(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 379 704**

(51) Int. CI.:
***G06F 17/30*** (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Número de solicitud europea: **99935478 .0**
(96) Fecha de presentación: **08.07.1999**
(97) Número de publicación de la solicitud: **1104568**
(97) Fecha de publicación de la solicitud: **06.06.2001**

(54) Título: **Sistema y método para corregir errores ortográficos en peticiones de búsqueda**

(30) Prioridad:
**15.07.1998 US 115662**

(45) Fecha de publicación de la mención BOPI:
**30.04.2012**

(45) Fecha de la publicación del folleto de la patente:
**30.04.2012**

(73) Titular/es:
**AMAZON.COM, INC.**
**SUITE 1200, 1200 12TH AVENUE SOUTH**
**SEATTLE, WA 98144, US**

(72) Inventor/es:
**ORTEGA, Ruben, Ernesto y**
**BOWMAN, Dwayne, Edward**

(74) Agente/Representante:
**Curell Aguilá, Mireia**

## DESCRIPCIÓN

Sistema y método para corregir errores ortográficos en peticiones de búsqueda.

**Apéndice y material bajo derechos de autor**

La presente memoria incluye como apéndice un listado C++ de una función de comparación ortográfica usada para comparar dos cadenas de caracteres. El contenido del apéndice esta sujeto a protección de derechos de autor. El propietario de los derechos de autor no plantea ninguna objeción a la reproducción en facsímil del documento de patente o de partes del mismo, tal como aparece en los archivos o registros de la Oficina de Patentes y Marcas de los Estados Unidos u otra oficina de patentes, aunque en cualquier otro caso se reserva todos los derechos.

**Campo**

La presente invención se refiere a la búsqueda y recuperación de información, y más específicamente, se refiere a métodos para procesar peticiones de búsqueda.

**Antecedentes**

Muchos sitios de la Malla Multimedia Mundial y servicios en línea proporcionan programas de motor de búsqueda ("motores de búsqueda") para ayudar a los usuarios en la localización de artículos de interés de entre un dominio de artículos. Por ejemplo, sitios Webs tales como AltaVista e Infoseek proporcionan motores de búsqueda para ayudar a los usuarios a localizar otros sitios Web, y servicios en línea tales como Lexis y Westlaw implementan motores de búsqueda para ayudar a los usuarios a localizar artículos y dictámenes de tribunales. Adicionalmente, los comercios en línea proporcionan comúnmente motores de búsqueda para ayudar a los clientes a localizar artículos de entre un catálogo en línea.

Para realizar una búsqueda usando un motor de búsqueda, un usuario presenta una petición que contiene uno o más términos de búsqueda. La petición también puede identificar explícita o implícitamente un campo de registro a buscar, tal como el título, el autor o la clasificación por materia del artículo. Por ejemplo, un usuario de un sitio de librería en línea puede presentar una petición que contiene términos que el usuario cree que aparecen en el título de un libro. Un programa del servidor de peticiones del motor de búsqueda procesa la petición para identificar cualesquiera artículos que coinciden con la petición. Al conjunto de artículos identificados por el programa del servidor de peticiones se le hace referencia como "resultado de la petición" y se le presenta comúnmente al usuario en forma de una lista de los artículos localizados. En el ejemplo de la librería, el resultado de la búsqueda sería típicamente el conjunto de títulos de libros que incluyen todos los términos de búsqueda, y comúnmente se le presentaría al usuario en forma de un listado hipertextual de estos artículos.

Cuando el usuario de un motor de búsqueda escribe incorrectamente un término de búsqueda dentro de una petición, por ejemplo porque lo ha tecleado mal o no ha conseguido recordar el término, el término que se ha escrito incorrectamente no coincidirá normalmente con ninguno de los términos de la base de datos que cubre la búsqueda. En este caso, muchos motores de búsqueda simplemente devolverán un resultado de búsqueda nulo (vacío). No obstante, la presentación de resultados de búsqueda nulos a los usuarios puede provocar una frustración significativa de estos últimos. Para reducir este problema, algunos motores de búsqueda ignorar eficazmente el(los) término(s) no coincidente(s) durante la búsqueda. Esta estrategia presenta la desventaja de no conseguir tener en cuenta información potencialmente importante especificada por el usuario y tiende a producir resultados de petición que contienen números relativamente grandes de artículos irrelevantes.

La patente US nº 5.694.592 da a conocer un proceso para medir el grado en el que documentos particulares son relevantes para una petición de búsqueda, tal como una petición de búsqueda de lenguaje natural. Este proceso conlleva el uso de un léxico semántico (esencialmente una correspondencia de palabras con categorías semánticas o significados asociados) para valorar los significados probables de palabras de peticiones específicas. El proceso valora también la importancia de cada palabra de búsqueda basándose en el número de documentos en los que aparece la palabra de petición. Si una palabra de petición dada no aparece en ninguno de los documentos en los que se está realizando la búsqueda, pero aparece en el léxico, el proceso puede sustituirla por un término relacionado seleccionado del léxico. En un ejemplo descrito en la patente US nº 5.694.592, el léxico se usa para sustituir la palabra de petición "partir", que no aparece en ningún documento, por la palabra "salir".

Un artículo titulado "Adapting a Full-text information Retrieval System to the Computer Troubleshooting Domain", publicado en ACM Press, "Online", Julio de 1994, de P.G Anick, describe la sintonización de un sistema de recuperación de información en el dominio de la resolución de problemas informáticos, en donde las peticiones de usuario tienden a ser breves y los términos en lenguaje natural se entremezclan con terminología de una variedad de sublenguajes técnicos. Anick menciona brevemente que se pueden usar datos de ficheros de registro de peticiones para detectar errores de ortografía pero no da a conocer cómo implementar la corrección ortográfica.

**Sumario**

La invención se define en las reivindicaciones.

La forma de realización de la invención afronta los problemas anteriores proporcionando un sistema y un método para corregir términos escritos incorrectamente dentro de peticiones de búsqueda. La forma de realización puede incluir una base de datos para datos de correlación que indican correlaciones entre términos de búsqueda. Los datos de correlación se pueden basar en las frecuencias con las cuales han aparecido históricamente juntos términos de búsqueda específicos dentro de la misma petición, y se pueden generar como, por ejemplo, a partir de un fichero de registro de peticiones. En el ejemplo, cada entrada dentro de la base de datos (implementada en forma de una tabla) comprende una palabra clave y una lista de “términos relacionados”, en donde la lista de términos relacionados está compuesta por los términos que han aparecido en combinación con la palabra clave con el grado más alto de frecuencia.

El método de corrección ortográfica se puede invocar cuando se presenta una petición de búsqueda que incluye por lo menos un término coincidente y por lo menos un término no coincidente. Usando la base de datos de correlación, se puede generar inicialmente una lista de términos que se consideran en relación con el término o términos coincidentes. Esto se puede lograr, por ejemplo, extrayendo la lista de términos relacionados para cada término coincidente, y si la petición incluye múltiples términos coincidentes, combinando estas listas en una única lista de términos relacionados.

Los términos relacionados se pueden comparar en cuanto a ortografía con el(los) término(s) no coincidente(s) para identificar cualesquiera sustituciones adecuadas. Las comparaciones ortográficas se pueden realizar usando una función de comparación ortográfica de tipo anagramático, que genera una puntuación que indica el grado de similitud entre dos cadenas de caracteres. En un ejemplo, si se encuentra un término relacionado, con una ortografía suficientemente similar a un término no coincidente, el término no coincidente se puede sustituir automáticamente por el término relacionado. Alternativamente, al usuario se le puede solicitar que seleccione el(los) término(s) de sustitución de entre una lista de términos. Una vez que se han sustituido el término o términos no coincidentes, la petición modificada se puede usar para realizar la búsqueda. Al usuario también se le puede notificar sobre la(s) modificación(es) realizada(s) en la petición.

Una ventaja importante del método de corrección ortográfica antes descrito con respecto a métodos convencionales de corrección ortográfica es que resulta significativamente más probable que los términos de sustitución seleccionados sean los términos que fueron introducidos por el usuario. Esta ventaja se puede potenciar adicionalmente a través del uso de datos de correlación de términos de búsqueda, y particularmente datos de correlación que reflejan las sumisiones de peticiones históricas. El método incrementa la probabilidad de que los resultados de la petición contengan artículos que sean de interés para el usuario. Otra ventaja es que el método resulta muy adecuado para corregir términos que no aparecen en el diccionario, tales como nombre propios de autores y artistas y términos peculiares dentro de títulos y nombres de productos.

En la forma de realización, los datos de correlación se pueden generar de tal manera que reflejen notablemente sumisiones de peticiones recientes, y que reflejen por lo tanto fuertemente las preferencias actuales de los usuarios. Esto se puede lograr, por ejemplo, generando periódicamente una tabla de correlación a partir de un número deseado (por ejemplo, 12) de los ficheros de registro de peticiones diarios más recientes. El uso de datos de correlación que reflejan notablemente sumisiones de peticiones recientes hace que aumente además la probabilidad de que las sustituciones realizadas por el proceso de corrección ortográfica sean aquellas deseadas por los usuarios.

**Breve descripción de los dibujos**

A continuación, se describirán estas y otras características de la invención en referencia a los dibujos que se resumen seguidamente. Estos dibujos y la descripción asociada se proporcionan para ilustrar la forma de realización de la invención, y no para limitar el alcance de la misma.

La Figura 1 ilustra un sitio Web que implementa un motor de búsqueda de acuerdo con la invención.

La Figura 2 ilustra una página de búsqueda de libros del sitio Web.

La Figura 3 ilustra el formato general de la tabla de correlación de la Figura 1.

La Figura 4 ilustra las etapas realizadas por el servidor de peticiones para procesar una sumisión de una petición.

La Figura 5 ilustra el formato general del fichero de registro de peticiones de la Figura 1.

La Figura 6 ilustra un proceso usado para generar periódicamente la tabla de correlación a partir del fichero

de registro de peticiones.

