Enhancing Dataset Discovery with Knowledge Graph Link Prediction Techniques

AMS 2024
Jan 28 - Feb 1 2024
seanhughes.srh@gmail.com

NASA/Goddard Earth Sciences Data and Information Services Center (GES DISC)

Sean Hughes[1, 2], Irina Gerasimov[2, 3], Armin Mehrabian[2, 3], Long Pham[2]

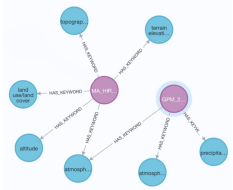
[1]University of Maryland, College Park, MD, USA [2]Code 619, NASA Goddard Space Flight Center, Greenbelt, MD, USA [3]ADNET Systems Inc., Lanham, MD, USA

Abstract

- In the evolving landscape of open science, the ability to navigate and discover pertinent datasets is increasingly significant. This primarily hinges on the presence of detailed metadata, delineating the dataset's content, and potential spheres of application.
- The GES DISC datasets are characterized by science keywords to enable dataset discovery in web search interfaces.
- A problem may arise where a dataset lacks a science keyword that it otherwise should have.
- Machine learning techniques such as link prediction can be used to detect these missing science keywords by estimating the probability of new links forming between dataset and keyword nodes.

Use Case/Link Prediction

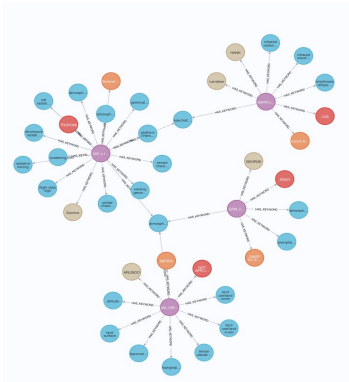
- The GES DISC archive consists of published datasets with metadata containing information about the principal investigator of the dataset, science keywords relating to the dataset, the mission and instruments that were used to create the dataset, and much more.
- This information can be utilized by a machine learning model to predict missing science keywords, as the datasets that utilize the same instruments or missions, have a higher chance of being about similar topics. Similarly, datasets authored by the same person have a higher chance of being about similar topics. These similarities are predicted upon to form new linkages that did not exist previously.



At the projects core, we try to create new dataset-keyword linkages like that shown. Above we have two datasets, with multiple keywords each.

Methodology

- Link prediction algorithms consist of: configuration of a knowledge graph, adjusting training splits, adding node properties/link features, training the link prediction pipeline, and modeling the pipeline.
- The GES DISC dataset metadata is first encoded as a knowledge graph consisting of nodes such as "Dataset", "Keyword", "Instrument", "Investigator", etc. We create this graph using the graph management software Neo4j, portraying relationships as linkages between nodes.
- The knowledge graph is then sliced into three sets: the feature-input set, the train set, and the test set.
- Node properties and link features are added to the graph to give more information for the machine learning model to train on. These properties manifest in graph algorithms such as: FastRP, Degree Centrality, Page Rank, Closeness centrality, etc.
- The machine learning model then begins to train on the data, and the sets created above. It generates the winning model of the training, as well as the training score generated using the AUCPR (area under curve precision recall) metric.
- Finally, the pipeline estimates on the data using the winning model. At this stage, the model generates a list of dataset-keyword pairs, and probabilities that they should have a linkage between them.



Knowledge graph generated using GES DISC dataset metadata, with "Dataset" nodes in purple, "Keyword" nodes in blue, "Investigator" nodes in beige, "Platform" nodes in orange, and "Instrument" nodes in red.

Results

- During the training step, the model generated a winning model that utilized a Random Forest regression.
- The model trained at an average AUCPR (area under curve precision recall) score of 91 percent.
- The model predicts hundreds of new dataset-keyword linkages at around 85-90 percent confidence ratings, which is much higher than the baseline score of around 50 percent for a non machine learning approach.

winningModel	avgTrainScore	outerTrainScore	testScore	validationScores
{maxDepth: 2147483647, criterion: 'GINI', minSplitSize: 2, minLeafSize: 1, numberOfSamplesRatio: 0.07, methodName: 'RandomForest', numberOfDecisionTrees: 100}	0.9304246460698485	0.9323961513568295	0.9162893145893125	[0.9149495645454436]

Winning model description, and average train score of 91 percent.

```
"DATASET_SHORT": "AIRH2RET",
"KEYWORD": "humidity",
"probability": 0.86,
```

Example predicted dataset-keyword linkage. The model predicts with 86 percent confidence that the AIRS/Aqua L2 Standard Physical Retrieval v006 (AIRH2RET) dataset should have the missing keyword "humidity".

Results

- Our work allowed for the creation of a streamlined machine learning model, that enables users to associate science keywords to datasets where linkages failed to appear.
- The software developed to detect missing science keywords helps the GES DISC archive achieve better dataset discovery and research, and will allow users to better utilize the search features of the archive to find similar datasets.
- By creating the machine learning pipeline, the GES DISC archive will have more cohesive metadata structure, and more relevant dataset search results.

Contact

- Sean Hughes: seanhughes.srh@gmail.com
- Irina Gerasimov: irina.v.gerasimov@nasa.gov
- Armin Mehrabian: armin.mehrabian@nasa.gov

- Software developed for this project can be accessed at: https://github.com/seanrhughes/NASA-GES-DISC-link-prediction