(Semi-) Automatic construction of knowledge graph metadata

Maryam Mohammadi [1][0000-0003-4850-8068]

[1] Institute of Data Science, Maastricht University, Maastricht, the Netherlands

m.mohammadi@maastrichtuniversity.nl

Abstract. Recently a huge number of knowledge graphs (KGs) has been generated, but there has not been enough attention to generate high-quality metadata to enable users to reuse the KGs for their own purposes. The main challenge is to generate standardized and high quality descriptive metadata which helps users understand the content of the large KGs. Some existing solutions make use of a combination of schema-level patterns derived from graph summarization with instance-level snippets. I will follow this trend and develop a method based on a combination of content-based patterns with user activity data such as SPARQL query logs to make generated metadata more informative and useful than other developed approaches. The problem of current models is generating complex, long or insufficient metadata which I plan to tackle by proposing a guideline to generate standard metadata during my Ph.D.

Keywords: Metadata, Knowledge Graph, FAIR, Graph Summarization, SPARQL Query Logs.

1 Motivation

Analyzing and mining Big Data provides a precious opportunity to solve a variety of problems. Increasingly, data is organized as a knowledge graph, in which data are represented as nodes and edges, and this representation facilitates data integration and knowledge discovery. Enormous amounts of Knowledge Graphs (KGs) have been published by researchers in academia and industry. However, there has not been enough attention to generate high-quality metadata to sufficiently describe these datasets which specify what these data are and how they were produced [1, 13]. The lack of standardized metadata also makes comparing possible datasets extremely difficult. Therefore, it is quite challenging for users to find pertinent datasets and reuse them for analyses or other aims without high quality metadata [10].

According to the FAIR principles [9], datasets and their metadata should be represented in a way that makes the dataset more Findable, Accessible, Interoperable, and Reusable. However, creating metadata manually is time-consuming, often incomplete,

and prone to error. Towards improving the findability and reusability of the KG datasets, there is an urgent need to provide a rich, structured and understandable description of the dataset in a cost-effective and scalable way [16].

My work aims to increase FAIRness of KGs by generating high quality descriptions of KGs. Increasing FAIRness will enable people and machines to discover relevant resources, and reuse them for new tasks instead of investing their time and money to create their own dataset. It is estimated that the European economy loses more than €10.2 billion euros per year [8] owing to a lack of FAIR data, based on five quantifiable indicators: time spent, cost of storage, license costs, research retraction and double funding. I believe that improving the amount and quality of metadata will improve overall productivity, reduce duplicative costs, and generate new opportunities. For instance, even assuming one has sufficient knowledge about SPARQL query language, it is important to understand the content of the dataset to write the query. Providing enriched metadata that present schemas as well as connections between entities of a KG extremely helps users in query writing tasks [1].

The rest of the paper is organized as follows. I review State-of-the-Art methods in section 2. I introduce problem formulation and research questions in section 3, research hypotheses and research steps are outlined in the section 4 followed by an evaluation plan in section 5. Section 6 presents preliminary results. Lessons learned and conclusions are described in section 7.

2 State-of-the-Art

Several approaches have been proposed to generate high quality descriptive metadata, but many of these rely on human curation, which is expensive, time consuming, and may be limited by the availability of experts. In contrast, semi-automatic and automatic methods offer another way forward. For example, the health care and life sciences community have proposed a generic profile for dataset descriptions and provided SPARQL-based templates for automatic data summarization [3]. Yike et al. [19] have described several graph summarization methods, as well as their different types of input and output. Safavi et al. [20] have proposed GLIMPSE for creating personalized knowledge graph summarization. Ayak et al. [2] have conducted an automatic method for predicting experimental metadata from scientific publications. Martínez-Romero M et al. [4] have developed a method for generating metadata with using ontology-based recommendations. Moreover, there are variant works based on graph summarization models [1, 5], which create schema-level patterns to represent content of a KG. Some works generate instance-level triples by the use of graph snippet generation methods [15]. Wang et al. [11] have proposed a Pattern-Coverage Snippet generation for RDF Datasets based on a combination of schema-level and instance-level data.