**Descripción detallada de las formas de realización preferidas**

La presente invención proporciona un método para corregir errores ortográficos en peticiones que son presentadas a motores de búsqueda. Brevemente, el método conlleva el uso de datos de correlación de términos de búsqueda para identificar términos de búsqueda que están relacionados con el(los) término(s) de búsqueda escrito(s) correctamente de la petición, y la evaluación de si cualquiera de estos términos relacionados presenta una escritura similar al(a los) término(s) de búsqueda escrito(s) incorrectamente. Los datos de correlación de términos de búsqueda se basan preferentemente en sumisiones de peticiones históricas, y más específicamente, en las frecuencias con las cuales han aparecido previamente términos de búsqueda juntos dentro de la misma petición. El método se puede implementar dentro de cualquiera de una variedad de tipos diferentes de motores de búsqueda, incluyendo, por ejemplo, motores de búsqueda de Internet, motores de búsqueda de investigaciones legales, y motores de búsqueda proporcionados por comercios en línea.

Con fines ilustrativos, el método se describe en la presente en el contexto de un motor de búsqueda que se usa para ayudar a clientes de Amazon.com Inc. a localizar artículos (libros, CDs, etcétera) de entre un catálogo en línea de productos. Durante toda la descripción, se hará referencia a varios detalles específicos de la implementación de Amazon.com. Estos detalles se proporcionan con el fin de ilustrar en su totalidad la forma de realización de la invención, y no limitan el alcance de la misma. El alcance de la invención se expone en las reivindicaciones adjuntas.

**I. Visión general del sitio Web y el motor de búsqueda**

La Figura 1 ilustra el sitio Web de Amazon.com 30, que incluye componentes usados para implementar un motor de búsqueda de acuerdo con la invención. Tal como es bien sabido en la técnica del comercio por Internet, el sitio Web Amazon.com incluye una funcionalidad para permitir que los usuarios busquen, exploren, y realicen compras en un catálogo en línea de títulos de libros, títulos de piezas musicales, y otros tipos de artículos. Puesto que el catálogo contiene millones de artículos explorables, es importante que el sitio proporcione un mecanismo eficaz para ayudar a los usuarios a localizar artículos.

Tal como se ilustra por medio de la Figura 1, el sitio Web 30 incluye una aplicación de servidor Web 32 (“servidor Web”) que procesa solicitudes recibidas a través de Internet desde ordenadores de usuarios 34. Estas solicitudes incluyen peticiones de búsqueda que son presentadas por los usuarios para buscar en el catálogo productos. El servidor Web 32 graba transacciones de usuarios, incluyendo sumisiones de peticiones, dentro de un fichero de registro de peticiones 36. En la forma de realización representada en la Figura 1, el fichero de registro de peticiones 36 consta de una secuencia de archivos de ficheros de registro de peticiones diarios 36, cada uno de los cuales representa un día de transacciones.

El sitio Web 30 incluye también el servidor de peticiones 38 que procesa las peticiones de búsqueda mediante búsqueda en una base de datos bibliográfica 40. La base de datos bibliográfica 40 incluye información sobre los diversos artículos que están disponibles para su compra en el sitio. Esta información incluye, por ejemplo, los títulos, autores, editores, descripciones de materias e ISBNs (Números Normalizados Internacionales de Libros) de títulos de libros, y los títulos, artistas, discográficas, y clasificaciones musicales de títulos de piezas musicales. La información para cada artículo se dispone dentro de campos (tales como un campo de “autor” y un campo de “título”), permitiendo buscar en la base de datos 40 de una manera acotada por campos. El sitio incluye también una base de datos 41 de contenido HTML (Lenguaje de Marcado de Hipertexto) que incluye, entre otras cosas, páginas de información del producto que muestran y describen los diversos artículos del catálogo.

La Figura 2 ilustra el formato general de una página de búsqueda de libros que se puede usar para buscar títulos de libros en la base de datos bibliográfica 40. La página incluye campos de autor, título y materia 42, 43, 44 y controles asociados que permiten que el usuario inicie búsquedas de títulos de libros acotadas por el campo. Los usuarios también pueden acceder a una página de búsqueda musical (no mostrada) para buscar títulos de piezas musicales usando los campos de artista, título y discográfica. Otras áreas del sitio permiten que el usuario presente peticiones de búsqueda sin limitar los términos de búsqueda a campos de la base de datos específicos.

Cuando el usuario presenta una petición de búsqueda desde la página de búsqueda de libros u otra página del sitio, el servidor de peticiones 38 (Figura 1) aplica la petición a la base de datos bibliográfica 40, teniendo en cuenta cualquier acotación por campo dentro de la petición. Si el resultado de la petición es un único artículo, se le presenta al usuario la página de información de producto del artículo. Si el resultado de la petición incluye múltiples artículos, se le presenta al usuario una lista de los artículos a través de una o más páginas de resultados de búsqueda (no mostradas) que incluyen enlaces hipertextuales a las páginas de información de producto respectivas de los artículos.

Para peticiones de múltiples términos, el servidor de peticiones 38 aplica eficazmente una operación AND lógica a los términos de búsqueda entre sí para ejecutar la misma. Por ejemplo, si el usuario introduce los términos “Java” y

"programación" en el campo de título, el servidor de peticiones 38 buscará y devolverá un listado de todos los artículos que presentan ambos términos mencionados dentro del título. De este modo, si cualquier término de búsqueda de la petición no produce una coincidencia (al cual se hace referencia en la presente como "término no coincidente"), la petición producirá un resultado de petición nulo. En este caso, al usuario se le puede presentar un listado de artículos que se le consideran como "casi coincidencias".

Aunque el motor de búsqueda descrito en la presente aplica una AND lógica a los términos de búsqueda entre sí, se reconocerá que la invención se aplicará a motores de búsqueda que utilicen otros métodos para combinar términos de búsqueda.

**II. Visión general de la característica de corrección ortográfica**

De acuerdo con la invención, cuando un usuario presenta una petición de búsqueda de múltiples términos que incluye términos tanto coincidentes como no coincidentes, un proceso de corrección ortográfica 48 (Figura 1) del servidor de peticiones 38 usa el(los) término(s) coincidente(s), en combinación con datos de correlación de términos de búsqueda, para intentar corregir la escritura del(de los) término(s) no coincidente(s). Esto se logra preferentemente usando una tabla de correlación de términos de búsqueda 50 para identificar términos adicionales que se considera que están en relación con el(los) término(s) de búsqueda coincidente(s), y a continuación comparando las escrituras de dichos términos relacionados con la(s) ortografía(s) del(de los) término(s) no coincidente(s). Por ejemplo, si un usuario presenta la petición

Java APPI,

y "APPI" es un término no coincidente, el proceso de corrección de escritura 48 usa la tabla de correlación para identificar una lista de términos que se consideran en relación con "Java". Esta lista podría ser la siguiente: café, programación, API, gestión, lenguaje. Al comparar ortografía de estos términos relacionados, con la escritura del término no coincidente, el proceso de corrección ortográfica 48 identificará "API" como término de sustitución candidato. Las comparaciones de escritura se realizan preferentemente usando una función de comparación ortográfica de tipo anagramático que genera una puntuación que indica el grado de similitud entre dos cadenas de caracteres.

Los datos contenidos dentro de la tabla de correlación 50 indican relaciones entre términos de búsqueda, y se usan para predecir de manera eficaz términos de búsqueda que es probable que aparezcan dentro de la misma petición. La incorporación de dichas predicciones en el proceso de corrección ortográfica hace que aumente significativamente la probabilidad de que un término de sustitución dado identificado por el proceso sea el término que era deseado por el usuario.

Si el proceso anterior identifica una sustitución adecuada para un término no coincidente dado, el término no coincidente se sustituye automáticamente por el término relacionado. Si no se halla ninguna sustitución adecuada para un término no coincidente dado, el término no coincidente se elimina preferentemente de la petición. Una vez que todos los términos no coincidentes o bien se han sustituido o bien se han eliminado, la petición modificada se usa para buscar en la base de datos bibliográfica 40 y el resultado se le presenta al usuario. Al usuario se le notifica(s) también la(s) modificación(es) realizada(s) en la petición de búsqueda.

Como alternativa a la sustitución automática del(de los) término(s) no coincidente(s), se puede solicitar al usuario que seleccione el(los) término(s) de sustitución de entre una lista de términos de sustitución candidatos. Esto se logra preferentemente presentando cada término de sustitución candidato en forma de un hiperenlace respectivo (no mostrado) que puede ser seleccionado por el usuario para iniciar la búsqueda modificada; de este modo, el usuario puede tanto seleccionar una petición modificada como iniciar una búsqueda nueva con una única acción.

De acuerdo con otro aspecto de la invención, la tabla de correlación 50 preferentemente contiene o refleja información histórica sobre las frecuencias con las cuales han aparecido juntos términos de búsqueda específicos dentro de la misma petición de búsqueda. Tal como se representa en la Figura 1, estos datos se extraen preferentemente del fichero de registro de peticiones 36 usando un proceso de generación de tablas 46. La incorporación de dicha información histórica en el proceso de corrección ortográfica hace que aumente además la probabilidad de que un término de sustitución localizado por el proceso sea el término que era deseado por el usuario.

En el ejemplo (descrito posteriormente), el proceso de generación de tablas 46 regenera la tabla de correlación 50 de una forma diaria a partir de los M archivos de ficheros de registro diarios más recientes 36(1)-36(M), en donde M es un número fijo tal como diez o veinte. Este planteamiento de "ventana deslizante" produce de manera ventajosa una tabla de correlación que se basa únicamente en sumisiones de peticiones recientes, y que por lo tanto refleja las preferencias actuales de los usuarios. Por ejemplo, si un número relativamente grande de usuarios ha buscado el libro *Into Thin Air* de Jon Krakauer durante la última semana, es probable que las correlaciones entre los términos "into", "thin", "air", y "Krakauer" sean de forma correspondiente elevadas; estas correlaciones elevadas a su vez harán que aumente la probabilidad de que un usuario que escribe incorrectamente un término (de una petición de

múltiples términos) mientras está buscando este libro sea dirigido al mismo. En el fichero de registro se puede aplicar cualquiera de una variedad de otros tipos de métodos de encauzamiento para lograr un resultado similar, incluyendo métodos que aplican un peso mayor a sumisiones de peticiones recientes que a sumisiones de peticiones antiguas.

