Informationssysteme
Prof. Dr. Felix Naumann
Masterprojekt WS 2020/21

Discovering Change Dependencies

Change Exploration

It is a fact that data and metadata of many public datasets change. When we monitor real-world datasets, we observe a large number of many different kinds of changes. The changes range from small-scale updates in a few tuples to large-scale schema changes, such as column deletions or bulk-inserts.

While it is easy to recognize that the data changes, it is much more difficult to determine why an individual change was observed. Often, there are hidden connections, possibly due to unknown integrity constraints or unknown automated processes that update the data. In the context of the Janus project (www.IANVS.org) we want to develop a scalable algorithm to automatically discover such hidden change dependencies on large datasets. Our goal is to prepare a submission to a top database conference together.

Change Dependencies

In this project, we want to mine change dependencies on a large scale. Change dependencies can come in many different forms of rules, such as

- Whenever there is a change in A, there is a change in B.
- Whenever there are *x* changes in A, there are at least *y* changes in B.
- Whenever there is a change in A, there is a change in B within *z* days.

These rules can shall be discovered on different levels of granularity: table, attribute, tuple or field. They can be mined within a single table with all its versions or across multiple tables from the same data source. We will evaluate the efficacy of known association rule mining algorithms with their various temporal extensions (rules in temporal) and sequential extensions (rules in sequences), adapt them if possible or develop a new approach if needed.

Change dependencies serve many different purposes. For individual values, they help determine the up-to-dateness of the data. For entities, the change dependencies reveal their relationship in the real world. For tables, the dependencies might uncover the data generation process.

Available Data

As a data basis for this project, we have collected (and are still collecting) over ten months of daily snapshots of structured US open-government data published on Socrata (https://dev.socrata.com/). The datasets are relational tables. A single snapshot typically encompasses around 40,000 tables and takes up roughly 12GB as uncompressed .json files. In total, we thus have more than 3.5TB of data to examine.

Programming experience is a must; we plan to use Java and Scala as programming languages in a possibly distributed setting. Experiences in databases and data mining are helpful but not required – the first weeks will be dedicated to getting up to speed in this research area.

For questions, please contact Leon Bornemann, Tobias Bleifuß, or Prof. Dr. Felix Naumann
leon.bornemann@hpi.de, tobias.bleifuss@hpi.de, felix.naumann@hpi.de