

SISTEMAS Y MÉTODOS PARA DESARROLLAR UNA BASE DE CONOCIMIENTOS COMPUESTA POR DATOS RECOPIADOS DE INNUMERABLES FUENTES

REFERENCIA CRUZADA A SOLICITUDES RELACIONADAS

[0001] Esta solicitud reclama el beneficio de y la prioridad sobre la Solicitud de Patente Provisional de EE. UU. N.º 63/511,773, presentada el 3 de julio de 2023, la cual se incorpora por referencia en el presente documento en su totalidad.

CAMPO TÉCNICO

[0002] La presente divulgación se refiere a sistemas de extracción de información y gestión de información, y más específicamente, a sistemas y métodos para enriquecer bases de datos con información de varias fuentes de datos, incluidas fuentes de datos no estructuradas e incompletas.

ANTECEDENTES

[0003] La información se puede recopilar con diferente granularidad. Las instituciones dependen de la información recopilada para tomar decisiones, por ejemplo, decisiones relacionadas con el desarrollo tecnológico, negocios, investigación, contratación, inversiones, etc. De manera similar, los individuos también dependen de la información para tomar decisiones, aunque a una escala menor. Con la explosión de información debido a la democratización del descubrimiento de conocimientos, datos o material importantes pueden ser descubiertos en un primer campo y, sin embargo, no estar disponibles en un segundo campo. Los individuos o instituciones en el primer campo pueden aprovechar los datos o el material descubiertos para tomar decisiones, pero los individuos o instituciones en el segundo campo no pueden. La asimetría en las bases de conocimientos entre los dos campos ilustra que, comparativamente, el campo con acceso a información incompleta puede estar en desventaja. Alternativamente, incluso cuando las bases de conocimientos son similares entre dos campos, la incapacidad de un campo para incorporar datos disponibles puede dejar a ese campo en desventaja. En algunas situaciones, un campo es incapaz de incorporar datos disponibles debido al mero tamaño de los datos que genera la explosión de información. La presente divulgación proporciona sistemas y métodos para construir bases de conocimientos y enriquecer estas bases de conocimientos con el tiempo para mejorar la precisión e informar mejor los procesos de toma de decisiones.

RESUMEN

[0004] El término modalidad y términos similares, p. ej., implementación, configuración, aspecto, ejemplo y opción, pretenden referirse ampliamente a todo el objeto de esta divulgación y a las reivindicaciones a continuación. Las declaraciones que contienen estos términos deben entenderse como que no limitan el objeto descrito en el presente documento ni limitan el significado o alcance de las reivindicaciones a continuación. Las modalidades de la presente divulgación cubiertas aquí están definidas por las reivindicaciones a continuación, no por este resumen. Este resumen es una visión general de alto nivel de varios aspectos de la divulgación e introduce algunos de los conceptos que se describen más detalladamente en la sección de Descripción Detallada a continuación. Este resumen no pretende identificar características clave o esenciales del objeto reivindicado. Este resumen tampoco pretende ser utilizado de forma aislada para determinar el alcance del objeto reivindicado. El objeto debe entenderse por referencia a las porciones apropiadas de toda la especificación de esta divulgación, cualquiera o todos los dibujos, y cada reivindicación.

[0005] De acuerdo con ciertos aspectos de la presente divulgación, se proporciona un sistema para construir una base de conocimientos. El sistema incluye un procesador. El sistema incluye además un medio de almacenamiento legible por computadora no transitorio que contiene instrucciones las cuales, cuando se ejecutan en el procesador, hacen que el sistema realice operaciones. Las operaciones incluyen recibir datos multimodales de una o más fuentes; analizar los datos para determinar características; almacenar las características en una base de datos; recibir una consulta de búsqueda para buscar las características almacenadas; analizar la consulta de búsqueda utilizando un gran modelo de lenguaje para extraer características de búsqueda; generar resultados de búsqueda a partir de las características de búsqueda; y mostrar los resultados de búsqueda en una interfaz gráfica de usuario estandarizada que incluye una leyenda que tiene al menos una o más de las características de búsqueda mostradas.

[0006] De acuerdo con ciertos aspectos de la presente divulgación, se proporciona un sistema para construir una base de conocimientos. El sistema incluye un procesador. El sistema incluye además un medio de almacenamiento legible por computadora no transitorio que contiene instrucciones las cuales, cuando se ejecutan en el procesador, hacen que el sistema realice operaciones. Las operaciones incluyen recibir datos multimodales de una o más fuentes, incluyendo los datos multimodales datos de texto; analizar los datos para determinar características; asociar las características determinadas con identificadores de fuente basados en la una o más fuentes; almacenar las características y los identificadores de fuente correspondientes en una base de datos; recibir una consulta de búsqueda para buscar las

características almacenadas; analizar la consulta de búsqueda utilizando un gran modelo de lenguaje para extraer características de búsqueda; generar resultados de búsqueda a partir de las características de búsqueda utilizando al menos las características almacenadas y los identificadores de fuente correspondientes; y mostrar los resultados de búsqueda en una interfaz gráfica de usuario estandarizada que incluye una leyenda que tiene al menos una o más de las características de búsqueda mostradas.

[0007] De acuerdo con ciertos aspectos de la presente divulgación, se proporciona un sistema para construir una base de conocimientos. El sistema incluye un procesador. El sistema incluye además un medio de almacenamiento legible por computadora no transitorio que contiene instrucciones las cuales, cuando se ejecutan en el procesador, hacen que el sistema realice operaciones. Las operaciones incluyen recibir datos multimodales de una o más fuentes, incluyendo los datos multimodales datos de texto; analizar los datos para determinar características; almacenar las características en una base de datos; recibir una consulta de búsqueda para buscar las características almacenadas; analizar la consulta de búsqueda utilizando un gran modelo de lenguaje para extraer características de búsqueda; generar resultados de búsqueda a partir de las características de búsqueda; y mostrar los resultados de búsqueda en una interfaz gráfica de usuario estandarizada que incluye (i) un mapa y (ii) una leyenda que tiene al menos una o más de las características de búsqueda mostradas.

[0008] De acuerdo con ciertos aspectos de la presente divulgación, un método para construir una base de conocimientos incluye recibir datos multimodales de una o más fuentes, incluyendo los datos multimodales datos de texto; analizar los datos para determinar características; almacenar las características en una base de datos; recibir una consulta de búsqueda para buscar las características almacenadas; analizar la consulta de búsqueda utilizando un gran modelo de lenguaje para extraer características de búsqueda; generar resultados de búsqueda a partir de las características de búsqueda; y mostrar los resultados de búsqueda en una interfaz gráfica de usuario estandarizada que incluye (i) un mapa y (ii) una leyenda que tiene al menos una o más de las características de búsqueda mostradas.

BREVE DESCRIPCIÓN DE LOS DIBUJOS

[0009] Las ventajas anteriores y otras de la presente divulgación se harán evidentes al leer la siguiente descripción detallada y al hacer referencia a los dibujos.

[0010] La FIG. 1 ilustra un diagrama de bloques de un sistema para construir bases de conocimientos y enriquecer bases de datos, de acuerdo con algunas implementaciones de la

presente divulgación.

[0011] La FIG. 2 es un diagrama de flujo que ilustra pasos de un proceso para construir bases de conocimientos y proporcionar información a dispositivos cliente, de acuerdo con algunas implementaciones de la presente divulgación.

[0012] La FIG. 3 es una primera interfaz gráfica de usuario, de acuerdo con algunas implementaciones de la presente divulgación.

[0013] La FIG. 4 es una segunda interfaz gráfica de usuario, de acuerdo con algunas implementaciones de la presente divulgación.

[0014] La FIG. 5 es una tercera interfaz gráfica de usuario, de acuerdo con algunas implementaciones de la presente divulgación.

[0015] La FIG. 6 es un diagrama de bloques que muestra información asociada con regiones de un mapa, de acuerdo con algunas implementaciones de la presente divulgación.

[0016] Si bien la presente divulgación es susceptible a varias modificaciones y formas alternativas, implementaciones específicas se han mostrado a modo de ejemplo en los dibujos y se describirán en detalle en el presente documento. Debe entenderse, sin embargo, que la presente divulgación no pretende limitarse a las formas particulares divulgadas. Más bien, la presente divulgación pretende cubrir todas las modificaciones, equivalentes y alternativas que caigan dentro del espíritu y alcance de la presente divulgación según lo definido por las reivindicaciones adjuntas.

DESCRIPCIÓN DETALLADA

[0017] Con la explosión de información, existe la necesidad de ingerir rápida y automáticamente datos de diferentes o varias fuentes y utilizar los datos ingeridos en varios sistemas o aplicaciones. Por ejemplo, en un escenario inmobiliario, las casas o propiedades inmobiliarias pueden estar listadas en un sitio web de listados. Las características de las propiedades pueden diferir. Por ejemplo, las propiedades comerciales pueden tener características diferentes de interés para los compradores en comparación con las propiedades residenciales. El acceso a servicios locales (p. ej., servicios de transporte, parques, etc.) puede diferir entre los compradores de viviendas. Se puede aprender nueva información sobre una casa listada entre el momento en que se lista la casa y el momento en que un comprador potencial ve el listado de la

casa. Sin formas de incorporar la información recién añadida, el listado permanece estático.

[0018] Las modalidades de la presente divulgación proporcionan un sistema de gestión de información que extrae datos de varias fuentes de datos y que puede manejar tanto datos estructurados como no estructurados. El sistema de gestión de información puede estandarizar y enriquecer los datos extraídos utilizando aprendizaje automático y modelos de aprendizaje profundo. El sistema de gestión de información proporciona varias ventajas que incluyen:

[0019] manejar descripciones de texto complejas y largas que pueden incluir información irrelevante;

[0020] identificar y extraer información relevante de las descripciones utilizando modelos de aprendizaje profundo mientras se supera el vocabulario limitado o las búsquedas de cadenas;

[0021] concatenar y estandarizar eficientemente datos extraídos de diversas fuentes, tales como PDF, imágenes y texto;

[0022] obtener datos de contextos de texto no estructurados, tales como títulos de documentos y comentarios de usuarios;

[0023] identificar con precisión categorías y subcategorías para la información proporcionada en los datos fuente, a pesar de posibles inexactitudes en los datos fuente;

[0024] corregir términos específicos del idioma y manejar errores tipográficos, faltas de ortografía y errores;

[0025] agilizar la extracción de información de textos descriptivos utilizando expectativas predefinidas para tipos de datos específicos;

[0026] abordar la falta de características estandarizadas a través de diferentes fuentes de datos;

[0027] evaluar y limpiar los datos extraídos para asegurar precisión y consistencia;

[0028] integrar con fuentes de datos externas para enriquecer las bases de conocimientos con información contextual relevante;

[0029] reentrenar y mejorar continuamente los modelos de aprendizaje profundo y las listas

predefinidas basándose en nuevos datos y comentarios de los usuarios;

[0030] asegurar la escalabilidad y la optimización del rendimiento para manejar grandes volúmenes de datos de varias fuentes de manera eficiente; e

[0031] implementar medidas robustas de seguridad y privacidad de datos para proteger información sensible y cumplir con las regulaciones relevantes y las mejores prácticas de la industria.

[0032] Las modalidades de la presente divulgación se pueden utilizar en varias situaciones. Por ejemplo, los estudiantes que se gradúan de la escuela secundaria tienen cientos, si no miles, de universidades para filtrar a un puñado o un subconjunto para un proceso de solicitud manejable. La información recopilada o generada sobre cada universidad continúa creciendo y puede estar aislada en grupos de expertos, revistas, etc. Las decisiones de los estudiantes estarían mejor informadas si la información aislada fuera accesible e incorporada en los procesos de toma de decisiones de los estudiantes. Por lo tanto, las modalidades de la presente divulgación permiten centralizar y enriquecer los datos de manera que los estudiantes puedan acceder a la información más reciente. Las modalidades de la presente divulgación se pueden utilizar en otras áreas también, por ejemplo, recopilando datos relacionados con la industria de seguros para ayudar a los proveedores de seguros a elaborar grupos de riesgo, recopilando datos relacionados con hipotecas para ayudar a comparar individuos en situaciones similares, recopilando datos relacionados con el ámbito académico para ayudar a los investigadores a obtener la información más reciente (p. ej., información de revistas, universidades, laboratorios de investigación, grupos de expertos, etc.) en su área de especialización.

[0033] La FIG. 1 ilustra un diagrama de bloques de un sistema 100 para construir bases de conocimientos y enriquecer bases de datos, de acuerdo con algunas implementaciones de la presente divulgación. El sistema 100 incluye un servidor 102, un dispositivo cliente 104, uno o más repositorios 106 para almacenar información, y/o uno o más sistemas externos 108 para obtener datos de entrada o proporcionar datos de salida. El servidor 102 y el dispositivo cliente 104 son dispositivos informáticos con al menos un procesador, memoria, dispositivo de almacenamiento e interfaz de red. Ejemplos del dispositivo cliente 104 incluyen una computadora portátil, una computadora de escritorio, un teléfono inteligente, una tableta, un phablet, un asistente digital personal (PDA), un televisor inteligente, un altavoz inteligente, etc. El servidor 102 puede incluir uno o más dispositivos informáticos para realizar funciones descritas en la presente divulgación. Los sistemas externos 108 pueden incluir otros sistemas informáticos.

[0034] El uno o más repositorios 106 pueden almacenar uno o más parámetros de modelo de lenguaje 110, una o más características de clasificación 112, y/o uno o más parámetros de modelo de imagen 114. El uno o más repositorios 106 pueden incluir además otros datos almacenados en una base de datos 116. El uno o más repositorios 106 pueden almacenar cálculos intermedios y otros datos utilizados por el servidor 102. El uno o más repositorios 106 son sistemas informáticos.

