(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 323 786**

(51) Int. Cl.:
G06F 17/27 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Número de solicitud europea: **04783836 .2**
(96) Fecha de presentación : **13.09.2004**
(97) Número de publicación de la solicitud: **1676211**
(97) Fecha de publicación de la solicitud: **05.07.2006**

(54) Título: **Sistemas y métodos para buscar utilizando preguntas escritas en un conjunto de caracteres y/o idioma distinto al de las páginas objetivo.**

(30) Prioridad: **30.09.2003 US 676724**

(45) Fecha de publicación de la mención BOPI:
24.07.2009

(45) Fecha de la publicación del folleto de la patente:
24.07.2009

(73) Titular/es: **Google Inc.**
1600 Amphitheatre Parkway
Mountain View, California 94043, US

(72) Inventor/es: **Mittal, Vibhu;**
Ponte, Jay, M.;
Sahami, Mehran;
Ghemawat, Sanjay y
Bauer, John, A.

(74) Agente: **Elzaburu Márquez, Alberto**

DESCRIPCIÓN

Sistemas y métodos para buscar utilizando preguntas escritas en un conjunto de caracteres y/o idioma distinto al de las páginas objetivo.

Antecendentes de la invención

1. **Campo de la invención**

La presente invención está relacionada en general con la búsqueda y recuperación de información. Más concretamente, se describen sistemas y métodos para realizar búsquedas utilizando preguntas o consultas que están escritas en un conjunto de caracteres o idioma que es distinto del conjunto de caracteres o idioma de al menos alguno de los documentos que se están buscando.

2. **Descripción de la técnica relacionada**

La mayor parte de los motores de búsqueda operan bajo la suposición de que el usuario final está introduciendo las preguntas o consultas de búsqueda, utilizando un teclado convencional, en donde no es difícil la entrada de cadenas alfanuméricas. Sin embargo, conforme llegan a ser ya comunes los pequeños dispositivos, esta suposición no es siempre válida. Por ejemplo, los usuarios pueden consultar motores de búsqueda con la utilización de teléfonos radioeléctricos que soporten el sistema WAP estándar (protocolo de aplicaciones radioeléctricas). Los dispositivos tales como los teléfonos radioeléctricos tienen típicamente una interfaz de entrada de datos, en donde una acción en particular por parte del usuario (por ejemplo, la pulsación de una tecla) puede corresponder a más de un carácter alfanumérico. La descripción detallada de la arquitectura WAP se encuentra disponible en http://wwwl.wapforum.org/tech/documents/SPEC-WAPArch-19980439.pdf (“*Especificación* de la Arquitectura del protocolo de aplicaciones radioeléctricas WAP 100”).

En el caso usual, el usuario de WAP navega hacia la página de consulta de búsqueda, y se le presenta un formato en el cual se introduce su pregunta o consulta de búsqueda. Con los métodos convencionales, se requiere al usuario que pulse múltiples teclas para seleccionar una letra en particular. En el teclado de un teléfono estándar, por ejemplo, el usuario seleccionaría la letra “b” mediante la pulsación de la tecla “2” dos veces, o bien seleccionaría la letra “s” por la pulsación de la tecla “7” cuatro veces. En consecuencia, para introducir una pregunta o consulta para “ben smith”, el usuario necesitaría normalmente la introducción de la siguiente cadena de pulsaciones de las teclas:

223366077776444844, las cuales se corresponderían con las letras según lo siguiente:

22 -> b

33 -> e

66 -> n

0 -> espacio

7777 -> s

6 -> m

444 -> i

8 -> t

44 -> h

Después de que el usuario haya introducido su pregunta o consulta de búsqueda, el motor de búsqueda recibe la palabra o palabras procedentes del usuario, y procederá de la misma forma que si recibiera la pregunta o consulta desde un navegador de sobremesa, en donde el usuario hubiera utilizado un teclado convencional.

Tal como puede observarse a partir del ejemplo anterior, está forma de introducción de los datos es ineficiente porque exige dieciocho pulsaciones de las teclas para poder introducir los nueve caracteres alfanuméricos (incluyendo el espacio) correspondientes a “ben smith”.

Pueden surgir unas dificultades similares al teclear preguntas o consultas con la utilización de teclados de idiomas de no objetivo. Por ejemplo, el texto japonés puede ser expresado con la utilización de una variedad de conjuntos de caracteres distintos, con la inclusión de los caracteres haragana, katakana, y kanji, en donde ninguno de los mismos pueden introducirse fácilmente utilizando un teclado típico ASCII, que esté basado en el alfabeto romano. En dicha situación, el usuario hará uso frecuente de un procesador de textos tal como el Ichitaro, producido por JustSystem Corporation de Tokushima City, Japón, que es capaz de convertir el texto escrito en romaji (una representación fonética

del alfabeto romano del japonés) a katakana, haragana, y kanji. Utilizando el procesador de textos, el usuario puede teclear una pregunta o consulta en romaji, y después cortar y pegar el texto traducido desde la pantalla del procesador de textos a un recuadro de búsqueda en el navegador. Un inconveniente de esta solución es que puede ser relativamente lenta y tediosa, y requiere tener acceso a una copia del procesador de textos, la cual puede no ser factible debido a las limitaciones de costo y/o memoria.

Queda pendiente, por tanto, la necesidad de métodos y aparatos para proporcionar unos resultados de búsqueda relevantes en respuesta a una pregunta o consulta de búsqueda eficiente.

El documento EP-A-597611 expone un sistema de análisis de documentos que gestiona los documentos en dos formatos.

La invención está expuesta en la reivindicación 1.

Los métodos y aparatos descritos aquí ampliamente, proporcionan unos resultados relevantes de la búsqueda, en respuesta a una pregunta o consulta de búsqueda ambigua. En forma compatible con la invención, dicho método incluye la recepción de una secuencia de componentes de información ambigua por parte del usuario. El método obtiene información de correspondencia que se corresponde con los componentes de información ambigua hacia unos componentes de información menos ambigua. Esta información de correspondencia se utiliza para traducir la secuencia de componentes de información ambigua en una o más secuencias correspondientes de componentes de información menos ambigua. Una o más de estas secuencias de información menos ambigua se proporcionan como una entrada a un motor de búsqueda. Los resultados de la búsqueda se obtienen del motor de búsqueda y son presentados al usuario.

Además de ello, se exponen sistemas y métodos para realizar búsquedas utilizando preguntas o consultas que se expresen en conjuntos de caracteres o idiomas que sean diferentes del conjunto de caracteres o idiomas de al menos algunos de los documentos en los que tenga que realizarse la búsqueda. Las realizaciones de la presente invención permitir al usuario el poder teclear las preguntas o consultas utilizando dispositivos estándar de entrada (por ejemplo, teclados ASCII), en donde se obtienen las consultadas traducidas a formatos relevantes en un servidor (por ejemplo, traducir una pregunta o consulta escrita en romaji a katakana, haragana, y/o kanji), y poder recibir los resultados de la búsqueda basándose en los formatos convertidos.

Se observará que la presente invención puede ser implementada de numerosas formas, incluyendo como un proceso, un aparato, un sistema, un dispositivo, un método, o bien un medio legible por ordenador, tal como un medio de almacenamiento legible por ordenador, onda portadora, o una red de ordenadores en donde las instrucciones del programa se envían a través de líneas de comunicación ópticas o electrónicas. Se describen más adelante varias realizaciones de la invención.

En una realización se describe un método para traducir automáticamente los términos de la pregunta o consulta desde un idioma y/o conjunto de caracteres a otro. Un primer conjunto de texto de anclaje conteniendo un termino de pregunta o consulta dado, son como un conjunto de documentos (por ejemplo, páginas Web) al cual apunta el texto de anclaje. Un segundo conjunto de texto de anclaje, escrito en un segundo formato y apuntando al mismo conjunto de documentos, queda de esta forma identificado. El segundo conjunto de texto de anclaje es entonces analizado, para poder obtener una probabilidad en donde una representación del término dado de la pregunta o consulta en el primer formato pueda corresponder a una representación del término dado de la pregunta o consulta en el segundo formato.

Incluso en otra realización, una pregunta o consulta provista en un primer idioma o conjunto de caracteres es traducida a una segunda lengua o conjunto de caracteres, mediante la comparación del texto de anclaje que contiene uno o más de los términos de la pregunta o consulta y que están escritos en el primer idioma o conjunto de caracteres con el texto de anclaje que corresponde al primer texto de anclaje y que está escrito en el segundo idioma o conjunto de caracteres.

En otra realización, se proporciona un producto de un programa de ordenador para traducir un término escrito en un primer formato a un segundo formato. El producto del programa de ordenador es operativo para provocar que un sistema de ordenadores identifique el texto de anclaje alineado, y para determinar una probabilidad de que una representación de un término dado en el primer formato se corresponda con uno o más términos en el segundo formato.

En otra realización, se proporciona un método para ejecutar búsquedas utilizando preguntas o consultas potencialmente ambiguas. Cuando un usuario introduzca una pregunta o consulta en un primer formato, se traducirá a un grupo de una o más variantes en un segundo formato. Se ejecuta entonces una búsqueda utilizando las variantes traducidas, y retornando la información sensible al usuario. Por ejemplo, el primer formato podría comprender una secuencia de números introducidos utilizando un teclado de teléfono, y el segundo formato podría comprender un texto alfanumérico (por ejemplo, inglés, romaji, romaja, pinyin, o similares). En algunas realizaciones, el grupo de una o más variantes se selecciona, mediante el descarte de variantes traducidas que no aparezcan en un léxico predefinido, y/o que contienen combinaciones de caracteres predefinidos de baja probabilidad. En algunas realizaciones, se utiliza un diccionario probabilística para traducir adicionalmente el grupo de una o más variantes en un tercer formato antes de ejecutar la búsqueda. Por ejemplo, el diccionario probabilística puede ser utilizado para traducir el grupo de una o más variantes

desde el romaji, romaja, o pinyin, a kanji, katakana, haragana, hangul, hanja, o caracteres chinos tradicionales, y la búsqueda puede ejecutarse entonces utilizando las variantes traducidas.

Estas y otras características y ventajas de la presente invención se presentarán con más detalle en la siguiente descripción detallada y en las figuras adjuntas, las cuales ilustran a modo de ejemplo los principios de la invención.

Breve descripción de los dibujos

Los dibujos adjuntos que se incorporan y que constituyen una parte de esta memoria descriptiva, ilustran las realizaciones de la invención, y que conjuntamente con la descripción, sirven para explicar las ventajas y principios de la invención. En los dibujos:

la figura 1 ilustra un diagrama de bloques de un sistema en el cual pueden implementarse métodos y aparatos compatibles con la presente invención;

la figura 2 ilustra un diagrama de bloques de un dispositivo de cliente, compatible con la invención;

la figura 3 ilustra un diagrama que describe tres documentos;

la figura 4a ilustra un índice alfanumérico convencional;

la figura 4b ilustra un diagrama de flujo para proporcionar los resultados de la búsqueda en respuesta a una pregunta o consulta de búsqueda alfanumérica convencional;

la figura 5a ilustra un diagrama de flujo, compatible con la invención, para proporcionar los resultados de la búsqueda, en respuesta a una pregunta o consulta de búsqueda ambigua;

la figura 5b ilustra un diagrama para la correspondencia de información alfanumérica con la información numérica; y

la figura 6 ilustra otro diagrama de flujo, compatible con la invención, para proporcionar resultados de la búsqueda en respuesta a una pregunta o consulta de búsqueda ambigua.

La figura 7 ilustra un método para ejecutar una búsqueda de acuerdo con las realizaciones de la presente invención.

La figura 8 ilustra un diccionario probabilística de traducciones de conjuntos de caracteres.

La figura 9 ilustra el uso de texto de anclaje paralelo para construir un diccionario probabilístico.

La figura 10 ilustra una recolección de documentos enlazados utilizando el texto de anclaje.

Las figuras 11A y 11B ilustran el cálculo de traducciones probables basándose en el texto de anclaje mostrado en la figura 10.

La figura 12 muestra una distribución de probabilidad asociada con una traducción de palabras ilustrativas.

