(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 471 142**

(51) Int. CI.:

G06K 9/00 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **19.12.2008 E 08866899 (1)**

(97) Fecha y número de publicación de la concesión europea: **12.03.2014 EP 2248071**

(54) Título: **Identificación basada en datos biométricos cifrados**

(30) Prioridad:

21.12.2007 FR 0760300

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
25.06.2014

(73) Titular/es:

MORPHO (100.0%)
11 Boulevard Galliéni
92130 Issy Les Moulineaux, FR

(72) Inventor/es:

KINDARJI, BRUNO;
CHABANNE, HERVÉ y
BRINGER, JULIEN

(74) Agente/Representante:

DE ELZABURU MÁRQUEZ, Alberto

DESCRIPCIÓN

Identificación basada en datos biométricos cifrados

La presente invención se refiere a un control basado en datos biométricos almacenados en una base de datos, y más particularmente, cuando estos datos están almacenados en una forma cifrada.

Los sistemas de identificación basados en datos biométricos comprenden en general una entidad de control que tiene acceso a una base de datos que contiene datos biométricos sobre la base de los cuales la entidad de control es capaz de identificar a una persona.

Los datos biométricos almacenados en esta base de datos proceden de una fase de inscripción (o 'enrollment' en inglés), en el curso de la cual una parte de un ser humano, tal como por ejemplo huellas o el iris, es captada bajo la forma de datos numéricos biométricos para ser finalmente almacenados y ser utilizados como datos de referencia para efectuar una identificación de una persona.

En efecto, cuando la entidad de control recibe nuevos datos biométricos, es entonces capaz de determinar la identificación de la persona a la que están asociados estos nuevos datos sobre la base de una comparación entre estos datos biométricos recibidos y los datos biométricos de referencia almacenados en la base de datos.

Conviene observar que es en general importante que tales datos biométricos permanezcan confidenciales, con el fin de poder proteger la vida privada de las personas a las cuales están asociados estos datos.

Así, puede preverse para ello almacenar estos datos biométricos bajo una forma cifrada en la base de datos.

No obstante, en este caso, la etapa de comparación consistente en comparar datos biométricos recibidos desde una entidad de control con datos biométricos almacenados en la base de datos puede entonces revelarse mucho más compleja que en el caso en el que estos datos biométricos son almacenados en la base de datos sin cifrado.

Se puede entonces imaginar que para hacer esta búsqueda en la base de datos, sea necesario descifrar los datos biométricos que están almacenados en ella. Pero una etapa de descifrado es susceptible de gravar enormemente la eficacia de tal búsqueda en la base de datos, así como de limitar la seguridad final.

Por otro lado, se conocen sistemas de búsqueda adaptados para efectuar búsquedas en una base de datos que comprende datos codificados.

En efecto, existen sistemas de codificación, o 'searchable encryption' en inglés, tal como el descrito en el documento de D. Boneh, G. Di Crescenzo, R. Ostrovsky y G. Persiano, Public Key Encryption with Keyword Search, EUROCRYPT 2004, que permiten hacer una búsqueda en una base de datos cifrados utilizando palabras claves que, ellos mismos, son determinados sobre la base de los mismos datos pero en su forma no cifrada. Tal sistema puede ser ventajoso particularmente para gestionar el archivado de e-mails por ejemplo. Tales búsquedas basadas en palabras claves evitan proceder a una búsqueda exhaustiva en la base de datos considerada y permiten aumentar la eficacia de la búsqueda.

Podría entonces resultar interesante utilizar tales sistemas en el contexto de la identificación basada en datos biométricos cifrados.

No obstante, la aplicación de este tipo de sistema para efectuar búsquedas por palabras claves en una base de datos biométricos cifrados puede revelarse errónea si no se procede al descifrado de los datos almacenados.

En efecto, es posible, e incluso muy probable, que los datos biométricos obtenidos en una fase de inscripción diferente substancialmente de los datos biométricos que son obtenidos para una misma persona tras una fase de identificación. En un contexto tal, es entonces posible que los resultados obtenidos mediante la aplicación de los principios de los sistemas de codificación anteriores no sean pertinentes.

En un artículo datado el 5 de Mayo de 2004 y titulado Secure Indexes (disponible en la dirección crypto.stanford.edu/~eujin/papers/secureindex/secureindex.pdf), Eu-Jin Goh describe un índice de seguridad y formula un modelo de seguridad para los índices, conocido bajo el nombre de seguridad semántica frente a un ataque adaptativo de palabra clave elegida (cuyo acrónimo es ind-cka). El autor describe una construcción de índice llamada z-idx. El autor la presenta como una construcción eficaz y de seguridad contra ind-cka. Esta construcción utiliza funciones pseudoaleatorias y filtros de Bloom. El artículo no se refiere a la biometría

El documento Public Key encryption that allows PIR queries (Dan Boneh et al. 2007) divulga un sistema de búsqueda adaptado para efectuar búsquedas en una base de datos que comprende datos codificados.

La presente invención se dirige a mejorar la situación.

Un primer aspecto de la presente invención propone un procedimiento de gestión de una base de datos que comprende datos biométricos almacenados en forma cifrada;

comprendiendo la citada base de datos un conjunto de filtros de indexación asociados respectivamente a identificadores de filtro;

el citado procedimiento comprende las etapas siguientes al nivel de una entidad de gestión:

/1/ recibir un dato biométrico;

/2/ almacenar, en una dirección dada en la base de datos, el citado dato biométrico bajo una forma codificada;

/3/ obtener palabras claves a partir de un primer conjunto de funciones de troceado y del citado dato biométrico, teniendo cada una de las funciones de troceado del primer conjunto por características proporcionar valores de entrada vecinos en un espacio métrico, valores muy vecinos, preferiblemente un mismo valor, a la salida en un segundo espacio métrico;

/4/ asociar un subconjunto de filtros a cada palabra clave seleccionando, para cada palabra clave, filtros en función de los identificadores de filtro respectivamente asociados, citadas palabras clave y un segundo conjunto de funciones de troceado; y

/5/ asociar la citada dirección dada en forma codificada a cada uno de los filtros del subconjunto de filtros.

Se comprende a continuación por los términos ‘filtro de indexación’, un filtro que permite representar relaciones entre datos por asociación a índices de filtros, como por ejemplo los filtros de Bloom.

La base de datos está por consiguiente estructurada en esta memoria sobre la base de filtros. Estos filtros están inicialmente vacíos. Después, cuando tienen lugar las recepciones de datos biométricos que se van a registrar, son actualizados.

Ventajosamente, esta actualización es efectuada de tal manera que los filtros considerados son respectivamente asociados a palabras claves que son obtenidas a partir del dato biométrico recibido.

Procediendo de este modo, se obtiene una base de datos en la cual es posible realizar búsquedas por palabra clave sobre la base de datos almacenados en forma codificada, sin tener que hacer una búsqueda exhaustiva en la base de datos.

En efecto, gracias a estas disposiciones, se asocia en esta base de datos, a través de la estructura de los filtros, al menos un dato biométrico a un subconjunto de filtros.

Además, las palabras claves que están respectivamente asociadas al conjunto de los filtros proceden ventajosamente del dato biométrico recibido. No existe por consiguiente ninguna necesidad de almacenar estas palabras claves de búsqueda para efectuar a continuación búsquedas ulteriores en la base de datos así construida.

