(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 775 603**

(51) Int. CI.:

G06F 16/33 (2009.01)
G06F 16/36 (2009.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **05.03.2015 PCT/EP2015/054634**

(87) Fecha y número de publicación internacional: **11.09.2015 WO15132342**

(96) Fecha de presentación y número de la solicitud europea: **05.03.2015 E 15707680 (3)**

(97) Fecha y número de publicación de la concesión europea: **04.12.2019 EP 3114597**

(54) Título: **Procedimiento de análisis de una pluralidad de mensajes, producto de programa informático y dispositivo asociados**

(30) Prioridad:

05.03.2014 FR 1400547

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
27.07.2020

(73) Titular/es:

THALES (100.0%)
Tour Carpe Diem, Place des Corolles, Esplanade Nord
92400 Courbevoie, FR

(72) Inventor/es:

GOUTTAS, CATHERINE;
DELAVALLADE, THOMAS y
CANUS, CHRISTOPHE

(74) Agente/Representante:

SALVÀ FERRER, Joan

DESCRIPCIÓN

Procedimiento de análisis de una pluralidad de mensajes, producto de programa informático y dispositivo asociados

[0001] La presente invención se refiere a un procedimiento de análisis de una pluralidad de mensajes, comprendiendo cada mensaje una pluralidad de campos distintos dos a dos, de ellos al menos un campo relativo al contenido del mensaje y un campo relativo al contexto de emisión del mensaje, incluyendo cada campo un valor, siendo procedimiento implementado mediante un ordenador.

[0002] La invención se refiere igualmente a un producto de programa informático que incluye instrucciones de software que, cuando son implementadas por un ordenador, implementan dicho procedimiento.

[0003] La invención se refiere igualmente a un dispositivo de análisis de una pluralidad de mensajes, capaz de implementar dicho procedimiento.

[0004] La invención se aplica por ejemplo al campo de la vigilancia y del análisis de los medios sociales.

[0005] Se sabe que se ha recurrido a procedimientos para el análisis de contenidos, principalmente textuales, publicados en medios sociales, por ejemplo, en el marco de la lucha contra la ciberdelincuencia. El documento EP 2 560 111 A2 describe un procedimiento, implementado por ordenador, de análisis de una pluralidad de mensajes.

[0006] Dichos procedimientos se basan generalmente en un análisis de los contenidos textuales basado en la búsqueda de palabras clave o en un análisis de las relaciones entre los autores de estos contenidos.

[0007] No obstante, estos procedimientos no producen una entera satisfacción. De hecho, no dan acceso, cuando son implementados, a un análisis fino de los contenidos publicados en los medios sociales. Por «análisis fino» se entiende por ejemplo el hecho, simultáneamente, de:

- seguir debates en el tiempo y en el espacio geográfico, y
- permitir la recogida de conocimientos en la estructura, los comportamientos y las prácticas de comunidades supervisadas.

[0008] Dichas limitaciones provienen principalmente del hecho de que los contenidos para su análisis representan cantidades de datos importantes.

[0009] Por lo tanto, un objeto de la invención es proponer un procedimiento de análisis de una pluralidad de mensajes que permita un análisis fino de una cantidad de datos importante.

[0010] Para este fin, la invención tiene por objeto un procedimiento según la reivindicación 1.

[0011] De hecho, cuando se implementa, el procedimiento extrae, de un conjunto inicial de datos para analizar, subconjuntos de datos, constituidos por los pares, cuyo tamaño es más reducido que el conjunto inicial de los mensajes.

[0012] Dicho tratamiento permite dividir los datos para analizar en datos de tamaño reducido, lo que permite un análisis más fino de los mensajes para analizar. Además, esta división permite una visualización más rápida de los datos para analizar.

[0013] Según otros aspectos ventajosos de la invención, el procedimiento es según cualquiera de las reivindicaciones 2 a 10.

[0014] Además, la invención tiene por objeto un producto de programa informático según la reivindicación 11.

[0015] Además, la invención tiene por objeto un dispositivo de análisis de una pluralidad de mensajes, según la reivindicación 12.

[0016] La invención se entenderá mejor con ayuda de la descripción que se ofrece a continuación, proporcionada únicamente a título de ejemplo no limitativo y realizada en referencia a los dibujos adjuntos en los que:

- la figura 1 es una representación esquemática de un dispositivo de análisis según la invención, antes de la implementación del procedimiento según la invención;
- la figura 2 es una representación del dispositivo de la figura 1, después de la implementación del procedimiento según la invención;
- la figura 3 es un organigrama del procedimiento de análisis según la invención;
- la figura 4 es una representación esquemática de una vista visualizada por el dispositivo de la figura 2;

- la figura 5 es un organigrama del tratamiento de una solicitud de un usuario por medio del dispositivo de la figura 2; y
- la figura 6 es una vista análoga a la de la figura 4, después del tratamiento de una solicitud de un usuario.

[0017] El dispositivo de análisis 2, representado en las figuras 1 y 2, incluye una base de datos 4, una unidad de tratamiento 6 para actuar sobre la base de datos 4 y una interfaz de interacción 7 para capturar las solicitudes relativas al contenido de la base de datos 4 y para visualizar las vistas relativas al contenido de la base de datos 4.

[0018] La base de datos 4 está adaptada para almacenar una pluralidad de tablas 8. En particular, la base de datos 4 está adaptada para almacenar una tabla principal 10.

[0019] La tabla principal 10 está adaptada para almacenar una pluralidad de mensajes 12.

[0020] Los mensajes 12 son, por ejemplo, mensajes recogidos desde los medios sociales, tales como foros, redes sociales tales como Twitter o Facebook o cualquier otra plataforma adaptada para una interacción social en línea, y la creación y la publicación en línea de contenido.

[0021] Cada mensaje 12 comprende una pluralidad de campos 14, que comprenden cada uno un valor. Los mensajes 12 incluyen los mismos tipos de campos 14, de un mensaje a otro.

[0022] Los campos 14 son distintos dos a dos. Los campos 14 de un mensaje 12 incluyen al menos un campo 16 relativo al contenido del mensaje 12 y un campo 18 relativo al contexto de emisión del mensaje 12.

[0023] Por ejemplo, los campos 16 relativos al contenido del mensaje 12 comprenden el cuerpo de un texto publicado en un medio social, o incluso el título de este texto.

[0024] Por ejemplo, los campos 18 relativos al contexto de emisión del mensaje 12 comprenden el autor de un contenido publicado en un medio social, la plataforma o la fuente, en los que se ha publicado el contenido, la fecha de publicación del contenido, la dirección de Internet de la página en la que se ha registrado el contenido o la geolocalización del terminal electrónico a partir del cual se ha publicado el contenido.

[0025] Los campos 16, 18 de un mismo mensaje 12 comprenden valores todos los cuales se refieren a un mismo contenido publicado en línea.

[0026] La base de datos 4 está adaptada además para almacenar una pluralidad de tablas secundarias 19, y una pluralidad de tablas de asociación 20, como se ilustra en la figura 2. Al menos una tabla secundaria 19 está asociada a un campo 14 de los mensajes 12.

[0027] Las tablas secundarias 19 y las tablas de asociación 20 se describirán más adelante.

[0028] La unidad de tratamiento 6 está adaptada para actuar sobre la base de datos 4, principalmente en las tablas 10, 19, 20 de la base de datos 4.

[0029] La unidad de tratamiento 6 comprende una memoria 22 y un procesador 24.

[0030] La memoria 22 es capaz de almacenar un software 26 de lectura-escritura en la base de datos 4, un software de comparación 28, un software 30 de análisis semántico de los valores de los campos 16 relativos al contenido de un mensaje 12 y un software 32 de agrupación temática.

[0031] El software de lectura-escritura 26 está adaptado para leer los valores de los campos 14 de las tablas 10, 19, 20 almacenados en la base de datos 4. El software de lectura-escritura 26 está adaptado igualmente para modificar el valor de un campo 14, para crear un campo 14 o para crear una tabla 10, 19, 20 en la base de datos 4.

[0032] El software de comparación 28 está adaptado para establecer las relaciones entre los campos 14 de los diferentes mensajes 12.

[0033] El software de análisis semántico 30 está adaptado para analizar el contenido de los campos 16 relativos al contenido de un mensaje 12 para extraer de ellos las palabras pertinentes en relación con el contenido de un mensaje 12, denominadas asimismo «palabras clave». Por ejemplo, se consideran pertinentes las palabras de los mensajes 12 que no aparecen en un primer diccionario predeterminado de palabras no pertinentes y que aparecen en un segundo diccionario predeterminado de palabras pertinentes. Como variante, todas las palabras que no aparecen en el primer diccionario se consideran pertinentes.

