(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **3 027 185**

(51) Int. CI.:

***G06F 16/30*** (2009.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **29.10.2018 PCT/JP2018/040056**

(87) Fecha y número de publicación internacional: **16.05.2019 WO19093172**

(96) Fecha de presentación y número de la solicitud europea: **29.10.2018 E 18876872 (5)**

(97) Fecha y número de publicación de la concesión europea: **16.04.2025 EP 3709183**

(54) Título: **Dispositivo de cálculo del índice de similitud, dispositivo de búsqueda de similitud y programa de cálculo del índice de similitud**

(30) Prioridad:

**07.11.2017 JP 2017214388**

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
**13.06.2025**

(73) Titular/es:

**FRONTEO, INC. (100.00%)**
**2-12-23, Kounan, Minato-ku**
**Tokyo 108-0075, JP**

(72) Inventor/es:

**TOYOSHIBA HIROYOSHI**

(74) Agente/Representante:

**GONZÁLEZ PECES, Gustavo Adolfo**

## DESCRIPCIÓN

Dispositivo de cálculo del índice de similitud, dispositivo de búsqueda de similitud y programa de cálculo del índice de similitud

### Campo técnico

La invención actual se relaciona con un aparato informático del valor del índice de similitud, un aparato de búsqueda de texto similar, y un programa informático del valor del índice de similitud, y se relaciona particularmente con una tecnología para calcular un valor del índice de similitud relacionada con un texto incluyendo una pluralidad de palabras y una tecnología para realizar la búsqueda de la similitud usando este valor del índice.

### Técnica antecedente

Convencionalmente, se ha utilizado mucho una tecnología para buscar otro texto similar a un texto introducido como clave de búsqueda a partir de un gran número de textos almacenados en una base de datos. En este tipo de tecnología de búsqueda, básicamente se calcula una determinada cantidad de características para cada texto y se busca un texto que tenga una cantidad de características similar. Se conoce una tecnología para calcular un vector de texto como una cantidad de características (por ejemplo, véanse los Documentos de Patente 1 y 2).

En un aparato de búsqueda de información descrito en el Documento de Patente 1, se analiza un documento de una respuesta de búsqueda para extraer palabras independientes, y se lee un vector de palabras a partir de las palabras independientes obtenidas para una palabra independiente registrada en un diccionario de generación de vectores. A continuación, se obtienen vectores de texto que representan características de los textos a partir de todos los vectores de palabras obtenidos en todos los textos, se obtiene una distancia entre los textos comparando los vectores de texto y se realiza la clasificación utilizando la distancia.

Un sistema de búsqueda por categoría correspondiente descrito en el Documento de Patente 2 busca un par de documentos japoneses e ingleses que tengan significados similares. El sistema de búsqueda de categorías correspondiente realiza un proceso de análisis morfológico de todos los documentos japoneses e ingleses incluidos en los datos de aprendizaje, y calcula un vector de palabra multidimensional correspondiente para cada una de las palabras japonesas e inglesas obtenidas como resultado. A continuación, se calcula un vector de documento en el que se normaliza la suma de los vectores de palabra correspondientes a todas las palabras incluidas en cada documento (la longitud del vector es 1), y se busca un par de documentos japoneses e ingleses que tengan una relevancia máxima (un valor del producto interno es grande) utilizando el vector de documento correspondiente al documento japonés y el vector de documento correspondiente al documento inglés.

Además, se ha conocido una tesis que describe la evaluación de un texto o un documento mediante un vector de párrafo (por ejemplo, véase el documento no patentado 1). En una tecnología descrita en el Documento sin Patente 1, de forma similar a los Documentos de Patente 1 y 2, se calcula un vector de palabra para una palabra incluida en un texto, y se calcula un vector de párrafo utilizando el vector de palabra.

Documento de patente 1: JP-A-7-295994

Documento de patente 2: JP-A-2002-259445

Documento no patentado

Documento no patentado 1: "Representaciones distribuidas de frases y documentos", de Quoc Le y Tomas Mikolov, Google Inc, Actas de la 31ª Conferencia Internacional sobre Aprendizaje Automático celebrada en Bejing (China) del 22 al 24 de junio de 2014.

### Resumen de la invención

Problema técnico

Cada una de las tecnologías descritas en los Documentos de Patente 1 y 2 y en el Documento sin Patente 1 tiene un mecanismo para calcular vectores de texto como cantidades de características de los textos, comparar los vectores de texto o calcular un producto interno de los vectores de texto, clasificando así los textos o buscando textos similares.

Sin embargo, un método convencional de evaluación de similitudes que utiliza sólo un vector de texto como índice tiene el problema de que no se puede aumentar suficientemente la precisión de la evaluación, ya que un texto incluye una combinación de una pluralidad de palabras, mientras que no se evalúa con precisión qué palabra contribuye a qué texto y en qué medida.

Obsérvese que los vectores de texto descritos en los Documentos de Patente 1 y 2 y en el Documento sin Patente 1 se obtienen mediante un cálculo predeterminado utilizando un vector de palabra. Sin embargo, el Documento de Patente 1 no divulga un método específico para determinar un vector de texto a partir de un vector de palabra. En la tecnología descrita en el Documento de Patente 2, dado que la suma de los vectores de palabras correspondientes a todas las palabras incluidas en el documento se normaliza simplemente para ser un vector de documento, el vector de palabra de cada palabra utilizada en el documento se redondea como la suma. En la tecnología descrita en el Documento no Patentado 1, aunque se utiliza un vector de palabra en un proceso de obtención de un vector de párrafo, el vector de palabra no se utiliza como índice para evaluar un texto o un documento.

La invención se ha hecho para resolver tal problema, y un objeto de la invención es hacer posible mejorar la exactitud de la evaluación de la similitud más que antes. Solución al problema

Para solucionar el problema antedicho, se proporcionan un aparato y un programa de búsqueda de similitud como se define en las reivindicaciones adjuntas.

Efectos ventajosos de la invención

Según la invención reivindicada

dado que se calcula un producto interno de un vector de texto calculado a partir de un texto y un vector de palabra calculado a partir de una palabra incluida en el texto para calcular un valor de índice de similitud que refleje una relación entre el texto y la palabra, es posible detectar qué palabra contribuye a qué texto y en qué medida como un valor de producto interno. Por lo tanto, es posible mejorar la precisión de la evaluación de la similitud más que antes utilizando un valor de índice de similitud de la invención obtenido de este modo.

**Breve descripción de las figuras**

La Figura 1 es un diagrama de bloques que ilustra un ejemplo de configuración funcional de un aparato de cálculo del valor del índice de similitud según una realización.

La Figura 2 es un diagrama de bloques que ilustra un ejemplo de configuración funcional de un aparato de búsqueda de similitud según la realización.

La Figura 3 es un diagrama de bloques que ilustra otro ejemplo de configuración funcional de un aparato de búsqueda de similitud según la realización.

La Figura 4 es un diagrama de bloques que ilustra otro ejemplo de configuración funcional de un aparato de búsqueda de similitud según la realización.

La Figura 5 es un diagrama de bloques que ilustra otro ejemplo de configuración funcional de un aparato de búsqueda de similitud según la realización.

**Modo de realización de la invención**

En lo sucesivo, se describirá una realización de la invención con referencia a las figuras. La Figura 1 es un diagrama de bloques que ilustra un ejemplo de configuración funcional de un aparato de cálculo del valor del índice de similitud según la presente realización. Un aparato de cálculo del valor del índice de similitud 10 de la presente realización introduce datos de texto relacionados con un texto, y calcula y emite un valor de índice de similitud que refleja una relación entre el texto y una palabra contenida en el mismo. El aparato de cálculo del valor del índice de similitud 10 incluye una unidad de extracción de palabra 11, una unidad de cálculo vectorial 12 y una unidad de cálculo del valor del índice 13 como configuración funcional del mismo. La unidad de cálculo vectorial 12 incluye una unidad de cálculo de vector de texto 12A y una unidad de cálculo de vector de palabra 12B como configuración funcional más específica.

