

Galhardas, et al. Declarative Data Cleaning (AJAX)

Discussion

- Do you buy the motivation? Key points?
- Sufficiently expressive?
- Benefit of being D.S.?
- The DSL approach.
 - Can you imagine solving the CiteSeer problem with this?
- Interactivity: Helpful? Scalable?
- Does the framework accomodate more modern extensions?
- Compare/contrast with Potter's Wheel

Experimental study

- Focused on runtime performance of matching
- Evaluate recall for MPN heuristic
- Conclusions specific to these matching heuristics
- No justification for appropriateness of language
- What are other important metrics? How would one evaluate them?

Approximate Match Techniques

- Thresholded distance functions
- Avoid cross-products?
- Prune candidates using a conservative distance function on mapped data (mapped to a coarser granularity)
 - e.g. strlen prunes for edit distance
- specific pruning algorithm and parameters expressed as "hints"

Exceptions

- Exception from user Java code tagged onto exceptional data inputs
- Users can examine exceptions, analyze provenance, interactively correct, and reintegrate into flow

Mapping

- 1-to-many
- can consider entire input relation
- Finally, a join of Lets followed by a tee to outputs
- Constraints on output
 - exceptions to constraints logged in input exception stream

View

- SQL query
- many-to-one
- constraints on output and exception handling

Matching

- approximate join between two relations
- cartesian product w/distance function
- + on input relation preserves unmatched tups (a la outer join)

Clustering

- Can be arbitrary distance-based
- Or, can take input distance pairs and cluster by graph properties like transitive closure to produce new pairs
- No exceptions
- Output depends on the style of clustering

Merging

- SQL-style GROUPBY/Agg with User-defined aggregates
- E.g. longest exemplar of cluster

Main operators

Key Points

- Main challenge: design/implement dataflow graphs
- Declarativity: Separate Logical from Physical
 - Logical = semantics, Physical = performance
 - Why is this so important?
 - Example: custom approximate match functions are black boxes
- Explain cleaning results (lineage)
- Interactive tuning of program

Example errors from CiteSeer

- Lack of Keys/IDs
- Abbreviations
- Inconsistent values
- Misspellings
- Missing values

Key contributions

- SQL-like language for data cleaning
- Exception handling to drive interactivity via data lineage
- Notation for approximate matching