(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 395 168**

(51) Int. CI.:

G06F 17/27 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **28.09.2005 E 05802157 (7)**

(97) Fecha y número de publicación de la solicitud europea: **27.06.2007 EP 1800224**

(54) Título: **Métodos y sistemas para seleccionar un idioma para segmentación de texto**

(30) Prioridad:

30.09.2004 US 955660

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
08.02.2013

(73) Titular/es:

GOOGLE, INC. (100.0%)
1600 AMPHITHEATRE PARKWAY
MOUNTAIN VIEW, CA 94043, US

(72) Inventor/es:

ELBAZ, GILAD ISRAEL y
MANDELSON, JACOB LEON

(74) Agente/Representante:

URÍZAR ANASAGASTI, José Antonio

DESCRIPCIÓN

Métodos y sistemas para seleccionar un idioma para segmentación de texto

CAMPO DE LA INVENCIÓN

[0001] La presente invención generalmente se refiere a segmentación de texto y, más particularmente, a seleccionar un idioma para segmentación de texto.

ANTECEDENTES DE LA INVENCIÓN

[0002] Existen métodos y sistemas de procesamiento de textos que intentan interpretar datos que representan texto. El procesamiento de textos se vuelve más difícil cuando se recibe texto que comprende una cadena de caracteres que no tiene interrupciones indicando palabras u otros identificadores. Cuando se procesan dichas cadenas de caracteres utilizando métodos y sistemas existentes, los caracteres pueden ser segmentados en identificadores a fin de interpretar la cadena. Identificadores pueden ser palabras, acrónimos, abreviaturas, nombres propios, nombres geográficos, símbolos de denominación abreviada del mercado de valores, u otros identificadores. Generalmente, una cadena de caracteres puede estar segmentada en múltiples combinaciones de cadenas segmentadas de caracteres usando métodos y sistemas existentes. Seleccionar el idioma correcto a usar cuando se segmenta el texto puede producir resultados más significativos.

[0003] US 6272456 revela un método de identificación de un idioma de una entrada de texto en la cual se utilizan múltiples conjuntos de perfiles de idioma n-grama. La clasificación de cada idioma está en base a parámetros de frecuencia de secuencias apareadas de letras de referencia n-gram a la entrada de texto.

RESUMEN

[0004] La presente invención puede definirse por las reivindicaciones 1 y 14.

[0005] Realizaciones de ejemplo son tratadas en la Descripción Detallada, y se proporciona allí una descripción adicional de la invención. Ventajas ofrecidas por las diversas realizaciones de la presente invención pueden además entenderse al examinar esta especificación.

BREVE DESCRIPCIÓN DE LOS DIBUJOS

[0006] Estas y otras características, aspectos, y ventajas de la presente invención se comprenden mejor cuando se lee la siguiente Descripción Detallada con referencia a los dibujos acompañantes, donde:

La Figura 1 ilustra un diagrama de un sistema adecuado para implementar un método de identificación de un idioma para segmentación de textos; y

La Figura 2 ilustra un diagrama de flujo de un método de identificación de un idioma para segmentación de textos.

DESCRIPCIÓN DETALLADA

Introducción

[0007] Las realizaciones de la presente invención comprenden métodos de selección de un idioma para segmentación de textos. Existen múltiples realizaciones de la presente invención. A modo de introducción y ejemplo, las realizaciones de la presente invención proporcionan un método de mejora de la segmentación de una cadena de caracteres, tales como un nombre de dominio, en múltiples identificadores o palabras seleccionando el idioma correcto para la cadena de caracteres. Puede seleccionarse una serie de idiomas potenciales o candidatos para la cadena de caracteres en base a una variedad de señales, tales como linguística asociada con la cadena de caracteres, una dirección IP asociada con el usuario, un conjunto de caracteres utilizado para la cadena de caracteres, configuración del navegador de un programa de aplicación del navegador asociado con el usuario, y cualquier dominio de primer nivel asociado con la cadena de caracteres. La cadena de caracteres puede segmentarse en muchos resultados segmentados usando cada idioma candidato. Cada resultado segmentado puede ser una combinación particular de palabras u otros identificadores. Por ejemplo, la cadena de caracteres “usedrugs” puede segmentarse en los siguientes resultados segmentados para el idioma Inglés: "used rugs", "use drugs", "us ed rugs", etc. A partir de esta serie de resultados segmentados para cada idioma candidato, puede identificarse un resultado segmentado operable y un idioma operable en base al número de documentos o solicitudes de búsqueda en el idioma operable que contienen el resultado operable segmentado.

[0008] Por ejemplo, pueden seleccionarse para cada idioma candidato resultados segmentados con la mayor probabilidad de ser el mejor resultado segmentado operable. Un motor de búsqueda puede determinar el número de documentos o solicitudes de búsqueda que contienen un resultado segmentado seleccionado y pueden hacer esto para cada resultado segmentado seleccionado en cada idioma candidato. En una realización, el resultado segmentado que sucede con mayor frecuencia en documentos o solicitudes de búsqueda en el particular idioma puede identificarse como el mejor resultado segmentado operable. El idioma asociado con el mejor resultado segmentado operable puede

identificarse como el mejor resultado segmentado operable. Las señales de idioma usadas para determinar los idiomas candidatos pueden también usarse para seleccionar el idioma operable. El resultado segmentado operable y el idioma operable pueden usarse para una variedad de funciones, incluyendo seleccionar anuncios en base a el idioma y el resultado.

[0009] Esta introducción se da para presentar al lector el objeto general de la solicitud. La invención no está limitada de forma alguna a dicho objeto. Realizaciones de ejemplo se describen más adelante.

Arquitectura del Sistema.

[0010] Pueden construirse diversos sistemas con la presente invención. La Figura 1 es un diagrama que ilustra un sistema de ejemplo en el que las realizaciones de ejemplo de la presente invención pueden operar. La presente invención puede operar, y ser realizada en, otros sistemas también.

[0011] Haciendo referencia ahora a los dibujos en los que número similares indican elementos similares en las diversas figuras, la Figura 1 es un diagrama que ilustra un entorno de ejemplo para la implementación de una realización de la presente invención. El sistema 100 mostrado en la Figura 1 comprende múltiples dispositivos cliente 102a-n en comunicación con un dispositivo servidor 104 y un dispositivo servidor 150 en una red 106. En un ejemplo la red 106 mostrada comprende la Internet. En otros ejemplos, pueden utilizarse otras redes, tal como una intranet, WAN, o LAN. Es más, los métodos según la presente invención pueden operar en un único ordenador.

