(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 453 343**

(51) Int. CI.:

G06F 17/30 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **26.04.2004 E 04750667 (0)**

(97) Fecha y número de publicación de la concesión europea: **25.12.2013 EP 1620816**

(54) Título: **Procedimientos, arquitecturas, sistemas y software de búsqueda distribuida**

(30) Prioridad:

25.04.2003 US 465585 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
07.04.2014

(73) Titular/es:

THOMSON REUTERS GLOBAL RESOURCES (100.0%)
Landis + Gyr-Strasse 3
6300 Zug , CH

(72) Inventor/es:

BLUHM, MARK

(74) Agente/Representante:

DE ELZABURU MÁRQUEZ, Alberto

DESCRIPCIÓN

Procedimientos, arquitecturas, sistemas y software de búsqueda distribuida

Campo técnico

Varias realizaciones de la presente invención se refieren a sistemas de recuperación de información y sistemas de gestión del conocimiento y, más particularmente, se refieren a funciones de búsqueda distribuida dentro de dichos sistemas.

Antecedentes

Típicamente, los proveedores modernos de información en línea para ordenadores requieren la capacidad de realizar búsquedas de grandes cantidades de datos. Por ejemplo, el ordenamiento jurídico de Estados Unidos, así como otros ordenamientos jurídicos de todo el mundo, dependen en gran medida de las opiniones judiciales escritas, declaraciones escritas de los jueces, para articular o interpretar las leyes que rigen la resolución de disputas. Como consecuencia, los jueces y abogados de nuestro ordenamiento jurídico están investigando continuamente un cuerpo cada vez mayor de dictámenes anteriores, o jurisprudencia, buscando los más relevantes para la resolución o la prevención de nuevas disputas. Se estudia la relevancia de los casos encontrados y, en última instancia, son citados y discutidos en documentos, denominados producto de trabajo, que, por ejemplo, aconsejan una acción judicial, aconsejan a los clientes sobre las probables acciones judiciales, o educan a los clientes y los abogados acerca del estado de la ley en determinadas jurisdicciones.

Además, los sistemas de gestión del conocimiento, los sistemas de gestión de documentos y otros proveedores de datos en línea requieren, típicamente, información a partir de conjuntos de datos que pueden variar en tamaño, desde grandes a pequeños. Los conjuntos de datos en el rango terabyte ya no son infrecuentes. Por ejemplo, algunos sistemas pueden utilizar registros públicos que comprenden aproximadamente 1,2 terabytes de datos únicos, y datos fiscales y contables (Tax & Accounting, TA) que incluyen aproximadamente 20 gigabytes (GB) de datos únicos. En los sistemas anteriores, se han producido problemas debido a que, típicamente, el sistema puede almacenar sólo el cinco por ciento de los datos únicos de registros públicos. Además, el sistema es demasiado grande para los datos TA únicos que, típicamente, comparten espacio en el servidor con otros proveedores de datos.

Dichas variaciones en los tamaños de los conjuntos de datos y del sistema tienen un impacto sobre el rendimiento del motor de búsqueda, especialmente en relación con las implementaciones empresa-servidor (incluyendo problemas de disponibilidad inherentes). Por ejemplo, si se produce un fallo de memoria dentro de la CPU de un sistema, típicamente, el sistema no puede ejecutar el servicio de búsqueda hasta que se resuelva el fallo, y los mecanismos de conmutación por error son problemáticos. Debido a que, típicamente, el servicio de búsqueda hace un uso intensivo de memoria y no está limitado a la CPU, se desperdician recursos en la resolución de estos problemas de error.

Además, a veces, el procesamiento de consultas fuerza al motor de búsqueda a acceder a un disco para obtener páginas de datos si estas no están disponibles en la memoria caché del sistema de archivos. Aunque en algunos casos, los datos pueden encontrarse típicamente en la memoria caché del sistema de archivos si el conjunto de datos es suficientemente pequeño para ser mantenido completamente en memoria RAM, frecuentemente sucede que los conjuntos de datos son tan grandes que el procesamiento de consultas se produce frecuentemente a nivel de disco en lugar de a nivel de caché del sistema de archivos. Además, típicamente, las arquitecturas actuales no aseguran que el mismo motor de búsqueda procesará los mismos datos de manera consistente, lo cual niega las ventajas del almacenamiento en caché del motor de búsqueda.

