(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 378 653**

(51) Int. Cl.:
G06F 17/30 (2006.01)
G06F 17/20 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Número de solicitud europea: **02799974 .7**
(96) Fecha de presentación : **20.12.2002**
(97) Número de publicación de la solicitud: **1474759**
(97) Fecha de publicación de la solicitud: **10.11.2004**

(54) Título: **Sistemas, métodos y software para hipervínculos automáticos de nombres de personas en documentos para directorios profesionales.**

(30) Prioridad: **21.12.2001 US 342956 P**
13.06.2002 US 171170

(45) Fecha de publicación de la mención BOPI:
16.04.2012

(45) Fecha de la publicación del folleto de la patente:
16.04.2012

(73) Titular/es: **Thomson Reuters Global Resources**
Landis + Gyr-Strasse 3
6300 Zug, CH

(72) Inventor/es: **Dozier, Christopher, C.**

(74) Agente/Representante:
Isern Cuyas, María Luisa

DESCRIPCIÓN

Sistemas, métodos, y software para hipervínculos automáticos de nombres de personas en documentos para directorios profesionales.

La presente invención se refiere a sistemas, métodos y software para establecer hipervínculos de nombres en documentos.

En los últimos años, el fantástico crecimiento de Internet y otras redes informáticas ha provocado un crecimiento igualmente fantástico en los datos accesibles a través de estas redes. Uno de los modos seminales de interaccionar con estos datos es mediante el uso de hipervínculos dentro de documentos electrónicos.

Los hipervínculos son elementos seleccionabas por el usuario, tales como texto resaltado o iconos, que vinculan una parte de un documento electrónico a otra parle del mismo documento o a otros documentos de una base de datos o red informática. Con un equipo informático y un acceso a la red adecuados, un usuario puede seleccionar o invocar un hipervínculo y ver casi instantáneamente el otro documento, que puede hallarse en casi cualquier parte del mundo. Además, el otro documento mismo puede incluir hipervínculos a otros documentos más que incluyan hipervínculos, permitiendo al usuario "brincar" por todo el mundo de documento en documento en documento buscando la información pertinente a voluntad.

Más recientemente se ha despertado el interés en establecer hipervínculos de unos documentos a otros basándose en los nombres de personas que aparecen en los documentos. Por ejemplo, para facilitar las investigaciones legales, la West Publishing Company de St. Paul, Minnesota, proporciona miles de resoluciones judiciales electrónicas con hipervínculos de los nombres de abogados y jueces a sus entradas biográficas en línea en el West Legal Directory, un directorio registrado de aproximadamente 1.000.000 de abogados estadounidenses y 20.000 jueces. Estos hipervínculos permiten a los usuarios acceder a resoluciones judiciales para lograr rápidamente el contacto y otra información específica de letrados y jueces mencionados en las resoluciones.

Los hipervínculos de estas resoluciones judiciales se generan automáticamente, utilizando un sistema que maneja nombres de pila, segundos nombres de pila y apellidos; nombre, ciudad y estado del bufete de abogados; e información del tribunal como indicaciones para vincular los abogados y jueces mencionados a sus correspondientes entradas en el directorio profesional. Véase Christopher Dozier y Robert Haschart, "Automatic Extraction and Linking of Person Names in Legal Text" (Proceedings of RIAO 2000: Content Based Multimedia Information Access. París, Francia. Páginas 1.305-1.321. Abril de 2000).

Aunque el sistema automatizado es muy eficaz, el presente inventor percibió que adolece de como mínimo dos limitaciones. En primer lugar, el sistema aprovecha características estructurales (organizativas) de las resoluciones judiciales, tales como los encabezamientos de caso, que no son comunes a otros documentos y limitan así su aplicación general a otros tipos de nombres y documentos. En segundo lugar, el sistema trata todos los nombres como igualmente ambiguos, o igualmente comunes, cuando, de hecho, algunos nombres son más o menos ambiguos que otros. Por ejemplo, el nombre David Smith es más común que el nombre Seven Drake y por lo tanto más ambiguo, o tiene mayor probabilidad de identificar a más de una persona.

Por consiguiente, el presente inventor ha identificado una necesidad de otros métodos para generar hipervínculos para nombres, o más en general de asociar datos que incluyan nombres.

Para abordar ésta y otras necesidades, el inventor ha ideado sistemas, métodos y software que facilitan el establecimiento de hipervínculos, o la asociación, de nombres que aparecen en documentos, tales como artículos informativos, a nombres que aparecen en otras estructuras de datos, tales como registros en directorios profesionales.

De acuerdo con un aspecto de la presente invención, se proporciona un método implementado en ordenador según lo reivindicado en la reivindicación 1.

De acuerdo con otro aspecto de la invención, se proporciona un sistema para añadir un hipervínculo a un documento según lo reivindicado en la reivindicación 8.

Un ejemplo de sistema incluye un módulo de descriptores y un módulo de vinculación. El módulo de descriptores desarrolla modelos descriptivos para seleccionar información que aparezca conjuntamente en el documento, útil para reconocer asociaciones entre nombres y categorías profesionales. El módulo de vinculación etiqueta nombres en un documento de entrada, extrae información que aparece conjuntamente utilizando los modelos descriptivos, clasifica cada nombre como perteneciente a una profesión concreta e intenta encontrar entradas correspondientes en directorios profesionales.

Para encontrar las entradas correspondientes, el módulo de vinculación determina una calificación en cuanto a la rareza (singularidad o ambigüedad) de cada nombre e introduce en una red de inferencia bayesiana esta calificación junto con el nombre y la información que aparece conjuntamente en el documento seleccionada. La red de inferencia mide las probabilidades de que el nombre se refiera a registros (o entradas) candidatos(as) concretos(as) en un directorio profesional determinado. El módulo de vinculación clasifica los registros candidatos basándose en las mediciones

de probabilidad y define un hipervínculo (u otra asociación lógica) basándose en el registro clasificado en la posición más elevada que sobrepase un umbral determinado.

El inventor ha ideado también sistemas, métodos y software que facilitan la búsqueda de datos que incluyan términos potencialmente ambiguos, tales como nombres de personas u otras entidades. Por ejemplo, un método implica recibir una consulta de un usuario, identificar uno o más nombres en la consulta, evaluar la ambigüedad o singularidad de los nombres y, si la ambigüedad es suficientemente grande, obtener información adicional y actualizar o complementar la consulta para ayudar a resolver o reducir la ambigüedad. La información adicional, que por ejemplo incluye un título profesional, una localización o una organización, puede obtenerse directamente del usuario o mediante una búsqueda suplementaria automática.

Breve descripción de los dibujos

La figura 1 es un diagrama de bloques de un ejemplo de un sistema 100 que incorpora enseñanzas de la presente invención.

La figura 2 es un organigrama de un ejemplo de un método para operar el sistema 100 con el fin de definir expresiones o descriptores para el uso en la clasificación y vinculación de nombres.

La figura 3 es un organigrama de un ejemplo de un método para operar el sistema 100 con el fin de definir un hipervínculo entre nombres que aparecen en un documento y nombres que aparecen en una base de datos, basado en la red de inferencia bayesiana formada según la figura 5.

La figura 4 es un diagrama de bloques de un ejemplo de un sistema de inferencia bayesiana utilizado para operar el sistema 100 con el fin de definir hipervínculos.

La figura 5 es un organigrama de un ejemplo de un método para operar el sistema 100 con el fin de formar una red de inferencia bayesiana para el uso en la medición de la probabilidad de que un nombre que aparece en un documento y un nombre que aparece en una base de datos se refieran a la misma persona.

La figura 6 es un organigrama de un ejemplo de un método de búsqueda que incorpora enseñanzas de la presente invención.

La siguiente descripción detallada, que alude a las figuras 1-6 y las incorpora, describe e ilustra uno o más ejemplos de realización de la invención. Estas realizaciones, ofrecidas no para limitar sino sólo para ejemplificar y enseñar la invención, se muestran y describen con un detalle suficiente para permitir a los técnicos en la materia llevar a cabo y utilizar la invención. Así, cuando resulte apropiado para no ofuscar la invención, la descripción puede omitir cierta información ya conocida por el técnico en la materia.

La descripción incluye muchos términos con significados derivados de su uso en la técnica o de su uso dentro del contexto de la descripción. Como ayuda adicional se ofrecen las siguientes definiciones de términos.

Los términos “un” y “una” se refieren a como mínimo uno o una.

El término “o” se utiliza en su sentido lógico booleano, a no ser que se utilice junto con “bien”.

El término “documento” se refiere a todo conjunto lógico o disposición lógica de datos legibles por máquina con un nombre de archivo.

El término “base de datos” incluye todo conjunto lógico o disposición lógica de documentos legibles por máquina.

El término “hipervínculo” incluye todo testigo en un documento que se ajuste estructural o funcionalmente a cualquier norma pasada, presente o futura relativa al Uniform Resource Locator (URL) (localizador uniforme de recursos). También incluye todo testigo que incluya información que identifique un sistema informático o dispositivo en red específico.

El término “nombre” incluye una o más palabras mediante las cuales una entidad, tal como una persona, un animal, un lugar, una cosa, un grupo, una organización o una entidad legal, se denomine y se distinga de otras.

El término “módulos de programa” incluye rutinas, programas, objetos, componentes, estructuras de datos e instrucciones o series de instrucciones, etc., que realicen tareas concretas o implementen tipos abstractos de datos concretos. El término no está limitado en cuanto a un soporte concreto.

Ejemplo de sistema informático para la realización de la invención

La figura 1 muestra un diagrama de un ejemplo de un sistema informático 100 que incorpora un sistema, un método y un software para el marcado automático de una o más partes de un documento y la definición de uno o más hipervínculos correspondientes para cada parte marcada. Aunque el ejemplo del sistema se presenta como un conjunto

interconectado de componentes separados, algunas otras realizaciones implementan su funcionalidad empleando un número mayor o menor de componentes. Además, algunas realizaciones interconectan uno o más componentes mediante redes de área local o redes de gran amplitud por cable o inalámbricas. Algunas realizaciones implementan una o más partes del sistema 100 utilizando uno o más servidores u ordenadores centrales. Así pues, la presente invención no está limitada a ninguna partición funcional en concreto.

En general, un sistema 100 incluye una base de datos de documentos de entrada 110, un subsistema de vinculación de nombres 120, directorios profesionales 130, una base de datos de documentos de salida 140 y dispositivos de acceso 150.

La base de datos de documentos de entrada 110 incluye uno o más documentos electrónicos, de los cuales se muestra como representante un documento 112. El documento 112 incluye uno o más nombres de personas, lugares, cosas o entidades legales (más en general nombres propios), tales como N1, N2, N3, N4 y N5, repartidos por todo el documento. En el ejemplo de realización, el documento 112 es una versión electrónica de un artículo informativo escrito u otro documento de texto, por ejemplo una resolución judicial u otro tipo de documento legal. Sin embargo, en otras realizaciones el documento 112 incluye una o más imágenes o datos multimedia que contienen uno o más nombres.

La base de datos 110 tiene conectado un sistema informatizado de vinculación de nombres 120. El sistema 120 incluye uno o más procesadores convencionales 121, un dispositivo de visualización 122, dispositivos de interfaz 123, dispositivos de comunicación en red 124, dispositivos de memoria 125, un software de procesamiento de documentos 126 y un software de marcado y vinculación 127. El software 126 y 127 incluye diversos componentes de software y de datos que pueden adoptar diversas formas, tales como instrucciones o datos codificados en un soporte eléctrico, magnético y/u óptico, y que pueden instalarse en el sistema 120 por separado o en combinación a través de una descarga de la red o a través de otros métodos de transferencia de software.

Entre los ejemplos de software de procesamiento de documentos se incluyen programas de procesamiento de texto, programas de edición de HTML, programas de hoja de cálculo, programas de correo electrónico, programas de desarrollo de presentaciones, programas de navegación, programas de gestión de documentos y programas de copia de seguridad de archivos. Así pues, la invención no está limitada a ningún género o especie de software de procesamiento de documentos en concreto.

En el ejemplo de realización, el software 127 es una herramienta adicional a un software de procesamiento de documentos 126. Sin embargo, en otras realizaciones funciona como un programa de aplicación independiente, tal como un programa accesible por red, o como parte del kernel o el shell de un sistema operativo. Más en concreto, el software 127 incluye un módulo de descriptores 1271, un módulo de vinculación 1272 y un módulo de formación 1273, descritos todos ellos más abajo con mayor detalle.