Descripción de realizaciones específicas

Se hará ahora referencia con detalle a las realizaciones de la presente invención según lo ilustrado en los dibujos adjuntos. Los mismo números de referencia pueden utilizarse a través de la totalidad de los dibujos y de la siguiente descripción para referirse a las mismas partes o similares. La siguiente descripción se presenta para permitir que cualquier persona en la técnica pueda hacer realizar y utilizar el cuerpo operativo de la invención. Las descripciones de las realizaciones específicas y aplicaciones se proporcionan solamente como ejemplos, y las distintas modificaciones podrán ser evidentes fácilmente para los técnicos especializados en la técnica. Por ejemplo, aunque muchos de los ejemplos se describen en el contexto de páginas Web de Internet, se comprenderá que las realizaciones de la invención presente podrían utilizarse para buscar otros tipos de documentos y/o información, tales como libros, periódicos, revistas o similares. De forma similar, aunque en aras de la ilustración muchos de los ejemplos describen la traducción de texto japonés de romaji a katakana, haragana, y/o kanji, los técnicos especializados en la técnica apreciarán que los sistemas y métodos de la presente invención podrán aplicarse a cualquier traducción adecuada. Por ejemplo, sin limitación alguna, las realizaciones de la presente invención podrían utilizarse para buscar texto escrito, por ejemplo, con caracteres chinos tradicionales o caracteres coreanos en hangul o hanja, basándose en las preguntas o consultas recibidas en algún otro formato (por ejemplo, pinyin o romaja). Los principios generales aquí descritos podrán aplicarse a otras realizaciones y aplicaciones sin desviarse del espíritu y alcance de la invención. Así pues, la presente invención tiene que estar de acuerdo con el alcance más amplio, abarcando numerosas alternativas, modificaciones y equivalentes compatibles con los principios y características aquí expuestos. Con los fines de la claridad, los detalles relacionados con el material técnico que es conocido en los campos relacionados con la invención, no han sido descritos con detalle, con el fin de no oscurecer innecesariamente la presente invención.

A. *Generalidades*

Los métodos y aparatos compatibles con la invención permiten a un usuario el proponer una pregunta o consulta de búsqueda ambigua, y recibir unos resultados de búsqueda potencialmente sin ambigüedad. En una realización, una secuencia de números recibidos de un usuario de un teclado de teléfono estándar se traduce a un conjunto de secuencias alfanuméricas potencialmente correspondientes. Estas secuencias alfanuméricas correspondientes se proporcionan como una entrada a un motor de búsqueda convencional, utilizando una expresión booleriana "O". De esta forma el motor de búsqueda se utiliza para ayudar a limitar los resultados de la búsqueda en la cual se interesó probablemente el usuario.

B. *Arquitectura*

La figura 1 ilustra un sistema 100 en el cual pueden ser implementados métodos y aparatos compatibles con la presente invención. El sistema 100 puede incluir múltiples dispositivos de cliente 110 conectados a múltiples servidores 120 y 130 por medio de una red 140. La red 140 puede incluir una red de área local (LAN), una red de área amplia (WAN), una red telefónica, tal como la red telefónica conmutada pública (PSTN), una intrared, Internet., o una combinación de redes. Se han ilustrado dos dispositivos 110 de cliente, y tres servidores 120 y 130, conectados a la red 140 en aras de la simplicidad. En la práctica, pueden existir más o menos dispositivos de clientes y servidores. Así mismo, en algunos casos, un dispositivo de cliente puede ejecutar las funciones de un servidor, y un servidor puede ejecutar las funciones de un dispositivo de cliente.

Los dispositivos de cliente 110 pueden incluir dispositivos, tales como grandes ordenadores, miniordenadores, ordenadores personales, ordenadores portátiles, asistentes personales digitales (PDA), o similares, capaces de conectar con la red 140. Los dispositivos de cliente 110 pueden transmitir datos a través de la red 140 o bien recibir datos de la red 140 a través de una conexión cableada, radioeléctrica o bien de tipo óptico.

La figura 2 ilustra un dispositivo 110 de cliente a modo de ejemplo, compatible con la presente invención. El dispositivo de cliente 110 puede incluir un bus 210, un procesador 220, una memoria principal 230, una memoria de solo lectura (ROM) 240, un dispositivo de almacenamiento 250, un dispositivo de entrada 260, un dispositivo de salida 270 y una interfaz de comunicaciones 280.

El bus 210 puede incluir uno o más buses convencionales que permitan la comunicación entre los componentes del dispositivo de cliente 110. El procesador 220 puede incluir cualquier tipo de procesador o microprocesador convencional que interprete y ejecute instrucciones. La memoria principal 230 puede incluir una memoria de acceso aleatorio (RAM) o bien otro tipo de dispositivo de almacenamiento dinámico, que almacene información e instrucciones para su ejecución por el procesador 220. La memoria ROM 240 puede incluir un dispositivo ROM convencional o bien otro tipo de dispositivo de almacenamiento estático, que almacene información e instrucciones estáticas para su utilización por el procesador 220. El dispositivo de almacenamiento 250 puede incluir un medio de grabación magnético y/u óptico y su unidad operativa correspondiente.

El dispositivo de entrada 260 puede incluir uno o más mecanismos convencionales que permitan a un usuario el poder introducir información en el dispositivo de cliente 110, tal como un teclado, un ratón, un lápiz, mecanismos de reconocimiento de voz y/o de tipo biométrico, etc. El dispositivo de salida 270 puede incluir uno o más mecanismos convencionales que proporcionen salida de la información hacia el usuario, incluyendo una pantalla, una impresora, un altavoz, etc. La interfaz de comunicaciones 280 puede incluir cualquier mecanismo similar a un transceptor que permita al dispositivo de cliente 110 el poder comunicar con otros dispositivos y/o sistemas. Por ejemplo, la interfaz de comunicaciones 280 puede incluir mecanismos para la comunicación con otro dispositivo o sistema a través de una red, tal como la red 140.

Tal como se describirá con detalle más adelante, los dispositivos de cliente 110, compatibles con la presente invención, ejecutan ciertas operaciones relacionadas con la búsqueda. Los dispositivos 110 de cliente pueden ejecutar estas operaciones en respuesta a un procesador 220 al ejecutar las instrucciones de software contenidas en un medio legible por ordenador, tal como la memoria 230. Un medio legible por ordenador puede ser definido como uno o más dispositivos de memoria y/o bien ondas portadoras. Las instrucciones de software pueden ser leídas en la memoria 230 desde otro medio legible por ordenador, tal como el dispositivo 250 de almacenamiento de datos, o desde otro dispositivo a través del dispositivo 280 de comunicaciones. Las instrucciones de software contenidas en la memoria 230 provocan que el procesador 220 ejecute las actividades relacionadas con la búsqueda descrita más adelante. Alternativamente, pueden utilizarse circuitos físicos cableados en lugar o en combinación con las instrucciones de software, para implementar procesos compatibles con la presente invención. Así pues, la presente invención no está limitada a cualquier combinación específica del circuito físico y del software.

Los servidores 120 y 130 pueden incluir uno o más tipos de sistemas de ordenadores, tal como un ordenador central, miniordenador, o bien un ordenador personal, capaz de conectar con la red 140, para habilitar a que los servidores 120 y 130 puedan comunicar con los dispositivos de cliente 110. En las implementaciones alternativas, los servidores 120 y 130 pueden incluir mecanismos para conectar directamente con uno o más dispositivos de cliente 110. Los servidores 120 y 130 pueden transmitir datos a través de la red 140, o bien recibir datos de la red 140 a través de una conexión cableada, radioeléctrica ú óptica.

Los servidores pueden estar configurados de una forma similar a la descrita anteriormente con referencia a la figura 2 par el dispositivo 110 de cliente. En una implementación compatible con la presente invención, el servidor 120 puede incluir un motor de búsqueda 125 utilizable por los dispositivos 110 de cliente. Los servidores 130 pueden almacenar documentos (o páginas Web) accesibles por los dispositivos 110 de cliente.

C. *Operación de la arquitectura de configuración*

La figura 3 ilustra un diagrama que describe tres documentos, los cuales pueden almacenarse por ejemplo en uno de los servidores 130.

Un primer documento (Documento 1) contiene dos entradas --"reparación de coches"-- y que está numerado con "3" en su parte inferior. Un segundo documento (Documento 2) contiene la entrada "Alquiler de vídeos". Un tercer documento (Documento 3) contiene tres entradas de --"vino", "champagne", y "artículos de bar"-- que incluye un enlace (o referencia) al Documento 2.

En aras de la simplicidad ilustrativa, los documentos mostrados en la figura 3 contienen solamente cadenas alfanuméricas de información (por ejemplo, "coche", "reparación", "vino", etc.). Los técnicos especializados en la técnica reconocerán, sin embargo, que en otras situaciones los documentos podrían contener otros tipos de información, tal como la información de fonética, o bien audiovisual.

La figura 4a ilustra un índice alfanumérico convencional, basado en los documentos mostrados en la figura 3. La primera columna del índice contiene una lista de términos alfanuméricos, y la segunda columna contiene una lista de los documentos correspondiente a dichos términos. Algunos términos, tales como el termino "3" alfanumérico, solo corresponde (por ejemplo, aparece) en un documento, en este caso en el Documento 1. Otros términos, tales como "alquiler", corresponde a múltiples documentos, en este caso en los Documentos 1 y 2.

La figura 4b ilustra la forma en que un motor de búsqueda convencional, tal como el motor de búsqueda 125, utilizaría el índice ilustrado en la figura 4a para proporcionar resultados de la búsqueda, en respuesta a una pregunta o consulta de búsqueda alfanumérica. La pregunta o consulta alfanumérica puede ser generada utilizando cualquier técnica convencional. Para los fines de la ilustración, la figura 4b describe dos preguntas o consultas alfanuméricas: "coche" y "vino". Bajo una solución convencional, el motor de búsqueda 125 recibe una pregunta o consulta alfanumérica, tal como "coche" (etapa 410), y utiliza el índice alfanumérico para determinar cuales son los documentos que corresponden a dicha pregunta o consulta (etapa 420). En este ejemplo, un motor 125 de búsqueda convencional utilizaría el índice ilustrado en la figura 4a, para determinar que "coche" corresponderá al Documento 1, y retornaría el Documento 1 (o una referencia al mismo) al usuario como un resultado de la búsqueda. De forma similar, un motor de búsqueda convencional determinaría que "vino" corresponderá al Documento 3 y retornaría el Documento 3 (o una referencia al mismo) al usuario (etapa 430).

La figura 5a ilustra un diagrama de flujo, compatible con la invención, de una técnica preferida para proporcionar resultados de búsquedas, en respuesta a una pregunta o consulta de búsqueda numérica, basándose en los documentos y el índice mostrados en las figuras 3 y 4a, respectivamente. En aras de la claridad de exposición, la figura 5a describe una técnica en particular para procesar una pregunta o consulta numérica, basándose en la correspondencia o mapeado con un teléfono de mano estándar; no obstante, los técnicos especializados en la técnica reconocerán que pueden utilizarse otras técnicas compatibles con la invención.

En la etapa 510, la secuencia "227" (consistente en los componentes numéricos "2", "2" y "7") es recibida procedente de un usuario. En la etapa 520, se obtiene la información de cómo los componentes numéricos se corresponden con letras. Suponiendo que el usuario introdujo la información desde un teclado de teléfono estándar, esta información de correspondencia o mapeado se muestra en la figura 5b. Tal como se muestra en la figura 5b, las letras "a", "b", y "c" se corresponden cada una con el numero "1", las letras "p", "q", "r", y "s" se corresponden cada una con el numero "7", y así sucesivamente.

En la etapa 530, utilizando esta información de mapeado o correspondencia, la secuencia "227" se traduce a sus equivalentes potenciales alfanuméricos. Basándose en la información mostrada en la figura 5b, existen 36 combinaciones posibles de letras, que corresponden a la secuencia "227", incluyendo las siguientes: aap, bap, cap, abp, bbp, ... bar ... coche ... ccs. Si los números están incluidos en las posibles combinaciones (por ejemplo, "aa7") entonces existirían 80 combinaciones posibles. En lugar de generar todos los posibles equivalentes alfanuméricos, puede ser deseable el limitar los equivalentes generados basándose en algún léxico. Por ejemplo, puede ser deseable generar solo aquellos equivalentes alfanuméricos que puedan aparecer en un diccionario, en un registro de motores de búsqueda de las preguntas o consultas de búsqueda previas, etc.; o bien por el contrario limitar los equivalentes alfanuméricos mediante la utilización de técnicas estadísticas conocidas (por ejemplo, la probabilidad de ciertas palabras que aparezcan conjuntamente).

