Improving Link Discovery using context-aware link specifications

PhD candidate
Andrea Cimmino

Supervised by
David Ruiz, University of Seville, Spain
Carlos R. Rivero, Rochester Institute of Technology, USA
Hi! My name is Andrea

- BARI
- ROCHESTER
- SEVILLE
To be or not to be ... the same

name: "Wei Wang"
email: "weiwang@cs.unc.edu"

name: "Wei Wang"
email: "wwang@unm.edu"

DATASET 1

DATASET 2
To be or not to be … the same

Link Specification (LS_{AR}): Levenshtein( name, full-name) \leq 0.42

name: “Wei Wang”

dataset 1

name: “Wei Wang”

dataset 2

full-name: “Wei Wang”
email: “weiwang@cs.unc.edu”

full-name: “Wei Wang”
email: “wwang@unm.edu”
To be or not to be … the same

Some publications in common?

Article

Paper

LS_{AR}

leads

supports

writes

Award
To be or not to be … the same

1. RDF, OWL

writes

Some publications in common?

LS_{AR}

leads

Awards

sponsors

Paper

Article
To be or not to be … the same

1. RDF, OWL
2. ≠ Vocabularies

Some publications in common?

Article

Paper

RDF, OWL

≠ Vocabularies

Some publications in common?

Article

Paper

To be or not to be … the same

1. RDF, OWL
2. ≠ Vocabularies

Some publications in common?

Article

Paper

To be or not to be … the same

1. RDF, OWL
2. ≠ Vocabularies

Some publications in common?

Article

Paper

To be or not to be … the same

1. RDF, OWL
2. ≠ Vocabularies

Some publications in common?

Article

Paper

To be or not to be … the same

1. RDF, OWL
2. ≠ Vocabularies

Some publications in common?

Article

Paper

To be or not to be … the same

1. RDF, OWL
2. ≠ Vocabularies

Some publications in common?

Article

Paper

To be or not to be … the same

1. RDF, OWL
2. ≠ Vocabularies

Some publications in common?

Article

Paper

To be or not to be … the same

1. RDF, OWL
2. ≠ Vocabularies

Some publications in common?

Article

Paper

To be or not to be … the same

1. RDF, OWL
2. ≠ Vocabularies

Some publications in common?

Article

Paper

To be or not to be … the same

1. RDF, OWL
2. ≠ Vocabularies

Some publications in common?

Article

Paper

To be or not to be … the same

1. RDF, OWL
2. ≠ Vocabularies

Some publications in common?

Article

Paper

To be or not to be … the same

1. RDF, OWL
2. ≠ Vocabularies

Some publications in common?

Article

Paper

To be or not to be … the same

1. RDF, OWL
2. ≠ Vocabularies

Some publications in common?

Article

Paper

To be or not to be … the same

1. RDF, OWL
2. ≠ Vocabularies

Some publications in common?

Article

Paper

To be or not to be … the same

1. RDF, OWL
2. ≠ Vocabularies

Some publications in common?

Article

Paper

To be or not to be … the same

1. RDF, OWL
2. ≠ Vocabularies

Some publications in common?

Article

Paper

To be or not to be … the same

1. RDF, OWL
2. ≠ Vocabularies

Some publications in common?

Article

Paper

To be or not to be … the same

1. RDF, OWL
2. ≠ Vocabularies

Some publications in common?

Article

Paper

To be or not to be … the same

1. RDF, OWL
2. ≠ Vocabularies

Some publications in common?

Article

Paper

To be or not to be … the same

1. RDF, OWL
2. ≠ Vocabularies

Some publications in common?

Article

Paper

To be or not to be … the same

1. RDF, OWL
2. ≠ Vocabularies

Some publications in common?

Article

Paper

To be or not to be … the same

1. RDF, OWL
2. ≠ Vocabularies

Some publications in common?

Article

Paper

To be or not to be … the same

1. RDF, OWL
2. ≠ Vocabularies

Some publications in common?

Article

Paper

To be or not to be … the same

1. RDF, OWL
2. ≠ Vocabularies

Some publications in common?

