

RECURSO DE REPOSICIÓN.

Señores

**Superintendencia de Industria y Comercio
División de Nuevas creaciones****E. S. D.**

SUPERINTENDENCIA DE INDUSTRIA Y COMERCIO



No. 05-076623- -00004-0000

Fecha: 2008-01-17 15:18:58 Dep. 2020 NUECREACIONE
Tra. 11 PATENTEIYII Eva: 379 FASENACIONALI
Act. 412 REPOSPRESRECU Folios: 154**Ref.: Expediente No. 05-076.623****Solicitud de patente de invención PCT "RETROALIMENTACION
CICLICA PARA LA PREVENCIÓN DE SPAM."****Solicitante: MICROSOFT CORPORATION**

Darío Cárdenas Navas, domiciliado en Bogotá D.C., identificado como aparece al pie de mi firma, abogado titulado con Tarjeta Profesional 1.250 expedida por el C.S.J., en mi condición de apoderado de la sociedad **MICROSOFT CORPORATION** domiciliada en One Microsoft Way, Redmond, Washington 98052, Estados Unidos de América, atentamente me dirijo a ese Despacho para formular, dentro del término legal correspondiente, el recurso legal de reposición que me asiste en contra de la Resolución No. 41845 de fecha 13 de Diciembre de 2007, la cual fue notificada por Edicto desfijado el 10 de Enero de 2008.

Objeto del Recurso:

Por medio del presente escrito, me permito solicitar a ese Despacho proceda como se señala a continuación:

1. **Revoque** la decisión contenida en la Resolución N. 41845 de fecha 13 de Diciembre de 2007, la cual fue notificada por Edicto desfijado el 10 de Enero de 2008.
2. Que de considerar que aún con la respuesta contenida en este memorial aún se requiere de alguna aclaración, el Despacho proceda a rendir un **examen de forma** ó requerimiento, donde se requiera al solicitante para la presentación de la información y aclaraciones a que haya lugar.
3. Que de manera subsidiaria y si se decide no requerir formalmente para la presentación de las reivindicaciones originales, se estudie a favor de **MICROSOFT CORPORATION** la patente de la referencia, respecto de las 44 reivindicaciones contenidas en el texto cuya traducción ya reposa en este expediente.

Fundamentos de la providencia recurrida.

Afirma el Despacho que la patente bajo estudio debe ser declarada retirada, con base

en los siguientes argumentos:

"SEXTO: Que de la simple revisión de la presente solicitud, se ha constatado que el solicitante no realizó los actos mencionados en el artículo 39)1) dentro del plazo aplicable. Es así como el solicitante, para efectos de entrar a la fase nacional no presentó una traducción al castellano de la solicitud internacional tal y como fue presentada, con el contenido prescrito en el mencionado punto 5.2.5., es decir, con la descripción, las reivindicaciones, copia de los dibujos, el resumen, y de ser el caso, las traducciones previstas en razón a modificaciones.

Fundamentos del Recurso

Solicito al Despacho tener en cuenta que conforme al artículo 52 numeral 3 y artículo 56, del Código Contencioso Administrativo Colombiano, al momento de interponer recurso de reposición es posible hacer valer todos los medios probatorios señalados en el Código de Procedimiento Civil Colombiano. Para rápida referencia, a continuación se transcriben las normas invocadas:

"Art. 52.-Los recursos deberán reunir los siguientes requisitos:

(...)

3) Relacionar las pruebas que se pretende hacer valer. "

(...)

"Art.-57.- Serán admisibles todos los medios de prueba señalados en el Código de Procedimiento Civil."

Ahora bien, no comparto las consideraciones de ese Despacho en razón a lo siguiente:

1. De acuerdo con la interpretación teleológica del texto del Tratado de Cooperación de Patentes, la sanción del retiro de la solicitud de patente de invención debe declararse cuando por inacción del solicitante, se puede concluir que no existe interés de parte en el trámite correspondiente. Por tanto el retiro de la solicitud se aplica cuando quiera que se evidencia la existencia de un desinterés de la parte.

En relación con la declaratoria de retiro de la solicitud, ésta procede entre otros en aplicación del Artículo 37 del PCT.

Esta sanción de "declaración de retirada", de conformidad con la anterior norma, surge por el desinterés del solicitante en continuar y contestar los requerimientos de la entidad. Este desinterés se hace evidente cuando el

ALICIA
N
DI
DEL
DE
CAIRO
APRO
2014

mismo solicitante no da respuesta en todo o en parte a algún requerimiento de la entidad que se permitió formularlo.

2. Lo anterior resulta además de la lectura del Artículo 26 del mencionado tratado PCT mediante el cual se señala lo siguiente:

Artículo 26

Oportunidad de efectuar correcciones ante las Oficinas designadas

Ninguna Oficina designada podrá rechazar una solicitud internacional por el incumplimiento de los requisitos del presente Tratado y su Reglamento, sin dar previamente al solicitante la oportunidad de corregir dicha solicitud, en la medida y de conformidad con el procedimiento previsto por la legislación nacional para situaciones idénticas o comparables respecto de solicitudes nacionales.

Conforme a esta disposición, existe prohibición expresa por parte del Tratado de rechazar una solicitud internacional por el incumplimiento de requisitos de este Tratado y su reglamento, cuando no se haya dado previamente al solicitante la oportunidad procesal de corregir dicha solicitud, en la medida y de conformidad con el procedimiento previsto por la legislación nacional para situaciones idénticas ó comparables respecto de solicitudes nacionales.

En este punto, señalo a la Administración que la corrección de cualquier error o incumplimiento de requisitos del Tratado ó Reglamento, debe ser advertido oportunamente al solicitante y debe dársele, previamente al rechazo, la oportunidad procesal de subsanarlo ó corregirlo según lo dispuesto en la legislación interna. En este caso, según lo establecido en el Artículo 39 de la Decisión 486, es decir concediendo al solicitante un término de dos meses contados a partir del momento en que se ha notificado del requerimiento por parte de la Administración.

3. Es importante también tener en cuenta lo dispuesto en el Artículo 27 del PCT, según el cual:

Artículo 27. 1. "Ninguna legislación nacional podrá exigir que la solicitud internacional cumpla, en cuanto a su forma o contenido, con requisitos diferentes de los previstos en el presente Tratado y su reglamento y con requisitos adicionales".

4. En este caso, se cumplió con todos los requisitos del tratado PCT en relación con la entrada en fase nacional es decir se presentó la siguiente documentación según lo establecido en el Artículo 22, así:

Artículo 22

Copia, traducción y tasa para las Oficinas designadas

1) El solicitante proporcionará a cada Oficina designada una copia de la solicitud internacional (salvo que ya se haya efectuado la comunicación prevista en el Artículo 20) y una traducción de la misma (en la forma prescrita) y pagará la tasa nacional (si procede) antes del vencimiento del plazo de 30¹ meses desde la fecha de prioridad. Cuando la legislación nacional del Estado designado exija la indicación del nombre y otros datos prescritos relativos al inventor, pero permita que esas indicaciones se proporcionen con posterioridad a la presentación de una solicitud nacional, a menos que esas indicaciones ya figurasen en el petitorio, el solicitante deberá proporcionarlas a la Oficina nacional de ese Estado o que actúe en su nombre antes del vencimiento del plazo de 30¹ meses desde la fecha de prioridad.

En ningún aparte de este artículo se contiene la frase "tal como fue presentada" según lo que reiteradamente se señala en la Resolución aquí recurrida.

- 3.1. La solicitud de entrada en fase nacional contenía: (i) copia de la solicitud internacional (en inglés), (ii) una traducción de la misma (en la forma prescrita) y (iii) el pago de la tasa nacional. Todos estos documentos fueron presentados dentro del término de 30/31 meses establecidos por dicho tratado.
- 3.2. Ahora bien, señala la administración que la traducción en referencia no fue presentada tal como fue presentada la solicitud internacional. Lo anterior, es cierto. La traducción de la solicitud fue presentada incluyendo modificaciones que limitan el alcance de la solicitud internacional.

En este caso, y pese a lo que se afirma en la Resolución No. 41845 de fecha 13 de Diciembre de 2007, no se aplica directamente la declaratoria de retiro de la solicitud de patente, puesto que además del claro contenido del Artículo 26 de dicho tratado, ya antes transcrito, el Reglamento del PCT establece en su regla 49.5 lo siguiente:

49.5 *Contenido y requisitos materiales de la traducción*

a) A los efectos de lo dispuesto en el Artículo 22, la traducción de la solicitud internacional contendrá la descripción (sin perjuicio de lo dispuesto en el párrafo a-bis)), las reivindicaciones, cualquier texto contenido en los dibujos y el resumen.

La traducción presentada cumple exactamente con lo prescrito en tal regla, que es precisamente la aplicable, en cuanto el mismo artículo 22 establece que la traducción deberá presentarse en la forma prescrita.

En otras palabras, para que una traducción cumpla con la forma prescrita por el Artículo 22, esta debe contener lo siguiente:

- 3.2.1. La descripción.
- 3.2.2. Las reivindicaciones.
- 3.2.3. Cualquier texto contenido en los dibujos.
- 3.2.4. El resumen.

De lo anterior, reitero, la traducción presentada cumple con los requisitos prescritos para las traducciones según lo señalado en el texto del PCT.

Ahora bien, señala la misma Regla 49.5., a renglón seguido:

Además, si la Oficina designada lo exige, sin perjuicio de lo dispuesto en los párrafos b), c-*bis*) y e), la traducción también:

- i) contendrá el petitorio,
- ii) si las reivindicaciones han sido modificadas conforme al Artículo 19, contendrá las reivindicaciones tal como fueron presentadas y las reivindicaciones tal como fueron modificadas, e
- iii) irá acompañada de una copia de los dibujos.

El requisito contenido en el numeral ii) arriba transcrito se aplica respecto de solicitudes que en su fase **internacional** hayan sido modificadas; sin embargo, en el presente caso la situación es distinta por cuanto a la solicitud de la referencia en fase internacional **no** le fueron introducidas dichas modificaciones; éstas modificaciones sólo se realizaron al entrar en su fase **nacional** en **Colombia**. Por lo tanto, tampoco resulta aplicable lo establecido en el Artículo 19 del PCT.

Sin embargo aún, entendiendo que éste precepto resulta aplicable, el mismo texto del PCT establece la solución, en la eventualidad en que el solicitante no aporte la traducción de las reivindicaciones originales. Y esta solución como apreciará el Despacho dista de la declaratoria de retiro, sanción que está siendo aplicada al solicitante en el presente caso.

c-*bis*) Cuando el solicitante entregue una sola de las dos traducciones requeridas a una Oficina designada que, en aplicación del párrafo a)ii), exija la traducción de las reivindicaciones tal como fueron presentadas y tal como fueron modificadas, la Oficina designada podrá ignorar las reivindicaciones cuya traducción no se haya entregado o requerir al solicitante para que entregue la traducción que falte en un plazo que deberá ser razonable habida cuenta de las circunstancias y que se fijará en el requerimiento. Si la Oficina designada decide requerir al solicitante para que entregue la traducción que falte y no se entrega en el plazo fijado en el requerimiento, la Oficina designada podrá

En aplicación al párrafo a) ii) mencionado anteriormente, señala el PCT que cuando el solicitante presente solo una de las traducciones (bien la modificada ó la original), a una Oficina designada que exija la presentación de la traducción de las dos versiones de las reivindicaciones (como es el caso de la Oficina Colombiana según el contenido del punto 5.2.4. del Capítulo Quinto, del Título X, de la Circular Única de la SIC), la Oficina designada podrá

- (i) ignorar las reivindicaciones cuya traducción no se haya entregado, ó
- (ii) requerir al solicitante para que entregue la traducción que falte en un plazo que deberá ser razonable habida cuenta de las circunstancias y que se fijará en el requerimiento.

Tal como se aprecia de lo anterior, existe un mecanismo previsto, en el evento en que se presente traducción de unas reivindicaciones modificadas a las originalmente presentadas. Siempre mediante requerimiento según lo establecido en el Artículo 26.

No se prevé, que cuando se trate de los requisitos adicionales, establecidos en la Regla 49.5. a) i- iii) conlleve la declaratoria de retiro de la solicitud de patente, por el contrario, establece según el 49.5. *cbis* que deberá requerirse al solicitante en tal sentido. En este caso particular, el Despacho ha debido requerirme mediante un Oficio debidamente notificado la presentación de la traducción de las reivindicaciones como fueron presentadas en la solicitud internacional.

5. La entrada en fase nacional, de la patente en referencia cumplió con los requisitos del Artículo 22, es decir se presentó i) copia de la solicitud internacional (en inglés), (ii) una traducción de la misma (en la forma prescrita) y (iii) el pago de la tasa nacional.

La traducción cumplió con la forma prescrita para las traducciones, según la Regla 49.5., es decir contenía (i) una descripción, (ii) las reivindicaciones, (iii) el texto contenido en los dibujos y (iv) el resumen.

Las reivindicaciones, no obstante, fueron modificadas para entrada en fase nacional, por lo cual no se cumplió con lo prescrito en el punto 5.2.4 del Capítulo Quinto, del Título X, de la Circular Única de la SIC, caso en el cual, era necesario que ese Despacho requiriera al solicitante para que éste procediera con las aclaraciones y/ó modificaciones del caso y para el pago de tarifas a que hubiere lugar.

6. Insistimos, en que de acuerdo con la legislación vigente en materia de patentes (Decisión 486), la presentación en fase nacional de la patente en cuestión cumplió con los requisitos mínimos establecidos por la norma supranacional vigente y aplicable, lo que demuestra que no hubo inacción tal imputable al solicitante que conlleve la sanción máxima al solicitante como lo es la declaratoria de retiro de la solicitud según lo anteriormente transcrito.
7. Para mayor entendimiento del Despacho, y para mejor proceder, adjunto nuevamente, el texto de las reivindicaciones originalmente presentadas, el texto de la modificación y su correspondiente traducción y el recibo de la misma.

Solicito igualmente, que en caso de requerir una traducción de las reivindicaciones originales y en cumplimiento al Artículo 26 del PCT y a la Regla 49.5.c) *bis* se emita un requerimiento para que la misma sea presentada.

8. Igualmente me permito informar al Despacho, que la declaratoria de retiro de la presente patente, está siendo emitida, sin previo requerimiento al solicitante dos años y medio después de la fecha de su entrada en fase nacional.
9. Lo anterior implica que, a lo largo de estos 2 años y medio, no se le ha comunicado en ningún momento al solicitante, requerimiento alguno, que le indique la necesidad de cumplir con requisitos adicionales a los establecidos en el Artículo 22 del PCT, por lo cual sin la realización de un examen de forma en este asunto se están violando los términos establecidos en los Artículos 38 y 39 de la Decisión 486, que establece un término de treinta días contados a partir de la fecha de presentación para cualquier reparo de forma, para que el solicitante pueda ejercer acciones que legítimamente le hubieren correspondido de haber emitido el Despacho un examen de forma expedito.

Además, y mediante el presente escrito me permito ejercer la petición de restitución del derecho establecida en la Regla 49.6 del PCT. Informo que razonablemente no fue posible presentar la documentación adicional pues no existió conocimiento de tal situación, la que fue notificada mediante la declaratoria de retiro de la solicitud (mediante la notificación de la Resolución 41845) por parte de ese Despacho, sin que hubiera mediado requerimiento alguno.

10. Ahora bien, teniendo en cuenta la anterior respuesta integral, y habiendo demostrado el solicitante su interés en el registro y ante los asuntos todavía en cuestión por parte del Despacho, procede llevar a cabo un examen de forma donde se requiera al solicitante presentar de manera completa los documentos de cesión necesarios conforme a los inventores que figuran en la solicitud de la patente de invención de la referencia.

DE C
OTARIA
EISEIS
CIRCUL
BOGOT
MARTIN
IS EDO
1994

11. Insistimos en que la declaratoria de retiro, no obstante estar siendo aplicada sin atender los preceptos de la Regla 49.5, en su integridad, viola

- el derecho del debido proceso, al no existir posibilidad de defensa;
- el principio procesal donde el fondo prevalece sobre la forma, y que en la medida en que tales defectos formales puedan ser subsanados debe admitirse la posibilidad a favor de la persona para que procedan con tal corrección;
- la prohibición contenida en el Artículo 26 del Tratado de Cooperación en materia de Patentes PCT.

Conclusión

Por todo lo anterior solicito a ese Despacho:

1. **Revoque** la decisión contenida en la Resolución N. 41845 de fecha 13 de Diciembre de 2007, la cual fue notificada por Edicto desfijado el 10 de Enero de 2008.
2. Que de considerar que aún con la respuesta contenida en este memorial aún se requiere de alguna aclaración, el Despacho proceda a rendir un **examen de forma** ó requerimiento, donde se requiera al solicitante para la presentación de la información y aclaraciones a que haya lugar.
3. Que de manera subsidiaria y si se decide no requerir formalmente para la presentación de las reivindicaciones originales, se estudie a favor de **MICROSOFT CORPORATION** la patente de la referencia, respecto de las 44 reivindicaciones contenidas en el texto cuya traducción ya reposa en este expediente.

Anexos al recurso

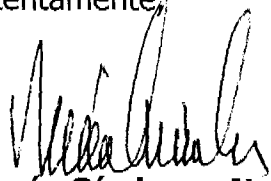
- Copia de las reivindicaciones en Inglés como se presentaron originalmente.
- Copia de la traducción al Español de las reivindicaciones originalmente presentadas.
- Copia del texto en Inglés de la solicitud modificada, incluyendo su (i) una descripción, (ii) las reivindicaciones, (iii) el texto contenido en los dibujos y (iv) el resumen de la misma.
- Copia del texto en traducido de la solicitud y reivindicaciones modificadas.
- Recibo oficial por concepto de petición de restablecimiento del derecho de la patente en referencia según lo establecido en la Regla 49.6.

Notificaciones

Recibiré notificaciones personalmente ante la Secretaría de ese Despacho ó en mi oficina de abogados situada en la siguiente Dirección:

Cárdenas & Cárdenas
Carrera 7 No. 71-52 Torre B. Piso 9
T. 312.36.00
Bogotá Colombia

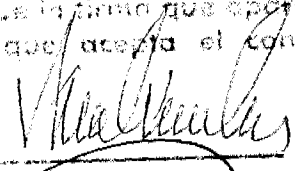
Atentamente,


Darío Cárdenas Navas
c.c. 17.066.629 de Bogotá
t.p. 1.250 C.S.J.

JGQ/recursos/P-344

COMPARACIÓN PERSONAL Y AUTENTICACION DE FIRMA

El Notario Dieciséis (16) (E) de Bogotá, da fe que el anterior escrito diligido a Supleniente de Ind. y Cio fue presentado personalmente por Darío Cárdenas Navas quien exhibió la c.c. 17.066.629 de Bta y 1.250 y manifestó que la firma que aparece en el presente documento es suya, y que acepta el contenido del mismo.



FIRMA
17 ENE. 2008
JOHN JAIRO MARTINEZ GONZALEZ
NOTARIO 16 ENCARGADO.



ESTA CERTIFICACION DE EXCEDENTE DE PAGO SOLO PODRA SER UTILIZADA
COMO PARTE DE LA CANCELACION DE UNA TARIFA VIGENTE Y NO PODRA SER
SUSTITUIDA POR NINGUN RECIBO.

190

U.S. DEPARTMENT OF COMMERCE
BUREAU OF ECONOMIC ANALYSIS
WASHINGTON, D. C. 20540

Fecha: 20/05/2014
Tra. 11
Act. 412

DEPOSITANTE	TIPO PAGO	BANCO	CUENTA	Nº. PAGO	FECHA PGO	Vr. PAGO
CARDENAS Y CARDENAS	CONSIGNACION	BANCO DE BOGOTA	062754387	2808794	03/01/2008	3,482,000.00

CANT.	RENTISTICO	CONCEPTO	TOTAL CONCEPTO
1	50005-01-04	PRESENTACION DE OBSERVACIONES	
		12 MARCAS Y LEMAS COMERCIALES	266,000.00
		TOTAL :	\$ 266,000.00

ay

RECIBO DE CAJA APLICADO AL EXPEDIENTE No. 05-0761673.

CARDENAS & CARDENAS
ABOGADOS

SUPERINTENDENCIA DE INDUSTRIA Y COMERCIO

NIT : 800176089-2

RECIBO OFICIAL DE CAJA : 07 - 99,581
FECHA : DICIEMBRE 6 DE 2007

SUPERINTENDENCIA DE INDUSTRIA Y COMERCIO



No. 05-076623- -00004-0000

Fecha: 2008-01-17 15:18:58 Dep. 2020 NUECREACIONI
Tra. 11 PATENTEIYII Eve: 379 FASENACIONALI
Act. 412 REPOSPRESRECU Folios: 154

**** CONSIGNACION ****

DEPOSITANTE	TIPO PAGO	BANCO	CUENTA	Nº. PAGO	FECHA PGO	Vr. PAGO
CARDENAS Y CARRENAS P.I. CONSIGNACION		BANCO POPULAR	050-00110-6	71023005	06/12/2007	11,808,100.00

**** CONCEPTO ****

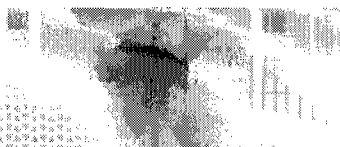
CANT. RENTISTICO	CONCEPTO	TOTAL CONCEPTO
1	30005-01-06 REGISTRO DE CESION, CANDID NOM	
	13 NOMBRES Y ENSEÑAS COMERCIALES, MAR	243,000.00
	TOTAL :	243,000.00

SON: DOSCIENTOS CUARENTA Y TRES MIL PESOS

RESPONSABLE : 
RECIBO DE CAJA APLICADO AL EXPEDIENTE No. 05-076623



IP SERVICES

[Home](#) [IP Services](#) [PATENTSCOPE®](#)

Search result: 1 of 1

(WO/2004/079514) FEEDBACK LOOP FOR SPAM PREVENTION

[Biblio. Data](#) [Description](#) [Claims](#) [National Phase](#) [Notices](#) [Documents](#)

Note: OCR Text

CLAIMS What is claimed is: 1. A system that facilitates classifying items in connection with spam prevention, comprising: a component that receives a set of the items; a component that identifies intended recipients of the items, and tags a subset of the items to be polled, the subset of items corresponding to a subset of recipients that are known spam fighting users; and a feedback component that receives information relating to the spam fighter's classification of the polled items, and employs the information in connection with training a spam filter and populating a spam list.

2. The system of claim 1, wherein the items comprise at least one of: electronic mail (email) and messages.
3. The system of claim 1, wherein the component that receives a set of the items is any one of an email server, a message server, and client email software.
4. The system of claim 1, wherein the subset of items to be polled comprises all of the items received.
5. The system of claim 1, wherein the subset of recipients comprises all recipients.
6. The system of claim 1, wherein the subset of recipients are randomly selected.
7. The system of claim 1, wherein the subset of recipients comprises paying users of the system.
8. The system of claim 1, wherein at least a subset of messages that would ordinarily be filtered are considered for polling.
9. The system of claim 1, wherein the subset of items tagged for polling is limited to at least one of the following: a number of the items selected per user; a number of the items selected per user per time period; and a probability of tagging an item corresponding to a known user.
10. The system of claim 1, wherein the tagged items are each assigned a unique ID, the unique ID corresponding to any one of the tagged item and contents of the tagged item.
11. The system of claim 1, further comprising a component that modifies an item tagged for polling to identify it as a polling item.
12. The system of claim 11, wherein the modified item comprises at least one of the following: a modified "from" address; a modified subject line; a polling icon; and a polling color for identification as a polling item.
13. The system of claim 11, wherein the modified item comprises the tagged item as an attachment.
14. The system of claim 11, wherein the modified item comprises a summary of the tagged item, the summary comprising at least one of a subject, a date, text of the message, and a first few lines of the text.
15. The system of claim 11, wherein the modified item comprises voting instructions and any one of at least two voting buttons and links which correspond to at least two respective classes of items facilitate classification of the item by the

~~193~~
193

user.

16. The system of claim 15, wherein the voting buttons correspond to respective links such that when any one of the voting buttons is selected by the user, information relating to the selected voting button, the respective user, and the item's unique ID assigned thereto is sent to a database for storage.

17. The system of claim 15, wherein the at least two voting buttons comprise a first voting button, the first voting button indicating "legitimate mail" and a second voting button, the second voting button indicating "spam".

18. The system of claim 15, wherein the voting buttons are implemented by modifying text of the item.

19. The system of claim 15, wherein the voting buttons are implemented by modifying a user interface of client email software.

20. The system of claim 1, further comprising a central database that stores information and data relating to user properties, item content and properties associated with tagged items, user classification and voting statistical data, frequency analysis data of polling per user and of polling per user per time period, spam lists, legitimate mail lists, and black hole lists.

21. The system of claim 1, wherein items tagged for polling that are marked as spam by an existing filter are delivered to the user's inbox and considered for polling.

22. The system of claim 1, wherein items tagged for polling are scanned for viruses such that one of the following occurs: detected viruses are stripped out and the items are polled; and infected items are discarded.

23. The system of claim 1 distributed across more than one spam-fighting company such that feedback from each company is sent to a central database operatively interfaced with each company, wherein some portion of the feedback is removed for privacy reasons.

24. The system of claim 23, wherein the company feedback comprises one of the following: only spam items, thereby excluding legitimate items; and spam items and sender name, domain name and IP address of legitimate items.

25. The system of claim 1, further comprising a user classification validation component that tests user reliability and trustworthiness.

26. The system of claim 25, wherein the user classification validation component is at least one of a cross-validation technique and a known result test message technique.

27. The system of claim 25, wherein the user classification validation component can be applied to one or more suspected users.

28. The system of claim 1, wherein the feedback component receives information relating to user feedback, honeypot feedback, and optionally, user recipient feedback of received items.

29. A server employing the system of claim 1.

30. An e-mail architecture employing the system of claim 1.

31. A computer readable medium having stored thereon the components of claim 1.

32. A method that facilitates classifying messages in connection with spam prevention comprising: receiving a set of the

16
184
194

messages; identifying intended recipients of the messages; tagging a subset of the messages to be polled, the subset of messages corresponding to a subset of the recipients that are known spam fighting users; receiving information relating to the user's classification of polling messages; and employing the information in connection with training a spam filter and populating a spam list.

33. The method of claim 32, wherein at least a subset of messages that would ordinarily be filtered are received by an email server and into a feedback loop system..

34. The method of claim 32, wherein all incoming messages are handled by client email software such that messages selected for polling are specific to preferences of an individual user.

35. The method of claim 32, wherein all messages received are considered for polling to mitigate bias of data.

36. The method of claim 32, wherein the subset of messages to be polled comprises all messages.

37. The method of claim 32, wherein the subset of recipients comprises all recipients.

38. The method of claim 32, wherein the subset of recipients which are known spam fighting users is determined by each recipient performing at least one of the following: opting in to provide feedback on messages to facilitate training a new spam filter; passively opting in to provide feedback on messages by not opting out; paying for email and message services provided by a participating message server; and opening an email account with a participating message server.

39. The method of claim 32, wherein the subset of users selected to participate in the message polling is selected at random.

40. The method of claim 32, wherein the subset of users selected to participate in the message polling are selected from all paying users, thereby making it more expensive for some spammers to subvert the spam filter training.

41. The method of claim 32, wherein the subset of messages tagged for polling are selected at random.

42. The method of claim 32, wherein the subset of messages tagged for polling is limited by one or more polling limits.

43. The method of claim 32, wherein the one or more polling limits comprise a per user limit and a per user per time period limit to mitigate bias of data.

44. The method of claim 32, further comprising modifying tagged messages to mark and identify them as polling messages.

45. The method of claim 44, wherein modifying the tagged messages comprises performing at least one of the following: moving the tagged message to a separate folder for polling messages; modifying the "from" address of the tagged message; modifying the subject line of the tagged message; using a polling icon on the tagged message to identify it as a polling message; and using a unique color to identify the tagged message as a polling message.

46. The method of claim 32, wherein the polling message comprises an attachment of the message as originally received and a set of instructions instructing the user on how to vote.

47. The method of claim 46, further comprising at least two voting buttons to facilitate classifying the message as spam and not spam.

48. The method of claim 47, wherein the voting buttons are links, which when selected by the user, create feedback that is employed in connection with training a spam filter and populating a spam list, the feedback comprising information relating

195
195

to the selected classification, the user, the message, and a unique ID assigned to one of the message and contents of the message.

49. The method of claim 47, further comprising a third voting button to opt out of future polling.

50. The method of claim 47, wherein the voting buttons are incorporated into the polling message by modifying text of the message before sending the polling message to the respective user.

51. The method of claim 47, wherein the voting buttons are implemented by modifying a user interface of client email software.

52. The method of claim 47, wherein the voting buttons are incorporated into the polling message.

53. The method of claim 32, further comprising scanning the tagged messages for viruses before they are downloaded for polling.

54. The method of claim 53, further comprising removing the viruses from any infected messages.

55. The method of claim 53, wherein tagged messages infected with a virus are discarded.

56. The method of claim 46, further comprising a summary of the message, the summary comprising at least one of a subject line, message sender, date the message was sent, date the message was received, and a first few lines of text from the message.

57. The method of claim 32, further comprising making a copy of each tagged message as originally received such that the respective users receive a first copy of the message in its original form and a second copy of the message in a form for polling.

58. The method of claim 32, wherein the tagged messages are individually assigned a unique ID corresponding to at least one the tagged message and contents of the tagged message.

59. The method of claim 58, wherein the tagged message and its associated ID are stored in a database in connection with training a spam filter and populating a spam list.

60. The method of claim 32, wherein a feedback component receives the information relating to the user's classification of the polling message, the feedback component comprising a central database.

61. The method of claim 60, wherein the database provides information in connection with training a spam filter and populating a spam list via a machine-learning technique.

62. The method of claim 32, wherein the spam filter is trained using messages classified as spam and not spam to mitigate bias of polling data and misclassification of the polling messages.

63. The method of claim 32, further comprising distributing the trained spam filter to one or more servers, the distribution occurring automatically and/or by request by at least one of an email message and a posting on a website for downloading.

64. The method of claim 60, wherein identifying users and tagging messages for polling is distributed across one or more mail servers and one or more client email software such that data generated by the mail servers and client email software are returned to a central database for storage in connection with training a spam filter and populating a spam list.

196

196

65. The method of claim 64, wherein key information is removed from any data that is sent to the central database by the mail servers and client email software for privacy reasons, such that only a portion of the data is sent to the central database to facilitate training the spam filter.

66. The method of claim 65, wherein the portion of data sent to the central database comprises at least one of the following: information relating to spam messages; domain names embedded in legitimate messages; and IP addresses embedded in legitimate messages.

67. The method of claim 64, wherein the data generated by the mail servers and the data generated by the client email software are aggregated into statistical data, respectively, corresponding to polling results and polling messages, thus mitigating bandwidth required to transmit the data to the central database.

68. The method of claim 32, wherein training the spam filter and populating the spam list is performed by machine learning techniques using data based on user classification feedback and optionally, data generated by one or more additional sources, the one or more sources comprising honeypots, recipient non-user classification feedback, and active learning techniques.

69. The method of claim 68, wherein data generated by the one or more sources is re-weighted proportionately with respect to the type of data generated by the source and relative to the user classification data to facilitate obtaining an unbiased sampling of data.

70. The method of claim 68, wherein honeypots correspond to email addresses disclosed in a restrictive manner such that it is known who is sending them legitimate messages, thereby facilitating immediate identification of spammers, verification of suspect merchants who distribute user subscriber information to spammers, and immediate classification of spam messages without waiting for user classification.

71. The method of claim 70, wherein information generated by the honeypots is down weighted selectively depending at least in part on a number of honeypots in use relative to a number of other sources including user classification feedback.

72. The method of claim 70, wherein data generated by the honeypots is integrated in real time into a central database, where information relating to user classifications and polling messages are also stored for later use in connection with training a spam filter and populating a spam list.

73. The method of claim 67, wherein the messages are selected using active learning techniques, i. e. techniques that select messages based on their estimated value to learning new or updated filters.

74. The method of claim 32 further comprising: monitoring incoming messages for their respective one or more positive features; determining a frequency of positive features received; determining whether one or more positive features received exceeds a threshold frequency based at least in part upon historical data; and quarantining suspicious messages, which correspond to the one or more positive features that exceed the threshold frequency, until further classification data is available to determine whether suspicious messages are spam.

75. The method of claim, 74 wherein the feature used is information about the sender comprising at least one of the sender's IP address and domain.

76. The method of claim 74, wherein quarantining suspicious messages is performed by at least one of the following acts: provisionally labeling the suspicious messages as spam and moving them to a spam folder; delaying delivery of the suspicious messages to the user (s) until further classification data is available; and storing the suspicious messages in a folder not visible to the user (s).

77. The method of claim 32, further comprising determining false positive and catch rates of the spam filter to facilitate optimization of the spam filter, wherein determining false positive and catch rates comprises: training the spam filter using a training data set, the training data set comprising a first set of polling results; classifying a second set of polling

197
197

messages using user feedback to yield a second set of polling results; running the second set of polling messages through the trained spam filter; comparing the second set of polling results to the trained spam filter results to determine false positive and catch rates of the filter to thereby evaluate and tune filter parameters according to optimal filter performance.

78. The method of claim 77, wherein more than one spam filter is built, each having different parameters and each being trained on the same training data set, such that the false positive and catch rates of each spam filter is compared to at least one other spam filter to determine optimal parameters for spam filtering.

79. The method of claim 32, further comprising building an improved spam filter using additional sets of incoming messages, subsets of which are subjected to polling to yield new information in connection with training the improved spam filter, wherein previously obtained information is re-weighted based at least in part upon how long ago it was obtained.

80. The method of claim 32, further comprising employing the information to build a legitimate sender list.

81. The method of claim 80, wherein the legitimate sender list comprises any one of IP addresses, domain names, and URLs that are substantially classified as sources of good mail according to a percentage of messages classified as good.

82. The method of claim 32, wherein the spam lists are used to generate a black hole list of addresses from whom no mail would be accepted.

83. The method of claim 32, further comprising employing the information to facilitate terminating accounts of spammers.

84. The method of claim 83, further comprising identifying a spammer who is using an ISP and automatically notifying the ISP of the spamming.

85. The method of claim 83, further comprising identifying a domain responsible for sending spam, and automatically notifying at least one of the domain's email provider and the domain's ISP of the spamming.

86. The method of claim 32, further comprising distributing at least one of the spam filter and the spam list to any one of mail servers, email servers, and client email software, wherein distributing comprises at least one of the following: posting a notification on a website notifying that the spam filter and spam list are available for downloading; automatically pushing the spam filter and spam list out to mail servers, email servers and client email software; and manually pushing the spam filter and spam list out to mail servers, email servers, and client email software.

87. A cross-validation method that facilitates verifying reliability and trustworthiness of user classifications comprising: excluding one or more suspected user's classifications from data employed to train the spam filter; training the spam filter using all other available user classifications; and running the suspected user's polling messages through the trained spam filter to determine how it would have classified the messages compared to the suspected user's classifications.

88. The method of claim 87, further comprising performing at least one of the following: discounting existing and future classifications provided by users who are determined to be untrustworthy until the users are determined to be trustworthy; discarding existing classifications provided by users determined to be untrustworthy; and removing the untrustworthy users from future polling.

89. A method that facilitates verifying reliability and trustworthiness in user classifications for training a spam filter via a feedback loop system comprising: identifying a subset of spam-fighting users as suspect users; providing one or more messages having a known result to the suspect users for polling; and determining whether the suspected users' classification of the one or more test messages matches the known classification to ascertain the reliability of the users' classifications.

90. The method of claim 89, wherein the subset of spam-fighting users identified as suspect users comprises all users.

28
198
198

91. The method of claim 89, wherein the message is a test message that is known to be at least one of spam and good mail and that is injected into a stream of incoming mail by the feedback loop system and delivered to the suspect users.

92. The method of claim 89, wherein the message received by the suspected users for polling is hand-classified by a system administrator to train the spam filter with a correct classification to identify untrustworthy users.

93. The method of claim 89, further comprising at least one of the following acts: discounting existing and future classifications provided by users who are determined to be untrustworthy until the users are determined to be trustworthy; discarding existing classifications provided by users determined to be untrustworthy; and removing the untrustworthy users from future polling.

94. A computer-readable medium having stored thereon the following computer executable components: a component that receives a set of messages; a component that identifies intended recipients of the messages, and tags a subset of the messages to be polled, the subset of messages corresponding to a subset of recipients that are known spam fighting users; a message modification component that modifies the tagged messages to identify them as polling messages to users; and a feedback component that receives information relating to the user's classification of the polled messages, and employs the information in connection with training a spam filter and populating a spam list.

95. A system that facilitates classifying messages in connection with spam prevention comprising: means for receiving a set of the messages; means for identifying intended recipients of the messages; means for tagging a subset of the messages to be polled, the subset of messages corresponding to a subset of the recipients that are known spam fighting users; means for receiving information relating to the user's classification of the polling messages; and means for employing the information in connection with training a spam filter and populating a spam list.

