然根據第一客戶端 704 的比率資料比較結果，預定給該第一客戶端 704 之一廣播的垃圾訊息會被要求由第一客戶 704 的第二過濾器 712 處理，但

根據所獲得的比率資料比較結果，亦同樣預定給該第二客戶 706 之垃圾訊息可被要求由訊息伺服器 702 的第一過濾器 710 來處理。

應進一步被瞭解的是，客戶端(704，706 及 708)獨立的新的比率資料可同時被伺服器 702 接收及處理以決定平均值。然後，此平均值可被用來決定是要獨立地或成群地使用該等客戶端的第一過濾器 710 或第二過濾器 712。或者，客端戶(704、706 及 708)的最佳比率資料可由伺服器 702 來決定，且被用來獨立地或成群地在第一過濾器 710 及第二過濾器 712 之間切換。

現參照第 8 圖，第 8 圖圖示使用本發明的過濾態樣之大規模的過濾系統 800 的另一實施例。在以一大規模的方式被全系統的郵件系統（如一網際網路服務提供者）所實施訊息過濾之更為堅實的應用中，多個過濾系統可被用來處理大量的進來的訊息。大量的進來的訊息 802 被接收並被送至許多不同的使用者目的地。訊息 802 經由 SMTP 閘道 804 進入該提供者系統，然後被發送至一系統訊息途徑(routing)元件 806，以繞經不同的過濾器系統 808、810 及 812(亦分別被標記為過濾器系統 1、過濾器系統 2...、過濾器系統 N)。

每一過濾器系統(808、810 及 812)都包括一途徑控制

元件、一第一過濾器、一第二過濾器、及一輸出緩衝器。過濾器系統 808 包括一途徑控制構件 814，該途徑控制構件 814 用來將訊息繞經第一系統過濾器 816 與第二系統過濾器 818 之間。第一及第二過濾器 (816 及 818) 的輸出被連接至一輸出緩衝器 820，在訊息被發送至一使用者收件匣途徑元件 822 之前暫時地儲存訊息。該使用者收件匣途徑元件 822 對從第一過濾器系統 808 的輸出緩衝器 820 接收到的每一訊息詢問使用者目的地地址，且將該訊息繞經多個使用者收件匣 824(亦被標記為收件匣 1、收件匣 2、...、收件匣 N) 中適當的使用者收件匣處。

該系統訊息途徑元件 806 包括一負荷平衡能力，根據過濾器系統 (808、810 及 812) 的可用頻寬將訊息繞經過濾器系統 (808、810 及 812) 之間來容納訊息處理。因此，如果第一過濾器系統 808 的一進來的訊息佇列 (未示出，但為途徑元件 814 的一部分) 被備份且無法容納系統 800 所需的產出時，此佇列的狀態資訊從該途徑控制元件 814 回授至該系統途徑元件 806，使得進來的訊息 802 繞至其它的過濾器系統 (810 及 812)，直到系統 814 之進來的佇列能夠接收進一步的訊息為止。其餘的每一個過濾器系統 (810 及 812) 都包括此進來的佇列回授能力，使得該系統途徑元件 806 可處理所有可用的過濾器系統 (過濾器系統 1、過濾器

系統 2、...、過濾器系統 N) 之間的訊息負荷。

第一系統過濾器 808 的可適性過濾能力現將詳細地加以說明。在此特定的系統實施中，藉由提供有關過濾器的正確性的回授來對訊息標記/去除標記，系統管理員將擔任起決定是什麼構成系統 800 的垃圾郵件的工作。亦即，管理員實施使用者更正以產生每一系統(808、810 及 812)的 FN 及 FP 資訊。由於進來的訊息的數量龐大，這可根據一統計取樣方法來實施，該取樣方法可數學上地提供該被取出的樣本有一高程度的或然率，可反應出各別過濾器系統(808、810 及 812)所實施在決定何者是一垃圾訊息及何者為非垃圾訊息的過濾上的正確性。

又，管理員將經由一系統控制元件 826 從該緩衝器 820 取出一樣本，並驗證在該樣本上之訊息標籤的正確性。該系統控制元件 826 可以是一硬體及/或軟體處理系統，並連接至該等過濾器系統(808、810 及 812)以監視及控制它們。任何錯誤地標記的訊息都將用來建立第一過濾器 816 的誤負(FN)及誤正(FP)率資料。然後，此 FN/FP 率資料用在第二過濾器 818 上。如果該第一過濾器 816 的比率資料落在一閾限值以下的話，則第二過濾器 818 即可用來提供至少與第一過濾器 816 一樣好的過濾。當管理員再次從緩衝器 820 實施使用者更正取樣時，如果第二過濾器 818 的

FN/FP 資料比第一過濾器 816 差的話，則該途徑控制元件 814 將會處理第二過濾器 818 的此 FN/FP 資料，並決定該訊息的途徑應被交換回到第一過濾器 816。

該系統控制元件 826 界接至該系統訊息途徑元件 806，以在系統控制元件 826 與系統訊息途徑元件 806 之間交換資料，並透過該管理員來提供管理。該系統控制元件 826 亦界接至其餘的系統（如過濾器系統 2、...、過濾器系統 N）的輸出緩衝器，以提供這些系統的取樣能力。管理員亦可透過系統控制元件 826 存取使用者收件匣途徑構件 822，以監督使用者收件匣途徑構件 822 之操作。

一如上文中參照第 1 圖所述之過濾器的正確性可延伸為複數個過濾器系統的正確性。該第一過濾器系統 808 的 FN/FP 率資料可用來訓練第二過濾器系統 810 及第三過濾器系統 812 的過濾器，以進一步加強整個系統 800 的過濾能力。相同地，可依據特定系統的 FN/FP 率資料來實施負荷控制。亦即，如果第一系統 808 的整體 FN/FP 資料比第二系統 810 的 FN/FP 資料差的話，則繞至第二系統 810 的訊息會比繞至第一系統 808 的多。

應被瞭解的是，過濾器系統 (808、810 及 812)可以是獨立的過濾演算法，每一演算法在專屬電腦或電腦組合上執行。或者，當有足夠的硬體能力時，演算法可一起在一

單一電腦上執行以使得所有的過濾功能被實施在一單一的堅實機器上。

現參照第 9 圖，第 9 圖圖示一可執行上述架構的電腦的方塊圖。為了要提供本發明的不同態樣的額外脈絡，第 9 圖及以下的討論的目的是要提供適合的計算環境 900 的簡短且一般性的描述，而本發明的不同態樣可在計算環境 900 上實施。雖然本發明已經在上文中用電腦可執行的指令的一般性內容來加以說明，而電腦可執行的指令可在一或多台電腦上運行，但熟習此技藝者將可瞭解到本發明亦可用與其它程式模組相組合及/或硬體與軟體的組合的形式來實施。大體上，可執行特定工作或實施特定的抽象資料種類之程式模組包括常式，程式，元件，資料結構等等。又，熟習此技藝者將可瞭解到，本發明的方法可用其它電腦系統配置來實施，包括單一處理器或多處理器電腦系統、迷你電腦、主機型電腦（即個人電腦）、手持式計算裝置、以微處理器為基礎或可程式的消費性電子裝置及類似者，上述的每一種裝置都可操作地耦合至一或多個相關的裝置上。本發明之所舉出的態樣亦可在分散式計算環境上實施，其中某些工作是由透過一通信網路連接之遠端處理裝置來實施。在一分散式計算環境中，程式模組可位在本地及遠端記憶儲存裝置上。

再次參照第 9 圖，用來實施本發明的不同態樣之該舉例性的環境 900 包括一電腦 902，該電腦 902 包括一處理單元 904、一系統記憶體 906 及一系統匯流排 908。該系統匯流排 908 將系統元件（包括但不侷限於系統記憶體 906）耦接至處理單元 904。處理單元 904 可以是任何市面上可購得的處理器。雙微處理器及其它多處理器架構亦可作為該處理單元 904。

