

Uso de técnicas de Inteligencia Artificial para la búsqueda de datos generados por sensores

Alberto Berenguer¹, Adriana Morejón¹, David Tomás¹, and Jose-Norberto Mazón¹

Universidad de Alicante, San Vicente del Raspeig 03690, España
aberenguer@dlsi.ua.es, adriana.morejon@ua.es, dtomas@dlsi.ua.es
jnmazon@dlsi.ua.es

Resumen El uso de datos generados por sensores es crucial en el desarrollo de gemelos digitales. Sin embargo, reutilizar algunos de estos datos no es sencillo debido a su heterogeneidad y el incumplimiento de los principios FAIR. En este artículo corto, se presenta una arquitectura preliminar basada en técnicas de Inteligencia Artificial (IA) para mejorar su reutilización. Nuestra arquitectura consta de dos componentes: uno hace uso de grandes modelos de lenguaje (LLMs) para extraer datos y transformarlos a un formato reutilizable; y el otro indexa estos datos, permitiendo usar un algoritmo de búsqueda basado en *word embeddings*. Nuestra arquitectura facilita la reutilización de datos procedentes de sensores para diversos fines, como el desarrollo de gemelos digitales.

Keywords: Sensores · LLMs · Word embeddings · Principios FAIR

1. Introducción

Los datos generados por sensores están cobrando cada vez más importancia en el desarrollo de servicios y productos informáticos de valor añadido [4]. Sin embargo, como se resalta en [2], muchos de estos sensores ofrecen sus datos a través de formatos poco reutilizables como páginas web con multitud de diseños, formatos y estructuras, alejándose de cumplir con los principios FAIR [6], es decir, datos encontrables (**F**indable), accesibles (**A**ccesible), interoperables (**I**nteroperable) y reutilizables (**R**eusable). Esto sugiere que, tal y como se expresa en [7], a pesar de que existen grandes cantidades de datos procedentes de sensores, acceder a ellos y procesarlos para el desarrollo de un gemelo digital puede ser un gran desafío. Por otro lado, los métodos existentes para la búsqueda de datos de sensores se basan generalmente en el uso de palabras clave o en puntos geográficos concretos [3], lo que requiere una inspección manual de los resultados devueltos para comprobar si satisfacen las necesidades de un gemelo digital, haciendo el proceso muy tedioso.

Para abordar estos problemas, este artículo presenta una arquitectura que permite el uso de técnicas de IA para la reutilización de datos. En concreto, nuestra propuesta (i) utiliza las capacidades de LLMs como GPT [5] para transformar datos generados por sensores y publicados en formato Web a un formato

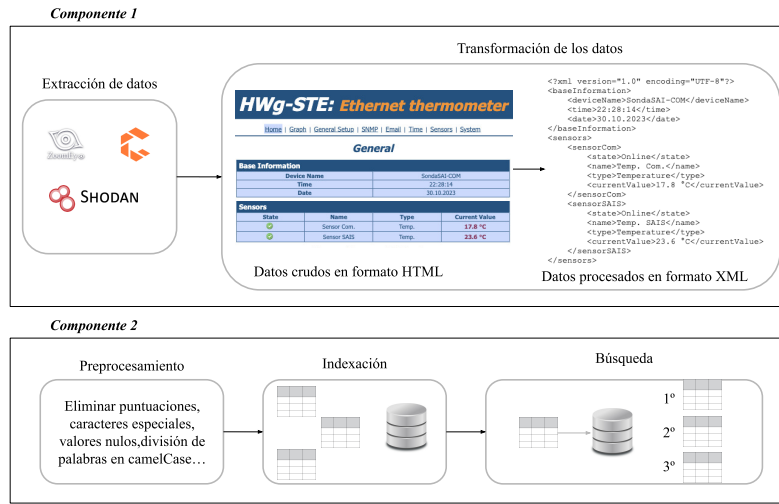


Figura 1. Arquitectura de recuperación y búsqueda de datos procedentes de sensores.

que cumpla con los principios FAIR, y (ii) indexa esos datos transformados para su posterior búsqueda mediante un proceso que emplea *word embeddings*, es decir, vectores que representan palabras en un espacio semántico, capturando su significado, y que nos sirven para realizar comparaciones a nivel semántico entre conjuntos de datos de sensores y permitir a los reutilizadores recuperar aquellos con mayor posibilidad de reutilización según sus intenciones, por ejemplo, en el enriquecimiento de un gemelo digital con nuevas fuentes de datos previamente no contempladas a partir de las que esté utilizando actualmente.

2. Arquitectura

La arquitectura propuesta (ver Figura 1) consta de dos componentes. El primero de ellos centrado en la extracción y transformación de datos, y el segundo centrado en indexar estos datos y facilitar su búsqueda para los usuarios.

2.1. Recolección de datos

Para recolectar datos procedentes de sensores, mejorando su reutilización, se aprovechan la comprensión contextual y las capacidades de generación de texto inherentes a los LLMs para transformar los datos. Esto requiere de dos pasos:

- **Extracción:** Se obtienen datos de fuentes como Shodan,¹ Censys,² ZoomEye³ o portales de datos abiertos, donde a menudo estos datos de sensores son compartidos directamente en formato Web, dificultando su reutilización [1].

¹ <https://www.shodan.io>.

² <https://search.censys.io>.

³ <https://www.zoomeye.org>.

- **Transformación de los datos:** Se convierten los datos no estructurados obtenidos en el paso anterior a un formato estructurado (en aras de su reusabilidad). Con el objetivo de alcanzar este propósito, se implementa el uso de LLMs a los cuales se le especifica la naturaleza de la transformación requerida. Esto permite transformar esos datos no procesados, extraídos directamente de una página web, en un formato estructurado, tal como archivos CSV o XML. Para ello, se emplean técnicas de ingeniería de *prompts* que optimicen los resultados de transformación.