REIVINDICACIONES

Lo que se reivindica es:

1. Un sistema que facilita la clasificación de ítems en conexión con la prevención de spam, y que comprende:

Un componente que recibe un conjunto de los ítems;

Un componente que identifica receptores pretendidos de los ítems, y marca un subconjunto de los ítems para ser consultados, el subconjunto de los ítems correspondientes a un subconjunto de receptores que son conocidos como usuarios luchadores contra el spam; y

Un componente de retroalimentación que recibe información relacionada con la clasificación del luchador contra el spam del ítem consultado, y utilizan la información en conexión con el entrenamiento de un filtro para spam y poblando una lista de spam.

2. el sistema de la reivindicación 1, donde los ítems comprenden por lo menos uno de: correo electrónico (e-mail) y mensajes.
3. El sistema de la reivindicación 1, donde el componente que recibe el conjunto de ítems es cualquier servidor de e-mail, servidor de mensajes, y software de e-mail de un cliente.
4. El sistema de la reivindicación 1, donde el subconjunto de ítems que se va a consultar comprende todos los ítems recibidos.
5. El sistema de la reivindicación 1, donde el subconjunto de receptores comprende todos los receptores.
6. El sistema de la reivindicación 1, donde el subconjunto de receptores se selecciona en forma aleatoria.
7. El sistema de la reivindicación 1, donde el subconjunto de receptores comprende los usuarios que hayan pagado por el sistema.
8. El sistema de la reivindicación 1, donde al menos un subconjunto de mensajes que ordinariamente hubieren sido filtrados se consideren para consulta.

9. El sistema de la reivindicación 1, donde el subconjunto de ítems marcados para consulta es limitado a por lo menos uno de los siguientes:

un número de los ítems seleccionados por el usuario;
un número de los ítems seleccionados por el usuario por período de tiempo; y
una probabilidad de marcar un ítem correspondiente a un usuario conocido.

10. El sistema de la reivindicación 1, donde los ítems marcados son asignados a un identificador único ID, el ID único corresponde a cualquiera de los ítems marcados y al contenido del ítem marcado.

11. El sistema de la reivindicación 1, que además un componente que modifica el ítem marcado para consulta para identificarlo como un ítem de consulta.

12. El sistema de la reivindicación 11, donde el ítem modificado comprende al menos una de las siguientes:

una dirección "de" modificada;
una línea de asunto modificada;
un color de consulta para identificación de ítem de consulta.

13. El sistema de la reivindicación 11, donde el ítem modificado comprende el ítem marcado como un anexo.

14. El sistema de la reivindicación 11, donde el ítem modificado comprende un resumen del ítem marcado, el resumen comprende por lo menos un asunto, una fecha, un texto de mensaje, y unas primeras líneas de texto.

15. El sistema de la reivindicación 11, donde el ítem modificado comprende las instrucciones de votación y uno cualquiera de por lo menos dos botones y vínculos que corresponden a por lo menos dos clases respectivas de ítem que facilita la clasificación del ítem por usuario.

16. El sistema de la reivindicación de la reivindicación 15, donde los botones de votación corresponden a los vínculos respectivos tal que cuando cualquiera de los botones de votación es seleccionado por el usuario, información relativa al botón de votación, el respectivo usuario, y el ID único del ítem asignado al mismo es enviado a una base de datos para almacenamiento.

- 17.El sistema de la reivindicación 15, donde al menos dos botones de votación comprenden un primer botón de votación, el primer botón de votación indicando "mail legítimo" y un segundo botón de votación, el segundo botón de votación indicando "spam".
- 18.El sistema de la reivindicación 15, cuando los botones de votación son implementados mediante modificación del texto del ítem.
- 19.El sistema de la reivindicación 15, donde los botones de votación son implementados por la modificación de un interfase de usuario de software de correo electrónico de cliente.
- 20.El sistema de la reivindicación 1, comprendiendo además una base de datos central que almacena información y datos relacionados con las propiedades del usuario, cuando del ítem, propiedades asociadas con los ítems marcados, clasificación del usuario y datos estadísticos de votación, datos de análisis de frecuencia de consulta por usuario y de consulta por usuario por período de tiempo, listas de spam, listas de correo ilegítimo, y listas de hoyo negro.
- 21.El sistema de la Reivindicación 1, donde los ítems marcados para consulta que son marcados como spam por un filtro existente son enviados al buzón de entrada del usuario y son considerados para consulta.
- 22.El sistema de la Reivindicación 1, donde los ítems marcados para consulta son escaneados para detección de virus manera tal que uno de los siguientes ocurran:

los virus detectados sean despojados y que el ítem sea consultado;
y los virus infectados sean descartados.
- 23.El sistema de la reivindicación 1 distribuído a través de mas de una compañía luchadora contra el spam de tal manera que cada compañía se envía a una base de datos central en interfase operativa con cada compañía, donde algunas porciones de la retroalimentación se remueven por razones de privacidad.
- 24.El sistema de la reivindicación 23, donde la retroalimentación de la compañía comprende una de las siguientes:

únicamente ítems de spam, por tanto excluyendo los ítems legítimos;
y sitem de spam y nombre de remitente, nombre de dominio y dirección IP de ítems legítimos.

25. El sistema de la reivindicación 1, donde el componente de validación de la clasificación del usuario se puede aplicar a uno o más usuarios sospechosos.
26. El sistema de la reivindicación 25, donde el componente de validación de la clasificación es al menos una de las técnicas de validación cruzada y un resultado de testeo conocido en la técnica.
27. El sistema de la reivindicación 25, donde el componente de validación de la clasificación del usuario se puede aplicar a uno o más usuarios sospechosos.
28. El sistema de la reivindicación 1, donde el componente de retroalimentación recibe información relacionada con la retroalimentación de un usuario, la retroalimentación de un "honeypot" y opcionalmente, la retroalimentación del usuario receptor de los ítems recibidos.
29. Un servidor que emplea el sistema de la reivindicación 1.
30. Una arquitectura de correo electrónico empleando el sistema de la reivindicación 1.
31. Un medio legible por computador en el que se haya almacenado en el mismo los componentes de la reivindicación 1.
32. Un método que facilita la clasificación de los mensajes en conexión con la prevención de spam que comprende:
 - recibir un conjunto de mensajes;
 - identificar los receptores pretendidos de los mensajes;
 - marcar un subconjunto de los mensajes que se van a consultar, un subconjunto de los mensajes correspondientes a un subconjunto de los receptores que son conocidos como usuarios que luchan contra el spam;
 - recibir información relacionada con la clasificación del usuario de mensajes de consulta;
 - y emplear la información en conexión con el entrenamiento de un filtro de spam y poblar una lista de spam.
33. El método de la reivindicación 32, en donde al menos un subconjunto de mensajes que podrían ordinariamente ser filtrados son recibidos por un servidor de correo electrónico y en un sistema de retroalimentación cíclica.
34. El método de la reivindicación 32, en donde todos los mensajes entrantes son manejados por el software de correo electrónico del cliente de

manera que los mensajes seleccionados para consulta son específicos a las preferencias de un usuario individual.

35. El método de la reivindicación 32, donde todos los mensajes recibidos son considerados para consulta de manera que mitigue la tendencia de los datos.
36. El método de la reivindicación 32, donde un subconjunto de mensajes a ser consultado comprende todos los mensajes.
37. El método de la reivindicación 32, donde el subconjunto de receptores comprende todos los receptores.
38. El método de la reivindicación 32, donde el subconjunto de receptores que son conocidos como luchadores contra el spam se determina por cada receptor ejecutando por lo menos uno de los siguientes:
 - adhiriéndose para suministrar retroalimentación en los mensajes para facilitar el entrenamiento de un nuevo filtro spam;
 - adhiriéndose en forma pasiva para suministrar retroalimentación en los mensajes al no retirarse;
 - pagar por servicios de inicio y mensajería suministrados por un servidor de mensajes participante; y
 - abrir una cuenta de correo electrónico en un servidor de mensajería participante.
39. El método de la reivindicación 32, en donde el subconjunto de usuarios seleccionado para participar en la consulta del mensaje de consulta es seleccionado de manera aleatoria.
40. El método de la reivindicación 32, donde el subconjunto de usuarios seleccionados para participar en la consulta del mensaje es seleccionado de todos los usuarios que pagan, así haciéndola mas costosa para algunos remitentes de spam para subvertir el entrenamiento de filtro spam.
41. El método de la reivindicación 32, donde el subconjunto de mensajes marcados para consulta son seleccionados aleatoriamente.
42. El método de la reivindicación 32, donde el subconjunto de mensajes marcado para consulta está limitado por uno o más límites de consulta.
43. El método de la reivindicación 32, donde el uno o más límites de consulta comprende un límite por usuario u un período de tiempo límite por usuario para mitigar la tendencia de los datos.

- 44.El método de la reivindicación 32, que comprende además los mensajes marcados para marcarlos e identificarlos como mensajes de consulta.
- 45.El método de la reivindicación 44, donde la modificación de los mensajes marcados comprende la ejecución de por lo menos uno de los siguientes:
- mover el mensaje marcado a un folder separado para mensajes de consulta;
 - modificar la dirección "de" del mensaje marcado;
 - modificar la línea del asunto del mensaje marcado;
 - utilizar un icono de consulta en el mensaje marcado para identificarlos como mensaje de consulta; y
 - utilizar un color único para identificar el mensaje marcado como mensaje de consulta.
- 46.El método de la reivindicación 32, donde el mensaje de consulta comprende un anexo del mensaje como fue originalmente recibido y un conjunto de instrucciones instruyendo al usuario sobre como votar.
- 47.El método de la reivindicación 46, que comprende además al menos dos botones de votación para facilitar la clasificación del mensaje como spam ó como no-spam.
- 48.El método de la reivindicación 47, donde los botones de votación son vínculos, que cuando son seleccionados por el usuario, crean una retroalimentación que es empleada en conexión con el entrenamiento del filtro spam y poblando una lista spam, la retroalimentación comprende información relacionada con la clasificación seleccionada, el usuario, el mensaje, y un ID único asignado a uno de los mensajes y contenidos del mensaje.
- 49.El método de la reivindicación 47, que además comprende un tercer botón de votación para otra fuera de futuras consultas.
- 50.El método de la reivindicación 47, donde los botones de votación son incorporados en el mensaje de consulta mediante la modificación del texto del mensaje antes de enviar el mensaje de consulta al respectivo usuario.
- 51.El método de la reivindicación 47 donde los botones de votación son implementados mediante la modificación de una interfase de usuario del software de correo electrónico del cliente.

27
~~205~~
205

- 52.El método de la reivindicación 47, donde los botones de votación son incorporados en el mensaje de consulta.
- 53.El método de la reivindicación 32, que además comprende el escaneo o barrido de los mensajes marcados para virus antes de que sean descargados para consulta.
- 54.El método de la reivindicación 53, que comprende además la remoción de los virus de un mensaje infectado.
- 55.El método de la reivindicación 53, donde los mensajes infectados marcados con un virus son descartados.
- 56.El método de la reivindicación 46, que además comprende un resumen del mensaje, el resumen comprendiendo al menos una de las líneas del asunto, remitente del mensaje, fecha en que el mensaje fue enviado, fecha en que el mensaje fue recibido, y unas primeras pocas líneas de texto del mensaje.
- 57.El método de la reivindicación 32, que comprende además hacer una copia de cada mensaje marcado como fue recibido originalmente de manera que los usuarios respectivos reciben una primera copia del mensaje en su forma original y una segunda copia del mensaje en una forma para consulta.
- 58.El método de la reivindicación 32, donde los mensajes marcados son asignados individualmente a una ID única correspondiente a al menos uno de los mensajes marcado y los contenidos del mensaje marcado.
- 59.El método de la reivindicación 58, donde los mensajes marcados y su ID asociado sean almacenados en una base de datos en conexión con el entrenamiento del filtro spam y la población de lista spam.
- 60.El método de la reivindicación 32, donde un componente de retroalimentación recibe la información relativa a la clasificación del usuario del mensaje de consulta, el componente de la retroalimentación comprendiendo una base de datos central.
- 61.El método de la reivindicación 60, donde la base de datos proporciona información en conexión con el entrenamiento del filtro spam y la población de la lista spam mediante una técnica de aprendizaje de máquina.

62. El método de la reivindicación 32, donde el filtro spam es entrenado usando mensajes clasificados como spam y no-spam para mitigar la tendencia de los datos de consulta y la clasificación errada de los mensajes de consulta.
63. El método de la reivindicación 32, que además comprende la distribución del filtro spam entrenado a uno ó más servidores, la distribución ocurre automáticamente y/o por solicitud de al menos un mensaje de correo electrónico y un aviso en un sitio web para la descarga.
64. El método de la reivindicación 60, donde los usuarios identificadores y los mensajes marcadores para consulta es distribuida a través de uno o más servidores de correo y uno o más softwares de correo electrónico de cliente de manera que los datos generados por los servidores de correo y el software de correo electrónico de cliente sean devueltos a una base de datos central para almacenamiento en conexión con el entrenamiento de un filtro spam y la población de una lista spam.
65. El método de la reivindicación 64, donde la información clave es retirada de cualesquiera datos que sean enviados a la base de datos central mediante los servidores de correo y el software de correo de cliente por razones de privacidad, de manera que solo una porción de los datos es enviada a la base de datos central para facilitar el entrenamiento del filtro spam.
66. El método de la reivindicación 65, donde la porción de los datos enviados a la base de datos central comprende al menos una de las siguientes:
- información relativa a los mensajes de spam;
 - nombres de dominio incrustados en mensaje legítimos;
 - direcciones IP incrustadas en mensajes legítimos.
67. El método de la reivindicación 64, donde los datos generados por los servidores de correo y los datos generados por el software de correo de cliente son agregados a los datos estadísticos, respectivamente, correspondiendo a resultados de consulta y mensajes de consulta, en consecuencia mitigando el ancho de banda requerido para transmitir los datos a la base de datos central.
68. El método de la reivindicación 32, donde el entrenamiento del filtro spam y la población de la lista spam es realizada por una técnica de aprendizaje de máquina usando datos basados en retroalimentación de la clasificación de usuario y opcionalmente, datos generados por una ó más fuentes adicionales, la una ó más fuentes comprendiendo

28
~~201~~
207

"honeypots", retroalimentación de clasificación de receptores no-usuarios, y técnicas de aprendizaje activo.

69. El método de la reivindicación 68, en donde los datos generados por la una o más fuentes son re-pesados proporcionalmente en relación con el tipo de dato generado por la fuente y relativo a los datos de clasificación de usuario para facilitar la obtención de una muestra de datos sin tendencia.
70. El método de la reivindicación 68, en donde los honeypots corresponden a direcciones de correo electrónico reveladas en una manera restrictiva tal que se conoce quien envía mensajes legítimos, así facilitando la inmediata identificación de los remitentes de spam, verificación de mercaderes sospechosos que distribuyen información de suscripción de usuario a remitentes de spam, y clasificación inmediata de los mensajes spam sin esperar a la clasificación de usuario.
71. El método de la reivindicación 70, en donde a la información generada por los honeypots se le baja el peso selectivamente dependiendo al menos en parte de un número de honeypots en uso relativo al número de otras fuentes incluyendo retroalimentación de clasificación de usuarios.
72. El método de la reivindicación 70, en donde los datos generados por los honeypots son integrados en tiempo real a una base de datos central, donde la información relativa a las clasificaciones de usuario y a mensajes de consulta son también almacenados para su uso posterior en conexión con el entrenamiento de un filtro spam y una población de lista spam.
73. El método de la reivindicación 67 donde los mensajes son seleccionados usando técnicas de aprendizaje activo, p.ej. técnicas que seleccionen mensajes basados en su valor estimado para aprender filtros nuevos ó actualizados.
74. El método de la reivindicación 32, que además comprende:
 - el monitoreo de mensajes entrante para sus respectivos uno o más rasgos positivos;
 - determinar la frecuencia de rasgos positivos recibidos;
 - determinar si uno o más rasgos positivos recibidos exceden un umbral de frecuencia basado por lo menos en parte sobre los datos históricos; y
 - poner los mensajes sospechosos en cuarentena lo que corresponde a uno o más datos positivos que exceden el umbral de frecuencia, hasta que datos de clasificación adicionales estén disponibles para determinar si el mensaje sospechoso es spam.

75.El método de la reivindicación 74, donde los rasgos utilizados son información acerca del remitente que comprenden por lo menos una de las direcciones IP del remitente y su dominio.

76.El método de la reivindicación 74, en donde la cuarentena de los mensajes sospechosos la ejecuta por lo menos uno de los siguientes actos:

marcar provisionalmente el mensaje sospechoso como spam y moverlo a un folder de spam;

demorar la entrega del mensaje sospechoso al usuario hasta que datos de clasificación estén disponibles; y

almacenar el mensaje sospechoso en un folder que no es visible al usuario.

77.El método de la reivindicación 32, que comprende además la determinación de falso positivo y tasa de captura del filtro del filtro spam para facilitar la optimización del filtro spam, donde determinar el falso positivo y la tasa de captura comprende:

entrenar al filtro spam utilizando un conjunto de datos de entrenamiento, el conjunto de datos de entrenamiento comprende un primer conjunto de resultados de la consulta;

clasificar un segundo conjunto de mensajes de consulta utilizando la retroalimentación del usuario para generar un segundo conjunto de resultados de consulta;

correr el segundo conjunto de mensajes de consulta a través del filtro entrenado de spam;

comparar el segundo conjunto de resultado de consulta con los resultados del filtro de spam entrenado para determinar el falso positivo y la tasa de captura del filtro evaluando de esta manera y sintonizando los parámetros del filtro de acuerdo con una ejecución óptima del filtro.

78.El método de la reivindicación 77, donde más de un filtro spam se construye, cada uno teniendo distintos parámetros y cada uno entrenándose con el mismo conjunto de datos de entrenamiento, de tal manera que el falso positivo y la tasa de captura de cada filtro spam se compara con por lo menos otros filtros para determinar los parámetros óptimos para la filtración de spam.

79.El método de la reivindicación 32, que además comprende la construcción de un filtro de spam mejorado utilizando cómputos adicionales de mensajes entrantes, subconjuntos de éstos que se someten a consulta

para obtener información nueva con relación al entrenamiento del filtro de spam mejorado, donde la información anterior obtenida se reevalúa basada en parte en el tiempo en el que se obtuvo.

- 80.El método de la reivindicación 32, que comprende además la utilización de información para construir una lista legítima de remitentes.
- 81.El método de la reivindicación 80, donde la lista legítima de remitentes comprende cualquiera de las direcciones IP, nombres de dominio y URL's que son clasificadas sustancialmente como fuentes de buen correo de acuerdo con el porcentaje de mensajes clasificados como buenos.
- 82.El método de la reivindicación 32, donde las listas spam son utilizadas para generar una lista de direcciones de hoyo negro desde la cual ningún correo será aceptado.
- 83.El método de la reivindicación 32, que además comprende el empleo de información para facilitar la terminación de cuentas de remitentes de spam.
- 84.El método de la reivindicación 83, que además comprende la identificación de un remitente de spam que esté utilizando una ISP y notificando automáticamente a la ISP del envío de spam.
- 85.El método de la reivindicación 83, que además comprende la identificación del dominio responsable del envío de spam, y automáticamente notifica al menos a uno de los proveedores de dominios de correo y al ISP de los dominios respecto del envío de spam.
- 86.El método de la reivindicación 32, que además comprende la distribución de por lo menos un filtro spam y una lista de spam a cualquier servidor de correo, servidor de correo electrónico, ó software de correo electrónico del cliente, donde la distribución comprende por lo menos uno de los siguientes:

 fijar una notificación en la red notificando que los filtros spam y la lista de spam están disponibles para descarga;
 automáticamente enviando el filtro spam y la lista spam al servidor de correo, servidor de correo electrónico, y software de correo electrónico del cliente; y
 manualmente enviando una lista de filtro spam al servidor de correo, servidor de correo electrónico y software de correo electrónico del cliente.

87. Un método de validación cruzada que facilita la verificación de la confiabilidad de la clasificación de un usuario y que comprende:

excluir una o mas de las clasificaciones de datos de un usuario sospechoso en el entrenamiento del filtro spam;
entrenar los filtros spam utilizando todas las demás clasificaciones disponibles de usuario;
y corriendo los mensajes de consulta del usuario sospechoso a través del filtro entrenado de spam para determinar como hubiera clasificado el mensaje comparado a la clasificación del usuario sospechoso.

88. El método de la reivindicación 87, comprende además la ejecución de por lo menos uno de los siguientes:

descontando clasificaciones existentes y futuras suministradas por usuarios que se determinan no confiables hasta que el usuario se determine confiable;
descartar las indicaciones existentes suministrada por usuarios determinados como no confiables; y
removiendo los usuarios no confiables de consultas futuras.

89. Un método que facilita la verificación de la confiabilidad de las clasificaciones de un usuario para entrenar el filtros spam a través de un sistema de retroalimentación cíclica que comprende:

identificar un subconjunto de usuarios luchadores contra el spam como usuarios sospechosos;
suministrar uno o más mensajes con resultados conocidos al usuario sospechoso para consulta; y
determinar si la clasificación del usuario sospechoso de uno o más mensajes de prueba concuerda con la clasificación conocida para verificar la confiabilidad de las clasificaciones del usuario.

90. El método de la reivindicación 89, donde el subconjunto de usuarios luchadores contra el spam identificados como sospechosos comprende a todos los usuarios.

91. El método de la reivindicación 89, donde el mensaje de prueba es un mensaje conocido como por lo menos uno spam y uno correo bueno y que se inserta en el flujo de correo entrante por el sistema de retroalimentación cíclica y se entrega al usuario sospechoso.

92. El método de la reivindicación 89, donde el mensaje que recibe el usuario sospechoso para consulta se clasifica a mano por un administrador del

sistema para entrenar el filtro de spam con una clasificación correcta para identificar a los usuarios no confiables.

93. El método de la reivindicación 89, comprende además por lo menos uno de los siguientes actos:

descontar clasificaciones existentes y futuras suministradas por un usuario que se determina que no es confiable hasta que el usuario se considere confiable;

descartar clasificaciones existentes suministradas por un usuario que se determina que no es confiable; y el usuario que no es confiable de consultas futuras.

94. Un medio legible por computador que tiene almacenado en el mismo los siguientes componentes ejecutables por computador:

un componente que recibe un conjunto de mensajes;

un componente que identifica los receptores pretendidos de los mensajes, y marca un subconjunto de los mensajes a ser consultados, un subconjunto de mensajes que corresponden al subconjunto de receptores que son conocidos como usuarios luchadores contra spam;

un componente de modificación de mensaje que modifica los mensajes marcados para identificarlos como mensajes de consulta para usuarios;

un componente de retroalimentación que recibe información relativa a la clasificación de usuario de los mensajes de consulta, y emplea la información en conexión con el entrenamiento de un filtro spam y una población de lista spam.

95. Un sistema que facilita la clasificación de mensajes en conexión con la prevención de spam que comprende:

medios para recibir un conjunto de mensajes; medios para identificar los receptores pretendidos de los mensajes; medios para marcar un subconjunto de mensajes a ser consultados, un subconjunto de mensajes que corresponden al subconjunto de los receptores que son conocidos como usuarios luchadores contra el spam;

medios para recibir información relativa a la clasificación de usuario de mensajes de consulta; y

medios para emplear la información en conexión con el entrenamiento del filtro spam y la población de una lista spam.

~~34~~
212
212

Atty. Docket No. MS303500.2 (MSFTP416WO)

FEEDBACK LOOP FOR SPAM PREVENTION

by

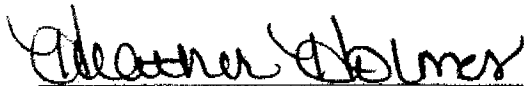
MICROSOFT CORPORATION

CERTIFICATION

I hereby certify that the attached patent application (along with any other paper referred to as being attached or enclosed) is being deposited with the United States Postal Service on this date February 25, 2004, in an envelope as "Express Mail Post Office to Addressee" Mailing Label Number EV330020761US addressed to the: MAIL STOP PCT, Commissioner for Patent, P. O. Box 1450, Alexandria, VA 22313-1450

Heather Holmes

(Typed or Printed Name of Person Mailing Paper)



(Signature of Person Mailing Paper)

213

Title: FEEDBACK LOOP FOR SPAM PREVENTION

TECHNICAL FIELD

5 This invention is related to systems and methods for identifying both legitimate (*e.g.*, good mail) and undesired information (*e.g.*, junk mail), and more particularly to classifying electronic mail correspondence for spam prevention.

BACKGROUND OF THE INVENTION

10 The advent of global communications networks such as the Internet has presented commercial opportunities for reaching vast numbers of potential customers. Electronic messaging, and particularly electronic mail ("email"), is becoming increasingly pervasive as a means for disseminating unwanted advertisements and promotions (also denoted as "spam") to network users.

15 The Radicati Group, Inc., a consulting and market research firm, estimates that as of August 2002, two billion junk e-mail messages are sent each day - this number is expected to triple every two years. Individuals and entities (*e.g.*, businesses, government agencies) are becoming increasingly inconvenienced and oftentimes offended by junk messages. As such, junk e-mail is now or soon will become a major threat to trustworthy computing.

20 A key technique utilized to thwart junk e-mail is employment of filtering systems/methodologies. One proven filtering technique is based upon a machine learning approach - machine learning filters assign to an incoming message a probability that the message is junk. In this approach, features typically are extracted from two classes of example messages (*e.g.*, junk and non-junk messages), and a learning filter is applied to
25 discriminate probabilistically between the two classes. Since many message features are related to content (*e.g.*, words and phrases in the subject and/or body of the message), such types of filters are commonly referred to as "content-based filters".

30 Some junk/spam filters are adaptive, which is important in that multilingual users and users who speak rare languages need a filter that can adapt to their specific needs. Furthermore, not all users agree on what is and is not, junk/spam. Accordingly, by

employing a filter that can be trained implicitly (*e.g.*, via observing user behavior) the respective filter can be tailored dynamically to meet a user's particular message identification needs.

One approach for filtering adaptation is to request a user(s) to label messages as junk and non-junk. Unfortunately, such manually intensive training techniques are undesirable to many users due to the complexity associated with such training let alone the amount of time required to properly effect such training. In addition, such manual training techniques are often flawed by individual users. For example, subscriptions to free mailing lists are often forgotten about by users and thus, are incorrectly labeled as junk mail. As a result, legitimate mail is blocked indefinitely from the user's mailbox. Another adaptive filter training approach is to employ implicit training cues. For example, if the user(s) replies to or forwards a message, the approach assumes the message to be non-junk. However, using only message cues of this sort introduces statistical biases into the training process, resulting in filters of lower respective accuracy.

Still another approach is to utilize all user(s) e-mail for training, where initial labels are assigned by an existing filter and the user(s) sometimes overrides those assignments with explicit cues (*e.g.*, a "user-correction" method)—for example, selecting options such as "delete as junk" and "not junk"—and/or implicit cues. Although such an approach is better than the techniques discussed prior thereto, it is still deficient as compared to the subject invention described and claimed below.

SUMMARY OF THE INVENTION

The following presents a simplified summary of the invention in order to provide a basic understanding of some aspects of the invention. This summary is not an extensive overview of the invention. It is not intended to identify key/critical elements of the invention or to delineate the scope of the invention. Its sole purpose is to present some concepts of the invention in a simplified form as a prelude to the more detailed description that is presented later.

The subject invention provides for a feedback loop system and method that facilitates classifying items in connection with spam prevention. The invention makes

uses of a machine-learning approach as applied to spam filters, and in particular, randomly samples incoming email messages so that examples of both legitimate and junk/spam mail are obtained to generate sets of training data. Pre-selected individuals serve as spam fighters and participate in categorizing respective replications (which optionally can be slightly modified) of the samples.

Generally, messages selected for polling are modified in various aspects to appear as polling messages. A unique aspect of the invention is that a copy of an incoming message selected for polling is made such that some users (*e.g.*, spam fighters) will receive the same message (*e.g.*, in terms of message content) twice: once in the form of a polling message and again, in its original form. Another unique aspect of the subject invention is all messages are considered for polling – including those which have been labeled as spam by existing filters. Spam-labeled messages are considered for polling and if selected, are not treated as spam according to specifications of the existing filter (*e.g.*, move to junk folder, delete...).

Unlike conventional spam filters, more accurate spam filters can be created by training spam filters in accordance with the feedback technique of the subject invention so as to learn to distinguish between good mail and spam, thereby mitigating biased and inaccurate filtering. The feedback is accomplished at least in part by polling any suitable number of users to obtain feedback on their incoming email. Users, identified as spam-fighters, are tasked with voting on whether a selection of incoming messages is either legitimate mail or junk mail. Both positive and negative classifications of incoming email are desired to mitigate improperly filtering out as spam mail that is good (*e.g.*, not spam) intended for a user. The respective classifications along with any other information associated with each mail transaction are moved to a database to facilitate training the spam filters. The database and related components can compile and store properties for selected message(s) (or selected mail transaction), which includes user properties, user voting information and histories, message properties such as unique identification numbers assigned to each selected message, message classifications, and message content summaries, or statistical data related to any of the above, to generate sets of training data for machine learning systems. Machine learning systems (*e.g.*, neural networks, Support Vector Machines (SVMs), Bayesian Belief Networks) facilitate

creating improved spam filters that are trained to recognize both legitimate mail and spam mail and further, to distinguish between them. Once a new spam filter has been trained in accordance with the invention, it can be distributed to mail servers and client email software programs. Furthermore, the new spam filter can be trained with respect to a specific user(s) to improve performance of a personalized filter(s). As new training data sets are built, the spam filter can undergo further training *via* machine learning to optimize its performance and accuracy. User feedback by way of message classification can also be utilized to generate lists for spam filters and parental controls, to test spam filter performance, and/or to identify spam origination.

Another aspect of the invention provides for a method of detecting untrustworthy users through cross-validation techniques and/or by known result test messages. Cross-validation involves training a filter from which the polling results of some users are excluded. That is, the filter is trained using polling results from a subset of users. On average, this subset of users will work well enough even with some mistakes to detect those who generally are not in agreement with them. The polling results from the excluded users are compared to those of the trained filter. This comparison essentially determines how the users from the training subset would have voted on the messages belonging to the excluded users. If the agreement between an excluded user's votes and the filter is low, then the polling results from that user can either be discarded or marked for manual inspection. This technique can be repeated as desired, excluding data from different users each time.

Mistakes on individual messages can also be detected such as a message on which the filter and the user vote strongly disagree. These messages can be flagged for either automatic removal and/or manual inspection. As an alternative to cross-validation, a filter can be trained on all or substantially all users. The user votes and/or messages that disagree with the filter can be discarded. Another alternative to cross-validation involves known result test messages in which the user(s) is asked to vote on a message(s) where the result is known. Accurate classification (*e.g.*, user vote matches filter action) of the message by the user verifies the user's trustworthiness and determines whether to remove the user's classifications from training, and whether to remove the user from future polling.

217
217

Yet another aspect of the invention provides for creating known spam targets (e.g., honeypots) to identify incoming mail as spam and/or to track specific merchant email address processing. A known spam target, or honeypot, is an email address where the set of legitimate mail can be determined and all other mail can be considered spam.

5 For instance, the email address can be disclosed on a website in a restrictive manner not likely to be found by people. Hence, any mail sent to this address can be considered spam. Alternatively, the email address may have only been disclosed to a merchant from whom legitimate mail is expected to be received. Thus, mail received from the merchant is legitimate mail, but all other mail received can safely be considered spam. Spam data

10 derived from honeypots and/or other sources (e.g., users) can be integrated into the feedback loop system, but because of the substantial increase in spam classification with honeypots, such data should be down weighted, as will be described *infra* in greater detail, to mitigate obtaining biased polling results.

Another aspect of the invention provides for quarantining messages which are

15 deemed uncertain either by the feedback loop system or by the filter. Such messages are held for any suitable period of time instead of being discarded or classified. This time period can be set in advance, or the message can be held until receipt of a determined number of poll results similar to the message, e.g., from the same IP address or with similar content.

To the accomplishment of the foregoing and related ends, certain illustrative

20 aspects of the invention are described herein in connection with the following description and the annexed drawings. These aspects are indicative, however, of but a few of the various ways in which the principles of the invention may be employed and the present invention is intended to include all such aspects and their equivalents. Other advantages

25 and novel features of the invention may become apparent from the following detailed description of the invention when considered in conjunction with the drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1A is a block diagram of a feedback loop training system in accordance with

30 an aspect of the present invention.

~~218~~
218

Fig. 1B is a flow diagram of an exemplary feedback loop training process in accordance with an aspect of the present invention.

Fig. 2 is a flow diagram of an exemplary method that facilitates mail classification by users to create spam filters in accordance with an aspect of the present invention.

5 Fig. 3 is a flow diagram of an exemplary method that facilitates cross-validation of users participating in the method of Fig. 2 in accordance with an aspect of the present invention.

Fig. 4 is a flow diagram of an exemplary method that facilitates determining whether users are untrustworthy in accordance with an aspect of the present invention.

10 Fig. 5 is a flow diagram of an exemplary method that facilitates catching spam and determining spam originators in accordance with an aspect of the present invention.

Fig. 6 is a block diagram of a client-based feedback loop architecture in accordance with an aspect of the present invention.

15 Fig. 7 is a block diagram of a server-based feedback loop system having one or more users that generate training data in accordance with an aspect of the present invention.

20 Fig. 8 is a block diagram of a cross-organizational server-based feedback loop system wherein the system includes an internal server with its own database to pull training data stored on external user databases in accordance with an aspect of the present invention.

Fig. 9 illustrates an exemplary environment for implementing various aspects of the invention.

Fig. 10 is a schematic block diagram of an exemplary communication environment in accordance with the present invention.

25 DETAILED DESCRIPTION OF THE INVENTION

The present invention is now described with reference to the drawings, wherein like reference numerals are used to refer to like elements throughout. In the following description, for purposes of explanation, numerous specific details are set forth in order to provide a thorough understanding of the present invention. It may be evident, however, that the present invention may be practiced without these specific details. In

~~219~~
219

other instances, well-known structures and devices are shown in block diagram form in order to facilitate describing the present invention.

As used in this application, the terms “component” and “system” are intended to refer to a computer-related entity, either hardware, a combination of hardware and software, software, or software in execution. For example, a component may be, but is not limited to being, a process running on a processor, a processor, an object, an executable, a thread of execution, a program, and/or a computer. By way of illustration, both an application running on a server and the server can be a component. One or more components may reside within a process and/or thread of execution and a component may be localized on one computer and/or distributed between two or more computers.

The subject invention can incorporate various inference schemes and/or techniques in connection with generating training data for machine learned spam filtering. As used herein, the term “inference” refers generally to the process of reasoning about or inferring states of the system, environment, and/or user from a set of observations as captured *via* events and/or data. Inference can be employed to identify a specific context or action, or can generate a probability distribution over states, for example. The inference can be probabilistic – that is, the computation of a probability distribution over states of interest based on a consideration of data and events. Inference can also refer to techniques employed for composing higher-level events from a set of events and/or data. Such inference results in the construction of new events or actions from a set of observed events and/or stored event data, whether or not the events are correlated in close temporal proximity, and whether the events and data come from one or several event and data sources.

It is to be appreciated that although the term message is employed extensively throughout the specification, such term is not limited to electronic mail *per se*, but can be suitably adapted to include electronic messaging of any form that can be distributed over any suitable communication architecture. For example, conferencing applications that facilitate a conference between two or more people (*e.g.*, interactive chat programs, and instant messaging programs) can also utilize the filtering benefits disclosed herein, since unwanted text can be electronically interspersed into normal chat messages as users exchange messages and/or inserted as a lead-off message, a closing message, or all of the

220
220

above. In this particular application, a filter could be trained to automatically filter particular message content (text and images) in order to capture and tag as junk the undesirable content (*e.g.*, commercials, promotions, or advertisements).

5 In the subject invention, the term "recipient" refers to an addressee of an incoming message or item. The term "user" refers to a recipient who has chosen, either passively or actively, to participate in the feedback loop systems and processes as described herein.

10 Referring now to Fig. 1A, there is illustrated a general block diagram of a feedback training system 10 in accordance with an aspect of the present invention. A message receipt component 12 receives and delivers incoming messages (denoted as IM) to intended recipients 14. The message receipt component can include at least one filter 16 as is customary with many message receipt components (*e.g.*, junk mail filter) to mitigate delivery of undesirable messages (*e.g.*, spam). The message receipt component 12 in connection with the filter 16 processes the messages (IM) and provides a filtered
15 subset of the messages (IM') to the intended recipients 14.

