(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 391 261**

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

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Número de solicitud europea: **01127768 .8**
(96) Fecha de presentación: **21.11.2001**
(97) Número de publicación de la solicitud: **1315096**
(97) Fecha de publicación de la solicitud: **28.05.2003**

(54) Título: **Método y aparato para recuperar información importante**

(45) Fecha de publicación de la mención BOPI:
22.11.2012

(45) Fecha de la publicación del folleto de la patente:
22.11.2012

(73) Titular/es:
VOEGELI, WERNER (100.0%)
QUELENSTRASSE 62
5330 ZURZACH, CH

(72) Inventor/es:
VOEGELI, WERNER

(74) Agente/Representante:
DE ELZABURU MÁRQUEZ, Alberto

DESCRIPCIÓN

Método y aparato para recuperar información importante.

El presente invento se refiere a un método para la recuperación de un conjunto de documentos electrónicos los que tienen un contenido relacionado con un documento electrónico usado como documento de entrada.

En la actualidad las personas se enfrentan a una cantidad siempre creciente de información. La cantidad de información producida en un solo día es tan grande que a una persona, incluso leyendo las 24 horas del día durante toda su vida, le sería imposible leer lo producido en un solo día.

Está claro que la información se produce a un ritmo tan rápido que de algún modo ha de ser tratada electrónicamente. El modo más convencional de tratar, almacenar, archivar y recuperar grandes cantidades de información son las denominadas bases de datos.

Las bases de datos almacenan fragmentos de información individuales ordenados de una manera denominada “indexación”.

La indexación significa que cada fragmento de información está etiquetada con al menos una etiqueta, siendo un tipo de “meta” información que describe algunos aspectos del contenido del fragmento de información que realmente ha sido almacenado.

Un ejemplo muy simple es un registro de dirección que puede comprender el apellido, el nombre dado, el nombre de la calle, código postal, y el nombre de una ciudad. Un campo de índice podría entonces, por ejemplo, ser el apellido que describe un aspecto del registro total de la dirección.

Suponiendo que se hayan almacenado muchas direcciones, entonces cada una tiene una etiqueta que contiene un apellido, y esas etiquetas podrían entonces ser ordenadas automáticamente y podrían ser usadas para acceder de una forma rápida a cada uno de los registros de direcciones, como es bien conocido en la técnica de la tecnología de las bases de datos.

Este método es muy rápido para acceder a cada uno de los registros de direcciones (que se denominan conjuntos de datos en la terminología de la base de datos), aunque tiene la desventaja de que solamente pueden ser buscados para su recuperación los fragmentos o tipos de información de los que se ha creado un índice. Esto significa que la información de acceso almacenada mediante índices no necesariamente dan al usuario una imagen total de todos los aspectos de la información que ha sido almacenada, debido a que solamente se abren para un usuario aquellos aspectos para los cuales ha sido creado un índice para ser usado como criterio de búsqueda.

Un enfoque completamente diferente es el almacenamiento y el acceso a información simulando una memoria denominada asociativa. De acuerdo con este enfoque cada documento de un conjunto de documentos es representado mediante un denominado “vector de características” el cual, de hecho, es una secuencia de bits. Cada bit de la secuencia de bits solicita la presencia o la ausencia de una cierta característica en el documento que representa.

Las características pueden, por ejemplo, ser la presencia o ausencia de ciertos unigramas, digramas, trigramas, o similares.

Cada documento almacenado es después representado mediante una secuencia de bits en la que cada bit individual indica si una cierta característica está presente o ausente en su correspondiente documento. El acceso a y la recuperación de los documentos así almacenados es después realizada mediante la introducción de un documento de consulta. Este documento de consulta después es también representado mediante una secuencia de bits que es generada de acuerdo con las mismas reglas que las secuencias de bits que representan el documento ya almacenado, es decir para generar un vector de características de consulta.

Después de haber realizado esto se realiza una operación lógica en términos de bits entre la secuencia de bits de consulta y las secuencias de bits que representan los documentos ya almacenados, por ejemplo una lógica AND en términos de bits entre la consulta y cada uno de los vectores de documentos almacenados. Basado en esta operación en términos de bits se ha formado algún tipo de medida de similitud entre la secuencia de bits de consulta y la secuencia de bits que representa uno respectivo de los documentos que forman la base de datos. Las características que están presentes en la consulta y en el documento almacenado dan un “1” lógico como resultado de la operación AND, cuantas más características están presentes en la consulta, y en el documento almacenado resultan más “1” lógicos AND de la operación. Como ejemplo, el número de “1” lógicos puede ser contemplado como una medida de similitud entre una secuencia de bits de consulta y la secuencia de bits (vector de características) de un documento almacenado. Las medidas de similitud (individuales) obtenidas de este modo para los diferentes documentos del conjunto pueden ser consideradas como componentes del vector de un vector de medida de la similitud.

Basado en el vector de medida de similitud que indica la similitud entre el documento de consulta y los documentos individuales almacenados en la base de datos se puede recuperar el documento más similar de la base de datos, que con respecto a las características codificadas en la secuencia de bits está más próximo al documento de consulta.

Con tal acceso asociativo resulta posible introducir cualquier consulta arbitraria en una base de datos y encontrar aquellos documentos que son más similares al documento de la consulta de entrada. Una de las posibilidades de tal búsqueda asociativa es que los documentos que son considerados como “similares” basada en una evaluación de los vectores de características realmente son similares en su contenido a la consulta de entrada. Esto significa que tal búsqueda asociativa puede realizar correctamente un acceso “asociativo” verdadero que es muy similar al que realiza el cerebro humano.

Por ejemplo, si uno lee un cierto libro sobre un tema determinado entonces un ser humano tiene automáticamente muchas asociaciones, a su mente vienen libros relacionados con los mismos o similares temas, sucesos y experiencias unidos a estos temas. Suponiendo que están todos representados mediante vectores de características en una base de datos, entonces uno podría realizar una consulta, basada por ejemplo en un libro que el usuario está leyendo en ese momento, después tal búsqueda asociativa electrónica sería capaz de producir unas asociaciones similares a las del cerebro humano, es decir se podrían recuperar libros y documentos similares, y –caso de estar almacenados y codificados- experiencias o sucesos.

Tal método de búsqueda asociativa o “Método de acceso asociativo (ASSA)” es por ejemplo conocido por Berkovich, S., El-Qawasmeh, E., Lapir, G.: “Organización de coincidencia próxima en una matriz de atributos aplicada a métodos de acceso asociativos en la recuperación de información”, Actas de la 16ª Conferencia Internacional de Informática Aplicada, 23. – 25.02.1998, Garmisch Partenkirchen, Alemania.

Correspondiendo, pero una exposición en algún modo más detallada, se conoce también de Lapir, G. M.: “Uso de un método de acceso asociativo para sistemas de recuperación de información” Actas de la 23ª Conferencia de Pittsburgh sobre modelización y simulación, 1992, páginas 951-958, XP009099840.

En consecuencia, se conoce un método para la recuperación de un conjunto de documentos electrónicos aquellos documentos que son próximos en contenido de un documento electrónico usado como un documento de entrada, el cual comprende:

- proporcionar un conjunto de documentos;
- generar unos vectores de característica de secuencias de bits para los documentos de dicho conjunto, en donde cada vector de características representa uno respectivo de dichos documentos e indica con cada uno de sus componentes vectoriales binarios la presencia o ausencia de una cierta característica dentro del respectivo documento;
- formar una matriz de atributos binaria a partir de los vectores de características de la secuencia de bits, constando la matriz de atributos binaria de bits individuales y representando cada bit un cierto atributo de un cierto documento mediante la indicación de la presencia o ausencia de una cierta característica dentro del respectivo documento;
- proporcionar un documento de entrada;
- generar un vector de características de la secuencia de bits de consulta que representa el documento de entrada y que indica con cada uno de sus componentes vectoriales binarios la presencia o ausencia de una cierta característica dentro del documento de entrada;
- efectuar una búsqueda asociativa de dicho documento de entrada usando dicha matriz de atributos binaria y el vector de características de la secuencia de bits de consulta que representa el documento de entrada, en el que la búsqueda asociativa de dicho documento de entrada se efectúa determinando para dicho documento de entrada y cada documento de dicho conjunto una medida de similitud individual entre dicho documento de entrada y el respectivo documento de dicho conjunto, en el que la respectiva medida de similitud individual se determina efectuando una operación lógica en términos de bits entre el vector de características de la secuencia de bits de consulta que representa el documento de entrada y la secuencia de bits que representa el respectivo documento en dicha matriz de atributos binarias, en el que dichas medidas de similitud individuales, estando cada una determinada por el documento introducido y un documento respectivo de dicho conjunto y que representa una similitud entre dicho documento de entrada y dicho documento respectivo de dicho conjunto son componentes vectoriales de un vector de medida de similitud asociados con el documento introducido y con todos los documentos de dicho conjunto de documentos;

- juzgar, sobre la base de dichas medidas de similitud individuales incluidas como componentes vectoriales en dicho vector de medida de similitud qué documento de dicho conjunto puede ser considerado como que tiene un contenido próximo al de dicho documento de entrada.

