(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 385 029**

(51) Int. CI.:
G06Q 10/00 (2012.01)
H04L 12/58 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Número de solicitud europea: **02737515 .3**
(96) Fecha de presentación: **11.06.2002**
(97) Número de publicación de la solicitud: **1397768**
(97) Fecha de publicación de la solicitud: **17.03.2004**

(54) Título: **Procedimiento y aparato para filtrar correos electrónicos**

(30) Prioridad:
14.06.2001 US 881986

(45) Fecha de publicación de la mención BOPI:
17.07.2012

(45) Fecha de la publicación del folleto de la patente:
17.07.2012

(73) Titular/es:
APPLE INC.
1 INFINITE LOOP
CUPERTINO, CA 95014, US

(72) Inventor/es:
BELLEGARDA, Jerome R.;
NAIK, Devang y
SILVERMAN, Kim E. A.

(74) Agente/Representante:
Fàbrega Sabaté, Xavier

DESCRIPCIÓN

Procedimiento y aparato para filtrar correos electrónicos

Campo de la invención

La presente invención se refiere en general a filtrado de mensajes. Más en particular, esta invención se refiere a filtrado de correos electrónicos usando análisis semántico latente.

Aviso/Permiso de derechos de autor

Antecedentes

A medida que el uso de ordenadores e Internet han proliferado, también lo ha hecho el uso del correo electrónico. Muchos negocios y usuarios utilizan el correo electrónico como un medio de comunicación prominente. De forma no sorpresiva, el crecimiento exponencial del medio también ha atraído el interés de anunciantes comerciales de correo electrónico. Los anunciantes comerciales de correo electrónico obtienen direcciones de correo electrónico de una variedad de fuentes, por ejemplo, de vendedores de correos electrónicos o de sitios web comerciales, a menudo sin el permiso de los dueños de los correos electrónicos. Estas direcciones de correo electrónico pueden usarse entonces para promover los productos y servicios de los anunciantes de correo electrónico o de las partes que representan.

El resultado es una avalancha de correo electrónico no solicitado recibido por usuarios de correo electrónico insatisfechos. Un procedimiento para encargarse con correos electrónicos no solicitados es que el usuario los seleccione y borre manualmente los correos electrónicos no solicitados. Otros procedimientos permiten reconocer un mensaje enviado de forma masiva a múltiples recipientes y/o descartar o marcar el mensaje como un posible mensaje no solicitado. Otros procedimientos más mantienen una base de datos de direcciones de remitentes conocidos de correo electrónico no solicitado y al recibir el correo electrónico, descartan automáticamente aquellos recibidos de los remitentes conocidos de correo electrónico no solicitado. Otros procedimientos más usan filtros de palabras clave. Este procedimiento permite escanear el asunto y/o el cuerpo del mensaje de correo electrónico en busca de palabras clave predeterminadas, y si se detectan, el mensaje puede ser o descartado o marcado como sospechoso.

A pesar de los procedimientos descritos con anterioridad, los anunciantes comerciales de correo electrónico utilizan métodos ingeniosos para frustrar los esfuerzos de los destinatarios de correos electrónicos. Por ejemplo, para vencer la detección de correos electrónicos masivos, los mensajes de correo electrónico pueden enrutarse a través de un laberinto de servidores de forma que en última instancia, el mensaje no parece ser un mensaje de correo electrónico masivo. Para vencer al sistema que rastrea las direcciones de remitentes conocidos de correo electrónico no solicitado, la dirección originaria de correo electrónico no solicitado puede cambiarse a menudo. Para confundir a los procedimientos de filtrado por palabras clave, el campo asunto del correo electrónico puede titularse de forma engañosa, por ejemplo, “En respuesta a su solicitud”. Además, el procedimiento de filtrado por palabras clave es víctima de otros problemas significativos, por ejemplo, al intentar filtrar mensajes de anunciantes de correo electrónico pornográficos usando la palabra “sexo”, pueden también eliminarse artículos legítimos anatómicos o biológicos que incluyen la palabra “sexo”.

Ejemplos de varias disposiciones del estado de la técnica se analizan en:

D1: PETER W. FOLTZ, SUSAN T. DUMAIS: “Personalized Information Delivery: An Analysis of Information Filtering Methods” COMMUNICATIONS OF THE ACM, volumen 35, número 12, Diciembre 1992 (1992-12), páginas 51-60, XP002219345, que divulga un método de filtrado de correo electrónico basado en análisis semántico de texto.

D2: PETER W. FOLTZ: “Using Latent Semantic Indexing for Information Filtering” PROCEEDINGS OF THE CONFERENCE ON OFFICE INFORMATION SYSTEMS” [Online] 1990, páginas 40-47, XP002219344 Cambridge, MA, EE.UU Recuperado de Internet: URL:http://www.psych.nmsu.edu/~pfoltz/cois.html>[recuperado el 31-10-2002];

D3: LI Y H ET AL: “classification of text documents” COMPUTER JOURNAL, OXFORD UNNIVERSITY PRESS; SURREY, GB, volumen 41, número 8, 1998, páginas 537-546, XP002116464 ISSN: 0010-4620.

Algunas realizaciones de la presente invención utilizan una técnica matemática conocida como análisis semántico latente. Esta técnica se explica por ejemplo, en un artículo de uno de los inventores, Jerome R. Bellegarda titulado "Exploiting Latent Semantic Information in Statistical Language Modelling". Este artículo enseña un entorno de trabajo para clasificación semántica automática en el contexto de modelos estadísticos de lenguaje para su uso en reconocimiento automático de voz.

Resumen de la invención

Un procedimiento y aparato para filtrar mensajes, en particular mensajes de correo electrónico se describe y reivindica tal y como se expone en las reivindicaciones independientes relativo a un procedimiento para filtrar mensajes, un medio legible por máquina que incluye instrucciones que, cuando son ejecutadas por una máquina, la máquina lleva a cabo el procedimiento, un sistema de procesamiento de datos que comprende medios para llevar a cabo el procedimiento y un sistema de ordenador como dicho sistema de procesamiento de datos para llevar a cabo el procedimiento.

Según un aspecto de la presente invención, el procedimiento comprende determinar una primera ancla semántica que corresponde a un primer grupo de mensajes, por ejemplo, mensajes de correo electrónico legítimos y una segunda ancla semántica que corresponde a un segundo grupo de mensajes, por ejemplo, mensajes de correo electrónico no solicitados. El procedimiento determina además un vector que corresponde a un mensaje entrante, compara el vector con al menos uno de la primera ancla semántica y una segunda ancla semántica para obtener al menos un valor de comparación y filtra el mensaje entrante en base al valor de comparación.

Se pueden representar realizaciones de la invención como un producto de software almacenado en un medio legible por ordenador (también denominado medio legible por ordenador o medio legible por procesador). Según un aspecto de la invención, el medio legible por máquina incluye instrucciones que, cuando son ejecutadas por una máquina hacen que la máquina lleve a acabo operaciones que comprenden determinar una primera ancla semántica que corresponde a un primer grupo de mensajes, por ejemplo, mensajes de correo electrónico legítimos. El medio legible por máquina incluye instrucciones adicionales para determinar una segunda ancla semántica que corresponde a un segundo grupo de mensajes, por ejemplo, mensajes de correo electrónico no solicitados. El medio accesible por máquina incluye instrucciones adicionales para determinar un vector correspondiente a un mensaje entrante, compara el vector con al menos uno de la primera ancla semántica y la segunda ancla semántica para obtener al menos un valor de comparación y filtra el mensaje entrante en base al valor de comparación.
Según un aspecto de la invención, la invención puede representarse como un aparato, por ejemplo, sistema de ordenador. El sistema de ordenador comprende un bus, un dispositivo de almacenamiento de datos acoplado al bus y un procesador acoplado al dispositivo de almacenamiento de datos, dicho procesador para llevar a cabo un procedimiento que comprende determinar una primera ancla semántica correspondiente a un primer grupo de mensajes. El procesador también determina una segunda ancla semántica correspondiente a un segundo grupo de mensajes. El procesador determina además un vector correspondiente a un mensaje entrante, compara el vector correspondiente al mensaje de entrada al menos con uno de la primera ancla semántica y la segunda ancla semántica para obtener un primer valor de comparación y un segundo valor de comparación. El procesador filtra el mensaje entrante en base al primer valor de comparación y el segundo valor de comparación.

BREVE DESCRIPCIÓN DE LOS DIBUJOS

La presente invención está ilustrada a título de ejemplo y no limitante en las figuras de los dibujos anejos, en los que referencias similares indican elementos similares y en los que:

La Figura 1 es un diagrama de bloques que ilustra un sistema de filtrado de correo electrónico según una realización de la presente invención;

La Figura 2 es un diagrama de matrices y vectores utilizados para encontrar anclas semánticas;

La Figura 3 es un diagrama de flujo que ilustra un procedimiento utilizado en filtrado de correo electrónico según una realización de la presente invención.

La Figura 4 ilustra un diagrama de bloques de un dispositivo de computación para su uso con una realización de la presente invención.

La Figura 5 ilustra un diagrama de bloques de una realización de la invención almacenado en un medio accesible por máquina.

DESCRIPCIÓN DETALLADA

Se describe un procedimiento y aparato para filtrar correo electrónico usando análisis semántico latente.