[0012] Cada uno de los dispositivos cliente 102a-n mostrados en la Figura 1 comprende un medio legible por ordenador, como una memoria de acceso aleatorio (RAM) 108 acoplada a un procesador 110. El procesador 110 ejecuta instrucciones de un programa informático ejecutable almacenadas en la memoria 108. Dichos procesadores pueden comprender un microprocesador, un ASIC, y máquinas de estado. Dichos procesadores comprenden, o pueden estar en comunicación con, medios, por ejemplo medios legibles por ordenador, que almacenan instrucciones que, cuando se ejecutan por el procesador, hacen que el procesador realice los pasos aquí descritos. Realizaciones de medios legibles por ordenador incluyen, pero no están limitados a, un dispositivo electrónico, óptico, magnético, o otro de almacenamiento o transmisión capaz de proporcionar a un procesador tal como el procesador 110 del cliente 102a, instrucciones legibles por ordenador. Otros ejemplos de medios adecuados incluyen, pero no están limitados a, un disquette, CD-ROM, DVD, disco magnético, chip de memoria, ROM, RAM, un ASIC, un procesador configurado, todos los medios ópticos, todas las cintas magnéticas u otros medios magnéticos, o cualquier otro medio adecuado del que un procesador informático pueda leer instrucciones. También, otras diversas formas de medios legibles por ordenador pueden transmitir o llevar instrucciones a un ordenador, incluyendo un router, una red privada o pública, u otro dispositivo o canal de transmisión, tanto por cable como inalámbrico. Las instrucciones pueden comprender códigos de cualquier idioma de programación informática adecuado, incluyendo, por ejemplo, , C, C++, C#, Visual Basic®, Java®, Python™, Perl®, y JavaScript ®.

[0013] Los dispositivos cliente 102a-n pueden también comprender una serie de dispositivos externos o internos tales como un ratón, un CD-ROM, DVD, un teclado, una pantalla, u otros dispositivos de entrada o salida. Ejemplos de dispositivos cliente 102a-n son ordenadores peronales, asistentes digitales, agendas electrónicas, teléfonos celulares, teléfonos móviles, teléfonos inteligentes, buscas, tabletas digitales, ordenadores portátiles, ordenadores de red, y otros dispositivos en base a procesadores. En general, un dispositivo cliente 102a puede ser un tipo adecuado de plataforma en base a un procesador que está conectada a una red 106 y que interactúa con uno o más programas de aplicación. Los dispositivos cliente 102a-n pueden operar en cualquier sistema operativo capaz de soportar un navegador o aplicación habilitada en un navegador, tal como Microsoft® Windows ® o Linux™. Los dispositivos cliente102a-n mostrados incluyen, por ejemplo, ordenadores personales que ejecutan un programa de aplicación de navegador tal como Internet Explorer™ de Microsoft Corporation, Netscape Navigator™ de Netscape Communication Corporation, y Safari™ de Apple® Computer, Inc.

[0014] Mediante los dispositivos cliente102a-n, los usuarios 112a-n pueden comunicarse por la red 106 entre sí y con otros sistemas y dispositivos acoplados a la red 106. Como se muestra en la Figura 1, un dispositivo servidor 104 y un dispositivo servidor 150 se acoplan también a la red 106.

[0015] El dispositivo servidor 104 puede comprender un servidor que ejecuta un programa de aplicación de motor de segmentación y un dispositivo servidor 150 puede comprender un servidor que ejecuta un programa de aplicación de motor de búsqueda. De forma similar a los dispositivos cliente 102a-n, el dispositivo servidor104 y el dispositivo servidor 150 mostrados en la Figura 1 comprenden un procesador 116 acoplado a una memoria legible por ordenador 118 y un procesador 152 acoplado a una memoria legible por ordenador 154, respectívamente. Los dispositivos servidores 104 y 150, descritos como sistemas informáticos únicos, pueden implementarse como una red de procesadores informáticos. Ejemplos de dispositivos servidores 104,150 son servidores, ordenadores centrales, ordenadores en red, un dispositivo en base a un procesador, y tipos similares de sistemas y dispositivos. El procesador cliente 110 y los procesadores servidores 116, 152 pueden ser cualquiera de una serie de procesadores informáticos, como antes se describe, como procesadores de Intel® Corporation de Santa Clara, California y Motorola® Corporation de Schaumburg los, Illinois.

[0016] La memoria 118 contiene un programa de aplicación de segmentación, también conocido como un motor de

segmentación 120. El dispositivo servidor 104, o dispositivo relacionado, puede acceder a la red 106 para recibir series de caracteres desde otros dispositivos o sistemas conectados a la red 106. Los caracteres pueden incluir, por ejemplo, marcas o símbolos utilizados en un sistema de escritura, que incluyen datos que representan un carácter, tales como ASCII, Unicode, ISO 8859-1, Shift-JIS, y EBCDIC o cualquier otro conjunto de caracteres adecuado. En un ejemplo, el motor de segmentación 120 puede recibir una cadena de caracteres, tal como un nombre de dominio, desde un dispositivo servidor en la red 106 cuando un usuario 112a intenta dirigir una aplicación de navegador web a un nombre de dominio que no está activo.

[0017] In un ejemplo, el motor de segmentación 120 identifica idiomas candidatos para la cadena de caracteres, segmenta la cadena de caracteres en combinaciones potenciales de identificadores para cada idioma candidato, y selecciona un idioma particular y combinación para asociar con la cadena de caracteres. Un identificador puede comprender una palabra, un nombre propio, un nombre geográfico, una abreviatura, un acrónimo, un símbolo de denominación abreviada del mercado de valores, u otros identificadores. El motor de segmentación 120 puede incluir un procesador de segmentación 122, un procesador de frecuencia 124, y un procesador de idioma 126. En el ejemplo mostrado en la Figura 1, cada uno comprende el código de ordenador que reside en la memoria 118.

[0018] El procesador de idioma 126 puede identificar un idioma o idiomas candidatos para la cadena de caracteres. En un ejemplo, el procesador de idioma 126 puede utilizar señales para identificar una serie de idiomas candidatos para la cadena de caracteres. Por ejemplo, el procesador de idioma puede usar linguística, la dirección IP del usuario, un conjunto de caracteres utilizado para la cadena de caracteres, configuración del navegador de un programa de aplicación de navegador asociado con el usuario, y un dominio de primer nivel asociado con la cadena de caracteres para determinar los idiomas candidatos para la cadena de caracteres.

[0019] El procesador de segmentación 122 puede determinar una lista de combinaciones potenciales de identificadores o resultados segmentados de la cadena de caracteres para cada idioma candidato. En un ejemplo, el procesador de identificador 124 determina una probabilidad para cada resultado segmentado de la lista y selecciona los mejores resultados segmentados para cada idioma en base a la probabilidad. La probabilidad para un resultado segmentado puede estar en base a valores de frecuencia asociados con los identificadores individuales en el resultado. En un ejemplo, la cadena de caracteres sin segmentar puede ser incluida como un resultado segmentado.

