User Driven Architecture for Data Discovery

NSF Award No. 1443070 PI: Giridhar Manepalli Co-PI: Allison Powell
{gmanepalli, apowell} @cnri.reston.va.us

How do you enable a uniform, consistent, and global discovery solution for datasets? Web search engines have solved this problem for textual content primarily by crawling the web via embedded links while indexing the text on those web pages. In the case of scientific materials, especially scientific datasets, there is an inherent lack of both links and "search-able" information; and even if those datasets are somehow made accessible to index engines, most of those datasets are just numbers. Despite the organized efforts of the semantic web (e.g., linked data) and registry approaches (e.g., re3data), the discovery challenge has remained largely an unsolved problem.

The biggest challenges in the dataset discovery space are social and organizational: what incentives do institutions have to coordinate across several other institutions to participate in any organized federation approach? How could you ensure that dataset producers and repository administrators across the world adhere to defined approaches and provide consistent metadata to enable discovery? How could you keep the barrier to entry as low as possible to encourage participation? We argue that the critical information that enables dataset discovery lies both in metadata, and more importantly, in dataset consumption patterns.

Our solution to these challenges is two-fold:

(1) Rely primarily on dataset ***usage*** rather than just on metadata (the quality and quantity of which vary substantially across communities). This enables ***personalized recommendations*** as a mode of discovery instead of just search, and

(2) Rely on programmatically accessing usage details from existing analytics software, e.g., Google Analytics and Kissmetrics. These analytics are often already utilized by repository websites and the number and varieties of analytics software are significantly fewer compared to the number and variety of repositories, reducing complexity. This also reduces impediments to participation, requiring minimal effort from repository administrators.

In this project, we are evaluating our solution by working with multiple dataset repositories, including the Vermont Monitoring Cooperative, which provides datasets pertaining to forested ecosystems. In particular, we have automatically extracted dataset usage information from VMC's Google Analytics data (with permission) and are computing personalized recommendations based on their user web clicks and downloads. Next steps are primarily about fine-tuning the recommendation algorithms, which are discussed next.

The second significant component of our research is adapting and improving existing algorithms in the recommendation space. While some dataset repositories allow users to rate datasets, not all users provide ratings, and the users who do may be inconsistent in their ratings. This has led us to focus on implicit user feedback. There are algorithms that work with implicit user behavior (website clicks versus explicit star ratings), but this research has focused on entertainment and e-commerce, not scientific datasets. There are few algorithms that consider the path a user takes when considering a dataset, e.g., the inherent difference of web clicks made by a given user while accessing different web pages that correspond to the same dataset. We are studying if we could take advantage of this difference between related web pages with the help of regression techniques. The outcome of our study will be published in conferences or journals and will be disclosed to the NSF community in our project reports.