EGU24-19358, updated on 03 Nov 2024
https://doi.org/10.5194/egusphere-egu24-19358
EGU General Assembly 2024

Research Notebook Retrieval with Explainable Query Reformulation

Na Li[1], Peide Zhu[1,2], Gabriel Gabriel Pelouze[1,2], Spiros Koulouzis[1,2], **Zhiming Zhao**[1,2], and Zhiming Zhao

[1]University of Amsterdam, Science Faculty, Multiscale Networked Systems, Amsterdam, Netherlands
[2]LifeWatch ERIC Virtual Lab Innovation Center (VLIC), Amsterdam, Nederalands

Data science and Machine learning techniques are increasingly important in tackling societal challenges and complex problems in environmental and earth sciences. Effectively sharing and (re)using research assets, including data sets, models, software tools, documents, and computing notebooks, are crucial for such data-centric research activities. Researchers can reproduce others' experiments with the available research assets to verify the results and to further develop new experiments. Computational notebooks, as an important asset, comprise free-form textual descriptions and code snippets. The notebook runtime environments, e.g., Jupyter, provide scientists with an interactive environment to construct and share the experiment descriptions and code as notebooks.

To enable effective research assets discovery, research infrastructures not only need to FAIRify the assets with rich meta information and unique identifiers but also provide search functionality and tools to facilitate the construction of scientific workflows for the data sciences experiments with research assets from multiple sources. The general-purpose search engines are helpful for initial coarse-grained search but often fail to find multiple types of research assets such as the data sets and notebooks needed by the research. The community-specific catalogues, e.g., in ICOS and LifeWatch, provide search capabilities for precisely discovering data sets, but they are often characterized by a specific type of asset. A researcher has to spend lots of time searching across multiple catalogues to discover all types of assets needed.

In the search process, user queries tend to be short and comprised of several key phrases that demand great efforts to understand users' information needs. Given the complexity of computational notebook contents and the mismatch between the form of user queries and the computational notebooks, it is critical to understand queries by augmentations and make explainable relevance judgments. To address these challenges, we developed a research asset search system for a Jupyter notebook-based Virtual Research Environment (called Notebook as a VRE) that supports scientific query understanding with query reformulation and explainable computational notebook relevance judgments via computational notebook summarization.

The proposed system includes four major components: the query reformulation module, the

notebook indexer and retriever, the summarization component, and the user interface. The query reformulation module performs active query understanding via query reformulation, where we extract scientific entities from user queries and search related entities from external knowledge graphs and resources as expansions and rank the reformulated queries for users to choose from. The system has been validated via a small user group and will be further developed in the coming ENVRI-HUB next project to support conversational search and recommendation.