[0020] El procesador de frecuencia 124 puede realizar una búsqueda de frecuencia o hacer que se realice una sobre los mejores resultados segmentados seleccionados de cada idioma candidato. El procesador de frecuencia 124 puede incluir una funcionalidad de revisión ortográfica o puede acudir a una funcionalidad de revisión ortográfica que reside en otro lado para que realice una revisión ortográfica sobre los resultados segmentados seleccionados. Cualquier resultado de ortografía corregida puede ser incluido en la búsqueda de frecuencia. En un ejemplo, el procesador de frecuencia envía los resultados segmentados seleccionados al dispositivo servidor 150 para realizar una búsqueda de frecuencia sobre los resultados segmentados seleccionados. Una búsqueda de frecuencia puede determinar la frecuencia de suceso para cada resultado particular segmentado como se describe antes. En base a la búsqueda de frecuencia un resultado segmentado mejor u operable puede ser identificado por el procesador de segmentación 122. El idioma asociado con el resultado operable puede ser identificado por el procesador de segmentación 122 como el idioma operable para la cadena de caracteres. En un ejemplo, el resultado segmentado operable y el idioma operable pueden ser envíados a un servidor de avisos que puede seleccionar avisos dirigidos en base a uno o ambos del idioma operable y el resultado segmentado. Otras funciones y características del procesador de segmentación 122, el procesador de frecuencia 124, y el procesador de idioma 126 se describen adicionalmente más adelante.

[0021] El dispositivo servidor 104 también proporciona acceso a otros elementos de almacenamiento, tal como un elemento de almacenamiento de identificador, en el ejemplo mostrado una base de datos de identificador 120. La base de datos de identificador 120 puede ser utilizada para almacenar identificadores e información de frecuencia asociados con cada identificador. La base de datos de identificador 120 puede también almacenar el idioma o idiomas asociados con cada identificador. Los elementos de almacenarmiento de datos pueden incluír uno cualquiera o la combinación de métodos para almacenar datos, incluyendo sin limitación, distribuciones, tablas hash, listas y pares. El dispositivo servidor 104 puede acceder a otros tipos similares de dispositivos de almacenamiento de datos.

[0022] El dispositivo servidor 150 puede incluír un sevidor que ejecuta un programa de aplicación de motor de búsqueda, tal como el motor de búsqueda Google™. En otros ejemplos, el dispositivo servidor 150 puede comprender un servidor de información relacionado o un servidor de publicidad. En otro ejemplo, puede existir múltiples dispositivos de servidor 150.

[0023] La memoria 154 contiene el programa de aplicación de motor de búsqueda, también conocido como un motor de búsqueda 156. El motor de búsqueda 156 puede localizar información relevante de la red 106 en respuesta a una petición de búsqueda de un usuario 112a y puede mantener un registro de solicitudes de búsqueda. El motor de búsqueda 156 puede también realizar una búsqueda de frecuencia en respuesta a una petición de búsqueda de frecuencia del procesador de frecuencia 124. El motor de búsqueda 156 puede proporcionar un conjunto de resultados de búsqueda a un usuario 112a o información de frecuencia al motor de segmentación 120 por la red 106.

[0024] En un ejemplo, el dispositivo servidor 150, o dispositivo relacionado, ha realizado previamente una inspección de la red 106 para localizar artículos, tales como páginas web, almacenadas en otros dispositivos o sistemas acoplados a

la red 106. Los artículos incluyen, por ejemplo, documentos, correos electrónicos, mensajes de mensajería instantánea, entradas de bases de datos, páginas web de varios formatos, tal como HTML, XML, XHTML, archivos PDF, y archivos de medios, tal como archivos de imagen, archivos de audio, y archivos de video, o cualquier otro documento o grupo de documentos o información de cualquier tipo adecuado de lo que sea. Un indexador 158 puede ser utilizado para indexar los artículos en la memoria 154 u otro dispositivo de almacenamiento de datos, tal como un índice 160. El índice puede incluír también el idioma o idiomas asociados con cada artículo. En una realización, existen múltiples índices conteniendo cada uno una parte del total de los artículos indexados. Debería apreciarse que puedan utilizarse otros métodos adecuados para indexar artículos en lugar de o en combinación con la inspección, tal como el depósito manual.

[0025] El motor de búsqueda 156 puede realizar una búsqueda de frecuencia en una serie de formas apropiadas. En un ejemplo, el motor de búsqueda 156 puede realizar una búsqueda web utilizando cada mejor resultado segmentado seleccionado como una solicitud de búsqueda y puede buscar artículos que contengan la solicitud de búsqueda en el idioma candidato del resultado segmentado. En este ejemplo, puede generarse un conjunto de resultados de búsqueda de frecuencia y puede comprender uno o más identificadores de artículos. Un identificador de artículo puede ser, por ejemplo un Localizador Uniforme de Recursos (URL), un nombre de archivo, un vínculo, un icono, una ruta para un archivo local, o cualquier otro que identifique un artículo. En un ejemplo, un identificador de artículos puede comprender una URL asociada con un artículo. El procesador de frecuencia 124 puede utilizar el número de identificadores de artículos en cada conjunto de resultados de búsqueda de frecuencia como una representación del número de sucesos del respectivo resultado segmentado.

[0026] En otro ejemplo, el procesador de frecuencia 124 puede interconectarse directamente con el indizador 158. El indizador 158 puede determinar, para cada mejor resultado segmentado seleccionado, el número de artículos en el idioma candidato asociado en el que aparece el resultado segmentado. Esta información puede enviarse al procesador de frecuencia 124. En otro ejemplo más, el motor de búsqueda 156 y/o el procesador de frecuencia 124 pueden determinar, para cada resultado segmentado, el número de sucesos en solicitudes de búsqueda en el idioma candidato asociado a partir del registro de búsqueda y el procesador de frecuencia 124 puede determinar una frecuencia de suceso en base a la información de registro de búsqueda. En una realización, el número de artículos o peticiones de búsqueda en una búsqueda de frecuencia asociada con un resultado segmentado pueden ser normalizados en base al número total de artículos o solicitudes de búsqueda en el idioma asociado.

[0027] Debiera observarse que los sistemas pueden tener distinta arquitectura que la que se muestra en la Figura 1. Por ejemplo, en algunos sistemas el dispositivo servidor 104 puede comprender un sólo servidor lógico o físico. El sistema 100 mostrado en la Figura 1 es meramente de ejemplo, y es utilizado para ayudar a explicar el método ilustrado en la Figura 2.

Proceso

[0028] Pueden llevarse a cabo diversos métodos de acuerdo a las realizaciones de la presente invención. Un método de ejemplo según la presente invención comprende la identificación de al menos un primer idioma candidato y un segundo idioma candidato asociados con una cadena de caracteres, determinar al menos un primer resultado segmentado asociado con el primer idioma candidato de la cadena de caracteres y un segundo resultado segmentado asociado con el segundo idioma candidato de la cadena de caracteres, determinar una primera frecuencia de suceso para el primer resultado segmentado y una segunda frecuencia de suceso para el segundo resultado segmentado, e identificar un idioma operable a partir del primer idioma candidato y el segundo idioma candidato en base al menos a la primera frecuencia de suceso y la segunda frecuencia de suceso. Pueden identificarse más de dos idiomas candidatos y pueden determinarse más de dos resultados segmentados. Por ejemplo, pueden identificarse tres idiomas candidatos y pueden determinarse cuatro resultados segmentados para cada idioma candidato.