El documento US 5.590.319 sobre el cual están caracterizadas las reivindicaciones independientes, describe un sistema de búsqueda en línea.

El documento US 2002/0143744 describe un procedimiento y un aparato para realizar búsquedas de información.

El documento WO 00/79436 describe una interfaz de motor de búsqueda.

En consecuencia, el presente inventor ha identificado una necesidad de mejores sistemas, herramientas y procedimientos para proporcionar funciones de búsqueda dentro de las plataformas de distribución en línea.

Sumario

Para abordar a esta y/u otras necesidades, el presente inventor ha ideado sistemas, procedimientos y software novedosos para proporcionar una función de búsqueda distribuida para plataformas de distribución en línea

usadas en los bufetes de abogados y otras empresas.

Según un primer aspecto de la presente invención, se proporciona un sistema de búsqueda en línea según se reivindica en la reivindicación 1.

Según un segundo aspecto de la presente invención, se proporciona un procedimiento de realización de una búsqueda según se reivindica en la reivindicación 4.

Por ejemplo, los sistemas, procedimientos y software proporcionan una pluralidad de conjuntos de datos. Los conjuntos de datos comprenden índices a otros conjuntos de datos. Al menos un motor de búsqueda está asociado con cada conjunto de datos. Un sistema que recibe una solicitud de búsqueda determina qué motores de búsqueda se usan para procesar la solicitud de búsqueda en base a los conjuntos de datos implicados en la solicitud de búsqueda. A continuación, la solicitud de búsqueda es reenviada a los motores de búsqueda identificados.

Particularmente, la realización ejemplar proporciona una función de búsqueda que es distribuida a través de múltiples motores de búsqueda en una manera en la que es probable que los datos de búsqueda estén almacenados en caché en la memoria RAM disponible, evitando de esta manera costosas búsquedas en disco.

Breve descripción de los dibujos

La Figura 1 es un diagrama de bloques de un sistema 100 de búsqueda distribuida ejemplar que corresponde a una o más realizaciones de la presente invención.

La Figura 2 es un diagrama de bloques que proporciona detalles adicionales de un sistema 200 de búsqueda distribuida ejemplar que corresponde a una o más realizaciones de la presente invención.

La Figura 3 es un diagrama de flujo que corresponde a uno o más procedimientos de funcionamiento ejemplares de un sistema de búsqueda distribuida ejemplar y los componentes asociados que conforman la presente invención.

Descripción detallada de las realizaciones ejemplares

La descripción siguiente, que incorpora las figuras y las reivindicaciones adjuntas, describe y/o ilustra una o más realizaciones ejemplares de una o más invenciones. Estas realizaciones, proporcionadas no para limitar sino sólo para ejemplificar y enseñar la invención o las invenciones, se muestran y describen con suficiente detalle para permitir que las personas con conocimientos en la materia realicen y usen la invención o las invenciones. De esta manera, cuando sea apropiado para evitar dificultar la comprensión de las una o más invenciones, la descripción puede omitir cierta información conocida por las personas con conocimientos en la materia relevante.

Sistema de Información Ejemplar

La Figura 1 representa un sistema 100 de búsqueda distribuida ejemplar que incorpora una o más enseñanzas de la presente invención. El sistema 100 incluye un controlador 102 de búsqueda, un conmutador 104 de mensajes, motores 106 de búsqueda, almacenamiento 110 conectado a red (Network Attached Storage, NAS) y la red 108 acopla, de manera comunicativa, los motores 106 de búsqueda al NAS 110. Los componentes indicados anteriormente pueden estar distribuidos a través de uno o más ordenadores servidores. En algunas realizaciones, los ordenadores servidores comprenden ordenadores de servicio basados en blades (blade: ordenador en una tarjeta plana, sin fuente de alimentación) de Sun Microsystems, Inc. Sin embargo, en realizaciones alternativas, pueden usarse servidores basados en arquitecturas de procesadores Intel.

