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)

Dr. Chaowei (Phil) Yang, Mr. Yongyao Jiang, Ms. Yun Li,
Geography and GeoInformation Science, George Mason University

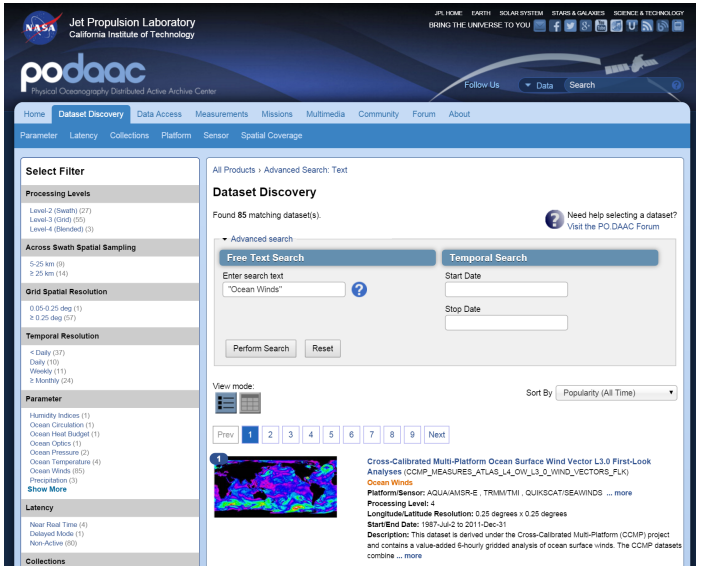
Mr. Edward M Armstrong, Mr. Thomas Huang, Mr. David Moroni,
Mr. Chris Finch, Dr. Lewis J. McGibbney, Mr. Frank Greguska, Mr. Gary Chen
Jet Propulsion Laboratory, NASA

Agenda

- Project Background
  - Problems
  - Objectives
  - Functions
- System
  - Log mining
  - Query Semantics
  - Ranking
  - Recommendation
- Results
- Next step

Data Discovery Problems

- Keyword-based matching (traditional search engines)
  - User query: **ocean wind**
  - Final query: ocean AND wind



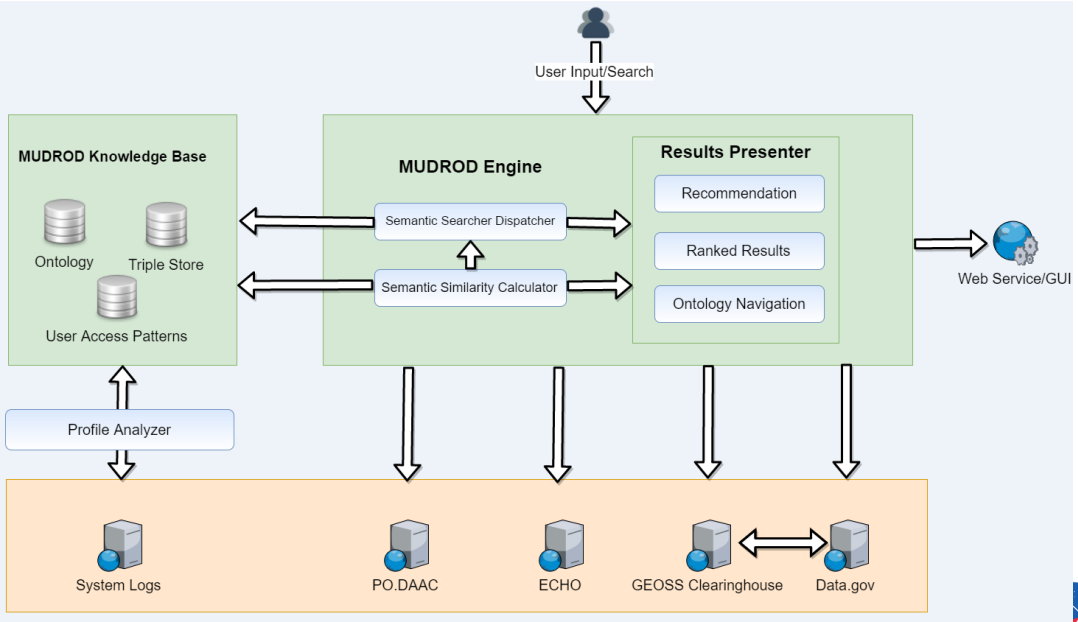
- Reveal the real intent of user query
  - ocean wind = “ocean wind” OR “greco” OR “surface wind” OR “mackerel breeze” …

