Towards a Scalable and Robust Entity Resolution - Blocking under Relational Constraints

Mingyuan Cui

Supervisor: Dr. Qing Wang, Dr. Huizhi Liang
COMP8740 Artificial Intelligence Project
Semester 1 May 28, 2014
Outline

- Introduction
- Motivation & Objectives
- Methodology
- Experiments
- Conclusion
Introduction

- **Entity Resolution (ER)**
  To identify data records over one or more datasets which refer to the same real world entity.
Introduction

Blocking

Blocking makes entity resolution scalable, by blocking a dataset into smaller blocks.

Entity resolution only needs to compare pairs in the same blocks.
## Introduction

### Blocking

<table>
<thead>
<tr>
<th>Last Name</th>
<th>First Name</th>
<th>Post Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roger</td>
<td>Jim</td>
<td>2602</td>
</tr>
<tr>
<td>Roger</td>
<td>James</td>
<td>2606</td>
</tr>
<tr>
<td>Tom</td>
<td>James</td>
<td>90024</td>
</tr>
<tr>
<td>Tom</td>
<td>Jimmy</td>
<td>90025</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Last Name</th>
<th>First Name</th>
<th>Post Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roger</td>
<td>Jim</td>
<td>2602</td>
</tr>
<tr>
<td>Roger</td>
<td>James</td>
<td>2606</td>
</tr>
<tr>
<td>Tom</td>
<td>James</td>
<td>90024</td>
</tr>
<tr>
<td>Tom</td>
<td>Jimmy</td>
<td>90025</td>
</tr>
</tbody>
</table>

10^5

120

230
Motivation

- Blocking usually performs based on textural similarity. (naïve method)
- Real world entities usually have relational semantic features to be distinguished from others.
- Semantic constraints could be a key to improve the performance of ER blocking. (new method)
  - Record 1: EM algorithm for Data mining (Book)
  - Record 2: EM algorithm for Data mining (Thesis)

Must not refer to the same entity
Objectives

Implement a blocking framework that ER can be more accurate while still keep high completeness

- To conduct a literature review on constraint-based entity resolution methods,
- To incorporate relational constraints into the blocking process.
- To analyze the efficiency and effectiveness of the developed approach, and evaluate the approach over real datasets.
Methodology

- **Step 1:** Block records based on textural similarity.
- **Step 2:** Set up semantic constraints according to the dataset.
- **Step 3:** Apply semantic constraints to the result of ER blocking.
Methodology

Overview

Domain Trees

Minhash functions

......

Semhash functions

Records

Minhash signature

......

Semhash signature

Locality-sensitive Hashing

......

Blocks
Step 1: Block a dataset based on textural similarity.

- **Similarity measure:** Jaccard similarity
  \[
  s(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}
  \]

- **Distance measure:** Jaccard distance
  \[
  d(S_1, S_2) = 1 - s(S_1, S_2)
  \]

- **Convert a text to set:** Q-gram
  Apple -> (ap, pp, pl, le) for q=2
Step 1: Block a dataset based on textural similarity.

**Similarity Preservation:**

- **Characteristic matrix**

<table>
<thead>
<tr>
<th>Element</th>
<th>$S_1$ = abc</th>
<th>$S_2$ = ab</th>
</tr>
</thead>
<tbody>
<tr>
<td>ab</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>bc</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

- **Minhash function**

<table>
<thead>
<tr>
<th>Minhash value</th>
<th>$S_1$ = abc</th>
<th>$S_2$ = ab</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h(S_1)$ = 1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$h(S_2)$ = 2</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

- **Minhash signature**

<table>
<thead>
<tr>
<th>Minhash function</th>
<th>$S_1$ = abc</th>
<th>$S_2$ = ab</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h_1$</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>$h_2$</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Step 1: Block a dataset based on textural similarity.

**Locality sensitive hashing (LSH):**

<table>
<thead>
<tr>
<th></th>
<th>$r_1$</th>
<th>$r_2$</th>
<th>$r_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h_1$</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>$h_2$</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>$h_3$</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
</tr>
<tr>
<td>$h_n$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$k$ rows $\rightarrow$ $l$ bands

$h_1, h_2$
Step2: Set up semantic constrains

- **Domain Tree**: Presents layered categories from domain knowledge.