Un método posterior para la recuperación a partir de un conjunto de documentos electrónicos aquéllos que tienen un contenido próximo al del documento electrónico usado como documento introducido es conocido a partir del artículo BO-REN BAI y otros: “Recuperación de documentos de texto/voz silábicos chinos usando consultas de recuperación de documentos de texto/voz” Diario internacional de reconocimiento de patrones y de inteligencia artificial, Agosto 2000, World Scientific, Singapur, volumen 14, nº 5, páginas 603-616.

De acuerdo con este enfoque se usan vectores de características no binarios que contienen la información de presencia, recuentos de frecuencia y registros acústicos de sílabas y pares de sílabas contiguas en un reticulado de sílabas. Tales vectores de características generados son almacenados en una base de datos de vectores de características. Mediante una consulta de texto o una consulta de voz se proporciona un documento de entrada en un subsistema de recuperación en línea directa. Se genera un correspondiente vector de características para el documento introducido sobre la base del mismo reticulado de sílabas. Sobre la base de la base de datos de vectores de características y del vector de características un módulo de recuperación efectúa una búsqueda asociativa dentro de la base de datos de vectores de características evaluando las medidas de similitud entre el vector de características y todos los vectores de características del documento de la base de datos. Se selecciona un conjunto de documentos con las medidas de similitud más altas como salida de la recuperación.

Un método para la recuperación de datos a partir de un conjunto de documentos electrónicos sobre la base del “vector de consulta” en la forma de una lista de identificadores de elementos o de identificadores de atributos y sus correspondientes valores, posiblemente con relación al contenido de un documento electrónico usado como un documento de entrada, es conocido a partir del documento US 5.778.362. Una matriz de datos bidimensional o “mapa” que es puesta en práctica como una “tabla asociativa” puede ser dispuesta como una matriz documentos-términos, y en consecuencia puede estar basada en un conjunto de documentos que se proporciona. Una entrada a la celda respectiva en la matriz da la frecuencia con la que se produce un término correspondiente en los respectivos documentos. Por lo tanto, las filas de la matriz corresponden a los vectores de características, los cuales son generados para los documentos del conjunto.

El documento US 5.778.362 expone además el proceso de efectuar una búsqueda asociativa dentro de la base de datos de vectores de características basada en un vector de características de consulta. Dicho vector de características de consulta podría estar basado en un documento electrónico usado como documento de entrada, por ejemplo un mensaje de correo electrónico o un documento para ser comparado con otros documentos para identificar otras similitudes.

El documento EP 0.750.266 A1 expone el uso de vectores distintivos (vectores de características) de documentos y de documentos de entrada. Usando identificadores conceptuales independientes de un lenguaje usado se generan los vectores de características conceptuales que se relacionan con identificadores conceptuales en lugar de identificadores basados en el lenguaje o en la fonética.

En el documento US 5.918.223El se expone el uso de vectores de características para clasificar y graduar la similitud entre ficheros de audio individuales. Los vectores de características se refieren a características que describen las características del sonido de los respectivos ficheros de audio respectivos tales como tono, brillantez y ritmo.

A partir del documento EP 1.049.030 A1El también es conocido el uso de vectores de características que representan las características de los respectivos documentos. Los componentes vectoriales de un vector que representa un cierto documento corresponden a la frecuencia con que ocurren los términos en dicho documento. Un esquema de clasificación se refiere a la separación entre los respectivos vectores de características en un espacio vectorial formado por las n dimensiones de los vectores de características, con subespacios en este espacio vectorial que corresponden a una cierta clase de documentos.

Con referencia otra vez a dos primeras citas de la técnica anterior puede existir un problema en el que la resolución de la búsqueda asociativa conocida convencionalmente esté limitada, ya que las características a las que el vector de características de consulta y los vectores de características del documento se refieren dan una información no útil pero solamente una gran cantidad de “ruido”. Éste puede ser en particular el caso en el que la dimensión de los vectores de características es relativamente alta y los documentos del conjunto así como el documento de entrada contienen muchos diferentes aspectos y cantidades de texto. Por ejemplo, cuando estos documentos, en particular el documento de entrada, son muy largos, muchas de las características en las que está basada la codificación binaria de características pueden a menudo ser encontradas en algún lugar en el texto, de modo que la fijación del respectivo bit en el vector de características respectivo no será una buena base para la búsqueda asociativa a fin de encontrar unos documentos similares.

Para aumentar la resolución de la búsqueda asociativa o para evitar un deterioro de la resolución de la búsqueda asociativa el invento proporciona el método definido en la reivindicación 1.

De acuerdo con el invento, el documento de consulta de entrada no solamente es dividido en fragmentos sino que cada uno de dichos fragmentos es después desplazado un determinado número de unidades basado en las que está compuesto el documento de consulta. Este uso de una denominada ventana de deslizamiento genera a continuación unos fragmentos adicionales que pueden ser usados como entradas para la búsqueda asociativa. Estas entradas adicionales dan nuevamente unas medidas de similitud adicionales que también son tenidas en cuenta al combinarlas para una medida final de similitud.

El invento está basado en el entendimiento de que tales fragmentos del documento obtenidos sobre la base del documento introducido están generalmente más bien enfocados con respecto al tema o aspecto tratado, y en consecuencia enfocados con respecto a las características representadas por los respectivos componentes vectoriales de los vectores de características de consulta generados de los fragmentos.

Por lo tanto, las medidas de similitud obtenidas para los fragmentos del documento introducido con respecto a todos los documentos del conjunto permitirán un mejor juicio pudiendo ser considerados los documentos del conjunto como que tienen un contenido próximo al del documento introducido, de modo que las medidas finales de similitud individuales incluidas como componentes vectoriales en las medidas finales de similitud incluidas como componentes vectoriales en el vector de medida de similitud pueden tener una mejor resolución.

Además, se pueden obtener unas ventajas y mejoras mediante las enseñanzas dadas en las subreivindicaciones.

De acuerdo con una realización particular, la consulta y los documentos del conjunto de documentos almacenados representan textos hablados por una voz humana.

De acuerdo con una realización particular adicional, la búsqueda asociativa puede ser usada para clasificar el documento de consulta de entrada como perteneciente a uno de una pluralidad de clases predefinidas. Con tal fin, los documentos del conjunto son respectivamente clasificados como pertenecientes a una de una pluralidad de clases, a continuación se realiza la búsqueda asociativa cuando se cuentan los aciertos en las clases individuales. La clase de la que se ha recuperado la mayor parte de los aciertos es después la clase a la que se asigna el documento de consulta de entrada.

De acuerdo con una realización particular adicional se usa un método de autooptimización para optimizar los parámetros de la búsqueda asociativa. Para tal fin se ha usado una base de datos ideal con dos o más clases que no contienen clasificaciones equivocadas. El método de clasificación de acuerdo con una realización del invento se usa después para tratar de reproducir esta clasificación ideal mediante la clasificación automática de todos los documentos contenidos en la base de datos. Este proceso se realiza repetidamente a la vez que se varían sistemáticamente los parámetros del proceso. Los métodos que no dan buenos resultados y para los cuales los parámetros aparentemente no están optimizados no son seguidos posteriormente, pero las variaciones de parámetros que tienen éxito son seguidas posteriormente. Finalmente, se ha conseguido un conjunto optimizado de parámetros que reproduce la base de datos ideal de una manera óptima mediante una clasificación automática.

Descripción detallada

A continuación se explica en principio una búsqueda asociativa simulando una memoria asociativa en conexión con la Figura 1. Para la búsqueda asociativa se almacena un conjunto de documentos D1 a DN. Cada uno de estos documentos está representado mediante un correspondiente vector de características 110 que es una secuencia de un predeterminado número de bits. Cada uno de estos bits representa o indica la presencia o la ausencia de una cierta característica en el documento que representa. Tal característica puede por ejemplo ser un trigrama, y el vector de características puede entonces tener una longitud de acuerdo con el número de trigramas posible. Naturalmente, también es posible la representación de otras características.

Como ya se ha mencionado, cada documento está representado mediante un vector de características correspondiente 110. Estos vectores de características conjuntamente forman una matriz de vectores de características 100 que también puede ser denominada matriz de atributos binaria debido a que consta de bits individuales y cada bit representa un cierto atributo de un cierto documento.

