Linking Geoscience Entity Mentions to the Web of Data

Xiaogang (Marshall) Ma (max7@rpi.edu), Jin Guang Zheng (zhengj3@gmail.com)
Tetherless World Constellation, Rensselaer Polytechnic Institute, 110 8th Street, Troy, NY, USA

Background
- Entity recognition and linking among heterogeneous datasets are of great value in data integration and reuse.
- Detecting and linking entity mentions in datasets can be facilitated by using knowledge bases on the Web, such as ontologies and vocabularies.
- The number of ontologies and vocabularies, including those in the field of geosciences, has been continually increasing. Those are valuable contributions to the Web of Data.

Technical Details

**Mention Extraction:**
- Uses publicly available name tagger and regular expressions to extract entity mentions.

**Context Analysis:**
- Sentence level: If terms appear in the same sentence, then we consider they are related to each other.
- Paragraph level: If terms appear in the same paragraph, then we consider they are related to each other.

**Candidate Retrieval:**
- Given a surface form of entity mentions, retrieve all entities with surface form that are similar to the mentions' surface form.
- Surface form is the textual appearance of entities/mentions. The Surface form dictionary is constructed based on the knowledge base.

**Non-collective Ranking:**
- Apply to the candidate entities.
- Assign entities with higher popularity a higher score, similar to PageRank.

**Collective Inference:**
- Analyze the mentions in a context simultaneously to determine the best reference entities.
- Both document graph and graph of candidate entities contain important contextual information about mentions and entities.

Initial Results
Our current demo system uses DBPedia as the knowledge base.

Nature of Efforts

**Challenge:** Existing supervised approaches require a large amount of manually-labeled training data. Such training data are limited in Earth and environmental sciences.

**Approach:** An un-supervised collective inference approach for entity linking.

Future Works

- Enrich the knowledge base: more ontologies and vocabularies in the field of Earth and environmental sciences.
- Semantic parsing: to improve the result of collective inference.
- Semantic reasoning: to improve the quality of linking.
- Propose collaboration with the ESIP portal for ontology and vocabulary registration and knowledge base construction.

Further Reading