En efecto, cuando se solicita una búsqueda en una base de datos tal, es suficiente determinar las palabras clave correspondientes a la búsqueda mediante la aplicación de una familia de funciones de troceado sobre el dato biométrico que se va a controlar.

En un modo de realización de la presente invención, la etapa /3/ es puesta en práctica de acuerdo con las etapas siguientes:

- obtener primeros valores troceados respectivos mediante la aplicación del primer conjunto de funciones de troceado al dato biométrico;
- obtener las citadas palabras clave combinando los citados primeros valores troceados respectivamente con identificadores respectivos de las funciones de troceado del primer conjunto.

Procediendo de este modo, se pueden determinar palabras claves a partir de funciones de troceado utilizadas y de un dato biométrico. Preferentemente, estas funciones de troceado del primer conjunto tienen cada una por características proporcionar valores de entrada vecinos en un espacio métrico, valores muy vecinos, preferiblemente un mismo valor, a la salida en un segundo espacio métrico.

Procediendo de este modo, se prepara una base de datos adaptada a una búsqueda sobre la base de palabras clave que pueden ser valores variables, alrededor del dato biométrico considerado. Dicho de otra forma, una búsqueda en esta base de datos adaptada para ciertas palabras claves puede ser realizada sobre palabras claves que no tienen de hecho los mismos valores pero que tienen valores vecinos. Así, tal base de datos codificados está

adaptada a la búsqueda de palabras claves procedentes de datos biométricos que pueden presentar valores diferentes entre dos capturas distintas de datos biométricos.

En un modo de realización de la presente invención, la etapa /4/ es puesta en práctica de acuerdo con las etapas siguientes:

- obtener segundos valores troceados mediante la aplicación del segundo conjunto de funciones de troceado respectivamente a las palabras clave;

- obtener el subconjunto de filtros de indexación, seleccionando cada filtro al cual está asociado un identificador de filtro que corresponde a los citados segundos valores troceados.

Procediendo de este modo, se construye una base de datos adaptada para una búsqueda sobre la base de palabras claves, estando las relaciones entre las palabras claves y los datos biométricos representadas bajo la forma de filtros. Este segundo conjunto de funciones de troceado aplicadas a las palabras claves permite repartir la representación de las palabras claves sobre diferentes filtros. Este segundo conjunto de funciones de troceado puede corresponder a funciones de troceado de tipo criptográfico, de manera que se obtenga un reparto ventajoso.

Se puede prever que una primera y una segunda familia de funciones de troceado sean inicialmente determinadas, siendo los conjuntos de funciones de troceado primero y segundo una sub parte de las citadas familias de funciones de troceado primera y segunda respectivamente.

La etapa /4/ puede ser puesta en práctica sobre la base de un umbral de tolerancia de error. En un modo de realización de la presente invención, puede resultar ventajoso tomar en consideración a la base de la obtención de las palabras claves, no solamente los primeros valores troceados sino además valores vecinos de estos valores troceados. Tal es el caso, particularmente, cuando se utilizan funciones de troceado del primer conjunto de funciones de troceado del tipo de las descritas en un documento ‘Efficient search for approximate nearest neighbor in high dimensional spaces’ de Eyal Kushilevitz, Rafail Ostrovsky y Yuval Rabani.

El primer conjunto de funciones de troceado puede ser de tipo LSH, por ejemplo como el descrito en el documento de P. Indyk y R. Morwani, ‘Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality’, STOC 1998, y el segundo conjunto de funciones de troceado puede ser de tipo criptográfico.

Ventajosamente los filtros del conjunto de filtros pueden ser filtros de Bloom, tales como los definidos en ‘Space/time trade-offs in hash coding with allowable errors’ de Burton H. Bloom de 1970. Permiten codificar una propiedad de pertenencia a un conjunto optimizando el espacio de almacenamiento, de manera probabilística.

Un segundo aspecto de la presente invención propone un procedimiento de identificación de dato biométrico en una base de datos gestionada de acuerdo con un procedimiento de gestión según el primer aspecto de la presente invención;

comprendiendo la citada base de datos una estructura basada en filtros (B_1, …, B_n);

comprendiendo el citado procedimiento las etapas siguientes al nivel de una entidad de identificación que tiene acceso a la base de datos:

/i/ recibir una solicitud indicando palabras clave troceadas;

/ii/ en la base de datos (10), determinar filtros asociados a las citadas palabras clave troceadas;

/iii/ obtener una lista de direcciones en forma codificada de la base de datos que están respectivamente asociadas a los citados filtros determinados;

/iv/ decidir que las palabras claves tienen una correspondencia en la base de datos, cuando una misma dirección de la citada lista de direcciones está asociada al menos a un número determinado de filtros.

Ventajosamente en esta memoria, se aprovecha la base de datos gestionada según un primer aspecto de la presente invención, haciendo una búsqueda con la ayuda de palabras claves basadas en propiedades de la estructura de filtros de esta base de datos, es decir, particularmente de los filtros de indexación.

El número determinado de filtros puede ser elegido en función de la tasa de errores que se decide tolerar.

Se puede ventajosamente prever que las palabras clave troceadas sean recibidas desde un emisor al nivel de la entidad de identificación;

habiendo obtenido el citado emisor la palabra clave troceada según las etapas siguientes:

- captar un dato biométrico;

- obtener valores troceados mediante la aplicación de un tercer conjunto de funciones de troceado respectivamente sobre el ciado dato biométrico captado;

- obtener las palabras clave combinando los citados primeros valores troceados respectivamente con identificadores respectivos de las funciones de troceado del tercer conjunto; y

- obtener las palabras clave troceadas aplicando un cuarto conjunto de funciones de troceado a las citadas palabras clave.

Procediendo de este modo, al nivel del emisor, las palabras clave utilizadas para la búsqueda son obtenidas a partir de un nuevo dato biométrico captado. Conviene por consiguiente observar aquí que las palabras claves utilizadas para la búsqueda en una base de datos según un modo de realización de la presente invención no están almacenadas sino que son obtenidas a partir de datos biométricos captados, ya sea en el momento de la inscripción con el fin de construir la base de datos, ya sea en el momento de una identificación de dato biométrico.

Los conjuntos de funciones de troceado primero y tercero pueden ser procedentes de una primera familia de funciones de troceado y los conjuntos de funciones de troceado segundo y cuarto pueden ser procedentes de una segunda familia de funciones de troceado.

Se puede observar que puede resultar ventajoso prever la aplicación de un número de funciones de troceado procedentes de las familias de funciones de troceado primera y segunda en el momento de la inscripción mayor que el número de funciones de troceado procedentes de las familias de funciones de troceado primera y segunda aplicadas en el momento de una identificación.

Un tercer aspecto de la presente invención propone una entidad de gestión de base de datos adaptada para la puesta en práctica de un procedimiento de gestión según el primer aspecto de la presente invención.

Un cuarto aspecto de la presente invención propone una entidad de identificación adaptada para la puesta en práctica de un procedimiento de identificación según el segundo aspecto de la presente invención.

Un quinto aspecto de la presente invención propone un sistema de identificación que comprende una entidad de gestión de base de datos de acuerdo con el tercer aspecto de la presente invención; una base de datos y una entidad de identificación según el cuarto aspecto de la presente invención.

Otros aspectos, objetos y ventajas de la invención resultarán evidentes con la lectura de la descripción de uno de sus modos de realización.