En la siguiente descripción, se exponen numerosos detalles específicos para proporcionar un entendimiento detallado de la presente invención. Será aparente, sin embargo, para un experto medio en el estado de la técnica que la presente invención se puede llevar a cabo sin estos detalles específicos. En otros casos, arquitecturas, etapas y técnicas sobradamente conocidas no han sido mostradas para no oscurecer innecesariamente la presente invención.

Se pueden presentar partes de la descripción utilizando terminología comúnmente utilizada por aquellos expertos en la materia para transmitir la sustancia de su trabajo a otros expertos en la técnica. Además, se pueden presentar partes de la descripción en términos de operaciones llevadas a cabo a través de la ejecución de instrucciones de programación. Como es sobradamente entendido por aquellos expertos en la técnica, estas operaciones a menudo toman forma de señales eléctricas, magnéticas, u ópticas susceptibles de ser almacenadas, transferidas, combinadas y aparte de eso manipuladas a través de, por ejemplo, componentes eléctricos.

La invención puede utilizar un entorno distribuido de computación. En un entrono distribuido de computación, se pueden ubicar físicamente módulos de programa en diferentes dispositivos de almacenamiento de memoria local y remota. La ejecución de los módulos de programa puede ocurrir localmente de forma autónoma o remotamente de forma cliente/servidor. Ejemplos de tales entornos de computación distribuidos incluyen redes de área local, redes de ordenadores de empresa y la Internet global.

Además, debería entenderse que los programas, procesos, procedimientos, etcétera descritos en este documento no están relacionados o limitados a cualquier ordenador o aparato particular ni están relacionados con o limitados a cualquier arquitectura de red de comunicación particular. Por el contrario, pueden usarse varios tipos de máquinas de propósito general con módulos de programa construidos según las enseñanzas descritas en este documento. De forma similar, puede ser ventajoso construir un aparato especializado para llevar a cabo las etapas descritas en este documento por medio de sistemas dedicados de ordenador en una arquitectura de red específica con lógica cableada o programas almacenados en memoria no volátil tal como memoria de sólo lectura.

Se describirán varias operaciones como múltiples etapas discretas llevadas a cabo en turnos de forma que es útil para entender la presente invención. Sin embargo, el orden de la descripción no debería interpretarse como que implica que estas operaciones se llevan a cabo en el orden en que se presentan o incluso dependientes del orden. Finalmente, el uso repetido de la frase “en una realización” no se refiere necesariamente a la misma realización, aunque puede.

El Análisis Semántico Latente (LSA) es un procedimiento que descubre automáticamente las relaciones semánticas salientes entre palabras y documentos en un cuerpo dado. Se mapean palabras discretas a un espacio vectorial semántico continuo, en el que se aplican técnicas de agrupamiento. El procedimiento para filtrar mensajes de correo electrónico comprende determinar una primera ancla semántica correspondiente a un primer grupo de mensajes de correo electrónico, por ejemplo, mensaje de correo electrónico legítimos y una segunda ancla semántica correspondiente a un segundo grupo de mensajes de correo electrónico, por ejemplo mensajes de correo electrónico no solicitados. Determinar un vector que corresponde a un mensaje entrante, comparando el vector con al menos uno de la primera ancla semántica y una segunda ancla semántica para obtener al menos un valor de comparación y filtrar los mensajes en base al valor de comparación.

La Figura 1 es un diagrama de bloques que ilustra un sistema de filtrado de correos electrónicos **100** según una realización de la presente invención. Aunque la descripción que sigue describe el filtrado de mensajes de correo electrónico, un experto en la técnica apreciará que el sistema puede usarse para filtrar archivos adjuntos de correo electrónico, mensajes de audio transcritos, programas de ordenador, por ejemplo, virus de ordenador, texto y similares. En una realización, el sistema de filtrado de correo electrónico **100** filtra mensajes de correo electrónico no solicitados de los mensajes de correo electrónico legítimos. Sin embargo, un experto en la técnica apreciará que otras realizaciones pueden clasificar mensajes en más de dos grupos.

El sistema de filtrado de correo electrónico **100** comprende una unidad de entrenamiento de correo electrónico **105** que incluye un cuerpo de entrenamiento de correo electrónico *T*, por ejemplo, una base de datos que comprende una colección de N_1 mensajes de correo electrónico legítimos y N_2 mensajes de correo electrónico no solicitados. En una realización, los mensajes de correo electrónico legítimos y no solicitados se obtienen de los correos electrónicos recibidos por un destinatario. Realizaciones alternativas pueden permitir a un usuario clasificar manualmente cada mensaje de correo electrónico entrante hasta que se ha establecido un cuerpo adecuado de entrenamiento de correo electrónico T. Las palabras utilizadas en la colección de mensajes de correo electrónico legítimos y en la colección de mensajes de correo electrónico no solicitados son de algún vocabulario subyacente v que comprende, pro ejemplo, las M palabras utilizadas más frecuentemente en el lenguaje. En una realización, M puede ser diez mil, y $1 \leq N_1 = N_2 \leq 150$.

La unidad de co-ocurrencias **110** del sistema de filtrado de correos electrónicos **100**, comprende una matriz W

bidimensional (M x 2) formada utilizando el cuerpo de entrenamiento de correo electrónico *T*. La matriz W esencialmente hace un seguimiento de qué palabra se encuentra en qué documento manteniendo un registro del número de veces que cada palabra aparece en cada correo electrónico legítimo y no solicitado. En particular, las entradas ω_{ii} de la matriz W refleja hasta qué punto cada palabra ω_i □ v apareció en el mensaje de correo electrónico legítimo (j = 1) o en un mensaje de correo electrónico no solicitado (j = 2). Se pueden utilizar varios procedimientos para mantener un registro del número de ocurrencias de una palabra en un documento, por ejemplo, una simple cuenta normalizada del número de ocurrencias de cada palabra. Sin embargo, en una realización, la unidad de co-ocurrencia **110** utiliza la función

$$\omega_{i,j} = (1-\varepsilon_i)\frac{c_{i,j}}{N_j} \qquad (1)$$

que normaliza la longitud del documento y la entropía de palabra para forma la matriz W. c_{ii} denota el número de veces que cada palabra ω_i ocurre en la colección de mensajes de correo electrónico no solicitado. En la ecuación (1) N_i, para j = 1 y j = 2, representa el número total de palabras en la colección de mensajes de correo electrónico legítimos y mensaje de correo electrónico no solicitados. ε_i es la entropía normalizada de ω_i en el cuerpo de correos electrónicos de entrenamiento T. (1 - ε_i) es simplemente un factor de ponderación, o una factor de distribución de palabras y es una medida de la distribución de una palabra particular en el cuerpo de correos electrónicos de entrenamiento *T*. Esto se explica a continuación.

En una realización, la unidad de co-ocurrencias **110** calcula ε_i utilizando la ecuación:

$$\varepsilon_i = -\frac{1}{\log N}\sum_{j=1}^{N}\frac{c_{i,j}}{t_i}\log\frac{c_{i,j}}{t_i} \qquad (2)$$

en donde $N = N_1 + N_2$. Por definición, $0 \le \varepsilon_i \le 1$, con igualdad si y solo si $c_{i,i} = t_i$ y $c_{i,i} = t_i / N$. Por lo tanto, un valor de ε_i cercano a 1 indica una palabra distribuida a lo largo de muchos correos electrónicos a lo largo de cuerpo de correos electrónicos de entrenamiento *T*. Sin embargo, un valor de ε_i cercano a 0 indica que la palabra está presente solo en unos pocos correos electrónicos. Por lo tanto, el factor de ponderación es una medida de la distribución de una palabra a lo largo del cuerpo de correos electrónicos de entrenamiento *T*. En particular, el factor de ponderación (1 - ε_i) es una medida de la potencia de indexación de la palabra ω_i.

Después de que la unidad de co-ocurrencias **110** construye la matriz W, la unidad **115** de Descomposición en Valores Singulares (SVD) descompone la matriz W, y subsecuentemente obtiene las anclas semánticas $\bar{v}_1$ y $\bar{v}_2$. Las anclas semánticas $\bar{v}_1$ **120** y $\bar{v}_2$ **125** son vectores derivados de la matriz W utilizando SVD. En una realización, los vectores $\bar{v}_1$ y $\bar{v}_2$ se derivan utilizando la siguiente ecuación:

$$W = USV^T \qquad (3)$$

en donde U es la matriz singular izquierda (M x 2) con vectores de fila u_i ($1 \le i \le M$), S es la matriz diagonal (2 x 2) de valores singulares $s1 \ge s2 \ge 0$, V es la matriz singular derecha (2 x 2) con vectores de fila v_j (j = 1,2), y *T* denota transposición de matriz. Por lo tanto, el vector $\bar{v}_1$ representa mensajes de correo electrónico legítimos y el vector $\bar{v}_2$ representa mensajes de correo electrónico no solicitados.

La Figura 2 es un diagrama de bloques de la SVD de la matriz W. Como se ilustra en la Figura 2, la SVD de la matriz W define un mapeado entre la representación matemática del conjunto de mensajes de correo electrónico legítimos y no solicitados **205** y **210** respectivamente y el espacio vectorial semántico latente extendido por los vectores singulares contenidos en U y V. El mapeado es escalado entonces por la matriz diagonal **230**, para asegurar una representación adecuada. A partir de este mapeado, la primera ancla semántica dada por

$$\bar{v}_1 = v_1 S \qquad (4)$$

y la segunda ancla semántica dada por

$$\bar{v}_2 = v_2 S \qquad (5)$$

se obtienen después de un escalado apropiado por la matriz diagonal S. Un experto en la técnica apreciará que V_1^T **215** y V_2^T **200** son anclas semánticas no escaladas en la matriz (2 x 2) V^T **235**, y pueden convertirse fácilmente a vectores bidimensionales $\bar{v}_1$ y $\bar{v}_2$ utilizando las ecuaciones (4) y (5) anteriores. Si se desean más de dos grupos de clasificación, es decir, grupos de clasificación distintos a legítimos y no solicitados, un experto en la técnica apreciará que las anclas semánticas correspondientes a cada grupo de clasificación pueden obtenerse como se describió anteriormente. La matriz U **240** se utiliza para calcular el vector correspondiente a un mensaje de correo electrónico entrante como se explica a continuación.

