(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 784 180**

(51) Int. CI.:

G06F 17/27 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **25.04.2012 PCT/US2012/034871**

(87) Fecha y número de publicación internacional: **01.11.2012 WO12148950**

(96) Fecha de presentación y número de la solicitud europea: **25.04.2012 E 12721633 (1)**

(97) Fecha y número de publicación de la concesión europea: **25.12.2019 EP 2705442**

(54) Título: **Representación de información de documentos**

(30) Prioridad:

29.04.2011 US 201113097619

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
22.09.2020

(73) Titular/es:

FINANCIAL & RISK ORGANISATION LIMITED (100.0%)
Five Canada Square, Canary Wharf
London, E14 5AQ, GB

(72) Inventor/es:

MALIK, HASSAN, H.;
BHARDWAJ, VIKAS, S.;
FIORLETTA, HUASCAR y
RAFAT, ARMUGHAN

(74) Agente/Representante:

IZQUIERDO BLANCO, María Alicia

DESCRIPCIÓN

Representación de información de documentos

CAMPO TÉCNICO

[0001] Esta descripción se refiere a la representación de la información de la información no estructurada, y más particularmente a sistemas y métodos para la representación de información automáticamente a partir de documentos no estructurados en un formato estructurado.

ANTECEDENTES

[0002] Hoy en día existe una cantidad creciente de información, predominantemente en forma de datos textuales no estructurados incluidos en los documentos, que es relevante para el proceso de toma de decisiones de un inversor. Si bien esta información es voluminosa, el esfuerzo por el cual un inversionista necesita identificar términos y comprender la semántica incluida en estos documentos puede ser laborioso. Aunque el almacenamiento electrónico de documentos ha simplificado el proceso de navegación a través de documentos múltiples y grandes, sigue siendo difícil y lento navegar a través de grandes volúmenes de texto para comprender y localizar rápidamente la información de interés.

[**0003]** Por ejemplo, comunicados de prensa corporativos identifican típicamente eventos financieros corporativos, tales como dividendos, las ganancias por acción, la gestión y estructura de propiedad, etc., en texto no estructurado (por ejemplo, forma libre), junto con información adicional. Analizar esta información para identificar elementos de interés es un proceso lento. Además, aunque la mayoría de las herramientas de procesamiento de texto proporcionan un mecanismo para buscar términos individuales en un documento, ninguna de estas herramientas proporciona información complementaria que acompaña a los elementos de interés.

[**0004]** El documento US 2004/0024769 A1 se refiere a un método para inducir un categorizador jerárquico de arriba hacia abajo. El método incluye proporcionar un conjunto de elementos de capacitación etiquetados. Cada elemento de entrenamiento etiquetado incluye una etiqueta asociada que representa una asignación de categoría única para el elemento de entrenamiento. Se proporciona un conjunto de elementos de capacitación sin etiquetar. Un anterior está asociado con el conjunto de elementos de entrenamiento no etiquetados que es independiente de cualquier función particular contenida en los elementos de entrenamiento no etiquetados. El anterior representa una pluralidad de posibles asignaciones de categoría para el conjunto de elementos de entrenamiento no etiquetados. Se induce un categorizador jerárquico de arriba hacia abajo con un algoritmo de aprendizaje automático basado en el conjunto de elementos de entrenamiento etiquetados, el conjunto de elementos de entrenamiento no etiquetados y el anterior.

[**0005]** US 2011/0066585 A1 se refiere a un "analizador de evento no estructurado". El analizador analiza un evento que está en forma no estructurada y genera un evento que está en forma estructurada. Una fase de mapeo determina, para una ficha de evento dado, los posibles campos del esquema de evento estructurado al que se podría mapear la ficha y las probabilidades de que la ficha se mapee a esos campos. Las fichas particulares se asignan a campos particulares del esquema de evento estructurado. Al usar el modelo de probabilidad Bayesiano ingenuo, un "mapeador probabilístico" determina, para una ficha particular y un campo particular , la probabilidad de que esa ficha se asigne a ese campo. El mapeador probabilístico también se puede usar en un "creador de expresiones regulares" que genera una expresión regular que coincide con un evento no estructurado y un "creador de archivos de parámetros" que ayuda a un usuario a crear un archivo de parámetros para uso con un generador de eventos normalizado parametrizado para generar un evento normalizado basado en un evento no estructurado.

[**0006]** Por consiguiente, existe la necesidad de sistemas y técnicas mejoradas para proporcionar información, tal como hechos y eventos, a partir de datos no estructurados.

SUMARIO

[0007] La invención se define en las reivindicaciones independientes 1 y 2. Los sistemas y técnicas se describen para la representación de información incluida en los documentos de texto no estructurados en un formato estructurado. Los sistemas y técnicas identifican eventos e información asociados con los eventos en documentos no estructurados, clasifican los eventos e información identificados y representan los eventos e información identificados en un formato estructurado basado en una puntuación de clasificación calculada. Los sistemas y técnicas también pueden asignar un puntaje de confianza a los eventos identificados, comparar el puntaje de confianza asociado con los eventos con un puntaje de confianza asociado con un modelo de confianza entrenado, y representar los eventos identificados y la información asociada con los eventos en un formato estructurado basado en la comparación.

[**0008]** Varios aspectos de los sistemas y técnicas relacionados con el cálculo de valores de probabilidad y la combinación de valores de probabilidad para generar una puntuación de clasificación.

[**0009]** Por ejemplo, según un aspecto, un método incluye la identificación de atributos de un evento incluido en un

documento de texto estructurado, cada uno de los atributos identificados similares a al menos un atributo de evento incluidos en un conjunto de atributos de eventos predefinidos, generando funciones del documento para cada uno de los atributos identificados y aplicar al menos uno de una pluralidad de clasificadores a cada una de las funciones generadas. El al menos un clasificador previamente entrenado usando un atributo de evento predefinido correspondiente al atributo de evento identificado.

[**0010]** El método incluye también el cálculo de un valor de probabilidad a partir de una puntuación de clasificador generada por el al menos un clasificador mediante un modelo de estimación de probabilidad, indicando el valor de probabilidad una probabilidad de que el atributo de evento identificado corresponde a uno del conjunto de atributos de evento predefinidos, combinando una pluralidad de valores de probabilidad calculados asociados con los atributos identificados para generar una puntuación de clasificación, y representando, a partir del documento de texto no estructurado , el evento y atributos identificados en un formato estructurado basado al menos en parte en la puntuación de clasificación.

[**0011]** En una realización, el método incluye además la asignación de una puntuación de confianza al evento utilizando al menos un modelo de confianza, la comparación de la puntuación de confianza asociada con el evento a una puntuación de confianza asociada con un modelo de confianza entrenado y que representa, a partir del documento de texto no estructurado, el evento y los atributos identificados en el formato estructurado basado en la comparación.

[**0012]** En aún otro aspecto, un método incluye acceder a un documento de texto estructurado para identificar un evento y un conjunto de atributos asociados con el evento, el conjunto de atributos que se relaciona con un conjunto de atributos de eventos predefinidos, y generar un conjunto de funciones del documento asociadas con el conjunto de atributos, el conjunto de funciones del documento que tiene un mayor número de elementos de conjunto que el conjunto de atributos. Para una primera función del documento en el conjunto de funciones del documento, el método incluye generar una primera puntuación de clasificación, generando la primera puntuación de clasificación con un clasificador previamente entrenado usando el conjunto de atributos de eventos predefinidos, y en base a la primera puntuación de clasificación, calcular un primer valor de probabilidad usando un modelo de estimación de probabilidad, indicando el primer valor de probabilidad una probabilidad de que un primer atributo de evento del conjunto de atributos de evento corresponda al conjunto de atributos de evento predefinidos.

[**0013]** El método también incluye, para una segunda función del documento en el conjunto de las funciones del documento, generar una segunda puntuación de clasificador, generándose la segunda puntuación de clasificador con el clasificador, y basándose en la segunda puntuación de clasificación, el cálculo de un segundo valor de probabilidad utilizando el modelo de estimación de probabilidad, indicando el segundo valor de probabilidad una probabilidad de que un segundo atributo de evento del conjunto de atributos de evento corresponda al conjunto de atributos de evento predefinidos.

[**0014]** El método incluye además la generación de una puntuación de clasificación utilizando un primer valor de probabilidad y el segundo valor de probabilidad, y en base a la puntuación de clasificación, representando de un documento de texto estructurado, el evento y el conjunto de atributos en un formato de datos estructurado.

[**0015]** Se describe un sistema, así como artículos que incluyen un medio legible por máquina que almacena instrucciones legibles por máquina para la aplicación de las diversas técnicas. Los detalles de diversas implementaciones se analizan con mayor detalle a continuación.

[**0016]** Funciones y ventajas adicionales serán fácilmente evidentes a partir de la siguiente descripción detallada, los dibujos adjuntos y las reivindicaciones.