El sistema 120 está conectado a directorios profesionales 130 y a una base de datos de documentos de salida 140.

Los directorios profesionales 130 incluyen uno o más directorios profesionales, tales como un directorio de abogados 132, un directorio de jueces 134, una base de datos de expertos 136 y un directorio de otros profesionales 138. Cada directorio (o más en general base de datos) incluye un juego de registros u otras estructuras de datos que contienen información asociada a una o más entidades nominadas o identificadas, tales como personas, lugares, cosas o entidades legales. Por ejemplo, un directorio de abogados 132 incluye cierto número de registros de abogados, tales como el ejemplo de registro de abogados A1; el directorio de jueces 134 incluye cierto número de registros de jueces, tales como el ejemplo de registro de jueces J1; el directorio de expertos 136 incluye cierto número de registros de expertos, tales como el ejemplo de registro de expertos E1; y el directorio de otros incluye cierto número de registros que contienen información asociada a otros individuos, tales como médicos, profesores, contables, profesores, celebridades, etc. Algunas realizaciones pueden incluir bases de datos de teléfonos y direcciones de correo electrónico, informes crediticios, informes fiscales, antecedentes penales, información médica, registros escolares, etc.

La base de datos de documentos de salida 140 incluye uno o más documentos procesados, tales como el ejemplo de documento 142. El documento 142 incluye nombres marcados N1, N2, N3, N4, N5 y los hipervínculos respectivos 1421, 1422, 1423, 1424 y 1425, que se refieren cada uno a un registro biográfico u otra estructura de datos dentro de como mínimo uno de los directorios profesionales 130, o a uno de los directorios profesionales sin indicar un registro concreto del directorio, o a un subconjunto de registros dentro de un directorio. Los hipervínculos 1421-1425, generados por el procesador de vinculación de nombres 120 e incrustados en el documento o asociados de otra manera al mismo, pueden seleccionarse para vincular las respectivas partes de nombre marcado N1, N2, N3, N4, N5 del documento 140 a bases de datos 130, 132 y 134 a través de una red de área local o una red de gran amplitud pública o privada o a través de una vía de transmisión dedicada (no mostrada). El ejemplo de realización presenta los nombres marcados en un color o una fuente que haga contraste, o de otro modo que pueda percibir el usuario, para indicar su asociación con un hipervínculo existente. A los documentos incluidos en la base de datos de salida 140 puede accederse a través de una red de área local o una red de gran amplitud por medio de los dispositivos de acceso 150.

El ejemplo de realización prevé la base de datos de salida 140 como una parte de un servidor web, por ejemplo un Microsoft Internet Information Server 4.0, que funcione en una red de varios servidores con procesadores y memoria

extendida y configuraciones de disco. La base de datos 140 puede tomar cualquier número de formas en diversas plataformas informáticas. Además, en algunas realizaciones, la base de datos 140 incluye un contenido redundante para permitir a más de un dispositivo, como los dispositivos de acceso 150, acceder simultáneamente a múltiples copias del mismo documento.

Los dispositivos de acceso 150 incluyen los ejemplos de dispositivo de acceso 152, 154, 156 y 158. Cada dispositivo de acceso incluye una pantalla, un procesador (uP) y software (SW). El término "dispositivo de acceso", tal y como se utiliza en el presente documento, abarca ordenadores personales equipados con navegador, equipos de red, asistentes digitales personales (PDA), teléfonos, teléfonos móviles, teléfonos web, televisores, televisión web, etc. También incluye monitores y equipos de otro tipo que puedan dar salida a datos en una forma con la que los usuarios u otros ordenadores puedan interaccionar. Así pues, la presente invención no está limitada a ninguna clase o forma concreta de dispositivo de acceso.

Ejemplo de operación del sistema 100

En general, el ejemplo de operación del sistema 100 implica la operación del módulo de descriptores 1271, el módulo de vinculación 1272 y el módulo de formación 1273. El módulo de descriptores 1271 genera una(o) o más estructuras o módulos de descriptores de nombres profesionales para su uso en la identificación de nombres con probabilidad de referirse a individuos dentro de una o más categorías profesionales concretas (o de satisfacer otros criterios predeterminados). El módulo de vinculación 1272 recibe un documento de entrada, por ejemplo el documento 110, y establece hipervínculos de uno o más nombres que aparecen en el documento de entrada a uno o más directorios profesionales, basándose en estructuras de descriptores de nombres profesionales y/u otros datos extraídos del documento 110 e introducidos en una red de inferencia bayesiana. El módulo de formación 1273 define las probabilidades condicionales en diversos nodos de la red de inferencia bayesiana utilizada por el módulo de vinculación 1272.

A. *Estructura y funcionamiento del módulo de descriptores*

Más en particular, la figura 2 muestra un organigrama 200 que ilustra un ejemplo de un método para operar el módulo de descriptores 1271 con el fin de generar descriptores de nombre para una profesión determinada. Una premisa del ejemplo de realización es que algunos nombres personales tienen una probabilidad mucho mayor de pertenecer a un único individuo que otros nombres y que, si tales nombres están también asociados a una profesión (u otra clasificación) común, es posible identificar automáticamente el lenguaje descriptivo común a los miembros de la profesión (o clasificación). Este lenguaje podría emplearse entonces para identificar a la mayoría de los miembros de la profesión (o clasificación) mencionados en el cuerpo.

El organigrama 200 incluye los bloques de proceso 210-260. Aunque estos bloques (y los de otros organigramas de este documento) están dispuestos en serie en el ejemplo de realización, otras realizaciones pueden reorganizar los bloques, omitir uno o más bloques y/o ejecutar dos o más bloques en paralelo empleando múltiples procesadores o un único procesador organizado como dos o más máquinas o subprocesadores virtuales. Además, otras realizaciones incluso implementan los bloques como uno o más módulos específicos de circuitos integrados o de hardware interconectados, con un control relacionado y señales de datos comunicadas entre y a través de los módulos. Así pues, éste y otros ejemplos de flujo de proceso de este documento son aplicables a software, firmware, hardware y otros tipos de implementación.

El bloque 210 implica identificar nombres que aparezcan en un conjunto de documentos (o cuerpos) que coincidan con nombres que aparezcan en un directorio profesional y nombres que no coincidan. La identificación de nombres coincidentes y no coincidentes, es decir nombres "en directorio" y nombres "fuera de directorio", implica identificar todos los nombres que aparecen en el conjunto empleando un programa de etiquetado de nombres y ejecutando a continuación una búsqueda en un directorio profesional, tal como uno de los directorios profesionales 130. Aunque la presente invención no está limitada a ningún género o especie de etiquetadores de nombres, entre los ejemplos de etiquetadores de nombres adecuados se incluye el software de análisis sintáctico NetOwl de IsoQuest, Inc. de Fairfax, Virginia. (El inventor considera un etiquetador de nombres basado en un modelo de entropía máxima para algunas realizaciones).

Una vez ejecutada la búsqueda, el ejemplo de realización identifica un subconjunto de los nombres "en directorio" como nombres "en directorio" poco comunes o únicos. Esto implica calcular una probabilidad de singularidad de nombre para cada nombre "en directorio", basándose la probabilidad de singularidad de nombre en un modelo de lenguaje para los nombres que aparecen en el directorio. El ejemplo de modelo de lenguaje se define en términos de probabilidad de nombre de pila y probabilidad de apellido, basándose la probabilidad de nombre de pila y la probabilidad de apellido en cada caso en la relación del número total de apariciones del nombre de pila y el apellido con respecto al número total de nombres que aparecen en una lista de nombres sacada de la población general.

La lista de nombres debería ser suficientemente grande para representar con precisión la distribución de nombres en la población general. Si el directorio profesional, u otra base de datos, es suficientemente grande, puede utilizarse como base para el modelo de lenguaje. Si el directorio profesional es pequeño, el modelo de lenguaje debería estar basado en alguna otra lista, tal como la lista de profesionales autorizados enumerados en registros públicos de los Estados Unidos. La descripción siguiente supone que el directorio profesional (o base de datos) en cuestión es suficientemente grande para ser representativo de los nombres que aparecen en la población general.

Una vez definido el modelo de lenguaje, se calcula la calificación de probabilidad de coincidencia de nombre para cada nombre "en directorio" empleando

$$P(\text{nombre}) = P(\text{nombre de pila}) \cdot P(\text{apellido}) \qquad (1)$$

donde P(nombre de pila) significa la probabilidad de sacar el nombre de pila al azar de entre todos los nombres de pila que aparecen en el directorio y P(apellido) significa análogamente la probabilidad de sacar el apellido al azar de entre todos los apellidos que aparecen en el directorio. Una probabilidad de singularidad o rareza de nombre se calcula entonces como

$$P(\text{singularidad de nombre}) = \frac{1}{(H \cdot P(\text{nombre}))+1} \qquad (2)$$

donde H significa el tamaño de la población humana con probabilidad de ser citada en el cuerpo. Por ejemplo, para un cuerpo consistente en artículos del Wall Street Journal, H se supone que es 300 millones, la población aproximada de los Estados Unidos. A continuación se utilizan como base todos los nombres "en directorio" con una probabilidad de singularidad de nombre que sobrepase un valor umbral, por ejemplo 0,07, junto con los nombres "fuera de directorio" para el procesamiento ulterior en el bloque 220.

El bloque 220 extrae información que aparece conjuntamente en el documento o asociada de otra manera a uno o más de los nombres "en directorio" identificados y uno o más de los nombres "fuera de directorio". En el ejemplo de realización, esto implica extraer texto o información dentro de cierto intervalo de texto (o región del documento) alrededor de cada uno de los nombres "en directorio" poco comunes y alrededor de todos los nombres "fuera de directorio".

Más en concreto, el ejemplo de extracción implica extraer unigramas y bigramas que aparezcan dentro de un intervalo de texto que se extiende ocho palabras antes y ocho palabras después de cada aparición de nombres "en directorio" y nombres "fuera de directorio" poco comunes identificados en el cuerpo. (Otras realizaciones utilizan otros tamaños y formas de intervalos de texto, tales como estructuras gramaticales u organizativas de documentos. Por ejemplo, algunas realizaciones definen el intervalo basándose en el número de caracteres, oraciones o subdivisiones). Los unigramas y bigramas asociados a nombres "en directorio" se definen como unigramas y bigramas "en directorio" de aparición conjunta, mientras que los asociados a nombres "fuera de directorio" se denominan unigramas y bigramas "fuera de directorio" de aparición conjunta. La ejecución continúa en el bloque 230.

El bloque 230 determina una probabilidad de que la información extraída aparezca con un nombre "en directorio" en lugar de con un nombre "fuera de directorio". En el ejemplo de realización, esto implica calcular la probabilidad de que cada unigrama y bigrama "en directorio" aparezca dentro de un intervalo de ocho palabras antes y después de los nombres "fuera de directorio" y la probabilidad de que cada unigrama y bigrama "fuera de directorio" aparezca dentro de un intervalo de ocho palabras antes y después de los nombres "en directorio". Estas probabilidades de aparición conjunta se calculan de la siguiente manera:

$$P(\text{unigrama/nombre poco común 'en directorio'}) = \frac{EU}{NE} \qquad (3)$$

donde EU = número de veces que un unigrama aparece en el intervalo con un nombre poco común "en directorio" y NE = número de nombres poco comunes "en directorio".

$$P(\text{unigrama/nombre poco común 'fuera de directorio'}) = \frac{FU}{NF} \qquad (4)$$

donde FU significa el número de veces que un unigrama aparece en el intervalo con un nombre "fuera de directorio" y NF significa el número de nombres "fuera de directorio".

$$P(\text{bigrama/nombre poco común 'en directorio'}) = \frac{EB}{NE} \qquad (5)$$

donde EB significa el número de veces que un bigrama aparece en el intervalo con un nombre poco común "en directorio" y NE el número de nombres poco comunes "en directorio".

$$P(\text{bigrama/nombre poco común 'fuera de directorio'}) = \frac{FB}{NF} \qquad (6)$$

donde FB significa el número de veces que un bigrama aparece en el intervalo con un nombre "fuera de directorio" y NF significa el número de nombres "fuera de directorio".

