Methodology to Build Scalable Knowledge Graphs for Earth Science

Rahul Ramachandran, Manil Maskey, Patrick Gatlin (NASA MSFC)

Jia Zhang, Xiaoyi Duan (CMU)

J.J. Miller, Kaylin Bugbee, Sundar Christopher (UAH)

1. What is a Knowledge Graph?

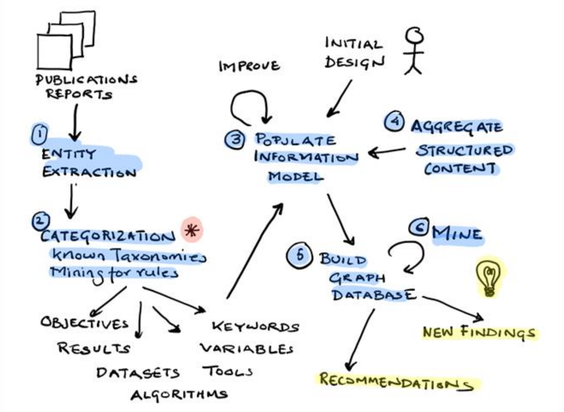
- Method developed by Google in 2012 to enhance the results of its search engine by semantically linking information from a wide variety of sources
- Links key entities in a specific domain with other entities via relationships
- Can be queried to obtain probabilistic recommendations and to infer new knowledge

2. Why build a Knowledge Graph?

- **Need**: Utilize the untapped knowledge resource for the Earth Science domain that is stored in papers and technical reports (unstructured data)
- **Challenge**: Difficult to extract and to infer knowledge at scale

3. Project Objectives

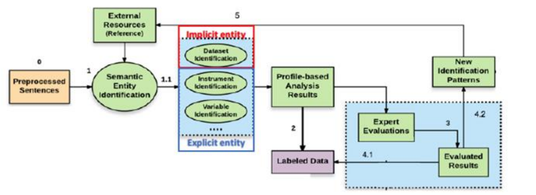
- Develop a methodology to extract important semantic entities from papers/reports
- Develop a software framework to implement the methodology to scale up
- Analyze results for new findings



4. Semantic Entity Extraction Framework

Key Steps:

1. Design and evaluate different heuristic algorithms for Semantic Entity Identification
2. Use heuristic algorithms to assist in creating labeled data
3. Apply Deep Learning algorithms to the labeled data to improve results



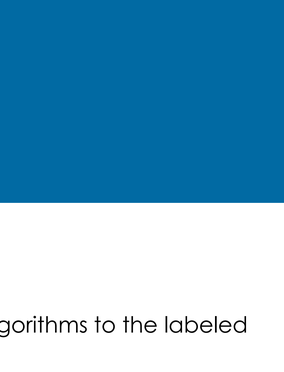
5. Heuristic Algorithms Development Strategy

- Explore the use of existing taxonomies/ control vocabularies (GCMD, CF, SWEET)
- Use a curated set of papers for a specific topic (e.g. - "Airborne Dust Retrieval from Satellites") as a benchmark use case
- Experts manually extract key entities from these papers
- Evaluate extraction results

5.1 Example Algorithm: GCMD Variable Extraction

- Match variable name; variables can appear multiple times in a collection
- Find the most related context:

0.7*topic_count + 0.3*term_count



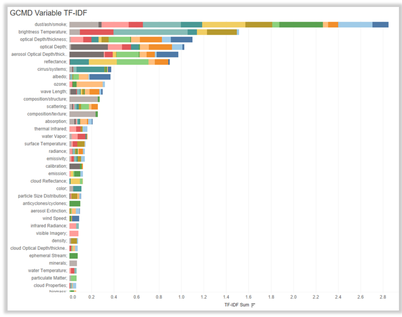
5.2 Extraction Results: GCMD

Good:

- TF/IDF better than total counts
  - Brightness temp ranked higher than in total counts result
- Errors uncovered in paper: "Dust has a higher albedo at 12 microns instead of 11"
  - Should be temperature, not albedo

Issues:

- GCMD does not differentiate between entity types: physical property, phenomena, etc.
- Emissivity and radiance are important properties but are ranked low
- Dust/ash/smoke gives big picture perspective but not very useful for analysis



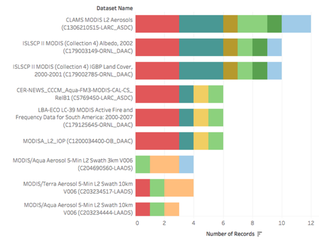
5.3 Example Algorithm: Dataset Extraction

Datasets are rarely mentioned verbatim in publications.
Approach:
- Identify potential sections in the paper describing the dataset
- Extract entities
- Construct profile of the entities in order to query CMR

5.4 Extraction Results: Datasets

Good:

- Most of the datasets are dust or aerosol related
- Extraction identifies all MODIS datasets (MODIS data is key for detecting dust events)



Issues:

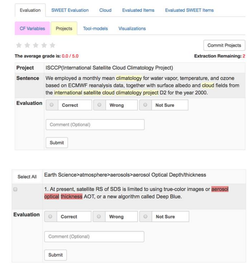
- Some datasets aren't relevant for dust studies
- Slight differences in the API query can provide very different results

6. Lessons Learned

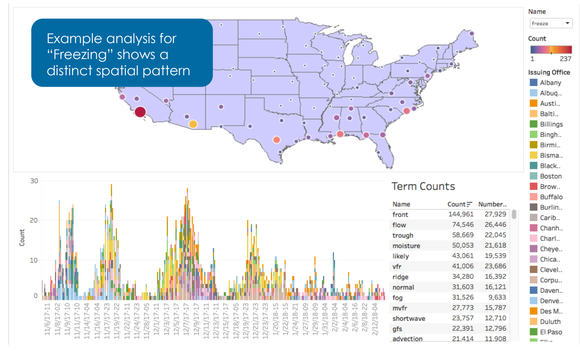
- Semantic entity identification is difficult, and heuristics based algorithms are brittle
- Existing taxonomies are helpful for specific well-defined entities (instruments/platforms); less helpful for others (physical property/phenomena, etc.)
  - Quality of the taxonomy impacts extraction results
  - SWEET covers the most concepts and has the best potential but also has noise
- Dataset profile approach is dependent on both the metadata and the entity extraction quality
  - Metadata authors assign dataset keywords differently from how researchers perceive or use the data

7. Next Steps

- Use algorithms for training set generation
- Have students evaluate extraction results from research papers in their area
- Train Deep Neural Networks for entity extraction



8. Other Applications



- Utilize the Semantic Entity Extraction algorithm to mine NWS forecast discussions to identify trends at different locations
- Analyze spatial/temporal distributions for "terms" of interest

Contact: rahul.ramachandran@nasa.gov

THE UNIVERSITY OF ALABAMA IN HUNTSVILLE