Volviendo a la Figura 1, cada vez que un mensaje de correo electrónico entrante es recibido por la unidad de correo electrónico entrante **150**, la ecuación 1 puede ser usada por la unidad de conversión de correo electrónico entrante **155** para convertir el correo electrónico entrante a un vector de columna d_3 de dimensión M. En una realización, el vector de columna resultante d_3 puede insertarse como una columna adicional en la matriz W, con lo que convierte la matriz W de dimensión (M x 2) en una matriz de dimensión (M x 3). Utilizando la SVD de la ecuación (2), se obtiene una representación no escalada del nuevo mensaje de correo electrónico. Por lo tanto, la representación de vector de correo electrónico de entrada se obtiene como sigue:

$$\bar{v}_3 = v_3 S = d_3^T U \qquad (6)$$

El vector bidimensional $\bar{v}_3$ de la ecuación (6) es la representación matemática del nuevo mensaje de correo electrónico y puede interpretarse como un punto en el espacio vectorial semántico latente expandido por los vectores $\bar{v}_1$ y $\bar{v}_2$.

Un experto en la técnica apreciará que la ecuación (6) es una representación aproximada del mensaje en el espacio existente LSA. Ya que el nuevo mensaje de correo electrónico no era parte de la extracción SVD original, las palabras en el nuevo mensaje de correo electrónico que no se encuentran en el cuerpo de entrenamiento *T*, pueden provocar que no se cumpla más la expansión SVD. Como tal, en una realización, un camino opcional de realimentación **180**, tal y como se ilustra en la Figura 1, puede usarse para añadir el nuevo mensaje de correo electrónico al cuerpo de entrenamiento *T*. Las anclas semánticas $\bar{v}_1$ y $\bar{v}_2$ pueden recalcularse periódicamente para tener en cuenta las nuevas palabras en los nuevos mensajes de correo electrónico, de forma que los mensajes de correo electrónico subsecuentes pueden ser clasificados de forma precisa como legítimos o no solicitados.

La invención contempla capturar asociaciones estructurales entre palabras. Por lo tanto, dos palabras cuyas representaciones sean “cercanas” (en alguna métrica adecuada) tienden a aparecer en el mismo tipo de documentos, ocurran o no dentro del mismo contexto de palabras en los documentos. Cada ancla semántica $\bar{v}_1$ y $\bar{v}_2$ puede verse como el centroide de las palabras en los mensajes de correo electrónico legítimos y en los mensajes de correo electrónico no solicitados respectivamente. Esto significa que palabras asociadas tales como sinónimos ocurren en proximidad cercana a otras palabras similares en cada categoría de los mensajes de correo electrónico legítimos y no solicitados en el espacio vectorial semántico S. Por ejemplo, si una palabra particular se encuentra con más frecuencia en los mensajes de correo electrónico no solicitados en comparación con los mensajes de correo electrónico legítimos en el cuerpo de entrenamiento, un correo electrónico entrante que contiene un sinónimo de la palabra estará más cercano a la categoría de mensajes de correo electrónico no solicitados en el espacio vectorial semántico S. Por lo tanto, el sistema de filtrado de correos electrónicos **100** clasifica apropiadamente los mensaje de correo electrónico entrantes que contienen sinónimos eliminando la necesidad de recalcular frecuentemente las anclas semánticas $\bar{v}_1$ y $\bar{v}_2$.

Después de calcular las anclas semánticas $\bar{v}_1, \bar{v}_2$, y la representación vectorial $\bar{v}_3$ del nuevo mensaje de correo electrónico, se calcula una medida de proximidad K. La medida de proximidad K es una medida de cuán cerca un nuevo mensaje de correo electrónico está de un mensaje de correo electrónico legítimo o de un mensaje de correo electrónico no solicitado. La medida de proximidad K es computada por la unidad de cálculo **160** y, en una realización, compara el ángulo que forman los vectores $\bar{v}_1$ y $\bar{v}_3$, con el ángulo formado entre los vectores $\bar{v}_2$ y $\bar{v}_3$. La medida de proximidad K puede calcularse utilizando:

$$K(\bar{v}_3, \bar{v}_j) = \cos(v_3 S, v_j S) = \frac{v_3 S^2 v_j^T}{\|v_3 S\| \|v_j S\|} \qquad (7)$$

para j = 1, 2. Se pueden emplear otros procedimientos para calcular la medida de proximidad K incluyendo, pero no limitados a, calcular la longitud de las normales entre los vectores $\bar{v}_1$, $\bar{v}_2$ y $\bar{v}_3$.

Después de calcular la medida de proximidad K, la unidad lógica **165** determina si el nuevo correo electrónico es no solicitado, **170**, legítimo **175** o ambiguo **180**. En una realización, si $\bar{v}_3$ es más cercano a $\bar{v}_1$, es decir, el ángulo entre $\bar{v}_3$ y $\bar{v}_1$ es más pequeño que el ángulo entre $\bar{v}_3$ y $\bar{v}_2$, el nuevo correo electrónico se considera un mensaje de correo electrónico legítimo **175**, un sistema de filtrado de correos electrónicos **100** puede permitir automáticamente que el nuevo correo electrónico sea visto por su destinatario deseado. Opcionalmente, el sistema de filtrado de correos electrónicos puede permitir al usuario incluir el mensaje de correo electrónico legítimo como parte del cuerpo de mensajes de correo electrónico de entrenamiento *T*. Alternativamente, si $\bar{v}_3$ está más cerca de $\bar{v}_2$, es decir, el ángulo entre $\bar{v}_3$ y $\bar{v}_1$ es más grande que el ángulo entre $\bar{v}_3$ y $\bar{v}_2$, el nuevo correo electrónico se considera no solicitado **170**. En una realización, los mensajes de correo electrónico no solicitados pueden ser descartados automáticamente por el sistema de filtrado de correos electrónicos. Realizaciones alternativas pueden mantener una copia del correo electrónico no solicitado de forma que un usuario puede, a conveniencia del usuario, descartar el correo electrónico no solicitado o incluirlo para formar parte del cuerpo de correos electrónicos de entrenamiento *T*.

Si el ángulo entre $\bar{v}_3$ y $\bar{v}_1$ es aproximadamente igual al ángulo entre $\bar{v}_3$ y $\bar{v}_2$, la unidad lógica **165** puede marcar el mensaje de correo electrónico como ambiguo **180**, por ejemplo, con un icono para indicar un mensaje de correo electrónico ambiguo. Realizaciones alternativas pueden marcar cada mensaje de correo electrónico entrante con una marca única para cada uno de las categorías no solicitado, legítimo y ambiguo, facilitando la ordenación y manejo de los mensajes de correo electrónico recibidos. Respecto a los mensajes de correo electrónico ambiguos, en una realización, el usuario puede determinar si el mensaje de correo electrónico es legítimo o no solicitado. Realizaciones alternativas pueden permitir a un usuario descartar el mensaje de correo electrónico ambiguo, o incluirlo, después de eliminar la ambigüedad, para formar parte del cuerpo de correos electrónicos de entrenamiento *T*, de forma que la ambigüedad asociada con mensajes similares futuros pueda ser gestionada de forma automática por el sistema de filtrado de correos electrónicos **100**.

A modo de ejemplo, considérense los siguientes mensajes de correo electrónico recibidos por una persona en el negocio de la pesca: (a) La pesca es excelente en la orilla (“bank” en inglés) sur del río y (b) El banco Mercante tiene altos tipos de interés. Aunque ambos mensajes incluyen la palabra “banco” en el texto del mensaje, el procedimiento descrito clasificará adecuadamente el mensaje (a) como un mensaje de correo electrónico legítimo y el mensaje (b) como un mensaje de correo electrónico no solicitado.

El cuerpo de correos electrónicos de entrenamiento *T* se desarrolla usando mensajes de correo electrónico existentes del usuario en el negocio de la pesca. Después de que el cuerpo de correos electrónicos de entrenamiento *T* es generado por la unidad de entrenamiento de correos electrónicos **105**, la unidad de co-ocurrencias **110** genera la matriz W utilizando el cuerpo de correos electrónicos de entrenamiento *T*. La unidad SVD **115** descompone la matriz W y obtiene las anclas semánticas $\bar{v}_1$ y $\bar{v}_2$. Cuando dos mensajes de correo electrónico son recibidos por el usuario en el negocio de la pesca, cada uno es convertido a un vector $\bar{v}_3$ utilizando la ecuación 6 anterior. En una realización, para cada mensaje de correo electrónico la medida de proximidad K entre $\bar{v}_1$ y $\bar{v}_3$ y entre $\bar{v}_2$ y $\bar{v}_3$ se calcula utilizando la ecuación 7. Para el mensaje de correo electrónico (a), la medida de proximidad K indica que $\bar{v}_3$ está más cerca de $\bar{v}_1$ en comparación con $\bar{v}_2$ con lo que indica que el mensaje es legítimo. Sin

embargo, para el mensaje no solicitado (b) la medida de proximidad K indica que el $\bar{v}_3$ está más cerca del vector no solicitado $\bar{v}_2$ en comparación con $\bar{v}_1$ indicando que el mensaje es no solicitado. Por lo tanto, a pesar de que la misma palabra “banco” está presente en ambos mensajes de correo electrónico, el contexto en el que aparecen es tomado en consideración para determinar si el mensajes de correo electrónico recibido es legítimo o no solicitado.

