Graph Retrieval Augmented Generation for Scientific Publications

AUTHOR: ANTONIO GASSNER

EXAM: TECHNOLOGIES FOR BIG DATA MANAGEMENT
PROFESSOR: MASSIMO CALLISTO DE DONATO

Introduction

01.

CEUR-WS is a free, open-access publication service focused primarily on computer science workshops.

02.

Each volume includes detailed **metadata** such as paper titles, author names and affiliations, abstracts, and publication details.

03.

We want to use this rich metadata to **improve search** functionality across their dataset of scientific publications

Project Objectives

Utilize **Graph-Retrieval Augmented Generation** to semantically search CEUR-WS content.

Model the relationships among **publications**, **authors**, and **topics** to provide context for queries.

Combine graph-based insights with the **natural language understanding** of large language models to improve search functionality.

Evaluate and validate the approach using the CEUR-WS dataset scraped using this Scraper developed by colleagues at UNICAM

Methodology

Retrieval Augmented Generation

A hybrid approach that combines **LLMs** with **information retrieval** to produce context-aware, accurate responses.

Knowledge Graphs

Structured representations of **entities** and their **relationships** that provide semantic context.

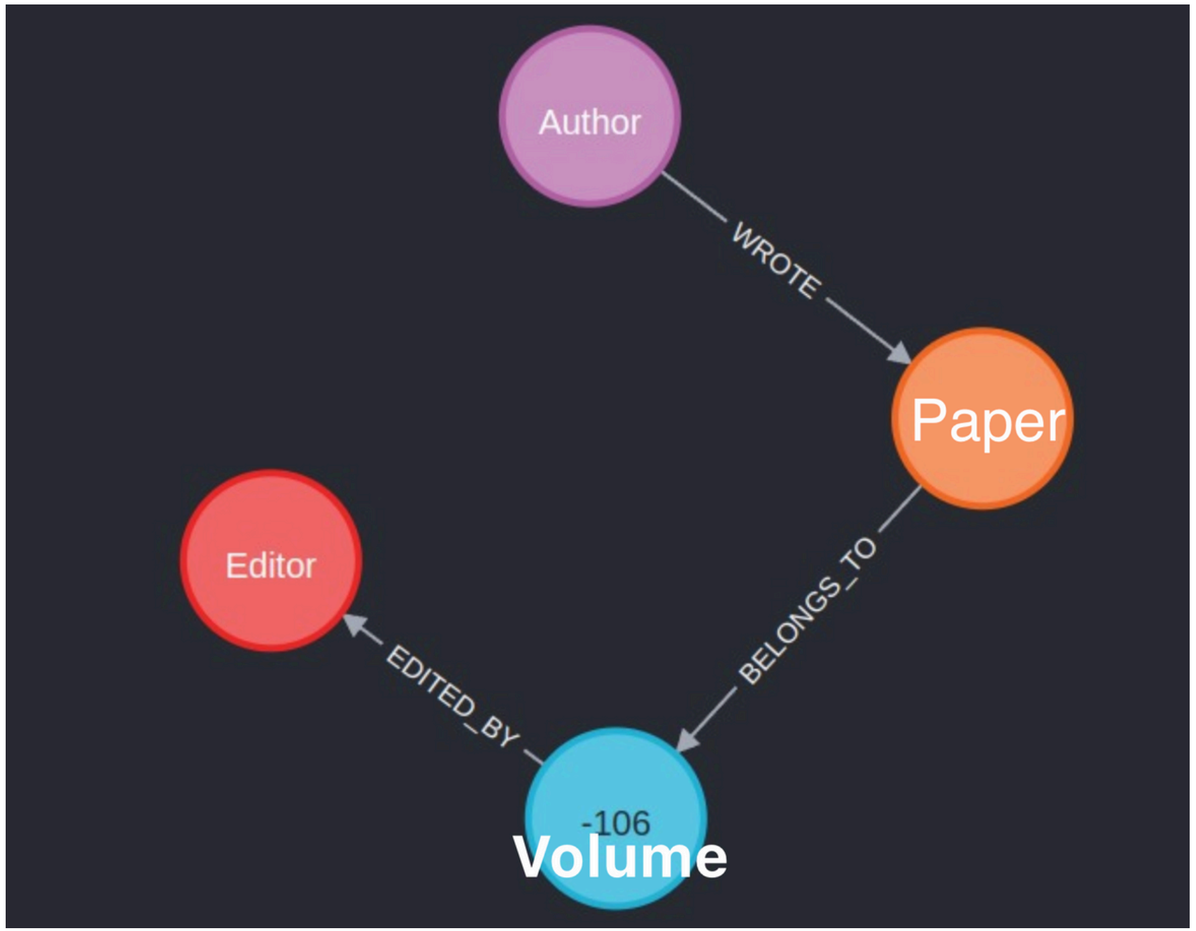
Combined Impact

Integrating RAG with knowledge graphs grounds the generated content, **reducing hallucinations** and **improving retrieval precision**

