Visualizing Session-Based Data Profiles (Abstract)

Andreas M. Wahl
FAU Erlangen-Nürnberg

Christian Sauerhammer
FAU Erlangen-Nürnberg

Richard Lenz
FAU Erlangen-Nürnberg

1 INTRODUCTION

In common analysis scenarios, data scientists face the challenge of discovering and combining data sources. Within an analysis session, they usually create a series of SQL queries building on each other to iteratively derive results. However, due to a lack of familiarity with data sources or the complexity of query results, it can be difficult to decide on the next query iteration solely based on the results of the last one.

Hence, we address the following challenge: *How can we assist data scientists with the iterative formulation of complex SQL queries through session-based data profile visualizations without disrupting established analysis workflows?*

Building upon existing data profiling tools (e.g. [2]) and recent research on visualization recommendation (e.g. [3]), we introduce *OCEANProfile* [4], a framework for session-based profiling of query results.

2 QUERY-DRIVEN DATA PROFILING

OCEANProfile seamlessly integrates with existing data analysis tools and workflows (Fig. 1). We provide a novel JDBC *Proxy Driver* containing proxy logic for JDBC statements and result sets. Data scientists can reference native JDBC drivers in the connection string to connect to data sources. Parallel to query processing, queries and result rows are streamed to the *OCEANProfile Server* using Apache Kafka to minimize the local performance impact of data profiling.

Our *Profiling Engine* processes incoming queries and result rows. We provide a plugin interface to embed existing algorithm implementations from other research projects. The *Profile Ranker* analyzes and ranks result profiles.

Data scientists can subsequently interact with the *OCEANProfile App*, a web-based companion interface which launches next to the analysis tool from which queries are created. The current query is displayed, along with a graph-based visualization of the current session. A query history allows inspecting previous profiling results. The app displays all session-based visualizations for the generated data profiles.

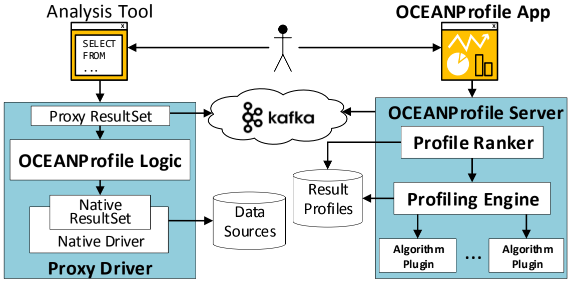

Figure 1: Framework Architecture

3 SESSION-BASED VISUALIZATIONS

For each numerical result column, statistics are visualized using a series of box plots, containing one box plot for each query in the session. The absolute and relative amounts of NULLs and distinct values are illustrated using multi-series line charts. For all non-numerical columns, we show color-coded tables of minimum and maximum values as well as the most frequent items. Unique column combinations (UCCs) are illustrated using bar charts. Each bar indicates in how many queries of the current session a UCC has been valid. UCCs, which became valid or invalid in the current query, are highlighted. OCEANProfile also uncovers functional dependencies, inclusion dependencies, multi-valued dependencies and order dependencies from the query results. These are visualized using heat maps. The color of each column combination reflects the amount of queries where these attributes have been part of the determinant or dependent set. The corresponding constraints are accessible through interactive tooltips. OCEANProfile also features a graph-based representation of normalized query result schemas and a prefix-tree visualization of denial constraints.

Visualizations are arranged according to their usefulness. We consider the usefulness of a result profile to be positively correlated with its deviation from a chosen reference profile. For purely numerical result profiles, we define this deviation as the percentage change to the reference profile. For result profiles consisting of sets of entities (e.g. constraints or schema elements), the deviation is determined by the dissimilarity between the sets according to the Jaccard distance.

4 RESULTS

OCEANProfile demonstrates the feasibility of a minimally-intrusive approach to support data scientists. Session-based visualizations of query results simplify targeted data exploration, data integration and plausibility assessment. A quantitative evaluation based on realistic benchmarks (e.g. [1]) and real-world workloads shows that OCEANProfile generates data profiles and their visualizations fast enough to allow its efficient usage through interactive analysis tools.

REFERENCES

[1] Eichmann et al. 2018. IDEBench: A Benchmark for Interactive Data Exploration. *CoRR* abs/1804.02593 (2018).

[2] Papenbrock et al. 2015. Data Profiling with Metanome. *PVLDB* 8, 12 (2015).

[3] Qin et al. 2018. DeepEye: An automatic big data visualization framework. *Big Data Mining and Analytics* 1, 1 (2018).

[4] Wahl et al. 2018. Query-Driven Data Profiling with OCEANProfile. In *BIRTE'18*.