Large Scale Services for Connecting and Integrating Hundreds of Linked Datasets

Michalis Mountantonakis
Institute of Computer Science, FORTH-ICS, Greece and
Computer Science Department, University of Crete, Greece
mountant@ics.forth.gr

Michalis Mountantonakis is a Postdoctoral Researcher of the Information Systems Laboratory at FORTH-ICS (Greece) and a Visiting Lecturer in the Computer Science Department at University of Crete (CSD-UoC), Greece. He obtained his PhD degree from the CSD-UoC in 2020. His research interests fall in the areas of large-scale semantic data integration, linked data and semantic data management. The results of his research have been published in more than 20 research papers. For his dissertation, he awarded a) the prestigious SWSA Distinguished Dissertation Award 2020, which is given to the PhD dissertation from the previous year with the highest originality, significance, and impact in the area of semantic web, and b) the Maria Michael Manasaki Legacy's fellowship, which is awarded to the best graduate student of CSD-UoC, once a year.

In his dissertation, supervised by Associate Professor Yannis Tzitzikas (Computer Science Department at University of Crete), Michalis Mountantonakis dealt with the problem of Linked Data Integration at large scale, which is a very big challenging problem. He factorized the integration process according to various dimensions, for better understanding the overall problem and for identifying the open challenges, and proposed novel indexes and algorithms for providing core services, which can be exploited for several tasks related to Data Integration, such as: for finding all the URIs and all the available information for an entity, for producing connectivity analytics, for discovering the most relevant datasets for a given task, for dataset enrichment, and many others.

1. MOTIVATION

Linked Data is a method for publishing structured data that facilitates their sharing, linking, searching and re-use. A big number of such datasets (or sources), has already been published and their number and size keeps increasing. Although the main objective of Linked Data is linking and integration, this target has not yet been satisfactorily achieved. Even seemingly simple tasks, such as finding all the available information for an entity is challenging, since this presupposes knowing the contents of all datasets and performing cross-dataset identity reasoning, i.e., computing the symmetric and transitive closure of the equivalence relationships that exist among entities and schemas. Another big challenge is Dataset Discovery, since current approaches exploit only the metadata of datasets, without taking into consideration their contents. Due to the aforementioned problems, the execution of various tasks that are related to data integration at large scale is quite difficult. These tasks are listed below and are depicted in Fig. 1.

• Task A (*Object Coreference*). Obtaining complete information about one particular en-

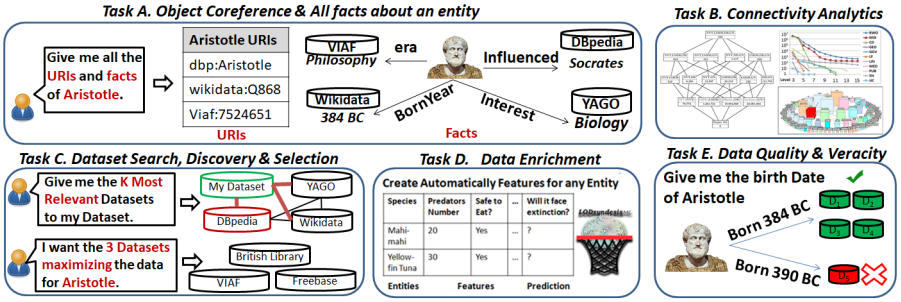

Fig. 1. LODsyndesis-based services for tasks A-E

tity or a set of entities by applying cross-dataset identity reasoning, e.g., "find all the URIs and facts of Aristotle" (see the upper left side of Fig. 1).

• Task B (*Connectivity Analytics*). Assessing the connectivity among any combination of datasets, and monitoring their evolution over time, e.g., see the upper right side of Fig. 1.

• Task C (*Dataset Search, Discovery and Selection*). Discovering the K most relevant datasets to a specific dataset or/and to a given task, by using the whole contents of datasets, and not only metadata, e.g., for finding "the most connected datasets to my dataset" or "the K datasets maximizing the available data for Aristotle" (see the lower left side of Fig. 1).

• Task D (*Data Enrichment*). Combining information from several datasets, e.g., for improving the execution of Machine-Learning tasks (see the lower middle side of Fig. 1).