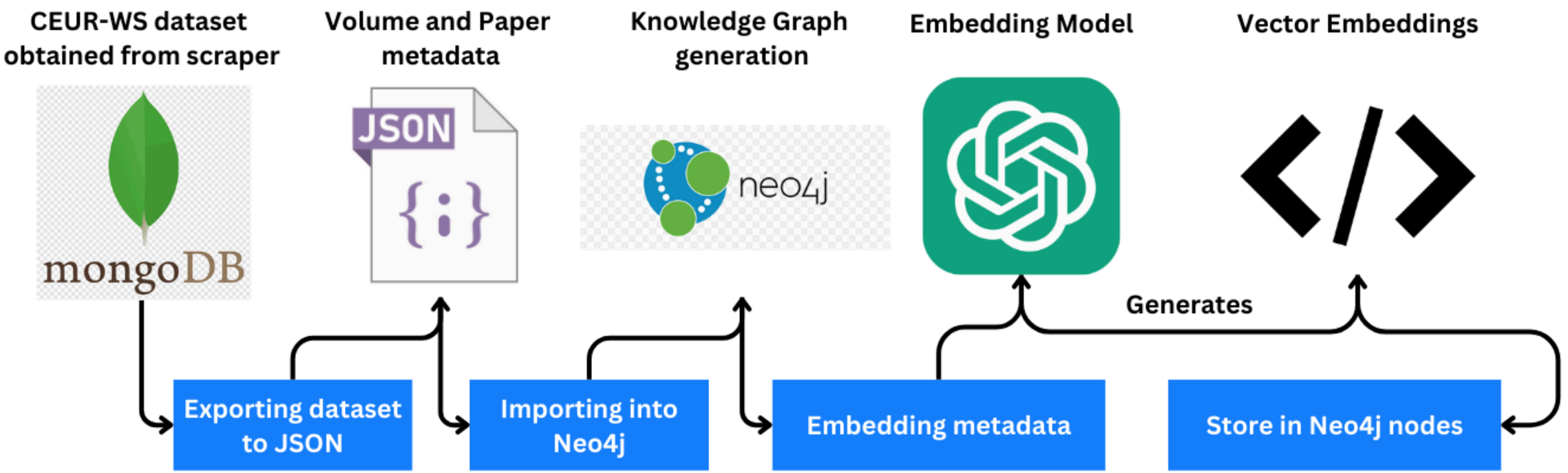
Technology

Neo4j	Graphrag	OpenAI Embedding Models & LLMs
A robust **graph database** that efficiently stores and queries interconnected data using graph structures.	A **high-level framework** by Neo4j that streamlines developing applications using graph retrieval augmented generation.	Advanced models that convert text into semantic **vectors** and generate context-aware responses

