Data Provenance Support in Entity Resolution

STUDENT: XU WANG
SUPERVISOR: DR. QING WANG
COURSE CODE: COMP4560
OUTLINE

- Background
- Project Goal
- Methods
- Evaluation
Background – Entity Resolution

• Definition:
The process of deciding which records from one or more databases correspond to the same real-world entity

• Applications:
  - National census
  - Crime detection
  - Medical practice and research

<table>
<thead>
<tr>
<th>id</th>
<th>name</th>
<th>age</th>
<th>address</th>
<th>postcode</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Xu Wang</td>
<td>21</td>
<td>5 Dunn Place</td>
<td>2602</td>
</tr>
<tr>
<td>2</td>
<td>Xu Wang</td>
<td>21</td>
<td>5 Dunn Pl</td>
<td>2602</td>
</tr>
<tr>
<td>3</td>
<td>X. Wang</td>
<td>21</td>
<td>5 Dunn Pl</td>
<td>2602</td>
</tr>
<tr>
<td>4</td>
<td>Xu Wang</td>
<td>33</td>
<td>1 Elliot Street</td>
<td>2612</td>
</tr>
<tr>
<td>5</td>
<td>Wang Xu</td>
<td>33</td>
<td>1 Elliot St</td>
<td>2612</td>
</tr>
<tr>
<td>6</td>
<td>Yang Wang</td>
<td>21</td>
<td>5 Dunn Place</td>
<td>2602</td>
</tr>
<tr>
<td>7</td>
<td>Y. Wang</td>
<td>21</td>
<td>5 Dunn Place</td>
<td>2602</td>
</tr>
<tr>
<td>8</td>
<td>Wang Yang</td>
<td>21</td>
<td>5 Dunn Place</td>
<td>2602</td>
</tr>
</tbody>
</table>
Background – Data Provenance

● **Definition:**
The process of tracing and recording the origins of data and its movement between databases.

● **Applications:**
  - Record repair
  - Reliability calculation
Background – Motivation

- **ER is important and widely applied in many fields**

- **ER still has problems:**
  - Hard to calculate reliability of results
  - Hard to fix inconsistent results

- **Solution: data provenance**
  - Features provided by data provenance are exactly right proper for ER problems
  - There are few studies about data provenance
Project Goal

- Develop **storage structures** that can be used to capture provenance information of entity resolution
- Develop **query methods** which can provide a detailed overview on the history of an entity
- Evaluate the **effectiveness and efficiency** of the developed methods over two real-world data sets.
Methods – ER workflow

Dataset A

Blocking/ Indexing

Comparison

Classification

Clustering

Matches

Non-matches

Result
Methods – Project Overview

Users

Provenance Query Methods

Provenance Storage

Dataset A

Blocking/ Indexing

Comparison

Classification

Matches

Clustering

Non-matches

Dataset B

Result
Methods – Provenance Storage

- **3 levels:**
  - Individual record level
  - Record pair level
  - Record cluster level
Methods – Provenance Storage

**Individual record level**

1 = [“block_1”]
2 = [“block_1”]
......
8 = [“block_3”]

**Record pair level**

1 =
   {“(1,2)” : ...}
   {“(1,3)” : ...}

2 =
   {“(1,2)” : ...}
   {“(2,3)” : ...}

**Record cluster level**

“cluster_1” : [“1”, “2”, “3”]
“cluster_2” : [“4”, “5”]
“cluster_3” : [“6”, “7”, “8”]

cluster_1 :
    pair: (1,2)
    pair: (2,3)
cluster_2 :
    pair: (4,5)
Methods – Provenance Query

- **3 levels:***
  - Individual record level
    - Find record
    - Find block
  - Record pair level
    - Find pair
  - Record cluster level
    - Find cluster
    - Reconstruct
Methods – Provenance Query

Find Record

Find Block

Find Pair

Find Cluster

Reconstruct
Methods – Graphical Query Tool
Evaluation – Experimental Setup

- **Environment:**
  - MAC OSX Yosmite, CPU: 2.5 GHz Intel Core i7, RAM: 16GBytes
  - Python 2.7.11

- **Datasets:**
  - CORA: 1878 records, 6 attributes, machine learning publication dataset
  - NCVR: 220,000 records, 19 attributes, North Carolina voter profile dataset
## Evaluation – Space and Time Efficiency

<table>
<thead>
<tr>
<th>Level</th>
<th>Storage Type</th>
<th>Size (MB)</th>
<th>Query Method</th>
<th>Loading (ms)</th>
<th>Execution (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Record</td>
<td>Original Dataset</td>
<td>0.256</td>
<td>FindRecord</td>
<td>0.0826</td>
<td>6.7</td>
</tr>
<tr>
<td></td>
<td>Block provenance</td>
<td>0.08</td>
<td>FindBlock</td>
<td>17.4</td>
<td>0.77</td>
</tr>
<tr>
<td>Pair</td>
<td>Pair dictionary</td>
<td>42.6</td>
<td>FindPair</td>
<td>14260</td>
<td>0.35</td>
</tr>
<tr>
<td>Cluster</td>
<td>Cluster dictionary</td>
<td>0.051</td>
<td>FindCluster</td>
<td>66</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td>Edge dictionary</td>
<td>0.456</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 6.3: Experiment result of CORA

<table>
<thead>
<tr>
<th>Level</th>
<th>Storage Type</th>
<th>Size (MB)</th>
<th>Query Method</th>
<th>Loading (ms)</th>
<th>Execution (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Record</td>
<td>Original Dataset</td>
<td>53.5</td>
<td>FindRecord</td>
<td>0.087</td>
<td>1117</td>
</tr>
<tr>
<td></td>
<td>Block provenance</td>
<td>14.3</td>
<td>FindBlock</td>
<td>2347</td>
<td>158</td>
</tr>
<tr>
<td>Pair</td>
<td>Pair dictionary</td>
<td>118.5</td>
<td>FindPair</td>
<td>45593</td>
<td>3.2</td>
</tr>
<tr>
<td>Cluster</td>
<td>Cluster dictionary</td>
<td>11.5</td>
<td>FindCluster</td>
<td>7190</td>
<td>41</td>
</tr>
<tr>
<td></td>
<td>Edge dictionary</td>
<td>37.2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 6.4: Experimental result of NCVR
Thank you!