RDF graph summarization: principles, techniques and applications

Haridimos Kondylakis
Institute Of Computer Science, FORTH
Heraklion, Greece
kondylak@ics.forth.gr

Dimitris Kotzinos
Lab. ETIS UMR 8051, University of Paris-Seine, University of Cergy-Pontoise, ENSEA, CNRS
Pontoise, France
Dimitrios.Kotzinos@u-cergy.fr

Ioana Manolescu
Inria and LIX (UMR 7161, CNRS and Ecole polytechnique)
Palaiseau, France
ioana.manolescu@inria.fr

ABSTRACT

The explosion in the amount of the RDF on the Web has lead to the need to explore, query and understand such data sources. The task is challenging due to the complex and heterogeneous structure of RDF graphs which, unlike relational databases, do not come with a structure-dictating schema. *Summarization* has been applied to RDF data to facilitate these tasks. Its purpose is to extract concise and meaningful information from RDF knowledge bases, representing their content as faithfully as possible. There is no single concept of RDF summary, and not a single but many approaches to build such summaries; the summarization goal, and the main computational tools employed for summarizing graphs, are the main factors behind this diversity.

This tutorial presents a structured analysis and comparison existing works in the area of RDF summarization; it is based upon a recent survey which we co-authored with colleagues [3]. We present the concepts at the core of each approach, outline their main technical aspects and implementation. We conclude by identifying the most pertinent summarization method for different usage scenarios, and discussing areas where future effort is needed.

1 INTRODUCTION

The explosion in the amount of the RDF on the Web has lead to the need to explore, query and understand such data sources. This need arises both for computer scientists and for scientists and practitioners in the many areas where Open Data is produced - ranging from agriculture to education, from cultural artefacts to criminality statistics. All users who need to tame, understand and analyze such complex RDF graphs are faced with several challenges.

Firstly, RDF graphs are often large compared with the human ability to understand and analyze them; even a “tiny” graph of e.g. 10.000 nodes is challenging for humans to comprehend. Secondly, unlike relational databases which come equipped with a prescriptive schema, RDF graphs lack regular structure or many times this structure exists but is unknown. Thirdly, size of the is challenging both for humans and for automated data processing tools. Fourthly, while RDF graphs may come equipped with ontologies, which specify the known relationships between the properties and classes present in the graph, the ontology itself is sometimes a source of complexity, especially if it is very large. In the presence of ontologies, graphs may contain implicit information, i.e. facts that hold in the graph despite not being physically present there. Reflecting the implicit facts of the ontology is in itself a challenge. Additionally specific parts of the ontology might not be used at all or very little in the specific Knowledge Base (KB).

Summarization has been applied to RDF data to facilitate these tasks. Its purpose is to extract concise and meaningful information from RDF knowledge bases, representing their content as faithfully as possible. There is no single concept of RDF summary, and not a single but many approaches to build such summaries.

Summarizing semantic graphs is a multifaceted problem with many dimensions, and thus many algorithms, methods and approaches have been developed to cope with it. As a result, there is now a confusion in the research community about the terminology in the area, further increased by the fact that certain terms are often used with different meanings in the relevant literature, denoting similar, but not identical research directions or concepts. We believe that this lack of terminology and classification hinders scientific development in this area.

Following up on a recent survey which we co-authored with colleagues [3], in this tutorial we present the main conceptual tools behind graph summarization, including some techniques developed prior to the advent of RDF, and show how all these techniques have been applied to the problems of summarizing semantic graphs. The goal of our tutorial is to acquaint the audience with the literature in this area, help them identify the tools and techniques most suited to the summarization problems they might have, and point out areas of interest for future work.

2 SCOPE

The tutorial aims at a broad range of researchers, students, IT professionals and practitioners, and developers. Anyone working with semantic graphs and RDF more specifically will benefit from this tutorial. Students and researchers will not only get a good introduction to the topic with a complete coverage of the state-of-the-art, but will also find a number of challenging research problems in these emerging technologies on which they may decide to focus their future research efforts. Practitioners will get a good overview of what the summarization algorithms, techniques and systems can offer nowadays and learn how they can use them to enhance their understanding of their available datasets. Developers of systems relying on semantic graphs will get helpful information that will help them improve further their products, enhancing query execution, data visualization and understanding.