El controlador 102 de búsquedas "escucha" las solicitudes de búsqueda. Utilizando un motor “división-combinación” ("split-merge"), el controlador de búsquedas recibe las solicitudes y las divide en solicitudes componentes (atendidas por los motores 106 de búsqueda). Cuando se reciben las respuestas desde los motores 106 de búsqueda, el controlador de búsqueda combina las respuestas y las envía al solicitante. Pueden realizarse solicitudes de división, a las que se ha referencia programáticamente como "SearchEngineRequest", a los diversos conjuntos de datos que comprenden o son generados a partir de la colección o el conjunto de recogida de datos. El conjunto de datos comprende una parte de un índice (denominado "IndexSet") a una colección o un conjunto de recogida de datos.

El conmutador 104 de mensajes opera para enrutar los mensajes desde el controlador 102 de búsqueda a uno o más motores 106 de búsqueda. Los mensajes pueden incluir solicitudes de búsqueda que deben ser realizadas por uno o más motores 106 de búsqueda. En algunas realizaciones de la invención, el conmutador 104 de mensajes proporciona una interfaz de servicios de mensajes Java (Java Message Service Interface, JMS). Además, en algunas realizaciones, los mensajes pueden ser enrutados usando software de cola de mensajes, tal como el

sistema de mensajería MQ disponible en IBM Corp. Sin embargo, se cree que ninguna de las realizaciones de la invención está limitada a un sistema de enrutamiento de mensajes particular y, en realizaciones alternativas, puede usarse el software de cola de mensajes SonicMQ de Sonic Software Corporation.

En algunas realizaciones, el motor 106 de búsqueda incluye un "contenedor" Java que pre- y post-procesa los datos buscados y encontrados por el servidor. En algunas realizaciones, este procesamiento puede ser llevado a cabo a través de una interfaz Java nativa ("Java Native Interface"). Los motores 106 de búsqueda reciben el componente SearchEngineRequest y el IndexSet específico y causan la ejecución de una búsqueda en el IndexSet especificado con la solicitud.

Los conjuntos de datos a ser buscados pueden residir en el almacenamiento 110 conectado a red que está acoplado, de manera comunicativa, al motor 106 de búsqueda a través de la red 108. El almacenamiento conectado a red puede ser cualquier tipo de dispositivo de almacenamiento accesible a través de una red. Los ejemplos de dicho almacenamiento conectado a red son conocidos en la técnica e incluyen servidores de archivos, servidores de almacenamiento y otros medios de almacenamiento conectados a la red.

La red 108 puede ser cualquier tipo de red, cableada o inalámbrica, capaz de soportar comunicación de datos. En algunas realizaciones de la invención, la red 108 comprende una red Gigabit Ethernet privada. Sin embargo se cree que ninguna realización de la invención está limitada a un tipo de red particular.

Los motores 106 de búsqueda pueden ser ejecutados en sistemas genéricos de Intel con el sistema operativo Linux instalado. Los datos para los IndexSets en algunas realizaciones pueden ser accedidos a través de un protocolo de sistema de archivos de red (Network File System, NFS) desde el servidor 110 de almacenamiento conectado a red (NAS). En cuanto la consulta inicial entra en el motor de búsqueda, el motor de búsqueda recibe el nombre IndexSet y los nombres de los archivos necesarios para satisfacer la consulta de búsqueda.

El motor 106 de búsqueda puede realizar llamadas NFS al servidor 110 NAS y puede solicitar datos para esos archivos. Típicamente, estos datos son estáticos y están almacenados en caché en el sistema cliente NFS. Posteriormente, cuando el motor de búsqueda accede a los datos para su IndexSet asignado, puede realizar una llamada meta-directorio al servidor NFS para obtener información de archivo. El motor 1-6 de búsqueda lee las páginas de datos desde la caché de memoria RAM local, lo que permite una búsqueda a velocidad de RAM de los términos de consulta.

La Figura 2 proporciona detalles adicionales de un sistema 200 de búsqueda distribuida ejemplar que incorpora una o más enseñanzas de la presente invención. El sistema 200 incluye los componentes descritos anteriormente con referencia a la Figura 1 y, además, incluye el producto/cliente 202, gestor 206 de recursos y el agente 208 de plataforma de distribución en línea.

