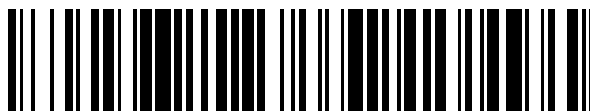


19



OFICINA ESPAÑOLA DE
PATENTES Y MARCAS

ESPAÑA



11 Número de publicación: **2 375 403**

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

12

TRADUCCIÓN DE PATENTE EUROPEA

T3

- 96 Número de solicitud europea: **01120429 .4**
96 Fecha de presentación: **27.08.2001**
97 Número de publicación de la solicitud: **1288792**
97 Fecha de publicación de la solicitud: **05.03.2003**

54 Título: **UN MÉTODO PARA LA INDEXACIÓN AUTOMÁTICA DE DOCUMENTOS.**

45 Fecha de publicación de la mención BOPI:
29.02.2012

45 Fecha de la publicación del folleto de la patente:
29.02.2012

73 Titular/es:
BDGB Enterprise Software Sàrl
Av. Du Vignoble 2 3rd Floor West
2009 Neuchâtel, CH

72 Inventor/es:
Goerke, Alexander y
Rabald, Matthias

74 Agente: **Ungría López, Javier**

ES 2 375 403 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

Un método para la indexación automática de documentos

5 Campo de la invención

La presente invención se refiere al procesamiento de documentos de texto, y en particular, a la indexación automática de documentos de texto.

10 Antecedentes de la invención

Un modo típico de estructurar grandes volúmenes de datos de modo que puedan accederse fácilmente es indexar los documentos. Esto significa que un documento o un grupo de documentos se referencia por un término de indexación. Una colección de tales términos de indexación forma entonces un índice. Esto se muestra en forma de ejemplo en la Fig. 4.

En la Fig. 4 los documentos de dirección 410, 420, 430 contienen datos de las direcciones sobre personas individuales. Los documentos pueden caracterizarse por los elementos individuales que contienen, uno de ellos es el nombre de familia de las personas a las que se refieren los documentos.

Esta información puede usarse a continuación para construir un índice 440 mostrado en la parte izquierda de la Fig. 4. Este índice contiene una lista de los nombres de la familia que están contenidos en los documentos, y cada uno de los elementos del índice se refiere a un documento individual como puede verse en la Fig. 4.

Este es un modo muy clásico y típico de organizar la información de un modo estructurado de modo que la información deseada o los documentos deseados que contienen un elemento de información buscado y deseado pueden recuperarse y accederse a partir de un gran volumen de documentos.

Los índices pueden construirse para varios elementos tales como el nombre de la familia, el nombre de pila, el nombre de la calle, etc. Lo que tienen en común los índices es que los elementos de un índice tienen todos en algún sentido el mismo "significado", tal como el "nombre de la familia", el "nombre de pila", o similares. Usando una terminología más rigurosa puede decirse que dichos índices pertenecen a una categoría común respectiva que está relacionada con un cierto significado para el ser humano. En lo siguiente, los términos "tener un cierto *significado*" y "pertenecer a una cierta *categoría*" se usan de forma sinónima.

Por lo tanto, los elementos individuales que se usan para construir un índice son de algún modo consistentes con respecto a la información que contienen cuando se ven desde un nivel más abstracto. En otras palabras, todos los elementos del índice tienen el mismo "significado".

Otro enfoque más general para el ordenamiento de documentos es caracterizar sólo uno o más documentos por un cierto término, y a continuación construir un índice a partir de los términos individuales usados de ese modo. En tal caso los elementos del índice no tienen que tener un "significado" consistente, aunque puede considerarse que tienen todos, el mismo significado, en el sentido de que cada uno de los elementos caracteriza o describe uno o más documentos a los que se refiere.

Después de que se ha construido un índice, puede usarse para preguntar y acceder al conjunto de documentos ordenados o estructurados por el índice construido de este modo. Puede introducirse directamente uno cualquiera de los términos de búsqueda, y si está contenido en el índice, a continuación se recuperan el documento o los documentos referenciados por el término del índice. Otra posibilidad es "navegar" por el índice, lo cual significa representar los elementos individuales del índice en algún orden (típicamente alfabético), como se muestra en el elemento 440 de la Fig. 4. Esto tiene la ventaja de que un usuario puede tener una visión rápida de qué elementos de índice se usan en total para organizar o "indexar" el conjunto de documentos.

Otro enfoque algo más sofisticado es usar una búsqueda llamada tolerante a fallos, lo que significa que se introduce un término de búsqueda y se recuperan los documentos en los que el valor del índice correspondiente es idéntico o al menos similar (en alguna medida, dependiendo del algoritmo de búsqueda tolerante a fallos utilizado) al término de búsqueda.

En cualquier caso, la construcción de un índice es un trabajo muy difícil y tedioso, que es el trabajo preparatorio que tiene que hacerse para hacer factible el acceso a grandes conjuntos de documentos de un modo ordenado y significativo.

