Improving Grouped-Entity Resolution using Quasi-Cliques

Byung-Won On, Ergin Elmacioglu, Dongwon Lee, Jaewoo Kang*, Jian Pei+

Penn State University, USA
*Korea University, Korea
+Simon Fraser University, Canada

IEEE ICDM 2006

The ER & GER Problems

- Entity Resolution Problem
  - Identifying matching entities that refer to the same real-world object
  - Main building block in many data applications
- Grouped-Entity Resolution
  - Entities have a group of repetitive elements
  - One can exploit the repetitive elements for improving ER accuracy
  - We propose to use Quasi-Clique

GER Example

In DL, identifying matching authors with their groups of citations

GER Examples

- In IRS, Identifying matching tax payers using their family information
- In Multimedia, retrieving matching images with m x n grids

Landscape

- Abundant research on related problems
- Also known as:
  - DB: approximate join, merge/purge, record linkage
  - DL: citation matching, author name disambiguation
  - AI: identity uncertainty
  - LIS: name authority control

Landscape

- In a nutshell, existing approaches often do:
  - For two entities, e1 and e2, capture their information in data structures, D(e1) and D(e2)
  - Measure the distance or similarity between data structures: \( \text{dist}(D(e1), D(e2)) = d \)
  - Determine for matching:
    - If \( d < \text{threshold} \), then \( e1 \) and \( e2 \) are matching entities
  - Work well for common applications
  - Ours do ER better when
    - Entities have structures (i.e., repetitive groups) that we can exploit using graphs
Using Graphs

False Positive Problem:
name variant of a

f = {d, e, k, g, z, y, x}
c = {d, e, x, g, z, y, j, t}

Jaccard (a, f) = 4/11 = 0.36
Jaccard (a, c) = 6/10 = 0.6 (<= name variant of a)

False positive problem:
V(a) V(f)
V(a) V(e)
V(d) V(e)
V(y)
V(a)
V(y)
V(z)
V(d)
V(e)
V(k)
V(g)
V(x)
V(f)
V(e)
V(k)
V(g)
V(æ)
V(œ)
V(ç)
V(é)
V(ê)
V(í)
V(î)
V(ï)
V(õ)
V(ô)
V(ö)
V(ò)
V(û)
V(ü)
V(ý)

Our graph-based approach:

- Overcome the limitation of existing distance metrics
- Unearth the hidden relationships in contents
- Use Quasi-Clique to measure the strong relationships

Graph(a)
Graph(b)
Graph(c)
Graph(d)
Graph(e)
Graph(f)
Graph(g)
Graph(h)

Using Graphs

- Represent entity e1 as graph g1 using common tokens
  - Author: co-author
  - Venue: common venues
  - Title: common keywords
- Superimpose the graph g1 onto base graph B1 to get a final graph representation G1
  - Author: entire collaboration graph as B1
  - Venue: entire venue similarity graph as B1
  - Title: entire token co-occurrence graph B1
- Measure the similarity of two entities e1 and e2 w.r.t. G1 and G2

Quasi-Clique

Graph G
V(G): set of vertices
E(G): set of edges
\( \Gamma \)-quasi-complete-graph \((0 \leq \Gamma \leq 1)\)
Every vertex in G has at least \(\Gamma \times (|V(G)|-1)\) degrees
V(S) \(\subseteq V(G)\)
\(G(S): \Gamma\)-Quasi-Clique
If V(S) forms the graph satisfying \(\Gamma\)-quasi-complete-graph
\(G(S): \text{Clique}\)
If \(\Gamma = 1\)
Use Quasi-Clique (QC) to measure contextual distances
E.g., Function QC(G(a), G(b), \(\Gamma = 0.3\), S=3)

Superimposition

ACM Dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>JC</td>
<td>0.8</td>
<td>0.9</td>
</tr>
<tr>
<td>JC+QC</td>
<td>0.8</td>
<td>0.9</td>
</tr>
<tr>
<td>TI</td>
<td>0.6</td>
<td>0.7</td>
</tr>
<tr>
<td>TI+QC</td>
<td>0.6</td>
<td>0.7</td>
</tr>
<tr>
<td>IC</td>
<td>0.7</td>
<td>0.8</td>
</tr>
<tr>
<td>IC+QC</td>
<td>0.7</td>
<td>0.8</td>
</tr>
</tbody>
</table>

IMDB Synthetic Dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>JC</td>
<td>0.7</td>
<td>0.8</td>
</tr>
<tr>
<td>JC+QC</td>
<td>0.7</td>
<td>0.8</td>
</tr>
<tr>
<td>TI</td>
<td>0.6</td>
<td>0.7</td>
</tr>
<tr>
<td>TI+QC</td>
<td>0.6</td>
<td>0.7</td>
</tr>
</tbody>
</table>
Conclusion

- Many ER problems have entities with a group of repetitive elements
- Our Quasi-Clique based method exploits them to achieve improved accuracy
- Further improvement using Groups instead of Graphs
  - Group Linkage
  - Will appear in ICDE 2007