A Pipeline for Population and Analysis of Personal Health Knowledge Graphs (PHKGs)

Dagmar Celuchova Bosanska[1,*], Michal Huptych[2] and Lenka Lhotská[1,2]

[1]Faculty of Biomedical Engineering, Czech Technical University in Prague, Czech Republic
[2]Czech Institute of Informatics, Robotics, and Cybernetics, Czech Technical University in Prague, Czech Republic

Abstract

Personal Health Knowledge Graphs (PHKGs) are not yet ubiquitous, even though they have a great potential to enrich general knowledge captured in various Knowledge Graphs by adding personal contexts. This poster paper presents work in progress about a pipeline for generating PHKGs from tree-structured Electronic Health Record (EHR) data by applying a hierarchical ontological approach. This pipeline could also be applied to other domains of Personal Knowledge Graphs. Moreover, this pipeline targets the intersection between the symbolic representation of knowledge used for computational semantics and numeric graph data representation used for graph analysis and machine learning. We present the first results from applying this pipeline to synthetic patient EHRs with the diagnosis of colorectal cancer (based on Synthea). The resulting numeric representation of PHKGs or their subgraphs can be used in many practical graph algorithms. Finally, our pipeline study uncovers future research on how this numeric representation of PHKGs should be embedded into continuous and low-dimensional vector space to utilize graph machine learning and deep learning methods.

Keywords

Personal Health Knowledge Graphs, Ontology, Graph algorithms, Machine Learning

1. Introduction

Personal health knowledge graphs (PHKGs) represent structured information about entities related to a patient's health and well-being, attributes, and relations between them. Unlike Knowledge Graphs (KGs), PHKGs are not yet ubiquitous. PHKGs should be generated for individual patients from numerous information sources such as electronic health records (EHRs), wearables and mobile health apps, sensors, and patient annotated texts and notes related to the patient's condition. In principle, PHKGs can be populated by all techniques mentioned in [1] if the information source contains personal data or data related to a patient and her health and well-being. However, no agreed representation and population of PHKGs exists. For instance, a KG for asthma can describe causes, symptoms, and treatments for asthma, and PHKG can be the subgraph containing just those causes, symptoms, and treatments that apply to a given patient [2]. Another point of view is that PHKGs can be used to add personal context to KGs and to help develop a personalized diagnosis, recommendations, and treatments [3].

SEMANTICS 2022 EU: 18th International Conference on Semantic Systems, September 13-15, 2022, Vienna, Austria
*Corresponding author.
celucdag@fbmi.cvut.cz (D. C. Bosanska)

This poster paper presents a work in progress about a pipeline on how to populate a PHKG from several data sources. First, the data must be harmonized according to a flexible and helpful data model for data analysis. Then, we apply this approach to the PHKG population from data stored in EHRs. In addition, ongoing work will provide evidence that this approach generalizes to PKGs and other data sources, and enables linking PHKGs with KGs and generating new KGs.

2. How to represent a PHKG and set up a pipeline

The proposed representation of a PHKG is composed of two elements: a domain graph and a mapping from the nodes and edge labels of the data graph to those of the domain graph in which they are called entities and relation-types, respectively [1]. The domain graph defines the schema of the PHKG – its high-level structure that can evolve more flexibly than a schema for a relational model. Using a harmonizing data model – the Simple Event Model Ontology [4] enables us to view EHR data in the HL7 FHIR RDF format as evolving chains of events and sub-events in time. This ontology represents a schema that simplifies further analysis and manipulation of PHKG graphs. Its practicality has been proven for Event-Centric Temporal Knowledge Graph [5]. In our case of PHKGs, events are central elements in representing a patient's experience with a concrete disease. This experience includes visits (encounters), reported complaints and symptoms (for example, in the form of observations), performed procedures, prescribed medications, finalized diagnostic reports, and even applied care plans. In terms of self-management, it can be, for example, exercising, self-measurements, or sleep monitoring. This ontology allows us to write easy-to-understand SPARQL queries without a more profound understanding of the domain ontologies such as HL7 FHIR RDF. The PROV-O ontology can capture the administrative part of health records, such as who created the entry, when, how, and in which institution.

Ontologies such as SNOMED CT and LOINC (also a part of UMLS) give identity to our nodes in PHKGs [6]. This identity denotes which nodes in PHKGs, or external KGs refer to the same real-world entity. Thanks to these ontologies, we can use the subsumption relationships to align the node labels to the same, more general term and thus improve the ontological graph union operation (see Section 3). In the clinical and self-care setting, the labels of the corresponding nodes would hardly be the same if there was no method of standardization and subsumption with the help of mentioned ontologies in place.

