



19



OFICINA ESPAÑOLA DE
PATENTES Y MARCAS

ESPAÑA

11 Número de publicación: **2 306 474**

51 Int. Cl.:
G06F 17/30 (2006.01)

12

TRADUCCIÓN DE PATENTE EUROPEA

T3

86 Número de solicitud europea: **98920069 .6**

86 Fecha de presentación : **29.04.1998**

87 Número de publicación de la solicitud: **0979470**

87 Fecha de publicación de la solicitud: **16.02.2000**

54 Título: **Método y dispositivo para buscar una base de datos de registros.**

30 Prioridad: **01.05.1997 US 846850**

45 Fecha de publicación de la mención BOPI:
01.11.2008

45 Fecha de la publicación del folleto de la patente:
01.11.2008

73 Titular/es: **Northern Light Technology, L.L.C.**
Suite 342, 222 Third Street
Cambridge, Massachusetts 02142, US

72 Inventor/es: **Krellenstein, Marc, F.**

74 Agente: **Elzaburu Márquez, Alberto**

ES 2 306 474 T3

Aviso: En el plazo de nueve meses a contar desde la fecha de publicación en el Boletín europeo de patentes, de la mención de concesión de la patente europea, cualquier persona podrá oponerse ante la Oficina Europea de Patentes a la patente concedida. La oposición deberá formularse por escrito y estar motivada; sólo se considerará como formulada una vez que se haya realizado el pago de la tasa de oposición (art. 99.1 del Convenio sobre concesión de Patentes Europeas).

DESCRIPCIÓN

Método y dispositivo para buscar una base de datos de registros.

5 Campo del invento

El invento se refiere generalmente a un método y dispositivo para buscar una base de datos de registros. Más particularmente, el invento se refiere a un método y aparato de búsqueda para buscar una base de datos que comprende tanto información de contenido de Internet como información de interés.

10 Antecedentes del invento

Internet atrae a millones de usuarios cada día. Se ha estimado que el número de usuarios de Internet crecería de 10 millones a finales de 1995 a 170 millones en el año 2000. La atracción principal de Internet es la promesa de enormes cantidades de información disponible sobre cualquier tema imaginable interesante. La investigación ha mostrado que los principales usos de Internet por los usuarios incluyen la búsqueda de información y la navegación (una forma de búsqueda) en busca de información.

Varias compañías ofrecen servicios de búsqueda para ayudar en la búsqueda de los datos masivos, rápidamente crecientes e infinitamente distribuidos en Internet. Un gran número de usuarios de Internet utilizan un servicio de búsqueda varias veces a la semana, y el veinte por ciento superior de usuarios de Internet utilizan un dispositivo de búsqueda varias veces al día.

Sin embargo, Internet no está exenta de fallos. Mientras que hay 250 gigabytes de información de texto en Internet accesible al público, muchos usuarios de Internet se sienten frustrados en su búsqueda de información en cuanto a que: (1) la información de calidad no se encuentra a menudo en Internet; (2) la información de calidad existe pero está dispersa en sitios de uso privado accesible mediante suscripción; (3) los servicios de búsqueda producen demasiada o muy poca información; y (4) los servicios de búsqueda no se anticipan a las solicitudes del usuario.

Internet es una excelente fuente del tipo de información encontrada en los catálogos de los productos. Sin embargo, Internet es una fuente notablemente pobre de información editorial, de información y de comentarios de referencia. Una razón de este inconveniente es que la información de calidad (es decir, el contenido de información de interés) la mayoría de las veces es creado y proporcionado por compañías que cobran por la información (es decir, propietarios de contenidos de interés). La tradición de la información gratis en Internet ha impedido que el propietario de contenidos de interés haga su información disponible a través de Internet. Otra razón ha sido la importante inversión financiera y de capital requerida para desarrollar el mercado y para mantener un contenido de interés en Internet. Los observadores de la industria no tienen claro qué modelos de negocio se materializarán finalmente para producir beneficios razonables al contenido de interés disponible en Internet. Como consecuencia de estos factores Internet no es actualmente considerada como una fuente principal del contenido más valorado sobre cualquier tema.

A pesar de las anteriores razones, algunos poseedores de contenidos de interés han comenzado a poner su información disponible en Internet, normalmente en forma de servicios de suscripción. Sin embargo, estos servicios tienen numerosos problemas y por lo tanto no siempre son una buena solución para los usuarios de Internet.

Un problema con respecto a los servicios de suscripción es que un usuario tiene que realizar muchas búsquedas y buscar muchos sitios (que a menudo incluyen muchas bases de datos en los sitios) para obtener información completa sobre el tema en cuestión. Para un resultado verdaderamente sólido los usuarios a menudo utilizan un dispositivo de búsqueda que puede generar volúmenes de información desde Internet. Con una forma no fácil de fusionar la información dada, los usuarios consideran el proceso demasiado engorroso y muy largo para ser ventajoso. Otro problema es que los usuarios pueden incurrir en grandes gastos en la contratación de muchos servicios de suscripción para satisfacer sus necesidades en cada área de tema de interés. Mientras que los usuarios normalmente tienen intereses variables, muchos se resisten a contratar muchas suscripciones sobre muchos temas. Otro problema es que los usuarios son requeridos para anticipar sus deseos para preguntar sobre un determinado tema con el fin de tener por adelantado todas las suscripciones necesarias. En realidad, los intereses de de información de muchos usuarios son específicas y de corta duración. Los servicios de abono no pueden satisfacer este tipo de necesidad de información al usuario.

Cuando un usuario accede a uno de los dispositivos de búsqueda destacados, la búsqueda puede producir cientos, incluso miles, de éxitos (es decir, registros). Por ejemplo, el dispositivo de búsqueda Alta Vista™ da cientos de miles de éxitos en respuesta a una búsqueda en el tema "ventanas". Este diluvio de información es a menudo demasiado material que revisar, escoger y seleccionar. Este problema está agravado por el fallo del dispositivo de búsqueda en agrupar los éxitos en la lista de resultados de búsqueda en cualquier forma significativa. En el ejemplo anterior, la información del producto de soporte lógico Windows™ 95 sería incluida junto con las ventanas arquitectónicas y las páginas personales en la lista de resultados de búsqueda. También, muchos de los dispositivos de búsqueda destacados ven cada página html como un éxito independiente, de forma que cien sitios de páginas Web pueden producir cien éxitos en la lista de resultados de búsqueda. Para hacer frente a este problema algunos dispositivos de búsqueda agrupan éxitos por sitio web.

Muchos dispositivos de búsqueda destacados usan rutinas primitivas de clasificación por importancia que dan lugar a listas de resultados de búsqueda con poca o ninguna clasificación de interés. Las listas de resultados de búsqueda mal clasificados son un problema importante para los consumidores. Si una búsqueda produce cien éxitos, el usuario debe navegar a través de veinte pantallas de información para encontrar la información más interesante. Se ha visto que la mayoría de los usuarios abandonan después de las primeras pantallas. Por lo tanto, caso de que la información más importante se encuentre en una pantalla posterior la mayoría de los usuarios nunca lo sabrán y dirán que la búsqueda fue un fallo.

Dos de los dispositivos de búsqueda destacados, Excite™ y Yahoo™, clasifican e indexan manualmente Internet. Este método produce índices de alta calidad y una apropiada clasificación de los sitios Web en la estructura del directorio. Sin embargo, el personal de redacción de estas compañías se encuentra a sí mismo en una competición que tiene perdida con el crecimiento de Internet. Incluso con un personal de cientos de redactores, estas compañías no pueden visitar suficientes sitios Web y no pueden volver a visitar cada sitio cada vez que cambia el sitio. Por lo tanto, estas compañías son incapaces de cubrir un gran porcentaje de Internet. Como resultado, las búsquedas que usan estos dispositivos de búsqueda pueden a menudo conseguir “demasiado poca” información útil.