Para determinar la probabilidad de que los unigramas y bigramas aparezcan conjuntamente con un nombre poco común "en directorio" en lugar de un nombre "fuera de directorio", el ejemplo de realización divide la probabilidad de aparición conjunta con un nombre poco común "en directorio" por la respectiva probabilidad de aparición conjunta "fuera de directorio". Estas fórmulas de probabilidad se expresan como

$$P(\text{unigrama}) = \frac{P(\text{unigrama/nombre poco común 'en directorio'})}{P(\text{unigrama/nombre poco común 'fuera de directorio'})} \qquad (7)$$

$$P(\text{bigrama}) = \frac{P(\text{bigrama/nombre poco común 'en directorio'})}{P(\text{bigrama/nombre poco común 'fuera de directorio'})} \qquad (8)$$

donde P(unigrama) es la probabilidad de que un determinado unigrama aparezca conjuntamente con un nombre poco común "en directorio" y P(bigrama) es la probabilidad de que un determinado bigrama aparezca conjuntamente con un nombre poco común "en directorio".

El bloque 240 clasifica la información extraída basándose en las probabilidades de aparición conjunta de unigramas y bigramas anteriores y posteriores. Con este fin, el ejemplo de realización clasifica u ordena los unigramas y bigramas de aparición conjunta "en directorio" en orden descendente según sus probabilidades de aparición conjunta con nombres poco comunes "en directorio" (otras realizaciones pueden calcular y usar probabilidades de aparición conjunta). En las dos tablas siguientes se muestran ejemplos de listas clasificadas de unigramas y bigramas anteriores y posteriores para profesionales legales, junto con sus calificaciones de probabilidad.

TABLA 1

Ejemplos de unigramas anteriores y posteriores

Unigrama anterior	Calificación de probabilidad	Unigrama posterior	Calificación de probabilidad
Solicitor	74,32	Judges	46,45
Judge	44,69	Attorney	22,40
Lawyer	18,94	Lawyer	18,34
Counsel	10,22	Prosecutor	12,38
congressman	10,13	Attorneys	9,29
Attorney	6,89	Counsel	8,37
Prosecutor	6,75	Judge	7,96

TABLA 2

Ejemplos de bigramas anteriores y posteriores

Bigrama anterior	Calificación de probabilidad	Bigrama posterior	Calificación de probabilidad
District judge	152,37	lead counsel	92,91
Court judge	106,85	Tax lawyer	55,75
General counsel	15,03	senior attorney	55,75
city attorney	14,86	Former congressman	55,75
democratic leader	14,45	u.s. attorney	37,16
District attorney	12,38	Law professor	18,58

El bloque 250 implica seleccionar uno o más conjuntos de la información extraída, basándose en las clasificaciones. En el ejemplo de realización, éste es un proceso de selección manual; sin embargo, otras realizaciones pueden aplicar criterios de selección automática basados, por ejemplo, en una clasificación mínima específica o una clasificación mínima específica en combinación con un umbral mínimo.

Más en concreto, el ejemplo de realización selecciona dos conjuntos de términos, denominados términos ancla. El primer conjunto de términos ancla incluye términos que tienen una gran probabilidad de aparecer antes de un nombre poco común "en directorio" y el segundo conjunto de términos ancla incluye términos que tienen una gran probabilidad de aparecer después de tal nombre. En muchos casos, los términos ancla son los nombres en oraciones o cláusulas apositivas que aparecen antes o después de un nombre determinado. El ejemplo de realización agrupa también términos en minúsculas que son sinónimos entre sí. Por ejemplo, los términos como "lawyer", "counsel" y "prosecutor" se consideran sinónimos de "attorney" y, por lo tanto, se agrupan para formar un único término ancla con el fin de reducir el número de expresiones o descriptores regulares generados en el bloque 260.

El bloque 260 implica generar automáticamente modelos (o expresiones) de descriptor de profesión que estén en correlación con la membresía en la profesión. Los ejemplos de descriptores representan modelos gramaticales en fragmentos del cuerpo que están delimitados por nombres poco comunes "en directorio" y los términos ancla seleccionados.

El ejemplo de realización genera los modelos de la siguiente manera: Para cada fragmento de oración delimitado por un término ancla y un nombre poco común, el ejemplo de realización deriva un modelo (o estructura) generalizado (a) dividiendo el fragmento en testigos separados de acuerdo con los espacios y las comas y normalizando a continuación cada testigo basándose en su categoría gramatical más comúnmente asociada en un diccionario de inglés. Los testigos no incluidos en el diccionario se clasifican como "otros" y las formas del verbo "to be" se clasifican como "is" para distinguirlos de otros verbos. Los signos de puntuación que no sean comas se clasifican como "otros". El ejemplo de realización utiliza un diccionario de inglés públicamente disponible de aproximadamente 90.000 palabras, estando cada palabra asociada a una o más categorías gramaticales, enumeradas por orden de frecuencia dentro de un cuerpo. En la tabla 3 se muestra un ejemplo de vocabulario para los modelos descriptivos.

TABLA 3

Ejemplo de vocabulario para modelos descriptivos

Símbolo de modelo	Rasgo gramatical	Ejemplos
$det-	determinante	a, an, the
$pos	posesivo	his, our
$adj	adjetivo	legal
$noun	nombre	attorney
$prep	preposición	for
$pronoun	pronombre	he
$properNoun	nombre propio	Johnsons
$adv	adverbio	legally
$inf	marcador de infinitivo	to
$isVerb	formas del verbo "to be"	is, was
$verb	verbo	said
$comma	coma	,
$other	todas las demás categorías gramaticales o puntuaciones	
/s	espacio	

A continuación, el ejemplo de realización crea una lista combinada de modelos únicos y cuenta la frecuencia con que cada uno se repite en el cuerpo. Después se ordenan los modelos basándose en sus recuentos de aparición y se desechan los modelos singulares. Los modelos restantes se toman entonces como indicativo de membresía en la profesión en cuestión.

El recuadro 260' muestra que un ejemplo de estructura de modelo incluye una estructura de nombre 261, estructuras gramaticales 262, una estructura de términos ancla 263 y datos de posición relativa 264 y 265. La estructura de nombre 261 es un marcador de posición para un nombre en un fragmento. Las estructuras gramaticales 262 incluyen la puntuación, identificadores de categoría gramatical e información posicional asociada indicativa de cualesquiera modelos estructurales gramaticales existentes entre el nombre y la estructura de términos ancla 263. La estructura de términos ancla 263 representa e incluye uno o más términos ancla.

Los datos de posición relativa 264 y 265 indican en cada caso la posición relativa de la estructura de nombre 261 y las estructuras gramaticales 262, y de la estructura gramatical 262 y los términos ancla 263. Aunque el ejemplo de realización implementa los datos de posición relativa implícitamente por lo que se refiere al orden de los datos dentro de la estructura de modelo, otras realizaciones indican explícitamente la posición relativa como "antes" o "después". Algunas realizaciones omiten la estructura de nombre y/o las estructuras gramaticales que intervienen y definen modelos en cuanto a distancia en palabras o caracteres entre los términos ancla y una estructura de nombre implícita.

En la tabla siguiente se muestran ejemplos de modelos descriptivos de profesión para identificar abogados (attorneys) en el cuerpo de artículos informativos. También se muestran fragmentos de oración correspondientes extraídos de un documento mediante el empleo de estos modelos.

TABLA 4

Ejemplos de modelos descriptivos de abogados y fragmentos de oración

Modelos descriptivos de abogados	Ejemplos de fragmentos recuperados
$Anchor $comma\s+ $Name	attorney, <name>
$Anchor \s+ $Name	Attorney general <name>
$Anchor \s+ $prep \s+ $det\s+ $other $comma \s+ $Name	attorney for the Johnsons, <name>
$Name $comma \s+ $det $Anchor	<name>, a lawyer
$Name $comma \s+ $pos\s+ $Anchor	<name>, his attorney
$Name $comma \s+ $det \s+ $adj \s+ $Anchor	<name>, a defense attorney

En esencia, el ejemplo de realización utiliza nombres poco comunes o menos ambiguos como etiquetas virtuales para identificar o extraer mediante filtrado un conjunto de ejemplos de fragmentos de oración que contienen descripciones de profesionales mencionados y que pueden emplearse como base para definir modelos gramaticalmente descriptivos. El filtro de rareza sirve para identificar buenos ejemplos con mucha más eficacia que la búsqueda de fragmentos de oración alrededor de nombres personales en general. A continuación se utilizan estos modelos como ayuda para la generación automática de hipervínculos en el módulo de vinculación 1272.

B. *Estructura y funcionamiento del módulo de vinculación*

En general, el módulo de vinculación 1272 (en la figura 1) recibe un documento de entrada, como el documento 110, y establece hipervínculos de uno o más nombres que aparecen en el documento de entrada a uno o más directorios profesionales, basándose en las estructuras de descriptores de nombres profesionales definidas por el módulo de descriptores 1271.

La figura 3 muestra un organigrama 300 de un ejemplo de un método realizado en el módulo de vinculación 1272. El organigrama 300 incluye los bloques de proceso 310-370.

La ejecución del ejemplo del método se inicia en el bloque 310, que implica recibir un documento, como por ejemplo un documento 112, de la base de datos de documentos de entrada 110. En algunas realizaciones, el módulo de vinculación 1272 ejecuta en el contexto de una o más sesiones iniciadas por el software de procesamiento de documentos 126 y el documento de entrada representa un documento completo hospedado por el software de procesamiento de documentos o una o más partes seleccionadas del documento dentro de una ventana de procesamiento activa en un programa de procesamiento de documentos. Sin embargo, en otras realizaciones el documento o la parte del documento se recibe o se recupera de una ventana de procesamiento actualmente activa en un programa de procesamiento de documentos. En algunas otras realizaciones el módulo de vinculación es una aplicación autónoma que interacciona con una o más bases de datos conectadas a un sistema informático central, como una estación de trabajo. La ejecución continúa en el bloque 320.

En el bloque 320, el módulo de vinculación utiliza un etiquetador genérico para etiquetar o marcar cada nombre de persona, lugar u organización en los documentos de entrada. El ejemplo de realización emplea el mismo etiquetador que el utilizado en el módulo de descriptores. En el ejemplo de realización, el etiquetado de nombres implica también resolver correferencias claras al mismo nombre dentro del documento de entrada.

Con este fin, el ejemplo de realización vincula de forma conjunta las referencias dentro de un documento al mismo nombre, utilizando una serie de reglas de comparación de nombres. Estos vínculos se califican de cadenas de correferencias “en documento”, o cadenas de nombre. Durante el proceso de etiquetado, el apellido de cada nuevo nombre encontrado se compara con los apellidos de los nombres ya encontrados. Si no se hallan coincidencias, el nuevo nombre encontrado se trata como nombre único en el documento. Si el nuevo apellido encontrado coincide con un apellido ya existente, el nombre de pila que acompaña al nuevo apellido encontrado se compara con el nombre de pila que acompaña a los apellidos coincidentes. Si un nombre de pila coincide o es compatible, se supone que los nombres son iguales. Si el nombre de pila es incompatible con el nombre de pila de cualquiera de los apellidos coincidentes, el nombre se trata como único en el documento. Si el nuevo nombre encontrado no tiene un nombre de pila claramente asociado, el nombre se trata como referencia al nombre encontrado más recientemente que tenga el mismo apellido. Otras realizaciones pueden utilizar otras técnicas para resolver correferencias “en documento”.

En el bloque 330, el módulo de vinculación determina cuál de la o las cadenas de nombre etiquetadas es probable que esté asociada a uno o más directorios profesionales. En el ejemplo de realización, esto implica aplicar uno o más modelos descriptivos de profesión generados por el módulo de descriptores 1271. Al aplicar los modelos descriptivos, el ejemplo de realización trata de emparejar o correlacionar cada modelo descriptivo para una profesión determinada con el texto que rodea a cada aparición de un nombre etiquetado en el documento de entrada. Si se correlaciona con éxito el modelo descriptivo con una aparición de un nombre, el nombre (y sus correferencias "en documento") se considera un profesional candidato y se añade a una lista de profesionales candidatos de un tipo determinado. (Otras realizaciones aprovechan la estructura regular o características clave de los documentos como evidencia adicional por lo que respecta a la naturaleza de los nombres en los documentos. Por ejemplo: las resoluciones judiciales (jurisprudencia) incluyen encabezamientos de caso, frases de abogados, resoluciones coincidentes y fechas que pueden utilizarse para identificar y/o distinguir abogados y jueces). La lista de cadenas de nombre candidato se envía a su procesamiento posterior en el bloque 340.

