Geophysical Research Abstracts
Vol. 19, EGU2017-5811, 2017
EGU General Assembly 2017

Enabling Open Research Data Discovery through a Recommender System

Anusuriya Devaraju (1), Gaya Jayasinghe (2), Jens Klump (1), and Dominic Hogan (3)
(1) CSIRO Mineral Resources, Australia (anusuriya.devaraju@csiro.au), (2) CSIRO Data61, Australia, (3) CSIRO Information Management & Technology, Australia

Government agencies, universities, research and nonprofit organizations are increasingly publishing their datasets to promote transparency, induce new research and generate economic value through the development of new products or services. The datasets may be downloaded from various data portals (data repositories) which are general or domain-specific. The Registry of Research Data Repository (re3data.org) lists more than 2500 such data repositories from around the globe. Data portals allow keyword search and faceted navigation to facilitate discovery of research datasets. However, the volume and variety of datasets have made finding relevant datasets more difficult. Common dataset search mechanisms may be time consuming, may produce irrelevant results and are primarily suitable for users who are familiar with the general structure and contents of the respective database. Therefore, we need new approaches to support research data discovery. Recommender systems offer new possibilities for users to find datasets that are relevant to their research interests.
This study presents a recommender system developed for the CSIRO Data Access Portal (DAP, http://data.csiro.au). The datasets hosted on the portal are diverse, published by researchers from 13 business units in the organisation. The goal of the study is not to replace the current search mechanisms on the data portal, but rather to extend the data discovery through an exploratory search, in this case by building a recommender system. We adopted a hybrid recommendation approach, comprising content-based filtering and item-item collaborative filtering. The content-based filtering computes similarities between datasets based on metadata such as title, keywords, descriptions, fields of research, location, contributors, etc. The collaborative filtering utilizes user search behaviour and download patterns derived from the server logs to determine similar datasets. Similarities above are then combined with different degrees of importance (weights) to determine the overall data similarity. We determined the similarity weights based on a survey involving 150 users of the portal. The recommender results for a given dataset are accessible programmatically via a RESTful web service. An offline evaluation involving data users demonstrates the ability of the recommender system to discover relevant and ‘novel’ datasets.