Open Linked Social Data for Scientific Benchmarking

Renato Fabbri[a,1,*], Osvaldo Novais de Oliveira Junior[a,1]

[a] *São Carlos Institute of Physics, São Paulo University, Brazil*

Abstract

The field of social network analysis and the topic of complex networks are widely researched. Recently, a myriad of results have been reported which are based in diverse datasets most often not accessible to other researchers. This work exposes an open dataset with diverse provenance and oriented to provide the scientific community a friendly and common repertoire. Current data was obtained from Facebook, Twitter, IRC, Email and the specific instances of ParticipaBR, AA and Cidade Democrática. These were translated to linked data format to homonenize access, conform to current best practices and ease analyzes which integrate third party and provided instances. This document presents an outline and overall statistics of given dataset which should favor subsequent work.

Keywords: Benchmark Data, Facebook, Twitter, IRC, Email, Complex Networks

1. Introduction

In recent years, the web of linked data [1] has attracted wide attention in both research and application realms. Linked data refers to data published in the web in such a way that it is machine readable and conforms to a set of best practices.

*Corresponding author

Email addresses: fabbri@usp.br (Renato Fabbri), chu@ifsc.usp.br (Osvaldo Novais de Oliveira Junior)

[1] *URL:* http://www.ifsc.usp.br/

1.1. Benchmarking in the analysis of complexity

The enormity of the digital data propels a rapid development of analysis methods from different perspectives. Karate club, what else?

2. Materials

2.1. Facebook data

2.2. Twitter data

2.3. IRC data

2.4. Email data

2.5. ParticipaBR data

2.6. AA data

2.7. Cidade Democrática data

3. Methods

3.1. Linked open data

The web of data is constructed with documents on the web such as the web of hypertext. In practice, the idea of linked data can me summarized by 1) the use of RDF to publish data on the web and 2) the use of RDF links to interlink data from different sources. The web is expected to be interconnected and to grow by the systematic application of four steps [1]:

- Use URIs to identify things [2].
- Use HTTP URIs.
- Provide useful information when an URI is accessed via HTTP.
- Provide other URIs in the description of resources so human and machine agents can perform discovery.

The Linked Open Data [3] builds an ever growing cloud of data which is usually conceived as centered around the DBPedia, a linked data representation of data from Wikipedia [4, 5].

3.2. RDF

The Resource Description Framework (RDF), a W3C recommendation, is a model for data interchange. It is based on the idea of making statements about resources in the form of triples, i.e. expressions in the form "subject - predicate - object". RDF can be serialized in several file formats, including RDF/XML, Turtle and Manchester all which, in essence, represent a labeled and directed multi-graph. RDF may be stored in a type of database called a triplestore [6].

As an example of an RDF statement, the following triple in the Turtle format asserts that "the paper has color white":

```
http://example.org/paper http://example.org/hasColor
http://example.org/White .
```

3.3. Data-driven ontology synthesis

OWL Ontologies are critical tools to describe taxonomies and the structure of knowledge. Most ontologies are created by domain experts even though the data they arrange is often given by a software system.

We developed an ontology synthesis method that probes the ontological structure in data with SPARQL queries and post-processing which can be divided in the following steps:

1. Obtaining all distinct classes with the query:
   ```
   SELECT DISTINCT ?class WHERE { ?s a ?class }
   ```

2. Obtaining all distinct properties with the query:
   ```
   SELECT DISTINCT ?p WHERE { ?s ?p ?o }
   ```

3. For each class, get distinct subject classes and predicates where the the object is an instance of the class:
   ```
   SELECT DISTINCT ?p ?cs WHERE { ?i a <class_uri> .  ?s ?p ?i .  ?s a
   ?cs .  }
   ```

4. For each class, get distinct predicates and object classes or datatypes where the subject is an instance of such class:

```
SELECT DISTINCT ?p ?co (datatype(?o) as ?do) WHERE { ?i a <class_uri>
.  ?i ?p ?o .  OPTIONAL { ?o a ?co .  } }
```

5. For each property, check if it is functional, i.e. if it occurs only once with each subject:
   ```
   SELECT DISTINCT (COUNT(?o) as ?co) WHERE { ?s <property_uri> ?o } GROUP
   BY ?s
   ```

6. For each property, find the incident range and domain with the queries:
   ```
   SELECT DISTINCT ?co (datatype(?o) as ?do) WHERE { ?s <property_uri>
   ?o .  OPTIONAL { ?o a ?co .  } }SELECT DISTINCT ?cs WHERE { ?s <property_uri>
   ?o .  ?s a ?cs .  }
   ```

7. For each instance of each class, get all distinct predicates. For each predicate, check if all instances of the class hold such relationship (existential restriction):
   ```
   SELECT DISTINCT ?p WHERE { ?s a <class_uri>.  ?s ?p ?o .  }
   SELECT DISTINCT ?s WHERE { ?s a <class_uri> }
   SELECT DISTINCT ?s ?co (datatype(?o) as ?do) WHERE {?s a <class_uri>.
   ?s <property_uri> ?o .  OPTIONAL {?o a ?co .  }}
   ```

8. and if all instances that hold such relationship are instances of the class (universal restriction):
   ```
   SELECT DISTINCT ?s WHERE { ?s <property_uri> ?o .  }
   ```

9. Draw each class, each property and the overall figure.

10. Make `rdfs:subClassOf` and `rdfs:subPropertyOf` statements to better organize knowledge and link to third party ontologies and data.

11.

References

[1] T. Berners-Lee, Design issues: Linked data (2006).

[2] L. Masinter, T. Berners-Lee, R. T. Fielding, Uniform resource identifier (uri): Generic syntax.

[3] J. Umbrich, S. Decker, M. Hausenblas, A. Polleres, A. Hogan, Towards dataset dynamics: Change frequency of linked open data sources.

[4] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, S. Hellmann, Dbpedia-a crystallization point for the web of data, Web Semantics: science, services and agents on the world wide web 7 (3) (2009) 154–165.

[5] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives, Dbpedia: A nucleus for a web of open data, in: The semantic web, Springer, 2007, pp. 722–735.

[6] R. Cyganiak, D. Wood, M. Lanthaler, Rdf 1.1 concepts and abstract syntax, W3C Recommendation 25 (2014) 1–8.