PROTOTIPO DE BÚSQUEDA INTELIGENTE CON LENGUAJE NATURAL PARA RECUPERACIÓN DE INFORMACIÓN EN REPOSITORIOS INSTITUCIONALES

Malgorzata Lisowska
CRAI Universidad del Rosario, Colombia | margarita.lisowska@urosario.edu.co
https://orcid.org/0000-0002-0697-5656

Blanco Castillo Humberto
CRAI – Universidad del Rosario, Colombia | humberto.blanco@urosario.edu.co
https://orcid.org/0000-0003-2339-0104

Sergio Abril Torres
CRAI – Universidad del Rosario, Colombia | sergio.abril@urosario.edu.co
https://orcid.org/0009-0009-1154-9289

DOI: *10.22477/xiv.biredial.372*

EJE TEMÁICO: *infraestructura tecnológica*

RESUMEN

Esta ponencia presenta el desarrollo de un prototipo de búsqueda inteligente basado en inteligencia artificial (IA) y procesamiento de lenguaje natural (PLN), diseñado para mejorar la recuperación de información en un repositorio institucional. Partimos del diagnóstico de una necesidad recurrente en muchos repositorios académicos: los resultados de búsqueda por palabras clave o aquellas con problemas ortográficos no siempre satisfacen las expectativas de los usuarios, dejando por fuera la recuperación de información relevantes. El objetivo fue construir un prototipo funcional que integrara un cajón de búsqueda semántica, capaz de interpretar consultas en lenguaje natural y devolver resultados pertinentes. La metodología llevada a cabo para la implementación del prototipo contempló: extracción de metadatos de los documentos almacenados en el repositorio institucional (autor, título, palabras clave, etc.); limpieza de caracteres especiales, espacios y registros incompletos; exportación de datos limpios en formato CSV; vectorización de textos usando modelos preentrenados de BERT en español para representar su contenido de manera semántica; indexación en FAISS, para optimizar el proceso de búsqueda y recuperación de información; integración de un cajón de búsqueda con el índice semántico. Como parte de los resultados se pretende: mejora en la relevancia de los resultados aprovechando que la solución puede entender el contexto e identificar documentos relevantes aun sin contener términos exactos en la consulta; reducción del tiempo de búsqueda, al ofrecer resultados más pertinentes desde el inicio; mejora en la visibilidad, debido a que podrían aparecer documentos que bajo las circunstancias actuales no aparecerían por tener limitaciones en los

metadatos o errores en los criterios de búsqueda. Este trabajo se convierte en una línea de innovación que puede ser replicable en otros repositorios instituciones y probablemente en un futuro podría extrapolarse para fortalecer la búsqueda desde redes de repositorios.

Palabras-clave: búsqueda semántica, repositorios institucionales, inteligencia artificial, procesamiento de lenguaje natural.

ABSTRACT

This paper presents the development of an intelligent search prototype based on artificial intelligence (AI) and natural language processing (NLP), designed to improve information retrieval in an institutional repository. The project started from diagnosing a recurring need in many academic repositories: search results based on keywords, or searches containing spelling errors, do not always meet user expectations, often failing to retrieve relevant information. The objective was to build a functional prototype that integrated a semantic search interface, capable of interpreting natural language queries and returning pertinent results. The methodology carried out for the prototype implementation included: extraction of metadata from documents stored in the institutional repository (author, title, keywords, etc.); cleaning of special characters, spaces, and incomplete records; export of cleaned data in CSV format; text vectorization using pre-trained BERT models in Spanish to represent their content semantically; indexing in FAISS to optimize the search and information retrieval process; integration of a search interface with the semantic index. The expected results include: improved relevance of results, leveraging the solution's ability to understand context and identify relevant documents even if they do not contain the exact terms from the query; reduced search time by offering more pertinent results from the start; improved visibility, as documents that would not appear under current circumstances due to metadata limitations or errors in search criteria could be retrieved. This work establishes an innovation pathway that can be replicated in other institutional repositories and could potentially be extrapolated in the future to enhance search capabilities across repository networks.

Keywords: semantic search, institutional repositories, artificial intelligence, natural language processing.

