(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 336 678**

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

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Número de solicitud europea: **04787046 .4**
(96) Fecha de presentación : **27.09.2004**
(97) Número de publicación de la solicitud: **1673704**
(97) Fecha de publicación de la solicitud: **28.06.2006**

(54) Título: **Recuperación de documentos asistida por ordenador.**

(30) Prioridad: **26.09.2003 GB 0322600**

(45) Fecha de publicación de la mención BOPI:
15.04.2010

(45) Fecha de la publicación del folleto de la patente:
15.04.2010

(73) Titular/es: **University of Ulster**
Cromore Road, Coleraine
County Londonderry BT52 1SA, GB
St. Petersburg State University

(72) Inventor/es: **Patterson, David y**
Dobrynin, Vladimir

(74) Agente: **Carvajal y Urquijo, Isabel**

DESCRIPCIÓN

Recuperación de documentos asistida por ordenador.

Campo de la invención

La presente invención hace referencia a la recuperación de documentos asistida por ordenador desde un corpus de documentos, en especial un corpus de documentos controlado. La invención hace referencia en particular al agrupamiento de documentos asistida por ordenador.

Antecedentes de la invención

La búsqueda de documentos asistida por ordenador, generalmente, incluye la utilización de uno o más programas informáticos para analizar un corpus de documentos y después realizar una búsqueda dentro de dicho corpus de documentos analizado. El análisis de un corpus de documentos puede incluir la organización de documentos en una pluralidad de corpus de documentos para facilitar el proceso de búsqueda. Generalmente, esto incluye la utilización de uno o más programas informáticos para implementar un algoritmo de agrupamiento. La búsqueda a través de un corpus de documentos normalmente se realiza mediante un programa informático comúnmente conocido como motor de búsqueda.

Una función con un impacto importante en el diseño arquitectónico de un motor de búsqueda es el tamaño del corpus de documentos. Otra consideración importante es si el mantenimiento del corpus de documentos (agregar y eliminar documentos) está abierto a todos los usuarios (un corpus no controlado como Internet), o si el mantenimiento es controlado, por ejemplo, por un administrador, (un corpus controlado como una Intranet). En forma más general, un corpus controlado comprende un conjunto de datos controlado por un administrador o un conjunto de datos que es completamente accesible.

Los algoritmos de búsqueda convencionales entregan, como resultado de la búsqueda, una lista de documentos que deben contener todo o una parte de un conjunto de palabras claves introducidas en la consulta de un usuario. Tales sistemas determinan la relevancia del documento basándose en la frecuencia de aparición de la palabra clave, o bien mediante la utilización de referencias y enlaces entre los documentos. A menudo aparecen muchos resultados de búsqueda y el usuario no puede determinar fácilmente qué resultados son relevantes de acuerdo con sus necesidades. Por lo tanto, aunque la cobertura puede ser alta, la gran cantidad de documentos que son devueltos para alcanzar dicho objetivo resulta en poca precisión y una búsqueda laboriosa para que el usuario pueda encontrar los documentos más relevantes.

De forma adicional, un motor de búsqueda convencional devuelve una lista de documentos sin clasificar. Si el tema de consulta es relativamente amplio, esta lista puede contener documentos pertenecientes a muchos subtemas específicos.

Para poder obtener los mejores resultados a partir de algoritmos de búsqueda convencionales, que se basan en estadísticas de palabras, un usuario necesita tener conocimientos estadísticos sobre el corpus de documentos antes de realizar la consulta. Este conocimiento nunca se conoce *a priori* y por lo tanto el usuario rara vez formula buenas consultas. Con una búsqueda temática, es posible ofrecer conocimientos sobre descripciones de agrupaciones al usuario, lo que le permite mejorar y especificar sus consultas de manera inteligente e interactiva.

Los motores de búsqueda convencionales a menudo utilizan información adicional tal como enlaces entre páginas Web, o referencias entre documentos para mejorar los resultados de búsqueda.

El concepto de búsqueda o exploración basadas en agrupamiento de documentos es conocido (por ejemplo, la herramienta de exploración Scatter-Gather [4]). Los principales problemas de este tipo de enfoque son su aplicabilidad a aplicaciones de la vida real y la eficiencia y efectividad de los algoritmos de agrupamiento. Los algoritmos de agrupamiento no supervisados se enmarcan dentro de paradigmas jerárquicos o separadores. En general, deben determinarse similitudes entre todos los pares de documentos lo que hace que estos enfoques sean no escalables. Los enfoques supervisados requieren un conjunto de datos de capacitación que no siempre está disponible, puede aumentar el coste de un proyecto y demandar mucho tiempo de preparación.

Se considera un enfoque diferente al problema de la recuperación con enfoque temático en [5]. Este sistema utiliza un conjunto de agentes para recuperar de Internet, o filtrar de entre un grupo de noticias, documentos relevantes sobre un tema específico. Los temas se describen manualmente en forma de texto. De forma adicional, un conjunto de reglas se genera manualmente en un lenguaje de reglas especial para describir cómo comparar un documento con un tema, es decir, qué palabras de la descripción de un tema deberían utilizarse y cómo influyen estas palabras en los pesos de las categorías. La categoría de documentos resultante se determina utilizando pesos de categoría calculados y lógica difusa. La principal desventaja de este enfoque es que todas las descripciones temáticas y reglas se definen de manera manual. Es imposible predecir por adelantado cuáles son las descripciones temáticas dadas y el grupo de reglas correspondiente, suficientes para recuperar los documentos relevantes, con gran precisión y cobertura. Por lo tanto, se requiere una gran cantidad de trabajo manual e investigación para generar descripciones temáticas y reglas efectivas. Por todo esto, este enfoque no puede considerarse escalable.

El descubrimiento automático de temas a través de la generación de grupos de documentos puede basarse en técnicas tales como Probabilistic Latent Semantic Indexing (Indexación Semántica Latente Probabilística) [6]. La Indexación Semántica Latente Probabilística utiliza un modelo probabilístico y los parámetros de este modelo se calculan utilizando el algoritmo de maximización de cálculos.

En [7], se presenta otro ejemplo de un motor de búsqueda basado en enfoques teóricos de información para descubrir información sobre temas que se encuentran en el corpus de documentos. La idea principal es generar un conjunto de los llamados hilos temáticos y utilizarlo para presentar el tema de cada documento en el corpus. El hilo temático es una secuencia de palabras de un sistema fijo de clases de palabras. Estas clases se forman como resultado de un análisis de un conjunto representativo de documentos seleccionados al azar del corpus de documentos (un conjunto de entrenamiento). Palabras de diferentes clases difieren por probabilidades de aparición en el conjunto de entrenamiento y, por lo tanto, representan temas en diferentes niveles de abstracción. Un hilo es una secuencia de esas palabras en la cual la siguiente palabra pertenece a una clase más delimitada, y las palabras vecinas de esta secuencia deben aparecer en el mismo documento con una probabilidad suficientemente alta. A cada documento del corpus de documentos se le asigna uno de los hilos temáticos posibles. Se utiliza entropía cruzada como medida para seleccionar un hilo temático que sea el más relevante según el tema del documento. Este hilo temático se almacena en el índice y se utiliza en la etapa de búsqueda en lugar del documento mismo. La principal desventaja de este enfoque es que sólo una parte relativamente pequeña de la información sobre un documento se almacena en el índice y se utiliza durante la búsqueda. Además, no pueden utilizarse los hilos temáticos para agrupar documentos dentro de agrupaciones temáticas y, por lo tanto, la información sobre la estructura temática del corpus de documentos está oculta para el usuario.

La publicación de la patente US2002/0042793 presenta agrupamiento en tiempo real sin utilizar la entropía de distribuciones de probabilidad para calcular los atractores de grupos. Sería deseable mitigar los problemas antes señalados.

Resumen de la invención

Un aspecto de la invención proporciona un método para determinar atractores de grupos en un aparato para determinar atractores de grupos a partir de una pluralidad de documentos, donde cada documento comprende al menos un término, y donde dicho método comprende: calcular, respecto de cada término, una distribución de probabilidad que indique la frecuencia de aparición del otro término, o de cada otro término, que co-aparezca con dicho término en al menos uno de dichos documentos; calcular, respecto de cada término, la entropía de la distribución de probabilidad respectiva; seleccionar al menos una de dichas distribuciones de probabilidad como un atractor de grupo, según el valor de entropía respectivo.

Cada distribución de probabilidad puede comprender, respecto de cada término que co-aparezca, un indicador que indique la cantidad total de casos del término que co-aparece en todos los documentos en donde co-aparece el término que co-aparece con el término respecto del cual se calcula la distribución de probabilidad. Cada distribución de probabilidad comprende, respecto de cada término que co-aparece, un indicador que comprende una probabilidad condicional de la aparición del término que co-aparece respectivo en un documento, según la aparición en dicho documento del término respecto del cual se calcula la distribución de probabilidad. En forma ventajosa, cada indicador se normaliza respecto de la cantidad total de términos en el, o en cada, documento en el cual aparece el término respecto del cual se calcula la distribución de probabilidad.