Other tutorials [10, 15] considers the broad topic of summarizing large graphs, mostly using data mining tools, and with an approach tailored specifically to social networks; our tutorial focuses on the particularities of RDF graphs (and of their summarization).

Series ISSN: 2367-2005

10.5441/002/edbt.2019.38

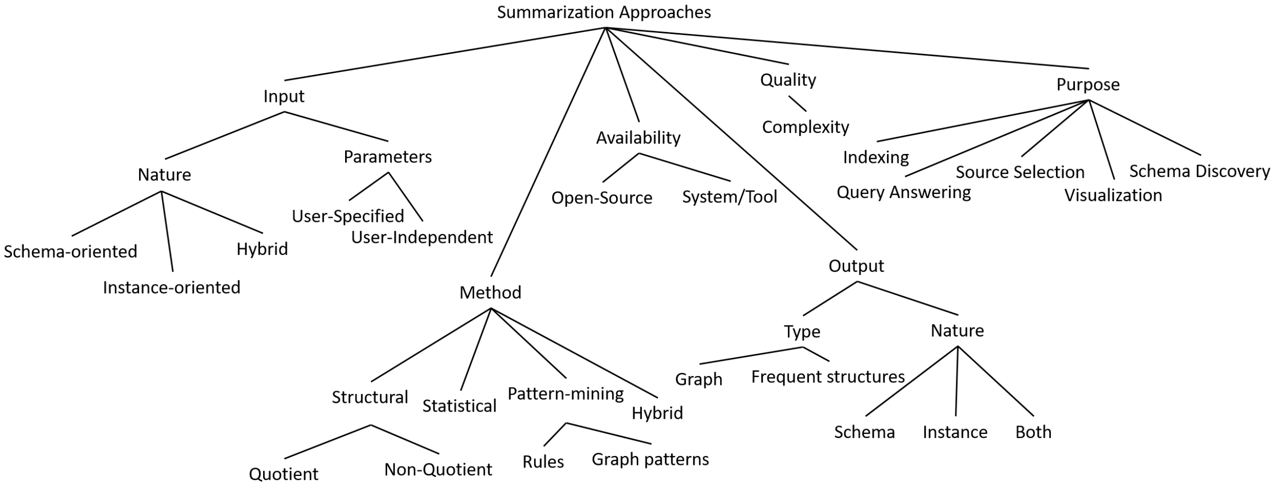

Figure 1: A taxonomy of the works in the area [3].

2.1 Tutorial goal

The goal of this tutorial is to introduce summarization notions and tools which are useful in order to concrete RDF data management applications. A broad set of techniques will be presented covering summarization of general RDF graphs, that contain or not ontological information, independent of their application domain.

3 OUTLINE

Our tutorial will be organized as follows.

3.1 Introduction and preliminaries

We will recall the basics of RDF graphs, RDFS and OWL ontologies, RDF queries (focusing in particular on conjunctive queries, the most frequently used in practice) and inference in RDF knowledge graphs in the presence of an ontology.

3.2 Applications

Next, we will present the main classes of application contexts which have justified the need for RDF summaries:

Indexing: The first and foremost application RDF summaries has been brought by the necessity of efficiently querying large and complex graphs. In this context, sets of nodes which are likely to be used together by queries are grouped together and their IDs are associated to a given summary node. Then, query processing proceeds in two stages: first, the summary nodes relevant to a given query are identified; then, from the summary node, an index lookup gives access directly to the respective data nodes.

Estimating the size of query results: To the same direction with indexing, summaries can be used to identify directly when no nodes are available for a specific query. More than this, summaries can also store statistical information on the available nodes, leading query optimizers to start query evaluation from the most selective conditions.

Source selection: Query evaluation across several graphs, in particular in a distributed setting where multiple graphs are each accessible behind its individual endpoint, can greatly benefit from the knowledge that one source (or one graph) does not have results matching a (sub)query. This is one facet of source selection: restricting query evaluation to avoid datasets on which it is guaranteed to have no answers. A variant consists of ordering the graphs to be explored (queried), so that in a finite time budget, the most interesting graphs are sure to be visited.

Graph visualization and schema discovery: Summarized information is easier to be visualized and comprehended. On the other hand when an ontology is not present, it can be extracted out of the available data, augmenting user understanding on the available information.