INTRODUCIÓN

Los repositorios institucionales se han consolidado como una herramienta esencial para la preservación y difusión de la producción académica y científica de las universidades. Sin embargo, aunque se trata de sistemas maduros, sostenidos por comunidades de desarrollo, aún presentan limitaciones en sus mecanismos de búsqueda, más aún cuando en el idioma español los datos contienen diacríticos o cuando las consultas incluyen errores ortográficos o no coinciden literalmente con los términos con los que se construyen los metadatos de los objetos **digitales** almacenados en dichos repositorios. Esta diferencia entre la manera como se estructuran las búsquedas y la manera como las máquinas las interpretan, frecuentemente invisibiliza los objetos o entrega resultados poco relevantes.

Luego de varias exploraciones que incluían intentos por ajustar o afinar indexadores como solr, y aprovechando el desarrollo acelerado de los últimos dos años de la inteligencia artificial, nos propusimos desarrollar un prototipo funcional de búsqueda que utilizara inteligen-

cia artificial y procesamiento de lenguaje natural, para mejorar la experiencia de los usuarios en cuanto a la recuperación de información. El resultado que esperábamos era contar con un cajón de búsqueda semántica que pueda interpretar las consultas de los usuarios y recuperar resultados relevantes, aún cuando los términos específicos no estén en los metadatos de los documentos.

Aunque actualmente se trata de un prototipo, el objetivo final es, además de mejorar la recuperación de resultados, la reducción del tiempo que invierten los usuarios del repositorio en realizar sus consultas y visibilizar documentos que dadas las condiciones actuales no aparecen en los resultados. Es probable que en el futuro este tipo de soluciones puedan escalar hacia redes de repositorios para la recuperación de información a nivel global.

MARCO TEÓRICO

Los métodos tradicionales de búsqueda por palabras clave, que aún se encuentran presentes en muchos sistemas de recuperación, son limitados al no capturar la intención del usuario, ni tener en cuenta el contexto semántico de las consultas. Esto ha motivado el surgimiento de nuevos enfoques donde se da prioridad a la interpretación del significado detrás de la palabra (búsqueda semántica), apoyados en tecnologías como la inteligencia artificial y el procesamiento de lenguaje natural.

Estudios recientes destacan que, en algunas bibliotecas académicas húngaras, por ejemplo, han adoptado soluciones basadas en IA para mejorar la recuperación de información (Winkler & Kiszl, 2022). Uno de los elementos centrales en esta evolución es la búsqueda semántica. A diferencia de los enfoques booleanos, la búsqueda semántica se basa en modelos de lenguaje capaces de representar el significado de textos y consultas en espacios vectoriales.

Modelos como BERT han demostrado ser especialmente eficaces al ofrecer representaciones contextuales profundas, lo que los hace ideales para tareas de comprensión y comparación semántica (Devlin et al., 2019). En otras palabras, este modelo entiende el significado de las palabras y cómo estas se conectan con las palabras que están alrededor. Por ejemplo, en una frase como “el banco está junto al río”, BERT interpreta que “banco” se refiere al objeto para sentarse, no a la entidad financiera, gracias a su análisis bidireccional del contexto (“junto al río”). Esta capacidad es esencial para ofrecer resultados de búsqueda más naturales y relevantes.

Por otra parte, FAISS (Facebook AI Similarity Search) es una biblioteca de código abierto desarrollada por Meta para realizar búsquedas eficientes de similitud en grandes colecciones de vectores. FAISS permite generar índices basados en estructuras como IndexFlatL2, que comparan rápidamente miles o millones de vectores en tiempo casi real. Este rendimiento es crítico en repositorios digitales de gran tamaño, donde se espera que las consultas sigan sien-

do rápidas aun cuando crezca la colección. Además, FAISS ofrece mecanismos para ajustar el equilibrio entre precisión y velocidad, adaptándose a distintos volúmenes de datos y requisitos de exactitud (Douze et al., 2025).

El desarrollo del prototipo de búsqueda semántica planteado en este proyecto se apoya en dos componentes fundamentales: la vectorización de textos mediante un modelo basado en BERT y el uso de la biblioteca FAISS como motor de indexación y recuperación. Esta combinación responde tanto a requerimientos funcionales como técnicos, ya que permite realizar búsquedas eficientes y escalables sobre grandes volúmenes de información:

1. **Comprensión contextual avanzada (BERT)**

- A diferencia de enfoques que se basan únicamente en frecuencias de términos (TF-IDF, Word2Vec), BERT genera embeddings que capturan el sentido de cada palabra según su posición en la oración.
- En el caso del repositorio EdocUR, donde las consultas pueden variar en forma y fondo, esta representación mejora la pertinencia de los resultados, al entender la intención real del usuario sin necesidad de operadores booleanos.

