(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 613 385**

(51) Int. CI.:

G06F 17/30 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **10.06.2005 E 05105110 (0)**

(97) Fecha y número de publicación de la concesión europea: **02.11.2016 EP 1643385**

(54) Título: **Sistema y procedimiento para clasificar resultados de búsqueda usando distancia de clic**

(30) Prioridad:

30.09.2004 US 955983

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
24.05.2017

(73) Titular/es:

MICROSOFT TECHNOLOGY LICENSING, LLC (100.0%)
One Microsoft Way
Redmond, WA 98052, US

(72) Inventor/es:

MEYERZON, DMITRIY y
ZARAGOZA, HUGO

(74) Agente/Representante:

CARPINTERO LÓPEZ, Mario

DESCRIPCIÓN

Sistema y procedimiento para clasificar resultados de búsqueda usando distancia de clic

Antecedentes de la invención

En una búsqueda de documentos de texto, un usuario introduce habitualmente una consulta en un motor de búsqueda. El motor de búsqueda evalúa la consulta en una base de datos de documentos indexados y devuelve la lista clasificada de documentos que mejor satisface la consulta. Una puntuación, que representa una medida de lo bien que el documento satisface la consulta, se genera algorítmicamente por el motor de búsqueda. Los algoritmos de puntuación habitualmente usados se basan en la división de la consulta en términos de búsqueda y el uso de información estadística sobre la aparición de términos individuales en el cuerpo de los documentos de texto a buscar. Los documentos se enumeran en orden de clasificación de acuerdo con sus puntuaciones correspondientes, de modo que el usuario puede ver los resultados de búsqueda que coinciden mejor en la parte superior de la lista de resultados de búsqueda.

Otra evaluación que ciertos motores de búsqueda pueden emplear para mejorar la calidad de los resultados es modificar la clasificación de los resultados mediante una función de clasificación seleccionada. Una función de clasificación de la técnica anterior a modo de ejemplo determina que cuando una página enlaza con otra página, se emite de manera eficaz un voto para la otra página. Cuantos más votos se emiten para una página, más importante es la página. La función de clasificación también puede tener en cuenta quién emitió el voto. Cuanto más importante es la página, más importantes son sus votos. Estos votos se acumulan y se usan como un componente de las valoraciones de las páginas en la red.

Una función de clasificación se usa para mejorar la calidad de la clasificación. Sin embargo, la eficacia de la función de clasificación puede verse afectada por la topología de la red. Por ejemplo, la función de clasificación que usa los votos descritos anteriormente puede ser menos eficaz en una configuración de intranet. Una intranet es una red que usa algunos de los mismos protocolos que internet, pero a la que solo puede accederse por un subconjunto de usuarios, tal como los empleados de una empresa. Las páginas de una intranet no se estructuran ni se conectan exactamente como internet, por lo que la pertinencia de los resultados producidos por una función de clasificación puede no reducirse en comparación con la configuración de internet.

Fagin R. y col.: “Searching the Workplace Web” WWW2003, del 20 al 24 de mayo de 2003, Budapest, Hungría, se refiere a la búsqueda en intranet, que se reconoce como muy diferente de una búsqueda en internet. Se centra en el uso de la agregación de clasificaciones, y permite examinar los efectos de diferentes heurísticas en la clasificación de los resultados de búsqueda. Los algoritmos de agregación de clasificaciones toman como entrada múltiples listas clasificadas de las diversas heurísticas y producen una ordenación de las páginas dirigida a minimizar el número de “desajustes” con respecto a la ordenación producida por las heurísticas de clasificación individuales.

Sumario de la invención

El objeto de la presente invención es proporcionar un sistema y un procedimiento para clasificar resultados de búsqueda de acuerdo con una nueva función denominada distancia de clic.

Este objeto se resuelve por el objeto de las reivindicaciones independientes.

Las realizaciones se proporcionan en las reivindicaciones dependientes.

La función de distancia de clic se aprovecha de la estructura jerárquica de una intranet. Una intranet normalmente sigue una estructura de árbol, con un nodo raíz y unas ramas subsiguientes que se extienden a otros nodos desde esa raíz. A menudo, el nodo raíz de la intranet se denomina su página principal. Otros sitios fuera de la configuración de intranet también pueden basarse en una estructura jerárquica y la distancia de clic para estos sitios también sería muy aplicable para clasificar las páginas del sitio.

La distancia de clic es una medida de pertinencia independiente de la consulta que mide el número de “clics” necesarios para llegar a una página determinada del sitio. En la estructura de árbol, el número de clics se representa por el número de ramas atravesadas en la ruta desde el nodo raíz. Una vez que se determina la distancia de clic para una página, la distancia de clic se incorpora en la puntuación de la página. La puntuación de la página que incorpora la distancia de clic determina la clasificación de la página entre las otras páginas dentro de los resultados de búsqueda.

En un aspecto de la presente invención, en primer lugar se “rastrea” la red para generar una tabla de propiedades asociadas con los enlaces y las páginas de la red. “Rastreo” se refiere a la recopilación automática de varios documentos (o cualquier unidad discreta análoga de información) en una base de datos denominada índice. El rastreo atraviesa múltiples documentos en la red siguiendo los enlaces de referencia de documentos dentro de ciertos documentos y, a continuación, procesando cada documento que haya encontrado. Los documentos se procesan identificando palabras clave o textos generales en los documentos para crear el índice.

Un índice a modo de ejemplo puede ser una lista invertida que tiene una columna de palabras y una columna que indica en qué documentos pueden encontrarse esas palabras. Cuando un usuario introduce uno o más términos de búsqueda, se obtienen los resultados y la presente invención aplica un algoritmo de clasificación que incluye la función de distancia de clic. La función de distancia de clic influye positiva o negativamente en la puntuación de ciertas páginas, perfeccionando los resultados devueltos al usuario.

En otro aspecto de la invención, se añade una propiedad de profundidad de URL (localizador de recursos uniforme) al algoritmo de clasificación para perfeccionar aún más los resultados. La propiedad de profundidad de URL mide el número de niveles en la URL para facilitar una comprobación con respecto a la función de distancia de clic y ajustar la puntuación de la página en consecuencia.

Breve descripción de los dibujos

La figura 1 ilustra un dispositivo informático a modo de ejemplo que puede usarse en una realización a modo de ejemplo de la presente invención.
La figura 2 ilustra un sistema para clasificar resultados de búsqueda de acuerdo con la distancia de clic de acuerdo con la presente invención.
La figura 3 ilustra una gráfica de red a modo de ejemplo de acuerdo con la presente invención.
La figura 4 ilustra una gráfica de red jerárquica a modo de ejemplo de acuerdo con la presente invención.
La figura 5 ilustra un diagrama de flujo lógico de un procedimiento a modo de ejemplo para calcular la distancia de clic de acuerdo con la presente invención.
La figura 6 ilustra un diagrama de flujo lógico de un procedimiento a modo de ejemplo para usar la distancia de clic en la clasificación de resultados de búsqueda de acuerdo con la presente invención.

Descripción detallada

A continuación, se describirá con todo detalle la presente invención con referencia a los dibujos adjuntos, que forman parte de la misma, y que muestran, a modo de ilustración, realizaciones a modo de ejemplo específicas para poner en práctica la invención. Sin embargo, la presente invención puede materializarse de muchas formas diferentes y no debe interpretarse como limitada a las realizaciones expuestas en el presente documento; por el contrario, estas realizaciones se proporcionan de manera que la presente divulgación será minuciosa y completa, y transmitirá con todo detalle el ámbito de la invención a los expertos en la materia. Entre otras cosas, la presente invención puede materializarse como procedimientos o dispositivos. En consecuencia, la presente invención puede adoptar la forma de una realización íntegramente de hardware, una realización íntegramente de software o una realización que combina aspectos de software y de hardware. Por lo tanto, la siguiente descripción detallada no debe interpretarse en un sentido limitante.