20 As part of the feedback aspect of the subject invention, a polling component 18 receives all of the incoming messages (IM) and identifies the respective intended recipients 14. The polling component selects a subset of the intended recipients 14 (referred to as spam fighters 20) to classify a subset of the incoming messages (denoted as IM'') as spam or not spam, for example. The classification-related information (denoted as VOTING INFO) are submitted to a message store/vote store 22, where the voting information as well as copies of the respective IM'' are stored for later use such as by a feedback component 24. In particular, the feedback component 24 employs machine
25 learning techniques (*e.g.*, neural networks, SVMs, Bayesian networks or any machine learning system suitable for employment with the subject invention) which make use of the voting information to train and/or improve the filter 16 (and/or build new filter(s)) with respect to identifying spam mail, for example. As new streams of incoming messages are processed through the newly trained filter 16, less spam and more legitimate messages (denoted as IM') are delivered to the intended recipients 14. Thus,
30 the system 10 facilitates the identification of spam and the training of improved spam filters by utilizing feedback generated by spam fighters 20. Such feedback aspect of the

~~22~~
~~221~~
221

subject invention provides for a rich and highly dynamic scheme for refining a spam detection system. Various details regarding more granular aspects of the subject invention are discussed below.

Referring now to Fig. 1B, there is illustrated a feedback loop training flow diagram 100 in connection with spam fighting and spam prevention in accordance with the subject invention. In preparation of and/or prior to the training process, users are selected to be spam-fighters (*e.g.*, from a master set comprising all email users) – the selection can be based on a random sampling, or level of trust, or any suitable selection scheme/criteria in accordance with the subject invention. For example, the selected subset of users can include all users, a randomly selected set of users, those who have opted in as spam fighters, or those who have not opted out, and/or any combination thereof, and/or based in part upon their demographic location and related information.

Alternatively, the master set of email users selected from can be limited to paying users which can make it more expensive for spammers to subvert the subject invention. Thus, a subset of users selected to participate in the spam fighting could comprise only paying users. A list or customer table including the names and properties of the selected users (*e.g.*, spam fighters) can then be created.

When an incoming stream of messages 102 is received, a recipient of each message is checked against a list of all spam fighters at 104. If the recipient is on the list, then the message is considered for polling. Next, a determination is made whether to select a message for polling. Unlike conventional spam filters, the invention does not delete any messages (*e.g.*, spam) until at least after all incoming mail is considered for polling. That is, the mail is classified before it is subjected to any labeling (*e.g.*, spam, non-spam) - this facilitates obtaining an unbiased sample of messages available for user polling.

A component for message selection (not shown) can be employed to select messages with some random probability to mitigate bias of data. Another approach involves using demographic information as well as other user/recipient attributes and properties. Thus, messages can be selected based at least in part upon the user/recipient. Other alternative algorithms exist for selecting messages. However, there may be limitations on the number of messages selected per user or per user per time period, or on

44
222
222

the probability of selecting a message from any given user. Without such limits, a spammer could create an account, send it millions of spam messages, and classify all such messages as good: this would allow the spammer to corrupt the training database with incorrectly labeled messages.

5 Some forms of spam filtering, notably referred to as black hole lists may not be skippable. Black hole lists prevent a server from receiving any mail from a list of Internet Protocol (IP) addresses. Therefore, the selection of messages can be chosen from the set of mail which is not from a black hole list.

10 A unique aspect of the invention is that messages selected for polling, which are marked as spam by filters currently in place, are not deleted or moved to a junk mail folder. Instead, they are placed in a usual inbox or mailbox where all other messages are received for polling consideration. However, if there are two copies of the message, and the message is considered as spam by the filter, then one copy is delivered to the spam folder or otherwise treated according to set parameters (*e.g.*, deleted, specially marked, or
15 moved to junk folder).

 When a message is selected, it is forwarded to the user and marked in some special way to indicate that it is a polling message. In particular, the selected message can be modified by a message modification component 106. Examples of message modification include, but are not limited to, locating the polling message in a separate
20 folder, changing the 'from' address or the subject line, and/or using a special icon or special color that would identify the message as a polling message to the user. The selected message can also be encapsulated within another message, which would provide instructions to the user on how to vote on and/or classify the encapsulated message. These instructions can include at least two buttons or links: one to vote the message as
25 spam and one to vote the message as not spam, for example.

 The voting buttons can be implemented by modifying the contents of the message before sending a copy of the polling message to the user. When the invention is employed with respect to client email software (as opposed to a mail server), the user interface can be modified to include the voting buttons.

30 Moreover, the polling message can contain instructions and voting buttons as well as the selected message attached thereto. The polling message can also comprise a

223
223

summary of the selected message such as the subject line, from address, date sent and/or received, and the text or at least the first few lines of the text. Another approach involves sending the message with the voting instructions and voting buttons pre-pended thereto. In practice, when a user opens and/or downloads a copy of the polling message, buttons (or links) including, but not limited to, "spam" and "not spam" buttons can pop up on the user interface or can be incorporated into the polling message. Thus, it is possible that each polling message contains a set of instructions and suitable voting buttons. Other modifications may be necessary, including possibly removing HTML background instructions (which could obscure the text of instructions or buttons.)

Another button such as a "solicited commercial email" button can also be provided, depending on the type of information that is desired. The message can also include a button/link to opt-out of future polling. The instructions are localized to the user's preferred language and can be embedded into the polling message.

Furthermore, messages selected for polling can be scanned for viruses by the message modification component 106 or by some other suitable virus scanning component (not shown). If a virus is found, the virus can either be stripped away or the message can be discarded. It should be appreciated that virus stripping can occur at any point of the system 100, including when the message is selected and right before the user downloads the message.

Following modification of the message, a message delivery component 108 delivers the polling message to the user for voting. User feedback (e.g., polling message, user's vote, and any user properties associated therewith) is assigned a unique identifier (ID) 110 (e.g., metadata). The ID 110 and/or the information corresponding thereto are submitted to a message store/vote store 112 (e.g., central database), where the user classifications/votes are compiled and stored.

At a database level, selected messages available for polling can be kept for later polling or use. In addition, the database can perform frequency analyses on a timed basis to make sure that a particular user is not being over sampled and that an amount of data is being collected from the user within limits as specified by the user. In particular, the feedback system 100 monitors a percentage limit of a user's mail as well as the sampling period to mitigate bias of both sampling and data. This is especially important where

users are selected from all available users, including both low usage and high usage users. For example, a low usage user typically receives and sends a significantly lower volume of mail as compared to a high usage user. Thus, the system 100 monitors the message selection process to be certain that the selected message is approximately one out of every T number of messages received by the user and no more than 1 message received every Z hours by the user. Accordingly, the system can poll 1 out of every 10 incoming messages to be sampled (*e.g.*, considered for polling), but no more than 1 every 2 hours, for example. The frequency, or percentage, limit mitigates sampling a disproportionate amount of messages for a low usage user as compared to a high usage user, and also mitigates overly annoying a user.

On a frequent basis, the central database 112 scans for messages which have been sampled by the system 100 for polling but that have not been classified. The database pulls these messages and localizes them relative to respective user's demographic properties and creates polling messages to request the user(s) to vote and classify the message(s). However, the spam filter may not be modified or trained immediately after receipt of every new incoming classification. Rather, offline training allows a trainer to continually look at the data received into the database 112 on a scheduled, ongoing, or daily basis. That is, the trainer starts from a prescribed starting point or at a set amount of time in the past and looks at all the data from that point forward to train the filter. For example, the prescribed time period can be from midnight to 6:00 AM.

The new spam filter can be trained on an ongoing basis by analyzing the message classifications maintained in the database 112 by way of machine-learning techniques 114 (*e.g.*, neural networks, support vector machines (SVMs)). Machine learning techniques require both examples of good mail and spam to learn from so that they can learn to distinguish between them. Even techniques based on matching known examples of spam can benefit from having examples of good mail, so that they can make sure they do not accidentally catch good mail.

Accordingly, it is important to have both positive and negative examples of spam, instead of just complaints. There are some domains that send out large amounts of both spam and legitimate mail such as free mailing lists. If one built a system based only on complaints, all mail from these domains may be filtered resulting in a large number of

mistakes. Hence, knowing that the domain also sends out large amounts of good mail is important. In addition, users often make mistakes such as forgetting that they have signed up on a free mailing list. For instance, a large legitimate provider such as the New York Times regularly sends out legitimate mail. A few users forget that they have signed up and complain, classifying these messages as spam. Without data that most users realize that this mail is legitimate, mail from this site can otherwise be blocked.

The new filter 116 can be distributed on an ongoing basis by a distribution component 118 across participating internet service providers (ISP), to the email or message servers, to individual email clients, to an update server, and/or to the central databases of individual companies. Moreover, the feedback system 100 functions on an ongoing basis such that samples of messages considered and utilized for polling can follow an actual distribution of email received by the system 100. As a result, training data sets employed to train new spam filters are kept current with respect to adaptive spammers. When new filters are built, polling data can be discarded or down weighted (*e.g.*, discounted) based on how long ago it was obtained.

The system 100 can be implemented when mail is received at a server such as a gateway server, email server, and/or message server. For instance, when mail comes into an email server, the server looks up the properties of the intended recipients to determine whether the recipients have opted in to the system 100. If their properties indicate as such, the recipients' mail is potentially available for polling. Client-only architectures also exist. For example, client email software can make the polling decisions for a single user and deliver the email either to a central database or use the polling information to improve the performance of a personalized filter. In addition to those described herein, other alternative architectures for this system 100 exist and such are contemplated to fall within the scope of the present invention.

Referring now to Fig. 2, there is illustrated a flow diagram of a basic feedback loop process 200 in accordance with one aspect of the present invention. While, for purposes of simplicity of explanation, the methodology is shown and described as a series of acts, it is to be understood and appreciated that the present invention is not limited by the order of acts, as some acts may, in accordance with the present invention, occur in different orders and/or concurrently with other acts from that shown and described

herein. For example, those skilled in the art will understand and appreciate that a methodology could alternatively be represented as a series of interrelated states or events, such as in a state diagram. Moreover, not all illustrated acts may be required to implement a methodology in accordance with the present invention.

5 The process 200 begins with mail coming into and being received by a component such as a server at 202. When mail arrives at the server, the server identifies properties of intended recipients to determine whether the intended recipients have previously opted in as spam fighters for polling (at 204). Thus, the process 200 utilizes a user property field where it can be indicated whether the recipient has opted in to the feedback system, or consults a list of users who have opted in. If the user is determined to be a participant in the feedback system and has been selected for polling at 206, the feedback system takes action by determining which messages are selected for polling (at 208). Otherwise, the process 200 returns to 202 until at least one intended recipient of an incoming message is determined to be a user (*e.g.*, spam fighter).

15 In practice, all messages are considered for polling including those messages which are designated (or would be) as spam by a currently employed filter (*e.g.*, personalized filter, Brightmail filter). Therefore, no messages are deleted, discarded, or sent to junk folders before they are considered for polling.

20 Each message or mail item received by the server has a set of properties corresponding to the mail transaction. The server compiles these properties and sends them along with the polling message to a central database. Examples of the properties include the recipient list (*e.g.*, as listed in "To:", "cc:", and/or "bcc:" fields), verdict of a currently employed filter (*e.g.*, whether filter identified message as spam), verdict of another optional spam filter (*e.g.*, Brightmail filter), and user information (*e.g.*, username, password, real name, frequency of messages polled, usage data,...). The polling message and/or its contents, as well as the corresponding user/recipient are each assigned a unique identifier. The identifier can also be sent to the database and subsequently updated as needed.

25 At 214, the message(s) selected for polling (*e.g.*, original message_{1-M}, where M is an integer greater than or equal to one) is modified to indicate to the user that the message_{1-M} is a polling message_{P1-PM} and then is delivered to the user for polling (at 216).

For example, the polling message can include the original message to be voted on as an attachment and a set of instructions on how to vote on the message. The set of instructions includes at least two buttons such as a “good mail” button and a “spam” button, for example. When the user clicks on one of the buttons (at 218) to classify the message as good mail or spam, the user is directed to a uniform resource locator (URL) that corresponds to a unique identifier for the classification that the user is submitting. This information is posted and the associated record in the central database for that original message_{1-M} is updated.

At 216 or at any other suitable time during the process 200, the original message can optionally be delivered to the user. Thus, the user receives the message twice – once in its original form and again in its modified polling form.

At some later time, a new spam filter is created and trained at 220 based at least in part upon user feedback. Once the new spam filter has been created and trained, the filter can be employed immediately on the email server and/or can be distributed to client servers, client email software, and the like (at 222). Training and distributing a new or updated spam filter is an ongoing activity. Thus, the process 200 continues at 204 when a new stream of incoming messages is received. When new filters are built, older data is discarded or down weighted based on how long ago they were obtained.

The feedback system 100 and process 200 rely on the feedback of its participating users. Unfortunately, some users cannot be trusted or are simply lazy and fail to provide consistent and accurate classifications. The central database 112 (Fig. 1a) maintains histories of user classifications. Thus, the feedback system 100 can track the number of contradictions, the number of times the user changed his/her mind, responses of the user to known good mail or known spam, as well as the number or frequency of user replies to polling messages.

When any one of these numbers exceeds a prescribed threshold, or simply for every user of the system, the feedback system 100 can invoke one or several validation techniques to assess the trustworthiness of a particular user or users. One approach is a cross-validation method 300 as illustrated in Fig. 3 in accordance with another aspect of the present invention.

The cross-validation technique begins at 302 with a central database receiving incoming data such as polling results and respective user information. Next, it must be determined whether cross-validation is desired to test a suitable number of users at 304. If it is desired, then, a new spam filter is trained using some portion of the incoming data at 306. That is, the data from the users which are being tested is excluded from the training. For example, the filter is trained with about 90% of the polled user data (denoted as the 90% filter), thereby excluding about 10% of the data (denoted as the 10% tested user) which corresponds to the data submitted by the tested user.

At 308, the 90% filter is run against the remaining 10% tested user data to determine how the 90% users would have voted on the tested user's messages. If the amount of disagreements between the 90% filter and the 10% tested user data exceeds a prescribed threshold (at 310), then the user's classifications can be manually inspected at 312. Alternatively or in addition, test messages can be sent to the suspicious or untrustworthy users and/or these particular users can be excluded from future polling, and/or their past data discarded. However, if the threshold is not exceeded, then the process returns to 306. In practice, the cross-validation technique 300 can be utilized with any suitable set of test users, excluding different users as necessary to determine and maintain the trustworthiness of the voting/classification data.

A second approach to assess user fidelity and reliability includes training a filter on all data gathered in a given period, and then testing on the training data, using the filter. This technique is known as test-on-training. If a message was included in the training, the filter should have learned its rating, *e.g.*, the learned filter should classify the message the same way that the user did. However, the filter may continue to make a mistake on it by labeling it as spam when the user labeled it is as not spam or vice versa. In order for a filter to disagree with its training data, the message has to strongly disagree with other messages. Otherwise, the trained filter would almost certainly have found some way to classify it correctly. Thus, the message can be discarded as having an unreliable label. Either this technique or cross validation may be used: cross-validation can yield more mistakes in classifications less reliably; conversely test-on-training finds fewer mistakes more reliably.

229
229

Both the test-on-training and the cross-validation technique 300 may be applied to individual messages wherein an individual user's classification or rating of a message is excluded by general agreement (*e.g.*, following the majority rating). Alternatively, both techniques can be used to identify potentially unreliable users.

5 In addition to, or instead of cross validation and/or test on training techniques, we can use the "known-results" technique to verify user trustworthiness (follow 314 to Fig. 4). Although the techniques of Figs. 3 and 4 are demonstrated separately, it should be appreciated that both approaches can be utilized at the same time. That is, information from known-good and known-spam messages can be combined with cross-validation or
10 test on training results to determine which users to discard.

Referring now to Fig. 4, there is illustrated a flow diagram of a process 400 to validate the fidelity of user voting in accordance with one aspect of the invention. The process 400 refers from 314 as shown in Fig. 3. At 402, a known result test message(s) is sent to suspicious user(s) (or all users). For example, a test message may be injected into
15 the incoming mail and then hand classified so that the database receives the "known" result. Otherwise, the process 400 can wait until a known result message is sent by a third party. The users are allowed to vote on the same test messages. The voting results are compared to the known results at 404. If the users' votes do not agree at 406, then their current and/or future and/or past classifications can be hand-inspected for a suitable
20 period of time (at 408) until they demonstrate consistency and reliability. Alternatively, their current or future or past classifications can be discounted or removed. Finally, the users can be removed from future polling. However, if their voting results do agree with the test message results, then the users can be considered trustworthy at 410. The process 400 returns at 412 to Fig. 3 to determine what type of validation technique is desired for
25 the next group of suspect users.

A fourth approach (not shown) to assess user reliability is active learning. With active learning techniques, messages are not picked at random. Instead, the feedback system can estimate how useful the message will be to the system. For instance, if the filter returns a probability of spam, one can preferentially select the messages which are
30 most uncertainly classified by the current filter for polling, *i.e.*, those whose probability of spam is closest to 50%. Another way to select messages is to determine how common

230
230

the message is. The more common the message, then the more useful it is to poll.

Unique messages are less useful because they are less common. Active learning can be employed by using the confidence levels of existing filters, using how common features of the message are, and using existing filter's confidence levels of its settings or content (e.g., metaconfidence). There are many other active learning techniques, such as query-by-committee, well known to those skilled in the art of machine learning, and any of these techniques can be used.

Referring now to Fig. 5, there is illustrated a flow diagram of a process 500 for incorporating honeypot feedback in addition to user feedback into spam filter training in accordance with one aspect of the subject invention. Honeypots are email addresses to which it is known who should be sending them email. For example, a newly created email address may be kept private and disclosed only to selected individuals (at 502). They may also be disclosed publicly but in restrictive ways not seen by people (e.g., putting it on a white background in white typeface as a mail link). Honeypots are particularly useful in dictionary attacks by spammers. A dictionary attack is one in which a spammer tries emailing a very large number of addresses, perhaps all addresses in a dictionary or made from pairs of words in a dictionary or similar techniques in order to find valid addresses. Any email sent to a honeypot (at 504) or any email not from the few selected individuals (at 506) is considered spam (at 508). An email address can also be signed up with a suspect merchant. Thus, any email received from the merchant is considered good mail (at 510) but all other mail is considered spam. The spam filter can be trained accordingly (at 512). Moreover, the suspect merchant is determined to sell or otherwise disclose the user's information (e.g., at least the email address) to third parties. This can be repeated with other suspect merchants and a list can be generated to warn users that their information could be distributed to spammers. These are just a few techniques of getting email sent to honeypots which can safely be considered spam. In practice, there are other alternative ways to get email sent to honeypots that can safely be considered spam.

Because honeypots are a good source of spam but a terrible source of legitimate mail, the data from honeypots can be combined with data from the feedback loop system (Fig. 1) to train new spam filters. Mail from different sources or different classifications

53
23T
231

can be weighed differently. For example, if there are 10 honeypots and 10 users who are polled on 10% of their mail, about 10 times as much spam is to be expected from the honeypots as from polling. Therefore, the legitimate mail from polling can be weighted at 10 or 11 times as much as the spam in order to make up for this difference.

Alternatively, honeypot data can be selectively down weighted. For example, about 50% of a user's mail is good mail and about 50% of it is spam. The same volume of spam is going to the honeypots. Therefore, it looks like the honeypot has 100% of spam, and all of it is sampled, not just 10%. In order to train with the correct ratios of spam and good mail in the combined system, the honeypot data is down weighted by 95% and the user spam is down weighted by 50% to result in a 1:1 overall ratio.

Other sources of spam reports include users who are not included as participants in the feedback loop system. For instance, there may be a "Report Spam" button available to all users for all mail, to report spam that has made it through the filter. This data can be combined with data from the feedback loop system. Again, this source of spam should be down weighted or weighted differently since it can be biased or untrustworthy in various aspects. Re-weighting should also be done to reflect the fact that only mail that was not filtered is subject to reporting by the "Report-as-spam" button.

In addition to the spam filter, a quarantine filter can be created and employed by the feedback loop system. The quarantine filter makes use of both positive and negative mail features. For example, mail from a popular online merchant is almost always good. A spammer exploits the system by mimicking an aspect of the good merchant mail in his spam. Another example is that the spammer intentionally tricks the feedback system by sending small amounts of good mail *via* an IP address. The feedback loop learns to classify this mail as good mail, when at such time, the spammer starts sending spam from the same IP address.

Thus, the quarantine filter notices a particular positive feature is being received in much greater quantities than the system is used to on the basis of historical data. This causes the system to be suspicious of the message and hence, quarantines it until sufficient poll results are obtained before choosing to deliver or mark the mail as spam. The quarantine filter can also be employed when mail is received from a new IP address, for which it is not known or certain whether the mail is spam or not spam and such will

not be known for a while. Quarantining can be performed in a number of ways, including provisionally marking the mail as spam and moving it to a spam folder or by not delivering it to the user or storing it somewhere where it will not be seen. Quarantining can be done for messages that are near the spam filter threshold: it can be assumed that additional information from polling will help make a correct decision. Quarantining can also be done when many similar messages are received: a few of the messages can be sent for polling with the feedback loop, and the retrained filter can be used to correctly classify the messages.

In addition to building filters, the feedback loop system as described herein can be utilized to evaluate them as well. That is, parameters of the spam filters can be tuned as needed. For example, a filter is trained up through midnight of last night. After midnight, take data that comes into the database to determine error rates of the spam filter as compared to the users' classifications. Further, the feedback loop can be employed to determine false positive and catch rates of the spam filter. For example, the user votes can be taken and the mail can be run through a potential filter to determine the false positive and catch rates. This information can then be used to tune and optimize the filter. Different parameter settings or different algorithms can be manually or automatically tried by building several filters, each one using a different setting or algorithm, to obtain the lowest false positive and catch rates. Thus, the results can be compared to select the best or optimal filter parameters.

The feedback loop can be utilized for building and populating lists of IP addresses or domains or URLs that are always voted as spam or always voted as good, or voted at least 90% good, etc. These lists can be used for spam filtering in other ways. For instance, a list of IP addresses voted at least 90% spam could be used for building a black-hole list of addresses from which to accept no mail. The feedback loop can also be used to terminate the accounts of spammers. For example, if a particular user of an ISP appears to be sending spam, the ISP can be automatically notified. Similarly, if a particular domain appears responsible for a large amount of spam, the domain's email provider can be automatically notified.

There are a number of architectures that can be used to implement the feedback loop system. One exemplary architecture is server based, as will be described in Fig. 7,

with the selection process happening when the mail reaches the email server. An alternate architecture is client based, as is described in Fig. 6. In a client-based feedback loop, polling information can be utilized to improve the performance of a personalized filter, or, in the exemplary implementation illustrated here, the information can be sent to a shared repository as training data for a shared filter (e.g. corporate wide, or global.) It should be appreciated that the following architectures described below are merely exemplary and can include additional components and features not depicted therein.

Referring now to Fig. 6, there is illustrated an exemplary general block diagram of a feedback loop technique in a client-based architecture. A network 600 is provided to facilitate communication of e-mail to and from one or more clients 602, 604, and 606 (also denoted as CLIENT₁, CLIENT₂... CLIENT_N, where N is an integer greater or equal to one). The network can be a global communication network (GCN) such as the internet, or a WAN (Wide Area Network), LAN (Local Area Network), or any other network configuration. In this particular implementation, an SMTP (Simple Mail Transfer Protocol) gateway server 608 interfaces to the network 600 to provide SMTP services to a LAN 610. An email server 612 operatively disposed on the LAN 610 interfaces to the gateway 608 to control and process incoming and outgoing email of the clients 602, 604, and 606. Such clients 602, 604, and 606 are also disposed on the LAN 610 to access at least the mail services provided thereon.

The client_i 602 includes a central processing unit (CPU) 614 that controls client processes. The CPU 614 can comprise multiple processors. The CPU 614 executes instructions in connection with providing any of the one or more data gathering/feedback functions described hereinabove. The instructions include, but are not limited to, the encoded instructions that execute at least the basic feedback loop methodology described above, at least any or all of the approaches that can be used in combination therewith for addressing client and message selection, polling message modification, data retention, client reliability and classification validation, reweighing of data from multiple sources including the feedback loop system, spam filter optimization and tuning, quarantine filters, creation of spam lists, and automatic notification of spammers to their respective ISPs and email providers. A user interface 616 is provided to facilitate communication

with the CPU 614 and client operating system such that the client₁ can interact to access the email and vote on polling messages.

A sampling of client messages retrieved from the server 612 can be selected for polling by a message selector 620. Messages are selected and modified for polling if the intended recipient (client) has previously agreed to participate. A message modifier 622 modifies the message to become a polling message. For example, the message(s) can be modified to include voting instructions and voting buttons and/or links according to the message modification descriptions provided hereinabove. Voting buttons and/or links are implemented by modifying the user interface 616 of the client email software. In addition, the message modifier 622 can remove any viruses in the messages (polling and non-polling messages) before they are opened or downloaded for viewing by the client 602.

In one implementation, the user of the spam fighting client 602 sees each message only once, with some messages specially marked as polling messages, and including voting buttons, etc. In the subject implementation, the user of the spam fighting client 602 may see some messages twice, wherein one is the normal message and the other is the polling message. This can be implemented in several ways. For instance, the polling message can be returned to the server 612 and stored in a polled message store. Alternatively, the client 602 can store an additional message in the E-Mail server 612. Alternatively, the client 602 can show the user each message twice, once as a normal message, and once in modified form.

Polling results 626 can be sent to the CPU 614 and then to a database 630 which can be configured to store data from one client or from more than one client, depending on the specific arrangement of the client feedback architecture. The central database 630 stores polling messages, polling results as well as the respective client-user information. Related components can be employed to analyze such information such as to determine polling frequency, client-user trustworthiness (*e.g.*, user validation 632), and other client statistics. Validation techniques can be employed particularly when the reliability of the client's voting is in question. Suspicion can arise from analyzing the number of contradictions, the number of changed minds, and the number of messages polled for a particular user or users; alternatively, validation techniques can be employed for every

5x
235-
235

user. Any suitable amount of data stored in the central database can be employed in machine learning techniques 634 to facilitate the training of a new and/or improved spam filter.

5 Clients 604 and 606 include similar components as described hereinabove to obtain and train a filter which is personalized to the particular client(s). In addition to what has been described, a polled message scrubber 628 can interface between the CPU 614 and the central database 630 such that aspects of the polled message may be removed for a variety of reasons such as data aggregation, data compression, etc. The polled message scrubber 628 can flush out extraneous portions of the polled message as well as
10 any undesired user information associated therewith.

Referring now to Fig. 7, there is illustrated an exemplary server-based feedback loop system 700 that facilitates multi-user logins and that obtains polling data in accordance with the feedback loop techniques of the present invention. A network 702 is provided to facilitate communication of e-mail to and from one or more users 704 (also
15 denoted as $USER_1$ 704₁, $USER_2$ 704₂ ... and $USER_N$ 704_N, where N is an integer greater or equal to one). The network 702 can be a global communication network (GCN) such as the internet, or a WAN (Wide Area Network), LAN (Local Area Network), or any other network configuration. In this particular implementation, an SMTP (Simple Mail Transfer Protocol) gateway server 710 interfaces to the network 702 to provide SMTP
20 services to a LAN 712. An email server 714 operatively disposed on the LAN 712 interfaces to the gateway 710 to control and process incoming and outgoing email of the users 704.

The system 700 provides multiple login capability such that user and message selection 716, message modification 718, and message polling (720, 722, 724) takes
25 place for each different user that logs into the system 700. Thus, there is provided a user interface 726 that presents a login screen as part of the boot-up process of the computer operating system, or as required, to engage an associated user profile before the user 704 can access his or her incoming messages. Thus, when a first user 704₁ ($USER_1$) chooses to access the messages, the first user 704₁ logs into the system *via* a login screen 728 by
30 entering access information typically in the form of a username and password. A CPU

730 processes the access information to allow the user access, *via* a message communication application (*e.g.*, a mail client) to only a first user inbox location 732.

When incoming mail is received on the message server 714, they are randomly selected for polling which means that at least one of the messages is tagged for polling. The intended recipient(s) of the tagged messages are looked at to determine whether any one of the recipients is also a designated spam fighting user. Recipient properties indicating such information can be maintained on the message server 714 or on any other component of the system 700 as appropriate. Once it is determined which of the intended recipients are also spam fighters, a copy of their respective mail as well as any other information regarding the mail transaction can be sent to a central database 734 for storage. Messages tagged for polling are modified by the message modifier 718 in any number of ways described hereinabove. Messages selected for polling may also be specific to the user 704. For example, the user 704 can indicate that only certain types of messages are available for polling. Since this can result in a biased sampling of data, such data can be re-weighted with respect to other client data to mitigate building disproportionate training data sets.

Virus scanning of the polling messages can also be performed at this time or at any other time before the polling message is downloaded and/or opened by the user 704. Once the messages have been modified in the appropriate manner, they are delivered to the respective user's inboxes which are denoted as $INBOX_1$ 732, $INBOX_2$ 736, and $INBOX_N$ 738, where they can be opened for polling. To facilitate the polling process, each polling message includes two or more voting buttons or links, which when selected by the user, generates information relating to the polling message and the polling result. The text of each polling message can be modified to incorporate the voting buttons or links therein.

Message poll results (denoted as $MESSAGE POLL_1$ 720, $MESSAGE POLL_2$ 722, and $MESSAGE POLL_N$ 724.), which include any information resulting from the classification (*e.g.*, polling message or ID associated therewith, user properties), are sent to the central database 734 *via* a network interface 740 on the LAN 712. The central database 734 can store polling and user information (720, 722, 724) from the respective users to apply to machine learning techniques to build or optimize a new and/or improved

58
237
237

spam filter 742. However, for privacy and/or security reasons, confidential information can be removed or stripped out of the information before it is sent to the central database 714. Information generated by the user(s) 704 *via* polling can also be aggregated into statistical data. Thus, less bandwidth is used to transmit the information.

5 The newly trained spam filter 742 can then be distributed to other servers (not shown) as well as client email software (not shown) interfacing with the LAN 712 on an ongoing basis, such as when a new filter is available, either by specific request or automatically. For example, the newest spam filter can be automatically pushed out to them and/or made available for downloading *via* a website. As new training data sets are
10 generated to build newer spam filters, older data sets (*e.g.*, information previously obtained and/or employed to train a filter) can be discarded or discounted depending on the age of the data.

 Consider now an alternate scenario wherein an organization devoted to spam fighting makes available a filter shared by many different filter-using organizations. In
15 one aspect of the invention, the filter provider is also a very large provider of email services (*e.g.*, paid and/or free email accounts). Rather than relying exclusively on email from its own organization, the filter provider chooses to also use some data from some of the filter-using organizations, so as to better capture the range of good mail and spam. The feedback loop system as described hereinabove can also be employed in such a
20 cross-organizational scenario, either in a server or client-based architecture. We will call the filter provider, who aggregates data from its own users and from the different filter-using organizations the "internal" organization and call the components residing at one of the participating filter using organizations "external." In general, the cross-organizational system includes a mail database server at the filter provider (internal), such as, but not
25 limited to, Hotmail and one or more message servers (external) such as those which may reside within one or more individual companies. In this case, the internal mail database server also stores substantial email feedback from its own customers. According to this aspect of the subject invention, training data sets may be generated based on information stored on an internal database (*e.g.*, free e-mail/messaging on a Hotmail or MSN server)
30 as well as information stored on one or more external databases associated with the respective external servers. Information maintained on the external databases can be

communicated to the internal server *via* a network such as the Internet, for example, for employment in machine learning techniques. Ultimately data from the external databases can be utilized to train new spam filters and/or improve existing spam filters located externally (*e.g.*, within the respective company) or associated with the internal mail server.

The data from one or more of the external databases should include at least one of polling messages, polling results (classifications), user information/properties, and voting statistical data per user, per group of users or on average for each company. The voting statistical data facilitate determining reliability of the information generated by the respective companies as well as mitigating bias of external data. Thus, the data from one or more external databases (companies) can be re-weighted or weighted differently from one or more of the other external databases. Moreover, the external entities can be tested for reliability and trustworthiness using similar validation techniques as described with hereinabove.

For company security, privacy and confidentiality, the information or data communicated across the Internet from each company to the e-mail server, for example, can be scrubbed, abbreviated, and/or condensed from its original form. The original form can be maintained on the respective external database and/or otherwise treated according to each company's preferences. Thus, the e-mail server or any other internal mail server receives only pertinent information necessary to generate training data such as spam classifications, sender domain, sender name, content of messages classified as spam, and the like.

Referring now to Fig. 8, there is illustrated an exemplary cross-organizational feedback system 800 where an internal database server and an external mail server can communicate and exchange database information *via* a network to facilitate the generation of training data sets used in machine learning techniques to build improved spam filters. The system 800 includes at least one external message server 802 (*e.g.*, associated with at least one company) and an internal database server 804. Due to the nature of the cross-organization system, the external server 802 and the internal e-mail server 804 respectively maintain their own databases. That is, the e-mail server 804 is associated with an internal database 806 that can also be used to train a new spam filter

6X
239
239

808. Likewise, the external server 802 is associated with an external database 810 which can be employed to train at least one new spam filter 812 as well as the spam filter 808 located internally with respect to the e-mail server 804. Thus, the information stored on the external database 810 can be utilized to train the spam filter 808 located on the e-mail server.

A GCN 814 is provided to facilitate communication of information to and from the internal e-mail server 804 and one or more external message servers 802. The external server(s) component of a cross-organizational system operates in a similar manner as does a server-based feedback loop system (*e.g.*, Fig. 7, *supra*). For example, the message server 802, external database 810 and filter 812 can be located on a LAN 815. In addition, there is provided a user interface 816 that presents a login screen 818 as part of the boot-up process of the computer operating system, or as required, to engage an associated user profile before the user(s) can access his or her incoming messages.

In this server-based system, one or more users (denoted as USER₁ 820, USER₂ 822, USER_N 824) can log into the system at the same time in order to make use of the available mail services. In practice, when a first user 820 (USER₁) chooses to access the messages, the first user 820 logs into the system *via* a login screen 818 by entering access information typically in the form of a username and password. A CPU 826 processes the access information to allow the user access to only a first user inbox location 828 *via* a message communication application (*e.g.*, a mail client).

When incoming mail is received on the message server 802, messages are randomly or specifically targeted for polling. Before messages can be selected for polling, the intended recipients of such targeted messages are compared to a spam-fighter user list to determine whether any one of the recipients is also a designated spam fighting user. Recipient properties indicating such information can be maintained on the message server 802, database 810, or on any other component of the system 800 as appropriate. Once it is determined which of the intended recipients are also spam fighters, the message(s) are selected for polling and a copy of polling message(s) as well as any other information pertaining to the mail transaction can be sent to the database 810.

Messages selected for polling are modified by a message modifier 830 in any number of ways described hereinabove. In practice, a unique identification (ID) can be

28
240
240

assigned to each polling message, to each spam fighter, and/or to each polling result and stored in the database 810. As previously mentioned, messages selected for polling can be randomly chosen or may be specific to the respective user(s) (820, 822, and 824). For example, the USER₁ 820 can indicate that only certain types of messages are available for polling (e.g., messages sent from outside of the company). Data generated from such specific messages is re-weighted and/or discounted to mitigate obtaining a biased sampling of data.

Virus scanning of the polling messages can also be performed at this time or at any other time before the polling message is downloaded and/or opened by the user. Once the messages have been modified in the appropriate manner, they are delivered to the respective user(s)'s inboxes which are denoted as INBOX₁ 828, INBOX₂ 832, and INBOX_N 834, where they can be opened for polling. To facilitate the polling process, each polling message includes two or more voting buttons or links, which when selected by the user, generates information relating to the polling message and the polling result. The text of each polling message can be modified to incorporate the voting buttons or links therein.

Message poll results (denoted as MESSAGE POLL₁ 836, MESSAGE POLL₂ 838, and MESSAGE POLL_N 840.), which include any information resulting from the classification (e.g., polling message or ID associated therewith, user properties), are sent to the database 810 via a network interface 842 located on the LAN 815. The database 810 stores polling and user information from the respective users for later use in machine learning techniques which are employed to build and/or optimize a new and/or improved spam filter(s) 812, 808.

For privacy reasons, each company may want to strip out key information before sending the polled message and/or user information to either its own database 810 and/or to the e-mail database 806 over the GCN 814, for example. One approach is to only provide feedback to the database (806 and/or 810) on spam messages, thereby excluding feedback on legitimate mail. Another approach is to only provide a partial subset of information on the legitimate mail such as the sender and the sender's IP address. Another approach is, for selected messages, such as those marked as good by the user that would be marked as bad by the filter, or vice versa, to explicitly ask for user

~~241~~
241
241

permission before sending them to the filter. Any of these approaches or a combination thereof facilitates maintaining privacy of confidential information for the participating clients while continually providing data to train the spam filter(s) (808 and/or 812).

5 User validation schemes such as those described hereinabove can also be applied to each company as well as to each user within the company. For example, the users can individually be subjected to cross-validation techniques wherein the classifications of a suspect user(s) are excluded from filter training. The filter is trained using the data from the remaining user(s). The trained filter then runs through the messages from the excluded user(s) to determine how it would have classified the messages. If the number
10 of disagreements exceeds a threshold level, then the suspect user(s) is considered untrustworthy. Future message classifications from the untrustworthy user(s) can be manually inspected before they are accepted by the database and/or filter. Otherwise, the user(s) can be removed from future polling.