2. **Indexación y recuperación de alta velocidad (FAISS)**

- FAISS está optimizado para buscar coincidencias en colecciones de vectores de gran tamaño en tiempo casi real, gracias a estructuras como IndexFlatL2 y otros índices aproximados.
- Ofrece parámetros de configuración para balancear rendimiento y precisión, permitiendo adaptar el sistema a crecimientos futuros del repositorio.

3. **Solución equilibrada y escalable**

- Aunque existen alternativas para cada componente (Annoy o ScaNN para indexación; FastText o Word2Vec para vectorización), la combinación BERT + FAISS ofrece el mejor compromiso entre calidad semántica y rendimiento técnico.
- Esta arquitectura sienta las bases de una solución robusta para sistemas de recuperación de información inteligentes, capaz de evolucionar con el repositorio institucional y las necesidades de sus usuarios.

En este contexto, el prototipo desarrollado es innovador al aplicar de forma práctica estas tecnologías en un entorno académico real, integrando limpieza, vectorización y búsquedas semánticas para ofrecer una experiencia de búsqueda más intuitiva y útil a los usuarios del repositorio.

METODOLOGÍA

La implementación del prototipo se estructuró en varias fases orientadas a construir un sistema de búsqueda semántica que fuera funcional y estuviera adaptado al contexto del repositorio institucional EdocUR de la Universidad del Rosario. Cada paso respondió a una necesidad específica para avanzar el siguiente paso que condujera al procesamiento de los datos y al diseño del motor de búsqueda inteligente.

1. *Extracción de metadatos desde el repositorio institucional*: Se realizó una consulta SQL directa sobre la base de datos PostgreSQL de DSpace para obtener los metadatos más relevantes de cada objeto digital (uuid, título, uri, autores, tutor – para trabajos de grado-, palabras clave, fecha de publicación, tipo de documento, doi, accesRights). Esta consulta generó un archivo csv estructurado que serviría de base para la representación semántica posterior.

2. *Obtención del contenido textual de los documentos*: Una vez los documentos del repositorio están disponibles al público, se ejecuta el proceso "filter-media" el cual extrae la caratula del archivo pdf y un archivo txt que es utilizado para la indexación del motor de búsqueda solr del repositorio. El siguiente paso consistió en identificar y enviar a una carpeta específica todos los archivos txt los cuales contienen el contenido de cada objeto del repositorio, con el fin de robustecer la descripción del registro que posteriormente será vectorizado por la IA

3. *Asociación entre metadatos y archivos txt*: Era necesario entonces establecer una relación directa entre los metadatos y su correspondiente archivo de texto, para ello, se realizó una consulta a la base de datos que vinculaba ambos conjuntos de datos (archivo de metadatos y archivos txt). Esta vinculación era esencial para mantener la coherencia entre la descripción bibliográfica y el contenido analizado.

4. *Limpieza de datos*: Se diseñó y ejecutó un script en lenguaje Python para eliminar elementos que pudieran distorsionar los resultados del análisis semántico, como espacios dobles, saltos de línea, signos especiales (por ejemplo, signos de admiración) y otros errores resultados de la conversión de PDF a texto.

5. *Vectorización del contenido textual*: Una vez se obtuvieron los textos limpios y asociados con sus metadatos, se aplicó un modelo BERT[1] en español, para convertir cada documento en un vector que representara su contenido semántico. Esta etapa era clave para permitir la búsqueda basada en similitud de significado y no únicamente por coincidencia textual.

6. *Indexación con FAISS*: Los vectores que resultaron del paso anterior, fueron indexados utilizando la biblioteca FAISS, con el fin de optimizar el proceso de búsqueda.

[1] **BERT** (*Bidirectional Encoder Representations from Transformers*), es un modelo de aprendizaje profundo que usa redes neuronales desarrollado por Google

7. *Integración del cajón de búsqueda*: Por último, se incorporó un cajón de búsqueda en la interfaz del prototipo, para permitir que los usuarios pudieran realizar sus consultas usando lenguaje natural. Para ello, este cajón se conecta directamente con el índice FAISS, retornando resultados ordenados por proximidad semántica.
8. El despliegue de los registros permite identificar el titulo del documento, los autores, la fecha de publicación y el doi, que lo enlaza con el repositorio institucional.