BREVE DESCRIPCIÓN DE LOS DIBUJOS

[**0017]**

La FIG. 1 es un esquema de un sistema informático ejemplar para representar información de un documento de texto no estructurado.
FIG. 2 ilustra un método ejemplar para entrenar el sistema basado en computadora mostrado en la FIG. 1.
FIG. 3 ilustra un método ejemplar para representar información de un documento de texto no estructurado.
FIG. 4 ilustra una interfaz de usuario ejemplar para entrenar el sistema basado en computadora de la FIG. 1.

[**0018]** Símbolos de referencia similares en los diversos dibujos indican elementos similares.

DESCRIPCIÓN DETALLADA

[0019] La presente invención incluye métodos y sistemas que facilitan la extracción automática (por ejemplo, la representación) de eventos (por ejemplo, hechos) y atributos identificados de eventos (por ejemplo, información relativa a los eventos) de datos no estructurados en un formato de datos estructurado. Los ejemplos de datos no estructurados que se pueden usar con la presente invención incluyen, entre otros, libros, revistas, documentos,

metadatos, registros de salud, registros financieros y texto no estructurado como informes de noticias, un comunicado de prensa corporativo, el cuerpo de un mensaje de correo electrónico, una página web, así como documentos de procesador de textos.

[**0020]** Formatos de datos estructurados especifican cómo se debe estructurar e incluir normas que estandarizan la estructura de datos y el contenido de la información. Los ejemplos de formatos de datos estructurados generados por la presente invención incluyen, entre otros, lenguaje de marcado extensible (XML), lenguaje de informes comerciales extensible (XBRL), lenguaje de marcado de hipertexto (HTML) y otros formatos de datos que tienen un documento de especificación publicado.

[**0021]** Los métodos y sistemas son particularmente beneficiosos en los escenarios en los que un evento financiero está incluido en el texto no estructurado, junto con varios otros hechos, algunos de los cuales se relacionan con el evento financiero y algunos de los cuales no se relacionan con el evento financiero.

[**0022]** Por ejemplo, un comunicado de prensa de empresas puede ser un evento como un anuncio de dividendo en acciones que tiene asociado un periodo de tiempo en el que el dividendo en acciones se paga y un nombre de entidad que identifica el negocio de la preocupación de pagar el dividendo en acciones, lo cual es de interés para un profesional del mercado. El comunicado de prensa también puede incluir información adicional no relacionada con el evento de dividendos, como información de beneficios para nuevos empleados, que puede ser de menor interés para el profesional del mercado. Utilizando la presente invención, el profesional del mercado no necesita pasar el tiempo leyendo el comunicado de prensa completo y analizando la nueva información de beneficios para empleados, ya que el dividendo y la información relacionada que es de interés para el profesional del mercado se pueden proporcionar automáticamente al mercado profesional en uno de varios formatos de datos estructurados.

[**0023]** Volviendo ahora a la FIG. 1, se describe un ejemplo de un sistema informático adecuado 10 dentro del cual se pueden implementar realizaciones de la presente invención. El sistema informático 10 es solo un ejemplo y no pretende sugerir ninguna limitación en cuanto al alcance de uso o funcionalidad de la invención. Tampoco debe interpretarse que el sistema informático 10 tiene una dependencia o requisito relacionado con cualquier componente o combinación de componentes ilustrados.

[**0024]** Por ejemplo, la presente invención es operativa con otros numerosos ejemplos de electrónica informática de consumo, PCs de red, miniordenadores, ordenadores centrales, ordenadores portátiles, así como entornos informáticos distribuidos de propósito general o de propósito especial que incluyen cualquiera de los sistemas o dispositivos anteriores, y similares.

[**0025]** La invención puede describirse en el contexto general de instrucciones ejecutables por ordenador, tales como el programa de módulos, siendo ejecutados por un ordenador. En general, los módulos de programa incluyen rutinas, programas, objetos, componentes, estructuras de datos, segmentos de código de bucle y construcciones, etc. que realizan tareas particulares o implementan tipos de datos abstractos particulares. La invención se puede practicar en entornos informáticos distribuidos donde las tareas son realizadas por dispositivos de procesamiento remoto que están vinculados a través de una red de comunicaciones. En un entorno informático distribuido, los módulos del programa se encuentran en medios de almacenamiento informáticos locales y remotos, incluidos los dispositivos de almacenamiento de memoria. Las tareas realizadas por los programas y módulos se describen a continuación y con la ayuda de figuras. Los expertos en la materia pueden implementar la descripción y las figuras como instrucciones ejecutables del procesador, que pueden escribirse en cualquier forma de un medio legible por computadora.

[**0026]** En una realización, con referencia a la FIG. 1, el sistema 10 incluye un dispositivo servidor 12 configurado para incluir un procesador 14, como una unidad de procesamiento central ('CPU'), memoria de acceso aleatorio ('RAM') 16, uno o más dispositivos de entrada-salida 18, como un dispositivo de visualización (no mostrado) y un teclado (no mostrado), y memoria no volátil 20, todos los cuales están interconectados a través de un bus común 22 y controlados por el procesador 14.

[**0027]** Como se muestra en el ejemplo de la FIG. 1, en una realización, la memoria no volátil 20 está configurada para incluir un módulo de normalización 24 para identificar, a partir de un documento de texto no estructurado, atributos de un evento, tales como monedas, calificadores financieros, períodos de tiempo, delimitadores, nombres de entidades, y otros elementos de importancia en el dominio financiero, un módulo de funciones 26 para generar funciones del documento (por ejemplo, vectores numéricos) que describen elementos, tales como palabras, términos, signos de puntuación, etc., que ocurren en el documento de texto no estructurado, un módulo de clasificación 28 para clasificar un conjunto de funciones del documento y asignar una puntuación de clasificación a los elementos que ocurren en el documento de texto no estructurado, un módulo de confianza 30 para determinar la precisión en la identificación del evento a partir del documento de texto no estructurado y un módulo de extracción 32 para representar el evento y cualquier atributo identificado del evento del documento de texto no estructurado en un formato de datos estructurados. Como se usa en este documento, las palabras “conjunto” y “conjuntos” se refieren a cualquier cosa, desde un conjunto nulo a un conjunto de elementos múltiples. Se discuten detalles adicionales de estos módulos 24, 26, 28, 30 y 32 en relación con las FIGS. 2, 3 y 4.

[**0028]** Una red 32 está provista que puede incluir diversos dispositivos tales como routers, servidores y elementos de conmutación conectados en una configuración de Intranet, Extranet o Internet. En una realización, la red 32 usa comunicaciones por cable para transferir información entre un dispositivo de acceso (no mostrado), el dispositivo servidor 12 y un almacén de datos 34. En otra realización, la red 32 emplea protocolos de comunicación inalámbrica para transferir información entre el dispositivo de acceso, el dispositivo servidor 12 y el almacén de datos 34. En otras realizaciones más, la red 32 emplea una combinación de tecnologías cableadas e inalámbricas para transferir información entre el dispositivo de acceso, el dispositivo servidor 12 y el almacén de datos 34.

[**0029]** El almacén de datos 34 es un repositorio que mantiene y almacena información utilizada por los módulos 24, 26, 28, 30 y 32 mencionados anteriormente. En una realización, el almacén de datos 34 es una base de datos relacional. En otra realización, el almacén de datos 34 es un servidor de directorio, tal como un protocolo ligero de acceso a directorio ('LDAP'). En aún otra realización, el almacén de datos 34 es un área de memoria no volátil 20 del servidor 12.

[**0030]** Como se muestra en el ejemplo de la FIG. 1, en una realización, el almacén de datos 34 incluye un conjunto de documentos de entrenamiento 36 que son utilizados por el módulo de clasificación 28 para entrenar múltiples clasificadores binarios en atributos de eventos, se proporcionan una pluralidad de esquemas de generación de funciones 38 que son aplicados por el módulo de funciones 26 para generar funciones del documento para el conjunto de documentos de capacitación 36 y el conjunto de documentos no estructurados 44, y se proporciona un conjunto de reglas predefinidas 40 que son aplicadas por el módulo de clasificación 28 si un atributo incluido en uno de un conjunto de los documentos no estructurados se identifican positivamente.

[**0031]** El almacén de datos 34 también incluye un conjunto de eventos predefinidos 42. Cada uno de los eventos predefinidos 42 incluye al menos un atributo de evento predefinido asociado con el mismo. Por ejemplo, en una realización, un evento predefinido titulado “dividendo” tiene asociado los siguientes atributos de evento predefinidos: una cantidad, un período y un calificador. En una realización, cada uno de los atributos de evento predefinidos está asociado con un identificador único en el sistema. El almacén de datos 34 también incluye uno o más modelos de confianza entrenados 46 que proporcionan una determinación precisa de los eventos identificados en el conjunto de documentos no estructurados 44, que en una realización, puede incluir uno o más artículos recibidos a través de una fuente de datos en tiempo real, y modelos de estimación de probabilidad 48 para calcular los valores de probabilidad de los puntajes de clasificación calculados por el módulo de clasificación 28. Se discuten en mayor detalle a continuación detalles adicionales de la información incluida en el almacén de datos 34.

[**0032]** Aunque el almacen de datos 34 mostrado en la FIG. 1 está conectado a la red 32, un experto en la materia apreciará que el almacén de datos 34 y/o cualquiera de la información 36-48 mostrada en la FIG. 1, pueden ser distribuidos a través de varios servidores y ser accesible para el servidor 12 sobre la red 32, ser acoplados directamente al servidor 12, o estar configurados en una zona de memoria no volátil 20 del servidor 12.