Entorno operativo ilustrativo

Con referencia a la figura 1, un sistema a modo de ejemplo para implementar la invención incluye un dispositivo informático, tal como el dispositivo 100 informático. El dispositivo 100 informático puede estar configurado como un cliente, un servidor, un dispositivo móvil, o cualquier otro dispositivo informático. En una configuración muy básica, el dispositivo 100 informático incluye habitualmente al menos una unidad 102 de procesamiento y una memoria 104 de sistema. Dependiendo de la configuración exacta y el tipo de dispositivo informático, la memoria 104 de sistema puede ser volátil (tal como RAM), no volátil (tal como ROM, memoria flash, etc.), o alguna combinación de las dos. La memoria 104 de sistema incluye habitualmente un sistema 105 operativo, una o más aplicaciones 106, y puede incluir unos datos 107 de programa. En una realización, la aplicación 106 incluye una aplicación 120 de clasificación de búsqueda para implementar la funcionalidad de la presente invención. Esta configuración básica se ilustra en la figura 1 por los componentes dentro de la línea 108 discontinua.

El dispositivo 100 informático puede tener características o funcionalidades adicionales. Por ejemplo, el dispositivo 100 informático también puede incluir dispositivos de almacenamiento de datos adicionales (extraíbles y/o no extraíbles) tales como, por ejemplo, discos magnéticos, discos ópticos o cintas. Este almacenamiento adicional se ilustra en la figura 1 mediante un almacenamiento 109 extraíble y un almacenamiento 110 no extraíble. Los medios de almacenamiento informático pueden incluir medios volátiles y no volátiles, extraíbles y no extraíbles implementados en cualquier procedimiento o tecnología para el almacenamiento de información, tales como instrucciones legibles por ordenador, estructuras de datos, módulos de programa u otros datos. La memoria 104 de sistema, el almacenamiento 109 extraíble y el almacenamiento 110 no extraíble son todos ejemplos de medios de almacenamiento informáticos. Los medios de almacenamiento informático incluyen, pero no se limitan a, una memoria RAM, ROM, EEPROM, memoria flash u otra tecnología de memoria, CD-ROM, discos versátiles digitales (DVD) u otro almacenamiento óptico, casetes magnéticos, cinta magnética, almacenamiento de disco magnético u otros dispositivos de almacenamiento magnético o cualquier otro medio que pueda usarse para almacenar la información deseada y al que pueda accederse mediante el dispositivo 100 informático. Cualquiera de estos medios de almacenamiento informáticos puede ser parte del dispositivo 100. El dispositivo 100 informático también puede tener un dispositivo(s) 112 de entrada como un teclado, un ratón, un puntero, un dispositivo de entrada de voz, un dispositivo de entrada táctil, etc. También puede incluirse un dispositivo(s) 114 de salida tal como una pantalla, unos altavoces, una impresora, etc.

El dispositivo 100 informático también contiene unas conexiones 116 de comunicación que permiten que el dispositivo se comunique con otros dispositivos 118 informáticos, tal como a través de una red. La conexión 116 de comunicación es un ejemplo de medio de comunicación. Los medios de comunicación pueden incorporar habitualmente instrucciones legibles por ordenador, estructuras de datos, módulos de programa u otros datos en una señal de datos modulada, tal como una onda portadora u otro mecanismo de transporte, e incluyen cualquier medio de suministro de información. La expresión "señal de datos modulada" significa una señal que tiene una o más de sus características establecidas o cambiadas de tal manera que codifican información en la señal. A modo de ejemplo, y no de limitación, los medios de comunicación incluyen medios cableados tales como una red cableada o una conexión directa por cable, y medios inalámbricos tales como medios acústicos, RF, infrarrojos y otros medios inalámbricos. La expresión medio legible por ordenador tal como se usa en el presente documento incluye tanto medios de almacenamiento como medios de comunicación.

Realizaciones ilustrativas para clasificar búsquedas por distancia de clic

Las realizaciones de la presente invención están relacionadas con una función de clasificación para un motor de búsqueda. La calidad de un motor de búsqueda se determina habitualmente por la pertinencia de los documentos de acuerdo con las clasificaciones asignadas por la función de clasificación. La función de clasificación puede basarse en múltiples características. Algunas de estas características pueden depender de la consulta, mientras que otras se consideran independientes de la consulta. La distancia de clic es el número de "clics" que un usuario tendrá que hacer desde la página principal de la intranet (la URL más autorizada en la intranet o una de las URL más autorizadas) a la página dada. En una gráfica web, la distancia de clic puede representarse como la ruta más corta entre la página principal y la página dada. En una realización, un algoritmo realiza en primer lugar un recorrido de amplitud y calcula la distancia entre un nodo dado y todos los otros nodos en la gráfica. El recorrido puede tomar N iteraciones para completarse, donde N es el diámetro de la gráfica (distancia máxima más corta), para calcular la distancia de clic para la intranet. La variable N en este caso es mucho menor que el número total de nodos en la gráfica. Por ejemplo, N para la presente invención puede estar entre 5 y 60 dependiendo de la red. Otras funciones de clasificación pueden requerir 40-50 iteraciones para cubrir la gráfica (por ejemplo, clasificación de página), haciendo que las otras funciones de clasificación sean varias veces más lentas que usar la distancia de clic.

La figura 2 ilustra un sistema para clasificar resultados de búsqueda de acuerdo con una distancia de clic de acuerdo con la presente invención. El motor 200 de búsqueda recibe una consulta que contiene múltiples términos de consulta. Cada término de consulta puede incluir múltiples términos de componente, como cuando el término de consulta es una frase (por ejemplo, la frase "sistema de gestión de documentos" puede considerarse un único término de consulta). Además, una consulta puede incluir uno o más operadores, tales como operadores booleanos, restricciones, etc., que se soportan habitualmente por motores de búsqueda conocidos.

Una pluralidad de documentos en una red distribuida, representada por los documentos 210, 212, 214, y 216, están disponibles para la búsqueda. En la práctica, un motor de búsqueda puede buscar cualquier número de documentos y, habitualmente, busca colecciones que contienen grandes números (por ejemplo, millones) de documentos. El volumen de documentos puede reducirse de la configuración de internet a la configuración de intranet, pero normalmente la reducción es de billones a millones, de manera que el número relativo de documentos todavía es bastante grande. Un módulo de indexación (no mostrado) genera estadísticas de documento individuales (por ejemplo, 218, 220, 222 y 224) para cada documento. Las estadísticas de documento se almacenan en un índice 226.

El motor 200 de búsqueda consulta el índice 226 para determinar una puntuación 228 de búsqueda para cada documento basándose en la consulta y las estadísticas de documento correspondientes. En la presente invención, una de las estadísticas de documento incluidas es la distancia de clic del documento. En otra realización, otra estadística de documento incluida es la profundidad de URL asociada con el documento. La distancia de clic y las profundidades de URL se combinan a continuación con las estadísticas dependientes de la consulta para formar la puntuación final de un documento. Habitualmente, las puntuaciones de documento 228 se clasifican a continuación en orden descendente para proporcionar al usuario una lista de documentos que se consideran por el algoritmo de búsqueda como los más relevantes para la consulta.

En el sistema ilustrado, el motor 200 de búsqueda representa un motor de búsqueda de clasificación de distancia de clic, que considera la distancia de clic de un documento en la determinación de la puntuación de búsqueda del documento. La valoración de distancia de clic de un documento aprovecha la presencia del documento en un sitio estructurado jerárquicamente (véase la figura 3), midiendo la distancia desde la página principal del sitio al documento. En un caso, la distancia de clic desde la página principal es una medida de la importancia de la página, donde las páginas más cercanas en la jerarquía a la página principal se consideran más importantes que las páginas inferiores en la jerarquía. Sin embargo, pueden existir otros escenarios donde ocurre lo contrario, donde los documentos inferiores en la jerarquía se consideran más que las páginas superiores en la jerarquía. Por lo tanto, la distancia de clic se considera una medida de pertinencia independiente de la consulta, ya que valora la importancia general del documento en lugar de la consulta (por ejemplo, una función de clasificación dependiente de la consulta contaría el número de veces que un término de búsqueda aparece en un documento).

La figura 3 ilustra una gráfica de red a modo de ejemplo de acuerdo con la presente invención. La gráfica de red está compuesta de nodos (por ejemplo, 310) y bordes o enlaces (por ejemplo, 320). Los nodos (por ejemplo, 310) representan las páginas y otros recursos que están en la red que pueden devolverse como resultados a una consulta de búsqueda. Los enlaces (por ejemplo, 320) conectan entre sí cada una de estas páginas a través del uso de los enlaces de navegación enumerados en las páginas. Puede recopilarse un conjunto de información de enlace para cada página que puede usarse en el cálculo de la distancia de clic para una página específica.

