An Effective Approach to Entity Resolution Problem Using Quasi-Clique and its Application to Digital Libraries

ACM/IEEE JCDL 2006

Byung-Won On, Ergin Elmacioglu, Dongwon Lee (Penn State Univ.)
Jaewoo Kang (North Carolina State Univ.)
Jian Pei (Simon Fraser Univ.)

Motivation @ ACM DL

Jeffrey D. Ullman @ Stanford Univ.

The same entities (authors) mistakenly appear under different name variants
Motivation @ DBLP

Each entity (author) has “a list of tuples” associated with it => Grouped Entity

Grouped Entity

Regular Entity

Grouped Entity

ID or Name

Contents or Metadata
Problem Definition

- Entity Resolution (ER) Problem
  - The process of detecting and merging duplicate entities that represent the same real-world object
- Grouped Entity Resolution (GER) Problem

Given a set of entities, \( E \), where each contains a group of elements, for each entity, \( e_c (\in \mathcal{E}) \), identify all variant entities, \( e_v (\in \mathcal{E}) \), such that \( \text{dist}(e_c, e_v) < \delta \)

Limitation to Distance Metrics

- Many previous approaches
  - SVM, Jaccard, TF/IDF
  - Cosine similarity
    => Best accuracy
  - SVM
    => Slow time
    => Training data set is required
- Suffer from false positives
- Need to unearth the hidden relationship beyond distance
False Positive Problem

\[ a = \{d, e, k, g, z, y, x\} \]
\[ f = \{d, e, k, g, p, r, t, q\} \]
\[ c = \{d, e, x, g, z, y, j, t\} \]

\[ \text{Jaccard} (a, f) = \frac{4}{11} = 0.36 \]
\[ \text{Jaccard} (a, c) = \frac{6}{10} = 0.6 (\leq \text{name variant of } a) \]
Superimposition

Graph(a) Collaboration graph

False Positive Problem

name variant of a

Graph(a) Graph(f) Graph(c)
Idea

- Represent entity $e_1$ as graph $g_1$ using common tokens
  - Author: co-author
  - Venue: common venues
  - Title: common keywords

- Superimpose the graph $g_1$ onto base graph $B_1$ to get a final graph representation $G_1$
  - Author: entire collaboration graph as $B_1$
  - Venue: entire venue similarity graph as $B_1$
  - Title: entire token co-occurrence graph as $B_1$

- Measure the similarity of two entities $e_1$ and $e_2$ w.r.t $G_1$ and $G_2$

Quasi-Clique

- Graph $G$
  - $V(G)$: set of vertices
  - $E(G)$: set of edges
  - $\Gamma$-quasi-complete-graph ($0<\Gamma\leq1$)
    - Every vertex in $G$ has at least $\Gamma$ degrees
  - $V(S) (\subseteq V(G))$
    - $G(S)$: $\Gamma$-Quasi-Clique
      - $\Gamma$: Quasi-Clique
        - $G(S)$: $\Gamma$-Quasi-Clique
          - If $V(S)$ forms the graph satisfying $\Gamma$-quasi-complete-graph
    - $G(S)$: Clique
      - $\Gamma=1$

- Use Quasi-Clique (QC) to measure contextual distances
  - E.g., Function $QC(G(a), G(b), \Gamma=0.3, S=3)$

  (1) $\Gamma=0.3$: each vertex has 2 degrees
  (2) Find cliques shared between $G(a)$ and $G(b)$
  (3) $S=3$: # of vertices appearing in cliques > 3

$b$ is the name variant of $a$ by (1), (2), and (3)
Algorithm 1: distQC

**Input:** A grouped-entity $e$, an ER method $\mathcal{M}$, and three parameters ($\alpha$, $\gamma$ and $\mathcal{S}$).

**Output:** $k$ variant grouped-entities, $e_v$ ($\in E$), such that $e_v \sim e$.

1. Using $\mathcal{M}$, find top $\alpha \times k$ candidate entities, $e_X$.
2. $G_c(e) \leftarrow$ context graph of $e$;
3. \textbf{forall} $e_i$ ($\in e_X$) \textbf{do}
4. \hspace{1em} $G_c(e_i) \leftarrow$ context graph of $e_i$;
5. \hspace{1em} $g_i \leftarrow$ QC($G_c(e)$, $G_c(e_i)$, $\gamma$, $\mathcal{S}$);
6. Sort $e_i$ ($\in e_X$) by $|g_i|$, and return top-$k$;

---

**Experimental Validation**

- 47 real cases of canonical and variant grouped entities (ACM)
  - Average 24 tuples
- Synthetic cases

<table>
<thead>
<tr>
<th>Data set</th>
<th>Domain</th>
<th># of grouped entities</th>
<th># of tuples in all entities</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACM</td>
<td>Computer Science</td>
<td>707,368</td>
<td>1,037,968</td>
</tr>
<tr>
<td>EconPapers</td>
<td>Economics</td>
<td>18,399</td>
<td>20,486</td>
</tr>
<tr>
<td>BioMed</td>
<td>Medical</td>
<td>24,008</td>
<td>6,160</td>
</tr>
<tr>
<td>IMDB</td>
<td>Entertainment</td>
<td>935,707</td>
<td>446,016</td>
</tr>
</tbody>
</table>
Parameter Tuning

- JC: Jaccard similarity
- JC+QC: JC + Quasi-Clique
- TI: TF/IDF Cosine similarity
- TI+QC: TI + Quasi-Clique
- IC: IntelliClean (venue hierarchy)
- IC+QC: IC + Quasi-Clique

ACM Real Dataset

Attributes:
- JC: Jaccard similarity
- JC+QC: JC + Quasi-Clique
- TI: TF/IDF Cosine similarity
- TI+QC: TI + Quasi-Clique
- IC: IntelliClean (venue hierarchy)
- IC+QC: IC + Quasi-Clique
IMDB Synthetic Dataset

Related Work

- Bhattacharya et al., DMKD (2004)
- Malin, LACS (2005)
- RelDC: Kalashnikov et al., TODS (2005)
Summary

- Using contextual distance by using a group of tuples in entity is beneficial in solving GER problem
- Quasi-Clique can be used to estimate the strength of core concepts of two entities
  - Reduce false positives by adjusting the results of distance-based metrics (eg, Jaccard, cosine)
- Extension to other domains (eg, actor with a movie list) to make the proposal generic and robust is underway