Los índices típicamente se crean "manualmente", al menos en el caso de documentos a indexar que están "sin estructurar" tal como los documentos de texto simples. Si los documentos a indexar están "estructurados", tal como en el caso de las tablas de bases de datos relacionales, entonces es relativamente fácil construir un índice. Sin embargo, si no se conoce el "significado" individual que tiene un elemento en un documento no estructurado,

entonces es extremadamente difícil y cansado seleccionar los elementos que pueden usarse para indexar este documento.

El documento "Extracción de información de HTML: aplicación de un enfoque general de aprendizaje de una máquina" de FREITAG D PROCEDIMIENTOS DE LA CONFERENCIA NACIONAL DECIMOQUINTA SOBRE INTELIGENCIA ARTIFICIAL (AAAI-98). DÉCIMA CONFERENCIA EN APLICACIONES INNOVADORAS DE INTELIGENCIA ARTIFICIAL, PROCEDIMIENTOS DE LA CONFERENCIA NACIONAL DECIMOQUINTA SOBRE INTELIGENCIA ARTIFICIAL, MADISON, WI, Estados Unidos, páginas 517-523, documento XP002197239 1998, Menlo Park, CA, Estados Unidos, Prensa AAAI, Prensa MIT, Estados Unidos ISBN: 0-262-51098-7 desvela un método para realizar la extracción de información de páginas Web. La extracción de información se causa como un problema de aprendizaje de una máquina estándar y se describe la implementación de un aprendizaje de relación de propósito general para la extracción de información.

En consecuencia, es altamente deseable mejorar el proceso de indexación de documentos.

Sumario de la invención

De acuerdo con la primera realización de la presente invención se proporciona un método para recuperar ciertos documentos a partir de un conjunto de documentos en base a un término de búsqueda introducido y en base también al "significado" correspondiente del término de búsqueda, como se define en la reivindicación 1. En primer lugar se buscan los documentos que contienen el término de búsqueda. A continuación, se evalúan los documentos encontrados (llamados documentos base) para ver si en esos documentos el término de búsqueda contenido en los mismos tiene un cierto significado. Esta evaluación comprende la generación de un documento (o disposición) de texto que representa el término de búsqueda y sus elementos circundantes y su correspondiente posición absoluta o relativa con respecto al término de búsqueda o dentro del documento base. Este documento de texto se introduce a continuación en un aparato de clasificación ejercitable que se ha entrenado para juzgar si el documento de texto introducido pertenece a una cierta categoría o no. Por lo tanto el juicio tiene que determinar si el término de búsqueda tiene el cierto "significado" deseado (es decir, si el documento de texto pertenece a la categoría entrenada) o si no lo tiene (el documento no pertenece a la categoría entrenada).

En base a este juicio puede comprobarse si en los documentos base encontrados los términos buscados tienen el cierto "significado" deseado.

Los términos de búsqueda y los significados correspondientes pueden usarse para construir un índice para los documentos base. Esto puede ser a continuación de utilidad para acceder a una gran reserva de documentos base en un modo ordenado.

De acuerdo con una realización particular dichas etapas de búsqueda, clasificación e indexación se repiten para una pluralidad de términos de búsqueda. Si esto se realiza para cada uno de los documentos base, a continuación puede determinarse para cada uno de los documentos, el valor individual correspondiente que tiene un cierto "significado" y que está contenido en este documento. Por ejemplo, para todos los documentos base de una pila de cartas puede determinarse o "extraerse" la fecha de envío y a continuación los elementos extraídos de este modo junto con su "significado" "fecha de envío" podrían usarse para formar un índice que referencia el conjunto de documentos base. Si esto se realiza para varios "significados" diferentes, tales como el destinatario, la dirección de origen, el número de referencia y así sucesivamente, a continuación podría generarse también una base de datos relacional que mapea el "contenido de información" del conjunto de documentos dentro de las tablas de la base de datos relacional.

De acuerdo con una realización adicional la búsqueda realizada puede comprender una búsqueda asociativa y/o una búsqueda difusa. Esto devolverá también documentos en los cuales los términos de búsqueda están mal escritos o que contienen cualesquiera otros errores de mecanografía.

El entrenamiento del aparato de clasificación puede realizarse seleccionando para el entrenamiento un conjunto de documentos cuyos elementos en los documentos respectivos tienen un cierto "significado", a continuación se generan los documentos de texto para los respectivos documentos base respectivos de la muestra de entrenamiento, y a continuación se usan los documentos de texto generados de este modo como un conjunto de entrenamiento para entrenar el aparato de clasificación funcionando el aparato de clasificación en el modo de entrenamiento.

Usando este enfoque, un aparato de clasificación usando sólo una pequeña muestra de los documentos de entrenamiento puede realmente "aprender" a evaluar si para ciertos documentos base, algunos elementos contenidos en el mismo tienen un cierto significado o no. Este "conocimiento" ejercitado de este modo puede usarse entonces más adelante para construir un índice en base a cierto "significado" que deberían tener los elementos del índice.

