Linked Data Quality Assessment through Network Analysis

Christophe Guéret[1], Paul Groth[1], Claus Stadler[2], and Jens Lehmann[2]

[1] Free University Amsterdam, De Boelelaan 1105, 1081HV Amsterdam
{c.d.m.gueret,p.t.groth}@vu.nl
[2] University of Leipzig, Johannisgasse 26, 04103 Leipzig
{cstadler,lehmann}@informatik.uni-leipzig.de

Abstract. Linked Data is at its core about the setting of links between resources. Links provide enriched semantics, pointers to extra information and enable the merging of data sets. However, as the amount of Linked Data has grown, there has been the need to automate the creation of links and such automated approaches can create low-quality links or unsuitable network structures. In particular, it is difficult to obtain an overall picture as to whether the links introduced improve or diminish the quality of Linked Data. In this work, we present an extensible framework that allows for the assessment of Linked Data quality from a global perspective. We test the framework on a set of known quality links and show that it effectively detects quality changes.

Keywords: linked data, quality assurance, network analysis

1 Introduction

Linked Data features a distributed publication model that allows for any data publisher to semantically link to other resources on the Web. Because of this open nature, several mechanisms have been introduced to semi-automatically link resources on the Web of Data (e.g. Silk [4]). This partially automated introduction of links raises the question as to which links are improving the *quality* of the Web of Data or are just adding clutter. The notion of link quality is important on the Web of Data, particularly, because unlike the regular Web, there is not a human deciding based on context whether a link is useful or not. Instead, automated agents (with currently less capabilities) must be able to make these decisions. The quality of a link can be assessed in a number of different ways. Here, we want to look at the global ramifications of link creation on the Web of Data (or subsets of it).

In order to address this, we first must define quality at a global scale. Given that the Web of Data is a network, we can asses its global properties using network measures. These statistical techniques provide *summaries* of the network along different dimensions. These dimensions can be used to get an overall perspective on the quality of the network. [1] analyzed a number of networks in

nature and noted similar characteristics for all, including the web graph. Naturally, analysing quality via network measures has limitations: It is always possible to create an artificial meaningless data set, which would score high in several network criteria. For this reason, we view our approach as an addition, rather than a replacement, to other quality assurance methods.

Unfortunately, many network measures that can be used to check the quality of the Web of Data are computationally complex. This limits the applicability of these measures as the Web of Data continues to expand. However, instead of computing the exact value of network measures, one can compute some *local approximations* of them. We measure the quality of a set of Linked Data under change along using approximate network measures and compare the results against a set of goal statistics. For each measure, we provide the capability to identify the particular links that are causing the deviation from the ideal. This allows designers of link creation mechanisms to adjust their approaches using fine-grained information. Importantly, these measures are encapsulated in a framework, LINK-QA, which allows for the definition and addition of new quality measures. We now describe the framework and some initial experiments using it.

2 LINK-QA analysis framework

LINK-QA is a framework for assessing the quality of the Web of Data through the analysis of its constituent parts. The framework is scalable and extensible: the metrics applied are generic and share a common set of basic requirements, making it easy to incorporate new metrics. Additionally, metrics are computed using only the local network of a resource and are thus parallelizable by design. It differs from other approaches [3,2] in that it takes a network centric approach. The framework consists of five components, "**Select**", "**Construct**", "**Extend**", "**Analyse**" and "**Compare**". These components are assembled together in the form of a workflow. We now discuss each of the five components.

Select This component is responsible for selecting the set of resources to be evaluated. This can be done through a variety of mechanisms including sampling the Web of Data, using a user specified set of resources, or looking at the set of resources to be linked by a link discovery algorithm. These set of resources define the network under consideration.

Construct Once a set of resources is selected, a local network (i.e. a small neighborhood around a node) is constructed for each resource. The local networks are created by querying the Web of Data. Practically, LINK-QA makes use of either SPARQL endpoints or data files to create the graph surrounding a resource. In particular, sampling is achieved by first sending a SPARQL query to a list of endpoints. If no data is found, LINK-QA falls back on de-referencing the resource.

Extend Optionally, a set of edges can be provided as input to the framework, whose impact on the overall network should be measured. This component adds these input edges to each local network where they apply, and computes a set

of new local networks around the original set of selected resources. The impact assessment is done by the Compare component.