El Hypertext 96 de Weiss R y otros, la 7ª Conferencia ACM sobre Hypertext, Washington, Marzo 16-20, 1996, 16 Marzo 1996, páginas 180-193. “Hypersuit: Un dispositivo de búsqueda de red jerárquica que explota la formación de agrupaciones de unidades de Hypertext de enlace de contenido” expone un dispositivo de búsqueda denominado Hypersuit que genera una lista de registros de una base de datos después de una petición de búsqueda, formando categorías candidatas para presentar dichos registros basadas en características asociadas a dichos registros generados. Sin embargo, en Hypersuit las categorías candidatas son agrupamientos de unidades formados iterativamente, agrupando/fusionando los dos registros/subagrupaciones más similares basadas en la similaridad de términos que contienen y la similaridad de las estructuras de hipervínculo, terminando la iteración cuando dichos grupos alcanzan un número predeterminado de agrupaciones de unidades esperado. La función de similaridad de términos por pares usa una función de ponderación de términos con el fin de evaluar si dos registros/agrupaciones deberían fusionarse conjuntamente, estando dicha función basada en la frecuencia de términos, factores de tamaño de documentos y atributos de términos (similar a metadatos), esto es, en características comunes a los dos documentos. Se escoge una etiqueta representativa resumen de la agrupación resultante, basándose en dicha ponderación. En Hypersuit se selecciona una agrupación candidata intermedia como agrupación final debido a que se ha alcanzado el número máximo de agrupaciones; no selecciona finalmente una agrupación debido a que tiene una ponderación muy alta en comparación con el de otra agrupación.

“Sense Maker: An Information-Exploration Interface Supporting the Contextual Evolution of a User’s Interests”, CHI 97, 22-27 Marzo 1997, páginas 11-18, de W. Baldano y T. Winograd, expone la herramienta del dispositivo de búsqueda SenseMaker, en la que un usuario introduce primero una petición de búsqueda en un lenguaje general de unidad de entrada. Los registros que coinciden con la petición son devueltos al usuario como conjuntos en una tabla de acuerdo con una vista dada, por ejemplo, por defecto, cada registro es organizado de acuerdo con su fuente URL característica: los conjuntos corresponden a grupos de registros seleccionados que tienen el mismo URL. El usuario puede alterar la presentación a otra vista, por ejemplo basada en el título unido a los registros; vistas diferentes pueden ser iterativamente cambiadas por el usuario.

“Navigation with a Web Compass”, de R. Baldazo, Byte, Vol. 21, Nº 3, 1 Marzo 1996, páginas 97-98, expone un dispositivo de metabúsqueda, el Web Compass, que consigue resultados de registros de diferentes dispositivos de búsqueda de acuerdo con una petición de usuario. El Web Compass presenta dichos registros al usuario en grupos de documentos similares, siendo obtenidos dichos grupos basado en una técnica AI denominada agrupamiento conceptual, que usa clasificaciones de frases.

Resumen del invento

El presente invento ofrece un método y dispositivo para buscar una base de datos que puede incluir registros de Internet y de contenidos de interés. El invento proporciona a los usuarios un acceso a la riqueza de información en Internet y a información de contenido de interés no en Internet. El invento usa métodos de categorización sofisticados además de criterios de importancia detallados para proporcionar una lista significativa de resultados de búsqueda en forma de un conjunto de categorías de resultados de búsqueda. Al usuario se le presenta un pequeño número de categorías junto con una lista de los registros más importantes. Cada categoría puede incluir categorías más reducidas y/o una lista de los registros más importantes. Organizando la lista de búsqueda en una jerarquía los usuarios pueden rápidamente enfocar la búsqueda hacia los pocos registros de interés sin verse abrumado por los resultados.

En un aspecto el invento incorpora un método para buscar una base de datos de registros tal como se reivindica en la reivindicación 1. La base de datos puede incluir Internet y registros de contenidos de interés. En respuesta a una instrucción de búsqueda por parte de un usuario, se busca la base de datos y se genera una lista de resultados de la búsqueda que incluye un conjunto seleccionado de los registros. Una parte de la lista de resultados de la búsqueda es procesada para crear dinámicamente un conjunto de categorías resultado de la búsqueda. A modo de ejemplo, la parte de la lista de resultados de búsqueda pueden ser los doscientos (o cien) primeros registros más importantes dentro del conjunto de registros seleccionados. Cada categoría de resultados de búsqueda se asocia con un subconjunto de los registros dentro de la lista de resultados de búsqueda.

El invento usa una metodología de categorización (o agrupación para recuperar registros almacenados en la base de datos para recopilar la lista de resultados de la búsqueda. La metodología tiene tres pasos principales: identificar las categorías candidatas, ponderar las categorías candidatas y visualizar un conjunto de categorías de resultados de la búsqueda seleccionados de las categorías candidatas.

Cada registro dentro de la lista de búsqueda puede tener características asociadas de tema, tipo, fuente e idioma. Se identifican las características comunes asociadas con los registros, y los registros que tienen características comunes se agrupan en categorías candidatas. Se recopila una lista de categorías candidatas, que son representativas de posibles resultados de búsquedas. Se pondera cada categoría candidata como una función de las características comunes identificadas de los registros dentro de esa categoría candidata. Se seleccionan una o más categorías candidatas como una función de las características comunes identificadas de los registros. Por ejemplo, se puede seleccionar aproximadamente de cinco a diez categorías de resultados de entre las categorías candidatas. Se provee una representación gráfica de las categorías para visualización de las categorías por parte del usuario. Las categorías pueden ser visualizadas como una pluralidad de carpetas en la pantalla del usuario.

En otro aspecto, el invento incorpora un dispositivo de búsqueda para buscar una base de datos de registros como la reivindicada en la reivindicación 12. La base de datos comprende una pluralidad de registros, que incluye registros de Internet y registros con contenidos de interés. El dispositivo incluye un procesador de búsqueda y un procesador de agrupamiento. El procesador de agrupamiento incluye un procesador de registros, un generador de candidatos, un procesador de ponderación, y un procesador de visualización. Cada uno de estos elementos es un módulo de soporte lógico. Alternativamente, cada elemento podría posiblemente ser un módulo de soporte físico o un módulo combinado de soporte físico/soporte lógico. El procesador de búsqueda recibe instrucciones de búsqueda por parte de un usuario. En respuesta a una instrucción de búsqueda, el procesador de búsqueda busca la base de datos para generar una lista de resultados de búsqueda que incluye un conjunto seleccionado de registros. El procesador de agrupamiento procesa una parte de la lista de resultados de la búsqueda para crear dinámicamente un conjunto de categorías de resultados de búsqueda. Cada categoría de resultados de búsqueda se asocia con un subconjunto de los registros en la lista de resultados de la búsqueda.

El dispositivo realiza una pluralidad de pasos de procesamiento para crear dinámicamente las categorías de resultados de la búsqueda. El procesador de registros que identifica las características de tema, tipo, fuente e idioma está asociado con cada registro dentro de la lista de resultados de la búsqueda. El generador de candidatos identifica las características comunes asociadas con los registros dentro de la lista de resultados de la búsqueda y recopila una lista de categorías candidatas. Cada categoría de candidatos es representativa de una categoría de resultados de búsqueda posibles. El procesador de ponderación pondera cada categoría de candidatos como una función de las características comunes identificadas de los registros dentro de la categoría candidata. El procesador de visualización selecciona una pluralidad de categorías de resultados de la búsqueda que corresponden a las categorías candidatas que tienen la mayor ponderación. El procesador de visualización proporciona una representación gráfica de las categorías del resultado de la búsqueda para visualizarla en el monitor del usuario.

El invento proporciona un método eficiente para ver y navegar entre grandes conjuntos de registros y tiene ventajas sobre listas lineales largas. El invento usa la categorización para guiar al usuario a través de un proceso de pasos múltiples de una forma humana y satisfactoria. Un usuario puede formular una petición compleja en pequeños pasos realizados uno a uno. Usando el invento, un usuario puede rápidamente realizar la búsqueda en unos pocos pasos sin tener que revisar largas listas lineales de registros.

A continuación, solamente a modo de ejemplo, se describen las realizaciones preferidas del invento, y con referencia a los dibujos anejos, en los que:

la Figura 1 es un diagrama de bloques que ilustra los elementos funcionales de un dispositivo de búsqueda que incorpora los principios del invento;

la Figura 2 es un diagrama de flujos que ilustra la secuencia de los pasos usada por el dispositivo de búsqueda en la realización de una búsqueda de acuerdo con el invento;

las Figuras 3A-3C son ilustraciones de una visualización por el usuario durante una búsqueda que usa el dispositivo de búsqueda.