Referring now to Fig. 9, an exemplary environment 910 for implementing various
15 aspects of the invention includes a computer 912. The computer 912 includes a processing unit 914, a system memory 916, and a system bus 918. The system bus 918 couples system components including, but not limited to, the system memory 916 to the processing unit 914. The processing unit 914 can be any of various available processors. Dual microprocessors and other multiprocessor architectures also can be employed as the
20 processing unit 914.

The system bus 918 can be any of several types of bus structure(s) including the memory bus or memory controller, a peripheral bus or external bus, and/or a local bus using any variety of available bus architectures including, but not limited to, 11-bit bus, Industrial Standard Architecture (ISA), Micro-Channel Architecture (MSA), Extended
25 ISA (EISA), Intelligent Drive Electronics (IDE), VESA Local Bus (VLB), Peripheral Component Interconnect (PCI), Universal Serial Bus (USB), Advanced Graphics Port (AGP), Personal Computer Memory Card International Association bus (PCMCIA), and Small Computer Systems Interface (SCSI).

The system memory 916 includes volatile memory 920 and nonvolatile memory
30 922. The basic input/output system (BIOS), containing the basic routines to transfer information between elements within the computer 912, such as during start-up, is stored

~~242~~
242

in nonvolatile memory 922. By way of illustration, and not limitation, nonvolatile memory 922 can include read only memory (ROM), programmable ROM (PROM), electrically programmable ROM (EPROM), electrically erasable ROM (EEPROM), or flash memory. Volatile memory 920 includes random access memory (RAM), which acts as external cache memory. By way of illustration and not limitation, RAM is available in many forms such as synchronous RAM (SRAM), dynamic RAM (DRAM), synchronous DRAM (SDRAM), double data rate SDRAM (DDR SDRAM), enhanced SDRAM (ESDRAM), Synchlink DRAM (SLDRAM), and direct Rambus RAM (DRRAM).

Computer 912 also includes removable/nonremovable, volatile/nonvolatile computer storage media. Fig. 9 illustrates, for example disk storage 924. Disk storage 924 includes, but is not limited to, devices like a magnetic disk drive, floppy disk drive, tape drive, Jaz drive, Zip drive, LS-100 drive, flash memory card, or memory stick. In addition, disk storage 924 can include storage media separately or in combination with other storage media including, but not limited to, an optical disk drive such as a compact disk ROM device (CD-ROM), CD recordable drive (CD-R Drive), CD rewritable drive (CD-RW Drive) or a digital versatile disk ROM drive (DVD-ROM). To facilitate connection of the disk storage devices 924 to the system bus 918, a removable or non-removable interface is typically used such as interface 926.

It is to be appreciated that Fig 9 describes software that acts as an intermediary between users and the basic computer resources described in suitable operating environment 910. Such software includes an operating system 928. Operating system 928, which can be stored on disk storage 924, acts to control and allocate resources of the computer system 912. System applications 930 take advantage of the management of resources by operating system 928 through program modules 932 and program data 934 stored either in system memory 916 or on disk storage 924. It is to be appreciated that the present invention can be implemented with various operating systems or combinations of operating systems.

A user enters commands or information into the computer 912 through input device(s) 936. Input devices 936 include, but are not limited to, a pointing device such as a mouse, trackball, stylus, touch pad, keyboard, microphone, joystick, game pad, satellite

~~243~~
243
243

dish, scanner, TV tuner card, digital camera, digital video camera, web camera, and the like. These and other input devices connect to the processing unit 914 through the system bus 918 *via* interface port(s) 938. Interface port(s) 938 include, for example, a serial port, a parallel port, a game port, and a universal serial bus (USB). Output device(s) 940 use some of the same type of ports as input device(s) 936. Thus, for example, a USB port may be used to provide input to the computer 912 and to output information from the computer 912 to an output device 940. Output adapter 942 is provided to illustrate that there are some output devices 940 like monitors, speakers, and printers among other output devices 940 that require special adapters. The output adapters 942 include, by way of illustration and not limitation, video and sound cards that provide a means of connection between the output device 940 and the system bus 918. It should be noted that other devices and/or systems of devices provide both input and output capabilities such as remote computer(s) 944.

Computer 912 can operate in a networked environment using logical connections to one or more remote computers, such as remote computer(s) 944. The remote computer(s) 944 can be a personal computer, a server, a router, a network PC, a workstation, a microprocessor based appliance, a peer device or other common network node and the like, and typically includes many or all of the elements described relative to computer 912. For purposes of brevity, only a memory storage device 946 is illustrated with remote computer(s) 944. Remote computer(s) 944 is logically connected to computer 912 through a network interface 948 and then physically connected *via* communication connection 950. Network interface 948 encompasses communication networks such as local-area networks (LAN) and wide-area networks (WAN). LAN technologies include Fiber Distributed Data Interface (FDDI), Copper Distributed Data Interface (CDDI), Ethernet/IEEE 1102.3, Token Ring/IEEE 1102.5 and the like. WAN technologies include, but are not limited to, point-to-point links, circuit switching networks like Integrated Services Digital Networks (ISDN) and variations thereon, packet switching networks, and Digital Subscriber Lines (DSL).

Communication connection(s) 950 refers to the hardware/software employed to connect the network interface 948 to the bus 918. While communication connection 950 is shown for illustrative clarity inside computer 912, it can also be external to computer

~~244~~
244

912. The hardware/software necessary for connection to the network interface 948 includes, for exemplary purposes only, internal and external technologies such as, modems including regular telephone grade modems, cable modems and DSL modems, ISDN adapters, and Ethernet cards.

5 Fig. 10 is a schematic block diagram of a sample computing environment 1000 with which the present invention can interact. The system 1000 includes one or more client(s) 1010. The client(s) 1010 can be hardware and/or software (*e.g.*, threads, processes, computing devices). The system 1000 also includes one or more server(s) 1030. The server(s) 1030 can also be hardware and/or software (*e.g.*, threads, processes, computing devices). The servers 1030 can house threads to perform transformations by employing the present invention, for example. One possible communication between a client 1010 and a server 1030 may be in the form of a data packet adapted to be transmitted between two or more computer processes. The system 1000 includes a communication framework 1050 that can be employed to facilitate communications
10 between the client(s) 1010 and the server(s) 1030. The client(s) 1010 are operably connected to one or more client data store(s) 1060 that can be employed to store information local to the client(s) 1010. Similarly, the server(s) 1030 are operably connected to one or more server data store(s) 1040 that can be employed to store information local to the servers 1030.
15

20 What has been described above includes examples of the present invention. It is, of course, not possible to describe every conceivable combination of components or methodologies for purposes of describing the present invention, but one of ordinary skill in the art may recognize that many further combinations and permutations of the present invention are possible. Accordingly, the present invention is intended to embrace all
25 such alterations, modifications, and variations that fall within the spirit and scope of the appended claims. Furthermore, to the extent that the term “includes” is used in either the detailed description or the claims, such term is intended to be inclusive in a manner similar to the term “comprising” as “comprising” is interpreted when employed as a transitional word in a claim.

~~245~~
245

CLAIMS

What is claimed is:

1. A system that facilitates classifying items in connection with spam prevention, comprising:
 - a component that receives a set of the items;
 - a component that identifies intended recipients of the items, and tags a subset of the items to be polled, the subset of items corresponding to a subset of recipients that are known spam fighting users; and
 - a feedback component that receives information relating to the spam fighter's classification of the polled items, and employs the information in connection with training a spam filter and populating a spam list.
2. The system of claim 1, wherein the items comprise at least one of: electronic mail (email) and messages.
3. The system of claim 1, wherein the component that receives a set of the items is any one of an email server, a message server, and client email software.
4. The system of claim 1, wherein the subset of items to be polled comprises all of the items received.
5. The system of claim 1, wherein the subset of recipients comprises all recipients.
6. The system of claim 1, wherein the subset of recipients are randomly selected.

28
246
256

7. The system of claim 1, wherein the subset of items tagged for polling is limited to at least one of the following:

- a number of the items selected per user;
- a number of the items selected per user per time period; and
- a probability of tagging an item corresponding to a known user.

8. The system of claim 1, wherein the tagged items are each assigned a unique ID, the unique ID corresponding to any one of the tagged item and contents of the tagged item.

9. The system of claim 1, further comprising a component that modifies an item tagged for polling to identify it as a polling item.

10. The system of claim 9, wherein the polling item comprises a summary of the tagged item, the summary comprising at least one of a subject, a date, text of the message, and a first few lines of the text.

11. The system of claim 9, wherein the polling item comprises voting instructions and any one of at least two voting buttons and links which correspond to at least two respective classes of items facilitate classification of the item by the user.

12. The system of claim 1, further comprising a central database that stores information and data relating to user properties, item content and properties associated with tagged items, user classification and voting statistical data, frequency analysis data of polling per user and of polling per user per time period, spam lists, legitimate mail lists, and black hole lists.

13. The system of claim 1 distributed across more than one spam-fighting company such that feedback from each company is sent to a central database operatively interfaced with each company, wherein some portion of the feedback is removed for privacy reasons.

88
247
247

14. The system of claim 1, further comprising a user classification validation component that tests user reliability and trustworthiness.

15. The system of claim 14, wherein the user classification validation component can be applied to one or more suspected users.

16. The system of claim 1, wherein the feedback component receives information relating to user feedback, honeypot feedback, and optionally, user recipient feedback of received items.

17. A method that facilitates classifying messages in connection with spam prevention comprising:

- receiving a set of the messages;
- identifying intended recipients of the messages;
- tagging a subset of the messages to be polled, the subset of messages corresponding to a subset of the recipients that are known spam fighting users;
- receiving information relating to the user's classification of polling messages; and
- employing the information in connection with training a spam filter and populating a spam list.

18. The method of claim 17, wherein the subset of recipients which are known spam fighting users is determined by each recipient performing at least one of the following:

- opting in to provide feedback on messages to facilitate training a new spam filter;
- passively opting in to provide feedback on messages by not opting out;
- paying for email and message services provided by a participating message server; and
- opening an email account with a participating message server.

~~248~~
248

19. The method of claim 17, wherein the subset of messages tagged for polling is limited by one or more polling limits.

20. The method of claim 17, further comprising modifying tagged messages to mark and identify them as polling messages.

21. The method of claim 20, wherein modifying the tagged messages comprises performing at least one of the following:

moving the tagged message to a separate folder for polling messages;

modifying the "from" address of the tagged message;

modifying the subject line of the tagged message;

using a polling icon on the tagged message to identify it as a polling message; and

using a unique color to identify the tagged message as a polling message.

22. The method of claim 17, further comprising scanning the tagged messages for viruses before they are downloaded for polling.

23. The method of claim 17, further comprising making a copy of each tagged message as originally received such that the respective users receive a first copy of the message in its original form and a second copy of the message in a form for polling.

24. The method of claim 17, further comprising distributing the trained spam filter to one or more servers, the distribution occurring automatically and/or by request by at least one of an email message and a posting on a website for downloading.

25. The method of claim 17, wherein training the spam filter and populating the spam list is performed by machine learning techniques using data based on user classification feedback and optionally, data generated by one or more additional sources,

the one or more sources comprising honeypots, recipient non-user classification feedback, and active learning techniques.

26. The method of claim 25, wherein data generated by the one or more sources is re-weighted proportionately with respect to the type of data generated by the source and relative to the user classification data to facilitate obtaining an unbiased sampling of data.

27. The method of claim 17 further comprising:
monitoring incoming messages for their respective one or more positive features;

determining a frequency of positive features received;

determining whether one or more positive features received exceeds a threshold frequency based at least in part upon historical data; and

quarantining suspicious messages, which correspond to the one or more positive features that exceed the threshold frequency, until further classification data is available to determine whether suspicious messages are spam.

28. The method of claim 27, wherein the feature used is information about the sender comprising at least one of the sender's IP address and domain.

29. The method of claim 27, wherein quarantining suspicious messages is performed by at least one of the following acts:

provisionally labeling the suspicious messages as spam and moving them to a spam folder;

delaying delivery of the suspicious messages to the user(s) until further classification data is available; and

storing the suspicious messages in a folder not visible to the user(s).

~~249~~
249

78
~~250~~
250

30. The method of claim 17, further comprising determining false positive and catch rates of the spam filter to facilitate optimization of the spam filter, wherein determining false positive and catch rates comprises:

training the spam filter using a training data set, the training data set comprising a first set of polling results;

classifying a second set of polling messages using user feedback to yield a second set of polling results;

running the second set of polling messages through the trained spam filter;

comparing the second set of polling results to the trained spam filter results to determine false positive and catch rates of the filter to thereby evaluate and tune filter parameters according to optimal filter performance.

31. The method of claim 30, wherein more than one spam filter is built, each having different parameters and each being trained on the same training data set, such that the false positive and catch rates of each spam filter is compared to at least one other spam filter to determine optimal parameters for spam filtering.

32. The method of claim 17, further comprising building an improved spam filter using additional sets of incoming messages, subsets of which are subjected to polling to yield new information in connection with training the improved spam filter, wherein previously obtained information is re-weighted based at least in part upon how long ago it was obtained.

33. The method of claim 17, further comprising employing the information to build a legitimate sender list.

34. The method of claim 17, further comprising employing the information to facilitate terminating accounts of spammers.

35. The method of claim 34, further comprising identifying a spammer who is using an ISP and automatically notifying the ISP of the spamming.

22
251
251

36. The method of claim 34, further comprising identifying a domain responsible for sending spam, and automatically notifying at least one of the domain's email provider and the domain's ISP of the spamming.

37. The method of claim 17, further comprising distributing at least one of the spam filter and the spam list to any one of mail servers, email servers, and client email software, wherein distributing comprises at least one of the following:

- posting a notification on a website notifying that the spam filter and spam list are available for downloading;

- automatically pushing the spam filter and spam list out to mail servers, email servers and client email software; and

- manually pushing the spam filter and spam list out to mail servers, email servers, and client email software.

38. A cross-validation method that facilitates verifying reliability and trustworthiness of user classifications comprising:

- excluding one or more suspected user's classifications from data employed to train the spam filter;

- training the spam filter using all other available user classifications; and

- running the suspected user's polling messages through the trained spam filter to determine how it would have classified the messages compared to the suspected user's classifications.

39. The method of claim 38, further comprising performing at least one of the following:

- discounting existing and future classifications provided by users who are determined to be untrustworthy until the users are determined to be trustworthy;

- discarding existing classifications provided by users determined to be untrustworthy; and

- removing the untrustworthy users from future polling.

28
252
252

40. A method that facilitates verifying reliability and trustworthiness in user classifications for training a spam filter *via* a feedback loop system comprising:

identifying a subset of spam-fighting users as suspect users;
providing one or more messages having a known result to the suspect users for polling; and

determining whether the suspected users' classification of the one or more test messages matches the known classification to ascertain the reliability of the users' classifications.

41. The method of claim 40, wherein the subset of spam-fighting users identified as suspect users comprises all users.

42. The method of claim 40, wherein the message is a test message that is known to be at least one of spam and good mail and that is injected into a stream of incoming mail by the feedback loop system and delivered to the suspect users.

43. The method of claim 40, wherein the message received by the suspected users for polling is hand-classified by a system administrator to train the spam filter with a correct classification to identify untrustworthy users.

44. The method of claim 40, further comprising at least one of the following acts:

discounting existing and future classifications provided by users who are determined to be untrustworthy until the users are determined to be trustworthy;

discarding existing classifications provided by users determined to be untrustworthy; and

removing the untrustworthy users from future polling.