La Figura 3 ilustra el formato general de la tabla de correlación 50. En el ejemplo representado en la Figura 3 y descrito de forma detallada en la presente, las correlaciones entre términos de búsqueda se basan meramente en la frecuencia de aparición dentro de la misma petición. Tal como se ha descrito anteriormente, se pueden usar de forma adicional o alternativa otros tipos de correlaciones de términos de búsqueda. Adicionalmente, aunque la implementación dada a conocer usa una tabla para almacenar las tablas de correlación, se pueden usar otros tipos de bases de datos.

Tal como se ilustra por medio de la Figura 3, cada entrada dentro de la tabla de correlación (se muestran 3 entradas) tiene preferentemente dos componentes principales: (1) una palabra clave 60, y (2) una lista de "términos relacionados" 62 para esa palabra clave. La lista de términos relacionados 62 es una lista de los N (por ejemplo, 20) términos de búsqueda que han aparecido dentro de la misma petición que la palabra clave, con el grado más alto de frecuencia, y está ordenada de acuerdo con la frecuencia. Por ejemplo, la entrada para la palabra clave COSMOS es:

COSMOS: ASTRONOMÍA, SAGAN, UNIVERSO, ESPACIO, CARL…

lo cual indica que ASTRONOMÍA ha aparecido junto con COSMOS con el grado más alto de frecuencia; SAGAN ha aparecido con COSMOS con el segundo grado más alto de frecuencia, y así sucesivamente. Cada término que aparece dentro de la parte de lista 62 se considera en relación con la palabra clave correspondiente 60 en virtud de la frecuencia relativamente alta con la que han aparecido los términos dentro de la misma petición. En la implementación descrita en el presente documento, las palabras claves y los términos relacionados se almacenan en la tabla sin tener en cuenta el orden alfabético, aunque alternativamente se puede conservar información sobre el orden alfabético.

En referencia adicionalmente a la Figura 3, cada término relacionado y cada palabra clave 60 dentro de la tabla 50 incluye preferentemente un prefijo de campo de un solo carácter (no mostrado) que indica el campo de búsqueda 42, 43, 44 con el cual se corresponde el término (basándose en los campos de búsqueda en los cuales fueron introducidos los términos por usuarios). Estos prefijos pueden ser, por ejemplo, los siguientes: A - autor, T = título, M - materia, R - artista, D - discográfica. De este modo, por ejemplo, si la palabra clave COSMOS en la Figura 3 tiene el prefijo "T" y el término relacionado SAGAN tiene el prefijo "A", esto indicaría que se presentó un número relativamente grande de peticiones que incluían COSMOS en el campo de título 43 junto con SAGAN en el campo de autor 42. Tal como se describe posteriormente, los prefijos de los términos relacionados son usados por el proceso de corrección ortográfica 48 para filtrar de manera eficaz los términos no correspondientes a campos de entre las listas de términos relacionados, de manera que un término no coincidente dentro de un campo de búsqueda dado se comparará únicamente con términos relacionados del mismo campo. De este modo, por ejemplo, un término no coincidente dentro del campo de autor 42 se comparará únicamente con otros términos que hayan sido introducidos históricamente en el campo de autor 42. El proceso de corrección ortográfica 48 usa de modo similar los prefijos de las palabras claves 60 para seleccionar entradas de la tabla que se corresponden con los campos de búsqueda respectivos de los términos coincidentes. Por ejemplo, si se recibe una petición errónea que incluye el término coincidente MONTAÑA dentro del campo de título 43, el proceso de corrección ortográfica 48 buscará una entrada de la tabla que tenga la palabra clave TMONTAÑA.

Tal como se representa adicionalmente en la Figura 3, la tabla de correlación 50 incluye también preferentemente puntuaciones de correlación 64 que indican el número de veces que ha aparecido cada término relacionado en combinación con la palabra clave. Por ejemplo, el término PROGRAMACIÓN tiene una puntuación de 320 en la entrada correspondiente a JAVA, lo cual indica que JAVA y PROGRAMACIÓN aparecieron dentro de la misma petición (dentro de los campos indicados por sus prefijos de campo respectivos) trescientas veinte veces. Tal como se describe posterior, el proceso de generación de tablas 46 ignora peticiones que produjeron un resultado de petición nulo, y por lo tanto las mismas no se reflejan en las puntuaciones de correlación 64. En otras implementaciones, las puntuaciones de correlación 64 pueden incorporar otros tipos de correlaciones. Tal como se describe posteriormente, las puntuaciones 64 se usan preferentemente para fusionar listas de términos relacionados cuando una petición tiene múltiples términos coincidentes.

Durante el funcionamiento, cuando el servidor de peticiones 38 determina que una petición contiene un término de búsqueda tanto coincidente como no coincidente, el proceso de corrección ortográfica 48 accede inicialmente a la tabla de correlación 50 para extraer la lista de términos relacionados asociadas 62. Si la petición incluye múltiples términos de búsqueda coincidentes, el proceso 48 obtiene la lista de términos relacionados 62 para cada término coincidente y fusiona estas listas entre sí (tal como se describe posteriormente) para generar una lista de términos relacionadas compuestas. Tal como se ha indicado anteriormente, a continuación el proceso compara las escrituras de los términos relacionados resultantes con la(s) escritura(s) del(de los) término(s) no coincidente(s), y el bien sustituye o bien elimina cada término no coincidente de la petición. Durante el proceso de comparación ortográfica,

cada término no coincidente se compara de forma preferente únicamente con aquellos términos relacionados que caen dentro del campo de búsqueda del término no coincidente.

Una ventaja importante de este método es que el mismo resulta particularmente muy adecuado para corregir escrituras incorrectas de términos que no aparecen en el diccionario. La técnica resulta por lo tanto particularmente útil para identificar artículos que tienden a caracterizarse por medio de términos que no son del diccionario. Dichos artículos incluyen, por ejemplo, productos vendidos por comercios en línea, opiniones de tribunales (identificadas comúnmente por los nombres de las partes), y negocios y su sitio Web. En el contexto de una tienda de libros/musical en línea, por ejemplo, el método es útil para corregir escrituras incorrectas de nombres propios de autores/artistas y términos peculiares que aparezcan dentro de los títulos. Por ejemplo, un usuario que busque un libro de Jon Krakauer podría encontrar el libro tecleando unas pocas palabras del título más una versión escrita incorrectamente de “Krakauer”. La probabilidad de que la búsqueda identifique el libro deseado será particularmente alta si un número relativamente grande de usuarios ha buscado recientemente el mismo libro.

Aunque el método usa preferentemente correlaciones de términos de búsqueda que se basan en sumisiones de peticiones anteriores, debería entenderse que los datos de correlación se pueden generar alternativamente a partir de otras fuentes. Por ejemplo, los datos de correlación se pueden generar procesando la información en la base de datos bibliográfica 40 para identificar términos que aparecen juntos dentro del mismo registro de artículo, título, descripción de artículo, revisión de libro u otro campo de la base de datos; al realizar este proceso, a cada artículo se le puede conceder un peso que sea proporcional, por ejemplo, al número de unidades de ese artículo vendido durante esa semana.

Adicionalmente, el método se puede modificar para incorporar otros tipos de correlaciones, incluyendo correlaciones basadas en otros tipos de acciones históricas del usuario. Por ejemplo, en la extracción de datos de correlación del fichero de registro de peticiones 36, se puede asignar un peso mayor a sumisiones de peticiones que, basándose en las acciones posteriores de los usuarios, se puede considerar que han producido un resultado de petición satisfactorio. Dicho resultado satisfactorio se puede presuponer, por ejemplo, si el usuario visiona, compra o añade a un carrito de la compra un artículo localizado por la búsqueda.

En referencia adicionalmente a la Figura 1, el servidor Web 32, el servidor de peticiones 38, el proceso de generación de tablas 46, y el software de la base de datos se ejecutan en uno o más servidores y estaciones de trabajo basados en Unix (no mostrados) del sitio Web 30. La tabla de correlación 50 se almacena en una RAM (memoria de acceso aleatorio) en la misma estación de trabajo que la usada para implementar el servidor de peticiones 38.

**III. Método de procesado de las consultas**

A continuación se describirá de forma más detallada el proceso de corrección ortográfica en referencia a la Figura 4, que es un diagrama de flujo de las etapas realizadas por el servidor de peticiones 38 (Figura 1) cuando un usuario presenta una petición. Para ilustrar este proceso, se supondrá que el usuario está buscando libros sobre senderismo (*hiking*) por la Ruta de los Apalaches (*Appalachian Trail*) y ha tecleado la siguiente petición en el campo de materia 44 (Figura 2): “*hike Appalatian trail*” (“senderismo ruta Appalaches”). Se supondrá también que “*Appalatian*” (“Appalaches”) es un término no coincidente (aunque “*Appalachian*” (“Apalaches”) es coincidente) y que “*hike*” (“senderismo”) y “*trail*” (“ruta”) son términos coincidentes que tienen las siguientes listas de términos relacionados:

HIKE: CAMPING (235), WALKS (160), TRAIL (150)
TRAIL: BIKE (200), APPALACHIAN (165), WALKS (50)

Se supondrá también que la totalidad de las palabras clave y términos relacionados anteriores tienen un prefijo de campo de “M” de “materia”.

Tal como se indica en la etapa 70, el servidor de peticiones 38 inicialmente aplica la petición a la base de datos bibliográfica 40. Tal como se representa mediante las etapas 72 y 74, si se hallan uno o más artículos, el servidor de peticiones devuelve una lista de estos artículos al servidor Web 32. El servidor Web 32 a su vez incorpora esta lista a una o más páginas de resultados de búsqueda, o, si solamente se localiza un artículo, devuelve la página de información de producto correspondiente a este artículo. Adicionalmente, el servidor Web registra el número de artículos hallados en el fichero de registro de peticiones 36 (véase Figura 5). En el presente ejemplo, no se hallaría ningún artículo puesto que “*Appalatian*” (“Appalaches”) no existe como palabra de materia en la base de datos bibliográfica 40.