La Figura 3 ilustra un procedimiento que puede ser usado para filtrar correos electrónicos según una realización de la invención. En **305** se accede al cuerpo de correos electrónicos de entrenamiento *T*, y en **310** los mensajes de correo electrónico en el cuerpo de correos electrónicos de entrenamiento *T* se utilizan para construir la matriz W (descrita con anterioridad) que esencialmente hace un seguimiento de qué palabra se encuentra en qué documento. En particular, la matriz W mantiene un registro del número de veces que cada palabra aparece en cada mensaje de correo electrónico legítimo y no solicitado. En una realización, la ecuación (1) se utiliza para construir la matriz W. Una vez construida la matriz W, en 315 se lleva a cabo SVD utilizando la ecuación (3) y las anclas semánticas $\bar{v}_1$ y $\bar{v}_2$ utilizando las ecuaciones (4) y (5).

En **320**, se recibe un mensaje de correo electrónico de entrada, y en **325**, se construye el vector $\bar{v}_3$ a partir del mensaje de correo electrónico entrante utilizando la ecuación (6). En **330**, se obtiene una medida de proximidad K utilizando la ecuación (7). Como se explicó con anterioridad, la medida de proximidad determina si el nuevo mensaje de correo electrónico es legítimo, no solicitado o ambiguo.

En **335**, se hace una determinación de si el nuevo mensaje de correo electrónico es legítimo. Si el ángulo entre $\bar{v}_3$ y $\bar{v}_1$ es menor que el ángulo entre $\bar{v}_3$ y $\bar{v}_2$, en **345**, el nuevo mensaje de correo electrónico se clasifica como legítimo. En una realización los mensajes de correo electrónico legítimos pueden transmitirse al destinatario deseado.

En **350**, se hace una determinación de si el nuevo mensaje de correo electrónico es no solicitado. Si el ángulo entre $\bar{v}_3$ y $\bar{v}_1$ es mayor que el ángulo entre $\bar{v}_3$ y $\bar{v}_2$, en **340**, el nuevo mensaje de correo electrónico se clasifica como no solicitado. En una realización, el nuevo mensaje de correo electrónico que es clasificado como no solicitado puede ser automáticamente descartado. Realizaciones alternativas pueden permitir que los mensajes de correo electrónico legítimos y no solicitados recientemente clasificados formen parte del cuerpo de correos electrónicos de entrenamiento *T*.

Sin embargo, si el ángulo entre $\bar{v}_3$ y $\bar{v}_1$ es aproximadamente igual al ángulo entre $\bar{v}_3$ y $\bar{v}_2$, en **355** el mensaje de correo electrónico puede clasificarse como ambiguo. En una realización los mensajes de correo electrónico ambiguos se transmiten al recipiente deseado del mensaje de correo electrónico para eliminar la ambigüedad y para clasificar el mensaje de correo electrónico como legítimo o no solicitado. En una realización, después de que un recipiente clasifique un mensaje de correo electrónico, el mensaje de correo electrónico se incluye en cuerpo de correos electrónicos de entrenamiento y se calculan nuevas anclas semánticas. Por lo tanto, la próxima vez que se recibe un mensaje de correo electrónico con mensaje similar al mensaje de correo electrónico ambiguo, el sistema de filtrado de correo electrónico clasifica automáticamente el correo electrónico como legítimo o no solicitado.

Pueden utilizarse realizaciones del sistema de filtrado de correos electrónicos de forma individual en una máquina para un usuario particular o en una máquina central, por ejemplo, un servidor de correos electrónicos, para filtrar mensajes de un grupo de destinatarios de correos electrónicos. Realizaciones alternativas pueden incluir utilizar el sistema de filtrado de correos electrónicos en un servidor u otro dispositivo que se comunica con un usuario remoto, por ejemplo, un usuario que utiliza un dispositivo inalámbrico tal como un asistente digital personal (PDA) u ordenador inalámbrico portátil, de forma que la memoria limitada del dispositivo portátil no se llena innecesariamente con mensajes de correo electrónico no solicitados. Realizaciones alternativas pueden utilizar el sistema de filtrado de correos electrónicos en la PDA y se pueden descartar mensajes no solicitados tan pronto como son recibidos.

La Figura 4 ilustra una realización de un aparato que puede ser utilizado para filtrar mensajes de correo electrónico. Aunque la realización descrita utiliza un ordenador personal, se pueden también usar otros dispositivos incluyendo dispositivos inalámbricos tales como teléfonos móviles y asistentes personales digitales. Se puede implementar una realización de la presente invención en una arquitectura de ordenador personal (PC). Será evidente para aquellos expertos medios en la técnica que se pueden también utilizar arquitecturas de sistemas de ordenadores alternativas u otros dispositivos basados en procesadores, programables o en electrónica.

En general, tales sistemas de ordenadores como se ilustra en la Figura 4 incluyen un procesador **402** acoplado a través de un bus **401** a una memoria de acceso aleatorio (RAM) **403**, una memoria de solo lectura (ROM) **404** y un dispositivo de almacenamiento masivo **407**. El dispositivo de almacenamiento masivo **407** representa un dispositivo de almacenamiento de datos persistente, tal como una disquetera de discos flexibles, disquetera de discos fijos (por ejemplo, magnética, óptica, magneto-óptica, o similar), o disquetera de cintas. El procesador **402** puede ser

cualquiera de una amplia variedad de procesadores o microprocesadores de propósito general (como el procesador Pentium® fabricado por la Intel® Corporation), un procesador de propósito especial o un dispositivo lógico específicamente programado.

El dispositivo de visualización **405** está acoplado al procesador **402** a través del bus **401** y proporciona salida gráfica al sistema de ordenadores **400**. Los dispositivos de entrada **406** tales como un teclado o un ratón se acoplan al bus **401** para comunicar información y selecciones de comando al procesador **402**. Acoplado también al procesador **402** a través del bus **401** está un interfaz de entrada/salida **410** que puede usarse para controlar y transferir datos a los dispositivos electrónicos (impresoras, otros ordenadores, etcétera) conectados al sistema de ordenadores **400**. El sistema de ordenadores **400** incluye dispositivos de red **408** para conectar el sistema de ordenadores **400** a una red **414** a través de la cual se pueden recibir mensajes de correo electrónico, por ejemplo, del dispositivo remoto **412**. Los dispositivos de red **408**, pueden incluir dispositivos Ethernet, conectores de teléfono y enlaces de satélite. Será evidente para un experto medio en la técnica que también se pueden utilizar otros dispositivos de red.

Una realización de la invención puede almacenarse completamente como un producto de software en el almacenamiento masivo **407**. Se puede embeber otra realización de la invención en un producto de hardware, por ejemplo, en un circuito impreso, en un procesador de propósito especial o en un dispositivo lógico programable específicamente programado acoplado de forma comunicativa al bus **401**. Otras realizaciones más de la invención pueden ser implementadas parcialmente como un producto de software y parcialmente como un producto de hardware.

Se pueden representar realizaciones de la invención como un producto software almacenado en un medio accesible por máquina (también denominado medio accesible por ordenador o medio accesible por procesador) como se ilustra en la Figura 5. El medio accesible por máquina puede ser cualquier tipo de medio de almacenamiento magnético, óptico o eléctrico que incluye un disquete, CD-ROM, dispositivo de memoria (volátil o no volátil) o mecanismo similar de almacenamiento. El medio accesible por máquina puede contener varios conjuntos de instrucciones, secuencias de código, información de configuración u otros datos. Aquellos expertos medio en la técnica apreciarán que otras instrucciones y operaciones necesarias para implementar la invención descrita también pueden ser almacenadas en el medio accesible por máquina. La Figura 5 ilustra un medio accesible por máquina que incluye instrucciones que cuando son ejecutadas por una máquina hacen que la máquina lleve a cabo operaciones que comprenden determinar una primera ancla semántica **520** que corresponde a un primer grupo de mensajes, por ejemplo, mensajes de correo electrónico legítimos. Determinar una segunda ancla semántica **525** correspondiente a un segundo grupo de mensajes, por ejemplo, mensajes de correo electrónico no solicitados. La primera y segunda anclas semánticas se determinan como se describió con anterioridad utilizando instrucciones que implementan la unidad de entrenamiento de correos electrónicos **505**, instrucciones que implementan la unidad de co-ocurrencias **510** e instrucciones que implementan la unidad de descomposición en valores singulares **515**. El medio accesible por máquina incluye instrucciones adicionales para determinar un vector correspondiente a un mensaje entrante e instrucciones para comparar el vector con al menos uno de la primera ancla semántica **520** y la segunda ancla semántica **525** para obtener al menos un valor de comparación. El vector correspondiente a un mensaje entrante se determina utilizando instrucciones para implementar la unidad de conversión de correos electrónicos **555**. Las instrucciones para comparar el vector **555** con al menos uno de la primera ancla semántica **502** y la segunda ancla semántica **525** para obtener al menos un valor de comparación comprenden instrucciones que implementan la unidad de cálculo **560**. El medio accesible por máquina incluye instrucciones adicionales para filtrar el mensaje entrante en base al valor de comparación. Las instrucciones para filtrar el mensaje de entrada en base al valor de comparación comprenden instrucciones para implementar la unidad lógica **565**. En particular, las instrucciones para filtrar el mensaje entrante comprenden instrucciones para determinar si el mensaje entrante es un correo electrónico no solicitado **570**, un correo electrónico legítimo **575** o un correo electrónico ambiguo **580**.