系統匯流排 908 可以是數種匯流排結構中的任何一種，包括使用市面上任何一種的匯流排架構之一記憶體匯流排或記憶體控制器、週邊匯流排及本地匯流排。該系統記憶體 906 包括唯讀記憶體 (ROM) 910 及隨機存取記憶體 (RAM) 912。一基本輸入 / 輸出系統 (BIOS) 被儲存在 ROM 910 中，該 BIOS 包含可在開機期間幫助傳輸資訊於電腦 902 內的部件之間的常式。

電腦 902 進一步包括一硬碟機 914、一磁碟機 916（如，從一可移除碟片 918 讀取或寫入）及一光碟機 920（如讀取一 CD-ROM 碟片 922 或從其它光學媒體讀取或寫入）。硬碟機 914、磁碟機 916 及光碟機 920 可分別藉由一硬碟機介面 924、一磁碟機介面 926 及一光碟機介面 928 而連接至該系統匯流排 908。該等裝置及與該等裝置相關的電腦可讀取媒體提供，資料、資料結構、電腦可執行的指令等

等的非揮發性儲存。對於電腦 902 而言，該等裝置及媒體容納以一適當的數位格式來儲存程式設計。雖然上文中有關電腦可讀取媒體的說明係指一硬碟、一可移除磁碟及一 CD，但熟習此技藝者可瞭解的是，其它可被電腦所讀取之媒體，ZIP 機、磁帶匣、快閃記憶卡、數位視訊碟片、卡匣、及類似者，亦可被使用在此舉例性的環境中，且任何這些媒體都可包含電腦可執行的指令用以實施本發明的方法。

許多電腦程式模組（包括作業系統 930、一或多個應用程式 932、其它程式模組 934 及程式資料 936）都可被儲存在該等裝置及 RAM 912 中。應被瞭解的是，本發明可用許多市面上可購得的作業系統或作業系統的組合來實施。

一使用者可經由一鍵盤 938 及一指標裝置，如一滑鼠 940，來將命令及資訊輸入電腦 902。其它的輸入裝置（未示出）可包括一麥克風、一紅外線（IR）遙控器、一搖桿、一遊戲盤、一衛星圓盤、一掃描器或類似者。這些及其它輸入裝置通常是經由串接埠介面 942 連接至處理單元 904，而該串接埠介面 942 耦接至該系統匯流排 908，但亦可經由其它界面來連接，如一平行埠、一遊戲埠、一萬用串接匯流排（“USB”）、一 IR 介面等。一監視器 944 或其它種類的顯示裝置亦經由一介面，如一視訊配接器 946，而

被連接至該系統匯流排 908。除了監視器 944 之外，一電腦典型地包括其它的週邊輸出裝置(未示出)，如喇叭、印表機等等。

電腦 902 可在一網路環境中操作，並使用邏輯連線至一或多個遠端電腦，如一遠端電腦 948。遠端電腦 948 可以是一工作站、一伺服器電腦、一路由器、一個人電腦、可攜式電腦、以微處理器為基礎的娛樂裝置、一同級裝置或其它共同的網路節點，且典型地包括許多或所有與電腦 902 相關的元件，然而，為了清晰起見，只有記憶儲存裝置 950 被示出。上述的邏輯連線包括一 LAN952 及一 WAN954。此等網路環境在辦公室中、在企業內的電腦網路中、在網際網路中是很常見的。

當在一 LAN 網路環境中使用時，電腦 902 是透過一網路介面或配接器 956 而連接至區域網路 952。當在 WAN 網路環境中使用時，電腦 902 典型地包括一數據機 958，或是連接至 LAN 上的一通信伺服器，或具有其它的構件來在 WAN954 (如網際網路) 上建立通信。數據機 958(可以是內建或是外接)是經由串接埠介面 942 而連接至系統匯流排 908。在一網路化的環境中，與電腦 902 相關的程式模組或程式模組的一部分可被儲存在遠端記憶儲存裝置 950 上。應被瞭解的是，所示的網路連線是舉例性，且建立通

信鏈於電腦之間的其它機構亦可被使用。

依據本發明的一個態樣，過濾器架構可適應使用過濾之系統的特定使用者所想要的過濾程度。然而，應被瞭解的是，此”可適性”之面向可從地區性使用者系統環境延伸至該系統供應商的製造過程，特定種類使用者的過濾程度可在工廠中被加以選擇用以實施在販售的系統中。例如，如果一購買者決定第一批採購的系統是要提供給不需要存取垃圾郵件的使用者的話，則此批系統在工廠的內定設定可設定為高，而供第二類別之使用者使用之第二批系統則可配置為較低的設定，以檢閱更多的垃圾郵件。在任何一種情形下，本發明之可適性的本質可被局部地建立，以允許任何類別的獨立使用者來調整過濾程度，或者如果禁止的話，可完全防止對內定的設定值作任何的變更。同樣應被瞭解的是，網路管理員執行比較存取權利來配置一或多個系統，以使得一或多個系統適當地與用本文所揭示之過濾器架構一起配置，而該網路管理員亦可局部地實施此一類別的配置。

上文中所揭示的包括了本發明的例子。當然不可能針對本發明之目的描述出所有可想出來的元件或方法組合，但熟習此技藝者可瞭解到本發明仍有許多其它進一步的組合及變更的可能。因此，本發明包含了落在由申請專利範

圍所界定的精神與範圍內的所有這些變化，修改及變更。又，在上文中或在申請專利範圍中所用之”包括”一詞，該詞之意義與”包含”的意義是相同的，而”包含”是申請專利範圍中的傳統用詞。

【圖式簡單說明】

第 1 圖顯示依據本發明的過濾器系統的一般方塊圖。

第 2 圖為性能取捨(tradeoff)與攔截率之間的關係的圖表。

第 3 圖為依據本發明的方法的流程圖。

第 4A 及 4B 顯示依據本發明之用來配置可適性垃圾郵件過濾系統的舉例性使用者界面。

第 5 圖顯示運用本發明之訊息處理架構的一般方塊圖。

第 6 圖顯示具有一或更多台客戶電腦以促進多個使用者登入的系統，並依據本發明的技術來過濾進來的訊息。

第 7 圖顯示依據本發明的系統，其中最初的過濾是實施在一訊息伺服器上，且輔助過濾係實施在一或多個客戶上。

第 8 圖顯示用在大規模應用上之可適性過濾系統的方塊圖。

第 9 圖顯示一電腦的方塊圖，該電腦可被操作用以執行本文所揭示的架構。

【元件代表符號簡單說明】

100	垃圾訊息偵測系統	102	訊息
104	過濾器控制構件	106	第一(種子)過濾器
108	第二(新的)過濾器	112	收件匣
114	使用者更正元件	400	使用者介面
200	接收者-操作者曲線(ROC)		
401	垃圾郵件頁面	402	選單列
404	連結列	408	設定子視窗
410	第一選項	412	第二選項
414	第三選項	416	第四選項
418	相關設定區	420	信箱視窗
422	電子郵件控制工具列	424	檔案夾選擇子視窗
426	訊息	428	訊息預覽子視窗
500	網路	502,504,506	客戶端
508	閘道伺服器	510	區域網路(LAN)
512	電子郵件伺服器	514	中央處理單元(CPU)
518	使用者介面	520	第一過濾器
522	第二過濾器		

524	電子郵件收件匣位置(檔案夾)
526	第二電子郵件收件匣位置(檔案夾)
600	系統
602	客戶端電腦
604	第一過濾器
606	第二過濾器
608	使用者介面
610	第一使用者
612	登入畫面
618	第二使用者
614	第一使用者收件匣位置(檔案夾)
616	第一使用者垃圾訊息位置(檔案夾)
620	第二使用者收件匣
624	第N個使用者
622	第二使用者垃圾訊息位置
626	第N個使用者收件匣
630	客戶端網路介面
628	第N個使用者垃圾訊息位置
632	客戶端
700	系統
702	訊息伺服器
704,706,708	客戶端
710	第一過濾器
712	第二過濾器
714	緩衝器
716	伺服器網路介面
800	大規模過濾系統
802	進來的訊息
804	SMTP閘道
806	系統訊息途徑元件
808,810,812	過濾器系統
814	途徑控制系統
816	第一系統過濾器
818	第二系統過濾器
820	輸出緩衝器
822	使用者收信匣途徑元件