El proceso de entrenamiento puede hacerse semiautomáticamente o puede hacerse en otra realización de un modo completamente automatizado. Para el entrenamiento semiautomático un usuario tiene que identificar en los

documentos base los elementos que tienen el "significado" deseado que se va a aprender o a entrenar. Después de que tiene lugar la identificación se proporciona la muestra de entrenamiento y puede realizarse el entrenamiento.

5 Para realizar un entrenamiento totalmente automático puede usarse un conjunto de documentos (tal como los documentos de una base de datos que ya existe) para el cual los elementos que ya tienen el "significado" son conocidos, por ejemplo como los campos en la base de datos. Esta pila de documentos para los cuales los elementos que tienen el "significado" deseado son ya conocidos puede usarse a continuación como muestras de entrenamiento.

10 Un conjunto de documentos base pueden indexarse automáticamente comprobando para todos los elementos (o al menos parte de ellos) contenidos en el conjunto de documentos base si esos elementos tienen un cierto "significado" o no. Con este enfoque de "bomba de arma de fuego" puede indexarse un conjunto de documentos base de forma totalmente automática, sin realizar siquiera ninguna búsqueda. En esta realización sólo se comprueban todos los elementos de los documentos base para ver si tienen un cierto significado deseado, y si es así, entonces esos
15 elementos se usan para construir el índice.

20 Sin embargo, la selección de elementos para los cuales se realiza la comprobación real, pueden contener alguna "selección inteligente". Por ejemplo, sólo los elementos que cumplen el mismo criterio de formato, podrían seleccionarse para el procedimiento de evaluación. Esto podría reducir a continuación la potencia de cálculo necesaria para realizar la indexación automática.

25 Para sacar ventaja del uno o más índices que se han construido pueden usarse para formar tablas de bases de datos relacionales. A continuación, los documentos base sin estructurar pueden convertirse automáticamente a una forma estructurada de una base de datos relacional.

Breve descripción de los dibujos

La Fig. 1 ilustra esquemáticamente la primera realización de la presente invención.

30 La Fig. 2 ilustra esquemáticamente una realización adicional de la presente invención.

La Fig. 3 ilustra esquemáticamente otra realización adicional más de la presente invención.

35 La Fig. 4 ilustra esquemáticamente la indexación de documentos.

Descripción detallada

La presente invención se explicará ahora por medio de relaciones ilustrativas en conexión con los dibujos adjuntos.

40 La Fig. 1 ilustra una primera realización de acuerdo con la presente invención. Asumamos que hay un conjunto de documentos que son no estructurados (solo documentos de texto normales) y que por consiguiente son difíciles de cuestionar para accederlos y ponerlos dentro de alguna forma ordenada.

45 Asumamos además que el usuario está interesado en los documentos de la pila de documentos base que a) contienen un cierto término, y b) en los cuales el cierto término tiene un cierto "significado", es decir en los que el cierto término pertenece a una cierta categoría que está relacionada con el significado del término para un ser humano. Como ejemplo, asumamos que un usuario está interesado en los documentos de la pila de documentos base que contienen el término "6 de Mayo de 2000", pero adicionalmente en los que el término (o elemento) "6 de Mayo de 2000" es una fecha de factura.

50 En primer lugar el usuario tiene que introducir entonces el término de búsqueda en el ordenador usando un método de entrada convencional. Además, el correspondiente "significado" que debería tener el término de búsqueda introducido también debería dirigirse de alguna forma al sistema de ordenador. El usuario podría bien introducir manualmente el "significado" correspondiente (tal como una entrada de "fecha de factura") o podría seleccionar de una diversidad de "significados" opcionales, o el "significado" correspondiente podría definirse a un valor por defecto. Después de que se ha clarificado el término de búsqueda y el "significado" correspondiente por la operación 100 en la Fig. 1, el método procede a la operación 110. En este documento la pila de documentos base se buscan los documentos que contienen el término de búsqueda introducido.

60 Esto recuperará cualesquiera documentos que contengan el elemento "6 de Mayo de 2000". Sin embargo, los documentos recuperados también pueden contener documentos en los que el término de búsqueda tiene un "significado" diferente del "significado" de la "fecha de factura", por ejemplo el "significado" podría ser una "fecha de nacimiento", una "fecha de expiración", o cualquier otra cosa.

65 Por lo tanto, en la operación 120 tiene que evaluarse si los documentos encontrados son documentos en los que el término de búsqueda tiene el significado deseado correspondiente o no.

