1. Overview

Linked Open Data has encouraged dataset publication on the Web. Data interoperability and integration are challenging in this vast data space. These issues should be addressed at different levels, and property alignment is one such level. We focus on object-type property alignment and their relative distribution in LOD.

We propose a novel technique using property extensions to match equivalent properties between linked datasets and it outperforms existing syntactic and semantic techniques.

Challenges
- Properties capture meaning in RDF triples and hence are complex.
- Name heterogeneities exist in LOD.
- Property alignment requires complex algorithms that go beyond mere name analysis of the properties.

2. Extension based approach

Equivalent properties have exactly the same property extensions (OWL definition). In practice, high overlap decides equivalence, which we call Statistical Equivalence. Match count \( \mu(P_1, P_2) \) and Co-appearance count \( \lambda(P_1, P_2) \) gives the overlap.

\[
\text{Statistically Equivalent property pair } (P_1, P_2) \text{ has: } \\
\mu(P_1, P_2)/\lambda(P_1, P_2) \geq \alpha \\
\text{where, } \mu(P_1, P_2) \geq 1, \text{ and } 0 < \alpha \leq 1, k > 1
\]

Example

<table>
<thead>
<tr>
<th>Property Pair</th>
<th>Subject-Object Type</th>
<th>Property Name</th>
<th>Extension</th>
<th>( \mu(P_1, P_2) )</th>
<th>( \lambda(P_1, P_2) )</th>
<th>( \mu(P_1, P_2)/\lambda(P_1, P_2) )</th>
<th>( \alpha )</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Software)</td>
<td>Software</td>
<td>Software</td>
<td></td>
<td>0.8157</td>
<td>0.8157</td>
<td>1.0000</td>
<td>1.00</td>
</tr>
<tr>
<td>(Person)</td>
<td>Person</td>
<td>Person</td>
<td></td>
<td>0.7656</td>
<td>0.7656</td>
<td>1.0000</td>
<td>1.00</td>
</tr>
<tr>
<td>(Film)</td>
<td>Film</td>
<td>Film</td>
<td></td>
<td>0.3018</td>
<td>0.3018</td>
<td>1.0000</td>
<td>1.00</td>
</tr>
<tr>
<td>(Article)</td>
<td>Article</td>
<td>Article</td>
<td></td>
<td>0.8809</td>
<td>0.8809</td>
<td>1.0000</td>
<td>1.00</td>
</tr>
<tr>
<td>(Book)</td>
<td>Book</td>
<td>Book</td>
<td></td>
<td>0.3333</td>
<td>0.3333</td>
<td>1.0000</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Evaluation results of the four techniques used for the selected datasets of 1000 instances each. * = miki estimated values because of the number of pairs to be manually evaluated. Balanced numbers represent the highest results for each dataset. Clearly, the extension based approach produces favorable results because it is able to capture semantics other than analyzing property names for alignment.

3. Property pair types

Two orthogonal ways.

(i) On the basis of semantics

(a) Equivalent properties: Have high syntactic similarity in the property names.

(b) Property - sub property

(ii) On the basis of techniques and tools required for processing

(a) Simple property pairs: Similarity cannot be determined using property names alone. Requires extra processing like domain and range, and extensions. Are ambiguous, e.g., birthPlace vs. placeOfBirth.

(b) Complex property pairs: Similarity cannot be determined using property names alone. Requires extra processing like domain and range, and extensions. Are ambiguous, e.g., birthPlace vs. placeOfBirth.

4. Analysis

Property distribution

Datasets
- 3000 instances samples from DBpedia, Freebase, LinkedMDB, DBLP_L3S, and DBLP_RKB.
- They belong to Person, Film, Software, and Articles.
- Sim, Comp, and Syn stand for simple, complex, and synonymous property pairs respectively.

Property alignment

- Majority of the property pairs are simple followed by complex and synonymous.
- We expect to achieve higher accuracy using simple matching techniques as many of the pairs are simple.
- Also, many do not have exactly the same string in the property names.

- Syntactic or dictionary based approaches failed to produce good results, even though majority of the pairs are simple.
- It shows that properties are harder to process for alignment using string manipulation techniques.
- Extension based analysis captures the meaning of properties for better alignment and outperformed others.

5. Conclusion

- We classify and study types of property pairs that exist in LOD considering techniques required for equivalent property alignment.
- Our analysis shows that the majority of the pairs are simple property pairs, but as string manipulation based simple techniques did not perform well, they should be extended to improve coverage.
- Property extensions could be analyzed to determine equivalent properties between two datasets effectively. It achieved higher precision and recall in every case showing its applicability to Linked Datasets.

6. Acknowledgments


Kalpa Gunaratna, Krishnaprasad Thirunarayan, and Amit Sheth

Kno.e.sis Center, Wright State University, Dayton, OH USA

Kno.e.sis, Wright State University, Dayton, OH USA

Kalpa Gunaratna, Krishnaprasad Thirunarayan, and Amit Sheth

Kno.e.sis Center, Wright State University, Dayton, OH USA

Kalpa Gunaratna, Krishnaprasad Thirunarayan, and Amit Sheth

Kno.e.sis Center, Wright State University, Dayton, OH USA

Kalpa Gunaratna, Krishnaprasad Thirunarayan, and Amit Sheth

Kno.e.sis Center, Wright State University, Dayton, OH USA

Acknowledgement: Sponsored by NSF Award #IIS-1143717, “III: EAGER Expressive Scalable Querying over Linked Open Data.”
