01. Introduction

This work presents a domain-aware cognitive search engine (SE) designed in EIFFEL H2020 project. It aims to exploit recent advances in Machine Learning (ML)-based Natural Language Processing (NLP) to overcome current challenges in the searching capabilities of Data Portals. The system includes an optimized AI Large Language Model (LLM) retrained with an extensive Climate Change (CC)-specific text corpus. Cognitive search adds language understanding to the search results, promoting the most semantically relevant results to the top. The use-case is the GEOSS Portal, but the same principles apply elsewhere.

02. Conventional vs Cognitive search data discovery experience

Conventional search engines, facts

- Based on exact or fuzzy term searches.
- Limited data discovery capabilities.
- The inherent limitations of such approaches are expressed in a higher degree in metadata querying since the available text is limited to, usually, title, description and keyword lists.

Cognitive search engines, facts

- Based on data-driven language models.
- Allow free-text querying.
- The language model inherently performs the semantic analysis. It understands individual words' meanings, the meaning of words within their context, the semantic relationships between individual words and even the meaning of whole sentences.

03. How does cognitive search works

- The LLM transforms all documents into mathematical vectors (Embeddings), inherently performing semantic analysis.
- Words/terms and sentences with similar concepts and meanings lie close in the embedding space.
- The distance between the document and the query embeddings measures the relevance between a document in the database and any query.
- Semantic search adds language understanding to search results, promoting the most semantically relevant results to the top.
- It can be domain-aware. In EIFFEL, we aim for CC domain specificity.

04. CC Domain-specific corpus collection for LLM training


05. LLM fine-tuning for domain-aware cognitive search

- The metadata (e.g., titles, descriptions, keywords) pass through the LLM to produce metadata embeddings (offline process).
- The user query passes through the language model to produce the query embedding (online process).
- The semantically similar data objects are returned in ranked order.
- Elastic search stores embeddings and calculates vector similarity fast.

06. EIFFEL Cognitive search pipeline

- Promising results with improved performance compared to conventional search engines.
- The system includes an optimized AI Large Language Model (LLM) retrained with an extensive Climate Change (CC)-specific text corpus. Cognitive search adds language understanding to the search results, promoting the most semantically relevant results to the top.
- It can be domain-aware. In EIFFEL, we aim for CC domain specificity.

References


Acknowledgement

This work is part of the EIFFEL H2020 project. The EIFFEL project has received funding from the European Union’s Horizon 2020 research and innovation programme under GA No 868485.