En una realización preferente, el método puede comprender asignar cada término a una pluralidad de subconjuntos de términos según la frecuencia de aparición del término; y seleccionar, como atractor de grupo, la distribución de probabilidad respectiva de uno o más términos de cada subconjunto de términos. Cada término puede asignarse a un subconjunto según la cantidad de documentos del corpus en donde aparece el término respectivo. Puede asignarse un umbral de entropía a cada subconjunto, y el método puede incluir seleccionar, como atractor de grupo, la distribución de probabilidad respectiva de uno o más términos de cada subconjunto que tenga una entropía que satisfaga el umbral de entropía respectivo. En forma ventajosa, el método comprende seleccionar, como atractor de grupo, la distribución de probabilidad respectiva de uno o más términos de cada subconjunto que tenga una entropía que sea menor o igual al umbral de entropía respectivo.

Cada subconjunto puede asociarse a un rango de frecuencia, en donde los rangos de frecuencia para subconjuntos respectivos están separados. Cada subconjunto puede asociarse a un rango de frecuencia, siendo el tamaño de cada rango de frecuencia sucesivo igual a una constante multiplicada por el tamaño del rango de frecuencia precedente en orden de frecuencia ascendente. En una realización, el umbral de entropía respectivo aumenta, por ejemplo, en forma lineal para subconjuntos sucesivos en orden de frecuencia ascendente.

Otro aspecto de la invención ofrece un producto de programa informático que comprende un código de programa informático para que un ordenador lleve a cabo el método de determinar atractores de grupo.

Otro aspecto de la invención ofrece un aparato para determinar atractores de grupo para una pluralidad de documentos, cada documento que comprende al menos un término, y el aparato que comprende: medios para calcular, respecto a cada término, una distribución de probabilidad que indica la frecuencia de aparición del término, o cada término, que co-aparece con dicho término en, al menos, uno de dichos documentos; medios para calcular, respecto

de cada término, la entropía de la distribución de probabilidad respectiva; y medios para seleccionar al menos una de dichas distribuciones de probabilidad como atractor de grupo, según el valor de entropía respectivo.

Otro aspecto de la invención ofrece un método para agrupar una pluralidad de documentos, el método incluye determinar los atractores de grupo según el método antes descrito.

En una realización, el método de agrupamiento comprende: calcular, respecto de cada documento, una distribución de probabilidad que indique la frecuencia de aparición de cada término en el documento; comparar la distribución de probabilidad respectiva de cada documento con cada distribución de probabilidad seleccionada como atractor de grupo; y asignar cada documento a al menos un grupo según la similitud entre las distribuciones de probabilidad comparadas.

El método de agrupamiento puede incluir organizar los documentos dentro de cada grupo mediante: la asignación de un peso respectivo a cada documento, el valor del peso dependerá de la similitud entre la distribución de probabilidad del documento y la distribución de probabilidad del grupo atractor; la comparación de la distribución de probabilidad respectiva de cada documento en el grupo con la distribución de probabilidad de cada otro documento en el grupo; la asignación de un peso respectivo a cada par de documentos comparados, el valor del peso dependerá de la similitud entre las distribuciones de probabilidad respectivas comparadas de cada documento del par; y el cálculo de un árbol de expansión mínimo para el grupo basado en los respectivos pesos calculados.

Otro aspecto de la invención ofrece un producto de programa informático que comprende un código de programa informático para que un ordenador lleve a cabo el método de agrupamiento antes mencionado.

En un aspecto, una realización preferente de la invención provee medios mediante los cuales un corpus de documentos puede indexarse de manera eficiente en grupos de documentos, basándose en contextos estrechos específicos que se identifican de manera automática del corpus en su totalidad. La entropía se utiliza de manera ventajosa para identificar estos contextos estrechos. Este proceso de indexación facilita la formación de grupos muy pequeños, lo cual permite una recuperación más acertada.

En otro aspecto, una realización preferente provee medios para descubrir una pluralidad de contextos estrechos relevantes para la consulta de un usuario y facilita la recuperación del corpus en base a ello.

La presente invención hace referencia al área de búsqueda dentro de un corpus de documentos. Ofrece un enfoque eficiente a la indexación de documentos dentro del corpus en base al concepto de contextos estrechos. Se producen grupos de documentos de tamaño pequeño y por lo tanto se facilita la visualización de similitud entre documentos, lo que permite una rápida identificación de los documentos más relevantes dentro de un grupo. Basada en el análisis detallado de un corpus de documentos controlado, la realización preferente es compatible con una búsqueda de información temática mediante la combinación de una capacidad de búsqueda por palabra clave con exploración. Como resultado, el usuario obtiene un conjunto de documentos relevantes a los temas de sus necesidades de información.

En la realización preferente, el corpus inicialmente se analiza y divide en grupos homogéneos por temas, utilizando un algoritmo de agrupamiento cuya complejidad es proporcional al tamaño del corpus. Una búsqueda por palabra clave puede utilizarse para identificar un conjunto de grupos relevantes a una consulta. La organización interna de documentos dentro de un grupo puede representarse como una estructura tipo gráfico, en la cual documentos con mucha proximidad tienen gran similitud. El conjunto de los documentos más relevantes a la consulta se identifica de manera ventajosa dentro de un grupo a través de un proceso de exploración guiado por ordenador.

En una realización preferente, un usuario recibe, como resultado de una consulta, una lista de grupos de documentos estrechos relacionados con un tema específico. Todos los documentos, en todos los grupos devueltos, se consideran mediante el algoritmo de búsqueda como relevantes para las necesidades de información del usuario. Muchos de los documentos devueltos no contienen palabras claves específicas de la consulta en sí misma pero el sistema los considera relevantes ya que su tema o temas corresponden a las necesidades de información del usuario según lo define la consulta mistral. Por lo tanto, pueden aparecer documentos que no pueden recuperarse a través de un algoritmo de búsqueda convencional por palabra clave. Esto resulta en una alta precisión y cobertura y en una menor búsqueda para el usuario.

En una realización preferente, durante el proceso de búsqueda se devuelve una lista de grupos clasificados según su relevancia a la consulta. Estos grupos cubren una gran variedad de temas. Por lo tanto, el usuario ha enfocado el acceso a información detallada sobre toda la estructura temática de esa parte del corpus de documentos que corresponde directamente a la consulta. El usuario puede seleccionar un grupo cuyo tema o temas sean más relevante para él o ella y continuar la búsqueda dentro de este grupo utilizando técnicas de exploración asistida por ordenador que, de manera ventajosa, se basan en una visualización de árbol de expansión mínima de los documentos dentro del grupo.

El sistema preferente no está supervisado, pero tiene una complejidad proporcional al tamaño del grupo de documentos y, por lo tanto, es efectivo en cuanto al coste y escalable a aplicaciones de la vida real, tales como una búsqueda con un corpus controlado.

El algoritmo de agrupamiento preferente, al aplicarlo a un corpus de documentos de temas heterogéneos, produce grupos de temas homogéneos de tamaño relativamente pequeño. Esto simplifica la etapa de recuperación, ya que los grupos recuperados son lo suficientemente compactos para permitir una exploración guiada por ordenador eficiente.

La realización preferente no utiliza un modelo probabilístico y como tal la cantidad de grupos se determina en forma automática. Además, el enfoque teórico de información preferente (entropía y divergencia de Jensen-Shannon) presenta una medida más natural para calcular la similitud entre documentos, documentos y centroides/atractores de grupos y la amplitud o estrechez de un contexto.

Puede presentarse al usuario información sobre la estructura temática de todo el corpus de documentos. Esta información puede utilizarse para mejorar la calidad de las consultas preparadas por motores de búsqueda convencionales de empresas. Además, es posible utilizar la presente invención en combinación con un motor de búsqueda convencional en Internet, tal como Google y Alta Vista, para aumentar la utilidad de los resultados del procedimiento de búsqueda, mediante un efecto sinergístico.

Otros aspectos ventajosos de la invención serán evidentes para los expertos en el arte al revisar la siguiente descripción de una realización preferente de la invención.

Breve descripción de los dibujos

A continuación se describirá una realización preferente de la invención mediante un ejemplo y con referencia a los dibujos adjuntos, en los cuales:

La figura 1 muestra un diagrama de bloques de un proceso de recuperación de documentos;

La figura 2 muestra un proceso de indexación del proceso de recuperación de documentos de la figura 1 en más detalle;

La figura 3 muestra un proceso de búsqueda por palabra clave que muestra cómo se identifican los grupos relevantes a partir de la consulta de un usuario;

La figura 4 muestra un proceso de exploración guiado por ordenador; y

La figura 5 muestra un proceso de recuperación de documentos basado en ordenador.