El producto/cliente 202 puede ser cualquier módulo de software cliente que usa la funcionalidad de búsqueda distribuida proporcionada según las enseñanzas de las realizaciones de la invención. Dicho software incluye navegadores, sistemas de gestión de documentos, sistema de gestión del conocimiento, sistemas de recuperación de documentos, sistemas de recuperación de jurisprudencia y similares. El producto/cliente 202 emite una o más solicitudes de búsqueda a un conmutador 104 de mensajes, que enruta las solicitudes a un controlador de servicios en base a los datos de la solicitud de búsqueda.

El proceso 208 del agente de plataforma de distribución en línea (Online Delivery Platform, ODP) inicia procesos en un servidor, tales como controladores 102 de búsqueda y motores 106 de búsqueda, y supervisa y gestiona estos procesos. En algunas realizaciones, el proceso 208 agente de ODP también realiza un seguimiento del proceso individual e informa sobre su estado de procesamiento a una base de datos que hace la función de tablón de anuncios. Además, en algunas realizaciones, el proceso 208 de agente de ODP reinicia los controladores de búsqueda o los motores de búsqueda cuando hay fallos o criterios "de transacción extensa". El agente 208 de ODP es considerado el proceso de agente de servidor que ejecuta los entornos ODP.

Cuando el sistema arranca, el agente 208 de ODP en el servidor es iniciado y consulta al gestor 206 de recursos (preconfigurado) para motores que asignan el agente 208 de ODP a un gestor de agentes (no mostrado). El gestor de agentes contiene información acerca de los agentes y los motores de búsqueda en un dominio y puede asignar cargas de trabajo, de manera dinámica, a los agentes 208 de ODP que lo consultan. En algunas realizaciones, el gestor de agentes comprende un agente LDAP (Lightweight Directory Access Protocol, protocolo ligero de acceso a directorios). En algunas realizaciones, a los motores de búsqueda se les asignan colas de mensajes con nombres que corresponden a los nombres del IndexSet asociado con el motor 106 de búsqueda.

En algunas realizaciones, si un motor 106 de búsqueda falla, su agente detectará el motor defectuoso y lo reiniciará. La consulta procesada en el momento del fallo del motor de búsqueda puede perderse y la solicitud del

controlador es descartada "con error". (Algunas realizaciones pueden transferir una copia de la consulta defectuosa a otro motor de búsqueda que opera en el conjunto de índice objeto). Sin embargo, el motor 106 de búsqueda defectuoso puede ser reiniciado de manera que las nuevas consultas puedan ser procesadas sin demora.

En algunas realizaciones, si el sistema de motor de búsqueda encuentra un fallo de CPU, RAM o de otro hardware, un agente conmutador de mensajes detecta que la cola IndexSet no tiene procesos que le den servicio. El agente alerta inmediatamente al gestor de agentes para reasignar los motores de búsqueda para dar servicio a esa cola IndexSet.

El diseño del sistema ejemplar de diversas realizaciones incorpora el despliegue de sistemas genéricos con una imagen fija del sistema operativo que "aprende" su papel en la arquitectura de búsqueda distribuida durante el proceso de arranque. La capacidad de recuperación del sistema puede soportar fallos de proceso o de hardware, y su flexibilidad permite la asignación de recursos adicionales para los componentes defectuosos.

Además, en algunas realizaciones, los recursos adicionales asignados para la conmutación por error no están "en espera o idle" (en espera de una conmutación por error). Pueden ser desplegados como "manipuladores de carga de trabajo", que proporcionan un procesamiento adicional si se detectan cuellos de botella de procesamiento. Esta carga de trabajo puede ser detectada a través del conmutador 104 de mensajes o el agente 208 de ODP, que pueden detectar e informar sobre los patrones de carga de trabajo de cada cola IndexSet.

Además, la arquitectura ejemplar de diversas realizaciones es propicia para un esquema de "supervisión relajada". No es necesario detectar y arreglar inmediatamente los fallos de los componentes. La detección y la notificación pueden suceder cuando ocurren eventos catastróficos, pero el arreglo de los componentes puede tener lugar en cualquier momento, siempre que haya recursos adicionales disponibles para asumir su carga de trabajo.