Descripción detallada

La Figura 1 diagrama de bloques que ilustra los elementos funcionales de un dispositivo de búsqueda que incorpora los principios del invento. El dispositivo 10 incluye un procesador de búsqueda 12 y un procesador de agrupamiento 14. El procesador de agrupamiento comprende un procesador de registros 16, un generador de candidatos 18, un procesador de ponderación 20, y un procesador de visualización 22. Estos elementos son módulos de soporte lógico y han sido así identificados solamente para ilustrar la funcionalidad del invento. El dispositivo 10 se comunica con un usuario 24 (es decir, un ordenador) y una base de datos 26, que incluye Internet y registros de contenidos de interés, a través de una barra colectora I/O 28. El dispositivo 10 es capaz de comunicar con una pluralidad de usuarios situados a distancia en una red de área ampliada (por ejemplo, Internet).

La Figura 2 es un diagrama de flujos que ilustra la secuencia de pasos usada por el dispositivo de búsqueda en la realización de la búsqueda. Con referencia a las Figuras 1 y 2, el dispositivo de búsqueda 12 recibe instrucciones de búsqueda (es decir, una petición) por parte de un usuario 24 a través de la barra colectora 28 (paso 30). El procesador de la búsqueda 12 busca la base de datos 26 y genera una lista de resultados de la búsqueda correspondiente a un conjunto seleccionado de los registros (paso 32). El conjunto seleccionado de los registros se clasifica de acuerdo con criterios de importancia. En una realización los criterios de importancia para clasificar los registros pueden incluir las siguientes reglas:

1. Si hay más “éxitos” (una palabra de un registro coincide con una palabra en los criterios de selección), el registro se clasifica más alto;

2. Si la frase de términos de la petición es un éxito en comparación con las palabras que independientemente son éxitos, el registro se clasifica más alto;

3. Si el empleo de letras mayúsculas es el mismo que en el término de la petición, el registro se clasifica más alto;

4. Si el término de la petición está en el título, el registro se clasifica más alto;

5. Si el término de la petición está en el resumen, el registro se clasifica más alto;

6. Si el término de la petición está en las palabras clave, el registro se clasifica más alto.

Si el número de registros es menor que un valor particular (por ejemplo, 20), se salta el procesador de agrupamiento 36 (paso 34). En caso contrario, el procesador de agrupamiento 14 procesa una parte de la lista de resultados de la búsqueda para crear dinámicamente un conjunto de categorías de resultados de búsqueda, en la que cada categoría de resultados de la búsqueda es asociada con un subconjunto de los registros en la lista de los resultados de la búsqueda. A modo de ejemplo solamente, la parte de la lista de los resultados de la búsqueda procesados pueden ser los doscientos (o cien) primeros registros más importantes dentro del conjunto seleccionado de registros.

El procesador de agrupamiento 14 realiza una pluralidad de pasos de procesamiento para crear dinámicamente el conjunto de categorías de resultados de la búsqueda. El procesador de registro 16 identifica las diversas características (por ejemplo, tema, tipo, fuente e idioma) asociadas con cada registro en la lista de resultados de la búsqueda (paso 36). El generador de candidatos 18 identifica las características comunes asociadas con los registros en la lista de resultados de la búsqueda y recopila una lista de categorías candidatas (paso 38). El generador de candidatos 18 utiliza diversas reglas, que se describen más adelante, para recopilar la lista. El procesador de ponderación 20 pondera la categoría de cada candidato como una función de las características comunes identificadas de los registros dentro de la categoría de candidatos (paso 40). También, el procesador de ponderación 20 utiliza diversas reglas de ponderación, que se describen más adelante, para ponderar las categorías candidatas. El procesador de visualización 22 selecciona una pluralidad de categorías de resultados de la búsqueda (por ejemplo, 5 a 10) que corresponden a las categorías candidatas que tienen la ponderación más alta (paso 42) y proporciona una representación gráfica de las categorías de los resultados de la búsqueda para visualización en el monitor del usuario (paso 44). Las categorías de los resultados de la búsqueda pueden ser visualizadas como una pluralidad de iconos en el monitor (por ejemplo, carpetas). Cuando el usuario selecciona una determinada categoría de resultados de la búsqueda, el procesador de visualización puede también proporcionar una representación gráfica del número de registros en la categoría de los resultados de la búsqueda, de categorías de resultados de búsqueda adicionales y una lista de los registros más importantes para visualización.

Como se ha observado antes, el usuario puede seleccionar una categoría de resultados de la búsqueda (paso 46) y visionar categorías de resultados de búsqueda adicionales (si el número de registros es mayor que un valor determinado) además de la lista de registros incluidos en esa categoría. Para hacer más corta la búsqueda, el usuario puede proporcionar unos términos de búsqueda adicionales (por ejemplo, una instrucción de afinación) (paso 48). A la recepción de los términos adicionales, el procesador de búsqueda 12 busca la base de datos 26 y genera otra lista de resultados de la búsqueda que corresponde a un conjunto afinado de los registros (paso 50). Alternativamente, el usuario puede (efectivamente) afinar la búsqueda simplemente abriendo sucesivamente categorías de resultados de la búsqueda adicionales.

Las Figuras 3A-3C son ilustraciones de muestra de una visualización del usuario durante una búsqueda usando el dispositivo 10. Estas ilustraciones son solamente a modo de ejemplo y están provistas solamente con fines explicativos. Por lo tanto, la disposición de las diversas claves, botones e iconos es inmaterial. Con referencia a las Figuras 3A-3C, el dispositivo 60 incluye un campo de búsqueda 62 en el que un usuario puede introducir instrucciones de búsqueda, y un icono de búsqueda 64 para ejecutar las instrucciones de búsqueda. La visualización también incluye un icono de señales 66 para proporcionar consejos de búsqueda, iconos de funciones variados (por ejemplo, un icono de búsqueda 68, un icono de directorios 70, un icono de apoyo 72 y un icono de leyes 74) e iconos de búsqueda (por ejemplo, búsqueda simple 76, búsqueda mecánica 78, búsqueda natural 80, búsqueda de empresa 82 y búsqueda por ordenador 84).

El usuario introduce las instrucciones de búsqueda (es decir, una petición) en el campo 62 y selecciona el icono de búsqueda 64 (véase la Figura 3B). El dispositivo de búsqueda 10 busca la base de datos 26 y crea dinámicamente un conjunto de categorías de resultados de la búsqueda (86a-86n) además de una lista de los registros más importantes

(88a-88m). Cada categoría (86a-86n) de resultados de búsqueda incluye un título del tema, y cada registro (88a-88m) incluye un título además de un indicador “honorarios/gratis”. El usuario puede ver una categoría seleccionando su icono o puede ver un determinado registro seleccionando su icono. Alternativamente, el usuario puede realizar una nueva búsqueda seleccionando el inicio en el icono 90 o puede afinar la petición introduciendo texto en el campo de búsqueda 62 y seleccionando el icono de búsqueda 64. Si el usuario selecciona una categoría, el dispositivo 10, crea otro conjunto de categorías de resultados de búsqueda y otra lista de los registros más importantes. El usuario puede repetir este proceso, además estrechando la búsqueda con cada iteración, hasta que el número de registros importantes desciende a un umbral predeterminado (por ejemplo, 20). En este punto, el dispositivo 10 proporciona al usuario solamente una lista de los registros más importantes.

El usuario puede usar una determinada lista de directorios (92a-92y) para centrar el proceso de búsqueda (véase la Figura 3C). El usuario introduce las instrucciones de búsqueda en el campo 62, selecciona uno o más directorios (por ejemplo, los directorios 92a, 92b) y selecciona el icono de búsqueda 64 (véase la Figura 3B). El dispositivo de búsqueda 10 busca la base de datos 26, centrándose en los registros que satisfacen la petición y caen dentro de los directorios seleccionados. El dispositivo proporciona un conjunto de categorías de resultados de búsqueda y de registros más importantes que están limitadas a estos directorios.

El procesador de agrupamiento ejecuta un algoritmo de categorización para crear dinámicamente el conjunto de categorías de resultados de búsqueda. El algoritmo incluye tres pasos principales: identificar las categorías candidatas, ponderar las categorías y visualizar una pluralidad de categorías con las ponderaciones más altas. Las reglas han sido organizadas alrededor de un número objetivo de siete (+/-2) categorías en la siguiente realización, pero generalmente son independientes de ese número.