Experimentos llevados a cabo usando una realización del procedimiento y aparato de la presente invención revelaron que para una base de datos que comprende un mensaje de correo electrónico legítimo N_1 y un mensaje de correo electrónico no solicitado N_2 en el cuerpo de entrenamiento T el sistema de filtrado de correos electrónicos funcionó razonablemente bien. Un incremento exponencial en el rendimiento del sistema de filtrado de correos electrónicos ocurrió cuando los valores de N_1 y N_2 se acercaron a 50. Incrementos subsecuentes en los valores de N_1 y N_2 revelaron una meseta relativa en el rendimiento del sistema de filtrado de correos electrónicos. En una realización, más del 95% de los mensajes de correo electrónico entrantes de un usuario fueron clasificados adecuadamente, con aproximadamente menos de un 3% de los mensajes de correo electrónico del usuario pasados al usuario para su desambiguación. Se observó una tasa de error de clasificación significativamente más baja en comparación con la tasa de error de clasificación de los procedimientos del estado de la técnica.

Mientras que se han ilustrado y descrito los que actualmente se consideran realizaciones de ejemplo de la presente invención, aquellos expertos en la técnica entenderán que pueden hacerse varias otras modificaciones y pueden sustituirse equivalentes, sin salirse del verdadero alcance de la invención. Adicionalmente, se pueden hacer muchas modificaciones para adaptar las enseñanzas de la presente invención a una situación particular sin salirse del

concepto inventivo central descrito en este documento. Por lo tanto, se pretende que la presente invención no esté limitada a las realizaciones particulares divulgadas sino que la invención incluya todas las realizaciones que entran dentro del alcance de las reivindicaciones adjuntas.

REIVINDICACIONES

1. Un procedimiento implementado por ordenador para filtrar mensajes que comprende:

determinar (315) una primera ancla semántica (120) que representa un primer vector en un espacio vectorial semántico correspondiente a un primer grupo de mensajes y una segunda ancla semántica (125) que representa un segundo vector en el espacio vectorial semántico correspondiente a un segundo grupo de mensajes, en el que el primer vector y el segundo vector fueron derivados a partir de un cuerpo de entrenamiento que comprendía el primer grupo de mensajes y el segundo grupo de mensajes, y en el primer grupo de mensajes y el segundo grupo de mensajes son diferentes;

determinar (325), en el espacio vectorial semántico, una representación matemática como un tercer vector correspondiente a un mensaje entrante que tiene texto reconocido por máquina;

comparar (330) la representación matemática correspondiente al mensaje entrante con la primera ancla semántica y la segunda ancla semántica para obtener un primer valor de comparación y un segundo valor de comparación; y

filtrar el mensaje entrante clasificando el mensaje entrante entre el primer y segundo grupos en base al primer valor de comparación y el segundo valor de comparación.

2. Un procedimiento según la reivindicación 1, en el que dicho segundo grupo de mensajes se define como mensajes no solicitados, y dicho primer grupo de mensajes se define como no pertenecientes a mensajes no solicitados, y en el que el primer grupo y el segundo grupo están predefinidos antes de determinar la representación matemática correspondiente al mensaje entrante.

3. Un procedimiento según la reivindicación 2, en el que la primera ancla semántica y la segunda ancla semántica son vectores obtenidos respectivamente a partir de mensajes no solicitados recibidos de un cuerpo de mensajes de entrenamiento y mensajes previamente recibidos como no pertenecientes a mensajes no solicitados del cuerpo de mensajes de entrenamiento.

4. Un procedimiento según la reivindicación 3, en el que el cuerpo de mensajes de entrenamiento se utiliza para obtener una matriz W que comprende un factor de distribución de palabras.

5. Un procedimiento según la reivindicación 4, en el que la matriz W se utiliza para generar la primera ancla semántica y la segunda ancla semántica utilizando descomposición en valores singulares.

6. Un procedimiento según la reivindicación 1, en el que el primer grupo de mensajes, el segundo grupo de mensajes y el mensaje entrante comprenden mensajes de al menos uno de: mensajes de correo electrónico, archivos adjuntos de correo electrónico y programas de ordenador.

7. Un procedimiento según la reivindicación 1, en el que determinar la representación matemática correspondiente comprende utilizar descomposición en valores singulares para generar el tercer vector correspondiente al mensaje entrante.

8. Un procedimiento según la reivindicación 1, en el que comparar la representación matemática correspondiente al mensaje entrante con la primera ancla semántica y la segunda ancla semántica comprende determinar un ángulo ente el tercer vector correspondiente al mensaje entrante y la primera ancla semántica y la segunda ancla semántica.

9. Un procedimiento según la reivindicación 1, en el que comparar la representación matemática correspondiente al mensaje entrante con la primera ancla semántica y la segunda ancla semántica comprende comparar la longitud de una normal entre la primera ancla semántica y el tercer vector correspondiente al mensaje entrante, y la longitud de una normal entre la segunda ancla semántica y el tercer vector correspondiente al mensaje entrante.

10. Un procedimiento según la reivindicación 1, en el que comparar la representación matemática correspondiente al mensaje entrante con la primera ancla semántica y la segunda ancla semántica para obtener un primer valor de comparación y un segundo valor de comparación comprende permitir a un usuario decidir si el mensaje entrante es del primer grupo de mensajes o del segundo grupo de mensajes cuando el primer valor de comparación es sustancialmente igual al segundo valor de comparación.

11. Un procedimiento según la reivindicación 10, en el que filtrar el mensaje entrante en base al primer valor de

comparación y al segundo valor de comparación comprende al menos uno de filtrar automáticamente los mensajes entrantes y marcar los mensajes entrantes.

12. Un procedimiento según la reivindicación 11, en el que marcar el mensaje entrante comprende al menos uno de marcar el mensaje entrante con una primera marca para un mensaje correspondiente al primer grupo de mensajes, marcar el mensaje entrante con una segunda marca para un mensaje correspondiente al segundo grupo de mensajes y marcar el mensaje entrante con una tercera marca cuando el primer valor de comparación es sustancialmente igual al segundo valor de comparación.

13. Un procedimiento según la reivindicación 1, en el que la segunda ancla semántica corresponde a un centroide de los mensajes no solicitados recibidos previamente de un cuerpo de mensajes de entrenamiento definidos como no pertenecientes a mensajes no solicitados del cuerpo de mensajes de entrenamiento en el espacio vectorial semántico.

14. Un procedimiento según la reivindicación 1, en el que cada uno de las primera y segunda anclas semánticas se determinan en base a primeros números de ocurrencias de un conjunto de palabras en el primer grupo y a segundos números de ocurrencias del conjunto de palabras en el segundo grupo.

15. Un procedimiento según la reivindicación 14, en el que dicho determinar la primera ancla semántica y la segunda ancla semántica comprende:

determinar una primera matriz, la matriz comprendiendo:

una primera columna determinada en base a los primeros números de ocurrencias de un conjunto de palabras en el primer grupo; y una segunda columna determinada en base a los segundos números de ocurrencias de un conjunto de palabras en el segundo grupo; y

determinar las primera y segunda anclas semánticas en base a una matriz singular derecha de la descomposición en valores singulares de la primera matriz.

16. Un procedimiento según la reivindicación 15, en el que:

la primera columna se determina en base a las frecuencias de ocurrencias del conjunto de palabras en el primer grupo; y la segunda columna se determina en base a las frecuencias de ocurrencias del conjunto de palabras en el segundo grupo.

17. Un procedimiento según la reivindicación 15, en el que dicho determinar la representación matemática correspondiente al mensaje entrante comprende:

determinar terceros números de ocurrencias del conjunto de palabras en el mensaje entrante; y

determinar la representación matemática correspondiente al mensaje entrante en base al tercer número de ocurrencias del conjunto de palabras en el mensaje entrante y una matriz singular izquierda de la descomposición en valores singulares de la primera matriz.

18. Un medio legible por máquina que tiene almacenado en él un programa de ordenador en donde dicho programa de ordenador comprende medios de código que, cuando se ejecutan en una máquina de procesado de datos hacen que la máquina lleve a cabo cada una de las etapas del procedimiento de la reivindicación 1.

19. Un medio legible por máquina según la reivindicación 18, en el que dicho segundo grupo de mensajes se define como mensajes no solicitados, y dicho primer grupo de mensajes se define como no pertenecientes a mensajes no solicitados, y en donde el primer grupo y el segundo grupo están predefinidos antes de determinar la representación matemática correspondiente al mensaje entrante.

20. Un medio legible por máquina según la reivindicación 19, en el que dichas instrucciones para obtener la primera ancla semántica y la segunda ancla semántica incluyen instrucciones adicionales para obtener vectores que utilizan un cuerpo de mensajes de entrenamiento que comprende mensajes no solicitados recibidos previamente y mensajes definidos como no pertenecientes a mensajes no recibidos previamente.