As I am using SPARQL query logs in my proposed method, I have done some literature review on SPARQL query logs analysis. Through an analysis on Bio2RDF

SPARQL query logs Carlos et al. [16] reported some statistics about SPARQL query keywords and triple patters. In addition, they found that there is a large amount of repeated queries and only 20 query patterns represent 90% of the whole Bio2RDF query logs. Saleem et al. [17] proposed The Linked SPARQL Queries Dataset (LSQ), which describes SPARQL queries issued to endpoints of four datasets with providing statistics of SPARQL features and classes. Claus et al. [18] developed LSQ 2.0: A Linked Dataset of SPARQL Query Logs and extended the work of [17] to 27 different endpoints. In contrast, in my work, I am focusing on each element of the triple patterns namely subjects, predicates, and objects rather than SPARQL keywords or features. For instance, I list the top 20 frequent subjects, predicates, and objects that indicate the interests of the users based on the query logs.

3 Problem Statement and Contributions

In this section, I introduce problem formulation and research questions that will be focused on during my Ph.D. and the expected contributions I aim to make by answering the research questions.

My PhD research focuses on developing computational approaches for generating high quality machine-readable and human-readable descriptive metadata for knowledge graphs. Generated metadata will help users in two tasks; 1) summarizing the content of a dataset and therefore, increasing the discoverability of a dataset, 2) helping users in SPARQL query writing. My work will explore ways to make generated metadata more informative and useful than other developed approaches. A key direction will be explored lies in the combination of content-based patterns data with external user activity data such as SPARQL query logs. I hypothesize that different metrics will be of value for different KG-related tasks, and intend to learn about these user preferences. The research will study the following research questions:

RQ1- To what extent do the analysis of external knowledge sources (e.g. query logs) inform users of its most relevant content?

RQ2- To what extent sensible natural language summaries could be generated from knowledge graphs?

4 Research Methodology and Approach

Graph summarization can generate metadata about the content of the graph by quantifying how many times certain patterns occur. For a very large graph, it may not be easy for somebody to quickly determine what the graph is about due to a high number of emerging patterns. On the other hand, I hypothesize the users' query logs (against the SPARQL endpoint of a KG) could potentially reveal valuable information about what is interesting to the users of that particular KG. To the best of our knowledge, none of the existing work has used SPARQL query logs in their proposed method. I aim to

propose a method to prioritize the patterns driven from graph summarization based on what users ask about the graph, in other words, based on the information from query logs. Additionally queries can contain constants, which are potentially informative to users. According to this idea, three hypotheses are shaped as below:

Hypothesis 1: frequently occurring SPARQL queries are more useful as metadata than frequently occurring graph summaries (frequently occurring concepts) for large graphs, as qualitatively evaluated by potential users of the graph.

Hypothesis 2: query filtered graph summaries are more useful as metadata than either ranked lists of SPARQL queries or graph summaries.

Hypothesis 3: frequently occurring patterns derived from SPARQL queries are more useful as metadata than ranked lists of SPARQL queries or ranked lists of graph summaries.

Research steps for exploring an answer for RQ1 and RQ2 and my progress in each step is as follows:

Step1: Retrieve SPARQL query logs from the endpoint or available resources for a KG (e.g. Bio2RDF Kg or Wikidata KG)

Step2: Remove personal data and prepare the SPARQL query data in a format that is clean for processing

For cleaning data I delete some basic automatic SPARQL queries that has been sent by machines such as “select* where {?s ?p ?o}”. However, I keep more complex SPARQL queries that has been sent by machines or web interfaces, which I believe they are interesting and informative.

Step 3. Isolate query patterns or keywords from SPARQL query logs using a library such as RDF4J, Apache Jena or RDFLib. Generated keywords for an example query is shown in Fig. 1.

Step 4. Apply a graph summarization algorithm [11] or a rule-mining algorithm to the KG and rank the output patterns based on their frequency. (We call these patterns, content-based patterns)

Step 5. Rank query patterns or keywords according to their frequency of use in the SPARQL queries.