Este procedimiento de evaluación se explicará ahora con mayor detalle en conexión con la Fig. 2. En la operación 200 se genera un documento de texto que codifica las posiciones de los elementos que rodean al término de búsqueda en el documento base encontrado. Esto se explica con mucho detalle en la Solicitud de Patente Europea 00103810.8, presentada el 23 de Febrero de 2000, por el solicitante de la presente solicitud y publicada como EP 1 128 278 A1. En particular, esta solicitud trata y describe un método que puede usarse para juzgar si ciertos elementos en un texto tienen un cierto "significado" o no. Esto se hace representando sus posiciones correspondientes a través del llamado "documento de disposición", y este documento de disposición se introduce a continuación dentro de un aparato de clasificación que se ha entrenado para reconocer si el documento de disposición introducido pertenece a cierta categoría o no.

El entrenamiento usualmente se realiza para posibilitar juzgar al aparato de clasificación si el documento de disposición representa el área de alrededor de un término de búsqueda para los documentos en los que el término de búsqueda tiene cualquier "significado" deseado o "entrenado" o no. Para detalles, se hace referencia a la Solicitud de Patente Europea anteriormente mencionada, es decir la EP 1 128 278 A1.

Este documento de disposición (o como lo llamamos en este punto, documento de texto) se usa a continuación como una entrada al aparato de clasificación como se ilustra en la operación 210 de la Fig. 2 de la presente solicitud. Este aparato de clasificación juzga a continuación si el documento de texto introducido pertenece a cierta "categoría" o no. La "categoría" significa en este punto si el término de búsqueda tiene un cierto "significado" o si no lo tiene (en este caso, si es una fecha de factura o no).

En la operación 220 el aparato de clasificación realiza a continuación su juicio de si el término de búsqueda tiene el "significado" introducido (o deseado) o no. Un aparato de clasificación particularmente adaptado se describe en la Solicitud de Patente Europea 99108354.4, que se presentó por el solicitante de la presente solicitud el 28 de Abril de 1999 y se publicó como EP 1 049 030 A1. En particular cualesquiera detalles relativos al aparato de clasificación descritos o mencionados en este documento pueden tomarse de esta solicitud.

Sin embargo, también puede usarse cualquier otro aparato de clasificación que es capaz de evaluar si los documentos de texto pertenecen a una cierta categoría o no y también que se puede entrenar. Por supuesto, en primer lugar el aparato de clasificación tiene que entrenarse usando cierta muestra de entrenamiento, y después de esto, los documentos base encontrados de nuevo pueden introducirse en el aparato de clasificación y evaluarse por lo tanto, para ver si el término de búsqueda tiene un cierto "significado" o no.

Debería observarse de nuevo que una descripción detallada del proceso de extracción de elementos a partir de documentos de texto que tienen un "significado" deseado se describe en el documento EP 1 128 278 A1 mencionado anteriormente. También se describen en este documento las particularidades concernientes a cómo se representa el área de alrededor de un elemento candidato por medio de un documento de disposición que se suministra a continuación a un aparato de clasificación ejercitable para juzgar si el elemento candidato realmente tiene el "significado" deseado.

Una realización adicional de la presente invención que se refiere a la generación automática de cualquier índice se explica ahora en conexión con la Fig. 3. En la operación 300 el cierto "significado" que los elementos de los documentos base debería tener se definen o seleccionan, como ya se ha explicado anteriormente.

A continuación, como ya se ha explicado también, para los elementos a comprobar, se genera el documento de texto (o documento de disposición).

Este documento de texto puede introducirse a continuación a un aparato de clasificación ya entrenado en la operación 320. Este aparato de clasificación en la operación 330 comprueba a continuación si los elementos a comprobar tienen el cierto "significado" definido o seleccionado.

Si la evaluación da como resultado un "si", a continuación en la operación 340 se usa el elemento evaluado de este modo para construir un índice junto con el significado definido o seleccionado.

Si la respuesta en la operación 330 es un "no", a continuación en la operación 350 los elementos comprobados se descartan o, en otras palabras no se usan para construir el índice.

Usando el proceso ilustrado en la Fig. 3, puede realizarse una indexación automática de documentos.

Esto puede hacerse de varias formas. Los elementos a comprobar podrían ser todos los elementos que están contenidos en los documentos base individuales. Para cada uno de los elementos individuales, podría comprobarse a continuación, por ejemplo, si este elemento es una "fecha de factura" o no, por ejemplo generando un documento de disposición correspondiente, y a continuación evaluando este documento de disposición a través de un aparato de clasificación.

Es fácilmente evidente, que este enfoque es más bien consumidor de tiempo y potencia de cálculo, por lo tanto, la selección de los elementos a comprobar podría hacerse de un modo algo inteligente. Por ejemplo, podrían comprobarse sólo los elementos que coinciden con un cierto criterio de formato. Tal criterio de orden puede ser una compilación de los formatos bien conocidos que puede asumir una fecha, y comprobarse los elementos para ver si coinciden con el criterio de formato y sólo se realiza la comprobación final en el caso de cartas si el elemento seleccionado de este modo (que puede asumirse que es una "fecha") realmente es una "fecha de factura" o no.