- PO.DAAC UWG Recommendation 2014-07

- NASA ESDSWG Search Relevance Recommendations 2016 & 2017

Objectives

- Analyze web logs to discover user knowledge (query and data relationships)
- Construct knowledge base by combining semantics and profile analyzer
- Improve data discovery by 1) better ranking; 2) recommendation; 3) ontology navigation



Functions/Modules

- Web log preprocessing
- Semantic analysis of user queries & Navigation
- Machine learning based search ranking
- Data Recommendation

Web log processing

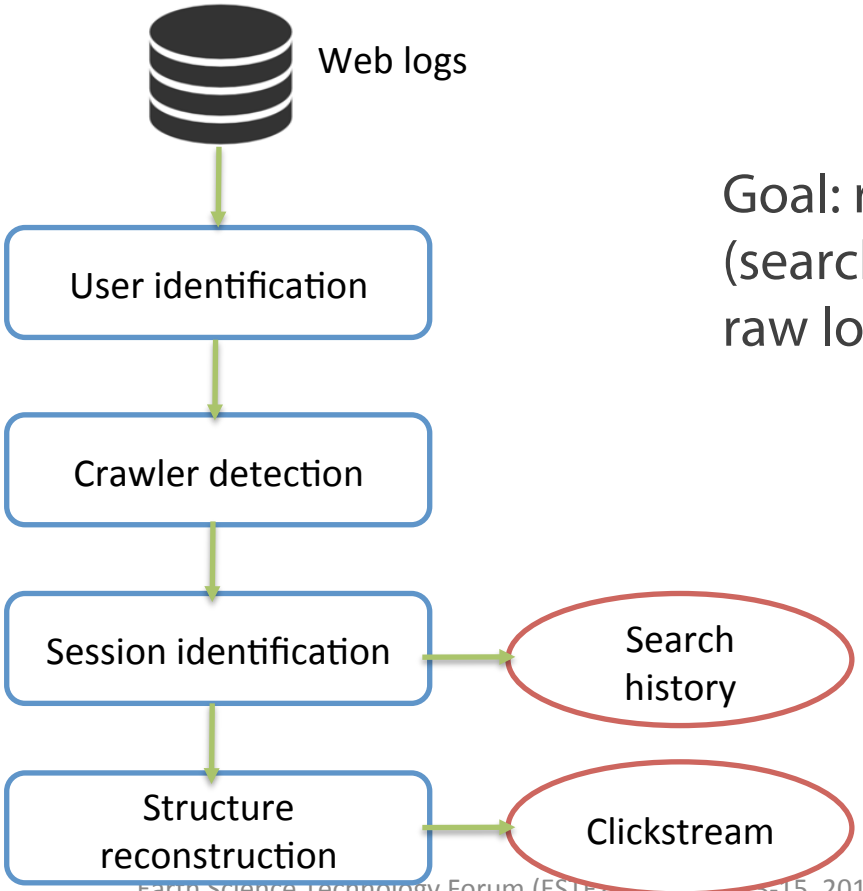
Web logs

- Requests sent from client e.g. browser, cmd line tool, etc. recorded by server
- Log files provided by PO.DAAC (HTTP(S), 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 ...

