Entity Resolution in the Presence of Constraints

ZEYU SHEN U4813857
SUPERVISED BY DR. QING WANG
Outline

- Introduction
- Objectives
- Dedupalog
- Methodology
- Experiments
- Future work
Introduction

Entity resolution

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

<table>
<thead>
<tr>
<th>Similar Strings</th>
<th>Zeyu Shen</th>
<th>ZEYU SHEN</th>
<th>Z. SHEN</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Different Strings</th>
<th>LBJ</th>
<th>King James</th>
<th>The Little Emperor</th>
</tr>
</thead>
</table>

Lebron James
Why constraints

Constraints exist everywhere

- “Each paper has a unique publication venue”
- “A person with age two cannot have a salary 200K”

If we can use constraints properly, the accuracy of entity resolution can be increased a lot.
Objectives

1. To develop a declarative framework by incorporating weighted rules into Dedupalog

2. To implement the algorithms used in Dedupalog for entity resolution

3. To analyse the effectiveness of the developed framework, and implement the algorithm using real datasets.
Dedupalog

Dedupalog program

- Soft-complete Rules
- Soft-incomplete Rules

VS

My method

- Weighted Soft Positive Rules
- Weighted Soft Negative Rules

- Hard Rules
- Complex Hard Rules

- Hard Positive Rules
- Hard Negative Rules
Weighted Soft rules

Positive_rule1  Weight: 0.8

Publication*(id,id’) <- Publication (id,t), Publication (id’,t’), TitleSimilar (t,t’)
If titles of two publications are similar, then these two publications should be put into one cluster.

Negative_rule3  Weight: -1.00

¬Publication*(id,id’) <- Publication (id,tech), Publication (id’,_ )
If one publication has value of tech and the other does not have, then these two publications should be put into different clusters
High level overview

Data Matching Process

Phase 1
- Data-preprocessing

Phase 2
- Indexing

Phase 3
- Record pair comparison
  - Include
  - Clerical review
    - Potential Match
      - Match
      - Not Match

Phase 4
- Evaluation
Phase 1: Preprocessing

Normalization

- Normalize Strings
  
  Quinlan (1993) "Combining instance-based and model-based learning",

- Normalize Integers

  (pp. 414-417).
Phase 1: Preprocessing

Data filling:

improve the completeness and maintain the correctness
Phase 2: Indexing

Remove the pairs which are apparently different and only keep the potential matching pairs, which are called \textit{candidate pairs}.

\textit{e.g.} Positive\_rule1 can filter out pairs like the follows

<table>
<thead>
<tr>
<th>id</th>
<th>authors</th>
<th>title</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>Steve Benford and Lennart E. Fahlen.</td>
<td>A spatial model of interaction in large virtual environments.</td>
</tr>
<tr>
<td>11</td>
<td>Benford, S., and Fahlen, L.</td>
<td>Viewpoints, Actionpoints and Spatial Frames for Collaborative User Interfaces,</td>
</tr>
</tbody>
</table>

Original pairs: 3.5 million

All positive rules

Candidate pairs: 74894
Phase 3: Classification

- **Step 1: Soften Rules**
  Each pair will be applied with all rules and gets a weight score, which determines whether it is a positive pair or negative pair.

- **Step 2: Harden Rules**
  add hard rules on the pairs generated by Step 1

- **Step 3: Conflicts Resolve**
  solve conflicts between hard negative pairs and clustering result.
Step 1: Soften rules

Threshold: 0.5

Positive_rule1: 0.8
Positive_rule2: 0.97
Negative_rule4: -0.97

Weight score: 0.8 > 0.5

Positive_rule1: 0.8
Negative_rule4: -0.97

Weight score: -1.17 < 0.5
Step 2: Harden rules

Cluster 1

- **Hard: positive_rule5**
  - 1
  - 2 positive
  - 3
  - 4
  - Weights: 0.8, 0.7, -1.17, -0.6

- **Hard: negative_rule2**
  - 2
  - 3 negative
  - 4
  - Weights: 0.8, 0.7, 0.6, 0.7
Step 3: Conflicts Resolve

1. Compare weight between (2,1) and (4,1)  ->  0.8 > 0.6
2. Compare weight between (2,3) and (4,3)  ->  (4,3) is hard
Phase 4: Evaluation

Experiment 1: Compare Quality between Each Step in Classification

Quality Improvement in each step

- Precision
- Recall
- F1-measure
Phase 4: Evaluation

Experiment 2: Compare Quality between Methods

**Method1**: My method

**Method2**: Dedupalog

![Comparison between method 1 & 2](chart.png)
Future work

1. Evaluate my method on large scale datasets to measure the efficiency and effectiveness
2. Investigate how rules can be specified across multiple entity types
3. Study the learning method for weights of constraints
Q&A