Step 6. Merge results of step 3 and 4 to produce metadata for the KG

One idea is to select high frequent patterns generated from executing step 4 only if they contain highly frequent keywords from step 3.

Step 7. Convert graph summaries (metadata) to sensible natural language summaries (metadata) (RQ2)

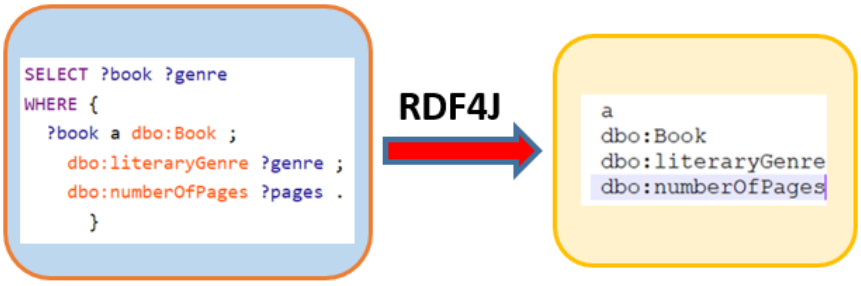

Fig. 1. Generated Subject, predicate and object keywords from an example query using RDF4J

Step 8. Evaluate the quality and utility of derived metadata with user study and FAIRness evaluation.

I proposed a workflow to generate descriptive metadata for a KG or a SPARQL endpoint by providing two use cases, Bio2RDF and Wikidata KGs. I chose these KGs because the goal of my work is to create a method which is able to support large scale KGs. I believe that this approach is generalizable; a limitation of method could be the difficulty of obtaining such SPARQL logs. Accessing SPARQL logs data might not always be possible. Depending on the tools used to store the logs, the process for obtaining the logs could be different. About our use cases, for Bio2RDF triplestore, because we had full control over it and we had added a service to store the logs, we were able to retrieve them. Wikidata logs were accessible through a public dataset from international center for computational logic.

5 Evaluation Plan

In this section, I explain three different experiments that I will conduct to evaluate my method. I intend to perform two types of user study mentioned in [1] and [11]. Spahiu et al. [1] have designed a SPARQL query formulation task for two groups of the people, control group and ABSTAT group to evaluate if generated metadata can help users in SPARQL query formulation. Formulating SPARQL queries is a task that requires prior knowledge about the dataset. In this work, five queries (with different complexity) in natural language format together with their incomplete SPARQL formulation were given to 20 participants. The participants were equally splitted into two groups and only one group could use their framework (ABSTAT) in order to get help and complete SPARQL queries. Authors measured the time spent to complete each query and the correctness of the answers and compared the performance of the groups based on completion time and accuracy of the answers. In an experiment by Wang et al. [11], participants of the user study are 20 computer science students that all have the essential

knowledge about RDF. Each participant were given ten RDF datasets together with metadata about each RDF dataset. Participants have been asked to rate the quality of two snippets generated by their method and a baseline method, in the range from 1 to 5 expressing how well that snippet exemplified the content of the RDF dataset and have asked to briefly explain the rating. Finally, I will introduce a new evaluation, based on assessing and comparing FAIRness of KG before and after importing generated metadata to the KG with different automatic or semi-automatic FAIR evaluation tools such as FAIR CHECKER [12] and F-UJI [14]. These tools take a resource as the input and check all the FAIR principles for that resource and assign a number indicating percentage of FAIRness of the resource.