De una forma análoga no solamente están representados los documentos sino que también puede ser representado el documento de consulta mediante una secuencia de bits 120 que está formada de acuerdo con las mismas reglas que son válidas para la formación de la matriz de atributos binaria. De acuerdo con un enfoque de la técnica anterior se genera una secuencia de bits de consulta de este tipo para representar el documento de consulta. La secuencia de bits 120 de consulta se combina a continuación con la matriz de atributos binaria 100 realizando unas operaciones lógicas en términos de bits de forma que para cada uno de los documentos D1 a DN se obtiene una medida de similitud que es después almacenada en un vector de medida de similitud 130. La forma más sencilla de calcular la medida de similitud es realizar una operación lógica AND entre la secuencia de bits de consulta y cada una de las filas de la matriz 100. Esto da a continuación para cada fila de la matriz 100 una fila resultante

correspondiente que tiene un cierto número de “1” lógicos. Cuanto más alto sea el número de “1” lógicos mayor es la similitud entre la secuencia de bits de consulta y un cierto vector de características. Basándose en esto se puede calcular la medida de similitud. Se pueden también imaginar otros modos más sofisticados de cálculo de la medida de similitud, por ejemplo además de solamente dirigir la operación de bits también es posible tener en cuenta los contenidos reales de los documentos D1 a DN, las similitudes fonéticas entre la secuencia de bits de consulta y los respectivos documentos. Las similitudes fonéticas pueden también ser tenidas en cuenta al calcular la medida final de similitud.

De acuerdo con el invento, no solamente se genera una secuencia de bits de consulta que representa todo el documento de consulta sino que se generan varias secuencias de bits de consulta, que representa cada una un respectivo fragmento del documento de consulta, tal como se explica en lo que sigue.

La Figura 2 explica cómo se puede procesar una consulta introducida con el fin de obtener un acceso asociativo o búsqueda asociativa más eficiente. La consulta de entrada 200 es primeramente dividida en fragmentos S1, S2, y S3 de un tamaño dado. El tamaño de los fragmentos es un parámetro que puede ser predefinido o que puede ser seleccionado por el usuario o puede ser fijado automáticamente usando un método de optimización que se usa para encontrar un conjunto de parámetros optimizado. La división se basa en unidades de elementos basados en los que el documento de consulta comprende, tal como por ejemplo basado en palabras. El documento de consulta de entrada puede por ejemplo comprender 15 palabras, y el tamaño dado del fragmento puede tener cinco, entonces los segmentos S1, S2, y S3 tiene cada uno una longitud de cinco palabras. Si la consulta comprendiera solamente 14 palabras, entonces el último fragmento S3 contendría solamente cuatro palabras.

Los segmentos individuales S1 a S3 son después convertidos en los correspondientes vectores de características F1 a F3, como se ha indicado en la Figura 2. Entonces para cada uno de ellos se realiza una consulta asociativa tal como la explicada en conexión con la Figura 1.

Como existen tres vectores de consulta de entrada, en consecuencia también existen tres vectores de medida de similitud o vectores de similitud que son obtenidos de la consulta. Los componentes vectoriales de los vectores de similitud representan la similitud entre la consulta de entrada, aquí el fragmento respectivo del que se codificó el respectivo vector de consulta introducido, y un respectivo documento individual del conjunto almacenado. Cada uno de los vectores de medida de la similitud SM1 – SM3 representa un aspecto de la similitud entre dicho documento de consulta de entrada del que se han obtenido los fragmentos, y el conjunto de documentos que está almacenado por una respectiva secuencia de bits para la búsqueda asociativa.

De acuerdo con este invento, se han de obtener unos vectores de consulta introducidos adicionales mediante la generación de unos fragmentos que usan un denominado “desplazamiento”, el cual se explica a continuación.

Como se muestra en la Figura 3, basado en los vectores individuales de medida de similitud SM1 a SM3 se calcula a continuación un vector FSM de medida de similitud. La Figura 3 ilustra esquemáticamente la adición vectorial de los vectores SM1 a SM3 individuales de medida de similitud.

Se debería tener en cuenta que la operación para obtener el vector FSM de medida final de similitud puede también comprender una ponderación de acuerdo con una cierta función de ponderación. Por ejemplo, los componentes vectoriales de los vectores de medida de similitud SM1 a SM3 que indican una mayor similitud pueden ser ponderados más alto que los que tienen una similitud más baja (por ejemplo, inferior al 50%). Esto da un peso mayor a los documentos del conjunto almacenado para los que se ha encontrado realmente una similitud, y con el fin de disminuir el peso de los documentos para los que la similitud realmente es sólo un tipo de “ruido”.

A continuación se explica una realización del invento en conexión con la Figura 4. Suponiendo que un documento de consulta de entrada tiene una longitud de 16 palabras. Suponiendo además que el tamaño del fragmento es de cuatro palabras, entonces la consulta introducida 400 se divide en cuatro bloques 410 a 440 como se muestra en la Figura 4, teniendo cada bloque cuatro palabras.

Además de los cuatro fragmentos de entrada que se generan por división de la consulta de entrada se generan otros fragmentos mediante un proceso de desplazamiento (ventana deslizante) como se ha ilustrado en la Figura 4. Tal desplazamiento significa que los límites del bloque generado mediante división de la consulta de entrada son desplazados una cierta cantidad de elementos, aquí dos palabras. Esto da lugar a cuatro bloques de fragmentos 450 a 480, como está ilustrado en la Figura 4, en donde los límites de los bloques individuales se desplazan cuando se comparan con los cuatro bloques originales.

Esto está además ilustrado en la Figura 5, en la que los bloques individuales y el número de palabras contenidas en ellos se muestran esquemáticamente. La división en fragmentos genera cuatro bloques, conteniendo cada uno cuatro palabras, y el desplazamiento genera cuatro bloques adicionales, en tanto que el último bloque en estos cuatro bloques 510 solamente contiene dos palabras, es decir las palabras 15 y 16 de la consulta original de entrada. Como consecuencia, mediante la división y el desplazamiento se han generado en total ocho fragmentos de la consulta de entrada que pueden ser usados como entrada para una búsqueda asociativa como se ha explicado

anteriormente. Por lo tanto, no solamente hay entonces tres vectores de medida de similitud como en la Figura 2 sino más bien habrá hasta ocho vectores de medida de similitud. Sobre la base de estos ocho vectores de medida de similitud se calculará un vector de medida final de similitud de una manera análoga a la explicada en conexión con la Figura 3.

Basadas en las medidas finales de similitud así obtenidas como componentes vectoriales del vector de medida final de similitud se realiza la recuperación. Para cada componente del vector FSM de medida final de similitud existe un documento correspondiente almacenado en el conjunto, y dependiendo del conjunto de criterios se recuperará solamente el documento más importante, o se recuperarán los documentos cuya medida de similitud se encuentre más allá de un cierto valor umbral, o se recuperarán los 10 documentos más importantes o similares. Esta clase de criterios de recuperación puede ser cambiada dependiendo de los deseos del usuario, será recuperado como una salida, ninguno, uno o más documentos.

Debido a que la recuperación ha sido realizada juzgando la similitud entre el documento de consulta de entrada y los documentos almacenados para acceso asociativo, la salida puede ser tal que sea similar a la consulta de entrada con respecto a los vectores de características.

Como los vectores de características, o en otras palabras, las características mismas representan el contenido real de la consulta de entrada así como de los documentos del conjunto, no solamente existe una similitud en esas características sino también realmente también una similitud en el contenido entre los documentos recuperados para los que existe una medida de similitud grande. Esto significa que con el método explicado antes un usuario puede generar o aplicar una consulta entrada arbitraria que define el campo de interés para el cual el usuario desea realizar algunas “asociaciones”, y la búsqueda asociativa realizará realmente tales asociaciones entregando los documentos que son similares en contenido al documento de consulta de entrada.

La división del documento de consulta de entrada en algunos casos “aumenta la resolución” del acceso de memoria asociativa. Además, como el acceso de memoria asociativa realiza “asociaciones verdaderas” en algún sentido, es ventajoso si la consulta introducida está enfocada sobre un cierto tema o aspecto de un tema más que contener muchos aspectos diferentes y partes de texto. Esto se consigue mediante la división de la consulta introducida en fragmentos individuales. Estos fragmentos son preferiblemente enfocados, y dan realmente una imagen buena y enfocada de la similitud entre los documentos representados por los vectores de características de la secuencia de datos en la matriz de atributos binaria para el acceso asociativo y los fragmentos individuales más bien que entregar una medida de similitud no enfocada que contenga gran cantidad de “ruido”, en caso de que la consulta introducida sea muy larga.

