12η επιστημονική ΔΙΗΜΕΡΙΔΑ ΙΤΕ 2019

Services for Connecting and Integrating Big Number of Datasets

Michalis Mountantonakis[1,2#], and **Yannis Tzitzikas** [1,2*]

[1] Institute of Computer Science, FORTH-ICS, Greece

[2] Computer Science Department, University of Crete, Greece

# Presenting author: Michalis Mountantonakis, email:mountant@ics.forth.gr
* Corresponding author: Yannis Tzitzikas, email: tzitzik@ics.forth.gr

ABSTRACT

Linked Data is a method for publishing structured data that allows them to be interlinked (by using URIs instead of simple values) for assisting their integration. A big number of such datasets (or sources), has already been published according to the principles of Linked data and their number and size keeps increasing. The large volume of the available linked data necessitates their connectivity, the preservation of their provenance, and the assessment of their quality and veracity, for fulfilling the requirement of e-science, which is, nowadays, one of the biggest challenges of computer science. The processing and the analysis of a large volume of data is crucial for any scientific field, for providing novel and accurate scientific results, thereby, it is necessary for the scientists to be able to discover fast reliable sources and data, which are connected and are related to their research.

Towards this direction, the Information Systems Laboratory of the Institute of Computer Science of FORTH has proposed methods and services, supported by special indexes and algorithms, for aiding the process of semantic data integration at large scale. More specifically, (i) we have studied the different dimensions of the integration process for large number of datasets from various aspects [1], (ii) we have proposed scalable methods and algorithms for constructing semantics-aware indexes (which takes into account the equivalences among several datasets), which enables the fast access to all the available information of any entity [2], (iii) we have proposed scalable methods for measuring the connectivity among two or more datasets, by using novel lattice-based incremental algorithms [3], and (iv) we have introduced data enrichment methods for improving several tasks (e.g., machine-learning based tasks).

For exploiting the aforementioned techniques and metrics, we offer through our research prototype, called LODsyndesis [2,3], a number of advanced global-scale services. In particular, we offer services, (a) for obtaining complete information about one particular entity (or a set of entities), (b) for discovering datasets which are relevant to another one or/and to discover which datasets are the most appropriate for a given task, (c) for assessing and improving Data Quality, i.e., assessing the connectivity between any set of datasets and monitoring their evolution over time, (d) for combining information from several datasets, e.g., for improving Machine-Learning based tasks, and (e) for estimating the reliability of a specific fact for any entity.

REFERENCES

[1] Mountantonakis, M., & Tzitzikas, Y. (2019). Large-scale Semantic Integration of Linked Data: A Survey. ACM Computing Surveys (CSUR), 52(5), 103

[2] Mountantonakis, M., & Tzitzikas, Y. (2018). High performance methods for linked open data connectivity analytics. Information, 9(6), 134.

[3] Mountantonakis, M., & Tzitzikas, Y. (2018). Scalable Methods for Measuring the Connectivity and Quality of Large Numbers of Linked Datasets. *Journal of Data and Information Quality (JDIQ)*, *9*(3), 15.