Descripción detallada de los dibujos

Con referencia a la figura 5, se muestra un ejemplo de un sistema de recuperación de documentos generalmente indicado con el número 10. El sistema 10 comprende al menos un ordenador 12 dispuesto para ejecutar uno o más programas informáticos que se representan en forma colectiva con el número 14.

El ordenador 12 se comunica con un dispositivo de almacenamiento 24 que contiene un corpus de documentos 16 (o cualquier forma de datos no estructurados heterogéneos) (figuras 1 y 2). El corpus de documentos 16 comprende una pluralidad de documentos que pueden leerse por ordenador (no se muestran). El corpus de documentos 16 puede ser un corpus controlado, lo que significa que el mantenimiento del corpus 16 es controlado por uno o más administradores (no se muestran). Además, los documentos en el corpus de documentos 16 normalmente son heterogéneos en contenido. El dispositivo de almacenamiento 24 comprende una base de datos o cualquier otro medio de reposición o almacenamiento de datos. Se comprenderá que el término “documento” como se utiliza en la presente abarca cualquier cuerpo de textos o términos que pueden leerse por ordenador o de manera electrónica, incluyendo correos electrónicos, mensajes de Servicio de Mensajes Cortos (SMS), documentos de procesador de texto, o cualquier otro archivo o mensajes de texto o archivos electrónicos generados por otra aplicación.

En una realización preferente, los programas informáticos 14 incluyen un módulo 18 para analizar, o indexar, el corpus de documentos 16 para poder formar grupos de documentos (cada grupo de documentos comprende un conjunto de uno o más documentos del corpus 16 que se han agrupado juntos). El módulo de análisis 18 incluye medios para implementar un algoritmo de agrupamiento, como se describirá en más detalle abajo. También se proporciona un módulo o motor de búsqueda 20 que es capaz de realizar búsquedas por palabras claves a través de una pluralidad de documentos que pueden leerse por ordenador y presentar un conjunto de resultados (normalmente comprenden uno o más documentos) al usuario a través de una interfaz de usuario 26. Preferentemente, también se proporciona un módulo de exploración 22 que es capaz de permitir a un usuario, a través de una interfaz de usuario 26, explorar los resultados provistos por el módulo de búsqueda 20. Se comprenderá que los módulos de análisis, búsqueda y exploración 18, 20, 22 pueden comprender, cada uno, uno o más programas informáticos.

El dispositivo de almacenamiento 24 también puede utilizarse para almacenar datos generados por uno o más de los módulos de análisis, búsqueda y exploración 18, 20, 22. En particular, los medios de almacenamiento 24 pueden utilizarse para almacenar contexto de términos, grupos de documentos y, en la realización preferente, árboles de expansión mínima, como se describe en más detalle más abajo. Queda entendido que el corpus de documentos 16, los grupos de documentos y otros datos no necesariamente deben almacenarse en el mismo dispositivo de almacenamiento.

El sistema 10 de la figura 5 se muestra de manera simplificada. Los componentes del sistema pueden comunicarse directamente unos con otros o a través de una red de ordenadores, tales como una intranet. Además, uno o más ordenadores, o clientes (no se muestran) pueden estar presentes en el sistema 10, cada uno capaz de ejecutar al menos el módulo de búsqueda 20, una interfaz de usuario 26 y normalmente también el módulo de exploración 22, o en algunos casos, para que un usuario respectivo pueda buscar a través de los contenidos agrupados del corpus de documentos 16. Normalmente, dichos ordenadores de clientes se comunican con los medios de almacenamiento 24 mediante una red de ordenadores.

La figura 1 ofrece un resumen del proceso de recuperación de documentos preferente. El proceso puede considerarse en tres etapas. La primera etapa es la *indexación*, la segunda la *búsqueda por palabra clave* y la tercera es la *exploración guiada por ordenador*. La etapa de indexación incluye el análisis de documentos en el corpus de documentos 16 para generar grupos de documentos y, en una realización preferente, también para organizar los documentos dentro de cada grupo, por ejemplo a través de un árbol de expansión mínima. La etapa de búsqueda por palabra clave incluye recuperar uno o más grupos de documentos y, donde corresponda, un árbol de expansión mínima respectivo, en respuesta a la búsqueda de un usuario que comprenda palabras claves. La etapa de exploración incluye explorar a través de los documentos de los grupos devueltos, en forma ventajosa utilizando el árbol de expansión mínima.

La siguiente descripción de una realización preferente de la invención se divide en dos partes. La primera, el análisis del corpus, describe cómo se indexa el corpus de documentos 16 para facilitar la recuperación (ver sección 1). La segunda sección, la recuperación, describe la búsqueda real (sección 2.1), y la exploración (sección 2.2) describe los procesos utilizados para encontrar los documentos más relevantes del corpus 16 basados en la consulta de un usuario.

1. *Análisis del corpus*

A continuación se describirá un método preferente de análisis del corpus que puede implementarse mediante el módulo de análisis 18.

El análisis del corpus se describe con referencia a la figura 2. La figura 2 ilustra cómo los grupos de documentos 26 se forman en base a perfiles generados 30 de los documentos y mediante la definición de uno o más contextos 32 para cada documento. Las figuras 3 y 4 muestran cómo puede utilizarse el concepto de un árbol de expansión mínima 40 para visualizar la similitud entre documentos dentro de los grupos.

El análisis del corpus puede dividirse en tres etapas: Descubrimiento de contextos estrechos; Agrupamiento de documentos; y Descubrimiento de estructura interna del grupo.

1. 1 *Descubrimiento de contextos estrechos*

Un problema con el agrupamiento de documentos es determinar una entidad respectiva que sirva como el enfoque, o atractor, para cada grupo. La selección de atractores de grupo puede tener un gran impacto en el tamaño y contenido del grupo resultante el cual a su vez tiene un gran impacto en la efectividad de una búsqueda por los grupos. Un aspecto de la presente invención proporciona un método para determinar atractores de grupo. En la realización preferente esto incluye identificar uno o más contextos que se consideren relativamente estrechos, como se describe en más detalle a continuación.

Cada documento en el corpus 16 comprende una o más instancias de al menos un término y normalmente comprende una pluralidad de términos. Un término comprende una o más palabras. Un contexto para un término se representa mediante una distribución de probabilidad condicional, o vector de indicadores de probabilidad, sobre todos, o una pluralidad de, los términos del corpus, donde la distribución de probabilidad se define por la frecuencia de los términos que co- aparecen con el término del contexto. En la realización preferente, la frecuencia de cada término es la cantidad total de apariciones del término en todos los documentos en donde el término co-aparece con el término del contexto. Preferentemente, cada frecuencia de término se normaliza con respecto a, o divide por, la cantidad total de términos en todos los documentos en donde aparece el término del contexto. En general, dado un término de contexto, la probabilidad condicional de un término que co-aparece es la cantidad total de veces normalizada que el término que co-aparece aparece en todos los documentos en donde aparece el término de contexto.

Por lo tanto, los contextos relativos pueden representarse mediante una distribución, o vector, de probabilidad respectivo, en una pluralidad de términos. Cada vector de probabilidad normalmente comprende al menos un término, cada término en el vector se asocia a un indicador de probabilidad, o frecuencia, respectivo. El indicador de probabilidad, o frecuencia, preferentemente indica la probabilidad de aparición o frecuencia de aparición de un término respectivo en documentos que contienen el término de contexto. Un contexto puede verse como amplio o estrecho en alcance. En la realización preferente, un contexto se describe como estrecho si la entropía es pequeña, y amplio si la entropía es grande.

Con referencia a la figura 2, se realiza un análisis de los documentos en el corpus 16 (módulo 31 en la figura 2) para generar perfiles de documento 30, como se describe en más detalle abajo. También se realiza un análisis de los documentos en el corpus 16 (módulo 33 en la figura 2) para calcular un contexto respectivo para cada término en el

corpus 16, como se describe en más detalle abajo. En la figura 2, los perfiles de documentos almacenados se indican con el número 30 y los contextos almacenados se indican con el número 32 y pueden almacenarse de cualquier manera conveniente, por ejemplo en los medios de almacenamiento 24.

A continuación se describe un método preferente para calcular contextos. X denota el conjunto de todos los documentos en el corpus de documentos 16 e Y denota el conjunto de todos los términos presentes en X, es decir, el conjunto de todos los términos presentes en uno o más de todos los documentos en el corpus 16. Para un término dado z, donde $z \in Y$, podemos definir un contexto para el término z como un contexto que comprende un tema de un conjunto de documentos en donde aparece el término z. Más específicamente, el contexto del término z, preferentemente, se representa en forma de distribución de probabilidad condicional $P(Y|z)$. Aquí la variable aleatoria Y adquiere valores de Y y $p(y|z)$ es la probabilidad de que, en un documento seleccionado al azar del corpus 16 que contiene el término z, un término seleccionado aleatoriamente es el término y. Sólo si la variable Y representa un término que co-aparece con el término z en un documento, su probabilidad respectiva contribuirá a la distribución de probabilidad general $P(Y|z)$. La distribución de probabilidad $P(Y|z)$ puede calcularse como:

$$p(y \mid z) = \frac{\sum_{x \in X(z)} tf(x,y)}{\sum_{x \in X(z), t \in Y} tf(x,t)},$$

donde $tf(x,y)$ es la frecuencia del término y en el documento x, y $X(z)$ es el conjunto de todos los documentos del corpus 16 que contienen el término z y donde t es un índice de términos.

