Building Semantic Descriptions of Linked Data

Craig Knoblock
University of Southern California

Joint work with
Rahul Parundekar and José Luis Ambite
Linked Open Data and Services

- Vast collection of interlinked information
- Various sources and services with different schemas
Where do the Semantics Come From?

- **Linked Open Data**
  - Populated by manually linking or writing procedures that define the links across sources
  - But we don’t know how the sources are related
  - In many cases there is no or very limited semantic descriptions of sources
- **Linked Open Services**
  - Manually constructed or built by wrapping existing Web services
  - Constructing the lifting and lowering rules that relate the services to existing ontologies is a difficult task
  - Even when done, it may only provide a partial description
    - e.g., descriptions of the inputs and outputs, but not the function of a service
Outline of the Talk

• Linked Open Data
  • Building and linking ontologies of linked data
• Linked Open Services
  • Building semantic web services from the Deep Web
• Discussion
  • Remaining challenges
Outline of the Talk

- Linked Open Data
  - Building and linking ontologies of linked data
- Linked Open Services
  - Building semantic web services from the Deep Web
- Discussion
  - Remaining challenges
Building and linking ontologies of linked data [Parundeckar et al., ISWC 2010]

<table>
<thead>
<tr>
<th>Source 1</th>
<th>Source 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Schema Level</strong></td>
<td><strong>Instance Level</strong></td>
</tr>
<tr>
<td>City</td>
<td>City</td>
</tr>
<tr>
<td>Los Angeles</td>
<td>City of Los Angeles</td>
</tr>
</tbody>
</table>

owl:sameAs
Disjoint Schemas

<table>
<thead>
<tr>
<th>Source 1</th>
<th>Source 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Schema Level</strong></td>
<td></td>
</tr>
<tr>
<td>City</td>
<td>City</td>
</tr>
<tr>
<td></td>
<td>NO LINKS!!</td>
</tr>
<tr>
<td><strong>Instance Level</strong></td>
<td></td>
</tr>
<tr>
<td>Los Angeles</td>
<td>City of Los Angeles</td>
</tr>
<tr>
<td><strong>owl:sameAs</strong></td>
<td></td>
</tr>
</tbody>
</table>
Objective 1: Find Schema Alignments

Source 1

<table>
<thead>
<tr>
<th>Schema Level</th>
<th>Source 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>City</td>
<td>City</td>
</tr>
</tbody>
</table>

Instance Level

Los Angeles \(\text{owl:sameAs}\) City of Los Angeles
Ontologies of Linked Data

- Ontologies can be highly specialized
  - e.g. DBpedia has classes for *Educational Institutions*, *Bridges*, *Airports*, etc.

- Ontologies can be rudimentary
  - e.g. in Geonames all instances only belong to a single class – ‘Feature’
  - Derived from RDBMS schemas from which Linked Data was generated

- There might not exist exact equivalences between classes in two sources
• Only subset relations possible with difference in class specializations

<table>
<thead>
<tr>
<th>Geonames</th>
<th>DBpedia</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Schema Level</strong></td>
<td></td>
</tr>
<tr>
<td>Feature</td>
<td>Educational Institution</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Instance Level</strong></td>
<td></td>
</tr>
<tr>
<td>University of Southern California</td>
<td>University of Southern California</td>
</tr>
</tbody>
</table>

owl:sameAs
• A specialized class can be created by restricting the value of one or more properties

• The following Venn diagram explains a restriction class in Geonames with a restriction on the value of the `featureCode` property as ‘S.SCH’

Set of all instances in Original Class - rdf:type=Feature

Set of all instances in Restricted Class - rdf:type=Feature & featureCode=S.SCH
Objective 2: Find Alignments Between Restriction Classes

- Find and model specialized descriptions of classes

<table>
<thead>
<tr>
<th>Geonames</th>
<th>DBpedia</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Schema Level</strong></td>
<td></td>
</tr>
<tr>
<td>$\text{rdf:type}=$Feature &amp; $\text{featureCode}=S.SCH$</td>
<td>$\text{rdf:type}=\text{Educational Institution}$</td>
</tr>
<tr>
<td><strong>Instance Level</strong></td>
<td></td>
</tr>
<tr>
<td>University of Southern California</td>
<td>University of Southern California</td>
</tr>
</tbody>
</table>