[0034] De forma opcional, el software de análisis semántico 30 es capaz de normalizar previamente el contenido de los campos 16 por procedimientos usuales de tratamiento automático del lenguaje natural, por ejemplo,

procedimientos de lematización o de radicación.

[0035] Por normalización se entiende por ejemplo la transformación de los verbos conjugados en su forma en infinitivo, la transformación de los nombres en su forma en singular y la transformación de los adjetivos en su forma en masculino singular.

[0036] En el sentido de la presente invención se entiende por «semántico» el análisis de la significación de las palabras de un mensaje.

[0037] De manera general se entiende por «semántico» lo que es relativo a la significación de las palabras.

[0038] El software de agrupación temática 32 está adaptado para crear grupos de mensajes 12 semánticamente homogéneos. Por ejemplo, el software de agrupación temática está adaptado para crear grupos de mensajes según un procedimiento estadístico clásico denominado «clasificación». Un grupo semánticamente homogéneo, también denominado «tema», es un conjunto de mensajes 12 cuyos contenidos semánticos son suficientemente similares entre sí y suficientemente diferentes de los contenidos semánticos de los mensajes 12 de los otros temas en relación con un umbral de semejanza predeterminado. Dichas medidas de semejanza se basan por ejemplo en una medida de distancia entre contenidos semánticos.

[0039] Para dos vectores que representan cada uno un mensaje 12 se entiende por ejemplo por «distancia» la norma euclídea de la diferencia de los dos vectores, o el producto escalar de los dos vectores. Dichos vectores se describen más adelante.

[0040] Para dos palabras clave separadas por una distancia D, la semejanza s viene dada por ejemplo por la función:

$$s = 1 - \frac{D}{D_{\max}}$$

en la que D_{max} es la distancia máxima entre dos palabras clave.

[0041] La semejanza entre dos palabras se obtiene por ejemplo por la función:

$$s = \exp\left[-\frac{D}{\alpha}\right]$$

en la que α es el número total de palabras clave.

[0042] La identificación de los temas pasa por la optimización de una función-objetivo en la que interviene la semejanza entre contenidos semánticos.

[0043] Dicha función-objetivo es por ejemplo la función-objetivo del algoritmo de las K-medias conocido clásicamente:

$$J = \sum_{i=1}^{K} \sum_{x_j \in C_i} \left\| x_j - c_i \right\|^2$$

en la que K es el número de temas para construir, C_i es el i-ésimo tema, x_j es un vector que representa el j-ésimo mensaje 12 del tema C_i, c_i es el baricentro de los mensajes 12 que pertenecen al tema C_i y $||x_j-c_i||$ es la norma euclídea, asimismo la norma 2, del vector $x_j - c_i$. $||x_j-c_i||$ es además la distancia D entre los vectores x_j y c_i.

[0044] En general, el vector x que representa un mensaje 12 es un vector que incluye tantas componentes como palabras clave existen en un diccionario que inventaría el conjunto de las palabras clave de todos los mensajes 12, estando cada componente asociada a una única palabra clave, siendo la asociación entre la componente de un vector x y una palabra clave del diccionario la misma para todos los vectores x. Cada componente del vector x que está asociado a un mensaje 12 dado tiene como valor el peso de la palabra clave correspondiente para el mensaje 12. Diversos procedimientos permiten el cálculo del peso de una palabra clave en un mensaje 12, es decir, del valor de la componente asociada en el vector x correspondiente al mensaje 12:

- una codificación binaria en la que una componente toma el valor 1 si el mensaje contiene la palabra clave en cuestión y 0 en caso contrario; o
- una codificación de frecuencia, denominada TF (del inglés «Term Frequency»), en la que una componente tiene como valor la frecuencia normalizada de aparición de la palabra clave en el mensaje, es decir, la proporción entre la frecuencia de aparición de las palabras clave en el mensaje y el número total de palabras clave del mensaje; o
- una codificación de frecuencia ponderada, denominada TF-IDF (del inglés «Term Frequency Inverse Document Frequency»), en la que una componente tiene como valor la frecuencia normalizada de aparición de la palabra clave en el mensaje, ponderada por un factor inversamente proporcional al número de mensajes 12 en los que aparece la palabra clave.

[0045] La implementación del algoritmo de las K-medias conduce a una partición de los mensajes 12 que minimiza la función-objetivo J. La partición así obtenida conduce a una dispersión de los mensajes 12 de cada tema alrededor del baricentro del tema, garantizando así la obtención de temas homogéneos.

[0046] Otra función-objetivo posible se escribe como:

$$F(C) = \sum_{i=1}^{n} \sum_{j=1}^{n} \left(s_{ij} c_{ij} + \left(\frac{s_{ii} + s_{jj}}{2} - s_{ij} \right) (1 - c_{ij}) \right)$$

en la que n es el número total de mensajes 12 para tratar, s_{ij} es el valor de la semejanza entre dos mensajes 12 identificados por identificadores i y j, c_{ij} es el término (i,j) de la matriz de la partición que se busca construir y vale 1 si i y j pertenecen al mismo tema y 0 en caso contrario. En general, la semejanza entre dos mensajes 12 se toma como igual al número de palabras clave que dichos mensajes tienen en común, en su caso normalizado por el número de palabras clave del mensaje 12 que incluye el menor número de palabras clave. s_{ij} es el valor de la semejanza para una distancia D igual al producto escalar de los vectores x_i y x_j que son respectivamente el vector que representa el mensaje i y el vector que representa el mensaje j.

[0047] Dicha función se describe por ejemplo en el documento EP-1960916-A1, página 10 línea 11 a página 11 línea 21.

[0048] La función-objetivo F(C) presenta la ventaja de que no necesita fijar el número de temas homogéneos para construir. Además, la optimización de esta función-objetivo se realiza por ejemplo por aproximaciones lineales. La función-objetivo F(C) hace así posible el tratamiento de un número muy elevado de mensajes 12.

[0049] El software de agrupación temática 32 es capaz además de determinar una etiqueta lingüística para cada tema. La etiqueta lingüística es un resumen, comprensible por un ser humano, del contenido de cada tema. La etiqueta lingüística de un tema está constituida por las palabras clave más características del tema. La etiqueta lingüística de un tema es por ejemplo un conjunto de cardinal predeterminado, que comprende las palabras clave que presentan la puntuación de pertinencia más elevada en relación con una medida de pertinencia P predeterminada. Un ejemplo de medida de pertinencia P, inspirado en la codificación TF-IDF, viene dado por:

$$P_c(t_i) = \frac{r_c(i)}{\max_j (r_c(j))} \times \log \frac{K}{\left| \{ c' | r_{c'}(i) > 0 \} \right|}$$

en la que K es el número de temas, t_i es la i-ésima palabra clave de la cual se desea estimar la pertinencia con respecto al tema c y $r_c(i)$ es la i-ésima componente de un vector que representa el tema c, es decir, un vector en el que cada componente está asociada a una palabra clave, y cuyo valor es igual al número de mensajes 12 del tema c que contienen la palabra clave correspondiente. Así, $r_c(i)>0$ para un tema dado significa que el tema contiene al menos un mensaje que usa una palabra clave t_i.

[0050] De forma opcional, un usuario modifica la etiqueta lingüística de un tema para permitir una mejor comprensión del contenido del tema.

[0051] El software de agrupación temática 32 está adaptado además para calcular una función de semejanza de dos temas distintos y comparar el resultado de este cálculo con un valor predeterminado, llamado «umbral entre temas». La función de semejanza de dos temas distintos es por ejemplo una función cuyas variables son las palabras clave asociadas a los dos temas, y cuyo resultado es un valor numérico, llamado «semejanza entre temas», representativo de la semejanza entre los dos temas. Por ejemplo, cuanto más elevado es el valor calculado de la semejanza entre temas, más próximos están los temas semánticamente. Si el valor calculado es superior al umbral

entre temas, se dice que los temas considerados son similares.

[0052] Un ejemplo de función de semejanza entre dos temas c y c' se escribe como:

$$S(c,c') = r_c \bullet r_{c'} - \frac{1}{4}\left(|c| \times |r_{c'}| + |c'| \times |r_c|\right)$$

en la que r_c es un vector representativo del tema c, $r_{c'}$ es un vector representativo del tema c', • designa el producto escalar, |c| es el número de mensajes (12) del tema c, |c'| es el número de mensajes (12) del tema c', $|r_c|$ es la norma 1 del vector r_c y $|r_{c'}|$ es la norma 1 del vector $r_{c'}$. Por «norma 1 de un vector» se entiende la suma de los valores absolutos de las componentes del vector.

