COALA
Correlation-Aware Active Learning of Link Specifications

Axel-Cyrille Ngonga Ngomo  Klaus Lyko  Victor Christen

ESWC 2013, Montpellier, France
Outline

1. Motivation
2. Approach
3. Evaluation
4. Conclusion and Future Work
Why Link Discovery?

1. Fourth principle
2. Links are central for
   - Cross-ontology QA
   - Data Integration
   - Reasoning
   - Federated Queries
   - ...
3. Current topology of the LOD Cloud
   - 31+ billion triples
   - \( \approx 0.5 \) billion links
   - owl:sameAs in most cases
**Why is it difficult?**

1. **Time complexity**
   - Large number of triples
   - Quadratic a-priori runtime
   - 69 days for mapping cities from DBpedia to Geonames

**Definition (Link Discovery)**

- Given sets $S$ and $T$ of resources and relation $\mathcal{R}$
- Find $M = \{(s, t) \in S \times T : \mathcal{R}(s, t)\}$
- Common approaches:
  - Find $M' = \{(s, t) \in S \times T : \sigma(s, t) \geq \theta\}$
  - Find $M' = \{(s, t) \in S \times T : \delta(s, t) \leq \theta\}$
Why is it difficult?

2 Complexity of specifications

- Combination of several attributes required for high precision
- Tedious discovery of most adequate mapping
- Dataset-dependent similarity functions
Learning Complex Specifications

- **Supervised** (mostly active, e.g., RAVEN, EAGLE, SILK)
- Unsupervised (e.g., KnoFuss, EUCLID, EAGLE)
Learning Complex Specifications

- **Supervised** (mostly active, e.g., RAVEN, EAGLE, SILK)
- **Unsupervised** (e.g., KnoFuss, EUCLID, EAGLE)
Learning Complex Specifications

- **Supervised** (mostly active, e.g., RAVEN, EAGLE, SILK)
- Unsupervised (e.g., KnoFuss, EUCLID, EAGLE)
Learning Complex Specifications

- **Supervised** (mostly active, e.g., RAVEN, EAGLE, SILK)
- Unsupervised (e.g., KnoFuss, EUCLID, EAGLE)
Learning Complex Specifications

- **Supervised** (mostly active, e.g., RAVEN, EAGLE, SILK)
- Unsupervised (e.g., KnoFuss, EUCLID, EAGLE)
Learning Complex Specifications

- **Supervised** (mostly active, e.g., RAVEN, EAGLE, SILK)
- **Unsupervised** (e.g., KnoFuss, EUCLID, EAGLE)
Learning Complex Specifications

Insight

- Choice of right example is key for learning
- So far, only use of informativeness
Learning Complex Specifications

Insight
- Choice of right example is key for learning
- So far, only use of informativeness

Question
- Can we do better by using more information?
Learning Complex Specifications

**Insight**
- Choice of right example is key for learning
- So far, only use of informativeness

**Question**
- Can we do better by using more information?
  - Higher F-measure
  - Often slower
Outline

1. Motivation
2. Approach
3. Evaluation
4. Conclusion and Future Work
Basic Idea

- Use similarity of link candidates when selecting most informative examples
Basic Idea

- Use similarity of link candidates when selecting most informative examples
Basic Idea

- Use similarity of link candidates when selecting most informative examples
Similarity of Candidates

- Link candidate $x = (s, t)$ can be regarded as vector $(\sigma_1(x), \ldots, \sigma_n(x)) \in [0, 1]^n$.
- Similarity of link candidates $x$ and $y$:

\[
\text{sim}(x, y) = \frac{1}{1 + \sqrt{\sum_{i=1}^n (\sigma_i(x) - \sigma_i(y))^2}}. \quad (1)
\]

- Allows exploiting both intra- and inter-class similarity
Graph Clustering

- **Rationale**: Use intra-class similarity
- **Approach**
  - Cluster elements of $S^+$ and $S^-$ independently
  - Choose one element per cluster as representative
  - Present oracle with most informative representatives
BorderFlow

- $G = (V, E, \omega)$ with $V = S^+$ or $V = S^-$
- $\omega(x, y) = sim(x, y)$
- Keep best $ec$ edges for each $x \in V$
**BorderFlow**

- Seed-based algorithm
- Goal: Maximize borderflow ratio $bf(X) = \frac{\Omega(b(X),X)}{\Omega(b(X),n(X))}$
**BorderFlow**

- Seed-based algorithm
- Goal: Maximize borderflow ratio \[ bf(X) = \frac{\Omega(b(X), X)}{\Omega(b(X), n(X))} \]
**BorderFlow**

- Seed-based algorithm
- Goal: Maximize borderflow ratio \( bf(X) = \frac{\Omega(b(X),X)}{\Omega(b(X),n(X))} \)

http://sourceforge.net/projects/cugar-framework/
BorderFlow

• Seed-based algorithm
• Goal: Maximize borderflow ratio $bf(X) = \frac{\Omega(b(X), X)}{\Omega(b(X), n(X))}$
BorderFlow