• Task E (*Data Quality and Veracity*). Assessing the quality of one or more datasets and estimating the reliability of a specific fact for an entity, e.g., "How many datasets agree about the birth date of Aristotle?" (see the lower right side of Fig. 1).

The main objective of this dissertation [Mountantonakis 2021] is to design and develop innovative indexes, methods, algorithms and services for assisting the mentioned tasks, i.e., the process of data discovery and semantic data integration at a large scale.

2. CHALLENGES AND RESEARCH QUESTIONS

Here, we provide the major challenges and the research questions of this dissertation.

Challenge 1. Cross-Dataset Identity Reasoning. The equivalence relationships that exist among entities and schemas of different datasets, such as `owl:sameAs`, `owl:equivalentProperty` and `owl:equivalentClass`, model an equivalence relation. For finding all the equivalent URIs for a given URI, it is a prerequisite to compute the transitive and symmetric closure of these relationships. However, it presupposes knowledge of all datasets and most of the algorithms that compute the transitive and symmetric closure of equivalence relationships require a lot of memory, i.e., one should keep in memory all the binary relationships during the computation of closure. The major research question (*RQ1*) is how to compute in an efficient way the cross-dataset identity reasoning, i.e., the computation of transitive and symmetric closure of such relationships, by using either a single or a cluster of machines, since the number of equivalence relationships at scale is large.

Challenge 2. Construction of Indexes at Large Scale by applying Cross-Dataset Identity Reasoning. We desire to index the whole contents of multiple datasets, for enabling fast access to all the triples of a given URI u (including the triples that contain any of the

equivalent URIs of u). Thereby, a major question (*RQ2*), is how to apply the result of the cross-dataset identity reasoning for constructing such semantics-aware indexes. Moreover, since there are many datasets (hundreds or even thousands) and some of them are very big, a key question (*RQ3*) is how to parallelize the construction of indexes in an efficient way.

Challenge 3. Dataset Discovery by using content-based measurements among two or more datasets. For offering dataset discovery through content-based measurements, e.g., "find the K most relevant datasets to a given one", it is a prerequisite to exploit the whole contents of datasets (and not only metadata) and to solve maximization problems. The main problem is that the possible combinations of datasets is exponential in number (the number of possible solutions for a given K is given by the binomial coefficient formula). Therefore, a question (*RQ4*) is whether a query language implementation (e.g., SPARQL) can be used for solving such maximization problems. Moreover, since set operations (i.e., intersection, union, complement) between large datasets are quite expensive, two key questions are the following: how can we reduce the number of set operations between different datasets (*RQ5*), and whether these content-based measurements can be parallelized (*RQ6*).

3. CONTRIBUTIONS

The key contributions of this dissertation [Mountantonakis 2021] for the above challenges and research questions are the following:

• It introduces a survey including a comprehensive and clear landscape of large scale semantic integration approaches [Mountantonakis and Tzitzikas 2019b] for better understanding the problem, i.e., by factorizing the integration process according to various dimensions, for identifying the research gaps as well as for identifying research directions.

• It proposes methods for performing cross-dataset identity reasoning at large scale (*RQ1*). These methods rely on special indexes and algorithms and can be executed by using either a single [Mountantonakis and Tzitzikas 2016] or a cluster of machines [Mountantonakis and Tzitzikas 2018c; 2018a]. Indicatively, by using 96 machines, 9 minutes are needed for computing the transitive and symmetric closure of 44 million equivalence relationships.

• It introduces scalable methods and algorithms that rely on MapReduce techniques [Mountantonakis and Tzitzikas 2016; 2018c; 2018a]. These algorithms apply the results of the closure for constructing dedicated global semantics-aware indexes (*RQ2* and *RQ3*) that cover the whole contents of datasets. Indicatively, 81.5 minutes are needed for constructing all the indexes for two billion triples by using 96 machines.

