TAA: A Platform for Triple Accuracy Measuring and Evidence Triples Discovering

Shuangyan Liu[1], Carlo Allocca[1], Mathieu d'Aquin[2], and Enrico Motta[1]

[1] Knowledge Media Institute, The Open University, UK
{shuangyan.liu, carlo.allocca, enrico.motta}@open.ac.uk
[2] Insight Center Galway, Ireland
mathieu.daquin@insight-centre.org

Abstract. Knowledge graph refinement is often a time consuming process due to the lack of domain knowledge and automatic tools to support maintainers to detect erroneous information. Few tools are available to support the task of measuring the accuracy of triples (source triples) in a knowledge graph. We developed a platform which we call TAA for measuring triple accuracy by discovering evidence triples from external knowledge graphs. It consists of a quality assessment pipeline which contains a series of components from fetching target triples from external knowledge graphs to finding matched triples among the target triples and calculating a score to represent the level of accuracy of a source triple. In addition, TAA represents the assessment result using an evidence graph and exposes its functionality to other applications via REST web service.

Keywords: Data Quality, Knowledge Graph Accuracy, Quality Assessment Pipeline, Evidence Graph, RESTful Web Service

1 Introduction

In recent years research communities and industrial stakeholders have constructed many large-scale knowledge graphs such as DBpedia[3], YAGO[4], Freebase[5], Wikidata[6], Google Knowledge Graph, Microsoft Satori, and others. They are intensively used in different application scenarios such as search, question answering, natural language processing, data integration and analytics, and for specialised areas such as digital humanities, business, life science and more. Due to the diversity of data sources and limitations of present knowledge graph construction methods, most knowledge graphs face a variety of quality issues such as noise and vague data, inconsistency, inaccurate and out-of-date data, incomplete information, and poor interlinking between KGs. To facilitate wide adoption and advanced usage, it is crucial to ensure the quality of knowledge graphs.

[3] http://wiki.dbpedia.org/
[4] http://yago-knowledge.org/
[5] https://developers.google.com/freebase/
[6] http://www.wikidata.org

Various studies have been conducted to investigate the quality of popular knowledge graphs and the quality measures along different dimensions [1] [3]. However, not enough tools have been developed for assisting the task of triple accuracy evaluation. We developed a system (TAA) to assist knowledge graph maintainers and developers to verify the accuracy of statements from a knowledge graph. The TAA platform allows automatically discovering evidence triples from external knowledge graphs, and assigns a confidence score to a source triple representing consensus of the evidence triples obtained.

The approach used in TAA consists of three phases. First, target triples which have equivalent subjects of a source triple are fetched by querying the *sameAs* links of the subject link of the source triple using *sameAs.org* service and the source knowledge graph, and then retrieving predicates and objects of the target triples from external knowledge graphs through content negotiation mechanism. Second, we developed a predicate matching mechanism based on predicate semantic similarity calculation and predicate type and value comparison to find matched triples of a source triple. Finally, a confidence score to represent the accuracy level of a source triple is calculated using the consensus of the matched triples. A detailed description of the approach is presented in [4]. In this paper, we focus on explaining the design principles, architecture and functionality of the TAA tool. We also present an analysis of other tools and systems for tackling similar quality problems.

2 The TAA Platform

The design principles of the TAA platform are to reduce the human workload in the process of knowledge graph accuracy assessment. We achieve this goal by (i) providing a pipeline of evidence triples discovering and accuracy measuring; (ii) minimising the user interactions with the platform for completing an evaluation task; (iii) representing the assessment results in an intuitive way. The architecture of TAA is shown in Figure 1. We designed and implemented the TAA platform using the Oracle Jersey RESTful Web Service framework[7]. Therefore, TAA not only provides a web application portal but also a REST web service to support the triple accuracy assessment process. The Grizzly HTTP server[8] is adopted in TAA for providing a container of REST web resources and non-REST web resources. The Grizzly server is chosen since it is fast, lightweight and supports features such as non-blocking IO and NIO buffers which makes TAA scalable to a large number of users. In addition, JAX-RS[9] resources are developed to contain methods to handle HTTP requests for performing triple accuracy assessment tasks.

The core functionality of TAA is discovering evidence triples for a source triple in external knowledge graphs. In order to achieve this objective, a pipeline is developed to contain a series of components adopting different technology

[7] https://jersey.github.io/
[8] https://javaee.github.io/grizzly/
[9] https://github.com/jax-rs

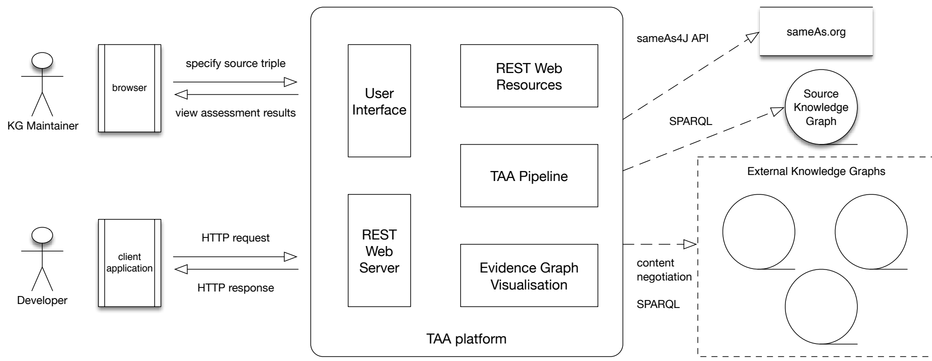

Fig. 1. Architecture of the TAA platform

including semantic web and linked data technology, predicate semantic similarity measures, natural language processing and outlier detection algorithms. The pipeline components are responsible for fetching and identifying matched triples and generating a confidence score of the source triple based on agreement among matched triples.

