Geophysical Research Abstracts
Vol. 15, EGU2013-6217, 2013
EGU General Assembly 2013

Two-step web-mining approach to study geology/geophysics-related open-source software projects

Knut Behrends and Ronald Conze
Deutsches GeoForschungsZentrum GFZ, Scientific Drilling, Potsdam, Germany (knb@gfz-potsdam.de)

Geology/geophysics is a highly interdisciplinary science, overlapping with, for instance, physics, biology and chemistry. In today's software-intensive work environments, geoscientists often encounter new open-source software from scientific fields that are only remotely related to the own field of expertise. We show how web-mining techniques can help to carry out systematic discovery and evaluation of such software.

In a first step, we downloaded ∼500 abstracts (each consisting of ∼1 kb UTF-8 text) from agu-fm12.abstractcentral.com. This web site hosts the abstracts of all publications presented at AGU Fall Meeting 2012, the world's largest annual geology/geophysics conference. All abstracts belonged to the category "Earth and Space Science Informatics", an interdisciplinary label cross-cutting many disciplines such as "deep biosphere", "atmospheric research", and "mineral physics". Each publication was represented by a highly structured record with ∼20 short data attributes, the largest authorship-record being the unstructured "abstract" field. We processed texts of the abstracts with the statistics software "R" to calculate a corpus and a term-document matrix. Using R package "tm", we applied text-mining techniques to filter data and develop hypotheses about software-development activities happening in various geology/geophysics fields. Analyzing the term-document matrix with basic techniques (e.g., word frequencies, co-occurences, weighting) as well as more complex methods (clustering, classification) several key pieces of information were extracted. For example, text-mining can be used to identify scientists who are also developers of open-source scientific software, and the names of their programming projects and codes can also be identified.

In a second step, based on the intermediate results found by processing the conference-abstracts, any new hypotheses can be tested in another webmining subproject: by merging the dataset with open data from github.com and stackoverflow.com. These popular, developer-centric websites have powerful application-programmer interfaces, and follow an open-data policy. In this regard, these sites offer a web-accessible reservoir of information that can be tapped to study questions such as: which open source software projects are eminent in the various geoscience fields? What are the most popular programming languages? How are they trending? Are there any interesting temporal patterns in committer activities? How large are programming teams and how do they change over time? What free software packages exist in the vast realms of related fields? Does the software from these fields have capabilities that might still be useful to me as a researcher, or can help me perform my work better? Are there any open-source projects that might be commercially interesting?

This evaluation strategy reveals programming projects that tend to be new. As many important legacy codes are not hosted on open-source code-repositories, the presented search method might overlook some older projects.