Si, tal como en el presente ejemplo, el número de artículos hallados es cero en la etapa 72, el servidor de peticiones 38 determina si la petición incluye términos tanto coincidentes como no coincidentes (etapa 76). Si los incluye, el servidor de peticiones 38 invoca su proceso de corrección ortográfica 48 (etapas 80 a 90 de la Figura 4) para intentar corregir el(los) término(s) no coincidente(s). Si no los incluye, se devuelve un mensaje al usuario (etapa 78) indicando que no se hallaron coincidencias exactas; en este caso, el servidor de peticiones 38 también puede generar y devolver una lista de “casi coincidencias”, que puede incluir artículos que contienen solamente un

subconjunto de los términos coincidentes. El servidor de peticiones 38 se podría configurar de manera adicional o alternativa para invocar un método de corrección ortográfica alternativo (no mostrado) con el fin de intentar corregir cualquier(cualesquiera) término(s) no coincidente(s). En el presente ejemplo, se invocaría el proceso de corrección ortográfica 48 puesto que la petición incluye términos tanto coincidentes como no coincidentes.

El proceso de corrección ortográfica comienza en la etapa 80 mediante la recuperación de la lista de términos relacionados para cada término coincidente a partir de la tabla de correlación 50. Si no se halla ninguna lista (entrada de tabla) de términos relacionados de la etapa 80, el proceso preferentemente devuelve un mensaje de resultado de petición nulo y finaliza (no se muestra). Alternativamente, el proceso podría usar un método alternativo de corrección ortográfica para intentar corregir el(los) término(s) no coincidente(s), o podría generar y devolver una lista de “casi coincidencias”.

Si, tal como en el presente ejemplo, la petición incluye múltiples términos coincidentes, las listas de términos relacionados para estos términos se fusionan entre sí en la etapa 80 para formar una lista compuesta de términos relacionados. Las listas se pueden fusionar, por ejemplo, combinando las listas mientras se suman las puntuaciones de correlación de cualesquiera términos cruzados (términos que aparecen en múltiples listas), y a continuación clasificando la lista compuesta según el orden de la puntuación de frecuencia de masa alta-a-más baja. En el presente ejemplo, este método produciría la siguiente lista de términos relacionados: CAMPING (235), WALKS (210), BIKE (200), APPALACHIAN (165), TRAIL (150). La lista compuesta clasificada se puede truncar para reducir la carga de procesado de las etapas posteriores.

A continuación el proceso de corrección ortográfica entra en un bucle (etapas 84 a 90) en el que se comparan las escrituras del(de los) término(s) no coincidente(s) y los términos relacionados. En cada pasada de este bucle, el proceso compara un término no coincidente con la lista de términos relacionados uno a uno (etapa 84), excluyendo cualesquiera términos relacionados que tengan prefijos de campo que no se correspondan con el campo de búsqueda del término no coincidente. Las comparaciones se realizan usando una función de tipo anagramático que compara dos cadenas de caracteres y devuelve una puntuación de similitud numérica. La puntuación de similitud indica el grado de similitud entre las escrituras de las dos cadenas, e indica por lo tanto la probabilidad de que el término relacionado dado sea una sustitución adecuada para el término no coincidente actual. Una puntuación de similitud de cero indica el grado más alto de similitud, y las puntuaciones progresivamente mayores indican grados de similitud progresivamente inferiores.

Las etapas realizadas por la función de comparación ortográfica para comparar CADENA1 con CADENA2 se exponen a continuación, en las cuales la variable RESULTADO representa la puntuación que devuelve la función. Se adjunta como apéndice un listado de código de una implementación en C++ del método (implementado como dos funciones independientes).

Etapa 1: ordenar CADENA1 y CADENA2 en orden alfabético.

Etapa 2: inicializar PUNTERO1 al primer carácter de CADENA1 ordenada; PUNTERO2 al primer carácter de CADENA2 ordenada y RESULTADO a cero.

Etapa 3: comparar caracteres respectivos a los que apunta PUNTERO1 y PUNTERO2 sin tener en cuenta mayúsculas y minúsculas. Si son iguales, hacer avanzar ambos punteros al siguiente carácter de la cadena; si no, hacer avanzar el puntero que apunta al carácter alfabéticamente inferior e incrementar RESULTADO en 1.

Etapa 4: repetir la etapa 3 hasta que un puntero se hace avanzar más allá del último carácter de su cadena respectiva.

Etapa 5: si uno de los punteros está apuntando todavía a un carácter, sumar a RESULTADO el número de caracteres (incluyendo el carácter al que se apunta) que quedan en esa cadena.

En el presente ejemplo, la etapa 84 da como resultado que el término no coincidente APPALATIAN se compare con los términos CAMPING, WALKS, BIKE, APPALACHIAN, y TRAIL. Las puntuaciones generadas por estas comparaciones se enumeran en la siguiente tabla.

TABLA

| TÉRMINO RELACIONADO | CADENA ORDENADA COMPARADA CON “AAAAILNPPT” | PUNTUACIÓN DE SIMILITUD |
|---|---|---|
| CAMPING | ACGIMNP | 9 |
| WALKS | AKLSW | 11 |
| BIKE | BEIK | 12 |
| APPALACHIAN | AAAACHILNPP | 3 |
| TRAIL | AILRT | 7 |

En la etapa 86, se evalúan las puntuaciones para determinar si cualquiera de los términos relacionados presenta una escritura suficientemente similar como para ser un término de sustitución candidato. En la forma de realización preferida, se considera que un término relacionado es similar (y por lo tanto una sustitución candidata) si su puntuación es menor que o igual a la mitad de la longitud del término no coincidente ("umbral de similitud"). En el presente ejemplo, el término "Appalachian" pasa la prueba de similitud (puesto que 3 5) y los restantes términos relacionados no.

Para incrementar la eficacia de la función de comparación ortográfica, el proceso en la etapa 3 se puede detener una vez que RESULTADO supera el umbral de similitud. La eficacia también se podría incrementar deteniendo el procesado de términos adicionales una vez que se ha identificado un término con una puntuación suficientemente baja (por ejemplo, 0).

Si por lo menos un término relacionado pasa la prueba de similitud en la etapa 86, el término no coincidente se sustituye por el término relacionado que presenta la puntuación más baja (etapa 88). Si múltiples términos relacionados comparten la puntuación más baja en la etapa 88, el término que se sitúa primero en la lista de términos relacionados (y que por lo tanto presenta la puntuación de frecuencia más alta) se usa como sustitución.

Si ninguno de los términos relacionados pasa la prueba de similitud en la etapa 86, el término no coincidente se elimina de la petición. Alternativamente, en este momento se le podría devolver al usuario un mensaje de resultado de petición nulo. Una vez que todos los términos no coincidentes o bien se han sustituido o bien se han eliminado, se vuelve a intentar la búsqueda usando la petición modificada y el resultado se devuelve al usuario (etapa 94).

Si la búsqueda que se ha intentado de nuevo en la etapa 94 produce uno o más artículos, los artículos se presentan junto con un mensaje que indica la(s) modificación(es) realizada(s) en la petición. En el presente ejemplo, este mensaje podría decir lo siguiente:

> *No se halló ningún libro que incluyera la palabra de materia "Appalatian". Sin embargo, se volvió a intentar su búsqueda usando el término "Appalachian" en lugar de "Appalatian" y se encontraron los siguientes títulos:*

El resultado de la página de búsqueda preferentemente también presenta y permite que el usuario edite la petición modificada de manera que el usuario puede rechazar eficazmente la(s) sustitución(es) del término de búsqueda y/o alternativamente revisar la petición.

Si la búsqueda que se ha intentado nuevamente en la etapa 94 no produce ninguna coincidencia exacta, al usuario se le puede presentar una lista de casi coincidencias, o simplemente se le puede notificar que no halló ninguna coincidencia. Como alternativa, se pueden intentar sustituciones y búsquedas adicionales.

Tal como se apreciará a partir de lo anterior, el proceso de la Figura 4 se puede modificar según cualquiera de una variedad de maneras para lograr un objetivo deseado. Por ejemplo, se podrían generar y usar tablas de correlación independientes para tipos diferentes de artículos (por ejemplo, libros con respecto a música) y/o tipos diferentes de campos (por ejemplo, títulos con respecto a materia). Además, el proceso de corrección ortográfica se podría usar únicamente para corregir escrituras incorrectas dentro de un campo específico (tal como el campo de autor 42), o se podría aplicar únicamente a términos no coincidentes que no aparezcan dentro de un diccionario de términos.

**IV. Generación de tabla de correlación**

A continuación se describirá en referencia a las Figuras 5 y 6 el proceso de generación de tablas 46 (Figura 1).

El proceso de generación de tablas 46 se implementa como un proceso fuera de línea que se genera periódicamente, por ejemplo, una vez por día, para generar una nueva tabla de correlación 50. Tal y como se ha descrito anteriormente, el proceso genera la tabla a partir de los M archivos de ficheros de registro de peticiones diarias más recientes 36(1) a 36(M). El uso de una M relativamente pequeña (por ejemplo, 5) tiende a producir datos de correlación que reflejan notablemente tendencias de compra de corto plazo (por ejemplo, versiones nuevas, *best-sellers* semanales, etcétera), mientras que el uso de una M mayor (por ejemplo, 100) tiende a producir una base de datos más exhaustiva. Se puede usar alternativamente un planteamiento híbrido en el que la tabla se genera a partir de un número grande de archivos de ficheros de registro, pero en el cual a los archivos de ficheros de registro más recientes se les asigna un peso mayor. Por ejemplo, las peticiones presentadas durante las últimas semanas se pueden contar tres veces cuando se generan las puntuaciones de correlación 64, mientras que las peticiones presentadas desde hace una semana hasta un mes se pueden contar solamente una vez.

La Figura 5 ilustra el formato general de los archivos de ficheros de registros de peticiones. Cada entrada en el fichero de registro (se muestran cuatro entradas) incluye información sobre una transacción HTTP (Protocolo de Transferencia de Hipertexto) particular. Por ejemplo, la entrada 100 indica que a las 2:23 AM del 13 de febrero de 1998, el usuario 29384719287 presentó la petición {autor=Seagal, título=Human Dynamics} desde la página de

búsqueda de libros y que se hallaron dos artículos que coincidían con la petición. Los valores de ARTÍCULOS_HALLADOS (*ITEMS_FOUND*) en el fichero de registro indican preferentemente el número de artículos que coincidieron exactamente con la petición original, y por lo tanto no reflejan ni "casi coincidencias" ni coincidencias resultantes de correcciones ortográficas.