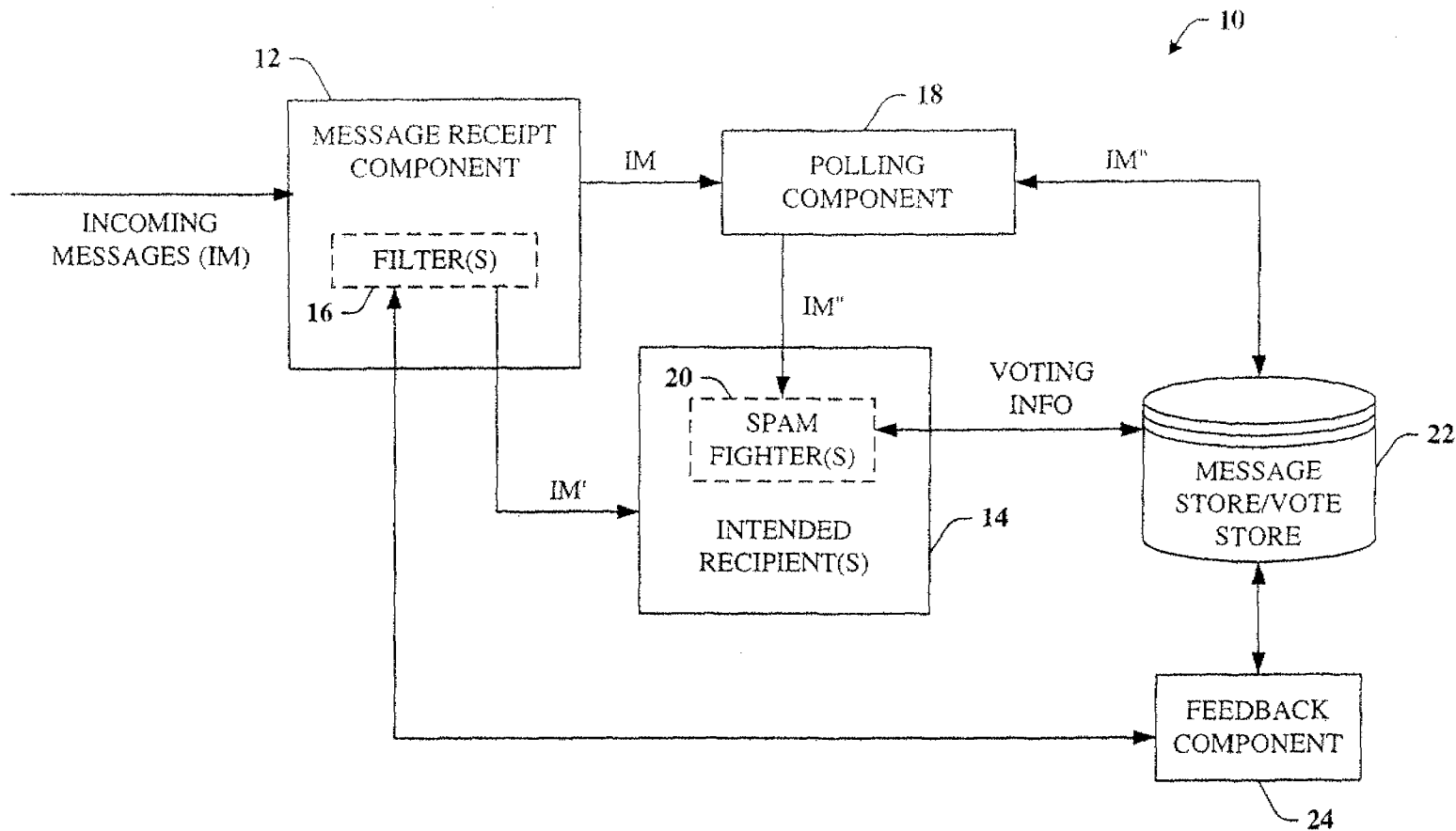


FIG. 1A

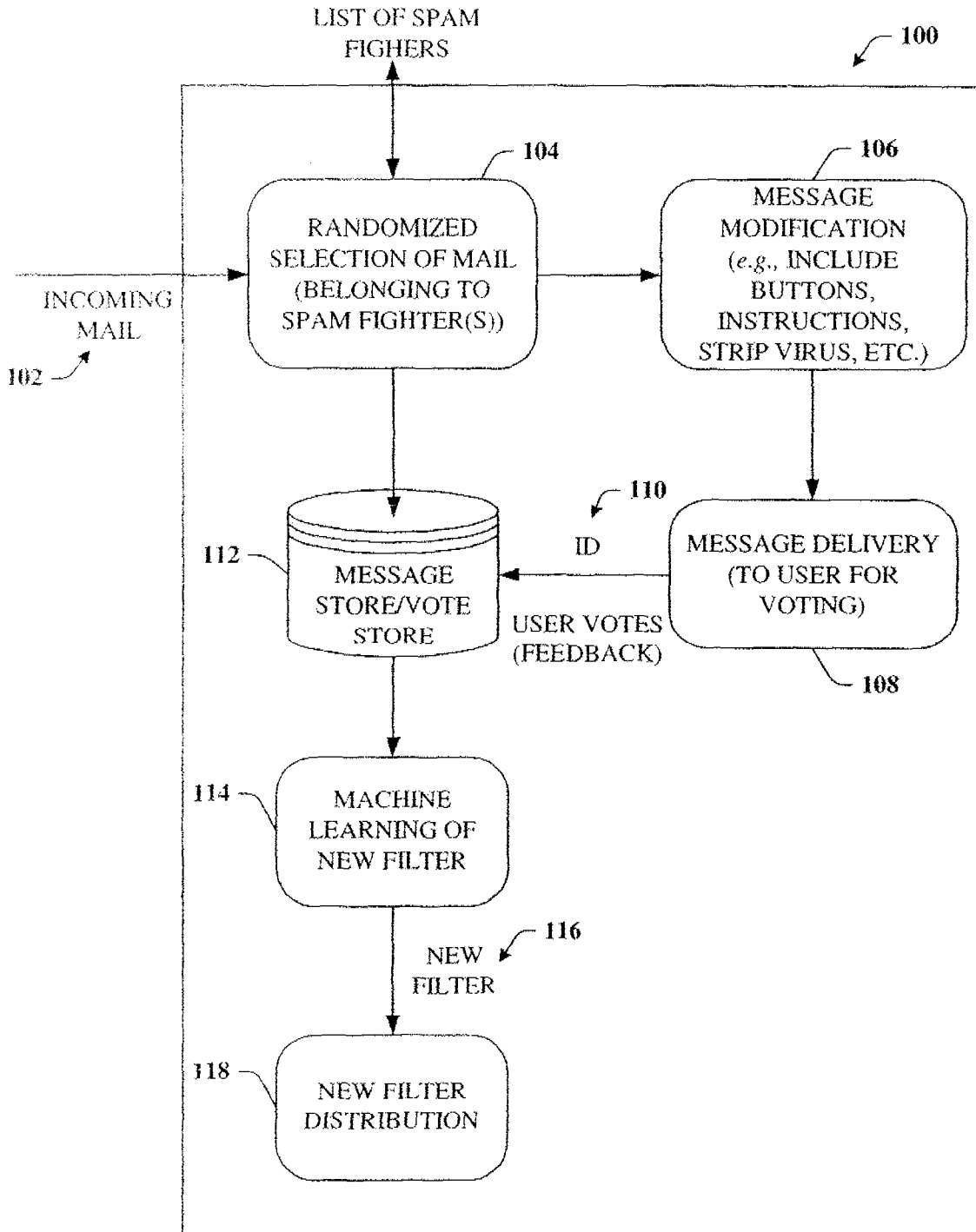


FIG. 1B

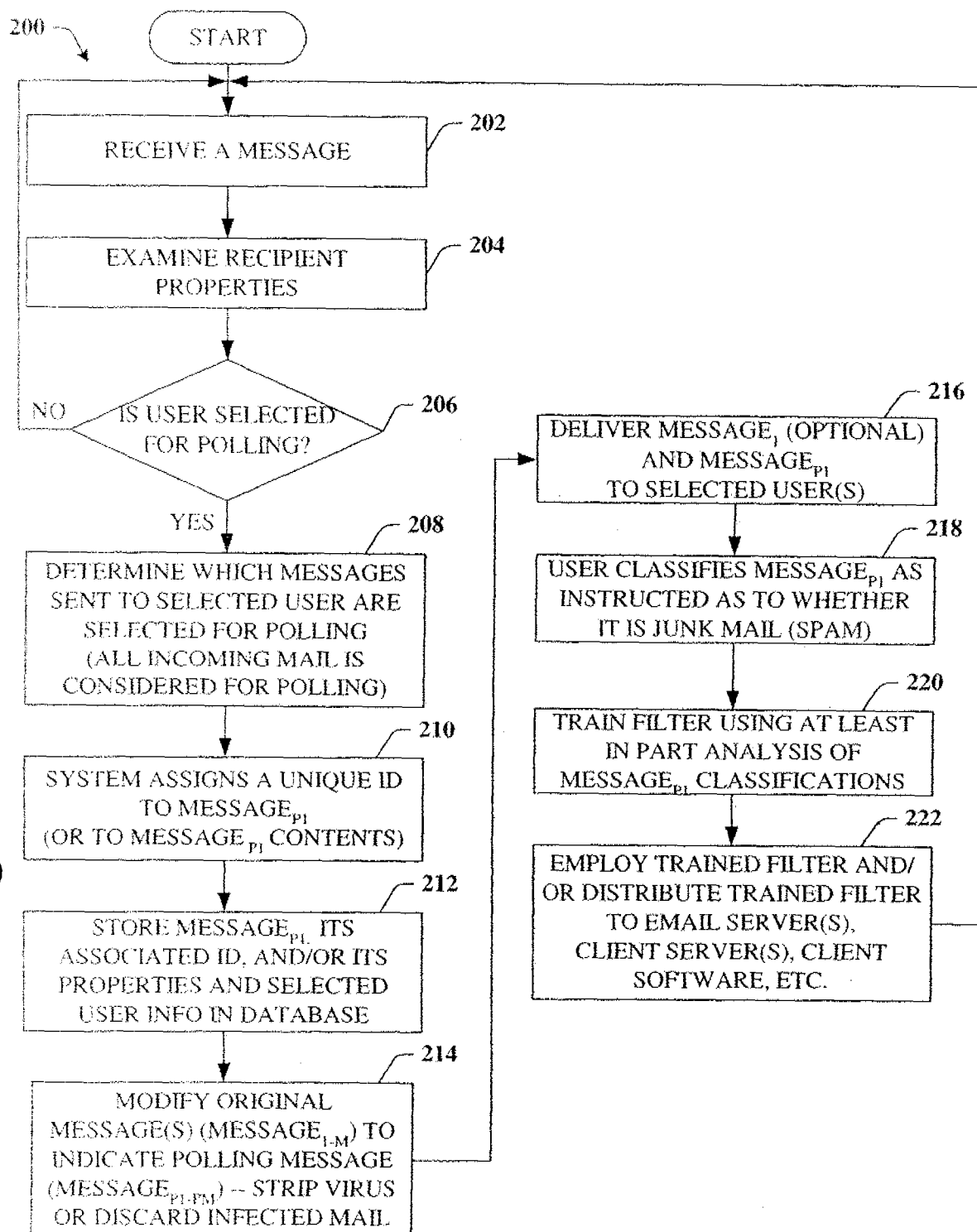


FIG. 2

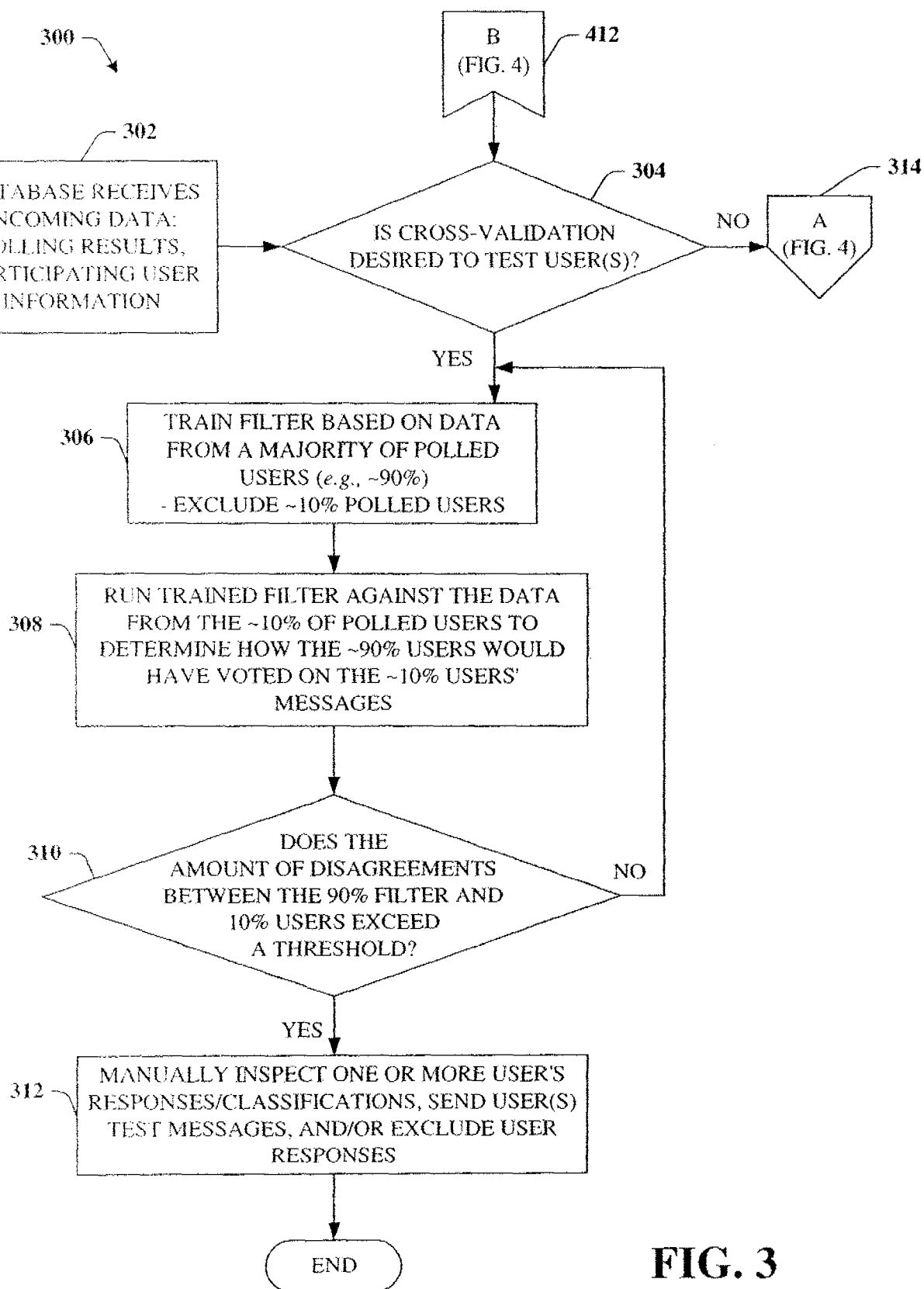


FIG. 3

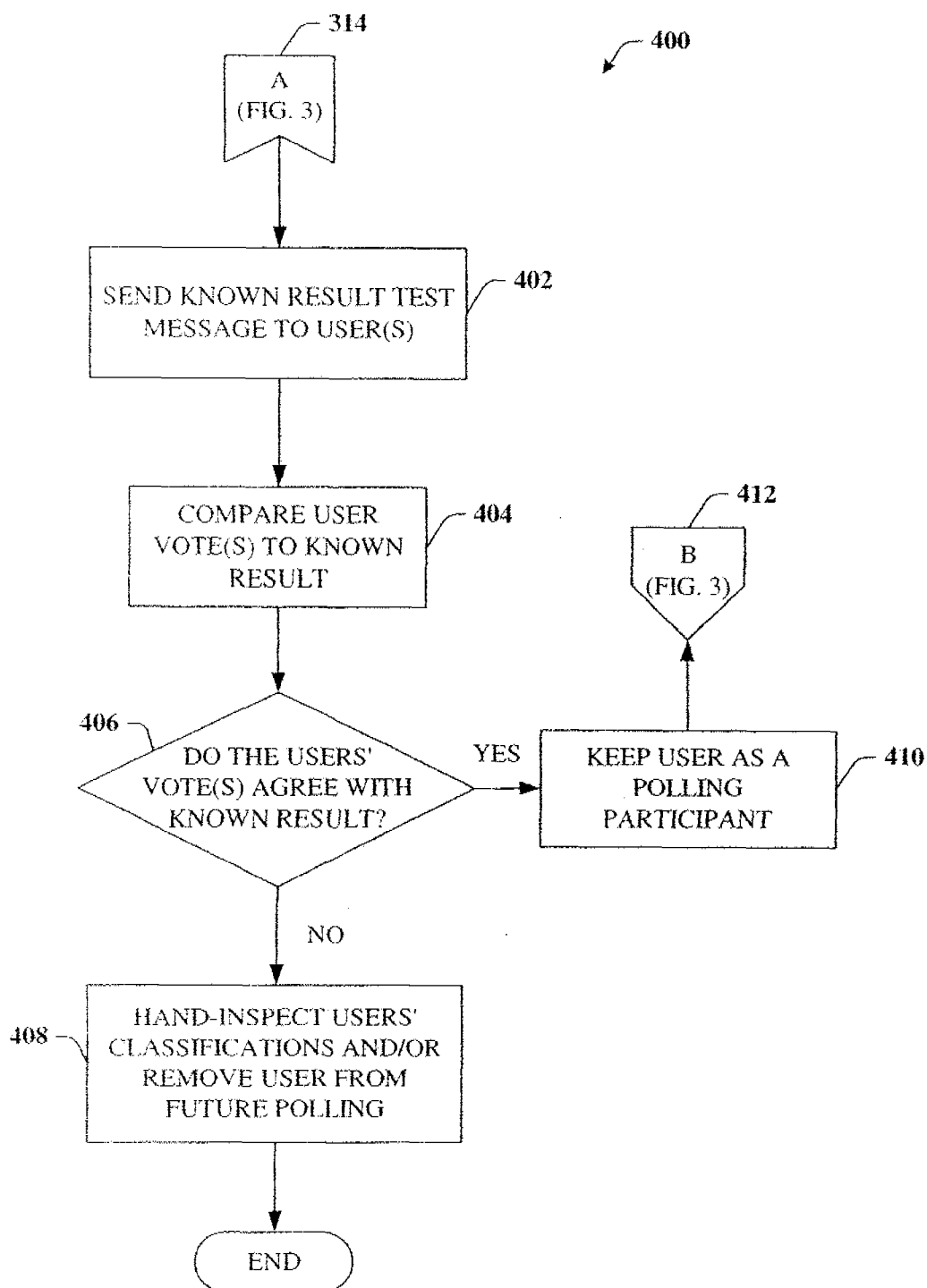


FIG. 4

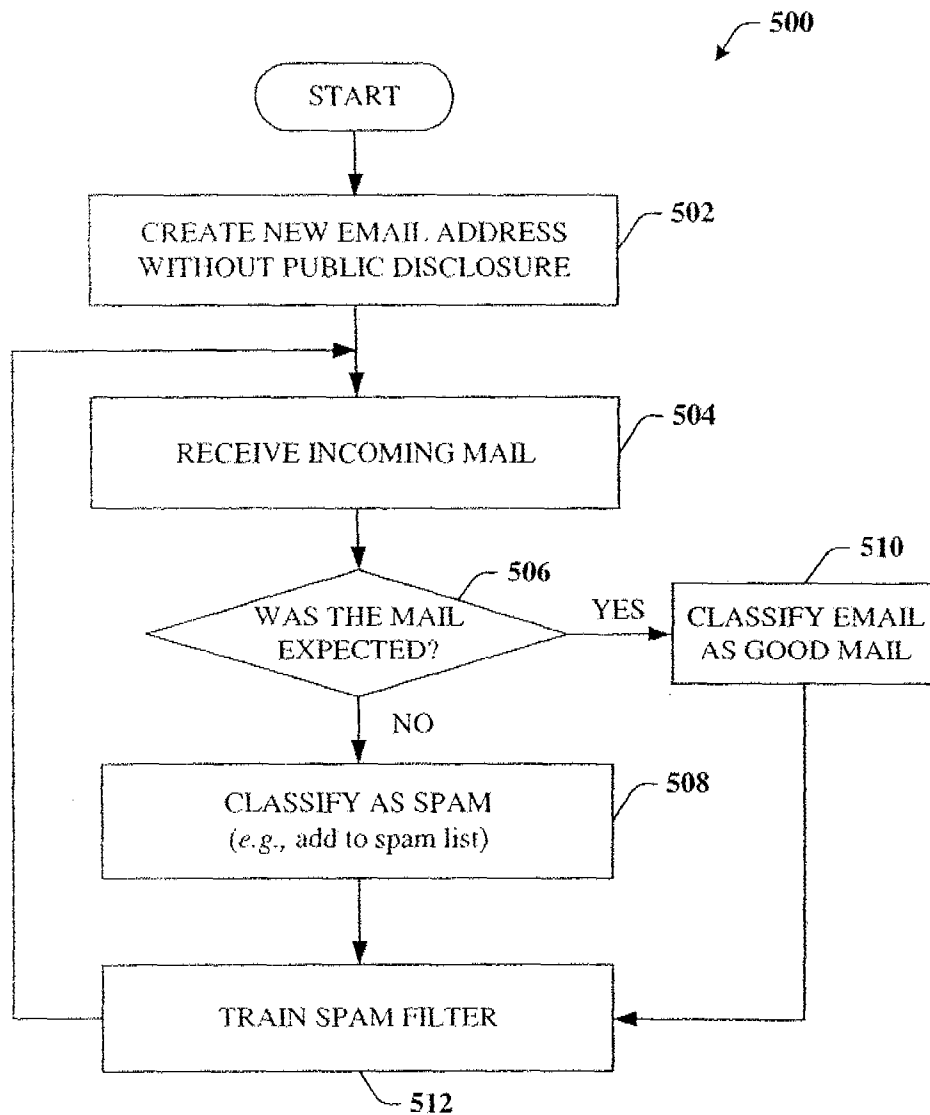


FIG. 5

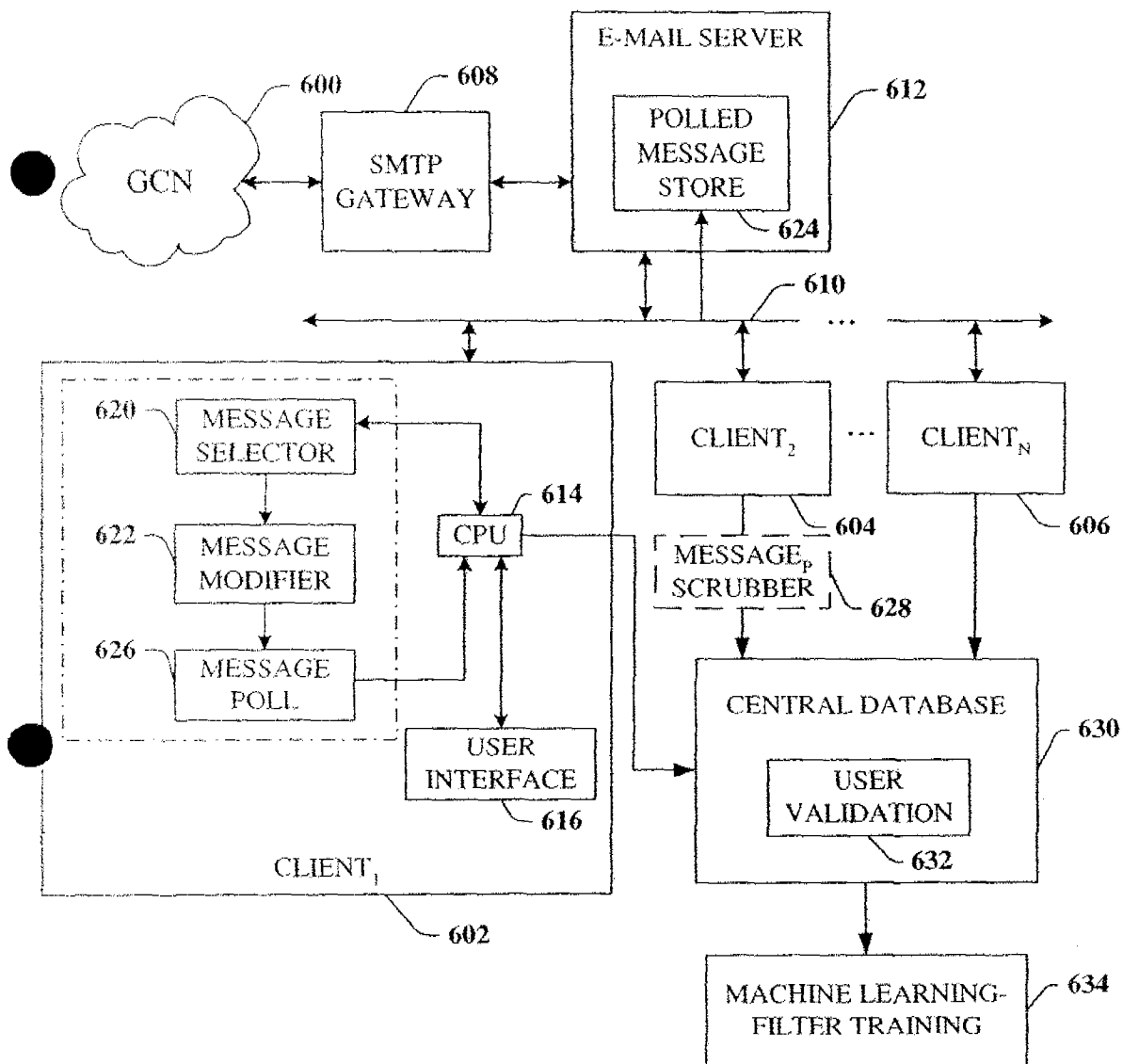


FIG. 6

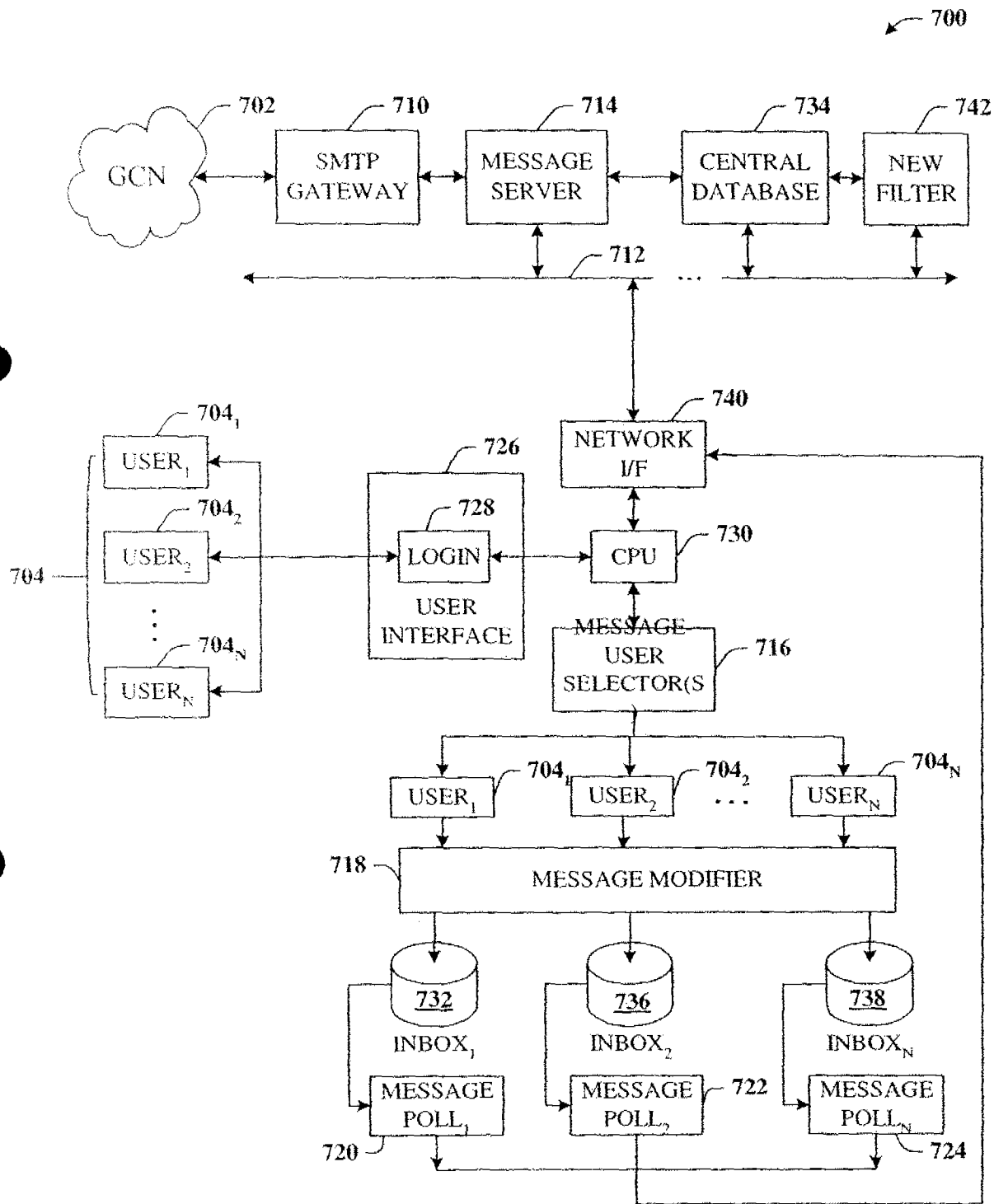


FIG. 7

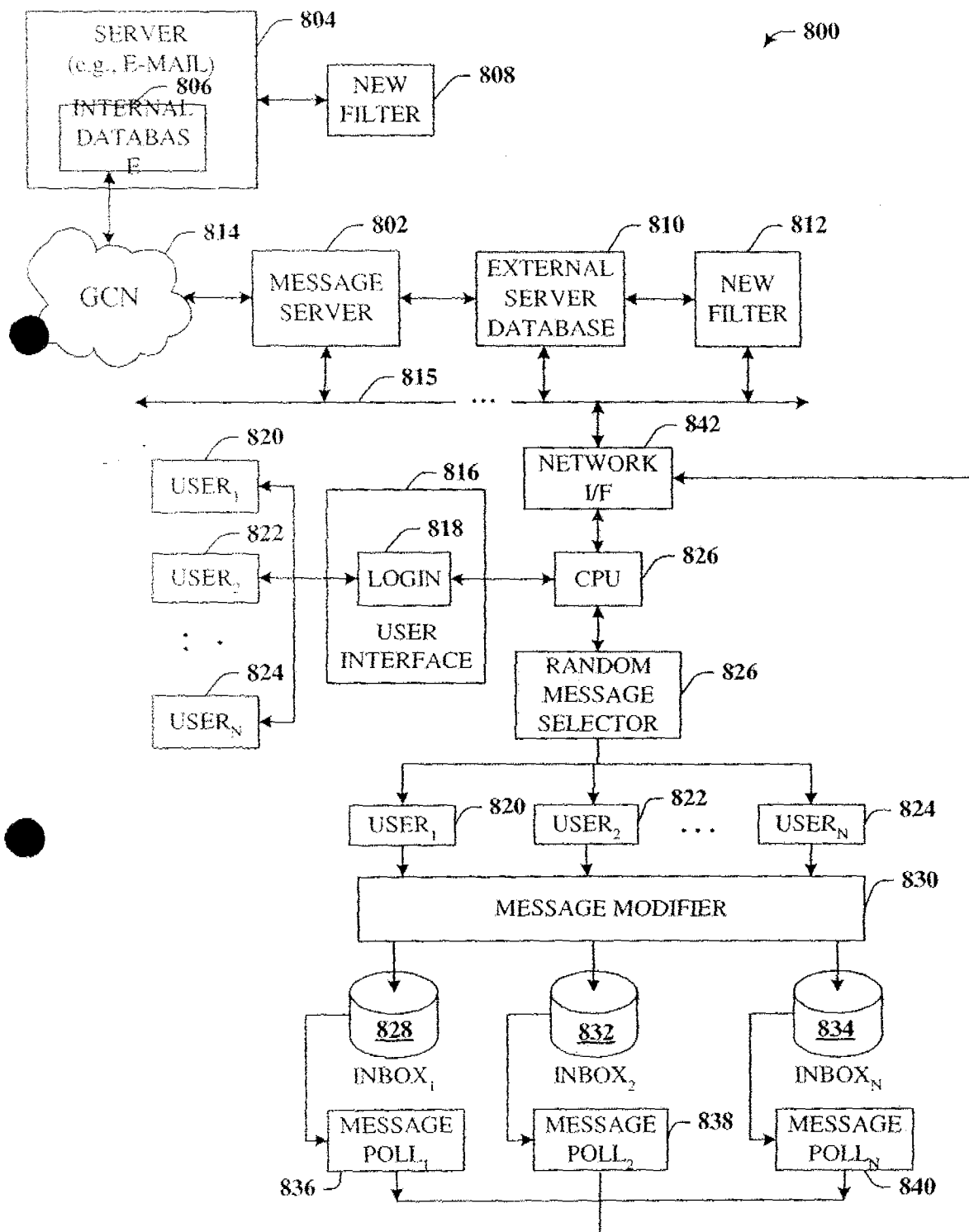


FIG. 8

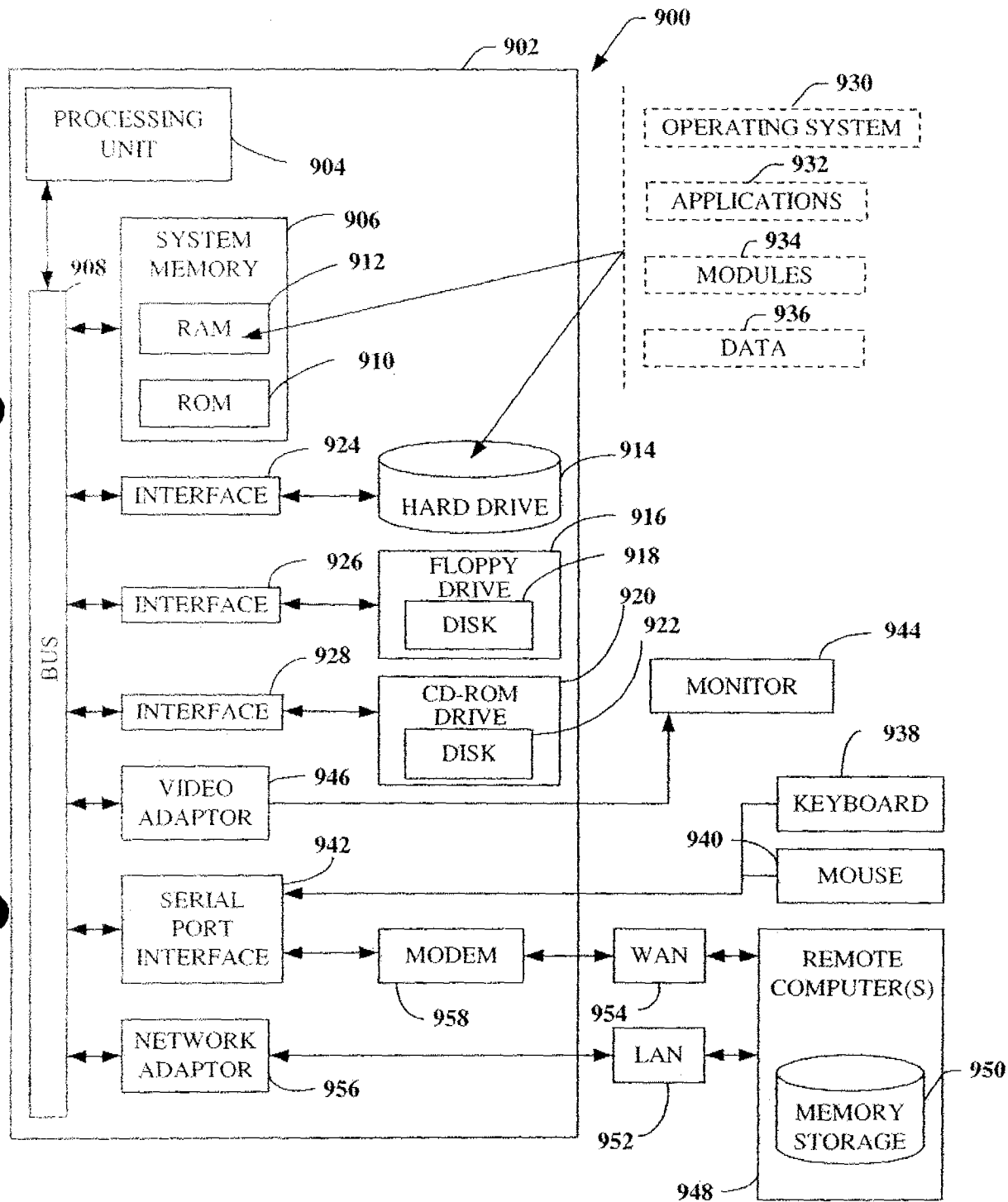
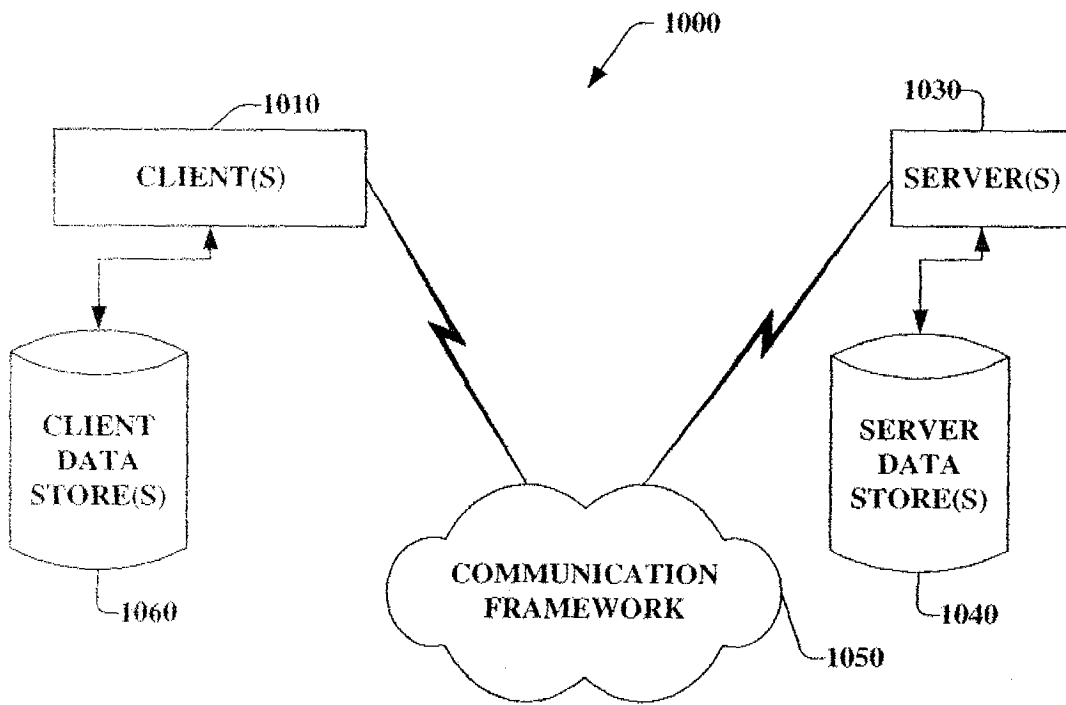


FIG. 9

**FIG. 10**

88
264
264

Atty. Radicado No. MS302992.2 (MSFTP416WO)

**RETROALIMENTACION CICLICA PARA LA PREVENDION DE
SPAM**

Por

CORPORACION MICROSOFT

CERTIFICACION

~~265~~
265

Yo certifico que la patente de la aplicación adjunta SOLICITUD CAPITULO II (así como cualquier otro documento referido adjunto o incluido) está siendo depositado al Servicio Postal de los Estados Unidos en esta fecha Febrero 25, 2004, en un sobre como "Oficina Postal de Correo Expreso Direccionado" Número de Etiqueta de Correo EV330020761US direccionado a: ESTACIÓN DE CORREO PCT, Comisionado de Patentes, P.O. Box 1450, Alexandria , VA 22313-1450

Heather Holmes

(Nombre de la Persona Escrito o Impreso que Envía el Documento)

(ilegible)

Firma de la Persona que Envía el Documento

266
266

Título: RETROALIMENTACIÓN CÍCLICA PARA LA PREVENCIÓN DE SPAM

CAMPO TÉCNICO

El invento se relación con sistemas y métodos para identificar tanto a la información legítima como a la indeseada (correo basura), y particularmente para clasificar correspondencia de correo electrónico para la prevención de spam.

ANTECEDENTES DEL INVENTO

El advenimiento de las redes de comunicaciones globales tales como Internet han presentado oportunidades comerciales para alcanzar un gran número de clientes potenciales. La mensajería electrónica, y en particular el correo electrónico ("e-mail"), ha ido convirtiéndose en forma persuasiva como un medio para diseminar propaganda indeseada y promociones (también llamadas "spam") a los usuarios de la red.

La Radicate Group Inc. una firma de consultoría a y mercadeo, estima que para agosto del 2002, 2 billones de mensajes de e-mail basura se envían cada día -se espera que este número se triplique cada dos años. Los individuos y las entidades (por ejemplo, agencias del gobierno, negocios) se encuentran cada vez más incómodos y a veces ofendidos por los mensajes basura. Como tal, los e-mails basura son ya o pronto serán una gran amenaza a la computación confiable.

Una técnica particular utilizada para prevenir e-mail basura es el uso de sistemas y metodologías de filtrado. Una técnica de filtrado probada se basa en un proceso de aprendizaje de la máquina -filtros que la máquina aprende se le asignan a los mensajes entrantes como una probabilidad de que el mensaje es basura. En éste enfoque, se extraen típicamente rasgos de dos clases de mensaje

~~267~~
267

ejemplo (por ejemplo, mensajes basura y no basura), y un filtro de aprendizaje se aplica para discriminar las probabilidades existentes entre estas dos clases. Como muchos rasgos de mensaje se relacionan con el contenido (por ejemplo, palabras y frases en el asunto y/o cuerpo de mensaje), estos tipos de filtro se conocen comúnmente como, "filtros basados en el contenido".

Algunos filtros basura/spam se pueden adaptar, lo que es importante cuando existen muchos usuarios de distintos idiomas, algunos de los cuales hablan lengua extrañas y necesitan un filtro que se puede adaptar a sus necesidades específicas. Aún más, no todos los usuarios están de acuerdo en que es y no es basura/spam. De esta manera, utilizando un filtro que se pueda entrenar implícitamente (por ejemplo, a través de la observación del comportamiento del usuario), el filtro respectivo se pueden diseñar en forma dinámica para atender las necesidades particulares de identificación de mensajes.

Un enfoque para filtrar las adaptaciones es solicitarle al usuario que marquen los mensajes como basura o no basura. Desafortunadamente, esta técnica de entrenamiento manual intensivo es indeseable para muchos usuarios por la complejidad asociada con este entrenamiento, y mucho más, el tiempo que se requiere para efectuar dicho entrenamiento en forma apropiada. Adicionalmente, esta técnica de entrenamiento manual frecuentemente falla a nivel del usuario individual. Por ejemplo, al usuario se le olvida frecuentemente la suscripción a listas gratuitas de correo y por lo tanto, estos mensajes se identifican incorrectamente como correo basura. Como resultado, un correo legítimo se bloquea indefinidamente de la casilla de correo del usuario. Otra técnica de filtrado que se puede adaptar es emplear implícitamente claves de entrenamiento. Por ejemplo, si un usuario responde o reenvía un mensaje, el enfoque asume que el mensaje no es basura. Sin embargo, el solo uso de claves de mensajes de este

268
268

tipo introduce una desviación estadística en el proceso de entrenamiento, resultando en filtros de más baja precisión.

Otro enfoque es el de utilizar para el entrenamiento todos los e-mails del usuarios, en donde un filtro existente asigna marcas iniciales mientras el usuario a veces descarta esa marcación con claves explícitas (por ejemplo, un método de "corrección del usuario") -por ejemplo, opciones de selección como "eliminar como basura" y "no basura"- y/o claves implícitas. Aunque estos enfoques son mejores que la técnica que se discutió anteriormente, todavía es deficiente comparado con el invento que se describe y se reivindica a continuación.

RESUMEN DEL INVENTO

Lo siguiente presenta un resumen simplificado del invento para proveer un entendimiento básico de algunos aspectos del invento. Este resumen no es una descripción extensiva del invento. Su intención no es identificar elementos críticos del invento o delinear su alcance. El único propósito es presentar algunos conceptos del invento en forma simplificada como un preámbulo a la descripción más detallada que se presenta más abajo.

Este invento provee un sistema y método de retroalimentación cíclica que facilita la clasificación de los items en conexión con la prevención de spam. El invento utiliza un enfoque de aprendizaje de máquina para ser aplicado a los filtros spam, y en particular, chequea en forma aleatoria los mensajes entrantes de tal manera que ejemplos de correo legítimo y basura/spam se obtienen para generar un conjunto de datos de entrenamiento. Individuos pre-seleccionados sirven como luchadores contra el spam y participan categorizando réplicas respectivas (que ocasionalmente se pueden modificar) de los ejemplos.

Generalmente, los mensajes seleccionados para revisión se modifican en varios aspectos para aparecer como mensajes de revisión. Un aspecto único del invento es que una copia del mensaje entrante se selecciona para revisión de tal manera que un usuario (por ejemplo, un luchador de spam) recibirá el mismo mensaje (por ejemplo, en términos del contenido del mensaje) dos veces: uno en la forma de un mensaje para revisión y luego, en su forma original. Otro aspecto único del presente invento es que todos los mensajes se consideran para revisión -incluyendo aquellos que se han marcado como spam por filtros existentes. Los mensajes marcados como spam se consideran para revisión y si se seleccionan, no se tratan como spam de acuerdo a las especificaciones del filtro existente (por ejemplo, mover aún folder basura, eliminar...).

A diferencia de los filtros de spam convencionales, se pueden crear filtros de spam más precisos con un filtro de entrenamiento de spam de acuerdo con la técnica de retroalimentación del presente invento, de tal manera que se aprende a distinguir entre correo legítimo y spam, y de esta manera se mitiga el filtrado impreciso. La retroalimentación se logra en parte revisando un número adecuado de usuarios para obtener una retroalimentación en su correo entrante. Los usuarios, identificados como luchadores contra el spam, tienen la tarea de escoger si una selección de mensajes entrantes son legítimos o basura. Tanto la clasificación positiva como la negativa del correo entrante mitigan el filtrado impreciso de correo legítimo en spam para un usuario determinado. La clasificación respectiva junto con otra información asociada con cada transacción de e-mail se mueve a una base de datos para facilitar el entrenamiento del filtro spam. La base de datos y los componentes relacionados pueden compilar y almacenar propiedades para un mensaje seleccionado (o transacciones de correo seleccionadas), que incluye las propiedades del usuario, la información de votación del usuarios y su historia, propiedades del mensaje tales como un número único de identificación asignado a cada mensaje seleccionado,

92
270
270

clasificaciones de mensajes, y resúmenes del contenido del mensaje, o datos estadísticos relacionados con cualquiera de los anteriores, para generar conjuntos de datos de entrenamiento para un sistema de aprendizaje de máquina.

Los sistemas de aprendizaje de máquina, (redes neurales, Máquinas de Soporte Vectorial (SVMs), Redes de menú Bayesian Belief) facilitan la creación de filtros spam mejorados que se pueden entrenar para reconocer tanto al correo legítimo como al spam y más aún, distinguir entre los dos. Una vez que un filtro spam nuevo se ha entrenado de acuerdo con este invento, se puede distribuir a servidores de correo y a programas de software para los clientes. Aún más, el nuevo filtro spam se puede entrenar con respecto a un usuario específico para mejorar la ejecución de un filtro personalizado. En la medida que nuevos conjuntos de datos de entrenamiento se construyen, el filtro spam puede sobrellevar entrenamiento adicional a través de aprendizaje de máquina para optimizar su ejecución y precisión. La retroalimentación del usuario a través de la clasificación de mensajes también se puede utilizar para generar listas para filtros de spam y control, para probar filtros spam, y/o identificar el origen del spam.

Otro aspecto del presente invento provee un método para detectar usuarios no confiables a través de técnicas de validación cruzada y/o mensajes de prueba y con resultado conocido. El validación cruzada incluye un filtro de entrenamiento de donde el resultado de la consulta de algunos usuarios se excluye. Esto es, el filtro se entrena utilizando resultados de consulta de un subconjunto de usuarios. En promedio, este conjunto de usuarios funcionará bien incluso con algunos errores para detectar aquellos que generalmente no están de acuerdo con ellos. El resultado de la consulta de los usuarios excluidos se compara con los del filtro entrenado. Esta comparación determina esencialmente como los usuarios del subconjunto de entrenamiento hubieran votado sobre los mensajes pertenecientes a los usuarios excluidos. Si el acuerdo entre un voto de un usuario excluido y el

92
27
281

filtro es bajo, el resultado de la consulta de ese usuario se puede descartar o marcar para inspección manual. Esta técnica se puede repetir como se desee, excluyendo cada vez datos de distintos usuarios.

También se pueden detectar errores en mensajes individuales tales como mensajes en los cuales el voto del usuario y del filtro están fuertemente en desacuerdo. Estos mensajes se pueden marcar para eliminación automática y/o inspección manual. Como una alternativa a la validación cruzada, un filtro se puede entrenar con todos o sustancialmente con todos los usuarios. Los votos y/o mensajes del usuario que están en desacuerdo con el filtro se pueden descartar. Otra alternativa a la validación cruzada tiene que ver con mensajes de una prueba de resultados conocidos con los cuales se le pide al usuario que vote sobre el mensaje donde el resultado es conocido. La clasificación precisa (el voto del usuario está de acuerdo con la acción del filtro) del mensaje por un usuario verifica la confiabilidad del usuario y determina si se remueve la clasificación del usuario del entrenamiento, y si remueve al usuario de consultas futuras.

Aún más, otro aspecto del presente invento facilita la creación de objetivos spam conocidos (por ejemplo, honeypots) para identificar correo entrante como spam y/o hacer un seguimiento al procesamiento de direcciones de e-mail de comerciantes específicos. Un spam objetivos conocido, o honeypot, es una dirección de e-mail donde un conjunto de correo legítimo se puede determinar y todo el resto de correo considerado como spam. Por ejemplo, la dirección de e-mail se puede presentar en la red en forma restrictiva de tal manera que no sea fácil encontrarla. Por lo tanto, cualquier correo enviado a esta dirección se puede considerar como spam. En forma alternativa, la dirección de e-mail puede haber sido presentada únicamente a un comerciante del cual se espera recibir correo legítimo. Por lo tanto, el correo recibido de este comerciante es legítimo, pero todo los demás correos se pueden considerar spam. Datos de spam derivados de un

74
272
272

honeypot y/o otras fuentes (por ejemplo, usuarios) se pueden integrar al sistema cíclico de retroalimentación, pero por el aumento sustancial en la clasificación de spam con honeypots, estos datos se deben aligerar, tal como se describirá *infra* en más detalle, para mitigar los resultados influenciados de la consulta.

Otro aspecto del presente invento facilita la cuarentena de mensajes que se consideran dudosos por el sistema cíclico de retroalimentación o por el filtro. Estos mensajes se mantienen por un período determinado de tiempo en vez de eliminarse o clasificarse. Este periodo de tiempo se puede fijar por adelantado, o el mensaje se puede retener hasta que la recepción de un determinado número de resultados de consulta similares al mensaje, por ejemplo, la misma dirección IP o con contenido similar.

Para lograr lo anterior y otros resultados relacionados, algunos aspectos ilustrativos del presente invento se mostrarán en relación con la descripción y dibujos anexos. Éstos aspectos son indicativos, sin embargo, de algunos de las varias maneras de los cuales los principios del invento se pueden utilizar y el presente invento incluye todos estos aspectos o sus equivalentes. Otras ventajas y rasgos novedosos del invento pueden ser aparentes de la siguiente descripción detallada del invento cuando se consideran en conjunto con los dibujos.

DESCRIPCIÓN RESUMIDA DE LOS DIBUJOS

La figura 1A es un diagrama de bloque de un sistema de entrenamiento de retroalimentación cíclica de acuerdo con un aspecto del presente invento.

La figura 1B es un diagrama de flujo de un proceso de entrenamiento de retroalimentación cíclica de acuerdo con un aspecto del presente invento.

85
273
273

La figura 2 es un diagrama de flujo de un método ejemplar que facilita la clasificación del correo por los usuarios para crear filtros de spam de acuerdo con un aspecto del presente invento.

La figura 3 es un diagrama de flujo de un método ejemplar que facilita la validación cruzada de usuarios que participan en el método de la figura 2 de acuerdo con un aspecto del presente invento.

La figura 4 es un diagrama de flujo de un método ejemplar que facilita la determinación de si un usuario es confiable de acuerdo con un aspecto del presente invento.

La figura 5 es un diagrama de flujo de un método ejemplar que facilita la detección de spam y determina el origen de acuerdo con un aspecto del presente invento.

La figura 6 es un diagrama de bloque de una arquitectura de retroalimentación cíclica basadas en el cliente de acuerdo con un aspecto del presente invento.

La figura 7 es un diagrama de bloque de un sistema de retroalimentación cíclica basada en un servidor que tiene uno o más usuarios para generar datos de entrenamiento de acuerdo con un aspecto del presente invento.

La figura 8 es un diagrama de bloque de un sistema de retroalimentación cíclica basada en un servidor a través de una organización donde el sistema incluye un servidor interno con su propia base de datos para generar datos de entrenamiento almacenados en una base de datos externos de acuerdo con un aspecto del presente invento.

La figura 9 muestra un ambiente ejemplar para implementar los varios aspectos del presente invento.

274
274

La figura 10 es un diagrama de bloque esquemático de un ambiente de comunicación ejemplar de acuerdo con el presente invento.

DESCRIPCIÓN DETALLADA DEL INVENTO

El invento presente ahora se describe en referencia a los dibujos, donde numerales de referencia iguales se utilizan para hacer referencia a elementos iguales. En la siguiente descripción, con el propósito de ilustración, numerosos detalles específicos se muestran para suministrar un entendimiento del presente invento. Puede ser evidente, sin embargo, que el presente invento se puede practicar sin estos detalles específicos en otras instancias. Estructuras bien conocidas y dispositivos se muestran en un diagrama de bloque para facilitar la descripción del presente invento.

Como se utiliza en esta aplicación, los términos "componente" y "sistema" tienen la intención de referirse a entidades relacionadas con los computadores, ya sea hardware, una combinación de hardware y software, software, o software que está en ejecución. Por ejemplo, un componente puede ser, pero no está limitado a ser, un proceso que corre en un procesador, un procesador, un objeto, un ejecutable, un hilo de ejecución, un programa, y/o un computador. Como ilustración, tanto una aplicación que corre en un servidor y el servidor pueden ser un componente. Uno o más componentes pueden residir dentro de un proceso y/o un hilo de ejecución y un componente puede estar localizado en un computador y/o distribuido entre dos o más computadores.

El presente invento puede incorporar varios esquemas de inferencia y/o técnicas relacionadas con la generación de datos de entrenamiento para el filtrado de spam aprendido por la máquina. Tal como se utiliza acá, el término "inferencia" se refiere generalmente al proceso de razonamiento sobre el sistema dado siendo

82
~~274~~
274

los estados, ambientes, y/o usuarios de un conjunto de observaciones como se reciben a través de los eventos y/o datos. La inferencia se puede utilizar para identificar una acción o contexto específico, puede generar una distribución de probabilidades sobre estados, por ejemplo. La inferencia puede ser probabilística - esto significa, el cómputo de una distribución de probabilidades sobre estados de interés basados en la consideración de eventos y datos. La inferencia también se puede referirse a técnicas que se utilizan para componer eventos de alto nivel de un conjunto de eventos y/o datos. Tales resultados de la inferencia son la construcción de nuevos eventos o acciones de un conjunto de eventos observados y/o almacenados, así los eventos estén correlacionados en proximidad temporal, y si los eventos y datos vienen de una o más fuentes de datos y eventos.

Se debe apreciar que aunque el término mensaje se utiliza en forma extensiva a través de la especificación, tal término no está limitado a correo electrónico per se, pero se puede adaptar en forma adecuada para incluir la mensajería electrónica de cualquier forma que se pueda distribuir a través de una arquitectura de comunicación adecuada. Por ejemplo, las aplicaciones de conferencias que facilita la conferencia entre dos o más personas (por ejemplo, programas de chat interactivo, o programas de mensajería instantánea) también se pueden utilizar para filtrado los beneficios descritos acá, puesto que texto indeseado se puede filtrar en forma electrónica en mensajes de chat en la medida que los usuarios intercambian mensajes y/o son insertados como mensajes iniciales, mensajes de cierre, o todos los anteriores. En esta aplicación particular, se puede entrenar un filtro para que automáticamente entre el contenido particular de un mensaje (texto e imágenes) para capturar y etiquetar el contenido indeseable como basura (por ejemplo, comerciales, promociones, o avisos).

En el presente invento, el término "receptor" se refiere el destinatario de un mensaje entrante o item. El término "usuarios" se refiere al receptor que ha

98
275
275

escogido, ya sea en forma pasiva o activa, participar en el sistema de retroalimentación cíclica y procesos que se describen acá.

En referencia ahora a la figura 1A, se muestran un diagrama de bloque de un sistema de entrenamiento de retroalimentación 10 de acuerdo con un aspecto del presente invento. Un componente de recepción de mensajes 12 recibe y envía el mensaje entrante (que se denota con IM) al receptor pretendido 14. El componente de recepción del mensaje puede incluir por lo menos un filtro 16 como se acostumbra con muchos componentes de recepción de mensajes (por ejemplo filtros de correo basura) para mitigar la distribución de mensajes indeseables (por ejemplo, spam). El componente de recepción de mensajes 12 en conexión con el filtro 16 procesa el mensaje (IM) y suministra el uso conjunto de mensajes filtrados (IM) al receptor pretendido 14.