824	使 用 者 收 件 匣	826	系 統 控 制 構 件
900	環 境	902	電 腦
904	處 理 單 元	906	系 統 記 憶 體
908	系 統 匯 流 排	910	唯 讀 記 憶 體 (ROM)
912	隨 機 存 取 記 憶 體 (RAM)	914	硬 碟 機
916	磁 碟 機	918	可 移 除 碟 片
920	光 碟 機	922	CD-ROM 碟 片
924	硬 碟 機 介 面	926	磁 碟 機 介 面
928	光 碟 機 介 面	930	作 業 系 統
932	應 用 程 式	934	程 式 模 組
936	程 式 資 料	938	鍵 盤
940	滑 鼠	942	串 接 埠 介 面
944	監 視 器	946	視 訊 配 接 器
948	遠 端 電 腦	950	記 憶 儲 存 裝 置
952	區 域 網 路 (LAN)	954	廣 域 網 路 (WAN)
956	網 路 配 接 器	958	數 據 機

十、申請專利範圍：

1. 一種資料過濾系統，該資料過濾系統包含：

一 第一過濾器，該第一過濾器經配置以至少部分地依據垃圾資訊來標記訊息為垃圾，該垃圾資訊與該等訊息相關聯，該第一過濾器具有與該第一過濾器相關聯的一誤正率及一誤負率；

一 或更多第二過濾器，該一或更多第二過濾器經配置以至少部分地依據垃圾資訊來標記該等訊息為垃圾，該垃圾資訊與該等訊息相關聯，該一或更多第二過濾器最初與該第一過濾器的該誤正率及該誤負率相關聯；

一 過濾器輸出，該過濾器輸出經配置以從該第一過濾器與該一或更多第二過濾器接收已標記與未標記之訊息；

一 使用者更正元件，該使用者更正元件經配置以接收與已傳送至該過濾器輸出的已標記與未標記之該等訊息相關之使用者動作，並依據與已傳送至該過濾器輸出的已標記與未標記之該等訊息相關之使用者動作來輸出誤正資料與誤負資料；以及

一 過濾器控制器，該過濾器控制器經配置以執行下列步驟：

接收該誤正資料與該誤負資料；

依據該一或更多第二過濾器之至少一者之該誤正資

料或該誤負資料或兩者，調整該一或更多第二過濾器之至少一者之該誤正率或該誤負率或兩者；以及

根據一閾限與該第一過濾器與該一或更多第二過濾器各別的誤正率、誤負率或兩者，將接收訊息依序繞經該第一過濾器與該一或更多第二過濾器之間。

2. 如申請專利範圍第1項所述之系統，該等使用者動作包含隱含地標記該等訊息。
3. 如申請專利範圍第1項所述之系統，該等使用者動作包含顯性地標記該等訊息為垃圾。
4. 如申請專利範圍第1項所述之系統，該第一過濾器與該一或更多第二過濾器更進一步經配置以至少部分地依據好的資料，標記該等訊息為垃圾。
5. 如申請專利範圍第1項所述之系統，其中該過濾器控制器更進一步經配置以依據其他使用者訊息之該內容，調整該至少一或更多第二過濾器之該誤正率與該誤負率。
6. 如申請專利範圍第1項所述之系統，其中經配置以根據一

閾限與該第一過濾器與該一或多更多第二過濾器各別的誤正率、誤負率或兩者，將接收訊息依序繞經該第一過濾器與該一或多更多第二過濾器之間之步驟，包含經配置以將該等訊息繞經具有一最佳誤正率之一過濾器之步驟。

7. 如申請專利範圍第1項所述之系統，其中該過濾器控制器更進一步經配置以在已標記一預定數量之訊息之後、已發生一預定時間之後或兩者皆發生之後，調整該至少一或多更多第二過濾器之該誤正率與該誤負率或兩者。
8. 如申請專利範圍第1項所述之系統，其中該系統更進一步經配置以從複數個產生閾限值中選取該閾限，該複數個產生的閾限值包含：超出合格閾限值的一平均閾限值、具有最低誤正率的一閾限值、以及最佳的或然率閾限(p^*)之一閾限值，其中 $p^*=N/(N+1)$ ，N 為一數量之訊息。
9. 如申請專利範圍第1項所述之系統，其中該系統更進一步經配置以從複數個閾限值中選取該閾限。
10. 一種電腦可讀取媒體，該電腦可讀取媒體具有電腦可執行指令，當該等電腦可執行指令藉由一電腦執行時，可實行

以下步驟：

藉由一第一過濾器，至少部分地依據垃圾資訊來標記訊息為垃圾，該垃圾資訊與該等訊息相關聯，該第一過濾器具有與該第一過濾器相關聯的一誤正率及一誤負率；

藉由一或更多第二過濾器，至少部分地依據垃圾資訊來標記該等訊息為垃圾，該垃圾資訊與該等訊息相關聯，該一或更多第二過濾器最初與該第一過濾器的該誤正率及該誤負率相關聯；

藉由一過濾器輸出，從該第一過濾器與該一或更多第二過濾器接收已標記與未標記之訊息；

藉由一使用者更正元件，接收與已傳送至該過濾器輸出的已標記與未標記之該等訊息相關之使用者動作，並依據與已傳送至該過濾器輸出的已標記與未標記之該等訊息相關之使用者動作來輸出誤正資料與誤負資料；以及

藉由一過濾器控制器，引導下列步驟：

接收該誤正資料與該誤負資料；

依據該一或更多第二過濾器之至少一者之該誤正資料或該誤負資料或兩者，調整該一或更多第二過濾器之至少一者之該誤正率或該誤負率或兩者；以及

根據一閾限與該第一過濾器與該一或更多第二過濾器各別的誤正率、誤負率或兩者，將接收訊息依序

繞經該第一過濾器與該一或更多第二過濾器之間。

11. 一種電腦，該電腦包含如申請專利範圍第1項所述的系統。

12. 一種電腦網路系統，該電腦網路系統包含如申請專利範圍第1項所述的系統。

13. 一種可攜式計算裝置，該可攜式計算裝置包含如申請專利範圍第1項所述的系統。

14. 如申請專利範圍第13項所述之裝置，該裝置係為個人資料助理、電話或膝上型電腦。

15. 一種促進可適性資料過濾之系統，該系統包含：

一處理器；

一記憶體，該記憶體通訊式耦接至該處理器，該記憶體具有儲存於該記憶體中之電腦可執行指令，該等電腦可執行指令經配置以實施該資料過濾系統，包括：

一第一過濾器，該第一過濾器經配置以依據垃圾資訊來標示訊息為垃圾，該垃圾資訊與該等訊息相關聯，其中該第一過濾器與一第一正確率相關聯；

一 第二過濾器，該第二過濾器經配置以依據垃圾資訊來標示該等訊息為垃圾，該垃圾資訊與該等訊息相關聯，其中該第二過濾器最初與該第一正確率相關聯；

一 過濾器輸出，該過濾器輸出經配置從該第一過濾器與該第二過濾器接收已標示與未標示之訊息；

一 使用者更正元件，該使用者更正元件經配置以接收使用者動作，並依據該等使用者動作來計算該第一正確率，且該等使用者動作覆蓋（override）在該過濾器輸出處接收之該等訊息之最初標示；以及