Procedimiento ejemplar de operación

La Figura 3 muestra un diagrama 300 de flujo de uno o más procedimientos ejemplares del funcionamiento de un sistema de gestión de información, tal como el sistema 100. El diagrama 300 de flujo incluye bloques 310-340, que están dispuestos y se describen en una secuencia de ejecución en serie en la realización ejemplar. Sin embargo, otras realizaciones pueden ejecutar dos o más bloques en paralelo usando múltiples procesadores o dispositivos similares a procesadores o un único procesador organizado como dos o más máquinas o sub-procesadores virtuales. Otras realizaciones alteran también la secuencia del proceso o proporcionan diferentes particiones funcionales para conseguir resultados análogos. Además, todavía otras realizaciones implementan los bloques como dos o más módulos de hardware interconectados con señales de datos y de control relacionadas comunicadas entre y a través de los módulos. De esta manera, el flujo del proceso ejemplar se aplica a implementaciones de software, hardware y firmware.

En el bloque 310, el procedimiento ejemplar comienza con la provisión de uno o más conjuntos de datos. Los conjuntos de datos comprenden partes de un índice a una colección de datos o un conjunto de colecciones de datos. El índice puede ser dividido en base a intervalos de índices de base de datos, en el que cada intervalo comprende un conjunto de datos. A continuación, los conjuntos de datos son almacenados en un dispositivo de almacenamiento, tal como NAS 110.

El bloque 320 implica la recepción de una solicitud de búsqueda. En el bloque 330, la solicitud de búsqueda es analizada para determinar cuáles son los conjuntos de datos requeridos.

En el bloque 340, a continuación, las solicitudes de búsqueda son reenviadas a los motores de búsqueda que corresponden a los conjuntos de datos identificados en el bloque 330. En algunas realizaciones de la invención, las solicitudes de búsqueda son reenviadas a los motores de búsqueda a través de colas de mensajes. Además, en algunas realizaciones, la cola de mensajes asociada a un motor de búsqueda particular recibe el mismo nombre que el IndexSet que está configurado para el motor de búsqueda.

Conclusión

Las realizaciones descritas anteriormente solo pretenden ilustrar y enseñar una o más maneras de realizar y usar la presente invención, no pretenden restringir su amplitud o alcance. El alcance real de la invención, que abarca todas las formas de practicar o implementar las enseñanzas de la invención, está definido solo por las reivindicaciones adjuntas.

REIVINDICACIONES

1. Un sistema (100) de búsqueda en línea que comprende:

una pluralidad de motores (106) de búsqueda;

medios para recibir una solicitud de búsqueda desde un solicitante;

medios para dividir la solicitud de búsqueda recibida en una pluralidad de solicitudes componentes y para asignar cada una de las solicitudes componentes a un motor de búsqueda correspondiente de entre los motores (106) de búsqueda;

medios para combinar los resultados de la búsqueda proporcionados por los motores (106) de búsqueda en respuesta a las solicitudes componentes en un resultado de búsqueda combinado; y

medios para proporcionar el resultado de la búsqueda combinada al solicitante;

caracterizado por que

un índice dividido en una pluralidad de partes índice, en el que a cada uno de entre la pluralidad de motores (106) de búsqueda se le asigna la búsqueda de datos usando al menos una de las partes índice; y por que

los motores (106) de búsqueda comprenden medios para recibir el componente de solicitud de búsqueda y el nombre de una parte índice específica y medios para causar la realización de una búsqueda en la parte índice especificada con la solicitud.

2. Sistema según la reivindicación 1, que comprende además una pluralidad de colas de mensajes, en el que cada cola de mensaje recibe solicitudes componentes y es asignada a un motor de búsqueda correspondiente de entre los motores (106) de búsqueda.

3. Sistema según la reivindicación 1 o 2, que comprende además un conmutador (104) de mensajes operable para enrutar cada solicitud componente a su motor (106) de búsqueda asignado.