Con el fin de evitar que el ruido sea después nuevamente introducido al calcular el vector de medida final de similitud, dicho vector de medida final de similitud se calcula usando alguna función de ponderación que da más peso a los documentos en los que se ha medido un valor de similitud mayor que en los que se ha medido un valor de similitud más bajo. Un ejemplo de tal función de ponderación está esquemáticamente ilustrado en la Figura 6. Como puede verse en la Figura 6, los documentos en los que la medida de la similitud es superior al 50% se les asigna un peso mayor, y aquéllos cuya medida de similitud es inferior al 50% se les asigna un peso inferior. Cerca del 100% y cerca del 0% la función de ponderación es más bien muy inclinada con el fin de dar un alto peso a los fragmentos de consulta introducidos en los que la correspondiente medida de similitud de los respectivos documentos ha sido calculada como relativamente alta. Al calcular la medida final de similitud como se ha ilustrado en la Figura 3, se tiene en cuenta la función de ponderación de tal manera que para cada uno de los componentes de los vectores individuales de medida de similitud SM1 a SM3 el valor de ponderación se coge de la función de ponderación mostrada en la Figura 6, y el componente del vector es multiplicado por el correspondiente valor de ponderación. Con esto se puede asignar un peso mayor a los valores de medida de similitud mayores al calcular el vector FSM de medida final de similitud.

El efecto que se consigue dividiendo la consulta en fragmentos y desplazando dichos fragmentos, como se ha explicado anteriormente, puede también conseguirse de un modo similar por un método que representa una realización del invento como se explica más adelante. La Figura 7 muestra un vector de características de consulta de entrada 700 que tiene N elementos individuales, aquí palabras. Basado en un tamaño de fragmento n que es predefinido o seleccionado o introducido por un usuario, se genera un primer fragmento 710. A continuación, de acuerdo con un parámetro adicional que puede ser denominado un parámetro de desplazamiento s se generan unos fragmentos adicionales de la siguiente manera. Un primer fragmento se genera a partir de la palabra número s y que se amplía a la palabra número n+s. Este fragmento se muestra como el fragmento 720 en la Figura 7. A continuación se genera un fragmento adicional partiendo en la palabra número 2s y que se amplía a la palabra número n+2s, manteniendo de este modo un tamaño de fragmento de n.

De este modo se generan varios fragmentos mediante el desplazamiento de la palabra de comienzo para cada fragmento s palabras hacia el lado derecho del documento de entrada 700 de la consulta. Esto da lugar a una pluralidad de fragmentos de la introducción de la consulta, y para cada uno de estos fragmentos se genera a continuación un correspondiente vector de características que se usa como una introducción de consulta para la búsqueda asociativa.

Lo que las realizaciones descritas hasta ahora tienen en común es que todas usan varios fragmentos generados basados en un documento de consulta de entrada, y basados en esos varios fragmentos se generan varios vectores de características de consulta que pueden ser usados para efectuar la búsqueda asociativa. Los parámetros individuales tales como el tamaño de los fragmentos individuales o el valor o número en el que se solapan los fragmentos individuales, tal como el valor s en la Figura 7, pueden ser predefinidos de acuerdo con un conjunto de parámetros adecuado, pueden ser seleccionados o elegidos por el usuario, o pueden ser optimizados mediante algún procedimiento que deduzca los parámetros y encuentre un grupo de valores que produzca un resultado óptimo, como ya se ha explicado anteriormente.

De acuerdo con una realización adicional, no solamente se usa para generar fragmentos el documento de consulta de entrada sino que también se usan los documentos almacenados en la memoria asociativa de una forma similar a la fragmentación del documento de consulta de entrada. Si se adopta tal medida, entonces esto da lugar realmente a varias matrices de atributos binarias, lo cual por supuesto aumenta la potencia de cálculo necesaria para realizar un acceso asociativo. Si se han generado tales matrices de atributos binarias por la generación de varios fragmentos de los documentos almacenados en la memoria asociativa y calculando los correspondientes vectores de características, entonces los fragmentos de consulta de entrada se aplican como consultas para todas las matrices de atributos binarias, aumentando de este modo el número de medidas de similitud obtenidas. Que tal enfoque sea factible en lo relativo a la potencia de cálculo puede depender del ordenador usado así como del número de documentos que realmente han de ser almacenados y representados.

Con respecto al desplazamiento de los fragmentos individuales con el fin de tener fragmentos solapados como se ha explicado en conexión con la Figura 7, tal desplazamiento puede realizarse basado en cualquiera de las unidades basadas en las que el documento de consulta realmente está compuesto. Si tal unidad puede por ejemplo ser una palabra, entonces esto significa que en el caso de la Figura 7, como ya se ha explicado el primer fragmento, contiene n palabras, el segundo fragmento comienza con la palabra $s^{\text{-ésima}}$ y llega a la palabra $(n+s)^{\text{-ésima}}$, y así sucesivamente. Otra unidad podría también ser una sílaba más bien que una palabra, o incluso un carácter. Se puede usar cualquier unidad que sea una unidad basada en la que el documento de consulta está compuesto. No obstante, como el proceso de acceder a la memoria asociativa simulada es con respecto a la obtención de asociaciones que de algún modo tengan sentido en cuanto a su contenido informativo, parece razonable que los elementos usados para generar los fragmentos y para generar los fragmentos solapados deberían ser elementos que ellos mismos contuvieran algún tipo de significado o al menos alguna información en ellos mismos, tal como palabras o al menos sílabas. Otra unidad de este tipo podría ser un fonema, lo que se explicará con más detalle en conexión con una realización posterior.

A continuación se explicará en conexión con la Figura 8 una realización del presente invento que puede ser usada para clasificar un documento de consulta de entrada en lo relativo a cuál de una pluralidad de clases de clasificación pertenece. En la Figura 8 se muestra la matriz de atributos binaria 800 que representa un número N de documentos D_1 a D_n por los respectivos vectores de características de la secuencia de bits. Cada uno de estos documentos está clasificado como que pertenece a una de tres clases de clasificación. Por ejemplo, los documentos D_1 a D_n pueden ser artículos, y están clasificados como que son artículos científicos (clase 1), artículos de prensa (clase 2), u otros artículos (clase 3). El vector 810 mostrado en la Figura 8 contiene la etiqueta que asigna una clase de clasificación a cada uno de los documentos D_1 a D_n.

Supongamos que el número total de documentos N sea mucho mayor de 3, de modo que cada una de las clases de clasificación contenga un número de documentos comparativamente grande. Supongamos además ahora que se usa un documento de consulta de entrada 820 como una consulta de entrada para la búsqueda asociativa. Un usuario puede ahora estar interesado no solamente en la recuperación de esos documentos entre D_1 a D_n, los cuales son los más similares a la consulta 820, pero también puede estar interesado en la clasificación si el documento de consulta 820 pertenece a la clase 1, la clase 2, o la clase 3.

Supongamos ahora que la consulta es realizada para una pluralidad de fragmentos, y supongamos que el criterio de recuperación es tal que son recuperados todos esos documentos para los que la medida de similitud es superior a un cierto umbral. Si el umbral de similitud se fija de modo que se obtenga una pluralidad de resultados de recuperación, entonces será una pluralidad de aciertos (documentos recuperados) para cada una de las diferentes clases 1, 2 ó 3. Sin embargo, basándose en la suposición de que los documentos que pertenecen a una cierta clase son en alguna manera autosimilares, habrá un significativamente gran número de documentos recuperados de la clase a la que el documento de consulta 820 realmente pertenece.

La Figura 9 ilustra las estadísticas de recuperación que pueden obtenerse al realizar una consulta basada en un documento de consulta de entrada en un conjunto de documentos almacenados, es decir documentos representados por un respectivo vector de características de la secuencia de bits en la matriz de atributos binaria, que pertenece a una de las tres clases. En el ejemplo mostrado en la Figura 9 existen 85 aciertos que pertenecen a la clase 2, esto significa que el documento de consulta de entrada más probablemente también pertenece a esta clase. Un “acierto” significa aquí que un resultado de la consulta satisface un cierto criterio tal como una medida de confidencialidad que está por encima de un cierto umbral, de modo que puede ser considerado como un “acierto”. El criterio particular y el valor umbral pueden ser seleccionados dependiendo de las circunstancias.

Basándose en tales estadísticas de recuento de aciertos se puede juzgar a qué clase de un conjunto de clases pertenece un documento entrada, y se puede realizar una clasificación automática. Esta clasificación automática será la mejor cuanto más precisa sea la clasificación original. Si el vector de clasificación 810 de la Figura 8 ya contiene algunas clasificaciones equivocadas, entonces las estadísticas de recuperación de la Figura 9 contendrán los correspondientes errores. No obstante, si la clasificación original es muy buena, entonces también es relativamente alta la probabilidad de que la estadística de recuperación sea una buena base para clasificar un documento de consulta desconocido y no clasificado.