Por lo tanto, dado un corpus de documentos 16 y un término z, podemos describir el contexto de este término como un conjunto ponderado, o vector, de términos que aparecen con el término z dado en uno o más documentos del corpus 16. La ponderación respectiva asociada a cada término, preferentemente, comprende una indicación de la probabilidad de aparición, o frecuencia de aparición, del término respectivo en todos los documentos en los que aparece el término z. En una realización, para generar este contexto, todos los documentos del corpus 16 en los que aparece el término z dado pueden combinarse en un solo prototipo de documento. La frecuencia respectiva de aparición de todos los términos en este nuevo prototipo de documento puede calcularse y, convenientemente, normalizarse dividiendo por la extensión del documento prototipo (es decir, la cantidad de términos en el documento prototipo). Estas frecuencias normalizadas de todos los términos del documento prototipo combinado representa el contexto para el término z. Por lo tanto, el contexto para el término z comprende una pluralidad o vector de términos que co-aparecen con el término z en al menos uno de los documentos del corpus 16, cada término de este conjunto se asocia con una ponderación respectiva que indica la frecuencia de aparición del término respectivo en el, o en cada, documento donde aparece el término z. Para cada término presente en los documentos del corpus 16 se calcula un contexto respectivo. Pueden utilizarse medidas alternativas de frecuencia de aparición.

En muchos casos, el contexto de un término z es demasiado amplio para presentar información útil sobre el corpus 16. Por lo tanto, lo deseable es identificar términos que aparecen en contextos más estrechos. La delimitación del contexto de un término z se calcula preferentemente como la entropía $H(Y|z)$ de la distribución de probabilidad $P(Y|z)$ respectiva, en donde:

$$H(Y \mid z) = -\sum_{y} p(y \mid z) \log p(y \mid z).$$

Si el valor de entropía es relativamente pequeño el contexto se considera “estrecho”, de lo contrario, se considera “amplio”.

En la realización preferente, un contexto respectivo se calcula para todos los términos significativos y no redundantes que aparecen en el corpus. Un valor de entropía respectivo se calcula para cada contexto. Basado en los valores de entropía respectivos, se seleccionan algunos de los contextos para proporcionar un centro, o atractor, para un grupo de documentos respectivo. Como se describirá con más detalle abajo, con este fin se seleccionan los contextos con entropía relativamente baja.

Y(z) denota el conjunto de todos los términos diferentes de los documentos de X(z). Donde hay una distribución uniforme de términos de Y(z) la entropía $H(Y|z)$ es igual al logaritmo |Y(z)|. Según la ley de Heaps [1] log |Y(z)| = O(log |X(z)|). Dado que la frecuencia de documentos para z es $df(z) \equiv |X(z)|$ hay una relación logarítmica entre la entropía y la frecuencia de documentos y, como tal, es razonable utilizar la frecuencia de documentos como medio para determinar los enlaces en los subconjuntos de contextos estrechos.

La estrechez, o amplitud, de un contexto puede determinarse no sólo en términos de entropía sino también por la frecuencia de aparición del término respectivo en forma colectiva por todos los documentos en el corpus 16. Es difícil establecer de antemano la cantidad de contextos estrechos que se requerirán para describir el contenido del corpus 16 en detalle, es decir, la cantidad de atractores de grupo que se requieren. En una realización preferente, cada término en el conjunto de términos Y se asigna a una pluralidad de subconjuntos de términos según la frecuencia de

aparición del término respectivo en el corpus 16, y preferentemente según la cantidad de documentos en el corpus 16 en los que aparece el término respectivo (aunque pueden utilizarse otras medidas de frecuencia). Los subconjuntos preferentemente están separados, o no se superponen, en sus rangos de frecuencia respectivos. Esto se expresa de la siguiente manera:

$$Y = \bigcup_i Y_i, Y_i = \{z : z \in Y, df_i \leq df(z) \leq df_{i+1}\}, i = 1, \ldots, r.$$

Aquí $df(z) \equiv |X(z)|$ se refiere a la frecuencia en el documento de un término z, la frecuencia en el documento es la cantidad de veces que aparece el término z en un documento y r es la cantidad de subconjuntos. El parámetro r puede tomar una variedad de valores y puede fijarse como un parámetro de sistema según el tamaño y la naturaleza del corpus 16.

En la realización preferente, todos los términos asignados a un subconjunto dado satisfacen los siguientes requisitos: la frecuencia de aparición del término respectivo está dentro del rango de frecuencia respectivo para el subconjunto respectivo; los rangos de frecuencia para los subconjuntos respectivos están separados, o no se superponen; y, más preferentemente, el tamaño del rango de frecuencia de un subconjunto dado es igual a una constante multiplicada por el rango del conjunto anterior. La constante puede tener cualquier valor mayor que uno. Por ejemplo, en una realización, la constante adquiere el valor 2. Por lo tanto, el rango de frecuencia de un subconjunto dado es dos veces mayor que el rango de frecuencia del subconjunto anterior y la mitad de grande que el rango de frecuencia del siguiente subconjunto. Así, por ejemplo, un primer subconjunto puede contener términos que aparecen, digamos, en entre 1 y 25 documentos en el corpus 16, el segundo subconjunto puede contener términos que aparecen en entre 26 y 75 documentos, el tercer subconjunto puede contener términos que aparecen en entre 76 y 175 documentos, y así sucesivamente.

Se asigna un umbral de entropía a cada subconjunto de términos. Los umbrales respectivos se preestablecen de manera conveniente como parámetros de sistema, como también puede hacerse con la definición de los subconjuntos de términos. Preferentemente, el umbral para un subconjunto dado es menor que el umbral para el siguiente subconjunto (donde el siguiente subconjunto contiene términos con una frecuencia mayor que el subconjunto anterior). Más preferentemente, el umbral respectivo para subconjunto sucesivos (con rango de frecuencia en aumento) aumenta en forma lineal. Más formalmente, en una realización preferente, los umbrales df_i pueden satisfacer la condición $df_{i+1} = \alpha\, df_i$, donde $\alpha > 1$ es una constante. Puede verse que $H(Y|z)$ se vincula desde arriba mediante una función lineal del índice i y por lo tanto es razonable fijar un umbral de función lineal $H_{max}(i)$ para seleccionar un conjunto Z de términos dentro de un contexto estrecho:

$$Z = \bigcup_i \{z : z \in Y_i, H(Y \mid z) \leq H_{max}(i)\}.$$

Por lo tanto, en una realización preferente, el contexto de un término se selecciona como estrecho si la entropía es menor o igual que el valor del umbral respectivo asociado con el subconjunto de términos a los cuales se asigna el término.

1. 2 *Agrupamiento de documentos*

El conjunto de contextos estrechos $\{P(Y|z)\}_{z \in Z}$ se considera atractores de grupo, es decir, cada contexto estrecho seleccionado puede utilizarse como una entidad central sobre la cual puede formarse un grupo de documentos respectivo.

Para agrupar documentos, cada documento x en el corpus 16 está representado en forma de un perfil de documento respectivo 30 que comprende una distribución de probabilidad $P(Y|x)$ respectiva, donde

$$p(y \mid x) = \frac{tf(x,y)}{\sum_{t \in Y} tf(x,t)}$$

Por lo tanto, el perfil de documento 30 para cada documento x comprende un conjunto ponderado, o vector, de términos que aparecen dentro del documento x. La ponderación respectiva asociada a cada término en el vector, preferentemente, comprende una indicación de la probabilidad de aparición, o frecuencia de aparición, del término respectivo en el documento x. De manera conveniente, la ponderación respectiva comprende la frecuencia de aparición del término respectivo en el documento x, normalizada con respecto a, o dividida por, la cantidad de términos en el documento x.

El agrupamiento de documentos (módulo 34 en la figura 2) puede realizarse comparando el perfil del documento respectivo 30 de cada documento en el corpus 16 con cada contexto 32 seleccionado como atractor de grupo. En la realización preferente, cada documento del corpus 16 se asocia con, o se asigna a, uno o más contextos 32 que se ajusta mejor, o se asemeja a, su perfil de documento 30. El resultado de asignar cada documento del corpus 16 a un

contexto del atractor de grupo 32 es crear una pluralidad de grupos de documentos (colectivamente representados con el número 36 en la figura 2), donde cada grupo comprende una pluralidad de documentos respectivos del corpus 16.