4. Procedimiento para realizar una búsqueda usando una pluralidad de motores (106) de búsqueda, en el que el procedimiento comprende:

recibir una solicitud de búsqueda desde un solicitante;

dividir la solicitud de búsqueda recibida en una pluralidad de solicitudes componentes;

asignar cada una de las solicitudes componentes a un motor de búsqueda correspondiente de entre los motores (106) de búsqueda;

combinar los resultados de búsqueda proporcionados por los motores (106) de búsqueda en respuesta a las solicitudes componentes en un resultado de búsqueda combinado; y

proporcionar el resultado de búsqueda combinado al solicitante; caracterizado por que proporciona un índice dividido en una pluralidad de partes índice, en el que cada parte índice es asignada a un motor de búsqueda correspondiente de entre los motores (106) de búsqueda; y por que

los motores (106) de búsqueda reciben el componente de solicitud de búsqueda y el nombre de una parte índice específica y causan la ejecución de una búsqueda sobre la parte índice especificada con la solicitud.

5. Procedimiento según la reivindicación 4, que comprende además comunicar las solicitudes componentes a una pluralidad correspondiente de colas de mensajes, en el que cada cola de mensajes es asignada a un motor de búsqueda correspondiente de entre los motores (106) de búsqueda.

6. Procedimiento según la reivindicación 4, que comprende además almacenar en caché los datos en base a cada una de las solicitudes componentes en una memoria de acceso aleatorio asociada con un motor de búsqueda correspondiente de entre los motores (106) de búsqueda.

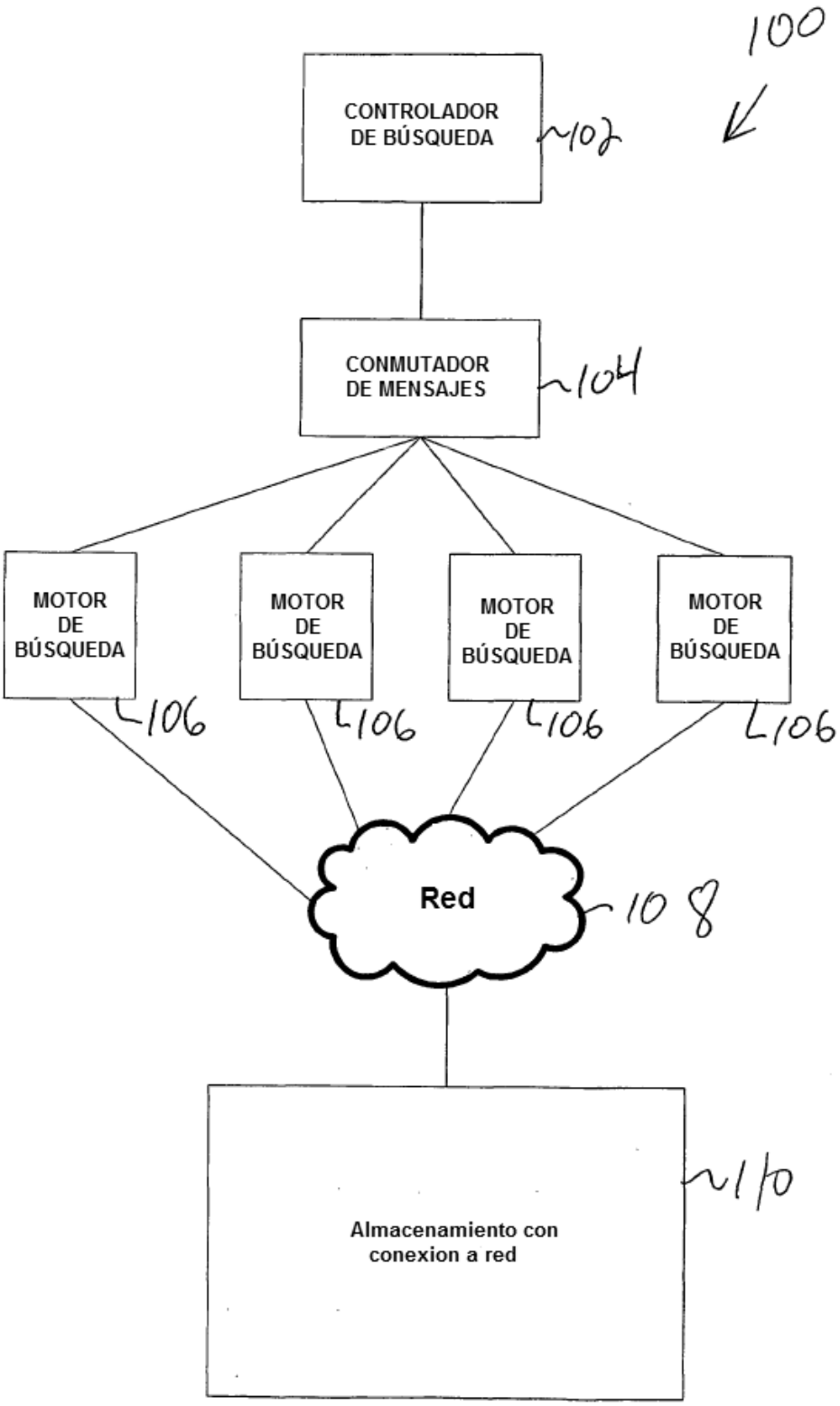
100
CONTROLADOR DE BÚSQUEDA
102
CONMUTADOR DE MENSAJES
104
MOTOR DE BÚSQUEDA
MOTOR DE BÚSQUEDA
MOTOR DE BÚSQUEDA
MOTOR DE BÚSQUEDA
106
106
106
106
Red
108
Almacenamiento con conexion a red
110