Con respecto a esto, se debería también tener en cuenta que el conjunto de documentos para los que la clasificación original es muy buena es la más adecuada como un conjunto de comienzo para encontrar un conjunto optimizado de parámetros tal como el tamaño de fragmentación y el desplazamiento o valor de solape explicado en conexión con la Figura 7. Supongamos que se usa una muy buena clasificación y un correspondiente conjunto de documentos ya clasificados, entonces para cada uno de los documentos de este conjunto se realiza una clasificación automática como ya se ha explicado antes. Como la clasificación original ya es buena o en algún caso ideal ya perfecta, las estadísticas de recuperación de la Figura 9 siempre deberían dar el valor correcto en el sentido de que la clasificación resultante también fuera correcta. Si éste no es el caso, entonces posiblemente algo está equivocado con los parámetros, y éstos deberían ser variados. Con los parámetros variados entonces nuevamente se puede llevar a cabo una nueva pasada basada en el conjunto de datos original, y puede ser comprobado si la clasificación con los nuevos parámetros da ahora un mejor resultado. Con tal método de una manera optimizada se puede encontrar un conjunto de parámetros ideal u óptimo.

Hasta ahora no se ha prestado atención alguna al tipo de documentos introducidos usados como el documento de consulta para consultar la memoria asociativa simulada. En principio, se puede usar cualquier tipo de documento de entrada, y cualquier tipo de documento puede ser representado por un vector de características para la búsqueda asociativa. Sin embargo, de ahora en adelante se explicará una realización particular en la que el documento de consulta de entrada así como los documentos representados para el acceso asociativo son documentos de voz.

Supongamos que el conjunto de documentos para ser representados para la búsqueda asociativa son ficheros que representan algunos textos hablados por una voz humana. Dichos textos pueden por ejemplo ser discursos radiofónicos, discursos televisivos, entrevistas, conversaciones telefónicas, o similares. Estos documentos de voz pueden ser convertidos en documentos de texto de acuerdo con un método convencional discurso-a-texto que los convierte en un correspondiente fichero de texto. Posiblemente, antes de la conversión de discurso-a-texto se podría haber realizado un registro de la voz que registra la voz en algún fichero de sonido, tal como un fichero .wav. Esto está ilustrado de forma esquemática en la Figura 10.

Por ejemplo, pueden ser muchos de tales ficheros de texto los que resulten de alguna voz humana, y dichos ficheros de texto podrían entonces ser almacenados como documentos en una memoria asociativa simulada a través de sus correspondientes vectores de características. Para tal fin cada uno de los ficheros de texto es convertido en su correspondiente vector de características, y el conjunto así resultante de vectores de características es dispuesto en una matriz para obtener la matriz de atributos binaria como ya se ha explicado en conexión con la Figura 1.

A continuación es posible usar también algún texto hablado por una voz humana como una introducción de consulta. Para tal fin, similar al proceso mostrado en la Figura 10, se genera un fichero de texto que representa la consulta. También este fichero de texto es representado en un correspondiente vector de características, y a continuación se puede obtener un acceso de memoria asociativa como ya se ha explicado antes.

Usando tal método es posible recuperar a partir de un archivo de sonido los elementos que son similares en contenido a cualquier sonido de consulta dado. Supongamos, por ejemplo, que los documentos representados para el acceso asociativo por un respectivo vector de características son textos de noticias, entonces sería posible para hablar un texto de consulta en un micrófono, registrarlo de acuerdo con la Figura 10, transferirlo en un fichero de texto, y recuperar de la memoria asociativa simulada esos ficheros de texto y a continuación los correspondientes ficheros de voz que son más similares en contenido a la consulta hablada.

A continuación se explicará con detalle una posterior realización del presente invento en conexión con la Figura 11. En una base de datos 900 se han almacenado muestras de ficheros de audio para permitir un acceso asociativo. Con tal fin las muestras son convertidas en ficheros de texto, los ficheros de texto son convertidos en vectores de características, y los vectores de características son dispuestos en una matriz de atributos binaria como ya se ha explicado.

Un fichero de audio 910, el cual puede por ejemplo adoptar la forma de un fichero .wav, es después introducido en el sistema y preprocesado acústicamente en el paso 920. El problema de los ficheros acústicos es a veces que la gama dinámica es demasiado grande, el nivel de ruido es demasiado alto, etc. Para mejorar la calidad en el paso de procesamiento previo 920 se normaliza la gama dinámica, se filtra el ruido, etc. Esto asegura que en el próximo paso de reconocimiento de voz 930 la máquina de reconocimiento de voz esté recibiendo un fichero de audio que es más adecuado para obtener mejores resultados en el proceso de conversión de voz-a-texto. El paso 930 devuelve como una salida un texto en el que ha sido convertido el fichero de audio.

En el paso 940 se usa a continuación un ordenador de secuencia / analizador para generar los fragmentos de la muestra de texto de entrada. Esto puede ser realizado de acuerdo con cualquiera de los métodos de generación de fragmentos explicados antes. Un resultado del ordenador de secuencia / analizador es una pluralidad de fragmentos del texto generado por la unidad de reconocimiento de voz 930.

Basado en esta pluralidad de fragmentos de texto se realiza a continuación el acceso asociativo directo en el paso 950 para cada uno de los fragmentos. Esto da como resultado una respectiva medida de la similitud de las muestras almacenadas por un respectivo vector de características en la base de datos 900 de la memoria, y dependiendo del criterio de restitución fijado se produce una o más muestras que tienen la medida de similitud más alta, o que tienen una medida de similitud por encima de un cierto umbral. Las muestras producidas pueden ser consideradas como que son las más similares a la muestra que ha sido introducida como un fichero de audio en el paso 910.

De acuerdo con una realización adicional, el método explicado en conexión con la Figura 11 puede ser incluso más afinado al tener dos conjuntos de muestras, un conjunto de muestras 900 mostrado en la Figura 11, y otro conjunto de muestras 905 que no está mostrado en la Figura 11. Supongamos que el primer conjunto de muestras 900 contiene unas muestras importantes en el sentido de que son muestras que pueden ser clasificadas como interesantes desde el punto de vista del usuario, y el otro conjunto de muestras 905 contiene unas muestras de voz o muestras de ficheros de voz que no son interesantes desde el punto de vista del usuario. En tal caso el acceso a la memoria asociativa puede ser combinado con una estadística de recuperación como se ha explicado en conexión con la Figura 9. Para un fichero de introducción de audio dado 910 habrá muchas muestras recuperadas del conjunto 900, otros muchos aciertos en el conjunto de muestras 905. Si el conjunto importante de muestras 900 entrega el mayor número de aciertos, entonces se puede suponer que el fichero de audio introducido es también de importancia para el usuario. Sin embargo, si el número de aciertos en el conjunto de muestras no importante 905 es más alto, entonces se puede suponer que el fichero de audio en su contenido no es de particular interés al usuario y no necesita ser sacado juntamente con los aciertos encontrados. Con tal método, la clasificación y la recuperación pueden ser combinadas con el fin de comprobar si un fichero de audio de entrada desconocido si, en todo caso, es importante, y si lo es, recuperar aquellas muestras ya almacenadas que son similares en contenido al fichero de audio introducido.

De acuerdo con una posterior realización, en el paso de reconocimiento de voz 930 la conversión no se realiza en un fichero de texto sino preferiblemente en un fichero cuyos elementos son fonemas codificados en una forma apropiada, tal como mediante las secuencias de caracteres ASCII. Se debería tener en cuenta que el reconocimiento de voz principalmente tiene dos pasos, primeramente una conversión de los ficheros de audio en fonemas (un reconocimiento de los fonemas de cada una de las palabras habladas), y después una conversión a partir de los fonemas reconocidos en las palabras finales del texto. Ocurre relativamente a menudo en el proceso del reconocimiento de voz que los fonemas son reconocidos correctamente, aunque el reconocimiento final de cada una de las palabras no se haya realizado correctamente. Si el paso 930 de reconocimiento de voz está limitado a realizar solamente la conversión en fonemas, y si la base de datos 930 de muestras almacena también la muestra en la forma de fonemas más bien que en forma de texto, entonces esta fuente de error puede ser evitada y se puede esperar un procesamiento más preciso.

Como resultará evidente a partir de la anterior explicación, el sistema explicado en conexión con la Figura 11 puede ser usado con dos fines. Primero, en el caso de que el usuario tenga una cierta idea del contexto en el que desea tener la asociación, o en otras palabras, si él ya conoce qué tipo de contenido deberían tener los documentos que han de ser recuperados, un usuario puede solamente componer un fichero de audio de entrada recitando un cierto texto que considere como interesante o importante. El proceso mostrado en la Figura 11 recuperará entonces como salida los documentos procedentes de las muestras almacenadas que sean más similares en su contenido al fichero de audio introducido generado por el usuario.