In addition, TAA represents the evidence triples discovered using an evidence graph. An evidence graph is a network of a source triple and its evidence triples, and the types of identity links between the source subject and matched subjects. We implemented the evidence graph using the D3 JavaScript API.[10] The data of the evidence graph is represented in the D3 graph JSON format, parsed and visualised using the D3 API. An example of the evidence graph visualisation is shown in Figure 2. Different colours are applied to nodes of the evidence graph for distinguishing a source triple (blue) and its evidence triples (orange).

Different types of users can benefit from TAA. Knowledge graph maintainers who need to verify the accuracy of triples can use TAA to collect evidence from external KGs and obtain the confidence level of how accurate the source triples are. Developers who build quality assessment applications can invoke the functionality of TAA using HTTP from their own applications, since TAA is developed as a REST web service that exposes the functionality of TAA over the web. The RESTful design also allows TAA to represent the assessment results in different formats such as JSON and XML depending on the type of client.

The TAA UI currently supports the validation of a single source triple at the same time. However, the core data processing pipeline of TAA is capable of validating a cluster of homogeneous triples simultaneously which have the same predicates. In order to verify a cluster of triples at the same time, the data input and the result visualisation components of the TAA platform can be extended correspondingly. The TAA platform source code is available on GitHub.[11]

[10] `https://github.com/d3/d3/wiki`
[11] `https://github.com/TriplesAccuracyAssessment/taa-demo`

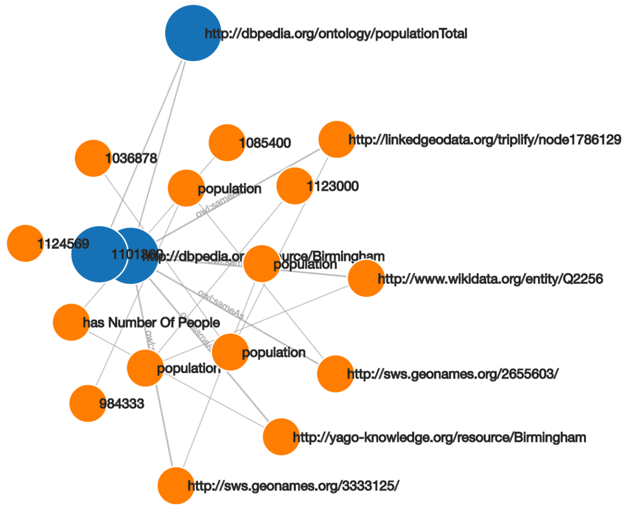

Fig. 2. An evidence graph as displayed in the TAA platform.

3 Related Work

Not enough research has been provided regarding developing quantitative measures of accuracy and automatic tools for assessing triple accuracy. DeFacto [3] was developed for RDF fact validation. The difference between DeFacto and TAA is that DeFacto retrieves web documents as evidence for verifying facts. [2] adopted an outlier detection method to identify numerical errors in KGs. Compared to TAA, [2] did not address the matching of similar properties for two triples. In the TAA pipeline, a predicate matching algorithm is used for matching similar properties of triples.

References

1. Färber, M., Ell, B., Menne, C., Rettinger, A., Bartscherer, F.: Linked data quality of dbpedia, freebase, opencyc, wikidata, and yago. Semantic Web Journal (to appear)
2. Fleischhacker, D., Paulheim, H., Bryl, V., Völker, J., Bizer, C.: Detecting errors in numerical linked data using cross-checked outlier detection. In: Mika, P., Tudorache, T., Bernstein, A., Welty, C., Knoblock, C., Vrandečić, D., Groth, P., Noy, N., Janowicz, K., Goble, C. (eds.) ISWC 2014, Part I. LNCS, vol. 8796, pp. 357–372. Springer, Cham (2014)
3. Gerber, D., Esteves, D., Lehmann, J., Bühmann, L., Usbeck, R., Ngomo, A.C.N., Speck, R.: Defacto—temporal and multilingual deep fact validation. Web Semantics: Science, Services and Agents on the World Wide Web 35, 85–101 (2015)
4. Liu, S., d'Aquin, M., Motta, E.: Measuring accuracy of triples in knowledge graphs. In: Gracia, J., Bond, F., McCrae, J., Buitelaar, P., Chiarcos, C., Hellmann, S. (eds.) International Conference on Language, Data and Knowledge. Lecture Notes in Computer Science, vol. 10318, pp. 343–357. Springer, Cham (2017)