21. Un medio legible por máquina según la reivindicación 20, en el que dichas instrucciones para obtener vectores utilizando un cuerpo de mensajes de entrenamiento comprende instrucciones adicionales para obtener una matriz W que comprende un factor de distribución de palabras.

22. Un medio legible por máquina según la reivindicación 21, en el que dichas instrucciones para obtener la matriz W comprenden instrucciones adicionales para generar la primera ancla semántica y la segunda ancla semántica usando descomposición en valores singulares.

23. Un medio legible por máquina según la reivindicación 18, en el que dicho primer grupo de mensajes, dicho segundo grupo de mensajes y dicho mensaje entrante comprenden mensajes de al menos uno de mensajes de correo electrónico, archivos adjuntos de correo electrónico y programas de ordenador.

24. Un medio legible por máquina según la reivindicación 18, en el que dichas instrucciones para determinar una representación matemática correspondiente a un mensaje entrante comprende instrucciones adicionales para utilizar descomposición en valores singulares para generar el tercer vector correspondiente al mensaje entrante.

25. Un medio legible por máquina según la reivindicación 18, en el que dichas instrucciones para comparar la representación matemática correspondiente al mensaje entrante con la primera ancla semántica y la segunda ancla semántica comprenden instrucciones adicionales para determinar un ángulo entre le tercer vector correspondiente al mensaje entrante y la primera ancla semántica y la segunda ancla semántica.

26. Un medio legible por máquina según la reivindicación 18, en el que dichas instrucciones para comparar la representación matemática correspondiente al mensaje entrante con la primera ancla semántica y la segunda ancla semántica comprenden instrucciones adicionales para comparar la longitud de una normal entre la primera ancla semántica y el tercer vector correspondiente al mensaje entrante, y la longitud de una normal entre la segunda ancla semántica y el tercer vector correspondiente al mensaje entrante.

27. Un medio legible por máquina según la reivindicación 18, en el que dichas instrucciones para comparar la representación matemática correspondiente al mensaje entrante con la primera ancla semántica y la segunda ancla semántica para obtener un primer valor de comparación y un segundo valor de comparación comprenden instrucciones adicionales para permitir a un usuario decidir si el mensaje entrante es del primer grupo de mensajes o del segundo grupo de mensajes cuando el primer valor de comparación es sustancialmente igual al segundo valor de comparación.

28. Un medio legible por máquina según la reivindicación 27, en el que dichas instrucciones para filtrar el mensaje entrante en base al primer valor de comparación y al segundo valor de comparación comprenden instrucciones adicionales para al menos uno de filtrar automáticamente los mensajes entrantes y marcar los mensajes entrantes.

29. Un medio legible por máquina según la reivindicación 28, en el que dichas instrucciones para marcar el mensaje entrante comprenden instrucciones adicionales para al menos uno de marcar el mensaje entrante con una primera marca para un mensaje correspondiente al primer grupo de mensajes, marcar el mensaje entrante con una segunda marca para un mensaje correspondiente al segundo grupo de mensajes y marcar el mensaje entrante con una tercera marca cuando el primer valor de comparación es sustancialmente igual al segundo valor de comparación.

30. Un sistema de procesado de datos que comprende medios para llevar a cabo cada una de las etapas de la reivindicación 1.

31. Un sistema de procesado de datos según la reivindicación 31, en el que dicho segundo grupo de mensajes se define como mensajes no solicitados, y dicho primer grupo de mensajes se define como no pertenecientes a mensajes no solicitados, y en donde el primer grupo y el segundo grupo están predefinidos antes de determinar la representación matemática correspondiente al mensaje entrante.

32. Un sistema de procesado de datos según la reivindicación 31, en el que la primera ancla semántica y la segunda ancla semántica son vectores obtenidos utilizando un cuerpo de mensajes de entrenamiento que comprende mensajes no solicitados recibidos previamente y mensajes definidos como no pertenecientes a mensajes no recibidos previamente.

33. Un sistema de procesado de datos según la reivindicación 32, en el que el cuerpo de mensajes de entrenamiento se utiliza para obtener una matriz W que comprende un factor de distribución de palabras.

34. Un sistema de procesado de datos según la reivindicación 33, en el que la matriz W se utiliza para generar la primera ancla semántica y la segunda ancla semántica utilizando descomposición en valores singulares.

35. Un sistema de procesado de datos según la reivindicación 30, en el que el primer grupo de mensajes, el segundo grupo de mensajes y el mensaje entrante comprenden mensajes de al menos uno de: mensajes de correo electrónico, archivos adjuntos de correo electrónico y programas de ordenador.

36. Un sistema de procesado de datos según la reivindicación 30, en el que determinar la representación matemática correspondiente comprende utilizar descomposición en valores singulares para generar el tercer vector correspondiente al mensaje entrante.

37. Un sistema de procesado de datos según la reivindicación 30, en el que comparar la representación matemática correspondiente al mensaje entrante con la primera ancla semántica y la segunda ancla semántica comprende determinar un ángulo ente el tercer vector correspondiente al mensaje entrante y la primera ancla semántica y la segunda ancla semántica.

38. Un sistema de procesado de datos según la reivindicación 30, en el que comparar la representación matemática correspondiente al mensaje entrante con la primera ancla semántica y la segunda ancla semántica comprende comparar la longitud de una normal entre la primera ancla semántica y el tercer vector correspondiente al mensaje entrante, y la longitud de una normal entre la segunda ancla semántica y el tercer vector correspondiente al mensaje entrante.

39. Un sistema de procesado de datos según la reivindicación 30, en el que comparar la representación matemática correspondiente al mensaje entrante con la primera ancla semántica y la segunda ancla semántica para obtener un primer valor de comparación y un segundo valor de comparación comprende permitir a un usuario decidir si el mensaje entrante es del primer grupo de mensajes o del segundo grupo de mensajes cuando el primer valor de comparación es sustancialmente igual al segundo valor de comparación.

40. Un sistema de procesado de datos según la reivindicación 39, en el que filtrar el mensaje entrante en base al primer valor de comparación y al segundo valor de comparación comprende al menos uno de filtrar automáticamente los mensajes entrantes y marcar los mensajes entrantes.

41. Un procedimiento según la reivindicación 40, en el que marcar el mensaje entrante comprende al menos uno de marcar el mensaje entrante con una primera marca para un mensaje correspondiente al primer grupo de mensajes, marcar el mensaje entrante con una segunda marca para un mensaje correspondiente al segundo grupo de mensajes y marcar el mensaje entrante con una tercera marca cuando el primer valor de comparación es sustancialmente igual al segundo valor de comparación.

42. Un sistema de ordenador según el sistema de procesado de datos de la reivindicación 30, los medios para llevarlo a cabo comprendiendo:

un bus;

un dispositivo de almacenamiento de datos acoplado a dicho bus;

un procesador acoplado a dicho dispositivo de almacenamiento de datos;

una unidad de descomposición en valores singulares acoplada de forma comunicativa al procesador para determinar la primera ancla semántica correspondiente al primer grupo de mensajes y la segunda ancla semántica correspondiente al segundo grupo de mensajes;

una unidad de conversión de correos electrónicos entrantes acoplada de forma comunicativa a la unidad de descomposición en valores singulares para determinar la representación matemática correspondiente al mensaje entrante;

una unidad lógica acoplada de forma comunicativa a la unidad de conversión de correos electrónicos entrantes y la unidad de descomposición en valores singulares para comparar la representación matemática correspondiente al mensaje entrante con la primera ancla semántica y la segunda ancla semántica para obtener el primer valor de comparación y el segundo valor de comparación y para filtrar el mensaje entrante en base al primer valor de comparación y al segundo valor de comparación.

43. Un sistema de ordenador según la reivindicación 42, en el que dicho segundo grupo de mensajes se define como mensajes no solicitados, y dicho primer grupo de mensajes se define como no perteneciente a mensajes no solicitados, y en donde el el primer grupo y el segundo grupo están predefinidos antes de determinar la representación matemática correspondiente al mensaje entrante.

44. Un sistema de ordenador según la reivindicación 42, en el que la primera ancla semántica y la segunda ancla semántica son vectores obtenidos utilizando un cuerpo de mensajes de entrenamiento que comprende mensajes no solicitados recibidos previamente y mensajes definidos como no pertenecientes a mensajes no recibidos previamente.

45. Un sistema de ordenador según la reivindicación 44, en el que el cuerpo de mensajes de entrenamiento se utiliza para obtener una matriz W que comprende un factor de distribución de palabras.

46. Un sistema de ordenador según la reivindicación 44, en el que la matriz W se utiliza para generar la primera ancla semántica y la segunda ancla semántica utilizando descomposición en valores singulares.

47. Un sistema de ordenador según la reivindicación 42, en el que el primer grupo de mensajes, el segundo grupo de mensajes y el mensaje entrante comprenden mensajes de al menos uno de: mensajes de correo electrónico, archivos adjuntos de correo electrónico y programas de ordenador.

48. Un sistema de ordenador según la reivindicación 42, en el que una unidad de conversión de correos electrónicos entrantes acoplada de forma comunicativa a la unidad de descomposición en valores singulares para determinar la representación matemática correspondiente al mensaje entrante comprende que la unidad de conversión de correos electrónicos entrantes utilice descomposición en valores singulares para generar el tercer vector correspondiente al mensaje entrante.