A continuación se expone lógicamente una realización del algoritmo de categorización empleado por el procesador de agrupamiento. Se ha observado que una aplicación real del algoritmo puede omitir pasos, realizar pasos en paralelo o arbitrariamente. En la descripción del algoritmo, se usan los siguientes términos. El término “nrecs” significa que los primeros 200 registros del número total de registros en la lista de resultados de la búsqueda, o el tamaño total de la lista de resultados, cualquiera que sea el número, es menor. El término nrecs se refiere a la intercambiabilidad de ese número de registros o de ese grupo de registros. El término “ncategories” significa el número de categorías deseado (+/-2 categorías). El término “ordenamiento del dominio interno” significa un ordenamiento de los dominios que resalta las capacidades de diferenciación importantes de los dominios. El ordenamiento puede ser como sigue: tipo; tema; fuente; e idioma. El término “ordenamiento de dominio del usuario” significa un ordenamiento de dominios que resalta accesibilidad del usuario/valor aparente del usuario de los dominios. El ordenamiento puede ser como sigue: tema; fuente; tipo; e idioma. El término “level” significa el nivel en un dominio jerárquico del valor único de tal dominio particular asignado a esa categoría. Los niveles jerárquicos se supone que están numerados de 1 (todos los elementos, por ejemplo todos los temas) hasta N (el nivel más bajo de la jerarquía, con el nivel “superior” de 6 o así elementos que son del nivel 2).

En respuesta a una petición, el procesador de búsqueda busca la base de datos y genera una lista de resultados de la búsqueda. El conjunto de registros en la lista está clasificado de acuerdo con los criterios de importancia descritos anteriormente. Todo el procesamiento posterior se realiza en nrecs. Si nrecs es menor de 20 (o de algún otro número predeterminado), la única categoría candidata es la categoría “todos los registros”, y el procesador salta a la categoría de ponderación (descrita más adelante).

Para todas las reglas de generación de candidatos, el conjunto de valores disponibles de tipo, tema, idioma o fuente está limitado por cualquier valor de subárboles de tal valor proporcionado en la petición (por ejemplo, peticiones limitadas a un resultado de tema particular en las categorías candidatas que solamente incluyen ese tema o más temas específicos en ese área de tema). Si no se han proporcionado valores para esos campos, están disponibles todos los dominios de esas características. Se supone que esos criterios especificados para campos múltiples son lógicamente AND'd juntamente en la petición.

El procesador de agrupamiento genera, como categorías candidatas, todas las combinaciones de tipo-tema que tienen más del 20% de nrecs y que usan todos los nodos disponibles en los dominios de tema y de tipo. El procesador de agrupamiento genera, como categorías candidatas, todos los agrupamientos de tema solamente y las fusiones de todos los nodos disponibles en el dominio de tema que tienen un 20% o más de nrecs. El procesador de agrupamiento genera, como categorías candidatas, todos los agrupamientos de tipo solamente y las fusiones de todos los nodos disponibles en el dominio de tipo que tiene el 20% o más de nrecs. El procesador de agrupamiento genera, como una categoría candidata, cualquier dominio en la jerarquía del idioma que contiene más del 20% pero menos del 80% de nrecs. El procesador de agrupamiento genera, como una categoría candidata, cualquier sitio web que contiene tres o más registros, o cualquier otro nodo en la jerarquía de fuente que contiene más del 20% de nrecs. El procesador de agrupamiento genera, como una categoría candidata, cualquier nodo de nivel superior en la jerarquía de fuente para la cual no ha sido todavía generado. Esto proporciona al menos un conjunto de categorías candidatas que son exhaustivas no solamente de nrecs sino de toda la lista de resultados de la búsqueda. El procesador de agrupamiento genera categorías candidatas con el 20% o más de nrecs no generados todavía que constan de todas las combinaciones por pares de todos los nodos disponibles de cualesquiera dos campos especificados en la petición (por ejemplo, una petición que especifica idioma y fuente tendrá candidatos generados para todas las combinaciones idioma-fuente con un 20% o más de nrecs). Finalmente, el procesador de agrupamiento elimina cualesquiera categorías con un valor de “Desconocido” de cualquier dominio en la categoría.

Segundo, el algoritmo pondera las categorías. Las reglas de ponderación indican que las ponderaciones se aplican de forma acumulativa a las categorías (es decir, la ponderación final de cada categoría es la suma de todas las ponderaciones recibidas). Una regla resalta el ordenamiento interno del dominio y el nivel de precisión dentro de un dominio. Esa regla prevé que todas las categorías reciban una ponderación para cada dominio que es el producto del factor de ese dominio y el nivel del valor de tal dominio. Los factores son los siguientes: tipo (10), tema (6), fuente (3) e idioma (1). Otra regla resalta las categorías que tienen un mayor número de registros. Esa regla prevé que todas las categorías reciban una ponderación que sea el 20% del porcentaje de nrecs contenidos en esa categoría. Tres reglas resaltan las categorías más importantes. La primera regla prevé que todas las categorías reciban una ponderación igual a diez veces el número de registros en la categoría que está entre los cinco registros de nrecs clasificados más altos. La segunda regla prevé que todas las categorías reciban una ponderación igual a cinco veces el número de registros que se encuentran en la categoría que está entre los segundos cinco registros de nrecs clasificados más altos. La tercera regla prevé que todas las categorías reciban una ponderación igual a dos veces el número de registros que se encuentran en la categoría que está entre el undécimo y vigésimo registro de nrecs clasificados.

Cuatro reglas de ponderación muestran que se extienden en un dominio, que aumenta la cobertura total y que minimiza la duplicación. La primera regla prevé que todas las categorías que contienen un valor en el nivel dos del dominio de ese valor, para el que no existen categorías de valores por debajo del nivel dos de ese dominio, reciban una ponderación de 15. Esto se aplica a cada dominio contenido dentro del tema de la categoría. La segunda regla prevé que todas las categorías que contengan un valor en el nivel tres de su dominio, para el que no existan categorías de valores por debajo del nivel tres de ese dominio, reciban una ponderación 8. Esto se aplica a cada dominio contenido dentro del tema de la categoría. La tercera regla prevé que si las categorías con las ponderaciones más altas no cubren exhaustivamente los valores de cualesquiera de los dominios, y no existen dos o menos categorías que puedan ser añadidas con valores de un único dominio para representar exhaustivamente ese dominio, se añada 25 a cada una de esas categorías. Sin embargo, si no hay más de dos categorías (en el mismo o diferente dominio) al que esto se aplique, se seleccionan las categorías para las que la suma de las dos categorías identificadas tengan la ponderación máxima. En el caso de un vínculo, se selecciona basándose en el ordenamiento interno del dominio y, si todavía hay un vínculo, se selecciona al azar. La cuarta regla prevé que todas las categorías que contengan registros, no encontrándose el 70% o más de nrecs en otras categorías, reciban una ponderación de 8. Se ha observado que se pueden usar esos otros porcentajes y valores de ponderación.

Otra regla resalta las categorías de los sitios web. Esa regla prevé que todas las categorías de sitios web solamente con el 20% o más de nrecs reciban una ponderación de 12. Otra regla más resalta los temas especificados dentro de una petición y prevé que todas las categorías que contengan un dominio para el cual el usuario especificó un valor reciban una ponderación de 10. Finalmente, una regla que resalta las categorías de combinación prevé que todas las categorías de combinación reciban una ponderación de 8.

Tercero, el algoritmo determina una pluralidad de categorías de resultados de búsquedas a partir de las categorías candidatas con las ponderaciones más altas. Primero, el procesador selecciona las categorías candidatas con la ponderación más alta. En el caso de vínculos, el ordenamiento del dominio del usuario se usa para seleccionar las categorías. Si la categoría con ponderación más baja o las dos categorías con ponderaciones más bajas en ncategorías representan una caída significativa desde la categoría ponderada más alta, las ncategorías se reducen en uno o dos. Sin embargo, si las dos categorías ponderadas más altas no están todavía en las ncategorías significativamente más bajas en ponderación que la categoría más baja ya en ncategorías, las ncategorías se incrementan en uno o dos. Se ha observado que se pueden usar otros porcentajes y valores de ponderación.