En una realización, el nodo 330 representa la página de autoridad más alta o el nodo de raíz en la red para un grupo de documentos. La distancia de clic para las páginas restantes de la red puede calcularse desde el nodo 330. Por ejemplo, el nodo 340 tiene una distancia de clic de dos “clics” desde el nodo 330. Como se ha indicado anteriormente, “clics” se refiere al número de ramas atravesadas en la ruta más corta desde el nodo de autoridad más alta. Podrían haberse elegido otras rutas desde el nodo 330 para alcanzar el nodo 340, pero la distancia de clic se refiere a la ruta más corta.

La gráfica 300 de red se muestra con unos nodos que no se ajustan a un orden específico, y pueden ser similares a internet en ese aspecto. Con la falta de orden, la aplicabilidad de la distancia de clic para las páginas de clasificación puede ser difícil de conceptualizar. Sin embargo, a menudo, la red de páginas y de recursos se ajusta a un orden aplicado como se muestra a continuación en la figura 4.

La figura 4 ilustra una gráfica de red jerárquica a modo de ejemplo de acuerdo con la presente invención. La gráfica 400 de red jerárquica es similar a la gráfica 300 de red mostrada en la figura 3 porque también incluye nodos (por ejemplo, 410) y enlaces (por ejemplo, 420). Sin embargo, la gráfica 400 de red jerárquica se basa en la jerarquía inherente de un sitio estructurado o intranet. En consecuencia, la gráfica 400 de red jerárquica puede conceptualizarse como una estructura de árbol con ramas que se extienden desde un nodo raíz.

Para la gráfica 400 de red jerárquica, la aplicabilidad y el cálculo de la distancia de clic es más reconocible. Por ejemplo, el nodo 330 corresponde al nodo de autoridad más alta o nodo raíz del árbol. Por lo tanto, el nodo 340 tiene una distancia de clic asociada de 3, siendo 3 los clics o las navegaciones de usuario desde el nodo raíz. Dicho de otra manera, puesto que se requiere que un usuario atraviese 3 ramas del árbol para navegar desde el nodo 330 al nodo 340, la distancia de clic también es 3.

Las gráficas de red representadas en las figuras 3 y 4 son ejemplos de gráficas que se construyen en la memoria durante la indexación de los documentos para calcular la distancia de clic. La construcción de una gráfica durante la indexación permite incluir la distancia de clic entre las estadísticas de documentos almacenadas en el índice y usarlas para clasificar las páginas. A continuación, se describen procedimientos a modo de ejemplo para generar la propiedad de distancia de clic y usar la propiedad de distancia de clic en los documentos de clasificación en relación con las figuras 5 y 6.

La figura 5 ilustra un diagrama de flujo lógico de un procedimiento a modo de ejemplo para calcular la distancia de clic (CD) de acuerdo con la presente invención. El procedimiento 500 comienza en el bloque 502 donde se han indexado documentos en una red distribuida y se ha generado la gráfica de red. En una realización, la gráfica de red se genera a partir de los datos recopilados mediante un procedimiento en el que se recopila información de texto de enlace y de anclaje y se atribuye a documentos de destino específicos del anclaje. El procesamiento continúa en el bloque 504.

En el bloque 504, la gráfica de red se carga en la memoria. Esta gráfica de red es la representación estructural de la identificación del documento (por ejemplo, el ID del documento) y la información de enlace recopilada de la red. En las figuras 3 y 4, se muestran ejemplos de la gráfica de red. La gráfica de red representa los nodos o páginas de un sitio o de intranet. Cada nodo tiene una propiedad de distancia de clic asociada que tiene un valor o peso. En una realización, esta propiedad de distancia de clic se concatena en el extremo del ID de documento. El procesamiento continúa en el bloque 506.

En el bloque 506, se inicializan los valores de distancia de clic (CD) de los nodos. Los nodos de autoridad más alta se denominan nodos asignados. A estos nodos se les asigna un valor de distancia de clic de 0 (cero). Puede indicarse más de un nodo de alta autoridad para una sola gráfica de red. Por ejemplo, un administrador puede clasificar manualmente un conjunto de 100 nodos e designarlos como nodos de alta autoridad. Además, los nodos de alta autoridad no necesitan tener una distancia de clic de 0 (cero), puede asignarse cualquier número por un administrador. Cambiar la distancia de clic de los nodos de alta autoridad no altera el algoritmo restante, sino que simplemente proporciona un procedimiento para designar manualmente la importancia de un nodo. Por ejemplo, un administrador puede mejorar la puntuación de distancia de clic de algunos nodos. En otros casos, el administrador puede disminuir la puntuación de distancia de clic (haciendo que la distancia de clic sea mayor que la calculada por el algoritmo de manera predeterminada). La distancia de clic para cada uno de los nodos no asignados se inicializa en un valor máximo. En una realización, el valor máximo establece esencialmente el valor de distancia de clic en infinito. Asignar el valor de infinito a un nodo lo hace fácilmente reconocible como un nodo cuya distancia de clic no se ha calculado. Con las inicializaciones de los valores de distancia de clic completadas, el procesamiento se mueve al bloque 508.

En el bloque 508, los nodos que tienen una distancia de clic asociada distinta del valor máximo se insertan en una cola. En un ejemplo, esta etapa solo se produce en una primera iteración. Los nodos insertados en la cola corresponden a los nodos de autoridad más alta ya que sus valores de distancia de clic se establecen en 0 (cero), un valor distinto del valor máximo. Una vez que los nodos con valor de distancia de clic distinto del máximo se añaden a la cola, el procesamiento continúa en el bloque 510 de decisión.

En el bloque 510 de decisión, se realiza una determinación de si la cola está vacía. Una cola vacía significa que no hay más nodos que necesiten calcular la distancia de clic de sus nodos destino. Si la cola está vacía, el procesamiento se mueve al bloque 512 donde termina el procedimiento 500. Sin embargo, si la cola no está vacía, el procesamiento continúa en el bloque 514.

En el bloque 514, se elimina un nodo de la cola. La eliminación del nodo de la cola inicia el cálculo de las distancias de clic para los nodos destino asociados con ese documento. Los nodos destino corresponden a los documentos que tienen un enlace con los mismos desde un documento de origen. En este caso, el documento de origen es el documento correspondiente al nodo eliminado de la cola. Una vez eliminado este nodo, el procesamiento se mueve al bloque 516.

En el bloque 516, se obtiene el siguiente nodo destino. El siguiente nodo destino hace referencia al documento siguiente entre los documentos vinculados por el documento de origen. Una vez que se obtiene el siguiente nodo destino, el procesamiento continúa al bloque 518 de decisión.

En el bloque 518 de decisión, se realiza una determinación de si la distancia de clic asociada con el nodo destino es mayor que la distancia de clic de la página actual más uno (CD + 1). En una realización, la única forma en que se cumple la condición en el bloque 518 es cuando el nodo destino tiene una distancia de clic de infinito (suponiendo que el nodo de alta autoridad se establezca en cero y un administrador no haya establecido manualmente una distancia de clic). Por ejemplo, si la distancia de clic actual es 1, entonces CD + 1 = 2. Una distancia de clic de 2 es menor que infinito y se cumple la condición. Determinar si la distancia de clic destino es mayor que la distancia de clic más uno evita que se cambien los documentos destino con una distancia de clic menor. Usando el ejemplo anterior, si la distancia de clic del nodo destino es 1 y la distancia de clic actual también es 1, entonces la distancia de clic destino no es mayor que CD + 1 = 2. En este caso, la ruta más corta hacia el nodo destino ya se ha registrado y, por lo tanto, no necesita actualizarse. En consecuencia, cuando la distancia de clic destino no es mayor que la distancia de clic actual más uno, el procesamiento avanza al bloque 522 de decisión. Sin embargo, si la distancia de clic destino es mayor que la distancia de clic actual más uno, el procesamiento se mueve al bloque 520.

En el bloque 520, se actualiza el valor de distancia de clic del nodo destino y se añade el nodo destino a la cola como un nodo cuando es necesario hacer el cálculo de distancia de clic de sus destinos. El nodo destino se actualiza con un nuevo valor de distancia de clic para eliminar el valor de infinito y establecer el valor de distancia de clic calculado de los nodos. En una realización, el valor de distancia de clic del nodo se establece en el valor de distancia de clic actual más uno (CD + 1). El procesamiento continúa en el bloque 522 de decisión.