Article

Paper

To be or not to be … the same

1. RDF, OWL
2. ≠ Vocabularies

Some publications in common?

Article

Paper

To be or not to be … the same

1. RDF, OWL
2. ≠ Vocabularies

Some publications in common?

Article

Paper

To be or not to be … the same

1. RDF, OWL
2. ≠ Vocabularies

Some publications in common?

Article

Paper

To be or not to be … the same

1. RDF, OWL
2. ≠ Vocabularies

Some publications in common?
To be or not to be … the same

1. RDF, OWL
2. ≠ Vocabularies
3. Rule generation

ls_{AR}

leads

Award

Paper

Some publications in common?

Article

writes
To be or not to be … the same

1. RDF, OWL
2. ≠ Vocabularies
3. Rule generation
4. Context

Some publications in common?

Article

Paper

Article writes LS_{AR} Paper leads Award supports

Some publications in common?
Overlap Factor

Contex-Aware Link Specification:

FOR ALL Levenshtein( name, full-name) ≤ 0.42
AND
EXISTS Levenshtein (title, title) < 1.20
Applying $L_{AR}$

name: "Wei Wang"

email: "weiwang@cs.unc.edu"

The same?

full-name: "Wei Wang"

email: "wwang@unm.edu"

The same?
Applying $\text{LS}_{\text{AR}}$

- wrongly linked
- correctly linked

name: "Wei Wang"

The same

full-name: "Wei Wang"
email: "wwang@unm.edu"

wrongly linked correctly linked
Applying CALS

- title: “Efficient computation …”
  year: “2007”

- title: “HolisticTtwig…”

- name: “Wei Wang”

- title: “Efficient computation …”
  date: “2007”

- full-name: “Wei Wang”
  email: “weiwang@cs.unc.edu”

- The same?

- title: “Direct Oxidative Conversion…”
  date: “2012”

- full-name: “Wei Wang”
  email: “wwang@unm.edu”

- The same?
Applying CALS

- **name**: "Wei Wang"
  - full-name: "Wei Wang"
  - email: "weiwang@cs.unc.edu"
  - email: "wwang@unm.edu"
- **date**: "2007"
  - title: "Efficient computation …"
  - title: "HolisticTtwig…"
- **date**: "2012"
  - title: "Direct Oxidative Conversion…"
Roadmap

Problem statement

Results

Future work
# Experiments

## Scenarios

### Scenario 1 – DBLP-NSF

<table>
<thead>
<tr>
<th></th>
<th>DBLP</th>
<th>NSF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author</td>
<td>764</td>
<td>235</td>
</tr>
<tr>
<td>Article</td>
<td>47,225</td>
<td>235</td>
</tr>
<tr>
<td>Paper</td>
<td></td>
<td>6,877</td>
</tr>
<tr>
<td>owl:sameAs</td>
<td>Author ~ Researcher</td>
<td>188</td>
</tr>
</tbody>
</table>

### Scenario 2 – DBLP-DBLP

<table>
<thead>
<tr>
<th></th>
<th>DBLP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author</td>
<td>58</td>
</tr>
<tr>
<td>Article</td>
<td>5,284</td>
</tr>
<tr>
<td>owl:sameAs</td>
<td>Author ~Author</td>
</tr>
</tbody>
</table>
DBLP-NSF improving precision

**Link Specification (LS₁)**

![Graph showing the relationship between Threshold and Effectiveness for Link Specification (LS₁).](image)

**Context-Aware Link Specification CALS**

![Graph showing the relationship between Threshold and Effectiveness for Context-Aware Link Specification CALS.](image)

**LS₁:** \( \text{Jaro(name, full-name)} < \text{Threshold} \)

**CALS:** \( \text{for all BEST(LS₁) and exists Jaro(title, title)} < \text{Threshold} \)
DBLP-DBLP improving recall

Link Specification (LS₁)

Context-Aware Link Specification CALS

```
LS₁: \text{Jaro(name, name)} < \text{Threshold}
CALS: \text{for all Jaro(title, title)} < \text{Threshold}
```
DBLP-NSF GenLink evaluation results

