Towards Linked Data Fact Validation through Measuring Consensus

Shuangyan Liu, Mathieu d’Aquín, Enrico Motta
Knowledge Media Institute (KMi), The Open University, UK
Shuangyan.Liu@open.ac.uk

Linked Data Quality Workshop @ ESWC
Linked Data Quality Assessment

• Data quality definition
  – “multi-dimensional construct with a popular definition ‘fitness for use’” (Zaveri et al. 2012)

• New challenges
  – Openness of the linked data
  – Diversity of the data
  – Dynamic set of autonomous data sources and publishers

• Linked data quality dimensions (Zaveri et al. 2012)
  – Accessibility dimensions
  – Representational dimensions
  – Contextual dimensions
  – Intrinsic dimensions

• Semantic accuracy dimension
  – “the degree to which data values correctly represent the real world facts” (Zaveri et al. 2012)
Problem Statement

• Assessing the semantic accuracy of the linked data
  – How to identify inaccurate facts or values in a linked dataset

• Existing approaches for detecting inaccurate values
  – SWIQA (Furber and Hepp, 2011)
    • functional dependency rules
    • limitation: dependencies must exist, can not be applied to arbitrary statement

  – DeFacto (Lehmann, 2012)
    • applied BOA framework to verbalise the RDF patterns from a training set. Then these patterns were used as queries made to search engines for retrieving relevant webpages and find proofs in the webpages
    • limitation: The training of BOA focused on object properties, no support of datatype properties
Approach

• **Measuring consensus** based on evidence triples collected from other linked datasets

• Following **owl:sameAs** links to collect triples describing same entities in other sources

• **Semantic relatedness** based predicate matching

• **Quantifying** the agreement among sources as an aggregated confidence score

• Compared to other approaches
  – no need to rely on dependency between values of properties (SWIQA)
  – Adequate support for datatype properties (complements DeFacto)
Overview

1. Collecting & Cleaning

2. Predicates & Objects Retrieving

3. Predicate Matching

4. Confidence Measuring

Target Triples

Subject Links

Data Set 1

Data Set 2

Data Set 3

sameas.org Service

Collect from

owl:sameAs
Subject Links Collecting and Cleaning

• Fetching equivalent subject links
  – Via the property `owl:sameAs` of the target fact triple
  – Via querying the [http://sameas.org](http://sameas.org) service

• Cleaning duplicated and non-resolvable subject links
  – By **pinging** the corresponding URIs
  – Removing **identical URIs**
  – Keeping **English** version of the same resource
Predicates and Objects Retrieving

- geonames.org
- ordnancesurvey.co.uk
- bbc.co.uk
- ... (more sources)

- yago-knowledge.org
- ... (more sources)

- wikidata.org
- ... (more sources)

Diagram:
- Jena API
- SPARQL query
- Knowledge Base Toolkit
- Subject Links
- Maps of DNs, PAOR modules

Collected Triples

DN = Domain Name
PAOR = Predicates and Objects Retrieving
Predicate Matching

• Semantic relatedness between predicates
  – Not based on string similarity
  – Different strings may have the same meaning (dbpedia-owl:populationTotal and yago:hasNumberOfPeople)

• The WuP [7] semantic relatedness measure is applied
  – considers the depths of the two synsets in the WordNet taxonomies

• The WS4J API is used to generate the pairwise similarity matrix for two input predicates (consisting of split compound words)
Predicate Matching

<table>
<thead>
<tr>
<th></th>
<th>has</th>
<th>Number of People</th>
</tr>
</thead>
<tbody>
<tr>
<td>population</td>
<td>0.0</td>
<td>0.4286</td>
</tr>
<tr>
<td>Total</td>
<td>0.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>

1. Pairwise semantic similarity matrix

\[
W(n) = \begin{cases} 
\max(S_{row}(n)) & \text{if } r \leq c \\
\max(S_{column}(n)) & \text{if } r > c 
\end{cases}
\]

2. Word Semantic Similarity

\[
P = \frac{\sum_{W \in \Phi(W)} W}{k} \text{ with } \exists W \in \Phi(W) \text{ and } W > \theta
\]

3. Predicate Similarity

Fact Validation | 1st June 2015 | Shuangyan Liu
Confidence Measuring

• Assign **weighting factors** to subject links to represent their reliability (based on provenance info)
• If the object of the target triple is string
  – **Aggregated score**: weighted average of string similarity between the object values of evidence triples and the target triple
• If the object of the target triple is numerical
  – **Subtracting the ratio** between difference in object values between target triple and the weighted average score and weighted avg score from 1

\[
\gamma = \frac{\sum_{i=1}^{m} \omega_i \cdot \nu_i}{\sum_{j=1}^{m} \omega_j}
\]

\[
\Gamma = 1 - \frac{\sqrt{(x - \gamma)^2}}{\gamma}
\]
This paper presents an approach for validating linked data facts by retrieving evidence triples from openly available knowledge bases. It incorporates the reliability of subject links and similarity of predicates and/or objects of evidence triples for representing the confidence of an input fact.

The rest of the paper is structured as follows. Section 2 presents the details of our approach. The method and results of an experiment with sample facts from DBpedia are described in Section 3. Finally, we conclude in Section 4 and provide an outlook for future work.

2 Approach

Pre-processing Fact Selection.