3.3 Classification & Dimensions

Then, we will present a classification of the available approaches according to the main algorithmic idea behind the summarization approach. We identify the following main categories; we also indicate a few of the relevant references (out of the 122 present in our survey [3]):

1. **Structural methods** are those which summarize semantic graphs, based mostly on the graph structure, i.e. the paths and sub-graphs available in the RDF graph. Techniques in this category can be further categorized as quotient and non-quotient.
   - **Quotient** approaches are based on the idea of characterizing selected graph nodes as "equivalent" in a certain way, and then summarizing the graph by assigning a representative summary node to each class of equivalent graph nodes; further, each edge between two graph nodes leads to the corresponding edge being present between the two graph node's representatives. In a quotient summary, each property (edge label) from the input RDF graph is guaranteed present in the summary graph; more a quotient summary is guaranteed to be at most as large as the original graph. Different quotient summaries result from different notions of equivalence among the RDF graph nodes. Sample works in this area include [2, 4, 11, 22, 27], while [17] and [9] are important sources of inspiration from before the RDF age;
   - **Non-quotient** approaches are the remaining methods that are based mostly, on specific measures according to which the most important nodes are identified and then linked to formulate the presented summary. Pioneered by Dataguides [6, 20], the area features many recent works such as [21, 24, 29, 30]; [28] has an information retrieval approach, as it aims at extracting the most interesting triples to be shown to a user about a

subject; [32] summarizes ontologies found on the web, through the prism of the salience (interestingness) of their concepts.

2. **Pattern mining methods** employee mining techniques to identify patterns appearing in the semantic graph. A pattern might be a set of instances having a certain set of properties, which are in exact or approximate terms representative of the graph or provide enough information on the graph using some cost function to determine that. We consider also as patterns the discovery of rules that can be used to reconstruct the graph and thus represent it adequately. Those patterns, together, compose the summary. Representative works in this area are, e.g., [8, 25, 33];
3. **Statistical methods** on the other hand try to qualitatively summarize the contents of a graph counting occurrences, building histograms, measuring frequencies and other statistical measures out of the available semantic graph. This class comprises notably works such as [5, 7, 23, 31];
4. Finally, several works combine techniques from several of the main areas listed above; these are **hybrid methods**, e.g., [1, 27] first and foremost aim at estimating the cardinality of query patterns, [26] summarizes RDF graphs and ontologies through the prism of statistics, while [19] aims at graph compression with bounded error, that is: a core (most regular) part of the graph is identified as comprising several copies of a same data pattern, and compressed into a single copy of this, whereas the rest is ignored from the summary and considered to be the summarization error (which the authors seek to minimize under certain constraints).

Further, we will characterize each of these proposal along a set of other informative dimensions:

1. **Input:** An interesting dimension of analysis is the input required by each summarization method, as different approaches have usually different requirements for the dataset they get as input. RDF data graphs are usually accepted, RDF/S and/or OWL are considered for some of the works for specifying graph semantics whereas very few works consider DL models. Some works are based only on the ontology part whereas others consider only instances. Hybrid approaches are also available consuming both instances and the ontology for producing summaries. In addition, many works in the area require additional user input of fine tuning (e.g. summary size, weights, equivalence relations etc.) whereas some others are completely user independent.
2. **Output:** Besides input, the available works might have also different output. The summary for example can be a graph or a selection of the most frequent structures such as nodes, paths, rules or queries. In addition we distinguish summaries that only output instances from those that output schema information as well.
3. **Availability:** Several approaches are available by the authors as complete system/tool and some others provide only the corresponding algorithms/theory. Finally some systems are available online and can be readily tested.
4. **Purpose:** As already explained in the applications of the summarization techniques, summaries can be build for indexing, source selection, visualisation, schema discovery or for facilitating query answering.
5. **Quality:** Finally an important dimension of study, for each summarization algorithm is its completeness in terms of coverage, precision and recall of the result if an "ideal" summary is available as golden standard and its corresponding computational complexity.

Figure 1 presents the various dimensions that will be used in order to present the works available in each category.

