Leveraging Ontologies for Entity Subgraph Extraction

Katrina Schweikert[1]

[1]School of Comp. and Inf. Science, Univ. of Maine, Orono, ME, USA

Abstract

Knowledge graphs are a structured form of knowledge representation that connects entities to other entities and to literal data values via relations. Combined with an ontology that captures relationships between concepts, knowledge graphs can provide a flexible structure that supports logical inference, aligns data, and adapts to a variety of knowledge tasks. A number of these tasks, such as question answering and entity summarization, often begin with understanding the context of one entity of interest (either an instance or a class). Querying the multi-hop neighborhood of entity to three or more hops can quickly become intractable in large knowledge graphs, but many existing ontology design patterns create meaningful entity relationships that span more than two relation hops in the graph. This research aims to use both ontology reasoning and graph analysis to build a contextually relevant subgraph around an entity of interest. This targeted subgraph extraction will have useful applications for a variety of upstream tasks related to human interaction with the knowledge graph via exploration and question answering.

Keywords

Knowledge Graph Subgraph, Ontology Modularization, Entity Summarization

1. Introduction

This research project investigates methods for extracting a subset of the neighborhood around one instance entity in a knowledge graph that contains both data (instance axioms i.e. ABox) and an ontology (TBox). The primary motivation and application for this task is to delineate the relevant context for a spatial entity in a geospatial knowledge graph. A geospatial knowledge graph, especially one that utilizes spatial entities like those in a discrete global grid, can contain a multitude of information that describes the environmental, social, and economic attributes of a region. For example, the SAWGraph project connects information about chemical contamination by PFAS in the environment of the continental U.S., with hydrology, soil properties, crop cover, census data, etc. . The graph uses topological enrichment [1] which explicitly relates all spatial entities to S2 Level 13 discrete global grid cells and Level 3 Administrative Regions (towns and townships). Relevant to the use cases for the development of this graph are a number of questions that involve describing the environmental or social context of a location. For this type of questions, what is interesting and reveling, is not just the entities that are directly *contained in* or otherwise *spatial related to* the area of interest, but also the attributes and related non-spatial entities that help describe those spatial entities. For example one S2 cell might contain a number of drinking water wells of various depths and types, and a number of streams or lakes with varying depths and water properties. The S2 cell also might have a number of observations tied to it that in turn have properties describing the percent coverage of particular crop types, or soil types, or census demographic categories. Alternately, the S2 cell also would have relation to a set of neighboring S2 cells as well as larger administrative region like the town, county and state, but following those directly linked entities any further to find their properties quickly explodes the size of the subgraph and is not directly describing the context of the S2 cell of interest. The subgraph extraction task seeks to leverage the ontology and graph statistics to delineate a subgraph around a single entity (or a template shape for all the entities in a class) which captures the relevant context for that entity. For each class of entities, this subgraph will take a different shape and follow different path lengths in various directions

FOIS with Proceedings of FOIS 2025 Satellite events co-located with the 15th International Conference on Formal Ontology in Information Systems (FOIS 2025), September 10-12, 2025, Catania, Italy

katrina.schweikert@maine.edu (K. Schweikert)

0000-0003-3271-6700 (K. Schweikert)

to reach literals and controlled vocabulary instances that would be considered relevant context for the entity being investigated. This task has additional relevance beyond just spatial entities and beyond the task of entity context summarization.

Subgraph extraction is a useful task for a variety of additional challenges related to knowledge graphs, including question answering on knowledge graphs, ontology exploration, entity summarization and semantic similarity between entities. These challenges often require a starting graph on which to run statistics, perform reasoning, or render a visualization to a user that is a less dense graph. An initial subgraph extraction can speed-up and simplify further processing. Filtering to only the most important statements around a focal point in the graph, and hiding other parts of the graph that are less strongly relevant to a task indeed has uses both for human interaction with knowledge graphs (ontology exploration and other graph visualization) as well as computation (efficient question answering from a starting entity, entity summarization, etc.). Complex question answering is a more challenging variation on the question answering task that can involve entities that are many hops (relations) away from the starting entity, and many question answering pipelines use the neighborhood of the starting entity as candidates from which to select the answer [2]. Similarly, most existing entity summarization techniques [3] use the one-hop neighborhood around the entity to derive the summary, and a multi-hop neighborhood could produce a more robust and informative summary. However, in both cases the unfiltered multi-hop entity neighborhood can quickly encompass a large part of the graph [4], and many knowledge graphs that include instance data contain millions of triples. Querying the full entity neighborhood, even to three hops, can quickly become intractable in large knowledge graphs. Additionally, tasks that involve humans, such as graph exploration and entity resolution, place a similar limitation on the amount of information that can be handled cognitively.