![Domain Tree Diagram]

- Research Output
  - Publication
    - Peer Reviewed
      - Journal
      - Conference
      - Book
    - Non-Peer Reviewed
      - Technical Report
      - Thesis
  - Patent
Step 2: Set up semantic constrains

- Semantic interpretation: a set of functions that can extract the information from the dataset to possible categories.
  - r1 = (journal, conference, thesis)
  - r2 = (book)

- Semantic signature: Project nodes down to leaves and use 1 and 0 to indicate whether a record may be in some category.
  - r1: [1, 1, 0, 0, 1, 0]
  - r2: [0, 0, 1, 0, 0, 0]
Step3: Apply semantic constraints to the result of ER blocking.

- Logical conjunction:

<table>
<thead>
<tr>
<th>Record ID</th>
<th>Semantic signature</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 1 1 1</td>
</tr>
<tr>
<td>2</td>
<td>1 1 1 1</td>
</tr>
<tr>
<td>3</td>
<td>1 0 1</td>
</tr>
</tbody>
</table>

candidate pair: (1,2)

- Logical disjunction:

<table>
<thead>
<tr>
<th>Record ID</th>
<th>Semantic signature</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 1 1 1</td>
</tr>
<tr>
<td>2</td>
<td>1 1 1 1</td>
</tr>
<tr>
<td>3</td>
<td>1 0 1</td>
</tr>
</tbody>
</table>

candidate pair: (1,2), (1,3), (2,3)
Experiments

Environment:
- Java with external packages & Postgres Database

Dataset:
- Publications table from the Cora dataset

Metrics:
- Pair completeness (PC)
- Pair precision (PP)
- Pair redundancy (PR)
Experiments

Comparison of blocking without constraint to blocking with constraints (logical disjunction with 3 sigs).

-PC values (\( PC = \frac{T_{PB}}{T_{PG}} \))

<table>
<thead>
<tr>
<th>k values</th>
<th>PC values</th>
</tr>
</thead>
<tbody>
<tr>
<td>k=1</td>
<td>0.7</td>
</tr>
<tr>
<td>k=2</td>
<td>0.8</td>
</tr>
<tr>
<td>k=3</td>
<td>1.0</td>
</tr>
<tr>
<td>k=4</td>
<td>0.9</td>
</tr>
<tr>
<td>k=5</td>
<td>0.8</td>
</tr>
<tr>
<td>k=6</td>
<td>0.7</td>
</tr>
</tbody>
</table>

- PC values (PC = \( \frac{T_{PB}}{T_{PG}} \))
Experiments

- **Comparison of blocking without constraint to blocking with constraints (logical disjunction with 3 sigs).**

- **PP values (PP = \( \frac{T_{PB}}{P_{NR}} \))**

```

<table>
<thead>
<tr>
<th>k values</th>
<th>PP values</th>
</tr>
</thead>
<tbody>
<tr>
<td>k=1</td>
<td></td>
</tr>
<tr>
<td>k=2</td>
<td></td>
</tr>
<tr>
<td>k=3</td>
<td></td>
</tr>
<tr>
<td>k=4</td>
<td></td>
</tr>
<tr>
<td>k=5</td>
<td></td>
</tr>
<tr>
<td>k=6</td>
<td></td>
</tr>
</tbody>
</table>
```

- **Graph:**

  - Blocking without constraints
  - Blocking with constraints
Experiments

Comparison of blocking without constraint to blocking with constraints (logical disjunction with 3 sigs).

-PR values (PR = 1 - \( \frac{P_{NR}}{P_R} \))

```
0  0.2  0.4  0.6  0.8  1
```

k values

- blocking without constraints
- blocking with constraints
Experiments

- Comparison of blockings with different constraints (logical disjunction with 3 sigs vs logical conjunction with 2 sigs).

- PC values (PC = $\frac{T_{PB}}{T_{PG}}$)

---

**Figure:**

- PC values

- k values

- logical disjunction with 3 sigs

- logical conjunction with 2 sigs
Experiments

- Comparison of blockings with different constraints (logical disjunction with 3 sigs vs logical conjunction with 2 sigs).

- PP values ($PP = \frac{T_{PB}}{P_{NR}}$)

![Bar chart showing PP values for different k values and logical disjunction vs conjunction](chart.png)
Conclusion

- No big loss of PC values
- PP values are improved a great deal
- PR indicates how two kinds of cost are related, i.e. the cost of comparisons and the cost of redundancy checks.
- Tradeoff between PP and PC values can be adjusted by applying different semantic constraints
Any questions?