[0053] El procesador 24 es capaz de cargar y ejecutar cada uno de los programas de software 26, 28, 30, 32, 34 almacenados en la memoria 22.

[0054] La interfaz de interacción 7 comprende medios de captura 36 para permitir la captura de solicitudes relativas al contenido de la base de datos 4. La interfaz de interacción 7 comprende igualmente medios de visualización 38 para visualizar vistas relativas al contenido de la base de datos 4.

[0055] A continuación, se describirá el funcionamiento del dispositivo de análisis 2, en referencia a las figuras 2 y 3.

[0056] Durante una etapa inicial 105, la base de datos 4 almacena una tabla principal 10, tablas secundarias 19 y tablas de asociación 20. Las tablas de asociación 20 no están pobladas, es decir, no están rellenas.

[0057] La tabla principal 10 incluye una pluralidad de mensajes 12.

[0058] Los mensajes 12 almacenados en la tabla principal 10 están formados por el conjunto de los mensajes recogidos, o una parte de los mensajes recogidos, siendo los mensajes 12 de la parte escogidos individualmente por un usuario o por medio de solicitudes capturadas por medio de un motor de búsqueda.

[0059] Al menos una tabla secundaria 19 está asociada a un campo predeterminado entre los campos 18 relativos al contexto de emisión de los mensajes 12 almacenados en la tabla principal 10.

[0060] Por ejemplo, dichos mensajes 12 de la tabla principal 10 son:

id	URL	título	Texto	id autor	id fuente	fecha	lugar
1	url1.com/1	Buen tiempo	Tiempo radiante hoy en París.	b1	s1	11 de junio de 2013, 7h32	48°51'N 2°21'E
2	url.1 com/2	Paseo por la ciudad	Confirmo: «Tiempo radiante hoy en París»	b2	s1	11 de junio de 2013, 7h32	48°51'N 2°21'E
3	url1.com/3	Asombroso	Pierre, ya no entiendo el clima de París.	b3	s1	11 de junio de 2013, 7h34	43°36' 16''N 1°26'E
4	url2.com/4	Pastel	Una buena receta de Pastel: gateau.fr/gateau	b4	s2	11 de junio de 2013, 7h35	48°51'N 2°21'E
5	url2.com/5	Vacaciones	Fin de las vacaciones, estoy de regreso en París.	b5	s2	26 de junio de 2013, 9h50	48°51'N 2°21'E

[0061] Los campos «título» y «texto» son los campos 16 representativos del contenido de los mensajes 12.

[0062] Los campos «URL», «fecha» y «lugar» son campos 18 representativos del contexto de emisión de los mensajes 12. Los campos «fecha» y «lugar» se denominan también «metadatos». Cada campo «id» incluye un identificador único del mensaje 12 correspondiente.

[0063] Por «URL» (del inglés «Uniform Resource Locator») se entiende una cadena de caracteres usada para hacer referencia a un recurso web, tal como se describe en la norma RFC 3986.

[0064] Los campos «id autor» e «id fuente» incluyen respectivamente los identificadores únicos de los campos «autor» y «fuente», que son campos 18 relativos del contexto de emisión de los mensajes 12.

[0065] En lo sucesivo, un mensaje 12 cuyo campo «id» incluye un valor k se designará como «mensaje-k». Además, durante la etapa 105, para cada mensaje 12, y para cada campo 18 predeterminado, el software de lectura-escritura 26 escribe en la tabla secundaria 19 correspondiente el identificador único y el valor del campo 18 asociado, así como opcionalmente informaciones complementarias asociadas al valor del campo 18. Dichas informaciones complementarias son por ejemplo el URL asociado al valor del campo 18.

[0066] En el ejemplo, para una tabla secundaria 19 asociada a los autores, en el curso de la etapa 105, el software de lectura-escritura 26 escribe:

id autor	autor	URL autor
b1	Pierre	url1.com/b1
b2	Paul	url1.com/b2
b3	Jean	url1.com/b3
b4	Jacques	url2.com/b4
b5	Matthieu	url2.com/b5

[0067] Cada campo «id autor» incluye un identificador único del autor correspondiente.

[0068] En el ejemplo, para una tabla secundaria 19 asociada a las fuentes, en el curso de la etapa 105, el software de lectura-escritura 26 escribe:

id fuente	fuente	URL fuente
s1	Foro 1	url1.com
s2	Foro 2	url2.com

[0069] Cada campo «id fuente» incluye un identificador único de la fuente correspondiente.

[0070] Además, en el curso de la etapa 105, el software de lectura-escritura 26 lee, para cada mensaje 12, el valor de cada uno de los campos 16 relativos al contenido del mensaje 12 y el software de comparación 28 compara este valor con el valor de cada uno de los campos 14 de los otros mensajes 12.

[0071] Como variante, el software de comparación 28 compara el valor de cada uno de los campos 16 relativos al contenido de un mensaje 12 con el valor de los campos 14 predeterminados de los otros mensajes 12.

[0072] Por ejemplo, el software de comparación 28 compara el valor de los campos «título» y «texto» de un mensaje 12 únicamente con el valor de los campos «URL», «título», «texto», «autor» de los otros mensajes 12.

[0073] En el ejemplo, el software de comparación 28 compara el valor de los campos «título» y «texto» de cada mensaje 12 con el valor de los campos «URL», «título», «texto», «autor» de los otros mensajes 12.

[0074] Si, en el curso de la etapa 105, el software de comparación 28 encuentra en un campo 14, también denominado «campo citante», de un primer mensaje 12 un valor contenido en un campo 16, también denominado «campo citado», relativo al contenido de un segundo mensaje 12, entonces el software de lectura-escritura 26 escribe, en el curso de una etapa siguiente 110, en una tabla de asociación 20, un par que incluye un identificador del valor del campo citante del primer mensaje 12 y un identificador del valor del campo citado del segundo mensaje 12.

[0075] Cada entrada de la tabla de asociación 20 corresponde a un par único.

[0076] En el ejemplo, en el curso de la etapa 105, el software de comparación 28 encuentra en el campo «texto» del mensaje-2 una frase igual al valor del campo «texto» del mensaje-1.

[0077] Además, en el curso de esta misma etapa 105, el software de comparación 28 encuentra en el campo «texto» del mensaje-3 una palabra igual al valor del campo «autor» del mensaje-1.

[0078] Así, el software de lectura-escritura 26 escribe, en el curso de la etapa 110, en una tabla de asociación 20 correspondiente a los mensajes 12:

id asoc mensaje	id mensaje citante	id mensaje citado
101	2	1

[0079] Cada campo «id asoc mensaje» incluye un identificador único del par correspondiente de mensajes 12.

[0080] La tabla de asociación 20 relativa a los mensajes 12 indica la existencia de una asociación entre un primer texto de un primer mensaje 12, cuyo identificador único de la tabla principal 10 se recoge en el campo «id mensaje citante», y un segundo texto de un segundo mensaje 12, cuyo identificador único de la tabla principal 10 se recoge en el campo «id mensaje citado». En su caso, el texto del mensaje-2 cita el texto del mensaje-1.

[0081] Además, el software de lectura-escritura 26 escribe, en el curso de la etapa 110, en otra tabla de asociación 20 correspondiente al campo «autor»:

id asoc autor	id autor citante	id autor citado
201	b3	b1

[0082] Cada campo «id asoc autor» incluye un identificador único del par de autores correspondiente.

[0083] La tabla de asociación 20 relativa al campo «autor» indica la existencia de una asociación entre un primer autor (Jean), cuyo identificador único de la tabla 19 secundaria de los autores se recoge en el campo «id autor citante», y un segundo autor (Pierre), cuyo identificador único de la tabla 19 secundaria de los autores se recoge en el campo «id autor citado». En su caso, Jean (de identificador b3) cita a Pierre (de identificador b1).

[0084] Además, en el curso de la etapa 110, el software de lectura-escritura 26 escribe, en una tabla de asociación 20 correspondiente a un segundo campo distinto del campo citado, un par que incluye un identificador del valor del segundo campo de cada uno entre el primer mensaje y el segundo mensaje.

[0085] En el ejemplo, en el curso de la etapa 110, en relación con el par identificado por «id asoc mensaje = 101», el primer mensaje es el mensaje-2 y el segundo mensaje es el mensaje-1. En su caso, el autor del mensaje-2 (Paul) cita el texto publicado por el autor del mensaje-1 (Pierre).

