Large-Scale Discovery of Research Datasets

"… all your communications will be useless to me unless you can propose some practicable way of supplying me with (your) observations …"

Isaac Newton to John Flamsteed in 1675

In the scientific community, the verification of the published research results and evaluation of new research methods, require access to the corresponding research data. Numerous of the research datasets available online can be discovered through open repositories and metasearch engines such as Google Dataset search [1], r3data, and others. Each dataset is typically accompanied by a little metadata, textual description, and a link to the original paper. To accelerate research and to avoid redundant work, two criteria are essential - broad recall and precise selection of as many available relevant datasets as possible. However, existing search engines currently provide only basic keyword and metadata-based search tools, turning the identification of relevant datasets into an investigative adventure.

In this thesis, we make use of the advances in Information Retrieval and Machine Learning for a scalable large-scale discovery of relevant datasets. To this end we: i) research, evaluate and develop a range of similarity metrics based on the full spectrum of available dataset descriptions and references, ii) ingest a large set of available datasets into appropriate indexing structures, capable to efficiently identify similar datasets given a seed dataset, iii) perform a thorough evaluation of the effectiveness of the system.

For the similarity metrics, we will collect many available datasets to train ML models. We will employ WordEmbeddings [2], citation-graphs, and Author-networks [3]. We will use KNN structures to develop a discovery index for broad recall and design simple ranking mechanisms to deliver a focused result set. Furthermore, we will evaluate the effectiveness of the approach in a brief user study.

Resources:

Dataset collections: datasetsearch.research.google.com, zenodo.org, re3data.org, kaggle.com. Citation Networks: crossref.org, opencitations.net, semanticscholar.org. Authors network: dblp.org

References:

[1] Brickley, D. Matthew B., and Natasha Noy "Google Dataset Search: Building a search engine for datasets in an open Web ecosystem." The World Wide Web Conference. 2019. DOI: https://doi.org/10.1145/3308558.3313685

[2] Tshitoyan, V., et al. "Unsupervised word embeddings capture latent knowledge from materials science literature." In Nature 571.7763 (2019): 95-98. DOI: 10.1038/s41586-019-1335-8

[3] Aung, T. T., et al. “Community detection in scientific co-authorship networks using neo4j.” In: 2020 IEEE Conference on Computer Applications (ICCA). IEEE, 2020. S. 1-6. DOI: 10.1109/ICCA49400.2020.9022826

[4] Mohamed Ben Ellefi, et al.: RDF dataset profiling - a survey of features, methods, vocabularies and applications. Semantic Web 9(5): 677-705 (2018) https://content.iospress.com/articles/semantic-web/sw294

Code:

https://github.com/anlausch/scientific-domain-embeddings, https://radimrehurek.com/gensim/
https://github.com/aaalgo/kgraph, https://www.kaggle.com/bkoseoglu/co-authorship-network-analysis

Contact: Dr. Sergej Zerr szerr@uni-bonn.de