[0029] El idioma operable puede ser identificado en parte en base a la identificación de un resultado segmentado operable del primer resultado segmentado y el segundo resultado segmentado en base al menos en parte a la primera frecuencia de suceso y la segunda frecuencia de suceso. Un primer idioma candidato y un segundo idioma candidato pueden ser identificados basándose en parte en una o más señales de idioma. Las señales de idioma pueden comprender al menos una de linguística asociada con la cadena de caracteres, una dirección IP de un usuario asociado con la cadena de caracteres, un conjunto de caracteres utilizado para la cadena de caracteres, configuraciones del navegador de un programa de aplicación de navegador asociado con un usuario, y un dominio de primer nivel asociado con la cadena de caracteres. En una realización, Identificar el idioma operable puede estar basada al menos en parte en señales de idioma.

[0030] En un ejemplo, Identificar el idioma operable del primer idioma candidato y del segundo idioma candidato basada al menos en parte en la primera frecuencia de suceso y la segunda frecuencia de suceso puede comprender seleccionar el primer idioma candidato si la primera frecuencia de suceso es mayor que la segunda frecuencia de suceso. La cadena de caracteres puede comprender un nombre de dominio. El primer resultado segmentado puede comprender una primera combinación de identificadores y el segundo resultado segmentado comprende una segunda combinación de identificadores.

[0031] En una realización, determinar la primera frecuencia de suceso para el primer resultado segmentado puede

comprender determinar una serie de artículos en el primer idioma candidato que contiene el primer resultado segmentado y normalizar el número de artículos en base a un número total de artículos en el primer idioma candidato y determinar el número de artículos en el primer idioma que contiene el primer resultado segmentado puede comprender determinar identificadores de artículos en un conjunto de resultados de búsqueda generado en respuesta a una petición de búsqueda que comprende el primer resultado segmentado.

[0032] En una realización, determinar el número de partículas en el primer idioma que contienen el primer resultado segmentado puede comprender acceder a un índice de artículos. En otra realización, determinar la primera frecuencia de suceso puede comprender determinar una serie de sucesos del primer resultado segmentado en una pluralidad de solicitudes de búsqueda en el primer idioma candidato y normalizar el número de sucesos en base a un número total de solicitudes de búsqueda en el primer idioma candidato.

[0033] El método puede también comprender seleccionar un artículo en base al menos en parte al idioma operable o el resultado segmentado operable (o ambos) y el artículo puede comprender un anuncio. En una realización, determinar el primer resultado segmentado puede comprender determinar una pluralidad de resultados segmentados en el primer idioma candidato a partir de la cadena de caracteres, y Identificar el primer resultado segmentado de la pluralidad de resultados segmentados en el primer idioma candidato. Identificar el primer resultado segmentado puede comprender calcular un valor de probabilidad para cada una de las pluralidades de resultados segmentados. Un primer valor de probabilidad asociado con el primer resultado segmentado puede estar basado al menos en parte en una frecuencia de cada identificador dentro del primer resultado segmentado.

[0034] Otro método de ejemplo comprende determinar un primer resultado segmentado en un primer idioma candidato y un segundo resultado segmentado en un segundo idioma candidato a partir de un nombre de dominio, determinar una primera frecuencia de suceso para el primer resultado segmentado en al menos uno de un índice de artículos, un índice de texto, yy un conjunto de resultados de búsqueda, determinar una segunda frecuencia de suceso para el segundo resultado segmentado, si la primera frecuencia de suceso es mayor que la segunda frecuencia de suceso, seleccionar después el primer idioma candidato como un idioma operable, si la segunda frecuencia de suceso es mayor que la primera frecuencia de suceso, seleccionar después el segundo idioma candidato como el idioma operable, seleccionar un anuncio basado al menos en parte en el idioma operable, donde el anuncio incluye texto en el idioma operable, y producir una visualización del anuncio en asociación con una página web asociada con el nombre de dominio.

[0035] La Figura 2 ilustra un método de ejemplo 200 para seleccionar un idioma para segmentación de texto, según una realización de la invención. Este método de ejemplo se proporciona a modo de ejemplo, puesto que hay una variedad de maneras de llevar a cabo los métodos según la presente invención. El método 200 mostrado en la Figura 2 puede ser ejecutado o realizado de otro modo por uno o una combinación de diversos sistemas. El método 200 se describe abajo como realizado por el sistema 100 mostrado en la Figura 1 a modo de ejemplo, y diversos elementos del sistema 100 se referencian al explicar el método de ejemplo de la Figura 2.

[0036] Haciendo referencia a la Fig. 2, en el bloque 202, el método de ejemplo comienza. El bloque 202 es seguido del bloque 204, en el que puede accederse a una cadena de caracteres por el motor de segmentación 120. Puede recibirse o accederse a una cadena de caracteres a partir de un dispositivo conectado a la red 106, por ejemplo, o desde otros dispositivo. En una realización, la cadena de caracteres puede ser un nombre de dominio asociado con una página web inactiva o inexistente recibida desde un servidor de publicidad asociado con el nombre de dominio.

[0037] El bloque 204 es seguido del bloque 205, en el que son identificados idiomas candidatos para la cadena de caracteres. En un ejemplo, el procesador de idioma 126 puede utilizar una o más señales de idioma para determinar una serie de idiomas candidatos para la cadena de caracteres. Por ejemplo, el procesador de idioma puede identificar, basado en las señales de idioma, Inglés, Francés y Español como los tres idiomas candidatos para la cadena de caracteres.

[0038] Algunas de las señales de idioma utilizadas pueden ser, por ejemplo, linguísitica asociada con la cadena de caracteres, la dirección IP de un usuario asociada con la cadena de caracteres, el conjunto de caracteres asociado utilizado para la cadena de caracteres, configuraciones de navegador de un programa de aplicación de navegador asociadas con el usuario asociado con la cadena de caracteres, y un dominio de primer nivel asociado con la cadena de caracteres. La linguística puede utilizarse, por ejemplo, para determinar si la estructura o naturaleza de la cadena de caracteres indica que está en un particular idioma. Por ejemplo, ciertos idiomas tienen una tendencia a empezar o terminar con un cierto grupo de caracteres y utilizar modelos generales. La dirección IP del usuario puede indicar la localización y país del usuario. A partir de la información del país un idioma o idiomas asociados con el país pueden usarse como idiomas candidatos. El conjunto de caracteres de la cadena de caracteres puede indicar un idioma o idiomas asociados con la cadena de caracteres. Por ejemplo, un conjunto de caracteres cirílicos puede indicar ruso o algún otro idioma eslavo. Las configuraciones del navegador para un programa de aplicación de navegador de un usuario asociado con la cadena de caracteres pueden indicar unas configuraciones de idioma y conjunto de caracteres del programa de aplicación del navegador del usuario que pueden ser pasadas en una cabecera HTTP junto con la cadena de caracteres. Un dominio de primer nivel asociado con la cadena de caracteres puede indicar un país. Un dominio de primer nivel puede ser el mayor nivel de la jerarquía después de la raíz. En un nombre de dominio, el dominio de alto nivel es la parte del nombre de dominio que aparece más lejos a la derecha. Por ejemplo, para el nombre de dominio "usedrugs.co.uk", el dominio de alto nivel es ".uk" y puede indicar el Reino Unido. El dominio de alto

nivel ".ru" puede indicar Rusia. El país asociado con el dominio de alto nivel puede ser utilizado al determinar un idioma candidato, tal como "ru" indica Rusia, lo cual indica que la cadena de caracteres asociada puede estar en el idioma ruso. Algunos dominios de alto nivel pueden indicar más de un idioma. Por ejemplo, ".ch" puede indicar Suiza y puede indicar que la cadena de caracteres puede estar asociada con francés, alemán, o Italiano. Pueden utilizarse otras señales y métodos adecuados de identificación de los idiomas candidatos para la cadena de caracteres.