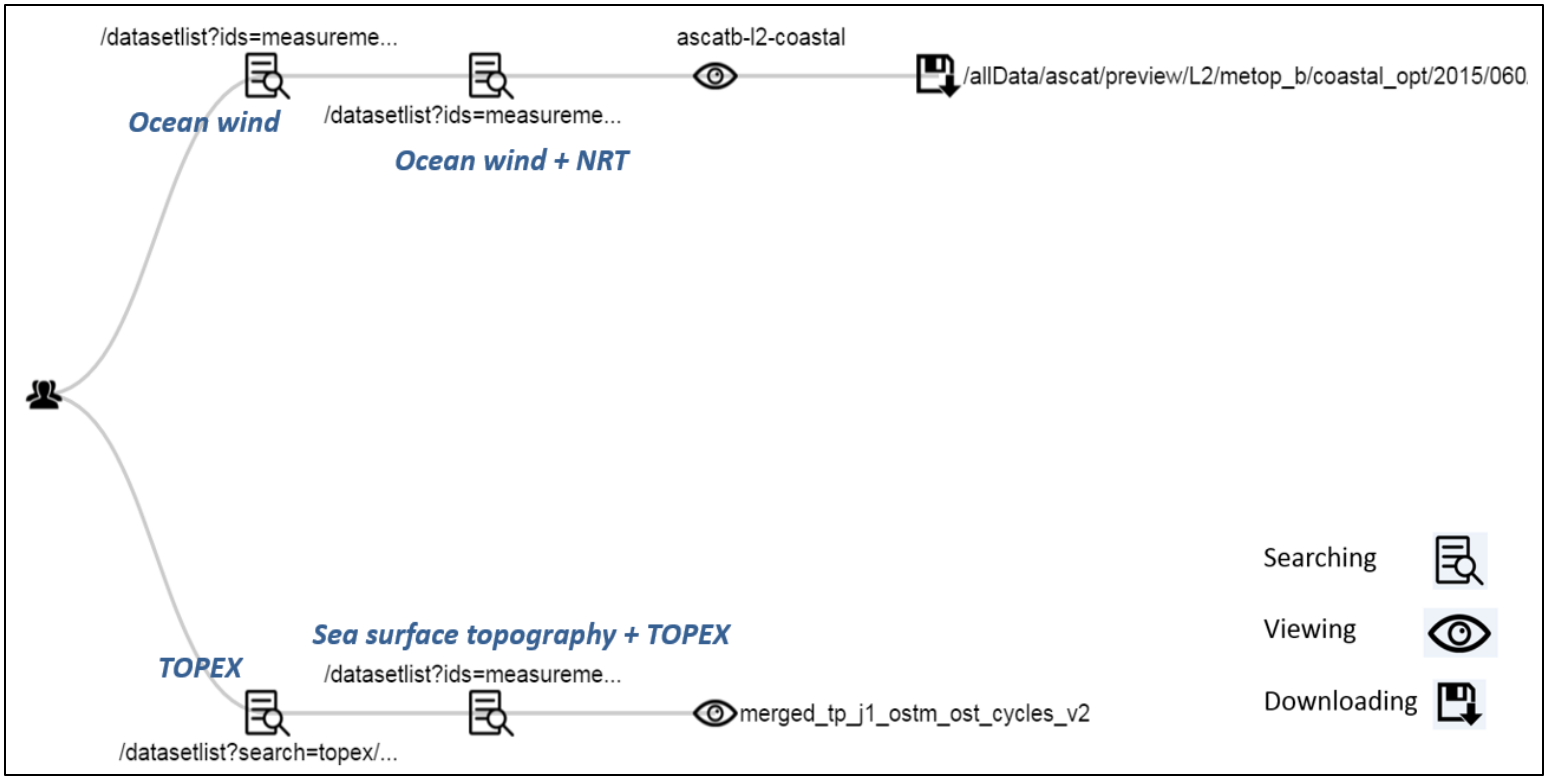
Data preprocess



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 preprocess results

1. User search history

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

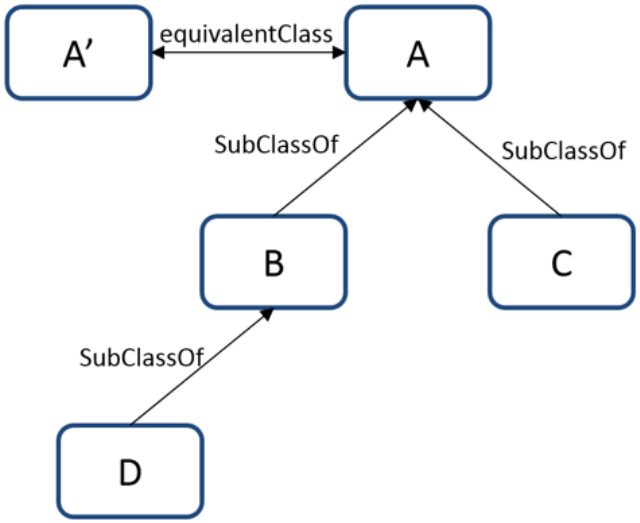
2. Clickstream

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

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.

Semantic similarity

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.