6 Preliminary Results

The results of the conducted research steps for each use case are described in this section. Bio2RDF SPARQL queries log has been dumped from ElasticSearch and its personal data has been removed. The code and data is accessible through kg-metadata-generation GitHub page. Organic SPARQL queries of Wikidata that are clean to process and do not contain personal data are downloaded from international center for computational logic. Using RDF4J package in java, all triple patterns of Wikidata queries has been extracted. Then each element of subject, predicate and object is extracted from the triple patterns. In the next step, frequency of the keywords has been calculated. Frequency of the keywords extracted from Organic SPARQL queries of Wikidata is shown in Table 1. and Table 2. These results suggest that another criterion such as validity of the keyword in addition to the frequency must be considered for step 5 of methodology to avoid meaningless keywords such as "string1". On the other hand, for generating content-based patterns different graph summarization models are explored. Rule base methods due to their scalability for large size of input will be studied.

Table 1. Frequency of the predicate keywords extracted from Organic SPARQL queries

Top 10 frequent predicate keywords (the labels)	frequency
Language	88072
instance of (P31)	63927
label	47823
Image (P18)	35315
coordinate location (P625)	28912
subclass of (P279)	27832
description	22334
about	14825

Commons category (P373)	12711
located in the administrative territorial entity (P131)	11717

Table 2. Frequency of the subject or object keywords extracted from Organic SPARQL queries of Wikidata

Top 10 frequent subject or object keywords (the labels)	frequency
"en"	35781
"string1"	15326
human (Q5)	9527
"en,en"	6420
"fr"	5449
"None"	4029
Wiki_Main_Page	3690
"POINT(9"	3017
“de”	2874
'41)"^^<http://www.opengis.net/ont/geosparql#wktLiteral>'	2813

7 Conclusions and Lessons Learned

Lack of high quality metadata hinders users to better understand existing datasets and reuse them for more analyses and other purposes. The informative, machine- and human-readable metadata (that describes relevant features of the data in the KGs) would increase reusability of the data from existing KGs. Generating metadata for KGs in automatic manner is essential in order to decrease socio-economic impact of not having metadata or high quality metadata. My work aims to propose computational methods to generate descriptive metadata for KGs based on the combination of their internal (e.g. content-based patterns) and external (e.g. user activity) data. Generated results on a sample of small size of the Wikidata dataset are very promising.

Acknowledgements. This research has been funded by the European Union’s Horizon 2020 research and innovation program under the Marie Skłodowska-Curie project Knowgraphs (grant agreement ID: 860801). I would like to express my special thanks of gratitude to my advisors and collaborators Prof. Michel Dumontier, Prof Christopher Brewster, Dr. Remzi Celebi, Chang Sun and Vincent Emonet.