El bloque 340 implica definir una o más plantillas de nombre u otras estructuras de datos basadas en las cadenas de nombre etiquetado de profesionales candidatos y el texto de aparición conjunta relacionado u otra información del documento. En el ejemplo de realización, definir las plantillas de nombre implica formar, para cada cadena de nombre etiquetado, un conjunto de oraciones del documento que contienen un nombre de la cadena y un conjunto de párrafos del documento que contienen un nombre de la cadena. En algunas realizaciones, los conjuntos de oraciones o párrafos son indicadores de posición en documento que denotan el principio y/o final de oraciones y párrafos.

Una vez formados los conjuntos de oraciones y párrafos, el módulo de vinculación define una plantilla para cada cadena de nombre, teniendo cada plantilla la forma del ejemplo de plantilla de nombre 340'. La plantilla de nombre 340' incluye un registro de nombre 341, un registro de datos de descriptor 342, un registro de datos de lugar 343, un registro de organización 344 y un registro de singularidad (o rareza) de nombre 345.

El registro de nombre 341 incluye un campo de nombre de pila (first), un campo de segundo nombre de pila (mid), un campo de apellido (last) y un campo de sufijo (suffix). Otras realizaciones incluyen otros datos relacionados con el nombre, tales como sobrenombres extraídos de una tabla de consulta u ortografías alternativas comunes. Y otras realizaciones excluyen uno o más de los campos utilizados en el ejemplo de realización, tales como el sufijo.

El registro de datos de descriptor 342 incluye uno o más campos de descriptor, tales como D1, D2, D3... Dn. En el ejemplo de realización, cada campo de descriptor incluye fragmentos de oraciones extraídos del documento de entrada mediante el empleo de uno o más de los modelos descriptivos de profesión generados por el módulo de descriptores 1271 para una profesión determinada. Para un documento en el que un nombre se repita o tenga correferencias en múltiples puntos, las estructuras de descriptor se aplican a cada aparición del nombre para montar el conjunto de campos de descriptor. Algunas realizaciones pueden aplicar incluso los modelos de descriptor a referencias a un nombre determinado hechas mediante pronombres (en otras realizaciones, los descriptores incluyen también verbos que aparecen conjuntamente dentro de cierto intervalo de distancia con respecto a los nombres).

La aplicación de los modelos implica alinear la parte del nombre de cada modelo con los nombres que aparecen en el conjunto de oraciones y determinar si la gramática de la parte (el fragmento) adyacente de la oración concuerda con la gramática del modelo. Si existe concordancia, el ejemplo de módulo de vinculación copia una o más partes del fragmento de oración correspondiente, por ejemplo el término ancla, en un campo de descriptor respectivo del registro de datos de descriptor 342.

El registro de datos de lugar 343 incluye un campo de "misma oración" y un campo de "mismo párrafo". El campo de "misma oración" incluye subcampos LS1, LS2,..., LSj, incluyendo cada subcampo LS un nombre de lugar que aparece conjuntamente en una oración con una referencia al nombre que aparece en el registro de datos de nombre 341. El campo de "mismo párrafo" incluye subcampos LP1, LP2,..., LPk. Cada subcampo LP incluye un nombre de lugar que aparece conjuntamente en un párrafo que contiene una referencia al nombre que aparece en el registro de datos 341. En el ejemplo de realización, la construcción de este registro implica buscar lugares en los conjuntos de oraciones y párrafos para un nombre determinado y copiar los lugares encontrados en los subcampos respectivos.

El registro de datos de organización 344 incluye un campo de "misma oración" y un campo de "mismo párrafo". El campo de "misma oración" incluye subcampos OS1, OS2,..., OSj, incluyendo cada subcampo una organización que aparece conjuntamente en una oración con una referencia al nombre que aparece en el registro de datos de nombre 341. El campo de "mismo párrafo" incluye subcampos OP1, OP2,..., OPk. Cada subcampo OP incluye una organización que aparece conjuntamente en un párrafo que contiene una referencia al nombre que aparece en el registro de datos 341.

En el ejemplo de realización se entiende que el concepto de oración incluye tanto oraciones gramaticales como oraciones tipográficas y que el concepto de párrafo incluye cualquier grupo de una o más oraciones delimitado o separado de otro grupo de una o más oraciones por signos de puntuación u otro recurso o técnica de señalización. Además, en el ejemplo de realización no es necesario que los lugares y organizaciones de "misma oración" y de "mismo párrafo" se excluyan mutuamente. Es decir que si un término aparece en una oración con la aparición de un nombre, también aparece dentro del mismo párrafo que contiene la oración. Sin embargo, en otras realizaciones los lugares de "misma oración" y "mismo párrafo" podrían definirse como mutuamente excluyentes.

Aunque el ejemplo de realización prevé información posicional implícita para los lugares y organizaciones a través de la estructura de los campos de "misma oración" y "mismo párrafo", otras realizaciones pueden utilizar otras técnicas para incorporar información de posición relativa a un nombre determinado en la plantilla de nombre. Por ejemplo: algunas realizaciones utilizan el desplazamiento de carácter, palabra, oración, párrafo o página a partir del nombre, o más exactamente la aparición de un nombre determinado. Otras realizaciones prevén información posicional en forma de un conjunto de códigos binarios o banderas, indicando cada bandera si un lugar o una organización lógicamente asociado(a) o correspondiente aparece dentro de una oración o párrafo con su nombre asociado. Otras realizaciones más prevén banderas que indican si los sitios se hallan dentro o fuera de cierto intervalo de texto o región del documento con relación al nombre. Y otras realizaciones prevén una posición relativa en documento o una posición absoluta en documento para cada organización o sitio.

El registro de singularidad (o rareza) de nombre 345 incluye un campo de calificación que contiene un indicador de singularidad o rareza de su nombre asociado. En el ejemplo de realización, este indicador es un indicativo numérico de cantidad de una probabilidad *a priori* de que un nombre coincida con un registro candidato sacado de un directorio en particular. Más en concreto, el ejemplo de realización define la cantidad como la probabilidad de singularidad de un nombre y la calcula mediante

$$\text{P(singularidad de nombre)} = \frac{1}{H' \cdot P(\text{nombre})) + 1} \qquad (9)$$

donde H' significa el tamaño de la categoría profesional indicada por la coincidencia de descriptor y P(nombre) se define como

$$P(\text{nombre}) = P(\text{nombre de pila}) \cdot P(\text{apellido}) \qquad (10)$$

donde P(nombre de pila) significa la probabilidad de sacar el nombre de pila al azar de entre todos los nombres de pila que aparecen en una lista de nombres representativa de la población general y P(apellido) significa análogamente la probabilidad de sacar el apellido al azar de entre todos los apellidos que aparecen en una lista de nombres representativa de la población general.

Aunque el ejemplo de realización utiliza una plantilla como la plantilla de nombre 340' para múltiples tipos de profesiones, algunas realizaciones pueden suprimir o añadir otras características de plantilla. Por ejemplo: las plantillas para jueces pueden omitir la información de lugar separada, dado que la información de la organización, por ejemplo el nombre del tribunal, contiene implícitamente información del lugar. Otras realizaciones pueden omitir información totalmente en lugar de sólo su forma explícita.

El bloque 350, que se ejecuta una vez definidas las plantillas de nombre en el bloque 340, recupera un conjunto de entradas candidatas de uno o más directorios profesionales 130. Con este fin, el ejemplo de realización busca entradas de directorio que tengan el mismo apellido que uno de los nombres profesionales candidatos. A continuación recupera el nombre completo, el título, la organización, el lugar y la información de identificación de entrada para estas entradas de directorio candidatas, para un procesamiento ulterior en el bloque 360.

El bloque 360 implica comparar y calificar la semejanza de cada plantilla de nombre con uno o más de los registros candidatos o estructuras de datos candidatas recuperados(as). En el ejemplo de realización, esto implica utilizar uno o más sistemas de inferencia bayesiana, tales como el mostrado en la figura 4.

La figura 4 muestra un ejemplo de un sistema de inferencia 400 que incluye una plantilla de nombre de entrada 410, un registro candidato de entrada 420 y uno o más motores de inferencia bayesiana, tales como el motor de inferencia bayesiana 430.

La plantilla de nombre de entrada 410 incluye datos de nombre 411, datos de descriptor 412, datos de lugar 413, datos de organización 414 y datos de singularidad de nombre 415. El registro candidato de entrada 420 incluye datos de nombre 421, datos de título 422, datos de lugar 423, datos de organización 424 y datos de identificación de registro 425.

El motor de inferencia 430 incluye los módulos de comparación de evidencia 431-434 y el módulo de cálculo 435. Los módulos de comparación 431-434 incluyen estructuras de datos y de reglas lógicas respectivas que definen diversos estados de comparación y probabilidades asociadas. En el ejemplo de realización, cada motor de inferencia está adaptado a un directorio profesional específico u otra base de datos específica. Además, el ejemplo de realización implementa cada motor empleando un módulo de software reconfigurable, con opciones de configuración para definir la lógica y los cálculos de comparación. Sin embargo, otras realizaciones pueden utilizar estructuras de motor de inferencia totalmente distintas.

Cada módulo de comparación incluye por lo general dos o más estados mutuamente excluyentes que indican un resultado potencial de comparación entre un ítem de datos candidato y un ítem de datos de entrada respectivo. Cada estado está asociado a una lógica de comparación específica y probabilidades condicionales para el estado suponiendo la coincidencia de un registro candidato con la plantilla de nombre de entrada y suponiendo la no coincidencia de los registros candidatos con la plantilla de nombre de entrada (algunas realizaciones incluyen múltiples conjuntos de estados, lógica y probabilidades condicionales, estando cada conjunto asociado a un directorio profesional o una profesión en concreto). Entre los ejemplos de estados se incluyen: una coincidencia exacta, una coincidencia muy aproximada, una coincidencia poco aproximada, una coincidencia desconocida (o no especificada) y una falta de coincidencia.

Una coincidencia exacta se produce cuando los ítems o elementos de datos coinciden exactamente. Una coincidencia muy aproximada se produce cuando los elementos no llegan a coincidir exactamente, pero son muy compatibles. Una coincidencia poco aproximada se produce cuando los elementos no llegan a coincidir exactamente y son poco compatibles. Una coincidencia desconocida se produce cuando no hay suficiente información para determinar si los datos coinciden o no. Y una falta de coincidencia se produce cuando los ítems no presentan compatibilidad.

Más exactamente, el ejemplo de realización define cada uno de los estados para cada módulo de comparación de evidencia de la siguiente manera: El módulo de comparación 431 tiene tres estados: un estado de coincidencia exacta (EX), un estado de coincidencia muy aproximada (SF) y un estado de coincidencia poco aproximada (WF). Para que se produzca una coincidencia exacta, todos los componentes de un nombre extraído deben coincidir exactamente con los de un registro candidato. Por ejemplo: Abraham Lincoln coincide exactamente con Abraham Lincoln, pero no con Abe Lincoln, Abraham Lincoln, Jr. o Abraham S. Lincoln. Para una coincidencia muy aproximada, el nombre de pila y el apellido que aparecen en el documento etiquetado deben coincidir con el nombre de pila y el apellido del registro, con todos los demás componentes del registro sin especificar o en blanco. Así pues, Abraham Lincoln es una coincidencia muy aproximada con Abraham Lincoln, Jr. y Abraham S. Lincoln. Para una coincidencia poco aproximada sólo coincide el apellido, mientras que todos los demás componentes quedan sin especificar o tienen formas variantes de coincidencia tales como sobrenombres. Así pues, Abraham Lincoln es una coincidencia poco aproximada con Abe Lincoln. (Aunque no se muestra en las figuras, la comparación 431 incluye o tiene acceso, por ejemplo, a una base de datos de nombres y sobrenombres o variantes comunes, a la que accede cuando la comparación de nombres revela que los nombres de pila no coinciden exactamente. Sin embargo, otras realizaciones pueden incluir tales sobrenombres en la plantilla de nombre misma o incluso omitir por completo la consideración del sobrenombre).

