Name: Leon Bornemann-Paulus
Thesis Title: Profiling Temporal Data

Abstract:

Data profiling is a research area that studies how statistics and metadata can be automatically and efficiently extracted from datasets. While various previous work exists in this area, it mostly deals with the discovery of metadata on snapshots of datasets. In practice, sometimes not only the current snapshot, but also older versions of the dataset are available. Older versions of the same dataset may look significantly different, because datasets may change significantly as time progresses. For example, attributes or tuples may be added or deleted, field values may be changed, schema elements can be renamed. All of these changes are data themselves, namely *change data*, which describe how the dataset has evolved. The availability of such change data makes it possible to not only study the current state of the dataset, but also past versions of specific data elements, such as individual fields or attributes, as well as entire relations, leading to temporal datasets.
The change data contained in temporal datasets is a potential source of information and insights that is unusable for traditional data profiling algorithms.

This thesis presents a framework to model change data and subsequently studies how data profiling tasks can benefit from it. The three main contributions of this thesis are three data profiling tasks, that utilize change data to discover metadata at different granularities. The first metadata that is discussed is the natural key of a relation, which is a metadata that enhances the knowledge about an entire relation. The problem of natural key discovery is modeled as a classification problem for which features are engineered from both snapshot and temporal data. The second metadata are temporal inclusion dependencies, which provide knowledge about containment-relationships between individual attributes. The problem of finding genuine inclusion dependencies in temporal data is solved by employing several relaxation criteria that allow for errors in the data, whereas the efficient search is solved by employing several index structures based on Bloom filters. Lastly, the third metadata are role matchings, which are equality constraints on fields and thus a very fine-grained metadata. Similarly to duplicate detection, the problem of discovering role matchings uses a blocking stage and a matching stage. For the blocking stage, this thesis presents and evaluates various methods, whereas for the matching stage, a large language model is employed.
When evaluating the discovery of metadata, this thesis is mainly concerned with the quality of the solutions, but as efficiency and feasibility are also required, runtime and scalability experiments are presented as well.

The experiments in this thesis show that change data is helpful to increase the accuracy with which natural keys and genuine inclusion dependencies can be discovered.
Furthermore, having change data available is a prerequisite for discovering role matchings, showing that change data can not only improve the results for known data profiling tasks, but can also enable the discovery of new types of metadata, that have not been considered in snapshot data.