一 過濾器控制元件，該過濾器控制元件經配置以執行下列步驟：

利用一閾限與該等使用者動作來訓練該第二過濾器，其中若一訊息為垃圾之或然率超過該閾限，則訓練該過濾器來標示該訊息為垃圾；

計算該第二過濾器之一第二正確率；

若該第二正確率優於該第一正確率，則將接收訊息依序繞經該第二過濾器，取代繞經該第一過濾器。

16. 如申請專利範圍第 15 項所述之系統，其中該第二過濾器係與該第一過濾器結合。

17. 如申請專利範圍第 15 項所述之系統，其中該垃圾資訊係包括下列之至少一者：傳送者資訊、來源 IP 位址、傳送者名稱、傳送者電子郵件位址、傳送者網域名稱、在識別子欄位中之難理解之文字與數字串、訊息本文中之用字與用詞、訊息本文中之特徵、或彈出式附加廣告之嵌入式鏈結。
18. 如申請專利範圍第 15 項所述之系統，其中該第一正確率及該第二正確率包含一誤正率及一誤負率。
19. 如申請專利範圍第 15 項所述之系統，其中該第二正確率係為該閾限之一函數。
20. 如申請專利範圍第 15 項所述之系統，其中該第一過濾器係為一種子過濾器，且該種子過濾器經配置以根據歷史資料來識別一般垃圾訊息。
21. 如申請專利範圍第 15 項所述之系統，其中該等使用者動作包括至少顯性或隱含地標記該訊息為一垃圾訊息或一非垃圾訊息中之一者。
22. 如申請專利範圍第 15 項所述之系統，其中該等使用者動作

包括下列至少一者：標記一訊息為一非垃圾訊息、閱讀並刪除一訊息、轉寄一訊息、或回覆一訊息。

23. 如申請專利範圍第 15 項所述之系統，其中該閾限係為一最佳的或然率閾限 (p^*)，其中 $p^* = N / (N + 1)$ ， N 為一數量之訊息。

24. 一種促進資料過濾的方法，該方法包含以下步驟：

根據一種子過濾器之一誤正率與一誤負率，自動地過濾輸入訊息；

接收關於至少一已過濾訊息之使用者更正資料；

依據關於至少一已過濾訊息之該使用者更正資料，決定該種子過濾器之一正確性；

使用該使用者更正資料來訓練一新過濾器；

決定該新過濾器的一誤正率及一誤負率；

依據該新過濾器的該誤正率及該誤負率，決定該新過濾器之一正確性；以及

若該新過濾器的正確性優於該種子過濾器的正確性，則用該新過濾器取代該種子過濾器。

25. 如申請專利範圍第 24 項所述之方法，其中該使用者更正資

料包含關於覆蓋至少一已過濾之該訊息之最初分類。

26. 如申請專利範圍第 24 項所述之方法，其中決定該新過濾器的該誤正率及該誤負率之步驟包含以下步驟：在藉由一使用者標記一預定數量的垃圾及非垃圾訊息之後、已發生一預定的時間之後、或兩者皆發生之後，決定該新過濾器的該誤正率及該誤負率。
27. 一種電腦可讀取媒體，該電腦可讀取媒體具有儲存在該電腦可讀取媒體上之電腦可執行指令，當該等電腦可執行指令藉由一電腦執行時，可實施如申請專利範圍第 24 項之方法。
28. 一種資料過濾系統，該資料過濾系統包含：
 用於過濾訊息的一第一構件，過濾該等訊息的一第一構件具有與該第一構件相關聯的一誤正率及一誤負率；
 過濾該等訊息之一新構件，過濾該等訊息之該新構件係根據與過濾該等訊息之該第一構件相關聯的該誤正率及該誤負率來加以訓練；
 一決定構件，該決定構件係用於決定與過濾該等訊息之該新構件相關聯之一新誤正率及一新誤負率與一閾限的函數關係；
 一決定閾限構件，該決定閾限構件係用於決定過濾該等

訊息之該新構件之一閾限；

一取代構件，該取代構件係用在對於過濾該等訊息之該新構件而言，若存在一閾限使得與過濾該等訊息之該新構件相關聯之該新誤正率及該新誤負率一起考量時優於過濾該等訊息之該第一構件之該誤正率及該誤負率的話，則用過濾該等訊息之該新構件取代過濾該等訊息之該第一構件。

29. 一種具有已儲存之電腦可執行指令之方法，已儲存之該等電腦可執行指令在一處理器上執行以促進可適性資料處理，該方法包含以下步驟：藉由一第一過濾器，依據垃圾資訊來標示訊息為垃圾，該垃圾資訊與該等訊息相關聯，其中該第一過濾器與一第一正確率相關聯；

