Informationssysteme
Prof. Dr. Felix Naumann
Masterprojekt Sommer 2015

Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

Approximate Data Profiling

Data Profiling

It is important to know and understand a dataset's structure and its inner dependencies in order to process the data or to gain knowledge from it. Such data about data is called metadata and the discipline of analyzing datasets for their metadata is called data profiling. Among various types of metadata, **functional dependencies** (FDs) are among the most important. They serve a wide range of data management tasks, such as data integration, schema normalization, or data cleansing.

Approximate Discovery

Research of the past decades has already developed several discovery algorithms for FDs, but none are applicable for larger real-world datasets, because either their runtime or their memory consumption grows exponentially in the number of attributes, and thus quickly exceeds the available resources. In this project we will develop **approximate discovery algorithms** for FDs and evaluate their runtime and result quality.

Approximate algorithms trade off the result correctness or completeness for efficiency, e.g., by employing sampling and/or heuristics. Thus, when applying such techniques, we aim to reduce the execution time significantly. This can be achieved by relaxing the requirements of the result in two different ways:

1. **Completeness**: Some FDs may be missing (relaxed recall)
   - Limited Size: Find only FDs that involve only a limited amount of attributes.
   - Unchecked Minimality: Find potentially non-minimal FDs.
   - Progressive Search: Find as many FDs as possible in a given time.
   - Focused Search: Find only FDs that fulfill a certain "interestingness" criterion.
2. **Correctness**: Some FDs may be incorrect (relaxed precision)
   - Subset Search: Search FD candidates only on a subset of the data.
   - Approximate Checks: Find FDs using heuristic data structures, e.g., Bloom filters.

For both relaxation aspects we need to measure their influence on the results' quality in order to quantify the trade-off between runtime savings and result quality. A possible result could, for instance, be an algorithm that is at least 100 times faster than all exact algorithms while reducing precision/recall by only 5%.

Informationssysteme
Prof. Dr. Felix Naumann
Masterprojekt Sommer 2015

Project Goals

1. Develop one (or more) approximate FD algorithms that relax the requirements for the result in different ways. The algorithms should achieve possibly fast execution times while sacrificing only little precision/recall.
2. To compare the algorithms, we need to formalize effectiveness, i.e., the trade-off between runtime and quality.
3. Evaluate the developed algorithms on different real-world and synthetic datasets for efficiency and effectiveness.
4. Prepare a submission-ready paper about approximate FD discovery aiming to submit to a scientific venue. In the paper, we categorize different types of approximate algorithms, describe our implementation(s) and report on our analysis. Finally, we aim to generalize our ideas on FD discovery for other types of metadata.

In this project we will extend the HPI Metanome framework (www.metanome.de). The algorithms and result management techniques should therefore be integrated into this project. A prerequisite of this master-project is the lecture “Data profiling and data cleansing”.

Contact

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