User search history

- Create query – user matrix
- Calculate binary cosine similarity

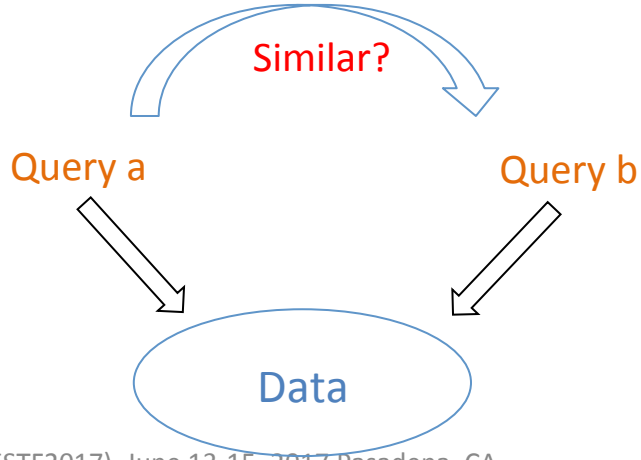
$$sim\,(t,s) = \frac{|t \cap s|}{\sqrt{|t| \cdot |s|}}$$

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

Conceptual example

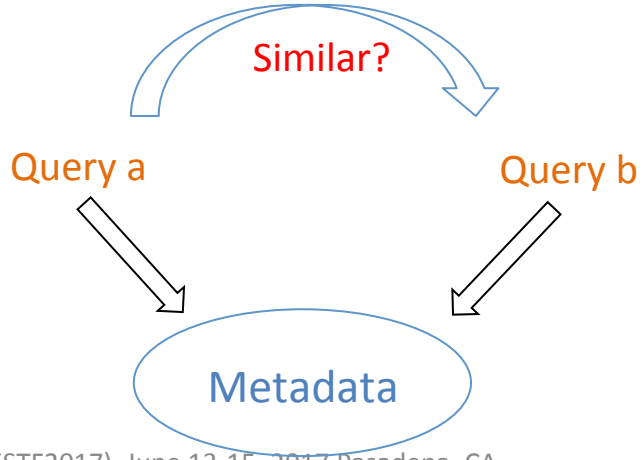
Clickstream

- 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



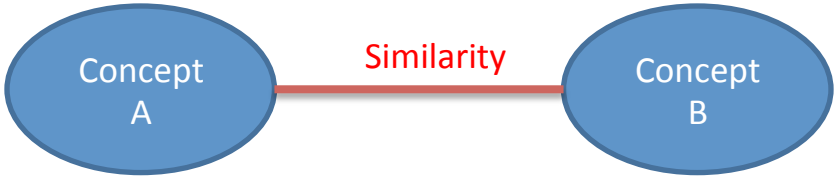
Metadata

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



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

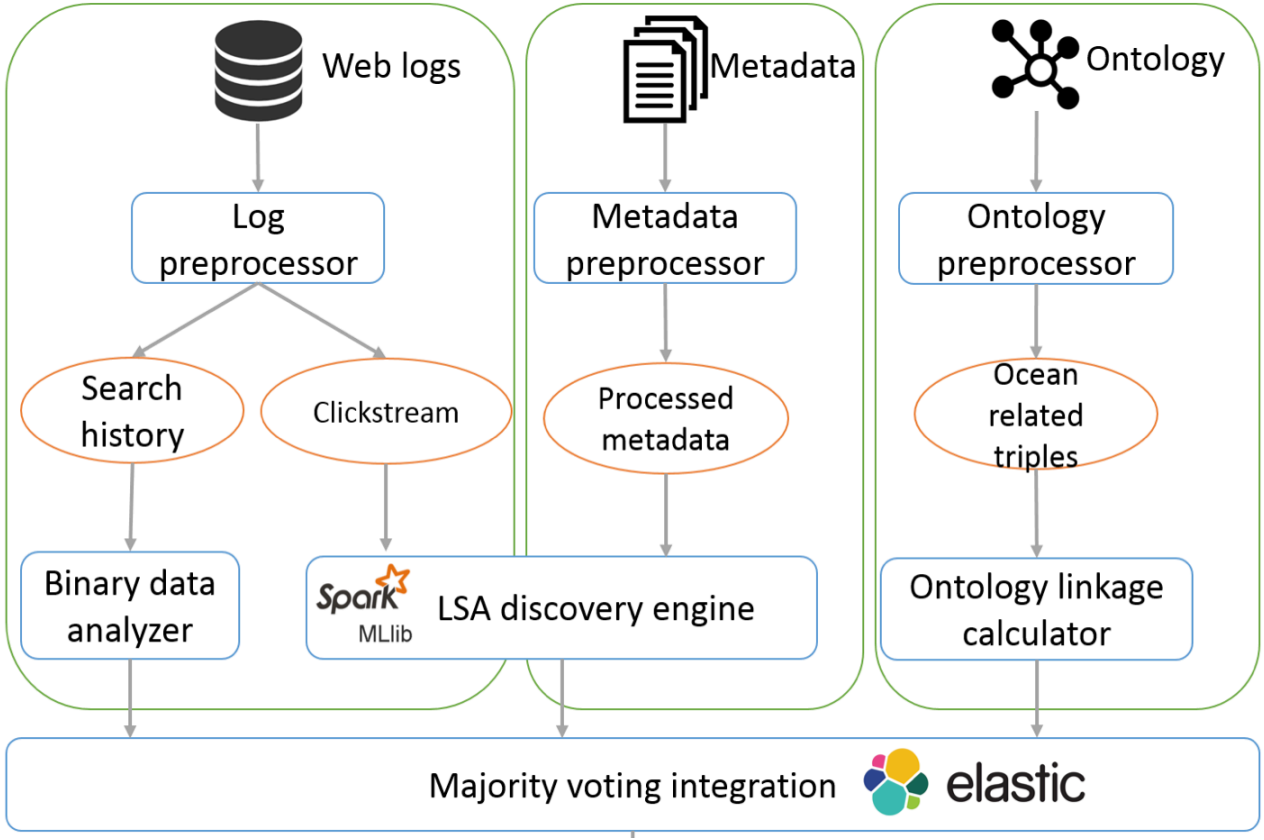
Integration

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

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.

- 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

Semantic Similarity Calculation Workflow



Results and evaluation

By domain experts

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)