Cada uno de los bloques funcionales 11 a 13 puede configurarse mediante hardware, un procesador de señal digital (DSP) y software. Por ejemplo, en el caso de estar configurado por software, cada uno de los bloques funcionales 11 a 13 incluye en realidad una CPU, una RAM, una ROM, etc. de un ordenador, y se implementa mediante el funcionamiento de un programa almacenado en un medio de grabación como una RAM, una ROM, un disco duro o una memoria semiconductora.

La unidad de extracción de palabras 11 analiza m textos (m es un número entero arbitrario de 2 o más) y extrae n palabras (n es un número entero arbitrario de 2 o más) de los m textos. Aquí, un texto a analizar puede incluir una frase (unidad dividida por un punto) o incluir una pluralidad de frases. Un texto que incluya una pluralidad de frases puede corresponder a algunos o a todos los textos incluidos en un documento.

Además, para el análisis de un texto, por ejemplo, puede utilizarse un análisis morfológico conocido. Aquí, la unidad de extracción de palabras 11 puede extraer morfemas de todas las partes de la oración divididas por el

análisis morfológico como palabras, o puede extraer sólo morfemas de partes específicas de la oración como palabras.

Tenga en cuenta que m textos pueden incluir una pluralidad de las mismas palabras. En este caso, la unidad de extracción de palabras 11 no extrae una pluralidad de palabras iguales, y extrae una sola palabra. Es decir, n palabras extraídas por la unidad de extracción de palabras 11 se refieren a n tipos de palabras.

La unidad de cálculo vectorial 12 calcula m vectores de texto y n vectores de palabra a partir de m textos y n palabras. Aquí, la unidad de cálculo de vector de texto 12A convierte cada uno de los m textos diana de análisis por la unidad de extracción de palabras 11 en un vector q-dimensional de acuerdo con una regla predeterminada, calculando así m vectores de texto que incluyen componentes del eje q (q es un número entero arbitrario de 2 o más). Además, la unidad de cálculo del vector de palabra 12B convierte cada una de las n palabras extraídas por la unidad de extracción de palabra 11 en un vector q-dimensional de acuerdo con una regla predeterminada, calculando así n vectores de palabras que incluyen componentes del eje q.

En la presente realización, a modo de ejemplo, un vector de texto y un vector de palabra se calculan como sigue. Ahora se considera un conjunto S = <d ∈ D, w ∈ W> que incluye los m textos y las n palabras. Aquí, un vector de texto $d_i$—> y un vector de palabra $w_j$—> (en lo sucesivo, el símbolo ">" indica un vector) están asociados con cada texto $d_i$ (i = 1, 2, ..., m) y cada palabras $w_j$ (j = 1, 2, ..., n), respectivamente. A continuación, se calcula una probabilidad P ($w_j$ | $d_i$) que se muestra en la siguiente ecuación (1) con respecto a una palabra arbitraria $w_j$ y un texto arbitrario $d_i$.

[Ecuación 1]

$$P(w_j \mid d_i) = \frac{\exp(\vec{w}_j \cdot \vec{d}_i)}{\sum_{k=1}^{n} \exp(\vec{w}_k \cdot \vec{d}_i)} \quad \cdots (1)$$

Obsérvese que la probabilidad P($w_j$ | $d_i$) es un valor que puede calcularse de acuerdo con una probabilidad p divulgada en el Documento sin Patente 1 descrito anteriormente. El Documento no Patentado 1 establece que, por ejemplo, cuando hay tres palabras "the", "cat" y "sat", se predice "on" como cuarta palabra, y se describe una fórmula de cálculo de la probabilidad de predicción p. La probabilidad p(wt | wt - k, ..., wt + k) descrita en el Documento no Patentado 1 es una probabilidad de respuesta correcta cuando se predice otra palabra wt a partir de una pluralidad de palabras wt - k, ..., wt + k.

Mientras tanto, la probabilidad P ($w_j$ | $d_i$) mostrada en la Ecuación (1) utilizada en la presente realización representa una probabilidad de respuesta correcta de que una palabra $w_j$ de n palabras sea predicha a partir de un texto $d_i$ de m textos. Predecir una palabra $w_j$ a partir de un texto $d_i$ significa que, concretamente, cuando aparece un determinado texto $d_i$, se predice la posibilidad de incluir la palabra $w_j$ en el texto $d_i$.

Nótese que dado que la ecuación (1) es simétrica con respecto a $d_i$ y $w_j$, se puede calcular la probabilidad P($d_i$|$w_j$) de que un texto $d_i$ de m textos se prediga a partir de una palabra $w_j$ de n palabras. Predecir un texto $d_i$ a partir de una palabra $w_j$ significa que, cuando aparece una determinada palabra $w_j$, se predice la posibilidad de incluir la palabra $w_j$ en el texto $d_i$.

En la ecuación (1), se utiliza un valor de función exponencial, donde e es la base y el producto interno del vector de palabra w-> y el vector de texto d-> es el exponente. A continuación, se calcula una relación entre un valor de función exponencial calculado a partir de una combinación de un texto $d_i$ y una palabra $w_j$ que se desea predecir y la suma de n valores de función exponencial calculados a partir de cada combinación del texto $d_i$ y n palabras $w_k$ (k = 1, 2, ..., n) como probabilidad de respuesta correcta que se espera de una palabra $w_j$ a partir de un texto $d_i$.

Aquí, el valor del producto interno del vector de palabra $w_j$-> y el vector de texto $d_i$-> puede considerarse como un valor escalar cuando el vector de palabra $w_j$-> se proyecta en una dirección del vector de texto $d_i$->, es decir, un valor del componente en la dirección del vector de texto $d_i$-> incluido en el vector de palabra $w_j$->, que puede considerarse que representa un grado en el que la palabra $w_j$ contribuye al texto $d_i$. Por lo tanto, la obtención de la relación entre el valor de la función exponencial calculado para una palabra $w_j$ y la suma de los valores de la función exponencial calculados para n palabras $w_k$ (k = 1, 2, n) utilizando el valor de la función exponencial calculado utilizando el producto interno corresponde a la obtención de la probabilidad de respuesta correcta de que se prediga una palabra $w_j$ de n palabras a partir de un texto $d_i$.

Observe que aquí se ha descrito un ejemplo de cálculo que utiliza el valor de la función exponencial utilizando como exponente el valor del producto interno del vector de palabra w-> y el vector de texto d->. Sin embargo, no se puede utilizar el valor de la función exponencial. Puede utilizarse cualquier fórmula de cálculo que utilice el valor del producto interno del vector de palabra w-> y el vector de texto d->. Por ejemplo, la probabilidad puede obtenerse a partir de la relación de los valores del producto interno.

A continuación, la unidad de cálculo vectorial 12 calcula el vector de texto $d_i$-> y el vector de palabra $w_j$-> que maximizan un valor L de la suma de la probabilidad $P(w_j | d_i)$ calculada por la Ecuación (1) para todo el conjunto S como se muestra en la siguiente Ecuación (2). Es decir, la unidad de cálculo del vector de texto 12A y la unidad de cálculo del vector de palabra 12B calculan la probabilidad $P(w_j | d_i)$ calculada por la Ecuación (1) para todas las combinaciones de los m textos y las n palabras, y calculan el vector de texto $d_i$-> y el vector de palabra $w_j$-que maximizan una variable diana L utilizando la suma de los mismos como variable diana L.

[Ecuación 2]

$$L = \sum_{d \in D} \sum_{w \in W} \#(w, d)\, p(w \mid d) \quad \cdots (2)$$

Maximizar el valor total L de la probabilidad $P(w_j|d_i)$ calculada para todas las combinaciones de los m textos y las n palabras corresponde a maximizar la probabilidad de respuesta correcta que predice una determinada palabra $w_j$ (j = 1, 2, ..., n) a partir de un determinado texto $d_i$ (i = 1, 2, m). Es decir, se puede considerar que la unidad de cálculo vectorial 12 calcula el vector de texto $d_i$-> y el vector de palabra $w_j$-> que maximizan la probabilidad de respuesta correcta.