• It introduces content-based intersection, union and complement metrics for dataset search and discovery over large number of linked datasets, which are formulated and tackled as maximization problems [Mountantonakis and Tzitzikas 2016; 2018c; 2018a; 2020]. It exploits the constructed indexes, lattice-based incremental algorithms and set-theory properties for tackling the exponential complexity of the problems (*RQ4*, *RQ5* and *RQ6*). Indicatively, these algorithms are even $5000\times$ faster than straightforward methods, and can compute the metrics for 1 million subsets of datasets even in 1 second. The parallel version of lattice-based algorithms [Mountantonakis and Tzitzikas 2018c] makes feasible the computation of metrics between 1 billion subsets in 1 minute by using 64 machines.

• It reports statistics and connectivity analytics for a big subset of the current LOD cloud

that comprises 400 datasets and 2 billion triples, which reveal the sparsity of LOD Cloud datasets [Mountantonakis and Tzitzikas 2016; 2018c; 2018b; 2018a; 2020].

• **The LODsyndesis Suite of Services.** By exploiting the semantics-aware indexes and the results of the content-based measurements, several services and research prototypes have been created. The key prototype is called *LODsyndesis*, which offers services for assisting the Data Integration process for the tasks A-E (see Fig. 1), for 400 RDF datasets from 9 different domains, including over 400 million entities and 2 billion triples. Concerning other prototypes, the tools LODsyndesis$_{ML}$ and LODvec [Mountantonakis and Tzitzikas 2017; 2019a] exploit the indexes for assisting Machine Learning tasks, by creating features and embeddings from multiple RDF datasets simultaneously. By using these tools, we managed to improve the accuracy of predictions (we identified even 13% increase) for machine learning classification problems. These tools and many others, including tools for Entity Extraction from texts and Question Answering, are accessible online[1].

• **More Information.** This dissertation and its presentation, including the slides and the video of the presentation of the PhD defense, are accessible online[2].

4. CONCLUSION AND FUTURE WORK

This dissertation [Mountantonakis 2021] focused on proposing novel methods and services, supported by special indexes and algorithms, for covering the needs of several tasks that are related to the connectivity and semantic integration among large number of Linked Datasets. As a future work, we desire to exploit the proposed methods and indexes for several other tasks, e.g., for evaluating the connectivity of datasets before their actual publishing to the LOD Cloud, for checking the quality of equivalence relationships and others.

REFERENCES

MOUNTANTONAKIS, M. 2021. *Services for Connecting and Integrating Big Numbers of Linked Datasets*. Studies on the Semantic Web, vol. 50. IOS Press.

MOUNTANTONAKIS, M. AND TZITZIKAS, Y. 2016. On measuring the lattice of commonalities among several linked datasets. *Proceedings of the VLDB Endowment 9,* 12, 1101–1112.

MOUNTANTONAKIS, M. AND TZITZIKAS, Y. 2017. How linked data can aid machine learning-based tasks. In *International Conference on Theory and Practice of Digital Libraries*. Springer, 155–168.

MOUNTANTONAKIS, M. AND TZITZIKAS, Y. 2018a. High performance methods for linked open data connectivity analytics. *Information 9,* 6, 134.

MOUNTANTONAKIS, M. AND TZITZIKAS, Y. 2018b. Lodsyndesis: global scale knowledge services. *Heritage 1,* 2, 335–348.

MOUNTANTONAKIS, M. AND TZITZIKAS, Y. 2018c. Scalable methods for measuring the connectivity and quality of large numbers of linked datasets. *Journal of Data and Information Quality (JDIQ) 9,* 3, 1–49.

MOUNTANTONAKIS, M. AND TZITZIKAS, Y. 2019a. Knowledge graph embeddings over hundreds of linked datasets. In *Research Conference on Metadata and Semantics Research*. Springer, 150–162.

MOUNTANTONAKIS, M. AND TZITZIKAS, Y. 2019b. Large-scale semantic integration of linked data: A survey. *ACM Computing Surveys (CSUR) 52,* 5, 1–40.

MOUNTANTONAKIS, M. AND TZITZIKAS, Y. 2020. Content-based union and complement metrics for dataset search over rdf knowledge graphs. *Journal of Data and Information Quality (JDIQ) 12,* 2, 1–31.

[1] http://demos.isl.ics.forth.gr/lodsyndesis/

[2] http://users.ics.forth.gr/mountant/dissertation.html