Un método preferente de comparar perfiles de documentos 30 con contextos 32 seleccionados como atractores de grupo es calcular la distancia, o similitud, entre un documento *x* y el contexto del término *z* utilizando, por ejemplo, la divergencia de Jensen-Shannon [2] entre las distribuciones de probabilidad p_1 y p_2, donde el documento *x* (es decir, su perfil de documento 30) y el contexto del término *z* representan:

$$JS_{\{0.5,0.5\}}[p_1, p_2] = H[\overline{p}] - 0.5H[p_1] - 0.5H[p_2],$$

donde $H[p]$ denota la entropía de la distribución de probabilidad p y $\overline{p}$ denota la distribución de probabilidad promedio $\overline{p} = 0.5\ p_1 + 0.5\ p_2$.

Por lo tanto, un documento x se asigna a un grupo 36 con contexto atractor del término z si:

$$z = \arg\min_{t \in Z} JS_{\{0.5,0.5\}}[P(Y \mid t), P(Y \mid x)].$$

Por lo tanto, un documento se asigna a un grupo donde la divergencia de Jensen-Shannon entre un documento y un atractor es un mínimo para todos los atractores.

1. 3 *Estructura interna de grupos*

En una realización preferente, los documentos dentro de cada grupo 36 se analizan, en forma ventajosa, para estructurar u organizar los documentos dentro de cada grupo 36. Con este fin, los documentos dentro de un grupo 36 se representan como un conjunto de vértices de gráficos hipotéticos, donde cada documento corresponde a un vértice respectivo. A cada vértice se asigna una ponderación que es igual a, o depende de, la distancia entre el documento respectivo y el contexto del atractor de grupo respectivo. Cada vértice se asocia a, o conecta con, cada otro vértice mediante un borde no orientado respectivo, cuya ponderación depende de, o es igual a, la distancia entre los respectivos perfiles de documentos 30 de los documentos correspondientes. La distancia entre los perfiles de documentos 30 puede determinarse en forma conveniente utilizando la divergencia de Jensen-Shannon de manera similar a la antes descrita. Puede utilizarse un algoritmo, por ejemplo, el algoritmo estándar de Kruskal [3] (ver módulo 38 en la figura 2) para construir un árbol de expansión mínima 40 que se extiende hasta todos los vértices de gráficos y tiene una ponderación promedio mínima para sus bordes. El árbol de expansión mínima puede presentarse a un usuario a través de una interfaz de usuario adecuada como una descripción completa de la estructura interna del grupo respectivo. Cabe destacar que pueden utilizarse otros algoritmos convencionales para construir el árbol de expansión mínima.

Las operaciones descritas en la sección 1 pueden realizarse en forma conveniente mediante el módulo de análisis 18. Los datos generados por el módulo de análisis 18 incluyendo, según corresponda, perfiles de documentos 30, contextos 32, grupos 36 y árboles de expansión mínima 40 pueden almacenarse de cualquier manera conveniente, por ejemplo en medios de almacenamiento 24. Las operaciones realizadas por el módulo de análisis 18 normalmente se realizan sin conexión, antes de que un usuario comience a buscar el corpus de documentos 16.

2. *Recuperación*

El algoritmo de recuperación preferente comprende medios para realizar una búsqueda por palabra clave (figura 3) en combinación con medios para exploración (figura 4), a través de los resultados de la búsqueda por palabra clave para encontrar los documentos más relevantes para un usuario. La búsqueda por palabra clave y exploración puede realizarse mediante módulos separados 20, 22 respectivamente, o puede realizarse mediante un solo módulo.

2. 1 *Búsqueda por palabras claves*

El objetivo de la fase de búsqueda por palabras claves es encontrar uno o más grupos 36 que sean relevantes a los temas de las necesidades de información de un usuario, según lo definen una o más palabras claves proporcionadas por el usuario en una consulta de búsqueda.

En la figura 3 puede verse que cada grupo de documentos 36 comprende una pluralidad de documentos (cada uno representado por un vértice respectivo 41), un contexto de atractor 32 y un centroide 42. El centroide 42 comprende una distribución de probabilidad promedio, o vector, $P_{avr}(Y|z)$, de términos que aparecen en los documentos del respectivo grupo de documentos C(z). $P_{avr}(Y|z)$ puede calcularse de manera similar a $P(Y|z)$ como se describe con anterioridad, utilizando la siguiente ecuación (que es similar a la ecuación antes dada para $p(y|z)$):

$$p_{avr}(y \mid z) = \frac{\sum_{x \in C(z)} tf(x,y)}{\sum_{x \in C(z), t \in Y} tf(x,t)}.$$

Por lo tanto, un grupo de documentos C(z) puede representarse mediante dos distribuciones de probabilidad para el conjunto de términos Y, a saber:

(i) el contexto $P(Y|z)$ del término z sirve como atractor para el grupo; y

(ii) la distribución de probabilidad promedio P_{avr} $(Y|z)$, o centroide 42, que presenta información sobre todos los documentos asignados al grupo.

Durante la fase de búsqueda por palabras claves, el usuario presenta sus necesidades de información en forma de un conjunto de palabras claves $Q = \{q_1, k, q_s\}$. Es deseable en una búsqueda en un corpus de documentos controlado que el sistema alcance la máxima cobertura, por lo tanto todos los grupos 36, C(z) que se calcula que son relevantes a la consulta se presentan preferentemente al usuario a través de la interfaz de usuario 26. Los criterios preferentes para determinar la relevancia del grupo $C(z)$ a la consulta Q son los siguientes:

$$p(q_i \mid z) \cdot p_{avr}(q_i \mid z) > 0, i = 1, \ldots, s.$$

Es decir, un grupo 36, $C(z)$ se considera relevante a la consulta Q si todas las palabras claves de la consulta están presentes en el atractor de grupo y al menos un documento de ese grupo (y por lo tanto en el centroide 42).

El grado de relevancia del grupo $C(z)$ a la consulta Q puede calcularse como una función de adición de los resultados de palabras claves, de la siguiente manera:

$$rel(Q, C(z)) = \sum_{q \in Q} p(q \mid z) p_{avr}(q \mid z).$$

Esto permite que se presenten al usuario grupos relevantes clasificados en orden de relevancia. En la realización preferente, se devuelve una lista de grupos relevantes al usuario en orden descendente de resultados de relevancia.

Para calcular el contenido de un grupo se puede proporcionar al usuario una breve descripción del resumen del grupo en forma de lista de los términos con más ponderación que aparecen en los documentos del grupo. Dado el término t, su ponderación dentro de un grupo $C(z)$ se calcula como una función de multiplicación de su ponderación respectiva en el vector de atractor de grupo $P(Y|z)$ y el vector de contenido del grupo P_{avr} $(Y|z)$ relevante. Por lo tanto:

$$peso \quad (t \mid C(z)) = p(t \mid z) p_{avr}(t \mid z).$$

2. 2 *Exploración de grupos*

El proceso de exploración preferente se ilustra en la figura 4. La figura 4 muestra un árbol de expansión mínima 40 dividido en subárboles utilizando un diámetro preestablecido, o parámetro de distancia, es decir, la distancia respectiva entre cada documento y cada otro documento en un grupo puede calcularse, por ejemplo, utilizando la divergencia Jensen-Shannon, y los documentos pueden agruparse en subárboles, cada documento en un subárbol se encuentra a una distancia de cada otro documento en el subárbol que es igual o menor que el parámetro de distancia predeterminado. Esto permite al usuario encontrar rápidamente el subárbol más relevante dentro del cual se encuentran los documentos más relevantes. El árbol de expansión mínima 40 puede utilizarse para calcular distancias entre documentos 41 en el grupo 36. Cualquier subárbol del árbol $T(z)$, dentro del diámetro predefinido, puede considerarse un subárbol o subgrupo 36' del grupo $C(z)$ principal. Esto permite la división eficiente de grandes grupos en subgrupos 36 más pequeños, generalmente homogéneos. Un breve resumen para cada subgrupo 36' puede generarse de la misma manera que el resumen para $C(z)$. El usuario puede revisar rápidamente los resúmenes de todos los subgrupos y seleccionar uno o más para explorar con más detalle.

Como puede observarse en la figura 4, el usuario puede navegar documentos en un grupo utilizando una representación visual del árbol de expansión mínima 40, $T(z)$ como guía. El árbol 40 representa gráficamente la similitud entre los documentos 41 en el grupo 36 relevante y otros documentos 41 en el grupo 36 y el atractor de grupo 32.