[**0033]** Además, debe observarse que el sistema 10 mostrado en la FIG. 1 es solo una realización de la divulgación. Otras realizaciones del sistema de la divulgación pueden incluir estructuras adicionales que no se muestran, tales como dispositivos computacionales de almacenamiento secundario y adicional. Además, varias otras realizaciones de la divulgación incluyen menos estructuras que las mostradas en la FIG. 1. Por ejemplo, en una realización, la divulgación se implementa en un único dispositivo informático en una configuración independiente no en red. La entrada de datos se comunica al dispositivo informático a través de un dispositivo de entrada, como un teclado y/o ratón. La salida de datos del sistema se comunica desde el dispositivo informático a un dispositivo de visualización, como un monitor de computadora.

[**0034]** Volviendo ahora a la FIG. 2, se describe un método de ejemplo para entrenar el sistema basado en computadora mostrado en la FIG. 1. Primero, en el paso 50, el módulo de no finalización 24 normaliza cada documento en el conjunto de documentos de capacitación 36. En una realización, la normalización de cada documento incluye la identificación de fichas de importancia (por ejemplo, palabras, frases, secuencias de letras, números y carácteres especiales) del dominio financiero para cada conjunto de documentos de capacitación.

[**0035]** A continuación, en el paso 52, el módulo de normalización 24 identifica atributos del candidato en cada uno de los documentos de formación. Como se usa en este documento, el término "atributo de candidato" se refiere a una palabra, frase u otra señal de importancia que puede relacionarse con un atributo predefinido asociado con uno de los eventos 42 predefinidos en el sistema. Por ejemplo, en una realización, los atributos de candidatos incluyen, entre otros, monedas, calificadores financieros, períodos de tiempo, delimitadores y nombres de entidades incluidos en cada uno de los documentos de capacitación. El módulo de normalización 24 asigna a cada ficha identificada de importancia un identificador único dentro de cada documento de entrenamiento.

[**0036]** Haciendo referencia a la FIG. 4, en una realización, el módulo de normalización 24 proporciona una interfaz de usuario que muestra cada documento de entrenamiento normalizado a un usuario, tal como un experto humano. El módulo de normalización 24 muestra cada atributo de candidato identificado como una porción de texto marcado/etiquetado dentro de cada documento de capacitación. Como se muestra en el ejemplo de FIG. 4, el experto puede identificar partes de texto marcado/etiquetado, representadas en el sistema por el identificador único, que son

positivas (por ejemplo, corresponden a) cualquier atributo en el conjunto de atributos de eventos predefinidos asociados con eventos 42. El módulo de normalización 24 genera entonces un par (MT_{ij}, S_k) que representa la parte marcada/etiquetada jº del texto M en el documento T_i que es positiva para un atributo de evento predefinido S_k. El conjunto de todos estos pares P se almacenan a continuación, por el módulo de normalización 24 en el almacén de datos 34.

[**0037]** En una realización, para cada atributo de evento pre-definido S_k, el módulo de normalización 24 identifica ejemplos positivos y ejemplos negativos del conjunto de documentos de entrenamiento 36. Los ejemplos positivos son todos los pares en el conjunto de pares P que corresponden a uno de los atributos de evento predefinidos S_k. Los ejemplos negativos son todos los pares en P que no corresponden al atributo de evento predefinido S_k,, pero tienen un tipo de atributo similar a S_k. Por ejemplo, si S_k es un valor de dividendo numérico, todos los demás valores numéricos se identifican como ejemplos negativos.

[**0038]** Con referencia de nuevo a la FIG. 2, una vez que se determinan los ejemplos positivos y negativos, en el paso 54, el módulo de funciones 26 genera una o más funciones del documento para cada uno de los ejemplos positivos y negativos identificados. En una realización, el módulo de funciones 26 genera una o más funciones del documentos (por ejemplo, vectores numéricos) en una porción de texto no estructurado (por ejemplo, el texto marcado/etiquetado) que rodea un atributo de evento potencial (por ejemplo, un candidato) de cada ejemplo positivo y negativo. El tamaño de la porción de texto no estructurado es configurable por el usuario. Por ejemplo, refiriéndose al siguiente ejemplo de texto no estructurado, la parte del texto no estructurado que rodea el atributo del evento de candidato "*0,45p*" es "*Junta recomienda, sujeto a la aprobación de los accionistas, un dividendo total para el año de 0,45p por acción (2009: 0,4p por acción*)".

[**0039]** El módulo de función 26 de la presente invención utiliza una pluralidad de esquemas de generación 38 de funciones (por ejemplo, algoritmos) para generar funciones del documento para ejemplos positivos y negativos. Por ejemplo, en una realización, los esquemas de generación de funciones incluyen, entre otros, los siguientes esquemas: “Bolsa de palabras”, “Distancia más lejana/Distancia más cercana”, “Antes o después”, “Calificador-presente”, ”Delimitador-presente”, “Figura-Valor-Umbral”, “N-Gramos”, “Palabras del título”, “Período en contexto”, “Etiqueta de ajuste único más cercana” y “Log del valor para atributos basados en la figura”.

[**0040]** El módulo de función 26 utiliza el esquema de bolsa de palabras para generar una función de documento para cada palabra única, frase o texto normalizado que se produce en una parte de texto no estructurado incluyendo la información marcada/etiquetada, y asigna un valor de función para la función de documento generada en función de una cantidad de veces que cada palabra, frase o texto normalizado, respectivamente, aparece en la porción de texto no estructurado. Por ejemplo, refiriéndose al ejemplo antes mencionado de texto no estructurado, unigrams extraídos incluyen “Junta”, “es”, “recomendar”, “sujeto”, etc.

[**0041]** El módulo de función 26 utiliza el esquema distancia más lejana/distancia más cercana para generar una función de documento para información marcada/etiquetada. En una realización, el módulo de funciones 26 compara la información etiquetada con una pluralidad de texto predefinido asociado con el conjunto de atributos de eventos predefinidos, y luego genera un documentopara la información etiquetada basada en la comparación. El módulo de funciones 26 luego asigna un valor de función a la función del documento generada que representa una distancia espacial entre la información marcada/etiquetada y un atributo de candidato.

[**0042]** Por ejemplo, en referencia al ejemplo antes mencionado de texto no estructurado, si las palabras "recomendando" y "dividendo" son parte del texto predefinido asociado con el conjunto de atributos de eventos predefinidos, valores de funciones asignados a la función del documento generada sería 11/21 y 5/21, donde 11 y 5 son distancias de palabras del atributo de candidato “.45p” y veintiuno (21) representa el número de palabras en el ejemplo antes mencionado de texto no estructurado.

[**0043]** El módulo de función 26 utiliza el esquema de antes o después para generar una función del documento para la información marcada/etiquetada que se produce en una lista de texto predefinido asociado con atributos de los eventos predefinidos. En una realización, el módulo de funciones 26 compara la información marcada/etiquetada con una pluralidad de texto predefinido asociado con el conjunto de atributos de eventos predefinidos, genera la función de documento para la información marcada/etiquetada basada en la comparación, y luego asigna un primer valor de función, por ejemplo, uno numérico (1), a la función del documento generada si la información marcada/etiquetada se incluye en la pluralidad de texto predefinido y la información marcada/etiquetada ocurre después del atributo de candidato en la parte del texto no estructurado. El módulo de funciones 26 asigna un segundo valor de función, por ejemplo uno negativo (-1), a la función del documento generada si la información marcada/etiquetada se incluye en la pluralidad de texto predefinido que ocurre antes del al menos un candidato en la porción de texto no estructurado, y asigna un tercer valor de función, por ejemplo un cero (0), a la función del documento generada si la información etiquetada no se incluye en la pluralidad de texto predefinido.

[**0044]** Por ejemplo, se hace referencia al ejemplo antes mencionado de texto no estructurado, si las frases "por acción" y "recomendar" son parte del texto predefinido asociado con un atributo de evento de figura, el módulo de función 26 asigna un valor de función de uno (1) y uno negativo (-1), respectivamente, ya que "por acción" aparece en el texto de

ejemplo después del atributo de candidato de figura y "ocurre" en el texto de ejemplo antes del atributo de candidato de figura.

[**0045]** El módulo de función 26 utiliza el esquema calificador-presente para generar una función del documento para los términos calificadores (por ejemplo, términos que diferencian, caracterizan o distinguen el atributo de candidato) que se producen en la porción de texto no estructurado. En una realización, el módulo de funciones 26 identifica el texto calificador incluido en la porción de texto no estructurado, genera una función del documento para el texto calificador identificado y luego asigna un valor de función a la función del documento generado que representa si el texto calificador identificado está incluido en una pluralidad de texto calificador predefinido asociado con el conjunto de atributos de evento predefinidos.

