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 aug-
mentation 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 lan-
guage 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.
Selected Publications


