Development of a Graph Model for the OMOP Common Data Model

Mengjia Kang, MS[1], Jose A. Alvarado-Guzman, MS[2], Luke V. Rasmussen, MS[1], Justin B. Starren, MD, PhD[1]

[1]Northwestern University, Feinberg School of Medicine, Chicago, Illinois; [2]Neo4j, Inc., San Mateo, California

Introduction

Current phenotyping and systems biology research requires not only integration of large volumes of Electronic Health Record (EHR) and multi-omics data, but also capturing the multitudes of relations among the concepts. Graph databases have emerged as a promising technology for such tasks, supporting not only local analysis but also global analysis leveraging graph algorithms like Centrality, Community Detection, Path Finding or Node Embeddings[1]. Unfortunately, EHR data is rarely available in a graph format. While a naïve row-to-node conversion is possible, the resulting graph is typically attribute-heavy, resulting in suboptimal performance. To address this limitation, we developed a modelling method to convert data form the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) to the Neo4j [www.neo4j.com] graph property model.

Methods

The Successful Clinical Response in Pneumonia Therapy (SCRIPT) is a five-year systems biology study that is integrating clinical, transcriptomic, metagenomic and bacterial genomic data to support Machine Learning on host pathogen interaction. Our modeling focused on nine OMOP Standardized Clinical Data Tables (PERSON, PROVIDER, OBSERVATION_PERIOD, VISIT_OCCURRENCE, CONDITION_OCCURRENCE, DRUG_EXPOSURE, PROCEDURE_OCCURRENCE, MEASUREMENT, OBSERVATION) and four Standardized Vocabularies Tables (CONCEPT, DOMAIN, VOCABULARY and CONCEPT_CLASS) which captured the SCRIPT clinical data. Our overall strategy was to encode as much information as possible in the edge topology to take advantage of the intrinsic strengths of the graph database. In general, nominal and categorical data were converted to nodes; foreign keys to edges; and numerical values to node or edge properties as appropriate. We also implemented self-directed relationships RELATED_TO and NEXT on the Concept and VisitOccurrence node separately. The former defines the nature and type of direct relationships between any two Concepts and the later builds up the patient journey.

Results

Our finalized graph property model was implemented using a local installation of Neo4j 4.0.2 Community Edition. It includes 16 types of nodes (entities) and 22 types of edges (relationships) as well as 55 node properties. This model contains on average 3.44 attributes per node. This work is available in both a markdown and Cypher query language format in our GitHub repository, https://github.com/NUSCRIPT/OMOP_to_Graph.

Discussion

Although more data preprocessing is required to load the data into our graph property model than the naïve row-to-node conversion method, previous work has demonstrated that the analytics performance will be greatly improved[2]. This model also reduces redundancy by eliminating the denormalization (Foreign Keys) that is often added to relational databases (e.g. person_id occurs in all other OMOP tables). The model was developed for the SCRIPT project, but the transforms can be applied to other OMOP CDM v5.x databases. Our current and future work will further demonstrate graph analytics examples using the SCRIPT EHR data.

References

1. Fabregat A, Korninger F, Viteri G, Sidiropoulos K, Marin-Garcia P, Ping P, et al. (2018) Reactome graph database: Efficient access to complex pathway data. PLoS Comput Biol 14(1): e1005968. https://doi.org/10.1371/journal.pcbi.1005968
2. Alvarado-Guzmán JA, MS, Keren I, MS https://www.ohdsi.org/web/wiki/lib/exe/fetch.php?media=resources:jose_alvarado_rd2gd_ohdsi_submission_2017.pdf