Natural connections exist between the families of RDF summaries and the applications they are best suited for. Structural quotient summaries are most applicable to indexing and query answering through graph reduction; this holds especially for quotients built through equivalence relations such as bisimilarity (possibly bounded). Non-quotient summaries mostly target visualization, schema discovery and data understanding. Pattern mining summaries provide in many cases logical rules besides the summary graph as part of the final result, so could be possibly more useful in RDF graph compression scenarios. Summaries could also be really useful in data integration scenarios [14], where instead of generating mappings [16], [18] between data source schemas, summaries could be used to drive the definition of the mapping. Extending this to a scenario where the sources can also evolve [13], [12], summaries can play a key role in schema understanding and mapping redefinition.

3.4 Open issues and future research directions

RDF graph summaries can be useful in different application and research scenarios. Each scenario brings each own specific requirements and the possibility of having more than one items being suitable is present. One open issue in this respect is whether one could use the provided taxonomy to further automate the selection of the appropriate algorithms in the different use cases.

Identifying the *quality* of the RDF summary is also a difficult and not really widely addressed problem. The main problem that remains is how could one compare the summaries produced by the different algorithms and take into account the specificities of the problem at hand and provide an RDF summary with some guarantees. Given that even human experts do not agree on the quality of different summaries in many cases, this remains a challenging task.

Finally, one important problem that has been looked up very little is the *updates* of the RDF summaries produced, given the dynamic nature of most RDF datasets as well as their size. It is an open issue how one could update the summary without having to recompute the whole summary every time; and this problem has also a temporal dimension since one should answer not only how but also when this update is pertinent.

4 PRESENTERS

Haridimos Kondylakis is a scientific collaborator at the Institute of Computer Science, FORTH. His research interests span the following areas: Semantic Integration; Knowledge Evolution; Big Data Management; Data Series Indexing and Querying. He has extensive experience in participating in more than 16 European Projects and he also acts as a regular reviewer and a PC member for a number of premier journals and conferences. He has more than 110 publications in international conferences, books and journals including ACM SIGMOD, VLDB, VLDB Journal, JWS, KER, EDBT, ISWC, ESWC etc.

Dimitris Kotzinos is a Professor at the Department of Computer Science of the University of Cergy – Pontoise, member of the ETIS Lab and head of the MIDI team of the lab. His main research interests include data management algorithms, techniques and tools; development of methodologies, algorithms and tools for web-based information systems, portals and web services; and the understanding of the meaning (semantics) of interoperable data and services on the web.

Ioana Manolescu is a senior researcher, and the lead of the CEDAR team, joint between Inria Saclay and the LIX lab (UMR 7161) of Ecole polytechnique, in France. The CEDAR team research focuses on rich data analytics at cloud scale. She is a member of the PVLDB Endowment Board of Trustees, and a co-president of the ACM SIGMOD Jim Gray PhD dissertation committee. Recently, she has been a general chair of the IEEE ICDE 2018 conference, an associate editor for PVLDB 2017 and 2018, and the program chair of SSDBBM 2016. She has co-authored more than 130 articles in international journals and conferences, and contributed to several books. Her main research interests include data models and algorithms for computational fact-checking, performance optimizations for semistructured data and the Semantic Web, and distributed architectures for complex large data.

REFERENCES

[1] Anas Alzogbi and Georg Lausen. 2013. Similar Structures inside RDF-Graphs. In *Proceedings of the WWW2013 Workshop on Linked Data on the Web, Rio de Janeiro, Brazil, 14 May, 2013.*

[2] Šejla Čebirić, François Goasdoué, Pawel Guzewicz, and Ioana Manolescu. 2017. *Compact Summaries of Rich Heterogeneous Graphs.* Research Report RR-8920. INRIA Saclay ; Université Rennes 1. https://hal.inria.fr/hal-01325900

[3] Sejla Cebirić, François Goasdoué, Haridimos Kondylakis, Dimitris Kotzinos, Ioana Manolescu, Georgia Troullinou, and Mussab Zneika. 2018. Summarizing Semantic Graphs: A Survey. *The VLDB Journal* (2018). https://hal.inria.fr/hal-01925496 Accepted for publication, to appear.

[4] Mariano P Consens, Valeria Fionda, Shahan Khatchadourian, and Giuseppe Pirro. 2015. S+ EPPs: construct and explore bisimulation summaries, plus optimize navigational queries; all on existing SPARQL systems. *Proceedings of the VLDB Endowment* 8, 12 (2015), 2028–2031.

