Mining and **U**tilizing **D**ataset **R**elevancy from **O**ceanographic **D**ataset (MUDROD) Metadata, Usage Metrics, and User Feedback to Improve Data Discovery and Access

NASA AIST (NNX15AM85G)

Yongyao Jiang, Chaowei (Phil) Yang, Yun Li,
George Mason University

Edward M Armstrong, Thomas Huang, David Moroni,
Chris Finch, Lewis Mcgibbney, JPL, NASA

GeoSemantics Symposium, 2017

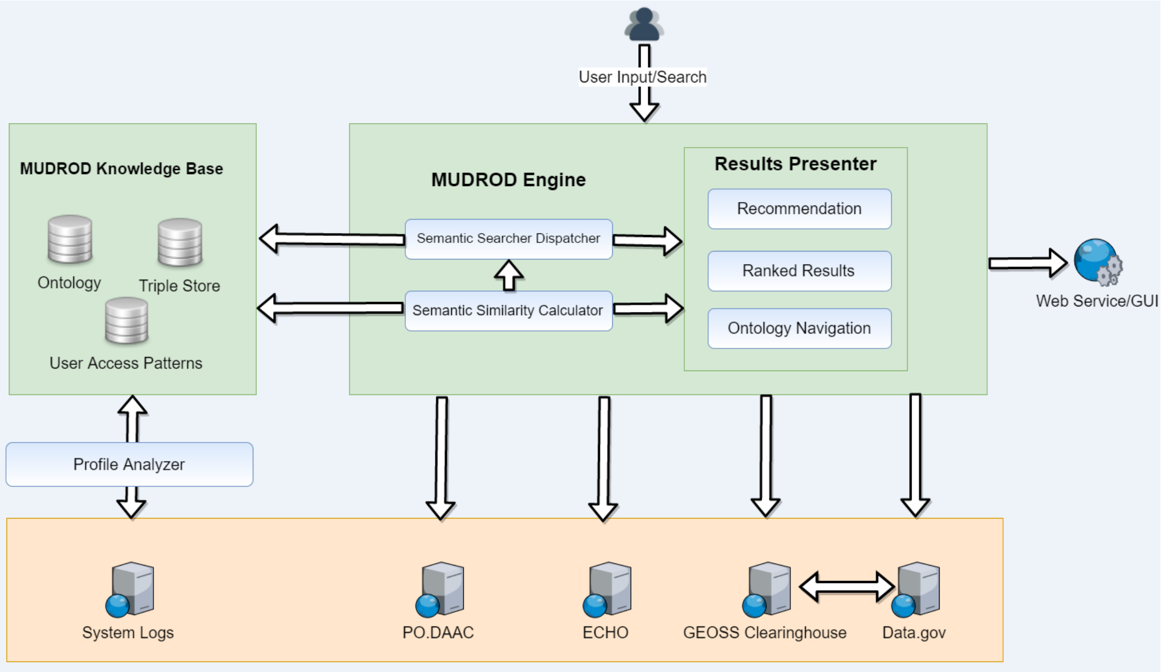
Background

- Traditional search engines (Keyword-based matching)
  - User query: *sea surface temperature*
  - Final query: *sea* AND *surface* AND *temperature*