The first task of validating the accuracy of a fact is knowing what the fact is and how it is retrieved from a dataset. A fact is an RDF statement which consists of three parts: subject, predicate, and object. For example, the following RDF statement represents the fact that Milton Keynes (a large town in southeast England) has a total population of 229,941.

Table 1. Example of linked data fact.

<table>
<thead>
<tr>
<th>Subject Predicate Object</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://dbpedia.org/">http://dbpedia.org/</a></td>
</tr>
<tr>
<td>resource/Milton_Keynes</td>
</tr>
<tr>
<td><a href="http://dbpedia.org/">http://dbpedia.org/</a></td>
</tr>
<tr>
<td>ontology/populationTotal</td>
</tr>
<tr>
<td>229941</td>
</tr>
</tbody>
</table>

It is straightforward to obtain a single fact from a dataset. The object value of a fact can be queried using the subject and predicate of the fact. In this case, we just need to know what the subject and predicate are in advance.

It becomes a little complicated when gathering a large number of facts from a dataset for validation. This requires in-depth knowledge of the ontology of the dataset and the appropriate query to make. For example, in order to validate whether DBpedia holds accurate population information for all large towns in Milton Keynes, we need to retrieve all relevant facts from DBpedia as the first step. We can use the DBpedia SPARQL endpoint and make the following SPARQL query:

```
SELECT ?s dbpedia-owl:populationTotal ?o
WHERE {?s rdf:type dbpedia-owl:PopulatedPlace . ?s dbpprop:postcode-Area "MK"@en . OPTIONAL {?s dbpedia-owl:populationTotal ?o .} FILTER (bound(?o)) FILTER (?o > 10000)}
```

The above query obtains all towns in the postcode area of Milton Keynes that have a population more than 10,000. The property `dbpedia-owl:PopulatedPlace` represents "a place or area with clustered or scattered buildings and a permanent human population (city, settlement, town, or village) referenced with ge-

Experiment

- establish the DBpedia test set

```
1 http://dbpedia.org/resource/Bletchley_and_Fenny_Stratford dbpedia-owl:populationTotal 15313
2 http://dbpedia.org/resource/Wolverton_and_Greenleys dbpedia-owl:populationTotal 12492
3 http://dbpedia.org/resource/Bedford dbpedia-owl:populationTotal 80000
4 http://dbpedia.org/resource/Wolverton dbpedia-owl:populationTotal 12492
5 http://dbpedia.org/resource/Shenley_Church_End dbpedia-owl:populationTotal 12961
6 http://dbpedia.org/resource/West_Bletchley dbpedia-owl:populationTotal 22213
7 http://dbpedia.org/resource/Milton_Keynes dbpedia-owl:populationTotal 229941
8 http://dbpedia.org/resource/Campbell_Park dbpedia-owl:populationTotal 16402
9 http://dbpedia.org/resource/Porttlebury dbpedia-owl:populationTotal 15118
10 http://dbpedia.org/resource/Walton,_Milton_Keynes dbpedia-owl:populationTotal 11923
11 http://dbpedia.org/resource/Flitwick dbpedia-owl:populationTotal 12700
12 http://dbpedia.org/resource/Kempston dbpedia-owl:populationTotal 19440
13 http://dbpedia.org/resource/Buckingham dbpedia-owl:populationTotal 12043
14 http://dbpedia.org/resource/Biggleswade dbpedia-owl:populationTotal 13774
15 http://dbpedia.org/resource/Shenley_Brook_End dbpedia-owl:populationTotal 25828
16 http://dbpedia.org/resource/Thorpe_Ash dbpedia-owl:populationTotal 10084
17 http://dbpedia.org/resource/Bletchley dbpedia-owl:populationTotal 33950
18 http://dbpedia.org/resource/Great_Linford dbpedia-owl:populationTotal 19350
```
Results

• Initially collected 1349 subject links; after the cleaning process, 172 subject links remained
  – 907 non-resolvable links, 18 duplicated links, 252 links in other languages

• Improve the cleaning of the subject links process
  – Examined the causes of the non-resolvable links
Results

- After the Predicates and Objects retrieving process, **1793** triples collected (for 172 subject links).
- For each target triple, dozens to several hundred triples were collected from other sources.
- Predicate Similarity Distribution
  - **60%** not matched by the algorithm, **40%** matched in different levels by the algorithm
Results

• Confidence measurement
  – 45% assigned with a confidence score while 55% not
  – 22% of the target triples were identified by the algorithm as highly reliable

• Manually examined the correctness of the confidence scores
  – Low confidence scores due to wrong subject links
  – Need to extend the approach to identify “fake” subject links
Confidence Score Distribution

-68.58
0.04
0.7 ~ 0.8
0.9 ~ 1

no matched triples found

Fact Validation | 1st June 2015 | Shuangyan Liu
Conclusion and Future Work

• Our main contributions include a novel approach that enables checking the accuracy of RDF triples or facts.
• The contributions also include identifying matched or relevant RDF triples to a target triple.
• The approach would become increasingly important due to the fast growth of LOD.
• We are planning to demonstrate the approach can be proficiently applied to arbitrary predicates.
• We are also going to evaluate the predicate similarity matching method with standard evaluation measures (Precision/Recall) on well-known datasets.
• We want to explore the correlation between the confidence score assigned with the proposed method and the accuracy of the input fact.
Q & A

Thank You!