Otro enfoque podría generar automáticamente los términos de búsqueda que a continuación se buscan en los documentos base.

Por ejemplo, podría seleccionarse un cierto intervalo de fechas (por ejemplo, desde el 1 de Enero de 2000 al 1 de Enero de 2001), y a continuación se buscan entonces todas las fechas que se encierran entre las mismas, en otras palabras, se aplican como términos de búsqueda para la pila de documentos base.

Para los términos de búsqueda en los que se encontrará una coincidencia, lo que significa que se encuentra un documento que contiene un término de búsqueda correspondiente, el documento de disposición se genera a continuación y se comprueba si el término de búsqueda tiene el cierto "significado" (en este caso: fecha de factura). Si es así, a continuación como ya se ha explicado anteriormente el término de búsqueda puede usarse para construir un índice.

Con este enfoque, no todos los elementos en los documentos base tiene que comprobarse, sin más bien sólo los elementos para los cuales la búsqueda ha suministrado una coincidencia.

El enfoque que es realmente más eficaz en términos de la potencia de cálculo necesaria puede depender de circunstancias individuales.

Debería estar claro que los ejemplos anteriores sólo son ejemplos ilustrativos, y que por supuesto el método explicado hasta ahora también puede aplicarse a cualesquiera otros "significados" que debería tener cierto término o elemento tal como por ejemplo un nombre, un nombre de calle, un nombre de familia, un nombre de pila, un nombre de ciudad, o cualquier otro.

El significado más general que podría tener cualquier elemento con respecto a cierto documento en el que está contenido podría ser que el elemento que "describe" justo de alguna manera el documento en el que está contenido. Por ejemplo, el término "factura" podría estar contenido en una factura, además, justo describe que el documento correspondiente realmente es una factura. Tales elementos que describen el contenido completo de un documento por supuesto puede usarse para construir un índice.

En una realización adicional el entrenamiento del aparato de clasificación se realiza de forma semiautomática. Por ejemplo un usuario identifica para cada uno de un conjunto de documentos los elementos que tienen el cierto "significado" deseado tal como por ejemplo, el número de factura. En base a esta selección introducida por el usuario se genera a continuación el documento de disposición que puede usarse como entrada de entrenamiento para el entrenamiento del aparato de clasificación. De este modo varios documentos de disposición que representan áreas de alrededor de los elementos que son números de facturas se introducen al aparato de clasificación y por lo tanto el aparato de clasificación se entrena ("aprende") para reconocer los elementos que son números de factura y a distinguirlos de los otros elementos que no tienen este significado.

En una realización adicional tal procedimiento de entrenamiento puede ejecutarse en un modo totalmente automático. Asumamos que hay un conjunto de documentos (por ejemplo desde cualquiera de las bases de datos ya existentes) que ya están almacenados, evaluados y clasificados en la forma de una base de datos relacional que tiene una columna "número de factura" en el cual para cada uno de los documentos ya está almacenado el número de factura. Entonces no hay necesidad de identificar manualmente el número de factura mediante el usuario sino que más bien la información desde la base de datos puede usarse para generar la entrada de entrenamiento. La base de datos proporciona el número de factura para cada uno de los documentos, este número puede buscarse a continuación en el propio documento y después de que se ha localizado en base al área de alrededor que rodea al número de factura, se genera el documento de disposición. Este proceso se puede realizar para cada uno de los documentos de la base de datos y a continuación da como resultado un gran conjunto de documentos de disposición que pueden usarse como ejemplos de entrenamiento para entrenar al aparato de clasificación. Por lo tanto puede realizarse un entrenamiento totalmente automático sin necesidad de identificar elementos por el usuario.

Debería estar claro para cualquier especialista en la técnica que este procedimiento puede realizarse no sólo para "números de factura" sino para cualesquiera otros elementos que tienen un cierto "significado" y que son ya conocidos. Este conocimiento acerca de los propios elementos y su "significado" no necesita venir de una base de datos, también puede estar almacenada por ejemplo en una tabla de una hoja de cálculo, o puede incluso estar escrita en papel en la forma de una tabla que a continuación puede escanearse. No importa desde donde viene el conocimiento acerca de los propios elementos y su "significado", siempre que exista y pueda utilizarse para automatizar el proceso de entrenamiento del aparato de clasificación en base a este conocimiento.

Aunque la presente invención se ha descrito anteriormente en este documento por medio de realizaciones de ejemplo, será fácilmente evidente para el lector experto que pueden realizarse modificaciones sin apartarse del alcance de la invención como se define en las reivindicaciones adjuntas.