El módulo de comparación 432 incluye cuatro estados: un estado de coincidencia exacta (EX), un estado de coincidencia muy aproximada (SF), un estado de coincidencia poco aproximada (WF) y un estado de no coincidencia (NO). Una coincidencia exacta se produce cuando coinciden todos los elementos de un descriptor completamente especificado. Por ejemplo: el descriptor “Tribunal Supremo de Justicia de los EE.UU.” coincide exactamente con el título “Tribunal Supremo de Justicia de los EE.UU.”. Una coincidencia muy aproximada se produce cuando coinciden algunos de los elementos de un descriptor y un título, pero no todos ellos. Por ejemplo: el descriptor “Tribunal Supremo de Justicia” es una coincidencia muy aproximada con el título “Tribunal Supremo de Justicia de los EE.UU.”. Un estado desconocido se produce cuando el descriptor identifica una profesión general que concuerda con la o las profesiones cubiertas por el directorio. Por ejemplo: si el descriptor es “juez” y el título en la entrada del directorio es “juez del 8º Tribunal Superior de los EE.UU.”, la coincidencia se considera desconocida o no especificada. Un estado de falta de coincidencia o “no coincidencia” se produce cuando el descriptor está en conflicto o no concuerda con lo especificado en el registro. Por ejemplo: si el descriptor es “juez del Distrito de Nueva York” y el título en la entrada del directorio es “juez del 8º Tribunal Superior de los EE.UU.”, el estado de coincidencia es una falta de coincidencia.

El módulo de comparación 433, que compara la evidencia de lugar, tiene cinco estados de coincidencia: un estado de coincidencia exacta (EX), un estado de coincidencia muy aproximada (SF), un estado de coincidencia poco aproximada (WF), un estado desconocido o no especificado (UN) y un estado de no coincidencia (NO). Una coincidencia exacta se produce cuando la evidencia de lugar de ciudad y estado que está explícitamente vinculada dentro del documento al nombre extraído coincide con la ciudad y el estado de un registro del directorio. Una vinculación explícita se produce, por ejemplo, cuando el lugar aparece conjuntamente en la misma oración. Una coincidencia muy aproximada se produce cuando la ciudad o el estado que aparecen en el mismo párrafo con el nombre extraído coinciden con la ciudad o el estado correspondientes en un directorio candidato. Una coincidencia poco aproximada se produce cuando la ciudad o el estado que aparecen en el mismo documento, pero fuera del mismo párrafo, que el nombre extraído coinciden con la ciudad o el estado que figuran en una entrada del directorio. El estado desconocido, o no especificado, se produce cuando el nombre extraído no está vinculado explícitamente a una ciudad o un estado en concreto y ninguno de los lugares que aparecen en el texto coincide con la información sobre la ciudad o el estado que aparece en el registro candidato. Una falta de coincidencia se produce cuando el nombre extraído está vinculado explícitamente a un nombre de ciudad o estado que no coincide con la información sobre la ciudad o el estado que aparece en el registro candidato.

El módulo de comparación 434, que compara datos de organizaciones, tiene cinco estados: un estado de coincidencia exacta (EX), un estado de coincidencia muy aproximada (SF), un estado de coincidencia poco aproximada (WF), un estado desconocido o no especificado (UN) y un estado de no coincidencia (NO). Una coincidencia exacta se produce cuando un nombre extraído está vinculado explícitamente a una organización concreta en el texto y dicha organización coincide con la organización que aparece en el registro candidato. Una vinculación explícita se produce,

por ejemplo, cuando el lugar aparece conjuntamente en la misma oración. Por ejemplo: existe una coincidencia exacta si el texto describe a un abogado que trabaja en el bufete de Smith & Jones y en la entrada del directorio figura Smith & Jones como un bufete de abogados u otra organización asociado(a). Una coincidencia muy aproximada se produce si el nombre extraído aparece en el mismo párrafo que la organización extraída y la organización extraída coincide con la organización que aparece en un registro candidato. Una coincidencia poco aproximada se produce si el nombre extraído aparece en el mismo documento, pero fuera del párrafo, y la organización extraída coincide con la organización del candidato. Una coincidencia desconocida se produce cuando el nombre extraído no está vinculado explícitamente a una organización concreta y ninguno de los nombres de organización que aparecen en el texto coincide con la información sobre la organización que aparece en el registro candidato. Y una falta de coincidencia se produce cuando el nombre extraído está vinculado explícitamente a un nombre de organización y dicho nombre de organización no coincide con el nombre de organización que aparece en el registro candidato.

La tabla siguiente resume los diversos estados vigentes para cada uno de los módulos de comparación de evidencia en el ejemplo de motor de inferencia bayesiana.

	Estados⇒ Evidencia⇓	Coincidencia exacta	Coincidencia muy aproximada	Coincidencia poco aproximada	Desconocida	Sin coincidencia
E1	Nombre	✓	✓	✓	-	-
E2	Descriptor	✓	✓	-	✓	✓
E3	Organización	✓	✓	✓	✓	✓
E4	Lugar	✓	✓	✓	✓	✓

Los módulos de comparación de evidencia 431-434 llevan a cabo sus respectivas comparaciones y transmiten sus resultados al módulo de cálculo 435 en forma de ocho probabilidades condicionales.

El módulo de cálculo 435 calcula una calificación de semejanza o probabilidad de coincidencia basándose en estas probabilidades condicionales y los datos de singuralidad o rareza del nombre para la plantilla de nombre de entrada. El ejemplo de cálculo utiliza la siguiente forma de la regla de Bayes:

$$P(M/E) = \frac{P(M)\prod_{i=1}^{4} P(E_i/M)}{P(M)\prod_{i=1}^{4} P(E_i/M) + P(\neg M)\prod_{i=1}^{4} P(E_i/\neg M)} \qquad (11)$$

donde P(M|E) significa la probabilidad de que una plantilla coincida con un registro candidato dado cierto conjunto de evidencias, tales como una plantilla de nombre de entrada y un registro candidato. P(M) significa la probabilidad *a priori* de que una plantilla y un registro biográfico coincidan (es decir que se refieran a la misma persona) y P(-M) significa la probabilidad *a priori* de que una plantilla y un registro biográfico no coincidan. El ejemplo de realización define P(M) como la probabilidad de singuralidad o rareza del nombre dentro de la población profesional y P(-M) como

$$P(-M) = 1 - P(M) \qquad (12)$$

$P(E_i|M)$ es la probabilidad condicional de que E_i adopte un estado en particular suponiendo que la plantilla de nombre de entrada coincida con el registro candidato. Por ejemplo, si E_3 significa evidencia de coincidencia de lugar, entonces $P(E_3|M)$ significa la probabilidad de que la información sobre el lugar que aparece en la plantilla de nombre y el registro candidato tenga el estado de coincidencia determinado por el módulo de comparación 433 (coincidencia exacta, coincidencia muy aproximada, coincidencia poco aproximada, coincidencia desconocida o falta de coincidencia), suponiendo que una plantilla de nombre y un registro candidato coincidan. $P(E_i|\text{-},M)$ significa la probabilidad condicional de que E_i adopte un estado en particular suponiendo que una plantilla de nombre no coincida con ningún registro del directorio profesional. Por ejemplo: $P(E_3|\text{-},M)$ significa la probabilidad de que la información sobre el lugar que aparece en una plantilla de persona y el registro candidato coincida, suponiendo que la plantilla y el candidato no coincidan (otras realizaciones incluyen un menor o mayor número de estados, así como otros tipos de información de aparición conjunta).

La calificación de cada comparación entre un registro candidato y una plantilla de nombre está asociada con una Identificación de registro candidato 425 para el registro candidato. En la figura 4, esta asociación está representada por la línea punteada entre la calificación 440 y la identificación de registro candidato 425. En la figura 3, la ejecución continúa en el bloque 370.

En el bloque 370, el módulo de vinculación vincula uno o más de los nombres profesionales etiquetados en el documento de entrada a uno o más de los directorios profesionales candidatos basándose en las calificaciones de comparación. En el ejemplo de realización, esto implica ejecutar los bloques de proceso 371-375 mostrados en el recuadro 370'.

El bloque 371 selecciona el mejor registro candidato entre los registros candidatos para una plantilla de nombre en particular. En concreto, esto implica seleccionar el candidato que tenga la mayor calificación de comparación. Si no hay un único candidato que tenga la calificación más alta, el ejemplo de realización avanza al bloque 372 sin seleccionar un candidato para la plantilla de nombre. Sin embargo, otras realizaciones podrían emplear algún tipo de "desempate" (por ejemplo uno basado en la cronología), o construir vínculos a cada uno de los registros candidatos con mayor puntuación con un mensaje que matice la incertidumbre en la precisión del vínculo, o construir un vínculo que presente al usuario un menú de los candidatos con mayor puntuación.

El bloque 372 determina si el registro candidato seleccionado cumple otros criterios. Con este fin, el ejemplo de realización determina si la calificación del registro candidato seleccionado satisface un determinado criterio umbral, por ejemplo mayor o igual que 0,05. Sin embargo, otras realizaciones utilizan criterios adicionales, tales como la relación comercial o cronológica con el directorio profesional que contiene el registro candidato. Si el registro candidato satisface la ejecución de los criterios de vinculación, continúa en el bloque 373.

El bloque 373 implica construir un hipervínculo que vincule como mínimo una aparición del nombre en cuestión en el documento de entrada a la entrada que aparece en el directorio profesional que hospeda el registro candidato. En el ejemplo de realización, esto incluye el marcado de todas las apariciones del nombre en el documento y la incrustación en el documento de un URL (uniform resource locator) que identifique el registro candidato (en algunas realizaciones, la identificación de referencia de candidato y un número de identificación de documento para el documento de entrada se escriben en un índice que puede utilizarse para facilitar la búsqueda basada en nombres y la posterior vinculación de un documento entre los directorios profesionales y las bases de datos de documentos). El marcado puede adoptar cualquier número de formas, tales como un cambio de fuente con relación al resto del texto en el documento. Otras realizaciones pueden insertar un hipervínculo que remita al directorio profesional, en lugar de a un registro concreto del directorio. Otras plantillas incluyen información del desplazamiento de carácter para el uso en el establecimiento de hipervínculos en el punto correcto dentro de una determinada oración, párrafo o documento. Y otras realizaciones pueden definir el hipervínculo en términos de uno o más destinos intermedios que remiten o encaminan de otro modo al directorio o al registro del directorio para lograr la vinculación deseada.

Una vez construido el vínculo en el bloque 373 (o después de determinar que no se cumplen los criterios de vinculación en el bloque 372), la ejecución pasa al bloque 374. El bloque 374 determina si el documento de entrada contiene otro nombre para una posible vinculación. Una determinación afirmativa devuelve la ejecución al bloque 371 para la selección de un registro candidato para otra plantilla de nombre y una determinación negativa deriva la ejecución al bloque 375. El bloque 375 devuelve la ejecución al bloque 310 para recibir otro documento de entrada para su posterior procesamiento.

C. *Estructura y funcionamiento del módulo de formación*

La figura 5 muestra un organigrama 500, que ilustra un ejemplo de estructura y funcionamiento del módulo de formación 1273 para definir las probabilidades condicionales utilizadas en el motor de inferencia bayesiana de la figura 4. El organigrama 500 incluye los bloques de proceso 510-560.

En el bloque 510, la ejecución comienza con la recepción de un conjunto de documentos de formación, que tienen nombres etiquetados de los cuales se sabe que coinciden con nombres que aparecen en un directorio profesional. En el ejemplo de realización, los documentos de formación están etiquetados manualmente; sin embargo, en otras realizaciones los documentos pueden etiquetarse automáticamente. Los documentos pueden guardarse en una base de datos local o remota y comunicarse al módulo de formación a través de diversas técnicas de transmisión.

El bloque 520 implica extraer datos de los documentos de formación basándose en los nombres etiquetados. Con este fin, el ejemplo de realización genera una plantilla de nombre, utilizando el módulo de extracción del módulo de vinculación 1272, para cada nombre etiquetado, incluyendo cada plantilla un nombre extraído, un texto de descriptor extraído, una lista de lugar extraída y una lista de organización extraída. La plantilla de nombre tiene una estructura similar a la de la plantilla de nombre 340' de la figura 3.

El bloque 530 implica buscar uno o más directorios profesionales, basándose en los nombres que aparecen en las estructuras de datos de formación. En el ejemplo de realización, esto implica buscar uno o más de los directorios profesionales y recuperar las entradas de directorio con apellidos que coincidan con los apellidos que aparecen en las plantillas de nombre para su posterior procesamiento.

