Entity Resolution in the Presence of Constraints

ZEYU SHEN U4813857
SUPERVISED BY DR. QING WANG
Outline

- Introduction
- Motivation
- Methods/Techniques
- Experiment
- Conclusion
- Q&A
**Introduction**

**Entity resolution**

Problem of identifying and linking/grouping different manifestations of the same real world object

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>LBJ</td>
<td></td>
<td></td>
</tr>
<tr>
<td>King James</td>
<td>Lebron James</td>
<td></td>
</tr>
<tr>
<td>L. James</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Introduction

Constraints

- Each paper has a unique publication venue
- Conferences in different cities, are in different years.

Two methods are going to adopted on constraints

- Dedupalog
- Weights
Motivation

In the prior works:

The computational cost for taking constraints into account is often very high.

This project:

Find an efficient way of handling constraints in the Entity Resolution process
Methods/Techniques

Dedupalog

**Soft rules:** are more likely to be clustered together.

\[
\text{Paper} \ast (id, id') \leftrightarrow \text{PaperRefs}(id, t, -, -, -), \text{PaperRefs}(id', t', -, -, -), \text{TitleSimilar}(t, t')
\]

*Papers with similar titles are likely duplicates*

**Hard rules:** must be satisfied by every legal clustering

\[
\text{Venue} \ast (id, id') \leq \text{Paper} \ast (id, id')
\]

*Each paper has a unique publication venue*
Methods/Techniques

Weights

<table>
<thead>
<tr>
<th>Rule</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Each paper has a unique publication venue</td>
<td>100%</td>
</tr>
<tr>
<td>Papers with similar titles are likely duplicates</td>
<td>80%</td>
</tr>
<tr>
<td>Author references that do not share any common coauthors do not refer to the same author.</td>
<td>60%</td>
</tr>
</tbody>
</table>

When conflicts happen, the rule with higher weight is adopted.
Experiment

Quality on Cora dataset

- Recall, precision
- Compare the differences between with constraints and without constraints
- Compare between using Dedupalog and using weighted constraints

Performance

- Efficiency of using constraints
Conclusion

1. To conduct a literature review on constrain-based entity resolution methods.

2. To implement the algorithms that support the framework for constraints.

3. To incorporate weights for constraints into the framework.

4. To analyse the efficiency and effectiveness of the developed approach using Cora datasets
Q&A