$\text{owl:sameAs}$
Nature of Restriction Classes

• Instances belonging to a restriction class also belong to parent restriction class
  • e.g. restrictions from Geonames below

\[
\begin{align*}
& (\text{rdf:type}=\text{geonames:Feature}) & (\text{featureCode}=\text{geonames:A.PCLI}) & (\text{featureClass}=\text{geonames:A}) \\
& (\text{rdf:type}=\text{geonames:Feature} \& \text{featureClass}=\text{geonames:A}) \\
& (\text{rdf:type}=\text{geonames:Feature} \& \text{featureCode}=\text{geonames:A.PCLI}) & (\text{featureCode}=\text{geonames:A.PCLI} \& \text{featureClass}=\text{geonames:A}) \\
& (\text{rdf:type}=\text{geonames:Feature} \& \text{featureCode}=\text{geonames:A.PCLI} \& \text{featureClass}=\text{geonames:A})
\end{align*}
\]

• This also results in a hierarchy in the alignments, which our algorithm exploits
Extensional Approach to Ontology Alignment

- Represents set of instances belonging to ClassA
- Represents set of instances belonging to ClassB

ClassA is disjoint from ClassB

ClassA is equivalent to ClassB

ClassA is subset of ClassB

ClassB is subset of ClassA
Alignment Hypotheses

- An alignment hypothesis considers aligning
  - a restriction class from ontology $O_1$
  - another restriction class from ontology $O_2$

- Find relation between the two restriction classes
  - using extensional comparison on set of instances belonging to each restriction class
  - Use instance pair identifiers from pre-processing step (combination of URIs of linked instances)
Exploration of Hypotheses Search Space

Seed hypotheses generation

Seed hypothesis pruning (owl:Thing covers all instances)

Prune as no change in the extension set

Pruning on empty set

$r_2 = \emptyset$
Example Alignments from LinkedGeoData, Geonames, and DBpedia

<table>
<thead>
<tr>
<th>#</th>
<th>LINKEDGEO DATA restriction</th>
<th>DBPEDIA restriction</th>
<th>Relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>rdf:type=lgd:node</td>
<td>rdf:type=owl:Thing</td>
<td>$r_1 = r_2$</td>
</tr>
<tr>
<td>2</td>
<td>rdf:type=lgd:aerodrome</td>
<td>rdf:type=dbpedia:Airport</td>
<td>$r_1 = r_2$</td>
</tr>
<tr>
<td>3</td>
<td>rdf:type=lgd:island</td>
<td>rdf:type=dbpedia:Island</td>
<td>$r_1 = r_2$</td>
</tr>
<tr>
<td>4</td>
<td>lgd:gnis_3AST_alpha=NJ</td>
<td>dbpedia:Place#type=<a href="http://dbpedia.org/resource/City_(New_Jersey)">http://dbpedia.org/resource/City_(New_Jersey)</a></td>
<td>$r_1 = r_2$</td>
</tr>
<tr>
<td>5</td>
<td>rdf:type=lgd:village</td>
<td>rdf:type=dbpedia:PopulatedPlace</td>
<td>$r_1 \subset r_2$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>#</th>
<th>GEONAMES restriction</th>
<th>DBPEDIA restriction</th>
<th>Relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>geonames:featureClass=geonames:P</td>
<td>rdf:type=dbpedia:PopulatedPlace</td>
<td>$r_1 = r_2$</td>
</tr>
<tr>
<td>7</td>
<td>geonames:featureClass=geonames:H</td>
<td>rdf:type=dbpedia:BodyOfWater</td>
<td>$r_1 = r_2$</td>
</tr>
<tr>
<td>8</td>
<td>geonames:parentFeature=<a href="http://sws.geonames.org/3174618/">http://sws.geonames.org/3174618/</a></td>
<td>dbpedia:City_region=<a href="http://dbpedia.org/resource/Lombardy">http://dbpedia.org/resource/Lombardy</a></td>
<td>$r_1 = r_2$</td>
</tr>
<tr>
<td>9</td>
<td>geonames:featureCode=geonames:S,SCH</td>
<td>rdf:type=dbpedia:EducationalInstitution</td>
<td>$r_1 = r_2$</td>
</tr>
<tr>
<td>11</td>
<td>geonames:featureCode=geonames:T,MT</td>
<td>rdf:type=dbpedia:Mountain</td>
<td>$r_1 \subset r_2$</td>
</tr>
</tbody>
</table>
Outline of the Talk

• Linked Open Data
  • Building and linking ontologies of linked data
• Linked Open Services
  • Building semantic web services from the Deep Web
• Discussion
  • Remaining challenges
Building semantic web services from the Deep Web [Ambite et al., ISWC 2009]