Otra aplicación podría ser que el contenido y la importancia del fichero de audio 910 propiamente dicho no se conozca todavía. Realizando un acceso asociativo directo como se ha explicado en conexión con las dos muestras 900 y 905 se puede primeramente comprobar si el fichero de audio es importante desde el punto de vista del usuario en todo caso comprobando si su clasificación lo clasificaría como importante o no importante, en otras palabras como que pertenece a las muestras 900 o a las muestras 905. Si el fichero introducido es considerado como importante, entonces se puede realizar una recuperación de acceso asociativo y documentos similares pueden ser sacados para el usuario del conjunto de muestras 900.

El método descrito en conexión con la realización anterior es una herramienta muy potente que puede usarse para reducir “asociaciones” como el cerebro humano. Basado en un documento de consulta introducido se han producido unos documentos de salida que son asociaciones relacionadas con el documento de entrada de tal manera que pueden tener un contenido similar al de dicho documento de entrada. “Contenido similar” es aquí muy difícil de describir en términos más exactos, un intento podría ser decir que los documentos encontrados tienen aspectos en los que pueden ser considerados similares al documento de consulta de entrada. En alguna medida estos aspectos dependen de la forma en la que el vector de características es producido sobre la base del documento de consulta y en los documentos almacenados para el acceso asociativo. Si los vectores de características se producen basados en bigramas o trigramas, entonces los documentos recuperados representan similitudes en términos de sílabas

consecutivas. Esto necesariamente refleja también alguna similitud en el contenido, debido a que las palabras similares que se producen en el documento de consulta y los segundos documentos almacenados dan lugar a unos vectores de características similares o idénticos.

Sin embargo, dependiendo de cómo se calcule realmente la medida de similitud se pueden tener en cuenta también otros aspectos de similitud, tal como la similitud fonética que podría ser más bien importante si se usaran fonemas para construir los vectores de características en vez de las secuencias de caracteres ASCI.

El presente invento en sus realizaciones es por lo tanto capaz de realizar “asociaciones verdaderas”, y esto es una herramienta muy potente para cualquier ser humano enfrentado al procesamiento del flujo de información que se encuentra cada día. Para muchas o todas las piezas de información que encontramos hoy en día tenemos que realizar la tarea de establecer asociaciones, es decir que tenemos que recordar o encontrar documentos/sucesos/decisiones etc similares para colocar la nueva información en el marco correcto con el fin de ser capaces de procesarlas de una forma apropiada. El cerebro humano es muy bueno para realizar estas asociaciones, aunque está muy limitado en el sentido de que su almacenamiento contiene un conjunto limitado de documentos de muestra. Este reducido conjunto de muestras es limitado debido a que solamente se almacenan en el cerebro humano las muestras que el usuario o el ser humano es capaz de leer y de oír. Sin embargo, la capacidad real de lectura y de escucha de un ser humano es muy pequeña en comparación con la cantidad de información producida cada día. Es por tanto muy útil si una máquina fuera capaz de almacenar muestras y a continuación recuperar de dichas muestras almacenadas las que pueden ser consideradas como una “asociación” con una muestra de entrada dada. Un usuario provisto de una máquina como la explicada en conexión con la Figura 11 por lo tanto es capaz de acceder de una manera asociativa a una muestra enorme de documentos que él realmente puede en cualquier caso no haber visto, leído, u oído, sin embargo, que están almacenadas en una base de datos 900 de la memoria para permitir un acceso asociativo como el explicado. Si un usuario después está interesado en un cierto aspecto de alguna parte de información, él puede crear su propio fichero de entrada 910 y después la máquina recuperará para él las asociaciones a partir del “conocimiento” previamente almacenado que son más interesantes para él.

De acuerdo con otro aspecto un usuario puede estar enfrentado a una nueva entrada que es incapaz de colocar en el marco correcto debido a que no tiene conocimiento” sobre ella. De una manera similar, un usuario puede entonces hacer uso de la máquina mostrada en la Figura 11 para recuperar muestras de un depósito o base de datos 900 que pueden ser consideradas como “asociaciones” del documento de entrada desconocido 910.

Tal proceso o máquina explicados en conexión con la Figura 11 es por lo tanto de sustancial ayuda para cualquier ser humano enfrentado al flujo de información que se encuentra cada día. Es muy útil al realizar asociaciones colocar la información desconocida en el marco correcto con el fin de recuperar muestras de una base de datos de conocimiento que de otro modo él nunca encontraría basándose en un documento autocreado, o con el fin de clasificar los datos de entrada no clasificados en uno de un conjunto de clases de clasificación predefinidas.

Considerando que un ser humano al procesar la información que encuentra no hace otra cosa más que realizar asociaciones y llevar a cabo clasificaciones, tal como una máquina explicada en conexión con la Figura 11 de hecho es capaz de ayudar al usuario a procesar información y realizar mucha parte del trabajo que hasta ahora tiene que ser hecho por un ser humano.

REIVINDICACIONES

1. Un método para recuperar de un conjunto de documentos electrónicos aquéllos que están próximos en contenido con un documento electrónico usado como documento de entrada, comprendiendo dicho método:

A) proporcionar un conjunto de documentos (D1,...DN);

B) generar unos vectores de características de secuencias de bits (110) para los documentos de dicho conjunto, en los que cada vector de características representa uno respectivo de dichos documentos e indica con cada uno de sus componentes vectoriales binarios la presencia o la ausencia de una cierta característica dentro del respectivo documento;

C) formar una matriz de atributos binaria (100; 800) a partir de los vectores de características (110) de la secuencia de bits, constando la matriz de atributos binarias de unos bits individuales y representando cada bit un cierto atributo de un cierto documento indicando la presencia o ausencia de una cierta característica dentro del respectivo documento;

D) proporcionar un documento de entrada (200; 400; 700; 820);

E) generar una pluralidad de fragmentos (S1, S2, S3; 410, 420, 430, 440, 450, 460, 470, 480; 710, 720, 730) de dicho documento de entrada (200; 400; 700; 820) que son partes de dicho documento de entrada;

F) generar para cada uno de dichos fragmentos un respectivo vector de características (F1; F2; F3) de la secuencia de bits de consulta que representa el respectivo fragmento y que indica con cada uno de sus componentes de vector binarios la presencia o ausencia de una cierta característica dentro del respectivo fragmento;

G) efectuar una búsqueda asociativa para cada uno de dichos fragmentos usando dicha matriz de atributos binaria y el respectivo vector de características (F1; F2; F3) de la secuencia de bits de consulta que representa el respectivo fragmento; en el que la búsqueda asociativa de un respectivo fragmento se efectúa determinando para el fragmento y para cada documento de dicho conjunto una medida de similitud individual entre dicho fragmento y el respectivo documento de dicho conjunto; en el que la respectiva medida de similitud individual se determina efectuando una operación lógica con bits entre el vector de características de secuencias de bits que representa el respectivo fragmento y la secuencia de bits que representa el respectivo documento en dicha matriz de atributos binaria; en el que dichas medidas de similitud individuales, estando cada una determinada para un determinado fragmento y un respectivo documento de dicho conjunto y que representa una similitud entre dicho fragmento particular y dicho documento respectivo de dicho conjunto, son componentes vectoriales de un vector de medida de similitud (SM1; SM2; SM3) asociados con los respectivos fragmentos particulares y con todos los documentos de dicho conjunto de documentos;

H) para cada uno de dichos documentos de dicho conjunto se calcula una medida final de similitud individual a partir de las medidas de similitud individuales, las cuales de acuerdo con el paso G) se determinan para el mismo documento respectivo con relación a dichos fragmentos, siendo dichas medidas finales de similitud componentes vectoriales de un vector (FSM) de medida final de similitud, el cual se obtiene con él sobre la base de dichos vectores de medida de similitud (SM1, SM2, SM3);

I) juzgar, sobre la base de dichas medidas finales individuales de similitud incluidas como componentes vectoriales en dicho vector (FSM) de medida final de similitud pudiéndose los documentos de dicho conjunto ser considerados como que tienen un contenido próximo al de dicho documento de entrada;

en el que dichos fragmentos (S1, S2, S3; 410, 420, 430, 440, 450, 460, 470, 480; 710, 720, 730) son generados segmentando dicho documento de entrada (200; 400; 700; 820) en segmentos, en donde de dichos segmentos los del primero al segundo últimos tienen un tamaño dado, y el último segmento tiene un tamaño dado o es más corto que el tamaño dado;

en el que los fragmentos del documento de entrada (450, 460, 470, 480; 720, 730) son generados por desplazamiento de uno o más fragmentos de dicho documento de entrada para generar unos fragmentos solapados (410, 420, 430, 440, 450, 460, 470, 480; 710, 720, 730) de dicho documento de entrada (400; 700) para cuyos vectores de características de secuencia de datos son generados y usados en dicha búsqueda asociativa.