Aquí, en la presente realización, como se describió anteriormente, la unidad de cálculo vectorial 12 convierte cada uno de los m textos di en un vector q-dimensional para calcular los m vectores de textos $d_i$-> incluyendo los componentes del eje q, y convierte cada una de las n palabras en un vector q-dimensional para calcular los n vectores de palabras $w_j$—> incluyendo los componentes del eje q, lo que corresponde a calcular el vector de texto $d_i$-> y el vector de palabra $w_j$-> que maximizan la variable diana L haciendo variables las direcciones del eje q.

La unidad de cálculo del valor del índice 13 toma cada uno de los productos internos de los m vectores de texto $d_i$-> y los n vectores de palabras $w_j$-> calculados por la unidad de cálculo vectorial 12, por lo tanto,

calculando un valor del índice de similitud que refleje la relación entre cada uno de los m textos $d_i$ y cada una de las n palabras $w_j$. En la presente

realización, como se muestra en la siguiente Ecuación (3), la unidad de cálculo del valor del índice 13 obtiene el producto de una matriz de texto D que tiene los respectivos componentes del eje q ($d_{11}$ a $d_{mq}$) de los m vectores de texto $d_i$-> como elementos respectivos y una matriz de palabra W que tiene los respectivos componentes del eje q ($w_{11}$ a $w_{nq}$) de los n vectores de palabras $w_j$-> como elementos respectivos, calculando así una matriz del valor del índice DW que tiene m x n valores de índice de similitud como elementos. Aquí, $W^t$ es la matriz transpuesta de la matriz de palabra.

[Ecuación 3]

$$D = \begin{pmatrix} d_{11} & d_{12} & \cdots & d_{1q} \\ d_{21} & d_{22} & \cdots & d_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ d_{m1} & d_{m2} & \cdots & d_{mq} \end{pmatrix} \quad W = \begin{pmatrix} w_{11} & w_{12} & \cdots & w_{1q} \\ w_{21} & w_{22} & \cdots & w_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{m2} & \cdots & w_{nq} \end{pmatrix}$$

$$DW = D * W^t = \begin{pmatrix} dw_{11} & dw_{12} & \cdots & dw_{1n} \\ dw_{21} & dw_{22} & \cdots & dw_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ dw_{m1} & dw_{m2} & \cdots & dw_{mn} \end{pmatrix} \quad \cdots (3)$$

Cada elemento de la matriz de valores del índice DW calculada de esta manera puede indicar qué palabra contribuye a qué

texto y en qué medida. Por ejemplo, un elemento $dw_{12}$ en la primera fila y la segunda columna es un valor que indica el grado en que la palabra $w_2$ contribuye a un texto $d_i$. De este modo, cada fila de la matriz de valores índice DW puede utilizarse para evaluar la similitud de un texto, y cada columna puede utilizarse para evaluar la similitud de una palabra. Los detalles se describirán más adelante.

A continuación, se describirá un aparato de búsqueda de similitudes que utiliza el aparato de cálculo del valor del índice de similitud 10 según la presente realización configurada como se ha descrito anteriormente. La Figura 2 es un diagrama de bloques que ilustra un ejemplo de configuración funcional del aparato de búsqueda de similitudes 20 según la presente realización. Como se ilustra en la Figura 2, además del aparato de cálculo de valor del índice de similitud 10 ilustrado en la Figura 1, el aparato de búsqueda de similitud 20 de la presente realización incluye una unidad de almacenamiento de datos de texto 21 como medio de almacenamiento y una unidad de designación de clave de búsqueda 22 y una unidad de búsqueda de texto similar 23 como configuración funcional.

Cada uno de los bloques funcionales 22 a 23 puede configurarse mediante hardware, DSP y software. Por ejemplo, en el caso de estar configurado por software, cada uno de los bloques funcionales 22 a 23 incluye en realidad una CPU, una RAM, una ROM, etc. de un ordenador, y se implementa mediante el funcionamiento de un programa almacenado en un medio de grabación como una RAM, una ROM, un disco duro o una memoria semiconductora.

La unidad de almacenamiento de datos de texto 21 almacena datos de texto relacionados con m textos junto con el valor del índice de similitud calculado por el aparato de cálculo del valor del índice de similitud 10. Aquí, la unidad de almacenamiento de datos de texto 21 almacena m x n valores de índice de similitud correspondientes a los valores de los elementos respectivos de la matriz de valor de índice DW calculada mediante la Ecuación (3) y datos de texto relacionados con m textos a partir de los cuales se calcula el valor del índice de similitud.

La unidad de designación de clave de búsqueda 22 designa un texto de los m textos almacenados en la unidad de almacenamiento de datos de texto 21 como clave de búsqueda. La designación de un texto se realiza cuando un usuario que desea buscar un texto similar opera una unidad de operación (un teclado, un ratón, un panel táctil, etc.) proporcionada en el aparato de búsqueda de similitudes 20. Específicamente, se obtiene una lista de textos almacenados en la unidad de almacenamiento de datos de texto 21 y se muestra en una pantalla, y el usuario selecciona un texto deseado de la lista para designar un texto como clave de búsqueda.

Obsérvese que la unidad de designación de claves de búsqueda 22 puede no estar incluida en el aparato de búsqueda de similitudes 20. Por ejemplo, el aparato de búsqueda de similitudes 20 puede configurarse como un aparato servidor conectado a una red de comunicación como Internet, la unidad de designación de claves de búsqueda 22 puede proporcionarse en otro terminal conectado a través de la red de comunicación, y la información que indica el contenido especificado puede transmitirse desde el terminal al aparato de búsqueda de similitudes 20.

Cuando la unidad de designación de clave de búsqueda 22 designa un texto de los m textos almacenados en la unidad de almacenamiento de datos de texto 21 como clave de búsqueda, la unidad de búsqueda de texto similar 23 establece los m - 1 otros textos excepto el texto designado como objetivo de búsqueda, busca en los m - 1 otros textos un texto similar al texto designado y extrae el texto. Específicamente, la unidad de búsqueda de texto similar 23 establece n valores de índice de similitud relacionados con un texto como un grupo de valor de índice de texto relacionado con la clave de búsqueda, establece n valores de índice de similitud relacionados con cada uno de los otros m - 1 textos como un grupo de valor de índice de texto relacionado con la diana de búsqueda, y determina una similitud entre el grupo de valor de índice de texto relacionado con la clave de búsqueda y el grupo de valor de índice de texto relacionado con la diana de búsqueda. A continuación, se extrae un número predeterminado de textos de los otros m - 1 textos, como resultados de la búsqueda en orden descendente de la similitud. El número predeterminado puede ser uno o más números arbitrarios.

En este caso, el grupo de valores de índice de texto relacionado con la clave de búsqueda que incluye los n valores de índice de similitud relacionados con un texto se refiere a n valores de índice de similitud incluidos en una fila relacionada con un texto entre las filas respectivas incluidas en la matriz de valores de índice DW mostrada en la Ecuación (3). Por ejemplo, cuando un texto $d_1$ se designa como un texto, n valores de índice de similitud $dw_{11}$ a $dw_{1n}$ incluidos en la primera fila de la matriz de valores de índice DW corresponden al grupo de valores de índice de texto relacionados con la clave de búsqueda. [0043]

Además, el grupo de valores de índice de texto relacionado con la diana de búsqueda que incluye los n valores de índice de similitud relacionados con los otros textos se refiere a n valores de índice de similitud incluidos en filas relacionadas con los otros textos. Por ejemplo, cuando el texto $d_1$ se designa como un texto, n valores de índice de similitud $dw_{21}$ a $dw_{2n}$, $dw_{31}$ a $dw_{3n}$, ..., $dw_{m1}$ a $dw_{mn}$ incluidos en cada una de las filas distintas de la primera fila de la matriz de valores de índice DW corresponden al grupo de valores de índice de texto relacionados con la diana de búsqueda. Aquí, n valores de índice de similitud $dw_{21}$ a $dw_{2n}$ incluidos en una segunda fila de la matriz de valores de índice DW corresponden a un grupo de valores de índice de texto relacionado con la diana de búsqueda relacionada con otro texto $d_2$. Además, n valores de índice de similitud $dw_{m1}$ a $dw_{mn}$ incluidos en una fila m-ésima de la matriz de valores de índice DW corresponden a un grupo de valores de índice de texto relacionado con la diana de búsqueda relacionada con otro texto $d_m$.