[0035] El servidor 102 está configurado para recibir datos o información del dispositivo cliente 104 y/o los sistemas externos 108. En algunas implementaciones, los datos recibidos son datos de listado de una agencia de listados inmobiliarios (p. ej., Zillow, Redfin, etc.). En algunas implementaciones, los datos recibidos son video y/o imágenes de Internet (p. ej., fuentes de redes sociales, agencias de noticias, etc.). El servidor 102 puede almacenar los datos recibidos en el repositorio 106 para actualizar el repositorio 106. El servidor 102 puede fusionar datos recibidos de varias fuentes para almacenar los datos en una estructura de datos más consultable (p. ej., una tabla, una tabla SQL, etc.) o presentar los datos en un formato más digerible (p. ej., un gráfico, una imagen, un informe, etc.). Un usuario del dispositivo cliente 104 puede proporcionar comentarios o actualizaciones a los datos almacenados por el servidor 102. Por ejemplo, el usuario del dispositivo cliente 104 puede proporcionar actualizaciones sobre un número de dormitorios en un listado; agregar amenidades no reflejadas previamente en el listado; eliminar detalles erróneos que han sido corregidos desde el listado original; etc. Los sistemas externos 108 pueden ser sondeados por el servidor 102 a intervalos regulares (p. ej., cada pocas horas, cada día, cada semana, etc.) para actualizar la información almacenada en el repositorio 106. Al utilizar información multimodal de diferentes fuentes, el servidor 102 puede verificar cruzadamente la información a través de las diferentes y diversas fuentes para mejorar la precisión de la información extraída.

[0036] El servidor 102 incluye una interfaz de programación de aplicaciones (API) 120, un motor de ingesta 122, un motor de recompensas 124, un motor de estandarización 126, un motor de extracción de características 128, un motor de identificación de idioma 130, un motor de gestión dinámica de campos 132, un motor de identificación contextual 134, y/o un motor de extracción de expresiones regulares 136. Cada uno de la API 120, el motor de ingesta 122, el motor de recompensas 124, el motor de estandarización 126, el motor de extracción de características 128, el motor de identificación de idioma 130, el motor de gestión dinámica de campos 132, el motor de identificación contextual 134, y/o el motor de extracción de expresiones regulares 136 identificados en la FIG. 1 es una combinación de hardware y software configurada para realizar una funcionalidad específica como se describe en los siguientes párrafos.

[0037] La API 120 del servidor 102 facilita la comunicación entre el dispositivo cliente 104 y el servidor 102. En algunas implementaciones, la API 120 también facilita la comunicación entre el servidor 102 y el uno o más repositorios 106. En algunas implementaciones, la API 210 facilita además la comunicación entre los sistemas externos 108 y el servidor 102. La API 120 empaqueta paquetes de datos hacia (y desde) el dispositivo cliente 104, de modo que haya un flujo de información bidireccional entre el servidor 102 y el dispositivo cliente 104. La API 120 puede empaquetar información (p. ej., imágenes, archivos de video, documentos en formato de documento portátil, entradas web, formatos de texto, etc.) recibida del dispositivo cliente 104 para que esta información proporcionada pueda ser procesada por el servidor 102. La API 120 puede empaquetar información procesada por el servidor 102 de manera que la información procesada pueda ser compartida con o sea legible por el dispositivo cliente 104 y/o los sistemas externos 108. En algunas implementaciones, la API 120 es un servicio web compatible con el protocolo de transferencia de hipertexto (HTTP) y formatos de archivo legibles por máquina, tales como el lenguaje de marcado extensible (XML) y la notación de objetos JavaScript (JSON). La API 120 puede permitir la conexión a servicios de mapeo, datos de mercado y otras bases de datos externas.

[0038] El motor de ingesta 122 del servidor 102 está configurado para recibir flujos de datos de la API 120 y/o realizar un formateo inicial de los datos. En algunas implementaciones, el formateo inicial incluye reescalado, normalización, recorte y aumento para imágenes; extracción de cuadros, muestreo temporal, flujo óptico para videos; tokenización, eliminación de palabras vacías (stop words), lematización o derivación, eliminación de puntuación y caracteres especiales, conversión a minúsculas para textos; extracción de texto y preprocesamiento como archivos de texto para archivos con formato de documento portátil (PDF); remuestreo, conversión de espectrograma, reducción de ruido para audio; remuestreo, manejo de valores faltantes, suavizado para datos de series temporales; etc.

[0039] El motor de recompensas 124 del servidor 102 está configurado para calcular recompensas o puntos para incentivar a los usuarios del dispositivo cliente 104 a subir y corregir información almacenada en el repositorio 106. Las recompensas pueden ser Red Points que se proporcionan a los usuarios del dispositivo cliente 104 basándose en criterios específicos en el programa de recompensas Red Points. En algunas implementaciones, las recompensas (es decir, los Red Points) se pueden proporcionar basándose en que el usuario del dispositivo cliente 104 complete información faltante en el repositorio 106. Por ejemplo, el usuario puede agregar una nueva amenidad (p. ej., una piscina) para un listado de una casa. Al agregar la nueva amenidad, el motor de recompensas 124 puede generar una primera recompensa (p. ej., un primer número de Red Points). En otro ejemplo, el usuario puede corregir información proporcionada en el listado (p.

ej., cambiar 2 baños a 1.5 baños). El motor de recompensas 124 puede generar una segunda recompensa (p. ej., un segundo número de Red Points). La segunda recompensa en algunos casos puede ser mayor que la primera recompensa. Es decir, corregir información puede ser valorado más alto por el motor de recompensas 124 en comparación con agregar nueva información. En algunas implementaciones, corregir información (o texto correctivo) puede ser valorado más bajo por el motor de recompensas 124. En algunas implementaciones, las recompensas generadas pueden basarse en la categoría de la información proporcionada. Por ejemplo, se puede proporcionar una primera recompensa por actualizar o agregar un número de dormitorios, y se puede proporcionar una segunda recompensa por actualizar o agregar una información de contacto. La segunda recompensa puede ser mayor que la primera recompensa basándose en que la categoría de información de contacto se valora más que la categoría de número de dormitorios.

[0040] El motor de recompensas 124 se puede utilizar para impulsar un programa de validación de datos de colaboración abierta (crowd-sourced). Se puede reclutar a usuarios del dispositivo cliente 104 para corregir inexactitudes de datos, mejorando así la integridad general de los datos. La información proporcionada por los usuarios del dispositivo cliente 104 puede ser confirmada o validada. En algunas implementaciones, algunos usuarios pueden ser designados como poseedores de privilegios especiales. Por ejemplo, los corredores (brokers) que publican listados de propiedades pueden tener una identificación específica en comparación con un usuario aleatorio. Además, los cambios de los usuarios del dispositivo cliente 104 pueden ser etiquetados o rastreados. Por lo tanto, se puede proporcionar un historial de versiones de la información almacenada en la base de datos 116 para su sondeo. En algunas implementaciones, los cambios del usuario del dispositivo cliente 104 o cualquier otro texto correctivo pueden desencadenar el control de versiones basándose en si la información se está sobrescribiendo de manera que se almacenen una versión actual y una versión anterior en la base de datos 116.

[0041] El motor de estandarización 126 del servidor 102 está configurado para transformar tipos de datos no estructurados en formatos estructurados y predecibles. El motor de estandarización 126 puede recibir datos no estructurados preprocesados del motor de ingesta 122 y poblar la base de datos 116. Las entradas de datos al motor de estandarización 126 pueden incluir archivos de varios formatos, archivos de video, documentos PDF, imágenes o fotografías, información en formato JSON, entradas web, información de listados, información de datos gubernamentales (p. ej., contratos, propiedad de la tierra, encuestas, seguros, etc.), información sobre hoteles, Airbnb, datos de parques, imágenes de espacios verdes cerca de una propiedad, información demográfica sobre un área, información de seguridad y protección sobre el área, espacio de uso de una propiedad, información de dirección, etc. El motor de estandarización 126 aprovecha

algoritmos para manejar varios estilos de representación de datos para asegurar que la salida de datos esté en un formato estructurado. Por ejemplo, el motor de estandarización 126 puede recibir información de dirección como una entrada única y moldear la información de dirección en un formato estructurado definido. En otro ejemplo, el motor de estandarización 126 recibe la información de dirección como múltiples entradas (p. ej., un número de casa, una calle, una localidad, etc.), y el motor de estandarización 126 concatena las múltiples entradas en un solo elemento y moldea el elemento único en el formato estructurado definido. Las salidas del motor de estandarización 126 pueden ser pares clave-valor en una o más tablas, tablas SQL, etc. En algunas implementaciones, las claves de los pares clave-valor están predefinidas y almacenadas como las características de clasificación 112. Las claves pueden ser categorías o conceptos, y los valores pueden indicar una propiedad o característica asociada con las categorías. En algunas implementaciones, las salidas del motor de estandarización 126 pueden ser informes estructurados o estructuras de datos para interactuar con una o más aplicaciones de software.

[0042] El motor de extracción de características 128 del servidor 102 está configurado para utilizar uno o más algoritmos para extraer información de los datos preprocesados del motor de ingesta 122. Los parámetros para el uno o más algoritmos son parámetros de modelo de lenguaje 110 almacenados en el repositorio 106 y/o parámetros de modelo de imagen 114 almacenados en el repositorio 106. Los parámetros de modelo de lenguaje 110 pueden ser parámetros asociados con uno o más grandes modelos de lenguaje. Los grandes modelos de lenguaje se pueden utilizar para destilar características de datos no estructurados (p. ej., descripciones de listados, contenido generado por usuarios de publicaciones de blogs, etc.). El motor de extracción de características 128 puede ser utilizado por el motor de estandarización 126 para analizar datos no estructurados para obtener información para poblar la una o más tablas proporcionadas por el motor de estandarización 126.

[0043] En algunas implementaciones, el motor de extracción de características utiliza técnicas simples de procesamiento de lenguaje natural como bolsa de palabras, n-gramas o frecuencia de término-frecuencia inversa de documento para procesar y extraer información del texto. Estos procesos se pueden utilizar además de modelos de aprendizaje profundo que capturan matices semánticos y sintácticos de manera más eficiente.

[0044] En algunas implementaciones, el motor de extracción de características 128 puede utilizar un formato de pregunta-respuesta para determinar qué información extraer de los datos preprocesados. El motor de extracción de características 128 puede recibir como entrada descripciones de listados (p. ej., listados inmobiliarios), documentos de compañías de seguros, documentos relacionados con hipotecas, documentos relacionados con la banca, documentos

relacionados con préstamos estudiantiles, documentos relacionados con el ámbito académico, etc., y una o más preguntas relacionadas con la información a extraer de los documentos. En algunas implementaciones, las preguntas incluyen parámetros para permitir que los grandes modelos de lenguaje realicen pasos adicionales. Por ejemplo, la pregunta puede ser "¿cuál es la superficie cuadrada de la habitación más grande en esta propiedad?". El documento que se busca puede no incluir la superficie cuadrada, pero puede incluir dimensiones de las habitaciones de la propiedad. La inclusión de "superficie cuadrada" puede ser un parámetro que desencadene el cálculo de un área de diferentes habitaciones en la propiedad y luego la selección del área más grande.

[0045] En algunas implementaciones, se proporciona una o más preguntas al motor de extracción de características 128 para extraer la misma información. Por ejemplo, se proporcionan diferentes formas de hacer la misma pregunta al motor de extracción de características 128. La pregunta que proporciona la mejor respuesta puede ser votada positivamente por el motor de extracción de características 128 o por un usuario del dispositivo cliente 104. Mediante el voto positivo, el motor de extracción de características 128 y/o el usuario del dispositivo cliente 104 proporciona una Indicación de que la pregunta que proporciona la mejor respuesta debe priorizarse la próxima vez que se generen preguntas para esa característica específica. Esta es una forma de reducir las mejores formas de determinar qué preguntas se utilizan mejor para extraer qué tipo de información. Por ejemplo, se pueden proporcionar las siguientes preguntas: (a) "¿cuántas habitaciones hay en este edificio?"; (b) "¿esta casa tiene cuántas habitaciones?"; (c) "¿cuántos dormitorios?". En algunas implementaciones, la clave "habitaciones" puede ser demasiado general, y el gran modelo de lenguaje puede generalizar y proporcionar una respuesta que incluya la sala de estar junto con los dormitorios, proporcionando así una respuesta deficiente. Por lo tanto, se preferiría la respuesta correcta que indique el número de dormitorios en lugar de habitaciones en general. La pregunta (c) puede ser votada positivamente sobre las otras dos preguntas después de mostrarse más representativa para extraer la información de interés de múltiples documentos.

[0046] Las preguntas con voto positivo se clasifican más alto que las preguntas sin voto positivo. Además, el número de votos positivos en las preguntas se puede comparar para determinar una clasificación para las preguntas, donde las preguntas con el mayor número de votos positivos se clasifican más alto que las preguntas con el menor número de votos positivos. El número total de votos positivos que tiene una pregunta se puede acumular con el tiempo de tal manera que la información histórica asociada con la eficacia de la pregunta para extraer una característica determinada sea capturada por el número total de votos positivos.

[0047] En algunas implementaciones, el motor de extracción de características 128 puede generar preguntas basadas en la característica a extraer. Por ejemplo, una declaración genérica o pregunta genérica como "¿cómo deberíamos preguntar sobre el número de dormitorios en esta propiedad?" puede ser utilizada por el motor de extracción de características 128 para generar una pregunta apropiada para sondear el documento. En algunas implementaciones, la declaración genérica o pregunta genérica puede ser proporcionada por el dispositivo cliente 104 para generar la pregunta que se utilizará como entrada al gran modelo de lenguaje. En algunas implementaciones, el motor de extracción de características 128 determina la pregunta a realizar basándose en palabras clave en el documento. Por ejemplo, si el documento menciona "dormitorios", "número de habitaciones" o términos relacionados, el motor de extracción de características 128 puede generar una o múltiples preguntas, p. ej., "¿Cuántos dormitorios tiene esta propiedad?". Este proceso es específico de la escritura y puede depender del conocimiento específico del dominio. El motor de extracción de características 128 puede seleccionar una de muchas o un subconjunto de preguntas basándose en palabras clave presentes en el documento. El motor de extracción de características 128 puede elaborar la pregunta basándose en las palabras clave en el documento.