5

10

REIVINDICACIONES

1. Un método para recuperar en base a un término de búsqueda junto con una cierta categoría a la que pertenece el término de búsqueda, a partir de un conjunto de documentos base, los documentos base que contienen dicho término de búsqueda y en los que dicho cierto término de búsqueda pertenece a dicha cierta categoría para posibilitar la construcción de un índice sobre dichos documentos recuperados, comprendiendo dicho método:
 - buscar (110) en los documentos base de entre dicho conjunto de documentos base los que contienen dicho cierto término de búsqueda;
 - evaluar (120) los documentos base encontrados para ver si dicho término de búsqueda contenido en dichos documentos base encontrados, pertenecen, respectivamente a dicha cierta categoría o no, comprendiendo dicha evaluación:
 - generar (200) un documento de texto para representar los elementos en el documento base encontrado alrededor del término de búsqueda y los elementos correspondientes a la posición absoluta o relativa con respecto a dicho término de búsqueda, codificando los elementos de dicho documento de texto, dichas posiciones absolutas o relativas de dichos elementos de alrededor.
 - introducir (210) dicho documento de texto en un aparato de clasificación ejercitable, que se ha entrenado para reconocer si dicho término de búsqueda en un documento de texto introducido, pertenece a cierta categoría o no, mientras que dicho entrenamiento se ha realizado en base a una muestra de entrenamiento de documentos de texto que se han generado para documentos de texto en los que el término de búsqueda rodeado por los elementos de alrededor pertenecen a dicha cierta categoría; y
 - clasificar dicho documento de texto introducido para juzgar (220) si dicho término de búsqueda pertenece a dicha cierta categoría introducida o no, e
 - indexar una pluralidad de dichos documentos base usando dicho término de búsqueda y dicha categoría correspondiente para construir un índice para dichos documentos base.
2. El método de la reivindicación 1, en el que los elementos de dicho documento de texto son codificaciones de dichas posiciones absolutas o relativas de dichos elementos de alrededor por las correspondientes cadenas de caracteres de texto.
3. El método de la reivindicación 1, que comprende además:
 - repetir dichas etapas de búsqueda, entrada y clasificación e indexación para una pluralidad de términos de búsqueda, generándose dichos términos de búsqueda automáticamente de acuerdo con una norma predefinida.
4. El método de una de las reivindicaciones de 1 a 3, en el que dicha búsqueda de dichos documentos base comprende una búsqueda asociativa; y/o una búsqueda difusa en base a dicho término de búsqueda.
5. El método de la reivindicación 1, que comprende además:
 - comprobar todos los términos de búsqueda de dichos documentos base, o
 - comprobar sólo los términos de búsqueda de dichos documentos base que coinciden con uno o más criterios predefinidos.
6. El método de reivindicaciones 1 ó 5, que comprende además:
 - generar una base de datos relacional a partir del término de búsqueda del índice y su categoría correspondiente.
7. El método de acuerdo con una de las reivindicaciones anteriores de 1 a 6, que comprende el entrenamiento del aparato de clasificación, comprendiendo dicho entrenamiento:
 - a) buscar los documentos base en los que el término de búsqueda pertenece a cierta categoría;
 - b) seleccionar dicho término de búsqueda por el usuario;
 - repetir las operaciones a) y b) hasta que se ha seleccionado el conjunto suficiente de documentos base para generar una muestra de entrenamiento;
 - generar los documentos de texto para los documentos base respectivos;
 - usar dichos documentos de texto generados como un conjunto de entrenamiento para entrenar dicho aparato de clasificación corriendo dicho aparato de clasificación en el modo de entrenamiento.
8. El método de acuerdo con una de las reivindicaciones anteriores de 1 a 6, que comprende el entrenamiento del aparato de clasificación, comprendiendo dicho entrenamiento:

- usar un conjunto de documentos base para los cuales los términos de búsqueda que pertenecen a la categoría predefinida deseada son ya conocidos para generar automáticamente documentos de texto como muestras de entrenamiento para entrenar dicho aparato de clasificación.

- 5 9. Un programa de ordenador que comprende un código de programa de ordenador para posibilitar que un ordenador realice un método de acuerdo con una de las reivindicaciones anteriores de 1 a 8 cuando dicho programa se corre sobre un ordenador.