La entrada 102 indica que el mismo usuario seleccionó un artículo que tenía un ISBN de 1883823064 aproximadamente veinte segundos más tarde, y que esta selección se realizó desde una página de resultados de búsqueda (tal como resulta evidente a partir de la línea de "ORIGEN_HTTP). Dentro de los archivos de los ficheros de registro se reflejan de modo similar otros tipos de acciones de usuario, tales como una solicitud de colocar un artículo en un carrito de la compra o de comprar un artículo. Tal como se ha indicado mediante el anterior ejemplo, una ruta de navegación de un usuario dado se puede determinar comparando entradas dentro del fichero de registro 36.

La Figura 6 ilustra la secuencia de etapas realizadas por el proceso de generación de tablas 46. En este ejemplo, se supone que el proceso se ejecuta una vez por día en la medianoche, justo después de que se haya cerrado el archivo del fichero de registro diario más reciente. Se supone también que ya se han procesado los M-1 archivos de ficheros de registro diarios más recientes usando las etapas 110 a 114 del proceso para generar archivos de resultados diarios respectivos.

En la etapa 110, el proceso analiza sintácticamente el nuevo archivo de fichero de registro diario para extraer todas las sumisiones de peticiones para las cuales ARTÍCULOS_HALLADOS>0. El hecho de ignorar las presentaciones de sumisiones que produjeron un resultado de petición nulo (ARTÍCULOS_HALLADOS=0) proporciona las ventajas importantes de (1) evitar que se añadan a la tabla de correlación términos no coincidentes o bien como palabras clave o bien como términos relacionados y (2) excluir la consideración de correlaciones potencialmente "débiles" entre términos coincidentes.

En la etapa 112, las entradas extraídas en la etapa 110 se procesan para la correlación de términos de búsqueda según la frecuencia de aparición dentro de la misma petición. Esta etapa conlleva el recuento, para cada par de términos de búsqueda que aparecieron dentro de la misma petición por lo menos una vez, del número de veces que aparecieron juntos los dos términos durante todo el día. Durante este proceso, los términos idénticos que se presentaron en campos de búsqueda diferentes se tratan como términos diferentes. Por ejemplo, el término TRAIL con un prefijo de campo de "T" se trataría como si fuera diferente de TRAIL con un prefijo de "M".

Tal como se ha indicado anteriormente, durante el proceso de la etapa 112 se puede tener en cuenta cualquiera de una variedad de otros factores. Por ejemplo, una sumisión de petición dada se puede contar dos veces si el usuario posteriormente seleccionó un artículo de la página de resultados de búsqueda, y se puede contar una tercera vez si el usuario a continuación compró el artículo o añadió el mismo a un carrito de compra personal. En el proceso también se pueden incorporar datos de correlación intrínsecos. Los resultados de la etapa 112, que se presentan en la forma general de la tabla de correlación de la Figura 3, se guardan como archivo de resultados diario.

En la etapa 116 el archivo de resultados diario creado en la etapa 114 se fusiona con los últimos M-1 archivos de resultados diarios para producir la tabla de correlación 50. Como parte de este proceso, las listas de términos relacionados se truncan a una longitud fija de N, y los datos de correlación resultantes se almacenan en una estructura de datos de árbol B para lograr una petición eficaz. A continuación, la nueva tabla de correlación 50 se escribe en la RAM en lugar de la tabla de correlación existente.

Aunque esta invención se ha descrito en términos de ciertos ejemplos preferidos, otros ejemplos que resultan evidentes para aquellos con conocimientos habituales en la técnica se sitúan también dentro del alcance de la invención. Por consiguiente, el alcance de la presente invención está destinado a quedar definido únicamente por referencia a las reivindicaciones adjuntas.

En las reivindicaciones que siguen, los caracteres de referencia usados para indicar etapas del proceso se proporcionan únicamente por comodidad descriptiva, y no para implicar un orden particular de realización de las etapas.

**Apéndice**

```
char *sort_string(char*string_to_sort) {
        qsort(strin_to_sort, strlen(string_to_sort),
                sizeof(char), qsort_char_compare);
        return string_to_sort;
}
int score_sorted_strings(char* string1, char* string2) {
        int result = 0;
        int finished = FALSE;
        int compare = 0;
        unsigned char* str1_ptr = (unsigned char*) string1;
        unsigned char* str2_ptr = (unsigned char*) string2;
        while (!finished) {
        -if (*str1_ptr = = 0 || *str2_ptr = = 0)
                finished = TRUE;
                /* La función de puntuación compara dos caracteres; si son iguales, se incrementan ambos
                punteros para avanzar, si no, se incrementa solamente el puntero menor y se realiza una
                comparación nuevamente */
                if (!finished) {
                        compare = qsort_char_compare((void*)str1_ptr, (void*)str2_ptr);
                                if ( compare < 0) {
                                        str1_ptr++;
                                                result+ +;
                                } else if (compare > 0) {
                                                str2_ptr + + ;
                                                result+ +;
                                } else if ( compare = = 0) {
                        str1_ptr++;
                                        str2_ptr+ +;
                }
        }
}
/* Añadir a la puntuación cualesquiera resultados restantes */
while (*str1_ptr++ !=NULL) {
        result + + ;
        }
        while (*str2_ptr++ != NULL) {
                result+ +;
        }
        return result;
}
```

**REIVINDICACIONES**

1. Método de procesado de una petición de búsqueda que incluye por lo menos un término de petición escrito incorrectamente en un sistema informático (30) que implementa un motor de búsqueda (38) que es accesible a los usuarios a través de una red de ordenadores, comprendiendo el método las siguientes etapas implementadas por ordenador:

(a) recibir la petición de búsqueda de un usuario a través de la red de ordenadores, comprendiendo la petición de búsqueda una pluralidad de términos de búsqueda y estando dirigida a una base de datos de información (40) en la que se va a realizar la búsqueda;

(b) identificar, dentro de la petición de búsqueda, un término de búsqueda no coincidente que no produce una coincidencia dentro de la base de datos de información, y por lo menos un término de búsqueda coincidente que produce una coincidencia dentro de la base de datos de información;

(c) usar datos de correlación de términos de búsqueda para identificar una pluralidad de términos adicionales que se consideran relacionados con dicho por lo menos un término de búsqueda coincidente, basándose los datos de correlación de términos de búsqueda por lo menos en sumisiones de peticiones históricas; y

(d) comparar los términos adicionales identificados en la etapa (c) con el término no coincidente usando una función de comparación ortográfica para identificar un término adicional que es una sustitución candidata escrita correctamente para el término no coincidente, estando adaptada la función de comparación ortográfica:

para comparar una primera y segunda cadenas de caracteres ordenando la primera y segunda cadenas y comparando carácter a carácter la primera y segunda cadenas ordenadas; y
para generar una puntuación que indica un grado de similitud entre la primera y segunda cadenas de caracteres, y

comparándose la puntuación con un valor umbral para determinar si un término adicional correspondiente es una sustitución candidata.

2. Método según la reivindicación 1, en el que los datos de correlación de términos de búsqueda se basan por lo menos en frecuencias con las cuales han aparecido previamente términos de búsqueda en la misma petición.

3. Método según la reivindicación 1 ó la reivindicación 2, que comprende además la etapa de procesar sumisiones de peticiones históricas dentro de un fichero de registro para generar los datos de correlación de términos de búsqueda.

4. Método según la reivindicación 3, en el que la etapa de procesar sumisiones de peticiones históricas comprende aplicar una función de ventana al fichero de registro.

5. Método de la reivindicación 3 ó la reivindicación 4, en el que la etapa de procesar sumisiones de peticiones históricas comprende ignorar peticiones que produjeron resultados de petición nulos.

6. Método según cualquiera de las reivindicaciones anteriores, en el que el valor umbral depende de una longitud del término no coincidente.

7. Método según cualquiera de las reivindicaciones anteriores, que comprende además las siguientes etapas implementadas por ordenador:

(e) sin requerir una entrada de datos por el usuario, sustituir el término no coincidente por el término de sustitución candidato dentro de la petición de búsqueda para generar una petición de búsqueda modificada;

(f) aplicar la petición de búsqueda modificada a la base de datos de información con el fin de realizar una búsqueda; y

(g) notificar al usuario el resultado de la petición de la etapa (f) y la sustitución realizada en la etapa (e).

8. Método según cualquiera de las reivindicaciones 1 a 6, en el que la etapa (d) comprende identificar una pluralidad de términos de sustitución candidatos escritos correctamente para el término no coincidente, y el método comprende además presentar al usuario una lista de los términos de sustitución candidatos para su selección.

9. Método según la reivindicación 8, en el que la etapa de presentar al usuario una lista comprende presentar cada término de sustitución candidato dentro de un respectivo hiperenlace que puede ser seleccionado por el usuario para

iniciar una búsqueda modificada.

10. Método según cualquiera de las reivindicaciones anteriores, en el que la etapa (b) comprende identificar una pluralidad de términos de búsqueda no coincidentes dentro de la petición de búsqueda, y el método comprende repetir la etapa (e) para cada uno de los términos de búsqueda no coincidentes con el fin de identificar un término de sustitución candidato para cada término de búsqueda no coincidente.

11. Método según cualquiera de las reivindicaciones anteriores, en el que la etapa (c) comprende:

(c1) para cada uno de entre una pluralidad de términos de búsqueda coincidentes, identificar una lista respectiva de términos relacionados; y

(c2) combinar las listas de términos relacionados identificadas en la etapa (c1).

12. Método según cualquiera de las reivindicaciones anteriores, que comprende además procesar peticiones de búsqueda sometidas al motor de búsqueda por una pluralidad de usuarios durante un periodo de tiempo para generar los datos de correlación (50), indicando los datos de correlación correlaciones entre términos de búsqueda sobre la base de por lo menos frecuencias de apariciones anteriores de los términos de búsqueda dentro de la misma petición de búsqueda.

