Pragmatic Semantics for the Web of Data

AlmWD -- Montpellier 2013
Stefan Schlobach
(based on work of and using slides from Christophe Gueret, Kathrin Denthler and Wouter Beek)
VU Amsterdam

Postulates

• The Web of Data requires semantics
• The Web of Data is not a database
• The Web of Data is a complex system
• Semantics for a database are not (always) suitable for complex systems
• We need new semantic paradigms
  — Voila: Pragmatic Semantics

Linked Data

► Graph/facts based knowledge representation
► Connect resources to properties / other resources
► Web-based: resources have a URI
  ● Try [http://dbpedia.org/resource/Amsterdam](http://dbpedia.org/resource/Amsterdam)

CLASSICAL SEMANTICS FOR THE WEB OF DATA

Model theory for Semantic Web

Languages: RDF, RDFS, OWL

• Ontology and Data: set of formulas S
• Model: formal structure satisfying all formulas in S
• Entailment: formula f entailed by S iff f in true in all models of S
• If contradiction, no models...
• No models, everything is entailed.

THE WEB OF DATA AS A COMPLEX SYSTEM
Since 2006, people are creating linked data

But publication and interpretation are distributed processes.

The Web of Data is a Complex System.
Not a database.
It is a Marketplace of ideas.
Key observations

- The Web of Data is more than the sum of its triples – it's a Complex System
- Different actors
- Different scales
- Dynamic

Evolution of the Web of Data

The WoD is a complex system!

- Countless extremely heterogeneous datasets
  - general-purposed datasets, such as DBpedia
  - domain-oriented datasets, such as Bio2RDF
  - government data, music data, geological data, social network data, etc.
- Hundreds of billions of RDF triples
  - Billions of links within the datasets
  - More than Million links between the datasets

Embedded rich semantics in the data
- data points are typed
- links are typed
- links is what makes the statements useful

Information has impact on different scales

A new way of seeing the WoD

Consider the WoD as network

Relevant (Network) Properties of WoD

- Average path length
- Degree distribution
- Strongly connected components
- Degree centrality
- Between centrality
- Closeness centrality
Scales of observation of the WoD

1. Graphs scale

- Each dataset is a node
- Edges are weighted, directed connections between the datasets
  - If there is at least one triple having a subject within dataset 1 and an object within dataset 2, then there is an edge between these two datasets.
  - The number of such triples is the weight of the edge.
- 110 nodes with 350 edges
- Average path length is 2.16
- 50 components

Top central nodes

<table>
<thead>
<tr>
<th>Node</th>
<th>Value</th>
<th>Node</th>
<th>Value</th>
<th>Node</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>DBpedia</td>
<td>0.332</td>
<td>DBpedia</td>
<td>0.762</td>
<td>DBpedia</td>
<td>0.505</td>
</tr>
<tr>
<td>DBLP Berlin</td>
<td>0.108</td>
<td>Geonames</td>
<td>0.614</td>
<td>UniProt</td>
<td>0.266</td>
</tr>
<tr>
<td>DBLP (RKB)</td>
<td>0.109</td>
<td>Drug Bank</td>
<td>0.576</td>
<td>DBLP (RKB)</td>
<td>0.206</td>
</tr>
<tr>
<td>DBLP Hannover</td>
<td>0.097</td>
<td>LinkedMDB</td>
<td>0.544</td>
<td>ACM (RKB)</td>
<td>0.229</td>
</tr>
<tr>
<td>FOAF profiles</td>
<td>0.075</td>
<td>Flickr wrappr</td>
<td>0.525</td>
<td>GeneID</td>
<td>0.211</td>
</tr>
</tbody>
</table>

Betweenness centrality | Closeness centrality | Degree centrality

Every centrality has a specific meaning...

Graph-scale WoD network

- Each dataset is a node
- Edges are weighted, directed connections between the datasets
  - If there is at least one triple having a subject within dataset 1 and an object within dataset 2, then there is an edge between these two datasets.
  - The number of such triples is the weight of the edge.
- 110 nodes with 350 edges
- Average path length is 2.16
- 50 components

The degree of 7 is a critical point after which the network is not scale-free any more.

Scales of observation of the WoD

2. Triple scale
Triple-scale WoD network

- We took the 10 million triples from the dataset crawled from the WoD, provided by the billion triple challenge 2009
- This "BTC" network is defined as $G=(V, (E, L))$, where
  - $V$ is a set of nodes, and each node is a URI or a literal
  - $E$ is a set of edges
  - $L$ is a set of labels, each label characterising a relation between nodes
- We applied a few strategies to aggregate data for comparison.

<table>
<thead>
<tr>
<th>Network</th>
<th>Nodes</th>
<th>Edges</th>
<th>Average path length</th>
<th>Components</th>
</tr>
</thead>
<tbody>
<tr>
<td>BTC</td>
<td>605K</td>
<td>860K</td>
<td>2.15</td>
<td>602K</td>
</tr>
<tr>
<td>BTC aggregated</td>
<td>14K</td>
<td>31K</td>
<td>2.80</td>
<td>7K</td>
</tr>
<tr>
<td>BTC aggregated + filter</td>
<td>37</td>
<td>91</td>
<td>1.88</td>
<td>17</td>
</tr>
</tbody>
</table>

Power-law distribution

BTC

BTC aggregated

Degree distribution

Monitoring and Improving the WoD

- Linked data is meant to be browsed, jumping from one resource to another
- The presence of Hubs is critical for the paths
- Create alternate paths to be used in case of failure

Guéret, Groth, van Harmelen, Schlobach, “Finding the Achilles Heel of the Web of Data: using network analysis for link recommendation”

Challenges:

- Multi-relations links
  - FOAF (social networks + personal information)
  - SIOC (relations characterising blogs)
  - SWRC (describing research work)
  - …
- Different filtering produce different networks
- Centrality status of nodes changes w.r.t the networks
- Dynamics
  - Data will be continuously added and linked.