[0086] La tabla de asociación 20 correspondiente al campo «autor» se completa entonces del modo siguiente:

id asoc autor	id autor citante	id autor citado
201	b3	b1
202	b2	b1

[0087] En el curso de una etapa siguiente 125, para cada mensaje 12, el software de lectura-escritura 26 lee el valor de los campos 16 relativos al contenido del mensaje 12, y después el software de análisis semántico 30 determina las palabras clave representativas del sentido del contenido del mensaje 12.

[0088] A continuación, el software de lectura-escritura 26 escribe, en una tabla de asociación 20 relativa a las palabras clave, un n-uplete que incluye un identificador del mensaje 12 y las palabras clave correspondientes, refiriéndose cada entrada de la tabla 20 a un único mensaje 12.

[0089] Por «n-uplete» se entiende un conjunto constituido por n elementos.

[0090] En el ejemplo, como resultado de las etapas 125, 130, la tabla de asociación 20 relativa a las palabras clave es, por ejemplo:

id asoc clave	id	palabras clave
301	1	tiempo; radiante; París; hoy
302	2	confirmar; tiempo; radiante; París; hoy
303	3	Pierre; entender; clima; París

(continuación)

id asoc clave	id	palabras clave
304	4	bueno; receta; pastel; gateau.fr/gateau
305	5	fin; vacaciones; regreso; París.

[0091] Cada campo «id asoc clave» incluye un identificador único del n-uplete que incluye un mensaje 12 y las palabras clave correspondientes.

[0092] En el curso de una etapa siguiente 135, el software de agrupación temática 32 agrupa los mensajes 12 semánticamente homogéneos en temas. Cada tema es identificado por la etiqueta lingüística correspondiente.

[0093] El software de lectura-escritura 26 escribe en una tabla secundaria 19 relativa a los temas:

id tema	etiqueta lingüística
c1	tiempo; radiante; clima
c2	París; ciudad
c3	pastel; receta

[0094] Cada valor del campo «id tema» incluye un identificador único del tema correspondiente. Cada valor del campo «etiqueta lingüística» incluye las palabras clave más características del tema correspondiente. Por «palabras clave características de un tema» se entiende las palabras clave que se emplean con mayor frecuencia en los mensajes 12 del tema a la vez que rara vez se emplean en los mensajes 12 de los otros temas.

[0095] El software de lectura-escritura 26 escribe además en una tabla de asociación 20 relativa a los temas y a los mensajes 12 que incluye:

id asoc tema	id tema	id
401	c1	1, 2, 3
402	c2	1, 2, 3, 5
403	c3	4

[0096] Como variante, el software de agrupación temática 32 es tal que un mensaje 12 pertenece a un solo y único tema.

[0097] Además, cada campo «id asoc tema» incluye un identificador único del n-uplete que comprende un tema y los mensajes 12 asociados al tema. Por ejemplo, cada campo «id tema» incluye el identificador del tema, en la tabla secundaria 19 relativa a los temas, asociado al identificador de los mensajes 12 correspondientes.

[0098] En el curso de una etapa siguiente 145, el software de agrupación temática 32 calcula, para una pluralidad de pares de temas, comprendiendo cada par un primer tema y un segundo tema, el valor de la función de semejanza entre temas para cada par de temas. A continuación, el software de cálculo 34 compara el valor obtenido con el umbral entre temas.

[0099] En el curso de una etapa siguiente 150, si, para un par de temas, el valor obtenido como resultado de la etapa 145 es superior al umbral entre temas, el software de lectura-escritura 26 escribe, en una segunda tabla de asociación 20 relativa a los temas, un par que incluye el identificador del primer tema y el identificador del segundo tema.

[0100] En el ejemplo, como resultado de las etapas 145, 150, el software de lectura-escritura 26 escribe, por ejemplo, en una segunda tabla de asociación 20 relativa a los temas:

id asoc inter	id tema 1	id tema 2
501	c2	c1

[0101] Cada campo «id asoc inter» incluye un identificador único del par que comprende el primer tema y el

segundo tema para los que el valor de la semejanza entre temas es superior al umbral entre temas.

[0102] En el curso de una etapa siguiente 155, la unidad de tratamiento 6 trata los datos contenidos en la tabla principal 10 y las tablas 19, 20 para elaborar los datos para visualizar.

[0103] A partir de los datos contenidos en la tabla principal 10 y las tablas secundarias 19 recogidas en el curso de las etapas 105 a 150, el software de lectura-escritura 26 puebla una tabla secundaria 19 de síntesis, denominada también «tabla de nodos», correspondiendo cada entrada de la tabla de nodos a un nodo, denotado por N_k.

[0104] En el sentido de la presente invención se entiende por «nodo» un mensaje 12, un tema o un valor tomado por un campo 14.

[0105] De forma opcional, los nodos correspondientes a al menos un campo 14 predeterminado de los mensajes 12, por ejemplo, los campos «autor», «fuente», incluyen metadatos relativos al mensaje 12 cuyo campo 14 toma el valor asociado al nodo considerado.

[0106] Preferentemente, cada nodo comprende una etiqueta de identificador si está asociado a un mensaje 12, a un tema o a un campo 14, y, en este caso, a qué campo 14 está asociado. Una etiqueta toma por ejemplo el valor «mensaje», «autor», «fuente» o «tema».

[0107] El conjunto de los nodos N_k forma un conjunto N.

[0108] Además, a partir de los datos contenidos en la tabla principal 10 y las tablas de asociación 20 recogidas en el curso de las etapas 105 a 150, el software de lectura-escritura 26 puebla una tabla de asociación 20 de síntesis, también denominada «tabla de arcos», correspondiendo cada entrada de la tabla de arcos a un arco.

[0109] En el sentido de la presente invención se entiende por «arco» una asociación entre un primer nodo N_k y un segundo nodo N_p, denotada por $A(N_k,N_p)$.

[0110] Dicha asociación viene dada por ejemplo por la relación entre el valor de un campo 14 citante, asociada a un primer nodo, y el valor correspondiente de un campo 14 citado, asociado a un segundo nodo. Por ejemplo, dicha asociación es asimismo la asociación entre dos temas similares. Además, dicha asociación es por ejemplo la asociación entre un mensaje 12 y los campos 14 correspondientes.

[0111] Dichas asociaciones se describen mediante la tabla principal 10 y las tablas de asociación 20 recogidas en el curso de las etapas 105 a 150.

[0112] Como variante, la tabla de nodos y la tabla de arcos se registran fuera de la base de datos 4, en un archivo almacenado en la memoria 22.

[0113] En el ejemplo, como resultado de la etapa 155, la tabla de nodos es:

id nodo	id orig	etiqueta
x1	1	mensaje
x2	2	mensaje
x3	3	mensaje
x4	4	mensaje
x5	5	mensaje
x6	b1	autor
x7	b2	autor
x8	b3	autor
x9	b4	autor
x10	b5	autor
x11	s1	fuente
x12	s2	fuente

(continuación)

id nodo	id orig	etiqueta
x13	c1	tema
x14	c2	tema
x15	c3	Tema

[0114] Cada campo «id nodo» incluye el identificador único del nodo correspondiente.

[0115] Además, la tabla de arcos es:

id arco	id nodo 1	id nodo 2
y1	x1	x6
y2	x1	x11
y3	x1	x13
y4	x1	x14
y5	x2	x7
y6	x2	x11
y7	x2	x13
y8	x2	x14
y9	x2	x1
y10	x3	x8
y11	x3	x11
y12	x3	x13
y13	x3	x14
y14	x3	x1
y15	x4	x9
y16	x4	x12
y17	x4	x15
y18	x5	x10
y19	x5	x12
y20	x5	x14
y21	x8	x6
y22	x7	x6
y23	x12	x15

[0116] En el curso de una etapa siguiente de espera 157, la unidad de tratamiento 6 espera la recepción de instrucciones de un usuario a través de la interfaz de interacción 7.

[0117] En el curso de esta misma etapa de espera 157, el dispositivo 2 visualiza, en los medios de visualización 38, una vista 200 que comprende una pluralidad de grafos 201, como se ilustra en la figura 4.

[0118] Cada grafo 201 es representativo de los diferentes valores que toma un campo 14 de los mensajes de la tabla principal 10, y de las asociaciones entre estos valores.