[5] Marek Dudás, Vojtech Svátek, and Jindrich Mynarz. 2015. Dataset Summary Visualization with LODSight. In *The Semantic Web: ESWC 2015 Satellite Events - ESWC 2015 Satellite Events Portorož, Slovenia, May 31 - June 4, 2015, Revised Selected Papers.* 36–40.

[6] Roy Goldman and Jennifer Widom. 1997. DataGuides: Enabling Query Formulation and Optimization in Semistructured Databases. In *VLDB'97, Proceedings of 23rd International Conference on Very Large Data Bases, August 25-29, 1997, Athens, Greece.* 436–445.

[7] Katja Hose and Ralf Schenkel. 2012. Towards benefit-based RDF source selection for SPARQL queries. In *Proceedings of the 4th International Workshop on Semantic Web Information Management, SWIM 2012, Scottsdale, AZ, USA, May 20, 2012.* 2.

[8] Amit Krishna Joshi, Pascal Hitzler, and Guozhu Dong. 2013. Logical Linked Data Compression. In *The Semantic Web: Semantics and Big Data, 10th International Conference, ESWC 2013, Montpellier, France, May 26-30, 2013. Proceedings.* 170–184.

[9] Raghav Kaushik, Philip Bohannon, Jeffrey F. Naughton, and Henry F. Korth. 2002. Covering indexes for branching path queries. In *Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data, Madison, Wisconsin, June 3-6, 2002.* 133–144.

[10] Arijit Khan, Sourav S. Bhowmick, and Francesco Bonchi. 2017. Summarizing Static and Dynamic Big Graphs. *PVLDB* 10, 12 (2017), 1981–1984.

[11] Shahan Khatchadourian and Mariano P. Consens. 2010. ExpLOD: Summary-Based Exploration of Interlinking and RDF Usage in the Linked Open Data Cloud. In *The Semantic Web: Research and Applications, 7th Extended Semantic Web Conference, ESWC 2010, Heraklion, Crete, Greece, May 30 - June 3, 2010, Proceedings, Part II.* 272–287.

[12] Haridimos Kondylakis and Dimitris Plexousakis. 2011. Ontology Evolution in Data Integration: Query Rewriting to the Rescue. In *Conceptual Modeling - ER 2011, 30th International Conference, ER2011, Brussels, Belgium, October 31 - November 3, 2011. Proceedings.* 393–401.

[13] Haridimos Kondylakis and Dimitris Plexousakis. 2012. Ontology Evolution: Assisting Query Migration. In *Conceptual Modeling - 31st International Conference ER 2012, Florence, Italy, October 15-18, 2012. Proceedings.* 331–344.

[14] Haridimos Kondylakis and Dimitris Plexousakis. 2013. Ontology evolution without tears. *J. Web Sem.* 19 (2013), 42–58.

[15] Shou-De Lin, Mi-Yen Yeh, and Cheng-Te Li. 2013. Sampling and Summarization for Social Networks (tutorial).

[16] Yannis Marketakis, Nikos Minadakis, Haridimos Kondylakis, Konstantina Konsolaki, Georgios Samaritakis, Maria Theodoridou, Giorgos Flouris, and Martin Doerr. 2017. X3ML mapping framework for information integration in cultural heritage and beyond. *Int. J. on Digital Libraries* 18, 4 (2017), 301–319.

[17] Tova Milo and Dan Suciu. 1999. Index Structures for Path Expressions. In *Database Theory - ICDT '99, 7th International Conference, Jerusalem, Israel, January 10-12, 1999, Proceedings.* 277–295.

[18] Nikos Minadakis, Yannis Marketakis, Haridimos Kondylakis, Giorgos Flouris, Maria Theodoridou, Gerald de Jong, and Martin Doerr. 2015. X3ML Framework: An Effective Suite for Supporting Data Mappings. In *Proceedings of the Workshop on Extending, Mapping and Focusing the CRM co-located with 19th International Conference on Theory and Practice of Digital Libraries (2015), Poznań, Poland, September 17, 2015.* 1–12.