La invención será igualmente mejor comprendida con la ayuda de los dibujos, en los cuales:

- la figura 1 ilustra las principales etapas de una gestión de base de datos de acuerdo con un modo de realización de la presente invención;

- la figura 2 ilustra un sistema de identificación de acuerdo con un modo de realización de la presente invención;

- la figura 3 ilustra una fase de inscripción de acuerdo con un modo de realización de la presente invención;

- la figura 4 ilustra una fase de identificación de acuerdo con un modo de realización de la presente invención.

En el contexto de la presente invención, se entiende por los términos ‘datos biométricos’ datos del ser humano que permiten identificarlo. Estos datos biométricos son en general obtenidos por un captador. Tal captador puede por consiguiente estar adaptado para captar datos biométricos de huella, de iris, de rostro, de grafología o incluso de firma vocal por ejemplo.

La figura 1 ilustra las principales etapas de un procedimiento de gestión de base de datos de acuerdo con un modo de realización de la presente invención.

Una entidad de gestión 11 está a cargo de gestionar una base de datos 10 adaptada para almacenar datos biométricos codificados de manera que permita una búsqueda por palabra clave de acuerdo con un modo de realización.

La entidad de gestión 11 recibe un mensaje que comprende un dato biométrico b para almacenar en la base de datos en forma codificada como dato de referencia para búsquedas ulteriores de identificación.

Este dato biométrico b está codificado al nivel de la entidad de gestión 11 en un dato biométrico codificado E(b), con la ayuda por ejemplo de una clave pública.

A continuación este dato biométrico codificado E(b) es transmitido a la base de datos 10 mediante un mensaje 102. Este último almacena este dato codificado en una etapa 103 y reenvía un mensaje 104 a la entidad de gestión 11 indicando la dirección de almacenamiento @b del dato biométrico codificado.

En una etapa 105, al nivel de la entidad de gestión 11, se obtienen palabras claves sobre la base de un primer conjunto de funciones de troceado h_α, para α entero comprendido entre 1 y μ, y del dato biométrico b. Así, en esta etapa, se puede prever aplicar las funciones de troceado h_α del primer conjunto sobre el dato biométrico b. Se obtienen entonces μ valores troceados denotados c_α y que verifican

$$c_\alpha = h_\alpha (b)$$

Tal etapa de construcción de la base de datos permitirá por consiguiente reconocer que dos datos biométricos, uno de los cuales es un dato biométrico codificado de referencia en la base de datos, proceden de una misma persona, incluso si estos dos datos biométricos son substancialmente diferentes.

Después, se puede prever combinar estos valores c_α con el valor α, por ejemplo concatenando α al valor c_α. El valor procedente de tal concatenación es a continuación denotado c_α// α. Después de tal etapa, se dispone entonces de μ palabras clave asociadas al dato biométrico b.

En una etapa 106, se prevé entonces asociar a cada una de estas palabras clave un subconjunto de filtros de indexación. Para constituir el subconjunto de filtros de indexación para ser asociado a cada una de estas palabras claves, se pueden seleccionar filtros en el conjunto de filtros en función de los identificadores de filtro asociados, de las palabras claves y de un segundo conjunto de funciones de troceado h'_β, para β comprendida entre 1 y v.

Para ello, se puede prever seleccionar, en la etapa 107, los filtros que tienen un identificador correspondiente a los valores troceados obtenidos mediante la aplicación del segundo conjunto de funciones de troceado a las palabras clave. Así, se asocia la dirección del dato biométrico a los filtros B_i que tienen los identificadores i que verifican:

$$i = h'_\beta (h_\alpha (b) // \alpha)$$

Después, en la etapa 108, en el subconjunto de los filtros seleccionados se almacena la dirección del dato biométrico b bajo una forma codificada E(@b). Se asocia así el dato biométrico b al subconjunto de filtros de indexación.

En un modo de realización de la presente invención, se ha previsto utilizar filtros de tipo de Bloom. Las secciones siguientes se refieren a este tipo de filtros.

Sea un conjunto D de tamaño relativamente grande, por ejemplo tal como el conjunto de las palabras binarias, y S un subconjunto de este conjunto D de tamaño l, entero positivo. Mediante la construcción de este tipo de filtro, se puede probar si un elemento x pertenece al conjunto S de manera probabilista. Se puede observar que es posible que algunas respuestas positivas sean falsas. Sin embargo, en el contexto de los filtros de Bloom no es posible que respuestas de pertenencia negativas sean falsas. Gracias a esta componente probabilista, se puede así obtener una estructura más compacta que una simple tabla de correspondencia.

Se considera una familia de k funciones de troceado h, k entero positivo, definidas sobre el conjunto D y de valores de llegada en el conjunto de los enteros comprendidos entre 1 y m.

En estas condiciones, un filtro (k; m) de Bloom asociado corresponde a una familia de elementos binarios (B_1, …, B_m), inicialmente inicializados al valor 0, definido como sigue:

para todo i entero comprendido entre 1 y l y j comprendido entre 1 y k, de define:

$$B_{h_j(a_i)} = 1$$

donde S está constituido por los elementos a_i para i entero comprendido entre 1 y l.

En este contexto, para verificar si un elemento x pertenece a S, es suficiente entonces verificar que para todo j entero comprendido entre 1 y k, se verifica la ecuación siguiente:

$$B_{h_j(x)} = 1.$$

La probabilidad de respuesta positiva falsa puede entonces ser despreciable, para una elección de m y k que depende de l.

Los filtros de Bloom pueden ventajosamente aplicarse a un contexto de almacenamiento en un modo de realización de la presente invención. Se busca asociar un conjunto de valores a cada elemento de S. Aquí, durante la gestión de

la base de datos, se busca asociar un conjunto de identificadores de filtro a cada palabra clave obtenida a partir de un dato biométrico de acuerdo con un modo de realización de la presente invención.

Para ello, se propone en el documento « *Public key encryption that allows PIR queries* » de D. Boneh, E. Kushilevitz, R. Ostrovsky, (Alfred Menezes, editor, CRYPTO, volumen 4622, páginas 50 – 67. Springer, 2007) una construcción de filtro de Bloom con almacenamiento.

Sea V un conjunto de valores y R una relación que asocia los elementos v del conjunto V a los elementos del conjunto S.

Los filtros (B1; …; Bm) son todos inicializados en el conjunto vacío y después actualizados como sigue:

para todo i entero comprendido entre 1 y l, donde l es un entero que representa el tamaño del conjunto S, para todo j entero comprendido entre 1 y k, y para todo v elemento del conjunto V tal que v tiene una relación R con a_i, entonces se asocia el elemento v al filtro $B_{h_j(a_i)}$. Así, si el conjunto X, incluido en el conjunto V, es asociado a un elemento a_i de S entonces se tiene que, para todo j entero comprendido entre 1 y k, que el conjunto X está incluido en el filtro $B_{h_j(a_i)}$.

Se tiene entonces, con una probabilidad elevada, el hecho de que un elemento a es elemento de S si y sólo si el resultado de la intersección de todos los filtros $B_{h_j(a)}$ para j comprendido entre 1 y k, es no vacío.

En el modo de realización de la presente invención, se ha previsto aplicar una o incluso varias familias de funciones LSH (para 'Locality Sensitive Hashing' en inglés) en la entrada de tales filtros de Bloom.

