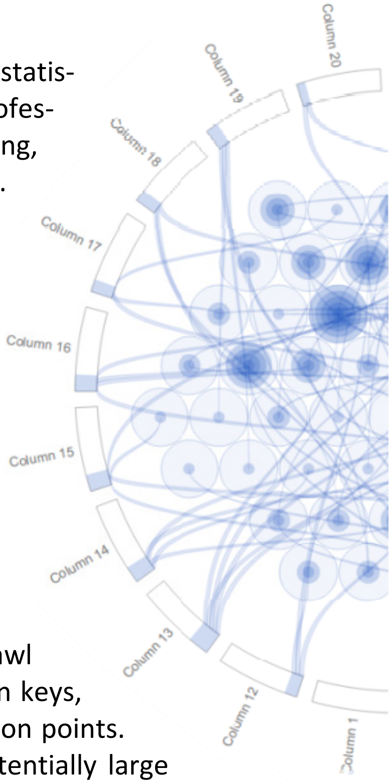
Informationssysteme
Prof. Dr. Felix Naumann
Masterprojekt WS 2014/15

Metadata Trawling – Interpreting Data Profiling Results

Data Profiling

Data profiling is the process of examining a given dataset for metadata and statistics about that data. It is an important and frequent activity for any IT professional or researcher and serves various use-cases, such as schema matching, database reverse engineering, query optimization, and data exploration. Among the simpler results are statistics, such as the number of null values and distinct values in a column, a column's data type, or the most frequent patterns in the columns' values. Metadata that are more difficult to compute usually involve multiple columns, such as inclusion dependencies (INDs), unique column combinations (UCCs), and functional dependencies (FDs).

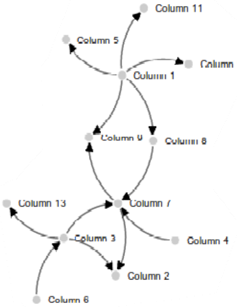


The Metadata Ocean

Data profiling can produce vast amounts of metadata. Typical numbers of unique column combinations are in the thousands, functional dependencies abound, … Once discovered, the metadata itself becomes a puzzle to solve: Experts need to analyze and interpret the discovered facts to gather useful insights into the structure of the datasets at hand. That is, they trawl this ocean of metadata. From INDs, for instance, we want to derive foreign keys, from UCCs we extract primary keys, and from FDs we identify decomposition points. This process requires novel techniques that automatically analyze the potentially large set of discovered metadata.

Project Goals

The objective of metadata trawling is to find, implement, and evaluate new techniques for the interpretation of data profiling results. For the automatic or semi-automatic metadata analysis several techniques might be useful:



- **Ranking:** Allows experts to focus on the most important/useful results.
- **Visualization:** Uncover new properties of the data and their dependencies.
- **Filtering:** Remove irrelevant or confusing information.
- **Clustering:** Combine results into groups of logically coherent dependencies.

In this project we will extend the HPI Metanome framework (www.metanome.de) with *data profiling result management* capabilities. For the experiments, we will use real-world datasets and their metadata. Besides the implementation and evaluation of the ideas developed in this project, we aim to prepare a demonstration paper for a major scientific conference.

Contact

Prof. Dr. Felix Naumann:
felix.naumann@hpi.uni-potsdam.de
Thorsten Papenbrock:
thorsten.papenbrock@hpi.uni-potsdam.de