Real-time collective entity resolution

Jingwei Zhang

Supervisor: Dr. Qing Wang
COMP 8780 Information & Human Centered Computing Project
Semester 2
7/11/2013
Outline

• Introduction
• Methodology
• Experiments
• Conclusion & Future works
Background

• Entity resolution
  – The problem of determining which records in a dataset refer to the same entities.
Background

• Three kinds of approaches:
  • Attribute-based entity resolution
    – real world entity attributes are often insufficient to determine entities’ similarity
    – sensitive to missing values
Background

• Three kinds of approaches:
  • Attribute-based entity resolution
  • Naïve relational entity resolution
    – poor performance when entity names are common and relationships are dense
Background

• Three kinds of approaches:
  • Attribute-based entity resolution
  • Naïve relational entity resolution
  • Collective entity resolution
    – resolve entities and their neighbors collectively
Motivation

• The shortness of collective entity resolution
  – high complexity

• The need for real-time collective entity resolution
  – fast response and low time cost
Objectives

• Develop a real-time collective entity resolution approach.

• Implement some techniques into this approach to improve the efficiency and effectiveness of entity resolution.

• Analyze experimental results of the developed approach.
## Methodology

<table>
<thead>
<tr>
<th>step</th>
<th>process</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Attribute-based blocking</td>
</tr>
<tr>
<td>2</td>
<td>Candidate representative determination</td>
</tr>
<tr>
<td>3</td>
<td>Collective entity resolution</td>
</tr>
<tr>
<td>4</td>
<td>Real-time threshold set determination</td>
</tr>
</tbody>
</table>
## Step 1 Attribute-based blocking

- **Single attribute: Entity name**
- **Name similarity comparison:**
  1. Comparing initials’ edit-distance
  2. Comparing last names’ and full names’ edit-distance
- **Generate a candidate set**

<table>
<thead>
<tr>
<th>Record</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z</td>
<td>Jingwei Zhang</td>
</tr>
<tr>
<td>A</td>
<td>Jing Zhang</td>
</tr>
<tr>
<td>B</td>
<td>Josh Zhang</td>
</tr>
<tr>
<td>C</td>
<td>Joshua Zhang</td>
</tr>
<tr>
<td>E</td>
<td>Mingshi Wu</td>
</tr>
<tr>
<td>F</td>
<td>John Doe</td>
</tr>
<tr>
<td>G</td>
<td>Osama Lade</td>
</tr>
<tr>
<td>H</td>
<td>George Bush</td>
</tr>
</tbody>
</table>
Step 2 Candidate representative determination

• Number of clusters is unknown

• Manually refinements
  – Let users determine one of the records in the candidate set to be the candidate representative
Step 3 Collective entity resolution

• Hierarchical clustering
  – Different thresholds, different cutting points
• Bottom up structure
  – Local optimum (collective cluster & one of the other clusters)
• Greedy algorithm
  – Merge the cluster with the highest similarity with collective cluster in each hierarchy
Example of Collective entity resolution
Step 3 Collective entity resolution

• Similarity formula:

\[ \text{Sim} (C_i, C_j) = (1 - \alpha) \times \text{Sim}_A(C_i, C_j) + \alpha \times \text{Sim}_N(C_i, C_j) \quad (0 \leq \alpha \leq 1) \]

  Attribute similarity    Neighborhood similarity

• Attribute similarity
  – Different weights

• Neighborhood similarity
  – Jaccard coefficient
Step 4 Real-time threshold set determination

• Threshold set determination
  - Different percentages of the highest combination similarity of the first merging
  - Gap observation

• Three thresholds for three results:
  - Fuzzy result with high recall (75%)
  - Balanced result with high balanced F-measure (80%)
  - Refine result with high precision (85%)
Example of Collective entity resolution
Experiments

• Experiment environment
  Java with external packages & Postgres Database

• Dataset
  One bibliographic dataset from Scopus (data volume 47,000+), contains authors & publications information
Experiments

• Effectiveness: 3 measures
  – Precision
  – Recall
  – Balanced F-Measure

• Efficiency: complexity analysis
Effectiveness comparison for different kinds of entity resolution

<table>
<thead>
<tr>
<th></th>
<th>Highest precision with over 90% recall</th>
<th>Highest recall with over 90% precision</th>
<th>Highest balanced F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attribute-based ER</td>
<td>0.8861</td>
<td>0.8579</td>
<td>0.8778</td>
</tr>
<tr>
<td>Naïve relational ER</td>
<td>0.7353</td>
<td>0.7206</td>
<td>0.7921</td>
</tr>
<tr>
<td>Collective ER</td>
<td><strong>0.9345</strong></td>
<td><strong>0.9372</strong></td>
<td><strong>0.9185</strong></td>
</tr>
</tbody>
</table>
## Threshold set evaluation

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Precision for threshold 0.75</th>
<th>Precision for threshold 0.8</th>
<th>Precision for threshold 0.85</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-100</td>
<td>0.8935</td>
<td>0.9291</td>
<td>0.9537</td>
</tr>
<tr>
<td>101-200</td>
<td>0.7712</td>
<td>0.8192</td>
<td>0.8417</td>
</tr>
<tr>
<td>201-300</td>
<td>0.8437</td>
<td>0.8951</td>
<td>0.9111</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Recall for threshold 0.75</th>
<th>Recall for threshold 0.8</th>
<th>Recall for threshold 0.85</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-100</td>
<td>0.9587</td>
<td>0.9272</td>
<td>0.8759</td>
</tr>
<tr>
<td>101-200</td>
<td>0.9868</td>
<td>0.9662</td>
<td>0.9022</td>
</tr>
<tr>
<td>201-300</td>
<td>0.9897</td>
<td>0.9801</td>
<td>0.9033</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Datasets</th>
<th>F-measure for threshold 0.75</th>
<th>F-measure for threshold 0.8</th>
<th>F-measure for threshold 0.85</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-100</td>
<td>0.9033</td>
<td>0.9088</td>
<td>0.8885</td>
</tr>
<tr>
<td>101-200</td>
<td>0.8146</td>
<td>0.8446</td>
<td>0.8229</td>
</tr>
<tr>
<td>201-300</td>
<td>0.8788</td>
<td>0.9082</td>
<td>0.8738</td>
</tr>
</tbody>
</table>
Efficiency evaluation

• Complexity changes
  – Resolution in real-time (N entities -> single entity)
  – Candidate representative (K clusters -> single cluster)

• Overall run time statistics
  – Average time needed: 2 seconds per resolution

<table>
<thead>
<tr>
<th>Threshold</th>
<th>0.85</th>
<th>0.8</th>
<th>0.75</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>1981.13ms</td>
<td>2004.38ms</td>
<td>2033.24ms</td>
</tr>
<tr>
<td>(average)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Conclusions and Future Works

• Multiple attributes for attribute-based blocking

• Data preprocessing into developed approach

• Automatic refinement techniques for candidate representative determination

• Domain-independent ability

• Experiments on data from different sources
Question