Graph-KD: Exploring Relational Information for Knowledge Discovery

Roland Roller[1], Gaurav Vashisth[1], Philippe Thomas[1], He Wang[1], Michael Mikhailov[1] and Mark Stevenson[2]

[1] DFKI, Berlin, Germany
[2] University of Sheffield, Sheffield, England

Abstract. This paper presents Graph-KD, a tool to navigate through large relational knowledge sources. Graph-KD provides methods to understand relationships between concepts using open discovery, closed discovery and knowledge inference. The purpose of the tool is the support of biomedical knowledge discovery and exploration. It is primarily intended to be used by medical researchers and presents a use case involving millions of relations from UMLS. Graph-KD is able to process even large graphs efficiently and can be accessed via a web-interface (`http://biomedical.dfki.de/graph-kd`).

1 Introduction

Relational knowledge bases and ontologies are rich sources of concepts and the relationships between them which often consist of large amounts of information. These resources generally include information about directly related concepts but the information about those related indirectly can also be extremely valuable. Exploring this information can provide further insights and can help to discover new knowledge.

Various tools exist to explore knowledge graphs such as UMLS. However, existing tools either have a different focus or/and cover only parts of the functionalities of Graph-KD. For instance, UMLS:Similarity [4] is a tool to measure semantic similarity and relatedness based on UMLS and provides a shortest path functionality. This covers only one aspect of Graph-KD and is only accessible via API. In k-neighborhood decentralization [8] methods for large scale knowledge discovery in context of UMLS are presented, which is related to the functionalities we provide, such as shortest path. Cantor et al. [2] offer a method to explore relationships between UMLS and the Gene Ontology. Using statistical and semantic relationships, it is possible to infer relationships between diseases and gene products. Gómez-Romero et al.[3] developed a Big Data graph processing and visualization pipeline in order to decrease processing time of large graphs. However, even if various tools exist to explore relational information of UMLS none of them provides easy-to-use functionalities techniques to explore and understand how long-range information are connected with each other.

2 Graph-KD

Graph-KD provides functionalities to explore a large knowledge graph using open and closed discovery as well as knowledge inference. Open and closed discovery base on the

idea of literature-based discovery (LBD) which was introduced by Swanson [6]. The author found that fish oil may have beneficial effects in patients with Raynaud syndrome, a fact which was not known beforehand. Swanson noticed that, although no connection between fish oil and Raynaud syndrome was known, they shared a number of common connections. It was known that fish oil lowers blood viscosity, inhibits platelet aggregation and causes vascular reactivity. Conversely, it was also known that patients with Raynaud syndrome have increased blood viscosity and platelet aggregation and suffer from impaired vascular reactivity [7].

One of the major problem of LBD is the enormous number of connections which effectively rules out checking all possibilities. Therefore, it is very important to concentrate on the most relevant connections during the early stages of research. Graph-KD supports this process by providing useful information in an easily navigable format. The following functionalities are provided:

Closed Discovery applies k-shortest path to find relevant connections between two concepts. Passing two concepts of interest to Graph-KB, the tool visualizes the connections between them. In case various relation paths exist, all shortest paths will be displayed. An example of the shortest paths between the class of pharmacologic substances (*nsaids*) and a particular disease (*kidney disease*) is presented in Figure 1.

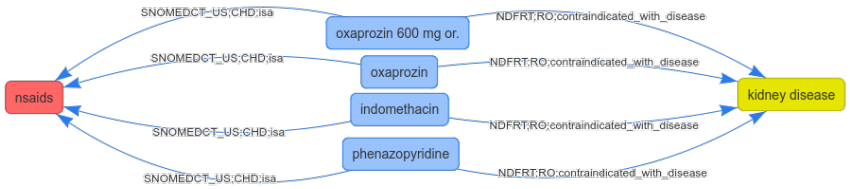

Fig. 1: Closed discovery between *nsaids* and *kidney disease*: There is no direct connection, however, the graph shows for instance that *kidney disease* can be a contraindication for several drugs which belong to NSAID family.

Open Discovery explores concepts and relations around a target concept. An example is provided in Figure 2. The graph shows the target concept *antipyretics* in combination with the target relation *may-be-treated-by*. The open discovery searches for the relation of interest around the target concept in closest distance. Since *antipyretics* is not linked to any *may-be-treated-by* relation in our example, the tool shows nodes in close distance which are connected via this relation.

Knowledge Inference is a technique which takes existing facts into account and tries to make assumptions about unknown information. As knowledge graphs tend to be incomplete this can be a useful feature to support the *open* and *closed discovery* process. Graph-KD integrates a rule-based inference at this point.

The backend of Graph-KD is written in python and has access to Neo4j, an open source NoSQL scalable graph database management system, which stores data in form of nodes and their corresponding typed edges. Using Neo4j's built-in functionality the open and closed discovery can be executed with a good performance.

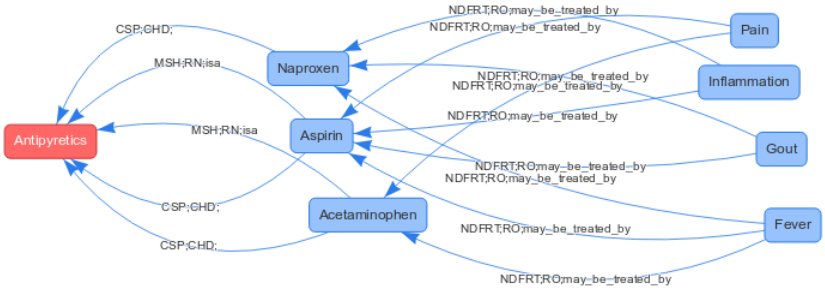

Fig. 2: Open discovery for node *antipyretics* and relation *may-be-treated-by*: As target node has no direct link to target relation, the graph shows nodes which connect to *may-be-treated-by* in closest distance.