[0048] Los grandes modelos de lenguaje pueden ser modelos preentrenados (p. ej., GPT-4, LLaMA, PaLM2, BERT, StableLM), grandes modelos de lenguaje específicos de dominio (p. ej., modelos específicos de bienes raíces, modelos específicos de compañías de seguros, modelos específicos de préstamos estudiantiles, modelos específicos de atención médica, modelos específicos de física, modelos específicos de matemáticas, etc.). En algunas implementaciones, las respuestas proporcionadas por el motor de extracción de características 128 se filtran y se proporcionan con un nivel de confianza. Por ejemplo, si el gran modelo de lenguaje proporciona que el número de habitaciones es de 500 habitaciones, la respuesta de 500 habitaciones puede no pasar una verificación de magnitud. Es decir, la verificación de magnitud puede establecerse en el Orden de 10 para dormitorios, por lo tanto, cualquier respuesta del motor de extracción de características 128 que devuelva un valor que sea significativamente mayor que 10 se marca como incorrecta o se marca como un valor atípico. En algunas implementaciones, un nivel de confianza de una característica específica puede basarse en una magnitud asociada con la característica, un tipo de datos asociado con la característica específica (p. ej., si el tipo de datos para el número de habitaciones es un número o una letra), etc. El nivel de confianza puede ser una bandera (p. ej., una bandera binaria que indica "no es necesaria revisión manual" vs. "se requiere revisión manual", "correcto" vs. "incorrecto", "ok" vs. "revisión adicional", etc.). El nivel de confianza puede citarse como una puntuación de confianza (p. ej., como un número entre 0 y 1 que indica una probabilidad de que el motor de extracción de características 128 haya obtenido un valor correcto para la característica extraída).

[0049] El motor de extracción de características 128 también puede trabajar con modelos de imagen utilizando parámetros de modelo de imagen 114. Las imágenes incluidas en los datos preprocesados del motor de ingesta 122 se proporcionan como entradas a los modelos de imagen. En algunas implementaciones, etiquetas y/o imágenes etiquetadas para entrenar los modelos de imagen también se proporcionan al motor de extracción de características 128. Las imágenes pueden ser imágenes en escala de grises, imágenes en blanco y negro y/o imágenes en color. En algunas implementaciones, las imágenes son imágenes en color proporcionadas como una matriz tridimensional para cada imagen. El motor de extracción de características 128 puede utilizarse para entrenar los modelos de imagen para actualizar los parámetros de modelo de imagen 114. En algunas implementaciones, los modelos de imagen son redes neuronales convolucionales.

[0050] Las características identificadas en imágenes pueden incluir amenidades (p. ej., piscinas, gimnasios, lavandería, etc.). El proceso de entrenamiento puede incluir el uso de un gran conjunto de datos de imágenes para entrenar y validar los modelos de imagen. El gran conjunto de datos de imágenes se puede dividir en imágenes de entrenamiento e imágenes de prueba. Por ejemplo, se pueden incluir 1,000 imágenes en el conjunto de datos con 800 de las imágenes siendo imágenes de entrenamiento y 200 siendo imágenes de prueba. Las imágenes se pueden intercambiar en un momento posterior con 800 imágenes siendo utilizadas para prueba y 200 imágenes siendo utilizadas para entrenamiento. Cada uno de los modelos de imagen se puede entrenar para reconocer una o más características. Se pueden utilizar diferentes modelos de imagen para el reconocimiento de diferentes características en imágenes. El motor de extracción de características 128 —cuando trabaja con modelos de imagen— proporciona salidas de características en las imágenes y/o métricas de precisión como puntuaciones de confianza en cuanto a si la característica está incluida en una imagen específica.

[0051] En algunas implementaciones, el motor de extracción de características 128 utiliza un modelo de "segmentar cualquier cosa" (segment anything model) para segmentar imágenes y clasificar píxeles de las imágenes en categorías. En algunas implementaciones, el modelo de segmentar cualquier cosa se puede utilizar en la fase de entrenamiento mientras se entrenan modelos de imagen. Por ejemplo, generar etiquetas para imágenes puede ser una tarea tediosa, que consume tiempo y recursos, por lo tanto, el modelo de segmentar cualquier cosa se puede utilizar para generar etiquetas para elementos contenidos en las imágenes. Una vez que esos elementos (o características) son identificados, las imágenes se etiquetan y se utilizan para entrenar modelos de imagen especializados en extraer esas características identificadas. Por ejemplo, se puede utilizar un modelo de segmentar cualquier cosa en imágenes para determinar si las imágenes contienen piscinas. Basándose en etiquetar algunas imágenes como que

contienen piscinas y algunas imágenes como que no contienen piscinas, los modelos de imagen se pueden entrenar para identificar piscinas en imágenes utilizando las imágenes etiquetadas.

[0052] El motor de estandarización 126 puede utilizar características identificadas en imágenes por el motor de extracción de características 128 para aumentar o enriquecer la información almacenada en la base de datos 116. Por ejemplo, un listado de propiedad puede incluir imágenes, descripción de la propiedad y algunas amenidades o una tabla de características de la propiedad. El listado de propiedad puede ser ingerido por el servidor 102, y el motor de estandarización 126 puede utilizar el motor de extracción de características 128 para emparejar pares clave-valor para almacenar en una tabla en la base de datos 116. Además, la descripción del listado puede incluir más información que no se proporciona en la tabla de características de la propiedad. Por ejemplo, la descripción puede incluir que la propiedad tiene lavandería en la unidad en lugar de un cuarto de lavado compartido. Este motor de estandarización 126 puede utilizar el motor de extracción de características 128 junto con los parámetros de modelo de lenguaje 110 para completar una clave "lavandería" en la base de datos 116 como "en la unidad". De manera similar, el motor de extracción de características 128 puede utilizar los parámetros de modelo de imagen 114 para identificar características en las imágenes incluidas en el listado de propiedad. Las características identificadas se pueden utilizar para verificar amenidades, por ejemplo, identificar una piscina en una imagen se puede utilizar para sugerir completar una clave "piscina" en la base de datos 116 como "sí". Por lo tanto, el servidor 102 se puede utilizar para enriquecer la información almacenada en la base de datos 116 y mejorar la tabla de características de la propiedad incluida en el listado de propiedad.

[0053] En algunas implementaciones, como se describió anteriormente, las características de imagen se pueden determinar a partir de imágenes asociadas con un primer elemento (p. ej., un listado de propiedad). Por ejemplo, una piscina es una característica de imagen extraída de una o más imágenes asociadas con un primer listado de propiedad. Las características de texto se pueden determinar a partir del texto asociado con el listado de propiedad para determinar si el texto en el listado de propiedad incluye una piscina. Si hay categorías (p. ej., amenidad de piscina) en las características de imagen faltantes en las características de texto, entonces la característica de imagen faltante (es decir, la piscina) se puede almacenar en la base de datos 116. Por ejemplo, como se describió anteriormente, una clave "piscina" se puede indicar como un "sí" en la base de datos 116. A la inversa, si hay características de texto faltantes en las imágenes y la base de datos 116, las características de texto se pueden almacenar en la base de datos 116. La base de datos 116 puede almacenar características en un formato Booleano "sí"/"no" o puede almacenar características en un formato más descriptivo (p. ej., una superficie cuadrada asociada con la característica).

[0054] En algunas implementaciones, se pueden recibir texto y/o imágenes actualizados en el servidor 102. El servidor 102 puede determinar características de imagen actualizadas asociadas con el primer elemento basándose en las imágenes actualizadas, y/o puede determinar características de texto actualizadas asociadas con el primer elemento basándose en el texto actualizado. Por ejemplo, una imagen actualizada puede incluir una nueva imagen que muestra una parte diferente de una propiedad o una nueva imagen que muestra una parte de la propiedad capturada previamente por una imagen anterior. El servidor 102 puede analizar la imagen actualizada para determinar si hay discrepancias asociadas con las características almacenadas asociadas con el primer elemento. En algunos casos, el servidor 102 puede solicitar al dispositivo cliente 104 que reciba una entrada para resolver la discrepancia. En algunos casos, el servidor 102 puede almacenar las características de imagen actualizadas como características almacenadas asociadas con el primer elemento. En algunos casos, el servidor 102 puede crear una nueva categoría asociada con el primer elemento y vincular la nueva categoría a las características de imagen. En algunos casos, un nivel de confianza asociado con una clave o categoría puede actualizarse basándose en las características de imagen. Por ejemplo, una nueva fotografía que muestra la piscina desde una perspectiva diferente puede aumentar un nivel de confianza asociado con la clave "piscina" mientras que una nueva fotografía que muestra un estanque puede reducir un nivel de confianza asociado con la clave "piscina".

[0055] El motor de identificación de idioma 130 del servidor 102 está configurado para determinar el idioma en que está escrito un documento. El motor de identificación de idioma 130 es utilizado por el motor de estandarización 126 para identificar el idioma. El idioma identificado puede ser utilizado por el motor de estandarización 126 para informar al motor de extracción de características 128 sobre qué modelo de lenguaje utilizar. Por ejemplo, los parámetros de modelo de lenguaje 110 pueden incluir parámetros para diferentes grandes modelos de lenguaje. Los diferentes grandes modelos de lenguaje pueden ser grandes modelos de lenguaje en diferentes idiomas. Los diferentes idiomas pueden incluir inglés, español, japonés, chino, portugués, etc. El motor de identificación de idioma 130 puede utilizar modelos de identificación de idioma preentrenados en datos no estructurados para identificar el idioma. Una vez identificado el idioma, el motor de extracción de características 128 puede limitar cuáles de los grandes modelos de lenguaje son apropiados.

[0056] En algunas implementaciones, el motor de identificación de idioma 130 está configurado para determinar el idioma de un documento. Una vez determinado el idioma, entonces el motor de ingesta 122 puede realizar traducción de idioma utilizando API de traducción de idioma o algún otro servicio proporcionado por los sistemas externos 108. Esto puede ser ventajoso donde no se tienen que entrenar múltiples modelos para diferentes idiomas, pero desventajoso donde la

información específica del dominio puede perderse en la traducción.

[0057] El motor de gestión dinámica de campos 132 del servidor 102 está configurado para generar nuevas claves, campos, categorías, características, etc. El motor de gestión dinámica de campos 132 genera los nuevos campos dinámicamente. En algunas implementaciones, las características de clasificación 112 no incluyen una clave específica, por lo tanto, se crea un campo "basurero" (dump field) para albergar valores que no encajan en ninguno de los campos o características predefinidos en las características de clasificación 112. El campo "basurero" puede incluir oraciones, párrafos, etc., de información del documento del cual el motor de estandarización 126 no ha extraído ninguna información. Por ejemplo, la descripción de una hermosa puesta de sol en un listado de propiedad puede no ser fácilmente importada al número de dormitorios, número de baños, precio de la propiedad, etc. El motor de gestión dinámica de campos 132 puede capturar partes de la descripción relacionadas con la hermosa puesta de sol y almacenar toda la descripción en un campo "basurero" para su procesamiento posterior. En algunas implementaciones, el motor de gestión dinámica de campos 132 puede encontrar patrones entre diferentes documentos para crear un nuevo campo. Por ejemplo, si múltiples documentos o descripciones de propiedades hacen referencia a techos altos, el motor de gestión dinámica de campos 132 puede crear un nuevo campo o una nueva clave llamada "techos altos" y poblar valores como "sí". El motor de gestión dinámica de campos 132 creando un nuevo campo puede desencadenar que el motor de estandarización 126 busque en campos "basurero" de información procesada previamente para determinar si estos campos incluyen descripciones que podrían proporcionar valores para el campo recién creado. Por ejemplo, antes de crear el campo "techos altos", se buscan los campos "basurero" de listados de propiedades procesados previamente para determinar si esas propiedades anteriores incluyen "techos altos". En algunas implementaciones, el motor de gestión dinámica de campos 132 puede confiar en métodos de aprendizaje no supervisado tales como agrupamiento o modelado de temas (como Asignación Latente de Dirichlet) para identificar temas o características comunes en descripciones de propiedades.

[0058] En algunas implementaciones, el motor de gestión dinámica 132 puede crear un nuevo campo o categoría basándose en una consulta de búsqueda. Por ejemplo, basándose en una frecuencia de palabras clave buscando un elemento específico, el motor de gestión dinámica 132 puede determinar que la palabra clave representa un campo o categoría. Por ejemplo, la base de datos 116 previamente no incluía una categoría para "puertas de vidrio" para los listados de propiedades almacenados. Después de que un número de usuarios del dispositivo cliente 104 busque "puertas de vidrio", entonces el motor de gestión dinámica 132 puede crear una nueva categoría en todos los listados de propiedades almacenados para "puertas de vidrio". En algunas

implementaciones, la nueva categoría se crea después de que un número de búsquedas alcanza un umbral de búsqueda dentro de un período de tiempo (es decir, un número de búsquedas durante un período de tiempo o una frecuencia de búsquedas). Por ejemplo, 300 búsquedas de la palabra clave dentro de tres días. El motor de gestión dinámica 132 creando la nueva categoría puede desencadenar que el motor de estandarización 126 busque en campos "basurero" de información procesada previamente para determinar valores Booleanos para "puertas de vidrio" para cada uno de los listados de propiedades almacenados.