Una función LSH es una función de troceado que tiene por característica proporcionar un resultado similar cuando se aplica a puntos que son vecinos en un espacio métrico. Tal función está particularmente definida en el documento 'Approximate nearest neighbors: Towards removing the cruse of dimensionality' de P. Indyk et R. Motwani, STOC 1998.

Sea B un espacio métrico, U un conjunto de valores de llegada de las funciones de troceado de la familia considerada, r1 y r2 dos números reales que verifican:

$$r1 < r2$$

sea p1 y p2 pertenecientes al conjunto [0, 1] y que verifican:

$$p1 > p2$$

Sea una familia H de funciones de troceado $h_1, \ldots, h_\mu$.

Toda función h_i, para i comprendido entre 1 y μ, que va de B hacia U, es una función (r1, r2, p1, p2) – LSH, si para toda función h_i en la familia H, y para todo elemento a y b del conjunto B, se tiene:

si $d_B(a, b) < r1$, entonces $Pr[h_i(a) = h_i(b)] > p1$

y

si $d_B(a, b) > r2$, entonces $Pr[h_i(a) = h_i(b)] < p2$

donde $d_B(a, b)$ es la distancia entre a y b elementos de B en el espacio métrico B.

Se puede, por ejemplo, tomar una familia H de funciones LSH correspondiente al conjunto de todas las proyecciones canónicas en el espacio de Hamming $\{0, 1\}^n$.

En este caso, cada función h_i hace corresponder un valor x_i a un conjunto de valores $(x_1, \ldots, x_n)$.

Después, para todos los r1 y r2 pertenecientes al conjunto de los enteros comprendidos entre 1 y n, y que verifican:

$$r1 < r2,$$

las funciones h_i se denominan funciones $(r_1, r_2, 1\text{-}r_1/n, 1\text{-}r_2/n)$ - LSH.

Se puede igualmente ventajosamente prever en el contexto de la presente invención poner en práctica una familia de funciones LSH tal como la que ha sido propuesta en el documento 'Efficient search for approximate nearest neighbor in high dimensional spaces'.

Aquí, por consiguiente, B es el conjunto de todas n-uplets de $\{0, 1\}^n$ y el vector que se va a trocear se denomina:

x = (x1, …, xn) en B.

La construcción se basa en vectores aleatorios de pesos elegidos. Así, se toma β en el conjunto [0, 1] y se construye un vector aleatorio r de B de tal manera que cada coordenada de r es igual a 1 con una probabilidad β.

La función de troceado h_r inducida entonces puede estar definida como sigue:

$$h_r : x \rightarrow h_r(x) = \sum_{i=1}^{n} x_i . r_i$$

Después, se elige un número t de vectores r_1 a r_t para obtener una función de troceado h tal que:

$$h = (h_{r1}, \ldots, h_{rt}) : B \rightarrow \{0,1\}^t$$

Así, sea x un elemento de B, y r1, … rt elementos de B, correspondientes a vectores aleatorios que verifican que cada bit de uno de estos elementos haya sido generado de manera aleatoria con una probabilidad de β.

Existe entonces un número δ_1 positivo tal que para todo ε positivo, a y b elementos de B, dos elementos de la base de datos tales que:

$d_B (x, a) \leq l$ y

$d_B (x, b) > (1 + \varepsilon).l$

donde $l = 1/2\beta$,

existe entonces una constante δ_2 tal que:

$\delta_2 = \delta_1 + \delta$

donde δ es positivo y dependiente sólo del valor de ε para el cual:

$$\Pr\left[d_B(h(x), h(a)) > \frac{(2\delta_1 + \delta_2)}{3} t \right] \leq e^{-\frac{2}{9}\delta^2 t}$$

$$\Pr\left[d_B(h(x), h(b)) < \frac{(2\delta_2 + \delta_1)}{3} t \right] \leq e^{-\frac{2}{9}\delta^2 t}$$

Un filtro de Bloom permite codificar una propiedad de pertenencia a un conjunto optimizando el espacio, de manera probabilista.

Aquí, en el contexto de la presente invención, se prevé basar tales filtros de Bloom en funciones de troceado de tipo LSH como las descritas anteriormente, seguidas por funciones de troceado de tipo criptográfico.

En un modo de realización de la presente invención, se prevé utilizar más espacio alrededor de los datos biométricos considerados con el fin de permitir una mayor tolerancia a los errores entre dos datos biométricos procedentes respectivamente de dos capturas diferentes pero efectuadas sobre una misma persona.

De manera más precisa, para una familia de funciones LSH tal como la definida anteriormente y en el documento 'Efficient search for approximate nearest neighbor in high dimensional spaces', se ha previsto calcular a partir de un conjunto de datos biométricos captado b, una pluralidad de conjuntos reducidos de datos biométricos c_1 a c_μ, tal que para i comprendido entre 1 y μ, se tiene:

$c_i = h_i (b)$

En este contexto, mediante la construcción de la familia de funciones utilizadas en esta memoria, es posible deducir, con una probabilidad importante, que conjuntos reducidos c_i obtenidos mediante el mismo método que el utilizado para los conjuntos reducidos c_i, pero a partir de una nueva captura de un dato biométrico b', están respectivamente a una distancia d de los conjuntos reducidos c_i, verificando esta distancia d:

$$d < \lambda t$$

donde λ es una constante, inferior a 1, que no depende más que de la primera familia de funciones de troceado H y de los errores que se estima poder tolerar entre el conjunto de datos biométricos b y el conjunto de datos biométricos b'.

Un filtro de Bloom (v, m) está asociado a una segunda familia de funciones de troceado $H' = \{h_1', \ldots, h_v'\}$, dirigiéndose esta segunda familia de funciones de troceado a establecer una buen reparto del almacenamiento de los datos en los diferentes filtros de Bloom B_i.

Se ha previsto entonces determinar valores troceados de todos los conjuntos reducidos c_i' que están a una distancia inferior a λt de c_i mediante la aplicación de un conjunto de funciones de troceado de la segunda familia H'.

Procediendo de este modo, es posible evitar falsas respuestas negativas durante una búsqueda por palabra clave en la base de datos, con una probabilidad elevada.

La figura 2 ilustra un sistema de identificación de acuerdo con un modo de realización de la presente invención. Tal sistema de identificación comprende por consiguiente una entidad de gestión 11, un emisor 12 que comprende un captador de datos biométricos, una base de datos 10 y una entidad de identificación 13.

La entidad de gestión 11 tiene una interfaz con la base de datos 10 para gestionar la inscripción. Puede también tener una interfaz con el emisor en el caso en el que los datos biométricos suscritos provengan de este emisor. No existe ninguna limitación asociada a la fuente de los datos biométricos que recibe la entidad de gestión 11 en la fase de inscripción.

La entidad de identificación 13 está, asimismo, adaptada para comunicarse por un lado con el emisor 12 y por otro lado con la base de datos 10 con el fin de gestionar una fase de identificación.

Se puede prever ya sea que las solicitudes de identificación basadas en palabras claves de acuerdo con un modo de realización de la presente invención sean directamente emitidas desde el emisor con destino a la base de datos 10. Para ello, se puede por consiguiente prever igualmente una interfaz entre el emisor 12 y la base de datos 10.

Sin embargo, en las secciones siguientes, a título de ejemplo, los mensajes entre el emisor y la base de datos pasan por la entidad de identificación 13.

Con el fin de aumentar el nivel de protección de la confidencialidad de los intercambios en el seno de este sistema de identificación, se puede ventajosamente prever utilizar enlaces de comunicación seguros.