3 Demo Use Case

For the demo use case, Graph-KD explores information from UMLS, a large biomedical knowledge base containing millions of medical terms and relations between them. UMLS defines medical concepts including their synonyms, and unifies them to a concept unique identifier (CUI). All those concepts are linked to at least one semantic type, such as *Body Part*, *Finding* or *Clinical Drug*. Moreover, UMLS defines relations between concepts which include for instance *isa*, *may-treat* or *contraindication-of*.

For our demo UMLS 2017AB was preprocessed and diminished. This included the removal of relations containing concepts related with itself, very general relations, inverted relations and concepts with less meaningful semantic types. Namely, semantic types of the semantic groups GEOG, OBJC, OCCU and ORGA, according to Bodenreider and McGray [1], were removed. The resulting data of more than 3 million different CUIs and 9.5 million relations were then imported into Neo4j.

The rule inference relies on transitivity rules, between hyponyms in combination with other relations (e.g. if A and B are related and B is a child of C, then we find a transitive relation between A and C).

Replication of Existing Discoveries In order to show the benefit of our tool, we explore long range dependencies by replicating existing discoveries as presented in Preiss et al. [5]. As shown in Table 1 Graph-KD is able to replicate former discoveries, such as *Raynaud disease* and *fish oil*. In all cases UMLS does not contain any direct connection. However, using Graph-KD it is possible to explore and understand, how information are connected within the complete graph. In most cases the distance (D) is 3. Furthermore, the table shows, that for all two target concepts pairs a large number of different shortest paths can be found (see #).

Runtime Table 2 shows the runtime for *open* and *closed discovery*. For both scenarios 200,000 randomly generated requests were sent via REST to the backend. The table shows that for approximately 8% (15,499) of all randomly selected CUI pairs a *shortest path* can be found (max distance 4). For those connections the average (mean) runtime is 0.15 seconds. The maximum response time is 175.02 seconds for closed and 0.51 for open discovery respectively. In addition to that, 75% of the requests are processed within less than a tenth of a second.

Discovery	D	#
Raynaud disease – Fish oil	3	127
Somatomedin C – Arginine	3	27
Migraine disorders – Magnesium	4	471
Magnesium deficiency – Neurologic disease	3	108
Alzheimer's disease – Indomethacin	3	105
Alzheimer's disease – Estrogen	3	100
Schizophrenia – Calcium-I. Phospholipase A2	4	22

Table 1: Exploring existing discoveries using Graph-KD

	closed discovery	open discovery
connections	15,499	6,793
mean	0.153920	0.006258
min	0.001387	0.001212
75%	0.060400	0.001599
max	175.018601	0.512222

Table 2: REST runtime test for 200k random requests

4 Conclusion

In this work we presented Graph-KD, a tool to explore large knowledge graphs. Graph-KD provides various functionalities for knowledge discovery and includes knowledge inference methods to gain further into the data. As our example in Table 1 showed, Graph-KD can be easily applied to support literature-based discovery. Moreover, other clinical use cases are possible in which physicians explore information in the knowledge graph in order to detect potential new links between medical concepts.

Acknowledgments

This project was funded by the European Union's Horizon 2020 research and innovation program under grant agreement No 780495 (BigMedilytics) and by the German Federal Ministry of Economics and Energy through the project MACSS (01MD16011F).

References

1. Bodenreider, O., McCray, A.T.: Exploring semantic groups through visual approaches. Journal of Biomedical Informatics **36**(6), 414 – 432 (2003), unified Medical Language System
2. Cantor, M.N., Sarkar, I.N., Bodenreider, O., Lussier, Y.A.: Genestrace: phenomic knowledge discovery via structured terminology. In: Biocomputing, pp. 103–114. World Scientific (2005)
3. Gómez-Romero, J., Molina-Solana, M., Oehmichen, A., Guo, Y.: Visualizing large knowledge graphs: A performance analysis. Future Generation Computer Systems **89**, 224 – 238 (2018)
4. McInnes, B., Pedersen, T., Pakhomov, S.: UMLS-Interface and UMLS-Similarity: Open Source Software for Measuring Paths and Semantic Similarity. In: Proceedings of the American Medical Informatics Association (AMIA) Symposium. San Fransico, CA (2009)
5. Preiss, J., Stevenson, M., Gaizauskas, R.: Exploring relation types for literature-based discovery. Journal of the American Medical Informatics Association **22**(5), 987–992 (05 2015)
6. Swanson, D.R.: Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspectives in biology and medicine **30**(1), 7–18 (1986)
7. Weeber, M., Kors, J.A., Mons, B.: Online tools to support literature-based discovery in the life sciences. Briefings in bioinformatics **6**(3), 277–286 (2005)
8. Xiang, Y., Lu, K., James, S.L., Borlawsky, T.B., Huang, K., Payne, P.R.: k-Neighborhood decentralization: a comprehensive solution to index the UMLS for large scale knowledge discovery. Journal of Biomedical Informatics **45**(2), 323–336 (2012)