En la etapa 540, estos equivalentes alfanuméricos se proporcionan como una entrada al motor de búsqueda convencional, tales como los descritos con referencia a las figuras 4a y 4b, utilizando una operación lógica "O". Por ejemplo, la pregunta o consulta de búsqueda proporcionada al motor de búsqueda podría ser "aap O bap O cap O abp ... O bar ... O coche". Aunque pueden proporcionarse al motor de búsqueda todos los posibles equivalentes alfanuméricos, un subconjunto puede en su lugar utilizarse mediante el uso de técnicas convencionales, para eliminar equivalentes impro-

bables. Por ejemplo, se podría generar una lista más estrecha de posibles combinaciones, mediante el uso de técnicas que utilizaran la información probabilística sobre el uso de letras o palabras: se podrían ignorar las combinaciones que empezaran con "qt", pero incluyendo (y favoreciendo) las combinaciones que comenzaran con "qu".

En la etapa 550, los resultados de búsqueda se obtienen a partir del motor de búsqueda. Debido a que los términos tales como "aap" y "abp" no aparecen en el índice del motor de búsqueda, se ignorarán realmente. En realidad, los únicos términos contenidos dentro del índice mostrado en la figura 4b son "coche" y "bar", y por tanto los únicos resultados de la búsqueda retornados son aquellos que hacen referencia a los Documentos 1 y 3. En la etapa 560, estos resultados de búsqueda se presentan ante el usuario. Los resultados de la búsqueda pueden ser presentados en el mismo orden proporcionado por el motor de búsqueda, o bien pueden reordenarse basándose en consideraciones tales como el idioma del usuario. Suponiendo que el usuario fuera el único interesado en los documentos que contuvieran el término "bar", el usuario recibiría un resultado no deseable (Documento 3) además del resultado deseado (Documento 1). Este puede ser un precio aceptable a pagar, no obstante, con la ventaja de que el usuario tenga solo que pulsar tres teclas para formular la pregunta o consulta de búsqueda.

La figura 6 ilustra otro diagrama de flujo, compatible con la invención, de una técnica preferida para proporcionar resultados de búsquedas, en respuesta a la pregunta o consulta de búsqueda numérica, basándose en los documentos y en el índice mostrados en las figuras 3 y 4a, respectivamente. Este diagrama de flujo demuestra la forma en donde al incrementar la dimensión de la secuencia recibida se puede ayudar a limitar los resultados de la búsqueda a los deseados por el usuario. En aras de la claridad de exposición ilustrativa, la figura 6 describe de nuevo una técnica en particular para procesar una pregunta o consulta numérica basándose en la correspondencia o mapeado de un teléfono de mano estándar; aunque los técnicos especializados en la técnica reconocerán que pueden utilizarse otras técnicas compatibles con la invención.

En la etapa 610, la secuencia "227 48367" (que consiste en los componentes numéricos "2", "2", "7", "4", "8", "3", "6", "7") es recibida procedente del usuario. En aras de la simplificación de la explicación, la secuencia "227" se denominará como "palabra numérica" y la secuencia completa "227 48367" se denominará como "frase numérica". Los posibles equivalentes alfanuméricos de una palabra numérica se denominarán como "palabras de letras" y los equivalentes posibles alfanuméricos de una frase numérica se denominarán como "frases de letras".

En la etapa 620, la información se obtiene en torno a como los componentes numéricos se corresponden o se mapean con las letras. Suponiendo que la misma información de correspondencia se utiliza tal como se muestra en la figura 5b, en la etapa 630, la frase numérica "227 48367" se traduce en las frases de letras potencialmente correspondientes. Basándose en la información mostrada en la figura 5b, existen 11664 posibles frases de letras que se corresponden con la secuencia "227 48367".

En la etapa 640, las frases de letras se proporcionan como una entrada a un motor de búsqueda convencional, tal como el descrito con referencia a las figuras 4a y 4b, utilizando una operación lógica "O". Por ejemplo, la pregunta o consulta de búsqueda que se proporciona al motor de búsqueda podría ser "'aap gtdmp' O 'aap htdmp' ... O 'artículos de bar'". Aunque todas las frases de letras posibles pueden ser suministradas al motor de búsqueda, puede en su lugar utilizarse un subconjunto mediante la utilización de técnicas convencionales para eliminar las frases de letras que sean improbables.

En la etapa 650, los resultados de la búsqueda se obtienen a partir del motor de búsqueda. Debido a que muchos motores de búsqueda están diseñados para que tengan un alto rango los documentos que contengan la frase exacta, el Documento 3 sería probablemente el resultado de búsqueda de más alto rango (es decir, debido a que contiene la frase exacta de "artículos de bar"). Ningún otro documento en el ejemplo contiene una de las demás frases de letras generadas en la etapa 620. Además de ello, muchos motores de búsqueda rebajan el rango (o lo eliminan) de los resultados de búsqueda que contengan partes individuales de una frase pero no la frase completa. Por ejemplo, el Documento 1 se rebajaría en su rango o se eliminaría debido a que contiene la palabra de letras "coche", la cual corresponde a la primera parte de la frase de letras, aunque no contiene ninguna palabra de letras que se corresponda con la segunda parte de la frase de letras. Finalmente, las frases de letras tales como "aap htdmp" se ignoran realmente porque no contienen palabras de letras que aparezcan en el índice del motor de búsqueda.

En la etapa 660, los resultados de la búsqueda se presentan ante el usuario. En el ejemplo mostrado, el primer resultado mostrado ante el usuario sería el Documento 3, el cual es probablemente el más relevante para la pregunta o consulta del usuario. El Documento 1 puede ser eliminado conjuntamente, porque no contiene una de las posibles frases de letras. De esta forma, el usuario está provisto con los resultados de búsqueda más relevantes.

Aunque las descripciones anteriores con referencia a las figuras 5 y 6 se realizan con referencia a la información numérica recibida, y en correspondencia con la información alfanumérica, los técnicos especializados en la técnica reconocerán que son posibles otras implementaciones compatibles con la invención. Por ejemplo, en lugar de recibir una secuencia de números correspondientes a las teclas pulsadas por un usuario, la secuencia recibida puede comprender las primeras letras correspondientes a las teclas pulsadas por el usuario. En otras palabras, en lugar de recibir "227", la secuencia recibida puede ser "aap". En forma compatible con la invención, las secuencias de letras equivalentes generadas en las etapas 530 ó 630 podrían ser otras secuencias de letras (por ejemplo, "bar") que correspondan a "aap". En realidad, la secuencia recibida puede contener elementos fonéticos, audiovisuales, o bien cualquier otro tipo de componentes de información.

Independientemente del formato en el cual se reciba la secuencia, se prefiere en general que la secuencia recibida sea traducida a una secuencia que corresponda l formato en el cual la información se almacene en el índice del motor de búsqueda. Por ejemplo, si el índice del motor de búsqueda se almacena en el formato alfanumérico, la secuencia recibida se traduciría a secuencias alfanuméricas.

Adicionalmente, se prefiere en general que la técnica de correspondencia o mapeado que se utilice para traducir la secuencia recibida de los componentes de información sea la misma técnica que se utilice en el dispositivo de usuario para realizar la correspondencia o mapeado de la entrada del usuario en la información generada por el dispositivo. No obstante, pueden ser casos en que sea preferible utilizar una técnica de correspondencia o mapeado distinta a la utilizada para la entrada del usuario.

Las realizaciones de la presente invención pueden habilitar a los usuarios para poder ejecutar búsquedas introducidas utilizando teclados de idiomas que no sean del objetivo perseguido. Por ejemplo, una página Web que contenga un texto japonés podrá ser escrita en kanji, mientras que un usuario que intente buscar dicha página puede solamente tener acceso a un teclado ASCII estándar (o teléfono de mano) basado en el alfabeto Romano.

La figura 7 ilustra un método para ejecutar dicha búsqueda. Tal como se muestra en la figura 7, un usuario teclea una pregunta o consulta, utilizando un dispositivo de entrada estándar (por ejemplo, un teclado ASCII, un teléfono de mano, etc.), y envía la pregunta o consulta al motor de búsqueda. La pregunta o consulta puede escribirse en un conjunto de caracteres (por ejemplo, romaji) que sea distinto del conjunto de caracteres en el cual algunos de los documentos sensibles estén escritos (por ejemplo, kanji). El motor de búsqueda recibe la pregunta o consulta (bloque 702), traduce la misma al formato(s) relevante (bloque 704), y ejecuta una búsqueda de los documentos sensibles a la pregunta o consulta traducida, atizando por ejemplo las técnicas de búsqueda convencionales (bloque 706). El motor de búsqueda retorna entonces una lista de documentos sensibles (y/o copias de los propios documentos) al usuario (bloque 708). Por ejemplo, los resultados podrían ser retornados al usuario de una forma similar a la descrita antes en relación con la figura 6.

Tal como se muestra en la figura 7, la pregunta o consulta del usuario se traduce preferiblemente en el servidor del motor de búsqueda, en oposición al cliente, liberando así al usuario de la necesidad de obtener un software de propósito especial para poder ejecutar la traducción. No obstante, se observará que en otras realizaciones, una parte o todas las traducciones podrían ser ejecutadas en el cliente. Además de ello, en algunas realizaciones la pregunta o consulta puede introducirse utilizando un dispositivo tal como un teclado de teléfono. En tales realizaciones, la pregunta o consulta inicial numérica puede convertirse primeramente a un formato alfanumérico (por ejemplo, romaji), utilizando las técnicas de mapeado o correspondencia anteriormente descritas en relación con las figuras 5 y 6, incluyendo por ejemplo la aplicación de un léxico y/o técnicas probabilísticas para descartar los mapeados o correspondencias de baja probabilidad (por ejemplo, los mapeados que incluyan combinaciones de letras que no tengan presencia en romaji). Una vez que se haya obtenido una traducción alfanumérica de la pregunta o consulta, podrían ser ejecutadas el resto de las etapas mostradas en la figura 7 (es decir, 704, 706 y 708).

La traducción de la pregunta o consulta desde un conjunto de caracteres a otro (es decir el bloque 704 en la figura 7) puede ejecutarse en distintas formas. Una técnica es utilizar un diccionario convencional estático de significados o traducciones de las palabras, para mapear o hacer corresponder cada término en la pregunta o consulta en un término correspondiente en el idioma del objetivo o conjunto de caracteres. No obstante, un problema existente con esta solución es que generará frecuentemente resultados no precisos, puesto que las palabras son frecuentemente ambiguas, y las preguntas o consultas serán con frecuencia demasiado cortas para poder proporcionar indicios contextuales adecuados para poder resolver esta ambigüedad. Por ejemplo, la palabra “banco” puede referirse a la orilla del río, o a una institución financiera, o a una maniobra de un aeroplano, haciendo así difícil el traducir con precisión en lo abstracto. Además de ello, si el diccionario no es relativamente grande y/o actualizado frecuentemente, podrá no contener entradas para todos los términos que el motor de búsqueda pueda encontrar, tal como palabras apenas utilizadas, argot, modismos, nombres propios o similares.

Las realizaciones de la presente invención pueden utilizarse para solucionar o atenuar algunos o todos estos problemas, mediante el uso de un diccionario probabilístico para traducir los términos de la pregunta o consulta desde un idioma o conjunto de caracteres (por ejemplo, ASCII) a otro (por ejemplo, kanji). En una realización preferida, el diccionario probabilístico mapea o hace corresponder un conjunto de términos a otro conjunto de términos, y asocia una probabilidad con cada uno de los mapeados o correspondencias. Por conveniencia, un “término” o “señal” se referirá a palabras, frases, y/o (más en general) a secuencias de uno o más caracteres que puedan incluir espacios.