El bloque 540 implica determinar las probabilidades condicionales para cada estado de cada variable de evidencia. En el ejemplo de realización, esta determinación implica determinar los recuentos de frecuencia para cada estado de comparación, basándose en las plantillas de nombre para el conjunto de documentos etiquetado manualmente. En concreto, para los casos en que un nombre etiquetado se haya emparejado manualmente con un registro candidato, la realización cuenta el número de veces que cada estado de evidencia concreto aparece para cada una de las variables

de evidencia: nombre, descriptor, lugar y organización. A continuación, la realización divide el recuento para cada estado de evidencia por el número total de coincidencias para obtener $P(E_i|M)$, es decir los valores de probabilidad condicional para cada estado suponiendo una coincidencia. Más exactamente, el ejemplo de realización determina las probabilidades condicionales utilizando una fórmula como

$$P(E_i = \text{estado concreto}/M) = a \frac{y}{z} + \frac{1-a}{x} \qquad (13)$$

donde x significa el número de estados de evidencia para la variable de evidencia E_i, por ejemplo tres estados para el nombre; y significa el número de registros de directorio para los cuales se ha producido el estado en particular, por ejemplo coincidencia exacta; z significa el número total de pares coincidentes de abogados; a es una constante de uniformidad, por ejemplo 0,999999.

Igualmente, el ejemplo de realización cuenta el número de veces que aparece cada estado de evidencia cuando el nombre candidato no coincide con el nombre de plantilla y lo divide por el número total de faltas de coincidencia para obtener $P(E_i|-,M)$ para cada estado, es decir la probabilidad de una coincidencia de estado de evidencia en particular suponiendo una falta de coincidencia en los nombres. Para reducir los cálculos, algunas realizaciones pueden muestrear los registros candidatos coincidentes y/o no coincidentes, por ejemplo, seleccionando uno de cada diez registros.

Otras aplicaciones

La figura 6 muestra que las enseñanzas de la presente revelación tienen aplicaciones más allá de facilitar la generación de hipervínculos para nombres. En particular, la figura 6 muestra un organigrama 600 de un ejemplo de un método para operar un sistema de recuperación de datos que incorpora enseñanzas de la presente revelación. El organigrama 600 incluye los bloques de proceso 610-670.

En el bloque 610, el ejemplo de método comienza con la recepción de una consulta de información. En el ejemplo de método, la consulta tiene una forma booleana o de lenguaje natural e incluye el nombre de una entidad, por ejemplo una persona. En algunos ejemplos, la consulta la efectúa el usuario de un ordenador cliente o un dispositivo de acceso, tal como uno de los dispositivos de acceso 150 de la figura 1, a un servidor, tal como el servidor de base de datos 140 de la figura 1, en un entorno cliente-servidor. En estos casos, el ejemplo de método se incorpora al software en un servidor. Sin embargo, en otros ejemplos, la consulta puede ser recibida y procesada (de acuerdo con este ejemplo de método) en el lado del cliente antes de la transmisión a un servidor para su ejecución. En tales casos, el ejemplo de método puede incorporarse a un navegador, un componente adicional para un navegador, un sistema operativo del lado del cliente o un software de búsqueda.

El bloque 620 implica determinar la ambigüedad de la consulta o de uno o más términos de la misma. En el ejemplo de método, esto implica identificar como mínimo un nombre en la consulta y calcular una probabilidad de singularidad del nombre según

$$P(\text{singularidad de nombre}) = \frac{1}{(H \cdot P(\text{nombre}))+1} \qquad (14)$$

donde H significa el tamaño estimado de la población humana con probabilidad de ser citada en el cuerpo o en la base de datos al o a la que va dirigida la consulta. Una manera de estimar H es tomar el tamaño del cuerpo de la colectividad con probabilidad de ser citada y aumentarlo proporcionalmente en un tanto por ciento para prever las inevitables referencias a personas fuera de la colectividad citada. P(nombre) se define como

$$P(\text{nombre}) = P(\text{nombre de pila}) \cdot P(\text{apellido}) \qquad (15)$$

donde P(nombre de pila) significa la probabilidad de sacar el nombre de pila al azar de entre todos los nombres de pila que aparecen en un universo de búsqueda pertinente, por ejemplo un directorio profesional, y P(apellido) significa análogamente la probabilidad de sacar el apellido al azar de entre todos los apellidos que aparecen en el universo.

El bloque 630 determina si pedir información adicional como ayuda para responder a la consulta, basándose en la ambigüedad determinada de un nombre (u otra parte) de la consulta. En el ejemplo de método, esto implica comparar la probabilidad calculada de singularidad de un nombre con respecto a un umbral. Si la probabilidad de singularidad del nombre está por debajo del umbral, la ejecución avanza al bloque 640, en caso contrario la ejecución continúa en el bloque 660.

El bloque 640 implica obtener información adicional en relación con la consulta. En un ejemplo de método, la obtención de la información adicional implica pedir al usuario información adicional relacionada con una o más

partes ambiguas de la consulta, por ejemplo un nombre que aparece en la consulta. La petición, en algunos ejemplos, se presenta como una ventana de diálogo que pide información relacionada con una profesión, un lugar y/o una organización asociada(o) al nombre.

En otro ejemplo, la obtención de información adicional implica formular automáticamente una o más consultas basadas en una o más partes ambiguas de la consulta recibida, tales como el nombre identificado o una parte del nombre, y realizar la consulta en una o más bases de datos, por ejemplo directorios profesionales u otras bases de datos que incluyan nombres asociados a otros datos. Por ejemplo: una consulta puede pedir registros o partes de registros que tengan apellidos que coincidan con el apellido de un nombre identificado en la consulta recibida. La partes de los registros, en un ejemplo, incluyen información sobre el lugar, la organización y/o el título profesional. Una vez obtenida la información adicional, la ejecución avanza al bloque 650.

El bloque 650 implica cambiar la consulta basándose en la información adicional. En un ejemplo que pide información adicional al usuario, el cambio de la consulta incluye añadir a la consulta una o más partes de la información adicional, por ejemplo en forma de una o más cadenas de texto añadidas. Sin embargo, algunos otros ejemplos cambian la consulta añadiendo un operador de búsqueda, por ejemplo un operador Y, y una o más partes de la información adicional. En otros ejemplos, en particular en algunos que obtienen la información adicional mediante el uso de subconsultas automáticas, el cambio de la consulta incluye añadir una o más subconsultas utilizando la información adicional, como por ejemplo una ciudad, un estado, una organización o un título profesional, obtenida en el bloque 650. Y en algunos otros ejemplos, el cambio de la consulta incluye cambiar el alcance de la búsqueda, por ejemplo añadiendo o borrando una o más bases de datos destino para la consulta, basándose en la información adicional.

El bloque 660 lleva a cabo una búsqueda basándose en la consulta original o la consulta cambiada. En un ejemplo de método, la búsqueda basada en la consulta cambiada implica realizar la búsqueda en una base de datos destino original y/u otra u otras bases de datos. La ejecución continúa en el bloque 670.

El bloque 670 implica emitir los resultados de las consultas. En el ejemplo de método, esto implica presentar los resultados en una pantalla. En un ejemplo que cambia la consulta recibida añadiendo subconsultas sobre la base de la información adicional, la emisión de los resultados incluye visualizar los resultados de la consulta recibida original y los resultados de las subconsultas en zonas separadas de una pantalla de visualización. En algunas variantes de este ejemplo, los resultados de ambas zonas están clasificados según su relevancia.

Entre otras aplicaciones más de las enseñanzas de la presente revelación se incluyen generar nuevos directorios de nombres para bases de datos sobre la base de nombres famosos, figuras políticas, celebridades, llenar lagunas en directorios actuales, identificar o descubrir lagunas en directorios. Otras aplicaciones incluyen la generación automática de expedientes y la referencia cruzada de individuos, empresas, bienes y registros públicos y privados.

Conclusión

Para permitir un avance de la técnica, el inventor ha presentado diversos ejemplos de sistemas, métodos y software que facilitan la asociación lógica de nombres en documentos u otras estructuras de datos a estructuras de datos, tales como registros, en directorios profesionales o bases de datos de otro tipo. Adicionalmente, el inventor ha presentado diversos sistemas, métodos y software para procesar y aumentar las consultas basadas en términos de consulta ambiguos, tales como nombres de entidades.

Las realizaciones arriba descritas están destinadas sólo a ilustrar y enseñar una o más maneras de poner en práctica o implementar la presente invención, no a restringir su amplitud o alcance. El alcance actual de la invención, que abarca todas las maneras de poner en práctica o implementar las enseñanzas de la invención, está definido sólo por las reivindicaciones siguientes.

REIVINDICACIONES

1. Método implementado en ordenador, que comprende:

identificar uno o más nombres en un documento;

seleccionar del o de los nombres identificados en el documento un nombre candidato en el documento correlacionando un modelo descriptivo predefinido de términos no referidos a personas con texto de alrededor de los nombres identificados en el documento, estando el modelo descriptivo basado en un conjunto de ejemplos de fragmentos de oración que contienen descripciones de profesionales mencionados;

definir una plantilla de nombre para el nombre candidato identificando uno o más términos no referidos a personas que aparezcan conjuntamente con el nombre candidato en el documento e incluyendo en la plantilla de nombre para el nombre candidato el o los términos no referidos a personas identificados;

determinar un indicador de rareza para el nombre candidato, siendo el indicador de rareza una cantidad basada en una probabilidad de sacar como mínimo una parte de palabra del nombre al azar de un conjunto de nombres de muestra representativos de una población humana pertinente;

identificar uno o más registros candidatos en una base de datos, basándose en como mínimo una parte de palabra del nombre candidato;

comparar los términos no referidos a personas para cada uno de los registros candidatos con los términos no referidos a personas que aparecen en la plantilla de nombre definida para el nombre candidato;

calcular una o más cantidades, basada cada una en el indicador de rareza del nombre candidato y la comparación de los términos no referidos a personas para uno de los registros candidatos; y

definir un hipervínculo para el nombre candidato basándose en la o las cantidades calculadas.

2. Método implementado en ordenador según la reivindicación 1, en el que el indicador de rareza es una cantidad basada en un tamaño de una población humana, una probabilidad de sacar una primera parte de palabra del nombre al azar y una probabilidad de sacar una segunda parte de palabra del nombre al azar del conjunto de nombres de muestra representativos de una población humana pertinente.

3. Método implementado en ordenador según la reivindicación 2, en el que la primera parte es una parte de nombre de pila del nombre y la segunda parte es una parte de apellido del nombre.

4. Método implementado en ordenador según la reivindicación 1, en el que el cálculo de una o más cantidades, basada cada una en el indicador de rareza del nombre de persona candidato y la comparación de los términos no referidos a personas para uno de los registros candidatos, incluye la utilización de un motor de inferencia bayesiana.

5. Método implementado en ordenador según la reivindicación 1, en el que la definición del hipervínculo basándose en la o las cantidades calculadas comprende:

comparar las cantidades con un umbral; y

definir el hipervínculo basándose en la mayor de las cantidades que sobrepasen el umbral.

6. Método implementado en ordenador según la reivindicación 5, en el que la definición del hipervínculo basándose en la mayor de las cantidades que sobrepasan el umbral comprende definir un hipervínculo para designar el registro candidato correspondiente a la mayor de las cantidades.

7. Medio legible por máquina que comprende instrucciones ejecutables por máquina para llevar a cabo el método implementado en ordenador según la reivindicación 1.

8. Sistema para añadir un hipervínculo a un documento que incluye un nombre de persona, comprendiendo el sistema:

como mínimo un procesador;

una memoria conectada al procesador, incluyendo la memoria instrucciones para:

identificar uno o más nombres en un documento;

seleccionar del o de los nombres identificados en el documento un nombre candidato en el documento correlacionando un modelo descriptivo predefinido de términos no referidos a personas con texto de alrededor de los nombres identificados en el documento, estando el modelo descriptivo basado en un conjunto de ejemplos de fragmentos de oración que contienen descripciones de profesionales mencionados;

definir una plantilla de nombre para el nombre candidato identificando uno o más términos no referidos a personas que aparezcan conjuntamente con el nombre candidato en el documento e incluyendo en la plantilla de nombre para el nombre candidato el o los términos no referidos a personas identificados;

determinar un indicador de rareza para el nombre candidato, siendo el indicador de rareza una cantidad basada en una probabilidad de sacar como mínimo una parte de palabra del nombre al azar de un conjunto de nombres de muestra representativos de una población humana pertinente;

identificar uno o más registros candidatos en una base de datos, basándose en como mínimo una parte de palabra del nombre candidato;

comparar los términos no referidos a personas para cada uno de los registros candidatos con los términos no referidos a personas que aparecen en la plantilla de nombre definida para el nombre candidato;

calcular una o más cantidades, basada cada una en el indicador de rareza del nombre candidato y la comparación de los términos no referidos a personas para uno de los registros candidatos; y

definir un hipervínculo para el nombre candidato basándose en la o las cantidades calculadas.

