Towards Automated Data Cleaning Workflows

Mohammad Mahdavi
TU Berlin
mahdavilahijani@tu-berlin.de

Felix Neutatz
DFKI
felix.neutatz@dfki.de

Larysa Visengeriyeva
TU Berlin
larysa.visengeriyeva@campus.tu-berlin.de

Ziawasch Abedjan
TU Berlin
abedjan@tu-berlin.de

Motivation

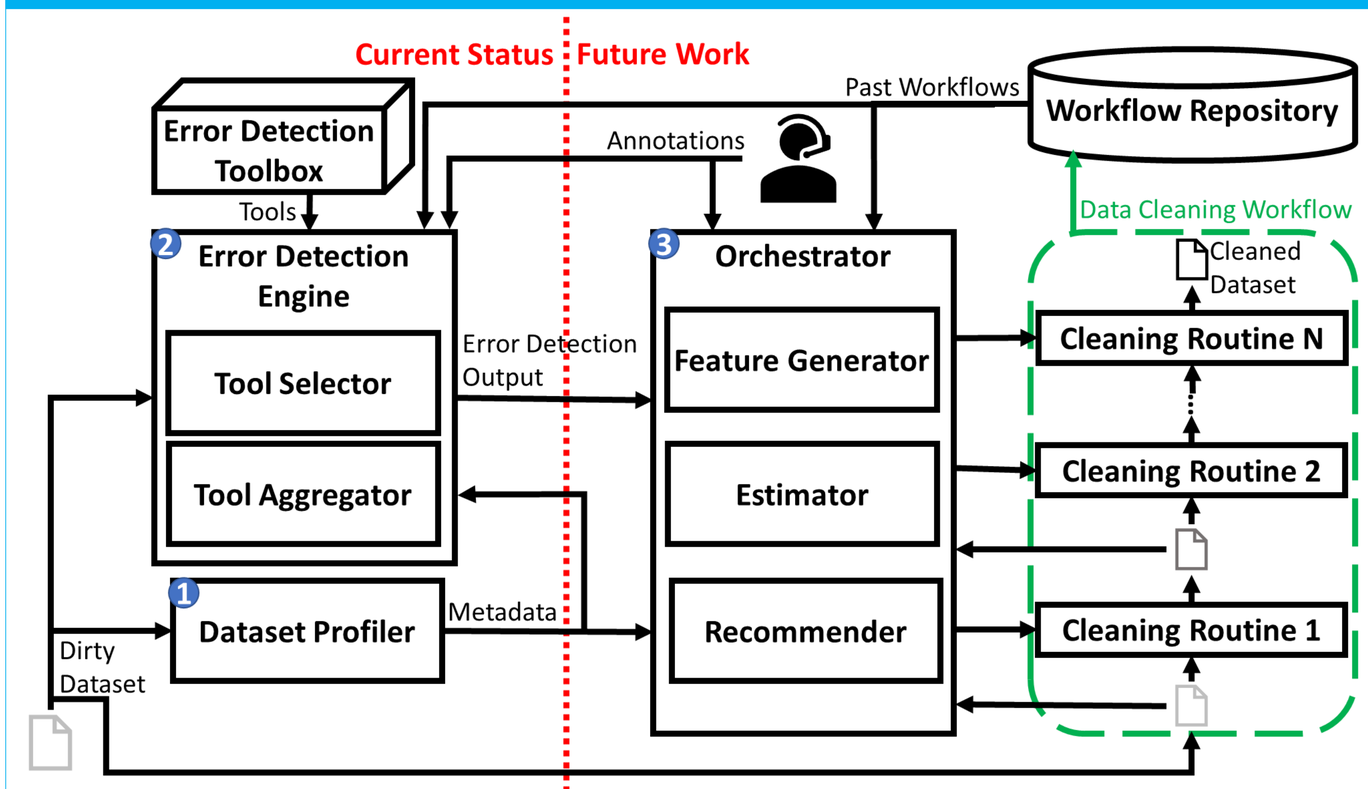
- Research has provided a variety of data cleaning tools [1]
  - Pattern-based [6]
  - Rule-based [7]
  - Statistical [8]
- But, there are still challenges in applying these tools
  - No one-size-fits-all solution
  - Iterative data cleaning
  - Trial-and-error parametrization

Research Question

- How can we leverage machine learning and data profiling techniques to automatically build data cleaning workflows?
  - How can we featurize data values to explain the context of a data error?
  - How can we capture similarities of data cleaning tasks to assess the effectiveness of each tool on a new dataset?
  - How can we aggregate the results of stand-alone cleaning strategies in a holistic manner?