藉由一第二過濾器，依據垃圾資訊來標示該等訊息為垃圾，該垃圾資訊與該等訊息相關聯，其中該第二過濾器最初與該第一正確率相關聯；

藉由一過濾器輸出，從該第一過濾器與該第二過濾器接收已標示與未標示之訊息；

藉由一使用者更正元件，接收使用者動作，並依據該等使用者動作來計算該第一正確率，且該等使用者動作覆蓋在該過濾器輸出處接收之該等訊息之最初標示；以及

包括一過濾器輸出，該過濾器輸出經配置以：

利用一閾限與該等使用者動作來訓練該第二過濾器，其中若一訊息為垃圾之或然率超過該閾限，則訓練該過濾器來標示該訊息為垃圾；

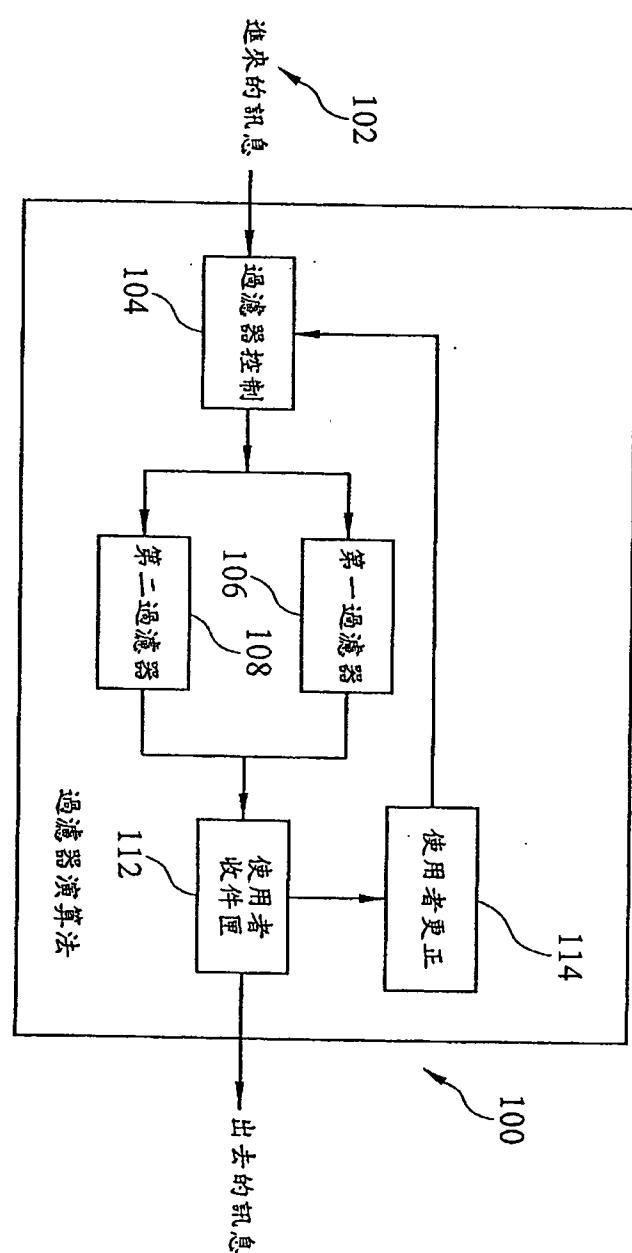
計算該第二過濾器之一第二正確率；以及

若該第二正確率優於該第一正確率，則將接收訊息依序繞經該第二過濾器，取代繞經該第一過濾器；

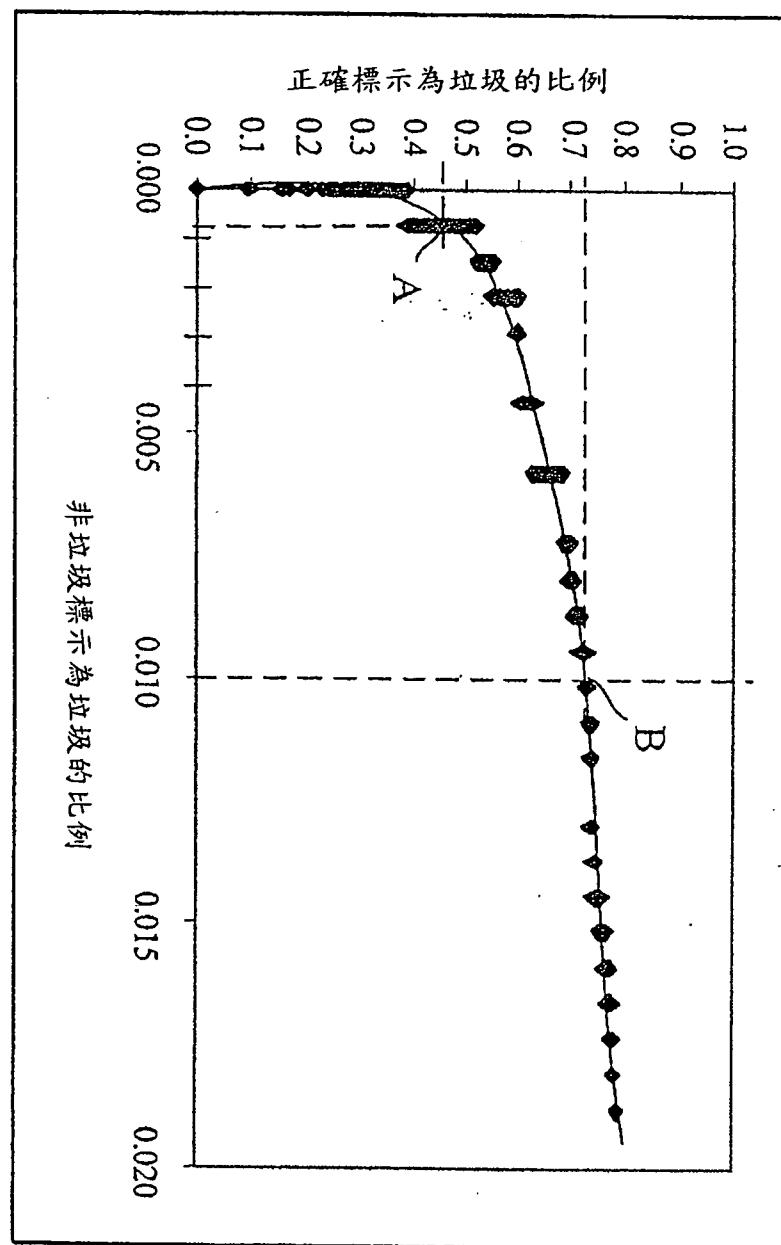
其中該垃圾資訊係包括下列之至少一者：傳送者資訊、來源 IP 位址、傳送者名稱、傳送者電子郵件位址、傳送者網域名稱、在識別子欄位中之難理解之文字與數字串、訊息本文中之用字與用詞、訊息本文中之特徵、或彈出式附加廣告之嵌入式鏈結。