La unidad de búsqueda de textos similares 23 calcula cada una de las similitudes entre el grupo de valores de índice de texto relacionado con la clave de búsqueda $dw_{11}$ a $dw_{1n}$ relacionado con un texto y m - 1 grupo de valores de índice de texto relacionado con la diana de búsqueda $dw_{21}$ a $dw_{2n}$, $dw_{31}$ a $dw_{3n}$, ..., $dw_{m1}$ a $dw_{mn}$ relacionado con los otros textos, y extrae un número predeterminado de textos de los otros m - 1 textos como resultados de la búsqueda en orden descendente de la similitud. En este caso, se puede utilizar una tecnología conocida para calcular la similitud. Por ejemplo, es posible aplicar un método de cálculo cualquiera de la distancia euclidiana, la distancia Mahalanobis, la distancia coseno, etc.

El aparato de búsqueda de similitudes 20 configurado como en la Figura 2 es útil para designar un texto arbitrario entre m textos para los que se han calculado previamente los valores del índice de similitud, y buscar otro texto similar al texto designado de entre los m - 1 textos restantes. Por ejemplo, el aparato de búsqueda de similitudes 20 es útil cuando se desea buscar otro documento que tenga un contenido similar al de un documento específico en una situación en la que los datos de documentos publicados anteriormente se almacenan en la unidad de almacenamiento de datos de texto 21 como m textos.

La Figura 3 es un diagrama de bloques que ilustra otro ejemplo de configuración funcional de un aparato de búsqueda de similitudes 30 que utiliza el aparato de cálculo del valor del índice de similitud 10 de la presente realización. Como se ilustra en la Figura 3, además del aparato de cálculo del valor del índice de similitud 10 ilustrado en la Figura 1, el aparato de búsqueda de similitud 30 según otro ejemplo de configuración incluye una unidad de almacenamiento de datos de texto 31 como medio de almacenamiento y una unidad de adquisición de clave de búsqueda 32 y una unidad de búsqueda de texto similar 33 como configuración funcional.

Cada uno de los bloques funcionales 32 a 33 puede configurarse mediante cualquier hardware, DSP y software. Por ejemplo, en el caso de estar configurado por software, cada uno de los bloques funcionales 32 a 33 incluye realmente una CPU, una RAM, una ROM, etc. de un ordenador, y se implementa mediante la operación de un programa almacenado en un medio de grabación como una RAM, una ROM, un disco duro o una memoria semiconductora.

La unidad de almacenamiento de datos de texto 31 almacena el valor del índice de similitud calculado por el aparato de cálculo del valor del índice de similitud 10 y una pluralidad de datos de texto. Aquí, la unidad de almacenamiento de datos de texto 31 almacena una pluralidad de valores de índice de similitud correspondientes a los valores de los elementos respectivos de la matriz de valor de índice DW calculada mediante la Ecuación (3) y datos de texto relacionados con una pluralidad de textos a partir de los cuales se calcula el valor del índice de similitud.

La unidad de adquisición de clave de búsqueda 32 adquiere datos de texto designados como clave de búsqueda. Los datos de texto adquiridos aquí son nuevos datos de texto diferentes de la pluralidad de piezas de datos de texto almacenados en la unidad de almacenamiento de datos de texto 31. La fuente de adquisición de los nuevos datos de texto es arbitraria. Además, un método de adquisición de los nuevos datos de texto es arbitrario. Por ejemplo, los datos de texto designados cuando el usuario que desea buscar un texto similar opera una unidad de operación se adquieren de un terminal externo, un servidor, un almacenamiento, etc. conectado al aparato de búsqueda de similitudes 30 a través de una red de comunicación.

Cuando la unidad de adquisición de la clave de búsqueda 32 adquiere un fragmento de datos de texto, el aparato de cálculo del índice de similitud 10 establece los datos de texto adquiridos por la unidad de adquisición de la clave de búsqueda 32 como un texto (texto de la clave de búsqueda), y establece una pluralidad de fragmentos de datos de texto almacenados en la unidad de almacenamiento de datos de texto 31 como otros m - 1 textos (textos a buscar), calculando así m x n valores del índice de similitud mediante la ecuación (3).

Los valores del índice de similitud calculados por el aparato de cálculo de valores de índice de similitud 10 se almacenan en la unidad de almacenamiento de datos de texto 31 junto con nuevos datos de texto. Es decir, se almacenan adicionalmente los nuevos datos de texto y se actualizan y almacenan los valores del índice de similitud. Obsérvese que cuando se adquieren nuevos datos de texto posteriores por la unidad de adquisición de claves de búsqueda 32, la pluralidad de piezas de datos de texto (datos de texto existentes y datos de texto añadidos) almacenados en la unidad de almacenamiento de datos de texto 31 de esta manera se utilizan como m - 1 piezas de datos de texto (donde un valor de m es uno mayor que el de una vez anterior).

Utilizando los valores de índice de similitud m x n calculados por el aparato de cálculo de valor de índice de similitud 10 y almacenados en la unidad de almacenamiento de datos de texto 31, la unidad de búsqueda de texto similar 33 busca un texto similar al texto adquirido como clave de búsqueda por la unidad de adquisición de clave de búsqueda 32 a partir de los textos existentes almacenados en la unidad de almacenamiento de datos de texto 31 y extrae el texto.

Específicamente, la unidad de búsqueda de texto similar 33 determina una similitud entre un grupo de valores de índice de texto relacionado con la clave de búsqueda que incluye n valores de índice de similitud relacionados con un texto adquirido por la unidad de adquisición de clave de búsqueda 32 y un grupo de valores

de índice de texto relacionado con la diana de búsqueda que incluye n valores de índice de similitud relacionados con otro texto existente almacenado en la unidad de almacenamiento de datos de texto 31. A continuación, se extrae un número predeterminado de textos de otros m - 1 textos almacenados en la unidad de almacenamiento de datos de texto 31 como resultados de búsqueda en orden descendente de la similitud.

Aquí, cuando un texto adquirido por la unidad de adquisición de clave de búsqueda 32 se establece en $d_1$, y otros textos existentes almacenados en la unidad de almacenamiento de datos de texto 31 se establecen en $d_2$ a $d_m$, n valores de índice de similitud $dw_{11}$ a $dw_{1n}$ incluidos en la primera fila entre las filas respectivas incluidas en la matriz de valor de índice DW computada por el aparato de cómputo de valor de índice de similitud 10 de acuerdo con la Ecuación (3) corresponden a un grupo de valor de índice de texto relacionado con la clave de búsqueda. Además, n valores de índice de similitud $dw_{21}$ a $dw_{2n}$, $dw_{31}$ a $dw_{3n}$, ..., $dw_{m1}$ a $dw_{mn}$ incluidos en cada una de la segunda fila y filas subsiguientes de la matriz de valores de índice DW corresponden a un grupo de valores de índice de texto relacionados con la diana de búsqueda.