En el bloque 522 de decisión, se realiza una determinación de si se han obtenido todos los nodos destino para el nodo actual eliminado de la cola. Si hay nodos destino a obtener para el nodo actual, el procesamiento vuelve al bloque 516 donde se obtiene el siguiente nodo destino. Sin embargo, si se han obtenido todos los nodos destino correspondientes al nodo actual, el procesamiento vuelve al bloque 510 de decisión para volver a comprobar si la cola está vacía. De nuevo, una vez que la cola está vacía, el procesamiento se mueve al bloque 512, donde termina el procedimiento 500.

Es posible que no todos los nodos de una red estén conectados a los nodos de alta autoridad iniciales. En consecuencia, en otra realización de la presente invención, se supone que los nodos que no están conectados a los nodos de alta autoridad tienen una importancia baja y se les asigna una distancia de clic que es menor que la media para la gráfica de red.

La figura 6 ilustra un diagrama de flujo lógico de un procedimiento a modo de ejemplo para usar la distancia de clic en la clasificación de los resultados de búsqueda de acuerdo con la presente invención. El procedimiento 600 comienza en el bloque 602 donde se ha solicitado una consulta y se ha calculado la distancia de clic para cada uno de los documentos de la red. El procesamiento continúa en el bloque 604.

En el bloque 604, el valor de distancia de clic para cada uno de los documentos se combina con las otras estadísticas de documento (véase la figura 2) en el índice. La combinación de los valores de distancia de clic con las otras estadísticas de documento permite un tiempo de respuesta de consulta más rápido ya que se agrupa toda la información relacionada con la clasificación. En consecuencia, cada documento enumerado en el índice tiene un valor de distancia de clic asociado después de la combinación. Una vez completada la combinación, el procesamiento se mueve al bloque 606.

En el bloque 606, una función de puntuación se rellena con el conjunto de estadísticas de documento, incluyendo la distancia de clic, para calcular una puntuación de un documento específico. La distancia de clic proporciona un factor independiente de la consulta a la función de puntuación. La otra parte de la función de puntuación corresponde a la parte dependiente de la consulta o relacionada con el contenido de la función de puntuación. En

una realización, la función de puntuación es una suma de funciones de puntuación dependientes de la consulta (QD) e independientes de la consulta (QID):

$$Puntuación = QD(doc, consulta) + QID(doc) \quad (1)$$

La función QD puede ser cualquier función de puntuación de documento. En una realización, la función de puntuación QD corresponde a la función de puntuación ponderada de campo descrita en la solicitud de patente número de serie 10/804.326, titulada "Field Weighting in Text Document Searching", presentada el 18 de marzo de 2004 e incorporada por referencia en el presente documento. Tal como se proporciona en la solicitud de patente 10/804.326, la siguiente es una representación de la función de puntuación ponderada de campo:

$$QD(doc, consulta) = \sum \frac{wtf(k_1 + 1)}{k_1((1-b) + b\frac{wdl}{avwdl}) + wtf} \times \log(\frac{N}{n}) \quad (2)$$

en donde los términos se definen de la siguiente manera: wtf es la frecuencia de término ponderada o la suma de las frecuencias de término de unos términos dados multiplicados por los pesos a través de todas las propiedades; wdl es la longitud de documento; avwdl es la longitud de documento ponderada promedio; N es el número de documentos en la red (es decir, el número de documentos rastreados); n es el número de documentos que contienen el término de consulta dado; y k_1 y b son unas constantes. Estos términos y la ecuación anterior se describen en detalle en la solicitud de patente 10/804.326.

La función QID puede ser cualquier transformación de la distancia de clic y otras estadísticas de documento (tales como la profundidad de URL). En una realización esta función es la siguiente:

$$QID(doc) = \sum w_{cd} \frac{k_{cd}}{k_{cd} + \frac{b_{cd}CD + b_{ud}UD}{b_{cd} + b_{ud}}} \quad (3)$$

en donde los términos de la función se definen de la siguiente manera: w_{cd} es el peso del componente independiente de la consulta; b_{cd} es el peso de la distancia de clic; b_{ud} es el peso de la profundidad de URL; CD es la distancia de clic; UD es la profundidad de URL; y K_{cd} es la constante de saturación de distancia de clic. Los términos ponderados (w_{cd}, b_{cd}, y b_{ud}) ayudan a definir la importancia de cada uno de sus términos relacionados y, finalmente, la forma de las funciones de puntuación. La profundidad de URL (UD) se añade al componente independiente de la consulta para suavizar el efecto de la distancia de clic en la función de puntuación. En algunos casos, un documento que no es muy importante (es decir, tiene una gran profundidad de URL) puede tener una corta distancia de clic. La profundidad de URL cuenta el número de barras en la URL de un documento. Por ejemplo, www.example.com\d1\d2\d3\d4.htm incluye cuatro barras y, por lo tanto, tendría una profundidad de URL de 4. Sin embargo, este documento puede tener un enlace directo desde la página principal www.example.com lo que da una distancia de clic de 1. Incluyendo el término de profundidad de URL en la función (3) y ponderándolo con respecto a la distancia de clic, se compensa la alta puntuación de distancia de clic para reflejar con mayor precisión la clasificación de la página dentro de la jerarquía. Dependiendo de la red, una profundidad de URL de 3 o más puede considerarse un enlace profundo. Para esta realización, la presente invención añade las dos funciones de (2) y (3) para recibir la función de puntuación (puntuación), de tal manera que la nueva función de puntuación se convierte en:

$$Puntuación = \sum \frac{wtf(k_1 + 1)}{k_1((1-b) + b\frac{wdl}{avwdl}) + wtf} \times \log(\frac{N}{n}) + w_{cd} \frac{k_{cd}}{k_{cd} + \frac{b_{cd}CD + b_{ud}UD}{b_{cd} + b_{ud}}} \quad (4)$$

En otras realizaciones, la profundidad de URL puede eliminarse de la función de puntuación o pueden añadirse otros factores a la función de puntuación para mejorar la precisión o del componente dependiente de la consulta o del componente independiente de la consulta. Además, el componente independiente de la consulta puede incorporarse en otras funciones de clasificación no mostradas para mejorar los resultados de clasificación. Una vez que la función de puntuación (4) se rellena con las estadísticas de documento para un documento específico, el procedimiento avanza al bloque 608.

En el bloque 608, se ejecuta la función de puntuación y se calcula la puntuación de pertinencia para el documento. Una vez que se calcula la puntuación de pertinencia, se almacena en la memoria y se asocia con ese documento específico. A continuación, el procesamiento se mueve al bloque 610 de decisión.

En el bloque 610 de decisión, se realiza una determinación de si se han calculado las puntuaciones de pertinencia de todos los documentos de acuerdo con la función (4) de puntuación. Las puntuaciones pueden calcularse en serie, como se muestra, o en paralelo. Si no se han calculado todas las puntuaciones, el procesamiento vuelve al bloque 606 donde la función de puntuación se rellena con el siguiente conjunto de estadísticas de documento. Sin embargo, si se han calculado todas las puntuaciones, el procesamiento continúa en el bloque 612.

En el bloque 612, los resultados de búsqueda de la consulta se clasifican de acuerdo con sus puntuaciones correspondientes. Las puntuaciones ahora tienen en cuenta la distancia de clic y la profundidad de URL de cada uno de los documentos. En consecuencia, se ha perfeccionado la clasificación de los documentos de manera que los documentos más altos en la jerarquía de una intranet o un sitio se clasifican más alto que los otros documentos donde todos los demás factores son los mismos. Una vez que se clasifican los resultados de búsqueda, el procesamiento avanza al bloque 614, donde termina el procedimiento 600.

Después de que se ha completado el procedimiento 600, los documentos clasificados pueden devolverse al usuario mediante las diversas operaciones asociadas con la transmisión y la visualización de los resultados por un motor de búsqueda. A continuación, los documentos correspondientes a los resultados de mayor precisión pueden seleccionarse y verse a discreción por el usuario.

La memoria descriptiva, los ejemplos y los datos anteriores proporcionan una descripción completa de la fabricación y el uso de la composición de la invención.