En la determinación del nombre de cada categoría de resultados de búsqueda, las categorías con combinaciones de dominios (por ejemplo, tema-tipo) se denominan con el valor de cada dominio separado por un guión. El orden de los dos dominios está determinado por el uso del ordenamiento del dominio. Para cada categoría de resultados de búsqueda, se visualiza un recuento del número de registros en cada categoría. Las categorías de sitios web se denominan por el dominio del sitio web, que puede ser un enlace directo a la página inicial del sitio por defecto. Independientemente de si cualquiera de las categorías visualizadas son puras o combinación de categorías de sitios web, todas las ocurrencias de registros individuales dentro de la lista de registros de cualquier categoría, distinta de una categoría de sitio web o de combinación de sitios web, son sustituidas por la categoría de sitio web que contiene esos registros. Para las categorías de sitios web así integradas dentro de una lista de resultados de búsqueda que tienen tres o menos registros, es posible mostrar los registros “en línea”, esto es, los registros individuales propiamente dichos pueden ser mostrados con el nombre de categoría, eliminando la necesidad de expandir explícitamente la categoría. Los registros dentro de las categorías de resultados de búsqueda se visualizan por defecto por orden de importancia, o, a deseo del usuario, en orden de fecha inverso (primero el más reciente). Las categorías web dentro de la lista de registros están clasificadas por el valor del registro clasificado más alto en el sitio web.

Los porcentajes numéricos, las ponderaciones asignadas y las reglas detalladas descritas anteriormente son a modo de ejemplo y pueden cambiar sin apartarse del espíritu y alcance del invento.

Para que funcione la metodología de categorización, cada registro dentro de la base de datos, que incluye registros de Internet y registros de contenidos de interés, está clasificado por características de tema, fuente, e idioma (esto es, atributos metadatos). En otras realizaciones, los registros pueden clasificarse por atributos adicionales de metadatos (por ejemplo, nivel de dificultad o popularidad), atributos basados en la petición, nombres propios, y características de

análisis de documentos de tiempo de ejecución. Debido a que tal tarea tiene mucho trabajo manual, el dispositivo de búsqueda autoclasifica sustancialmente todos los registros en las categorías apropiadas.

Cada registro es asignado a uno o más tipos (por ejemplo, artículo, reseña de libros, carta) y un único valor de fuente (por ejemplo, PC Week, páginas web personales) a través de un proceso más bien automático (completamente automático para datos de Internet), a pesar de que existe alguna asignación editorial para ciertos datos de contenido de interés. También, cada registro es generalmente asignado a uno o más temas (por ejemplo, biología molecular) e idiomas (por ejemplo, francés) a través de un proceso más bien automático (completamente automático para datos de Internet). Ocasionalmente, un registro no está asignado a un tema y/o idioma. En tales casos, a los registros se les asigna un valor de “desconocido” para estos elementos metadatos particulares (o atributos, o campos, o dominios). En el caso de tema, idioma y tipo (pero no para fuente), es posible para un registro tener más de un valor (por ejemplo, debido a que en realidad se refiere a dos o más áreas de temas, o debido a que contiene texto en más de un idioma).

Además, todos los valores en estos dominios están dispuestos jerárquicamente (por ejemplo, “biología molecular” pertenece a “biología”, y “reseña de libros” pertenece a “reseña”). A pesar de que los registros se clasifican automática o manualmente en solamente uno (o quizás dos o más) valores objetivamente específicos para un dominio dado (por ejemplo, “reseña de libros” para tipo, “biología molecular” para tema), reciben inherentemente todos los valores que son mayores que esos valores en sus respectivas jerarquías de dominio respectivas. Por ejemplo, a un registro clasificado como “biología molecular” también se le da el tema de “biología” (el padre de la biología molecular) y el tema de “ciencia” (que es el padre de la biología). Esto puede dar lugar a unos valores de clasificación adicionales de 5 ó 6 para ese registro. Los niveles superiores de cada jerarquía son objetivamente pequeños, por ejemplo, aproximadamente seis áreas de temas para el dominio de temas (que incluye humanidades y sociedad, negocios, etc), cada uno de los cuales está dividido en tres o cuatro más, que hacen 18 en el nivel 2, cada uno de los cuales está dividido en aproximadamente 35, que hacen aproximadamente 600 en el nivel 3, etc.

Se ha desarrollado un sistema de clasificación para determinar automáticamente los cuatro atributos de los datos (es decir, tema, tipo, idioma y fuente) cuando tales atributos no están editorialmente disponibles del editor o de la fuente de registro. Los sistemas de clasificación incluyen dos componentes principales: (1) un programa de clasificación basado en una petición; y (2) un conjunto de programas individuales.

El programa de clasificación basado en petición realiza eficientemente la clasificación de atributos seleccionados y valores de atributos, que incluyen más de 20.000 términos de temas. Se necesita una petición para cada atributo y valor de atributo (por ejemplo, atributo = tema, valor de atributo = bioquímica). Las peticiones se ejecutan con respecto a todos los registros, y las puntuaciones de clasificación que representan la fuerza de la comparación se computan para cada registro y petición. Los registros se clasifican entonces en los dos o tres peticiones/valores de atributos para los que ese registro tiene la puntuación de clasificación más alta.

El programa de clasificación basado en la petición produce los siguientes subcomponentes: (1) un idioma de clasificación para especificar las peticiones de clasificación; (2) un medio de y fuentes para producir automáticamente las peticiones de clasificación usadas por el programa; y (3) varias peticiones de clasificación formadas manualmente usadas por el programa.

Los medios de y las fuentes para producir automáticamente las peticiones de clasificación genera peticiones de aproximadamente 5 líneas de largo. Cada petición se produce analizando un registro ejemplar o modelo de ese valor de atributo (por ejemplo, un artículo de enciclopedia sobre bioquímica) y extrayendo automáticamente los términos más significativos para el registro. La “petición” resultante se usa para hacer coincidir y recuperar otros registros similares (es decir, clasificables con el mismo valor). La significancia del término se determina mediante cuánta frecuencia el término está dentro del registro (es decir, más frecuente es igual a más importante) y qué infrecuente es en el cuerpo particular de los ejemplos de registros que se han usado, por ejemplo, la enciclopedia en su conjunto (es decir, menos frecuente es igual a más importante). Exactamente qué valores de frecuencia/infrecuencia usar se determina empíricamente y fijado para cada fuente particular de registros ejemplares. Se pueden usar varias fuentes. También se han desarrollado varias herramientas del programa relacionado (por ejemplo, para hacer coincidir automáticamente artículos de enciclopedia con términos en la jerarquía de los temas).

El número de peticiones de clasificación formadas manualmente es como sigue: aproximadamente 2.000 de tales peticiones de temas; aproximadamente 50 peticiones de tipo manual; y aproximadamente 6 peticiones de idioma manuales. La media de peticiones manuales es de aproximadamente 15 líneas de longitud, excepto para peticiones de idioma, que son considerablemente más largas.

El segundo componente del sistema de clasificación es un conjunto de programas individuales y un programa de control de nivel más alto que se usan para clasificar datos con respecto a ciertos valores particulares (por ejemplo, receta o “página web personal”) de uno de los atributos de datos (“tipo” en el caso de receta, “fuente” en el caso de páginas personales) cuando la aproximación basada en una petición se considera inadecuada. Estos programas están compuestos de varios cientos de líneas de Perl. Estos programas buscan no solamente la presencia de ciertas palabras sino para formatear señales (por ejemplo, el formato particular de un registro de receta tipo, o de entrevista tipo). También existe un conjunto de herramientas de ensayo para evaluar los resultados de estas clasificaciones.

ES 2 306 474 T3

Una base de datos de clasificación crea y mantiene las taxonomías de los datos, jerarquías, referencias cruzadas y peticiones de clasificación asociadas. La base de datos incluye un editor de clasificación multiusuario, y unos medios para generar los informes y ficheros de datos necesitados por otros componentes del sistema (por ejemplo, el dispositivo de búsqueda) y se aplica usando Microsoft Access, formularios Microsoft, Data Access Objects, SQL y Visual Basic. La base de datos incluye aproximadamente 40 tablas, 15 formularios, 25 informes y 5.000 líneas de Visual Basic, y produce 12 ficheros intermedios de otras partes del sistema.