Figura 1

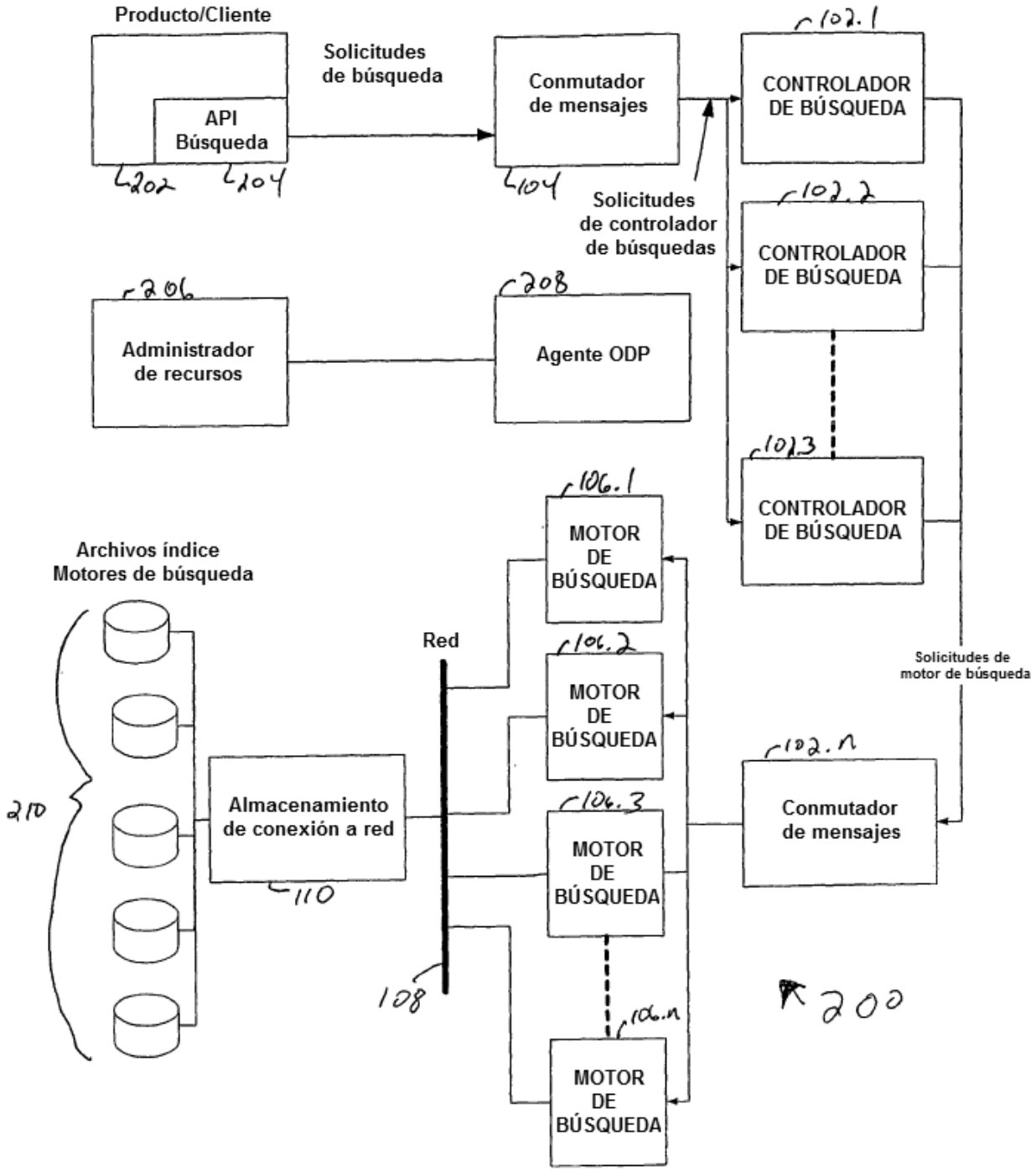
Producto/Cliente
API Búsqueda
202
204
Solicitudes de búsqueda
Conmutador de mensajes
104
102.1
CONTROLADOR DE BÚSQUEDA
Solicitudes de controlador de búsquedas
102.2
CONTROLADOR DE BÚSQUEDA
206
Administrador de recursos
208
Agente ODP
102.3
CONTROLADOR DE BÚSQUEDA
106.1
MOTOR DE BÚSQUEDA
Archivos índice Motores de búsqueda
Red
106.2
MOTOR DE BÚSQUEDA
Solicitudes de motor de búsqueda
102.n
Conmutador de mensajes
210
Almacenamiento de conexión a red
110
106.3
MOTOR DE BÚSQUEDA
108
200
106.n
MOTOR DE BÚSQUEDA

