Data Curation at Scale: The Data Tamer System

Stonebraker et al., CIDR 2013

Presenter: Kevin Chang

15-799

11/24/2013
Data Curation

• Find/ingest a data source(s) of interest
• Clean
• Transform
• Deduplicate/consolidate
Example

- Web aggregator: 80,000 URLs

Courtesy: Prof. Stonebraker’s slides
Motivation

• **Problem**: Little work on an end-to-end system that collects all the curation components

• **Proposal**: *Data Tamer*
  – End-to-end curation system with machine learning and statistics to make automatic decisions
  – Transform, clean, **deduplicate** incoming data
Data Tamer
Data Tamer

• GUI console for the DTA (Data Tamer administrator) to specify actions:
  – Ex: Sites or sources of data (e.g., URL)
  – Ex: Store incoming data into a Postgres database
Data Tamer

- Integrates data sources based on specified schemas (partial, complete, or nothing)
- Compares an attribute from a data source to a collection of other attributes
- Uses a **collection of algorithms** (experts)
  - Ex: Fuzzy string comparisons, Jaccard similarity, etc
Data Tamer

- **Deduplication:**
- 1. Obtains a training set of duplicates
- 2. Data categorization (k-means++)
  - Ex: western vs. eastern ski areas
Data Tamer

3. Duplicate-tuple clustering
   - Ensures transitive deduplication results

(t1, t2) (t2, t3)
• **Crowd sourcing mode** to increase confidence on the correctness of results
  – Asks DE (Domain Experts) for responses
  – Additional quality rating on their responses
  – *Economic incentive* to increase response rate
Data Tamer

- Displays data source
Evaluation

• Web aggregator data:

<table>
<thead>
<tr>
<th></th>
<th>Current</th>
<th>Data Tamer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Found duplicates</td>
<td>4%</td>
<td>98.9%</td>
</tr>
<tr>
<td>Precision</td>
<td>97%</td>
<td>100%</td>
</tr>
</tbody>
</table>