Many common ontology design patterns create multiple sub-entities to represent what in a traditional tabular database would constitute one record, or one data view in a relational database. For example PROV-O [5], a widely used ontology for representing provenance information contains three main classes, *Activity*, *Entity* and *Agent*. These are connected directly by a number of object properties that represent different relationships. However, also critical to the PROV-O ontology is the qualified relation pattern [6], where intermediate entities can be added to hold additional attributes that qualify the relation. For example an *Activity* can have a qualified usage of an *Entity* that includes the time that the usage occurred. This same data in a traditional tabular database may just be one *Activity* record with attributes containing the time and the *Entity* used. If *Activity* is the main entity of interest, *datetime* would not be directly connected to it, and neither would any additional properties about the *Entity* such as that it derived from another *Entity*, but this information is likely relevant parts of the *Activity* context. The creation of atomic entities is ontologically justified, but leads to challenges in delineating the proper extent of relevant facts related to an entity for both algorithms and humans interacting with the graph. For example, a sample or feature is at least two hops from its observed value in SOSA [7], or a quantity to it's measured value in QUDT[8]. Owl-Time [9] is another example where an entity and it's literal *datetime* value can be separated by multiple intermediary entities. While its possible to design the ontology to include property chains that shortcut and add direct relations between an instance and its identifying datatype property while still preserving the more ontologically robust and atomic design patterns, property chains are typically used in moderation to avoid exploding the graph and to preserve the ontological distinctiveness of different concepts within the graph.

What is needed is a principled method to extract a subgraph around an entity of interest, reaching multiple hops away from the starting entity, but terminating at different hop distances in each direction once the connected entities can no longer be considered an extension of the description of the entity of interest. The goal of this research is to extract a meaningful subgraph that captures the distinguishing relationships of a starting entity, without traversing too far into neighboring entities, by leveraging both the ontology and the graph structure. One strategy would be to capture the pattern of the ontology in reference to a starting class and map it to a query on the instances to build a subgraph. Alternately, querying the instances directly in the full graph can also reveal a different relationship between types of entities that might not be explicit in the ontology or may span multiple ontologies. Another strategy is to handle different semantic relationships between entities in different ways, for relationship types like

mereology, subsumption, topology, and functional relations. Additionally, comparing graph statistics calculated on classes in the ontology to statistics calculated on instances in the ABox can inform the role that some related entities may play.

2. Background / Related Work

Entity summarization focuses on providing a short summary of an entity in a Knowledge Graph, mostly for the purpose of presenting a user with a digestible amount of information [3]. Existing entity summarization methods generally only look at entities, literals and classes directly related to the entity of interest, and struggle to handle even blank nodes in a knowledge graph. They also focus on more heuristic measures for ranking the most important facts to add to the summary, such as informativeness and diversity. By contrast the proposed subgraph extraction task focuses on selecting multiple triples, not just those directly connected to the entity of interest, to comprehensively describe the entity context.

RDF graph, Knowledge Graph and ontology summarization tasks tackle summarization of an entire graph to aid in user exploration and understanding [10] [11]. These summarization tasks differs from the proposed entity subgraph extraction because they don't focus on a particular starting entity or class of interest. The tasks vary in whether they include just TBox ontology statements, or also ABox instance assertions. This area of study does include a number of logical and statistical methods that could be repurposed for focusing on a single entity of interest, such as pattern mining, decomposition, quotients [10] , and methods for identifying key concepts in an ontology [12].

The entity subgraph extraction task is very similar to ontology modularization, especially those modularization methods that extract a section of an ontology to perform a specific task, referred to as a knowledge hiding motivation [13] or as forgetting [14]. However few ontology modularization techniques operate on both the TBox ontology assertions and the ABox data assertions. Additionally many knowledge graphs utilize multiple ontologies (in whole or in part), and often the areas where the ontologies connect or where the useful information resides only becomes visible from the ABox instance data, not in the TBox axioms alone.

Another closely related concept is sub-graph extraction for question answering. In order to constrain the search space for question answering on knowledge graphs, methods such as personal page rank information propagation are applied to the graph to prune to the most relevant parts for the question [4]. Some question answering systems rely on a single starting entity, but other systems that tackle complex questions may also include additional entities identified from the question[2]. The two main differences between this task and the proposed entity subgraph extraction are the additional guidance that the question can provide, and the lack of methods that incorporate logic or ontology reasoners in the existing work on sub-graph extraction for question answering.

