Database Schemas "in-the-wild": What Can We Learn from a Large Corpus of Relational Database Schemas?

Till Döhmen
University of Amsterdam

1 The Case for a Large-Scale Schema Dataset

Tabular data collections such as GitTables [3] are important sources of real-world tabular data, among others, providing training data for table representation learning approaches that push the state-of-the-art for problems like semantic annotation, data imputation, and automated error detection [4]. However, such datasets are limited to individual tables and do not contain schema information about database constraints (uniqueness, not nulls, etc.), and relationships to other tables. We expect that *schema* representation learning, similar to *table* representation learning, can lead to significant advances in existing problems like foreign key detection, and constraint prediction. As real-world database schemas are hard to come by - the largest public repository of databases we are aware of contains about 150 relational databases - there is a need in the community for a new dataset. We hence created GitSchemas [1] - a large corpus of database schema information extracted from *SQL* scripts from public code repositories, that contains highly accurate schema information for, as of now, more than 165k schemas, 1M+ tables (including column names, data types and database constraints), and almost 700k foreign key relationships. We believe that schema information alone (without data) at this scale will be suitable for solving a variety of interesting and relevant problems, apart from presenting an opportunity to learn more about how database systems are used in practice.

2 Current State & Ongoing Work

Compared to the initial release, we could increase the number of extracted schemas by 250%, by using a new SQL parser, and we plan to further increase the number with a specifically developed, more robust, and dialect-independent SQL parsing approach. In addition to demonstrating the usefulness of the dataset for automatic data augmentation in ML pipelines [1], we conducted first experiments with recognizing CSV headers. Detecting the correct header in CSV files is a relevant problem for large-scale data integration and particularly challenging for heuristics-based approaches, as it sometimes requires a semantic understanding of the file content [2]. Using a large language model (LLM), fine-tuned with GitSchemas, achieved an unprecedented accuracy of 99.98% in detecting the correct header rows on a balanced sample of header and data rows. The semantic understanding of schemas and tables is also expected to be a *key driver* for improving the accuracy of existing foreign key detection and constraint prediction approaches. First experiments showed encouraging results. Finally, we plan to explore ways to enable new use cases such as schema completion and assisted schema design with Transformer-based schema representation models, trained on GitSchemas.

References

[1] T. Döhmen, M. Hulsebos, C. Beecks, and S. Schelter. Gitschemas: A dataset for automating relational data preparation tasks. In *Workshop on Databases and Machine Learning (DBML)*. IEEE, 2022.

[2] T. Döhmen, H. Mühleisen, and P. Boncz. Multi-hypothesis csv parsing. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, pages 1–12, 2017.

[3] M. Hulsebos, Ç. Demiralp, and P. Groth. Gittables: A large-scale corpus of relational tables. *arXiv preprint arXiv:2106.07258*, 2021.

[4] D. Vos, T. Döhmen, and S. Schelter. Towards parameter-efficient automation of data wrangling tasks with prefix-tuning. In *NeurIPS, Table Representation Learning Workshop*, 2022.