La unidad de búsqueda de textos similares 33 calcula cada una de las similitudes entre un grupo de valores de índice de texto relacionado con la clave de búsqueda $dw_{11}$ a $dw_{1n}$ relacionado con un texto y m - 1 grupos de valores de índice de texto relacionados con el objetivo de búsqueda $dw_{21}$ a $dw_{2n}$, $dw_{31}$ a $dw_{3n}$, ..., $dw_{m1}$ a $dw_{mn}$ relacionados con otros textos, y extrae un número predeterminado de textos de otros m - 1 textos como resultados de búsqueda en orden descendente de similitud.

El aparato de búsqueda de similitudes 30 configurado como en la Figura 3 es útil para buscar un texto similar a un nuevo texto adquirido como clave de búsqueda a partir de m - 1 textos para los que se han calculado previamente valores de índice de similitud. Por ejemplo, el aparato de búsqueda de similitudes 30 es útil cuando se desea buscar un artículo que tenga un contenido similar al de un artículo recién adquirido en una situación en la que los datos de artículos publicados anteriormente se almacenan en la unidad de almacenamiento de datos de texto 31 como m - 1 textos.

Obsérvese que en la realización de la Figura 2, se ha descrito una configuración en la que el aparato de búsqueda de similitudes 20 incluye el aparato de cálculo de valores de índice de similitud 10 y la unidad de almacenamiento de datos de texto 21. Sin embargo, la invención no se limita a ello. Es decir, el aparato de cálculo del valor de índice de similitud 10 y la unidad de almacenamiento de datos de texto 21 pueden configurarse como aparatos diferentes de un aparato de búsqueda de similitud que tenga la unidad de designación de clave de búsqueda 22 y la unidad de búsqueda de texto similar 23. La Figura 4 es un diagrama que ilustra un ejemplo de configuración de este caso.

Como se ilustra en la Figura 4, el aparato de cálculo del valor del índice de similitud 10 y la unidad de almacenamiento de datos de texto 21 están incluidos en un aparato servidor 100 conectado a una red de comunicación como Internet. El aparato servidor 100 incluye además una unidad de comunicación 101 y una unidad de provisión de datos 102, lee datos de texto y un valor de índice de similitud de la unidad de almacenamiento de datos de texto 21, y proporciona los datos de texto leídos y el valor de índice de similitud al aparato de búsqueda de similitud 40 en respuesta a una solicitud de adquisición de datos del aparato de búsqueda de similitud 40 conectado a la red de comunicación.

Además de la unidad de designación de claves de búsqueda 22 y la unidad de búsqueda de textos similares 23, el aparato de búsqueda de similitudes 40 incluye además una unidad de comunicación 41 y una unidad de adquisición de datos 42. La unidad de adquisición de datos 42 adquiere datos de texto y un valor de índice de similitud de la unidad de almacenamiento de datos de texto 21 del aparato servidor 100 mediante la transmisión de una solicitud de adquisición de datos al aparato servidor 100 a través de la unidad de comunicación 41. El valor del índice de similitud almacenado en la unidad de almacenamiento de datos de texto 21 es calculado por el aparato de cálculo del valor del índice de similitud 10 y almacenado por adelantado.

La unidad de adquisición de datos 42 adquiere, como grupo de valores de índice de texto relacionados con la clave de búsqueda, n valores de índice de similitud relacionados con un documento designado como clave de búsqueda por la unidad de designación de clave de búsqueda 22 y adquiere n valores de índice de similitud relacionados con cada uno de los otros m - 1 documentos como grupo de valores de índice de texto relacionados con la diana de búsqueda. Obsérvese que, por ejemplo, la designación de la clave de búsqueda por parte de la unidad de designación de claves de búsqueda 22 se realiza accediendo al aparato servidor 100 desde el aparato de búsqueda de similitudes 40 para adquirir una lista de textos almacenados en la unidad de almacenamiento de datos de texto 21, mostrando la lista en una pantalla y seleccionando un texto deseado de la lista por parte del usuario.

Cuando un texto es designado como clave de búsqueda por la unidad de designación de clave de búsqueda 22 de entre m textos almacenados en la unidad de almacenamiento de datos de texto 21 como se ha descrito anteriormente, la unidad de búsqueda de texto similar 23 determina una similitud entre un grupo de valores de índice de texto relacionado con la clave de búsqueda que incluye n valores de índice de similitud relacionados con un texto y un grupo de valores de índice de texto relacionado con la diana de búsqueda que incluye n valores de índice de similitud relacionados con cada uno de los otros m - 1 textos utilizando los valores de

índice de similitud adquiridos por la unidad de adquisición de datos 42 del aparato servidor 100, y extrae un número predeterminado de textos de los otros m - 1 textos como resultados de búsqueda en orden descendente de la similitud.

Además, en las realizaciones descritas anteriormente, se ha descrito un ejemplo en el que cada fila de la matriz de valores de índice DW calculada por el aparato de cálculo de valores de índice de similitud 10 se utiliza como una unidad, y n valores de índice de similitud se utilizan como un grupo de valores de índice de texto para buscar un texto similar. Sin embargo, la invención no se limita a ello. Por ejemplo, cada columna de la matriz de valores de índice DW calculada por el aparato de cálculo de valores de índice de similitud 10 puede utilizarse como una unidad, y m valores de índice de similitud pueden utilizarse como un grupo de valores de índice de palabra para buscar una palabra similar.

La Figura 5 es un diagrama de bloques que ilustra un ejemplo de configuración funcional de un aparato de búsqueda de similitudes 50 configurado para buscar una palabra similar. En la Figura 5, los componentes denotados por los mismos números de referencia que los ilustrados en la Figura 2 tienen las mismas funciones, por lo que aquí se omitirá una descripción redundante. Como se ilustra en la Figura 5, además del aparato de cálculo de valor del índice de similitud 10 ilustrado en la Figura 1, el aparato de búsqueda de similitud 50 incluye una unidad de almacenamiento de datos de texto 21 como medio de almacenamiento y una unidad de designación de clave de búsqueda 52 y una unidad de búsqueda de palabra similar 53 como configuración funcional.

Cada uno de los bloques funcionales 52 a 53 puede configurarse mediante cualquier hardware, DSP y software. Por ejemplo, en el caso de estar configurado por software, cada uno de los bloques funcionales 52 a 53 incluye realmente una CPU, una RAM, una ROM, etc. de un ordenador, y se implementa mediante la operación de un programa almacenado en un medio de grabación como una RAM, una ROM, un disco duro o una memoria semiconductora.

La unidad de designación de clave de búsqueda 52 designa una palabra como clave de búsqueda de entre n palabras incluidas en los datos de texto almacenados en la unidad de almacenamiento de datos de texto 21. La designación de una palabra se realiza cuando el usuario que desea buscar una palabra similar acciona una unidad de operación provista en el aparato de búsqueda de similitudes 50. Específicamente, una lista de palabras incluidas en un texto almacenado en la unidad de almacenamiento de datos de texto 21 se adquiere y se muestra en una pantalla, y una palabra deseada por el usuario se selecciona de la lista, designando así una palabra como clave de búsqueda. Tenga en cuenta que para mostrar una lista de palabras de esta manera, n piezas de datos de palabras pueden almacenarse en la unidad de almacenamiento de datos de texto 21 por separado de m piezas de datos de texto.

Obsérvese que la unidad de designación de claves de búsqueda 52 puede no estar incluida en el aparato de búsqueda de similitudes 50. Por ejemplo, el aparato de búsqueda de similitudes 50 puede configurarse como un aparato servidor conectado a una red de comunicación como Internet, la unidad de designación de clave de búsqueda 52 puede proporcionarse en otro terminal conectado a través de la red de comunicación, y la información que indica el contenido de la designación puede transmitirse desde el terminal al aparato de búsqueda de similitudes 50.