Como parte del aspecto de retroalimentación del presente invento, un componente de consulta 18 recibe todos los mensajes entrantes (IM) e identifica al receptor pretendido respectivamente 14. El componente de consulta selecciona un subconjunto de los receptores pretendidos 14 (referidos como luchadores contra el spam 20) para clasificar un subconjunto de los mensajes entrantes (denotados como IM 18) como spam o no spam, por ejemplo. La información relacionada con la clasificación (que se denota como INFORMACION DE VOTACION) se entrega a un almacén 22 de mensajes/votos, donde la información de votación como también copias de la respectivamente IM se almacenan para uso más tarde como lo hace el componente de retroalimentación 24. En particular, el componente de retroalimentación 24 utiliza técnicas de aprendizaje de máquina (por ejemplo redes neurales, SVMs, redes Bayesianas o cualquier sistema de aprendizaje de máquina adecuado para uso con el presente invento) que utiliza la información de votación para entrenar y/o mejorar al filtro 16 (y/o construir nuevos filtros) con respecto a la identificación de correo no spam, por ejemplo. En la medida que flujos nuevos de

276
276

mensajes entran y son procesados a través de un nuevo filtro entrenado 16, menos spam y más mensajes legítimos (de anotados como IM□) se entregan al receptor pretendido 14. Así, el sistema 10 facilita la identificación de spam y el entrenamiento mejorado de filtros spam utilizando la retroalimentación generada por los luchadores contra el spam 20. Este aspecto de retroalimentación del presente invento suministra un esquema rico y altamente dinámico para refinar los sistemas de detección de spam. Varios detalles en cuanto los aspectos más particulares del invento se discuten a continuación.

En referencia ahora a la figura 1B se ilustra un diagrama de flujo de un entrenamiento de retroalimentación cíclica 100 en conexión con la lucha y prevención contra el spam de acuerdo con el presente invento. En preparación para y/o antes del proceso de entrenamiento, se selecciona a los usuarios para ser luchadores contra spam (por ejemplo, de un conjunto maestro que consta de todos los usuarios de e-mail) -la selección se puede basar en una muestreo aleatorio, o nivel de confianza, o cualquier selección de esquema/criterios adecuada de acuerdo con el presente invento. Por ejemplo, el subconjunto seleccionado de usuarios puede incluir a todos los usuarios, un conjunto aleatorio de usuarios seleccionados, aquellos que se han inscrito como luchadores contra el spam, o aquellos que no se han retirado, y/o cualquier combinación de los anteriores, y/o basada en parte en su localidad demográfica e información relacionada.

En forma alterna, el conjunto maestro de usuarios de e-mail seleccionado puede estar limitado a usuarios que pagan lo que hace más costoso a los spammers infringir en el presente invento. Así, un subconjunto de usuarios seleccionado para participar en la lucha contra el spam puede constar únicamente de usuarios que pagan. Una lista o una tabla de clientes que incluyen los nombres y las

100
277
277

propiedades de los usuarios seleccionados (por ejemplo luchadores contra spam) se puede crear.

Cuando un flujo nuevo de mensajes entrantes 102 se recibe, se chequea al receptor de cada mensaje contra una lista de luchadores contra spam 104. Si al receptor está en la lista, el mensaje se considera para consulta. Luego, se determina si se selecciona un mensaje para consulta. A diferencia de los filtros spam convencionales, el invento no elimina ningún mensaje (por ejemplo, spam) hasta que todo el correo entrante se considere para consulta. Esto significa, el correo se clasifica antes de ponerle una etiqueta (por ejemplo spam, no spam) - esto facilitar la obtención de unos mensajes de prueba no prejuizados que el usuario tiene disponibles para consulta.

Un componente para la selección de mensajes (que no se muestra) se puede utilizar para seleccionar los mensajes con alguna probabilidad aleatoria para mitigar el error en los datos. Otro enfoque tiene que ver con el uso de información demográfica como también atributos usuario/receptor y propiedades. Así, los mensajes se pueden seleccionar basados en parte en el usuario/receptor. Otro algoritmo alternativo existe para seleccionar mensajes. Sin embargo, puede haber limitantes en el número de mensajes seleccionados por usuario o por usuario por periodo de tiempo, o en la probabilidad de seleccionar un mensaje de un usuario dado. Sin estos límites, un spamer puede crear una cuenta, enviar millones de mensajes spam, y clasificar todos estos mensajes como buenos: esto le permitiría al spamer corromper la base de datos de entrenamiento con mensajes marcados en forma incorrecta.

Algunas formas de filtrado de spam, llamadas listas de hoyo negro no se pueden evadir. Las listas de hoyo negro previenen a un servidor de recibir cualquier correo de una lista de direcciones de Protocolo de Internet (IP). Por lo

701
~~278~~
278

tanto, la selección de mensajes se puede hacer de un conjunto de correo que no es de una lista de hoyo negro.

Un aspecto único en invento que es que los mensajes seleccionados para consulta, y que están marcados como spam por filtros actualmente funcionando, no se eliminan o se mueven a un folder de correo basura. A cambio, se colocan en una casilla de correo entrante donde todos los mensajes que se reciben se consideran para consulta. Sin embargo, si hay dos copias del mensaje, quien mensaje se considera como spam por un filtro, entonces una copia se le entrega al folder de spam o de lo contrario se trata de acuerdo con parámetros que se han fijado (por ejemplo, eliminados, marcados en forma especial, o movidos a un folder basura).

Cuando un mensaje se selecciona, se le reenvía al usuario y se marca de alguna forma especial para indicar que es un mensaje de consulta. En particular, los mensajes seleccionados se pueden modificar por componente de modificación de mensajes 106. Ejemplos de modificación de mensajes incluyen, pero no están limitados a, guardar el mensaje en consulta en un folder separado, cambiar la dirección "de" o la línea del asunto, y/o utilizando un icono especial o un color especial que identifique los mensajes como mensajes de consulta al usuario. El mensaje seleccionado luego se puede encapsular dentro de otro mensaje, que le da instrucciones al usuario de cómo debe votar y/o clasificar el mensaje encapsulado. Estas instrucciones pueden incluir por lo menos dos botones o vínculos: por ejemplo, uno que elige el mensaje como spam y uno que elige el mensaje como no spam.

Los botones de votación se pueden implementar modificando el contenido del mensaje antes de enviar una copia del mensaje que se va a consultar al usuario. Cuando el invento se utiliza con respecto a un software de e-mail del

279
279

cliente (a diferencia de un servidor de correo), la interfase del usuario no se puede modificar para incluir los botones de votación.

Aún más, el mensaje que se consulta puede contener instrucciones y botones de votación como también el mensaje seleccionado anexo. El mensaje que se consulta también puede contener un resumen del mensaje seleccionado tal como la línea de asunto, dirección remitente, fecha de envío y/o recibido, y el texto o por lo menos las primeras líneas del texto. Otro enfoque involucra el envío del mensaje con instrucciones de votación y botones de votación anexos. En la práctica, cuando un usuario abre y/o baja una copia del mensaje de consulta, botones (o vínculos) incluyendo, pero no limitado a, botones "spam" y "no spam" pueden aparecer en la interfase del usuario o se pueden incorporar en el mensaje que se consulta. Por lo tanto, es posible que cada mensaje de consulta contenga una serie de instrucciones y botones adecuados para la votación. Otras modificaciones pueden ser necesarias, incluyendo la posibilidad de remover las instrucciones HTML que se encuentran en el trasfondo (que podrían oscurecer el texto de las instrucciones o botones.)

Otro botón tal como el botón "e-mail comercial solicitado" también se pueden suministrar, dependiendo del tipo de información que se desea. El mensaje también puede incluir una botón/vínculo para editar consultas futuras. Las instrucciones se localizan en el lenguaje preferido del usuario y se pueden incrustar en el mensaje de consulta.

Aún más, los mensajes seleccionados para consulta se pueden revisar para virus por el componente de modificador del mensaje 106 o por otro componente de revisión de virus adecuado (que no se muestra). Si un virus se encuentra, el virus se puede remover o el mensaje se puede eliminar. Se debe apreciar que la remoción de un virus puede ocurrir en cualquier punto del sistema 100, inclusive cuando el mensaje se selecciona o antes de bajar el mensaje.

~~785~~
~~285~~
286

Siguiendo a las modificaciones del mensaje, un componente de entrega de mensajes 108 entrega el mensaje que se va a consultar al usuario para la votación. La retroalimentación del usuario (por ejemplo mensaje de consulta, el voto del usuario, o cualquier otra propiedad asociada) se le asigna un identificador único (ID) 110 (por ejemplo, de datos). El ID 110 y/o la información correspondiente se envían a un almacén de mensajes/votos 112 (por ejemplo, una base de datos en total), donde la clasificación/votos del usuario se compilan y almacenan.

En el nivel de la base de datos, los mensajes seleccionados disponibles para consulta se pueden mantener para consultas posteriores o su utilización. Adicionalmente, la base de datos puede ejecutar análisis frecuentes en un tiempo determinado para asegurarse que un usuario particular no está siendo consultado en exceso y que la cantidad de datos que se están recolectando del usuario dentro de ciertos límites especificados por el usuario. En particular, el sistema de retroalimentación 100 monitorea un porcentaje límite del correo del usuario como también períodos de prueba para mitigar los prejuicios tanto de las muestras como de los datos. Esto es especialmente importante donde los usuarios se selecciona de todos los usuarios disponibles, incluyendo a los usuarios de bajo y alto uso. Por ejemplo, un usuario de bajo uso recibe típicamente y envía un volumen significativamente menor de correo comparado con un usuario de alto uso. Por lo tanto, el sistema 100 monitorea el proceso de selección para asegurarse que el mensaje seleccionado es aproximadamente uno de un número T de mensajes recibidos por el usuario y no más de un mensaje recibido cada Z horas por el usuario. De esta manera, el sistema puede consultar uno de cada diez mensajes entrantes para probar (por ejemplo, considerado para consulta), pero no más de 1 cada 2 horas, por ejemplo. El límite de la frecuencia, o porcentaje, mitiga la prueba desproporcionada de mensajes para un usuario de bajo uso comparado con un usuario de alto uso, y también mitiga la molestia frecuente al usuario.

784
281
281

En una base frecuente, la base de datos central 112 busca mensajes que han sido probados por el sistema 100 para consulta pero que no se han clasificado. La base de datos localiza estos mensajes relativos a las propiedades demográficas del usuario y crea mensajes de consulta para solicitarle el voto al usuario y clasificar el mensaje(s). Sin embargo, el filtro spam puede no ser modificado o entrenado inmediatamente luego de la recepción de cada clasificación entrante. Más bien, el entrenamiento fuera de línea le permite al entrenador buscar continuamente en los datos recibidos en la base de datos 112 en una base diaria, permanente o programada. Esto quiere decir, que el entrenador comienza de un punto de partida prescrito o en un tiempo determinado en el pasado y chequea todos los datos posteriores a ese punto para entrenar al filtro. Por ejemplo, el tiempo prescrito puede ser desde la medianoche hasta las 6 a.m.

El nuevo filtro de spam se puede entrenar continuamente analizando la clasificación del mensaje que se mantiene en la base de datos 112 a través de técnicas de aprendizaje de máquina 114 (por ejemplo, redes neurales, máquinas de soporte de vectores (SVM)). Las técnicas de aprendizaje de máquina requieren ambos ejemplos de correo bueno y spam para aprender de tal manera que puedan distinguir entre ellos. Inclusive técnicas que se basan en la comparación de ejemplos conocidos de spam pueden beneficiarse con ejemplos de correo bueno, de tal manera que se aseguren que accidentalmente no confundan un buen correo.

De esta manera, es importante tener ejemplos tanto de spam positivo como negativo, en vez de sólo quejas. Hay algunos dominios que envían grandes cantidades tanto de spam como de correo legítimo tales como las listas gratuitas de correo. Si uno construye un sistema basado únicamente en quejas, todos los correos de estos dominios se pueden filtrar lo que resultaría en una gran cantidad

765
282
282

de errores. Por lo tanto, sabiendo que el dominio también envía gran cantidad de correo bueno es importante. Adicionalmente, los usuarios frecuentemente cometen errores tales como olvidarse que se han inscrito a una lista gratuita de correo. Por ejemplo, un proveedor grande legítimo tal como el New York Times envía correo legítimo en forma regular. Algunos usuarios olvidan que se han afiliado y luego se quejan, clasificando estos mensajes como spam. Sin los datos suficientes para darse cuenta que este correo legítimo, correo de este sitio puede ser bloqueado.

El nuevo filtro 116 se puede distribuir en forma continua por el componente de distribución 118 a través de los servicios de Internet (ISP) participantes, a los servidores de correo, a clientes individuales, para actualizar un servidor, y/o a la base de datos centros de compañías individuales. Aún más, el sistema de retroalimentación 100 funciona en una base continua de tal manera que copias de los mensajes considerados y utilizados para consulta pueden seguir una distribución actual de e-mail recibido por el sistema 100. Como resultado, conjuntos de datos de entrenamiento utilizados para entrenar el nuevo filtro spam se mantienen actualizados con respecto a spammers que se adaptan. Cuando se construyen nuevos filtros, los datos de consulta se pueden descartar o aligerar (por ejemplo, de descontar) basándose en el periodo de tiempo desde que se obtuvo.

El sistema 100 se puede implementar cuando se recibe correo de un servidor tal como un servidor de entrada, un servidor de e-mail, y/o un servidor de mensajes. Por ejemplo, cuando correo entra en un servidor de e-mail, el servidor mira las propiedades del receptor pretendido para determinar si el receptor es apropiado por el sistema 100. Si las propiedades así lo indican, el receptor de correo potencialmente está disponible para consulta. Arquitecturas de sólo-cliente también existe. Por ejemplo, software de e-mail del cliente puede tomar las

706
~~285~~
783

decisiones de consulta para un usuario individual y entregar el e-mail ya sea a una base de datos centros o utilizar la información de consulta para mejorar la ejecución del filtro personalizado. Adicionalmente a los que se describen acá, otras arquitecturas alternativas para este sistema 100 existen y se considera que ellas caen dentro del alcance del presente invento.

En referencia ahora a la figura 2, se muestra un diagrama de flujo de un proceso de retroalimentación cíclica básico 200 de acuerdo con un aspecto del presente invento. En mientras que la metodología que se muestra que se describe como una serie de actos para simplicidad en la explicación se debe entender y apreciar que el presente invento no está limitado por el orden de los actos, como algunos actos pueden, de acuerdo con el presente invento, ocurrir en distinto orden y/o concurrentes con otros actos de aquellos que se muestran y se describen acá. Por ejemplo, aquellos diestros en el arte entenderán y apreciarán que una metodología se podría representar alternativamente como una serie de estados o eventos interrelacionados, tales como en un diagrama de estado. Aún más, no todos los actos frustrados se pueden requerir para implementar la metodología de acuerdo con el presente invento.

El proceso 200 comienza con correo que entra y lo recibe un componente tal como el servidor en 202. Cuando el correo llega al servidor, el servidor identificar las propiedades del receptor pretendido para determinar si el receptor pretendido ha optado por enterrar como un luchador contra el spam para consulta (en 204). Así el proceso 200 utiliza un campo de propiedad del usuario donde se puede indicar si el receptor ha optado por el sistema de retroalimentación, o consulta una lista de usuarios que han entrado. Si se determina que el usuario es un participante en el sistema de retroalimentación y ha sido seleccionado para consulta en 206, el sistema de retroalimentación, acción determinando en mensajes se seleccionan para consulta (en 208). De lo contrario, el proceso 200

7/88
285
285

"spam", por ejemplo. Cuando el usuario hace clic en un botón (en 218) para clasificar el mensaje como correo bueno o spam, al usuario se le dirige a un localizador uniforme de recursos (URL) que corresponde a un identificador único para la clasificación del usuario solicita. Esta información y el récord asociado se almacenan en la base de datos central para ese mensaje original que se actualiza.

En 216 o en cualquier otro momento propicio durante el proceso 200, el mensaje original se puede enviar opcionalmente al usuario. De esta manera, del usuario recibe el mensaje dos veces -una en su forma original y luego en la forma modificada de consulta.

Luego más tarde, un filtro de spam nuevo se crea y se entrena en 220 basándose en parte en la retroalimentación del usuario. Una vez que el nuevo filtro de spam se ha creado y entrenado, se puede utilizar inmediatamente en el servidor de e-mail y/o distribuir a un servidor, software de e-mail del cliente, y demás (en 222). Entrenar y distribuir un nuevo filtro actualizado de spam es una actividad recurrente de esta manera, el proceso 200 continúa en 204 cuando un nuevo flujo de mensajes entrantes se recibe. Cuando se construyen filtros nuevos, los datos viejos se descartan o se aligeran basándose en el tiempo desde cuando se obtuvieron.

El sistema de retroalimentación 100 y el proceso 200 dependen en la retroalimentación de los usuarios participantes. Desafortunadamente, no se puede confiar en algunos usuarios o son demasiado perezosos y no proveen una clasificación consistente y precisa. La base de datos central 112 (figura 1a) mantiene las historias de la clasificación de los usuarios. De esta manera, el sistema de retroalimentación 100 puede hacer un seguimiento del número de contradicciones, en número de veces que el usuario ha cambiado su mente, respuestas del usuario a e-mail conocido como bueno o como spam, como

704
286
936

también en número por la frecuencia en el usuario contesta las encuestas de mensajes.

Cuando cualquiera de estos números excede un umbral prescrito, o simplemente para cada usuario del sistema, el sistema de retroalimentación 100 puede invocar una o más técnicas de validación para evaluar la confiabilidad de un usuario particular. Un enfoque es un método de validación cruzada 300 cómo se muestra en la figura 3 de acuerdo con otro aspecto del presente invento.

Las técnicas de validación cruzada comienzan en 302 con una base de datos central que recibe datos entrantes tales como los resultados de las consultas y la información del usuario respectivo. Luego, se debe determinar si se desea que la validación cruzada pruebe un número adecuado de usuarios en 304. Si se desea, luego, un nuevo filtro de spam se entrena utilizando alguna porción de los datos entrantes en 306. Lo que esto significa, es que los datos del usuario que se están probando se excluyen del entrenamiento. Por ejemplo, el filtro se entrena con el 90% de los datos de consulta del usuario (llamado un filtro de 90%), excluyendo a sí cerca de 10% de los datos (llamado ser 10% probado por usuario) que corresponde a los datos suministrados por el usuario que se aprueba.

En 308, el filtro de 90% corre contra los datos del 10% probado por usuario para determinar cómo el 90% de usuarios habría votado sobre el mensaje del usuario probado. Si el desacuerdo entre el filtro de 90% y el 10% de los datos probados por usuario exceden un umbral prescrito (en 310), entonces la clasificación del usuario se puede inspeccionar manualmente en 302. En forma alternativa o adicionalmente, mensajes de prueba se pueden enviar a los usuarios sospechosos o no confiables y/o estos usuarios particulares se pueden excluir de consultas futuras, y/o sus datos pasados eliminados. Sin embargo, si el umbral no se exceden, entonces el proceso regresa a 306. En la práctica, la técnicas de validación cruzada 300 se pueden utilizar con cualquier cantidad

~~46~~
~~287~~
287

adecuada de usuarios, excluyendo distintos usuarios como sea necesario para determinar mantener la confiabilidad en la votación/clasificación de datos.

Un segundo enfoque para evaluar la fidelidad del usuario y la confiabilidad incluye un filtro de entrenamiento en todos los datos que se recogen en un determinado período de tiempo, y luego probar con los datos de entrenamiento utilizando el quinto. Esto técnicas se conoce como prueba-en-entrenamiento. Si un mensaje se incluyó en el entrenamiento, el filtro debe haber aprendido su rating, por ejemplo, el filtro que ha aprendido debe clasificar el mensaje de la misma manera que el usuario lo hizo. Sin embargo, el filtro puede continuar cometiendo errores cuando lo considera que es spam y el usuario no o viceversa. Para que un filtro este en desacuerdo con los datos de entrenamiento, el mensaje debe estar fuertemente en desacuerdo con otros mensajes. De lo contrario, el filtro entrenado habría casi con certeza encontrado alguna forma de clasificarlo correctamente. Así, el mensaje se puede descartar como teniendo una etiqueta no confiable.

Se puede utilizar ya sea esta técnica o la validación cruzada: la validación cruzada puede cometer más errores en clasificaciones menos confiables; en forma converso la prueba-en-entrenamiento encuentra menos errores en clasificación es más confiables.

Tanto la prueba-en-entrenamiento y la técnicas de validación cruzada 300 se pueden aplicar a mensajes individuales donde la clasificación individual y un usuario se excluye del compromiso general (por ejemplo, siguiendo la evaluación mayoritaria). En forma alterna como ambas técnicas se pueden utilizar para identificar a usuarios potenciales no confiables.

Adicionalmente a, o envés de la validación cruzada y/o técnicas de prueba-en-entrenamiento, podemos utilizar la técnicas de "resultado-conocido" para verificar la confiabilidad de un usuario (de 314 en la figura 4). Aunque las técnicas de la figura 3 y 4 se muestran en forma separada, se debe apreciar que ambos

74
~~288~~
288

enfoques se pueden utilizar al mismo tiempo. Esto es, la información de mensajes conocidos como buenos o spam se puede combinar con la validación cruzada o los resultados de la prueba-en-entrenamiento para determinar qué usuario se debe descartar.

En referencia ahora a la figura 4, se ilustra un diagrama de flujo del proceso 400 para validar la fidelidad de la votación del usuario de acuerdo con un aspecto del presente invento. El proceso 400 se refiere a 314 cómo se muestra en la figura 3. En 400 todos, un mensaje prueba de resultado conocido se envía al usuario sospechoso (o a todos los usuarios). Por ejemplo, un mensaje de prueba se puede incrustar en el correo entrante y luego clasificar mano de tal manera que la base de datos reside el resultado "conocido". De lo contrario, el proceso 400 puede esperar hasta que un mensaje de resultado conocido se ha enviado por un tercer partido. A los usuarios se les permite votar en el mismo mensaje de prueba. Los resultados de la votación se comparan con los resultados conocidos en 404. Si los votos de los usuarios no están de acuerdo en 406, entonces su actual y/o futuras y/o clasificaciones pasadas se pueden inspeccionar a mano por un período adecuado de tiempo (en 408) hasta que se demuestre la consistencia y confiabilidad. En forma alterna en su clasificación actual futuras o pasada se puede descartar puede eliminar. Finalmente, el usuario se puede eliminar de consultas futuras. Sin embargo, si los resultados de la votación están de acuerdo con los resultados del mensaje de prueba, entonces el usuario se considera confiable en 410. El proceso 400 regresa en 412 a la figura 3 para determinar qué tipo de técnica de validación se desea para el grupo siguiente de usuarios sospechosos.

Un cuarto enfoque (que no se muestra) para evaluar la confiabilidad de un usuario es el aprendizaje activo. Con las técnicas de aprendizaje activo, los mensajes no se escogen en forma aleatoria. En cambio, el sistema de

42
~~289~~
289

retroalimentación puede estimar que tan útil es el mensaje para el sistema. Por ejemplo, si el filtro evalúa una probabilidad de spam, uno puede en forma preferencial seleccionar los mensajes que el filtro actual probablemente no clasifique para consulta, por ejemplo, aquellos cuyo probabilidad de spam es cerca a 50%. Otra manera de seleccionar los mensajes es determinar qué tan común es el mensaje. Entre más común sea el mensaje, más sutil es para la consulta. Mensajes únicos son menos útiles porque son menos comunes. El aprendizaje activo se puede emplear utilizando los niveles de confianza de los filtros actuales, utilizando los rasgos comunes de los mensajes, y utilizando los niveles de confianza de los filtros existentes del contenido (por ejemplo, meta confianza). Hay muchas otras técnicas de aprendizaje activo, tales como indagación-por-comité, bien conocida a aquellos diestros en el arte del aprendizaje de máquina, cualquiera de estas técnicas se puede utilizar.

En referencia ahora a la figura 5, se muestra un diagrama de flujo del proceso 500 para incorporar retroalimentación de honeypot adicional a la retroalimentación del usuario en el entrenamiento del filtro de spam de acuerdo con un aspecto del presente invento. Los honeypot son direcciones de correo en las cuales se conoce quién envía los e-mail. Por ejemplo, una dirección de e-mail nueva creada se puede mantener privada y entregarse únicamente a individuos seleccionados (en 502). También se pueden entregar en forma pública pero de manera restrictiva que la gente no ve (por ejemplo, con un tipo de letra blanca en un fondo blanco y como un vínculo del correo). Los honeypot son particularmente útiles en ataques de diccionario por los spammers. Un ataque de diccionario es uno en donde un spammer trata de enviar e-mails a una gran cantidad de direcciones, tal vez todos los direcciones en un diccionario o hechos de pares de palabras en un diccionario o técnicas similares para tratar de encontrar direcciones válidas. Cualquier e-mail enviado a un honeypot (en 504) o cualquier e-mail que no sea de dos individuos seleccionados (en 506 se considera como spam (en 508). Una

~~290~~
290

dirección de email también puede estar firmada por un comerciante sospechoso. Por lo tanto, cualquier e-mail que se recibe de un comerciante se considera un buen correo (en 510) pero todos los otros correos se consideran spam. El filtro de spam se puede entrenar de acuerdo (en 512). Aún más, se determina que el comerciante sospechoso vende o dará a conocer información del usuario (por ejemplo, por lo menos la dirección de e-mail) a terceras personas. Esto se puede repetir con otros comerciantes sospechosos y generar una lista para prevenir a los usuarios que su información podría ser distribuida por spammers. Estas son sólo algunas técnicas de enviar un e-mail a un honeypot que se puede considerar en forma confiable como spam. En la práctica, hay otras formas alternativas de enviar un e-mail a un honeypot que se puede considerar en forma confiable un spam.

Porque los honeypots son una buena fuente de spam pero una terrible fuente de correo legítimo, los datos de los honeypots se pueden combinar con

los datos de un sistema de retroalimentación cíclica (figura 1) para entrenar a un nuevo filtro spam. Correo de distintas fuentes o distinta clasificación se puede evaluar en forma diferente. Por ejemplo, si hay 10 honeypots y 10 usuarios que se consultan en 10% de su correo, se espera 10 veces más spam del honeypot que de la consulta. Por lo tanto, el correo legítimo de la consulta se puede evaluar 10 o 11 veces más que el spam para compensar la diferencia. En forma alternativa, los datos de la honeypot se pueden evaluar en forma selectiva. Por ejemplo, el 50% del correo de un usuario es buen correo y el 50% es spam. El mismo volumen de spam va a los honeypots. Por lo tanto, parece que el honeypot tiene 100% de spam y todo se consulta, no sólo el 10%. Para entrenar con la tasa correcta de spam y buen correo en un sistema combinado, los datos del honeypot se evalúan en un 95% y el spam del usuario se evalúa en un 50% para resultar una tasa general de 1:1.

~~28~~
~~291~~
291

Otras fuentes de reportes de spam incluye usuarios que nos son participantes en el sistema de retroalimentación cíclica. Por ejemplo, puede haber un botón de "Reporte de Spam" disponible a todos los usuarios para todo el correo, para reportar spam que ha logrado pasar el filtro. Estos datos se pueden combinar con datos del sistema de retroalimentación cíclica. De nuevo, esta fuente de spam se debe evaluar en forma diferente puesto que puede no ser confiable en varios aspectos. Un nuevo chequeo también se debe hacer para reflejar el hecho que únicamente correo que no se filtro está sujeto a ser reportado por el botón de "Reportar-cómo-spam".

Adicional al filtro spam, un filtro de cuarentena se puede crear y lo puede utilizar el sistema de retroalimentación cíclica. El filtro de cuarentena utiliza tanto rasgos de correo positivos como negativos. Por ejemplo, correo de un comerciante en línea popular casi siempre es bueno. Un spamer aprovecha en sistema mimetizando un aspecto del correo del buen comerciante en su spam. Otro ejemplo es que el spamer engaña al sistema de retroalimentación enviando pequeñas cantidades de correo bueno a través de una dirección IP. La retroalimentación cíclica aprende a clasificar este correo como buen correo, en ese momento el spamer comienza a enviar el spam de la misma dirección IP.

Así, el filtro de cuarentena se da cuenta que se está recibiendo un rasgo positivo particular en cantidades más grandes de lo que el sistema estaba acostumbrado basado en los datos históricos. Esto hace que el sistema sospeche del mensaje y por lo tanto, lo ponga en cuarentena hasta que suficientes resultados de consulta se obtengan antes de decidir entregarlo o marcarlo como spam. El filtro de cuarentena también se puede emplear cuando el correo se recibe de una dirección IP nueva, de la cual no se sabe con certeza si los correos son spam o no spam y esto no se sabrá por algún tiempo. La cuarentena se puede hacer de distintas maneras, incluyendo la marcación provisional del correo como

7/5
~~292~~
292

spam y moviéndolo a un folder de spam o no entregándolo al usuario y guardándolo donde no se vea. La cuarentena se le puede hacer a mensajes que están cerca al umbral de spam del filtro: se puede asumir que información adicional de las consultas ayudara tomar la decisión correcta. La cuarentena también se puede hacer cuando muchos mensajes similares se reciben: unos cuantos mensajes se envían para consulta con la retroalimentación cíclica, y el filtro que se ha entrenado de nuevo se puede utilizar para clasificar el mensaje en forma correcta.

Adicionalmente a la construcción de filtros, el sistema de retroalimentación cíclica como se describe a acá se puede utilizar también para evaluarlos. Lo que esto significa es que los parámetros del filtro spam se pueden adaptar como se necesiten. Por ejemplo, un filtro se puede entrenar hasta la medianoche de la última noche. Después de la medianoche, toma los datos que entran a la base de datos para determinar la proporción de errores en el filtro spam comparado con la clasificación del usuario. Aún más, la retroalimentación cíclica se puede utilizar para determinar positivos falsos y proporción de capturas del filtro spam. Por ejemplo, los votos del usuario sobre el correo se toman en cuenta y se pueden correr a través de un filtro potencial para determinar el positivo falso y la proporción de capturas. Esta información luego se puede utilizar para sintonizar y optimizar el filtro. Distintas configuraciones para los parámetros o distintos algoritmos se pueden ensayar manualmente o automáticamente construyendo distintos filtros, cada uno utilizando una configuración o algoritmos diferentes, para obtener el positivo falso menor y proporción de capturas. Así, el resultado se puede comparar para seleccionar los mejores u óptimos parámetros de filtro.

La retroalimentación cíclica se puede utilizar para construir y poblar listas de direcciones IP o dominios o URLs que siempre se votan como spam o siempre se votan como buenos, o se votan por lo menos en un 90% buenos, etc. Estas listas

778
~~293~~
293

se pueden utilizar para filtros de spam en otras maneras. Por ejemplo, una lista de direcciones IP que se vota como 90 por ciento spam se puede utilizar para construir una lista de hoyo negro de direcciones de las cuales no se acepta ningún correo. La retroalimentación cíclica también se puede utilizar para terminar las cuentas de los spammers. Por ejemplo, si un usuario de una ISP aparece como enviando spam, se le puede notificar automáticamente a la ISP. En forma similar, si un dominio particular aparece como responsable de una gran cantidad de spam, se le puede notificar al proveedor del dominio del e-mail automáticamente.

Existen un número de arquitecturas que se pueden implementar con el sistema de retroalimentación cíclica. Una arquitectura ejemplar está basada en un servidor, como se describe en la figura 7, con el proceso de selección sucediendo cuando el correo llega al servidor. Una arquitectura externa está basada en el cliente, como se describe en la figura 6. En una retroalimentación cíclica basada en el cliente, la información de consulta se puede utilizar para mejorar la ejecución de un filtro personalizado, o, en la implementación ejemplar que se ilustra, la información se puede enviar a un depósito compartido como datos de entrenamiento para un filtro que se comparte (por ejemplo, a lo ancho de la corporación o global.) Se debe apreciar que las siguientes arquitecturas descritas a continuación simplemente son ejemplares y pueden incluir componentes adicionales y rasgos que no se muestran acá.

En referencia ahora a la figura 6, se muestran un diagrama de bloque general y ejemplar de una técnica de retroalimentación cíclica en una arquitectura basada en el cliente. Una red 600 se suministra para facilitar la comunicación de e-mail de y para uno o más clientes 602, 604, y 606 (llamados también cliente uno, cliente dos... Cliente N donde N es un entero más grande o igual a 1). La red puede ser una red de comunicación global (GCN) como Internet, o un WAN (Red de Área Amplia), LAN (Red de Área Local), o cualquier otra configuración de red.

7/2
294
294

En esta implementación particular, un servidor de entrada SMTP 608 (Protocolo de Transferencia de Correo Simple) sirve de interfase a la red 600 para suministrar servicios SMTP al LAN 610. Un servidor de correo 612 dispuesto en forma operativa en el LAN 610 sirve de interfase al servidor de entrada 608 para controlar y procesar los e-mail entrantes y salientes de los clientes 600 todos, 604 y 606. Estos clientes 602, 604 y 606 están dispuestos en el LAN 610 para acceder por lo menos a los servicios de correo que se suministran allí.

El cliente 602 incluye una unidad de procesamiento central (CPU) 614 que controla los procesos del cliente. La CPU 614 puede comprender múltiples procesadores. La CPU 614 ejecuta instrucciones en relación con el suministro de una o más funciones de recolección/retroalimentación de datos descritos anteriormente. Las instrucciones incluyen, pero no están limitadas a, las instrucciones codificadas que ejecutan por lo menos la metodología de retroalimentación cíclica descrita anteriormente, por lo menos cualquiera o todos de los enfoques que se pueden utilizar en combinación para dirigirse a la selección de clientes y mensajes, modificaciones de los mensajes de consulta, retención de datos, confiabilidad de los clientes y clasificación de la validación, reevaluación de los datos de múltiples fuentes que incluyen el sistema de retroalimentación cíclica, la optimización de filtros spam y sincronización, filtros de cuarentena, creación de listas spam, y notificación automática de spammers a sus respectivos ISPs y proveedores de correo. Una interfase de usuario 616 se suministra para facilitar la comunicación con la CPU 614 y el sistema operativo el cliente de tal manera que el cliente puede interactuar para acceder al e-mail y votar sobre los mensajes de consulta.

Un muestreo de los mensajes de cliente recuperados del servidor 612 se puede seleccionar para consulta por un selector de mensajes 620. Los mensajes se seleccionan y se modifican para consulta si el receptor pretendido (clientes) ha

42
295
295

acordado previamente participar. Un modificador de mensajes 622 modifica el mensaje para convertirlo en un mensaje de consulta. Por ejemplo, el mensaje (s) se puede modificar para incluir las instrucciones de votación y botones de votación y/o vínculos de acuerdo con las modificaciones del mensaje suministradas anteriormente. Los botones de votación y/o vínculos se implementan modificando la interfase del usuario 616 en el software de e-mail del cliente. Adicionalmente, el modificador de mensajes 622 se puede eliminar cualquier virus en los mensajes (mensajes de consulta y no consulta) antes de que sean abiertos o descargados por el cliente 602.

En una implementación, el usuario del cliente luchador en contra del spam 602 ve cada mensaje únicamente una vez, con algunos mensajes marcados especialmente como mensajes de consulta, e incluyendo botones de votación, etc. En la presente implementación, el usuario del cliente luchador contra spam 602 puede ver algunos mensajes dos veces, donde uno es el mensaje normal y el otro es el mensaje de consulta. Esto se puede implementar de varias maneras. Por ejemplo, el mensaje de consulta se puede devolver al servidor 612 y almacenar en un almacén de mensajes de consulta. En forma alterna, el cliente 602 puede almacenar un mensaje adicional en el servidor de e-mail 602. De otra manera, el cliente 602 puede mostrarle al usuario cada mensaje dos veces, una como mensajes normal y otra de su forma modificada.

Los resultados de la consulta 626 se pueden enviar a la CPU 614 y luego a la base de datos 630 que se puede configurar para almacenar los datos de un cliente o de más de un cliente, dependiendo en los arreglos específicos de la arquitectura de retroalimentación del cliente. La base de datos central 630 almacena los mensajes de consulta, lo resultados de consulta como también la información respectiva al cliente. Se pueden utilizar componentes relacionados para analizar tal información de tal manera que se determina la frecuencia de

119
~~296~~
296

consulta, la confiabilidad del cliente-usuario (por ejemplo, la validación del usuario 632), y otras estadísticas del cliente. Las técnicas de validación se pueden emplear particularmente cuando hay sospechas en la confiabilidad de la votación del cliente. Las dudas pueden surgir del análisis de las contradicciones, el número de veces que se ha cambiado de opinión, y el número de mensajes consultados para usuario o usuarios particulares; en forma alterna, las técnicas de validación se pueden utilizar para cada usuario. Cualquier cantidad de datos adecuados almacenados en la base de datos central se pueden utilizar en las técnicas de aprendizaje de máquina 634 para facilitar el entrenamiento de un nuevo y/o mejorado filtro de spam.

Los clientes 604 y 606 incluyen componentes similares a los descritos anteriormente para obtener y entrenar un filtro que es personalizado para un cliente(s) particular(es). Adicionalmente a lo que se ha descrito, un limpiador de mensajes de consulta 628 puede servir de interfase entre la CPU 614 y la base de datos central 630 de tal manera que aspectos del mensaje consultado se pueden eliminar por distintas razones tales como, agregar datos, compresión de datos, etc. el limpiador de mensajes 628 puede ser eliminar porciones extrañas del mensaje consultado como también cualquier información asociada del usuario no deseada.