La figura 8 muestra un ejemplo de un diccionario probabilístico 800 tal como el descrito anteriormente. El diccionario probabilístico 800 del ejemplo mostrado en la figura 8 mapea o hace corresponder las palabras escritas en romaji (una representación alfabética del alfabeto Romano del Japonés) a palabras escritas en kanji (un conjunto romano de caracteres japoneses basado en ideogramas). Para facilitar la explicación, la figura 8 representa términos en romaji como “<término>$_{romaji}$”, y términos kanji como “<término>$_{kanji}$”. Se observará que en un diccionario actual de romaji a kanji, se usarían los términos actuales de romaji y kanji, en lugar de las traducciones en Inglés mostradas en la figura 8. Por tanto, se observará que la figura 8 está provista para facilitar una explicación de las realizaciones de la presente invención, y no para ilustrar las características actuales y el significado del texto japonés.

El diccionario 800 contiene las entradas 808, 810, 812, 814 para varios términos romaji 802. El diccionario contiene también las representaciones potenciales de cada uno de estos términos en kanji 804, junto con la probabilidad correspondiente 806 de cada representación es correcta. Por ejemplo, el término romaji "banco" podría corresponderse o mapearse con el término kanji de significado "pendiente escarpada" con probabilidad 0,3, a un termino de significado "institución financiera" con probabilidad 0,4, y a un termino de significado "maniobra del aeroplano" con probabilidad 0,2. Con probabilidad 0,1, el término podría corresponderse o mapearse con "otro", lo cual es una forma genérica de permitir que cada término se corresponda con los términos que puedan no estar en el diccionario.

De nuevo, se observará que el ejemplo mostrado en la figura 8 ha sido construido para ilustrar que un término dado (por ejemplo, la palabra "banco") en un primer conjunto de caracteres o idioma, puede mapearse o corresponderse con más de un término en otro conjunto de caracteres o idioma. El técnico especializado en la técnica observará, no obstante, que mientras que en aras de la claridad el ejemplo particular de la figura 8 ilustra este principio, utilizando palabras y significados ingleses, la representación en romaji actual de la palabra "banco", por ejemplo, podría no ser ambigua en el mismo formato que en el equivalente inglés (por ejemplo, puede no existir ambigüedad en romaji entre la palabra para la institución financiera y la palabra para la maniobra de aeroplano). Se observará que para facilitar la explicación, el diccionario mostrado en la figura 8 se ha simplificado en otros aspectos también. Por ejemplo, un diccionario probabilístico actual podría contener muchas correspondencias o mapeados potenciales para cada término, o podría contener solo las correspondencias que excedieran de un umbral de probabilidad predefinido.

Las realizaciones preferidas de la presente invención utilizan dicho diccionario probabilístico para la traducción de preguntas o consultas expresadas en un idioma 7/o conjunto de caracteres, habilitando por tanto a los usuarios a encontrar documentos escritos en un conjunto diferente de caracteres y/o en un idioma distinto al de la pregunta o consulta original. Por ejemplo, si el usuario introduce una pregunta o consulta para "coches" en romaji, el diccionario probabilístico podrá ser utilizado para la correspondencia del término romaji para "coches", por ejemplo, para el término kanji para "coches". De esta forma, los usuarios pueden encontrar documentos relacionados con sus preguntas o consultas, incluso aunque el conjunto de caracteres de las preguntas o consultas (por ejemplo, romaji) y el conjunto de caracteres de los documentos iguales (por ejemplo, kanji) no sean los mismos. Se observará que en este ejemplo en particular, el idioma actual de la pregunta o consulta no se ha cambiado (tanto el romaji como el kanji se utilizan para expresar el Japonés), y si solo la codificación de los caracteres.

Como otro ejemplo adicional, el término "cansado" en inglés ASCII podría realizar la correspondencia o mapeado con el término "müde" en Alemán, utilizando la codificación de los caracteres de Latín 1, puesto que el carácter umlaut-u no existe en ASCII. Se observará que en este ejemplo el diccionario proporciona tanto una traducción a otro idioma (Inglés a Alemán) y una traducción en otra codificación de caracteres (ASCII a Latín 1).

El texto de anclaje comprende el texto asociado con un hiperenlace entre las páginas Web (o lugares dentro de una página Web dada). Por ejemplo, en el idioma de marcas de hipertexto (HTML), la orden: "<A ref="http://www.abc.com">Bancos y Ahorros y Préstamos</A>" provoca que el texto "Bancos y Ahorros y Préstamos" sea visualizado como un hiperenlace que apunte a la página Web encontrada en http://www.abc.com. El texto "Bancos y Ahorros y Préstamos" se denomina como texto de anclaje, y típicamente proporciona una corta descripción de la pagina Web a la cual apunte (por ejemplo, www.abc.com). En realidad, el texto de anclaje proporcionará con frecuencia una descripción de mayor precisión de la página Web que la propia página en sí, y por tanto puede ser particularmente útil al determinar la naturaleza de la página Web a la cual apunte. Además de ello, el uso de la palabra y la distribución en el texto de anclaje está más cerca en el espíritu que el encontrado en las preguntas o consultas de usuario. Es también el caso de que muchos los anclajes que apuntan a una página dada pueden contener el mismo texto altamente similar. Por ejemplo, los anclajes que apunten a www.google.com mostrarán frecuentemente en forma simple "Google", o al menos utilizarán este término a lo largo de otros textos. Así pues, mediante el examen de todo ello, por ejemplo, katakana, los anclajes que apunten a www.google.com, la traducción en katakana de "Google" podrá inferirse con un grado relativamente alto de confianza, simplemente buscando el término que aparezca con la frecuencia más alta (posiblemente después de descartar ciertos anclajes de contenido de baja información, tal como aquellos que expresan sencillamente "hacer clic aquí"). Las realizaciones preferidas de la presente invención aprovechan la ventaja de estas características del texto de anclaje para proporcionar unas traducciones más precisas.

Con referencia a la figura 9, al recibir una pregunta o consulta que contenga un término escrito en un primer conjunto de caracteres (por ejemplo, ASCII) (bloque 902), el servidor identifica un conjunto de texto de anclaje en donde el término pueda aparecer (bloque 904). Por ejemplo, el servidor puede examinar un índice de todos los anclajes conocidos, para identificar dichos anclajes que contengan el término. A continuación, las páginas Web para las cuales están identificados los anclajes (bloque 906), serán los anclajes escritos en el idioma de objetivo o conjunto de caracteres de objetivo (por ejemplo, haragana, katakana, y/o kanji) que apunten a estas páginas (bloque 908). El sistema tendrá ahora dos conjuntos de documentos (en donde el texto de anclaje se considere como un formato del documento). La distribución del término de pregunta o consulta en un conjunto de documentos (por ejemplo, los anclajes que contengan la pregunta o consulta ASCII original) se utilizará entonces para identificar los candidatos más probables para la frase traducida en el otro conjunto de documentos (por ejemplo, los anclajes en paralelo). Las estadísticas pueden ser calculadas con respecto a la frecuencia con la que aparecen los términos de texto de anclaje, y estas estadísticas pueden ser utilizadas para determinar las frecuencias relativas o probabilidades de los términos encontrados en el texto de anclaje que comprendan la traducción correcta de la pregunta o consulta original (bloque 910). Para las preguntas o consultas de múltiples palabras, el proceso descrito anteriormente podrá ser repetido para cada palabra, o bien la pregunta o consulta completa puede ser tratada simplemente como un único término, o podría

utilizarse una agrupación adecuada de las palabras. Por ejemplo, si la pregunta o consulta es “casas grandes”, podría construirse unas posibles traducciones mediante la localización del texto de anclaje alineado que contenga dicha frase (o al menos una de las palabras en la frase). De forma similar, si la pregunta o consulta contuviera más de dos términos, podrían construirse experimentos para determinar cualquier mapeado o correspondencia, mediante la selección de los subconjuntos apropiados de los términos de la pregunta o consulta y generando los resultados de dichos términos.

Una ventaja de la realización de una traducción de la forma mostrada en la figura 9 es el sistema de traducción no precise del conocimiento previo de las correspondencias o mapeados entre los términos en un idioma o conjunto de caracteres y los correspondientes en el conjunto de objetivo. En su lugar, las correspondencias o mapeados pueden determinarse dinámicamente, basándose en el cuerpo de datos que está disponible para ejecutar el análisis estadístico. Así pues, por ejemplo, es posible descubrir traducciones precisas de términos de argot, modismos, nombres propios, y similares, sin incurrir en el esfuerzo o gastos (por ejemplo, análisis lingüístico e investigación) del mantenimiento de un diccionario estático convencional.

Se describirá a continuación una realización ilustrativa de las técnicas de traducción anteriores en relación con las figuras 10-12. En este ejemplo, se supondrá que el usuario ha introducido el término de pregunta o consulta “casa”, y que desea obtener los resultados de la búsqueda escritos en español (o simplemente una traducción del término de la pregunta o consulta). El servidor intentará por tanto traducir el término inglés “house” al español equivalente.

Con referencia a la figura 10, la variedad de páginas Web 959, 961, 965 se enlazan por medio del texto de anclaje 960, 962, 964, 966 a las páginas 972 y 974. Algunas de estas páginas, y su texto de anclaje asociado, están escritas en inglés (es decir, las páginas 959a-e y 963a-t) y algunas están escritas en español (es decir, las páginas 961a-e y 965a-j). El servidor localiza primeramente todos los anclajes que utilicen el término “house”. Estos anclajes pueden estar situados, por ejemplo, mediante la búsqueda de un índice del texto de anclaje almacenado en el servidor. Utilizando dicho índice, el servidor podría primero encontrar los cinco anclajes 960 que utilicen la frase “big house” (“casa grande”, y que apunten a la página Web 972. el servidor determina a continuación que existen también cinco anclajes 962 del idioma del objetivo (es decir, español) que apunten a la página 972 también. En el ejemplo mostrado en la figura 10, estos anclajes contienen el texto “casa grande”. Los anclajes que apuntan a la misma página (tales como los anclajes 960 y anclajes 962) o a las páginas que soportan una relación predefinida, se dice que están “alineados”, en donde en un sentido más general el alineamiento se refiere típicamente (o una equivalencia probable) a la equivalencia de las unidades alineadas.

La figura 11A muestra la frecuencia con la cual aparece el termino en los anclajes 962 del idioma del objetivo. Tal como se muestra en la figura 11A, los términos “casa” y “grande” aparecen cada uno cinco veces (es decir, una vez en cada anclaje 962). Así pues, aparte de los diez términos en total que aparecen en los anclajes de objetivo 962 (es decir, dos términos por anclaje en cada uno de los cinco anclajes), “casa” cuenta por la mitad, y “grande” cuenta por la otra mitad. Así pues, tal como se muestra en la figura 11A, en este punto el término “house” podría mapearse o corresponderse bien con “casa” o “grande” con igual probabilidad, puesto que ambos términos aparecen con igual frecuencia.

No obstante, tal como se muestra en la figura 10, el sistema encuentra también veinte anclajes 964 ingleses que contienen el término “house” y que apuntan a la pagina 974, y diez anclajes españoles 966 que contienen el término “casa” y que apuntan también a la página 974. Tal como se muestra en la figura 11B, el término “house” se corresponderá o se mapeará con “casa” con la probabilidad de 0,75 (es decir, 15/20), y con “grande” con probabilidad de 0,25 (es decir, 5/20). Estas probabilidades se calculan sencillamente dividiendo el numero total de presencias de cada término en los anclajes del idioma del objetivo (es decir, quince, en el caso de “casa”) por el numero total de términos, incluyendo los duplicados, en los anclajes del idioma del objetivo (es decir, veinte términos: diez contenidos en los anclajes 962, y diez contenidos en los anclajes 964). Alternativamente, o adicionalmente, podrían utilizarse otras técnicas para calcular y/o refinar las probabilidades de una traducción o correspondencia dada. Por ejemplo, los técnicos especializados en la técnica observarán que podrían utilizarse una diversidad de técnicas bien conocidas para reducir el error de variancia y las estimaciones de la probabilidad, tales como los métodos Bayesianos, alisamiento de histogramas, alisamiento Kernel, estimadores de contracción, y/o bien otras técnicas de estimación.

En caso de encontrar disponible más texto, las probabilidades podrían refinarse incluso adicionalmente. Por ejemplo, una distribución de la probabilidad final podría ser similar a la mostrada en la figura 12, en la cual el termino “house” se mapea o se corresponde con una probabilidad relativamente alta con respecto a “casa” y su forma diminutiva “casita”, y con una probabilidad algo menor con los términos similares a “casino” y “mansión” (la palabra en español para mansión), y con una probabilidad despreciable con los términos similares a “grande”. Así pues, puede obtenerse una traducción correcta, así como también la identificación de sinónimos probables, sin el conocimiento de los idiomas y/o conjuntos de caracteres que estén traduciéndose.