[0059] El motor de gestión dinámica de campos 132 se puede utilizar con imágenes también. Por ejemplo, las características de clasificación 112 se pueden actualizar para incluir una nueva amenidad, categoría, clave, característica, etc. Si las imágenes para listados ya procesados están almacenadas en la base de datos 116, entonces el motor de extracción de características 128 puede reprocesar esas imágenes utilizando los parámetros de modelo de imagen 114 para identificar qué imágenes incluyen la nueva característica.

[0060] El motor de identificación contextual 134 del servidor 102 puede utilizar la extracción de características basada en el contexto del texto de descripción o texto del documento para identificar y extraer características relevantes, claves, amenidades, etc. Por ejemplo, el motor de identificación contextual 134 puede identificar todos los sustantivos en el texto del documento y clasificar los diferentes sustantivos como características. El motor de identificación contextual 134 puede asumir que todos los sustantivos son características (p. ej., amenidades). El motor de identificación contextual 134 puede identificar partes del discurso y, basándose en las partes del discurso, puede definir palabras clave. Las palabras clave pueden, por ejemplo, ser utilizadas por el motor de gestión dinámica de campos 132 para determinar nuevos campos, el motor de extracción de características 128 para generar preguntas basadas en palabras clave como se describió anteriormente, etc. En un ejemplo de listados inmobiliarios, el motor de identificación contextual 134 puede ser utilizado junto con el motor de gestión dinámica de campos 132 para extraer nuevas amenidades no encontradas previamente. Nuevas amenidades o palabras clave se pueden utilizar para actualizar las características de clasificación 112. Esto permite que las características de clasificación 112 sean cambiadas o actualizadas con el tiempo.

[0061] El motor de extracción de expresiones regulares 136 del servidor 102 puede ser utilizado por el motor de estandarización 126 para emparejar expresiones regulares en descripciones de texto. En un ejemplo, el motor de extracción de características 128 puede utilizar un gran modelo de lenguaje 110 para desarrollar una respuesta o contestación a una pregunta. El motor de extracción de expresiones regulares 136 puede emparejar la respuesta con expresiones regulares como un método para verificar la corrección. Por ejemplo, si las respuestas a un área de una

propiedad deben estar en pies cuadrados o un área de tierra debe estar en acres, entonces la respuesta proporcionada por el motor de extracción de características 128 utilizando el gran modelo de lenguaje 110 debe incluir las unidades esperadas. En algunas implementaciones, el motor de extracción de expresiones regulares 136 se utiliza directamente en descripciones de texto para reducir la complejidad computacional y el tiempo de procesamiento. Por ejemplo, si un listado de propiedad incluye una tabla de características como el número de dormitorios, precio, etc., entonces emparejar la moneda puede proporcionar una forma rápida de extraer el precio del texto y emparejar "dormitorios" puede proporcionar una forma rápida de extraer un número en la vecindad de dormitorios. En algunas implementaciones, modismos específicos o lenguaje específico del dominio pueden ser extraídos o emparejados directamente. En algunas implementaciones, el motor de extracción de expresiones regulares 136 puede utilizar emparejamiento de cadenas y/o confiar en heurísticas basadas en la densidad de palabras clave.

[0062] Refiriéndose a la FIG. 2, se proporciona un proceso 200 para construir bases de conocimientos y enriquecer bases de datos, de acuerdo con algunas implementaciones de la presente divulgación. El proceso 200 puede ser realizado por el sistema 100, y en algunos casos por el servidor 102 en cooperación con el repositorio 106. En el paso 202, el servidor 102 puede recibir datos multimodales de una o más fuentes de datos. Como se describió anteriormente en conexión con la FIG. 1, los datos multimodales pueden obtenerse de los sistemas externos 108 y/o el dispositivo cliente 104. Los datos pueden ser video, texto, imágenes, etc.

[0063] En el paso 204, el servidor 102 analiza los datos recibidos para determinar características. El motor de estandarización 126 en cooperación con diferentes motores del servidor 102 puede analizar los datos para determinar una o más características presentes en los datos recibidos.

[0064] En el paso 206, el servidor 102 almacena las características en la base de datos 116. Como se describió anteriormente en conexión con la FIG. 1, la base de datos 116 puede incluir pares clave-valor donde los valores para claves predefinidas se pueblan utilizando características extraídas de los datos recibidos. El motor de extracción de características 128, el motor de identificación de idioma 130, el motor de gestión dinámica de campos 132, el motor de identificación contextual 134 y/o el motor de extracción de expresiones regulares 136 pueden aprovecharse para extraer características.

[0065] En el paso 208, el servidor 102 genera un informe concerniente a las características en la base de datos 116 para mostrar en el dispositivo cliente 104. El informe se puede publicar como un resultado web. Por lo tanto, el servidor 102 puede actualizar la información en el informe a

medida que haya nuevos datos disponibles, independientemente del modo o formato de los nuevos datos.

[0066] Las modalidades de la presente divulgación pueden permitir varias aplicaciones. Por ejemplo, en un contexto de listado de propiedades inmobiliarias, se pueden identificar ciertas ventajas. El servidor 102 puede implementar una tubería integral de extracción, procesamiento y enriquecimiento de datos que aprovecha técnicas avanzadas de aprendizaje automático, protocolos robustos de limpieza de datos e integraciones con fuentes de datos externas para entregar datos de listado de propiedades estandarizados y de alta calidad. El motor de estandarización 126 puede analizar datos de listados inmobiliarios. El motor de estandarización 126 puede utilizar una variedad de algoritmos, cada uno adaptado para extraer un tipo específico de información como dirección, números de teléfono, amenidades, etc. Estos algoritmos utilizan expresiones regulares y técnicas de aprendizaje automático para analizar texto no estructurado e identificar patrones recurrentes de fuentes de datos.

[0067] El motor de estandarización 126 puede aprovechar el motor de extracción de características 128 para realizar análisis de imágenes y extraer información de imágenes. Las redes neuronales convolucionales pueden permitir etiquetar y categorizar imágenes relacionadas con listados inmobiliarios. A través del uso de extracción de características y agrupamiento de imágenes, imágenes similares se pueden agrupar juntas para identificar amenidades clave como piscinas, vistas, etc.

[0068] El servidor 102 puede facilitar la extracción de amenidades vinculando imágenes en la base de datos 116 a sus amenidades correspondientes, proporcionando un recurso integral para que los modelos de imagen aprendan. Se pueden utilizar modelos de aprendizaje profundo que pueden trabajar con datos multilingües, permitiendo una detección de amenidades más completa a través de una variedad de descripciones de listados.

[0069] El servidor 102 facilita el entrenamiento de modelos. Utilizando información almacenada en la base de datos, las redes neuronales pueden entrenarse para reconocer patrones y categorizar listados inmobiliarios basándose en sus amenidades. Este entrenamiento implica ajustes iterativos a los atributos de los modelos, minimizando los valores posibles hasta que las salidas de los modelos se alineen con las expectativas. El modelo se puede almacenar en los parámetros de modelo de imagen 114.

[0070] Los modelos de imagen y lenguaje entrenados pueden ser utilizados por el motor de extracción de características 128 para clasificar nuevos listados. Los modelos entrenados

pueden facilitar la clasificación de nuevos listados, almacenando la salida en una forma estructurada que incluye tanto el listado como las amenidades detectadas. Los modelos pueden sugerir mejoras al listado, como solicitar al usuario del dispositivo cliente 104 que agregue imágenes de amenidades mencionadas en la descripción pero no mostradas en las imágenes.

[0071] El servidor 102 permite la verificación de datos verificando cruzadamente los datos extraídos con fuentes oficiales externas como registros de transacciones y permisos. Esto se puede realizar comparando información geográfica y detalles de propiedad de estas fuentes con la información en nuestros listados. Las transacciones provienen de fuentes oficiales (p. ej., documentos de arrendamiento, registros gubernamentales, fuentes comerciales, etc.). La información en estos documentos de transacción se puede verificar cruzadamente contra la información proporcionada por corredores, sitios web, etc. Por ejemplo, un listado de propiedad proporcionado por un corredor puede decir que la propiedad es un apartamento, y el servidor 102 puede utilizar fuentes oficiales para confirmar esta información. Además, la verificación cruzada puede permitir confirmar si un espacio específico fue utilizado previamente como un negocio operativo basándose en el registro comercial. La información de transacciones se puede utilizar para aumentar la información en el listado.

[0072] En algunas implementaciones, verificar cruzadamente los datos extraídos con fuentes oficiales externas puede ayudar a reducir una cantidad de información almacenada por el servidor 102. Por ejemplo, la información de fuentes más confiables se puede utilizar para reemplazar información de fuentes menos confiables. De esa manera, para una característica determinada, el servidor 102 puede eliminar la información de fuentes menos confiables para conservar espacio de almacenamiento. En algunas implementaciones, la información se almacena junto con un identificador de fuente, y a medida que los detalles de la propiedad se pueblan para un listado determinado con el tiempo, la fuente de los detalles de la propiedad está fácilmente disponible. En algunas implementaciones, en lugar de un identificador de fuente, una puntuación de confianza asociada con la fuente de los detalles de la propiedad se almacena junto con los detalles de la propiedad.

[0073] En algunas implementaciones, el identificador de fuente se puede utilizar para tratar con información duplicada obtenida de diferentes fuentes. Por ejemplo, la información de un identificador de fuente más confiable se puede retener mientras que la información de un identificador de fuente menos confiable se puede eliminar o desindexar. Por ejemplo, un corredor que proporciona información para un número de dormitorios es más confiable que un usuario anónimo que proporciona la información. La información del corredor puede reemplazar la información proporcionada por el usuario anónimo basándose parcialmente en que el

identificador de fuente para el corredor es más confiable que el identificador de fuente para el usuario anónimo. Por lo tanto, los resultados de búsqueda que devuelven, por ejemplo, el número de dormitorios proporcionarán resultados asociados con la información obtenida por el corredor en lugar del usuario anónimo.

[0074] Los identificadores de fuente se pueden utilizar para reemplazar una cantidad de información almacenada en la base de datos reteniendo solo la información más confiable. En algunos casos, la información menos confiable se elimina de la base de datos 116. En algunos casos, la información menos confiable se comprime y desindexa. La desindexación evita que la información aparezca en los resultados de búsqueda, pero la información puede retenerse en un estado comprimido para fines de control de versiones.

[0075] En alguna implementación, el motor de extracción de características 128 utiliza modelos de procesamiento de lenguaje natural para procesar y simplificar descripciones de texto. Por ejemplo, los modelos de procesamiento de lenguaje natural se pueden utilizar para resumir texto largo para mejorar la eficiencia de extracción de información del texto. En algunas implementaciones, los pasos de preprocesamiento por el motor de ingesta 122 y/o pasos de preprocesamiento por el motor de extracción de características 128 y/o el motor de estandarización 126 se pueden utilizar para limpiar datos, simplificando así el texto, eliminando palabras innecesarias y proporcionando una salida más condensada y útil del motor de estandarización 126. En algunas implementaciones, se utilizan modelos de aprendizaje automático tradicionales en lugar de/o además de modelos de aprendizaje profundo. Los modelos de aprendizaje automático tradicionales incluyen árboles de decisión, máquinas de vectores de soporte o k-vecinos más cercanos.

[0076] El motor de recompensas 124 se puede utilizar para incentivar a los usuarios a subir y corregir datos. Los Red Points se pueden proporcionar a los usuarios para incentivar su adición o corrección de información proporcionada en los listados. Esto no solo aumenta la cantidad de datos para uso de los modelos de lenguaje e imagen para aprender, sino que también mejora la precisión de la información del listado en la base de datos 116.

[0077] El servidor 102 está configurado para utilizar algoritmos de aprendizaje automático multimodal que pueden manejar diferentes tipos de fuentes de datos, incluyendo texto, imágenes y datos estructurados. El servidor 102 puede aplicar modelos específicos para cada tipo. Esto permite la creación de "super listados" que combinan datos de varias fuentes en un solo listado integral. En algunas implementaciones, a medida que nueva información se vuelve disponible, el "super listado" se actualiza en tiempo real. El servidor 102 permite un camino más fácil hacia la

escalabilidad y la optimización, siendo capaz de manejar grandes volúmenes de datos y mejorar continuamente basándose en nuevos datos y comentarios de los usuarios.

[0078] El servidor 102 permite el enriquecimiento de datos de la información almacenada en la base de datos 116. Los listados se mejoran aún más integrándolos con fuentes de datos externas (p. ej., de los sistemas externos 108) que proporcionan información contextual relevante, como tendencias del mercado o datos demográficos.

[0079] Seguridad de Datos y Cumplimiento: Se implementan medidas robustas para proteger la información sensible y asegurar el cumplimiento con las regulaciones de protección de datos relevantes y las mejores prácticas de la industria. En algunas implementaciones, se utilizan técnicas de encriptación para salvaguardar los datos tanto almacenados en el repositorio 106 y/o en tránsito a cualquier componente en el sistema 100 (p. ej., el dispositivo cliente 104). Los datos encriptados solo pueden ser accedidos con las claves de desencriptación apropiadas. En algunas implementaciones, la información de identificación personal (PII) se anonimiza para proteger aún más la privacidad individual, lo que implica eliminar o alterar información identificable.

[0080] En algunas implementaciones, la información del listado o los resultados de reseñas de una propiedad o un vecindario se pueden utilizar para realizar análisis de sentimiento por el servidor 102. Los grandes modelos de lenguaje se pueden entrenar para analizar el sentimiento de las reseñas y comentarios relacionados con los listados o propiedades de unidades múltiples específicas. Esto ayuda a medir el valor percibido y la reputación de la una o más propiedades.

[0081] El servidor 102 puede emplear modelos estadísticos (p. ej., modelos de valoración automatizada (AVM)) para estimar el valor de mercado de una propiedad. Los modelos utilizan los puntos de datos extraídos de los listados y datos del mercado local para hacer estas estimaciones. Cuando la información de precios no está disponible para una propiedad específica, el AVM se puede utilizar para proporcionar a los usuarios estimaciones basadas en la localidad de la propiedad.

