Exploiting Linked Data Datasets: a Summarization Based Approach

Honghan Wu[1], Boris Villazon-Terrazas[2], Jeff Z. Pan[1], and Jose Manuel Gomez-Perez[2]

[1] Department of Computing Science, University of Aberdeen, King's College, Aberdeen, AB24 3UE, UK,
{honghan.wu,jeff.z.pan}@abdn.ac.uk
[2] iSOCO, Intelligent Software Components S.A., Av. del Partenon, 16-18, 1-7, 28042, Madrid, Spain,
bvillazon,jmgomez@isoco.com

Abstract. In the last years, we have witnessed vast increase of Linked Data datasets not only in the volume, but also in number of various domains and across different sectors. However, due to the nature and techniques used within Linked Data, it is non-trivial work for normal users to quickly understand what is within the datasets, and even for tech-users to efficiently exploit the datasets. In this paper, we propose a summarisation based approach to guide the exploitation of Linked Data. Firstly, we introduce the details of the summarisation definition and generation. Then, we present the good properties of this summarisation and propose a set of useful services from the summarisation, both of which can be utilised to guide data exploitation tasks like query writing, ontology reasoning, data compression, and data diagnosis. Finally, we evaluate our approach in several typical data exploitation scenarios. Experiments on real word datasets show that our approach can guide very efficient data exploitation.

1 Introduction

Lately there is an increasing variety of Linked Data datasets coming from different data sources. However, these datasets are of varying quality ranging from extensively curated datasets to crowd-sourced or extracted data to often relatively low quality. This lack of data quality poses problems to developers aiming to seamlessly consume and integrate Linked Data in their applications.

There are some In [] authors group the linked data quality features into six main dimensions (1) conextual dimensions, (2) trust dimensions, (3) intrinsic dimensions, (4) accessibility dimensions, (5) accessibility dimensions, (6) representational dimensions, and (7) dataset dynamicity

Verifiability.- Verifiability refers to the degree by which a data consumer can access the correctness of a dataset and as consequence its trustworthiness

however, data in the LOD cloud has not been necessarily curated, and tools are required to detect possible ambiguities and quality problems ...

Linked Data has a number of challenges

There are many research works that try to define a set of key points for data quality [].

One important aspect of the Linked Data community is to measure the quality of a datasets. In spite *quality* is a subjective factor, we can list some key points regarding the quality of the dataset [], namely

- Accuracy - are facts actually correct?
- Intelligibility - are there human readable labels on things?
- Referential correspondence - are resources identified consistently without duplication?
- Completeness - do you have all the data you expect?
- Boundedness - do you have just the data you expect or is it polluted with irrelevant data?
- Typing - are nodes properly typed as resources or just string literals?
- Modeling correctness - is the logical structure of the data correct?
- Modeling granularity - does the modeling capture enough information to be useful?
- Connectedness - do combined datasets join at the right points?
- Isomorphism - are combined datasets modeled in a compatible way?
- Currency - is it up to date?
- Directionality - is it consistent in the direction of relations?
- Attribution - can you tell where portions of the data came from?
- History - can you tell who edited the data and when?
- Internal consistency - does the data contradict itself?
- Licensed - is the license for use clear?
- Sustainable - is there a credible basis for believing the data will be maintained?
- Authoritative- is the provider of the data a credible authority on the subject?

Linked Data quality measures might include:

- incoming and outgoing links
- used vocabularies, and properties
- adherence to property range restrictions, their values, etc.

However, the first step to get qualiaty of the data is get the snapshot, a summary, of the linked data dataset

2 Related Work

3 The Data Summarisation Approach

3.1 The EDP Graph

- EDP definition
  - The EDP graph construction
  - The statistics information

3.2 The Interactive User Interface

- visualisation (The online demo: the link needs to be given)
  - Interaction Operations

4 The Summary Base Data Exploitations

- Query Generation
  - Reasoning (DL-Lite materialisation of EDP graph)
  - Data Set Enrichment

5 Conclusions and Future Work

Acknowledgments

This work was supported by K-Drive (http://www.kdrive-project.eu/).

References