Informationssysteme
Prof. Dr. Felix Naumann
Masterprojekt WS 2016/17

Profiling Dynamic Data

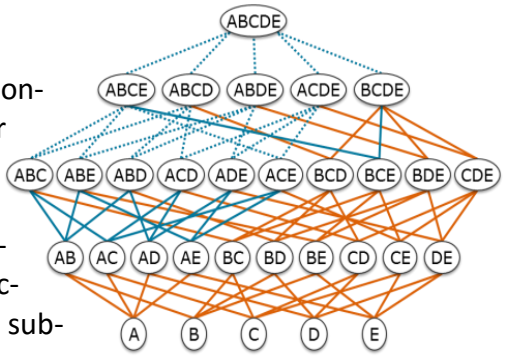
Data Profiling

Data profiling is the process of examining a given dataset for its structural metadata. These metadata include simple statistics, such as value distributions, which are easy to compute. More complex metadata usually involve multiple columns, which makes them significantly harder to discover. Particularly important multi-column metadata types are *unique column combinations*, *inclusion dependencies*, and *functional dependencies*. Because of their many use-cases, such as schema matching, data cleansing, query optimization, or data exploration, data profiling is a frequent activity for any IT professional.

Dynamic Data

Real-world data constantly change in daily business, thereby rendering existing metadata out of date. To keep up with the changes, profiled metadata must be updated continuously or at least frequently. So far, database research has proposed many profiling algorithms to efficiently discover certain types of metadata for fixed datasets, but re-executing these algorithms for every change is a too costly and time-consuming process. In consequence, new profiling algorithms are needed to efficiently maintain the metadata of such ever-changing datasets.

Project Goals



Given a relational dataset and its metadata, our objective is to monitor insert, update, and delete operations on the dataset in order to update the metadata accordingly. The metadata-updates need to be fast enough to cope with possibly high change rates of the data. While incremental metadata updates are an algorithmic challenge for every type of metadata, we shall focus on functional dependencies (FDs). The project consists of the following subgoals:

- **Literature research:** Review different profiling algorithms from previous research and consider their suitability for making them incremental.
- **Algorithm development:** Develop a novel incremental FD profiling algorithm that includes finding appropriate index structures and clever look-up strategies.
- **Evaluation:** Evaluate the correctness and efficiency, i.e., throughput, of our solution on different real-world datasets in the incremental setting.
- **Presentations:** In addition to regular project meetings, we will have a midterm and a final presentation to gather feedback from the research community.
- **Paper Preparation:** We conclude our work in a submission-ready 12-page paper, describing incremental data profiling, our algorithm, and our experimental results.

With the HPI Metanome data profiling framework (www.metanome.de), we have access to many existing profiling algorithms and can probably reuse previous work for our new task. Ultimately, we aim to publish our results at a major scientific conference.

Contact

Thorsten Papenbrock, Sebastian Kruse, Prof. Dr. Felix Naumann
thorsten.papenbrock@hpi.de, sebastian.kruse@hpi.de, felix.naumann@hpi.de