101年5月10日修正本

第 93/01775 號專利案 / 01 年 5 月修正

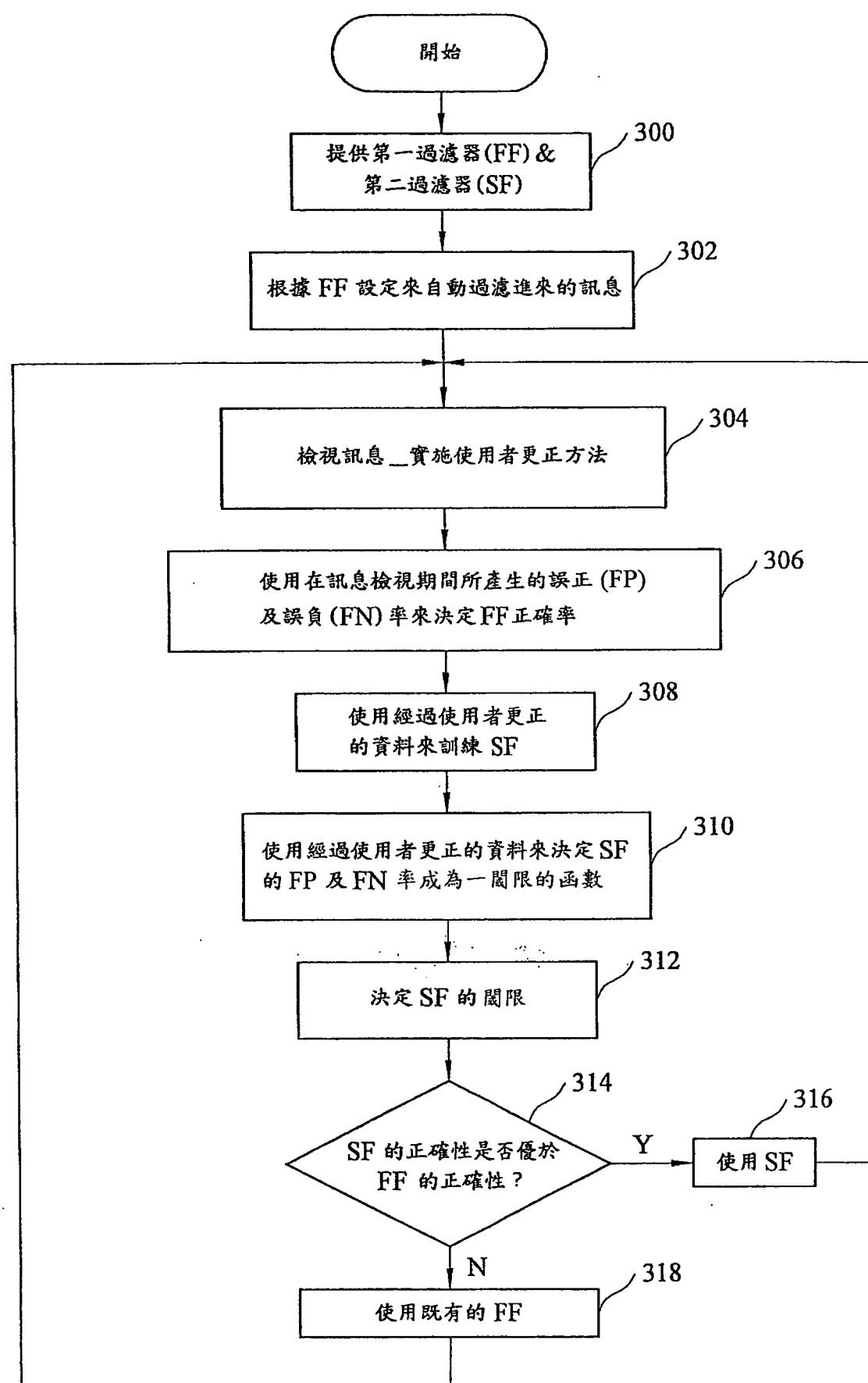


第 1 圖



第 2 圖

200



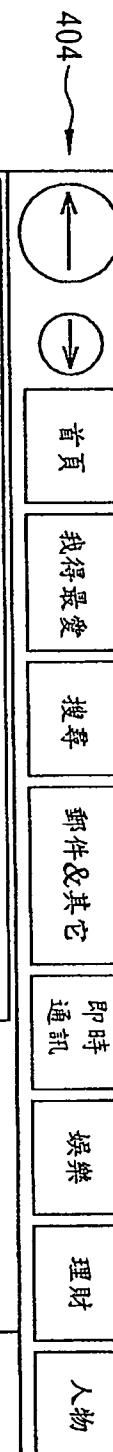
第 3 圖

401

400

歡迎

402 檔案 編輯 檢視 登出 說明&設定



404 406 設定

進階垃圾郵件設定

設定首頁

電子郵件

相關設定

垃圾郵件過濾器

安全名單

封鎖傳送者名單

顯示所有設定

418 開始

選擇是否要動垃圾郵件保護

啟動 — 當你使用垃圾及非垃圾按鈕時，過濾會更為正確
選擇你的垃圾郵件過濾器等級

使用者偏好
資訊

內定 — 明顯的垃圾郵件被攔截

加強型 — 更多垃圾郵件被攔截

414
412

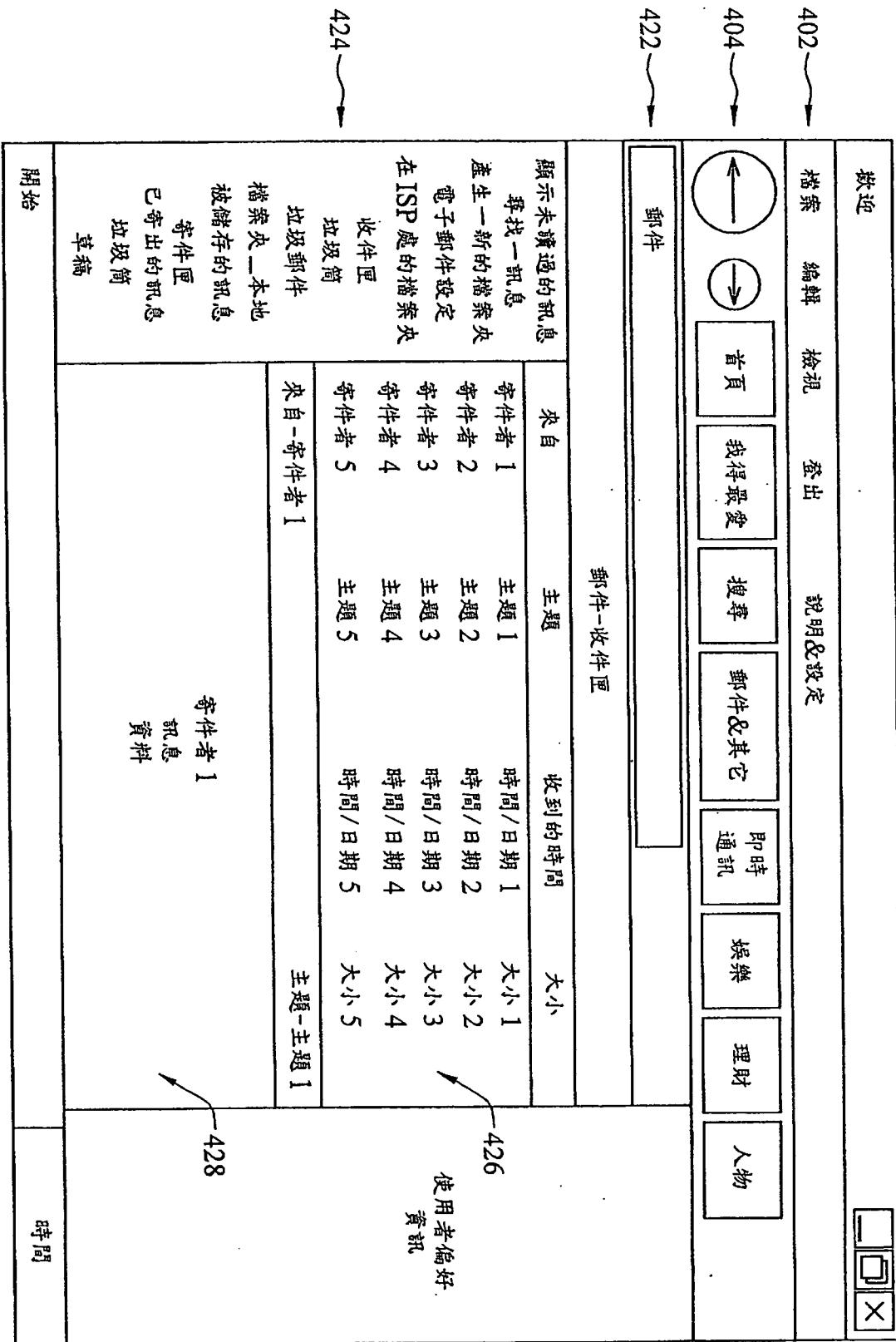
410

唯一型 — 你將只會接到來自於在你的通訊錄上或安全名單
上的人的郵件注意在加強及唯一等級中你必需定期檢查的
郵件檔案夾，因為你想要的郵件可內被標記為垃圾郵件並
被放在其內