Esta metodología permitió desplegar en corto tiempo un prototipo robusto, que al final se traduce en una interfaz pensada para mejorar la experiencia de búsqueda dentro del repositorio.

RESULTADOS Y DISCUSIÓN

El conjunto de datos (dataframe) utilizado para la indexación correspondió a 37324 objetos digitales almacenados en el repositorio institucional EdocUR (ver imagen 1).

El conjunto de datos (dataframe) utilizado para la indexación correspondió a 37324 objetos digitales almacenados en el repositorio institucional EdocUR (ver imagen 1).

Imagen 1 - Conjunto de datos utilizado

```
Información del DataFrame:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 37324 entries, 0 to 37323
Data columns (total 9 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   uuid                   37324 non-null  object
 1   Titulo                 37324 non-null  object
 2   URI                    37324 non-null  object
 3   Fechadedisponibilidad  37324 non-null  object
 4   Tutor                  12774 non-null  object
 5   Autor                  37265 non-null  object
 6   type                   37321 non-null  object
 7   dc.identifier.doi      27053 non-null  object
 8   dc.rights.accesRights  37324 non-null  object
dtypes: object(9)
memory usage: 2.6+ MB
None
```

Fuente: (ver com os autores)

Sobre estos registros se hizo necesario aplicar algunos ajustes para eliminar por ejemplo las filas duplicadas o revisar los valores que aparecían nulos (imagen 2)

Imagen 2 - Calidad de los datos

```
--- Reporte de Calidad de Datos ---
total_filas: 37324
total_columnas: 9
conteo_nulos_por_columna: {'uuid': 0, 'Titulo': 0, 'URI': 0, 'Fechadedisponibilidad': 0, 'Tutor': 24
550, 'Autor': 59, 'type': 3, 'dc.identifier.doi': 10271, 'dc.rights.accesRights': 0}
conteo_filas_duplicadas: 118
```

Fuente: (ver com os autores)

Durante las diferentes fases del proceso también se identificaron algunos errores sobre los datos, por ejemplo, durante el proceso de limpieza y asociación de metadatos con los textos extraídos, se detectó que algunos registros incluían listas de autores o palabras clave separadas por caracteres no estándar (‘¬’), con lo cual fallaba el procesamiento automatizado. También se encontraron textos con errores derivados de la conversión PDF a texto, en estos casos se realizó la corrección de forma manual.

También, durante el proceso de consulta, se identificaron otro tipo de ajustes requeridos en torno a las proximidades de algunos términos. Por ejemplo, el sistema presentaba asociaciones semánticas que generaban resultados incorrectos o poco pertinentes. Por ejemplo, consultas que incluían la palabra “vanguardia” arrojaban documentos relacionados con “guardia” o “salvaguarda”, evidenciando que los vectores semánticos compartían cercanías no deseables desde el punto de vista de la experiencia de usuario. Esto fue mitigado mediante el ajuste del umbral de similitud a 85%. De este modo se redujeron los falsos positivos sin afectar la recuperación de documentos valiosos.

A pesar de estos retos, el sistema pudo ser ajustado fácilmente y se realizaron múltiples pruebas utilizando diferentes criterios y consultas en lenguaje natural, lo cual permitió observar mejoras en la recuperación de información cuando se comparó contra el sistema tradicional de búsqueda del repositorio institucional.

Algunos de los resultados que pueden ser de especial interés fue la capacidad del prototipo para identificar documentos pertinentes que no estaban literalmente en los términos de la consulta. Por ejemplo, búsquedas como “conflicto armado” lograron recuperar documentos que mencionaban temas relacionados al desplazamiento forzado.

En general, los resultados indican que el uso de modelos de lenguaje para las búsquedas académicas no solo mejora la experiencia del usuario, sino que amplía el alcance de la recuperación de los objetos digitales del repositorio, fortaleciendo la visibilidad la producción institucional alojada en el repositorio.

