Query-time Entity Resolution over Graphs

BINGXIN LI
SUPERVISED BY DR. QING WANG
Outline

• Background
• Query-time ER
• ER Techniques
• Evaluation
• Questions
Entity Resolution (ER) is to identify which records in one or more databases refer to the same real world entity.
Query-time ER

- Given a query: find all records for the student Ming Chen at ANU (r1)
ER Techniques

- Attribute-based ER Techniques
  - String similarity

- Collective ER Techniques
  - Relational similarity + String similarity

- Connectivity ER Techniques (to be explored in the project)
  - Graph properties like connectivity + ...
Attribute-based ER

<table>
<thead>
<tr>
<th>Name</th>
<th>University</th>
<th>Email</th>
</tr>
</thead>
<tbody>
<tr>
<td>r1</td>
<td>Ming Chen</td>
<td>ANU</td>
</tr>
<tr>
<td></td>
<td></td>
<td><a href="mailto:mc@gmail.com">mc@gmail.com</a></td>
</tr>
<tr>
<td>r2</td>
<td>M. Chen</td>
<td>ANU</td>
</tr>
<tr>
<td></td>
<td></td>
<td><a href="mailto:mc@anu.edu.au">mc@anu.edu.au</a></td>
</tr>
</tbody>
</table>

Similarity

- r1: 0.7
- r2: 0.6
Collective ER

<table>
<thead>
<tr>
<th></th>
<th>Name</th>
<th>Uni</th>
<th>Email</th>
</tr>
</thead>
<tbody>
<tr>
<td>r1</td>
<td>Ming Chen</td>
<td>ANU</td>
<td><a href="mailto:mc@gmail.com">mc@gmail.com</a></td>
</tr>
<tr>
<td>r2</td>
<td>M. Chen</td>
<td>UC</td>
<td><a href="mailto:mc@uc.edu.au">mc@uc.edu.au</a></td>
</tr>
<tr>
<td>sim1</td>
<td>0.7</td>
<td>0</td>
<td>0.5</td>
</tr>
</tbody>
</table>

String sim1:

relational sim2:
(using Jaccard Coefficient):

\[
\text{JaccardCoef } f(c_1,c_2) = 0.8
\]

\[
\text{Sim}(c_1,c_2) = (1 - \alpha) \times \text{sim1}_{\text{total}} + \alpha \times \text{sim2}
\]

(0 <= \alpha <= 1)
Connectivity ER

ANU

Mc@ gmail.com

Abraham

Tom

Ming Chen

James

Lily

UC

Mc@uc.edu.au
Evaluation

- Experiment environment
  - Java with external packages, Database(PostgreSQL)

- Data sets
  - One data sets from Cora that contains authors, publications and venues information

- Expected results
  - Evaluate the effectiveness of connectivity ER
  - Compare the efficiency between connectivity and other ER techniques
  - Application independent (use other dataset)
Questions?