[**0046]** Por ejemplo, en referencia al ejemplo antes mencionado de texto no estructurado, si el texto de clasificación predefinido incluye las palabras "total", "final", "intermedio" y "básico", el módulo de función 26 puede asignar valores de función para las funciones del documento generadas de uno (1), cero (0), cero (0) y cero (0), respectivamente, ya que solo la palabra "total" está presente en el texto no estructurado de ejemplo.

[**0047]** El módulo de función 26 utiliza el esquema delimitador-presente para generar una función de documento para cada delimitador (por ejemplo, coma, dos puntos, paréntesis, período, etc.) que se produce en la parte de texto no estructurado. En una realización, el módulo de funciones 26 identifica un delimitador incluido en la porción de texto no estructurado, genera una función del documento para el delimitador identificado, y luego asigna un valor de función a la función del documento generada que representa si el delimitador identificado está incluido en una pluralidad de delimitadores predefinidos asociados con el conjunto de atributos de eventos predefinidos.

[**0048]** El módulo de función 26 utiliza el esquema Figura-Valor-Umbral para generar funciones del documento para atributos de eventos numéricos. En una realización, el módulo de funciones 26 identifica un atributo de evento numérico incluido en la porción de texto no estructurado, genera una función del documento para el atributo de evento numérico identificado, compara el atributo de evento numérico con un valor umbral predefinido; y asigna un valor de función a la función del documento generada en función de la comparación. El módulo de funciones 26 puede asignar un valor de función de uno (1) si el atributo de evento numérico no excede el valor umbral y asignar un valor de función de cero (0) si el atributo de evento numérico excede el valor de umbral.

[**0049]** El módulo de función 26 utiliza el esquema N-Gramos para generar una función de documento para cada N-Gramo único (por ejemplo, bi-gramo, tri-gramo, etc..) que se produce en la parte de texto estructurado y utiliza la cantidad de veces que el N-Gramo ocurre en la porción de la ventana de texto no estructurado como una frecuencia de función del documento. En una realización, el módulo de funciones 26 identifica cada N-Gramo único incluido en la porción de texto no estructurado, genera una función de documento para cada uno de los N-Gramos identificados, y luego asigna un valor de función a la función de documento generada en función de una frecuencia en que se produce cada N-gramo único identificado en la porción de texto no estructurado.

[**0050]** Por ejemplo, en referencia al ejemplo antes mencionado de texto no estructurado y el uso de Bi-gramos, la función de módulo 26 utilizando el esquema N-gramos generaría lo siguiente como funciones del documento: "Junta es", "recomienda", "por acción", etc.

[**0051]** El módulo de función 26 utiliza el esquema de palabras de título para generar una función del documento para la información marcada/etiquetada que se produce tanto en el título del texto estructurado y la porción de texto no estructurado. Por ejemplo, en una realización, el módulo de funciones 26 genera una función del documento para la información marcada/etiquetada, y asigna un valor de función a cada función del documento generada que representa si la información etiquetada está incluida en un título asociado con el documento de texto no estructurado y también se incluye en una pluralidad de textos predefinidos asociados con el conjunto de atributos de eventos predefinidos.

[**0052]** El módulo de función 26 utiliza el esquema de período en contexto para generar funciones de documento para el periodo dependiente de tipos de datos, y asigna un valor de la función a funciones de documento generado basadas en si un período identificado a partir de un contexto de documento (por ejemplo, un título de documento o metadatos) corr-spoiids al período especificado en la parte del texto no estructurado. En una realización, el módulo de funciones 26 identifica un atributo dependiente del período a partir de un contexto del documento de texto no estructurado, el contexto definido por uno de un título asociado con el documento de texto no estructurado y los metadatos asociados con el documento de texto no estructurado, genera una función del documento para el atributo dependiente del período, y asigna un primer valor de función a la función del documento generado si el atributo dependiente del período se incluye en la parte del texto no estructurado.

[**0053]** El módulo de función 26 utiliza el esquema más cercano individual de coincidencia de etiqueta para generar una función del documento para la información marcada/etiquetada que se produce más cercana al atributo de candidato, a su izquierda de la derecha, respectivamente. Por ejemplo, en una realización, el módulo de funciones 26 genera una función del documento para información marcada/etiquetada más cercana a un atributo de candidato incluido en la porción de texto no estructurado, y asigna un valor de función a la función del documento generada en base a un índice numérico de la información etiquetada más cercana al al menos un atributo de candidato.

[**0054]** El módulo de función 26 utiliza el esquema de log del valor para atributos basados en la figura para generar valores de funciones que representan el registro del valor real de los atributos de candidatos a base de la figura. En una realización, el módulo de funciones 26 identifica un atributo de evento numérico incluido en la porción de texto no estructurado, genera una función del documento para el atributo de evento numérico identificado y asigna un valor de función a la función del documento generada en base a un logaritmo del atributo del evento numérico.

[**0055]** En una realización, una vez que se genera una pluralidad de funciones del documento, el módulo de función 26 normaliza los valores de funciones obtenidos utilizando todos o algunos de los esquemas de generación de funciones anteriormente descritas. En una realización, el módulo de funciones 26 normaliza los valores de funciones asignados usando Term Frequency-Inverse Document Frequency (TF-IDF). En otra realización, el módulo de funciones 26 normaliza los valores de funciones asignados usando otros esquemas de normalización.

[**0056]** Haciendo referencia a la FIG. 2, una vez que el módulo de funciones 26 genera las funciones del documento para ejemplos positivo y negativo, en el paso 56, el módulo de clasificación 28 usa los ejemplos positivos y negativos para entrenar múltiples clasificadores binarios para cada tipo de atributo de evento predefinido. En una realización, cada uno de los clasificadores binarios usa un algoritmo de clasificación diferente, un conjunto de funciones del documento generadas y/o un subconjunto diferente de documentos de entrenamiento. A continuación, en el paso 58, para cada clasificador entrenado, el módulo de clasificación 28 entrena un modelo de estimación de probabilidad utilizando uno de varios esquemas existentes. Por ejemplo, en una realización, el módulo de clasificación 28 entrena el modelo de estimación de probabilidad usando una técnica de regresión isotónica. En otra realización, el módulo de clasificación 28 entrena el modelo de estimación de probabilidad usando un esquema de estimación de probabilidad.

[**0057]** A continuación, en el paso 60, para cada evento en el conjunto de eventos 42, el módulo de confianza 60 construye un modelo de confianza. En una realización, el módulo de confianza 60 construye el modelo de confianza calculando primero los recuentos de n-gramos, siendo n configurable, para cada n-gramo único que ocurre en cualquiera de las porciones de texto no estructurado en el conjunto de documentos de capacitación 36 que corresponden a atributos de evento predefinidos en el conjunto de eventos 42. A continuación, el módulo de confianza 60 asigna un puntaje de confianza a cada parte del texto no estructurado. La puntuación de confianza es un promedio de todos los recuentos de n-gramo asociados con cada parte del texto no estructurado. A continuación, el módulo de confianza 60 calcula las propiedades estadísticas para cada una de las porciones de texto no estructurado utilizando las puntuaciones de confianza. Las propiedades estadísticas incluyen, entre otras, una desviación promedio, máxima, mínima y estándar de todas las puntuaciones de confianza. El módulo de confianza 60 genera entonces un primer cuerpo de documentos y un segundo corpus de documentos basado en estas propiedades estadísticas. El primer cuerpo incluye porciones de texto no estructurado del conjunto de documentos de entrenamiento 36 que son verdaderamente positivos para atributos de eventos predefinidos. El segundo cuerpo de documentos incluye porciones de texto no estructurado del conjunto de documentos de entrenamiento 36 que son instancias de falsos positivos para atributos de eventos predefinidos.

[**0058]** Con referencia ahora a la FIG. 3, se divulga un método ejemplar para representar información de un documento de texto no estructurado. Como se muestra en el ejemplo de FIG. 3, en el paso 61, el módulo de normalización 24 normaliza al menos uno del conjunto de documentos no estructurados 44. Como se describió anteriormente, el conjunto de documentos no estructurados puede ser un documento de texto no estructurado D recibido a través de una fuente de noticias en tiempo real. En una realización, el módulo de normalización 24 normaliza el documento D identificando un atributo de candidato incluido en el documento de texto no estructurado, asociando un identificador único con el atributo de candidato, comparando el atributo de candidato con cada uno de los conjuntos de atributos de evento predefinidos y almacenando el atributo de candidato, el identificador único y al menos uno de los atributos de evento predefinidos según la comparación. Los atributos de candidatos pueden ser palabras clave, secuencias de letras, números y caracteres, que se definen en un dominio financiero.

[**0059]** A continuación, en el paso 62, el módulo de normalización 24 identifica atributos de un evento incluido en el documento de texto no estructurado D. Cada uno de los atributos identificados es al menos similar a al menos un atributo de evento incluido en un conjunto de atributos de eventos predefinidos definidos en el conjunto de eventos 42. A continuación, en el paso 64, el módulo de funciones 26 genera funciones del documento a partir del documento de texto no estructurado utilizando uno o más de los esquemas de generación de funciones discutidos anteriormente.