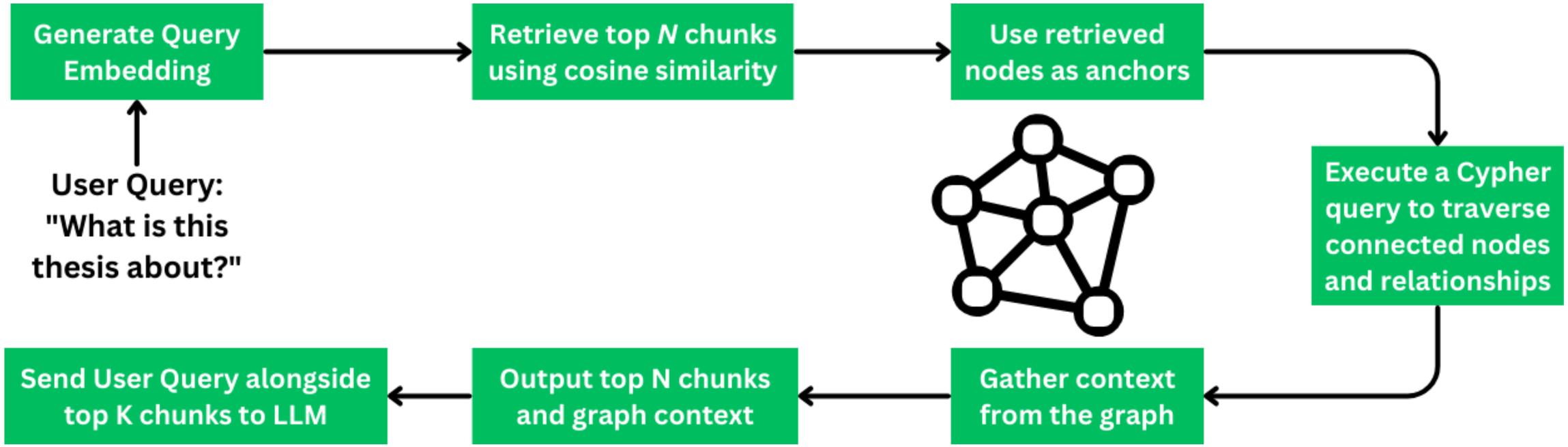
Graph Structure



Architecture - Ingestion



Architecture - Retrieval



Results

Retrieved Chunk

[RETRIEVERRESULTITEM(CONTENT="{'VOLTITLE': 'JOINT PROCEEDINGS OF THE SIGIR-AP 2024 WORKSHOPS EMTCIR 2024 AND UM-CIR 2024',
'ID': '67DBF8C44E4E9E66638FBD2A',
'URN': 'URN:NBN:DE:0074-3854-4',
'TITLE': ' CEUR-WS.ORG/VOL-3854 - JOINT PROCEEDINGS OF THE SIGIR-AP 2024 WORKSHOPS EMTCIR 2024 AND UM-CIR 2024',
'LOCTIME': 'TOKYO, JAPAN, DECEMBER 12, 2024',
'VOLACRONYM': 'EMTCIR_UM-CIR 2024',
'VOLNR': 'VOL-3854',
'FULLTITLE': 'JOINT PROCEEDINGS OF THE 1ST WORKSHOP ON EVALUATION\\N METHODOLOGIES,TESTBEDS AND COMMUNITY FOR INFORMATION ACCESS RESEARCH\\N (EMTCIR 2024) AND THE 1ST WORKSHOP ON USER MODELLING IN CONVERSATIONAL\\N INFORMATION RETRIEVAL (UM-CIR 2024)',
'EMBEDDING': NONE,
'PUBYEAR': '2024'}",
METADATA={
'SCORE': 0.7629427909851074,
'NODELABELS': ['VOLUME'],
'ID': '4:A90A02F2-7D0D-4906-B34C-1B60A4E5DD25:5952'}),

RAG Answer

CEUR HAS PUBLISHED SEVERAL PROCEEDINGS FOR VARIOUS WORKSHOPS AND CONFERENCES. HERE ARE SOME EXAMPLES:

1. **JOINT PROCEEDINGS OF THE SIGIR-AP 2024 WORKSHOPS EMTCIR 2024 AND UM-CIR 2024**: THIS INCLUDES THE 1ST WORKSHOP ON EVALUATION METHODOLOGIES, TESTBEDS, AND COMMUNITY FOR INFORMATION ACCESS RESEARCH (EMTCIR 2024) AND THE 1ST WORKSHOP ON USER MODELLING IN CONVERSATIONAL INFORMATION RETRIEVAL (UM-CIR 2024). IT TOOK PLACE IN TOKYO, JAPAN, ON DECEMBER 12, 2024.

2. **THE 1ST INTERNATIONAL WORKSHOP ON ADVANCED APPLIED INFORMATION TECHNOLOGIES (ADVAIT 2024)**: THIS WORKSHOP WAS HELD IN KHMELNYTSKYI, UKRAINE, AND ZILINA, SLOVAKIA, ON DECEMBER 5, 2024. THE PROCEEDINGS ARE PUBLISHED IN VOL-3899.

Future Improvements

Develop a dedicated **index** for *Papers* to enable semantic search at the document level alongide the *Volume* level.

Implement an **automated scraping pipeline** that augments metadata with vector embeddings and integrates this enriched data into the index.

Increase the richness of metadata (by incorporating abstracts) to address the limitations posed by **small text chunks** and improve overall data quality.

Enhance usability by developing a **frontend**, **containerizing** the application for easier deployment, and releasing the service for public access.

Thank You!

https://github.com/AntonioGassner/TBDM_CEUR_Graph_RAG

https://github.com/AronOehrli/TBDM-CEUR-WS