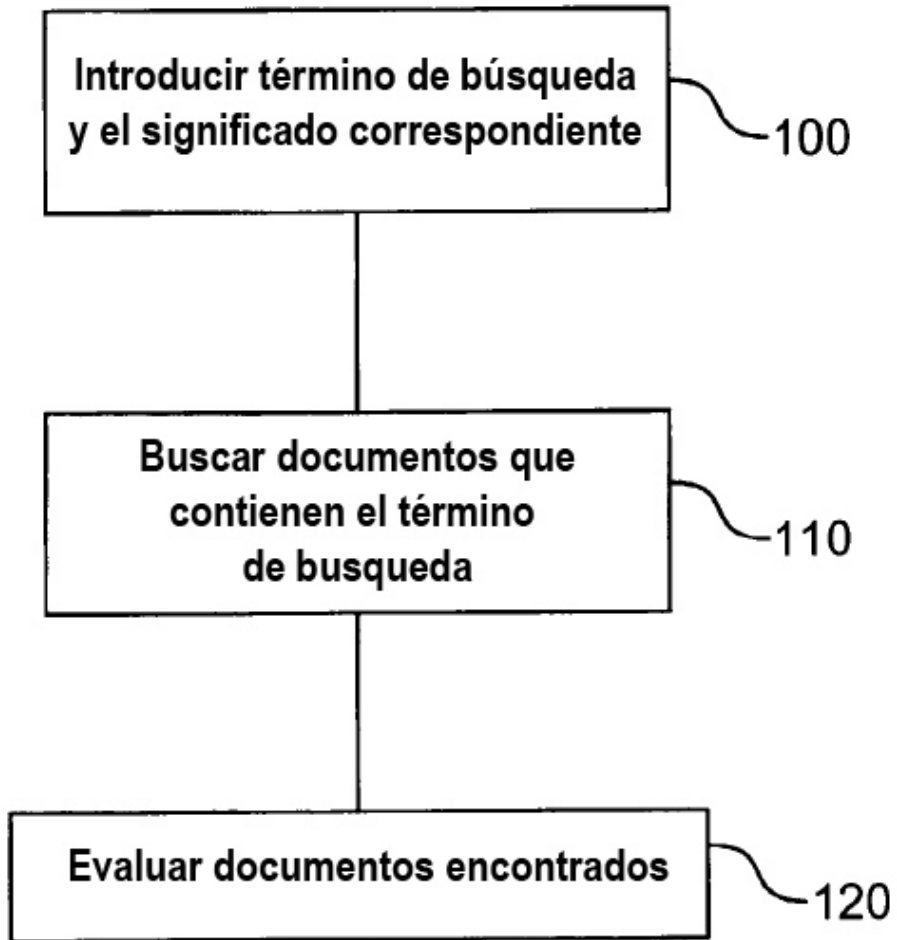


Fig. 1

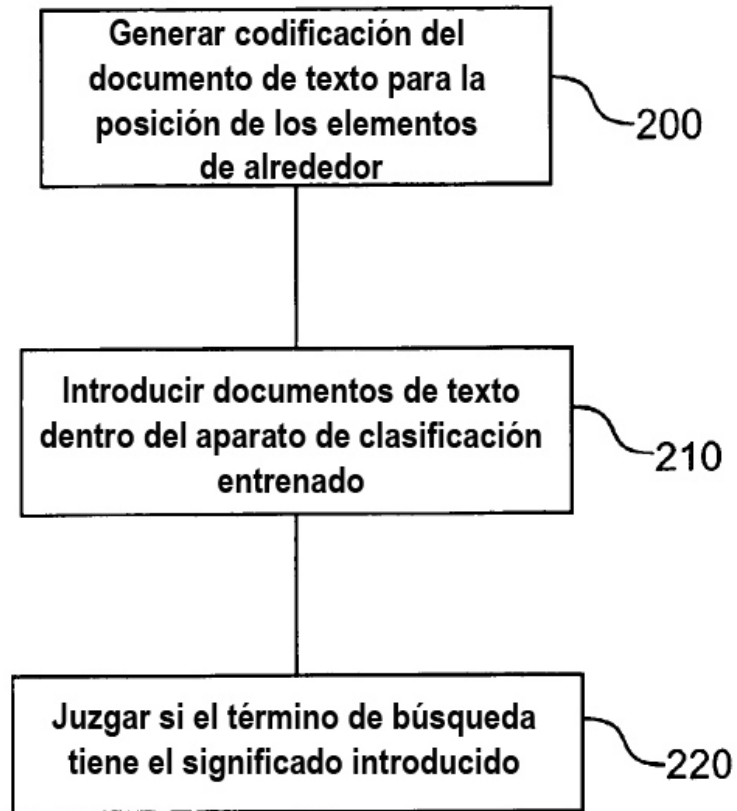


Fig. 2