Para ello, es posible prever la codificación y la firma de todos los intercambios efectuados entre las entidades de este sistema. Se pueden por ejemplo utilizar los principios de mecanismos de codificaciones tales como los descritos en el documento « *A public key cryptosystem and a signature scheme based on discrete logarithms* » de Taher El Gamal (en CRYPTO, páginas 10 – 18, 1984). Debe denotarse a continuación E una función de codificación asociada a este sistema y D la función de descodificación correspondiente.

Conviene observarse que allí donde se enuncia una función de codificación, respectivamente de descodificación, se puede fácilmente aplicar una función de cifrado, respectivamente de descifrado.

Por otra parte, igualmente con el objetivo de aumentar el nivel de protección de la confidencialidad de los datos manipulados en este sistema, se puede prever utilizar un protocolo de tipo 'Private Block Retrieval protocol' para una comunicación entre la base de datos 10 y la entidad de gestión 11 y/o para una comunicación entre la base de datos 10 y la entidad de identificación 13, tal como se define en el documento « *Private information retrieval* » de Benny Chor, Eyal Kushilevitz, Oded Goldreich, Madhu Sudan (J. ACM, 45(6): 965 - 981, 1998). Procediendo así, la base de datos 10 no tiene acceso a ninguna información relativa al usuario que se está suscribiendo o incluso en proceso de identificación.

En este sistema, se efectúa una captura sobre un usuario al nivel del emisor 12. Después, el dato biométrico b' procedente de esta captura es transmitido a la entidad de identificación 13 que, ella, solicita a la base de datos 10 un reconocimiento o identificación de este dato biométrico b' captado.

Si este dato biométrico no es reconocido a la vista del contenido de la base de datos 10, eso significa que este usuario no está registrado en esta base de datos y el emisor 12 puede ser informado.

En este sistema, una fase de inscripción es gestionada al nivel de la entidad de gestión 11, y se dirige a conseguir datos biométricos de referencia para los usuarios registrados, y una fase de identificación está gestionada al nivel de la entidad de identificación 13 y se dirige a determinar si un usuario es previamente conocido por el sistema, sobre la base de una comparación entre datos biométricos de referencia y datos biométricos nuevamente captados.

La tabla siguiente describe una estructura de la base de datos 10 de acuerdo con un modo de realización de la presente invención.

Filtros	Contenido
B_1	$\{E(a_{1,1}), \ldots, E(a_{p1,1})\}$
…	
B_m	$\{E(a_{1,1}), \ldots, E(a_{pm,m})\}$

Los diferentes filtros corresponden a filtros de Bloom respectivos que son utilizados en el contexto de la presente invención.

Los elementos $E(a_{i,\ j})$ contenidos en los filtros B_1 a B_m corresponden a direcciones de datos biométricos almacenados en la estructura de la tabla siguiente, en forma codificada.

La tabla siguiente describe la estructura del almacenamiento de los datos biométricos.

Etiqueta	Datos
@ (b_1)	$E(b_1)$
…	…
@ (b_N)	$E(b_N)$

En un modo de realización de la presente invención, se determina previamente una primera familia de funciones de troceado H de tipo LSH y una segunda familia de funciones de troceado de tipo criptográfico H'. El conocimiento de estas familias es compartido por el emisor 12 y la entidad de gestión 11. En cambio, no se requiere que estas familias de funciones de troceado se conozcan al nivel de la entidad de identificación 13.

Las funciones de gestión utilizadas para la fase de inscripción al nivel de la entidad de gestión 13 no son necesariamente las mismas que las utilizadas en el curso de la fase de identificación al nivel del emisor, sino que provienen de las mismas dos familias de funciones de troceado compartidas H y H'. Se puede por otra parte prever aplicar un mayor número de funciones de troceado en la fase de inscripción que en la fase de identificación, con el fin de permitir limitar los errores que ocurran durante una identificación.

En el curso de la fase de inscripción, se determina un número μ de funciones de troceado LSH, h_1 a h_μ, que van desde el espacio B al espacio $\{0, 1\}^t$ de la primera familia de funciones H.

Se determina igualmente un número v de funciones de la segunda familia H', es decir dedicadas a un filtro de Bloom con almacenamiento, denotadas h'_1 a h'_v tales que:

$$\{0,1\}^t \times [1, \mu] \rightarrow [1, m]$$

Inicialmente, la base de datos 10 está vacía y los filtros de Bloom B_j, para j comprendido entre 1 y m son inicializados en el conjunto vacío.

La figura 3 ilustra tal fase de inscripción de acuerdo con un modo de realización de la presente invención.

Cada usuario del sistema proporciona un conjunto de datos biométricos b_i captados en la entidad de gestión 11 que está adaptada para gestionar la base de datos. No existe ninguna limitación ligada al método utilizado para efectuar tal etapa. Se puede particularmente prever que esta etapa sea realizada por el emisor 12, mediante la entidad de gestión 11.

Después, la entidad de gestión 11 codificada b_i, en la etapa 21, con la ayuda de la función de codificación E y de una clave pk asociada a la entidad de gestión 11. El conjunto de datos así codificado denotado $E(b_i)$ es almacenado en la base de datos 10.

La base de datos 10 es así rellenada con conjuntos de datos biométricos codificados $E(b_i)$ respectivamente almacenados en direcciones @ b_i en la base de datos en la etapa 22.

Después, las direcciones @ b_i son transmitidas desde la base de datos 10 a la entidad de gestión 11.

Sobre la base de estas direcciones @ b_i, la entidad de gestión es entonces capaz de actualizar los filtros de Bloom que están almacenados en la estructura tal como la descrita anteriormente en la base de datos 10.

Para ello, la entidad de gestión 11 calcula valores c_α para todo α entero comprendido entre 1 y μ, de acuerdo con la ecuación siguiente:

$$c_\alpha = h_\alpha (b_i)$$

Cuando se utilizan funciones de troceado del tipo de las definidas en el documento 'Efficient search for approximate nearest neighbor in high dimensional spaces' citado anteriormente, se puede ventajosamente prever una etapa suplementaria para obtener las palabras clave.

Esta etapa suplementaria consiste en deducir conjuntos C_α correspondientes de acuerdo con la ecuación siguiente:

$$C_\alpha = c_\alpha + \{e \in \{0,1\}^t\}$$

donde e es un vector de $\{0, 1\}^t$ de longitud t que comprende un número de 1 inferior a λt.

En este modo de realización, los elementos de C_α representan las palabras clave en el sentido de la presente invención.

Conviene observar que esta etapa suplementaria es opcional. Es posible en efecto contentarse con aplicar la etapa 106 sobre las palabras claves c_α// α.

Después, se aplican las funciones de troceado de la segunda familia H' sobre estas palabras clave, en la etapa 106.

Para todo β número entero comprendido entre 1 y v, y c_e perteneciente a C_α, un valor γ_e es determinado de acuerdo con la ecuación siguiente:

$$\gamma_e = h'_\beta (c_e // \alpha)$$

donde c_e // α representa una concatenación del valor de c_e y del valor de α.

Finalmente, la entidad de gestión 11 actualiza los filtros de Bloom B_1 a B_m almacenando las direcciones de los datos biométricos b_i bajo su forma codificada E(@ b_i) (etapa 108); para todo i, en los filtros correspondientes B_{γ_e}, es decir los filtros que tienen un identificador, es decir, un índice, que corresponde al resultado de la aplicación de las funciones de la segunda familia de funciones de troceado H' a las palabras clave.

