Name Disambiguation in AMiner:

Jing Zhang*, Jie Tang+

*Information School, Renmin University
+Computer Science Department, Tsinghua University
An Example in AMiner

All the researchers named “Jing Zhang”

Assign papers with author name “Jing Zhang” to right authors
Three Scenarios

Full ND
- Name disambiguation when the system is built from scratch

Continuous ND
- Name disambiguation when persons’ profiles are continuously updated

Error
- Error detection upon existing persons’ profiles
1st Scenario: Full ND

- Stage One: Constructing candidate sets of potential matchings.

Image Reference: Xin Luna Dong and Divesh Srivastava. Big data integration. Tutorial in ICDE’13, VLDB’13
Stage Two: Similarity Matching

- Constructing similarity matrix for each candidate set.

“Jing Zhang”

“Yi Li”

“Xiao Yang”

“Wei Zhang”

Image Reference: Xin Luna Dong and Divesh Srivastava. Big data integration. Tutorial in ICDE’13, VLDB’13
Stage Three: Clustering

- Partitioning each candidate set based on similarity.

Image Reference: Xin Luna Dong and Divesh Srivastava. Big data integration. Tutorial in ICDE'13, VLDB'13
**Feature-based Matching**

**Local features**
- Similarity between a paper and its cluster centroid according to papers’ attributes

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_i$.title</td>
<td>title of $p_i$</td>
</tr>
<tr>
<td>$p_i$.pubvenue</td>
<td>published conference/journal of $p_i$</td>
</tr>
<tr>
<td>$p_i$.year</td>
<td>published year of $p_i$</td>
</tr>
<tr>
<td>$p_i$.abstract</td>
<td>abstract of $p_i$</td>
</tr>
<tr>
<td>$p_i$.authors</td>
<td>authors name set of $p_i$ ${a^{(0)}_i, a^{(1)}_i, ..., a^{(k)}_i}$</td>
</tr>
<tr>
<td>$p_i$.references</td>
<td>references of $p_i$</td>
</tr>
</tbody>
</table>

**Correlation features**
- Similarities between two papers according to their relationships.

<table>
<thead>
<tr>
<th>Relation</th>
<th>Relation Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r_1$</td>
<td>CoPubVenue</td>
<td>$p_i$.pubvenue = $p_j$.pubvenue</td>
</tr>
<tr>
<td>$r_2$</td>
<td>CoAuthor</td>
<td>$\exists r, s &gt; 0, a^{(r)}_i = a^{(s)}_j$</td>
</tr>
<tr>
<td>$r_3$</td>
<td>Citation</td>
<td>$p_i$ cites $p_j$ or $p_j$ cites $p_i$</td>
</tr>
<tr>
<td>$r_4$</td>
<td>Constraint</td>
<td>feedback supplied by users</td>
</tr>
<tr>
<td>$r_5$</td>
<td>$\tau$-CoAuthor</td>
<td>$\tau$-extension co-authorship ($\tau &gt; 1$)</td>
</tr>
</tbody>
</table>
Embedding-based Matching

• Calculating similarities between embeddings.

**Global embeddings**

• Map every paper to a **unified representation space**.
• **Share supervision** across different candidate sets.

**Local embeddings**

• For each candidate set: build a graph by linking each two similar papers.
• Train a **graph auto-encoder** to learn **separate representation space**.
2nd Scenario: Continuous ND

- Papers come in a streaming fashion (500,000/month).
- Assigning new papers to right persons continuously.
- Stage One: Constructing candidates related to the author of the target paper.

Image Reference: Xin Luna Dong and Divesh Srivastava. Big data integration. Tutorial in ICDE'13, VLDB'13
Stage Two: Similarity Matching

• Matching the target paper and each candidate.
• Assigning the paper to the candidate with the largest matching score.

Image Reference: Xin Luna Dong and Divesh Srivastava. Big data integration. Tutorial in ICDE'13, VLDB'13
Interaction-based matching

- Calculating the similarities between the embeddings of each pairs of tokens in $p$ and $c$.
- Capturing both the exact and the soft matches.
3rd Scenario: Error Detection

• The accuracy of ND algorithms cannot be 100%.
• An additional error detection function is needed.
• Stage One: Extract patterns that can distinguish the right (normal) and wrong (abnormal) assigned papers.

Image Reference: Xin Luna Dong and Divesh Srivastava. Big data integration. Tutorial in ICDE'13, VLDB'13
Pattern Extraction

• Construct a **multi-relation egonet** for each candidate c.
  • Co-author, co-venue, citation relationships

• For each relation, extract:
  • Number of neighbors of ego c
  • Number of edges in c’s egonet
  • Total weight of c’s egonet
  • Principal eigenvalue of the weighted adjacency matrix of c’s egonet

User Feedbacks

Merge Function:
Merge person profiles

Add function:
Assign new papers to persons

Remove function:
Remove the wrongly assigned papers
Name Disambiguation in AMiner:

Jing Zhang*, Jie Tang+

*Information School, Renmin University
+Computer Science Department, Tsinghua University

Dataset: https://www.aminer.cn/na-data
AMiner: https://www.aminer.cn/