[**0060]** Por ejemplo, en una realización, el módulo de función 26 puede aplicar el esquema de generación de funciones de bolsa de palabras mediante la generación de una función del documento para cada palabra única, frase o texto normalizado que aparece en una parte del documento de texto no estructurado, y la asignación de un valor de función a la función del documento generado en función de una cantidad de veces que cada palabra, frase o texto normalizado, respectivamente, aparece en la porción del documento de texto no estructurado.

[**0061]** El módulo de función 26 también puede aplicar el esquema de generación de funciones de distancia más lejana/distancia más cercana por la identificación de texto adyacente a uno de los atributos identificados a partir de una pluralidad de texto predefinido asociado con el conjunto de atributos de eventos predefinidos, generando una función de documento para el texto adyacente identificado y la asignación de un valor de función a la función de

documento generado que representa una distancia espacial entre el texto adyacente identificado y uno de los atributos identificados.

[**0062]** En una realización, por ejemplo, el módulo de función 26 puede aplicar el esquema de generación de funciones de antes o después mediante la identificación de texto adyacente a uno de los atributos identificados, generando una función del documento para el texto adyacente identificado, la asignación de un primer valor de función a la función de documento generado si el texto adyacente identificado se incluye en una pluralidad de texto predefinido asociado con el conjunto de atributos de evento predefinidos y el texto adyacente identificado ocurre después del atributo identificado en la porción de texto no estructurado.

[**0063]** El módulo de función 26 también puede asignar un segundo valor de la función para la función de documento generado si el texto adyacente identificado está incluido en la pluralidad de texto predefinido asociado con el conjunto de atributos de eventos predefinidos y se produce el texto adyacente identificado antes del atributo identificado en la parte del texto no estructurado. El módulo de funciones 26 puede asignar un tercer valor de función a la función del documento generada si el texto adyacente identificado no está incluido en la pluralidad de texto predefinido asociado con el conjunto de atributos de evento predefinidos.

[**0064]** El módulo de función 26 puede aplicar el esquema de generación de funciones de calificador-presente mediante la identificación de calificador de texto incluido en la parte de texto no estructurado, la generación de una función del documento para el calificador de texto identificado, y la asignación de un valor de función a la función de documento generado que representa si el texto calificador identificado se incluye en una pluralidad de texto calificador predefinido asociado con el conjunto de atributos de evento predefinidos.

[**0065]** En una realización, el módulo de función 26 puede aplicar el esquema de generación de funciones de delimitador-presente por la identificación de un delimitador incluido en la parte de texto no estructurado, la generación de una función del documento para el delimitador identificado, y asignando un valor de función a la función de documento generado que representa si el delimitador identificado está incluido en una pluralidad de delimitadores predefinidos asociados con el conjunto de atributos de eventos predefinidos.

[**0066]** El módulo de funciones 26 puede aplicar el esquema de generación de funciones de umbral de valor de figura identificando un atributo de evento numérico incluido en la porción de texto no estructurado, generando una función del documento para el atributo de evento numérico identificado, comparando el atributo de evento numérico con un valor umbral predefinido y asignación de un valor de función a la función del documento generada en función de la comparación.

[**0067]** En una realización, el módulo de función 26 puede aplicar el esquema de generación de funciones N-gramos mediante la identificación de cada N-Gramo único incluido en la porción de texto no estructurado, la generación de una función del documento para cada uno de los n-gramos identificados, y asignando un valor de función para la función del documento generado basado en una frecuencia en que cada N-gramo único identificado ocurre en la porción de texto no estructurado.

[**0068]** El módulo de funciones 26 puede aplicar el esquema de generación de funciones de palabras de título mediante la identificación de texto adyacente a uno de los atributos identificados, generando una función del documento para el texto adyacente identificado, y la asignación de una función de valor a la función del documento generado que representa si el texto adyacente identificado se incluye en un título asociado con el documento de texto no estructurado y una pluralidad de texto predefinido asociado con el conjunto de atributos de evento predefinidos.

[**0069]** En una realización, por ejemplo, el módulo de función 26 puede aplicar el esquema de generación de funciones de período en contexto mediante la identificación de un atributo dependiente de período de un contexto del documento de texto estructurado, el contexto definido por un título asociado con el documento de texto no estructurado o metadatos asociados con el documento de texto no estructurado, generando una función de documento para el atributo dependiente del período y asignando un primer valor de función a la función del documento generado si el atributo dependiente del período se incluye en la porción de texto no estructurado.

[**0070]** El módulo de función 26 puede aplicar el esquema de generación de funciones de etiqueta de ajuste único más cercano mediante la generación de una función del documento para el texto adyacente más cercano al atributo identificado en la parte del texto no estructurado, y la asignación de un primer valor de función a la función de documento generado basada en un índice numérico del texto adyacente más cercano al atributo identificado.

[**0071]** En otra realización más, el módulo de función 26 puede aplicar el log del valor para el esquema de generación de funciones de atributos basados en figura mediante la identificación de un atributo de evento numérico incluido en la parte de texto no estructurado, generando una función del documento para el evento numérico identificado y asignar un valor de función a la función del documento generado en función de un logaritmo del atributo de evento numérico.

[**0072]** A continuación, como se muestra en el paso 66 de la FIG. 3, el módulo de clasificación 28 aplica al menos uno de una pluralidad de clasificadores a cada una de las funciones del documento generado. El al menos un clasificador

previamente entrenado usando un atributo de evento predefinido correspondiente al atributo de evento identificado. A continuación, en el paso 68, el módulo de clasificación 28 calcula un valor de probabilidad a partir de una puntuación de clasificador generada por el al menos un clasificador usando uno de los modelos de estimación de probabilidad previamente entrenados. El valor de probabilidad calculado que indica una probabilidad del atributo de evento identificado correspondiente a uno del conjunto de atributos de evento predefinidos.

[0073] Como se muestra en el paso 70, el módulo de clasificación 28 luego calcula una puntuación de clasificación para cada atributo identificado en D utilizando los valores de probabilidad calculados. En una realización, el módulo de clasificación 28 calcula la puntuación de clasificación combinando los resultados de los clasificadores. Por ejemplo, en una realización, el módulo de clasificación 28 normaliza y/o convierte puntuaciones brutas asignadas por los clasificadores a probabilidades usando un esquema de normalización o estimación de probabilidad. En una realización, el módulo de clasificación 28 usa la regresión isotónica en la normalización de los puntajes brutos, pero el módulo de clasificación 28 también puede utilizar otros esquemas de estimación conocidos en la técnica. Estos puntajes normalizados se combinan en un solo puntaje como una combinación lineal ponderada. En una realización, el módulo de clasificación 28 determina los pesos empíricamente. En otra realización, el módulo de clasificación 28 determina los pesos aplicando validación cruzada en cada atributo identificado.

[**0074]** A continuación, en el paso 72, el módulo de clasificación 28 determina si el atributo identificado en D ha sido positivamente identificado como un atributo en el conjunto de atributos de eventos predefinidos. Si el módulo de clasificación 28 determina que lo identificado en D se identifica positivamente, en el paso 74, el módulo de clasificación aplica al menos uno del conjunto de reglas predefinidas 40 al atributo identificado. Cada uno del conjunto de reglas predefinidas 40 identifica patrones en porciones de texto adyacente al evento en D.

[**0075]** Por ejemplo, refiriéndose a la porción ejemplar siguiente de texto adyacente al atributo de evento de figura de "1.1p por acción", según lo identificado por un clasificador:
"En este período ***se pagó*** un dividendo de **1,1p por acción por un** total de £ 2,1m con respecto al período que finalizó el 1 de octubre de 2006". A continuación se establece una regla predefinida de ejemplo:
".*__Ficha candidato.__*(*era* |*anteriormente)[ ] + (pagado*|*propuesto*|*declarado*|*recomendado).**".
En una realización, la regla predefinida de ejemplo es una *regla de expresión regular* que identifica cifras numéricas para dividendos que se han pagado o declarado anteriormente y, por lo tanto, el sistema no los considera noticias. En una realización, la regla predefinida devuelve un valor verdadero si el atributo de evento de figura (1,1p por acción) es seguido por las palabras "se pagó, se declaró, se propuso o se recomendó".

[**0076]** Las reglas condicionales también pueden incluirse en el conjunto de reglas predefinidas 40. Por ejemplo, en una realización, las fechas son identificadas en el contexto de atributos identificados y se comparan con la fecha o período de noticias de texto publicado. Si la fecha pertenece a un período anterior, la regla devuelve verdadero, lo que indica que las fechas se relacionan con información anterior.

[**0077]** A continuación, en el paso 76, si el módulo de clasificación 28 determina que el atributo identificado satisface una o más reglas aplicadas, en el paso 78, el módulo de clasificación 28 identifica cualesquiera atributos de evento predefinidos adicionales que corresponden al atributo identificado.