Por consiguiente, la dirección de b_i bajo una forma cifrada es almacenada en cada uno de los filtros de Bloom que tienen un índice γ_e que verifica la ecuación:

$$\gamma_e = h'_\beta (c_e // \alpha)$$

O incluso, en el caso en el que la etapa suplementaria descrita anteriormente no se aplica:

$$\gamma_\alpha = h'_\beta (c_\alpha // \alpha)$$

Tras esta etapa de inscripción, el sistema está listo para efectuar una fase de identificación sobre la base de los datos biométricos codificados almacenados y de un nuevo dato biométrico captado.

La figura 4 ilustra tal fase de identificación de acuerdo con un modo de realización de la presente invención.

En esta fase de identificación, un nuevo dato biométrico b' es captado al nivel del emisor 12. Este último elige entonces un conjunto de funciones de troceado entre la primera familia de funciones de troceado H. Aplica las funciones de este conjunto H al dato biométrico b'. Después, concatena cada uno de los resultados así obtenido en el índice de la función de troceado correspondiente para generar palabras clave. Estas palabras clave son a continuación sometidas a un conjunto de funciones de la segunda familia de funciones de troceado H' para obtener

palabras clave troceadas con el fin de requerir una búsqueda en la base de datos 10 sobre la base de estas palabras clave troceadas.

El principio de la aplicación de las funciones de troceado de la primera familia H, después de la aplicación de las funciones de troceado de la segunda familia H', al nivel del emisor, es casi similar al que es aplicado al nivel de la entidad de gestión en el curso de la fase de inscripción. No obstante, en aras de la simplificación, a continuación, el resultado obtenido, es decir, las palabras clave troceadas, son denotadas $h_{i_j}^{c}(b')$, para i comprendido entre 1 y k, y para j comprendido entre 1 y s, donde k es el número de funciones de troceado aplicadas procedentes de la primera familia de funciones y s es el número de funciones de troceado aplicadas procedentes de la segunda familia de funciones, representando la notación h^c el principio de aplicación de las funciones de troceado descritas anteriormente.

Después, al nivel del emisor 12, se calcula, para α comprendido entre i_1 e i_s el valor de $h_{\alpha}^{c}(b')$ y para al menos uno de los citados entero i.

Todos estos valores $h_{\alpha}^{c}(b')$ son a continuación enviados a la entidad de identificación 13 en un mensaje 41.

Sobre la base de estos valores, la entidad de identificación 13 realiza una solicitud 42 a la base de datos 10 de manera que recupere datos asociados a los filtros B_u de índice u, tal que u es igual a los valores $h_{\alpha}^{c}(b')$ recibidos.

La entidad de identificación 13 recibe a continuación la respuesta 43 a esta solicitud 42 desde la base de datos, para todo u igual a $h_{\alpha}^{c}(b')$.

Así, en esta etapa, la entidad de identificación 13 dispone direcciones de los conjuntos de datos biométricos cifrados que están próximos al dato biométrico b'.

Entonces, la entidad de identificación 13, sobre la base de la clave utilizada para codificar, puede encontrar las direcciones de los datos biométricos almacenados y compararlas entre sí.

A continuación, puede ventajosamente decidir que, si al menos una dirección aparece en todos los B_u recuperados, las direcciones obtenidas son relativas a uno o a varios datos biométricos codificados que corresponden al dato biométrico nuevamente captado b'. En este caso, el usuario en curso de identificación es entonces considerado como identificado.

Puede resultar ventajoso prever restringir la información obtenida por la entidad de identificación 13. Para ello, en un modo de realización de la presente invención, la entidad de identificación 13 no recupera más que las direcciones que están presentes en un número mínimo de B_u. Este modo de realización está por ejemplo basado en la utilización de un esquema de compartición de secreto (o 'secret sharing scheme' en inglés) tal como el que se describe por ejemplo en el documento de A. Shamir, 'How to share a secret', Commun. ACM, 1979.

En un modo de realización, se puede también prever no utilizar más que filtros de Bloom clásicos, sin la característica de almacenamiento, y/o no almacenar los datos biométricos cifrados si no se efectúan etapas ulteriores en el procedimiento de control.

En el caso en el que no haya ninguna dirección común que aparezca en todos los B_u recuperados, se decide entonces que el usuario en curso de identificación no ha sido identificado.

Así, la búsqueda por palabras clave en una base de datos de acuerdo con la presente invención, permite obtener una lista de direcciones de almacenamiento con respecto a candidatos potencialmente próximos de los datos biométricos de identificador b'.

A veces, en ciertos casos, puede resultar ventajoso consolidar esta lista efectuando una verificación final.

Para ello, se puede prever solicitar, a la base de datos, el envío de datos almacenados a las citadas direcciones de la lista, con el fin de poder efectuar una verificación sobre los datos biométricos recuperados y descifrados.

Se puede para ello prever un sistema de codificación que permite cifrar bits de tal manera que sea posible calcular de manera cifrada un XOR (+) de dos mensajes m y m', a partir de mensajes cifrados de m y m'.

Se puede para ello basar en el esquema descrito en el documento 'Evaluating 2-DNF Formulas on Cipher-texts, Theory of Cryptography, Second Theory of Cryptography Conference' de Boneh, Goh y Nissim, LNCS3378, 2005, o

incluso en el sistema de codificación en 'Probabilistic encryption and how to play mental poker keeping secret all partial information' de Goldwasser y Micali, ACM, 1982. Este último corresponde a un esquema homomórfico tal como para una pareja de claves (pk; sk) y dos mensajes m; m' comprendidos en el conjunto {0, 1}, mientras que se verifica la ecuación siguiente:

$$D(E(m; pk) \times E(m'; pk); sk) = m (+) m'$$

donde E y D son respectivamente las funciones de codificación y de descodificación con la clave secreta sk y la clave pública pk.

Los datos biométricos b almacenados en la base de datos 10 están almacenados de manera cifrada bit a bit de acuerdo con la ecuación:

$$E(b; pk) = (E(b_1; pk); \ldots ; E(b_n; pk))$$

En este contexto, una etapa de verificación de acuerdo con un modo de realización de la presente invención, corresponde a un cálculo de distancia de Hamming.

Se puede entonces prever:

- enviar un mensaje que comprende E(b'; pk) a la base de datos 10;
- calcular al nivel de la base de datos 10 para todo dato biométrico b de la base de datos 10:

  E(b; pk) x E(b'; pk);

- efectuar una solicitud desde la entidad de identificación 13 para recuperar los resultados del cálculo anterior;
- recibir desde la base de datos 10 una distancia de Hamming entre el dato biométrico b y el b_i almacenado en la dirección @ b_i recuperada en el filtro B_u, bajo una forma cifrada que la entidad de identificación 13 es capaz de descifrar.

Así, ventajosamente, la entidad de identificación no obtiene al final más que una distancia de Hamming entre el dato biométrico captado y un candidato de la base de datos. Además, en esta memoria, la base de datos no aprende nada con respecto a las operaciones efectuadas.