9. Sistema según la reivindicación 8, en el que el indicador de rareza es una cantidad basada en un tamaño de una población humana, una probabilidad de sacar una primera parte de palabra del nombre al azar y una probabilidad de sacar una segunda parte de palabra del nombre al azar del conjunto de nombres de muestra representativos de una población humana pertinente.

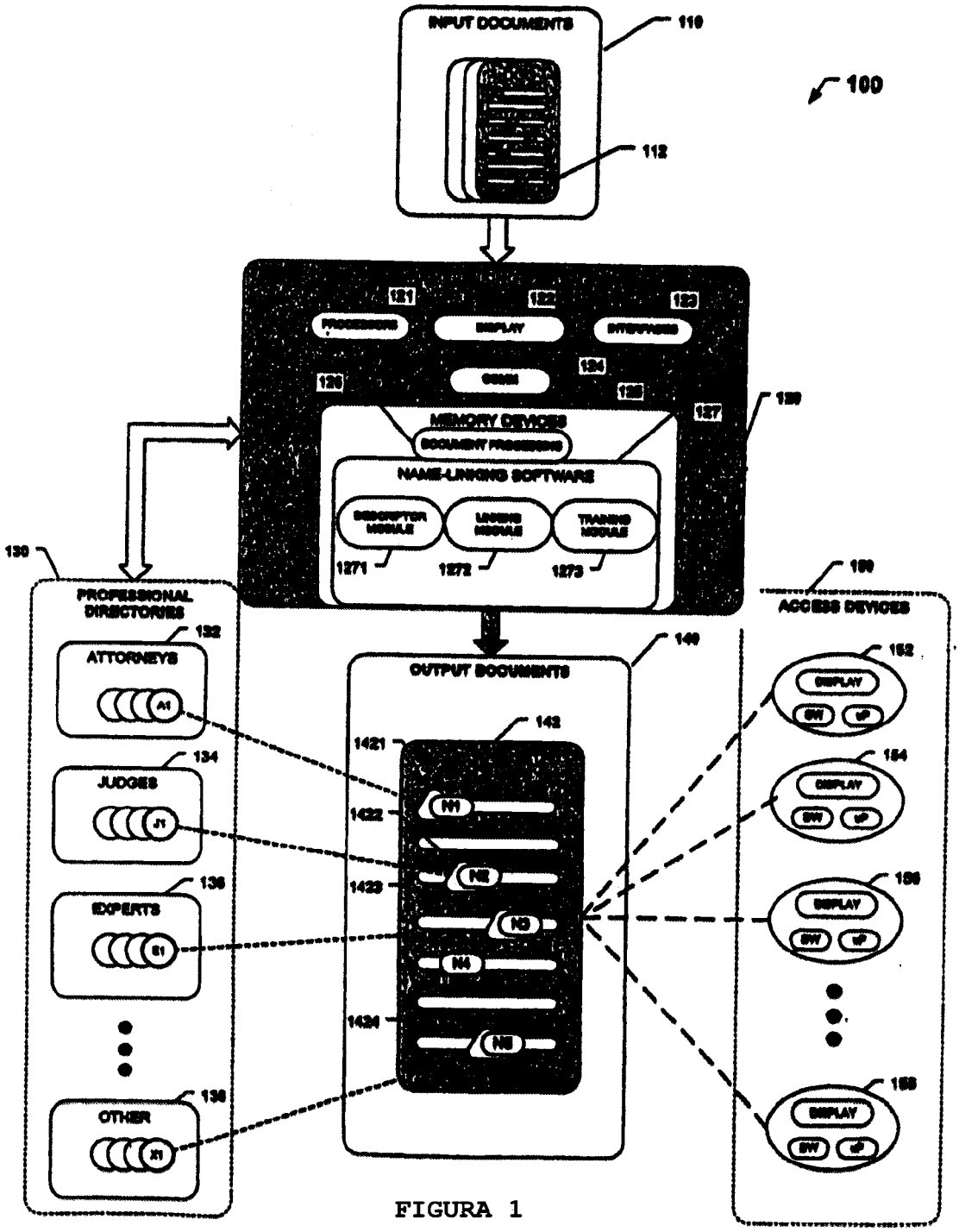
INPUT DOCUMENTS
110
100
112
121
122
123
DISPLAY
124
126
125
127
MEMORY DEVICES
NAME-LINKING SOFTWARE
1271
1272
1273
130
PROFESSIONAL DIRECTORIES
ATTORNEYS
132
JUDGES
134
EXPERTS
136
OTHER
138
OUTPUT DOCUMENTS
140
142
1421
1422
1423
1424
ACCESS DEVICES
152
154
156
158
DISPLAY
SW
HP

FIGURA 1

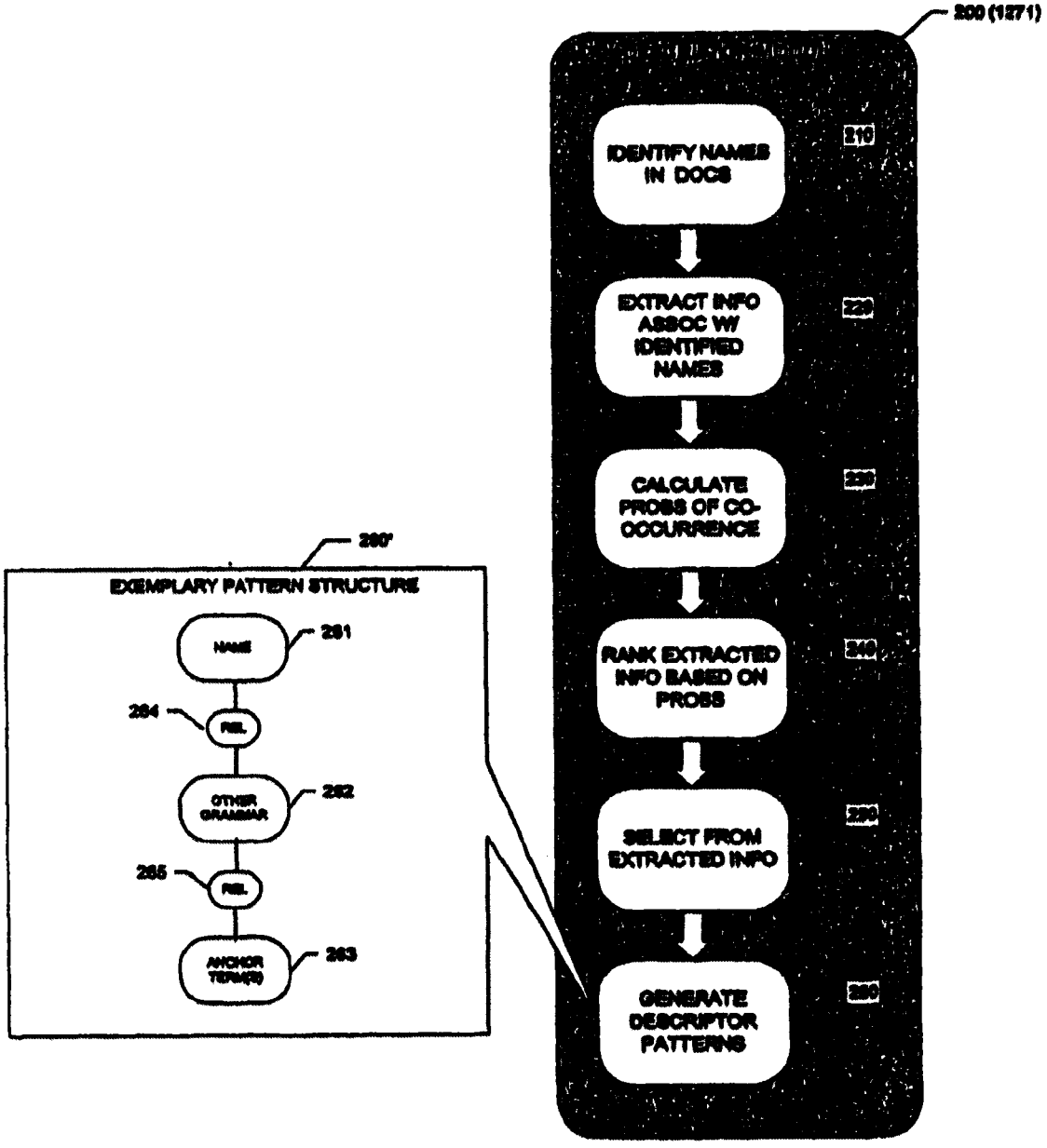
200 (1271)
IDENTIFY NAMES IN DOCS
210
EXTRACT INFO ASSOC W/ IDENTIFIED NAMES
220
CALCULATE PROBS OF CO-OCCURRENCE
230
RANK EXTRACTED INFO BASED ON PROBS
240
SELECT FROM EXTRACTED INFO
250
GENERATE DESCRIPTOR PATTERNS
260
260'
EXEMPLARY PATTERN STRUCTURE
NAME
261
264
REL
OTHER GRAMMAR
262
265
REL
ANCHOR TERM(S)
263

FIGURA 2

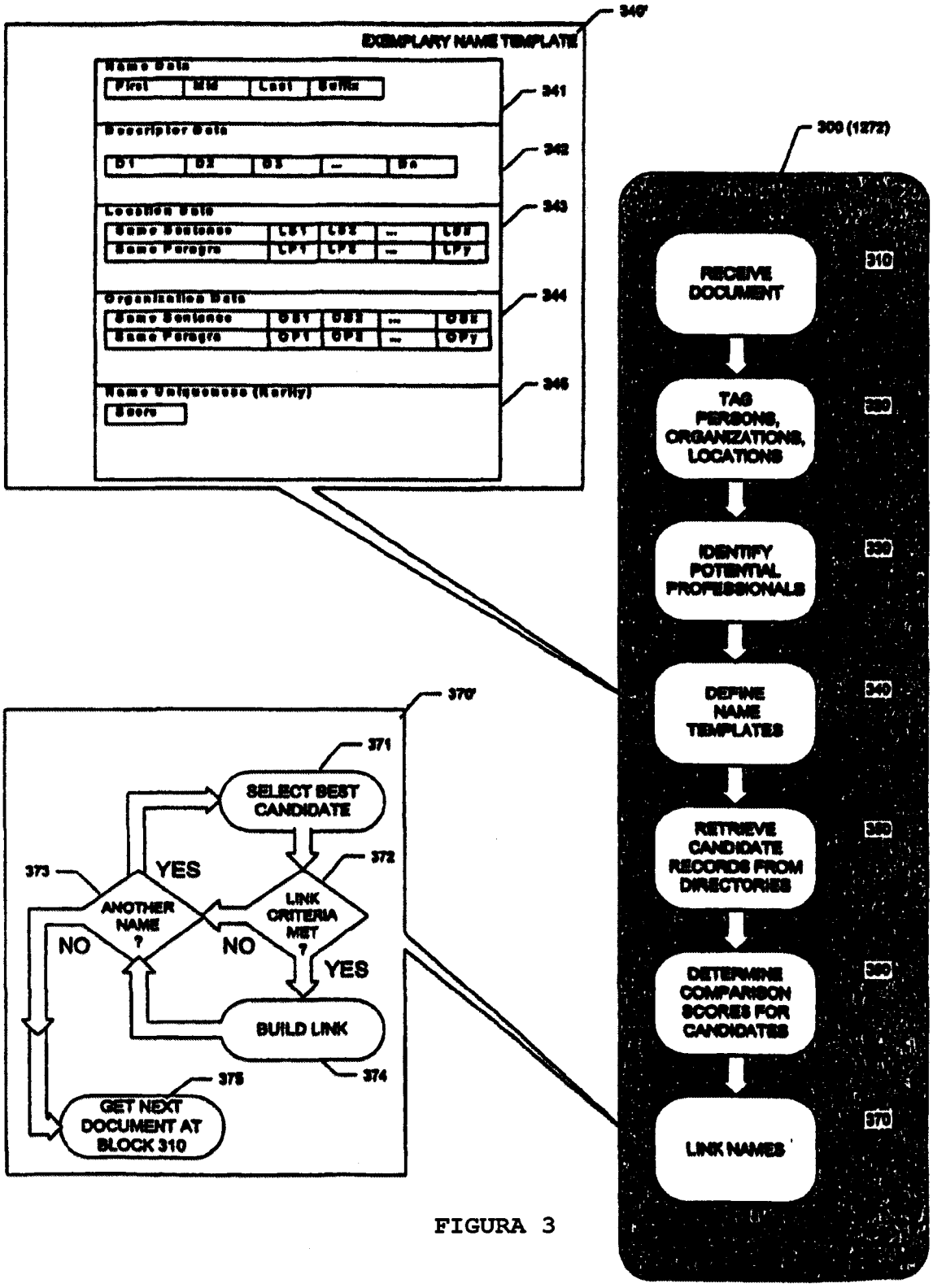
EXEMPLARY NAME TEMPLATE
340'
Name Data
First
MI
Last
Suffix
341
Descriptor Data
D1
D2
D3
...
Dn
342
Location Data
Same Sentence
LS1
LS2
...
LSz
Same Paragra
LP1
LP2
...
LPy
343
Organization Data
Same Sentence
OS1
OS2
...
OSz
Same Paragra
OP1
OP2
...
OPy
344
Name Uniqueness (Rarity)
Score
345
300 (1272)
RECEIVE DOCUMENT
310
TAG PERSONS, ORGANIZATIONS, LOCATIONS
320
IDENTIFY POTENTIAL PROFESSIONALS
330
DEFINE NAME TEMPLATES
340
RETRIEVE CANDIDATE RECORDS FROM DIRECTORIES
350
DETERMINE COMPARISON SCORES FOR CANDIDATES
360
LINK NAMES
370
370'
371
SELECT BEST CANDIDATE
372
LINK CRITERIA MET ?
YES
NO
373
ANOTHER NAME ?
YES
NO
BUILD LINK
374
375
GET NEXT DOCUMENT AT BLOCK 310