[0039] El bloque 206 está seguido por el bloque 208, en el que una pluralidad de resultados segmentados se genera a partir de la cadena de caracteres segmentando la cadena de caracteres para cada uno de los idiomas candidatos. La segmentación de la cadena de caracteres puede incluir el análisis sintáctico de los caracteres en la cadena en una pluralidad de combinaciones de identificadores y puede ser realizada por el procesador de segmentación 122. El procesador de segmentación 122 puede desarrollar una lista de resultados segmentados para cada idioma candidato. Cada resultado segmentado puede ser una combinación particular de identificadores o un sólo identificador. Por ejemplo, la cadena de caracteres "assocomunicazioni" puede ser segmentada en italiano en "asso comunicazioni" y otros resultados segmentados y puede ser segmentada en francés en "asso com uni cazioni" y otros resultados segmentados. En otro ejemplo, la cadena de caracteres "maisonblanche" puede ser segmentada en francés en "maison blanche" y otros resultados segmentados y puede ser segmentada en inglés en "mai son blanc he" y otros resultados segmetnados. En otro ejemplo, la cadena de caracteres "usedrugs" puede ser segmentada en inglés en los resultados segmentados incluyendo "used rugs", "use drugs", "us ed rugs", "u sed rugs", "usedrugs", etc. Resultados segmentados pueden ser generados para los otros idiomas candidatos, tal como, en el ejemplo anterior, francés y español. La cadena de caracteres sin segmentar puede ser incluida como un resultado segmentado.

[0040] El procesador de segmentación 122 puede utilizar identificadores de la base de datos de identificadores 126 en el proceso de segmentación. Pueden utilizarse diversos métodos para segmentar la cadena de caracteres, tales como las técnicas de segmentación descritas en la publicación de Patente Internacional PCT Nº WO2005/069199 titulada "Métodos y Sistemas de Segmentación de Texto " presentada el 30 de diciembre de 2003.

[0041] El bloque 208 está seguido del bloque 210, en el que se determinan los mejores resultados segmentados para cada idioma candidato. Los mejores resultados segmentados pueden ser determinados por el procesador de segmentación 122 y pueden ser los resultados con mayor probabilidad de ser el resultado segmentado mejor u operable. En una realización, los resultados segmentados pueden ser clasificados en base a un valor de probabilidad determinado para cada resultado segmentado. En una realización, un valor de probabilidad puede ser determinado sumando los valores de frecuencia asociados con los identificadores individuales en cada resultado individual segmentado. En otra realización, un valor de probabilidad puede ser determinado por una función compleja que implica sumar los logaritmos de los valores de frecuencia asociados con los identificadores individuales en cada resultado individual segmentado. Puede seleccionarse luego una serie de los resultados segmentados mejor clasificados. Por ejemplo, pueden clasificarse los resultados segmentados para cada idioma candidato y pueden seleccionarse los tres mejores resultados de cada idioma candidato.

[0042] El bloque 210 está seguido del bloque 212, en el que se realiza una búsqueda de frecuencia para los resultados segmentados mejor seleccionados para cada idioma candidato. La búsqueda de frecuencia puede ser realizada por el procesador de frecuencia 124 conjuntamente con el motor de búsqueda 156. En un ejemplo, el procesador de segmentación 122 puede pasar los resultados segmentados seleccionados al procesador de frecuencia 124, el que puede determinar la frecuencia de suceso para cada uno de los resultados segmentados en una colección de artículos o solicitudes de búsqueda.

[0043] Por ejemplo, el procesador de frecuencia 124 puede determinar la frecuencia de suceso para los resultados segmentados en base a los artículos indizados por un motor de búsqueda 156. En un ejemplo, el procesador de frecuencia 124 puede enviar los mejores resultados segmentados seleccionados al motor de búsqueda 156 por la red 106. El motor de búsqueda 156 puede realizar una búsqueda para cada uno de los resultados segmentados en los artículos indizados utilizando cada resultado segmentado como una solicitud de búsqueda. Por ejemplo, el procesador de frecuencia 124 puede enviar cada resultado segmentado para cada idioma candidato rodeado por comillas al motor de búsqueda 156 como una solicitud de búsqueda, de manera que el motor de búsqueda 156 realice la búsqueda sobre la frase exacta segmentada en artículos en el idioma concreto. En un ejemplo, para cada resultado segmentado, el motor de búsqueda 156 puede generar un conjunto de resultados de búsqueda que contiene una serie de identificadores de artículo en respuesta a la solicitud de búsqueda. El motor de búsqueda 156 puede enviar el conjunto de resultados de búsqueda para cada uno de los resultados segmentados de vuelta al procesador de frecuencia 124 por la red 106. El procesador de frecuencia 124 puede determinar a partir de cada conjunto de resultados de búsqueda, en base al número de identificadores de artículo, la frecuencia con la que sucede cada resultado segmentado.

[0044] En otro ejemplo, el procesador de frecuencia 124 puede enviar los mejores resultados segmentados seleccionados al indizador 158 por la red 106. El indizador 158 puede acceder al índice 160 para determinar el número de artículos en el particular idioma en el que se presenta un resultado segmentado y puede hacerlo para cada uno de los resultados segmentados seleccionados. En un ejemplo, el índice 160 pueden ser múltiples índices y el indizador 158 puede revisar una fracción del índice total para cada resultado segmentado. El indizador 158 puede después pasar el número de sucesos asociados con cada resultado segmentado al procesador de frecuencia 124 por la red 106.

[0045] En otro ejemplo más, el procesador de frecuencia 124 puede enviar los mejores resultados segmentados

seleccionados al motor de búsqueda 156 por la red 106 para determinar el número de sucesos de los resultados segmentados en las solicitudes de búsqueda. Por ejemplo, el motor de búsqueda 156 puede, para cada resultado segmentado en el idioma asociado, determinar el número de veces que el resultado segmentado fue utilizado como una solicitud de búsqueda o parte de una solicitud de búsqueda. El número de sucesos en las solicitudes de búsqueda para cada resultado segmentado puede enviarse por el motor de búsqueda 156 al procesador de frecuencia 124 por la red 106.