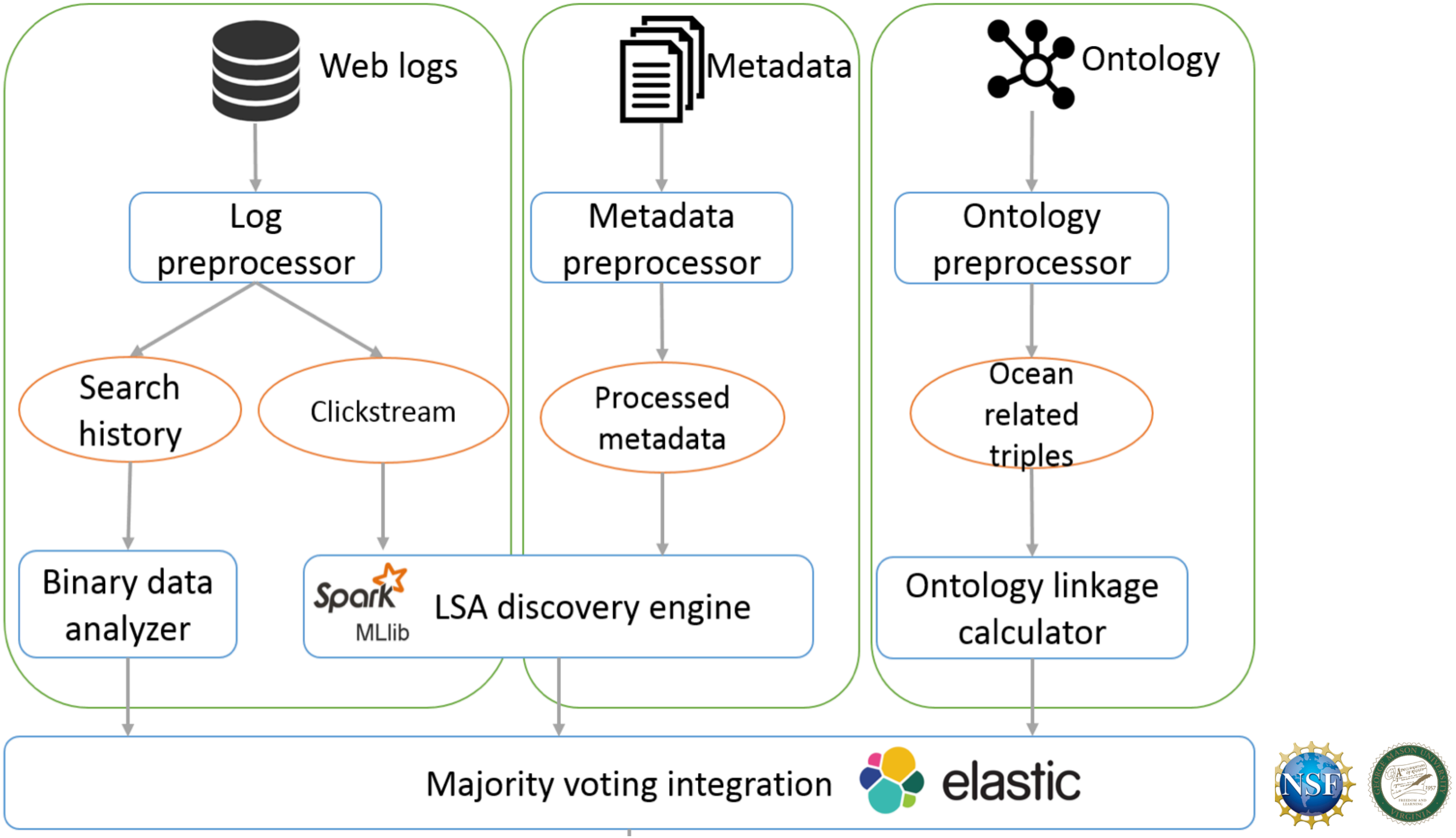
- The real intent of user
  - *“sea surface temperature”* OR *“sst”* OR *“ghrsst”* OR …

Objectives

- Mine web logs to discover user access pattern
- Build a knowledge base by combining user access pattern, existing ontology, and metadata
- Improve data discovery by providing 1) better ranked results; 2) recommendation; 3) ontology navigation



