Data Recommender System: Improving the Discovery of Environmental Datasets through Text Analytics and Usage Mining

Anusuriya Devaraju, Uwe Schindler, Robert Huber, Michael Diepenbroek
PANGAEA, MARUM – Center for Marine Environmental Sciences,
University of Bremen, Leobener Str. 8, D-28359 Bremen, Germany.
{adevaraju,uschindler,rhuber,mdiepenbroek}@marum.de

Abstract

In Earth and Environmental Sciences, as observation data in digital repositories grows, it can become hard for users to find the relevant and novel datasets that they need for their applications. Existing data portals primarily support data finding through full text, faceted or map-based search, but they lack innovative data discovery tools. A number of user studies on digital repositories have revealed that data portals failed to deliver meaningful search results that users may expect (Gregory et al., 2017), (Kern & Mathiak, 2015), (Maier et al., 2014). The data search mechanisms employed by the portals may be suitable for users who perform known-item searches (e.g., search by author, title or DOI) and the users who are familiar with the nature and structure of the repositories. However, these mechanisms may be insufficient for users who are unable to clearly articulate their needs or simply seek test datasets. We need innovative data discovery tools that complement existing search functionalities on the portals to improve user experience. PANGAEA is a digital repository for archiving and publishing environmental datasets. It is jointly managed by the Alfred Wegener Institute, Helmholtz Center for Polar and Marine Research (AWI) and the Center for Marine Environmental Sciences (MARUM), at the University of Bremen. The infrastructure holds more than 380,000 datasets (e.g., time series, spatial, images) from individual researchers, projects, data centers, and research infrastructures. This presentation describes the development and evaluation of a recommender system to improve the data discovery on the PANGAEA portal. The recommender system applies text analytics and usage mining to produce two types of data recommendations on the portal. First, "Datasets with similar metadata"- this resembles text-matching search; however, we utilize the metadata (e.g., title, author, location, time, publication, etc.) of a target dataset to produce its similar datasets. Second, "Users that were interested in this dataset were also interested in" – we produce this recommendation based on web log analysis (e.g., data search and download patterns). We evaluated the recommender system in online settings, and the evaluation results (i.e., click-through rate) signify the overall effectiveness of the recommender system in improving data discovery.

Keywords: Recommender System, Environmental Data, Content-based Filtering, Usage Mining

References

Gregory, K., Groth, P., Cousijn, H., Scharnhorst, A., and Wyatt, S. (2017). Searching Data: A Review of Observational Data Retrieval Practices. *Journal of Information Science*, (2019). Retrieved from https://doi.org/10.1177%2F0165551519837182

Kern D., and Mathiak., B. (2015). Are There Any Differences in Data Set Retrieval Compared to Well-Known Literature Retrieval? In Kapidakis S., Mazurek C., Werla M. (Eds), *Research and Advanced Technology for Digital Libraries*. TPDL 2015. Lecture Notes in Computer Science, vol 9316. Springer, Cham, Retrieved from https://doi.org/10.1007/978-3-319-24592-8_15

Maier D., Megler V.M., Tufte K. (2014) Challenges for Dataset Search. In Bhowmick S.S., Dyreson C.E., Jensen C.S., Lee M.L., Muliantara A., Thalheim B. (Eds), *Database Systems for Advanced Applications*. DASFAA 2014. Lecture Notes in Computer Science, vol 8421. Springer, Cham. Retrieved from https://doi.org/10.1007/978-3-319-05810-8_1