A PHKG is thus represented as a directed edge-labeled graph that enables querying and reasoning. However, most graph data analysis techniques do not apply to this representation. Therefore, we need to transform it into an undirected or directed graph without edge labels (i.e., predicate names). A directed graph is thus projected by optionally selecting a sub-graph from the data graph from which all edge labels can be dropped. The proposed pipeline to populate a PHKG from various data sources and use it for graph analysis is as follows:

1. **Find or create an ontology for data harmonization**: Our research suggests that the mentioned Simple Event Model Ontology is a universal and useful upper-ontology that can be extended by more specific models for various event types.
2. **Choose a standard or format for harmonized data transformation into graph dataset**: As our chosen example of HL7 FHIR data was available in a tree-structured

format (JSON) and a tool to convert the data into HL7 FHIR RDF format existed (see Section 3), the choice was straightforward. But for different categories of data other formats such as KGTK [7] or property graphs may be easier to apply.

3. **Assert relations** between nodes/edges in the graph dataset and nodes (classes) / edges (properties) from the ontology chosen for data harmonization (or even for provenance): It means that more facts stated as triples will be added to the graph dataset representing the PHKG. Additional information about events will be available in the data from various sources in their original data model, such as HL7 FHIR RDF.
4. **Create a subgraph** from the source PHKG with the help of SPARQL based on the data harmonization ontology.
5. **Convert the symbolic representation of the subgraph into its numeric representation for a directional graph** (for example into an adjacency matrix) for further analysis and transformation: Make use of the subsumption provided by linked ontologies (SNOMED CT and LOINC in our case) to unite node and/or edge labels. If possible, store edge and node labels and other meta data within the (sub)graph structure. The symbolic representation is thus as follows:
   - $\mathbf{V}$ is a vertex set - a set of nodes {a, b, c, d, ...}.
   - $\mathbf{A}$ is an $|\mathbf{V}| \times |\mathbf{V}|$ adjacency matrix (assume binary if there are no edge weights).
   - $\mathbf{X} \in \mathbf{R}^{\mathbf{m} \times |\mathbf{V}|}$ is a matrix of node features, such as node identity (URI and the SNOMED-CT code) and the begin timestamp.
6. **Apply graph algorithms**, such as operators, distance measures, and shortest paths to one or more PHKGs in their numeric form. Even if the edge labels and other edge and node meta data are dropped for further graph analysis, it is a good practice to create a possibility to retrieve this information even for the outputs of the graph analysis.
7. If needed, convert the result of the graph analysis back into the symbolic form to expand the original PHKG by new knowledge.

This choice of a strategy to convert any source PHKG into its symbolic representation for analysis is not trivial. It requires empirical validation and iterations of the steps 5 and 6 to develop new insights from the graph analysis. In addition, more study is needed to understand the effects of such strategies more generally on the results of different analytical techniques.

3. The usage example

We applied the methods and pipeline proposed in the Section 2 on synthetic data generated for the colorectal cancer diagnosis according to [8]. The output data from this generator are in the JSON format based on the HL7 FHIR standard. In addition, datasets for individual patients were converted to the HL7 FHIR RDF format using the FHIR JSON to RDF conversion utility[1] (on the HL7 FHIR webpage, other open source implementations can be found). We populated, visualized, and analyzed the PHKGs for individual patients using Python libraries RDFlib[2],