Pueden utilizarse diferentes enfoques para ayudar al usuario a seleccionar eficientemente la parte del árbol 40 con los documentos más relevantes. Por ejemplo, una búsqueda por palabras claves convencional puede utilizarse para encontrar un conjunto de documentos que contienen el conjunto dado de palabras claves. Estos documentos pueden considerarse como puntos de partida para explorar utilizando el árbol de expansión mínima 40 para visualizar la estructura interna del grupo 36. Dado que cada grupo 36 contiene una cantidad relativamente pequeña de documentos 41, es preferible utilizar un vocabulario controlado, generado automáticamente a partir de los propios documentos (es decir, que comprende términos que aparecen en el corpus de documentos 16 en su totalidad, o sólo en los documentos del grupo o de cada grupo 36 que se está explorando), para ayudar al usuario a generar buenas consultas.

Las siguientes características y ventajas de la realización preferente de la invención serán evidentes a partir de la siguiente descripción. Los contextos estrechos pueden distinguirse, o identificarse, mediante palabras, o términos, que sólo aparecen en documentos que pertenecen a, o se asocian con, el contexto estrecho respectivo, donde la estrechez se mide por entropía condicional.

Durante el proceso de agrupamiento, la similitud entre documentos no se calcula directamente sino a través de su similitud con un atractor de grupo, utilizando, por ejemplo, la divergencia de Jensen-Shannon. Esto produce cálculos de similitud más precisos comparados con los enfoques convencionales en los que se comparan todos los documentos directamente con cada uno de los otros para determinar similitudes. Este enfoque convencional es irrelevante e innecesario en muchos casos, ya que a menudo los documentos se asemejan poco unos a otros y por lo tanto no hay necesidad de calcular la similitud. En la realización preferente, la similitud entre documentos se determina sólo para documentos dentro del mismo grupo donde se sabe que tienen temas comunes. Esto produce un enfoque más relevante, eficiente y escalable para la determinación de similitud. Las técnicas de agrupamiento descritas en el presente documento no están supervisadas y por lo tanto no necesitan conocimientos contextuales. Si se aplica el método de indexación preferente a un corpus de documentos de temas heterogéneos, se generan grupos de temas homogéneos de tamaño relativamente pequeño lo que permite explorar los documentos que contienen con la ayuda de un ordenador. Una característica adicional del método de indexación preferente es que los temas similares a la consulta del usuario, pero que no se mencionan directamente en la consulta, se identifican automáticamente como relevantes. Esto puede verse como creatividad por parte del algoritmo de indexación. Por lo tanto, durante una búsqueda por palabras claves los documentos agrupados por estos temas similares se devuelven como parte de la consulta, lo que mejora la cobertura. El árbol de expansión mínima describe la estructura del grupo de manera tal que cualquier subárbol del árbol dentro de un diámetro pequeño dado puede considerarse como un subgrupo temático estrecho del grupo dado. Esto mejora la eficiencia de la búsqueda para el usuario, ya que se pueden generar resúmenes breves de estos subgrupos y utilizarlos para encontrar los documentos más relevantes.

Referencias

[1] **Baeza-Yates** and **Ribeiro-Neto**, Modern Information Retrieval, *ACM Press*, 1999.

[2] **Lin**, J. Divergence Measures Based on the Shannon Entropy, *IEEE Transactions on Information Theory*, 37(1), pp. 145-151, 1991.

[3] **Kruskal**, J.B. On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem, *Proc. Amer. Math. Soc.*, 7:1, pp. 48-50, 1956.

[4] **Pedersen** Jan. O., **Karger** D., **Cutting** D. R., **Tukey** J. W. Scatter-gather: a cluster-based method and apparatus for browsing large document collections. US Patent 5,442,778. August 15, 1995.

[5] **Swannack** C. M., **Coppin** B.K., **McKay Grant** C.A., **Charlton** C.T. Data acquisition system. Jul. 4, 2002. US Patent Application US 2002/0087515 A1.

[6] **Hofmann** T., **Pusicha** J. C. System and method for personalized search, information filtering, and for generating recommendations utilizing statistical latent class models. US Patent Application 2002/0107853 A1, Aug.8, 2002

[7] **Wing** S. **Wong**, An Qin. Method and apparatus for establishing topic word classes based on an entropy cost function to retrieve documents represented by the topic words. US Patent 6,128,613. Oct. 3, 2000.

REIVINDICACIONES

1. Método para determinar atractores de grupo (32) en un aparato (34) para determinar atractores de grupo (32) para una pluralidad de documentos (41), donde cada documento comprende al menos un término, donde dicho método comprende: calcular, respecto de cada término, una distribución de probabilidad que indica la frecuencia de aparición de otro término, o de cada otro término que coaparece con dicho término en al menos uno de dichos documentos; calcular, respecto de cada término, la entropía de la distribución de probabilidad respectiva; seleccionar al menos una de dichas distribuciones de probabilidad como un atractor de grupo (32) según el valor de entropía respectivo.

2. Método según la reivindicación 1, en donde cada distribución de probabilidad comprende, respecto de cada término que coaparece, un indicador que indica la cantidad total del término respectivo que coaparece en todos los documentos (41) en los cuales coaparece el término respectivo que coaparece con el término respecto del cual se calcula la distribución de probabilidad.

3. Método según la reivindicación 1 ó 2, en donde cada distribución de probabilidad comprende, respecto de cada término que coaparece, un indicador que comprende una probabilidad condicional de la aparición del término respectivo que coaparece en un documento (41) si aparece en dicho documento el término respecto del cual se calcula la distribución de probabilidad.

4. Método según cualquiera de las reivindicaciones 1 a 3, en donde cada indicador se normaliza respecto de la cantidad total de términos en el documento, o en cada documento (41) en el cual aparece el término respecto del cual se calcula la probabilidad de distribución.

5. Método según la reivindicación 1, que comprende asignar cada término a un subconjunto de una pluralidad de subconjuntos de términos según la frecuencia de aparición del término; y seleccionar, como atractor de grupo (32), la distribución de probabilidad respectiva de uno o más términos de cada subconjunto de términos.

6. Método según la reivindicación 5, en donde cada término se asigna a un subconjunto según la cantidad de documentos del corpus donde aparece el término respectivo.

7. Método según la reivindicación 5 ó 6, en donde un umbral de entropía se asigna a cada subconjunto, el método comprende seleccionar, como atractor de grupo (32), la distribución de probabilidad respectiva de uno o más términos de cada subconjunto que tiene una entropía que satisface el umbral de entropía respectiva.

8. Método según la reivindicación 7, que comprende seleccionar, como atractor de grupo (32), la distribución de probabilidad respectiva de uno o más términos de cada subconjunto que tiene una entropía que es menor o igual al umbral de entropía respectiva.

9. Método según cualquiera de las reivindicaciones de 5 a 8, en donde cada subconjunto se asocia a un rango de frecuencia y en donde los rangos de frecuencia de los subconjuntos respectivos están separados.

10. Método según cualquiera de las reivindicaciones de 5 a 9, en donde cada subconjunto se asocia a un rango de frecuencia, el tamaño de cada rango de frecuencia sucesivo es igual a una constante multiplicada por el tamaño del rango de frecuencia precedente en orden de frecuencia ascendente.

11. Método según cualquiera de las reivindicaciones 7 a 10, en donde cada umbral de entropía respectivo aumenta para subconjuntos sucesivos en orden de frecuencia ascendente.

12. Método según la reivindicación 11, en donde el umbral de entropía respectivo para subconjuntos sucesivos aumenta de manera lineal.

13. Programa informático que comprende un código de programa informático que hace que un ordenador lleve a cabo el método de la reivindicación 1.

14. Aparato (34) para determinar atractores de grupo (32) para una pluralidad de documentos (41), donde cada documento comprende al menos un término, donde el aparato comprende: medios para calcular, respecto de cada término, una distribución de probabilidad que indica la frecuencia de aparición del otro término o cada otro término que comparece con dicho término en al menos uno de dichos documentos (41); medios para calcular, respecto de cada término, la entropía de la distribución de probabilidad respectiva; y medios para seleccionar al menos una de dichas distribuciones de probabilidad como un atractor de grupo según el valor de entropía respectivo.