13. Método según la reivindicación 12, en el que dicho procesado de peticiones de búsqueda comprende actualizar los datos de correlación sustancialmente en tiempo real a medida que se reciben peticiones de búsqueda de los usuarios.

14. Método según la reivindicación 12 ó la reivindicación 13, en el que dicho procesado de peticiones de búsqueda comprende analizar sintácticamente un fichero de registro que incluye peticiones sometidas al motor de búsqueda.

15. Método según la reivindicación 14, en el que dicho procesado de peticiones de búsqueda comprende además aplicar una función de sesgado basada en el tiempo al fichero de registro para favorecer las sumisiones de peticiones de búsqueda recientes con respecto a las sumisiones de peticiones de búsqueda antiguas.

16. Método según la reivindicación 15, en el que dicha aplicación de una función de sesgado basada en el tiempo comprende aplicar una función de ventana al fichero de registro.

17. Método según cualquiera de las reivindicaciones 12 a 16, en el que dicho procesado de peticiones de búsqueda comprende ignorar peticiones de búsqueda que producen un resultado de petición nulo.

18. Método según cualquiera de las reivindicaciones 12 a 17, en el que dicho procesado de peticiones de búsqueda comprende además evaluar acciones de usuarios posteriores a la sumisión de peticiones para identificar peticiones de búsqueda que se considera que han producido un resultado satisfactorio, y ponderar con mayor intensidad las peticiones de búsqueda que produjeron un resultado satisfactorio en la generación de los datos de correlación.

19. Método según cualquiera de las reivindicaciones anteriores, en el que el sistema informático forma parte de un sitio Web, y la base de datos de información incluye información sobre productos que están disponibles para su compra a través del sitio Web.

20. Motor de búsqueda para permitir que los usuarios de una red de ordenadores realicen búsquedas de una base de datos (40) de artículos, comprendiendo el motor de búsqueda:

un sistema informático (30) que tiene almacenados en una memoria del mismo datos de correlación de términos de búsqueda, indicando los datos de correlación de términos de búsqueda (50) correlaciones entre términos de búsqueda sobre la base por lo menos de sumisiones de peticiones anteriores de usuarios; y

un servidor de peticiones (38) que se ejecuta en el sistema informático, estando adaptado el servidor de peticiones para buscar en la base de datos de artículos usando peticiones de búsqueda recibidas de los usuarios, estando configurado el servidor de peticiones para procesar una petición de búsqueda de múltiples términos que incluye términos tanto coincidentes como no coincidentes mediante por lo menos:

el acceso a los datos de correlación de términos de búsqueda para identificar una pluralidad de términos adicionales que se consideran relacionados con el(los) término(s) coincidente(s) de la petición de búsqueda;

la comparación de la ortografía de los términos adicionales con la escritura de por lo menos un término no coincidente de la petición de búsqueda usando una función de comparación ortográfica para determinar si algunos de los términos adicionales es una sustitución candidata para el término no coincidente, estando adaptada la función de comparación ortográfica:

para comparar una primera y segunda cadenas de caracteres ordenando la primera y segunda cadenas y comparando carácter a carácter la primera y segunda cadenas ordenadas; y

para generar una puntuación que indica un grado de similitud entre la primera y segunda cadenas de caracteres; y

comparándose la puntuación con un valor umbral para determinar si un término adicional correspondiente es una sustitución candidata.

21. Motor de búsqueda según la reivindicación 20, en el que los datos de correlación de términos de búsqueda se basan por lo menos en frecuencias con las cuales han aparecido previamente términos de búsqueda dentro de la misma petición.

22. Motor de búsqueda según la reivindicación 20 ó la reivindicación 21, en el que el servidor de peticiones está configurado para, sin requerir una entrada de datos por el usuario:

sustituir un término no coincidente por un término adicional para generar una petición modificada; y

buscar en la base de datos de artículos con la petición modificada.

23. Programa de ordenador que comprende elementos de programa de ordenador legibles por ordenador o por máquina para configurar un sistema informático con el fin de implementar el método según cualquiera de las reivindicaciones 1 a 19.

24. Medio portador de programa de ordenador que comprende un programa de ordenador según la reivindicación 23.

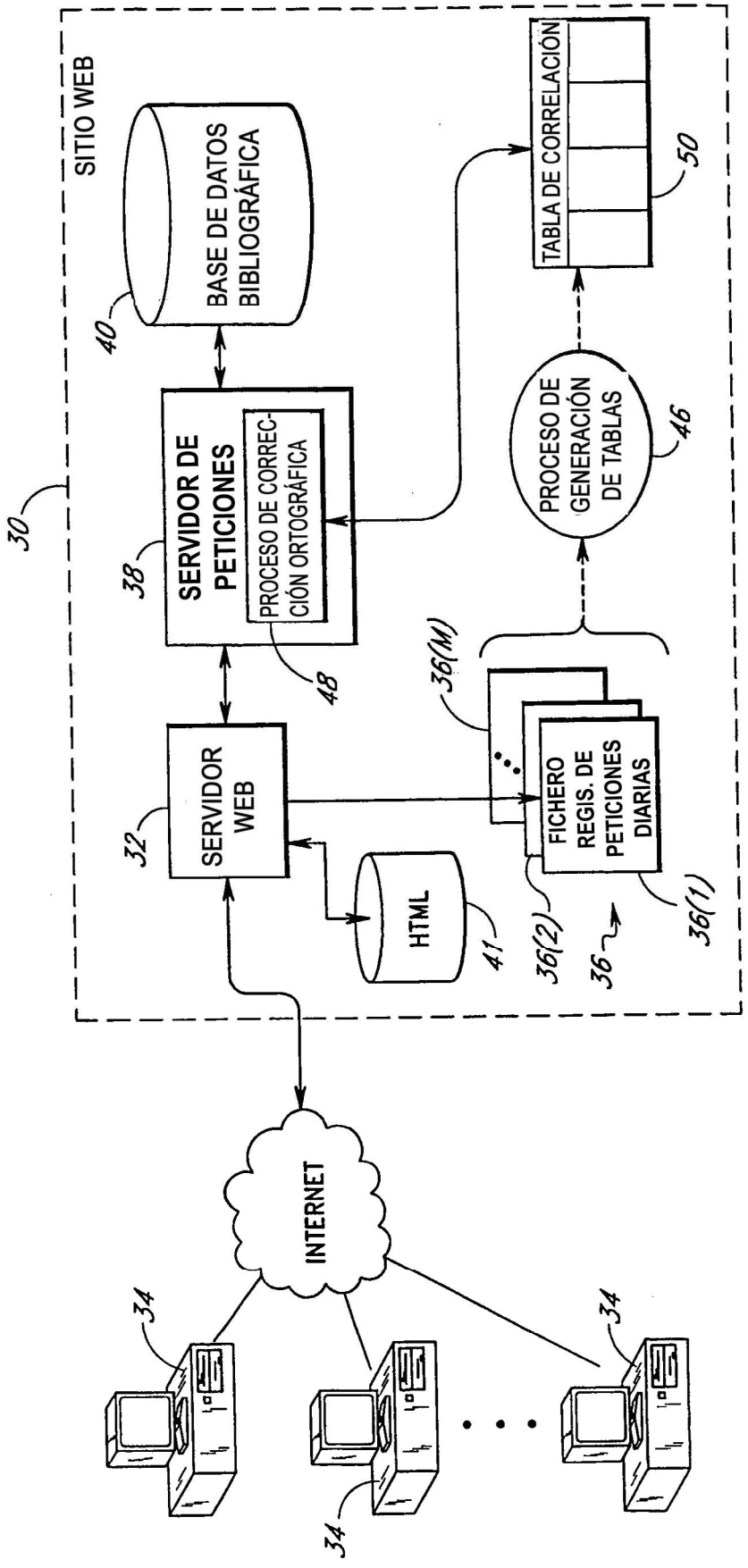
SITIO WEB
30
BASE DE DATOS BIBLIOGRÁFICA
40
TABLA DE CORRELACIÓN
50
SERVIDOR DE PETICIONES
38
PROCESO DE CORRECCIÓN ORTOGRÁFICA
48
PROCESO DE GENERACIÓN DE TABLAS
46
SERVIDOR WEB
32
FICHERO REGIS. DE PETICIONES DIARIAS
36(M)
36(2)
36
36(1)
HTML
41
INTERNET
34
34
34