Se observará que el ejemplo descrito en relación con las figuras 10-12 se proporciona para los fines de la ilustración, y no de limitación, y que pueden realizarse muchos cambios en la metodología aquí descrita. Por ejemplo, podrían utilizarse distintas técnicas estadísticas para alcanzar las probabilidades, y/o modificaciones para las técnicas básicas anteriormente descritas. Además de ello, aunque el ejemplo precedente describe el proceso de traducción tal como tiene lugar después de la recepción de la pregunta o consulta del usuario, se observará en las otras realizaciones el proceso de mapeado o correspondencia podría ejecutarse antes de que se reciba la pregunta o consulta del usuario. Tales correspondencias o mapeados pre-calculadas podrían almacenarse en un diccionario tal como el descrito en la

figura 8, el cual se aplicaría entonces para traducir las preguntas o consultas de usuario tal como pudieran recibirse. Finalmente, se comprenderá que el texto distinto al texto de anclaje alineado podría utilizarse par la traducción. Por ejemplo, las sentencias alineadas o bien otros datos podrían utilizarse de una forma similar. En muchos países existe más de un idioma oficial o reconocido, y los periódicos y revistas contendrán con frecuencia el mismo artículo escrito en cada uno de estos idiomas. Estas traducciones paralelas pueden utilizarse de la misma manera que el texto de anclaje previamente descrito, para preparar diccionarios probabilísticas de traducciones de palabras.

Así pues, las realizaciones preferidas permiten ventajosamente a que usuarios introduzcan preguntas o consultas de búsqueda y/o peticiones de traducción de una forma conveniente (por ejemplo, utilizando un teclado ASCII), y proporcionar una traducción y búsqueda precisa y automática y su búsqueda. En algunas realizaciones, pueden hacerse refinamientos adicionales con el modelo básico anteriormente descrito. Por ejemplo, en algunas realizaciones puede darse una preferencia (ponderación) a los anclajes que contengan varios términos que sean similares al número de términos en la pregunta o consulta original y/o en otros anclajes alineados. Por ejemplo, en el sistema mostrado en la figura 10, l preferencia podría darse a los anclajes que apunten a la página 974, al igual que la pregunta o consulta original, conteniendo cada uno un término único. De forma similar, si un anclaje que contenga el texto "la casa grande" está apuntado también a la pagina 972, su ponderación podría disminuir en un factor apropiado, puesto que contendrá más términos (es decir, 3) que los demás anclajes con los cuales esté alineado. Dicho esquema de ponderación podría estar reflejado en el calculo de probabilidades mostrado en la figura 11B, por la multiplicación de las frecuencias asociadas con estos términos de anclaje mediante un factor adecuado.

D. *Conclusión*

Tal como se ha descrito anteriormente, los métodos y sistemas compatibles con la invención pueden utilizarse para proporcionar los resultados de la búsqueda en respuesta a las preguntas o consultas de búsqueda ambiguas y/o para traducir términos en otro conjunto de caracteres y/o idiomas. Se han descrito varias técnicas y sistemas de traducción y búsqueda. No obstante, se observará que la descripción anterior se ha presentado para los fines de la ilustración, y que son posibles muchas modificaciones y variaciones a la luz de las descripciones anteriores, o a través de la puesta en práctica de la invención. Por ejemplo, aunque la descripción anterior está basada en una arquitectura de cliente-servidor, los técnicos especializados en la técnica reconocerán que puede utilizarse una arquitectura de entidades pares (P2P), compatible con la invención. Además de ello, aunque la implementación descrita incluye software, la invención puede ser implementada como una combinación de hardware y software o solo con hardware. Adicionalmente, aunque los aspectos de la presente invención están descritos como almacenados en la memoria, el técnico especializado en la técnica apreciará que estos aspectos pueden ser almacenados también en otros tipos de medios legibles por ordenador, tales como en dispositivos de almacenamiento secundarios, similares a discos duros, discos flexibles, o CD-ROM; una onda portadora de Internet; o bien otras formas de RAM o ROM. El alcance de la invención está definido por tanto por las reivindicaciones y sus equivalentes.

REIVINDICACIONES

1. Un método que comprende:

la identificación (904) de un primer conjunto de texto de anclaje escrito en un primer formato y conteniendo un término dado;

la identificación (906) de un conjunto de documentos hacia los cuales apunta el primer conjunto de texto de anclaje;

la identificación (908) de un segundo conjunto de texto de anclaje escrito en un segundo formato, y apuntando al conjunto identificado de documentos;

el análisis (910) del segundo conjunto de texto de anclaje para determinar que una representación del término dado en el primer formato se corresponde a la representación de un término dado en el segundo formato.

2. El método de la reivindicación 1, en donde el primer formato comprende un primer conjunto de caracteres, y el segundo formato comprende un segundo conjunto de caracteres.

3. El método de la reivindicación 1, en donde el primer formato comprende un primer idioma y el segundo formato comprende un segundo idioma.

4. El método de la reivindicación 1, en donde el análisis del segundo conjunto del texto de anclaje incluye la identificación de un término que aparece en el segundo conjunto de texto de anclaje, y la designación del termino más frecuente como la representación del termino dado en el segundo formato.

5. El método de la reivindicación 1, en donde el análisis del segundo conjunto del texto de anclaje comprende:

calcular una probabilidad de que el termino dado corresponde a un término en el segundo conjunto de texto de anclaje.

6. El método de la reivindicación 5, en donde la probabilidad se obtiene utilizando al menos unos medios Bayesianos, alisamiento de histogramas, alisamiento Kernel, y estimadores de contracción.

7. El método de la reivindicación 5, en donde la probabilidad de que un termino dado corresponda a un término en el segundo conjunto del texto de anclaje se obtiene por la división del numero de presencias del término en el segundo conjunto del texto de anclaje por el numero total de presencias de todos los términos en el segundo conjunto del texto de anclaje.

8. El método de la reivindicación 1, en donde el análisis del segundo conjunto del texto de anclaje comprende:

el cálculo de un probabilidad de que el termino dado se corresponda con cada termino en el segundo conjunto del texto de anclaje.

9. El método de la reivindicación 1, en donde el análisis del segundo conjunto de texto de anclaje comprende:

la identificación de un término que aparece más frecuentemente en el segundo conjunto del texto de anclaje.

10. El método de la reivindicación 2, en donde se selecciona el primer formato a partir del grupo que comprende: formato, romaja y pinyin; y en donde el segundo conjunto de caracteres se selecciona a partir del grupo que comprende: katakana, haragana, kanji, hangul, hanja, y los caracteres chinos tradicionales.

11. El método de la reivindicación 1, en donde los documentos comprenden páginas Web.

12. El método de la reivindicación 1, que comprende además:

la obtención de una pregunta o consulta escrita en el primer formato y conteniendo el término dado;

traducción de la pregunta o consulta en el segundo formato basándose al menos en parte del mencionado paso de análisis;

búsqueda de una base de datos para la información escrita en el segundo formato que sea sensible a la pregunta o consulta traducida.

13. El método de la reivindicación 12, en donde las etapas se ejecutan en el orden expuesto.

14. Un producto de un programa de ordenador incluido en un medio legible por ordenador, en donde el programa de ordenado incluye instrucciones, las cuales se ejecutan mediante un sistema por ordenador, que son operativas para hacer que el sistema por ordenador ejecute acciones, que comprenden:

la identificación (904) de un primer conjunto de texto de anclaje escrito en un primer formato y conteniendo un término dado;

la identificación (906) de un conjunto de páginas Web a las cuales apunta el primer conjunto de texto de anclaje;

la identificación (908) de un segundo conjunto de texto de anclaje escrito en un segundo formato, y apuntando a un conjunto identificado de páginas Web;

determinación de la probabilidad de que una representación de un término dado en el primer formato se corresponda a una representación de un término dado en el segundo formato.

15. El producto del programa de ordenador de la reivindicación 14, que incluye además instrucciones, las cuales al ejecutarse por el sistema de ordenador, son operativas para provocar que el sistema de ordenador ejecute acciones que comprenden:

modificar la probabilidad de que una representación del termino dado en el primer formato se corresponda con una representación del término dado en el segundo formato, basándose al menos en parte en un análisis de la selección del usuario de los resultados de la búsqueda.

16. El producto del programa de ordenador de la reivindicación 14, que incluye además instrucciones, las cuales al ser ejecutadas por el sistema de ordenador son operativas para hacer que el sistema de ordenador ejecute acciones, que comprenden:

modificar la probabilidad de que una representación del término dado en el primer formato se corresponda con una representación del término dado en el segundo formato, basándose al menos en parte, en un análisis de las preguntas o consultas previas del usuario.

17. El producto del programa de ordenador de la reivindicación 14, en donde la probabilidad se determina al menos en parte, utilizando al menos uno de los métodos Bayesianos, alisamiento del histograma, alisamiento kernel, y estimadores de contracción.