2. El método de la reivindicación 1, en el que dicho solape entre dichos fragmentos de solapamiento es un solape en unidades de uno de los siguientes:

caracteres;

sílabas;

fonemas;

palabras.

3. El método de una de las anteriores reivindicaciones, en el que dicho documento de entrada se obtiene a partir de la realización de un proceso de reconocimiento de voz sobre un texto hablado por una voz humana.

4. El método de una de las anteriores reivindicaciones, en el que cada uno de dichos documentos (D1, ..., DN) de dicho conjunto es clasificado como que pertenece a una de una pluralidad de clases (810), comprendiendo además dicho método:

realizar dicha búsqueda asociativa,

basado en el número de documentos similares al documento de entrada (820) que ha sido encontrado en las clases individuales, que clasifican dicho documento de entrada como perteneciente a una de dichas clases.

5. El método de la reivindicación 4, en el que el número de clases es dos, una de un conjunto importante de documentos y otra de un conjunto no importante de documentos;

se restituyen los resultados de la recuperación a partir de dicha búsqueda asociativa a un usuario solamente si dicho documento de entrada ha sido clasificado como perteneciente a dicha clase importante.

6. El método de una de las anteriores reivindicaciones, que además comprende:

optimizar los parámetros de dicho método como el tamaño del fragmento y el número de elementos de solape basado en un esquema de clasificación dado, comprendiendo dicha optimización:

a) usar los documentos de dicho esquema de clasificación de los que su clasificación es ya conocida;

b) clasificarlos de acuerdo con la reivindicación 6;

c) variar dichos parámetros y repetir los pasos a) y b); repitiendo el paso c) hasta haber encontrado un conjunto óptimo de parámetros.

7. El método de una de las anteriores reivindicaciones, en el que los datos de entrada de voz están representados por fonemas para proporcionar dicho documento de entrada, y dichos documentos de dicho conjunto, para los cuales se generan dichos vectores de características de la secuencia de datos, contienen los datos del fonema;

o se usan unos datos de entrada escritos, y dichos documentos de dicho conjunto, para los que se generan dichos vectores de características de la secuencia de datos, contienen unos datos escritos con caracteres;

o se convierten los datos de entrada escritos en datos de fonemas para proporcionar dicho documento de entrada, y dichos documentos de dicho conjunto, para los que se generan dichos vectores de características de la secuencia de datos, contienen datos de fonemas.

8. El método de acuerdo con una de las anteriores reivindicaciones, en el que de acuerdo con el paso H) se calcula la respectiva medida final de similitud a partir de dichas medidas de similitud individuales usando una función de ponderación que da un peso más alto a unas medidas de similitud individuales que indican una similitud más alta entre el respectivo fragmento y un respectivo documento de dicho conjunto y que da un peso más bajo a unas medidas de similitud individuales que indican una similitud más baja entre el respectivo fragmento y un documento respectivo de dicho conjunto.

9. El método de acuerdo con una de las anteriores reivindicaciones, en el que dicha operación lógica en términos de bits es una operación lógica AND en términos de bits.

10. El método de acuerdo con la reivindicación 9, en el que las características presentes están representadas por un “1” lógico y el número de “1” lógicos que resulta de la operación lógica AND en términos de bits se usa como medida de similitud individual.

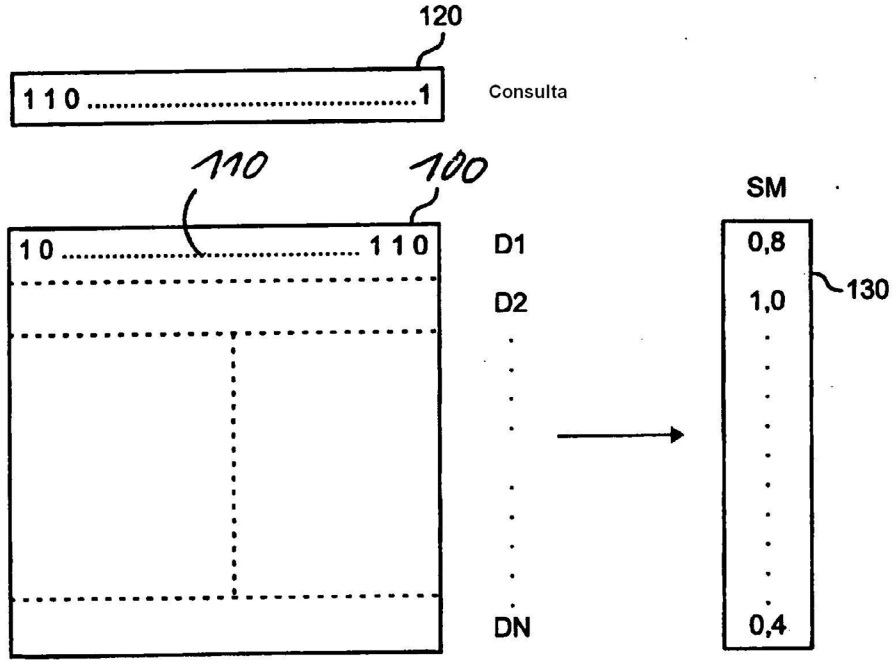

Fig. 1

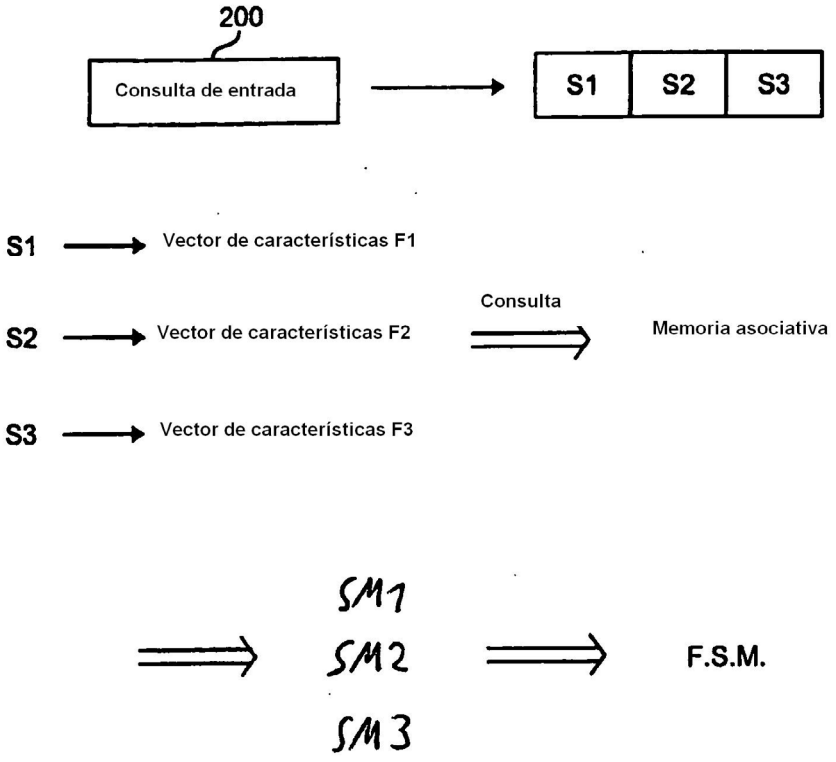
200
Consulta de entrada
S1
S2
S3
S1
Vector de características F1
S2
Vector de características F2
Consulta
Memoria asociativa
S3
Vector de características F3
SM1
SM2
SM3
F.S.M.

