The Development of a Data Recommender System for Improving the Discovery of Environmental and Biological Scientific Datasets

***Anusuriya Devaraju*[1], *Uwe Schindler*[2], *Michael Diepenbroek*[2], *Robert Huber*[2], *Jens Klump*[1], *Markus Stocker*[3]**

[1]*CSIRO Mineral Resources, Australia*; [2]*MARUM, University of Bremen, Germany*; [3]*German National Library of Science and Technology (TIB), Germany*

Corresponding author(s) e-mail: anusuriya.devaraju@googlemail.com

ABSTRACT:

In biological and environmental science, there has been good progress in setting up several data repositories to provide greater access to scientific data. However, users cannot realize their values if they cannot quickly locate the datasets required for their scientific research and applications. Recent studies on data retrieval practices [1,2] have revealed that current data portals lack effective data discovery solutions. Text or keyword search provided by the data portals matches user queries and data descriptions to rank the relevant datasets. This type of search depends on how well the data owners described the datasets, or how the users expressed their information needs. It may yield either empty results or too many almost identical datasets. Additionally, users need to be familiar with the structure and terminology of the portal to obtain meaningful results. The text-based search may produce top-ranked search results, which may be retrieved from the same data collection, sharing common attributes. As a result, users are unlikely to discover novel datasets. Therefore, we need an innovative data discovery solution that complements the existing search tools on the portals to deliver relevant and new datasets to users. To address this challenge, we developed a data recommender system for scientific datasets. We describe the system in the context of the PANGAEA data portal. PANGAEA is a data publisher for Earth and Environmental Sciences and hosts more than 370,000 datasets with more than 12 billion measurements from various disciplines. It uses Elasticsearch to index and to support full-text search on the datasets. We present the design and development of the recommender system and describe how a data search engine can be used to build a scalable data recommender system. The data recommender system uses the metadata of datasets and the user interactions extracted from the data server logs to deliver two types of recommendations, i.e., 'similar datasets' and 'users who are interested in this item are also interested in..'. In addition to PANGAEA, we provide some insights on how we may apply the system in the context of other scientific data portals such as the German Federation for Biological Data (GFBio). Building a data recommender on top of Elasticsearch enhances the scalability and maintainability of the recommender system. Our work is an essential contribution towards developing a real-world recommender system for improving scientific dataset discovery.

KEYWORDS: Recommender System, Open Scientific Data, Digital Library, PANGAEA, Data Discovery

REFERENCES:

1. Kathleen Gregory, Paul T. Groth, Helena Cousijn, Andrea Scharnhorst, and Sally Wyatt. 2017. Searching data: A review of observational data retrieval practices. CoRR, http://arxiv.org/abs/1707.06937.
2. Dagmar Kern and Brigitte Mathiak. 2015. Are there any differences in data set retrieval compared to well-known literature retrieval? Springer International Publishing, Cham. 197–208. https://doi.org/10.1007/978-3-319-24592-8_15