The INDE lab Team

INDELab started Nov. 5, 2018
Research Topics at INDE lab

• Design systems to support people in working with data from diverse sources
• Address problems related to the preparation, management, and integration of data

• Automated Knowledge Graph Construction
  (e.g., predicting and adding new links in datasets such as Wikidata based on text)

• Data Search & Reuse
  (e.g., studies on GitHub hosted data; algorithms for making data FAIR)

• Data Management for Machine Learning
  (e.g., scalable concept drift detection for ML training data, integrated in AWS SageMaker Model Monitor; using data provenance for ML debugging)

• Causality-Inspired Machine Learning
  (e.g., using ideas from causal inference to improve the robustness and generalization of ML algorithms, especially in cases of distribution shift; domain adaptation)
Highlights for the DBDBD audience

About GitTables  
gittables.github.io

GitTables is a dataset of currently 1.7M relational tables extracted from CSV files in GitHub. Our continuing curration aims at growing the dataset to at least 20M tables. Table columns in GitTables have been annotated with more than 2k different semantic types from Schema.org and DBpedia. Our column annotations consist of semantic types, hierarchical relations, range types and descriptions.

The high-level pipeline in Figure 1 illustrates how GitTables was created.

Figure 1: high-level pipeline of the process of constructing GitTables.

Analysis of how data reuse and search in the wild

Other highlights:

- FYI:
  - Hiring: assistant prof. in data management methodologies
  - Special Issue JWS on Knowledge Engineering for large scale knowledge graphs