Equivalentes

Mientras que el invento ha sido particularmente mostrado y descrito con referencia a realizaciones preferidas específicas, los expertos en la técnica que se pueden realizar diversos cambios en la forma y detalle pueden ser hechos en ellas sin apartarse del alcance del invento definido por las reivindicaciones anejas.

15

20

25

30

35

40

45

50

55

60

65

REIVINDICACIONES

1. Un método para buscar una base de datos de registros, que comprende:

5 buscar la base de datos (26), en respuesta a una instrucción de búsqueda (30), para generar una lista de resultados de búsqueda que incluye un conjunto seleccionado de los registros (32);

10 procesar al menos una parte de la lista de resultados de la búsqueda para identificar dinámicamente (i) las características comunes asociadas con los registros (36) dentro de la lista de resultados de la búsqueda, (ii) agrupar los registros en la lista de resultados de la búsqueda que tienen características comunes en categorías candidatas (38), y (iii) recopilar una lista de categorías candidatas, siendo cada categoría candidata representativa de una categoría de resultados de búsqueda posibles;

15 **caracterizado** por ponderar cada categoría candidata (40) como una función de las características comunes identificadas de los registros dentro de la categoría candidata para seleccionar una pluralidad de categorías de resultados de la búsqueda correspondientes a las categorías candidatas que tienen la ponderación más alta (42).

20 2. Un método de acuerdo con la reivindicación 1 que además comprende clasificar los registros dentro de la lista de resultados de búsqueda de acuerdo con los criterios de importancia preseleccionados.

3. Un método de acuerdo con la reivindicación 1 ó 2 que además comprende identificar características de tema, tipo, fuente e idioma asociadas con cada registro dentro de la lista de resultados de la búsqueda.

25 4. Un método de acuerdo con la reivindicación 1, en el que la ponderación de cada categoría candidata se realiza basándose al menos en parte en varios registros en la categoría candidata.

30 5. Un método de acuerdo con la reivindicación 2, en el que la ponderación de cada categoría candidata se realiza basándose al menos en parte en varios registros en la categoría candidata que tienen una clasificación dentro de un alcance preseleccionado.

6. Un método de acuerdo con cualquier reivindicación anterior, que además comprende seleccionar categorías candidatas como una función de las características comunes identificadas de los registros.

35 7. Un método de acuerdo con la reivindicación 6, que además comprende seleccionar entre aproximadamente cinco a diez categorías resultado de la búsqueda de entre las categorías candidatas.

8. Un método de acuerdo con la reivindicación 3, que además comprende agrupar las categorías resultado de la búsqueda en respuesta a un valor seleccionado por el usuario para una de las características.

40 9. Un método de acuerdo con cualquier reivindicación anterior, en el que la base de datos incluye registros de Internet y registros de contenidos de interés.

10. Un método de acuerdo con cualquier reivindicación anterior, que además comprende proporcionar una representación gráfica de las categorías.

45 11. Un método de acuerdo con cualquier reivindicación anterior, que además comprende identificar características de metadatos asociados con registros dentro de la lista de resultados de la búsqueda.

12. Un dispositivo de búsqueda (10) para buscar una base de datos (26) de registros, que comprende

50 un procesador de búsqueda (12), en respuesta a una instrucción de búsqueda (30) para buscar la base de datos para generar una lista de resultados de la búsqueda que incluye un conjunto seleccionado de los registros (32),

55 un procesador de agrupamiento (14) para procesar al menos una parte de la lista de resultados de la búsqueda para identificar dinámicamente (i) las características comunes asociadas con los registros (36) dentro de la lista de resultados de la búsqueda, (ii) agrupar los registros que tienen características comunes en categorías candidatas (38), (iii) recopilar una lista de categorías candidatas, siendo cada categoría candidata representativa de un resultado de búsqueda posible,

60 **caracterizado** por un procesador de ponderación (20) para ponderar cada categoría candidata (40) como una función de las características comunes identificadas de los registros dentro de la categoría candidata.

13. Un dispositivo (10) de acuerdo con la reivindicación 12, que además comprende medios para clasificar los registros dentro de la lista de resultados de la búsqueda de acuerdo con criterios de importancia preseleccionados.

65 14. Un dispositivo (10) de acuerdo con la reivindicación 12 ó 13, que además comprende un procesador de registros (16) para identificar características de tema, tipo, fuente e idioma asociadas con cada registro dentro de la lista de resultados de la búsqueda.

ES 2 306 474 T3

15. Un dispositivo (10) de acuerdo con la reivindicación 12, en el que el procesador de ponderación (20) prevé la ponderación de cada categoría candidata basándose al menos en parte en varios registros en la categoría candidata.

5 16. Un dispositivo (10) de acuerdo con la reivindicación 13, en el que el procesador de ponderación (20) prevé la ponderación de cada categoría candidata basándose al menos en parte en varios registros en la categoría candidata que tienen una clasificación dentro de un alcance preseleccionado.

10 17. Un dispositivo (10) de acuerdo con cualquiera de las reivindicaciones 12 a 16, que además comprende medios para seleccionar entre aproximadamente cinco a diez categorías de resultados de la búsqueda de entre las categorías candidatas.

18. Un dispositivo (10) de acuerdo con cualquiera de las reivindicaciones 12 a 17, que además comprende un procesador de visualización (22) para proporcionar una representación gráfica de las categorías.

15 19. Un dispositivo (10) de acuerdo con la reivindicación 13, que además comprende medios para agrupar los registros dentro de la lista de resultados de la búsqueda en respuesta a un valor seleccionado por el usuario para una de las características.

20 20. Un dispositivo (10) de acuerdo con cualquiera de las reivindicaciones 12 a 19, que además comprende medios para generar como una función de una de las categorías, una instrucción de afinación que es representativa de una instrucción adicional para buscar la base de datos (26) de los registros asociados con la categoría y la instrucción adicional.

25 21. Un dispositivo (10) de acuerdo con la reivindicación 14, que además comprende medios para clasificar las características comunes identificadas de los registros en un orden jerárquico.

22. Un dispositivo (10) de acuerdo con cualquiera de las reivindicaciones 12 a 21, en el que la base de datos incluye registros de Internet y registros de contenidos de interés.

30 23. Un dispositivo de búsqueda (10) reivindicado en la reivindicación 12, que además comprende:

una base de datos (26) para almacenar una pluralidad de registros, que incluyen registros de Internet y registros de contenidos de interés; y

35 un procesador de visualización (22) para proporcionar una representación gráfica de las categorías al usuario, en el que la instrucción de búsqueda procede de un usuario.