FIG. 1

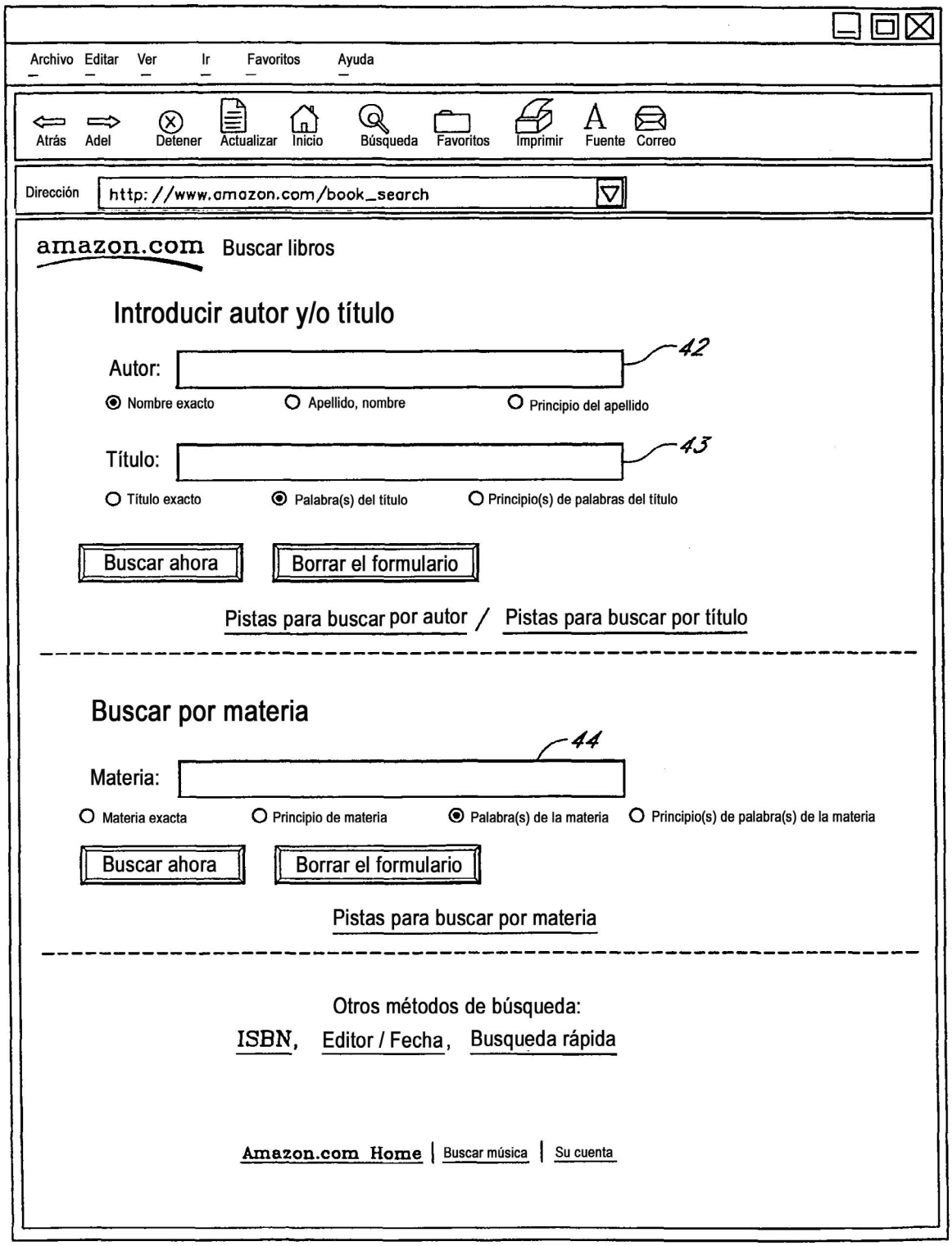
Archivo Editar Ver Ir Favoritos Ayuda
Atrás Adel Detener Actualizar Inicio Búsqueda Favoritos Imprimir Fuente Correo
Dirección http://www.amazon.com/book_search
amazon.com Buscar libros
Introducir autor y/o título
Autor:
42
Nombre exacto
Apellido, nombre
Principio del apellido
Título:
43
Título exacto
Palabra(s) del título
Principio(s) de palabras del título
Buscar ahora
Borrar el formulario
Pistas para buscar por autor / Pistas para buscar por título
Buscar por materia
44
Materia:
Materia exacta
Principio de materia
Palabra(s) de la materia
Principio(s) de palabra(s) de la materia
Buscar ahora
Borrar el formulario
Pistas para buscar por materia
Otros métodos de búsqueda:
ISBN, Editor / Fecha, Busqueda rápida
Amazon.com Home | Buscar música | Su cuenta

FIG.2

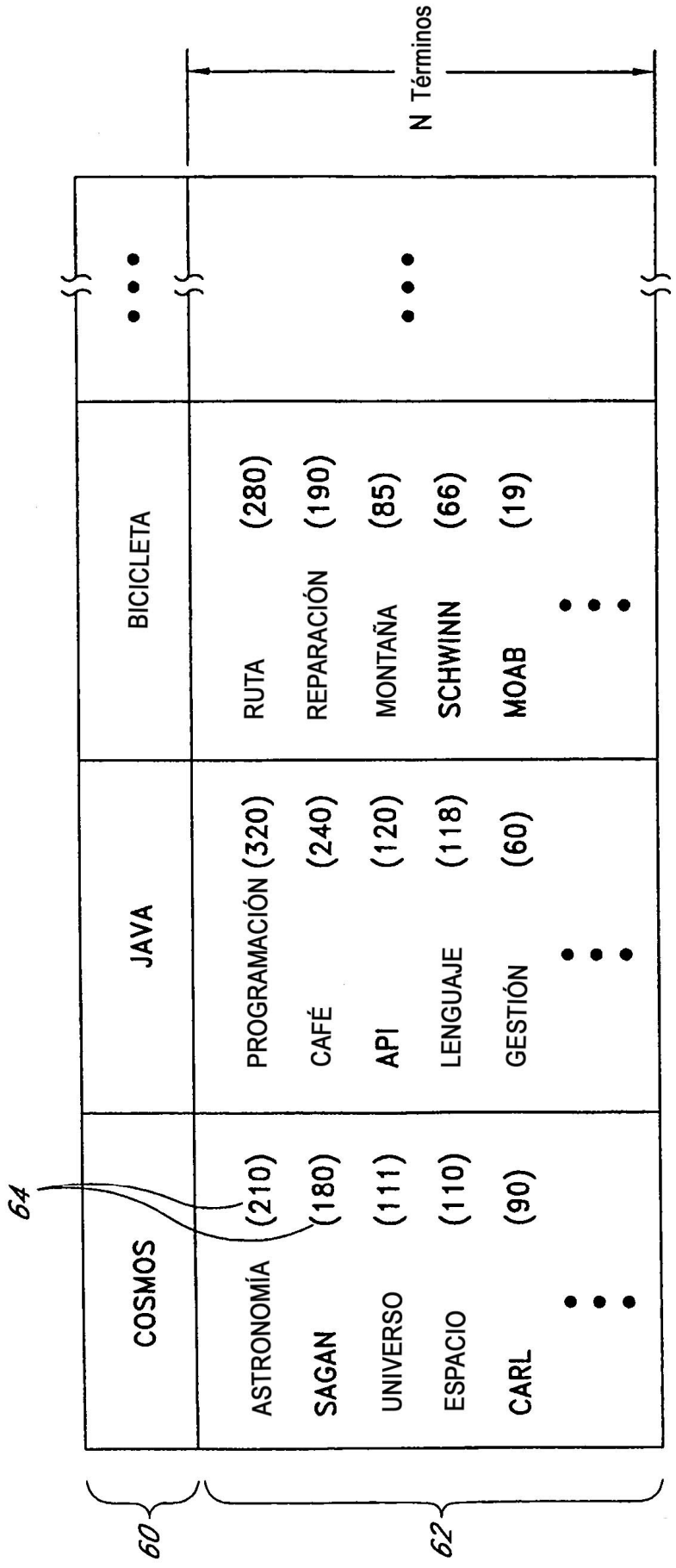
N Términos
64
60
62
COSMOS
JAVA
BICICLETA
ASTRONOMÍA (210)
SAGAN (180)
UNIVERSO (111)
ESPACIO (110)
CARL (90)
PROGRAMACIÓN (320)
CAFÉ (240)
API (120)
LENGUAJE (118)
GESTIÓN (60)
RUTA (280)
REPARACIÓN (190)
MONTAÑA (85)
SCHWINN (66)
MOAB (19)

FIG. 3

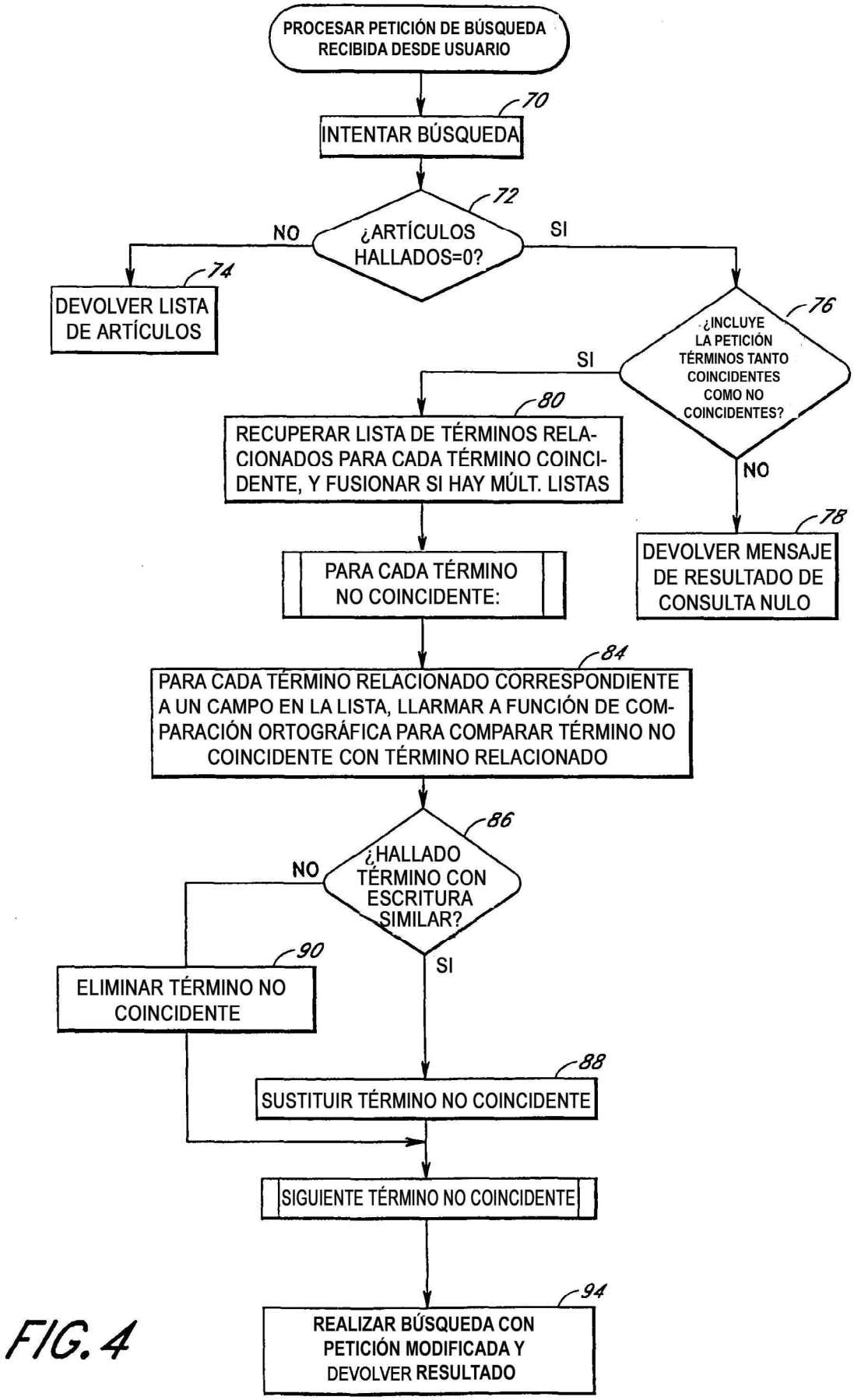
PROCESAR PETICIÓN DE BÚSQUEDA RECIBIDA DESDE USUARIO
70
INTENTAR BÚSQUEDA
72
¿ARTÍCULOS HALLADOS=0?
NO
SI
74
DEVOLVER LISTA DE ARTÍCULOS
76
¿INCLUYE LA PETICIÓN TÉRMINOS TANTO COINCIDENTES COMO NO COINCIDENTES?
SI
NO
80
RECUPERAR LISTA DE TÉRMINOS RELACIONADOS PARA CADA TÉRMINO COINCIDENTE, Y FUSIONAR SI HAY MÚLT. LISTAS
78
DEVOLVER MENSAJE DE RESULTADO DE CONSULTA NULO
PARA CADA TÉRMINO NO COINCIDENTE:
84
PARA CADA TÉRMINO RELACIONADO CORRESPONDIENTE A UN CAMPO EN LA LISTA, LLARMAR A FUNCIÓN DE COMPARACIÓN ORTOGRÁFICA PARA COMPARAR TÉRMINO NO COINCIDENTE CON TÉRMINO RELACIONADO
86
¿HALLADO TÉRMINO CON ESCRITURA SIMILAR?
NO
SI
90
ELIMINAR TÉRMINO NO COINCIDENTE
88
SUSTITUIR TÉRMINO NO COINCIDENTE
SIGUIENTE TÉRMINO NO COINCIDENTE
94
REALIZAR BÚSQUEDA CON PETICIÓN MODIFICADA Y DEVOLVER RESULTADO