REIVINDICACIONES

1. Un procedimiento implementado por ordenador para clasificar resultados de búsqueda, que comprende:

almacenar (502) información de documentos y de enlaces para documentos (210, 212, 214, 216) en una red;
generar (504) una representación de la red a partir de la información de documentos y de enlaces almacenada, incluyendo la representación de la red unos nodos (310, 320, 330, 340) que representan los documentos y en el que más de un nodo dentro de la representación de la red se designa como un nodo de alta autoridad;
inicializar (506) los valores de distancia de clic para los nodos, incluyendo:

asignar a cada nodo designado como un nodo de alta autoridad un valor de distancia de clic establecido por un administrador; e
inicializar el valor de distancia de clic de cada nodo no asignado en un valor máximo;

calcular (520) una distancia de clic para cada uno de los nodos en la representación de la red, siendo la distancia de clic para un nodo de alta autoridad dado el valor de distancia de clic establecido por el administrador durante la inicialización, y midiéndose la distancia de clic para un nodo dado no designado como un nodo de alta autoridad a partir del nodo de alta autoridad más próximo al nodo dado; y
usar (612) la distancia de clic calculada asociada con cada uno de los documentos como una medida de pertinencia independiente de la consulta en la clasificación de los documentos para producir los resultados de búsqueda clasificados.

2. El procedimiento implementado por ordenador de la reivindicación 1, en el que generar una representación de la red comprende, además, generar una gráfica (300, 400) de red y almacenar la gráfica de red en la memoria.

3. El procedimiento implementado por ordenador de la reivindicación 1, que comprende, además, almacenar un nodo actual de la representación de la red en una cola de nodos hasta que se calcula la distancia de clic de los nodos destino asociados con el nodo actual.

4. El procedimiento implementado por ordenador de la reivindicación 3, en el que la distancia de clic de uno de los nodos destino se establece en la distancia de clic del nodo actual más una variable cuando la distancia de clic del uno de los nodos destino es mayor que la distancia de clic del nodo actual más la variable.

5. El procedimiento implementado por ordenador de la reivindicación 1, en el que la distancia de clic calculada asociada con cada uno de los documentos se combina con un índice que incluye otras estadísticas que corresponden a cada uno de los documentos.

6. El procedimiento implementado por ordenador de la reivindicación 5, en el que una función de puntuación se rellena con la distancia de clic calculada y las otras estadísticas para producir una puntuación por la que se clasifican los documentos.

7. El procedimiento implementado por ordenador de la reivindicación 1, en el que usar la distancia de clic calculada asociada con cada uno de los documentos como una medida de pertinencia independiente de la consulta comprende, además, usar un componente correspondiente a la distancia de clic en una función de puntuación para determinar una puntuación de pertinencia para cada uno de los documentos.

8. El procedimiento implementado por ordenador de la reivindicación 7, en el que la puntuación de pertinencia se compensa con una propiedad de profundidad de localizador de recursos uniforme que suaviza el efecto de la distancia de clic en la puntuación de pertinencia.

9. El procedimiento implementado por ordenador de la reivindicación 1, que comprende, además, permitir que la distancia de clic se cambie manualmente después que se calcule la distancia de clic.

10. El procedimiento implementado por ordenador de la reivindicación 1, que comprende, además, clasificar los documentos de acuerdo con una función de puntuación, puntuación, que se determina de acuerdo con al menos: la distancia de clic calculada, CD, un peso de un componente independiente de la consulta, w_{cd}, un peso de la distancia de clic, b_{cd}, un peso de una profundidad de URL, b_{ud}, la profundidad de URL, UD, y una constante de saturación de distancia de clic, K_{cd}.

11. El procedimiento implementado por ordenador de la reivindicación 1, que comprende, además, clasificar los documentos de acuerdo con una función de puntuación, puntuación, que se determina de acuerdo con al menos: la distancia de clic calculada, CD, una frecuencia de término ponderada, wtf, una longitud de documento ponderada, wdl, una longitud de documento ponderada promedio, avwdl, un número de documentos en la red, N; un número de documentos que contienen un término de consulta, n, un peso de un componente independiente de la consulta, w_{cd}, un peso de la distancia de clic, b_{cd}, un peso de una profundidad de URL, b_{ud}, la profundidad de URL, UD, una constante de saturación de distancia de clic, K_{cd}, y otras constantes, k_1, b.

12. El procedimiento implementado por ordenador de la reivindicación 11, en el que la función de puntuación, puntuación, está dada por:

$$puntuación = \sum \frac{wtf(k_1+1)}{k_1((1-b)+b\frac{wdl}{avwdl})+wtf} \times \log(\frac{N}{n}) + w_{cd}\frac{k_{cd}}{k_{cd}+\frac{b_{cd}CD+b_{ud}UD}{b_{cd}+b_{ud}}}$$

13. Un sistema para clasificar resultados de búsqueda, que comprende:

un motor (200) de búsqueda incluido en un dispositivo (100) informático, estando el motor de búsqueda configurado para ejecutar instrucciones ejecutables por ordenador, comprendiendo las instrucciones ejecutables por ordenador:

descubrir documentos (210, 212, 214, 216) en una red;
registrar información de documentos y de enlaces para cada uno de los documentos en la red;
generar una representación de la red a partir de la información de documentos y de enlaces registrada, en el que la representación de la red incluye unos nodos (310, 320, 330, 340) que representan los documentos, y en el que más de un nodo dentro de la representación de la red se designa como un nodo de alta autoridad;
inicializar los valores de distancia de clic para los nodos, incluyendo:

asignar a cada nodo designado como un nodo de alta autoridad un valor de distancia de clic establecido por un administrador; e
inicializar el valor de distancia de clic de cada nodo no asignado en un valor máximo;

calcular una distancia de clic para cada uno de los nodos en la representación de la red, en el que la distancia de clic para un nodo de alta autoridad dado es el valor de distancia de clic establecido por el administrador durante la inicialización, y midiéndose la distancia de clic para un nodo dado no designado como un nodo de alta autoridad a partir del nodo de autoridad más próximo al nodo dado;
asociar la distancia de clic calculada para cada nodo con el documento que corresponde a ese nodo; y
usar la distancia de clic calculada asociada con cada uno de los documentos como una medida de pertinencia independiente de la consulta en la clasificación de los documentos para producir los resultados de búsqueda clasificados.

14. El sistema de la reivindicación 13, en el que generar una representación de la red comprende, además, generar una gráfica de red y almacenar la gráfica de red en la memoria.

15. El sistema de la reivindicación 13, en el que asociar la distancia de clic calculada a cada nodo con el documento que corresponde a ese nodo comprende, además, combinar la distancia de clic calculada asociada con cada uno de los documentos con un índice que incluye otros valores de clasificación que corresponden a cada uno de los documentos.

16. El sistema de la reivindicación 15, en el que una función de puntuación se rellena con la distancia de clic calculada y los otros valores de clasificación para producir una puntuación por la que se clasifican los documentos.

17. El sistema de la reivindicación 13, en el que usar la distancia de clic calculada asociada con cada uno de los documentos como una medida de pertinencia independiente de la consulta comprende, además, usar un componente correspondiente a la distancia de clic en una función de puntuación para determinar una puntuación de pertinencia para cada uno de los documentos.

18. El sistema de la reivindicación 17, en el que la puntuación de pertinencia se compensa con una propiedad de profundidad de localizador de recursos uniforme que suaviza el efecto de la distancia de clic en la función de puntuación cuando la distancia de clic para un nodo es desproporcionada con respecto a la profundidad del nodo en la representación de la red.

19. El sistema de la reivindicación 13, que comprende, además, clasificar los documentos de acuerdo con una función de puntuación, puntuación, que se determina de acuerdo con al menos: la distancia de clic calculada, CD, un peso de un componente independiente de la consulta, w_{cd}, un peso de la distancia de clic, b_{cd}, un peso de una profundidad de URL, b_{ud}, la profundidad de URL, UD, y una constante de saturación de distancia de clic, K_{cd}.

20. El sistema de la reivindicación 13, que comprende, además, clasificar los documentos de acuerdo con una función de puntuación, puntuación, que se determina de acuerdo con al menos: la distancia de clic calculada, CD, una frecuencia de término ponderada, wtf, una longitud de documento ponderada, wdl, una longitud de documento ponderada promedio, avwdl, un número de documentos en la red, N; un número de documentos que contienen un término de consulta, n, un peso de un componente independiente de la consulta, w_{cd}, un peso de la distancia de clic, b_{cd}, un peso de una profundidad de URL, b_{ud}, la profundidad de URL, UD, una constante de saturación de distancia de clic, K_{cd}, y otras constantes, k_1, b.