Cuando una de las n palabras es designada como clave de búsqueda por la unidad de designación de clave de búsqueda 52, la unidad de búsqueda de palabras similares 53 establece las otras n - 1 palabras excepto la palabra única como diana de búsqueda, busca una palabra similar a la palabra única entre las otras n - 1 palabras, y extrae la palabra. Específicamente, la unidad de búsqueda de palabras similares 53 establece m valores de índice de similitud relacionados con una palabra como un grupo de valor de índice de palabra relacionado con la clave de búsqueda, establece m valores de índice de similitud relacionados con cada una de las otras n-1 palabras como un grupo de valor de índice de palabra relacionado con la diana de búsqueda, y determina una similitud entre el grupo de valor de índice de palabra relacionado con la clave de búsqueda y el grupo de valor de índice de palabra relacionado con la diana de búsqueda. A continuación, se extrae un número predeterminado de palabras de las otras n-1 palabras como resultados de la búsqueda en orden descendente de la similitud.

El aparato de búsqueda de similitudes 50 configurado como en la Figura 5 es útil para designar una palabra arbitraria de entre n palabras incluidas en m textos para los que se han calculado previamente valores de índice de similitud, y buscar otra palabra similar a la palabra designada de entre las n-1 palabras restantes. La palabra similar aquí mencionada puede corresponder a un equivalente o a un sinónimo de la palabra de la clave de búsqueda, o puede no corresponder a ella. Según la presente realización, es posible buscar una palabra que tenga una tendencia similar a ser utilizada en un texto como una palabra similar.

Además, la realización es meramente un ejemplo de una realización específica para llevar a cabo la invención, y el alcance técnico de la invención no debe interpretarse de forma limitada. Es decir, la invención puede llevarse a la práctica de diversas formas sin apartarse de la esencia o de las características principales de la misma.

**Lista de signos de referencia**

10 Aparato de cálculo del valor del índice de similitud

11 Unidad de extracción de palabras

12 Unidad de cálculo vectorial

12A Unidad de cálculo de vector de texto

12B Unidad de cálculo de vector de palabra

13 Unidad de cálculo del valor del índice

20, 30, 40, 50 Aparato de búsqueda de similitudes

21, 31 Unidad de almacenamiento de datos de texto

22, 52 Unidad de designación de teclas de búsqueda

23, 33 Unidad de búsqueda de texto similar

32, Unidad de adquisición de claves de búsqueda

42, Unidad de adquisición de datos

## REIVINDICACIONES

1. Un aparato de búsqueda de similitud (20,30) que comprende un aparato de cálculo del valor del índice de similitud (10) y una unidad de búsqueda de texto similar (23,33), el aparato de cálculo del valor del índice de similitud (10) que comprende:

una unidad de extracción de palabras (11) que analiza m - cuando m>=2 -

textos y extractos n - donde n>=2 -

palabras de los textos m;

una unidad de cálculo de vector de texto (12A) que convierte cada uno de los m textos en un vector de dimensión q - donde q>=2 - según una regla predeterminada, calculando así m vectores de texto que incluyen los componentes de eje q;

una unidad de cálculo de vector de palabra (12B) que convierte cada una de las n palabras en un vector de dimensión q de acuerdo con una regla predeterminada, calculando así n vectores de palabras que incluyen los componentes del eje q; caracterizados por

una unidad de cálculo del valor del índice (13) que toma un producto de una matriz de texto que tiene componentes del eje q de cada uno de los m vectores de texto como elementos respectivos y la transposición de una matriz de palabras que tiene componentes del eje q

de cada uno de los n vectores de palabras como elementos respectivos, calculando así una matriz de valor de índice que tiene m x n valores de índice de similitud que reflejan una relación entre los m textos y las n palabras como elementos respectivos, en la que

la unidad de búsqueda de textos similares (23,33) establece otros m - 1 textos excepto uno de los m textos como dianas de búsqueda cuando el único texto se designa como clave de búsqueda entre los m textos,

determina una similitud entre un grupo de valores de índice de texto relacionado con la clave de búsqueda que incluye n valores de índice de similitud relacionados con un texto y un grupo de valores de índice de texto relacionado con la diana de búsqueda que incluye n valores de índice de similitud relacionados con cada uno de los otros m - 1 textos, y extrae un número predeterminado de textos de los otros m - 1 textos como resultados de búsqueda en orden descendente de la similitud.

2. El aparato de búsqueda de similitud según la reivindicación 1, caracterizado porque la unidad de cálculo de vector de texto (12A) y la unidad de cálculo de vector de palabra (12B) calculan una probabilidad de que uno de los m textos se prediga a partir de una de las n palabras o una probabilidad de que una de las n palabras se prediga a partir de uno de los m textos para todas las combinaciones de los m textos y las n palabras, establecen un valor total de los mismos como variable diana, y calculan un vector de texto y un vector de palabra maximizando la variable diana.

3. El aparato de búsqueda de similitudes según la reivindicación 1 o 2, que comprende además

una unidad de almacenamiento de datos de texto (21) que almacena los valores de índice de similitud calculados por el aparato de cálculo de valores de índice de similitud (10) y los datos de texto relacionados con los m textos, caracterizada por que la unidad de búsqueda de texto similar (23) establece los otros m - 1 textos excepto el texto único como dianas de búsqueda cuando el texto único se designa como clave de búsqueda a partir de

los m textos almacenados en la unidad de almacenamiento de datos de texto (21), determina una similitud entre un grupo de valores de índice de texto relacionado con la clave de búsqueda que incluye n valores de índice de similitud relacionados con un texto y un grupo de valores de índice de texto relacionado con la diana de búsqueda que incluye n valores de índice de similitud relacionados con cada uno de los otros m - 1 textos, y extrae un número predeterminado de textos de los otros m - 1 textos como resultados de búsqueda en orden descendente de la similitud.

4. El aparato de búsqueda de similitudes (30) según la reivindicación 1 o 2, que comprende además:

una unidad de almacenamiento de datos de texto (31) que almacena los valores de índice de similitud calculados por el aparato de cálculo de valores de índice de similitud (10) y una pluralidad de datos de texto; y

una unidad de adquisición de clave de búsqueda (32) que adquiere datos de texto designados como clave de búsqueda, caracterizada porque el aparato de cálculo del valor del índice de similitud (10) calcula los valores del índice de similitud utilizando datos de texto adquiridos por la unidad de adquisición de clave de búsqueda como un texto y datos de texto almacenados en la unidad de almacenamiento de datos de texto (31) como los otros m - 1 textos, y

la unidad de búsqueda de texto similar (33) determina una similitud entre un grupo de valores de índice de texto relacionado con la clave de búsqueda que incluye n valores de índice de similitud relacionados con el texto buscado por la unidad de adquisición de clave de búsqueda y un grupo de valores de índice de texto relacionado con la diana de búsqueda que incluye n valores de índice de similitud relacionados con cada uno de los otros m - 1 textos almacenados en la unidad de almacenamiento de datos de texto (31), y extrae un número predeterminado de textos como resultados de búsqueda de los otros m - 1 textos almacenados en la unidad de almacenamiento de datos de texto (31) en orden descendente de la similitud.

5. Un aparato de búsqueda de similitudes (40) según la reivindicación 1 o 2, que comprende además:

una unidad de adquisición de datos (42) que adquiere, de una unidad de almacenamiento de datos de texto (21) que almacena los valores de índice de similitud calculados por el aparato de cálculo de valor de índice de similitud (10) y datos de texto relacionados con los m textos, los datos de texto y los valores de índice de similitud,

en la que la unidad de búsqueda de texto similar (23) utiliza los datos adquiridos por la unidad de adquisición de datos (42) para establecer otros m - 1 textos excepto uno de los m textos como dianas de búsqueda cuando el texto se designa como clave de búsqueda de los m textos, determina una similitud entre un grupo de valores de índice de texto relacionado con la clave de búsqueda que incluye n valores de índice de similitud relacionados con el texto y un grupo de valores de índice de texto relacionado con la diana de búsqueda que incluye n valores de índice de similitud relacionados con cada uno de los otros m - 1 textos, y extrae un número predeterminado de textos como resultados de búsqueda de los otros m - 1 textos en orden descendente de la similitud.