- Automatically build semantic models for data and services available on the larger Web
- Construct models of these sources that are sufficiently rich to support querying and integration
  - Build models for the vast amount of structured and semi-structured data available
    - Not just web services, but also form-based interfaces
    - E.g., Weather forecasts, flight status, stock quotes, currency converters, online stores, etc.
  - Learn models for information-producing web sources and web services
Approach

- Start with some initial knowledge of a domain
  - Sources and semantic descriptions of those sources
- Automatically
  - Discover related sources
  - Determine how to invoke the sources
  - Learn the syntactic structure of the sources
  - Identify the semantic types of the data
  - Build semantic models of the source
  - Construct semantic web services
Integrated Approach

- Seed URL
- Sample input values

unisys(Zip, Temp, ...)
:- weather(Zip, ..., Temp, Hi, Lo)

http://wunderground.com

- Definition of known sources
- Sample values
- Patterns
- Domain types

unisys(Zip, Temp, Humidity, ...)

Source

Modeling

Semantic

Typing

Invocation

&

Extraction
Semantic Typing
[Lerman, Plangprasopchok, & Knoblock]

✓ Idea: Learn a model of the content of data and use it to recognize new examples

<table>
<thead>
<tr>
<th>Person</th>
<th>Address</th>
<th>Work</th>
</tr>
</thead>
<tbody>
<tr>
<td>E Lewis</td>
<td>3518 Hilltop Rd</td>
<td>(419) 531 - 0504</td>
</tr>
<tr>
<td>Andrew Lewis</td>
<td>3543 Larchmont Pkwy</td>
<td>(518) 474 - 4799</td>
</tr>
<tr>
<td>C. S. Lewis</td>
<td>555 Willow Run Dr</td>
<td>(612) 578 - 5555</td>
</tr>
<tr>
<td>Carmen Jones</td>
<td>355 Morgan Ave N</td>
<td>(612) 522 - 0555</td>
</tr>
<tr>
<td>John Jones</td>
<td>3574 Brookside Rd</td>
<td>(555) 531 - 9566</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Location</th>
<th>State_prov</th>
<th>Postal_code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Toledo</td>
<td>OH</td>
<td>64325-3000</td>
</tr>
<tr>
<td>Toledo</td>
<td>OH</td>
<td>64356</td>
</tr>
<tr>
<td>Seattle</td>
<td>WA</td>
<td>8422</td>
</tr>
<tr>
<td>Seattle</td>
<td>WA</td>
<td>8435</td>
</tr>
<tr>
<td>Omaha</td>
<td>NE</td>
<td>52456-6444</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>FullName</th>
<th>StreetAddress</th>
<th>Telephone</th>
</tr>
</thead>
<tbody>
<tr>
<td>E Lewis</td>
<td>3518 Hilltop Rd</td>
<td>(419) 531 - 0504</td>
</tr>
<tr>
<td>Andrew Lewis</td>
<td>3543 Larchmont Pkwy</td>
<td>(518) 474 - 4799</td>
</tr>
<tr>
<td>C. S. Lewis</td>
<td>555 Willow Run Dr</td>
<td>(612) 578 - 5555</td>
</tr>
<tr>
<td>Carmen Jones</td>
<td>355 Morgan Ave N</td>
<td>(612) 522 - 0555</td>
</tr>
<tr>
<td>John Jones</td>
<td>3574 Brookside Rd</td>
<td>(555) 531 - 9566</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>City</th>
<th>State</th>
<th>Zipcode</th>
</tr>
</thead>
<tbody>
<tr>
<td>Toledo</td>
<td>OH</td>
<td>64325-3000</td>
</tr>
<tr>
<td>Toledo</td>
<td>OH</td>
<td>64356</td>
</tr>
<tr>
<td>Seattle</td>
<td>WA</td>
<td>8422</td>
</tr>
<tr>
<td>Seattle</td>
<td>WA</td>
<td>8435</td>
</tr>
<tr>
<td>Omaha</td>
<td>NE</td>
<td>52456-6444</td>
</tr>
</tbody>
</table>
Inducing Source Definitions

Step 1: classify input & output semantic types

- **Known Source 1**
- **Known Source 2**
- **Known Source 3**

- **New Source 4**

source1($zip, lat, long) :-
    centroid(zip, lat, long).

source2($lat1, $long1, $lat2, $long2, dist) :-
    greatCircleDist(lat1, long1, lat2, long2, dist).

