Strategies to connect RDF graphs for link prediction using Drug-Disease Knowledge Graphs

Sophie Hallstedt[1], Nikita Makarov[1], Hossein Samieadel[1], Maria Pellegrino[2,3], Martina Garofalo[2,3], and Michael Cochez[1,2,4][0000−0001−5726−4638]

[1] Information Systems and Databases, RWTH Aachen University, Germany
{sophie.hallstedt,nikita.makarov,hossein.semieadel}@rwth-aachen.de
[2] Fraunhofer FIT, Sankt Augustin, Germany michael.cochez@fit.frauenhofer.de
[3] Department of Computer Science, University of Salerno, Italy
{mariaangelapellegrino94,margar1994}@gmail.com
[4] Faculty of Information Technology, University of Jyvaskyla, Finland

Abstract. Traditionally, drug development is a time-consuming and costly process. Using the vast amount of available data, it is hoped that new information can be mined or inferred automatically, reducing this cost. In this work, we present steps towards completing the ReDrugS KB, which others have used to predict interactions between various drugs and diseases. Our goal is to further complete this graph, without human intervention in the process, aiming at a high recall. For the link prediction, we used state-of-the-art embedding techniques for RDF graphs. The embeddings are fed into binary classifiers which predict the relation existence between entities. The ReDrugS knowledge graph is the combination of the results of many studies organised in 8 million named graphs. These graphs are not entirely disjoint and might even contain contradictory assertions. Hence, a significant challenge is making the named graphs suitable for graph embedding, by combining them into a single graph. In this work, we report on classification performance on the merged KG.

Keywords: Knowledge Graph · Embedding · Link prediction

1 Introduction

Traditionally, the drug discovery process has been long and time-consuming. Drug repositioning is a technique where known drugs are used to treat diseases other than the one they were originally designed for, improving efficiency and safety. Prediction of drug-disease interactions is promising for either drug repositioning or disease treatment fields. With the explosion of healthcare information, there has been a tremendous amount of heterogeneous knowledge which plays an essential role in healthcare information systems. The Knowledge Graph (KG) created by McCusker et. al [2] is an example of a KG containing data from multiple different sources. The ReDrugS KB contains 6.180 drugs, 3.820 diseases, 69.279 proteins and 899.198 interactions, all in 8 million named graphs. To deal with missing information, which could lead to forecast new possible methods, we apply machine learning methods to perform link prediction.

2 Exploring Link Prediction

Current methods for embedding of graphs assume a single KG. Unfortunately, the ReDrugS KB comprises 8 million individual named graphs, which need to be merged together. We tried two methods. In the first, we merged the triples from all named graphs into a single graph without including contextual information about the statements. In the second we sought to address the challenge where information in different graphs can sometimes be contradictory. For this, we applied an Intermediate-Node merging approach that works by adding additional nodes to the merged graph for entities which occur in more than one named graph. After we created a single KG from the merging step, either through the first direct merge or the second intermediate node merge, we applied a biased variation of the RDF2Vec embedding algorithm. [1]

For the link prediction, we trained a separate binary classifier for each of the most frequent relation types. For our dataset, we selected 1.8 million triples from the KG and another 0.2 million were randomly chosen entity pairs. For each binary classifier, the data set was transformed by replacing each subject and object with their respective embeddings from the previous step, and was then split into 90/10 training/test datasets. Subsequently, each of the classifiers was trained such that it learns to predict whether its relation exists between the given pair. To evaluate the classification, precision, recall and accuracy scores were computed. An accuracy of over 95% was achieved, albeit somewhat highly misleading due to the highly imbalanced data. The best performing classifier, based on recall, was Naive Bayes, which had a recall score between 30%-50% for all except two link types, with low precision being under 15% for all link types. Precision and recall revealed that the classifiers were sub-optimal, however, as the goal was to complete the graph, recall was considered more important than precision. While we have not identified any putative drug-disease candidates in our work, validation by experts would be the routine next step.

Outlook With this work we explored the capabilities of using KGs for the discovery of possible new drug-disease relations. It is apparent that ReDrugS KB link prediction needs more in-depth work until it can be applied practically, but there is a substantial amount of information which can be inferred and used in the medical field.

Acknowledgment This work was conducted as part of the Knowledge Graphs Lab offered by the RWTH Aachen University Informatik 5 department in collaboration with OSTHUS. We thank OSTHUS for providing student travel grants.

References

1. Cochez, M., Ristoski, P., et al.: Biased graph walks for RDF graph embeddings. In: Proceedings of the 7th International Conference on Web Intelligence, Mining and Semantics. pp. 21:1–21:12. WIMS '17, ACM, New York, NY, USA (2017)
2. McCusker, J.P., Dumontier, M., Yan, R., He, S., Dordick, J.S., McGuinness, D.L.: Finding melanoma drugs through a probabilistic knowledge graph. PeerJ Computer Science **3**, e106 (2016)