2.2. Búsqueda de datos

Este componente permite indexar y recuperar los datos transformados para que puedan integrarse a partir de una entrada específica suministrada por un usuario que quiera utilizar datos. Esta operación permite identificar y recuperar aquellos conjuntos de datos indexados que son más pertinentes según las intenciones de las personas reutilizadoras, es decir, los datos que guardan relación directa con la información provista inicialmente por el usuario. Todo este proceso queda resumido en la Figura 2 y explicado en los siguientes pasos:

- **Preprocesamiento:** La limpieza de datos involucra procesos clave como la eliminación de signos de puntuación, descomposición de palabras en formato *camelCase*, y la eliminación de valores nulos, entre otros.
 - **Enriquecimiento:** Para compensar la falta de metadatos en nuestra arquitectura y enriquecer la búsqueda de datos con contexto relevante, se extrae la cobertura espacial y temporal directamente de la tabla.
 - **Indexación:** Este paso implica segmentar cada tabla previamente recuperada, calcular los *word embeddings* que representen cada columna, y posteriormente indexarlos utilizando una base de datos especializada en vectores.
- Búsqueda:** Se identifica, entre las tablas indexadas, aquellas que guardan relación con los datos de entrada introducidos por los usuarios. Para ello, se calcula la similitud promedio en función del *embedding* que representan cada columna. Posteriormente, las tablas se ordenan según su puntuación de similitud, ofreciendo una lista clasificada por su relevancia para el usuario de acuerdo a su contexto espacial y temporal específico.

3. Conclusiones

Este artículo propone un enfoque innovador para usar diversas técnicas de IA en (i) la extracción y transformación de datos de sensores en un formato reusable; y en (ii) la indexación y recuperación de datos relevantes para facilitar el desarrollo de gemelos digitales.

Como trabajo futuro, se implementará completamente la arquitectura propuesta, buscando la eficiencia en la transformación de datos, ya que el uso de modelos como GPT resulta significativamente costoso temporal y económicamente. Se plantea también una experimentación completa para poder validar

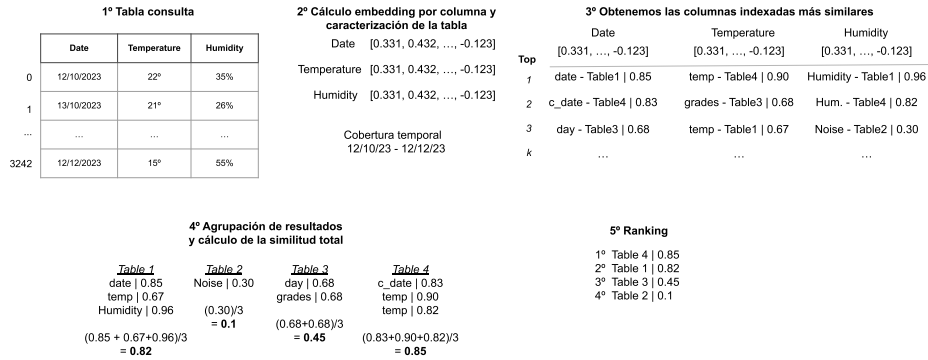


Figura 2. Proceso de búsqueda basado en *word embeddings*.

nuestra arquitectura en el desarrollo de gemelos digitales, haciendo hincapié en entornos donde existan datos heterogéneos que no cumplan los principios FAIR.

Agradecimientos Investigación financiada por la Generalitat Valenciana mediante el proyecto CIAICO/2022/019, así como por la Agencia Española de Investigación MCI-N/AEI/10.13039/501100011033 y por los fondos de la Unión Europea Next Generation EU/PRTR cómo parte de los proyectos TED2021130890B-C21 y PID2021-122263OB-C22. Alberto Berenguer tiene un contrato predoctoral con la Generalitat Valenciana y el Fondo Social Europeo, financiado mediante la subvención número ACIF/2021/507.

Referencias

- Berenguer, A., Morejón, A., Tomás, D., Mazón, J.N.: Using large language models to enhance the reusability of sensor data. *Sensors* **24**(2) (2024). <https://doi.org/10.3390/s24020347>, <https://www.mdpi.com/1424-8220/24/2/347>
- Liu, M., Li, D., Xu, C., Zhou, J., Huang, W.: Discovery of multimodal sensor data through webpage exploration. *IEEE Internet of Things Journal* **6**(3), 5232–5245 (2019)
- Liu, M., Li, D., Zeng, Y., Huang, W., Meng, K., Chen, H.: Combinatorial-oriented feedback for sensor data search in internet of things. *IEEE Internet of Things Journal* **7**(1), 284–297 (2020)
- McCreadie, R., Albakour, D., Manotumruksa, J., Macdonald, C., Ounis, I.: Searching the internet of things. In: *Building Blocks for IoT Analytics Internet-of-Things Analytics*, pp. 39–80. River Publishers (2022)
- Radford, A., Narasimhan, K.: Improving language understanding by generative pre-training (2018), <https://api.semanticscholar.org/CorpusID:49313245>
- Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., et al.: The FAIR guiding principles for scientific data management and stewardship. *Scientific data* **3**(1), 1–9 (2016)
- Zhang, M., Tao, F., Huang, B., Liu, A., Wang, L., Anwer, N., Nee, A.: Digital twin data: methods and key technologies. *Digital Twin* **1**, 2 (2022)