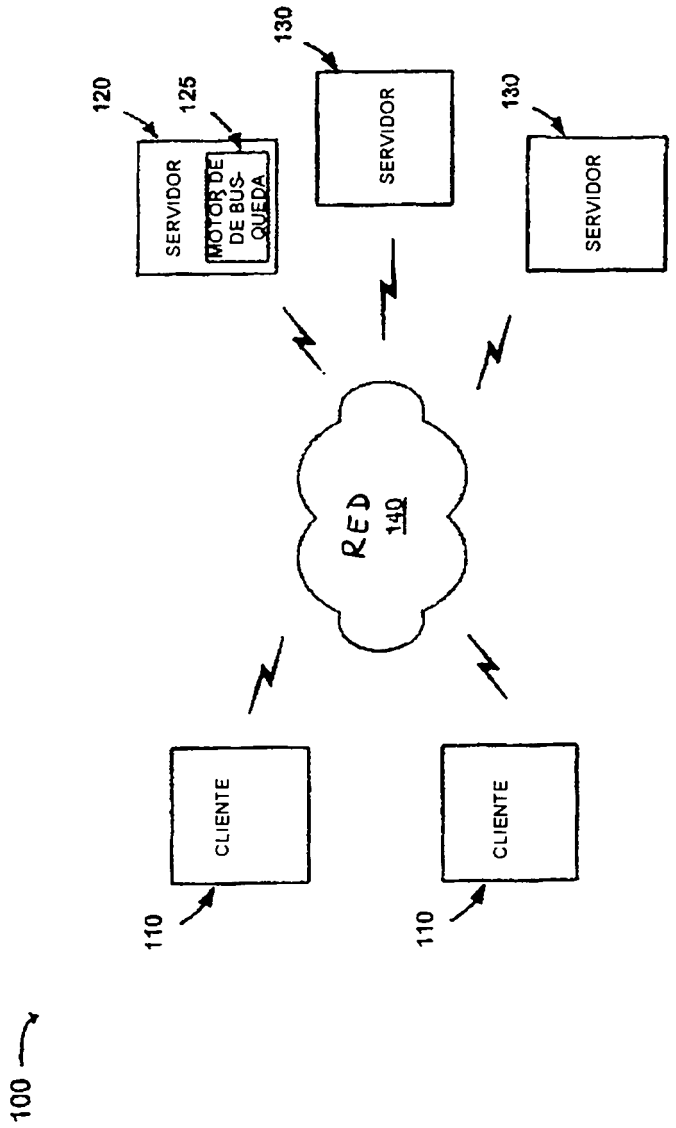

100
110
CLIENTE
110
CLIENTE
RED
140
120
SERVIDOR
125
MOTOR DE BUSQUEDA
130
SERVIDOR
130
SERVIDOR

FIG. 1

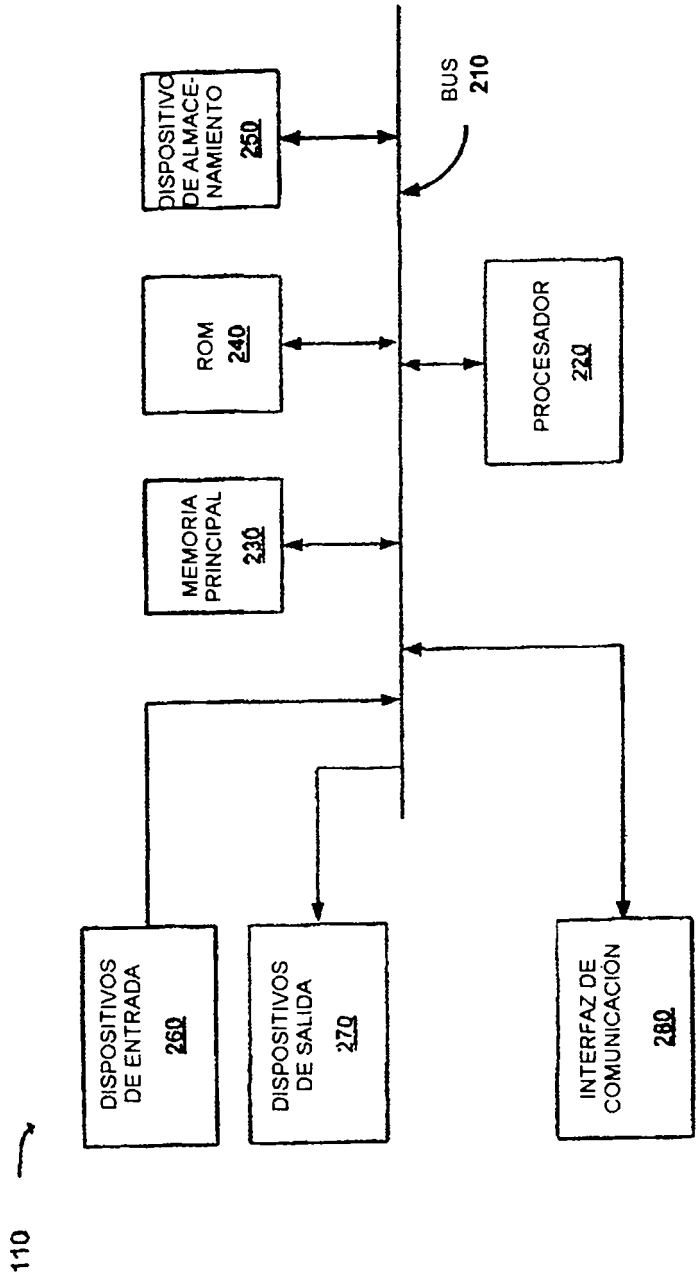

110
DISPOSITIVOS DE ENTRADA 260
DISPOSITIVOS DE SALIDA 270
INTERFAZ DE COMUNICACIÓN 280
MEMORIA PRINCIPAL 230
ROM 240
DISPOSITIVO DE ALMACE-NAMIENTO 250
BUS 210
PROCESADOR 220

FIG. 2

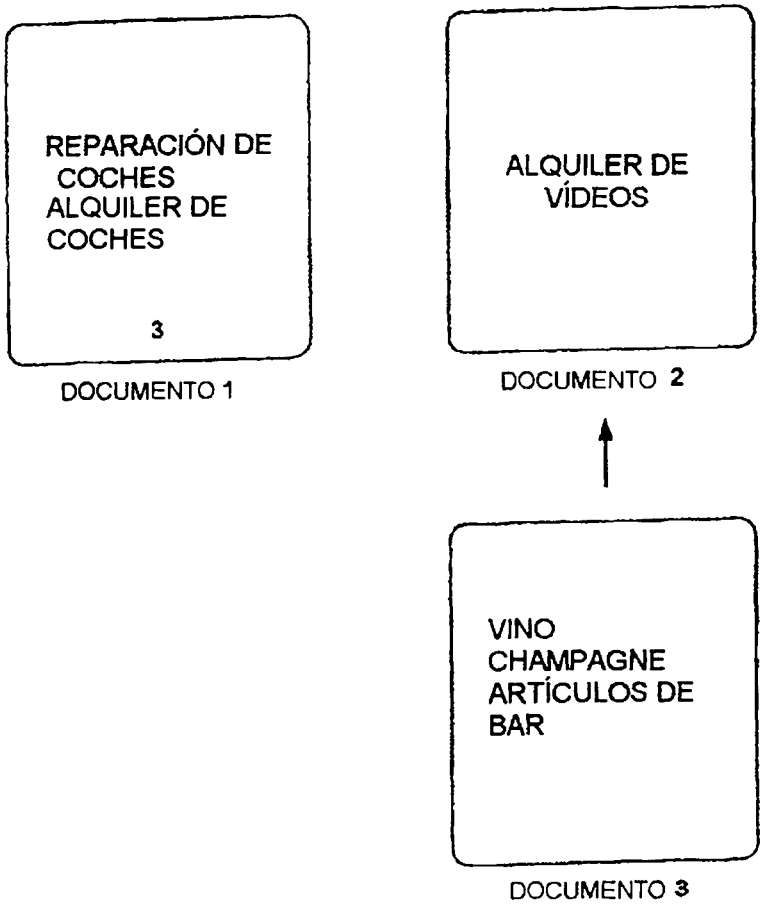

REPARACIÓN DE COCHES
ALQUILER DE COCHES
3
DOCUMENTO 1
ALQUILER DE VÍDEOS
DOCUMENTO 2
VINO
CHAMPAGNE
ARTÍCULOS DE BAR
DOCUMENTO 3

FIG. 3

TÉRMINO	LOCALIZACIÓN (DOCUMENTO)
3	DOCUMENTO 1
BAR	DOCUMENTO 3
COCHE	DOCUMENTO 1
CHAMPAGNE	DOCUMENTO 3
ARTÍCULOS	DOCUMENTO 3
ALQUILER	DOCUMENTOS 1 Y 2
REPARACIÓN	DOCUMENTO 1
VÍDEO	DOCUMENTO 2
VINO	DOCUMENTO 3

FIG. 4A

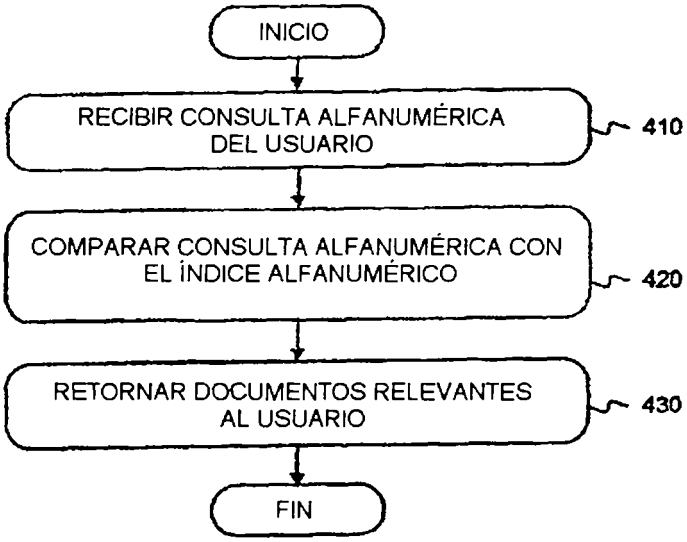

INICIO
RECIBIR CONSULTA ALFANUMÉRICA DEL USUARIO
410
COMPARAR CONSULTA ALFANUMÉRICA CON EL ÍNDICE ALFANUMÉRICO
420
RETORNAR DOCUMENTOS RELEVANTES AL USUARIO
430
FIN

FIG. 4B

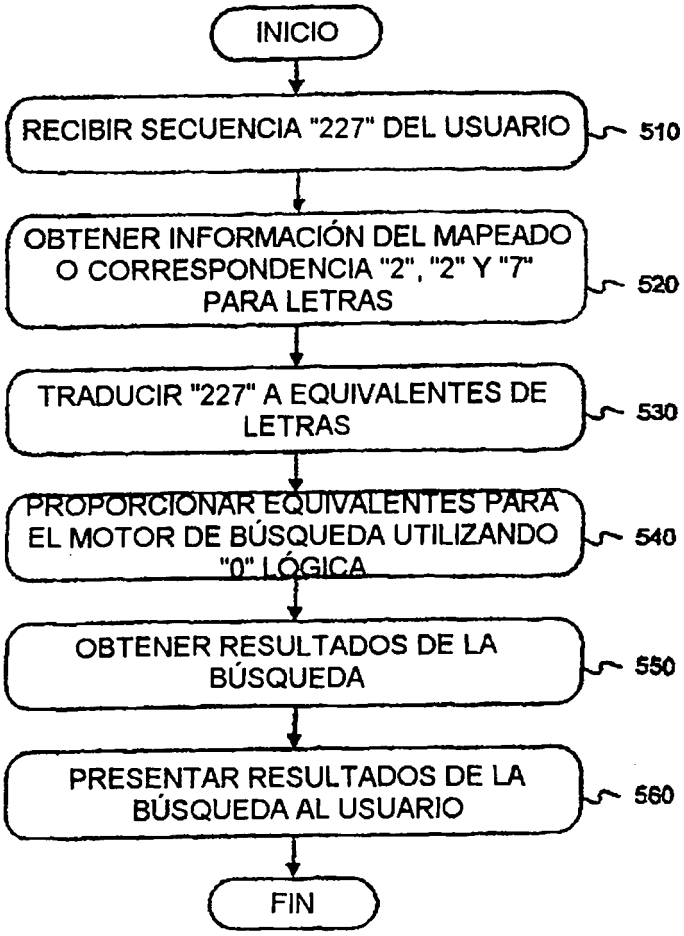

INICIO
RECIBIR SECUENCIA "227" DEL USUARIO
510
OBTENER INFORMACIÓN DEL MAPEADO O CORRESPONDENCIA "2", "2" Y "7" PARA LETRAS
520
TRADUCIR "227" A EQUIVALENTES DE LETRAS
530
PROPORCIONAR EQUIVALENTES PARA EL MOTOR DE BÚSQUEDA UTILIZANDO "0" LÓGICA
540
OBTENER RESULTADOS DE LA BÚSQUEDA
550
PRESENTAR RESULTADOS DE LA BÚSQUEDA AL USUARIO
560
FIN

FIG. 5A

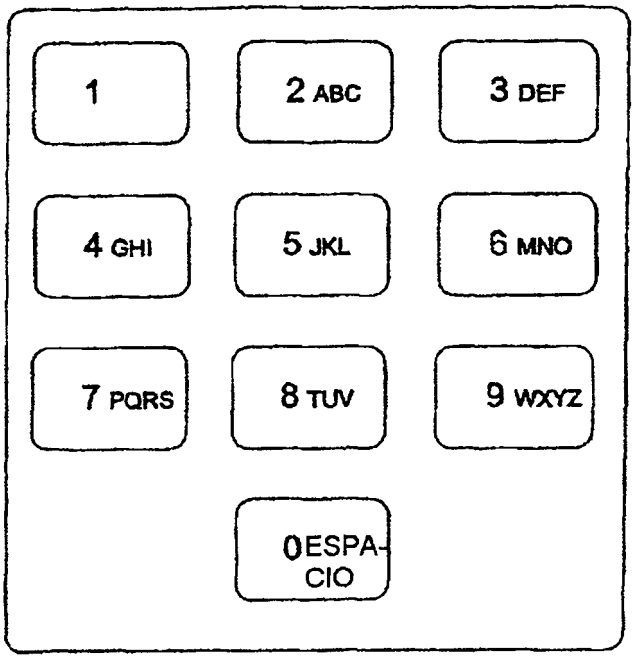


FIG. 5B

TÉRMINO	LOCALIZACIÓN (DOCUMENTO)
3	DOCUMENTO 1
227	DOCUMENTOS 1 Y 3
242672463	DOCUMENTO 3
48367	DOCUMENTO 3
736825	DOCUMENTOS 1 Y 2
737247	DOCUMENTO 1
84336	DOCUMENTO 2
8463	DOCUMENTO 3