REIVINDICACIONES

1. Procedimiento de gestión de una base de datos (10), caracterizada por que la citada base de datos comprende datos biométricos almacenados en forma cifrada así como un conjunto de filtros (B_1, …, B_m) asociados respectivamente a identificadores de filtro,

el citado procedimiento comprende las etapas siguientes al nivel de una entidad de gestión:

/1/ recibir (101) un dato biométrico (b_i);

/2/ almacenar, en una dirección dada (@ b_i) en la base de datos, el citado dato biométrico bajo una forma codificada ($E(b_i)$);

/3/ obtener palabras claves a partir de un primer conjunto de funciones de troceado (H) y del citado dato biométrico, teniendo las funciones de troceado del primer conjunto cada una por característica proporcionar valores de entrada vecinos en un espacio métrico, valores muy vecinos, preferiblemente un mismo valor, a la salida en un segundo espacio métrico;

/4/ asociar un subconjunto de filtros de indexación a cada palabra clave seleccionando, para cada palabra clave, filtros en función de los identificadores de filtro respectivamente asociados, de las citadas palabras clave y de un segundo conjunto de funciones de troceado (H'); y

/5/ asociar la citada dirección dada en forma codificada a cada uno de los filtros del subconjunto de filtros de indexación.

2. Procedimiento de gestión de acuerdo con la reivindicación 1, en el cual la etapa /3/ es puesta en práctica de acuerdo con las etapas siguientes:

- obtener primeros valores troceados respectivos mediante la aplicación del primer conjunto de funciones de troceado (H) al dato biométrico (b);

- obtener las citadas palabras clave combinando los citados primeros valores troceados respectivamente con identificadores respectivos de las funciones de troceado del primer conjunto (H).

3. Procedimiento de gestión de acuerdo con la reivindicación 1 ó 2, en el cual la etapa /4/ es puesta en práctica de acuerdo con las etapas siguientes:

- obtener segundos valores troceados mediante la aplicación del segundo conjunto de funciones de troceado (H') a las palabras clave respectivamente;

- seleccionar cada filtro del subconjunto de filtros de indexación al cual está asociado un identificador de filtro que corresponde a los citados segundos valores troceados.

4. Procedimiento de gestión de acuerdo con una cualquiera de las reivindicaciones precedentes, en el cual una primera y una segunda familia de funciones de troceado son inicialmente determinadas, y en el cual el primer y el segundo conjunto de funciones de troceado son una sub parte de las citadas familias primera y segunda de funciones de troceado respectivamente.

5. Procedimiento de gestión de acuerdo con una cualquiera de las reivindicaciones precedentes, en el cual la etapa /4/ es puesta en práctica además sobre la base de un umbral de tolerancia de error.

6. Procedimiento de gestión de acuerdo con una cualquiera de las reivindicaciones precedentes, en el cual el primer conjunto de funciones de troceado es de tipo LSH, y en el cual el segundo conjunto de funciones de troceado es de tipo criptográfico.

7. Procedimiento de gestión de acuerdo con una cualquiera de las reivindicaciones precedentes, en el cual los filtros son filtros de Bloom.

8. Procedimiento de identificación de datos biométricos en una base de datos (10) gestionada de acuerdo con un procedimiento de gestión de acuerdo con una cualquiera de las reivindicaciones precedentes; comprendiendo la citada base de datos una estructura basada en filtros (B_1, …, B_m);

comprendiendo el citado procedimiento las etapas siguientes al nivel de una entidad de identificación (13) que tiene acceso a la base de datos:

/i/ recibir un solicitud indicando palabras clave troceadas;

/ii/ en la base de datos (10), determinar filtros asociados a las citadas palabras clave troceadas;

/iii/ obtener una lista de direcciones en forma codificada de la base de datos que están respectivamente asociadas a los citados filtros determinados;

/iv/ decidir que las palabras claves tienen una correspondencia en la base de datos, cuando una misma dirección de la citada lista de direcciones es asociada al menos a un número determinado de filtros.

9. Procedimiento de identificación de una base de datos de acuerdo con la reivindicación 8, en el cual las palabras clave troceadas son recibidas desde un emisor,

habiendo el citado emisor obtenido la palabra clave troceada de acuerdo con las etapas siguientes:

- captar un dato biométrico (b');

- obtener valores troceados mediante la aplicación de un tercer conjunto de funciones de troceado respectivamente sobre el citado dato biométrico captado;

- obtener las palabras clave combinando los citados primeros valores troceados respectivamente con identificadores respectivos de las funciones de troceado del tercer conjunto (H); y

- obtener las palabras clave troceadas aplicando un cuarto conjunto de funciones de troceado a las citadas palabras clave.

10. Procedimiento de identificación de acuerdo con la reivindicación 9, en el cual los conjuntos de funciones de troceado primero y segundo proceden de una primera familia de funciones de troceado y los conjuntos de funciones de troceado segundo y cuarto proceden de una segunda familia de funciones de troceado.

11. Entidad de gestión de datos adaptada para la puesta en práctica de un procedimiento de gestión de base de datos de acuerdo con una cualquiera de las reivindicaciones 1 a 7.

12. Entidad de identificación adaptada para la puesta en práctica de un procedimiento de identificación de acuerdo con una cualquiera de las reivindicaciones 8 a 10.

13. Sistema de identificación que comprende una entidad de gestión de base de datos de acuerdo con la reivindicación 11; una base de datos (10) y una entidad de identificación de acuerdo con la reivindicación 12.

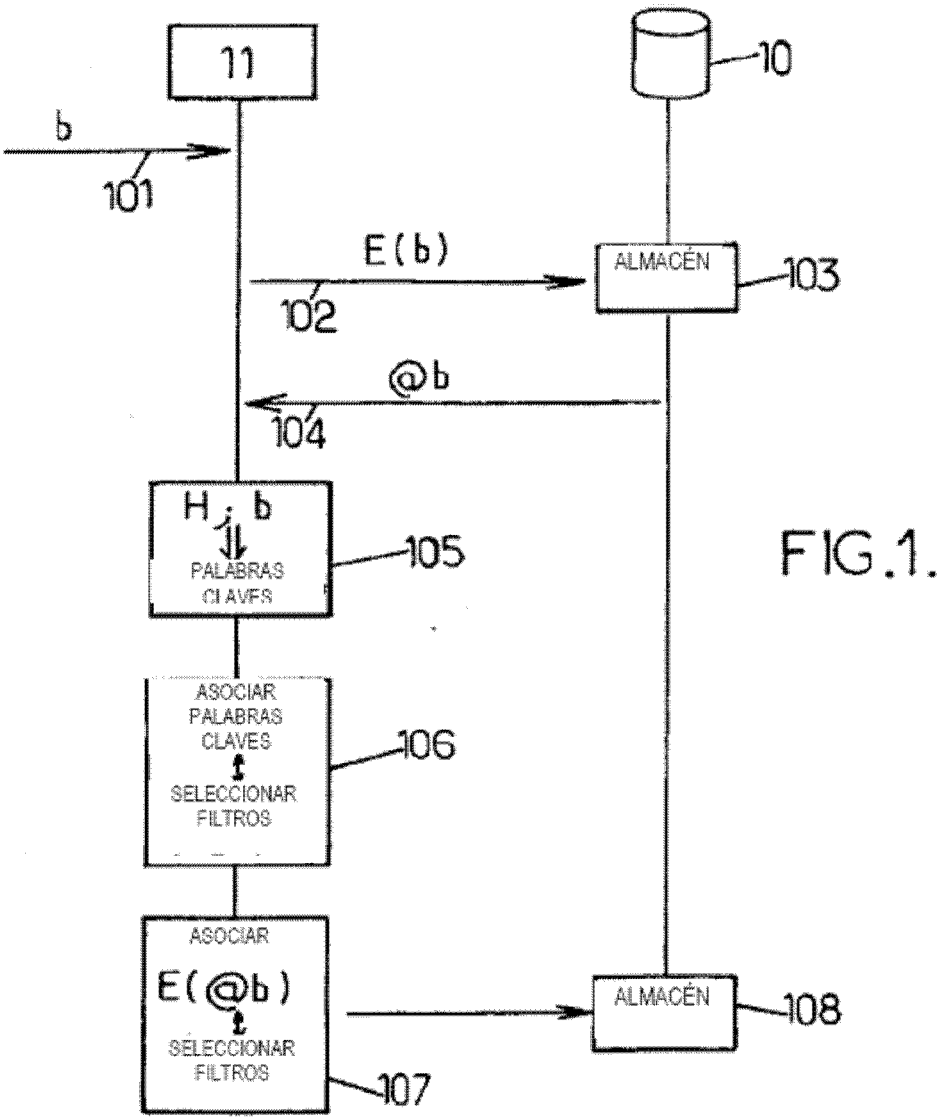
11
10
b
101
E(b)
102
ALMACÉN
103
@b
104
H ; b
PALABRAS CLAVES
105
ASOCIAR PALABRAS CLAVES
SELECCIONAR FILTROS
106
ASOCIAR
E(@b)
SELECCIONAR FILTROS
107
ALMACÉN
108