Analyse Once the original local network and its extended local networks have been created, an analysis consisting of two parts is performed: First, a set of metrics m is performed on each node v_i within each local network. This produces a set of metric results m_i for each node v_i. Then, these results are aggregated into a distribution. Currently, the following five metrics are used: degree, clustering coefficient, the number of open `owl:sameAs` chains in the local network, centrality and a measure of the richness of a resource description.

Compare The results coming from both analyses, before and after the addition of the new edges by the Extend component, are compared to ideal distributions for the different metrics. The comparison component assesses whether the set of new links globally improves the quality of the network. Note, that we can also run the framework without any additional input edges. This would just provide a quality assessment of the current network as represented by the selected resources.

The implementation of LINK-QA is available as free software on https://github.com/cgueret/LinkedData-QA and takes as input a set of resources, information from the Web of Data (*i.e.* SPARQL endpoints and/or dereferencable resources) and a set of new triples to perform quality assessment on. The output is an HTML report (see Figure 1).

Fig. 1. Example of report generated. Both this report and the links analysed are available at https://github.com/cgueret/LinkedData-QA/tree/master/example

3 Experimental Results and Conclusion

The framework is designed to analyse the potential impact of a set of link candidates prior to their publication on the Web of Data. The European project LOD Around the Clock (LATC) aims to enable the use of the Linked Open Data cloud for research and business purposes. LATC created and manually checked a set of reference linking specifications for the engine Silk [4]. The specifications along with the link sets they produce are publicly available[3]. The goal of the experiments is to see whether the links created improve, worsen or don't affect the quality of the WoD as defined by the different metrics.

Table 1 shows the detection of changes achieved by the metrics over 14 heterogeneous linking specifications, establishing links among entities from DBpedia, GHO, LinkedCT and Eunis. The metrics are sensitive for at least 80% of the tests. The changes detected contribute to obtaining a power-law distribution of the degree and increasing the overall descriptive richness. From these results, we conclude that LINK-QA is able to detect global changes related to the addition of a set of heterogeneous links.

	Degree	**Clustering**	**sameAs**	**Centrality**	**Description**
Red	35.7% ($1.87 \pm 1.38\%$)	64.3% ($0.21 \pm 0.17\%$)	0.1% ($6.67 \pm 0\%$)	71.4% ($0.21 \pm 0.15\%$)	0%
Green	42.9% ($-2.14 \pm 1.87\%$)	14.3% ($-0.05 \pm 0.02\%$)	0%	0.1% ($-0.05 \pm 0\%$)	78.6% ($-71.78 \pm 12.60\%$)
N.A.	21.4%	21.4%	99.9%	28.5%	21.4%

Table 1. Average changes detected by the metrics reported in terms of status with respect to the ideal. Red: distance to ideal increased. Green: decreased. N.A.: no change. The values in parenthesis are the actual average value returned by the metrics.

LINK-QA is an extensible framework for performing quality assessment on the Web of Data. We analysed a set of links provided by well-known link creation services and showed how the framework can be used to detect change in quality. Going forward, we aim to develop a live service for running quality assessment over the whole of the Web of Data on a periodic basis.

Acknowledgements This work was supported by the EU 7th Framework Programme within the projects LOD2 (GA no. 257943) and LATC (GA no. 256975). The authors would like to thank Peter Mika for his input.

References

1. Barabási, A.L.: Linked. (Perseus, Cambridge, Massachusetts) (2002)
2. Bizer, C., Cyganiak, R.: Quality-driven information filtering using the WIQA policy framework. Web Semantics: Science, Services and Agents on the World Wide Web 7(1), 1–10 (Jan 2009)
3. Niu, X., Wang, H., Wu, G., Qi, G., Yu, Y.: Evaluating the stability and credibility of ontology matching methods. In: 8th Extended Semantic Web Conference (ESWC2011) (June 2011)
4. Volz, J., Bizer, C., Gaedke, M., Kobilarov, G.: Silk: A link discovery framework for the web of data. In: Bizer, C., Heath, T., Berners-Lee, T., Idehen, K. (eds.) 2nd Linked Data on the Web Workshop LDOW2009. pp. 1–6. CEUR-WS (2009)

[3] https://github.com/LATC/24-7-platform/tree/master/link-specifications