Using Knowledge Graphs to train Graph Neural Networks to find Data Similarity

National Center for Supercomputing Applications

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

1. Department of Crop Sciences, College of Agricultural, Environmental, and Consumer Economics, University of Illinois Urbana-Champaign
2. Department of Computer Science, Grainger College of Engineering, University of Illinois Urbana-Champaign
3. National Center for Supercomputing Applications, University of Illinois Urbana-Champaign
4. Center for Digital Agriculture, University of Illinois Urbana-Champaign
5. To whom correspondence should be addressed: ese@illinois.edu

Aniruddha Lappathi [1,2,3], Dr. Elhan Ersoz [1,3,4,5]

INTRODUCTION

Knowledge bases have gained significant attention due to the surge in popularity of large language models. Central to the creation of high-quality knowledge bases is the need for assessing data topics accurately. In our research, we hypothesized the use of knowledge graphs, to capture the relationships between keywords, provides higher context to data. We hope this approach not only allows for precise evaluation of data similarity but also contributes to the construction of more effective knowledge bases. By leveraging knowledge graphs and SimGNN, a fast graph similarity computation, we can enhance the quality and of knowledge bases, ultimately leading to improved knowledge representation and retrieval in various domains.

OBJECTIVES

Knowledge Graph Creation

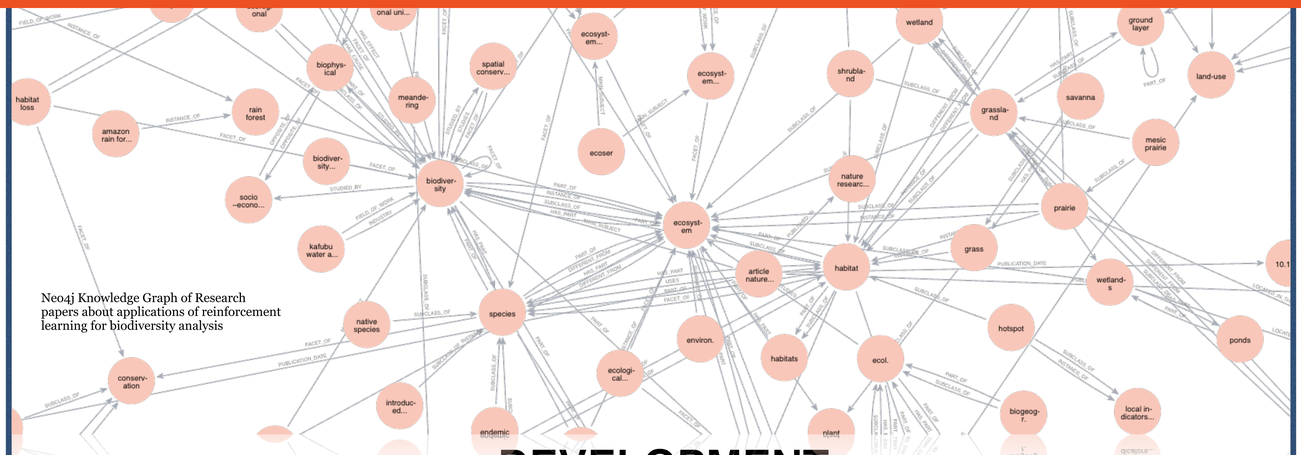
Create a data pipeline that can convert text data into a knowledge graph efficiently that can be used as training and testing data for a GNN.

Model Development

Develop an accurate SimGNN model that can classify data based on its keywords and their contexts.

Knowledge Base Deployment

Produce a topical knowledge base using the data that was above a given threshold after being passed through the SimGNN model.



Neo4j Knowledge Graph of Research papers about applications of reinforcement learning for biodiversity analysis

DEVELOPMENT

Data Collection

Download PDF files of research papers within a chosen topic (Selected "applications of reinforcement learning for biodiversity analysis" for this research).

Data Clean Up

Input PDF files for the text to be cleaned up before tokenizing it into chunks of text data for later processing.

