STRATEGIES TO CONNECT RDF GRAPHS FOR LINK PREDICTION USING DRUG-DISEASE KNOWLEDGE GRAPHS

BACKGROUND

- ReDrugS Knowledge Base by McCusker et al. [2]
  - 8 million named Knowledge Graphs containing information on drugs and diseases
  - 6180 drugs
  - 3820 diseases
  - 69279 proteins
  - 899198 interactions
- Information may be contradictory or missing

GOALS

- Predict new links between entities in the KB
  - Use graph embeddings and machine learning
- Long term: find new possible treatments for diseases based on existing drugs

METHOD

1. Merge individiual graphs together into a single KG (context sensitive vs context insensitive merging)
2. Perform graph embeddings using RDF2Vec [1]
3. Apply machine learning to perform link prediction
   - One binary classifier per link type
   - Classifiers: Gradient Boosting Classifier (GB), SVM, Naive Bayes Classifier (NB)

RESULTS

- Accuracy high (≥ 94%) but not representative due to skewed data, precision and recall ≤ 50%
- NB has highest recall
- Context insensitive merge performs better on *rdf:type*, *prov:wasQuotedFrom* and *rdfs:subClassOf* but worse on the other predicates e.g. *sio:has-component-part*, *sio:has-participant*, *sio:has-agent*, *sio:is-located-in*
- Embedding visualisations:

OUTLOOK

- Explore hybrid context sensitive and insensitive merging based on link type
- Examine larger embeddings
- Include more link types in classification

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)

Sophie Hallstedt (sophie.hallstedt@rwth-aachen.de), Nikita Makarov (nikita.makarov@rwth-aachen.de), Hossein Samieadel (hossein.semieadel@rwth-aachen.de), Maria Pellegrino (mariaangelapellegrino94@gmail.com), Martina Garofalo (margar1994@gmail.com), Michael Cochez (michael.cochez@fit.frauenhofer.de)

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.

RWTH AACHEN UNIVERSITY