Interactive and Iterative Automated Data Wrangling in Big Data Applications

Abstract

Identifying data, extracting data and data cleansing is time consuming and may be technically challenging. Data engineers may nearly spend 80% of their time on this process rather than analyzing or using the data. Therefore user centered and more cost-effective workflows are needed. These workflows should allow automatic data profiling, extraction and cleansing processes which take user preferences into account and also give the user the freedom to explore data in an iterative manner. Also, the processes should improve in each iteration by using the feedback provided by the user.

In each iteration, users have the chance to review the processed data. For example, data can be visualized in data analysis tools (e.g. Tableau or Qlik). Additionally, an interactive interface (Wrangler) can be used to visualize distributions and inconsistencies and provide recommendations to improve the processes. The user then has the possibility to provide feedback, data context, user context and to make interactive modifications to the processes. The resulting processes can be exported into big data applications.

Iterative Improvement

Research Topic:
How to improve the processes without a lot of data provided by users?

Research Topic:
How can it be described in a way that users can formulate easily but is also effective for data processes?

Research Topic:
How to do effective (explorative) quality analysis in big data sets?

Validation by Users
Detection of Outliers / missing values / incorrect types

Wrangler visualizes distributions/inconsistencies & generates recommendations

Interaction

Data Context
Definition of desired features (target schema) and assignment of data context (e.g. reference data or properties) to features.

Feedback
Annotation of result data as correct or incorrect.

User Context
Relative importance of different features.

Manual Changes
Manual creation of data processes (also accepted recommendation by the wrangler)

necessary in first iteration

New Generated Processes

Resulting Data

data sample

export processes

Data Sources

Storage
Files
Api's
...

Data Extraction

Data Integration

E.g. parts could be exported into Apache Flink/Spark

Data Analysis

E.g.: tableau, Qlik

Storing Data

E.g.: amazon REDSHIFT

Victor Künstler

Master Student
Hasso Plattner Institute, Potsdam, Germany

E-Mail: victor.kuenstler@hpi.de

Sources:

- Konstantinou, Nikolaos, et al. "The VADA architecture for cost-effective data wrangling." Proceedings of the 2017 ACM International Conference on Management of Data. ACM, 2017.
- Furche, Tim, et al. "Data Wrangling for Big Data: Challenges and Opportunities." EDBT. 2016.
- Kandel, Sean, et al. "Wrangler: Interactive visual specification of data transformation scripts." Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 2011.
- Verbruggen, Gust, and Luc De Raedt. "Towards automated relational data wrangling." Proceedings of AutoML 2017@ ECML-PKDD: Automatic selection, configuration and composition of machine learning algorithms. 2017.

HPI Hasso Plattner Institut