Context-Aware Notebook Search in a Jupyter-Based Virtual Research Environment

Li, N.; Farshidi, S.; Bianchi, R.; Koulouzis, S.; Zhao, Z.

DOI
10.1109/eScience55777.2022.00054

Publication date
2022

Document Version
Author accepted manuscript

Published in
eScience '22 : Democratizing science : 2022 IEEE 18th International Conference on e-Science

License
CC BY

Citation for published version (APA):
Context-Aware Notebook Search in a Jupyter-Based Virtual Research Environment

Na Li*, Siamak Farshidi†, Riccardo Bianchi‡, Spiros Koulouzis†, Zhiming Zhao*†
*Multiscale Networked System, Informatics Institute, University of Amsterdam, Netherlands,
†LifeWatch ERIC Virtual Lab and Innovation Center (VLIC), Amsterdam, Netherlands
{n.li | s.farshidi | r.bianchi | s.koulouzis | z.zhao}@uva.nl

Abstract—Computational notebook environments such as the Jupyter play an increasingly important role in data-centric research for prototyping computational experiments, documenting code implementations, and sharing scientific results. Effectively discovering and reusing notebooks available on the web can reduce repetitive work and facilitate scientific innovations. However, general-purpose web search engines (e.g., Google Search) do not explicitly index the contents of notebooks, and notebook repositories (e.g., Kaggle and GitHub) require users to create domain-specific queries based on the metadata in the notebook catalogs, which fail to capture the working contexts in the notebook environment. This poster presents a Context-aware Notebook Search Framework (CANSF) to enable a researcher to seamlessly discover external notebooks based on semantic contexts of the literate programming activities in the Jupyter environment.

Index Terms—computational notebook, Jupyter notebook, notebook search, data science, code reuse

I. INTRODUCTION

A computational notebook, e.g., Jupyter notebook, provides a literate programming [1] environment for researchers to prototype and execute computational experiments. It can effectively support exploratory and iterative computational processes, and the documented notebooks can be easily shared for scholarly communication and code reuse [2].

During past years, more than a billion notebooks have been created,[1] which become an important type of web resources for researchers to check related work, reuse existing implementations, and expedite scientific innovations. However, general search engines mainly focus on HTML documents and multimedia and do not explicitly index notebooks. Moreover, existing search tools rely on queries explicitly generated by users and fail to capture the programming contexts in the notebook environment.

Several barriers hamper the finding, accessing, and utilizing (namely FAIRness [3]) of notebooks. First, notebooks available on the web have a diverse quality of documentation and source codes. Most notebooks published on GitHub are “not narratives but collections of scripts with loose notes” [4], which makes the execution difficult and thus impairs the reproducibility of the experiments contained in these notebooks. Second, only a small portion of notebooks created by scientific communities have been curated and collected by public notebook repositories or catalogs. Many research infrastructures, e.g., ENVRI [5], still focus on the assets of research data and web services in their catalogs; less attention has been paid to notebooks. Yet most catalogs treat the notebook as a single digital object or part of a project and do not provide searching over notebook contents.

This paper aims to tackle notebook search challenges in the context of the research activities conducted in a Jupyter environment. We propose a Context-aware Notebook Search Framework (CANSF), a semantic notebook search system that can derive context aware queries from users’ working notebooks and search for related notebooks. We prototype the proposed system and integrate it with the Jupyter environment.

II. PROBLEM STATEMENT AND RELATED WORK

Notebooks group content into chunks called cells. There are two basic types of cells: code cells containing code fragments, and Markdown cells for narrative descriptions. Notebooks can be searched based on metadata, code in code cells, or textual descriptions in Markdown cells. Research on notebook search is still in its infancy, with little prior work. NBSearch [6] supports semantic code search in an extensive notebook collection and visual result exploration. It treats code cells as search units and does not consider Markdown cells. Jupyter [7] is a content-based notebook search system that takes codes, tabular data and libraries as queries. However, these systems work as standalone search tools, do not have seamless integration with the Jupyter environment, and most of all, do not consider contextual information.

III. CONTEXT-AWARE NOTEBOOK SEARCH FRAMEWORK

To address the challenges we highlighted we propose the Context-aware Notebook Search Framework (CANSF). The main focuses of the system are: 1) context-awareness during searching and 2) support for semantic notebook representation. Users that extend their Jupyter environments with CANSF achieve increased efficiency in their notebook programming.

A. System Design

Fig. 1 illustrates the high-level architecture of the framework. It consists two parts: a search agent to be installed on the client side, e.g., a Jupyter environment, and a search backend.

The Search agent is integrated in the notebook environment and observes the activities conducted by the user. By analyzing

---

1https://blog.jetbrains.com/datalore/2020/12/17/we-downloaded-10-000-000-jupyter-notebooks-from-github-this-is-what-we-learned/
Solid and rich indexes are essential to support context matches of the retrieved notebooks with the user operational status. To measure this we tested the performance of our semantic indexing pipeline. We collected 37 queries from 7 researchers from the CLARIFY project. In this case the contextual information is primarily about medical imaging and related AI models. At the same time we collected and indexed 1250 notebooks crawled from the Kaggle platform. Our experiments show that CANSF can provide the users with highly relevant external notebooks that well match their query and context.

IV. CONCLUSION AND FUTURE WORK

In this poster, we present CANSF, Context-aware Notebook Search Framework, that increases the discover-ability of computational notebooks. In the future, we will improve notebook search performance by introducing computing-related factors into notebook ranking, e.g., execution time. Furthermore, given that notebooks can be seen as a connection for different research objects, we will investigate a combined search paradigm, e.g., searching for datasets and machine learning models simultaneously. Finally, it is our goal to include CANSF into the general architecture of the NaavRE - Notebook-as-a-Virtual-Research-Environment [8].

ACKNOWLEDGMENT

This work has been partially funded by the European Union’s Horizon 2020 research and innovation program by the project CLARIFY under the Marie Skłodowska-Curie grant agreement No 860627, by the ARTICONF project grant agreement No 825134, by the ENVRI-FAIR project grant agreement No 824068, by the BLUECLOUD project grant agreement No 862409, by the LifeWatch ERIC.

REFERENCES


2https://www.kaggle.com/