LOKI

Poonam Kumari
Supervised by Dr. Oliver Kennedy
State University of New York at Buffalo, Buffalo, NY, USA
{poonamku,okennedy}@buffalo.edu

ABSTRACT

It has become very easy to obtain a large dataset for experimental analysis or personal use. But most of these datasets are unlabeled or poorly labeled. Absence of labels, leads to difficulty in accessing the data.

We propose the design of a system LOKI, which would serve as a knowledge base for storing column-naming heuristics, as well as an interactive tool: the LOKI editor for populating the knowledge-base.

Poonam:paraphrase, taken from HILDA paper

The LOKI editor primes the knowledge base by learning from example data (e.g., from open data portals), and assists domain experts in reviewing and refining the resulting heuristic naming schemes. We identify specific issues arising from training and show how the LOKI editor streamlines the process of manually repairing these issues.

PVLDB Reference Format:
Ben Trovato, G. K. M. Tobin, Lars Thørväld, Lawrence P. Leipuner, Sean Fogarty, Charles Palmer, John Smith, Julius P. Kumquat, and Ahmet Sacan. A Sample Proceedings of the VLDB Endowment Paper in LaTeX Format. *PVLDB*, 11 (8): xxxx-yyyy, 2018.
DOI: https://doi.org/TBD

1. MOTIVATION

Big datasets are available in abundance which are being used by data scientists for and database community for research purpose. But we cannot use these datasets in their raw form, since they might be missing labels, values, etc. The most common issue faced while using these datasets is unlabeled or poorly labeled data. Without a system in place to label the data, user can perform following operations (1) Guess and label based on the data provided. (2) Write a script to automate the process of guessing. Case 1 would prove to be hectic if we have large data in hand, and is error prone. In case 2, same script might not work for other datasets, user would need to modify it. The other problem with case 2 is related to the documentation of the script. Once the initial script writer leaves without any documentation, the next user has to put in a lot of effort in reverse-engineering it.

Poonam:rewrite, taken from HILDA paper

We propose to end the suffering with Label Once, and Keep It (LOKI), a data-ingest middleware for incremental, re-usable schema recovery. When a user first points LOKI at a new tabular data set, LOKI proposes a schema for it. It then collects feedback, both learning and also preserving schema metadata for later use. In short, LOKI will allow users to assemble schemas on-demand, both (re-)discovering and incrementally refining schema definitions in response to changing data needs.

1.1 Terms

- Domain:
- Concept:
- Unit:
- Name:
- Signature:
- Column:

Signature consists of

1. Domain
2. Range
3. Distribution
4. Set of values
5. type

1.2 Research Questions

Ways to describe a column

1. Type of distribution (uniform, lognormal, zipfian)
2. Range of values
3. Type of data (numerical, categorical, date)
4. Given column name, guess the domain
5. Mean, max, min, std
6. Mathematical units, guess what the unit represents

Proceedings of the VLDB Endowment, Vol. 11, No. 8
Copyright 2018 VLDB Endowment 2150-8097/18/4.
DOI: https://doi.org/TBD

Challenges in labeling a column

1. does the column signature match more than one column
2. domain of the data, would help narrow down the search
3. There might be many column names with primitive data types like year, date, etc
4. column cannot be described effectively by a distribution
5. column has generic data

2. BACKGROUND AND RELATED WORK

3. EXPERIMENTS

Describe how KB was created for datasets. System design from HILDA paper

4. RESEARCH PLAN

Tackling the challenges of describing and labeling a column

1. Column might match on multiple signatures
2. Similar concepts with different signatures
3. Similar signatures for different concepts
4. Insufficient signal for signature based matching. -¿types of signature insufficient
5. different signatures combine differently to id concepts
6. performance

5. CONCLUSION

6. ACKNOWLEDGMENTS

7. REFERENCES