49. Un sistema de ordenador según la reivindicación 42, en el que la unidad lógica acoplada de forma comunicativa a la unidad de conversión de correos electrónicos entrantes y la unidad de descomposición en valores singulares para comparar la representación matemática correspondiente al mensaje entrante con la primera ancla semántica y la segunda ancla semántica para obtener un primer valor de comparación y un segundo valor de comparación comprende la unidad lógica para determinar un ángulo entre el tercer vector correspondiente al mensaje entrante y la primera ancla semántica y la segunda ancla semántica.

50. Un sistema de ordenador según la reivindicación 42, en el que la unidad lógica acoplada de forma comunicativa a la unidad de conversión de correos electrónicos entrantes y la unidad de descomposición en valores singulares para comparar la representación matemática correspondiente al mensaje entrante con la primera ancla semántica y la segunda ancla semántica para obtener el primer valor de comparación y el segundo valor de comparación comprende la unidad lógica para comparar la longitud de una normal entre la primera ancla semántica y el tercer vector correspondiente al mensaje entrante, y la longitud de una normal entre la segunda ancla semántica y el tercer vector correspondiente al mensaje entrante.

51. Un sistema de ordenador según la reivindicación 42, en el que la unidad lógica acoplada de forma comunicativa a la unidad de conversión de correos electrónicos entrantes y la unidad de descomposición en valores singulares para comparar la representación matemática correspondiente al mensaje entrante con la primera ancla semántica y la segunda ancla semántica para obtener el primer valor de comparación y el segundo valor de comparación comprende la unidad lógica para permitir a un usuario decidir si el mensaje entrante es del primer grupo de mensajes o del segundo grupo de mensajes cuando el primer valor de comparación es sustancialmente igual al segundo valor de comparación.

52. Un sistema de ordenador según la reivindicación 51, en el que la unidad lógica para filtrar el mensaje entrante en base al primer valor de comparación y al segundo valor de comparación comprende la unidad lógica para al menos uno de filtrar los mensajes entrantes y marcar los mensajes entrantes.

53. Un sistema de ordenador según la reivindicación 52, en el que la unidad lógica para marcar el mensaje entrante comprende al menos uno de marcar el mensaje entrante con una primera marca para un mensaje correspondiente al primer grupo de mensajes, la unidad lógica para marcar el mensaje entrante con una segunda marca para un mensaje correspondiente al segundo grupo de mensajes y la unidad lógica para marcar el mensaje entrante con una tercera marca cuando el primer valor de comparación es sustancialmente igual al segundo valor de comparación.

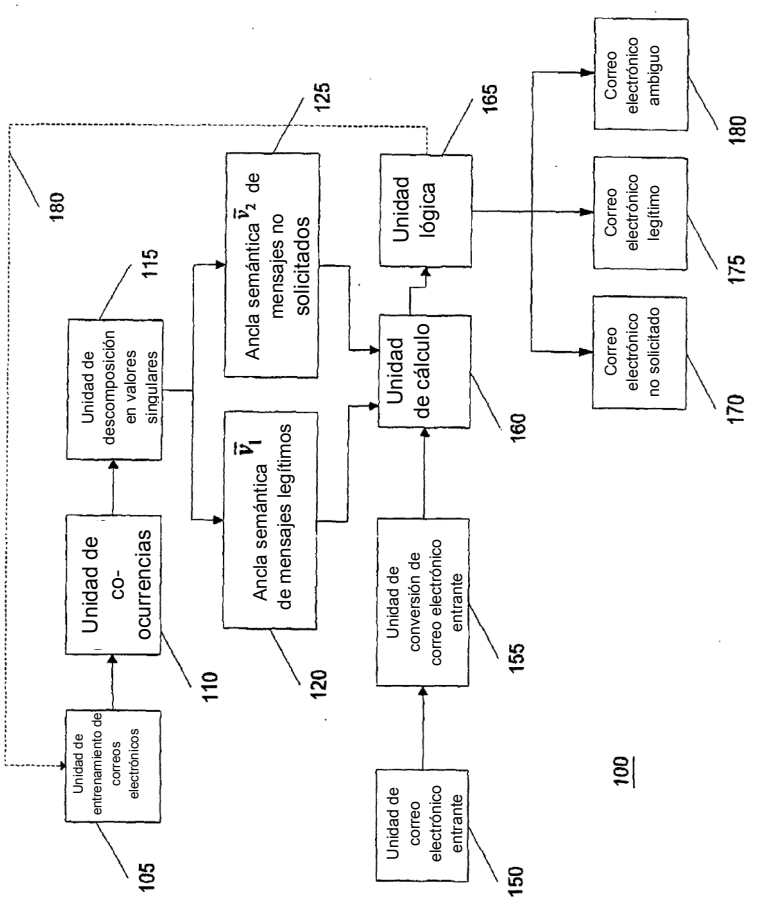
Unidad de entrenamiento de correos electrónicos
105
Unidad de co-ocurrencias
110
Unidad de descomposición en valores singulares
115
Ancla semántica $\tilde{v}_1$ de mensajes legítimos
120
Ancla semántica $\tilde{v}_2$ de mensajes no solicitados
125
180
Unidad de correo electrónico entrante
150
Unidad de conversión de correo electrónico entrante
155
Unidad de cálculo
160
Unidad lógica
165
Correo electrónico no solicitado
170
Correo electrónico legítimo
175
Correo electrónico ambiguo
180
100

Fig. 1

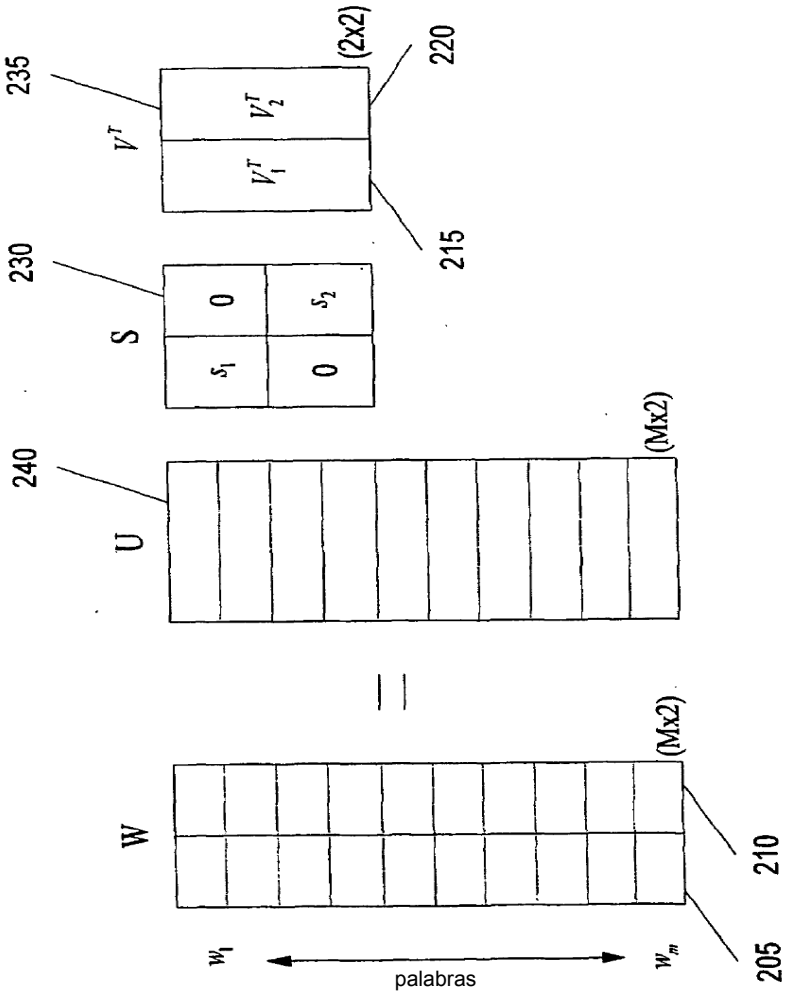
235
V^T
(2x2)
220
V_2^T
V_1^T
215
230
S
0
s_2
s_1
0
240
U
(Mx2)
=
W
(Mx2)
210
205
w_1
palabras
w_m