FIG. 5C

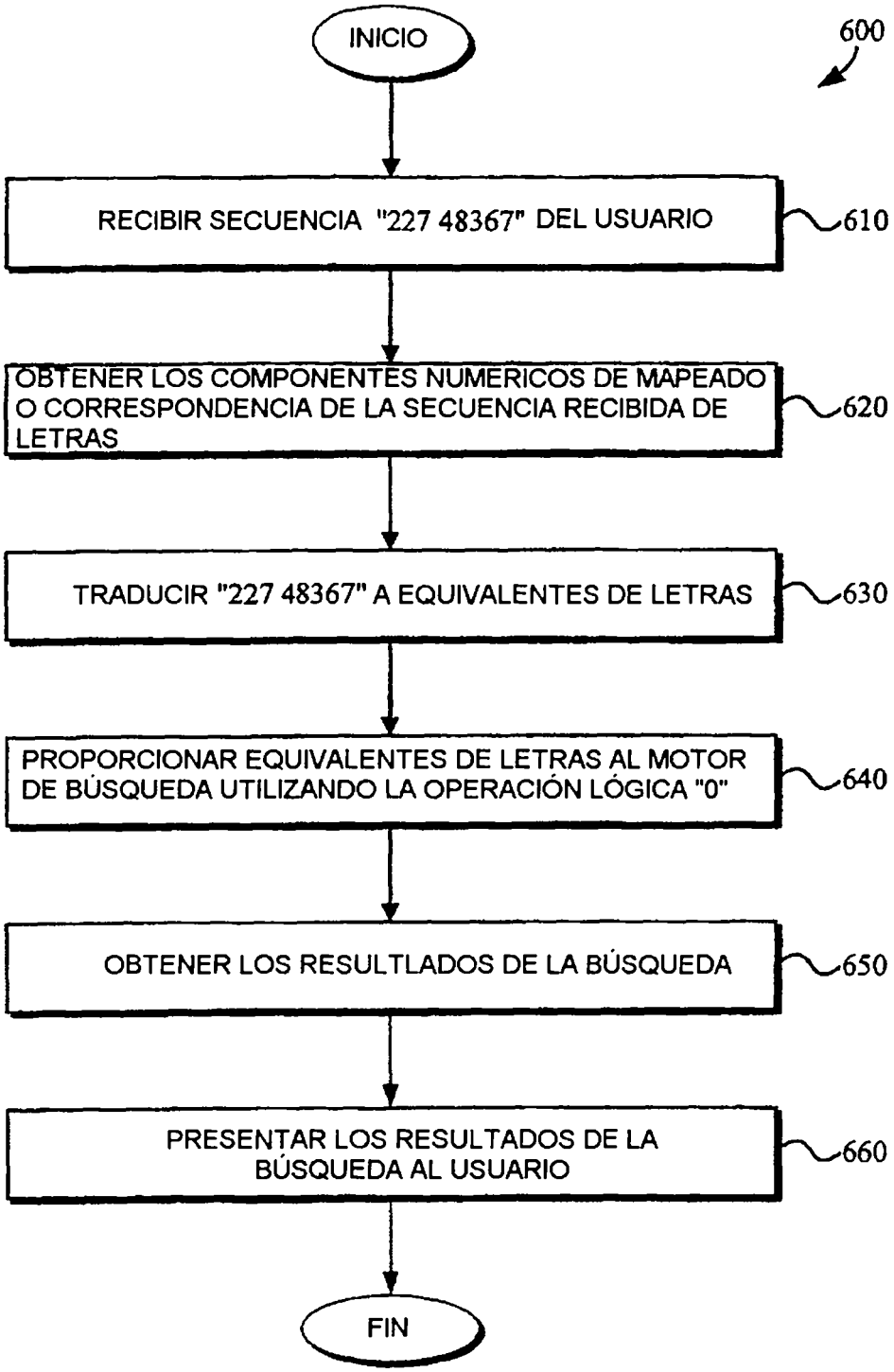

INICIO
600
RECIBIR SECUENCIA "227 48367" DEL USUARIO
610
OBTENER LOS COMPONENTES NUMERICOS DE MAPEADO O CORRESPONDENCIA DE LA SECUENCIA RECIBIDA DE LETRAS
620
TRADUCIR "227 48367" A EQUIVALENTES DE LETRAS
630
PROPORCIONAR EQUIVALENTES DE LETRAS AL MOTOR DE BÚSQUEDA UTILIZANDO LA OPERACIÓN LÓGICA "0"
640
OBTENER LOS RESULTLADOS DE LA BÚSQUEDA
650
PRESENTAR LOS RESULTADOS DE LA BÚSQUEDA AL USUARIO
660
FIN

FIG. 6

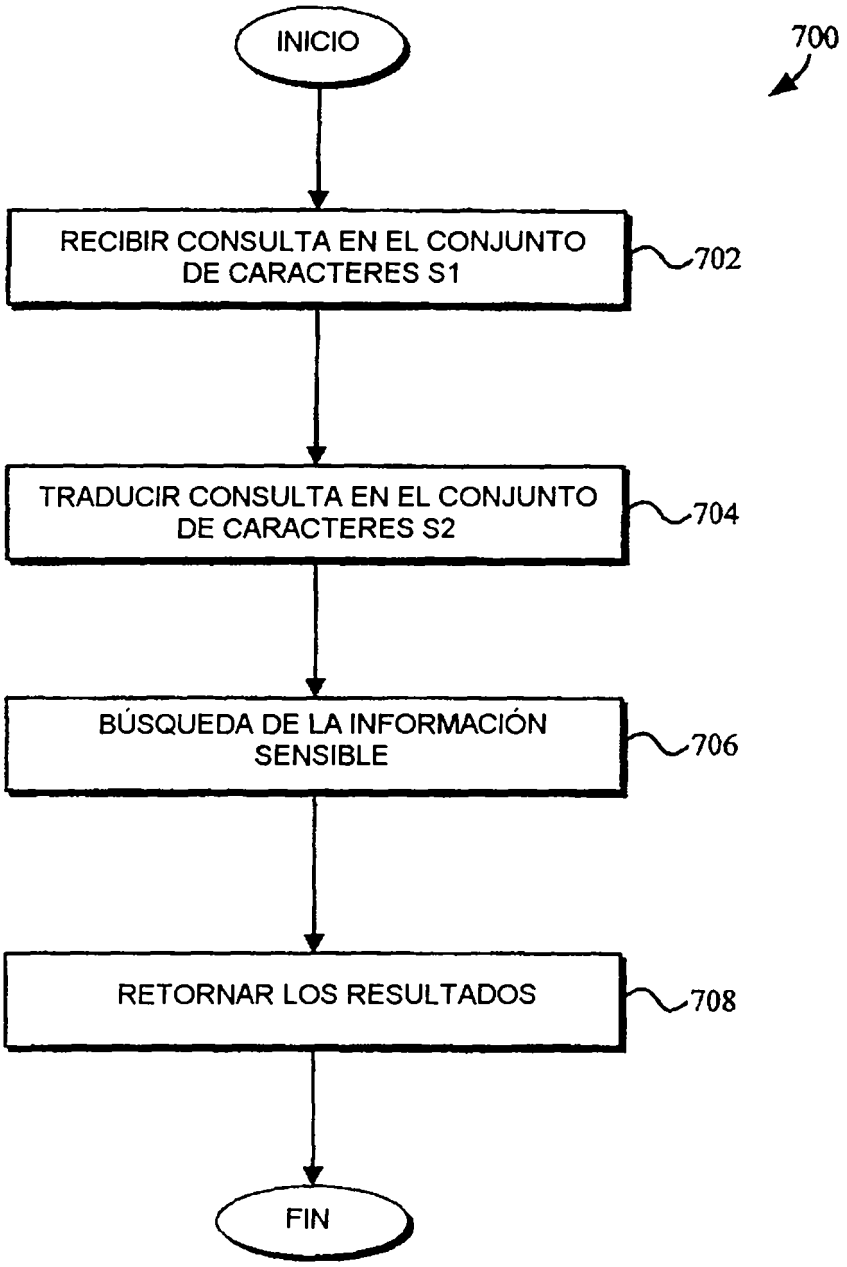
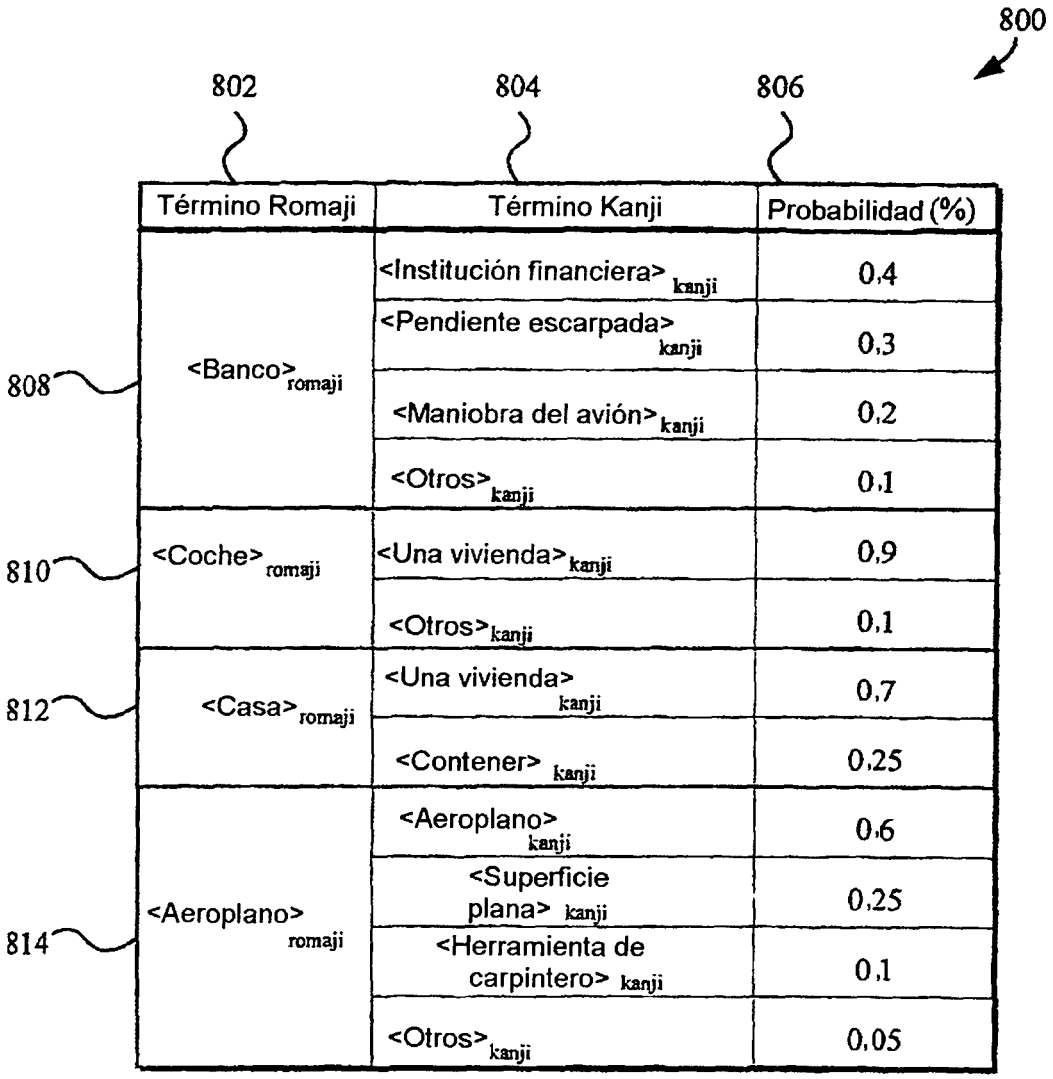

INICIO
700
RECIBIR CONSULTA EN EL CONJUNTO DE CARACTERES S1
702
TRADUCIR CONSULTA EN EL CONJUNTO DE CARACTERES S2
704
BÚSQUEDA DE LA INFORMACIÓN SENSIBLE
706
RETORNAR LOS RESULTADOS
708
FIN

FIG. 7



Término Romaji	Término Kanji	Probabilidad (%)
<Banco>$_{romaji}$	<Institución financiera>$_{kanji}$	0,4
	<Pendiente escarpada>$_{kanji}$	0,3
	<Maniobra del avión>$_{kanji}$	0,2
	<Otros>$_{kanji}$	0,1
<Coche>$_{romaji}$	<Una vivienda>$_{kanji}$	0,9
	<Otros>$_{kanji}$	0,1
<Casa>$_{romaji}$	<Una vivienda>$_{kanji}$	0,7
	<Contener>$_{kanji}$	0,25
<Aeroplano>$_{romaji}$	<Aeroplano>$_{kanji}$	0,6
	<Superficie plana>$_{kanji}$	0,25
	<Herramienta de carpintero>$_{kanji}$	0,1
	<Otros>$_{kanji}$	0,05