Jiang, Y., Y. Li, C. Yang, K. Liu, E. M. Armstrong, T. Huang & D. Moroni (2017) A Comprehensive Approach to Determining the Linkage Weights among Geospatial Vocabularies - An Example with Oceanographic Data Discovery. International Journal of Geographical Information Science (minor revision)

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

What can we use it for?

- Query suggestion
- Query modification

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

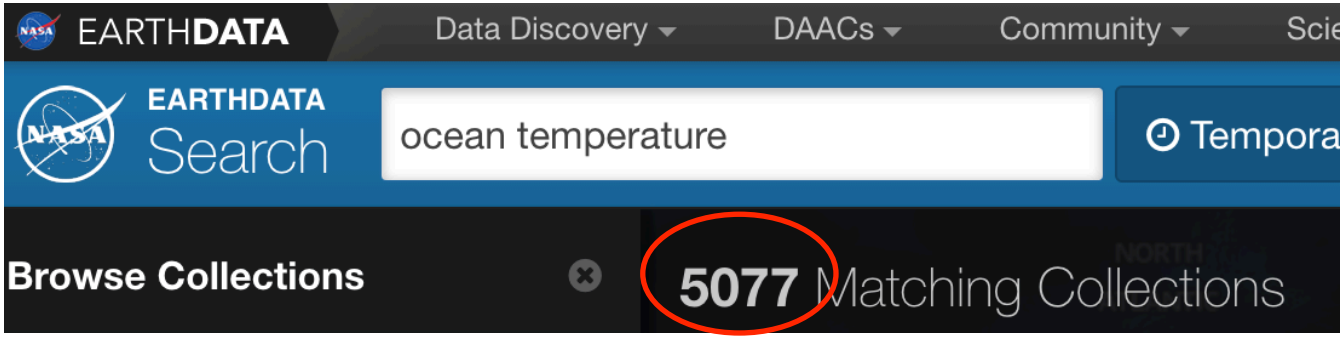
Standard full-text query

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

Semantic boosting query

Search ranking

Background



- Ranking is a long-standing problem in geospatial data discovery
- Typically, hundreds, even thousands of matches
- Can get larger as more Earth observation data is being collected

Objective and Methods

- Put the most desired data to the top of the result list
- What features can represent users' search preferences for geospatial data?
- How can the ranking function reach a balance of all these features?