- Seed-based algorithm
- Goal: Maximize borderflow ratio \( bf(X) = \frac{\Omega(b(X),X)}{\Omega(b(X),n(X))} \)

\[ X \quad \text{X} \quad \text{X} \quad \text{C}_f(X) \]

http://sourceforge.net/projects/cugar-framework/
**Spreading Activation**

- **Rationale**: Use both inter- and intra-class similarity
- **Approach**
  - $M_0 : m_{ij} = \text{sim}(x_i, x_j)$ with $(x_i, x_j) \in (S^+ \cup S^-)^2$
  - $A_0 : a_i = \text{ifm}(x_i)$
Spreading Activation

- **Rationale**: Use both inter- and intra-class similarity
- **Approach**
  - \( M_0 : m_{ij} = \text{sim}(x_i, x_j) \) with \((x_i, x_j) \in (S^+ \cup S^-)^2\)
  - \( A_0 : a_i = \text{ifm}(x_i) \)
  - \( A_t = A_{t-1} + M_{t-1}A_{t-1} \) (**spread activation**)
  - \( A_t = A_t / \max(A_t) \) (**normalize**)
  - \( M_t = M_{t-1} \) (**weight decay**)

\[
\begin{array}{c|c|c|c|c|c|c}
\text{Iterations} & 0 & 1 & 2 & 3 & \\
\text{Effect on \( M_t \)} & 0.8 & 0.9 & 0.7 & 0.8 & \\
\text{Effect on \( A_t \)} & 0.25 & 0.5 & 0.25 & 0.5 & \\
\end{array}
\]
**Spreading Activation**

- **Rationale**: Use both inter- and intra-class similarity
- **Approach**
  - $M_0 : m_{ij} = \text{sim}(x_i, x_j)$ with $(x_i, x_j) \in (S^+ \cup S^-)^2$
  - $A_0 : a_i = ifm(x_i)$
  - $A_t = A_{t-1} + M_{t-1}A_{t-1}$ (spread activation)
  - $A_t = A_t / \max(A_t)$ (normalize)
  - $M_t = M_{t-1}$ (weight decay)

![Diagram of Spreading Activation](image-url)
Outline

1. Motivation
2. Approach
3. Evaluation
4. Conclusion and Future Work
**Experimental Setup**

- Used EAGLE as active learning approach
  - Mutation and crossover rate $= 0.6$
  - Selection rate $= 0.7$
  - Not deterministic $\Rightarrow$ Ran each experiment 5 times
  - 5 queries to oracle per iteration
  - 10 iterations overall
  - 2 populations sizes: 20 and 100
  - 50 generations between iterations
- Two real-world and three synthetic datasets
- Single thread of a server (JDK1.7, Ubuntu 10.0.4, AMD Opteron 2GHz, 2GB/Experiment)
Parameters for WD

- Ran experiments on DBLP-ACM
- Population $= 20$
- $r \in \{2, 4, 8, 16, 32\}$
Parameters for CL

- Ran experiments on DBLP-ACM
- Population = 20
- ec ∈ \{1, 2, 3, 4, 5\}
**F-Scores**

- **Population = 100, final values**
- **Better results, yet unclear when to use WD or CL**

<table>
<thead>
<tr>
<th>DataSet</th>
<th>EAGLE</th>
<th>WD</th>
<th>CL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abt</td>
<td>0.19±0.04</td>
<td>0.25±0.04</td>
<td>0.23±0.04</td>
</tr>
<tr>
<td>DBLP</td>
<td>0.91±0.03</td>
<td>0.96±0.01</td>
<td>0.96±0.02</td>
</tr>
<tr>
<td>Person1</td>
<td>0.86±0.02</td>
<td>0.89±0.01</td>
<td>0.81±0.18</td>
</tr>
<tr>
<td>Person2</td>
<td>0.74±0.03</td>
<td>0.71±0.08</td>
<td>0.77±0.03</td>
</tr>
<tr>
<td>Restaurant</td>
<td><strong>0.89±0.0</strong></td>
<td>0.86±0.02</td>
<td><strong>0.89±0.0</strong></td>
</tr>
</tbody>
</table>
Outline

1. Motivation
2. Approach
3. Evaluation
4. Conclusion and Future Work
Conclusion and Future Work

- **Conclusion**
  - Presented correlation-aware learning of link specifications
  - Improved F-measures for both WD and CL
  - Longer runtimes (up to $2\times$)

- **Future Work**
  - Evaluation on other datasets
  - Effect of combination of CL with other graph clustering approaches
Conclusion

Thank You!

Questions?

Axel Ngonga
Augustusplatz 10
D-04109 Leipzig
ngonga@informatik.uni-leipzig.de
http://aksw.org/AxelNgonga
http://limes.sf.net
Demo at http://saim.aksw.org