[0082] En algunas implementaciones, el informe generado por el servidor 102 puede aprovechar la geocodificación. Es decir, los servicios de geocodificación se pueden utilizar para convertir direcciones en coordenadas geográficas. Las coordenadas geográficas permiten trazar los listados en un mapa para realizar análisis espaciales, lo que permite visualizar, por ejemplo, la proximidad de los listados a varios puntos de interés. Estos puntos de interés incluyen escuelas, hospitales, parques y/o centros de transporte. Los análisis espaciales se pueden utilizar para identificar áreas con altas concentraciones de propiedades o regiones con características

similares utilizando agrupamiento espacial. Los análisis espaciales se pueden utilizar para interpolar datos entre puntos conocidos para estimar valores en ubicaciones no medidas utilizando interpolación espacial. El análisis espacial puede ayudar a identificar tendencias, patrones y relaciones dentro de los datos basándose en la ubicación geográfica. Ejemplos incluyen cálculos de distancia, análisis de puntos calientes (hotspot) y agrupamiento.

[0083] En algunas implementaciones, el mapa se puede segmentar en una o más regiones. Las regiones pueden tener puntuaciones asociadas con las mismas. Por ejemplo, una primera región que incluye una escuela, hospital y parques a poca distancia puede tener una primera puntuación, y una segunda región que incluye una escuela y parques a poca distancia puede tener una segunda puntuación. La primera puntuación puede ser mayor que la segunda puntuación basándose en que la primera región incluye una escuela y la segunda región no tiene una escuela. Las puntuaciones asociadas con las regiones pueden basarse en una suma de puntuaciones compuestas de amenidades públicas y/o privadas. Por ejemplo, una puntuación compuesta se puede asociar con tiendas de comestibles, transporte público, hospitales, restaurantes, iglesias, gimnasios, granjas, nueva construcción, etc. En algunas implementaciones, las puntuaciones compuestas pueden ser valores negativos que restan de la puntuación asociada con una región. Por ejemplo, la tierra situada en una zona de inundación puede tener una puntuación compuesta negativa.

[0084] En algunas implementaciones, la FIG. 6 proporciona un diagrama de bloques 600 para comparar regiones. El diagrama de bloques 600 es indicativo de un mapa de región donde la información proporcionada en el diagrama de bloques 600 se puede ver en un mapa, una tabla o alguna otra interfaz gráfica. El diagrama de bloques 600 incluye un mapa 602. El mapa 602 es indicativo de un área, por ejemplo, una ciudad, un condado, un vecindario, un área definida por una superficie específica en un mapa, etc. Las cualidades asociadas con la región 1 se pueden identificar utilizando indicadores 1 a i. Por ejemplo, el indicador 1 puede ser un hospital, el indicador 2 puede ser una escuela, el indicador 3 puede ser un parque, etc. La puntuación general asociada con la región 1 se identifica como "Puntuación de Región 1". En algunas implementaciones, las puntuaciones compuestas asociadas con cada indicador se pueden incluir también. Una descripción similar se puede hacer para la región 2 y la región 3.

[0085] En algunas implementaciones, las regiones del mapa se pueden determinar basándose en una consulta de búsqueda. Por ejemplo, la consulta de búsqueda incluyendo "Ciudad de Nueva York" desencadenará que el mapa de Nueva York se segmente en regiones. En otro ejemplo, la consulta de búsqueda incluyendo una dirección específica puede desencadenar que la dirección específica se convierta en un centro de una región. Las formas de las regiones pueden

basarse en palabras clave tales como "transitable a pie" (walkable). Transitable a pie se puede definir como dos millas desde un centro, en algunos casos. En algunas implementaciones, una región corresponde a una sola propiedad y cubre un área del mapa incluyendo la propiedad única. En algunas implementaciones, las regiones del mapa se pueden determinar basándose en amenidades. Por ejemplo, las regiones se pueden segmentar basándose en hospitales en una ciudad de modo que cada región solo contenga un hospital. En algunas implementaciones, las regiones en el mapa se pueden determinar basándose en precios asociados con listados de propiedades. Por ejemplo, propiedades de precios similares tienden a estar situadas juntas, por lo tanto, un área que tiene precios de propiedad similares se puede agrupar como una sola región. Los valores atípicos se pueden ignorar al hacer tales agrupaciones.

[0086] En algunas implementaciones, los planos de planta se pueden utilizar para determinar la utilización del espacio. Por ejemplo, posiciones de escaleras, pasillos, columnas, etc., se pueden utilizar para calcular un espacio efectivo que un usuario realmente puede acceder o utilizar en una propiedad. Por ejemplo, dependiendo del ancho de un pasillo, el usuario puede no ser capaz de almacenar o colocar artículos en el pasillo. Dependiendo de la ubicación de las escaleras, áreas específicas alrededor o debajo de las escaleras son inutilizables por el usuario. Por lo tanto, aunque el listado puede tener una superficie cuadrada específica listada, el servidor 102 puede identificar características en el plano de planta para ajustar la superficie cuadrada y proporcionar una superficie cuadrada de espacio efectivo que indique espacio utilizable dentro del plano de planta de la propiedad que el usuario realmente puede usar. En algunas implementaciones, el espacio efectivo se divide en interior/exterior, y si la superficie cuadrada en el listado incluye el espacio exterior o de patio, entonces el espacio efectivo puede listar el espacio efectivo del patio por separado del espacio efectivo debajo del techo de la propiedad. Como se proporciona, el servidor 102 puede realizar cálculos para determinar el espacio utilizable y/o inutilizable con información extraída de los datos multimodales.

[0087] En algunas implementaciones, los informes creados incluyen tableros interactivos e informes que presentan los datos procesados de una manera amigable para el usuario. Los usuarios del dispositivo cliente 104 pueden filtrar, ordenar y visualizar los datos de varias maneras. En algunas implementaciones, los informes creados pueden enfocarse en una región específica, pueden enfocarse en un grupo de regiones, o pueden contrastar dos o más regiones. En algunas implementaciones, basándose en una puntuación asociada con una región, el servidor 102 puede sugerir una o más características para mejorar la región. Por ejemplo, basándose en nueva construcción dentro de una región, el servidor 102 puede sugerir qué tipos de amenidades mejorarían la puntuación asociada con la región. En algunas implementaciones, el servidor 102 puede comparar una primera región con una región situada de manera similar, una

segunda región. El servidor 102 puede sugerir amenidades en la segunda región que faltan en la primera región para mejorar la puntuación asociada con la primera región.

[0088] En algunas implementaciones, el control de versiones se puede utilizar no solo para la información almacenada en la base de datos 116, sino que el control de versiones se puede utilizar para rastrear cambios en la base de código y uno o más de los modelos almacenados como los parámetros de modelo de lenguaje 110 y los parámetros de modelo de imagen 114. El control de versiones también se puede utilizar para rastrear cambios en las características de clasificación 112. Rastrear cambios en los modelos asegura la reproducibilidad y la responsabilidad en el proceso de desarrollo para permitir la reversión de cambios si un modelo se sobreentrena o exhibe nuevos errores debido a un error evolutivo. En algunas implementaciones, el control de versiones se puede utilizar para monitorear modelos del sistema 100. Monitorear continuamente los modelos de aprendizaje automático asegura su rendimiento y precisión. Los modelos se pueden monitorear rastreando métricas clave y reentrenando regularmente los modelos con nuevos datos. Estas métricas pueden cambiar de acuerdo con la tarea de interés. En algunas implementaciones, las métricas pueden incluir exactitud, precisión, recuperación, puntuación F1, Área Bajo la Curva (AUC), Error Absoluto Medio (MAE) o Error Cuadrático Medio (MSE), curva de Característica Operativa del Receptor (ROC), matriz de confusión, pérdida de validación, o cualquier combinación de los mismos.

[0089] Las modalidades de la presente divulgación son compatibles con prácticas de desarrollo de integración continua/despliegue continuo (CI/CD). Las prácticas de CI/CD se pueden utilizar para automatizar la prueba y el despliegue de la base de código y los modelos para asegurar que el servidor 102 siempre esté ejecutando las versiones más recientes y eficientes. De manera similar, se admiten pruebas A/B, permitiendo realizar regularmente pruebas A/B para comparar diferentes versiones de modelos para identificar las soluciones más efectivas para extraer información de las diferentes fuentes de datos.

[0090] Además, se pueden emplear algoritmos de detección de anomalías para identificar y marcar cualquier comportamiento inusual o datos dentro del sistema 100. En algunas implementaciones, los algoritmos de detección de anomalías pueden monitorear las métricas de rendimiento de los modelos desplegados en tiempo real. Caídas repentinas en la precisión, patrones de predicción inusuales o variaciones inesperadas en el comportamiento del modelo pueden marcarse como anomalías para una investigación adicional. En algunas implementaciones, los algoritmos de detección de anomalías pueden evaluar si hay cambios significativos en la distribución o características de los datos de entrada en comparación con los datos con los que se entrenó el modelo. Desviaciones de los patrones de datos esperados

pueden indicar deriva de datos, desencadenando alertas o la necesidad de reentrenamiento o recalibración del modelo. Esto puede ayudar a mantener la integridad y seguridad de la información almacenada en el repositorio 106. La política de gobernanza de datos se puede hacer cumplir en el servidor 102 para asegurar que los datos extraídos sean precisos, confiables y se utilicen apropiadamente. Esto incluye auditorías regulares y el uso de herramientas de gestión de calidad de datos.

[0091] En algunas implementaciones, el servidor evalúa y limpia los datos extraídos utilizando bibliotecas de validación de datos y métodos tradicionales de limpieza de datos tales como eliminar valores atípicos y manejar valores faltantes a través de imputación. En algunas implementaciones, para integrar con fuentes de datos externas, se utilizan técnicas tradicionales de ETL (Extraer, Transformar, Cargar).

[0092] En algunas implementaciones, para manejar la escalabilidad y la optimización del rendimiento, el servidor 102 utiliza procesamiento por lotes en lugar de procesamiento en tiempo real, o confía en sistemas de gestión de bases de datos tradicionales en lugar de marcos de computación distribuida como Hadoop o Spark.

[0093] En lugar de utilizar IA para reconocimiento de imágenes, se podrían utilizar técnicas de procesamiento de imágenes más simples como detección de bordes, comparaciones de histogramas de color o coincidencia de plantillas para extraer información de imágenes. Sin embargo, estos métodos podrían no ser tan efectivos en reconocer patrones y objetos complejos.

[0094] En algunas implementaciones, el servidor 102 puede resolver conflictos o inconsistencias en datos de diferentes fuentes empleando técnicas estadísticas tradicionales como pruebas de hipótesis o pruebas de chi-cuadrado para identificar diferencias significativas entre características extraídas de diferentes fuentes. En algunas implementaciones, las fuentes de datos se priorizan. Por ejemplo, la información de un sitio de listado de propiedades se prioriza sobre la información de una publicación de blog.

[0095] En algunas implementaciones, el servidor 102 corrige errores simples como errores tipográficos, faltas de ortografía y errores en los datos del listado utilizando una función de verificación ortográfica o verificación gramatical.

[0096] En algunas implementaciones, en lugar de identificar tipos y subtipos de propiedades a través de análisis de texto, el servidor 102 puede confiar en la categorización proporcionada por la fuente del listado. En algunas implementaciones, el servidor 102 confía en otros métodos para

verificar cruzadamente la información del tipo de propiedad a través de diversas fuentes. El servidor 102 puede recuperar información del tipo de propiedad del texto de descripción cuando esté disponible, mejorando la precisión de la clasificación de propiedades. La información del tipo de propiedad incluye propiedad comercial, propiedad residencial, etc.

[0097] Las modalidades de la presente divulgación proporcionan un sistema para construir una base de conocimientos (ver p. ej., FIGS. 1 y 2 y descripción acompañante). El sistema está configurado para recibir datos multimodales de una o más fuentes. El sistema puede analizar los datos multimodales para determinar características. Las características, en este caso, pueden ser valores asociados con claves o categorías predefinidas (p. ej., "dormitorio"), o pueden ser valores asociados con claves recientemente identificadas (p. ej., "techo alto"). Las características determinadas pueden almacenarse en una base de datos, por ejemplo, en un par clave-valor. En algunas implementaciones, las características determinadas se indexan para búsqueda.

[0098] El sistema puede recibir una consulta de búsqueda para buscar características almacenadas. La consulta de búsqueda puede ser una búsqueda textual en formato de oración, formato de párrafo o fragmento de oración. Por ejemplo, la consulta de búsqueda puede incluir una dirección, un código postal, un rango de precios, un número de dormitorios, etc. La consulta de búsqueda puede analizarse utilizando un gran modelo de lenguaje para extraer características de búsqueda. Por ejemplo, si la consulta de búsqueda está en formato de oración, se pueden proporcionar una serie de preguntas preaprobadas junto con la consulta de búsqueda a un gran modelo de lenguaje. Las respuestas a la serie de preguntas preaprobadas son las características de búsqueda. En algunos casos, si las respuestas a las preguntas devuelven una respuesta inválida, entonces se considera que la información no está incluida en la consulta de búsqueda.

[0099] Por ejemplo, la consulta de búsqueda proporcionada puede ser "encontrar un apartamento en la Ciudad de Nueva York en el rango de precios de \$700k a \$1.5 millones", y las preguntas preaprobadas incluyen: (a) "¿cuántos dormitorios?" (b) "¿cuál es el rango de precios?" (c) "¿qué código postal?" (d) "¿qué ciudad?" (e) "¿qué tipo de unidad?" (f) "¿qué amenidades?". El gran modelo de lenguaje puede no proporcionar resultados significativos para las preguntas (a) y (f) para la consulta de búsqueda de ejemplo. Por lo tanto, las características de búsqueda extraídas de la consulta de búsqueda pueden incluir "Ciudad de Nueva York", uno o más códigos postales asociados con la Ciudad de Nueva York, "\$700k a 1.5 millones" y "apartamento". Estas características extraídas se pueden utilizar para establecer límites y/o preferencias para buscar en la base de datos propiedades específicas. Los criterios de búsqueda como número de dormitorios y amenidades no tendrían preferencias.