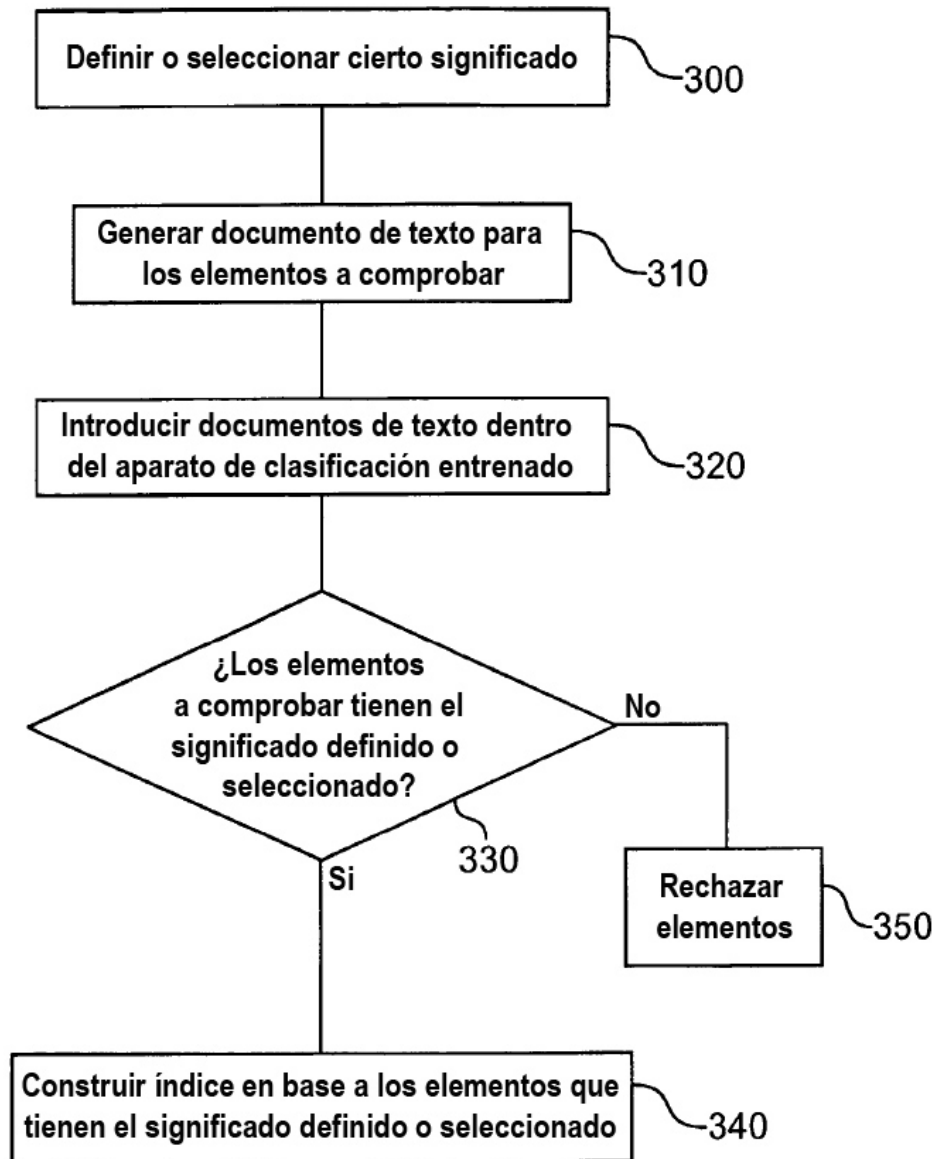


Fig. 3

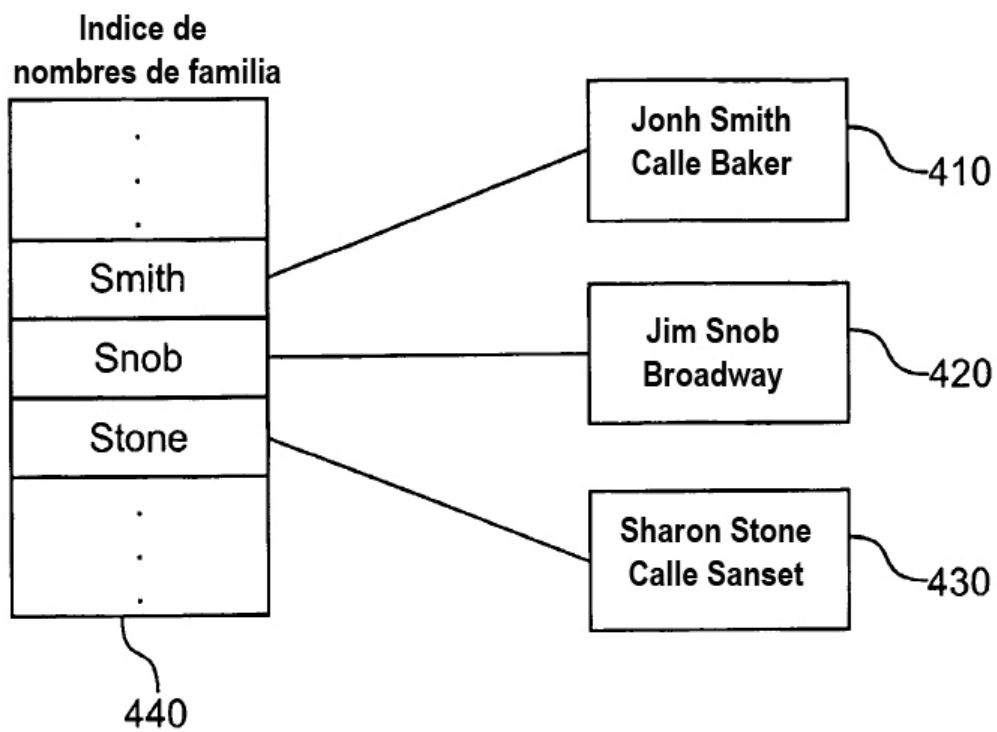


Fig. 4