[**0078]** A continuación, en el paso 80, el módulo de confianza 30 asigna una puntuación de confianza para el evento en D usando uno de los modelos de confianza previamente entrenados. Una vez que se asigna la puntuación de confianza, en el paso 82, el módulo de confianza 30 compara la puntuación de confianza asignada al evento con una puntuación de confianza asociada con un modelo de confianza entrenado. En base a la comparación, en el paso 84, el módulo de extracción 32 representa el evento del documento de texto no estructurado D y uno o más atributos identificados en un formato estructurado basado en la puntuación del clasificador y la puntuación de confianza. En una realización, el módulo de confianza 30 calcula la puntuación de confianza asociada con el evento promediando todos los recuentos de N-gramos derivados de una porción de texto no estructurado adyacente e incluyendo el evento en D. El módulo de confianza 30 luego compara la puntuación de confianza calculada asociada con el evento promedio estimado asociado con al menos un atributo de evento incluido en el conjunto de atributos de evento predefinidos. En una realización, el módulo de confianza 30 determina cuántas desviaciones estándar por encima o por debajo del promedio estimado previamente es la puntuación de confianza calculada. El módulo de confianza 30 luego asigna la puntuación de confianza al evento basándose en la comparación.

[**0080]** En otra realización, el módulo de confianza 30 determina, si la puntuación de confianza sobrepasa un umbral de valor, si un atributo de evento identificado incluido en la parte de texto no estructurado es probable que se identifica por un modelo M entrenado en el primer cuerpo o segundo cuerpo antes mencionados de los documentos. Como se discutió anteriormente, el primer cuerpo de documentos incluye texto no estructurado del conjunto de documentos de capacitación 36 previamente determinado como un verdadero positivo para el atributo de evento y el segundo cuerpo de documentos incluye porciones de texto no estructurado del conjunto de documentos de capacitación 36 que son instancias falsas positivas para atributos de eventos predefinidos.

[**0081]** En una realización, el módulo de confianza 30 calcula la probabilidad de que el atributo de evento $P_M(c)$ se identifique utilizando el primer cuerpo o el segundo cuerpo utilizando la siguiente fórmula:

$$P_M(c) = \sum_{\forall n-gramo\ n \in c} \log\left(pgen_M(n)\right)$$

donde $pgen_M(n)$ es una probabilidad de un modelo M entrenado en el primer cuerpo de texto no estructurado para generar el n-gramo n y se calcula mediante:

$$pgen_M(n) = \frac{S(\text{recuento}_M(n))}{\sum_{\forall i \in M} recuento(i)}$$

donde $S()$ es una función de suavizado de Good-Turing para representan 0 ocurrencias n-gramos.

[**0082]** Si la probabilidad calculada del atributo de evento es menor que un valor umbral de probabilidad asociado con el modelo M entrenado en el primer cuerpo de texto no estructurado, el módulo de confianza 30 disminuye el valor de la puntuación de confianza computarizada. De lo contrario, el módulo de confianza 30 mantiene el valor de la puntuación de confianza calculada.

[**0083]** En otra forma de realización, el módulo de confianza 30 aumenta la puntuación de confianza calculada para el atributo de evento si un clasificador binario clasifica la parte de texto estructurado como positivo para el atributo de evento, y disminuye la puntuación de confianza calculada para el atributo de candidato si el clasificador binario clasifica la porción de texto no estructurado como negativa para el atributo de evento.

[**0084]** Varias funciones del sistema pueden implementarse en hardware, software, o una combinación de hardware y software. Por ejemplo, algunas funciones del sistema pueden implementarse en uno o más programas de computadora que se ejecutan en computadoras programables. Cada programa puede implementarse en un lenguaje de programación orientado a objetos o de procedimientos de alto nivel para comunicarse con un sistema informático u otra máquina. Además, cada uno de estos programas de computadora puede almacenarse en un medio de almacenamiento tal como memoria de solo lectura (ROM) legible por una computadora o procesador programable de propósito general o especial, para configurar y operar la computadora para realizar las funciones descritas anteriormente.

REIVINDICACIONES

1. Un sistema (10) para extracción automática de datos no estructurados introduce un formato de datos estructurado que comprende:
un servidor (12) que incluye un procesador (14) y memoria (16, 20) que almacena instrucciones que, en respuesta a recibir una primera solicitud para acceder a un servicio, hacen que el procesador (14):

identifique (62) mediante un módulo de normalización (24) los atributos de un evento incluido en un documento de texto no estructurado, siendo cada uno de los atributos identificados similar a al menos un atributo de evento incluido en el conjunto de atributos de eventos predefinidos;
genere (64) mediante un módulo de funciones (26) funciones del documento para cada uno de los atributos identificados;
aplique (66) mediante un módulo de clasificación (28) al menos uno de la pluralidad de clasificadores a cada una de las funciones del documento generado, el al menos un clasificador previamente entrenado utilizando el atributo de evento predefinido correspondiente al atributo de evento identificado;
calcule (68) mediante el módulo de clasificación (28) un valor de probabilidad a partir de una puntuación de clasificación generada por el al menos un clasificador usando un modelo de estimación de probabilidad, el valor de probabilidad indicando una probabilidad de que el atributo de evento identificado corresponda a uno del conjunto de atributos de eventos definidos;
combine (70) mediante el módulo de clasificación (28) una pluralidad de valores de probabilidad calculados asociados con los atributos identificados para generar una puntuación de clasificación; y
extraiga (84) mediante un módulo de extracción (32), del documento de texto no estructurado, el evento y los atributos identificados, en donde el módulo de extracción (32) se extrae automáticamente del documento de texto no estructurado en un formato estructurado basado, al menos en parte, en la puntuación de clasificación.

2. Un método para uso en un sistema de extracción automática de datos no estructurados introduce un formato de datos estructurados que comprende:

identificar (62) atributos de un evento incluido en un documento de texto no estructurado, cada uno de los atributos identificados similar a al menos un atributo de evento incluido en un conjunto de atributos de eventos predefinidos;
generar (64) funciones del documento para cada uno de los atributos identificados; aplicar (66) al menos uno de una pluralidad de clasificadores a cada una de las funciones del documento generadas, el al menos un clasificador previamente entrenado usando el atributo de evento predefinido correspondiente al atributo de evento identificado;
calcular (68) un valor de probabilidad a partir de una puntuación de clasificación generada por el al menos un clasificador usando un modelo de estimación de probabilidad, indicando el valor de probabilidad una probabilidad del atributo de evento identificado correspondiente a uno del conjunto de atributos de evento predefinidos;
combinar (70) una pluralidad de valores de probabilidad calculados asociados con los atributos identificados para generar una puntuación de clasificación; y
representar (84), a partir del documento de texto no estructurado, el evento y los atributos identificados, en donde un módulo de extracción se extrae automáticamente del documento de texto no estructurado en un formato estructurado basado al menos en parte en la puntuación de clasificación.

3. El método de la reivindicación 2, que comprende además:

aplicar (74) al menos una regla de una pluralidad de reglas predefinidas a cada uno de los atributos identificados; y
determinar (76) si cada uno de los atributos identificados es similar a al menos un atributo de evento incluido en el conjunto de atributos predefinidos basado en al menos una regla.

4. El método de la reivindicación 2, que comprende además:

asignar (80) una puntuación de confianza al evento utilizando al menos un modelo de confianza; comparar (82) la puntuación de confianza asociada con el evento con una puntuación de confianza asociada con un modelo de confianza entrenado; y representar (84), a partir del documento de texto no estructurado, el evento y los atributos identificados en el formato estructurado basado en la comparación.

5. El método de la reivindicación 4, en el que identificar los atributos del evento comprende normalizar (61) el documento de texto no estructurado.

6. El método de la reivindicación 5, en el que la normalización (61) del documento de texto no estructurado comprende:

identificar (52) un atributo de candidato incluido en el documento de texto no estructurado; asociar un identificador único con el atributo de candidato;

comparar el atributo de candidato con cada uno de los conjuntos de atributos de eventos predefinidos; y
almacenar el atributo de candidato, el identificador único y al menos uno de los atributos de evento predefinidos en función de la comparación.

7. El método de la reivindicación 6, en el que los atributos candidatos son una de las palabras clave, secuencias de letras, números y caracteres, los atributos candidatos definidos en un dominio financiero.

8. El método de la reivindicación 4, que comprende además:

identificar una porción de texto no estructurado adyacente e incluir el evento, la porción de texto no estructurado que tiene un tamaño de texto configurable por el usuario;
calcular la puntuación de confianza asociada con el evento promediando todos los recuentos de N-gramos derivados de la porción de texto no estructurado;
comparar la puntuación de confianza calculada asociada con el evento con un promedio estimado anterior asociado con al menos un atributo de evento incluido en el conjunto de atributos de evento predefinidos; y
asignar la puntuación de confianza al evento basado en la comparación.

9. El método de la reivindicación 8, que comprende además determinar, si la puntuación de confianza excede un valor umbral, si un atributo de candidato incluido en la porción de texto no estructurado es probable que sea identificado por un modelo M capacitado en un primer cuerpo de texto no estructurado, el primer cuerpo de texto no estructurado es una porción de texto no estructurado que se determina que es un verdadero positivo para el atributo de evento.

10. El método de la reivindicación 9, en el que la probabilidad de que el atributo de candidato sea identificado por el modelo M entrenado en el primer cuerpo de texto no estructurado P_M (c) se calcula por:

$$P_M(c) = \sum_{\forall n-gramo\ n \in c} \log\left(pgen_M(n)\right)$$

donde $pgen_M(n)$ es una probabilidad del modelo M se entrena en texto no estructurado para generar el n-gramo n y se calcula mediante:

$$pgen_M(n) = \frac{S(\text{recuento}_M(n))}{\sum_{\forall i \in M} recuento(i)}$$

donde S ( ) es una función de suavizado para justificar n gramos de 0 ocurrencia.