En referencia ahora a la figura 7, se muestra un sistema de retroalimentación cíclica 700 ejemplar basado en un servidor que facilita la conexión de múltiples usuarios y que obtiene datos de consulta de acuerdo con las técnicas de retroalimentación cíclica del presente invento. Una red 702 se suministra para facilitar la comunicación de e-mail de y para uno o más usuarios 704 (también llamados usuario₁ 704₁, usuario₂ 704₂... Usuario_N 704_N donde N es un entero más grande o igual a uno). La red 702 puede ser una red de comunicación global (GCN) tal como Internet, o WAN, LAN o cualquier otra

~~200~~
~~295~~
297

configuración de red. En esta implementación particular a un SMTP (Protocolo de Transferencia de Correo Simple) el servidor de entrada 710 hace interfase con la red 702 para suministrar servicios SMTP al LAN 712. Un servidor de e-mail 714 dispuesto en forma operativa en el LAN 712 hace interfase con la puerta de entrada 710 para controlar y procesar e-mails entrantes y salientes del usuario 704.

El sistema 700 suministra capacidad de conexión múltiple de tal manera que la selección de usuarios y mensajes 716, modificación de mensajes 718, y consulta de mensajes (720, 722, 724) tiene lugar para cada usuario que se conecta al sistema 700. De esta manera se suministra una interfase de usuario 726 que muestra una pantalla de conexión como parte del proceso de arranque del sistema operativo del computador, o como se requiera, para conectarse a un perfil de usuario asociado antes que el usuario 704 pueda acceder a sus mensajes entrantes. Por lo tanto, cuando un primer usuario 704₁ (usuario 1) decide acceder a los mensajes, el primer usuario 704₁ se conectará al sistema a través de la pantalla de conexión 728 entrando información de acceso típicamente en la forma de un nombre de usuario y una clave. Una CPU 730 procesa la información de acceso para permitir el acceso el usuario, a través de una aplicación de comunicación de mensajes (por ejemplo, un cliente de correo) para una primera casilla de entrada del usuario 732.

Cuando se recibe correo entrante en el servidor de mensajes 714, se seleccionan en forma aleatoria para consulta lo que significa que por lo menos uno de los mensajes se marca para consulta. El receptor pretendido del mensaje marcado se investiga para determinar si uno de los receptores también está designado como un luchador contra el spam. Las propiedades del receptor que indican tal información se pueden mantener en el servidor de mensaje 714 o en cualquier otro componente del sistema 700 como sea apropiado. Una vez que se

~~298~~
298

determina cuál de los receptores pretendidos es también un luchador contra el spam, una copia de su respectivo correo como también otra información concerniente a la transacción de correo se puede enviar a la base de datos central 734 para almacenar. Los mensajes marcados para consulta se modifican con el modificador de mensajes 718 en cualquiera de las formas descritas anteriormente. Los mensajes seleccionados para consulta también pueden especificar el usuario 704. Por ejemplo, el usuario 704 puede indicar que únicamente cierto tipo de mensajes están disponibles para consulta. Como esto puede resultar en una prueba de datos prejuzgada, estos datos se pueden reevaluar con respecto a otros datos de cliente para mitigar la construcción desproporcionada de conjuntos de datos de entrenamiento.

La búsqueda de virus en los mensajes de consulta también se puede ejecutar en este momento o en cualquier otro momento antes de que mensaje de consulta se descargue y/o lo abra el usuario 704. Una vez que el mensaje se ha modificado de la manera apropiada, se entregan a la casilla de entrada del usuario respectivo llamadas $INBOX_1$ 732, $INBOX_2$ 736, y $INBOX_N$ 738, donde se pueden abrir para consulta. Para facilitar el proceso de consulta, cada mensaje de consulta incluye dos o más botones de votación o vínculos, que el usuario puede seleccionar y generar información relacionada con el mensaje de consulta y el resultado de la consulta. El texto de cada mensaje de consulta se puede modificar para incorporar los botones de votación o los vínculos.

Los resultados de la consulta de mensaje (llamados mensaje de consulta₁ 720, mensaje de consulta₂ 722, y mensaje de consulta_N 724) que incluyen cualquier información que resulta de la clasificación (por ejemplo, consulta de mensajes o id asociado, propiedades del usuario), se envían a la base de datos central 734 a través de la interfase de red 740 en el LAN 712. La base de datos central 734 puede almacenar la información de consulta y del usuario (720, 722,

~~125~~
~~299~~
299

724) respectivo para aplicar a las técnicas de aprendizaje de máquina y construir y optimizar un nuevo y/o filtro de spam mejorado 742. Sin embargo, por razones de privacidad y/o seguridad, la información confidencial se puede eliminar de la información antes de enviarse a la base de datos central 714. La información generada por el usuario 704 a través de la consulta o también se puede agregar a los datos estadísticos. De esta manera se requiere un menor ancho de banda para transmitir la información.

El nuevo filtro de spam entrenado 742 se puede entonces distribuir a otros servicios (que no se muestran) como también al software de e-mail del cliente (que no se muestran) que hace interfase con el LAN 712 de forma continua, de tal manera que cuando un filtro nuevo está disponible, ya sea por solicitud específica o en forma automática. Por ejemplo, el filtro de spam más reciente puede sacarse y/o hacerse asequible para descargar a través de un sitio de la red. En la medida que nuevos conjuntos de datos de entrenamiento se entregan para construir nuevos filtros spam, los conjuntos de datos viejos (información que sea obtenido previamente y/o empleado para entrenar el filtro) se pueden eliminar o descartar dependiendo de la antigüedad de los datos.

Considere ahora el escenario alternativo donde una organización dedicada a la lucha contra el spam hace disponible un filtro compartido por muchas organizaciones que utilizan filtros. En un aspecto del invento, el proveedor del filtro también es un gran proveedor de servicios de e-mail (por ejemplo cuentas de correo pagas y/o gratuitas). En vez de depender exclusivamente en el e-mail de su propia organización, el filtro suministrado decide también utilizar algunos datos de las organizaciones que utilizan filtros, de tal manera que mejore la captura de spam y buen correo. El sistema de retroalimentación cíclica como se describió anteriormente también se puede utilizar en escenarios a través de las organizaciones, ya sea una arquitectura basada en el cliente o en el servidor.

~~725~~
~~300~~
300

Llamaremos al proveedor del filtro, que agrega datos de su propio uso y de las distintas organizaciones que utilizan filtros la organización "interna" y llamaremos a los componentes que residen en una de las organizaciones que utilizan filtros, "externa". En general, el sistema a través de las organizaciones incluye un servidor de base de datos de correo en el proveedor del filtro (interno), de tal manera, pero no limitado a, Hotmail y uno o más servidores de mensajes (externo) tales como aquellos que pueden residir dentro de una o más compañías individuales. En éste caso, el servidor interno de base de datos de correo también almacena una sustancial retroalimentación de e-mail de sus propios clientes. De acuerdo con este aspecto del presente invento, el entrenamiento de datos se puede generar basándose en información almacenada en la base de datos interna (por ejemplo, e-mail/mensajería gratuita en Hotmail o un servidor MSN) como también información almacenada en una o más bases de datos externa asociada con el respectivo servidor externo. La información que se mantiene en la base de datos externa se puede comunicar al servidor interno a través de una red como Internet, por ejemplo, para ser usada en técnicas de aprendizaje de máquina. Finalmente datos de la base de datos externa se pueden utilizar para entrenar un nuevo filtro de spam y/o mejorar los filtros de spam existentes localizados externamente (por ejemplo, dentro de la respectiva compañía) o asociados con el servidor de correo interno.

Los datos de una o más de las bases de datos externas deben incluir por lo menos uno de los mensajes de consulta, resultados de la consulta (clasificaciones), información/propiedades del usuario, y datos estadísticos de votación por usuario, por un grupo de usuarios promedio para cada compañía. Los datos estadísticos de votación facilitan la determinación de la confiabilidad de la información generada por la compañía respectiva como también reducir los errores en los datos externos. De esta manera, los datos de una o más bases de datos externas (compañías) pueden ser reevaluados o evaluados en forma

724
301
301

diferente por una o más de las bases de datos externa. Aún más, las entidades externas se pueden poner a prueba para determinar su confiabilidad utilizando una técnica de validación similar a la descrita anteriormente.

Para la seguridad de la compañía, privacidad y confidencialidad, la información o datos que se comunicaron a través de Internet de cada compañía al servidor de e-mail, por ejemplo, se pueden limpiar, abreviar, y/o condensar de su forma original. La forma original se puede mantener en la base de datos externa respectiva y/o de otra forma tratarse de acuerdo con las preferencias de cada compañía. De esta manera, el servidor de e-mail o cualquier otro servidor interno de correo recibe únicamente la información pertinente necesaria para generar los datos de entrenamiento tales como clasificación de spam, dominio del remitente, nombre del remitente, contenido del mensaje clasificado como spam, y por el estilo.

En referencia ahora a la figura 8, se muestra un sistema de retroalimentación a través de distintas organizaciones ejemplar 800 donde un servidor de base de datos interna y un servidor de correo externo se pueden comunicar e intercambiar información de la base de datos a través de la red a para facilitar la generación de datos de entrenamiento utilizados en las técnicas de aprendizaje de máquinas para construir filtros de spam mejorados. El sistema 800 incluye por lo menos un servidor externo de mensajes 802 (por ejemplo, asociado con por lo menos una compañía) y un servidor de base de datos interno 804. Dada la naturaleza de los sistemas a través de distintas organizaciones, el servidor externo 802 y el servidor interno de e-mail 804 respectivamente mantienen su propia base de datos. Lo que esto significa este servidor de e-mail 804 está asociado con una base de datos interna 806 que también se pueden utilizar para entrenar al nuevo filtro de spam 808. De la misma manera, el servidor externo 802 está asociado con una base de datos externa 810 que se puede utilizar para

~~125~~
~~302~~
302

entrenar por lo menos a un nuevo filtro de spam 812 como también al filtro de spam 808 localizado internamente con respecto al servidor de e-mail 804. Por lo tanto, la información almacenada en la base de datos externa 810 se puede utilizar para entrenar al filtro de spam 808 localizado en el servidor de e-mail.

Se suministra un GCN 814 para facilitar la información de comunicación para ir de servidor interno de e-mail 804 y uno o más de los servidores de mensaje externos 800 todos. El componente de servidor externo de un sistema a través de las organizaciones funciona en forma similar a como lo hace un sistema de retroalimentación cíclica basada en una servidor (por ejemplo la figura 7, *supra*). Por ejemplo, el servidor de mensajes 802, la base de datos externa 810 y el filtro 812 pueden estar localizados en un LAN 815. Adicionalmente, se suministra una interfase de usuario 816 que muestra una pantalla de conexión 818 como parte del proceso de arranque del sistema operativo del computador, o como se requiera, para enganchar el perfil de un usuario asociado antes de que el usuario pueda acceder a sus mensajes entrantes.

En este sistema basado en un servidor, uno o más usuarios (llamados usuario₁ 820, usuario₂ 822, usuario_N 824) pueden conectarse al sistema al mismo tiempo para hacer uso de los servicios de correo disponibles. En la práctica, cuando un primer usuario 820 (usuario₁) decide acceder a los mensajes, el primer usuario 820 se conecta al sistema a través de la pantalla de conexión 818 entrando la información de acceso típicamente en la forma de un nombre de usuario y una clave. Una CPU 826 procesa la información de acceso para permitir el acceso del usuario a una primera localidad de entrada 828 a través de una aplicación de comunicación de mensajes (por ejemplo, un cliente de e-mail).

Cuando un correo entrante se recibe en el servidor de mensajes 802, los mensajes se escogen en forma aleatoria o específica para consulta. Antes de que los mensajes se puedan seleccionar para consulta, el receptor pretendido de tales

~~42~~
~~38~~
303

mensajes se compara con una lista de usuarios luchadores contra el spam para determinar si uno de los receptores también está designado como un usuario luchador contra el spam. Las propiedades de receptor indicando esta información se pueden mantener en el servidor de mensajes 802, en la base de datos 810, o en cualquier otro componente del sistema 800 como sea apropiado. Una vez se determina cuál de los receptores pretendidos también es un luchador contra el spam, los mensajes se seleccionan para consulta y una copia de los mensajes de consulta también con el resto de información perteneciente a la transacción de correo se envían a la base de datos 810.

Los mensajes seleccionados para consulta se modifican con el modificador de mensajes 830 en una serie de maneras descritas anteriormente. En la práctica, un identificador único (ID) se le puede asignar a cada mensaje de consulta, a cada luchador contra el spam, y/o a cada resultado de la consulta y almacenado en la base de datos 810. Como se mencionó previamente, los mensajes seleccionados para consulta se pueden escoger en forma aleatoria o específica a un usuario respectivo (820, 822, y 824). Por ejemplo, el usuario uno 820 puede indicar que únicamente cierto tipo de mensajes están disponibles para consulta (por ejemplo, mensajes enviados de una compañía externa). Los datos generados de estos mensajes específicos se revalúan y/o descartan para mitigar obtener errores en las muestras de los datos.

La búsqueda de virus en los mensajes de consulta se puede ejecutar en este momento o en cualquier otro momento antes de que el mensaje de consulta lo descargue y/o abra el usuario. Una vez los mensajes se han modificado de la manera apropiada. Se entregan al usuario respectivo que se denota como $INBOX_1$ 828, $INBOX_2$ 832, y $INBOX_N$ 834, donde se pueden abrir para consulta. Para facilitar el proceso de consulta, cada mensaje de consulta incluye dos o más botones de votación o vínculos, que cuando el usuario los selecciona, generan

~~124~~
~~304~~
304

información relacionada con el mensaje de consulta y el resultado de la consulta. El texto de cada mensaje consultado se puede modificar para incorporar los botones de votación o los vínculos.

Los resultados de la consulta de mensajes (llamados CONSULTA DE MENSAJES₁ 836, CONSULTA DE MENSAJES₂ 838, y CONSULTA DE MENSAJES_N 840), que incluyen cualquier información que resulta de la clasificación (por ejemplo, la consulta de mensajes o el ID asociado, las propiedades del usuario), se envían a la base de datos 810 a través de la interfase de red 842 localizado a en el LAN 815. La base de datos 810 almacena la información de consulta y del usuario de los respectivos usuarios para uso más tarde en las técnicas de aprendizaje de máquina que se emplean para construir y/o optimizar y/o mejorar un nuevo filtro de spam 812, 808.

Por razones de privacidad, cada compañía puede desear sacar información clave antes de enviar el mensaje consulta y/o información del usuario a su propia base de datos 810 y/o a la base de datos de e-mail 806 a través del GCN 814, por ejemplo. Un enfoque es suministrar únicamente la retroalimentación a la base de datos (806 y/o 810) con mensajes spam, de esta manera excluyendo la retroalimentación con el correo legítimo. Otro enfoque es suministrar únicamente un subconjunto de la información en el correo legítimo tal como la dirección IP del remitente. Otro enfoque para mensajes seleccionados, como aquellos marcados como buenos por el usuario que serían marcados como malos por un filtro, o viceversa, es solicitar explícitamente al usuario permiso antes de enviarlos al filtro. Cualquiera de estos enfoques o combinaciones por lo tanto facilitan mantener la privacidad y confidencialidad de la información para los clientes participantes mientras se suministran datos para entrenar el filtro de spam (808 y/o 812).

Los esquemas de validación del usuario tales como los descritos anteriormente también se pueden aplicar a cada compañía como también a cada

~~128~~
~~305~~
305

usuario dentro de la compañía. Por ejemplo, los usuarios pueden estar sujetos individualmente a técnicas de validación cruzada donde la clasificación de un usuario sospechoso se excluye del entrenamiento del filtro. El filtro se entrena utilizando datos de los usuarios restantes. El filtro entrenado luego correr los mensajes del usuario excluido para determinar cómo hubiera clasificados los mensajes. Si el número de desacuerdos excede un nivel de umbral, el usuario sospechoso se considera no confiable. La clasificación futura de mensajes del usuario no confiable se puede inspeccionar manualmente antes de aceptarse dentro de la base de datos y/o filtro. De otra manera, el usuario se puede eliminar de consulta futuras.

En referencia ahora a la figura 9, un ambiente ejemplar 910 para implementar varios aspectos del invento incluye a un computador 912. El computador 912 incluye una unidad procesadora 914, una memoria del sistema 916, y un bus de sistema 918. El bus del sistema 918 se acopla a los componentes incluyendo, pero no limitado a, la memoria del sistema 916 a la unidad procesadora 914. La unidad procesadora 914 puede ser cualquiera de varios procesadores disponibles. Arquitecturas de microprocesadores duales y otros multiprocesadores también se pueden utilizar como la unidad de procesamiento 914.

El bus del sistema 918 puede ser cualquiera de distintos tipos de estructuras de bus incluyendo el bus de la memoria o controlador de la memoria, un bus periférico o bus externo, y/o un bus local utilizando una variedad de arquitecturas disponibles incluyendo, pero no limitado a, el bus de 11-bits, el Estándar Industrial de Arquitectura (ISA), la Arquitectura de Micro-canal (MSA), la ISA Extendida (EISA), el Controlador Electrónico Inteligente (IDE), el Bus Local VESA (VLB), el Componente de Interconexión Periférica (PCI), el Bus Serial Universal(USB), el Puerto Grafico Avanzado(AGP), la tarjeta Internacional de Memoria de

~~127~~
~~386~~
306

Computador Personal(PCMCIA), y la Interfase de Pequeños Computadores(SCSI).

La memoria del sistema 916 incluye a la memoria volátil 920 o la memoria no volátil 922. El sistema básico de entrada/salida (BIOS), contiene las rutinas básicas para la transferencia de información entre los elementos dentro del computador 912, tales como durante el inicio, se almacenan en la memoria no volátil 922. Como ilustración, y no como limitante, la memoria no volátil 922 puede incluir a la memoria de sólo lectura (ROM), la programable ROM (ROM), la programable eléctricamente ROM (EPROM), la que se puede borrar eléctricamente ROM (EEPROM), o la memoria flash. La memoria volátil 920 incluye a la memoria de acceso aleatorio (RAM), que actúa como memoria CACHE externa. Como ilustración y no como limitante, la RAM está disponible en muchas formas tales como la RAM sincronizada (SRAM), RAM dinámica (DRAM), la DRAM sincronizada (SDRAM), la SDRAM de doble velocidad (DDR SDRAM), la SDRAM (ESDRAM), Synchlink DRAM (SLDRAM), y la Rambus directa RAM (DRRAM).

El computador 912 también incluye medios de almacenaje en computador removibles/no removibles, volátiles/no volátiles. La figura 9 muestra, por ejemplo el almacenaje en disco 924. El almacenaje en disco 924 incluye, pero no esta limitado a, dispositivos como discos magnéticos, discos floppy, cintas, drives Jaz, drives Zip, drives LS-100, tarjetas de memoria flash, o barras de memoria. Adicionalmente, el almacenaje en disco 924 puede incluir medios de almacenaje separados o en combinación con otros medios de almacenaje incluyendo, pero no limitados a, discos ópticos tales como discos compactos ROM (CD-ROM), discos regrabables de CD (CD-R), drives de discos regrabables CD (CD-RW) o discos versátiles digitales (DVD-ROM). Para facilitar la conexión entre los dispositivos de

307
307

almacenaje en el disco 924 al bus del sistema 918, se usa típicamente una interfase removibles o no removibles tal como la interfase 926.

Se debe apreciar que la figura 9 describe software que actúa como un intermediario entre el usuario y los recursos básicos del computador descritos en un ambiente de operación adecuado 910. Este software incluye un sistema operativo 928. El sistema operativo 928, que se puede almacenar en disco 924, actúa para controlar y disponer de los recursos del sistema de cómputo 912. Las aplicaciones del sistema 930 aprovecha el manejo de los recursos por el sistema operativo 928 a través de módulos de programa 932 y datos de programa 934 almacenados ya sea en la memoria del sistema 916 o en disco 924. Se debe apreciar que el presente invento se puede implementar en varios sistemas operativos o en combinaciones de sistemas operativos.

Un usuario entra comandos o información en el computador 912 a través de dispositivos de entrada 936. Los dispositivos de entrada 936 incluyen, pero no están limitados a, dispositivos de señalamiento tales como ratón, bolígrafos, trackballs, almohadillas táctiles, teclados, micrófonos, joysticks, control de juegos, discos satelitales, escáneres, tarjetas de sintonización de televisión, cámaras digitales, cámaras de vídeo, cámaras web, y demás. Éstos y otros dispositivos de entrada se conectan a la unidad de procesamiento 914 a través del bus del sistema 918 a través de puertos de interfase 938. Los puertos de interfase 938 incluyen, por ejemplo, puertos seriales, puertos paralelos, puertos de juegos, el bus serial universal (USB). Los dispositivos de salida 940 utilizan algún tipo de puerto como los de entrada 936. Así, por ejemplo, un puerto USB se puede utilizar para entradas al computador 912 y salidas de información del computador 912 a un dispositivo de salida 940. El adaptador de salida 942 se suministra para mostrar que hay algunos dispositivos de salida 940 como monitores, parlantes, impresoras entre otros dispositivos de salida 940 que requieren un adaptador

734
208
300

especial. Los adaptadores de salida 942 incluyen, como ilustración y no limitación, tarjetas de vídeo y de sonido que suministran una forma de conexión entre un dispositivo de salida 940 y el bus del sistema 918. Se debe anotar que otros dispositivos y/o sistemas de dispositivos suministran tanto capacidades de entrada como de salida tales como el computador remoto 944.

El computador 912 puede operar en un ambiente de red utilizando conexiones lógicas a uno o más computadores remotos, tales como el computador remoto 944. El computador remoto 944 puede ser un computador personal, un servidor, un controlador, un PC en red, una estación de trabajo, un dispositivo basado en un microprocesador, el dispositivo de un colega o cualquier otro modo en red y demás, y típicamente incluye muchos o todos de los elementos descritos relativos al computador 912. Con el propósito de brevedad, sólo se demuestra un dispositivo de almacenaje 946 con el computador remoto 944. El computador remoto 944 está conectado lógicamente al computador 912 a través de una interfase de red 948 y luego físicamente conectado a través de la conexión de comunicación 950. La interfase de red 948 comprende las redes de comunicación tales como una red de área local (LAN) y una red de área amplia a (WAN). La tecnología LAN incluye la interfase de datos de fibra distribuidos (FDDI), la interfase de cobre de datos distribuidos (CDDI), Ethernet/IEEE 1102.3, Token Ring/IEEE 1102.5 y demás. La tecnología WAN incluye, pero no esta limitada a, vínculos, redes de cambio de circuitos como las Redes Integradas de Servicios Digitales (ISDN) y demás variaciones, redes de conmutación de paquetes, y líneas de suscripción digital (DSL).

La conexión de comunicación 950 se refiere al hardware/software utilizado para conectar la interfase de red 948 al bus 918. Mientras la conexión de comunicación 950 se muestra por razones ilustrativas dentro del computador 912, puede también ser externa al computador 912. El hardware/software necesario

132
309
309

para conectar la interfase de red 948 incluye, por razones ejemplares únicamente, tecnologías internas y externas tales como, módems incluyendo módems de teléfono regular, módems de cable y módems DSL, adaptadores ISDN, y tarjetas Ethernet.

La figura 10 es un diagrama de bloque esquemático de un ambiente de computación 1000 ejemplar con el cual el presente invento puede interactuar. El sistema 1000 incluye uno más clientes 1010. Los clientes 1010 pueden ser hardware y/o software (por ejemplo, hilos, procesos, dispositivos de cómputo). El servidor 1030 puede albergar hilos, por ejemplo para ejecutar transformaciones empleando el presente invento. Una comunicación posible entre el cliente 1010 y el servidor 1030 puede ser en la forma de paquetes de datos adaptados para ser transmitidos entre dos o más procesos de computador. El sistema 1000 incluye el marco de trabajo de comunicación 1050 que se puede utilizar para facilitar la comunicación entre los clientes 1010 y los servidores 1030. Los clientes 1010 se conectan en forma operativa a uno o más almacenes de datos de clientes 1060 que se pueden utilizar para almacenar información local para el cliente 1010. En forma similar, el servidor 1030 esta conectado en forma operativa a uno o más almacenes de datos de servidor 1040 que se pueden utilizar para almacenar información local al servidor 1030.

Lo que se ha descrito anteriormente incluye ejemplos del presente invento. No es posible, claro está, describir cada combinación concebible de componentes o metodologías con el propósito de describir el presente invento, pero uno de destreza común en el arte puede reconocer que muchas más combinaciones y permutaciones del presente invento son posibles. De esta forma, el presente invento pretende abarcar tales alteraciones, modificaciones, y variaciones que caen dentro del espíritu y alcance de las reivindicaciones anexas. Aún más, en la medida que el término "incluye" se utiliza tanto en la descripción detallada o en las

733
~~310~~
310

reivindicaciones. tal término pretende ser inclusivo de una manera similar al término "comprendiendo" como "comprendiendo" se interpreta cuando se utiliza como una palabra transicional en una reivindicación.

134
311
311

REIVINDICACIONES

Lo que se reivindica es:

1. Un sistema que facilita la clasificación de items en conexión con la prevención de spam, y que comprende:

un componente que recibe un conjunto de los items;

un componente que identifica receptores pretendidos del items, y marca un subconjunto de los items para ser consultados, el subconjunto de los items correspondientes a un subconjunto de receptores que son conocidos como usuarios luchadores contra el spam; y

un componente de retroalimentación que recibe información relacionada con la clasificación del luchador contra el spam del items consultado, y utilizan la información en conexión con el entrenamiento de un filtro para spam y poblando una lista de spam.

2. El sistema de la reivindicación 1, donde los items comprenden por lo menos uno de : correo electrónico (e-mail) y mensajes.

3. El sistema de la reivindicación 1, donde el componente que recibe el conjunto de items es cualquier servidor de e-mail, servidor de mensajes, y software de e-mail de un cliente.

4. El sistema de la reivindicación 1, donde el subconjunto de items que se va a consultar comprende todos los items recibidos.

5. El sistema de la reivindicación 1, donde el subconjunto de receptores comprende todos los receptores.

6. El sistema de la reivindicación 1, donde el subconjunto de receptores se selecciona en forma aleatoria.

~~13~~
~~312~~
312

7. El sistema de la reivindicación 1, donde el subconjunto de items marcados para consulta está limitado a por lo menos uno de los siguientes:

un número de los items seleccionados por el usuario;

un número de los items seleccionados por el usuario por periodo de tiempo;

y

una probabilidad de marcar un item correspondiente a un usuario conocido.

8. El sistema de la reivindicación 1, donde los items marcados son asignados un identificador único ID, el ID único correspondiente a cualquiera de los items marcados y al contenido del item marcado.

9. El sistema de la reivindicación 1, comprende además un componente que modifica el item marcado para consulta para identificarlo como un item de consulta.

10. El sistema de la reivindicación 9, donde el item de consulta comprende un resumen del item marcado, el resumen comprende por lo menos un asunto, una fecha, un texto de mensaje, y unas primeras líneas de texto.

11. El sistema de la reivindicación 9, donde el item de consulta comprende las instrucciones de votación y uno cualquiera de por lo menos dos botones de votación y vínculos que corresponden a por lo menos dos clases respectivas de ítem que facilita la clasificación del ítem por usuario.

12. El sistema de la reivindicación 1, comprende además una base de datos que almacena información y datos relacionados con las propiedades del usuario, contenido del ítem, propiedades asociadas con los items marcados, clasificación del usuario y datos estadísticos de votación, datos de análisis de frecuencia de consulta por usuario y de consulta por usuario por periodo de tiempo, listas de spam, listas de correo ilegítimo, y listas de hoyo negro.

13. El sistema de la reivindicación 1 distribuido a través de más de una compañía luchadora contra el spam de tal manera que la retroalimentación de

cada compañía se envía a una base de datos central en interfase operativa con cada compañía, donde algunas porciones de la retroalimentación se remueven por razones de privacidad.

14. El sistema de la reivindicación 1, comprende además un componente de validación de la clasificación del usuario que prueba la confiabilidad y dependabilidad del usuario.

15. El sistema de la reivindicación 14, donde el componente de validación de la clasificación del usuario se puede aplicar a uno o más usuarios sospechosos.

16. El sistema de la reivindicación 1, donde el componente de retroalimentación recibe información relacionada con la retroalimentación de un usuario, la retroalimentación de un honeypot, y ocasionalmente, la retroalimentación del usuario receptor de los items recibidos.

17. Un método que facilita la clasificación de los mensajes en conexión con la prevención del spam que comprende:

- recibir un conjunto de mensajes;

- identificar los receptores pretendidos del mensaje;

- marcar un subconjunto de los mensajes que se van a consultar, un subconjunto de los mensajes correspondiente a un subconjunto de los receptores que son conocidos como usuarios que luchan contra el spam;

- recibir información relacionada con la clasificación del usuario de mensajes de consulta; y

- emplear la información en conexión con el entrenamiento de un filtro de spam y poblar una lista de spam.

18. El método de la reivindicación 17, donde el subconjunto de receptores que son conocidos como luchadores contra el spam se determina por cada receptor ejecutando por lo menos uno de los siguientes:

adhiriéndose para suministrar retroalimentación en los mensajes para facilitar el entrenamiento de un nuevo filtro spam;

adhiriéndose en forma pasiva para suministrar retroalimentación en los mensajes al no retirarse;

pagar por servicios de inicio y mensajería suministrados por un servidor de mensajes participante; y

abrir una cuenta de e-mail en un servidor de mensajes participante.

19. El método de la reivindicación 17, donde el subconjunto de mensajes marcados para consulta está limitado por uno o más límites de consulta.

20. El método de la reivindicación 17, comprende además la modificación de los mensajes marcados para señalarlos e identificarlos como mensajes de consulta.

21. El método de la reivindicación 20, donde la modificación de los mensajes marcados comprende la ejecución de por lo menos uno de los siguientes:

mover el mensaje marcado a un folder separado para mensajes de consulta;

modificar la dirección "de" del mensaje marcado;

modificar la línea del asunto del mensaje marcado;

utilizar un icono de consulta en el mensaje marcado para identificar los como mensaje de consulta; y

utilizar un color único para identificar el mensaje marcado como mensaje de consulta.

22. El método de la reivindicación 17, comprende además la revisión de los mensajes marcados para virus antes de descargarlos para consulta.

23. El método de la reivindicación 17, comprende además hacer una copia de cada mensaje marcado como se recibió originalmente de tal manera que el

usuario respectivo recibe una primera copia del mensaje en su forma original de una segunda copia del mensaje para consulta.

24. El método de la reivindicación 17, comprende además distribuir el filtro de spam entrenado a uno o más servidores, distribución que ocurre en forma automática y/o por petición de por lo menos uno de los mensajes de e-mail y el anuncio en un sitio de la red para descarga.

25. El método de la reivindicación 17, donde el entrenamiento del filtro spam y la población de la lista spam se ejecuta con una técnica de aprendizaje de máquinas utilizando datos basados en la retroalimentación de clasificación del usuario y ocasionalmente, datos generados por una o más fuentes adicionales, fuentes que comprenden honeypots, retroalimentación de clasificación de receptores no-usuarios, y técnicas activas de aprendizaje.

26. El método de la reivindicación 25, donde los datos generados por uno o más fuentes se reevalúan proporcionalmente con respecto al tipo de datos generados por la fuente y relativos a los datos de clasificación del usuario para facilitar la obtención de unos datos de prueba influenciados.

27. El método de la reivindicación 17 comprende además:

el monitoreo de mensajes entrantes para sus respectivos rasgos positivos;

determinar la frecuencia de rasgos positivos recibidos;

determinar si uno o más rasgos positivos recibidos exceden un umbral de frecuencia basado por lo menos en parte sobre los datos históricos; y

poner los mensajes sospechosos en cuarentena, lo que corresponde a uno o más rasgos positivos que exceden el umbral de frecuencia, hasta que datos de clasificación adicionales estén disponibles para determinar si el mensaje sospechoso es spam.

28. El método de la reivindicación 27, donde los rasgos utilizados son información acerca del remitente que comprenden por lo menos una de las direcciones IP del remitente y su dominio.

29. El método de la reivindicación 27, donde la cuarentena de los mensajes sospechosos la ejecuta por lo menos uno de los siguientes actos:

marcar provisionalmente el mensaje sospechoso como spam y moverlo a un folder de spam;

demorar la entrega del mensaje sospechoso al usuario hasta que datos de clasificación estén disponibles; y

almacenar el mensaje sospechoso en un folder que no es visible al usuario.

30. El método de la reivindicación 17, comprende además la determinación de positivo falso y tasa de captura del filtro de spam para facilitar la optimización del filtro spam, donde determinar el positivo falso y la tasa de captura comprende:

entrenar al filtro de spam utilizando un conjunto de datos de entrenamiento, el conjunto de datos de entrenamiento comprende un primer conjunto de resultados de la consulta;

clasificar un segundo conjunto de mensajes de consulta utilizando la retroalimentación del usuario para generar un segundo conjunto de resultados de consulta;

correr el segundo conjunto de mensajes de consulta a través del filtro entrenado de spam;

comparar el segundo conjunto de resultados de consulta con lo resultados del filtro de spam entrenado para determinar el positivo falso y la tasa de captura del filtro evaluando de esta manera y sintonizando los parámetros del filtro de acuerdo a una ejecución optima del filtro.

740
317
317

31. El método de la reivindicación 30, donde más de un filtros spam se construye, cada uno teniendo distintos parámetros y cada uno entrenándose con el mismo conjunto de datos de entrenamiento, de tal manera que el positivo falso y la tasa de captura de cada filtro spam se compara con por lo menos otros filtros para determinar los parámetros óptimos para la filtración de spam.

32. El método de la reivindicación 17, comprende además la construcción de un filtro de spam mejorado utilizando cómputos adicionales de mensajes entrantes, subconjuntos de estos que se someten a consulta para obtener información nueva con relación al entrenamiento del filtro de spam mejorado, donde la información anterior obtenida se revalúa basada en parte en el tiempo desde que se obtuvo.

33. El método de la reivindicación 17, comprende además la utilización de información para construir una lista legítima de remitentes.

34. El método de la reivindicación 17, comprende además la utilización de información para facilitar la determinación de las cuentas de los spammers.

35. El método de la reivindicación 34, comprende además la identificación de un spammer que utiliza una ISP y automáticamente notifica a la ISP del envío de spam.

36. El método de la reivindicación 34, comprende además la identificación del dominio responsable por el envío del spam, y automáticamente notifica a por lo menos uno del dominio del proveedor de e-mail y del dominio ISP de donde se envió el spam.

37. El método de la reivindicación 17, comprende además la distribución de por lo menos un filtro spam y una lista de spam a cualquier servidor de correo, servidor de e-mail, o software de e-mail del cliente, donde la distribución comprende por lo menos uno de los siguientes:

38
318

fijar una notificación en la red notificando que el filtros spam y la lista de spam están disponibles para descarga;

automáticamente enviando el filtro spam y la lista spam al servidor de correo, servidor de e-mail y software de e-mail del cliente; y

manualmente enviando una lista spam al servidor de correo, servidor de e-mail, y software de e-mail del cliente.

38. Un método de validación cruzada que facilita la verificación de la confiabilidad de la clasificación de un usuario y que comprende:

excluir una o más de las clasificaciones de datos de un usuario sospechoso en el entrenamiento del filtro spam;

entrenar el filtros spam utilizando todas las demás clasificaciones disponibles de usuarios; y

corriendo los mensajes de consulta del usuario sospechoso a través del filtro entrenado de spam para determinar cómo hubiera clasificado el mensaje comparado a la clasificación del usuario sospechoso.

39. El método de la reivindicación 38, comprende además la ejecución de por menos uno de los siguientes:

descontando clasificaciones existentes y futuras suministradas por usuarios que se determinan no confiables hasta que el usuario se determine confiable;

descartar las indicaciones existentes suministradas por usuarios determinados como no confiables; y

removiendo los usuarios no confiables de consultas futuras.

40. Un método que facilita la verificación de la confiabilidad de las clasificaciones de un usuario para entrenar el filtros spam a través de un sistema de retroalimentación cíclica que comprende:

identificar un subconjunto de usuarios luchadores contra el spam como usuarios sospechosos;

suministrar uno o más mensajes con resultados conocidos al usuario sospechoso para consulta; y

determinar si la clasificación del usuario sospechoso de uno o más mensajes de prueba concuerda con la clasificación conocida para verificar la confiabilidad de las clasificaciones del usuario.

41. El método de la reivindicación 40, donde el subconjunto de usuarios luchadores contra el spam identificados como sospechosos comprende a todos los usuarios.

42. El método de la reivindicación 40, donde el mensaje de prueba es un mensaje conocido como por lo menos uno spam y uno correo bueno y que se inserta en el flujo de correo entrante por el sistema de retroalimentación cíclica y se entrega al usuario sospechoso.

43. El método de la reivindicación 40, donde el mensaje que recibe el usuario sospechoso para consulta se clasifica a mano por un administrador del sistema para entrenar el filtros de spam con una clasificación correcta para identificar a los usuarios no confiables.