21. El sistema de la reivindicación 20, en el que la función de puntuación, puntuación, está dada por:

$$puntuación = \sum \frac{wtf(k_1+1)}{k_1((1-b)+b\frac{wdl}{avwdl})+wtf} \times \log(\frac{N}{n}) + w_{cd}\frac{k_{cd}}{k_{cd}+\frac{b_{cd}CD+b_{ud}UD}{b_{cd}+b_{ud}}}$$

22. El sistema de la reivindicación 13, en el que generar la representación de la red comprende, además, generar una representación de la red, designándose más de un nodo dentro de la representación de la red como un nodo de alta autoridad.

23. El sistema de la reivindicación 13, que comprende, además, permitir que la distancia de clic se cambie manualmente después que se calcule la distancia de clic.

24. Un medio legible por ordenador que incluye instrucciones ejecutables por ordenador para clasificar resultados de búsqueda, comprendiendo las instrucciones:

almacenar información de documentos y de enlaces para documentos (210, 212, 214, 216) en una red, de tal manera que una gráfica (300, 400) de red que representa la red se genera en la memoria, designándose más de un nodo dentro de la representación de la red como un nodo de alta autoridad;
inicializar los valores de distancia de clic para los nodos, incluyendo:

asignar a cada nodo designado como un nodo de alta autoridad un valor de distancia de clic establecido por un administrador; e
inicializar el valor de distancia de clic de cada nodo no asignado en un valor máximo;

almacenar cada documento representado en la gráfica de red en una cola cuando el documento tiene un valor de distancia de clic que es diferente del valor máximo;

cuando la cola no está vacía:

eliminar un documento de la cola,
calcular una distancia de clic para cada documento destino asociado con el documento eliminado, en el que cada documento destino se actualiza con un nuevo valor de distancia de clic distinto del valor máximo cuando cada distancia de clic del documento destino es mayor que la distancia de clic asociada con el documento eliminado más una variable, correspondiendo el nuevo valor de distancia de clic de un documento destino dado a un nodo dado que se mide a partir del nodo de alta autoridad más próximo al nodo dado, y
añadir cada uno de los documentos destino a la cola que se ha actualizado; y
usar la distancia de clic calculada asociada con cada uno de los documentos como una medida de pertinencia independiente de la consulta en la clasificación de los documentos para producir los resultados de búsqueda clasificados.

25. El medio legible por ordenador de la reivindicación 24, en el que usar la distancia de clic calculada asociada con cada uno de los documentos como una medida de pertinencia independiente de la consulta comprende, además, usar un componente correspondiente a la distancia de clic en una función de puntuación para determinar una puntuación de pertinencia para cada uno de los documentos.

26. El medio legible por ordenador de la reivindicación 24, en el que generar la representación de la red comprende, además, generar una representación de la red, designándose más de un nodo dentro de la representación de la red como un nodo de alta autoridad.

27. El medio legible por ordenador de la reivindicación 24, que comprende además permitir que la distancia de clic se cambie manualmente después de que se calcule la distancia de clic.

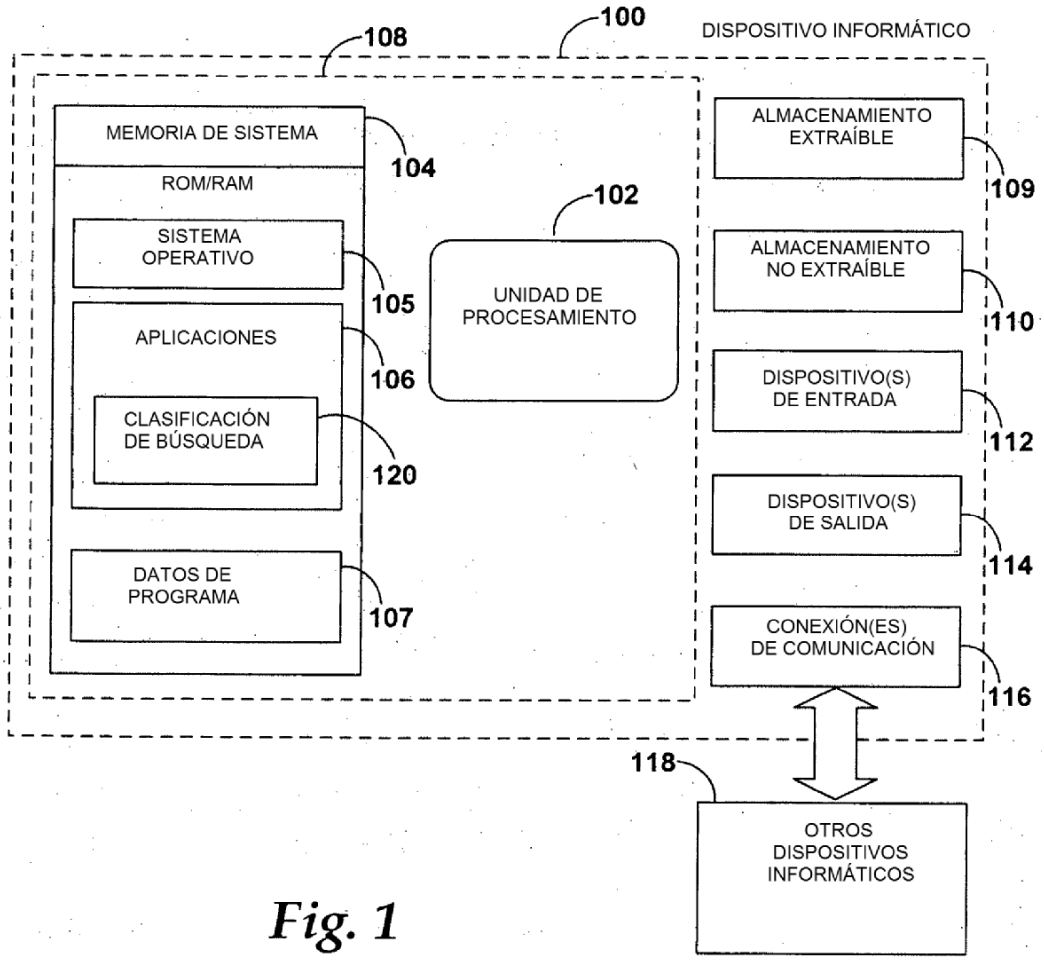
100
DISPOSITIVO INFORMÁTICO
108
MEMORIA DE SISTEMA
104
ROM/RAM
SISTEMA OPERATIVO
105
APLICACIONES
106
CLASIFICACIÓN DE BÚSQUEDA
120
DATOS DE PROGRAMA
107
102
UNIDAD DE PROCESAMIENTO
ALMACENAMIENTO EXTRAÍBLE
109
ALMACENAMIENTO NO EXTRAÍBLE
110
DISPOSITIVO(S) DE ENTRADA
112
DISPOSITIVO(S) DE SALIDA
114
CONEXIÓN(ES) DE COMUNICACIÓN
116
118
OTROS DISPOSITIVOS INFORMÁTICOS