[0119] Por ejemplo, un primer grafo 202 es representativo de las asociaciones entre los textos de los mensajes

12. Un segundo grafo 204 es representativo de las asociaciones entre los autores de los mensajes 12. Un tercer grafo 206 es representativo de las asociaciones entre las fuentes de los mensajes 12. Un cuarto grafo 208 es representativo de las asociaciones entre los temas de los mensajes 12.

[0120] Cada grafo 201 incluye una pluralidad de nodos 210 y una pluralidad de arcos 212 que conectan cada uno dos nodos 210.

[0121] Preferentemente, los arcos 212 están orientados. Para cada arco 212 de un grafo 201 correspondiente a un campo 14 de los mensajes 12, el arco 212 está orientado desde un nodo correspondiente a un campo citante de un primer mensaje hacia un nodo correspondiente a un campo citado de un segundo mensaje.

[0122] La vista 200 incluye además un eje temporal 214, que comprende un cursor de inicio 216 y un cursor de fin 218.

[0123] El eje temporal 214 es representativo del intervalo temporal durante el cual se han emitido los mensajes 12 de la tabla principal 10.

[0124] En el ejemplo, el intervalo temporal se extiende del 11 de junio de 2013 a las 7h32 al 26 de junio de 2013 a las 9h50.

[0125] El cursor de inicio 216 y el cursor de fin 218 permiten que un usuario regule la fecha de inicio y la fecha de fin, respectivamente, de un intervalo temporal personalizado.

[0126] El intervalo temporal personalizado es tal que los nodos correspondientes a mensajes publicados fuera del intervalo temporal personalizado pueden ser filtrados fácilmente en la vista 200. Por ejemplo, los nodos 210 correspondientes a mensajes publicados fuera del intervalo temporal personalizado aparecen en un color diferente del color de los otros nodos 210.

[0127] Si un usuario selecciona uno o varios nodos en la vista 200, a través de los medios de captura 36, el dispositivo 2 visualiza, en la vista 200, únicamente los nodos para los que existe un arco con el o los nodos seleccionados.

[0128] Como variante, si un usuario selecciona uno o varios nodos en la vista 200, a través de los medios de captura 36, el dispositivo 2 visualiza, en la vista 200, únicamente los nodos que tienen una etiqueta diferente para los que existe un arco con el o los nodos seleccionados. Este proceso permite así al usuario facilitar la navegación entre grafos.

[0129] A continuación, se detallará el procedimiento relativo a la visualización, en referencia a la figura 5.

[0130] En el curso de una etapa 300, si el usuario ha seleccionado un intervalo temporal personalizado, solo se conserva el conjunto T de los nodos asociados a un mensaje 12 cuya fecha de publicación se encuentra en el intervalo temporal personalizado. El conjunto T es un subconjunto del conjunto N de los nodos.

[0131] En el curso de una etapa siguiente 305, el usuario selecciona al menos un nodo N_k.

[0132] En el curso de una etapa siguiente 310, solo se conserva un subconjunto C del conjunto de los nodos T. El conjunto C es el conjunto de los nodos del conjunto T tal que existe un arco que los asocia al nodo N_k seleccionado.

[0133] En el curso de una etapa opcional siguiente 315, solo se conserva un subconjunto S del conjunto C. El conjunto S es el conjunto de los nodos cuya etiqueta comprende un valor diferente del valor de la etiqueta del nodo N_k seleccionado.

[0134] En el curso de una etapa siguiente 320, los nodos del conjunto S se visualizan en la vista 200.

[0135] Como variante, los nodos del conjunto S se visualizan con un color diferente del color de los otros nodos. Como variante, los nodos del conjunto S se visualizan en brillo resaltado. Como variante, los nodos del conjunto S se visualizan con una forma diferente de la forma de los otros nodos. Como variante, los nodos del conjunto S se visualizan con un parpadeo cuya frecuencia es diferente de la de los otros nodos.

[0136] La elección del intervalo temporal personalizado puede realizarse igualmente como resultado de la selección del al menos un nodo N_k.

[0137] En la vista 200, dos nodos, cuyo valor de etiqueta es el mismo y que están asociados por un arco, están conectados entre sí por un segmento representativo del arco.

[0138] Se entiende así que para una pluralidad de mensajes 12 que van a analizarse almacenados en una tabla principal 10, el conjunto de los mensajes 12 para su análisis se segmenta en una pluralidad de datos registrados en las tablas de asociación 20 cuyo tamaño es inferior al tamaño de la tabla principal 10. Este hecho tiene como consecuencia una reducción de los tiempos de cálculo ligados al análisis de los mensajes 12. Este resultado se obtiene buscando en al menos un campo 16 relativo al contenido de un primer mensaje 12, el valor de un campo 14, llamado campo objeto, de un segundo mensaje 12, y después registrando, en una tabla de asociación 20, un par que incluye un identificador del valor del campo objeto del primer mensaje 12 y un identificador del valor del campo objeto del segundo mensaje 12.

[0139] Además, la memorización de un par que incluye un identificador del valor de un segundo campo, distinto del campo objeto, de cada uno de los mensajes fuente, permite establecer, con tiempos de cálculo menores, enlaces entre los valores de campos 14 distintos.

[0140] Dichos tratamientos permiten además reducir al mínimo los tiempos de cálculo y mejorar la rapidez del sistema después de las solicitudes de un usuario.

REIVINDICACIONES

1. Procedimiento de análisis de una pluralidad de mensajes (12), comprendiendo cada mensaje (12) una pluralidad de campos (14, 16, 18) distintos dos a dos, de ellos al menos un campo (16) relativo al contenido del mensaje (12) y un campo (18) relativo al contexto de emisión del mensaje, incluyendo cada campo (14, 16, 18) un valor, de manera que el procedimiento se implementa mediante un ordenador y está **caracterizado porque** incluye:

- una etapa de búsqueda (105), en al menos un campo relativo al contenido de un primer mensaje, del valor de un campo objeto (14, 16, 18) de un segundo mensaje (12);
- una etapa de memorización (110), en una memoria (20), de un par que incluye un identificador del valor del campo objeto (14, 16, 18) del primer mensaje (12) y un identificador del valor del campo objeto (14, 16, 18) del segundo mensaje (12);
- una etapa (135) de agregación de mensajes (12) por temas, comprendiendo cada tema mensajes (12) que presentan una semejanza superior a un valor predeterminado según una función de semejanza predeterminada; -
- para cada tema, una etapa de memorización (140) de un n-uplete que comprende un identificador del tema y un identificador de cada mensaje (12) correspondiente;
- una etapa (145) de cálculo de una función de semejanza de un primer tema y de un segundo tema, verificando la función de semejanza:

$$S(c,c') = r_c \bullet r_{c'} - \frac{1}{4}\left(|c| \times |r_{c'}| + |c'| \times |r_c|\right)$$

en la que c y c' son dos temas, r_c es un vector representativo del tema c, $r_{c'}$ es un vector representativo del tema c', • designa el producto escalar, |c| es el número de mensajes (12) del tema c, |c'| es el número de mensajes (12) del tema c', $|r_c|$ es la norma 1 del vector r_c y $|r_{c'}|$ es la norma 1 del vector $r_{c'}$, siendo la norma 1 de un vector la suma de los valores absolutos de las componentes del vector; y
- una etapa de memorización (150) de un par que comprende un identificador del primer tema y un identificador del segundo tema, si el resultado de la etapa de cálculo es superior a un umbral predefinido.

2. Procedimiento según la reivindicación 1, **caracterizado porque** el campo objeto (14, 16, 18) del segundo mensaje es diferente del campo (16) relativo al contenido del primer mensaje.

3. Procedimiento según la reivindicación 1 o 2, **caracterizado porque** incluye una etapa de memorización (110), en la memoria (20) de un par que incluye un identificador del valor de un segundo campo (14, 16, 18), distinto del campo objeto (14, 16, 18), del primer mensaje (12) y del segundo mensaje (12).

4. Procedimiento según cualquiera de las reivindicaciones anteriores, **caracterizado porque** la etapa (135) de agregación de mensajes (12) comprende la optimización de una función-objetivo.

5. Procedimiento según la reivindicación 4, **caracterizado porque** la etapa (135) de agregación de mensajes (12) comprende la minimización de una primera función-objetivo que verifica:

$$J = \sum_{i=1}^{K} \sum_{x_j \in C_i} \left\| x_j - c_i \right\|^2$$

en la que K es el número de temas para construir, C_i es el i-ésimo tema, x_j es un vector que representa el j-ésimo mensaje (12) del tema C_i y c_i es el baricentro de los mensajes (12) que pertenecen al tema C_i y $||x_j-c_i||$ es la norma euclídea del vector $x_j - c_i$.