Fig. 2

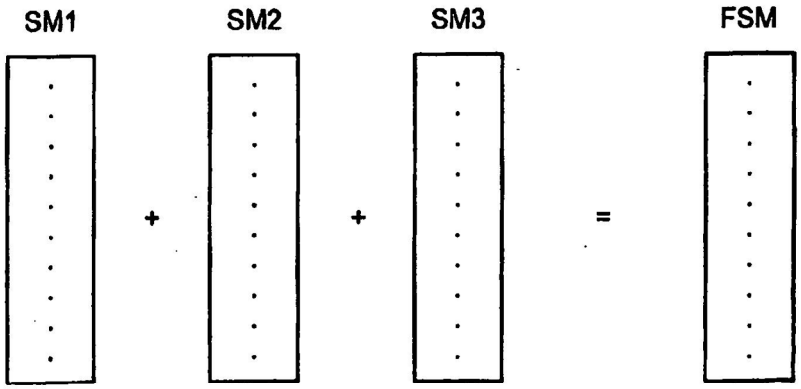
SM1
SM2
SM3
FSM
+
+
=

Fig. 3

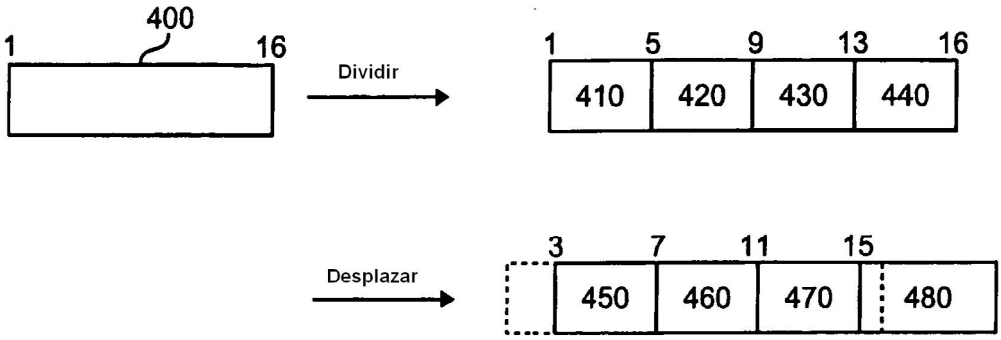

Fig. 4

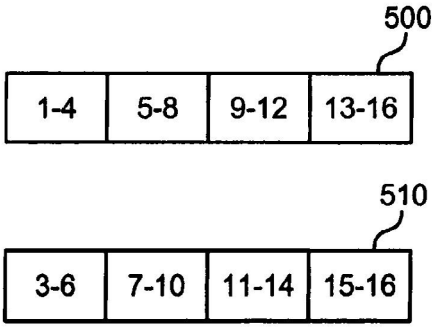

Fig. 5

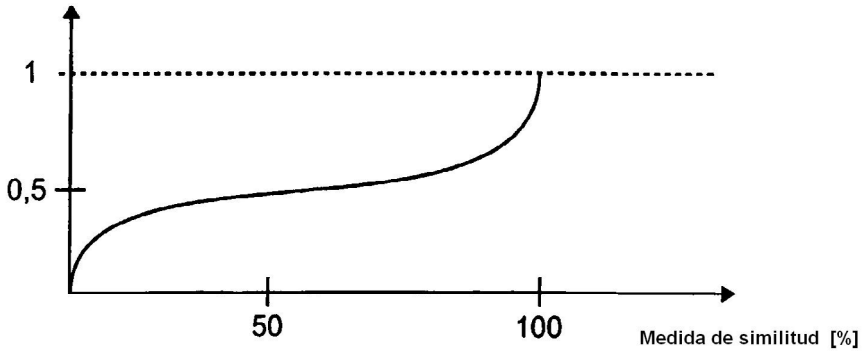

Fig. 6

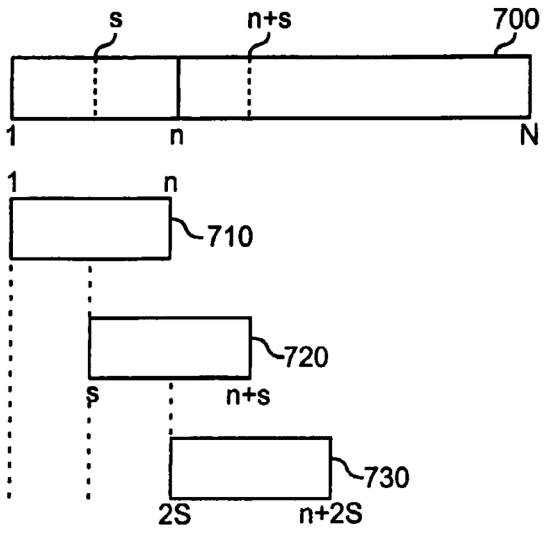

Fig. 7

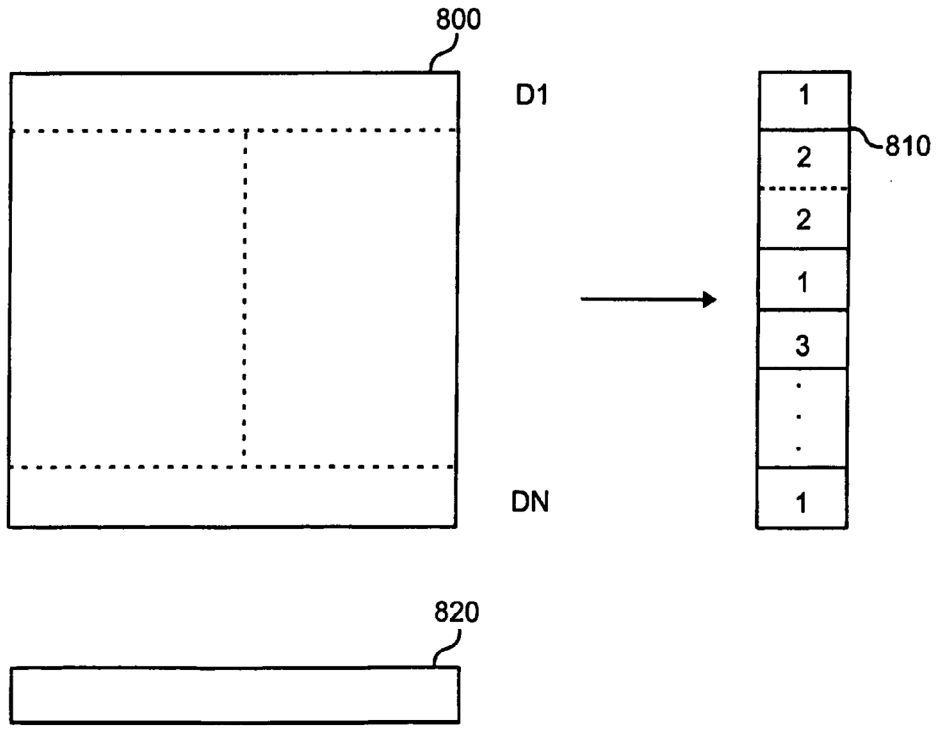

Fig. 8

Estadística de recuperación

Clase	1	2	3
Número de desplazamientos	5	85	12

Fig. 9

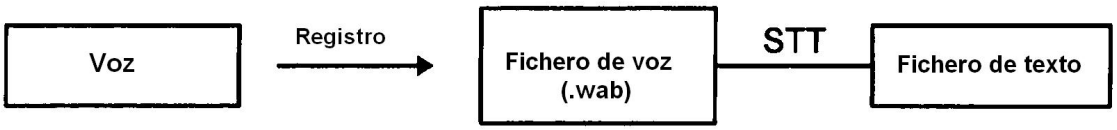

Fig. 10

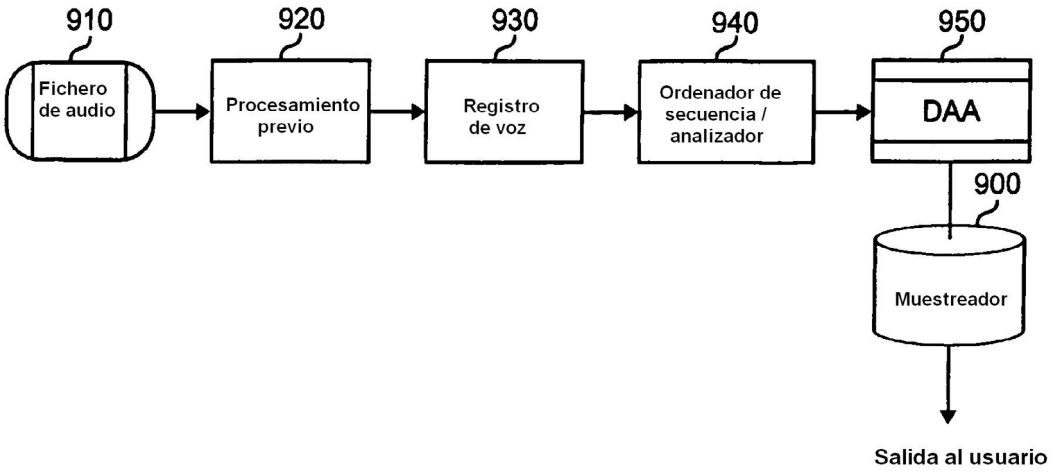
910
920
930
940
950
Fichero de audio
Procesamiento previo
Registro de voz
Ordenador de secuencia / analizador
DAA
900
Muestreador
Salida al usuario

Fig. 11