[0046] Por ejemplo, si el procesador de segmentación 122 determina que los resultados segmentados seleccionados para la cadena de caracteres "usedrugs" en inglés son "used rugs", "use drugs", y "us ed rugs", el procesador de frecuencia 124 puede enviar estos resultados segmentados y los resultados segmentados asociados con otros idiomas candidatos al motor de búsqueda 156. El motor de búsqueda 156 puede, por ejemplo, utilizar estos resultados como solicitudes de búsqueda y generar conjuntos de resultados de búsquedas para cada resultado segmentado. Por ejemplo, el motor de búsqueda 156 puede utilizar "used rugs" como una solicitud de búsqueda y determinar el conjunto de resultados de búsqueda para la solicitud de búsqueda que contiene los identificadores de artículos asociados con los artículos en inglés que contienen la frase "used rugs". El motor de búsqueda 156 puede hacer lo mismo para los resultados segmentados asociados con otros idiomas candidatos. En otro ejemplo, el motor de búsqueda 156 puede determinar, a partir de registros de búsqueda asociados que contienen solicitudes anteriores de búsqueda recibidas, el número de veces que se recibieron solicitudes de búsqueda que contienen los resultados segmentados. Por ejemplo, el motor de búsqueda 156 puede buscar en sus registros de búsqueda el número de veces que fue recibida una solicitud de búsqueda que contiene la frase "used rugs". En otro ejemplo más, el indizador 158 del motor de búsqueda 156 puede recibir los resultados de búsqueda y determinar el número de artículos en el índice 160 o una parte del índice 160 que contienen los resultados segmentados. Por ejemplo, el indizador 158 puede buscar por el índice 160 o una parte del índice 160 el número de artículos del idioma inglés que contengan "used rugs".

[0047] Puede incluirse una función de corrección ortográfica en la búsqueda de frecuencia. Por ejemplo, el procesador de frecuencia 124 puede incluir o acudir a una función de corrección ortográfica, de modo que los mejores resultados segmentados seleccionados puedan ser corregidos ortográficamente. La función de corrección ortográfica puede determinar escrituras preferidas o correctas para los identificadores individuales en cada resultado segmentado. El procesador de frecuencia 124 puede realizar una búsqueda de frecuencia sobre los resultados segmentados así como cualquier resultado segmentado con escritura corregida para determinar la frecuencia de suceso para ambos resultados. Por ejemplo, si un resultado segmentado es “basebal game” y el resultado corregido de la escritura es "baseball game", puede realizarse una búsqueda de frecuencia para ambos resultados.

[0048] En una realización, cada frecuencia de suceso para los resultados segmentados es un valor normalizado basado en el número de artículos o solicitudes totales de búsqueda en el particular idioma. Por ejemplo, si un resultado segmentado en el idioma inglés sucede en 70 artículos o solicitudes de búsqueda de idioma inglés y existe un número total de 1000 artículos o solicitudes de búsqueda en inglés, la frecuencia de suceso para este resultado segmentado en inglés es 0,07 (70/1000). Similarmente, si un resultado segmentado en francés sucede en 60 artículos o solicitudes de búsqueda de idioma francés y existe un número total de 400 artículos o solicitudes de búsqueda en idioma francés, la frecuencia de suceso para este resultado segmentado en francés es 0,15 (60/400). De este modo la frecuencia de suceso tiene en cuenta la prevalencia del particular idioma en el cuerpo de los artículos o resultados de búsqueda y no se sopesa intrínsecamente con idiomas más comunes.

[0049] El bloque 212 es seguido del bloque 214, en el que se identifican el idioma operable y los resultados segmentados operables. En un ejemplo el procesador de frecuencia 124 puede identificar el idioma operable y el resultado segmentado operable. Por ejemplo, el procesador de frecuencia 124 puede seleccionar el resultado segmentado que tiene la frecuencia asociada de suceso más elevada. Como se explica antes, la frecuencia de suceso puede ser un valor normalizado basado en el número de artículos o solicitudes de búsqueda que contienen el resultado segmentado y el número total de artículos o solicitudes de búsqueda en el particular idioma. Pueden utilizarse señales adicionales para determinar el resultado segmentado operable. Por ejemplo, el procesador de frecuencia 124 puede tener en cuenta una clasificación objetiva (tal como el algoritmo de clasificación PageRank™ para artículos web) de los artículos que contienen cada resultados segmentado y utilicen la clasificación objetiva para valorar los artículos que contienen cada resultado segmentado. El número de veces que el resultado segmentado sucede en un artículo y la localización del resultado segmentado en los artículos pueden usarse para valorar los artículos que contienen un resultado segmentado. El idioma candidato asociado con el resultado segmentado operable puede seleccionarse como el idioma operable.

[0050] En una realización, las señales de idioma utilizadas para identificar los idiomas candidatos en el bloque 206 pueden usarse para determinar el idioma operable. Si las señales de idioma indican que la cadena de caracteres es muy probablemente un idioma particular, estas señales pueden usarse para valorar más este idioma. Por ejemplo, las señales de idiomas, tal como linguística, la dirección IP de un usuario asociado, el conjunto de caracteres utilizado para la cadena de caracteres, configuración del navegador del programa de aplicación del navegador asociado con un usuario, y el dominio de alto nivel asociado con la cadena de caracteres, pueden indicar que el idioma asociado con la cadena de caracteres es un particular idioma, tal como el francés, por ejemplo. La frecuencia de información de suceso para un resultado segmentado en otro idioma, tal como inglés, por ejemplo, puede estar cerca de o exceder la frecuencia de la información de suceso para otro resultado segmentado en francés. Las señales de idioma pueden usarse para valorar el idioma de francés con el fin de causar la selección del francés como el idioma operativo en este

ejemplo. En 216, el método 200 finaliza.

[0051] El idioma operable y el resultado segmentado operable pueden usarse en una variedad de modos. El idioma operable y/o resultado segmentado operable pueden usarse para seleccionar anuncios. Por ejemplo, un usuario 112a puede intentar navegar en su aplicación de navegador a la página "usedrugs.com" introduciendo esta cadena de caracteres en la aplicación del navegador. Si no existe dicha página web en el nombre de dominio "usedrugs.com", la aplicación del navegador del usuario puede redireccionarse a una página web de un tercero. La página web de un tercero puede desear colocar anuncios y/o vínculos relevantes para el nombre de dominio introducido por el usuario en una página web que ve el usuario. La página web del tercero puede enviar el nombre de dominio "usedrugs.com" al motor de segmentación 120. El motor de segmentación 120 puede utilizar los métodos y sistemas descritos antes para devolver un idioma operable y un resultado segmentado operable a la página web del tercero o un servidor de publicidad asociado con la página web. Por ejemplo, el resultado segmentado operable puede ser "used rugs" y el idioma operable puede ser el inglés. La página web o servidor de publicidad del tercero puede causar la visualización de anuncios y/o vínculos relevantes para la frase "used rugs" en inglés en la página web vista por el usuario y puede asegurar que el idioma utilizado en la website es el inglés. El idioma operable puede utilizarse en la selección del idioma utilizado en mensajes de estado mostrados al usuario.