44. El método de la reivindicación 40, comprende además por lo menos uno de los siguientes actos:

descontar clasificaciones existentes y futuras suministradas por un usuario que se determina que no es confiable hasta que el usuario se considere confiable;

descartar clasificaciones existentes suministradas por un usuario que se determina que no es confiable; y el usuario que no es confiable de consultas futuras.

443
320
390

RESUMEN

El presente invento presenta un sistema de retroalimentación cíclico que facilita la clasificación de items en relación con la prevención de spam en servidores y/o arquitecturas basadas en el cliente. El invento hace uso de los enfoques de aprendizaje de máquina tal como se aplican a los filtros spam, en particular, chequea en forma aleatoria los mensajes de e-mail de tal manera que tanto los legítimos como los basura/spam se obtienen para generar conjuntos de datos de entrenamiento. A los usuarios que se les identifica como luchadores contra el spam se les pide que voten sobre una selección de sus mensajes de e-mail entrantes para determinar si son correo legítimo o basura. Una base de datos almacena las propiedades para cada correo y la transacción de votación, tal como la información del usuario, las propiedades del mensaje, un resumen del contenido, y los resultados de la consulta para cada mensaje para generar datos de entrenamiento para los sistemas de aprendizaje de máquina. El sistema de aprendizaje de máquina facilita la creación de filtros spam mejorados que se pueden entrenar para reconocer tanto al correo legítimo como al correo spam y distinguir entre ellos.

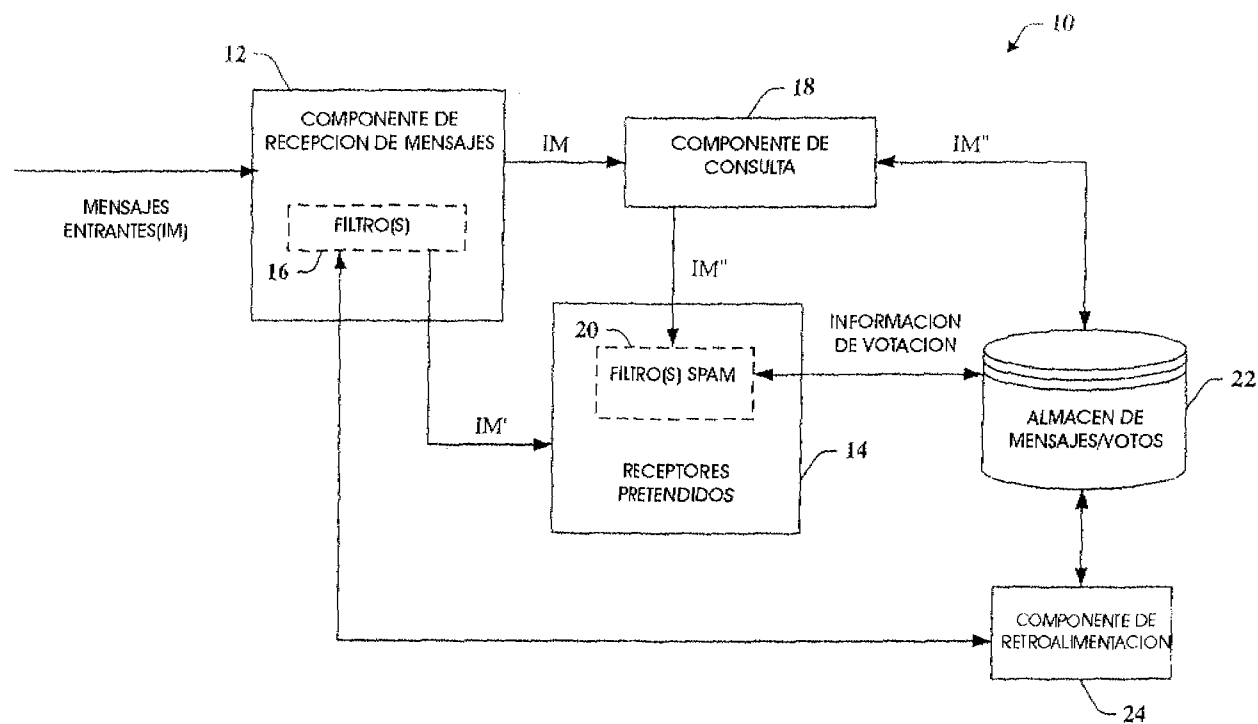


FIG. 1A

1/11

221
321
171

322 795
322

2/11

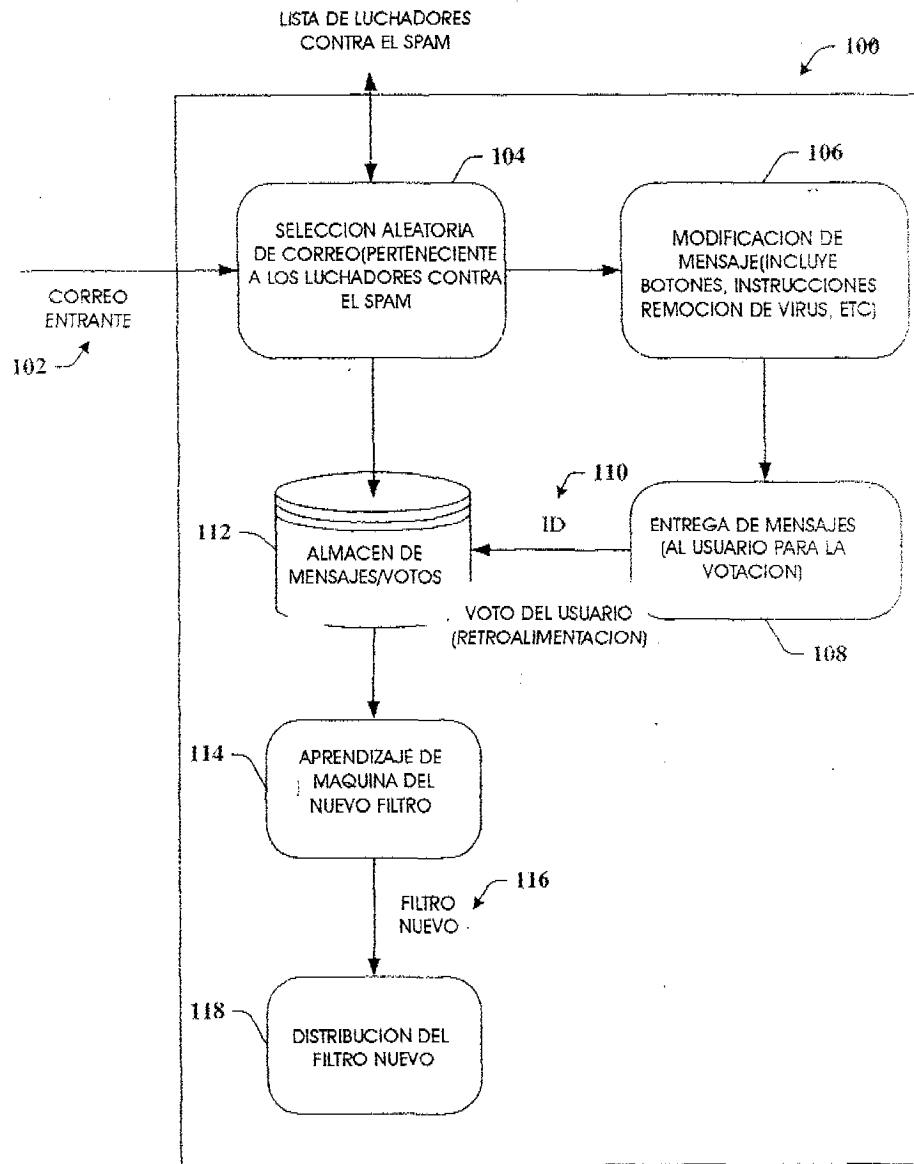


FIG. 1B

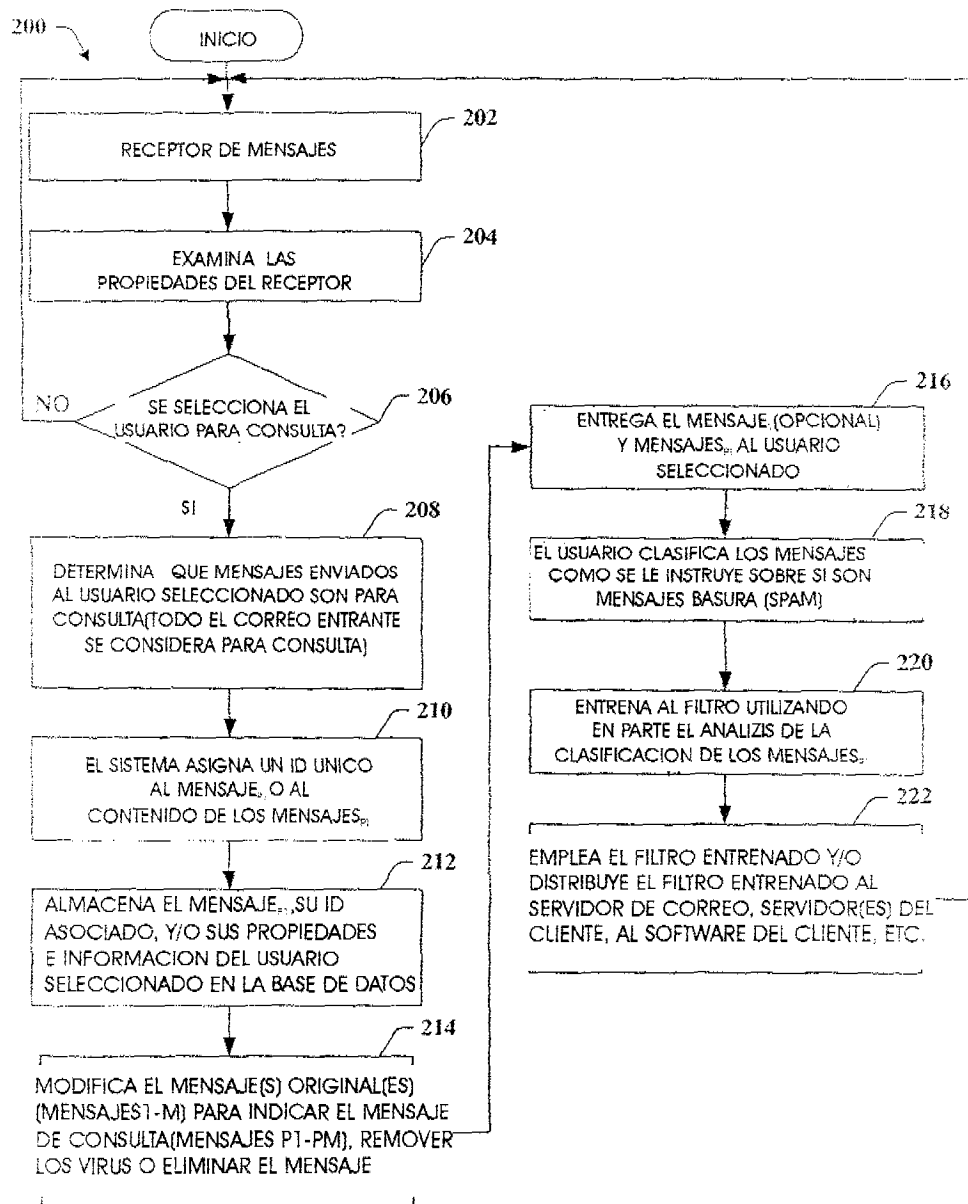


FIG. 2

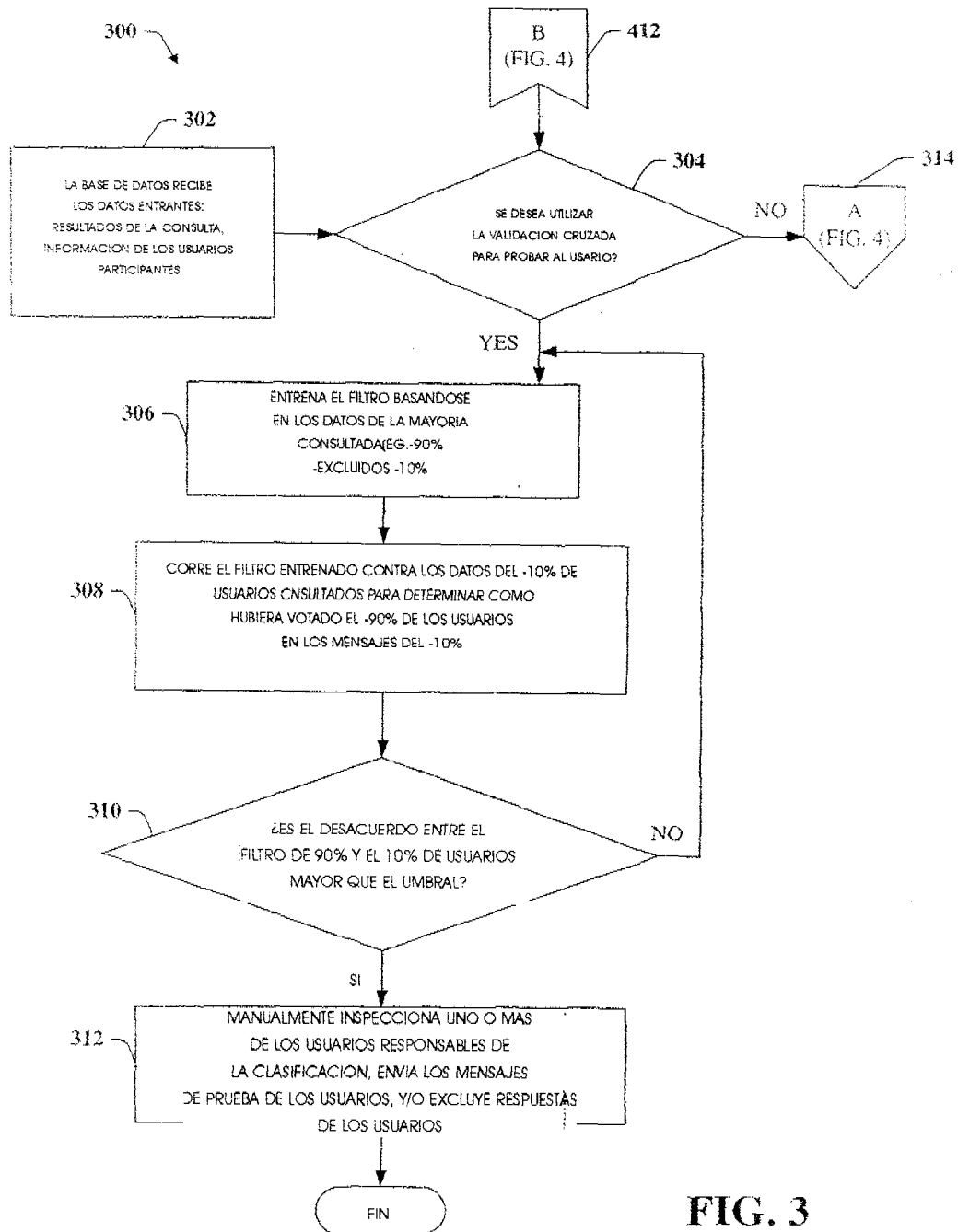


FIG. 3

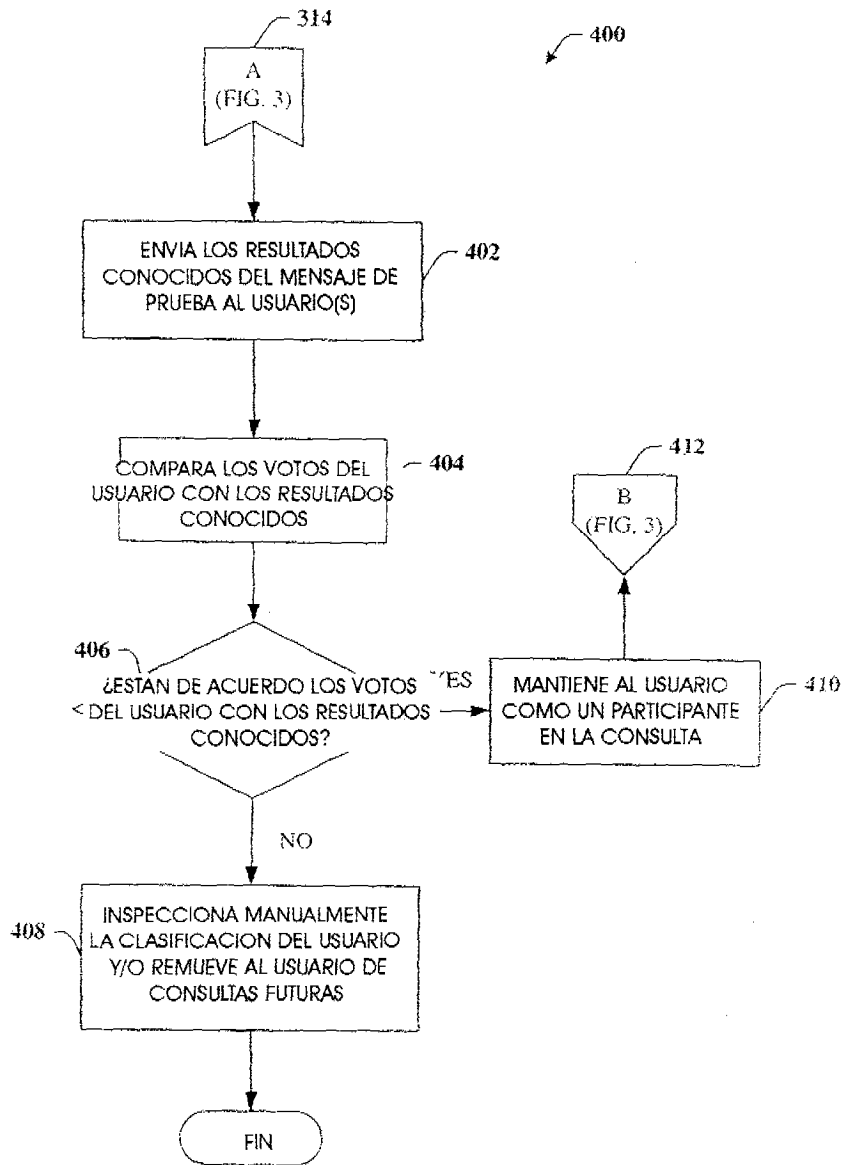


FIG. 4

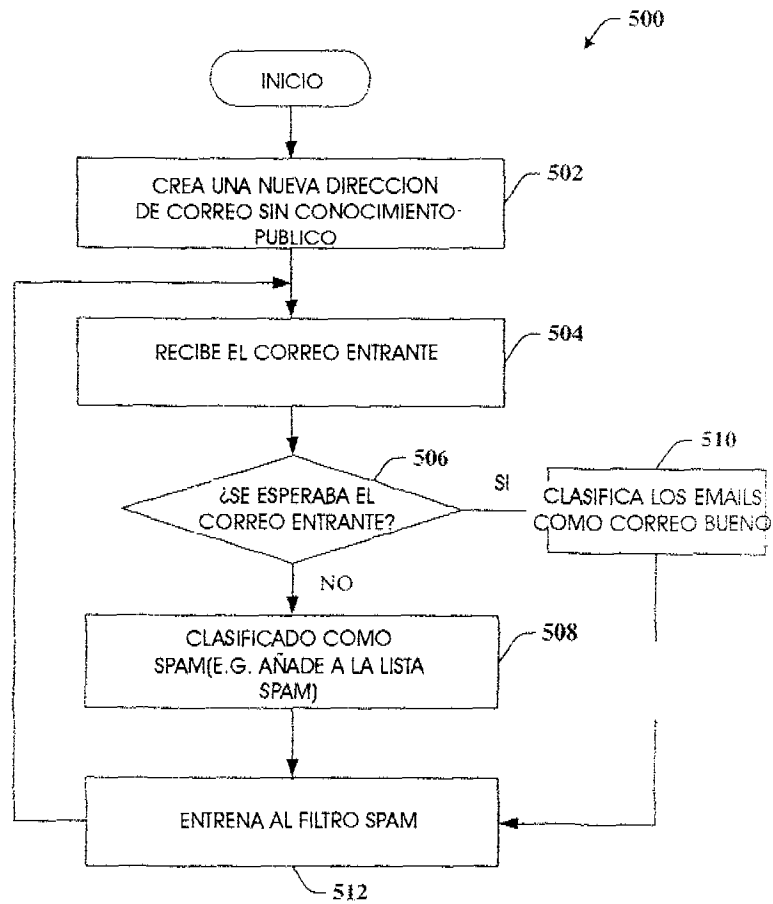


FIG. 5

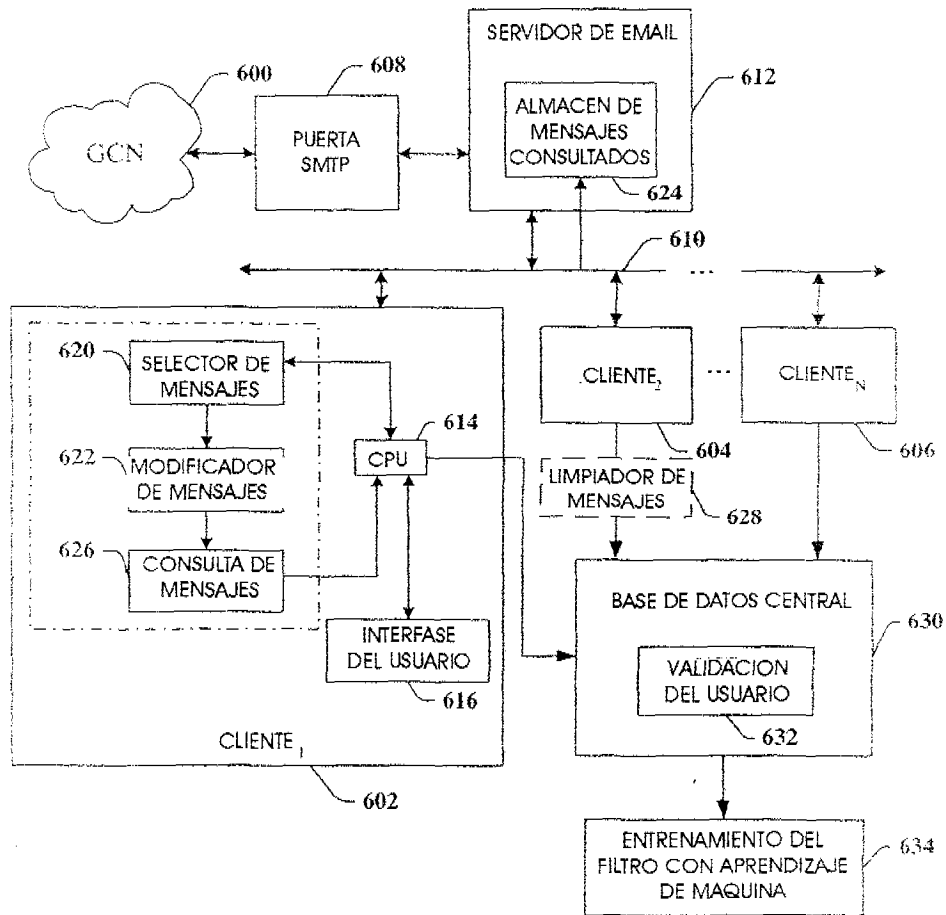


FIG. 6

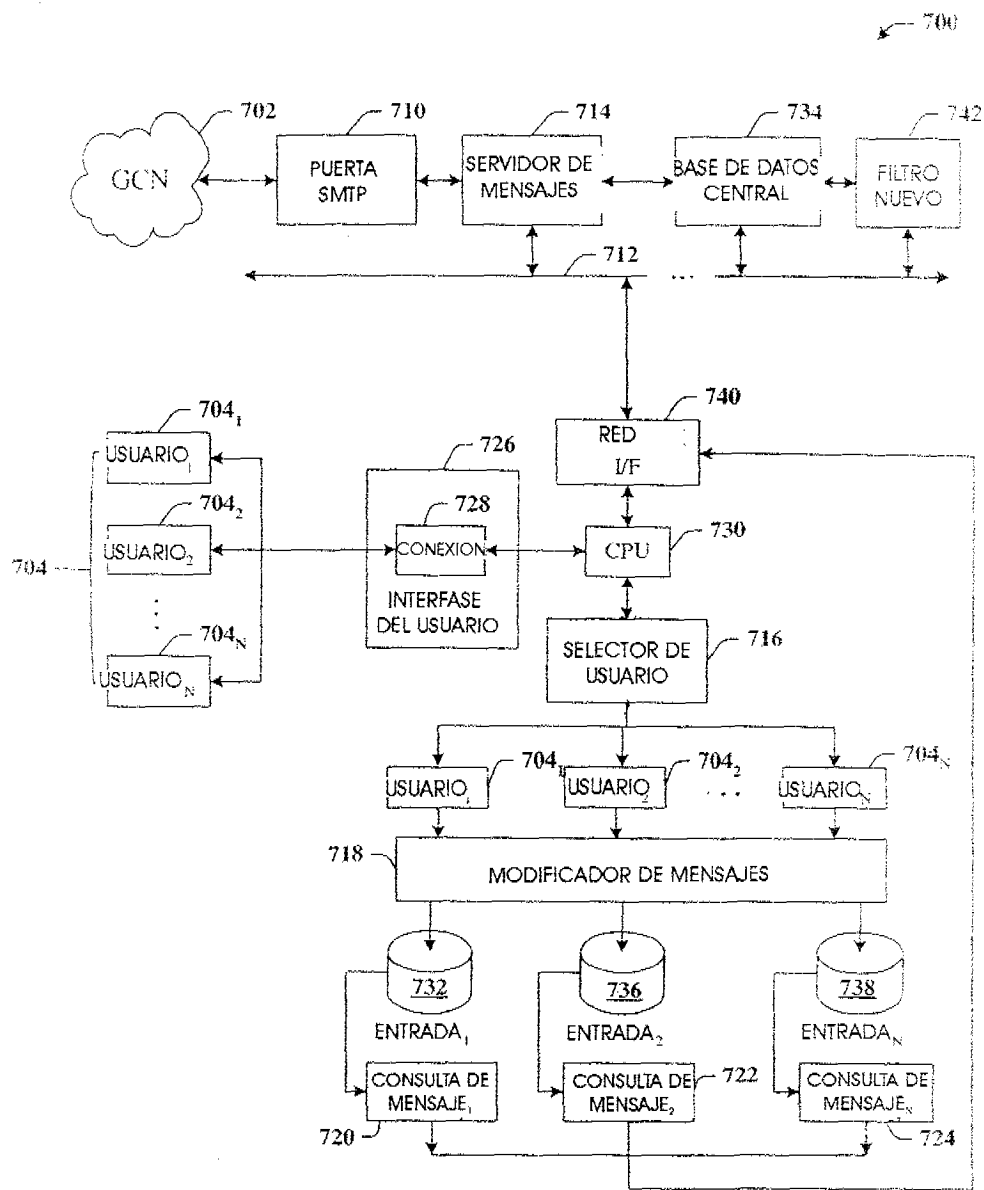


FIG. 7

752
329
329

9/11

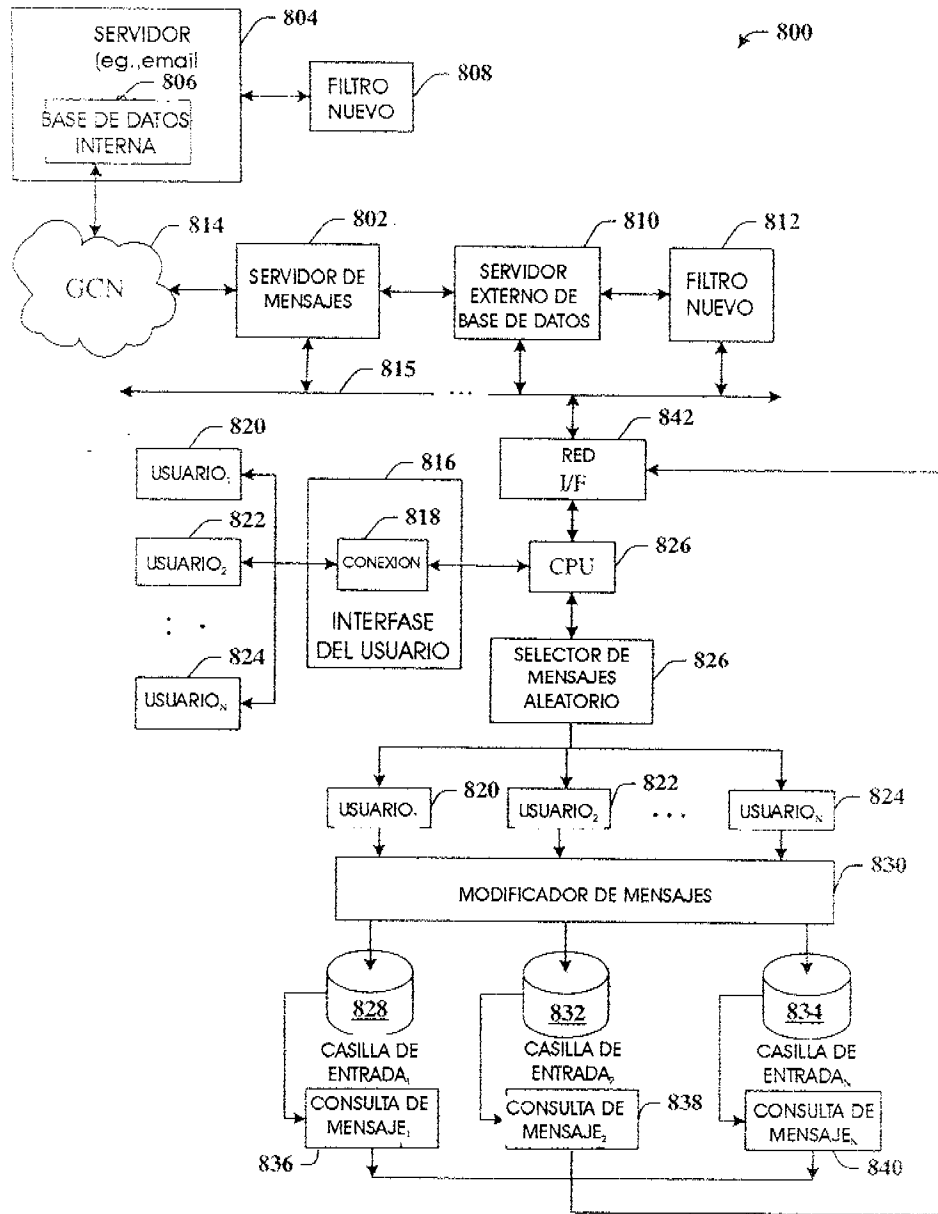


FIG. 8

153
330
330

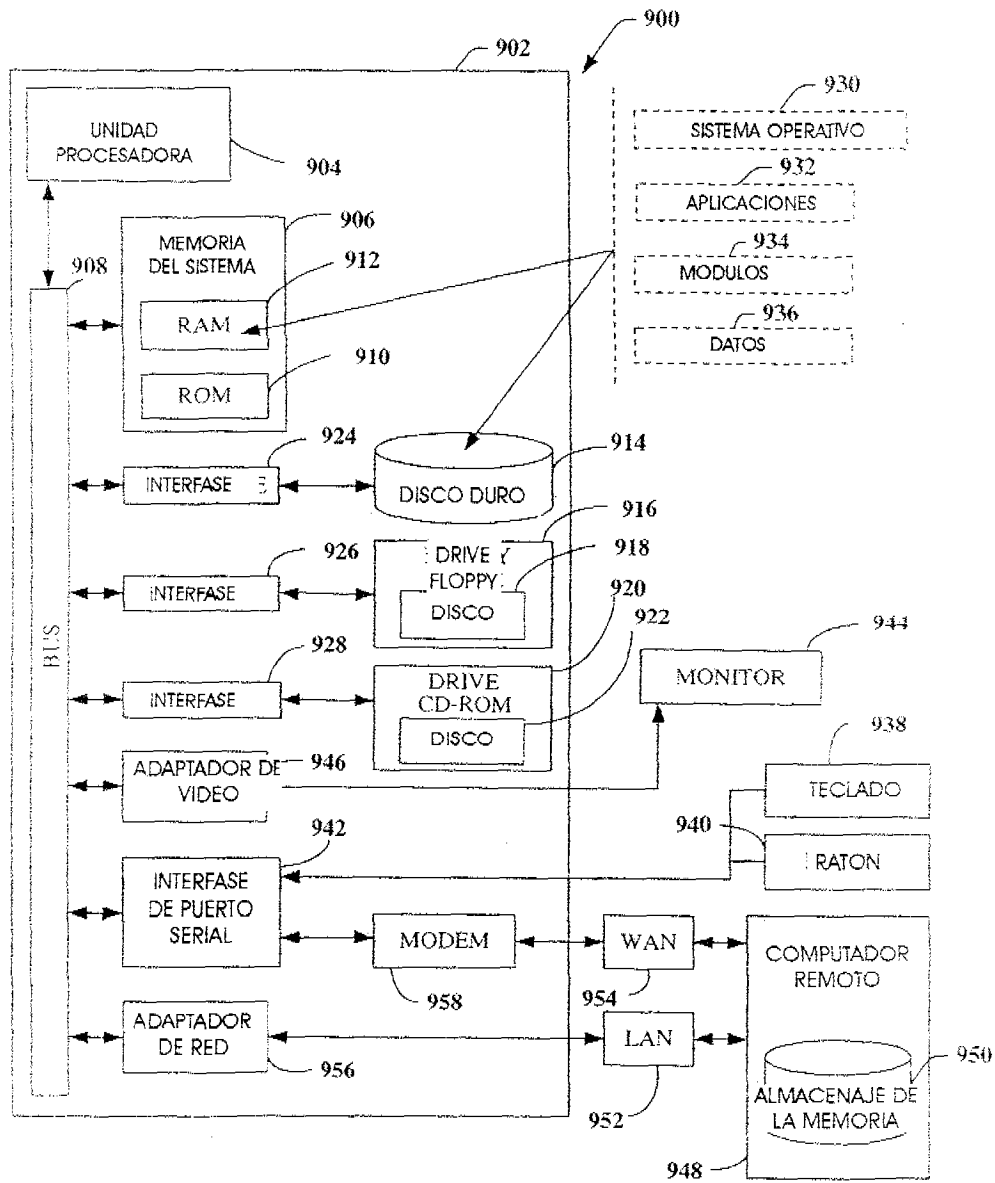


FIG. 9

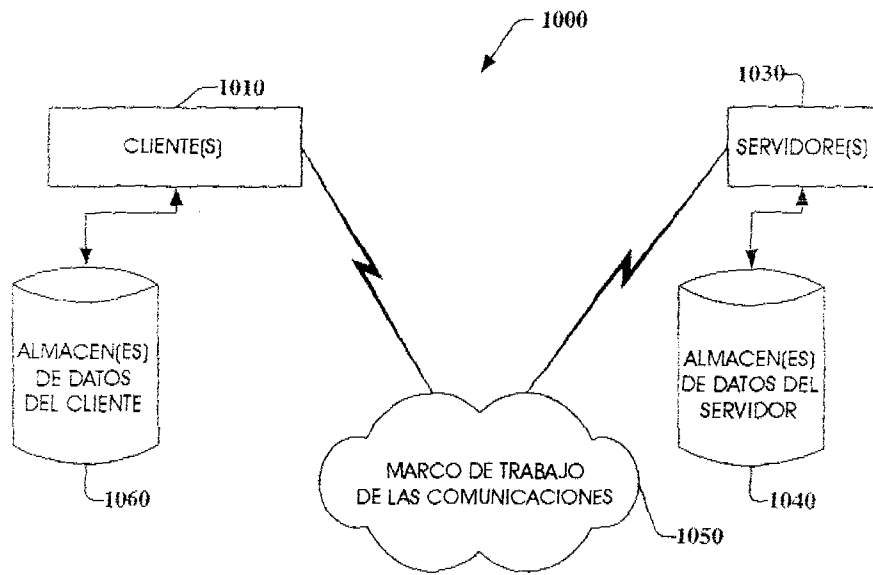


FIG. 10

INFORME INTERNO

Referencia:	Radicación No.	05-76623
	Solicitud internacional	PCT/US2004/005501
	Fecha de Inicio Fase Nacional	Octubre 3 de 2005
	Trámite	11

La solicitud no fue presentada en fase nacional tal y como fue presentada en fase internacional, ya que el capitulo original presentado en fase internacional tiene 95 reivindicaciones, mientras que el pliego radicado para entrar en fase nacional cuenta con solo 45 reivindicaciones.

El 17 de enero de 2008, el representante de la firma MICROSOFT CORPORATION radico recurso de reposición el cual contiene una copia de las 95 ~~reivindicaciones~~ reivindicaciones originales con su respectiva traducción al español. Adicionalmente se radico una solicitud modificada que contiene 44 reivindicaciones (Nota: el solicitante no presenta recibos correspondientes al pago por modificaciones a la solicitud)

Este informa no supe ni total ni parcialmente el examen de forma a que hacen referencia los artículos 38 y 39 de la decisión 486.

Bogotá D.C., 16 de julio de 2008

** folios 199 a 211*

cgomez