<table>
<thead>
<tr>
<th>LS for DBLP-NSF</th>
<th>Examples(+-)</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LSN_1</td>
<td>(+1, -1)</td>
<td>Author ~ Researcher</td>
</tr>
<tr>
<td>LSN_5</td>
<td>(+5, -5)</td>
<td>Author ~ Researcher</td>
</tr>
<tr>
<td>LSN_{10}</td>
<td>(+5, -5)</td>
<td>Author ~ Researcher</td>
</tr>
<tr>
<td>LST_1</td>
<td>(+1, -1)</td>
<td>Article ~ Paper</td>
</tr>
<tr>
<td>LST_5</td>
<td>(+5, -5)</td>
<td>Article ~ Paper</td>
</tr>
<tr>
<td>LST_{10}</td>
<td>(+10, -10)</td>
<td>Article ~ Paper</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LS for DBLP-NSF</th>
<th>P</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSN_1</td>
<td>0.76</td>
<td>1.0</td>
</tr>
<tr>
<td>LSN_5</td>
<td>0.76</td>
<td>1.0</td>
</tr>
<tr>
<td>LSN_{10}</td>
<td>0.76</td>
<td>1.0</td>
</tr>
</tbody>
</table>

<p>| CALS for DBLP-NSF for link Author~ Researcher |</p>
<table>
<thead>
<tr>
<th>ID</th>
<th>P</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>for all LSN_1 and exists LST_1</td>
<td>0.94</td>
<td>1.0</td>
</tr>
<tr>
<td>for all LSN_5 and exists LST_5</td>
<td>1.0</td>
<td>0.38</td>
</tr>
<tr>
<td>for all LSN_{10} and exists LST_{10}</td>
<td>1.0</td>
<td>0.95</td>
</tr>
<tr>
<td>Best improvement</td>
<td>0.24</td>
<td></td>
</tr>
</tbody>
</table>
### DBLP-DBLP GenLink evaluation results

#### LS for DBLP-NSF

<table>
<thead>
<tr>
<th>ID</th>
<th>Examples (+/-)</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSN₁</td>
<td>(+1, -1)</td>
<td>Author ~ Author</td>
</tr>
<tr>
<td>LSN₅</td>
<td>(+5, -5)</td>
<td>Author ~ Author</td>
</tr>
<tr>
<td>LSN₁₀</td>
<td>(+5, -5)</td>
<td>Author ~ Author</td>
</tr>
<tr>
<td>LST₁</td>
<td>(+1, -1)</td>
<td>Article ~ Article</td>
</tr>
<tr>
<td>LST₅</td>
<td>(+5, -5)</td>
<td>Article ~ Article</td>
</tr>
<tr>
<td>LST₁₀</td>
<td>(+10, -10)</td>
<td>Article ~ Article</td>
</tr>
</tbody>
</table>

#### LS

<table>
<thead>
<tr>
<th>ID</th>
<th>P</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSN₁</td>
<td>1.00</td>
<td>0.26</td>
</tr>
<tr>
<td>LSN₅</td>
<td>1.00</td>
<td>0.30</td>
</tr>
<tr>
<td>LSN₁₀</td>
<td>1.00</td>
<td>0.26</td>
</tr>
</tbody>
</table>

#### CALS for DBLP-NSF for link Author ~ Author

<table>
<thead>
<tr>
<th>ID</th>
<th>P</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>for all LST₁</td>
<td>1.00</td>
<td>0.84</td>
</tr>
<tr>
<td>for all LST₅</td>
<td>1.00</td>
<td>0.84</td>
</tr>
<tr>
<td>for all LST₁₀</td>
<td>1.00</td>
<td>0.84</td>
</tr>
<tr>
<td>Best impr.</td>
<td></td>
<td>0.58</td>
</tr>
</tbody>
</table>
Roadmap

Problem statement

Results

Future work
Current work

- Article
- Paper

*WWW2017 Australia*
Future work
Future Future work

c-co-author  
\[\text{owl:sameAs}\]  
writes  
\[\text{owl:sameAs}\]  
leads  
\[\text{co-leads}\]  
\[\text{Award}\]  
supports  
\[\text{Paper}\]  
\[\text{Article}\]  
\[\ldots\]
THANKS! Queries?