CONCLUSIONES Y LÍNEAS FUTURAS

Los resultados de búsqueda obtenidos, permiten reforzar la hipótesis inicial que la incorporación de herramientas de inteligencia artificial en los repositorios institucionales tiene un alto potencial de fortalecer los procesos de recuperación de información. La incorporación de un cajón de búsqueda de lenguaje natural que ofrece, además la posibilidad de recuperar resultados con base en la semántica del contenido, puede sustituir los sistemas de búsqueda tradicionales por palabras clave.

Además, se destaca la importancia de la estandarización y limpieza de los metadatos en los repositorios, lo cual se puede extrapolar a cualquiera de las plataformas de gestión de conocimiento

Teniendo en cuenta que el prototipo busca implementar una solución que entienda el contexto de la búsqueda e identifique los documentos relevantes y como consecuencia de ello ayude a reducir el tiempo de búsqueda y visibilizar resultados que actualmente podrían permanecer ocultos cuando se busca de forma tradicional por palabras clave se plantea que posteriormente se ejecuten las siguientes líneas de acción:

- Realización de pruebas con usuarios reales: Para evaluar que esta implementación representa un verdadero impacto, es necesario realizar una evaluación que permita determinar la usabilidad del sistema y la calidad percibida por parte de los usuarios.
- Construcción de API´s: para interoperar con otros repositorios permitiendo búsquedas semánticas distribuidas entre distintas redes de repositorios.

Por último, es importante mencionar que este piloto no solo ha permitido explorar una solución técnica para la recuperación de contenido, sino que esperamos que aporte una semilla para reflexionar sobre el futuro de los repositorios académicos dentro de este nuevo contexto de la inteligencia artificial.

BIBLIOGRAFÍA

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805v2). arXiv. https://doi.org/10.48550/arXiv.1810.04805

Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.-E., Lomeli, M., Hosseini, L., & Jégou, H. (2025). *The Faiss library* (arXiv:2401.08281). ArXiv. https://doi.org/10.48550/arXiv.2401.08281

Winkler, B., & and Kiszl, P. (2022). *Views of Academic Library Directors on Artificial Intelligence: A Representative Survey in Hungary*. New Review of Academic Librarianship, 28(3), 256-278. https://doi.org/10.1080/13614533.2021.1930076

ANEXO 1

RESUMEN BIOGRÁFICO DE LOS AUTORES

Malgorzata Lisowska, directora del Centro de Recursos para el Aprendizaje y la Investigación – CRAI, de la Universidad del Rosario en Bogotá.

Magister en Bibliotecología e Información Científica, Universidad Jagiellona de Cracovia, Polonia. Especialista en Administración de Empresas, Universidad del Rosario. Especialista en Gerencia y Gestión Cultural, Universidad del Rosario. Amplia experiencia en bibliotecas públicas y universitarias, con énfasis en gestión y evaluación bibliotecaria y en implementación de nuevas tecnologías. investigadora en el proyecto de la Creación De La Biblioteca Digital Colombiana BDCOL, coordinación de proyectos internacionales como CoLaBoRa (Comunidad Latinoamericana de Bibliotecas y Repositorios Digitales) y en "LAReferencia" patrocinado por de la RedClara y el BID.

Humberto Blanco Castillo, jefe de Innovación y Proyectos del Centro de Recursos para el Aprendizaje y la Investigación – CRAI, de la Universidad del Rosario en Bogotá.

Ingeniero de sistemas, especialista en gerencia de proyectos TIC. Experto en el desarrollo de proyectos enfocados a la implementación, visibilidad e interoperabilidad de repositorios institucionales, así como el desarrollo de soluciones basadas en software libre para la gestión de bibliotecas. Actualmente lidera las estrategias para promover la visibilidad de la producción institucional en acceso abierto, la generación iniciativas y gestión proyectos de base tecnológica que apoyan a los procesos de innovación del CRAI.

Sergio Abril Torres, profesional en arquitectura de información del Centro de Recursos para el Aprendizaje y la Investigación – CRAI, de la Universidad del Rosario en Bogotá.

Ingeniero de sistemas, Maestría en Business Analytics. Con experiencia en el diseño de arquitecturas de información y el desarrollo de aplicaciones para la gestión y análisis de recursos académicos. Experiencia en la implementación y fortalecimiento de plataformas digitales para bibliotecas académicas, desarrollo de soluciones tecnológicas que promueven la interoperabilidad y el acceso abierto al conocimiento.