Semantic Research Workflow



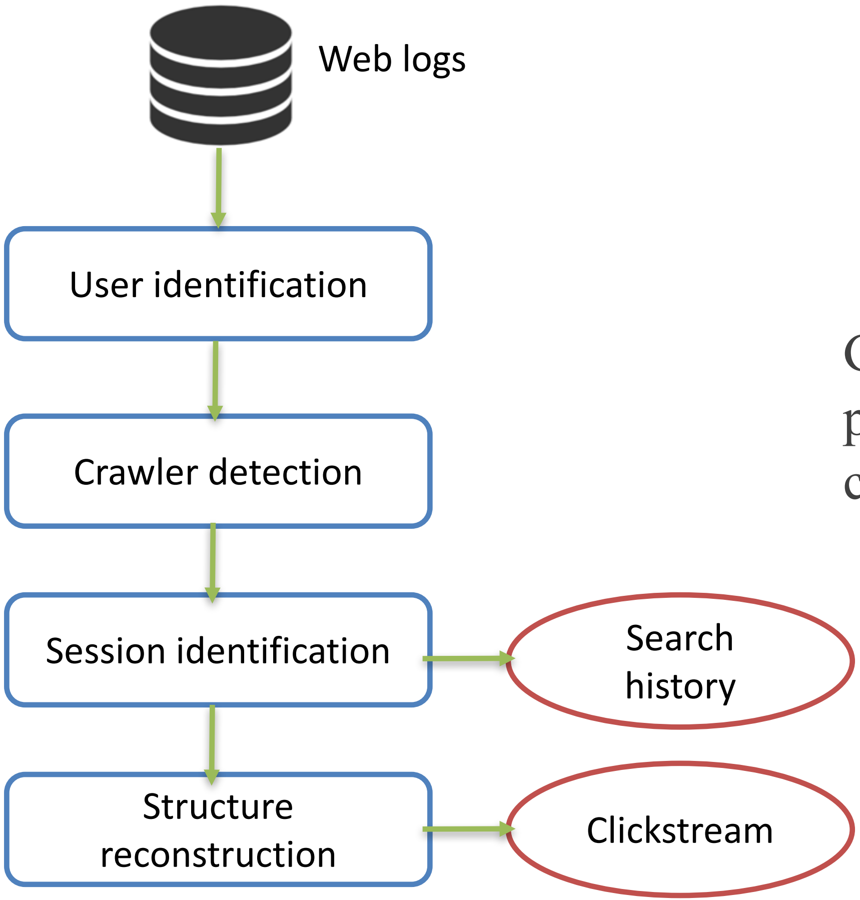
Web logs

- Requests sent from browser, recorded by server
- Log files provided by PO.DAAC (HTTP, FTP)

68.180.228.99 - - [31/Jan/2015:23:59:13 -0800] "GET /datasetlist/... HTTP/1.1" 200 84779 "/ghrsst/" "Mozilla/5.0 ..."

Client IP: 68.180.228.99
Request date/time: [31/Jan/2015:23:59:13 -0800]
Request: " GET /datasetlist/... HTTP/1.1 "
HTTP Code: 200
Bytes returned: 84779
Referrer/previous page: “/ghrsst/"
User agent/browser: "Mozilla/5.0 ...