Figura 2

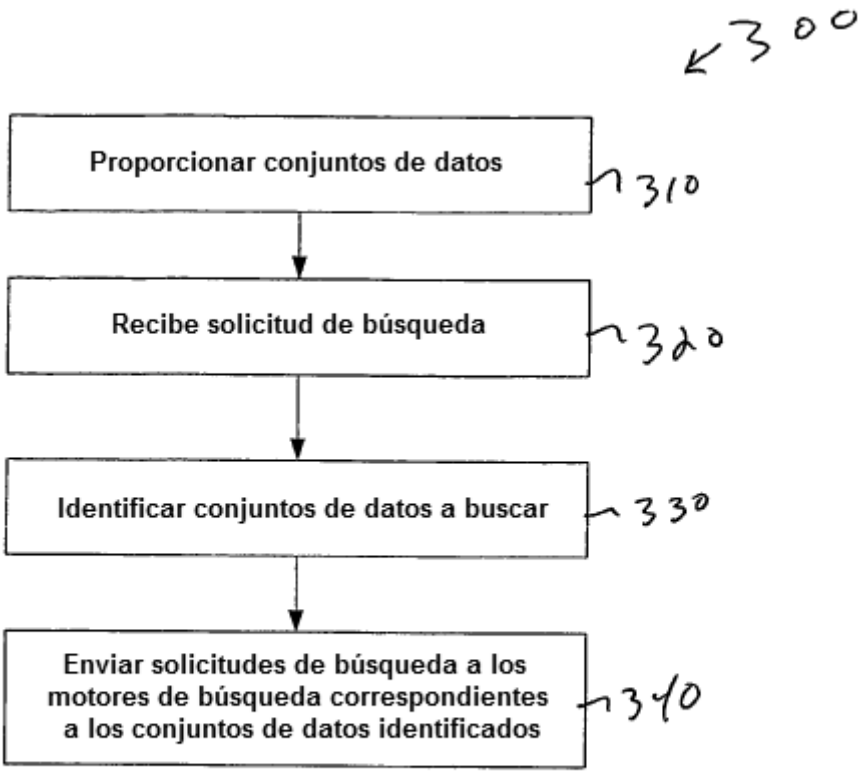
300
Proporcionar conjuntos de datos
310
Recibe solicitud de búsqueda
320
Identificar conjuntos de datos a buscar
330
Enviar solicitudes de búsqueda a los motores de búsqueda correspondientes a los conjuntos de datos identificados
340

Figura 3