11. El método de la reivindicación 10, en el que si la probabilidad calculada del atributo de candidato es menor que un valor de probabilidad umbral asociado con el modelo entrenado en el primer cuerpo de texto no estructurado, disminuyendo el valor de la puntuación de confianza calculada.

12. El método de la reivindicación 10, que comprende además:

aplicar un clasificador binario a la porción de texto no estructurado;
aumentar la puntuación de confianza calculada para el atributo de candidato si el clasificador binario clasifica la porción de texto no estructurado como positiva para el atributo de evento; y
disminuir la puntuación de confianza calculada para el atributo de candidato si el clasificador binario clasifica la porción de texto no estructurado como negativa para el atributo de evento.

13. El método de la reivindicación 2, en el que el modelo de estimación de probabilidad utiliza regresión isotónica o un esquema de estimación de probabilidad y la puntuación de clasificación generada es una combinación lineal ponderada de la pluralidad de valores de probabilidad calculados.

14. El método de la reivindicación 2, en el que generar las funciones del documento para cada uno de los atributos identificados comprende aplicar una pluralidad de esquemas de generación de funciones a los atributos identificados.

15. El método de la reivindicación 2, que comprende además entrenar (56) la pluralidad de clasificadores que usan una pluralidad de esquemas de generación de funciones, un conjunto de documentos de entrenamiento que incluyen cada uno al menos un evento candidato y el conjunto de atributos de evento predefinidos.

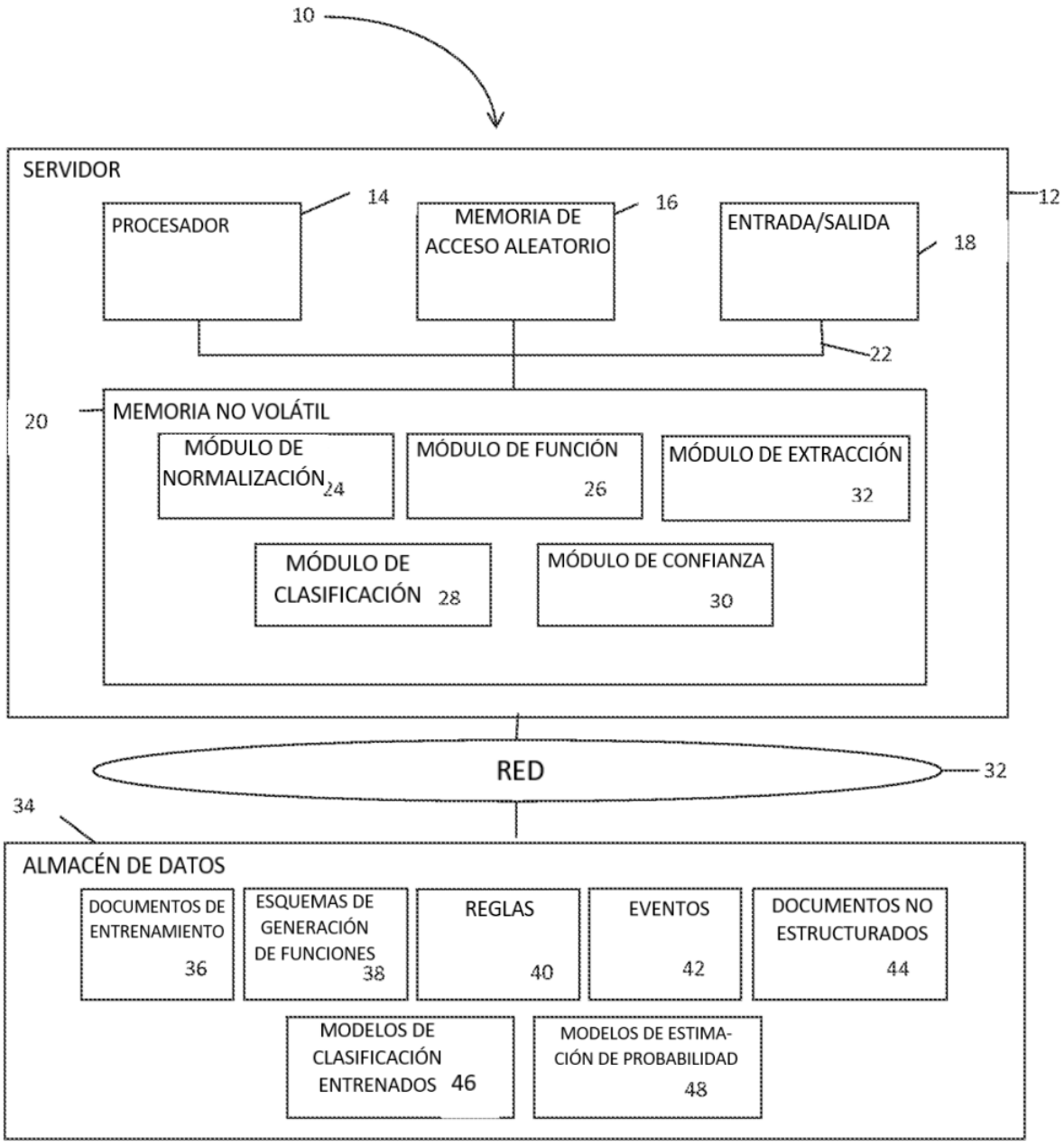
10
SERVIDOR
12
PROCESADOR
14
MEMORIA DE ACCESO ALEATORIO
16
ENTRADA/SALIDA
18
22
20
MEMORIA NO VOLÁTIL
MÓDULO DE NORMALIZACIÓN 24
MÓDULO DE FUNCIÓN 26
MÓDULO DE EXTRACCIÓN 32
MÓDULO DE CLASIFICACIÓN 28
MÓDULO DE CONFIANZA 30
RED
32
34
ALMACÉN DE DATOS
DOCUMENTOS DE ENTRENAMIENTO 36
ESQUEMAS DE GENERACIÓN DE FUNCIONES 38
REGLAS 40
EVENTOS 42
DOCUMENTOS NO ESTRUCTURADOS 44
MODELOS DE CLASIFICACIÓN ENTRENADOS 46
MODELOS DE ESTIMACIÓN DE PROBABILIDAD 48

FIG. 1

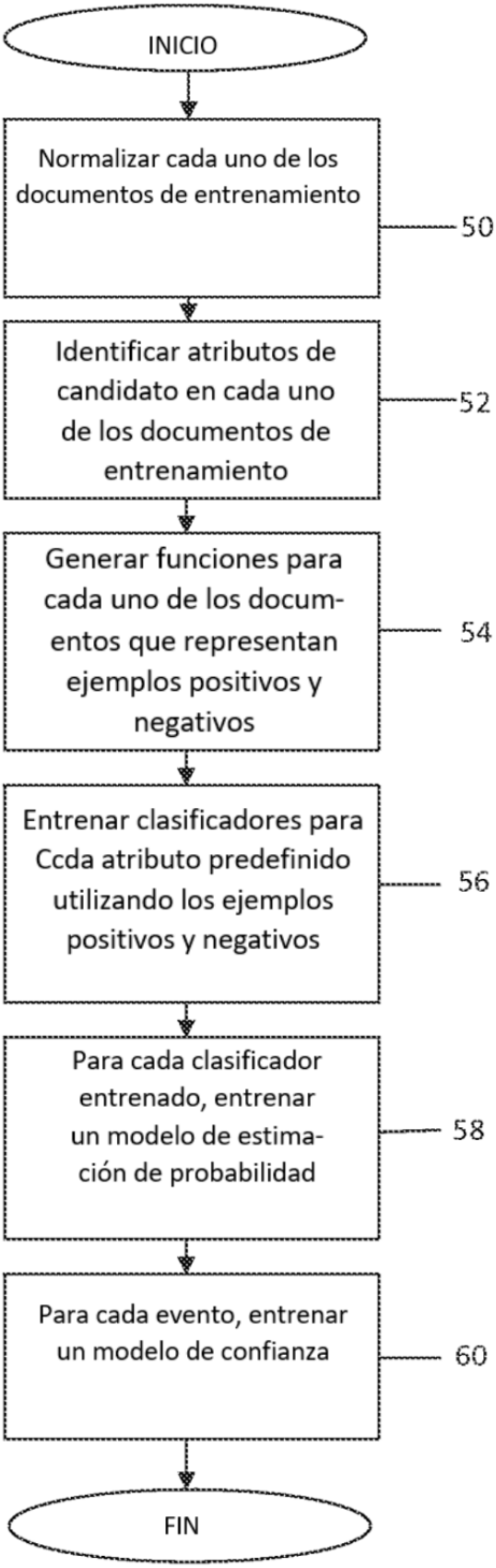
INICIO
Normalizar cada uno de los documentos de entrenamiento
50
Identificar atributos de candidato en cada uno de los documentos de entrenamiento
52
Generar funciones para cada uno de los documentos que representan ejemplos positivos y negativos
54
Entrenar clasificadores para Ccda atributo predefinido utilizando los ejemplos positivos y negativos
56
Para cada clasificador entrenado, entrenar un modelo de estimación de probabilidad
58
Para cada evento, entrenar un modelo de confianza
60
FIN