[00100] Después de que se generan los resultados de búsqueda, los resultados se pueden mostrar en una interfaz gráfica de usuario (GUI) estandarizada. La GUI estandarizada incluye una leyenda que tiene al menos una o más de las características de búsqueda mostradas. En un ejemplo, se proporciona una primera GUI en la FIG. 3 que muestra un listado de propiedad de ejemplo. En la primera GUI, se proporciona una leyenda 312. La leyenda 312 incluye ejemplos "multifamiliar", "activa", "en venta", "no en zona de inundación". La leyenda 312 puede tener diferentes colores o indicadores para las diferentes características de búsqueda. La primera GUI proporciona además cualidades adicionales 310 que incluyen ejemplos de precio por pie cuadrado, área construida de la propiedad, número de dormitorios, número de baños, número de espacios de estacionamiento, etc. En algunas implementaciones, la primera GUI proporciona además una sección sobre amenidades que se pueden obtener de la extracción de imagen y/o texto como se discute en el presente documento. Por ejemplo, las amenidades 306 incluyen jardín, vista al mar, complejo cerrado y paisaje. En algunas implementaciones, la primera GUI proporciona una descripción adicional de características asociadas con habitaciones. Las habitaciones 308 incluyen una descripción de que hay una habitación estilo atrio.

[00101] En algunas implementaciones, los resultados de búsqueda se pueden proporcionar en una vista de mapa en una segunda GUI donde un texto de búsqueda 402 de "apartamento con vista al mar" devuelve un mapa 404 de un área junto con listados. Los listados se superponen en el mapa 404 y se muestran en una vista de galería. La vista de galería tiene una imagen de propiedad 406, una fuente 408 asociada con una propiedad específica, una leyenda 412 superpuesta en las imágenes de propiedad. En algunas implementaciones, la información se obtiene de más de una fuente, y se utilizan múltiples fuentes 410 como tal indicación. El texto de búsqueda 402 "apartamento con vista al mar" es analizado por el servidor 102 para obtener la característica "multifamiliar" y mostrando propiedades que incluyen la característica "vista al mar". Las características que no están restringidas son, por ejemplo, "en alquiler" o "en venta", por lo tanto, se proporcionan resultados de búsqueda que incluyen propiedades en venta y propiedades en alquiler. Además, se proporcionan propiedades en una zona de inundación vs. no en una zona de inundación.

[00102] Refiriéndose a la FIG. 5, se puede proporcionar un listado más detallado en una tercera GUI. El texto de búsqueda 502 "apartamento con vista al mar" se utiliza para establecer filtros 504 que muestran "multifamiliar" y "vista al mar". Las leyendas 512 incluyen características asociadas con las propiedades en los resultados de búsqueda, permitiendo identificar rápidamente cualidades asociadas con cada una de las propiedades listadas. "Precio de venta"; "precio por área"; "área construida"; "tamaño del lote"; "dormitorios"; "baños"; "amenidades" indicando número de amenidades; "estado del listado" indicando activa, contingente, inactiva, etc.; "días en

el mercado"; fuente 506 identificando fuentes únicas y múltiples; y "enlace" para ver el listado se proporcionan. En algunos ejemplos, "precio por área" es una categoría creada basada en "área construida".

[00103] De acuerdo con algunas modalidades de la presente divulgación, los procesos descritos anteriormente con referencia a diagramas de flujo o esquemas de flujo (p. ej., en la FIG. 2) pueden implementarse en un programa de software informático. Por ejemplo, algunas modalidades de la presente divulgación incluyen un producto de programa informático, que incluye un programa informático que se lleva en un medio legible por computadora. El programa informático incluye códigos de programa para ejecutar el proceso 200. El programa informático puede descargarse e instalarse desde una red (p. ej., Internet, una red local, etc.) y/o puede instalarse desde un medio extraíble (p. ej., un disco duro extraíble, una unidad flash, una unidad externa, etc.). El programa informático, cuando es ejecutado por una unidad central de procesamiento implementa las funciones anteriores definidas por métodos y diagramas de flujo proporcionados en el presente documento en la presente divulgación.

[00104] Un medio legible por computadora de acuerdo con la presente divulgación puede ser un medio de señal legible por computadora o un medio de almacenamiento legible por computadora o cualquier combinación de los dos anteriores. Ejemplos del medio de almacenamiento legible por computadora pueden incluir sistemas, elementos, aparatos eléctricos, magnéticos, ópticos, electromagnéticos, infrarrojos o semiconductores, o una combinación de cualquiera de los anteriores. Ejemplos más específicos del medio de almacenamiento legible por computadora incluyen un disco informático portátil, un disco duro, una memoria de acceso aleatorio (RAM), una memoria de solo lectura (ROM), una memoria de solo lectura programable borrable (EPROM o memoria flash), una fibra óptica, una memoria de solo lectura de disco compacto portátil (CD-ROM), una memoria óptica, una memoria magnética o cualquier combinación adecuada de los anteriores.

[00105] El medio de almacenamiento legible por computadora de acuerdo con algunas modalidades puede ser cualquier medio tangible que contenga o almacene programas, que pueden ser utilizados por, o utilizados en combinación con, un sistema, aparato o elemento de ejecución de comandos. En algunas modalidades de la presente divulgación, el medio de señal legible por computadora puede incluir una señal de datos en la banda base o propagándose como parte de una onda portadora, en la cual se llevan códigos de programa legibles por computadora. La señal de datos que se propaga puede tomar varias formas, incluyendo, pero no limitado a una señal electromagnética, una señal óptica o cualquier combinación adecuada de las anteriores. El medio de señal legible por computadora también puede ser cualquier medio legible por

computadora excepto el medio de almacenamiento legible por computadora. El medio legible por computadora es capaz de transmitir, propagar o transferir programas para uso por, o utilizados en combinación con, un sistema, aparato o elemento de ejecución de comandos. Los códigos de programa contenidos en el medio legible por computadora pueden transmitirse con cualquier medio adecuado, incluyendo, pero no limitado a: inalámbrico, cableado, cable óptico, medio de RF, etc., o cualquier combinación adecuada de los anteriores.

[00106] Un código de programa informático para ejecutar operaciones en la presente divulgación puede compilarse utilizando uno o más lenguajes de programación o combinaciones de los mismos. Los lenguajes de programación incluyen lenguajes de programación orientados a objetos, tales como Java o C++, y también incluyen lenguajes de programación procedimentales convencionales, tales como lenguaje "C" o lenguajes de programación similares. El código de programa puede ejecutarse completamente en la computadora de un usuario, ejecutarse parcialmente en la computadora de un usuario, ejecutarse como un paquete de software separado, ejecutarse parcialmente en la computadora de un usuario y ejecutarse parcialmente en una computadora remota, o ejecutarse completamente en una computadora remota o dispositivo electrónico. En la circunstancia que involucra una computadora remota, la computadora remota puede estar conectada a la computadora de un usuario a través de cualquier red, incluyendo red de área local (LAN) o red de área amplia (WAN), o estar conectada a una computadora externa (por ejemplo, conectada a través de Internet utilizando un proveedor de servicios de Internet).

[00107] Los diagramas de flujo y diagramas de bloques en los dibujos adjuntos ilustran arquitecturas, funciones y operaciones que pueden implementarse de acuerdo con los sistemas, métodos y productos de programa informático de las diversas modalidades de la presente divulgación. Cada uno de los bloques en los diagramas de flujo o diagramas de bloques puede representar un segmento de programa o código que incluye una o más instrucciones ejecutables para implementar funciones lógicas especificadas. Debe observarse además que, en algunas implementaciones alternativas, las funciones denotadas por los diagramas de flujo y diagramas de bloques también pueden ocurrir en una secuencia diferente de las secuencias mostradas en las figuras. Por ejemplo, dos bloques cualesquiera presentados en sucesión pueden ejecutarse sustancialmente en paralelo, o a veces ejecutarse en una secuencia inversa, dependiendo de las funciones involucradas. Debe observarse además que cada bloque en los diagramas de bloques y/o diagramas de flujo, así como una combinación de bloques en los diagramas de bloques y/o diagramas de flujo, pueden implementarse utilizando un sistema basado en hardware dedicado que ejecuta funciones u operaciones especificadas, o mediante una combinación de hardware dedicado e instrucciones informáticas.

[00108] Los motores, manejadores, generadores, gestores o cualquier otro bloque de software o bloque híbrido de hardware-software identificado en algunas modalidades de la presente divulgación pueden implementarse mediante software, o pueden implementarse mediante hardware. Los bloques descritos también pueden proporcionarse en un procesador.

IMPLEMENTACIONES ALTERNATIVAS

[00109] Implementación alternativa 1. Un sistema para construir una base de conocimientos, comprendiendo el sistema un procesador y un medio legible por computadora no transitorio que almacena instrucciones tales que cuando las instrucciones son ejecutadas por el procesador, el sistema está configurado para: (a) recibir datos multimodales de una o más fuentes; (b) analizar los datos para determinar características; (c) almacenar las características en una base de datos; (d) recibir una consulta de búsqueda para buscar las características almacenadas; (e) analizar la consulta de búsqueda utilizando un gran modelo de lenguaje para extraer características de búsqueda; (f) generar resultados de búsqueda a partir de las características de búsqueda; y (g) mostrar los resultados de búsqueda en una interfaz gráfica de usuario estandarizada que incluye una leyenda que tiene al menos una o más de las características de búsqueda mostradas.

[00110] Implementación alternativa 2. El sistema de la implementación alternativa 1, en donde los datos multimodales incluyen video, texto no estructurado, texto estructurado y/o imágenes.

[00111] Implementación alternativa 3. El sistema de la implementación alternativa 1 o 2, en donde los datos multimodales incluyen descripción de listados, documentos de compañías de seguros, documentos relacionados con hipotecas, documentos relacionados con la banca, documentos relacionados con préstamos estudiantiles, documentos relacionados con el ámbito académico, o cualquier combinación de los mismos.

[00112] Implementación alternativa 4. El sistema de cualquiera de las implementaciones alternativas 1 a 3, en donde el sistema está configurado para analizar los datos para determinar características mediante: (a) recibir una clasificación de preguntas para extraer una primera característica, incluyendo la clasificación de preguntas al menos dos preguntas asociadas con la extracción de la primera característica; y (b) extraer la primera característica utilizando el gran modelo de lenguaje, en donde las entradas al gran modelo de lenguaje incluyen (i) una de las al menos dos preguntas y (ii) los datos multimodales.

[00113] Implementación alternativa 5. El sistema de la implementación alternativa 4, en donde las al menos dos preguntas incluyen una clave y la primera característica extraída incluye un valor

asociado con la clave.

[00114] Implementación alternativa 6. El sistema de la implementación alternativa 4 o 5, en donde la clasificación de preguntas se basa en información histórica asociada con una métrica de rendimiento del gran modelo de lenguaje, incluyendo la información histórica un voto positivo recibido de un dispositivo cliente.

[00115] Implementación alternativa 7. El sistema de cualquiera de las implementaciones alternativas 1 a 6, en donde el sistema está configurado para analizar los datos para determinar características mediante: (a) determinar una palabra clave a partir de los datos multimodales utilizando el gran modelo de lenguaje; (b) generar una pregunta a partir de la palabra clave; y (c) extraer una primera característica utilizando el gran modelo de lenguaje, en donde las entradas al gran modelo de lenguaje incluyen (i) la pregunta generada y (ii) los datos multimodales.

[00116] Implementación alternativa 8. El sistema de la implementación alternativa 7, en donde el sistema está configurado para generar un nivel de confianza asociado con la primera característica extraída.

[00117] Implementación alternativa 9. El sistema de la implementación alternativa 8, en donde el nivel de confianza se basa en una magnitud asociada con la primera característica extraída.

[00118] Implementación alternativa 10. El sistema de cualquiera de las implementaciones alternativas 1 a 9, en donde la característica se almacena como un par clave-valor, identificando la clave una categoría asociada con la característica e indicando el valor una propiedad asociada con la categoría.

[00119] Implementación alternativa 11. El sistema de la implementación alternativa 10, en donde la clave es un número de dormitorios y el valor es un entero.

[00120] Implementación alternativa 12. El sistema de cualquiera de las implementaciones alternativas 1 a 11, en donde la consulta de búsqueda es una oración, un fragmento de oración o un párrafo.

[00121] Implementación alternativa 13. Un sistema para construir una base de conocimientos, comprendiendo el sistema un procesador y un medio legible por computadora no transitorio que almacena instrucciones tales que cuando las instrucciones son ejecutadas por el procesador, el sistema está configurado para: (a) recibir datos multimodales de una o más fuentes, incluyendo

los datos multimodales datos de texto; (b) analizar los datos para determinar características; (c) almacenar las características en una base de datos; (d) recibir una consulta de búsqueda para buscar las características almacenadas; (e) analizar la consulta de búsqueda utilizando un gran modelo de lenguaje para extraer características de búsqueda; (f) generar resultados de búsqueda a partir de las características de búsqueda; y (g) mostrar los resultados de búsqueda en una interfaz gráfica de usuario estandarizada que incluye una leyenda que tiene al menos una o más de las características de búsqueda mostradas.

[00122] Implementación alternativa 14. El sistema de la implementación alternativa 13, en donde los datos multimodales incluyen video, texto no estructurado, texto estructurado y/o imágenes.

