Incremental Schema Recovery

Poonam Kumari
University at Buffalo, SUNY
poonamku@buffalo.edu

Gourab Mitra
University at Buffalo, SUNY
gourabmi@buffalo.edu

Oliver Kennedy
University at Buffalo, SUNY
okennedy@buffalo.edu

ABSTRACT

Abstract goes here

KEYWORDS

Stuff

ACM Reference Format:
Poonam Kumari, Gourab Mitra, and Oliver Kennedy. 1997. Incremental Schema Recovery. In *Proceedings of ACM Woodstock conference (WOODSTOCK'97).* ACM, New York, NY, USA, Article 4, 2 pages. https://doi.org/10.475/123_4

1 INTRODUCTION

It's a story as old as time: Student gathers data, makes a graph with the data, writes a paper about the data, and then graduates. With the student gone, the data languishes. Without so much as a wiki page or README file documenting it, anyone who wants to re-use the data needs to spend hours, days, or even weeks reverse-engineering it. If we're lucky, that person documents their efforts. If not, the entire process repeats.

To break this cycle of data abandonment, we propose *Label Once, and Keep It* (**LOKI**), a data-ingest middleware for incremental, re-usable schema recovery. Fundamentally, **LOKI** allows users to assemble schemas on-demand, both discovering and incrementally refining schema definitions in response to changing data needs. To accomplish this, **LOKI** maintains a knowledge-base of both approximate, as well as exact schema labelings. First, approximate labelings derived from existing open-data sets, user-feedback, and expert-provided heuristics, jump-start the labeling process. When a user first points **LOKI** at a new tabular data set, **LOKI** provides users with a preliminary, default schema. As users confirm and/or override parts of the proposed schema, **LOKI** registers the mappings for anyone who uses the same dataset in the future.

In this paper, we detail on our initial efforts to prime the **LOKI** knowledge-base with existing governmental open-data. Specifically...

1.1 Notes

One more thought regarding a pitch for the work. We could wrap the idea in the context of a larger system for importing / querying initially unlabeled data. Specifically, when someone first loads an unlabeled (or only partially labeled) CSV file into a database/spark, they have two problems:

1) They need to label a subset of the columns that pertain to the specific analysis they want to do now. 2) They don't need to label *all* of the columns (might be 10s, 100s, or 1000s of columns that they don't care about).

However, at some point in the future, more labeling might be helpful. For example: 1) They pose a query and randomly discover that they are missing a column that *could* potentially exist in the source data. 2) Someone else wants to use the same data set, but with a different selection of columns. 3) The knowledge-base is updated and more automatic labelings become available.

I'm going to suggest that we present our contribution in the context of a system that: 1) Auto-suggests names for columns based on existing heuristics 2) Saves labeling efforts, making it possible to incrementally label a data-set and re-use effort across analyses 3) Allows you to ask whether a particular column name *could* exist in a given data set, and identify the data column that most-likely represents it.

Specifically, in this paper, we're conducting a case study evaluating one particular approach to task (1).

2 SYSTEM DESIGN

System Overview

Things we need to address:

- How is this a "middleware" Come up with a system diagram
- How to uniquely identify *specific* source datasets (provenance, UUIDs, URLs, etc...)?
- (possibly future work) Distributed feedback: Merging (conflicting?) labels from multiple users.

3 KNOWLEDGE-BASE

Overview sketching data for similarity.

Data type taxonomy

- Numeric Data (Sketch = Distribution)
- Textual Data (N-Gram distribution?,)
- Enum Types (Overlap, Concept-Similarity)
- ...?

3.1 Numeric Data

Focus on the challenges of sketching numeric types

4 EXPERIMENTS

Experiments

5 RELATED WORK

Related work

6 FUTURE WORK

Future work...

- Meta-queries for columns that *could* be in a query.
- Discovery of meta-data (e.g., units)
- Automatic translation/transformation (units)

WOODSTOCK'97, July 1997, El Paso, Texas USA
2016. ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/10.475/123_4