6. Procedimiento según la reivindicación 4 o 5, **caracterizado porque** la etapa (135) de agregación de mensajes (12) comprende la maximización de una segunda función-objetivo que verifica:

$$F(C) = \sum_{i=1}^{N} \sum_{j=1}^{N} \left(s_{ij} c_{ij} + \left(\frac{s_{ii} + s_{jj}}{2} - s_{ij} \right) \left(1 - c_{ij}\right) \right)$$

en la que N es el número total de mensajes (12) para tratar, s_{ij} es el valor de la semejanza entre dos mensajes (12) identificados por identificadores i y j, c_{ij} es un término que vale 1 si el i-ésimo y el j-ésimo mensajes (12) pertenecen al

mismo tema y 0 en caso contrario.

7. Procedimiento según cualquiera de las reivindicaciones anteriores, **caracterizado porque** incluye:

- una etapa (125) de lectura del valor de al menos un campo (16) relativo al contenido de un mensaje (12);
- una etapa (125) de análisis semántico del o de cada valor leído y de extracción de palabras clave relativas al o a cada valor leído.

8. Procedimiento según cualquiera de las reivindicaciones anteriores, **caracterizado porque** incluye:

- la visualización de una vista (200) que incluye una pluralidad de grafos (201, 202, 204, 206, 208), estando cada grafo asociado a un elemento entre los mensajes (12), los campos (14, 16, 18) y los temas, siendo un tema un conjunto que comprende una pluralidad de mensajes (12) cuyos contenidos presentan, según una función de semejanza predeterminada, una semejanza superior a un valor predeterminado, comprendiendo cada grafo (201, 202, 204, 206, 208) al menos un nodo (210), estando cada nodo de un grafo (201, 202, 204, 206, 208) asociado a un identificador de un mensaje (12) o de un tema o de un valor del campo (14, 16, 18) correspondiente al grafo (201, 204, 206, 208);
- la selección de al menos un nodo (210) por un usuario;
- la visualización de nodos (210) acoplados al o a cada nodo (210) seleccionado y del o de los nodos (210) seleccionados de manera distintiva con respecto a los otros nodos (210), siendo los nodos (210) acoplados nodos (210) asociados a identificadores que forman cada uno con el identificador asociado a un nodo (210) seleccionado uno de los pares memorizados.

9. Producto de programa informático que incluye instrucciones de software que, cuando son implementadas por un ordenador, implementan el procedimiento según cualquiera de las reivindicaciones 1 a 8.

10. Dispositivo (2) de análisis de una pluralidad de mensajes (12), comprendiendo cada mensaje (12) una pluralidad de campos (14, 16, 18) distintos dos a dos, de ellos al menos un campo (16) relativo al contenido del mensaje y un campo (18) relativo al contexto de emisión del mensaje, incluyendo cada campo (14, 16, 18) un valor, **caracterizado porque** es capaz de implementar el procedimiento de análisis según cualquiera de las reivindicaciones 1 a 8.

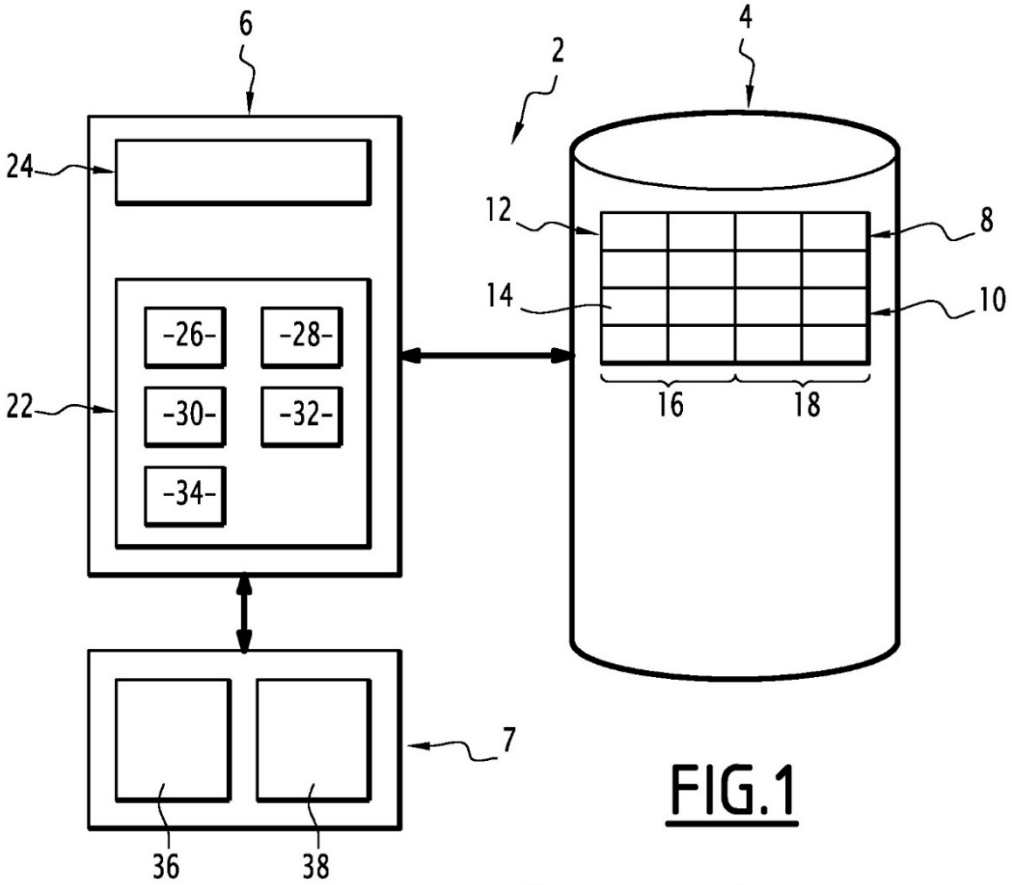
6
2
4
24
12
8
14
10
-26-
-28-
22
-30-
-32-
-34-
16
18
7
36
38

FIG.1

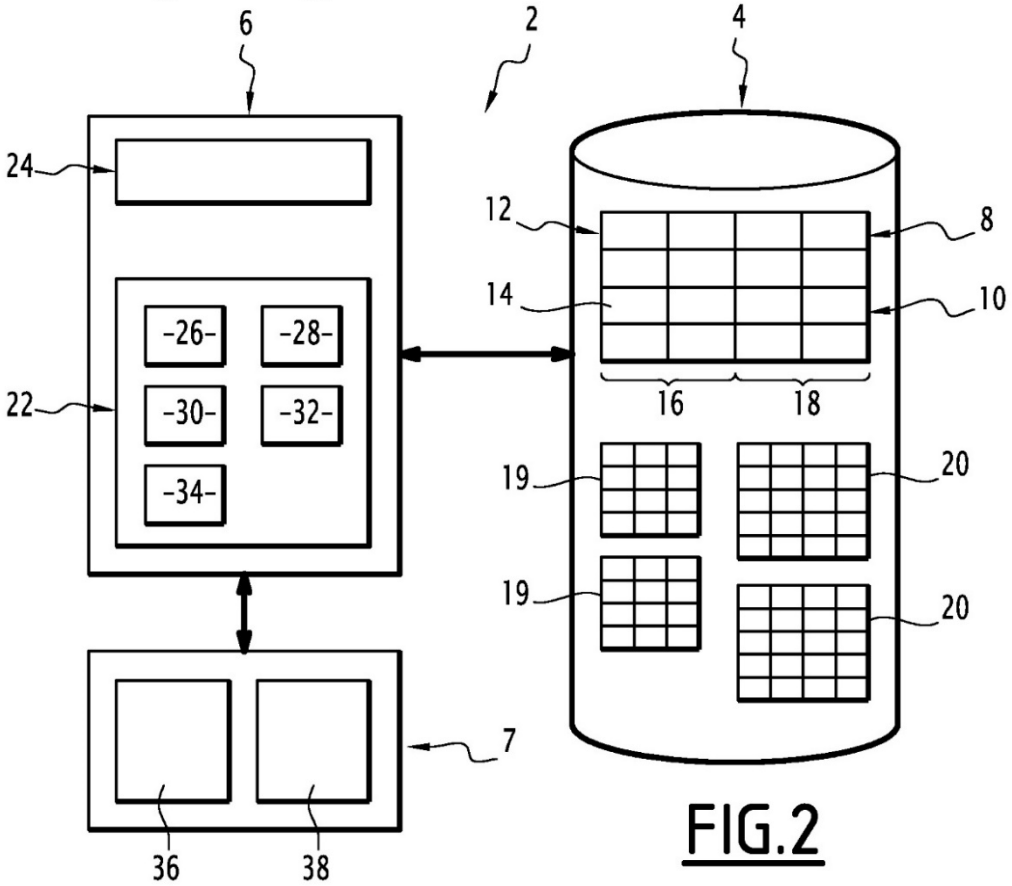
6
2
4
24
12
8
14
10
-26-
-28-
22
-30-
-32-
-34-
16
18
19
20
19
20
7
36
38