40

45

50

55

60

65

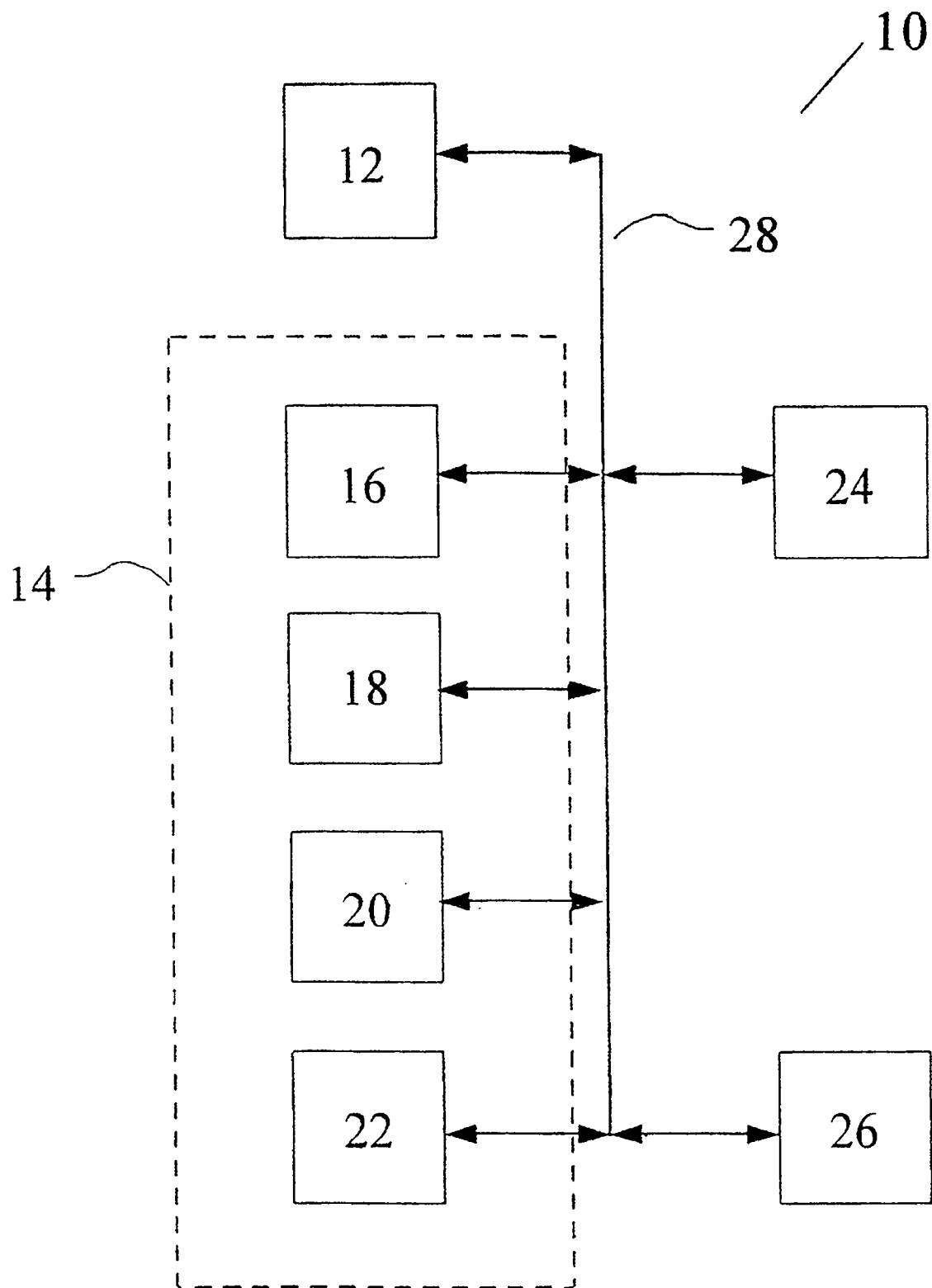


Fig. 1

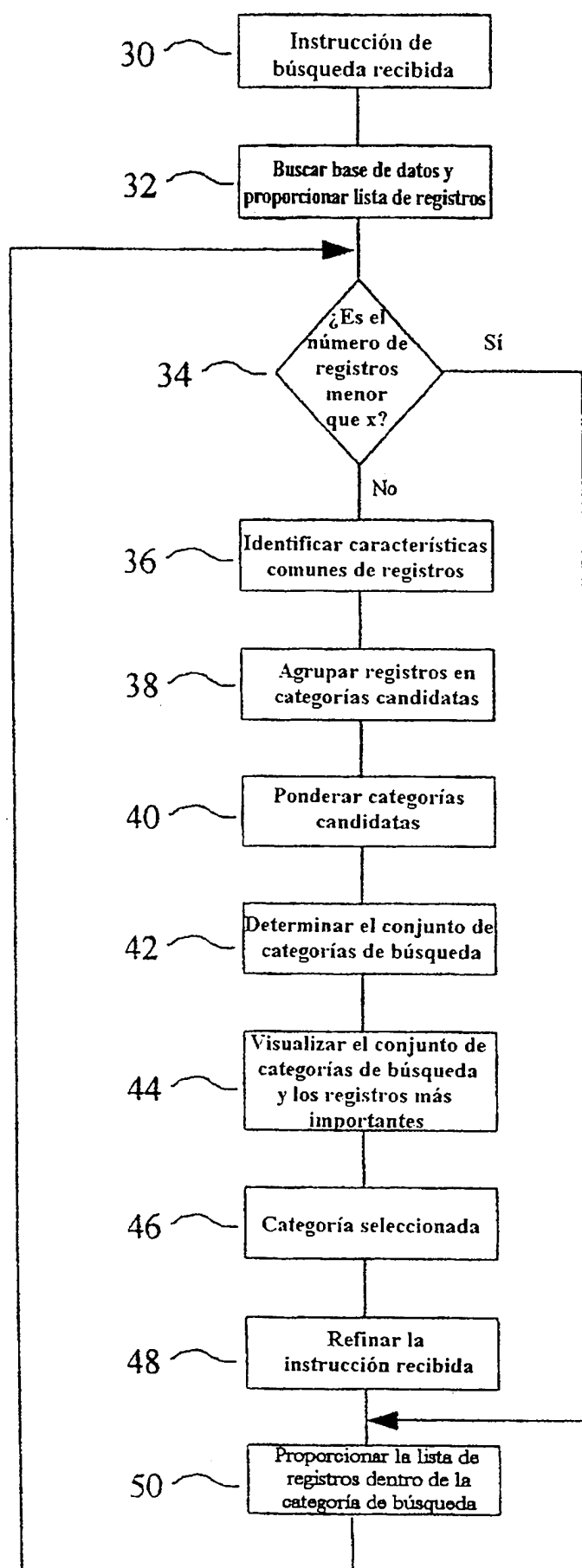


Fig. 2

