Proposed Master's Thesis Project

Discovery of Attribute Relationships with Data-Driven Sampling

Thesis Information

Type of Thesis: Method development, implementation, experimental evaluation

Thesis Start Date: 01. January at the earliest

Duration: Six months, full-time

Language: English or German

Prerequisites: Solid programming skills, knowledge of statistics is highly desirable, prior Python experience is helpful but not required.

Motivation

Data profiling provides statistics and metadata that help characterize datasets and support data cleaning. Beyond describing individual attributes, relationships between attributes can provide useful context for subsequent processing like cleaning decisions.

Consider an online shop whose customer data are processed in daily batches. An upstream schema change introduces a new categorical attribute, `customer_type`, distinguishing private and business customers. The existing cleaning pipeline replaces missing values in the numerical attribute `annual_spending` with a global mean. If spending differs systematically between the new customer groups, detecting and quantifying this relationship could support the subsequent use of group-specific imputation values. The profiler would capture the relationship and the corresponding group statistics. These statistics could include the number of observed spending values, their mean, and their variation within each customer group. They would describe how spending differs between private and business customers and provide the group-specific values that a later cleaning component could use for imputation. Storing this information in the data profile would preserve it even after the raw batch data have been discarded. Adapting the cleaning operation is a separate task.

This example motivates a more general profiling capability: users should be able to configure attribute pairs or sets of attributes and obtain quantitative descriptions of their pairwise relationships. This requires identifying which existing detection methods provide a useful set of complementary capabilities across different attribute types. For large datasets, profile-guided sampling may additionally reduce the computational cost of applying these methods while preserving the quality of their results.

Problem

The existing Python-based profiler *EDaM* produces data profiles, but does not currently provide statistics describing relationships between attribute pairs. Profiling takes place on uncleaned batch data. Raw records are available for the current batch, whereas only aggregated profiles are retained for historical batches.

The thesis focuses on pairwise relationships, initially considering numerical, categorical, and free-text attributes. Selecting a feasible set of relationship types, detection methods is part of the work. Additional types, such as dates, may be considered where appropriate. The challenge is to identify and justify a complementary set of established methods suitable for this profiling context, considering the relationships they capture, their assumptions, computational cost, and robustness to data errors. A further challenge is to investigate how profile-guided sampling can make their application more efficient.

Research question: Which combination of established methods is suitable for discovering and quantitatively characterizing pairwise attribute relationships in uncleaned batch data, and to what extent can profile-guided sampling reduce computational cost while maintaining detection quality and robustness to data errors?

Goal

The goal is to develop and experimentally evaluate a modular extension of the existing profiler. The extension should reuse available profiling functionality where appropriate and remain independently testable. The work should provide a justified selection of complementary relationship-detection methods from the literature and a profile-guided sampling strategy for their efficient application. Both the suitability of the selected methods and the benefits and limitations of sampling are to be evaluated.

The work comprises the following tasks:

1. Review and categorize existing methods for detecting and quantifying relationships across the chosen attribute types. Compare their coverage, assumptions, redundancy, computational requirements, and handling of missing or erroneous values. Justify a feasible, complementary selection for the profiler.

2. Design a sampling strategy informed by profile metadata. Possible approaches include adapting sample sizes or identifying candidate relationships on a small sample and selectively validating them using additional data. The strategy and its stopping or validation criteria are to be developed within the thesis.

3. Implement the extension in Python, extending the profiler's configuration to support explicit attribute pairs and pairwise comparisons within selected attribute sets.

4. Extend the profile output to retain quantitative descriptions of detected relationships, including method-specific statistics, observation counts, and relevant sampling information. The stored information should support interpretation and potential subsequent use after raw data have been discarded.

5. Evaluate the approach primarily on available real datasets. Assess reference-data suitability and supplement the evaluation with synthetic data containing known relationships where necessary. Evaluate the selected detection methods across different relationship and attribute types, including their complementary strengths and limitations. Compare full-data analysis, simple random sampling, and the proposed strategy across sample sizes, controlled error types and levels, and data scales.

Evaluation should address missed and incorrectly detected relationships where ground truth is available, accuracy of quantitative estimates, runtime, and memory consumption. Runtime measurements should include sampling and any additional profiling overhead.

Expected outcomes are a structured comparison and justified selection of relationship-detection methods, the documented profiler extension, a profile-guided sampling method, and reproducible experiments. The results should establish which methods are suitable under which conditions and when sampling is beneficial or insufficient. The profiler repository, documentation, configuration examples, profile examples, and real datasets are available, together with supervision by a researcher familiar with its implementation. The project is particularly suitable for students in Data Science or Computer Science with an interest in statistics and data engineering.

Automatic orchestration following schema changes, adaptation of cleaning operators, and detection of data drift or semantic shifts are outside the scope. Relationships involving three or more attributes jointly are also excluded.

Next steps

If you are interested in the proposed thesis project, please provide the following information and attachments:

- A short statement explaining your motivation for applying to this project.
- Relevant skills and knowledge you possess that would benefit the project.
- Your current transcript of grades.
- An estimated timeframe for completing the project.

References

[1] Ziawasch Abedjan, Lukasz Golab, and Felix Naumann. Profiling relational data: a survey. *VLDB J.*, 24(4):557–581, 2015.

[2] Kevin M. Kramer, Valerie Restat, and Uta Störl. Evolving gracefully: Building robust and self-adaptive data cleaning pipelines for schema evolution and uncertainty. In *VLDB Workshops*. VLDB.org, 2025.

[3] Thorsten Papenbrock and Felix Naumann. A hybrid approach to functional dependency discovery. In *SIGMOD Conference*, pages 821–833. ACM, 2016.