Dataset Search and Augmentation

Zhiyu Chen
Lehigh University
USA
zhc415@lehigh.edu

Abstract

Data has become an indispensable part of our life. However, current mainstream commercial search engines do not support specialized functions for dataset search. A dataset usually consists of both metadata and data content. Existing information retrieval models designed for Web search cannot efficiently extract semantic information inside structured datasets, even when they contain textual content. Developing new algorithms for next-generation search engines to efficiently find datasets can benefit data practitioners in their data discovery experience.

In this dissertation, we consider how to effectively perform dataset search and augmentation. We start by providing an end-to-end description of a dataset search engine following the lifecycle of datasets. Our review includes web dataset acquisition techniques, dataset profiling and augmentation methods, and dataset search tasks and corresponding methods. In order to extract datasets from research articles, we present an information extraction framework to determine triples of interest which can be used for academic dataset search. We propose a feature-based method to augment tabular datasets with additional schema labels to help users and systems to better understand the datasets. We develop three methods for tabular dataset search: the first utilizes generated schema labels to enhance the search results; the second adopts pretrained language models to learn matching features; the third models the complex relations in the datasets as one or more graphs and uses graph neural networks to learn representations of queries and tables. To support dataset search in which a query is also a dataset, we propose universal dataset encoders which regard a dataset as a point set so that the encoded dataset representations can be used to search for similar datasets. Extensive experiments across multiple tasks demonstrate the superiority of our proposed methods over the state of the art.

Awarded by: Lehigh University, Bethlehem, USA **on** 10 May 2022.
Supervised by: Brian D. Davison.
Available at: https://github.com/Zhiyu-Chen/Dissertation/blob/main/Dissertation_Dataset_Search.pdf.

Selected Publications

Zhiyu Chen, Haiyan Jia, Jeff Heflin, and Brian D. Davison. Generating schema labels through dataset content analysis. In *Companion Proceedings of The Web Conference 2018*, pages 1515–1522, 2018.

Zhiyu Chen, Haiyan Jia, Jeff Heflin, and Brian D. Davison. Leveraging schema labels to enhance dataset search. In *European Conference on Information Retrieval*, pages 267–280. Springer, 2020a.

Zhiyu Chen, Mohamed Trabelsi, Jeff Heflin, Yinan Xu, and Brian D. Davison. Table search using a deep contextualized language model. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 589–598, 2020b.

Zhiyu Chen, Mohamed Trabelsi, Jeff Heflin, Dawei Yin, and Brian D. Davison. MGNETS: Multi-graph neural networks for table search. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 2945–2949, 2021a.

Zhiyu Chen, Shuo Zhang, and Brian D. Davison. WTR: A test collection for web table retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2514–2520, 2021b.