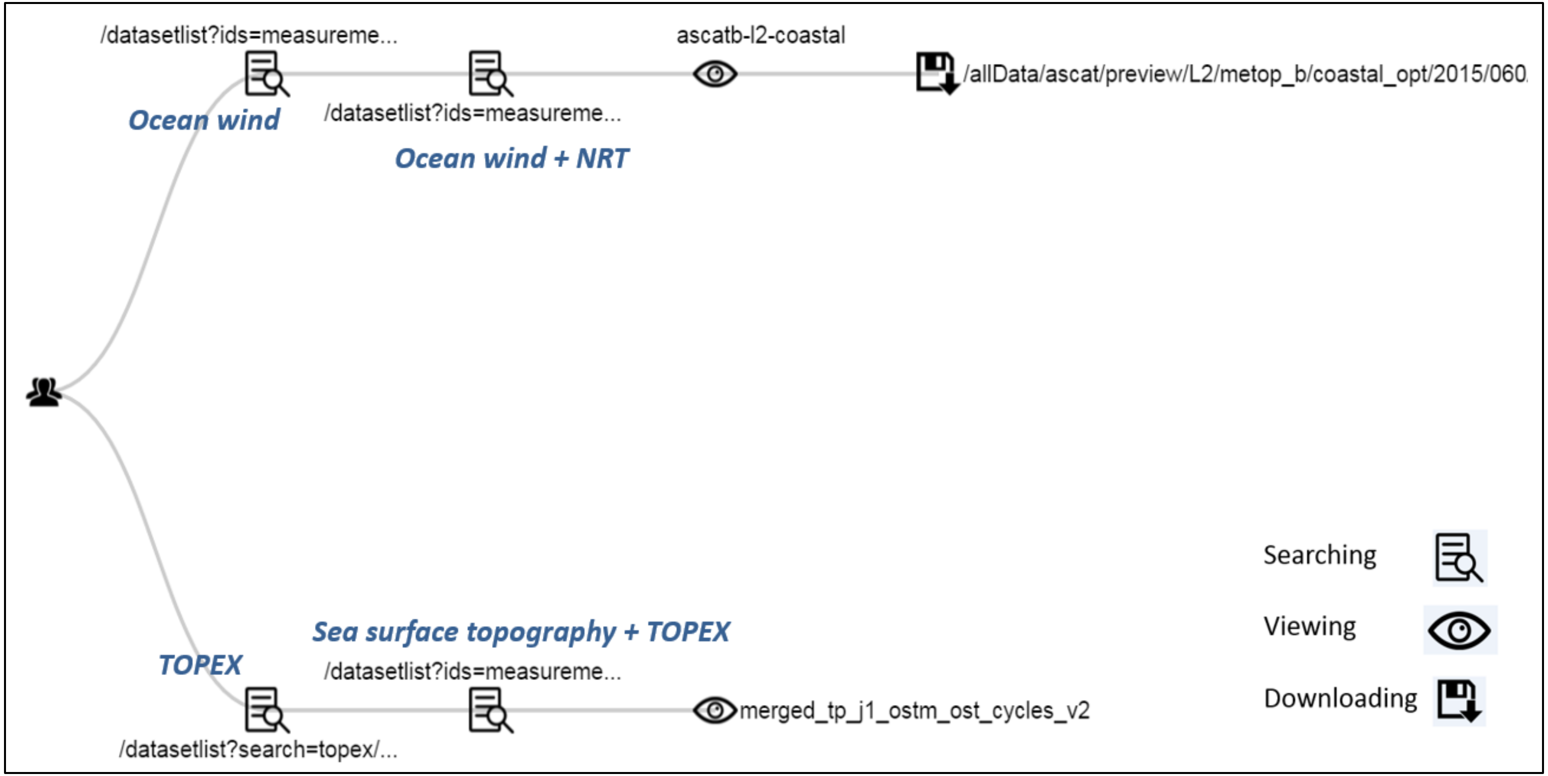
Data preparation



Goal: reconstruct user browsing pattern (search history & clickstream) from a set of raw logs

Additional steps include: word normalization, stop words removal, and stemming

Reconstructed session structure



Data preparation results

1. User search history


```
{
    "User A": [
        "modis",
        "sst",
        "ocean winds",
        "surface wind"
        ...
    ]
}
```


2. Clickstream


```
{
    "Query": "sst"
    "View": "navo-l2p-avhrr19_g"
    "Download": "navo-l2p-avhrr19_g"
}
```

User search history

- Hypothesis: the more frequent two queries co-occur in distinct users' search history, the more similar they are.

User A	User B
ocean temperature,	*quikscat,*
ocean wind,	*sea surface topography,*
sea surface topography,	*sea surface temperature,*
sea surface temperature,	*amsr,*
quikscat,	*oscar,*
cross calibrate multi platform ocean surface wind vector analysis field,	*suomi npp,*
grace,	*altika,*
aquarius project,	*dmsp f17*
saline density	*grace,*
……	*ocean temperature,*
	aquarius,
	……

User search history

- Create query – user matrix
- Calculate binary cosine similarity (collaborative filtering)

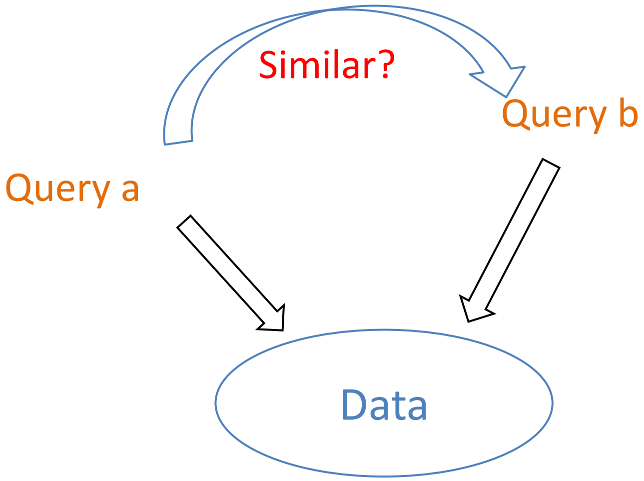
$$sim\ (t,s) = \frac{|t \cap s|}{\sqrt{|t| \cdot |s|}}$$