[1] https://github.com/BD2KOnFHIR/fhirtordf
[2] https://github.com/RDFLib/rdflib

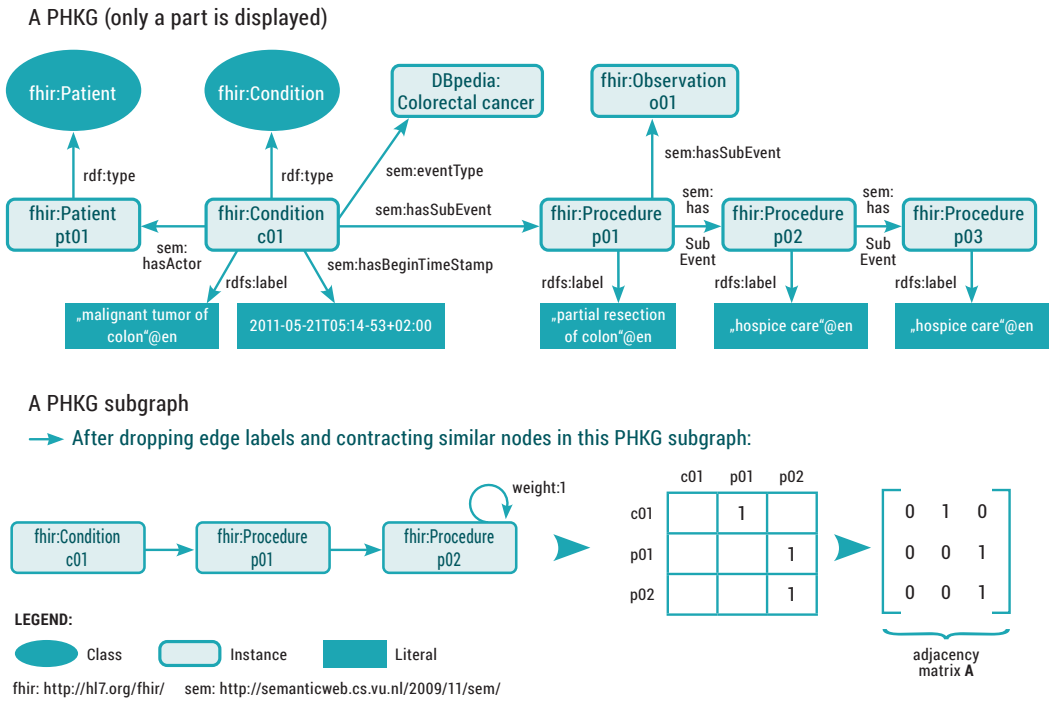

Figure 1: The method of data harmonization transformation into numeric form for graph analysis

Owlready2 with its PyMedTermino2 for easy access to domain ontologies[3] and NetworkX [9]. We needed to implement tools to switch between different graph data structures of these libraries to implement the whole pipeline for graph data analysis.

Once we had the numeric representation of PHKG subgraphs for our 325 synthetic patients, performed according to the Figure 1, we developed an algorithm for an ontological graph union to create a KG containing different patient pathways from the point of diagnosis to the outcome. This algorithm can contract nodes of the same type (in our case, the same SNOMED CT code of the underlying procedure) across patient PHKGs while preserving all node features. In theory, if the set of PHKGs were representative enough, this KG would cover all possible ways of treatment and their outcomes. We analyzed the most connected nodes - procedures with the help of the degree centrality. They represent the key decision points in the patients' treatment or life events.

It is also possible to analyze the shortest simple paths from the principal diagnosis to the outcome, considering the weights. For our group of patients, the shortest simple path from the diagnosis to the unfortunate event of death is length four because the diagnosis happened at a very late stage of the disease.

[3]https://owlready2.readthedocs.io/en/v0.37/

4. Conclusions and future work

In our future work, we will further explore the best combination of ontologies, knowledge, and data harmonization to create a universal pipeline for the PKG population. In addition, as we can see in Figure 1, the adjacency matrix is sparse (the sparsity equals 2/3 in this case), and this fact about the numeric representation holds in general. Therefore, a more efficient representation in continuous and low-dimensional vector space with the help of node and graph embedding should be researched.

Finally, in our pipeline, we found a more straightforward representation of the partial PHKG knowledge in which we could drop the edge labels (the nodes representing the events (procedures) were connected only by a sub-event relationship to capture the sequence of events in time). However, the more properties the graph embedder encodes, the better results can be retrieved in later tasks. Therefore, we can generate a more complex subgraph from our source PHKG with edge labels.

References

[1] A. Hogan, E. Blomqvist, M. Cochez, et al., Knowledge graphs, ACM Computing Surveys 54 (2021). doi:10.1145/3447772. arXiv:2003.02320.

[2] A. Gyrard, M. Gaur, S. Shekarpour, K. Thirunarayan, A. Sheth, Personalized health knowledge graph, in: CEUR Workshop Proceedings, volume 2317, 2018, pp. 1–6.

[3] O. Seneviratne, J. Harris, C.-h. Chen, D. L. McGuinness, Personal Health Knowledge Graph for Clinically Relevant Diet Recommendations, Workshop on Personal Knowledge Graphs Co-located with the 3rd Automatic Knowledge Base Construction Conference (AKBC'21) (2021). URL: http://arxiv.org/abs/2110.10131. arXiv:2110.10131.

[4] W. R. Van Hage, V. Malaisé, R. Segers, et al., Design and use of the Simple Event Model (SEM), Journal of Web Semantics 9 (2011) 128–136. doi:10.1016/j.websem.2011.03.003.

[5] S. Gottschalk, E. Demidova, EventKG: A Multilingual Event-Centric Temporal Knowledge Graph, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 10843 LNCS (2018) 272–287. doi:10.1007/978-3-319-93417-4_18. arXiv:1804.04526.

[6] M. Ivanović, Z. Budimac, An overview of ontologies and data resources in medical domains, Expert Systems with Applications 41 (2014) 5158–5166. doi:10.1016/j.eswa.2014.02.045.

[7] F. Ilievski, D. Garijo, H. Chalupsky, et al., KGTK: A Toolkit for Large Knowledge Graph Manipulation and Analysis, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 12507 LNCS (2020) 278–293. doi:10.1007/978-3-030-62466-8_18. arXiv:2006.00088.

[8] J. Walonoski, M. Kramer, J. Nichols, et al., Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record, Journal of the American Medical Informatics Association 25 (2018) 230–238. URL: https://academic.oup.com/jamia/article/25/3/230/4098271. doi:10.1093/jamia/ocx079.

[9] A. A. Hagberg, D. A. Schult, P. J. Swart, Exploring network structure, dynamics, and function using NetworkX, 7th Python in Science Conference (SciPy 2008) - (2008) 11–15.