FIG.2

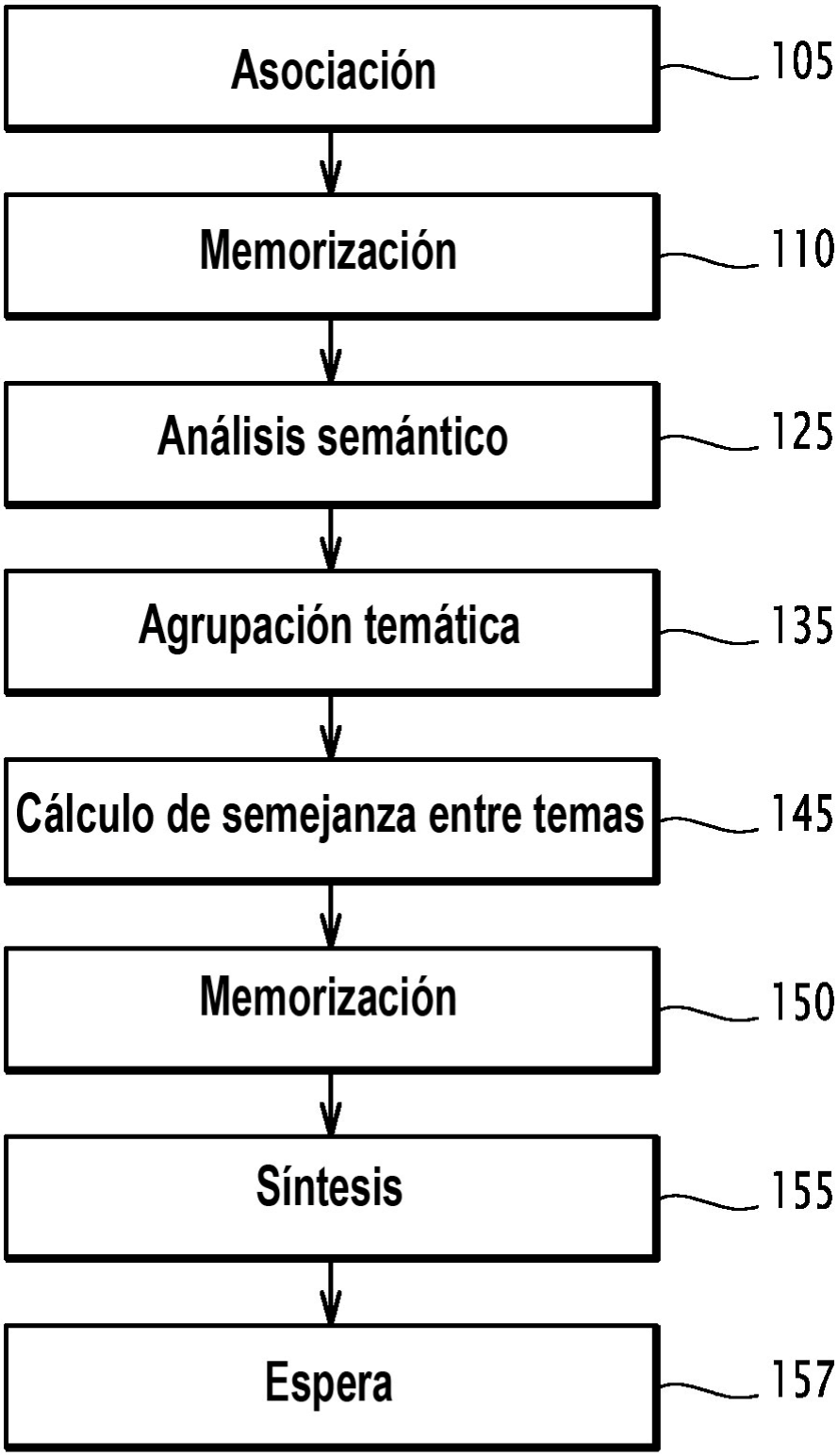
Asociación
105
Memorización
110
Análisis semántico
125
Agrupación temática
135
Cálculo de semejanza entre temas
145
Memorización
150
Síntesis
155
Espera
157

FIG.3

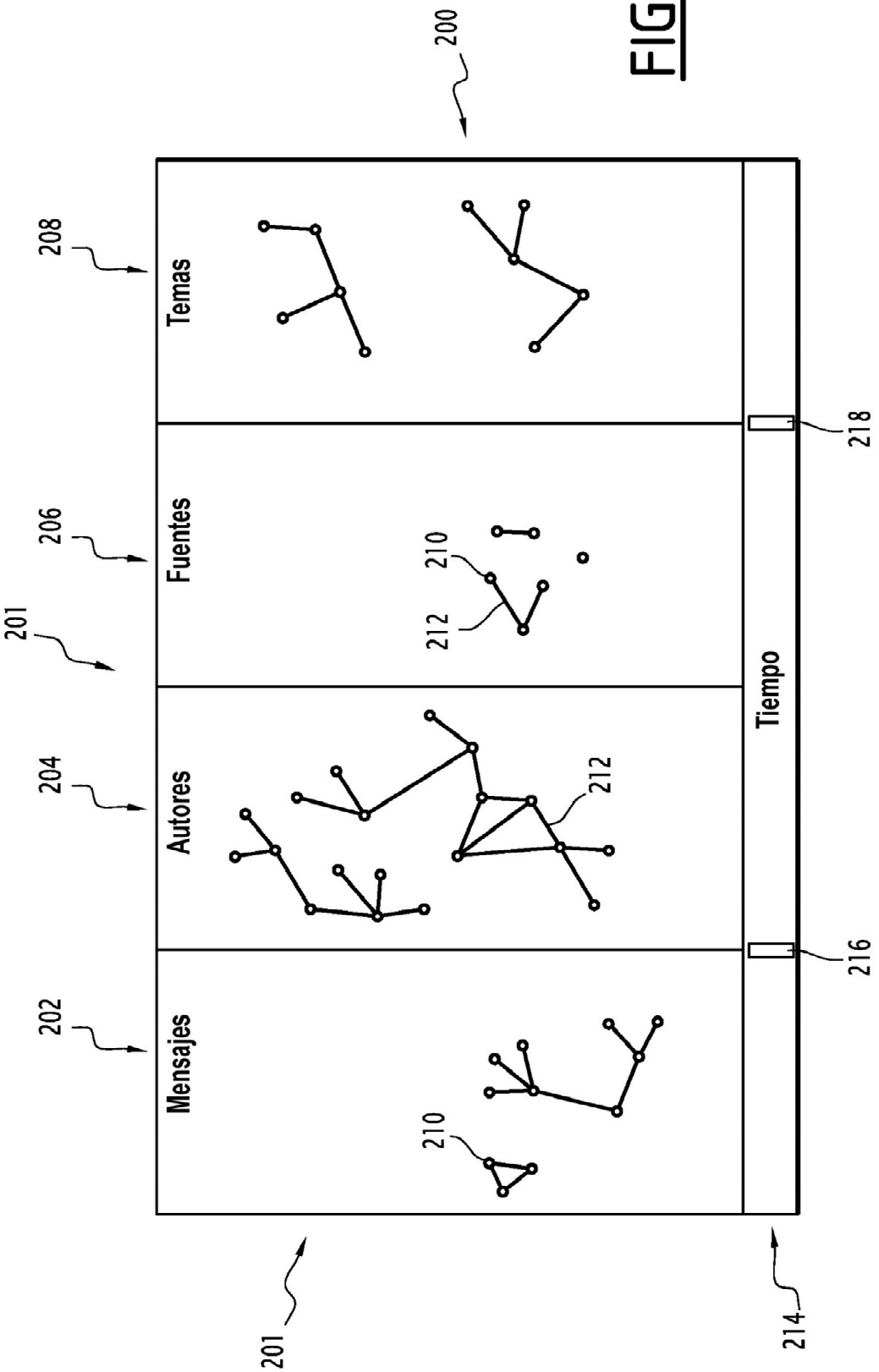
200
208
Temas
206
Fuentes
201
204
Autores
202
Mensajes
Tiempo
210
212
212
210
216
218
201
214

