

ARQUITETURA PARA ANÁLISE E RECUPERAÇÃO SEMÂNTICA DE DADOS CIENTÍFICOS EM LARGA ESCALA

ARCHITECTURE FOR SEMANTIC ANALYSIS AND RETRIEVAL OF LARGE-SCALE SCIENTIFIC DATA

Heitor Rodrigues Carvalho

Centro Federal de Educação Tecnológica de Minas Gerais, CEFET-MG
E-mail: heitorrodrigues2909@gmail.com

Thiago Magela Rodrigues Dias

Centro Federal de Educação Tecnológica de Minas Gerais, CEFET-MG
E-mail: thiagomagela@gmail.com
ORCID: <https://orcid.org/0000-0001-5057-9936>

Patrícia Mascarenhas

Universidade do Estado de Minas Gerais, UEMG
E-mail: patriciamdias@gmail.com

RESUMO

A crescente disponibilidade de grandes volumes de dados científicos exige o desenvolvimento de soluções capazes de integrar análise estruturada, recuperação de informação e interpretação semântica. Este trabalho propõe uma arquitetura computacional para exploração de dados curriculares acadêmicos baseada na integração entre técnicas de análise bibliométrica e modelos de linguagem avançados. A solução utiliza dados extraídos da Plataforma Lattes, processados por meio de etapas de transformação, normalização e indexação em banco de dados relacional com suporte a vetores semânticos. A abordagem combina análises quantitativas tradicionais com mecanismos de recuperação aumentada por geração (RAG), permitindo consultas em linguagem natural e síntese automatizada de informações científicas. Os resultados demonstram que a integração entre visualização interativa e inteligência artificial amplia significativamente a capacidade de exploração e interpretação dos dados, oferecendo suporte tanto a análises exploratórias quanto a consultas complexas. O estudo contribui para o avanço de infraestruturas de organização da informação científica, destacando o potencial da inteligência artificial na construção de sistemas mais interativos, escaláveis e orientados à descoberta de conhecimento.

Palavras-chave: Organização da informação científica, Inteligência artificial, Recuperação aumentada (RAG), Análise bibliométrica, Plataforma Lattes.

ABSTRACT

The growing availability of large volumes of scientific data requires the development of solutions capable of integrating structured analysis, information retrieval, and semantic interpretation. This paper proposes a computational architecture for exploring academic curriculum vitae data based on the integration of bibliometric analysis techniques and advanced language models. The solution uses data extracted from the Lattes Platform, processed through transformation, normalization, and indexing stages in a relational database with support for semantic vectors. The approach combines traditional quantitative analyses with retrieval-augmented generation (RAG) mechanisms, enabling natural language queries and automated synthesis of scientific information. The results demonstrate that the integration of interactive visualization and artificial intelligence significantly expands the capacity to explore and interpret data, supporting both exploratory analyses and complex queries. The study contributes to the advancement of infrastructures for organizing scientific information, highlighting the potential of artificial intelligence in building more interactive, scalable systems oriented toward knowledge discovery.

Keywords: Scientific information organization, Artificial intelligence, Retrieval-augmented generation (RAG), Bibliometric analysis, Lattes Platform.

REFERÊNCIAS

DIAS, T. M. R. **Um estudo sobre a produção científica brasileira a partir de dados da Plataforma Lattes**. 2016. Tese (Doutorado em Modelagem Matemática e Computacional) – Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte, 2016.

LANE, J. Let's make science metrics more scientific. **Nature**, Nature Publishing Group, v. 464, n. 7288, p. 488, 2010. Disponível em: <<https://doi.org/10.1038/464488a>>. Acesso em: 10 mar. 2026.