6. El aparato de búsqueda de similitudes (50) según una cualquiera

de las reivindicaciones 1 a 5, que comprende además

una unidad de búsqueda de palabras similares (53) que establece otras n - 1 palabras distintas de una de las n palabras como diana de búsqueda cuando se designa una palabra como clave de búsqueda de las n palabras, determina una similitud entre un grupo de valores de índice de palabra relacionados con la clave de búsqueda que incluye m valores de índice de similitud relacionados con la palabra y un grupo de valores de índice de palabra relacionados con la diana de búsqueda que incluye m valores de índice de similitud relacionados con cada una de las otras n - 1 palabras, y extrae un número predeterminado de palabras como resultados de búsqueda de las otras n - 1 palabras en orden descendente de similitud.

7. Un programa de búsqueda de similitudes que hace que un ordenador funcione como:

medio de extracción de palabras que analiza m - donde m> = 2 -

textos y extrae n - donde n> = 2 -

palabras de los m textos; medio de cálculo vectorial que convierte cada uno de los m textos en un vector de dimensión q - donde q> = 2 -

según una regla predeterminada, y convierte cada una

de las n palabras en un vector de dimensión q de acuerdo con una regla predeterminada, calculando así m vectores de texto que incluyen componentes del eje q y n vectores de palabras que incluyen componentes del eje q; caracterizado por

un medio de cálculo del valor del índice que toma un producto de una matriz de texto que tiene componentes de eje q de cada uno de los m vectores de texto

como elementos respectivos y la transposición de

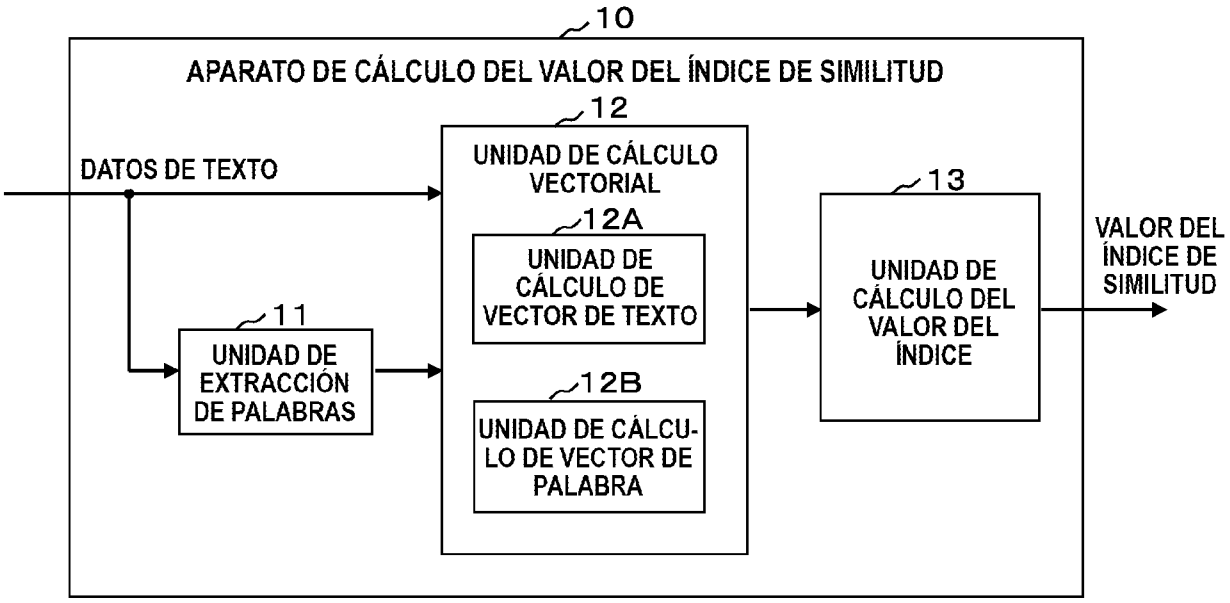
una matriz de palabra con componentes del eje q

componentes de cada uno de los n vectores de palabras como elementos respectivos, calculando así una matriz de valores de índice que tiene m x n valores índice de similitud que reflejan una relación entre los m textos y las n palabras como elementos respectivos; y

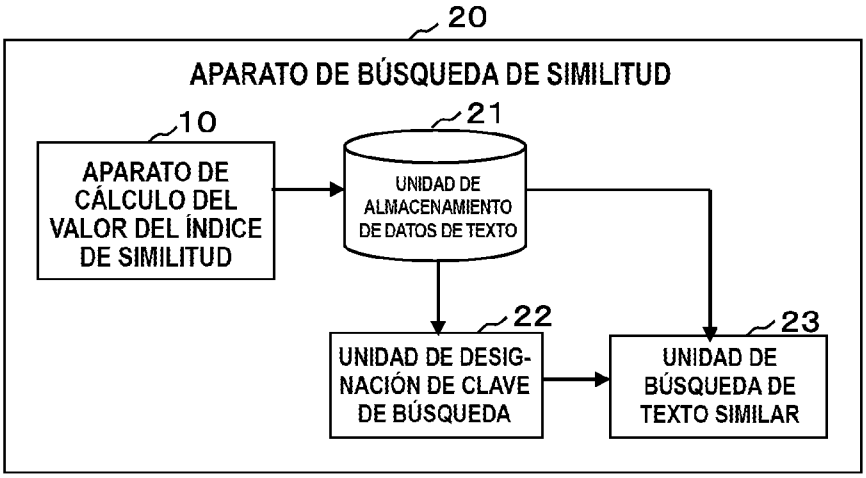
medio de búsqueda de textos similares que establecen otros m - 1 textos excepto uno de los m textos como dianas de búsqueda cuando el único texto se designa como clave de búsqueda entre los m textos, determina una similitud entre un grupo de valores de índice de texto relacionados con la clave de búsqueda que incluye n valores de índice de similitud relacionados con el único texto y un grupo de valores de índice de texto relacionados con la diana de búsqueda que incluye n valores de índice de similitud relacionados con cada uno de los otros m - 1 textos, y extrae un número predeterminado de textos de los otros m - 1 textos como resultados de búsqueda en orden descendente de la similitud.

8. El programa de búsqueda de similitud según la reivindicación 7, caracterizado porque el medio de cálculo vectorial calcula una probabilidad de que uno de los m textos se prediga a partir de una de las n palabras o una probabilidad de que una de las n palabras se prediga a partir de uno de los m textos para todas las combinaciones de los m textos y las n palabras, establece un valor total del mismo como variable diana, y calcula un vector texto y un vector palabra maximizando la variable diana.