[00123] Implementación alternativa 15. El sistema de la implementación alternativa 13, en donde los datos multimodales incluyen imágenes y texto y en donde el sistema configurado para analizar los datos para determinar características incluye: (a) determinar características de imagen a partir de imágenes asociadas con un primer elemento; (b) determinar características de texto a partir de texto asociado con el primer elemento; (c) comparar las características de texto con las características de imagen para determinar si hay categorías en las características de imagen faltantes en las características de texto; y (d) en respuesta a determinar que hay categorías en las características de imagen faltantes en las características de texto, almacenar las características de imagen faltantes en la base de datos.

[00124] Implementación alternativa 16. El sistema de la implementación alternativa 15, en donde el sistema configurado para analizar los datos para determinar características incluye además: en respuesta a determinar que hay categorías en las características de imagen presentes en las características de texto, almacenar las características de texto en la base de datos.

[00125] Implementación alternativa 17. El sistema de la implementación alternativa 15 o 16, en donde una propiedad asociada con las características de imagen faltantes es un valor Booleano.

[00126] Implementación alternativa 18. El sistema de cualquiera de las implementaciones alternativas 15 a 17, en donde un nivel de confianza está asociado con las características de imagen faltantes.

[00127] Implementación alternativa 19. El sistema de la implementación alternativa 13 a 18, en donde el sistema está configurado además para: entrenar un modelo de imagen para extraer características de imagen.

[00128] Implementación alternativa 20. El sistema de la implementación alternativa 13 a 19, en donde el sistema está configurado además para: (a) recibir una imagen actualizada asociada con un primer elemento; (b) determinar características de imagen asociadas con el primer elemento; (c) comparar características almacenadas en la base de datos contra las características de imagen para determinar discrepancias en las características almacenadas asociadas con el primer elemento y las características de imagen; y (d) en respuesta a determinar que hay al menos una discrepancia, resolver la al menos una discrepancia.

[00129] Implementación alternativa 21. El sistema de la implementación alternativa 20, en donde el sistema está configurado para resolver la al menos una discrepancia solicitando al dispositivo cliente.

[00130] Implementación alternativa 22. El sistema de la implementación alternativa 20 o 21, en donde el sistema está configurado para resolver la al menos una discrepancia almacenando las características de imagen asociadas con el primer elemento como parte de las características almacenadas asociadas con el primer elemento.

[00131] Implementación alternativa 23. El sistema de cualquiera de las implementaciones alternativas 20 a 22, en donde el sistema está configurado para resolver la al menos una discrepancia creando una nueva categoría asociada con el primer elemento y vinculando una propiedad asociada con las características de imagen asociadas con el primer elemento a la nueva categoría.

[00132] Implementación alternativa 24. El sistema de cualquiera de las implementaciones alternativas 20 a 23, en donde el sistema está configurado para resolver la al menos una discrepancia reduciendo un nivel de confianza de una categoría asociada con las características almacenadas.

[00133] Implementación alternativa 25. El sistema de cualquiera de las implementaciones alternativas 20 a 23, en donde el sistema está configurado para resolver la al menos una discrepancia aumentando un nivel de confianza de una categoría asociada con las características almacenadas.

[00134] Implementación alternativa 26. Un sistema para construir una base de conocimientos, comprendiendo el sistema un procesador y un medio legible por computadora no transitorio que almacena instrucciones tales que cuando las instrucciones son ejecutadas por el procesador, el sistema está configurado para: (a) recibir datos multimodales de una o más fuentes; (b) analizar

los datos para determinar características; (c) almacenar las características en una base de datos; (d) recibir una consulta de búsqueda para buscar las características almacenadas; (e) analizar la consulta de búsqueda utilizando un gran modelo de lenguaje para extraer características de búsqueda; (f) generar resultados de búsqueda a partir de las características de búsqueda; y (g) mostrar los resultados de búsqueda en una interfaz gráfica de usuario estandarizada que incluye una leyenda que tiene al menos una o más de las características de búsqueda mostradas.

[00135] Implementación alternativa 27. El sistema de la implementación alternativa 26, en donde el sistema está configurado para crear una nueva categoría basada en la consulta de búsqueda.

[00136] Implementación alternativa 28. El sistema de la implementación alternativa 27, en donde la nueva categoría se basa en palabras clave incluidas en los datos multimodales.

[00137] Implementación alternativa 29. El sistema de la implementación alternativa 28, en donde la nueva categoría se basa en partes del discurso de las palabras clave incluidas en los datos multimodales.

[00138] Implementación alternativa 30. El sistema de cualquiera de las implementaciones alternativas 27 a 29, en donde la nueva categoría se basa en un cálculo realizado sobre la información contenida en los datos multimodales.

[00139] Implementación alternativa 31. El sistema de cualquiera de las implementaciones alternativas 27 a 30, en donde la nueva categoría se basa en una frecuencia de palabras clave incluidas en consultas de búsqueda de una pluralidad de usuarios.

[00140] Implementación alternativa 32. El sistema de cualquiera de las implementaciones alternativas 27 a 31, en donde la nueva categoría se basa en un emparejamiento de expresiones regulares de descripciones de texto en los datos multimodales.

[00141] Implementación alternativa 33. El sistema de cualquiera de las implementaciones alternativas 27 a 32, en donde la nueva categoría se basa en una comparación de diferentes documentos incluidos en los datos multimodales.

[00142] Implementación alternativa 34. El sistema de cualquiera de las implementaciones alternativas 27 a 33, en donde la nueva categoría se basa en analizar imágenes en los datos multimodales.

[00143] Implementación alternativa 35. El sistema de la implementación alternativa 34, en donde la nueva categoría está asociada con espacio utilizable y en donde las imágenes están asociadas con un plano de planta de una unidad.

[00144] Implementación alternativa 36. El sistema de la implementación alternativa 35, en donde el espacio utilizable es un ajuste de un espacio total basado en un conjunto de características.

[00145] Implementación alternativa 37. El sistema de la implementación alternativa 36, en donde el conjunto de características incluye espacio exterior y/o escaleras.

[00146] Implementación alternativa 38. El sistema de cualquiera de las implementaciones alternativas 26 a 37, en donde la leyenda incluye una o más características diferentes de las características de búsqueda.

[00147] Implementación alternativa 39. Un sistema para construir una base de conocimientos, comprendiendo el sistema un procesador y un medio legible por computadora no transitorio que almacena instrucciones tales que cuando las instrucciones son ejecutadas por el procesador, el sistema está configurado para: (a) recibir datos multimodales de una o más fuentes, incluyendo los datos multimodales datos de texto; (b) analizar los datos para determinar características; (c) asociar las características determinadas con identificadores de fuente basados en la una o más fuentes; (d) almacenar las características y los identificadores de fuente correspondientes en una base de datos; (e) recibir una consulta de búsqueda para buscar las características almacenadas; (f) analizar la consulta de búsqueda utilizando un gran modelo de lenguaje para extraer características de búsqueda; (g) generar resultados de búsqueda a partir de las características de búsqueda utilizando al menos las características almacenadas y los identificadores de fuente correspondientes; y (h) mostrar los resultados de búsqueda en una interfaz gráfica de usuario estandarizada que incluye una leyenda que tiene al menos una o más de las características de búsqueda mostradas.

[00148] Implementación alternativa 40. El sistema de la implementación alternativa 39, en donde los identificadores de fuente se seleccionan del grupo que consiste en: corredores, sitios web gubernamentales, usuarios autenticados, usuarios anónimos, documentos de arrendamiento, registros gubernamentales, fuentes comerciales, registros de transacciones y permisos.

[00149] Implementación alternativa 41. El sistema de la implementación alternativa 39, en donde los identificadores de fuente incluyen corredores y usuarios anónimos, y en donde los resultados de búsqueda generados clasifican las características correspondientes a corredores más alto que

los resultados de búsqueda correspondientes a usuarios anónimos de tal manera que los resultados de búsqueda generados devuelven las características correspondientes a corredores para ser mostradas en la interfaz gráfica de usuario estandarizada.

[00150] Implementación alternativa 42. El sistema de cualquiera de las implementaciones alternativas 39 a 41, en donde los identificadores de fuente se utilizan para reducir una cantidad de información almacenada en la base de datos.

[00151] Implementación alternativa 43. El sistema de la implementación alternativa 42, en donde la primera información de una primera fuente es información duplicada de la segunda información de una segunda fuente, y en donde la segunda información asociada con la segunda fuente se elimina de la base de datos basándose en que la segunda fuente se determina como menos confiable que la primera fuente basándose en los identificadores de fuente de la primera fuente y la segunda fuente.

[00152] Implementación alternativa 44. El sistema de la implementación alternativa 42, en donde la primera información de una primera fuente y la segunda información de una segunda fuente incluyen información conflictiva sobre una característica de la propiedad, en donde los identificadores de fuente de la primera fuente y la segunda fuente indican que la segunda fuente es menos confiable que la primera fuente, y en donde la segunda información de la segunda fuente se elimina de la base de datos.

[00153] Implementación alternativa 45. El sistema de la implementación alternativa 42, en donde la primera información de una primera fuente y la segunda información de una segunda fuente incluyen información asociada con una misma característica de propiedad, en donde los identificadores de fuente de la primera fuente y la segunda fuente indican que la segunda fuente es menos confiable que la primera fuente, y en donde la segunda información de la segunda fuente se comprime y desindexa.

[00154] Implementación alternativa 46. El sistema de la reivindicación 39 a 45, en donde el sistema está configurado además para: (a) recibir texto correctivo de un dispositivo cliente para aumentar al menos una característica; y (b) generar una recompensa asociada con el texto correctivo.

[00155] Implementación alternativa 47. El sistema de la implementación alternativa 46, en donde la recompensa generada es un primer valor diferente de un segundo valor, estando asociado el segundo valor con agregar una nueva característica, y siendo el primer valor mayor que el

segundo valor cuando el texto correctivo está asociado con cambiar una característica mostrada de una primera entrada a una segunda entrada.

[00156] Implementación alternativa 48. El sistema de la implementación alternativa 46, en donde la recompensa generada es un primer valor diferente de un segundo valor, siendo el primer valor mayor que el segundo valor cuando el texto correctivo está asociado con una primera característica que tiene un valor más alto en comparación con una segunda característica asociada con el segundo valor.

[00157] Implementación alternativa 49. El sistema de la implementación alternativa 48, en donde la primera característica es información de contacto y la segunda característica es un número de dormitorios.

[00158] Implementación alternativa 50. El sistema de cualquiera de las implementaciones alternativas 46 a 49, en donde la recompensa generada asociada con el texto correctivo desencadena un control de versiones mediante el cual un valor anterior que está siendo sobrescrito por el texto correctivo se asocia con una versión anterior y el texto correctivo se asocia con una versión actual.

[00159] Implementación alternativa 51. Un sistema para construir una base de conocimientos, comprendiendo el sistema un procesador y un medio legible por computadora no transitorio que almacena instrucciones tales que cuando las instrucciones son ejecutadas por el procesador, el sistema está configurado para: (a) recibir datos multimodales de una o más fuentes, incluyendo los datos multimodales datos de texto; (b) analizar los datos para determinar características; (c) almacenar las características en una base de datos; (d) recibir una consulta de búsqueda para buscar las características almacenadas; (e) analizar la consulta de búsqueda utilizando un gran modelo de lenguaje para extraer características de búsqueda; (f) generar resultados de búsqueda a partir de las características de búsqueda; y (g) mostrar resultados de búsqueda en una interfaz gráfica de usuario estandarizada que incluye (i) un mapa y (ii) una leyenda que tiene al menos una o más de las características de búsqueda mostradas.

[00160] Implementación alternativa 52. El sistema de la implementación alternativa 51, en donde la interfaz gráfica de usuario incluye además indicaciones en el mapa que muestran uno o más puntos de interés.

[00161] Implementación alternativa 53. El sistema de la implementación alternativa 51 o 52, en donde la interfaz gráfica de usuario incluye además indicaciones en el mapa que muestran

puntuaciones asociadas con regiones en el mapa.

[00162] Implementación alternativa 54. El sistema de la implementación alternativa 53, en donde las regiones del mapa se determinan basándose en la consulta de búsqueda.

[00163] Implementación alternativa 55. El sistema de la implementación alternativa 54, en donde la consulta de búsqueda incluye palabras clave tales como transitable a pie.

[00164] Implementación alternativa 56. El sistema de la implementación alternativa 53, en donde las regiones del mapa se determinan basándose en amenidades proporcionadas en las características almacenadas.

[00165] Implementación alternativa 57. El sistema de la implementación alternativa 53, en donde las regiones del mapa se determinan basándose en precios proporcionados en las características almacenadas.

[00166] Implementación alternativa 58. El sistema de la implementación alternativa 53, en donde las puntuaciones asociadas con las regiones del mapa se determinan basándose en características incluidas en regiones correspondientes.

[00167] Implementación alternativa 59. El sistema de cualquiera de las implementaciones alternativas 53 a 58, en donde el sistema está configurado además para generar un informe comparando cada una de las regiones en el mapa.

[00168] Implementación alternativa 60. El sistema de la implementación alternativa 59, en donde cada una de las regiones cubre un área del mapa incluyendo una sola propiedad.

[00169] Implementación alternativa 61. El sistema de la implementación alternativa 59, en donde el sistema está configurado además para generar una o más características para mejorar al menos una de las regiones en el mapa.