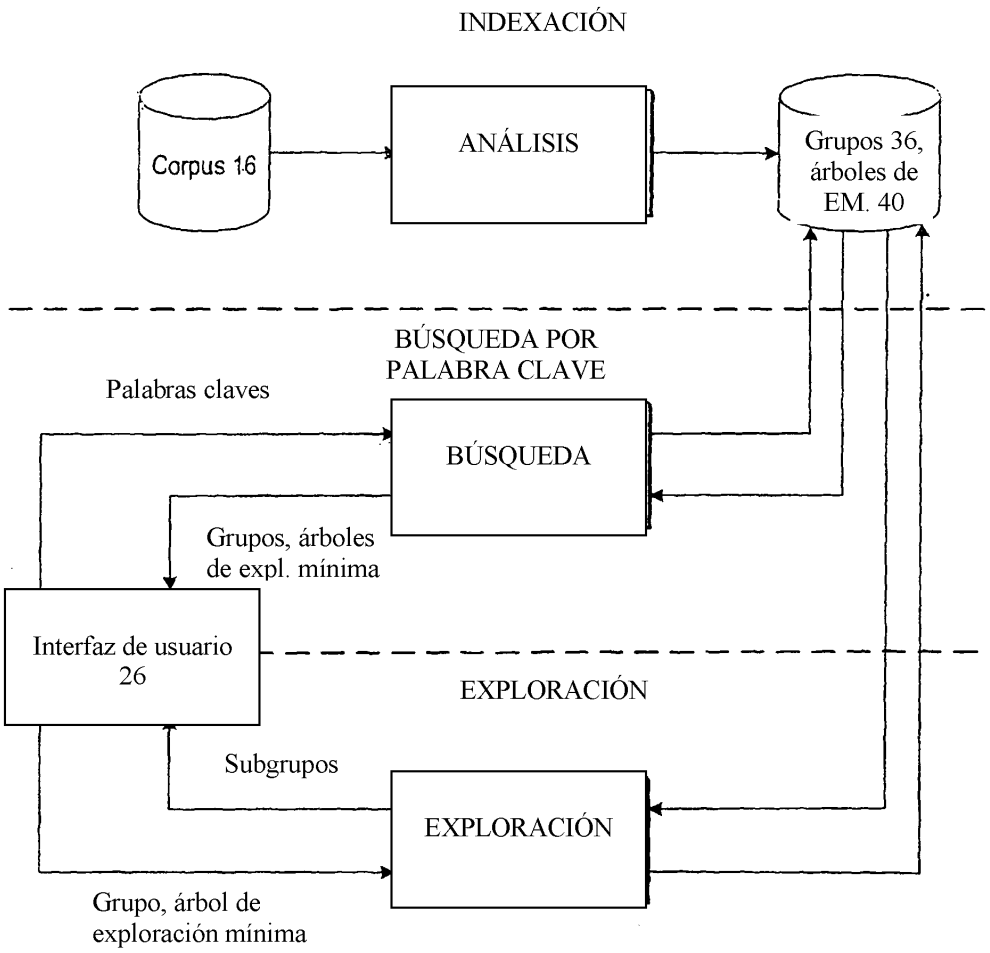
INDEXACIÓN
Corpus 16
ANÁLISIS
Grupos 36, árboles de EM. 40
BÚSQUEDA POR PALABRA CLAVE
Palabras claves
BÚSQUEDA
Grupos, árboles de expl. mínima
Interfaz de usuario 26
EXPLORACIÓN
Subgrupos
EXPLORACIÓN
Grupo, árbol de exploración mínima

Fig. 1

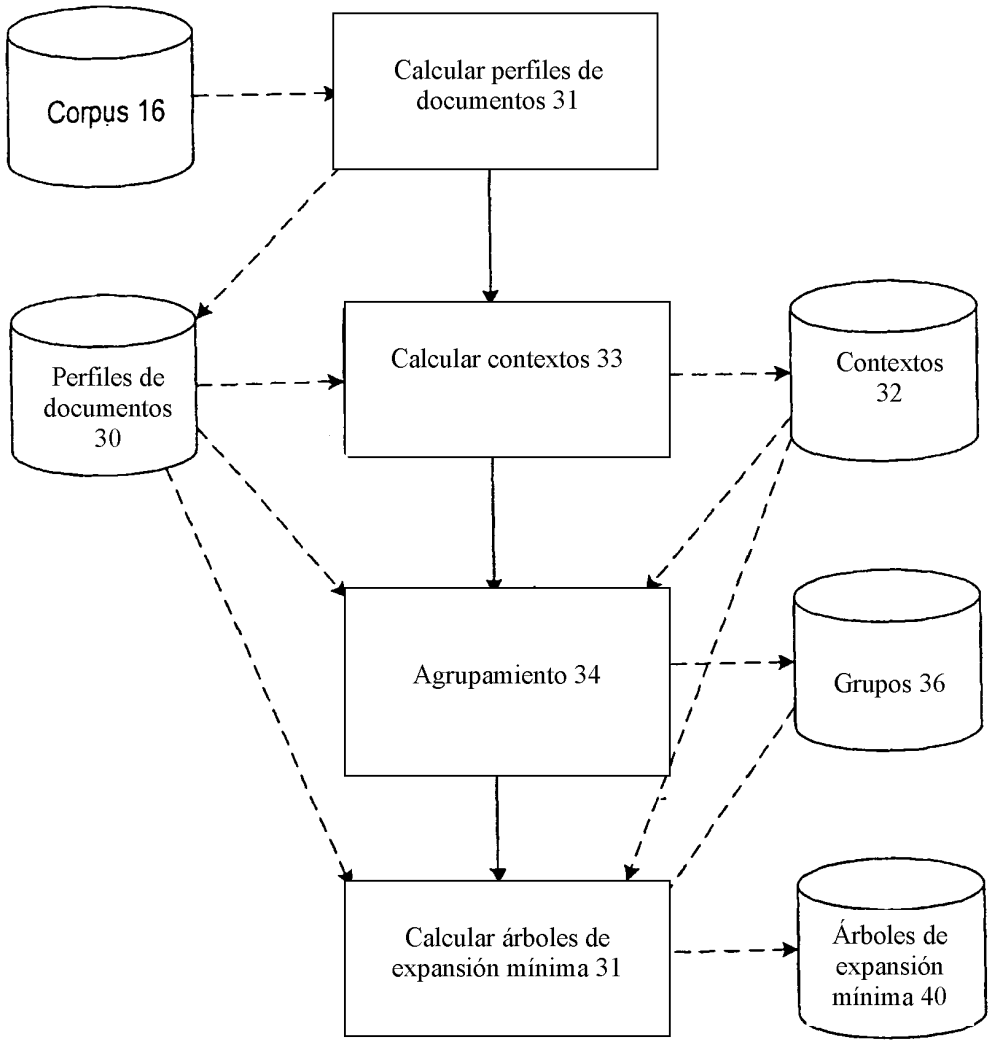
Corpus 16
Calcular perfiles de documentos 31
Perfiles de documentos 30
Calcular contextos 33
Contextos 32
Agrupamiento 34
Grupos 36
Calcular árboles de expansión mínima 31
Árboles de expansión mínima 40

Fig. 2

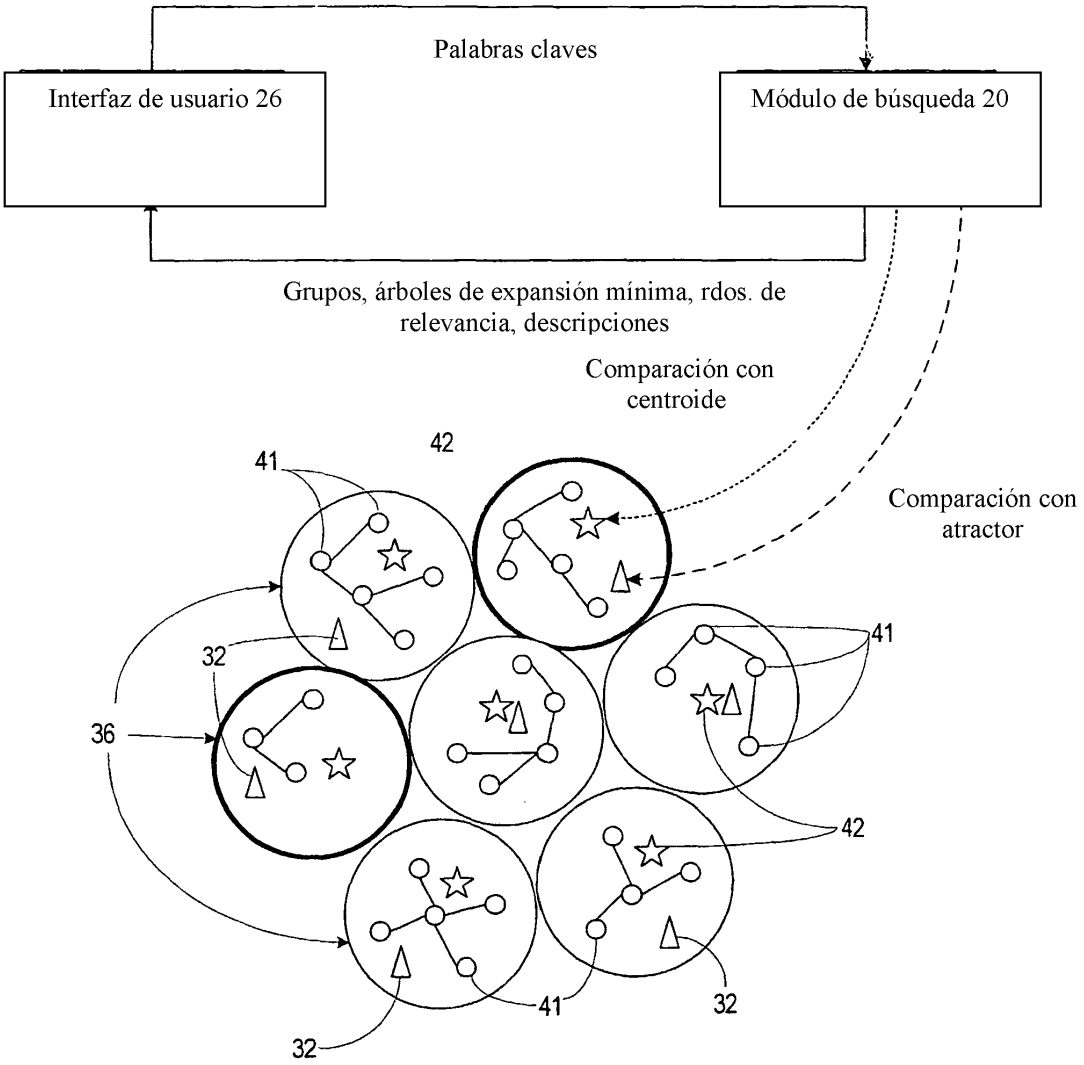
Palabras claves
Interfaz de usuario 26
Módulo de búsqueda 20
Grupos, árboles de expansión mínima, rdos. de relevancia, descripciones
Comparación con centroide
Comparación con atractor
41
42
32
36
41
42
41
32
32