We need a workflow orchestrator that learns from previous tasks to propose promising data cleaning workflows for a new dataset.

Architecture



1) Dataset profiler
   - Generates metadata to describe data quality problems of datasets
2) Error detection engine
   - Leverages the metadata to compare the similarity of datasets
   - Selects and aggregates the promising error detection tools
3) Orchestrator
   - Leverages error detection results and metadata to generate dataset-specific cleaning workflows

Current Status and System Artifacts

- **MDED**, a system that learns to aggregate error detection strategies via metadata [3]
- **REDS**, a system that estimates the performance of error detection strategies via metadata [2]
- **ED2**, an active learning-driven error detection system [4]
- **Raha**, a configuration-free error detection system to detect data errors holistically [5]

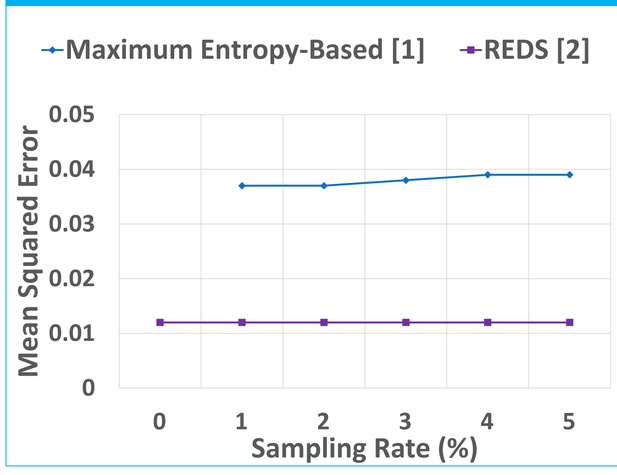
http://bit.ly/systems-aggregation

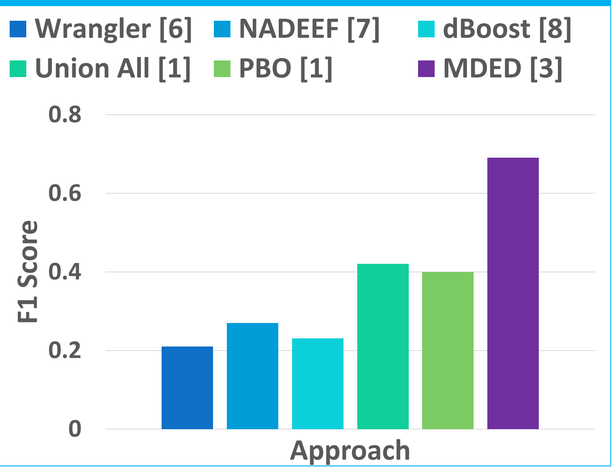
https://github.com/bigdama/reds

http://bit.ly/2mjyiTO

https://github.com/BigDaMa/raha

Experimental Results





References

[1] Ziawasch Abedjan et al. 2016. Detecting data errors: Where are we and what needs to be done? PVLDB 9, 12, 993–1004.
[2] Mohammad Mahdavi et al. 2019. REDS: Estimating the performance of error detection strategies based on dirtiness profiles. SSDBM, 193–196.
[3] Larysa Visengeriyeva et al. 2018. Metadata-driven error detection. SSDBM, 1–12.
[4] Felix Neutatz et al. 2019. ED2: A case for active learning in error detection. CIKM.
[5] Mohammad Mahdavi et al. 2019. Raha: A configuration-free error detection system. SIGMOD, 865–882.
[6] Sean Kandel et al. 2011. Wrangler: Interactive visual specification of data transformation scripts. SIGCHI, 3363–3372.
[7] Michele Dallachiesa et al. 2013. NADEEF: A commodity data cleaning system. SIGMOD, 541–552.
[8] Clement Pit-Claudel et al. 2016. Outlier detection in heterogeneous datasets using automatic tuple expansion. Technical Report. CSAIL, MIT.

Acknowledgement

This project has been supported by the following three grants: The German Research Foundation (DFG) under grant agreement 387872445, the German Federal Ministry of Education and Research as BBDC II (01IS18025A), and the German Federal Ministry of Transport and Digital Infrastructure as Day stream (19F2013).