	$user_1$	$user_2$	$user_3$
ocean temperature	1	1	1
sea surface temperature	1	1	1
ocean wind	0	0	1

Conceptual example

Clickstream

- Hypothesis: similar queries can result in similar clicking behavior
- If two queries are similar, the data that get clicked after they are searched would be more likely to be similar



Clickstream

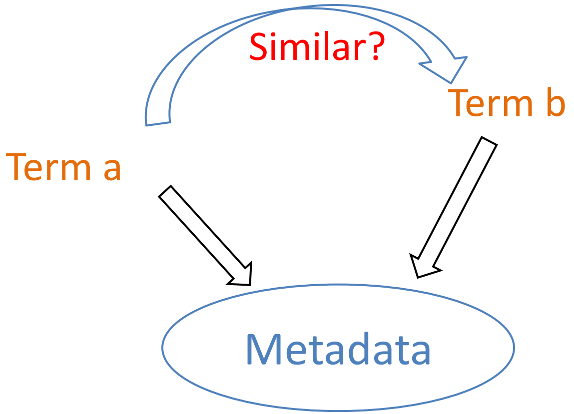
- Create query – data matrix
- Perform Latent Semantic analysis (LSA, feature normalization and reduction)
- Calculate cosine similarity

$$sim(t,s) = \frac{\vec{t} \cdot \vec{s}}{|\vec{t}| * |\vec{s}|}$$

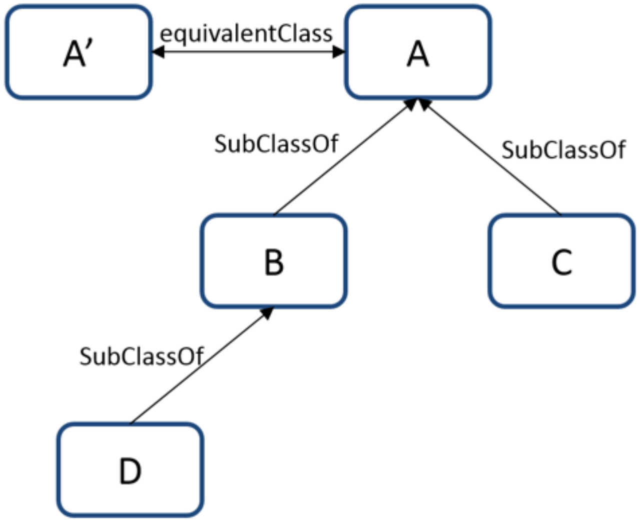
	$data_1$	$data_2$	$data_3$
ocean temperature	2	5	3
sea surface temperature	2	5	4
ocean wind	5	0	0

Metadata

- Hypothesis: semantically related terms tend to appear in the same metadata more frequently
- Essentially the same as the clickstream analysis
- Perform LSA over the *term – metadata* matrix



Existing ontology (SWEET)



- SWEET (Raskin and Pan 2003)
- Focus on only two relations
- The closer, the more similar

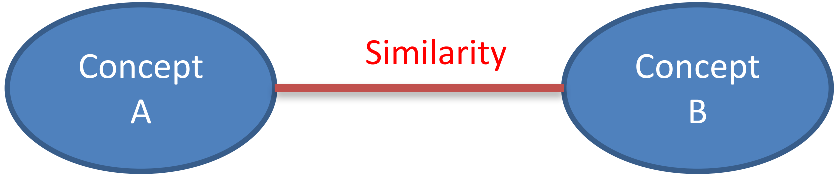
$$sim(X \rightarrow Y) = \frac{e}{Dist(X \rightarrow Y) + e} \quad (9)$$

$$Dist(X \rightarrow Y) = \sum_{i} Edge(Type_i) \quad (10)$$

Where e is a constant used to adjust the final similarity, $Dist(X \rightarrow Y)$ is the distance from X to Y, and $Edge(Type)$ is a function: if the relation type is "SubClassOf", it returns 1; if the relation type is "equivalentClass", it returns 0; if the relation type does not exist, it returns infinity. The resulting value ranges from 0 meaning no relation, to 1 meaning exactly the same.