3. Approach

The proposed approach to entity subgraph extraction is a hybrid logical and graphical methodology. The method performs both reasoning on the ontologies and instances, and also computing graph statistics on both the TBox and ABox graphs. These methods are combined to arrive at a subgraph around an entity of interest in a efficient and repeatable manner.

By first reasoning on the core ontology in the graph to which the instance belongs, we can obtain graph closure under logical entailment, and begin to identify walks through the ontology from the starting class via domain and range restrictions, universal and existential quantification, and property chain axioms. The starting class is defined by the entity of interest, and there may be more than one depending on the ontology complexity. Transitive reduction may also assist in simplifying the possible pathways. Next we turn to a sampling of the instances in the graph to do further reasoning. Preliminary testing shows this needs to be limited to a subset of instance data for performance reasons. Reasoning on the instances may identify pathways that connect different ontologies, such as instances from a class in one ontology connected by relations from another ontology. Additionally instance data can

potentially predict some domain and range values for relations that either cross ontologies or are not explicitly defined in the TBox of one ontology. The method for querying the full graph is similar to that employed in [15] for extracting validating SHACL shapes on very large knowledge graphs. However to build the full context shape for the class, multiple class shapes will be combined, and then translated into a sparql query that extracts the subgraph, rather than a SHACL shape.

The last proposed step in the approach is using graph analysis techniques such as calculating the degree of nodes in both the ontology only graph (TBox as a graph) and the instances graph (ABox sample as a graph). By comparing graph statistics between the ontology class, and a summary of instances of the class we may be able to further identify triples which should be included in or excluded from the subgraph. One such example is controlled vocabularies that are implemented as class instances. These are akin to rdf:type / 'is a' statements, and traversing the graph past them generally leads to 'sister' instances that may not be particularly useful for a summative understanding of the entity of interest. For example in SAWGraph, a PFAS sample has an object property relation to an instance of the sample material type class, which if followed further would lead to all other samples of the same type. This implementation of a controlled vocabulary is useful for practical querying of the graph rather than directly translating every type attribute to a subclass structure, however it becomes one of the key challenges for delineating an appropriate subgraph.

4. PhD Research Context and Future Work

This research is part of SAWGraph project, to build an open knowledge network for environmental contamination data about PFAS contamination. It represents an essential task in support of a couple of different future work goals related to the project.

The first goal relates to a user interface for the graph, which allows users to build interactive queries and provides map-based visualization of the results. We would like to present filters to the user based on selected starting classes of entities, and then allow them to further refine the query. The interface is similar to faceted search in terms of the filtering, but allows connecting two different entities via a spatial relation. However we don't want to present the user with all possible data values as filter options, and therefore allow them to build a complex query which has no result in the knowledge graph. We need a mechanism to dynamically update the filters relating to the possible range of values of connected attributes to the set of selected features, and achieve this result dynamically at near real-time to create a responsive interface. Creating an entity subgraph for the possible relevant connections to instances of entities will facilitate building an efficient query programmatically for any starting entity class, rather than manually for each possible starting entity type in the interface. It would also allow the interface to adapt as the graph grows.

The second use of entity subgraphs is for future work related to multi-hop entity summarization. For a knowledge graph on environmental data, one goal of the graph is to elucidate the environmental context in which the observed contamination is occurring, and potentially to be able to compare and pinpoint the differences in environmental context. In the geospatial context especially, single-hop entity summaries don't capture sufficient contextual information, where there is often a nesting of spatial entities. Measurement are associated with spatial entities, but represent their own distinct entities. Design patterns for representing observations and measurements, as described above, also have complex multi-hop contexts. An ideal contextual summary would include not just a filtering of directly related facts, but also an aggregation of associated entities and their range of values. For example the range of PFAS concentration for all samples in an area, together with the range of hydrology features and their properties in the same area, and properties of facilities in the area that are emitting or suspected of using PFAS chemicals together builds a picture of the environmental context. This can be compared to the environmental context at another location, to help elucidate what properties might be impacting the observations. This can lead to new hypothesis about what environmental factors in these complex systems could be contributing to higher or lower accumulation of PFAS concentrations.