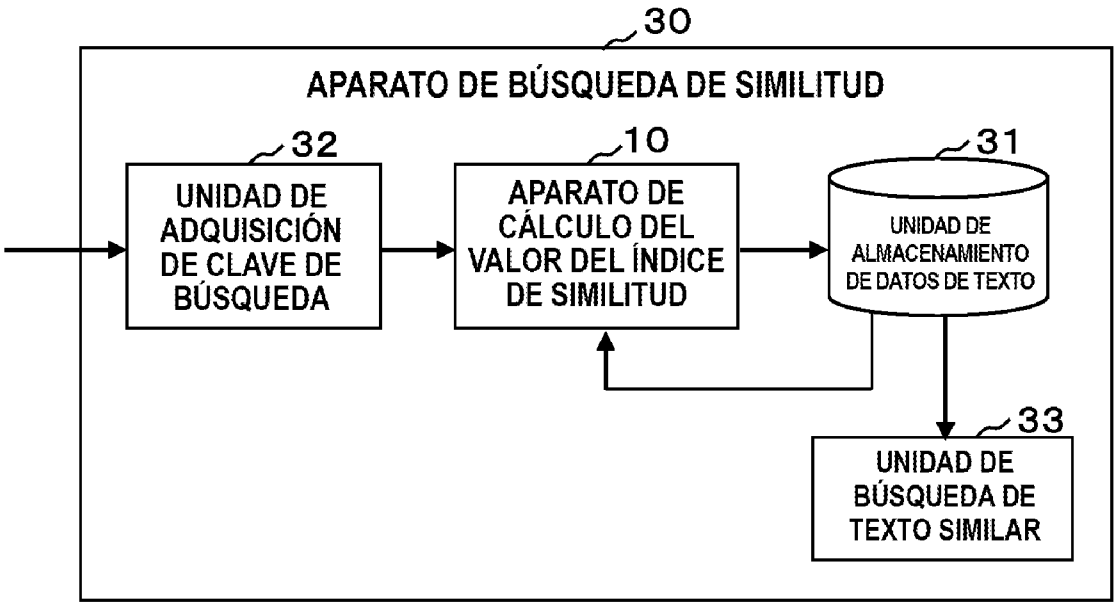
## Figura 1



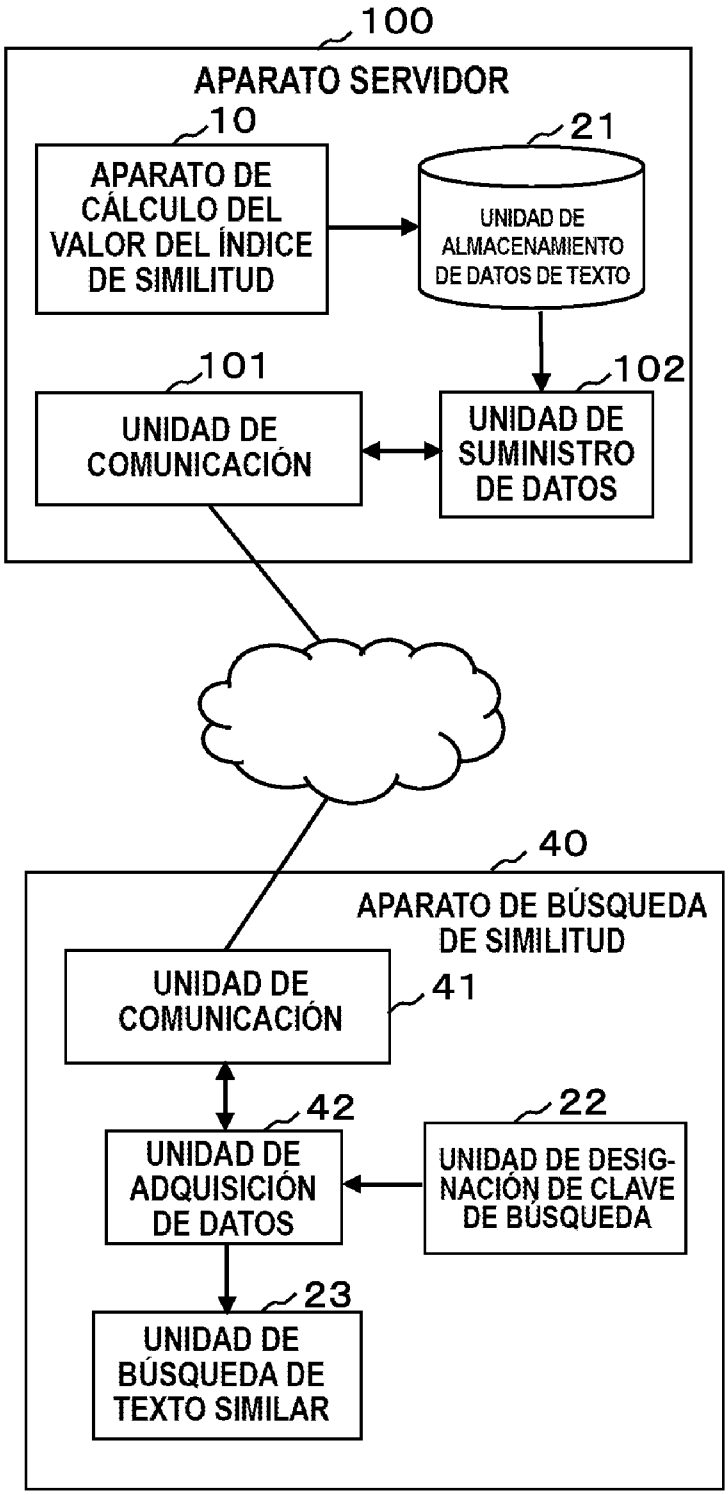

## Figura 2



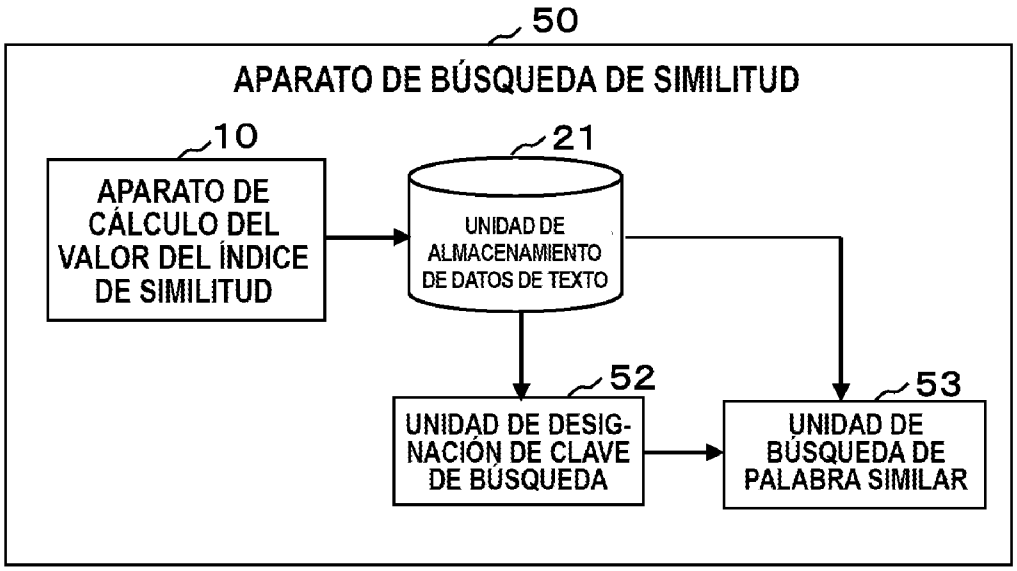
# Figura 3


30
APARATO DE BÚSQUEDA DE SIMILITUD
32
UNIDAD DE ADQUISICIÓN DE CLAVE DE BÚSQUEDA
10
APARATO DE CÁLCULO DEL VALOR DEL ÍNDICE DE SIMILITUD
31
UNIDAD DE ALMACENAMIENTO DE DATOS DE TEXTO
33
UNIDAD DE BÚSQUEDA DE TEXTO SIMILAR

# Figura 4


100
APARATO SERVIDOR
10
APARATO DE CÁLCULO DEL VALOR DEL ÍNDICE DE SIMILITUD
21
UNIDAD DE ALMACENAMIENTO DE DATOS DE TEXTO
101
UNIDAD DE COMUNICACIÓN
102
UNIDAD DE SUMINISTRO DE DATOS
40
APARATO DE BÚSQUEDA DE SIMILITUD
UNIDAD DE COMUNICACIÓN
41
42
UNIDAD DE ADQUISICIÓN DE DATOS
22
UNIDAD DE DESIGNACIÓN DE CLAVE DE BÚSQUEDA
23
UNIDAD DE BÚSQUEDA DE TEXTO SIMILAR

# Figura 5


50
APARATO DE BÚSQUEDA DE SIMILITUD
10
APARATO DE CÁLCULO DEL VALOR DEL ÍNDICE DE SIMILITUD
21
UNIDAD DE ALMACENAMIENTO DE DATOS DE TEXTO
52
UNIDAD DE DESIGNACIÓN DE CLAVE DE BÚSQUEDA
53
UNIDAD DE BÚSQUEDA DE PALABRA SIMILAR