References

1. Spahiu, Blerina, Riccardo Porrini, Matteo Palmonari, Anisa Rula, and Andrea Mau-rino. "ABSTAT: ontology-driven linked data summaries with pattern minimaliza-tion." In European Semantic Web Conference, pp. 381-395. Springer, Cham, 2016.
2. Stuti Nayak, Amrapali Zaveri, Pedro Hernandez Serrano, and Michel Dumontier. 2021. Experience: Automated Prediction of Experimental Metadata from Scientific Publications. J. Data and Information Quality 13, 4, Article 21 (December 2021), 11 pages. DOI:https://doi.org/10.1145/3451219
3. Dumontier, Michel, Alasdair JG Gray, M. Scott Marshall, Vladimir Alexiev, Peter Ansell, Gary Bader, Joachim Baran et al. "The health care and life sciences commu-nity profile for dataset descriptions." PeerJ 4 (2016): e2331.
4. Martínez-Romero M, O'Connor MJ, Shankar RD, et al. Fast and Accurate Metadata Authoring Using Ontology-Based Recommendations. *AMIA Annu Symp Proc*. 2018;2017:1272-1281. Published 2018 Apr 16.
5. Q. Song, Y. Wu and X. L. Dong, "Mining Summaries for Knowledge Graph Search," 2016 IEEE 16th International Conference on Data Mining (ICDM), 2016, pp. 1215-1220, doi: 10.1109/ICDM.2016.0162.
6. Safavi, Tara, Caleb Belth, Lukas Faber, Davide Mottin, Emmanuel Müller and Danai Koutra. "Personalized Knowledge Graph Summarization: From the Cloud to Your Pocket." *2019 IEEE International Conference on Data Mining (ICDM)* (2019): 528-537.
7. Dudáš M., Svátek V., Mynarz J. (2015) Dataset Summary Visualization with LOD-Sight. In: Gandon F., Guéret C., Villata S., Breslin J., Faron-Zucker C., Zimmermann A. (eds) The Semantic Web: ESWC 2015 Satellite Events. ESWC 2015. Lecture Notes in Computer Science, vol 9341. Springer, Cham. https://doi.org/10.1007/978-3-319-25639-9_7
8. European Commission, Directorate-General for Research and Innovation, *Cost-benefit analysis for FAIR research data : cost of not having FAIR research data*. Publications Office; 2019. Available from: doi/10.2777/02999 https://op.europa.eu/en/publication-detail/-/publication/d375368c-1a0a-11e9-8d04-01aa75ed71a1/language-en
9. Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg et al. "The FAIR Guiding Principles for scientific data management and stewardship." *Scientific data* 3, no. 1 (2016): 1-9. https://www.nature.com/articles/sdata201618
10. Pietriga, Emmanuel, Hande Gözükan, Caroline Appert, Marie Destandau, Šejla Če-birić, François Goasdoué, and Ioana Manolescu. "Browsing linked data catalogs with LODAtlas." In International Semantic Web Conference, pp. 137-153. Springer, Cham, 2018.
11. Wang, Xiaxia, Gong Cheng, Tengteng Lin, Jing Xu, Jeff Z. Pan, Evgeny Kharlamov, and Yuzhong Qu. "PCSG: pattern-coverage snippet generation for RDF datasets." In International Semantic Web Conference, pp. 3-20. Springer, Cham, 2021.
12. Rosnet, T., F. de Lamotte, M. D. Devignes, V. Lefort, and A. Gaignard. "FAIR-Checker–Supporting the findability and reusability of digital life science resources."
13. Palmonari, Matteo, Anisa Rula, Riccardo Porrini, Andrea Maurino, Blerina Spahiu, and Vincenzo Ferme. "ABSTAT: linked data summaries with abstraction and statistics." In European Semantic Web Conference, pp. 128-132. Springer, Cham, 2015.
14. Huber, Robert, and Anusuriya Devaraju. "F-UJI: An Automated Tool for the Assessment and Improvement of the FAIRness of Research Data." In EGU General Assembly Conference Abstracts, pp. EGU21-15922. 2021.

15. Liu, Daxin, Gong Cheng, Qingxia Liu, and Yuzhong Qu. "Fast and practical snippet generation for RDF datasets." *ACM Transactions on the Web (TWEB)* 13, no. 4 (2019): 1-38.
16. Buil-Aranda, Carlos, Martın Ugarte, Marcelo Arenas, and Michel Dumontier. "A preliminary investigation into SPARQL query complexity and federation in Bio2RDF." In Alberto Mendelzon International Workshop on Foundations of Data Management, p. 196. 2015.
17. Saleem, Muhammad, Muhammad Intizar Ali, Aidan Hogan, Qaiser Mehmood, and Axel-Cyrille Ngonga Ngomo. "LSQ: the linked SPARQL queries dataset." In International semantic web conference, pp. 261-269. Springer, Cham, 2015.
18. Stadlera, Claus, Muhammad Saleema, Qaiser Mehmoodb, Carlos Buil-Arandac, Michel Dumontierd, Aidan Hogane, and Axel-Cyrille Ngonga Ngomoa. "LSQ 2.0: A Linked Dataset of SPARQL Query Logs."
19. Liu, Yike, Tara Safavi, Abhilash Dighe, and Danai Koutra. "Graph summarization methods and applications: A survey." ACM computing surveys (CSUR) 51, no. 3 (2018): 1-34.
20. Safavi, Tara, Caleb Belth, Lukas Faber, Davide Mottin, Emmanuel Müller, and Danai Koutra. "Personalized knowledge graph summarization: From the cloud to your pocket." In 2019 IEEE International Conference on Data Mining (ICDM), pp. 528-537. IEEE, 2019.