Acknowledgements This work and the development of SAWGraph have been supported by the National Science Foundation (NSF) under Grant No. 2333782 as part of the Proto-OKN initiative (https://www.proto-okn.net/) and by a Non-Assistance Cooperative Agreement with the USDA-ARS New England Center for Sustained Soil and Water Health. I would like to thank my advisor, Torsten Hahmann, and the entire SKAI Lab and SAWGraph teams.

Declaration on Generative AI The author has not employed any Generative AI tools

References

[1] S. Stephen, M. Faulk, K. Janowicz, C. Fisher, T. Thelen, R. Zhu, P. Hitzler, C. Shimizu, K. Currier, M. Schildhauer, et al., The S2 hierarchical discrete global grid as a nexus for data representation, integration, and querying across geospatial knowledge graphs, arXiv preprint arXiv:2410.14808 (2024).

[2] Y. Lan, G. He, J. Jiang, J. Jiang, W. X. Zhao, J.-R. Wen, Complex knowledge base question answering: A survey (2021). URL: http://arxiv.org/abs/2108.06688.

[3] Q. Liu, G. Cheng, K. Gunaratna, Y. Qu, Entity summarization: State of the art and future challenges, Journal of Web Semantics 69 (2019) 100647. URL: http://arxiv.org/abs/1910.08252.

[4] S. Aghaei, K. Angele, A. Fensel, Building Knowledge Subgraphs in Question Answering over Knowledge Graphs, Fensel, 2022, pp. 237–251. URL: https://link.springer.com/10.1007/978-3-031-09917-5_16. doi:10.1007/978-3-031-09917-5_16.

[5] T. Lebo, S. Sahoo, D. e. a. McGuinness, PROV-O: The PROV Ontology, Technical Report, 2013. URL: http://www.w3.org/TR/prov-o/.

[6] L. Dodds, I. Davis, Linked data patterns: a pattern catalogue for modelling, publishing, and consuming linked data, L. Dodds, I. Davis, 2011.

[7] K. Janowicz, A. Haller, S. J. Cox, D. Le Phuoc, M. Lefrançois, Sosa: A lightweight ontology for sensors, observations, samples, and actuators, J. Web Semantics 56 (2019) 1–10.

[8] R. Hodgson, P. J. Keller, J. Hodges, J. Spivak, QUDT: Quantities, units, dimensions and types, https://qudt.org/, 2022.

[9] J. R. Hobbs, F. Pan, Time ontology in OWL, https://www.w3.org/TR/2006/WD-owl-time-20060927/, 2006.

[10] Šejla Čebirić, F. Goasdoué, H. Kondylakis, D. Kotzinos, I. Manolescu, G. Troullinou, M. Zneika, Summarizing semantic graphs: a survey, The VLDB Journal 28 (2019) 295–327. URL: http://link.springer.com/10.1007/s00778-018-0528-3. doi:10.1007/s00778-018-0528-3.

[11] S. Pouriyeh, M. Allahyari, K. Kochut, H. R. Arabnia, A comprehensive survey of ontology summarization: Measures and methods, CoRR (2018). URL: http://arxiv.org/abs/1801.01937.

[12] S. Peroni, E. Motta, M. d'Aquin, Identifying Key Concepts in an Ontology, through the Integration of Cognitive Principles with Statistical and Topological Measures, volume 5367, Springer Berlin Heidelberg, 2008, pp. 242–256. URL: http://link.springer.com/10.1007/978-3-540-89704-0_17. doi:10.1007/978-3-540-89704-0_17.

[13] A. L. Clair, A. Marinache, H. E. Ghalayini, W. Maccaull, R. Khedri, A review on ontology modularization techniques - a multi-dimensional perspective, IEEE Transactions on Knowledge and Data Engineering 35 (2022) 1–1. URL: https://ieeexplore.ieee.org/document/9721157/. doi:10.1109/TKDE.2022.3152928.

[14] T. Eiter, G. Ianni, R. Schindlauer, H. Tompits, K. Wang, Forgetting in managing rules and ontologies, in: 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings)(WI'06), IEEE, 2006, pp. 411–419. URL: http://ieeexplore.ieee.org/document/4061405/. doi:10.1109/WI.2006.83.

[15] K. Rabbani, M. Lissandrini, K. Hose, Extraction of validating shapes from very large knowledge graphs, Proceedings of the VLDB Endowment 16 (2023) 1023–1032. URL: https://dl.acm.org/doi/10.14778/3579075.3579078. doi:10.14778/3579075.3579078.