The links have explicit semantics, which brings implicit links deduced after the reasoning process
FORMAL INTERACTIONS WITH THE WEB OF DATA

Interacting with Linked Data

When solutions do not (quite) fit the problem ...

Motivation

In the context of Web data?
- Issues with scale
- Issues with lack of consistency
- Issues with contextualised views over the World

Revise the goals
- As many answers as possible (or needed)
- Answers as accurate as possible (or needed)

From logic to optimisation

- Optimise towards the revised goals
- Need methods that cope with uncertainty, context, noise, scale, ...

Nature inspired methods for interacting with complex systems

- Advantageous properties
  - Adaptation
  - Simplicity
  - Interactivity: Anytime, user in the loop
  - Scalability and robustness
  - Good for dealing with dynamic information
- Studied for different interaction types
Answering queries over the data

The problem
- Match a graph pattern to the data
- Most common approach
  - Join partial results for each edge of the query

Solving approaches
- Logic-based
  - Find all the answers matching all of the query pattern
- Optimisation
  - Find answers matching as much of the query as possible
- Important implications of the optimisation
  - Only some of the answers will be found
  - Some of the answers found will be partially true

Evolutionary Computing
- Competition to survive in an environment with limited resources
- Inspired by theory of evolution (only best adapted can survive)

ERDF: An evolutionary algorithm under the hood

1. Input: Set of property/value pairs
2. Query: Web of Data
3. Initial Population
4. Results

Initial Population
- Randomly chosen to fit the query graph
**Properties of eRDF**

- Scalable
- Lean
- Robust
- Anytime
- Approximate
- Arbitrary SPARQL endpoints
- Join-free, so scaling to more endpoints is comparably pain free

**Some results**

- Tested on queries with varied complexity
- Works best with more complex queries
- Find exact answers when there are some
Finding implicit facts in the data

The problem
- Deduce new facts from others
- Most common approach
  - Centralise all the facts, batch process deductions

Solving approaches
- Logic-based
  - Find all the facts that can be derived from the data

- Optimisation
  - Find as many facts as possible while preserving consistency

- Important implications of the optimisation
  - Only some of the facts will be found
  - Unstable content

Collective Intelligence
- Individuals showing intelligence when acting as a group. Notion of emerging behaviour.
- Swarms inspired by flocks of birds, social insects (ants, bees, ...), schools of fish, ...

An optimisation approach: Swarms
- Swarm of micro-reasoners
  - Browse the graph, applying rules when possible
  - Deduced facts disappear after some time

Some results
- If they stay, most of the implicit facts are derived
- Ants need to follow each other to deal with precedence of rules
- Several ants per rule are needed
Related findings and approaches

- Storage optimisation using swarms (SwarmLinda from FU Berlin)
- Join optimisation with swarms (RCQ-ACS Erasmus Rotterdam)
- Emergent Semantics (eXascale Infolab Fribourg)
- Previous speaker (argumentation based semantics)

The day Semantics died…. ?

Part 4

PRAGMATIC SEMANTICS FOR THE WEB OF DATA

There is meaning in the structure

Requirements

- Standard languages
- Standard semantics still valid (for simple data)
- Integrate structural properties
  - Popularity of nodes/triples
  - “Distance” between triples
  - Frequency of triples

Semantics not strict, but pragmatic

Intuitively: a statement twenty times made is more true than a statement once made
Approach

- Entailment defined through optimality over different (possibly competing) notions of truth
- Make as much information in the data explicit, and turn it into first-class semantics citizens (truth orderings)
- Pragmatic entailment is defined through multi-objective optimisation.
- Interoperability is then achieved by enriching an ontology with meta-information about semantic orderings, as well as agreement on the weighting of orderings.

Subset based truth orderings

- the size of the minimal entailing subontology
- ratio of sub-models in which a formula is satisfied versus the total number of sub-models
- ratio between sub-ontologies of O in which a formula holds holds versus the number of all sub-ontologies

Graph-based truth orderings

- A shortest path ordering (diameter of the induced sub-graphs). Such a notion is a proxy for confidence of derivation. A
- A random-walk distance or edge-weights, induce orderings that are clustering-aware, with sub-ontologies entailing a formula have more cohesion than others.
- PageRank orderings can be used as proxies for popularity

Pragmatic Entailment

- A pragmatic closure C for an ontology O and orderings f1 to fn is then a set of formulas that is Pareto-optimal w.r.t. the optimisation problem max[f1 (C),...,fn (C)].

Deal with Open World Assumption

- Project title: Pragmatic Semantics for the Web of Data
- Acronym: PraSem
- Main researcher: Wouter Beek
- People involved: Stefan Schlobach, Christophe Gueret, Kathrin Denthler, Pepijn Kroes, Frank van Harmelen, and hopefully more people soon.
Deal with incompleteness

Formalise approximations

"Either Gödel has come up with a great new theory, or the most elaborate excuse in mathematical history for not finishing something."

Take home message

- The Web of Data requires semantics
- The Web of Data is not a database
- The Web of Data is a complex system
- Semantics for a database are not (always) suitable for complex systems
- We need new semantic paradigms
  - Voila: Pragmatic Semantics