Andrea Cimmino
cimmino@us.es
http://tdg-seville.info/acimmino
Features

♦ R1: Input RDF, not OWL.
♦ R2: Handle different schemas/vocabularies
♦ R3: Rule based (LS)
♦ R4: Context aware
♦ R5: Efficient context
## Related Work

<table>
<thead>
<tr>
<th>Technique</th>
<th>R1</th>
<th>R2</th>
<th>R3</th>
<th>R4</th>
<th>R5</th>
</tr>
</thead>
<tbody>
<tr>
<td>RiMOM</td>
<td>-</td>
<td>✓</td>
<td>-</td>
<td>✓</td>
<td>-</td>
</tr>
<tr>
<td>Nikolov et al.</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
</tr>
<tr>
<td>AgreementMaker</td>
<td>-</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
</tr>
<tr>
<td>GenLink</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CODI</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>✓</td>
<td>-</td>
</tr>
<tr>
<td>EAGLE</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>LOGMAP</td>
<td>-</td>
<td>✓</td>
<td>-</td>
<td>✓</td>
<td>-</td>
</tr>
<tr>
<td>Zhishi.links</td>
<td>-</td>
<td>✓</td>
<td>-</td>
<td>✓</td>
<td>-</td>
</tr>
<tr>
<td>SLINT+</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SignoProakis</td>
<td>✓</td>
<td>~</td>
<td>-</td>
<td>✓</td>
<td>~</td>
</tr>
<tr>
<td>SERIMI</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Song and Heflin</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PARIS</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>✓</td>
<td>-</td>
</tr>
<tr>
<td>Hassanzadeh et al.</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
<td>✓</td>
<td>-</td>
</tr>
</tbody>
</table>
# DBLP-NSF GenLink LS

<table>
<thead>
<tr>
<th>dblp:name, nsf:name</th>
<th>Jaccard ≤ 0.37</th>
<th>LSN&lt;sub&gt;1&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Jaccard ≤ 0.37</td>
<td>LSN&lt;sub&gt;5&lt;/sub&gt;</td>
</tr>
<tr>
<td></td>
<td>Jaccard ≤ 0.21</td>
<td>LSN&lt;sub&gt;10&lt;/sub&gt;</td>
</tr>
<tr>
<td>dblp:title, nsf:title</td>
<td>Levenshtein ≤ 29.48</td>
<td>LST&lt;sub&gt;1&lt;/sub&gt;</td>
</tr>
<tr>
<td></td>
<td>Levenshtein ≤ 0.59</td>
<td>LST&lt;sub&gt;5&lt;/sub&gt;</td>
</tr>
<tr>
<td></td>
<td>Levenshtein ≤ 7.05</td>
<td>LST&lt;sub&gt;10&lt;/sub&gt;</td>
</tr>
</tbody>
</table>
### DBLP-DBLP GenLink LS

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Jaccard ≤ 0.15</th>
<th>Levenshtein ≤ 1.48</th>
<th>Levenshtein ≤ 1.15</th>
<th>Levenshtein ≤ 1.76</th>
<th>Levenshtein ≤ 1.46</th>
<th>Levenshtein ≤ 1.76</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>dblp:name, nsf:name</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>LSN(_1)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>LSN(_5)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>LSN(_{10})</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>dblp:title, nsf:title</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>LST(_1)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>LST(_5)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>LST(_{10})</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Link Specification model

- LinkSpecification
  - source: Set<Class>
  - target: Set<Class>

- Condition
  - SameAsCondition
    - f: Function
    - threshold: Double
  - ConditionComposite
    - f: Aggregation

- Operand
  - LeafNode
    - prop: DataProperty
    - dataset: {SRC, TRG}
  - OperandComposite
    - f: Transformation
Link Specification extended (context)