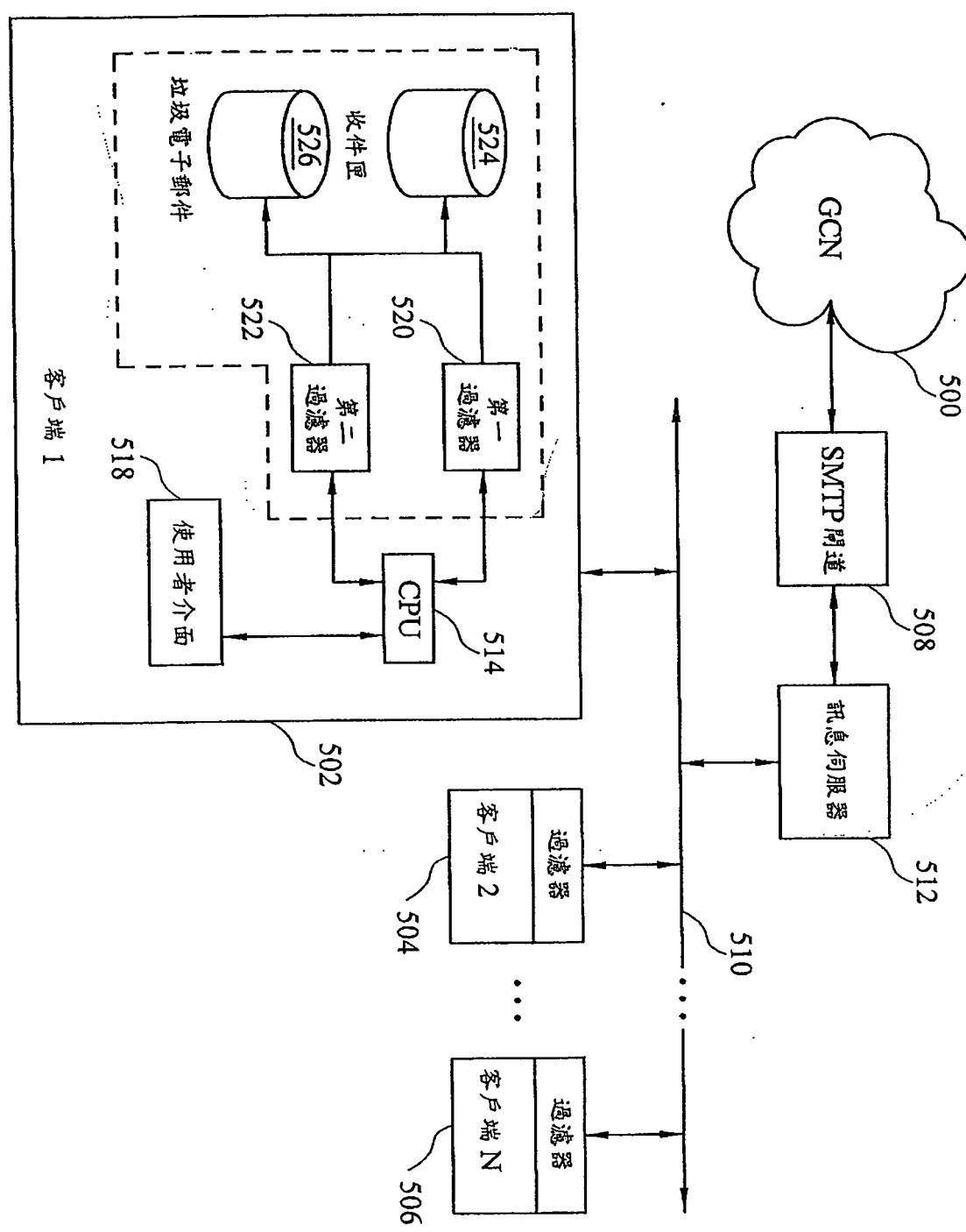
時間

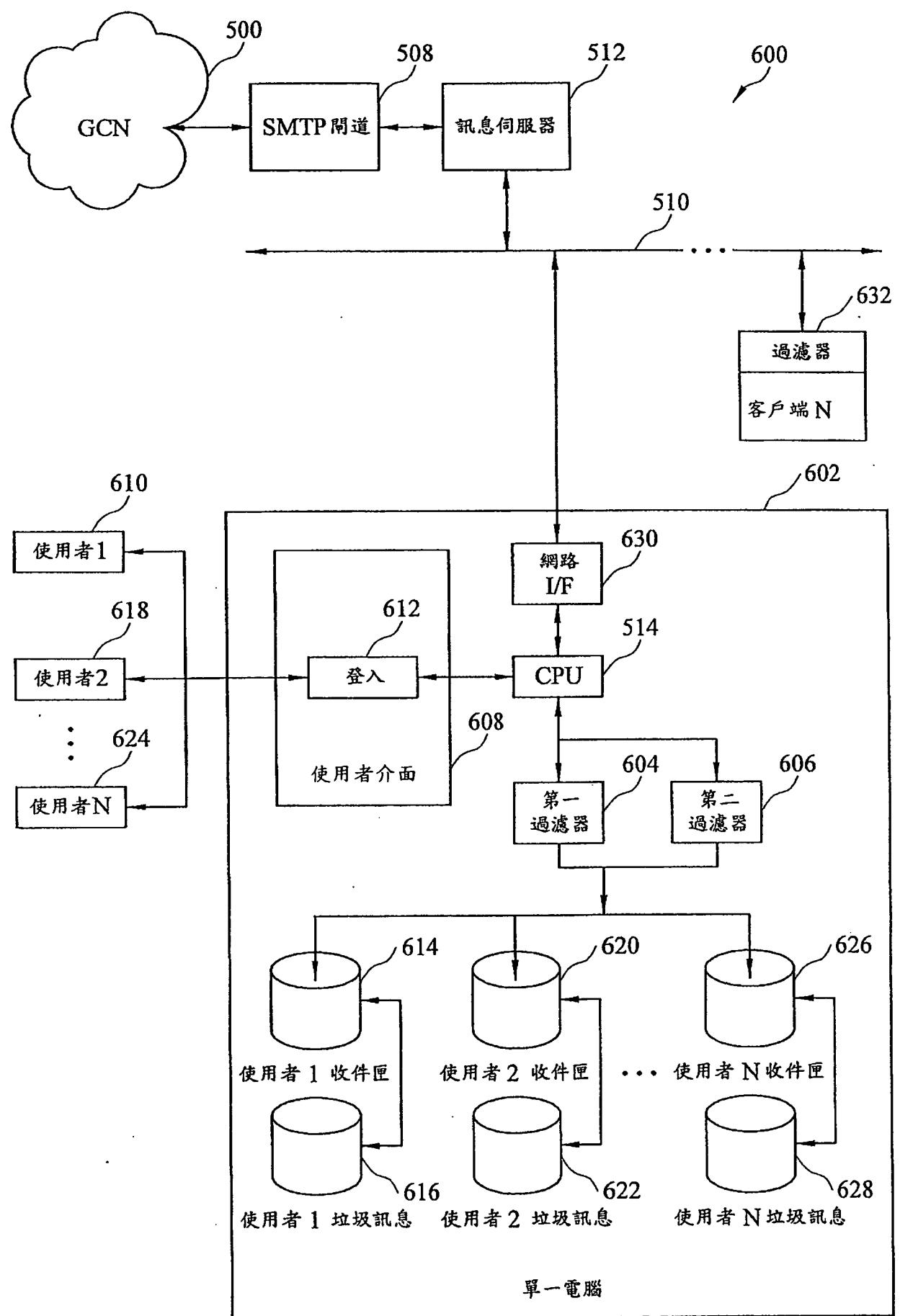
I393391

420

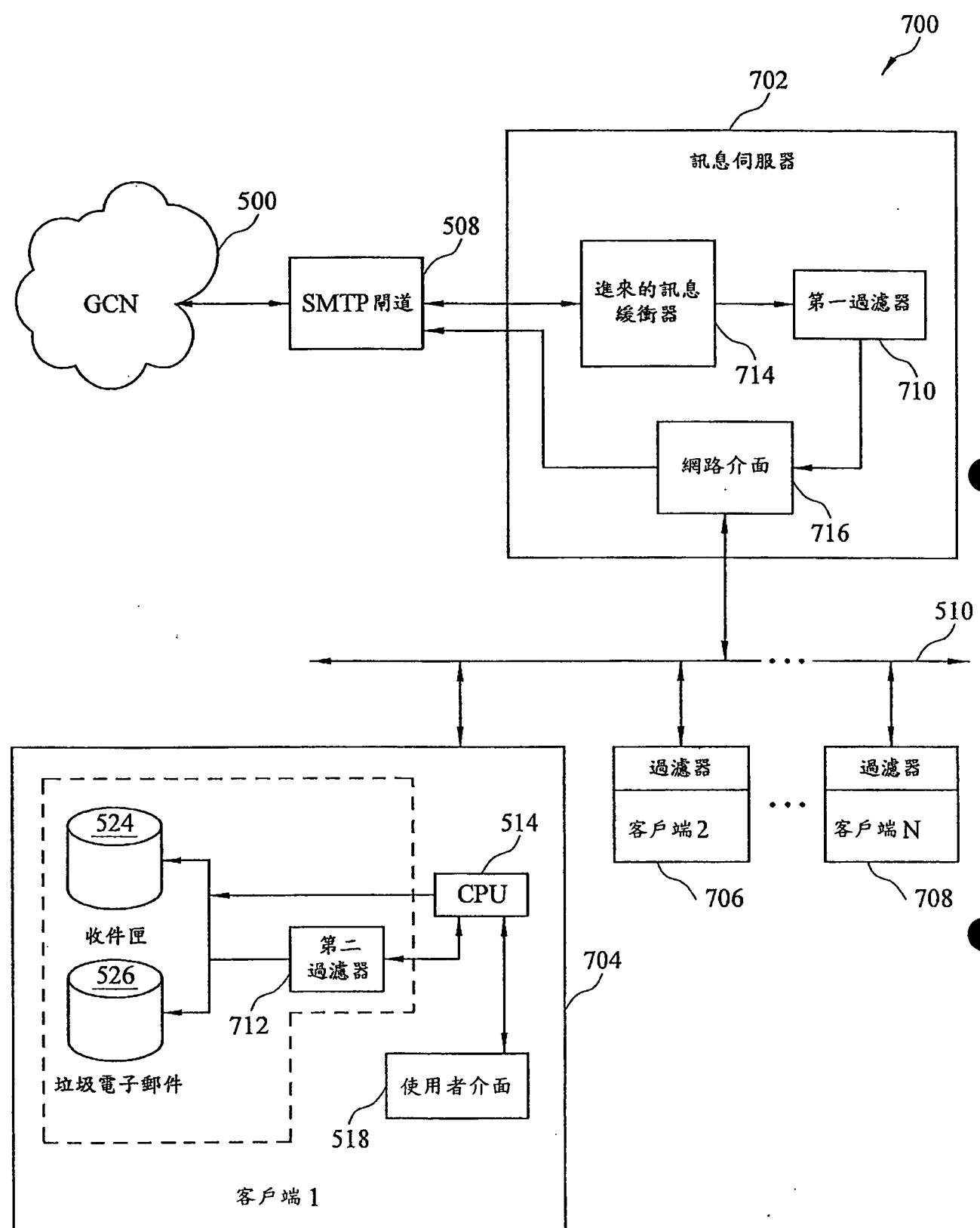