Fig. 1

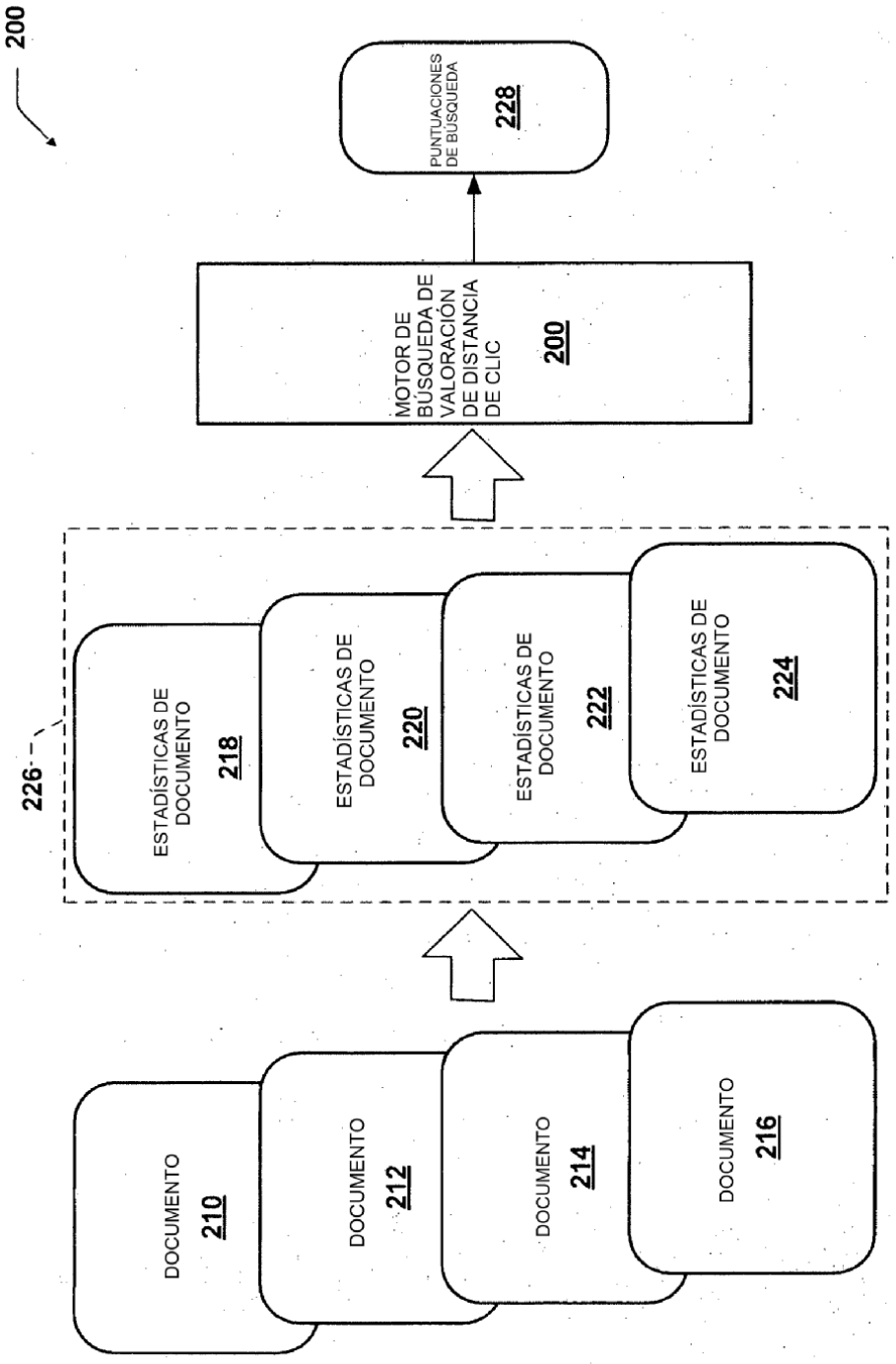
200
PUNTUACIONES DE BÚSQUEDA
228
MOTOR DE BÚSQUEDA DE VALORACIÓN DE DISTANCIA DE CLIC
200
226
ESTADÍSTICAS DE DOCUMENTO
218
ESTADÍSTICAS DE DOCUMENTO
220
ESTADÍSTICAS DE DOCUMENTO
222
ESTADÍSTICAS DE DOCUMENTO
224
DOCUMENTO
210
DOCUMENTO
212
DOCUMENTO
214
DOCUMENTO
216