Fig. 3

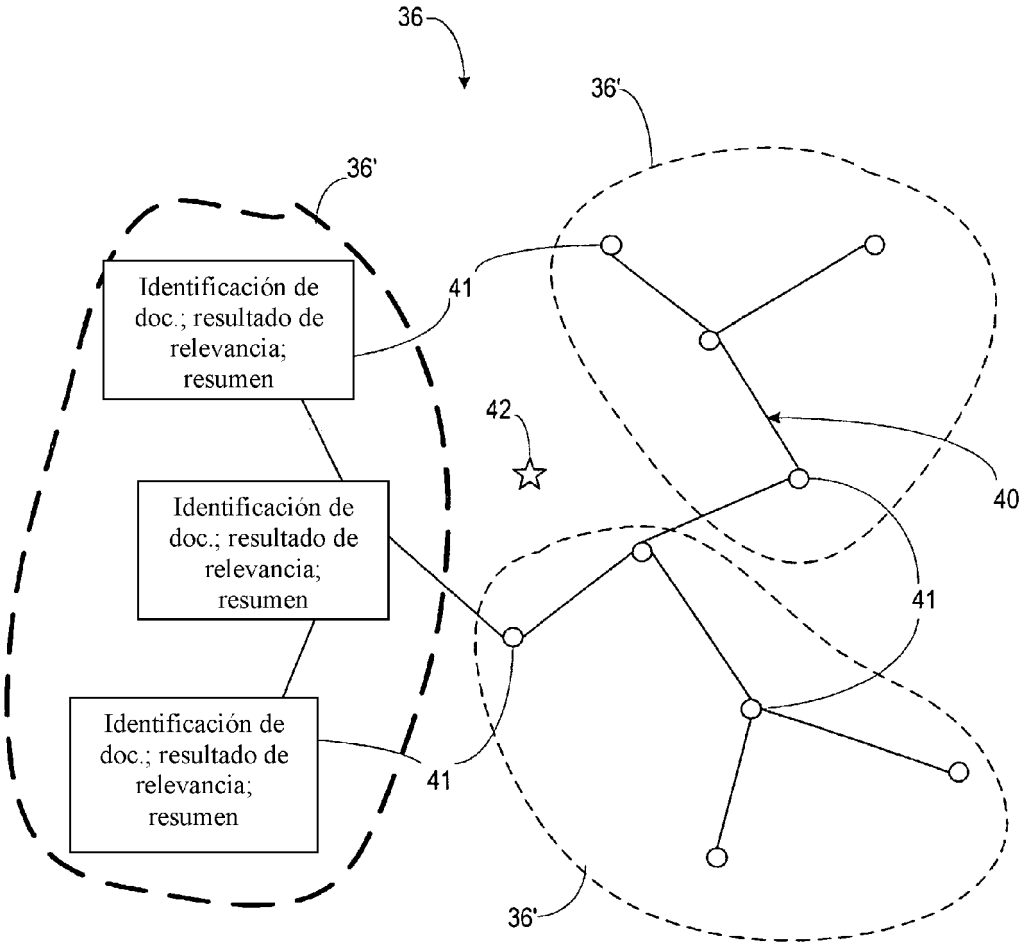
36
36'
36'
Identificación de doc.; resultado de relevancia; resumen
41
42
40
Identificación de doc.; resultado de relevancia; resumen
41
Identificación de doc.; resultado de relevancia; resumen
41
36'

Fig. 4

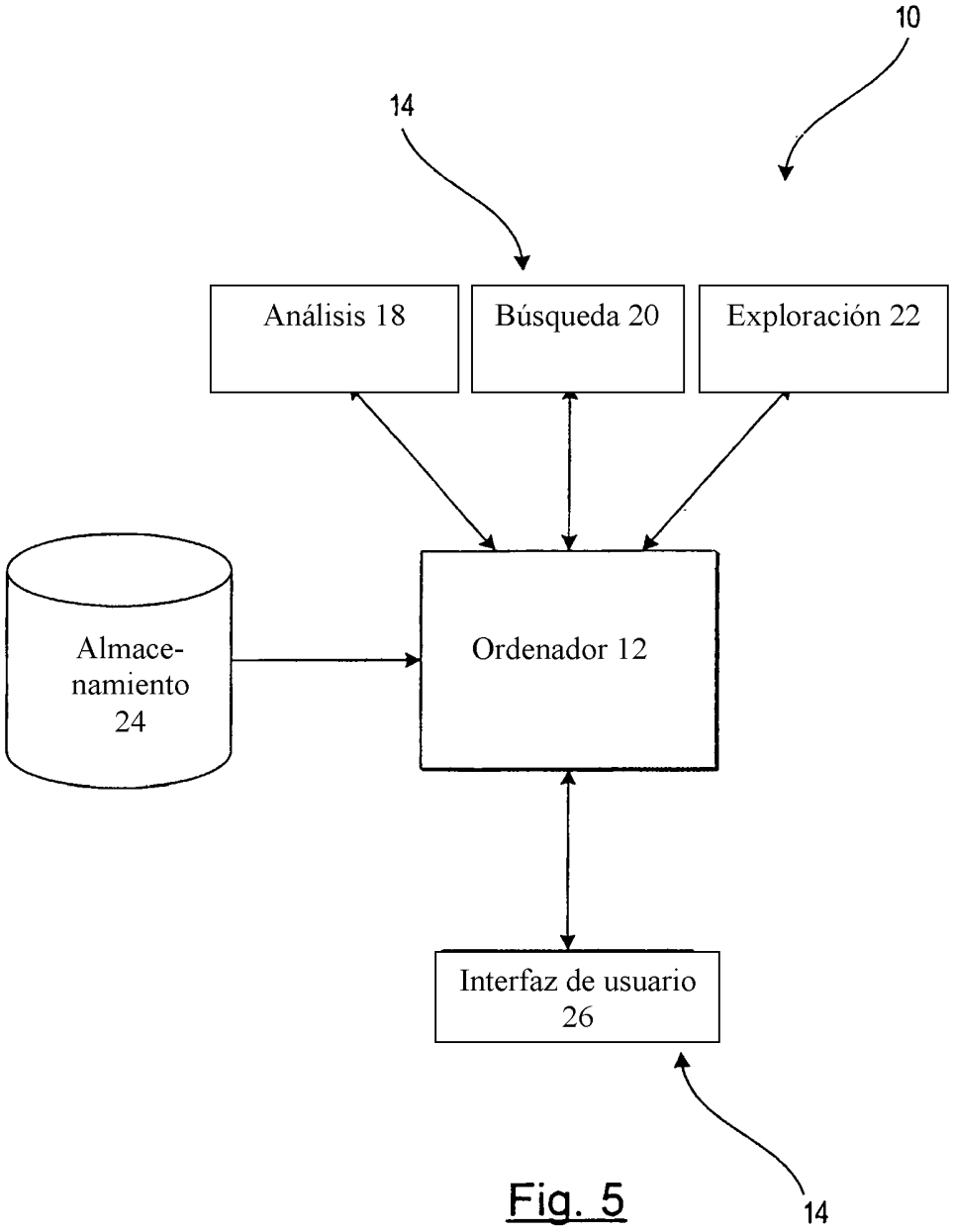
10
14
Análisis 18
Búsqueda 20
Exploración 22
Almace-
namiento
24
Ordenador 12
Interfaz de usuario
26
14

Fig. 5