FIG. 4

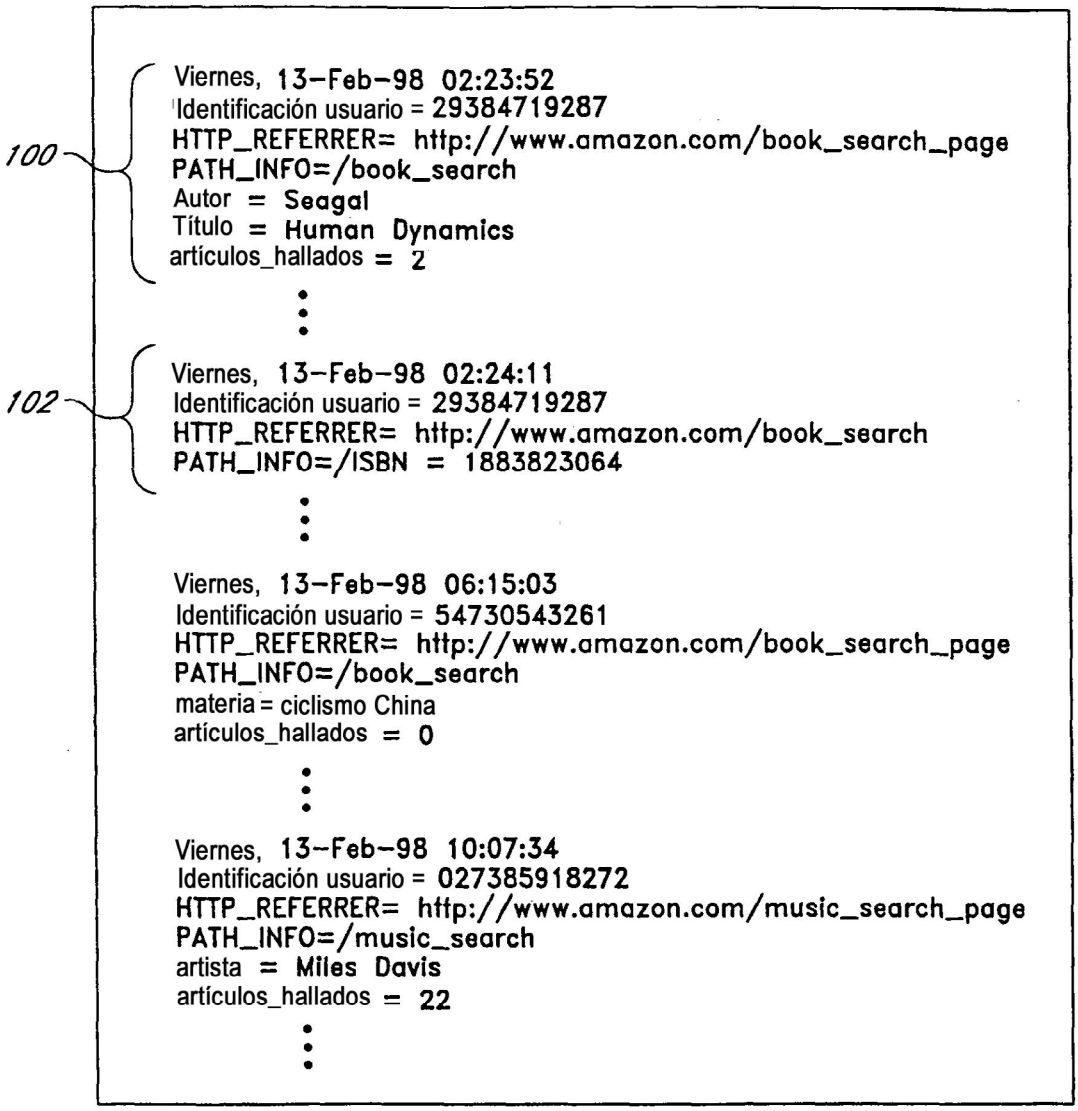
100
Viernes, 13-Feb-98 02:23:52
Identificación usuario = 29384719287
HTTP_REFERRER= http://www.amazon.com/book_search_page
PATH_INFO=/book_search
Autor = Seagal
Título = Human Dynamics
artículos_hallados = 2
102
Viernes, 13-Feb-98 02:24:11
Identificación usuario = 29384719287
HTTP_REFERRER= http://www.amazon.com/book_search
PATH_INFO=/ISBN = 1883823064
Viernes, 13-Feb-98 06:15:03
Identificación usuario = 54730543261
HTTP_REFERRER= http://www.amazon.com/book_search_page
PATH_INFO=/book_search
materia = ciclismo China
artículos_hallados = 0
Viernes, 13-Feb-98 10:07:34
Identificación usuario = 027385918272
HTTP_REFERRER= http://www.amazon.com/music_search_page
PATH_INFO=/music_search
artista = Miles Davis
artículos_hallados = 22

FIG. 5

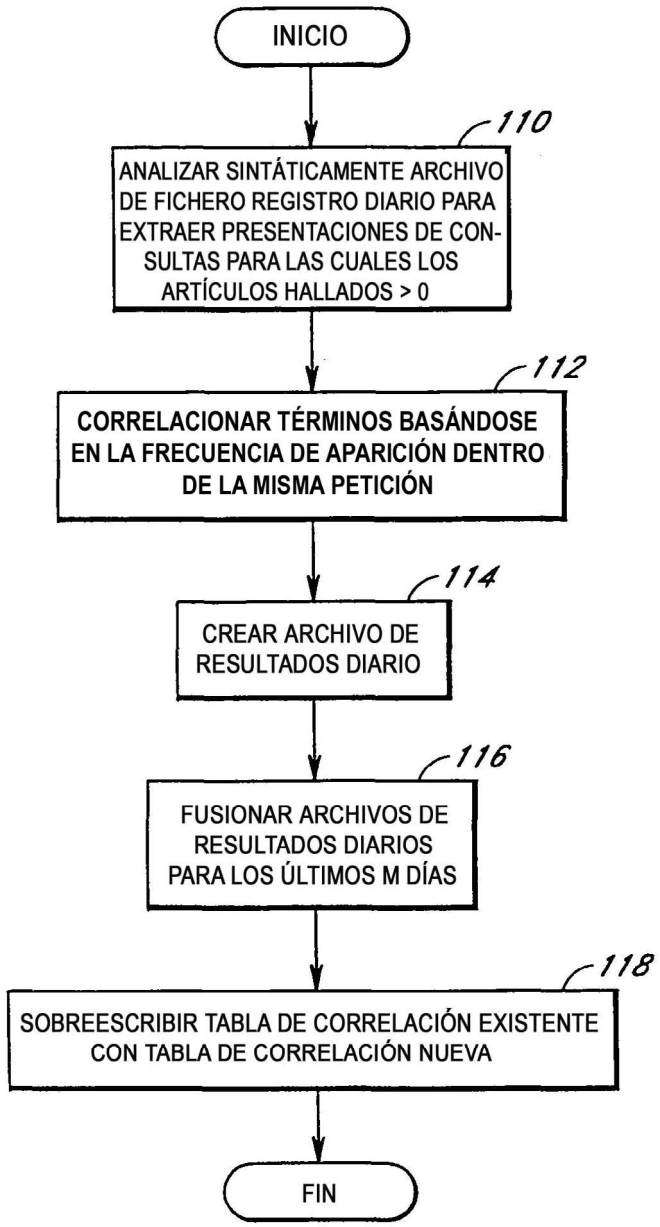
INICIO
110
ANALIZAR SINTÁTICAMENTE ARCHIVO DE FICHERO REGISTRO DIARIO PARA EXTRAER PRESENTACIONES DE CONSULTAS PARA LAS CUALES LOS ARTÍCULOS HALLADOS > 0
112
CORRELACIONAR TÉRMINOS BASÁNDOSE EN LA FRECUENCIA DE APARICIÓN DENTRO DE LA MISMA PETICIÓN
114
CREAR ARCHIVO DE RESULTADOS DIARIO
116
FUSIONAR ARCHIVOS DE RESULTADOS DIARIOS PARA LOS ÚLTIMOS M DÍAS
118
SOBREESCRIBIR TABLA DE CORRELACIÓN EXISTENTE CON TABLA DE CORRELACIÓN NUEVA
FIN

FIG. 6