Fig. 2

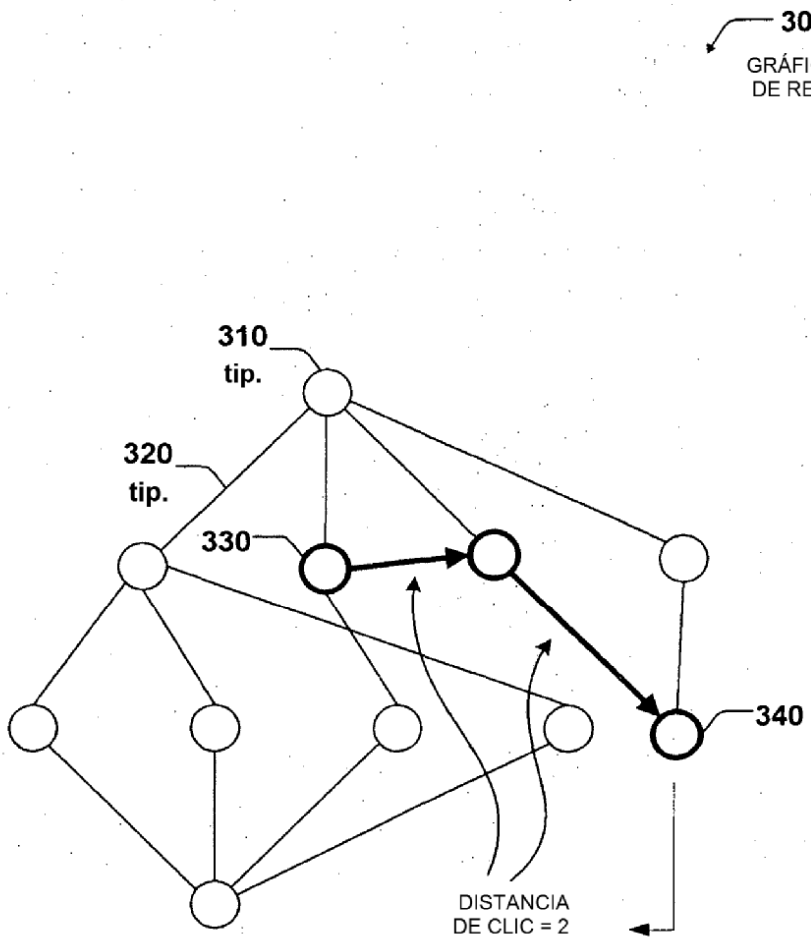

Fig. 3

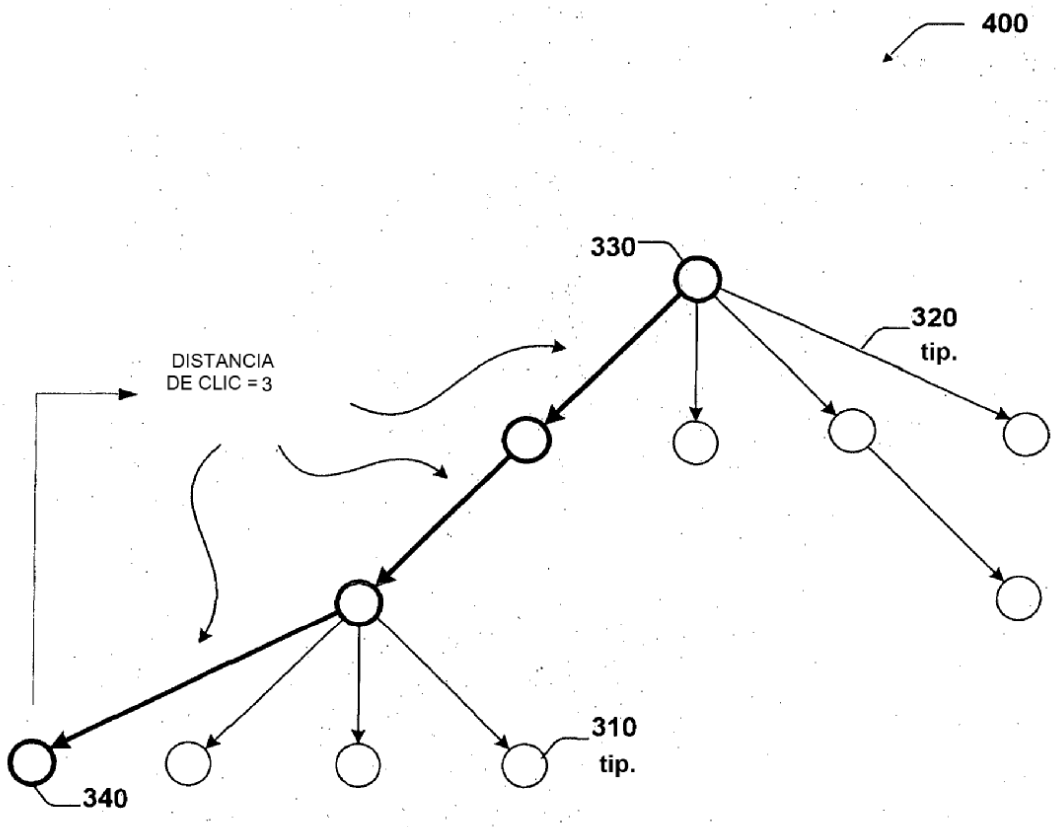

Fig. 4

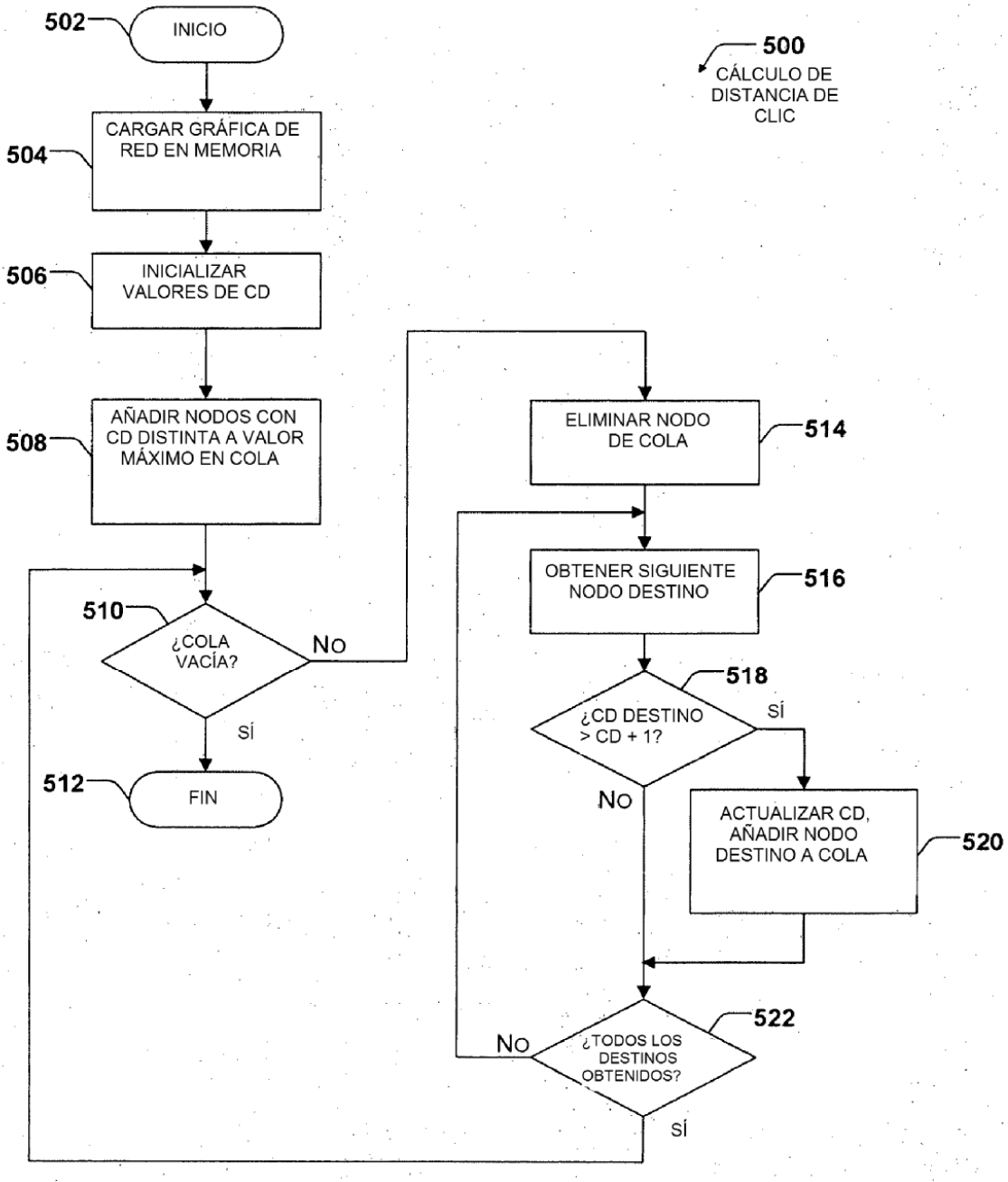
502
INICIO
500
CÁLCULO DE DISTANCIA DE CLIC
504
CARGAR GRÁFICA DE RED EN MEMORIA
506
INICIALIZAR VALORES DE CD
508
AÑADIR NODOS CON CD DISTINTA A VALOR MÁXIMO EN COLA
510
¿COLA VACÍA?
No
SÍ
512
FIN
514
ELIMINAR NODO DE COLA
516
OBTENER SIGUIENTE NODO DESTINO
518
¿CD DESTINO > CD + 1?
SÍ
No
520
ACTUALIZAR CD, AÑADIR NODO DESTINO A COLA
522
¿TODOS LOS DESTINOS OBTENIDOS?
No
SÍ

Fig. 5

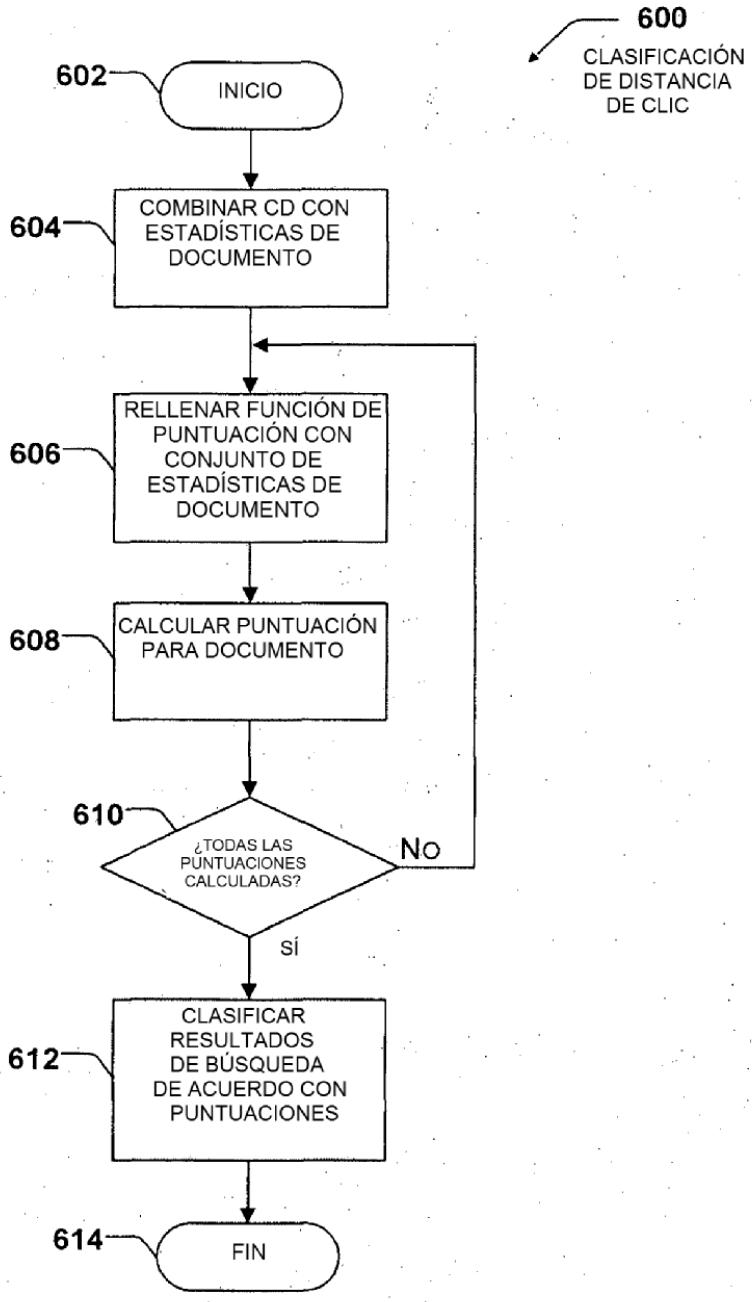
600
CLASIFICACIÓN DE DISTANCIA DE CLIC
602
INICIO
604
COMBINAR CD CON ESTADÍSTICAS DE DOCUMENTO
606
RELLENAR FUNCIÓN DE PUNTUACIÓN CON CONJUNTO DE ESTADÍSTICAS DE DOCUMENTO
608
CALCULAR PUNTUACIÓN PARA DOCUMENTO
610
¿TODAS LAS PUNTUACIONES CALCULADAS?
NO
SÍ
612
CLASIFICAR RESULTADOS DE BÚSQUEDA DE ACUERDO CON PUNTUACIONES
614
FIN

Fig. 6