第 4b 圖

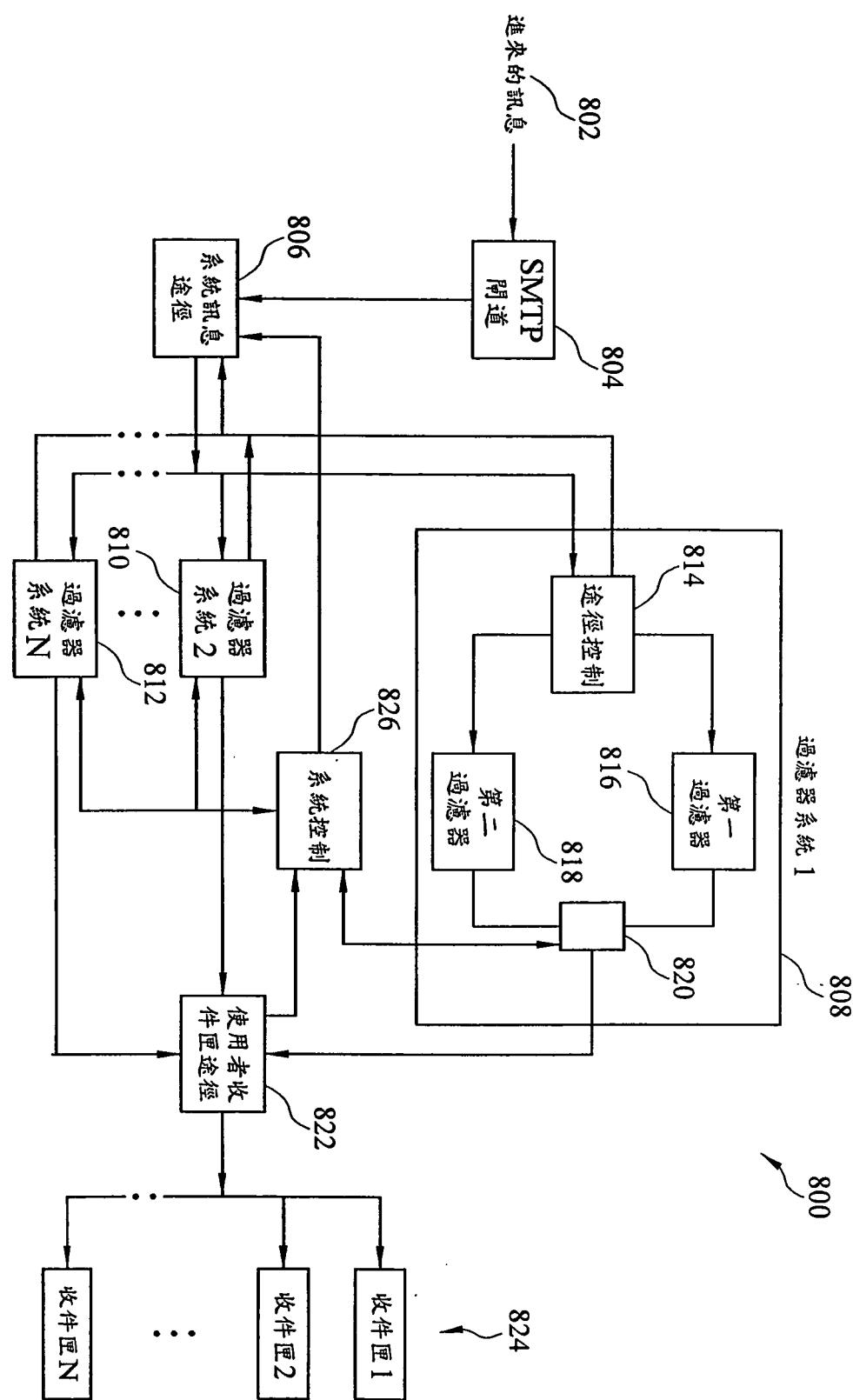




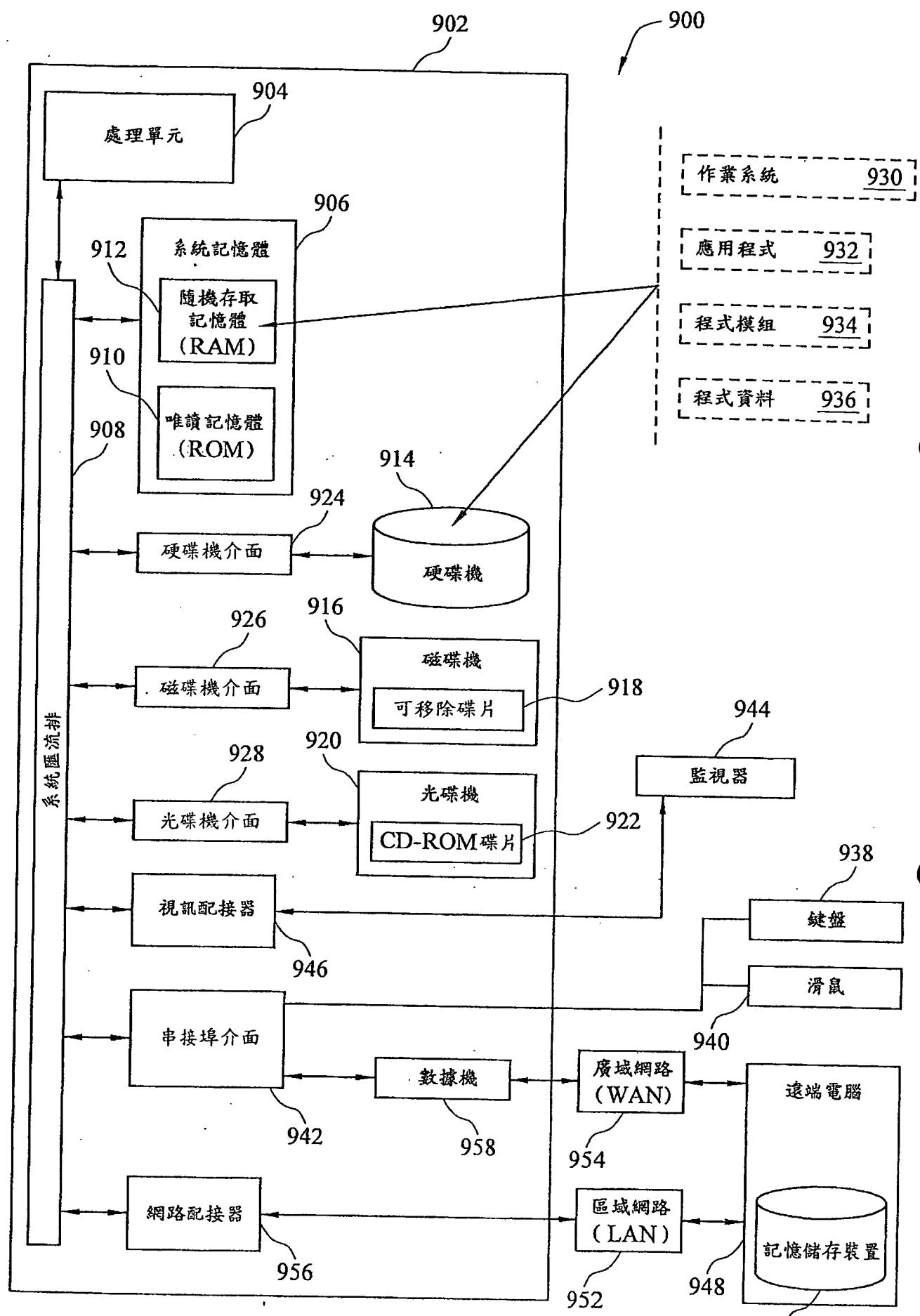
第 6 圖



第 7 圖



第 8 圖



第 9 圖