Integration

- All four results could be converted to



- Problem:
  - None of them are perfect (uncertainty in data, hypothesis and method)
  - Metadata and ontology might have unknown terms to search engine end users
  - Sometimes, similarity values from different methods are inconsistent

Integration

$$sim(X,Y) = \boxed{\max(sim_1, \dots, sim_i)} + \boxed{\frac{(\sum_i w_i - \theta) \cdot \beta}{\theta}} \qquad (8)$$

Where method i is the method that has the linkage of (X,Y), w_i is the weight of method i, sim_i is the similarity of (X,Y) in method i, θ is the threshold that represents the minimum sum of methods weights that makes the linkage a majority, and β is a constant that represents the majority rule change rate.

constant that represents the majority rule change rate.

- The maximum similarity of all of the components (large similarity appears to be more reliable)
- The adjustment increment becomes larger when the similarity exists in more sources

Results and evaluation

Query	Search history	Clickstream	Metadata	SWEET	Integrated list
ocean temperature	sea surface temperature(0.66), sea surface topography(0.56), ocean wind(0.56), aqua(0.49)	sea surface temperature(0.94), sst(0.94), group high resolution sea surface temperature dataset(0.89), ghrsst(0.87)	sst(0.96), ghrsst(0.77), sea surface temperature(0.72), surface temperature(0.63), reynolds(0.58)	None	sst(1.0), sea surface temperature(1.0), ghrsst(1.0), group high resolution sea surface temperature dataset(0.99), reynolds sea surface temperature(0.74)

Sample group	Overall accuracy
Most popular 10 queries	88%
Least popular 10 queries	61%
Randomly selected 10 queries	83%

By domain experts

What can we use it for?

- Query suggestion
- Query modification

```
"bool": {
    {
      "match": {
        "_all": {
          "query": "ocean temperature"
        }
      }
    }
}
```

Standard full-text query

```
"bool": {
  "should": [
    {
      "match": {
        "_all": {
          "query": "ocean temperature",
          "boost": 1
        }
      }
    },
    {
      "match": {
        "_all": {
          "query": "sea surface temperature",
          "boost": 1
        }
      }
    }
  ]
}
```

Semantic boosting query

Conclusion

- Can be used to discover domain-specific semantic relationships
- Can be updated periodically as user behavior changes
- Existing ontology is a just supplement, not a requirement
- It is an automatic approach, and involves certain level of noise
- The result is reasonable and more importantly saves huge amount of time from manually creating ontology

Resources

- Part of the MUDROD project

- Related papers
- Jiang, Y., Y. Li, C. Yang, E. M. Armstrong, T. Huang & D. Moroni (2016) Reconstructing Sessions from Data Discovery and Access Logs to Build a Semantic Knowledge Base for Improving Data Discovery. ISPRS International Journal of Geo-Information, 5, 54.
- Y. Li, Jiang, Y., C. Yang, K. Liu, E. M. Armstrong, T. Huang & D. Moroni (2016) Leverage cloud computing to improve data access log mining. IEEE Oceans 2016.
- Jiang, Y., Y. Li, C. Yang, K. Liu, E. M. Armstrong, T. Huang & D. Moroni (2016) A Comprehensive Approach to Determining the Linkage Weights among Geospatial Vocabularies - An Example with Oceanographic Data Discovery. International Journal of Geographical Information Science (under review)
- Jiang, Y., Y. Li, C. Yang, K. Liu, E. M. Armstrong, T. Huang, D. Moroni & L. Mcgibbney (2016) Towards intelligent geospatial discovery: a machine learning ranking framework. Remote Sensning (under review)

- Available on GitHub: https://github.com/mudrod/mudrod

Demo

- http://199.26.254.164:8080/mudrod-service/