FIGURA 3

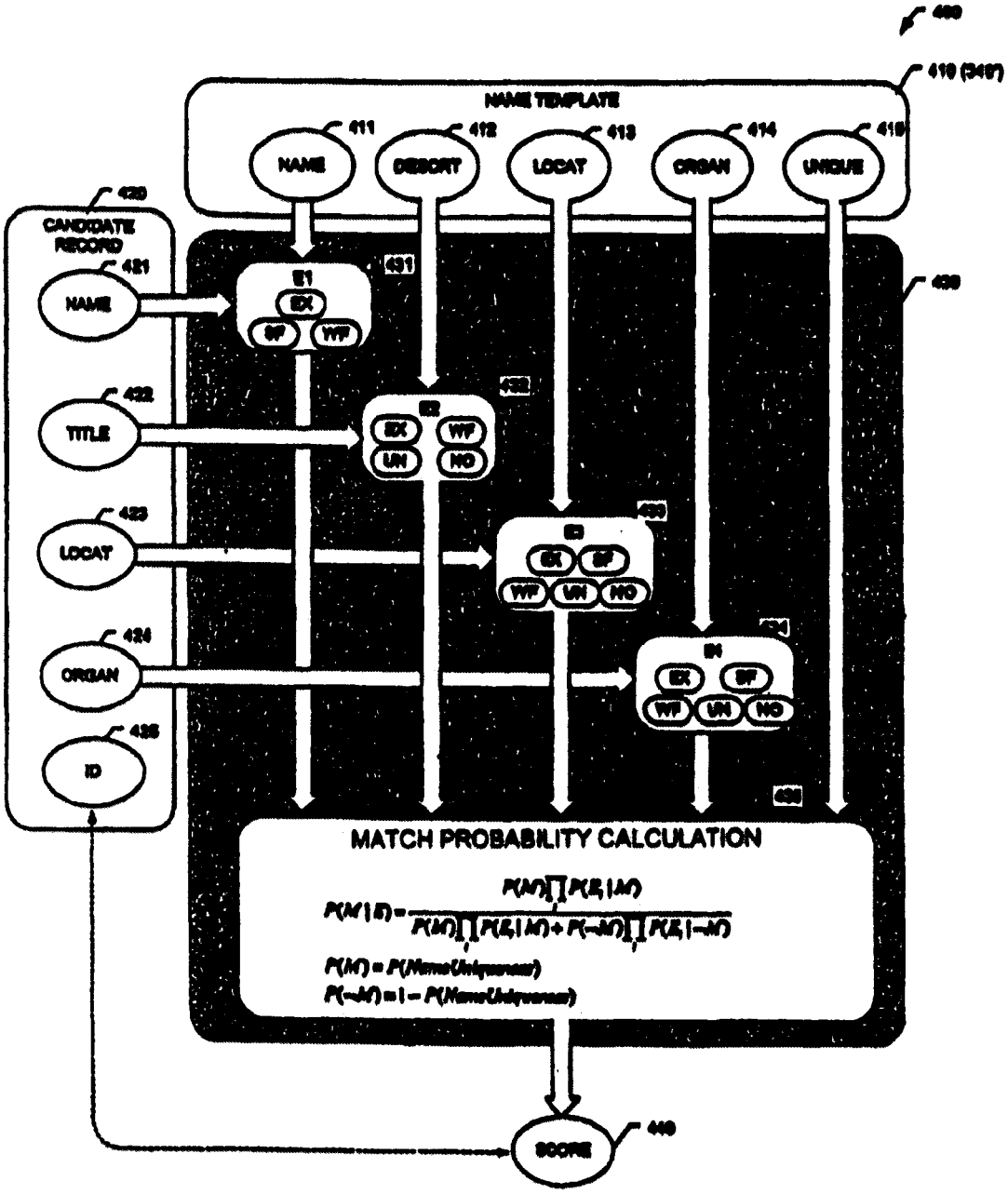
NAME TEMPLATE
NAME
DESCRIT
LOCAT
ORGAN
UNIQUE
CANDIDATE RECORD
NAME
TITLE
LOCAT
ORGAN
ID
MATCH PROBABILITY CALCULATION
SCORE

FIGURA 4

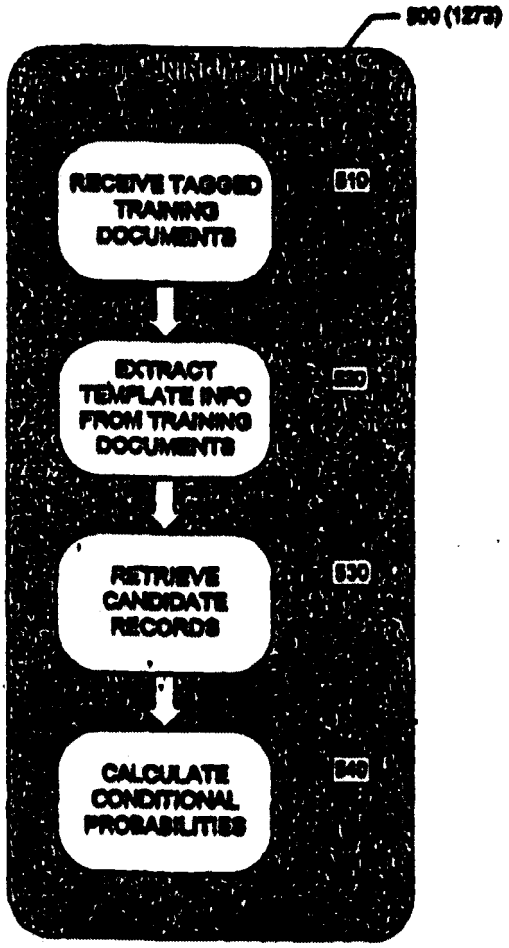
RECEIVE TAGGED TRAINING DOCUMENTS
EXTRACT TEMPLATE INFO FROM TRAINING DOCUMENTS
RETRIEVE CANDIDATE RECORDS
CALCULATE CONDITIONAL PROBABILITIES

FIGURA 5

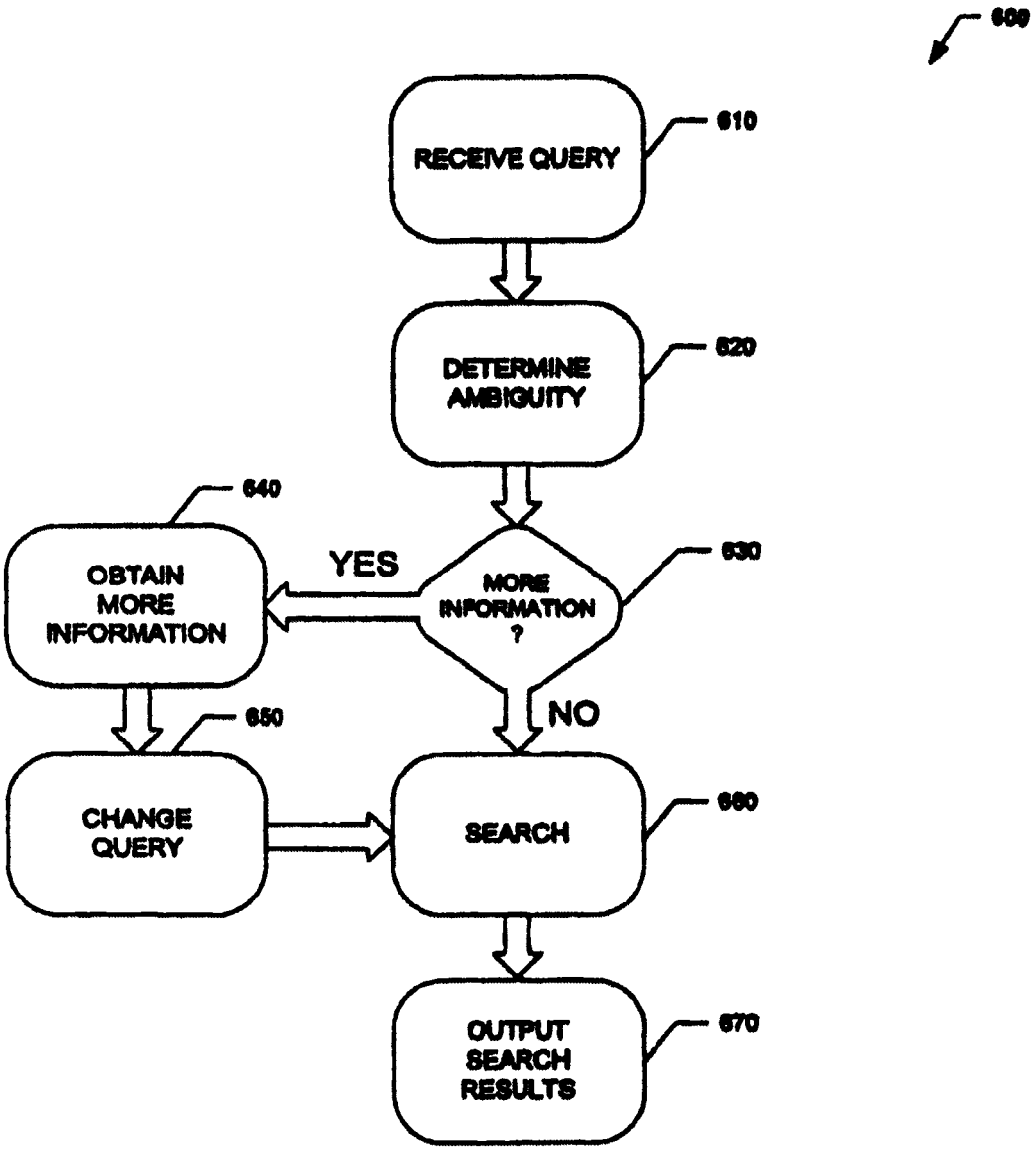
600
RECEIVE QUERY
610
DETERMINE AMBIGUITY
620
MORE INFORMATION ?
630
YES
OBTAIN MORE INFORMATION
640
NO
CHANGE QUERY
650
SEARCH
660
OUTPUT SEARCH RESULTS
670

FIGURA 6