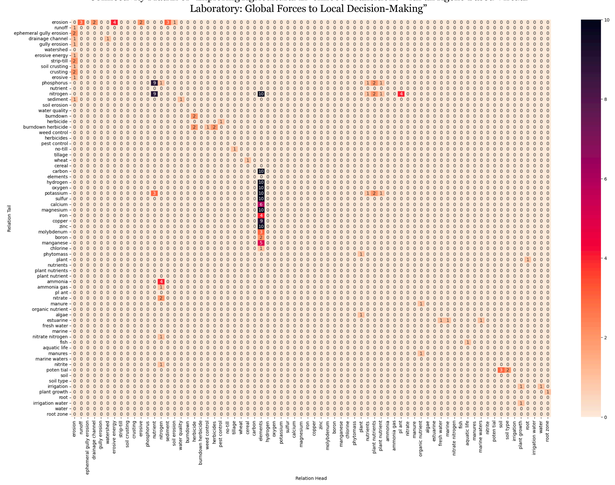
Knowledge Graph Creation

Extract relations using Babelscape's Relation Extraction By End-to-end Language generation model (REBEL). Convert relations to connectivity matrices.

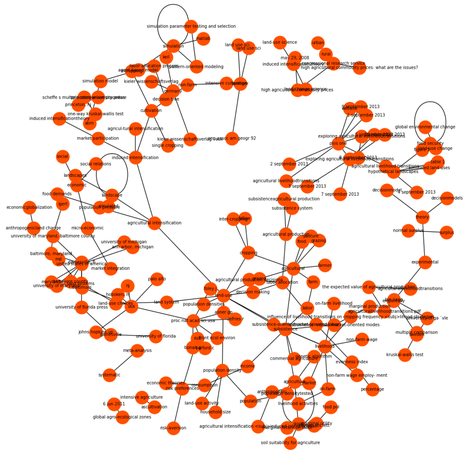
Convert to SimGNN Training Data

SimGNN requires testing data to be a JSON containing the adjacency matrix of two matrices and the graph edit distance (GED). Rather than using the conventional GED, we favored using the cosine similarity value between the matrices. This gives more importance to bigger differences in keyword frequencies.

Connectivity Matrix of "Exploring Agricultural Livelihood Transitions with an Agent-Based Virtual Laboratory: Global Forces to Local Decision-Making"



Knowledge Graph of "Exploring Agricultural Livelihood Transitions with an Agent-Based Virtual Laboratory: Global Forces to Local Decision-Making"



RESULTS

We were able to develop a data pipeline that can produce training and testing data for a graph neural network from any PDF data. The pipeline is also parallelized for computation done with NCSA's DELTA supercomputer. The procurement of this knowledge graph data is fundamental to the later classification processes to produce knowledge bases.

Some challenges faced during this process is ensuring all the connectivity matrices contain all other seen keywords. Other issues faced were parallelization and ensuring efficient data processing and analysis with sparse matrices.

FUTURE WORK

We hope to continue developing a pipeline that can generate SimGNN models that can classify similar topics with high accuracy. We also hope to use this to produce knowledge bases that can be later in large language models, academia, and precise data context classification. We hope to optimize the pipeline to compress the sparse matrices. This project is working towards helping create knowledge bases to categorize research papers for improved human learning.

REFERENCES

1. SimGNN: A Neural Network Approach to Fast Graph Similarity Computation. Yunsheng Bai, Hao Ding, Song Bian, Ting Chen, Yizhou Sun, Wei Wang. WSDM, 2019.
2. Huguet Cabot, P.-L., & Navigli, R. (2021). REBEL: Relation extraction by end-to-end language generation. *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2370-2381.
3. Magliocca NR, Brown DG, Ellis EC (2013) Exploring Agricultural Livelihood Transitions with an Agent-Based Virtual Laboratory: Global Forces to Local Decision-Making. PLoS ONE 8(9): e73241. https://doi.org/10.1371/journal.pone.0073241

ACKNOWLEDGEMENTS

Thank you to Dr. Elhan Ersoz and the NCSA for aiding in the research and funding access to the resources for the progression of this project.

anirudl2@illinois.edu