General

[0052] Mientras la descripción anterior contiene muchas particularidades, estas particularidades no deberían ser interpretadas como limitaciones del ámbito de la invención, sino meramente como ejemplificaciones de las realizaciones reveladas. Los expertos en la técnica imaginarán cualesquiera otras posibles variaciones que estén dentro del ámbito de la invención. Los términos primero y segundo se utilizan aquí meramente para diferenciar un artículo de otro artículo. Los términos primero y segundo no son utilizados para indicar primero o segundo en el tiempo, primero o segundo en una lista, u otro orden, a menos que se indique explícitamente. Por ejemplo, el "segundo" puede venir en el tiempo o en una lista antes que el “primero”, a menos que se indique explícitamente de otro modo.

REIVINDICACIONES

1. Un método implementado en ordenador (200), que comprende:

recibir (204) una cadena de cadena de caracteres que no tiene interrupciones de delimitación de identificadores;

identificar (206) al menos un primer idioma candidato y un segundo idioma candidato para la cadena de caracteres;

determinar (208) al menos un primer resultado segmentado que comprende una primera pluralidad de identificadores asociados con el primer idioma candidato que incluye la cadena de caracteres y un segundo resultado segmentado que comprende una segunda pluralidad de identificadores asociados con el segundo idioma candidato que incluye la cadena de caracteres;

determinar (210, 212) una primera frecuencia de suceso del primer resultado segmentado en al menos uno de índices o registros del motor de búsqueda de las solicitudes de búsqueda recibidas por un motor de búsqueda (156) y una segunda frecuencia de suceso del segundo resultado segmentado en al menos uno de un índice o registros de motor de búsqueda de solicitudes de búsqueda recibidas por un motor de búsqueda (156); e

identificar (214) un idioma operable del primer idioma candidato y el segundo idioma candidato en base al menos en parte a la primera frecuencia de suceso y la segunda frecuencia de suceso.

2. El método implementado en ordenador (200) de la reivindicación 1, donde identificar (206) el primer idioma candidato comprende Identificar el primer idioma candidato en base al menos a una señal de idioma seleccionada del grupo compuesto de una dirección IP de un usuario asociado con la cadena de caracteres, un conjunto de caracteres asociado con la cadena de caracteres, una configuración de navegador de un programa de aplicación del navegador asociado con un usuario asociado con la cadena de caracteres, y un dominio de primer nivel asociado con la cadena de caracteres.

3. El método implementado en ordenador (200) de la reivindicación 2, donde identificar (206) el idioma operable se basa al menos en parte en al menos una señal de idioma.

4. El método implementado en ordenador (200) de la reivindicación 1, donde determinar (212) la primera frecuencia de suceso del primer resultado segmentado en el índice del motor de búsqueda comprende normalizar la primera frecuencia en base a una serie de entradas de índice del motor de búsqueda correspondientes al primer idioma candidato y determinar la primera frecuencia de suceso del primer resultado segmentado en registros de solicitudes de búsqueda recibidas por el motor de búsqueda comprende normalizar la primera frecuencia en base a una serie de solicitudes de búsqueda en los registros correspondientes al primer idioma candidato.

5. El método implementado en ordenador (200) de la reivindicación 1, que además comprende suministrar un anuncio seleccionada en base al idioma operable.

6. El método implementado en ordenador (200) de la reivindicación 1, donde determinar (212) el primer resultado segmentado comprende:

determinar una pluralidad de resultados segmentados en el primer idioma candidato a partir de la cadena de caracteres, donde cada resultado segmentado contiene una pluralidad diferente de identificadores que cada uno de los otros resultados segmentados; e

identificar el primer resultado segmentado a partir de la pluralidad de resultados segmentados en base a un valor de probabilidad asociado con cada una de la pluralidad de resultados segmentados.

7. El método implementado en ordenador (200) de la reivindicación 6, donde un primer valor de probabilidad asociado con el primer resultado segmentado está basado al menos en parte en una frecuencia de cada identificador en cada primer resultado segmentado.

8. El método implementado en ordenador (200) de la reivindicación 1, que además comprende suministrar un anuncio seleccionado en base al idioma operable.

9. El método implementado en ordenador (200) de la reivindicación 1, donde determinar (212) la primera frecuencia de suceso comprende emplear el motor de búsqueda (156) para identificar una serie de artículos en el primer idioma candidato que responden a una primera solicitud que contiene el primer resultado segmentado y determinar la segunda frecuencia de suceso que comprende el empleo del motor de búsqueda (156) para identificar una serie de artículos en el segundo idioma candidato que responden a una segunda solicitud que contiene el segundo resultado segmentado.

10. El método implementado en ordenador (200) de la reivindicación 9, donde determinar (212) la primera frecuencia comprende normalizar la primera frecuencia en base a un número de artículos totales en el primer idioma candidato que están indizados por el motor de búsqueda.

11. El método implementado en ordenador (200) de la reivindicación 9, donde emplear el motor de búsqueda (156) para identificar el número de artículos en el primer idioma candidato comprende ejecutar en el motor de búsqueda una solicitud de búsqueda que contiene el primer resultado segmentado y determinar una serie de identificadores de artículo en un conjunto de resultados generado por el motor de búsqueda como resultado de la ejecución de la solicitud de búsqueda.

12. El método implementado en ordenador (200) de la reivindicación 9, donde emplear el motor de búsqueda (156) para identificar el número de artículos en el primer idioma candidato comprende determinar una serie de entradas en un índice (160) asociado con el motor de búsqueda (156) que corresponde a uno o más de la primera pluralidad de identificadores.

13. El método implementado en un ordenador (200) de la reivindicación 1, que comprende:

seleccionar un anuncio en base al menos en parte al idioma operable, donde el anuncio incluye texto en el idioma operable; y

causar una visualización del anuncio en asociación con una página web asociada con un nombre de dominio.

14. Un medio legible por ordenador que contiene código de programa adaptado, cuando dicho programa se carga en un ordenador, para hacer que el ordenador ejecute el procedimiento (200) de cualquiera de las reivindicaciones 1 a 13.