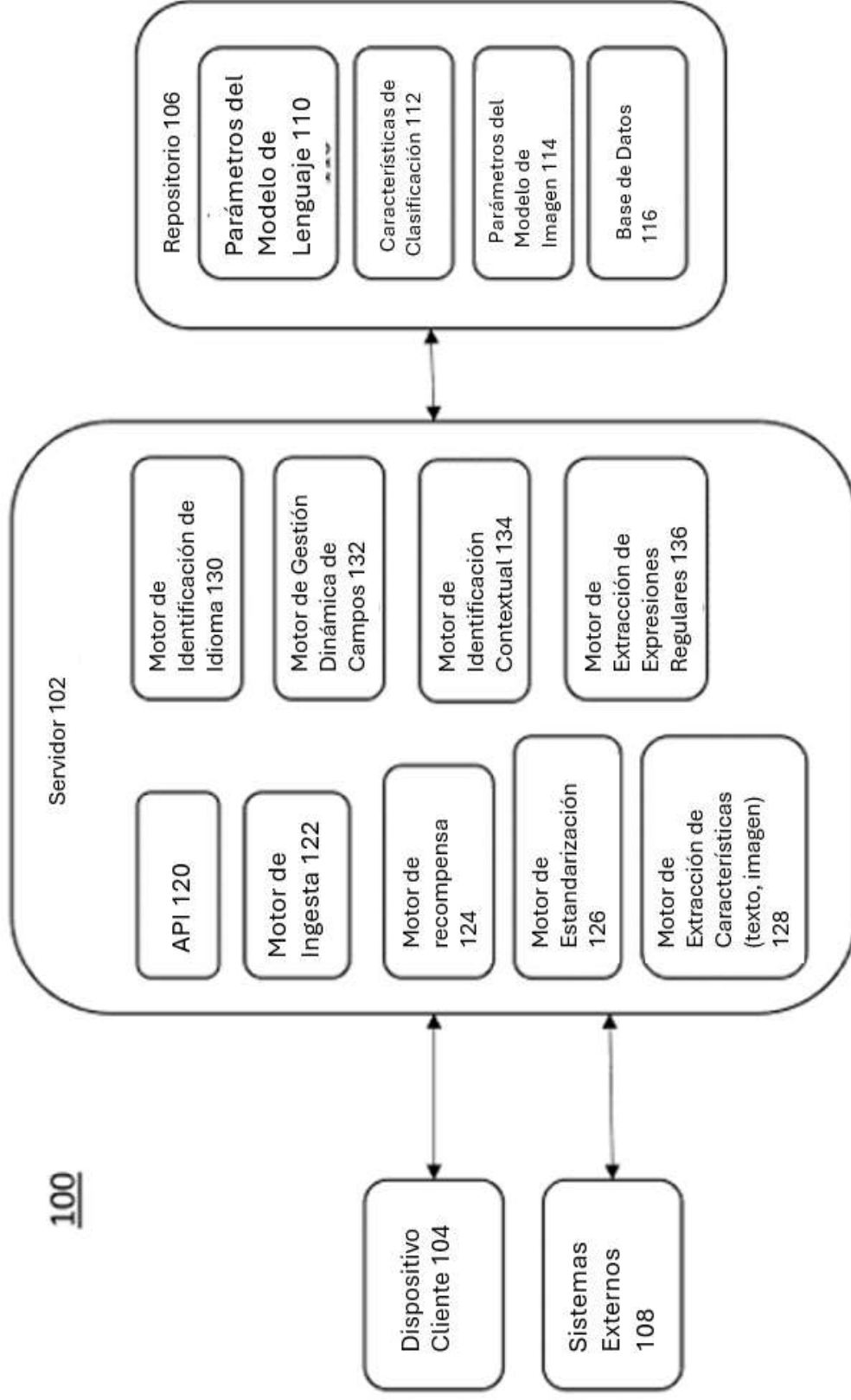


Fig. 1

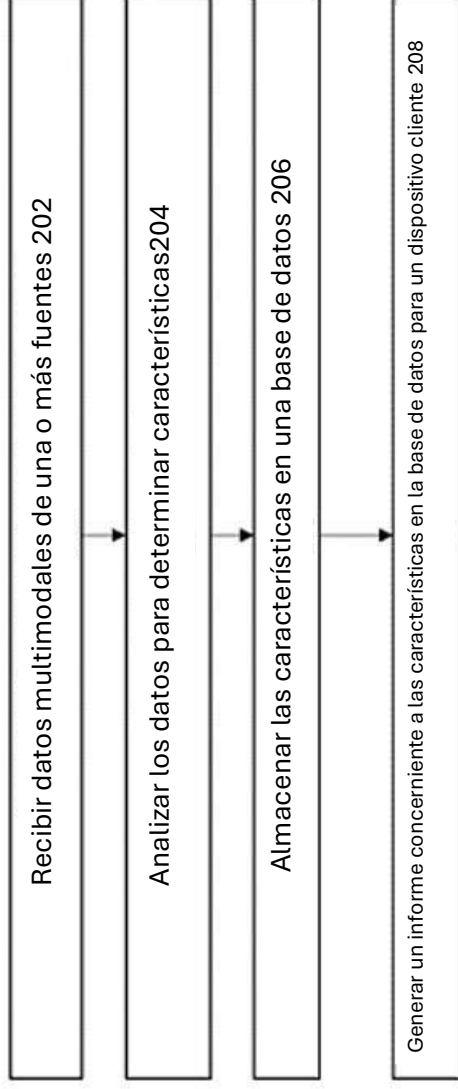


Fig. 2

\$ 235,000
 Multi-Family Active For Sale **Not for Flood zone**
 \$195.83 per Sqft | Built Area: 1,200 Sqft | 3 Bedrooms | 2 Bathrooms | 2 Parking Spaces

312 **310** **302** **304**

VIEW LISTING

Days on Market: 3
View: Ocean View
Date Added: 06/24/2024
Condition: C-2 (Fair, deferred maintenance)
Year Built: 1995

Condominio Costa Brava, Machos, Ceiba, 00735
 Garden apartment with fenced in back yard and ocean views. 3 bedrooms and 2 bathrooms. Gated community with pool, gym, tennis and basketball courts. Just minutes away from Puerto del Rey Marina, Macho Macho beach and shops and Ceiba Island Lanes.

AMENITIES
 • Garden • Ocean view
 • Enclosed complex • Landscape

ROOMS
 • Allium

306

CONTACT
 Maria Arria
 📞 787.690.3970

308

LOCATION

MUNICIPALITY	NEIGHBORHOOD	TOWN NEIGHBORHOOD	ZIP CODE	BUSINESS NAME	GPS
Ceiba	Machos	—	00735	Condominio Costa Brava	—

16.27702575N - 65.5458546W

Generate Report (Beta)

No Unlited Property

Fig. 3

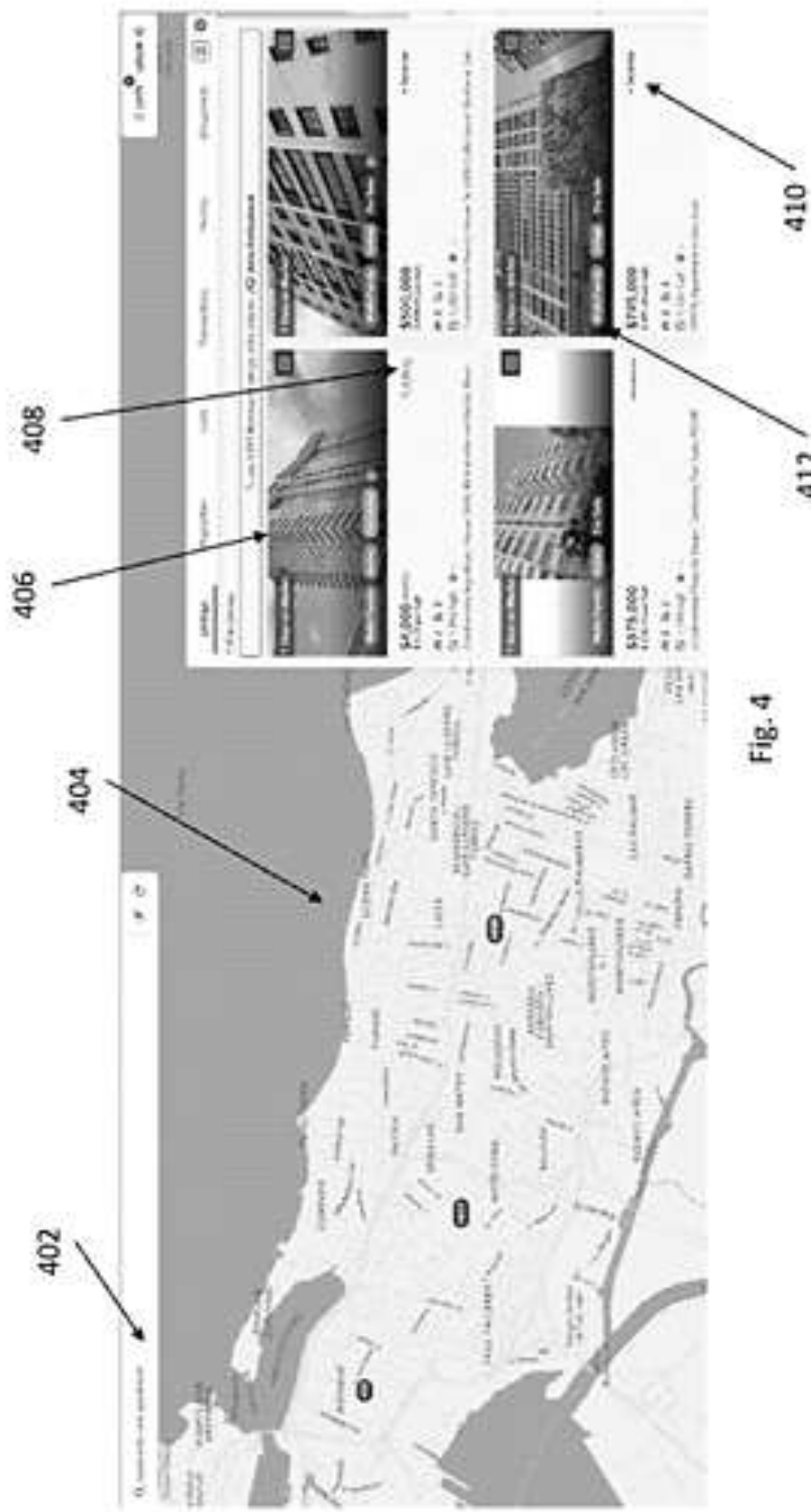


Fig. 4

502

506

504

502	504	506
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R
R	R	R

Fig. 5

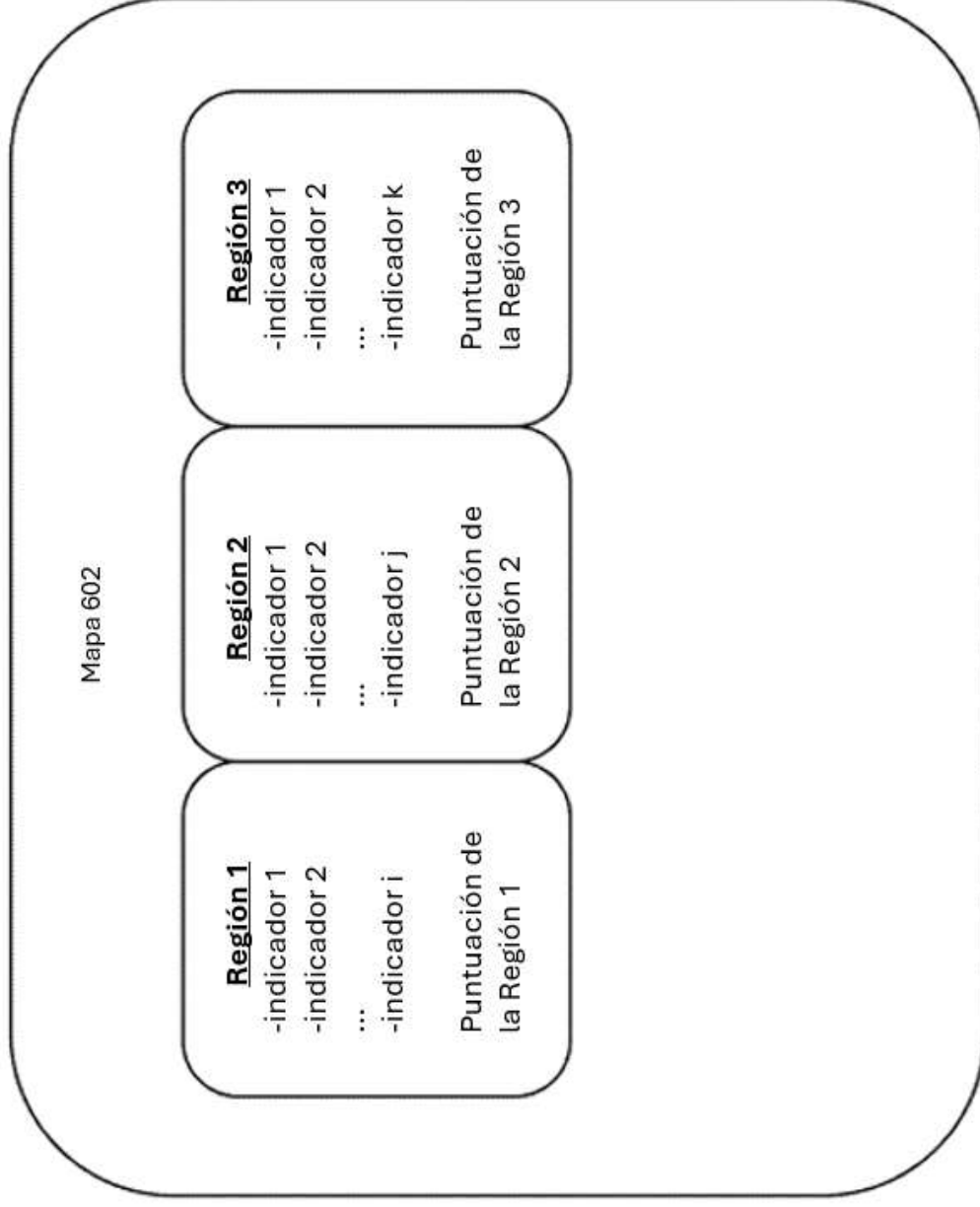


Fig. 6