FIG. 2

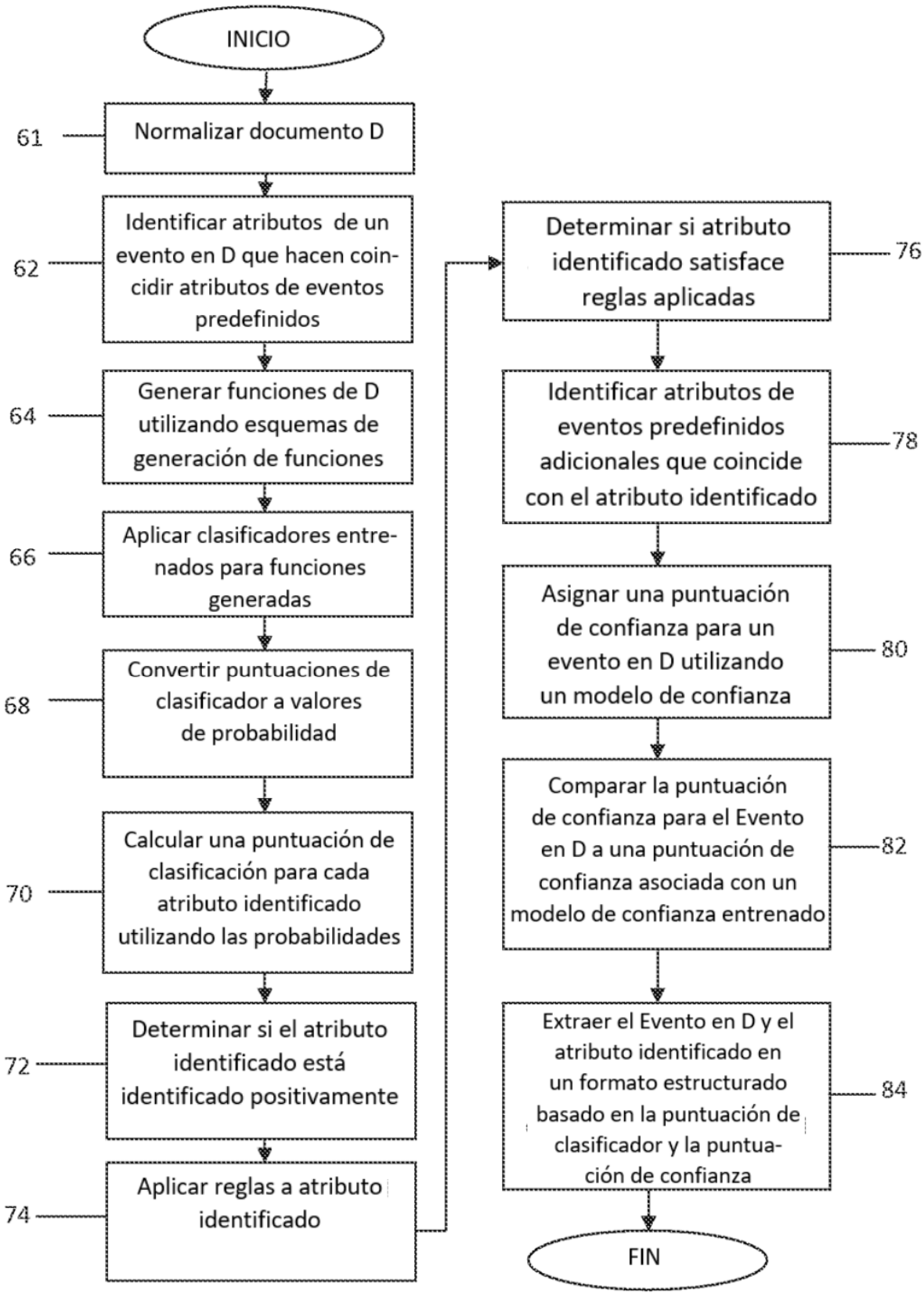
INICIO
61
Normalizar documento D
62
Identificar atributos de un evento en D que hacen coincidir atributos de eventos predefinidos
64
Generar funciones de D utilizando esquemas de generación de funciones
66
Aplicar clasificadores entrenados para funciones generadas
68
Convertir puntuaciones de clasificador a valores de probabilidad
70
Calcular una puntuación de clasificación para cada atributo identificado utilizando las probabilidades
72
Determinar si el atributo identificado está identificado positivamente
74
Aplicar reglas a atributo identificado
76
Determinar si atributo identificado satisface reglas aplicadas
78
Identificar atributos de eventos predefinidos adicionales que coincide con el atributo identificado
80
Asignar una puntuación de confianza para un evento en D utilizando un modelo de confianza
82
Comparar la puntuación de confianza para el Evento en D a una puntuación de confianza asociada con un modelo de confianza entrenado
84
Extraer el Evento en D y el atributo identificado en un formato estructurado basado en la puntuación de clasificador y la puntuación de confianza
FIN

FIG. 3

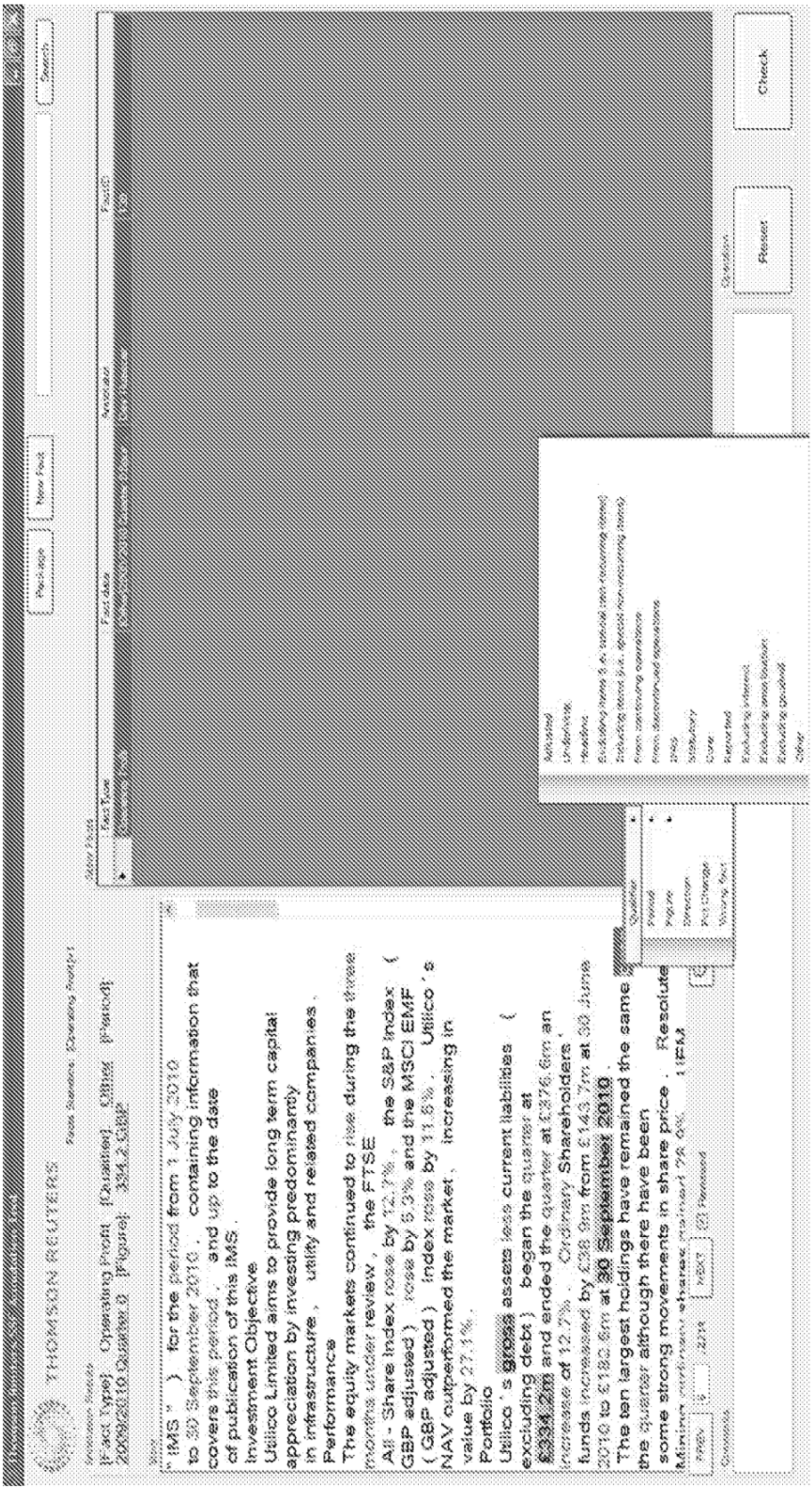

Fig. 4