Fig. 2

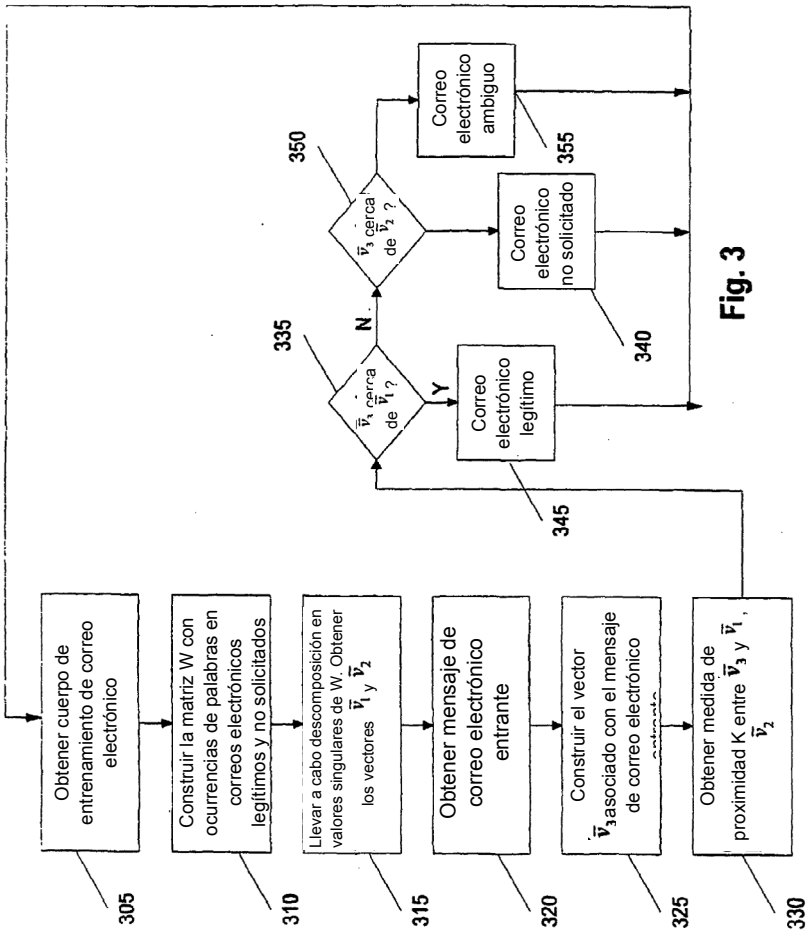
Obtener cuerpo de entrenamiento de correo electrónico
305
Construir la matriz W con ocurrencias de palabras en correos electrónicos legítimos y no solicitados
310
Llevar a cabo descomposición en valores singulares de W. Obtener los vectores $\bar{v}_1$ y $\bar{v}_2$
315
Obtener mensaje de correo electrónico entrante
320
Construir el vector $\bar{v}_3$ asociado con el mensaje de correo electrónico
325
Obtener medida de proximidad K entre $\bar{v}_3$ y $\bar{v}_1$, $\bar{v}_2$
330
$\bar{v}_3$ cerca de $\bar{v}_1$?
335
Y
Correo electrónico legítimo
345
N
$\bar{v}_3$ cerca de $\bar{v}_2$?
350
Correo electrónico no solicitado
340
Correo electrónico ambiguo
355

Fig. 3

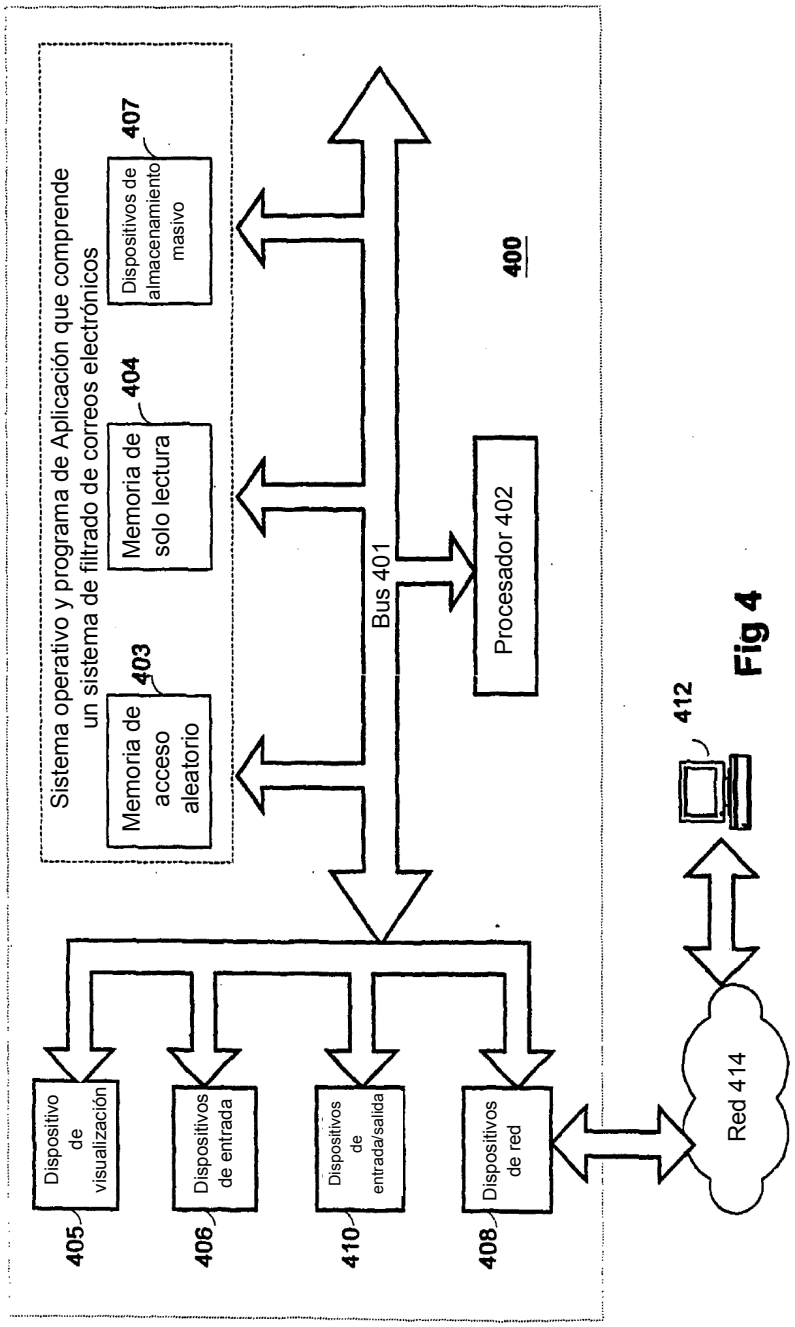
Sistema operativo y programa de Aplicación que comprende un sistema de filtrado de correos electrónicos
Memoria de acceso aleatorio
403
Memoria de solo lectura
404
Dispositivos de almacenamiento masivo
407
Bus 401
Procesador 402
400
Dispositivo de visualización
405
Dispositivos de entrada
406
Dispositivos de entrada/salida
410
Dispositivos de red
408
412
Red 414

Fig 4

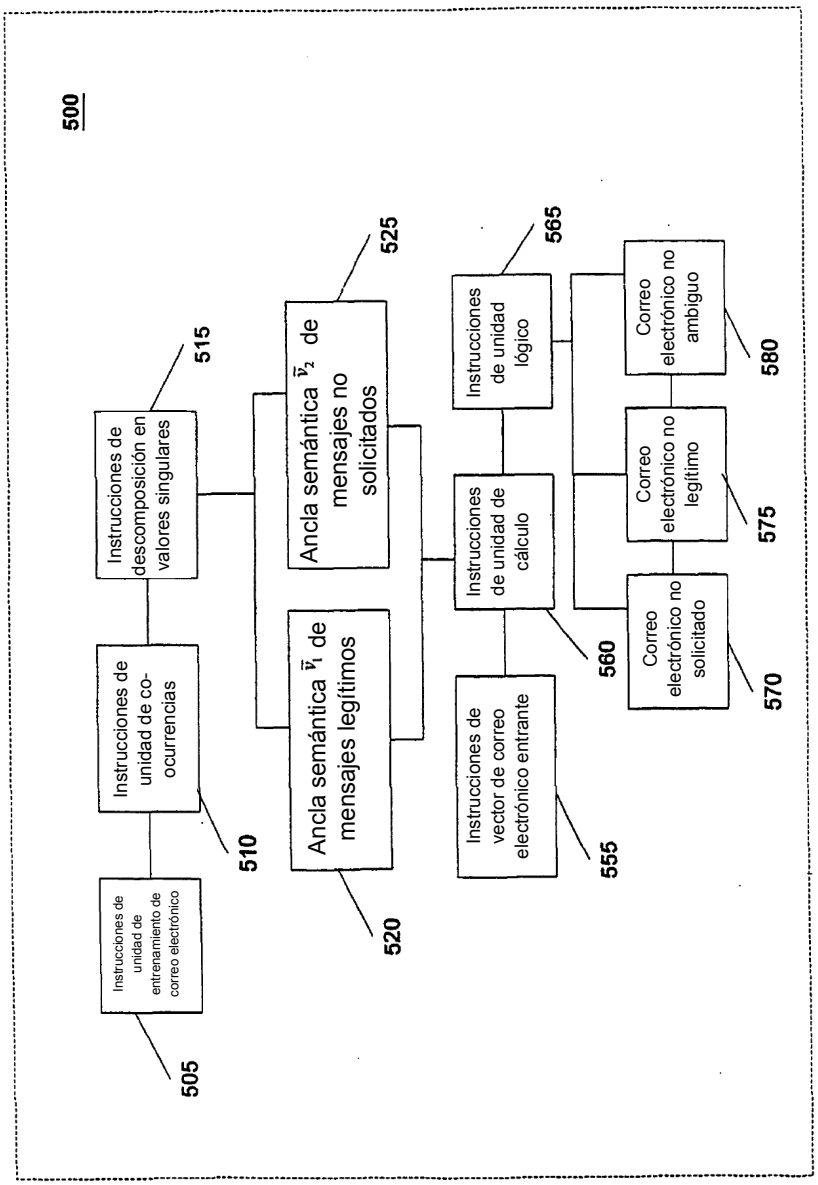
500
Instrucciones de unidad de entrenamiento de correo electrónico
505
Instrucciones de unidad de co-ocurrencias
510
Instrucciones de descomposición en valores singulares
515
Ancla semántica $\bar{v}_1$ de mensajes legítimos
520
Ancla semántica $\bar{v}_2$ de mensajes no solicitados
525
Instrucciones de vector de correo electrónico entrante
555
Instrucciones de unidad de cálculo
560
Instrucciones de unidad lógico
565
Correo electrónico no solicitado
570
Correo electrónico no legítimo
575
Correo electrónico no ambiguo
580

Fig. 5