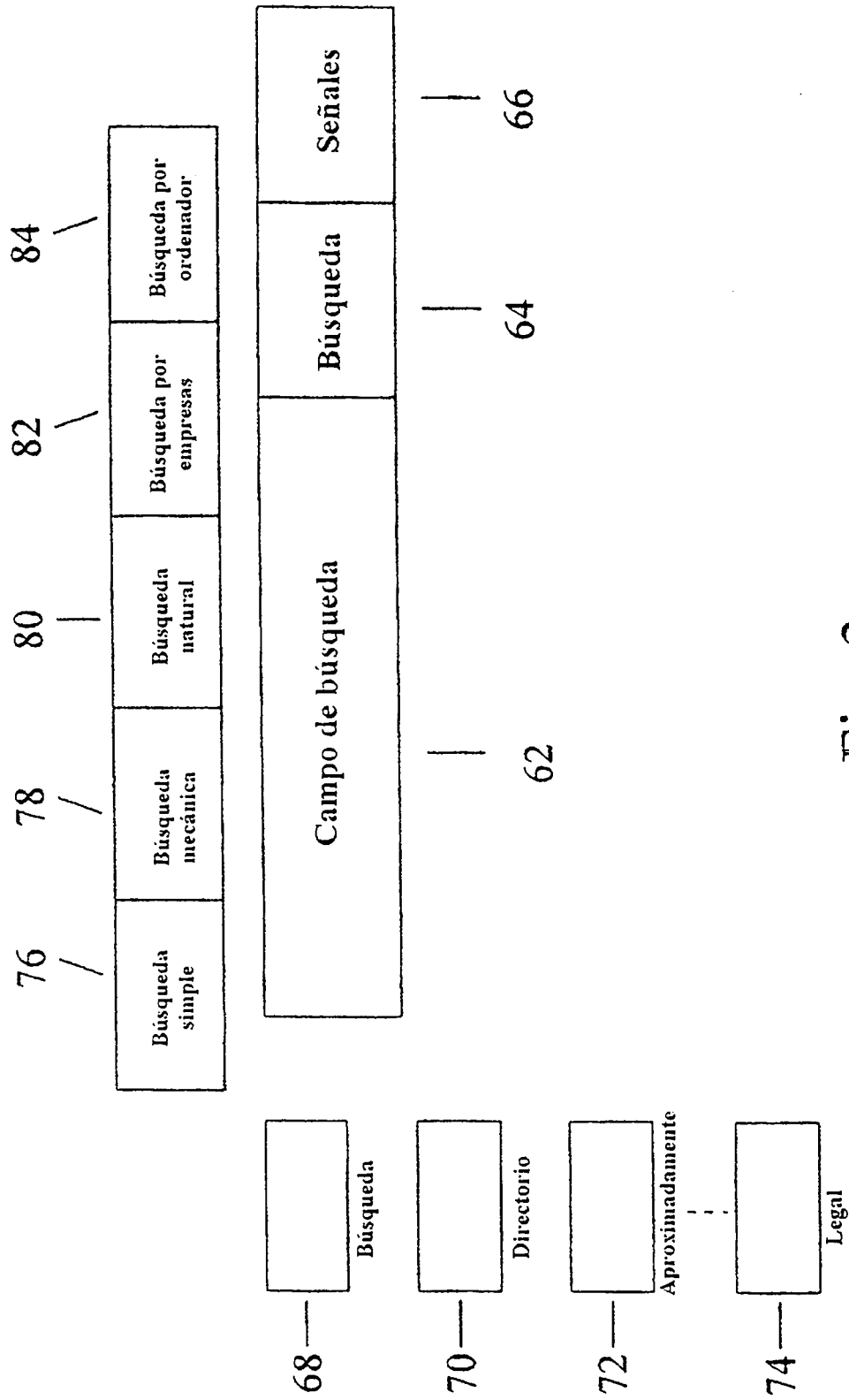


Fig. 3a

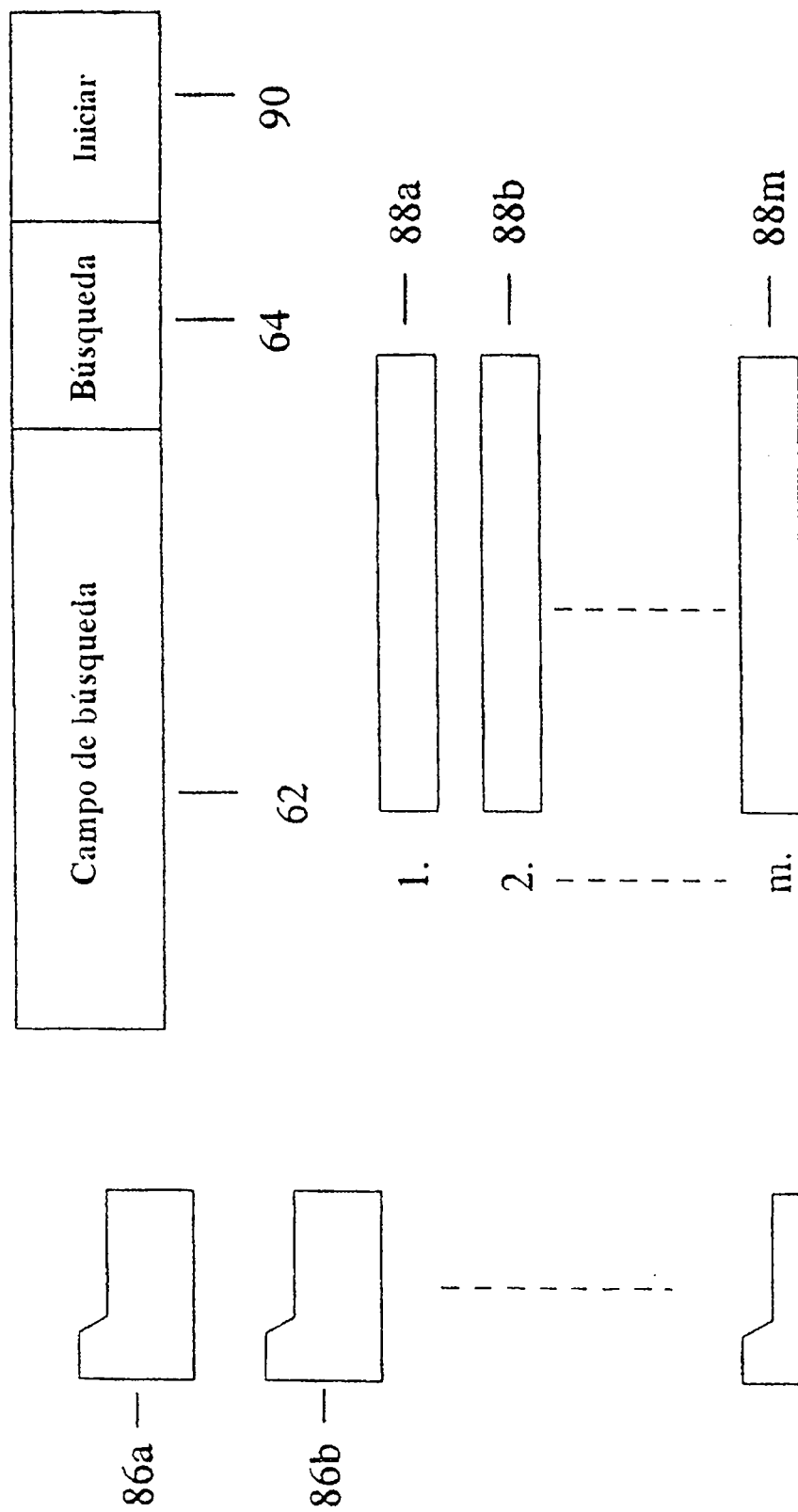


Fig. 3b

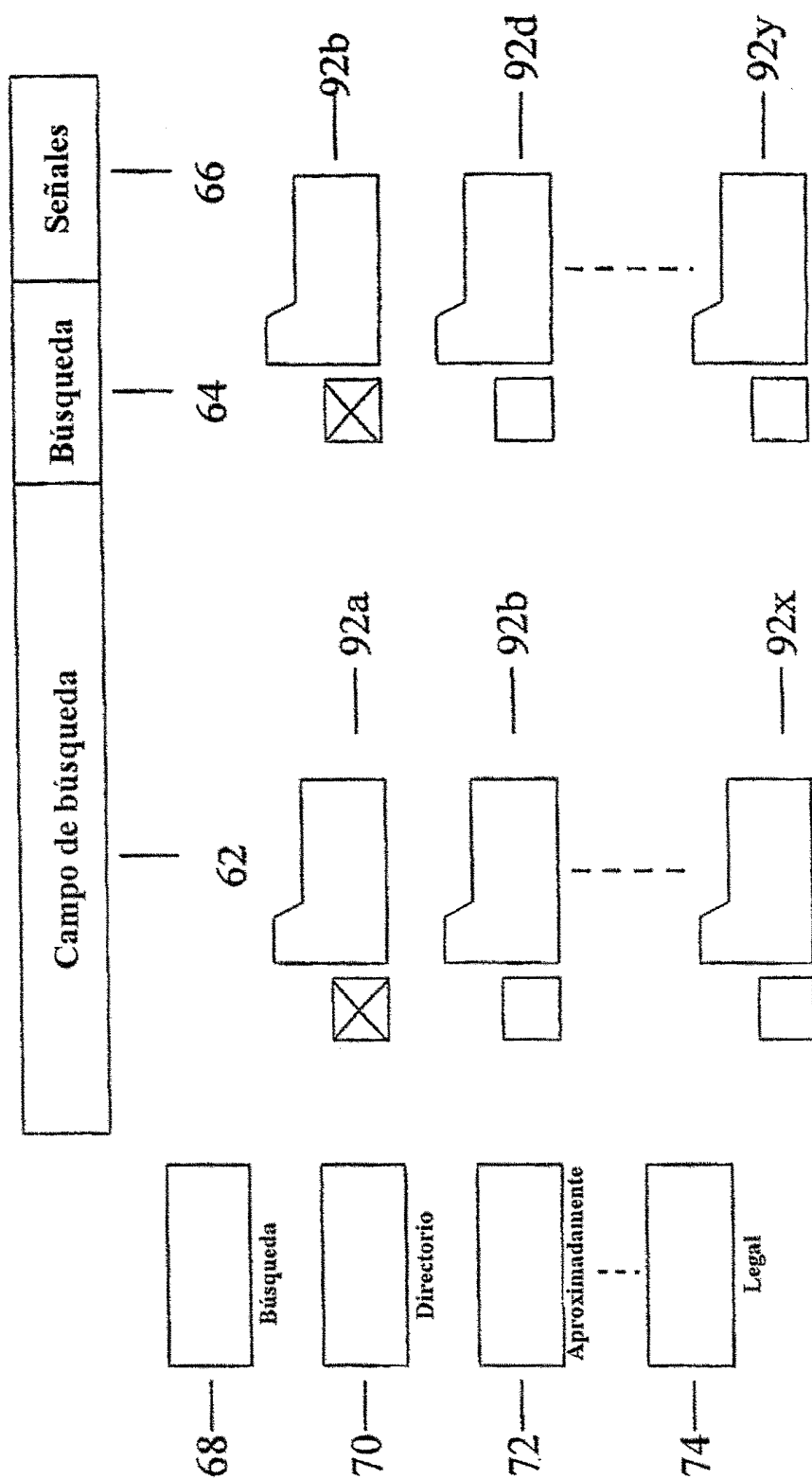


Fig. 3c