FIG.4

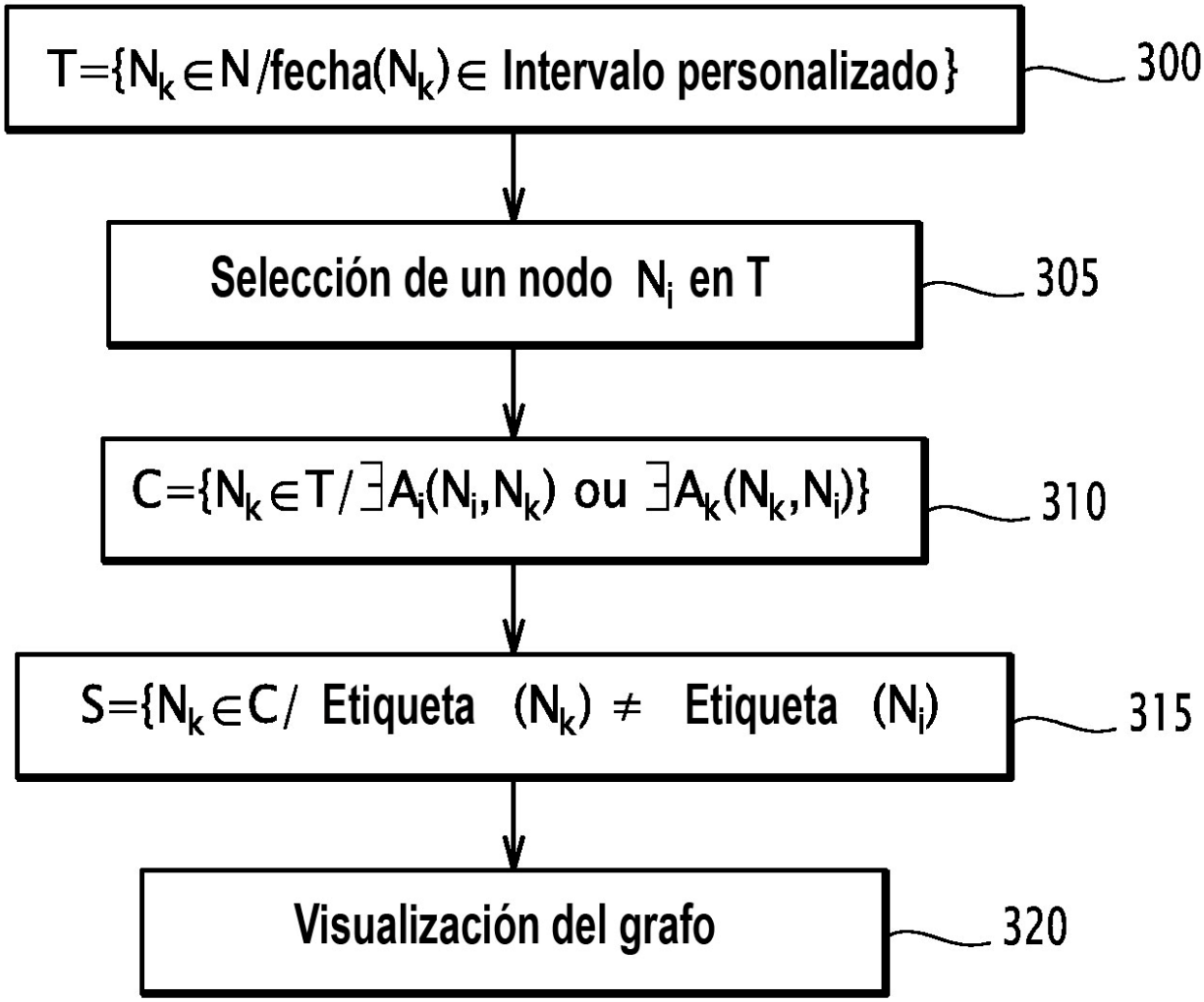
T={Nk∈N/fecha(Nk)∈ Intervalo personalizado}
300
Selección de un nodo Ni en T
305
C={Nk∈T/∃Ai(Ni,Nk) ou ∃Ak(Nk,Ni)}
310
S={Nk∈C/ Etiqueta (Nk) ≠ Etiqueta (Ni)
315
Visualización del grafo
320

FIG.5

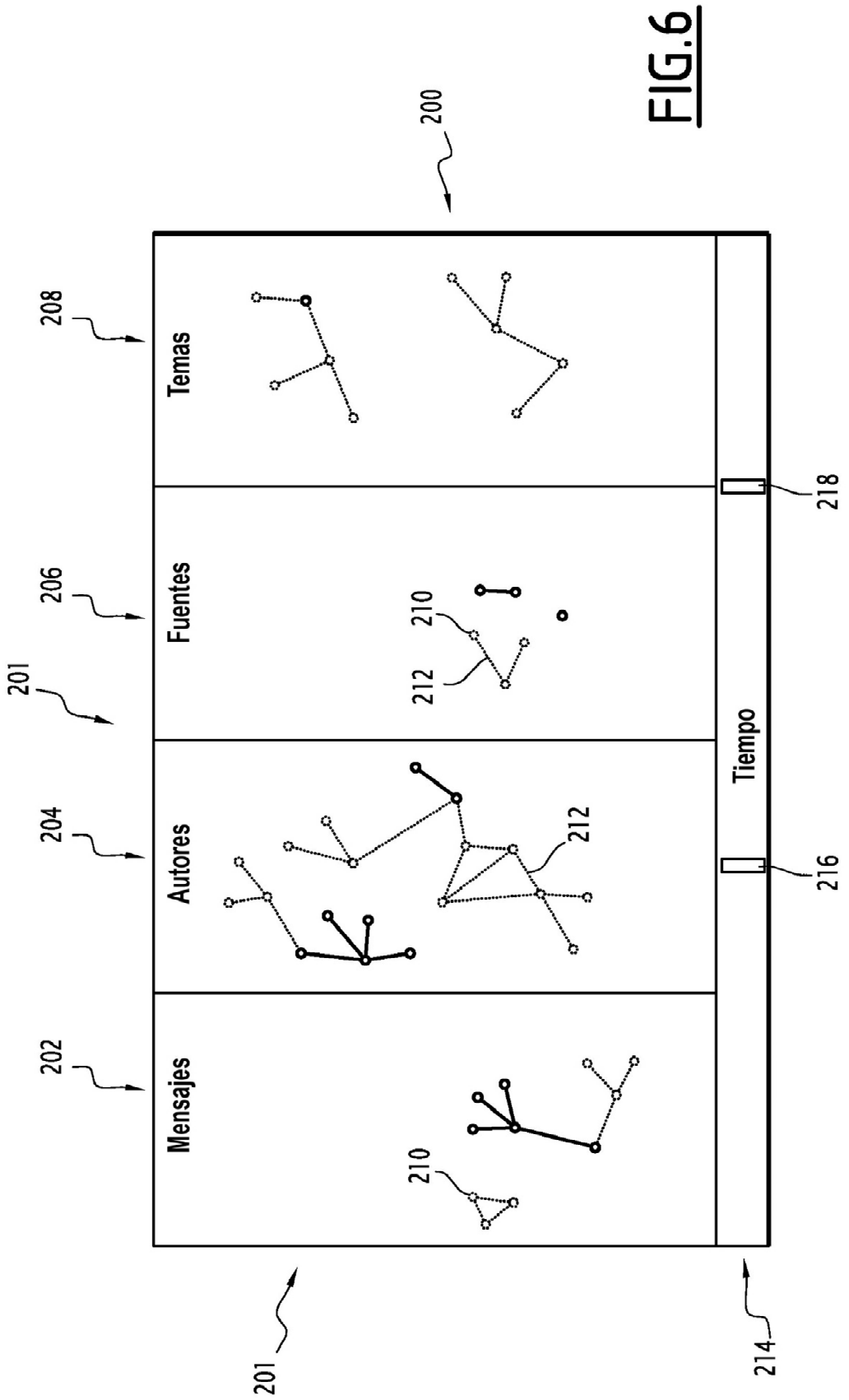
200
208
Temas
206
Fuentes
201
210
212
204
Autores
212
Tiempo
218
216
202
Mensajes
210
201
214

FIG.6