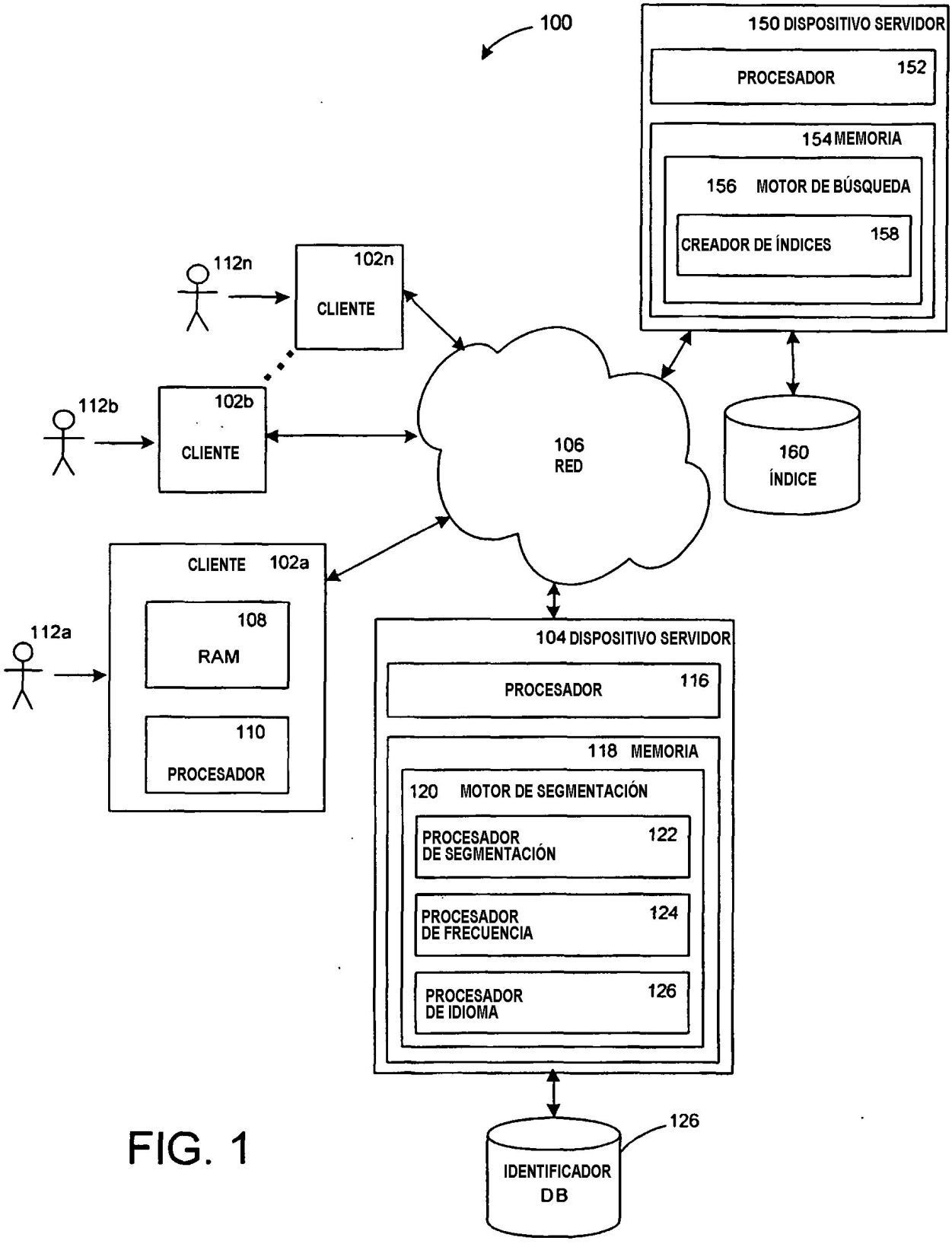
100
150 DISPOSITIVO SERVIDOR
PROCESADOR
152
154 MEMORIA
156 MOTOR DE BÚSQUEDA
CREADOR DE ÍNDICES
158
112n
102n
CLIENTE
112b
102b
CLIENTE
106
RED
160
ÍNDICE
CLIENTE 102a
108
RAM
110
PROCESADOR
112a
104 DISPOSITIVO SERVIDOR
PROCESADOR
116
118 MEMORIA
120 MOTOR DE SEGMENTACIÓN
PROCESADOR DE SEGMENTACIÓN
122
PROCESADOR DE FRECUENCIA
124
PROCESADOR DE IDIOMA
126
126
IDENTIFICADOR DB

FIG. 1

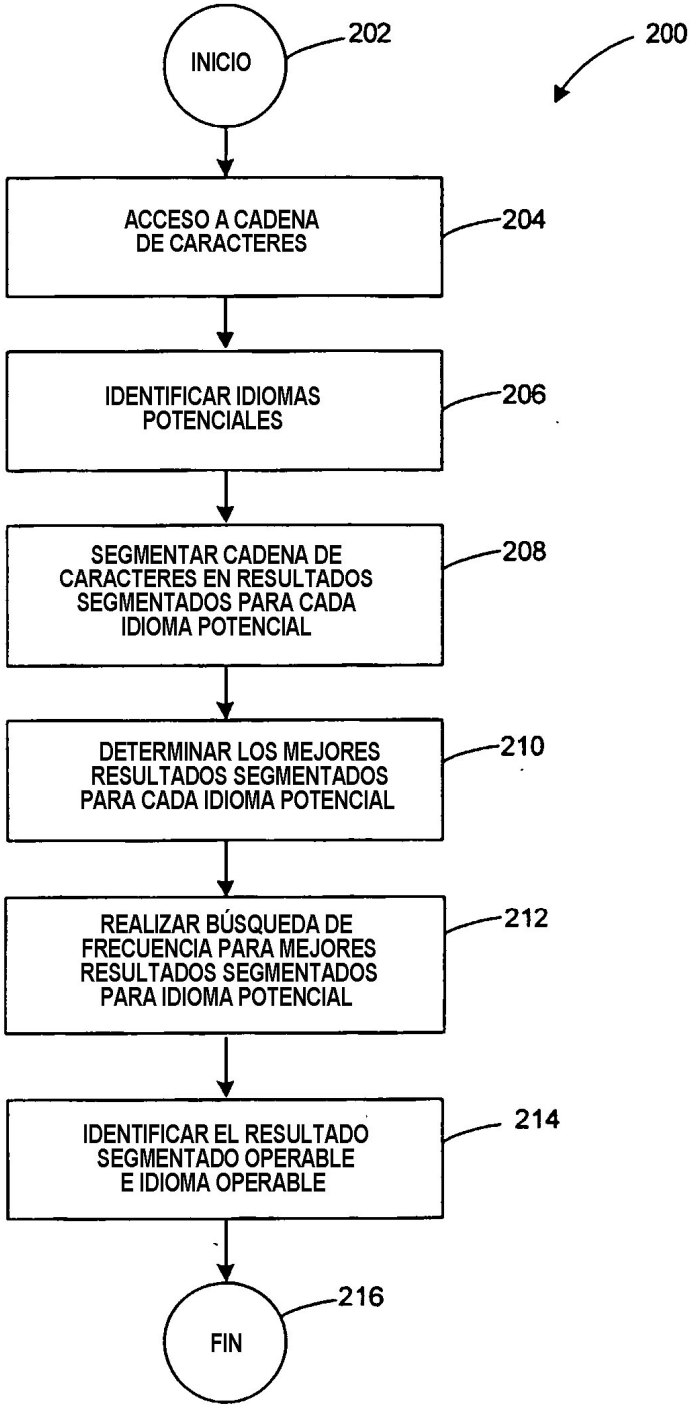
INICIO
202
200
ACCESO A CADENA DE CARACTERES
204
IDENTIFICAR IDIOMAS POTENCIALES
206
SEGMENTAR CADENA DE CARACTERES EN RESULTADOS SEGMENTADOS PARA CADA IDIOMA POTENCIAL
208
DETERMINAR LOS MEJORES RESULTADOS SEGMENTADOS PARA CADA IDIOMA POTENCIAL
210
REALIZAR BÚSQUEDA DE FRECUENCIA PARA MEJORES RESULTADOS SEGMENTADOS PARA IDIOMA POTENCIAL
212
IDENTIFICAR EL RESULTADO SEGMENTADO OPERABLE E IDIOMA OPERABLE
214
FIN
216

FIG. 2