[19] Saket Navlakha, Rajeev Rastogi, and Nisheeth Shrivastava. 2008. Graph summarization with bounded error. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2008, Vancouver, BC, Canada, June 10-12, 2008.*

[20] Svetlozar Nestorov, Jeffrey D. Ullman, Janet L. Wiener, and Sudarshan S. Chawathe. 1997. Representative Objects: Concise Representations of Semistructured, Hierarchical Data. In *ICDE.*

[21] Alexandros Pappas, Georgia Troullinou, Giannis Roussakis, Haridimos Kondylakis, and Dimitris Plexousakis. 2017. Exploring Importance Measures for Summarizing RDF/S KBs. In *The Semantic Web - 14th International Conference, ESWC 2017, Portorož, Slovenia, May 28 - June 1, 2017, Proceedings, Part I.* 387–403.

[22] François Picalausa, Yongming Luo, George H. L. Fletcher, Jan Hidders, and Stijn Vansummeren. 2012. A Structural Approach to Indexing Triples. In *The Semantic Web: Research and Applications - 9th Extended Semantic Web Conference, ESWC 2012, Heraklion, Crete, Greece, May 27-31, 2012. Proceedings.*

[23] Carlos Eduardo S. Pires, Paulo Orlando Queiroz-Sousa, Zoubida Kedad, and Ana Carolina Salgado. 2010. Summarizing ontology-based schemas in PDMS. In *Workshops Proceedings of the 26th International Conference on Data Engineering, ICDE 2010, March 1-6, 2010, Long Beach, California, USA.* 239–244.

[24] Paulo Orlando Queiroz-Sousa, Ana Carolina Salgado, and Carlos Eduardo S. Pires. 2013. A Method for Building Personalized Ontology Summaries. *JIDM* 4, 3 (2013), 236–250.

[25] Qi Song, Yinghui Wu, and Xin Luna Dong. 2016. Mining Summaries for Knowledge Graph Search. In *IEEE 16th International Conference on Data Mining, ICDM 2016, December 12-15, 2016, Barcelona, Spain.* 1215–1220.

[26] Blerina Spahiu, Riccardo Porrini, Matteo Palmonari, Anisa Rula, and Andrea Maurino. 2016. ABSTAT: Ontology-driven Linked Data Summaries with Pattern Minimalization. In *SumPre.*

[27] Giorgio Stefanoni, Boris Motik, and Egor V. Kostylev. 2018. Estimating the Cardinality of Conjunctive Queries over RDF Data Using Graph Summarisation. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018, Lyon, France, April 23-27, 2018.* 1043–1052. https://doi.org/10.1145/3178876.3186003

[28] Marcin Sydow, Mariusz Pikula, and Ralf Schenkel. 2013. The notion of diversity in graphical entity summarisation on semantic knowledge graphs. *J. Intell. Inf. Syst.* 41, 2 (2013), 109–149.

[29] Georgia Troullinou, Haridimos Kondylakis, Evangelia Daskalaki, and Dimitris Plexousakis. 2017. Ontology understanding without tears: The summarization approach. *Semantic Web* 8, 6 (2017), 797–815.

[30] Georgia Troullinou, Haridimos Kondylakis, Kostas Stefanidis, and Dimitris Plexousakis. 2018. Exploring RDFS KBs Using Summaries. In *The Semantic Web - ISWC 2018 - 17th International Semantic Web Conference, Monterey, CA, USA, October 8-12, 2018, Proceedings, Part I.* 268–284.

[31] Gang Wu, Juanzi Li, Ling Feng, and Kehong Wang. 2008. Identifying Potentially Important Concepts and Relations in an Ontology. In *The Semantic Web - ISWC 2008, 7th International Semantic Web Conference, ISWC 2008, Karlsruhe, Germany, October 26-30, 2008. Proceedings.* 33–49.

[32] Xiang Zhang, Gong Cheng, Weiyi Ge, and Yuzhong Qu. 2009. Summarizing Vocabularies in the Global Semantic Web. *J. Comput. Sci. Technol.* 24, 1 (2009), 165–174.

[33] Mussab Zneika, Claudio Lucchese, Dan Vodislav, and Dimitris Kotzinos. 2016. Summarizing Linked Data RDF Graphs Using Approximate Graph Pattern Mining. In *EDBT 2016.* 684–685.