source3($dist1, dist2) :-
    convertKm2Mi(dist1, dist2).

source4($startZip, $endZip, separation)
Generating Plausible Definition

[Carman & Knoblock, 2007]

- Step 1: classify input & output semantic types
- Step 2: generate plausible definitions

source1($zip, lat, long) :-
    centroid(zip, lat, long).

source2($lat1, $long1, $lat2, $long2, dist) :-
    greatCircleDist(lat1, long1, lat2, long2, dist).

source3($dist1, dist2) :-
    convertKm2Mi(dist1, dist2).

source4($zip1, $zip2, dist) :-
    source1(zip1, lat1, long1),
    source1(zip2, lat2, long2),
    source2(lat1, long1, lat2, long2, dist2),
    source3(dist2, dist).

source4($zip1, $zip2, dist) :-
    centroid(zip1, lat1, long1),
    centroid(zip2, lat2, long2),
    greatCircleDist(lat1, long1, lat2, long2, dist2),
    convertKm2Mi(dist1, dist2).
Invoke and Compare the Definition

- Step 1: classify input & output semantic types
- Step 2: generate plausible definitions
- Step 3: invoke service & compare output

```prolog
source4($zip1, $zip2, dist):-
  source1(zip1, lat1, long1),
  source1(zip2, lat2, long2),
  source2(lat1, long1, lat2, long2, dist2),
  source3(dist2, dist).
```

```prolog
source4($zip1, $zip2, dist):-
  centroid(zip1, lat1, long1),
  centroid(zip2, lat2, long2),
  greatCircleDist(lat1, long1, lat2, long2, dist2),
  convertKm2Mi(dist1, dist2).
```

<table>
<thead>
<tr>
<th>$zip1</th>
<th>$zip2</th>
<th>dist (actual)</th>
<th>dist (predicted)</th>
</tr>
</thead>
<tbody>
<tr>
<td>80210</td>
<td>90266</td>
<td>842.37</td>
<td>843.65</td>
</tr>
<tr>
<td>60601</td>
<td>15201</td>
<td>410.31</td>
<td>410.83</td>
</tr>
<tr>
<td>10005</td>
<td>35555</td>
<td>899.50</td>
<td>899.21</td>
</tr>
</tbody>
</table>
**Constructing Semantic Web Services**

ForecastDay = one-of(0,1,2,3,4,5) ;;
0 is today, 1 is tomorrow, ...

DEIMOS generated
Web Service

z90292 hasName 90292 .

w0 hasZIP z90292 .
w0 hasTemp 61° F .
...
w1 hasZIP z90292 .
w1 hasTemp 61° F .

 Legend:

ontology

RDF Input

RDF output
Evaluation on Multiple Domains

Number of URLs

- Flight
- Geospatial
- Weather
- Currency
- Mutual Funds

Categories:
- Discovery
- Extraction
- Typing
- Modeling
### Accuracy of the Models

<table>
<thead>
<tr>
<th>domain</th>
<th>Precision</th>
<th>Recall</th>
<th>F₁-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>weather</td>
<td>0.64</td>
<td>0.29</td>
<td>0.39</td>
</tr>
<tr>
<td>geospatial</td>
<td>1.00</td>
<td>0.86</td>
<td>0.92</td>
</tr>
<tr>
<td>flights</td>
<td>0.69</td>
<td>0.35</td>
<td>0.46</td>
</tr>
<tr>
<td>currency</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>mutualfund</td>
<td>0.72</td>
<td>0.30</td>
<td>0.42</td>
</tr>
</tbody>
</table>
Outline of the Talk

• Linked Open Data
  • Building and linking ontologies of linked data
• Linked Open Services
  • Building semantic web services from the Deep Web
• Discussion
  • Remaining challenges
Discussion

- Initial work described here just scratches the surface of the problem
  - Goal is to both populate the Web of linked data and have rich semantic models of the data
  - Building semantic descriptions of linked open data will allow us to better understand the available sources and use the sources in a broad range of applications
  - Methods for automatically constructing linked open services will improve the coverage and quality of the sources available
Some Challenges

- **Linked Open Data**
  - How do we build an overall class hierarchy for a source
  - How do the relations map across sources
  - What do we do about missing and extraneous links

- **Linked Open Services**
  - How do we improve the accuracy of the learned semantic descriptions
  - How can we learn semantic descriptions that go beyond the current sources
  - How do we learn mappings between enumerated types (e.g., “Arrived” vs. “Landed”)