FIG.1.

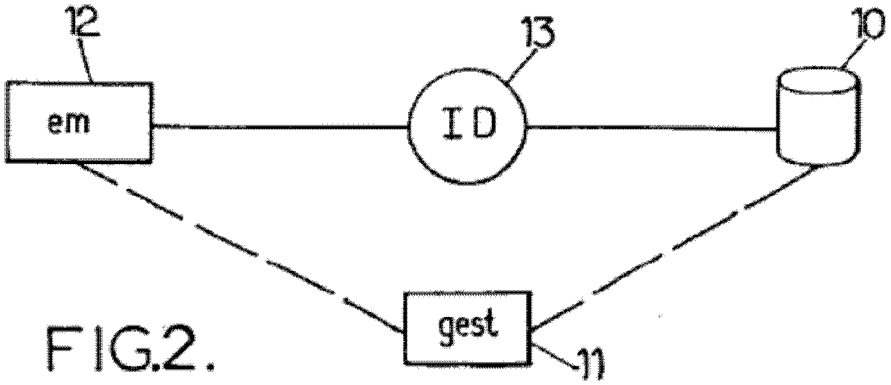
12
em
13
ID
10
gest
11

FIG.2.

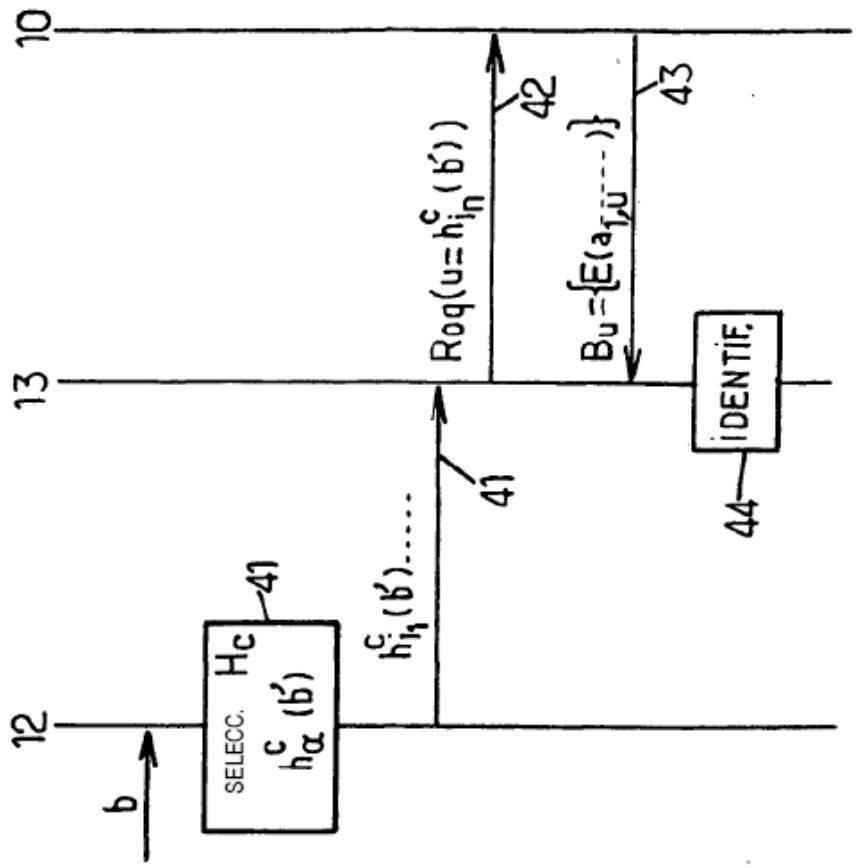
10
13
12
Roq(u=h^c_in(b'))
42
43
Bu={E(a_r,u...)}
IDENTIF.
44
41
h^c_i1(b').....
SELECC. Hc
h^c_α(b')
41
b

FIG.4.

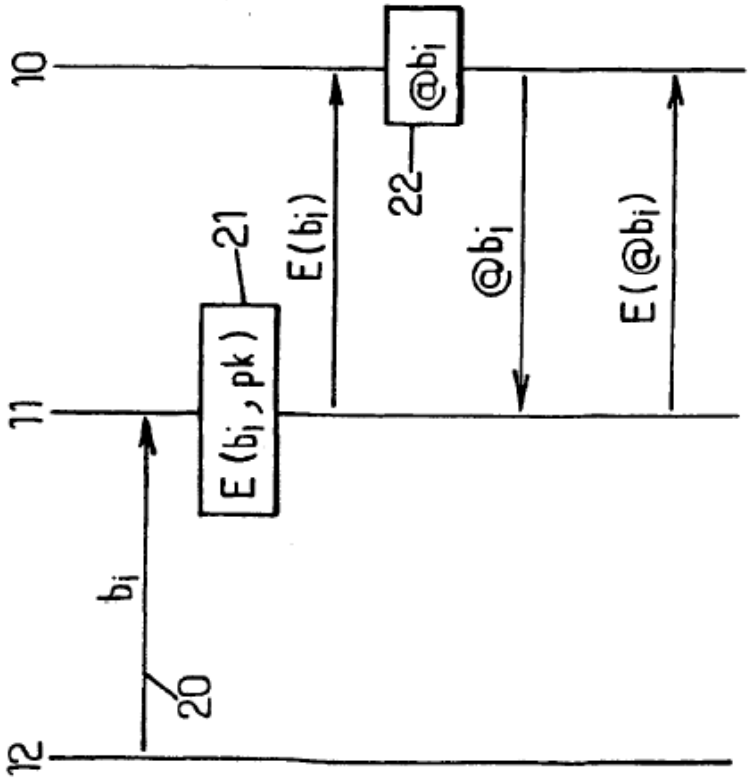
10
11
12
@bi
22
E(bi)
@bi
E(@bi)
E(bi, pk)
21
bi
20

FIG.3.