FIG. 8

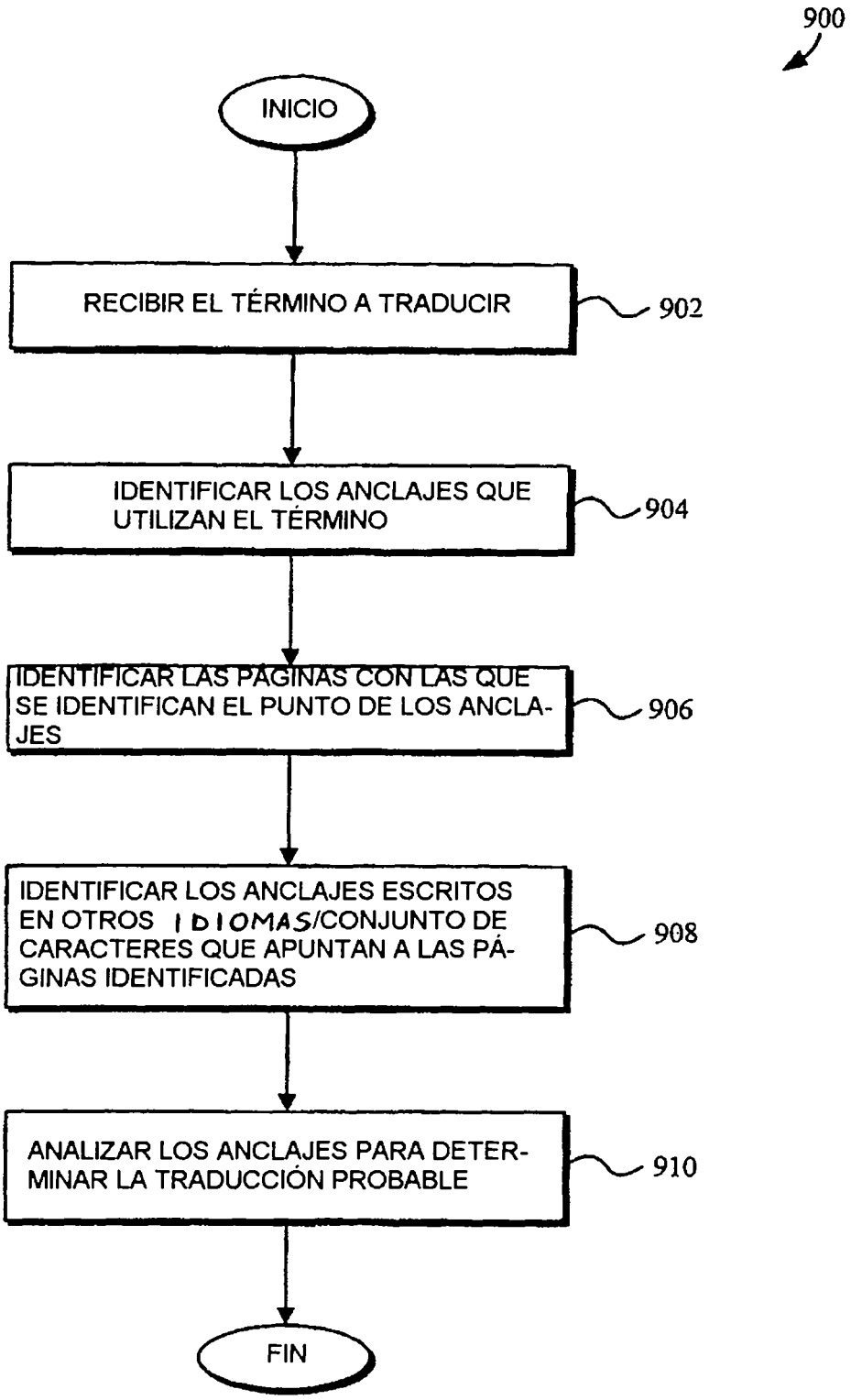

900
INICIO
RECIBIR EL TÉRMINO A TRADUCIR
902
IDENTIFICAR LOS ANCLAJES QUE UTILIZAN EL TÉRMINO
904
IDENTIFICAR LAS PÁGINAS CON LAS QUE SE IDENTIFICAN EL PUNTO DE LOS ANCLAJES
906
IDENTIFICAR LOS ANCLAJES ESCRITOS EN OTROS IDIOMAS/CONJUNTO DE CARACTERES QUE APUNTAN A LAS PÁGINAS IDENTIFICADAS
908
ANALIZAR LOS ANCLAJES PARA DETERMINAR LA TRADUCCIÓN PROBABLE
910
FIN

FIG. 9

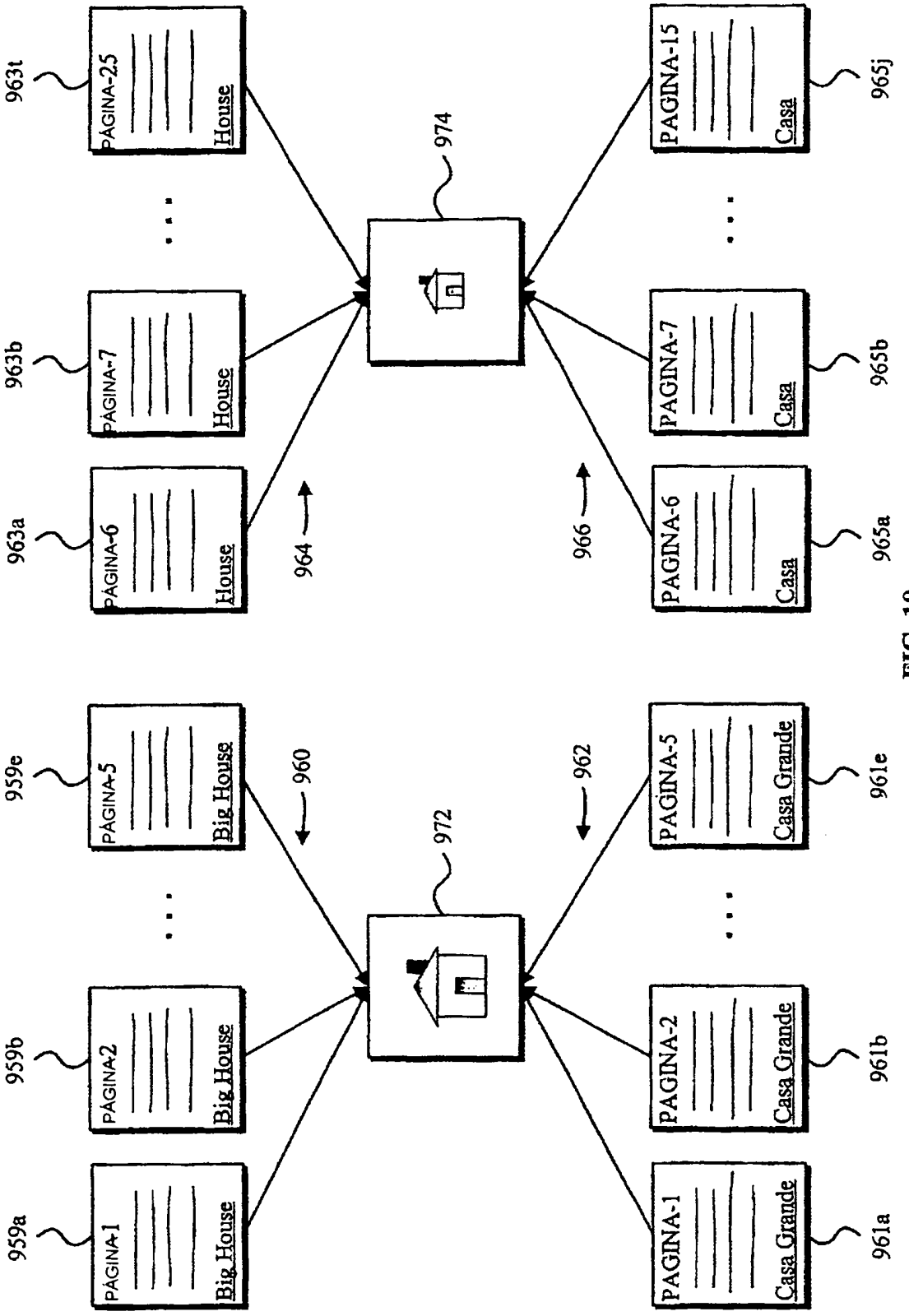

963t
PÁGINA-25
House
963b
PÁGINA-7
House
963a
PÁGINA-6
House
964
974
966
PAGINA-15
Casa
965j
PAGINA-7
Casa
965b
PAGINA-6
Casa
965a
959e
PÁGINA-5
Big House
959b
PÁGINA2
Big House
959a
PÁGINA-1
Big House
960
972
962
PAGINA-5
Casa Grande
961e
PAGINA-2
Casa Grande
961b
PAGINA-1
Casa Grande
961a

FIG. 10

Casa

1	Casa
2	Casa
3	Casa
4	Casa
5	Casa
6	Grande
7	Grande
8	Grande
9	Grande
10	Grande

Probabilidad (house=casa): 5/10 = 0 5

Probabilidad (house=grande): 5/10 = 0 5

FIG. 11A

Casa

1	Casa
2	Casa
3	Casa
. . .	Casa
13	Casa
14	Casa
15	Casa
16	Grande
17	Grande
18	Grande
19	Grande
20	Grande

Probabilidad (house=casa): 15/20 = 0 75

Probabilidad (house=grande): 5/20 = 0 25

FIG. 11B

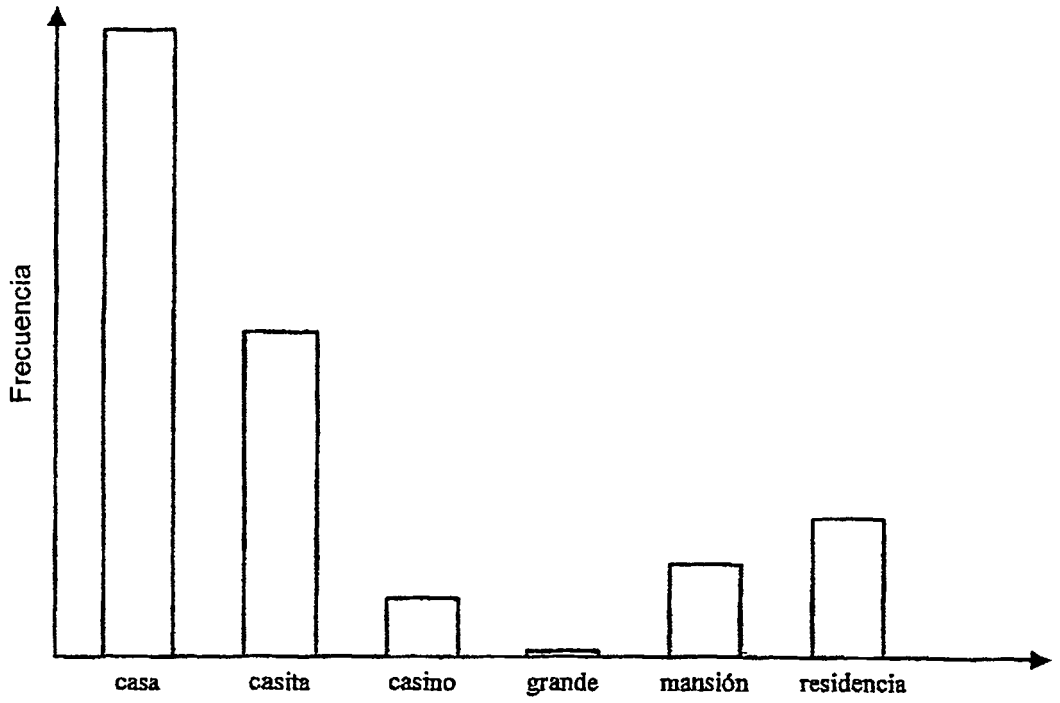


FIG. 12