- Identified eleven features from
  - Geospatial metadata attributes
  - Query – metadata content overlap
  - User behavior from web logs

Ranking features – Metadata attributes

Features	Description
Release date	The date when the data was published
Processing level (PL)	The processing level of image products, ranging from level 0 to level 4.
Version number	The publish version of the data
Spatial resolution	The spatial resolution of the data
Temporal resolution	The temporal resolution of the data

- Five metadata features
- Verified by domains experts
- Query-independent: static, depends on the data itself, won't change with the query

Spatial query-metadata overlap

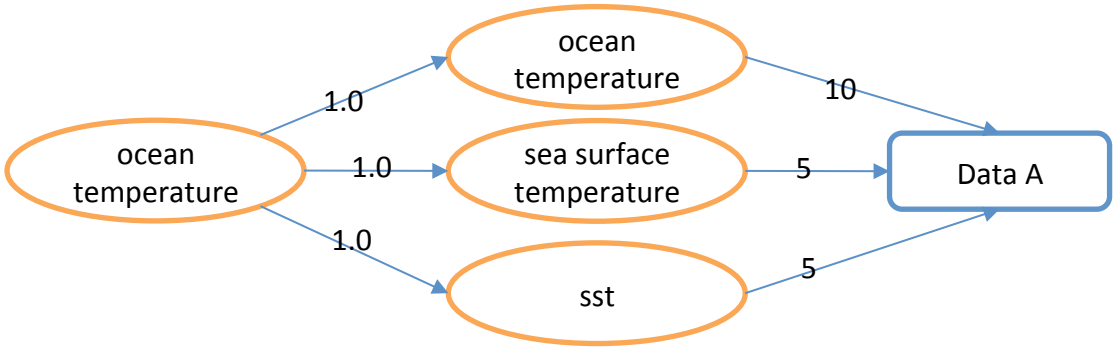
- Spatial similarity between query area and the coverage of a particular data

$$Sim(q,d) = \left(\frac{area(q \cap d)}{area(q)} + \frac{area(q \cap d)}{area(d)}\right) * 0.5$$

- Overlap area normalized by the original area of query and data

Ranking features − User behavior

- All-time, monthly, user popularity, and semantic popularity (retrieved from web logs)
- Semantic popularity: the number of times that the data has been clicked after searching a particular query *and its highly related ones* (query-dependent)

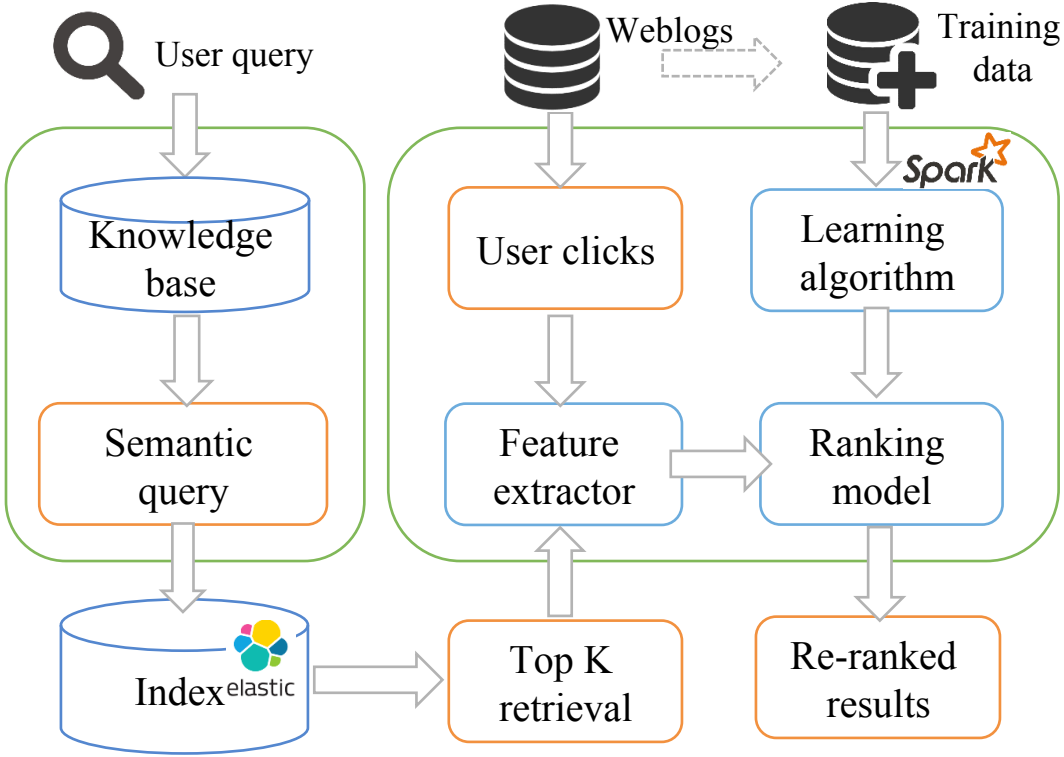


$$\sum_q sim_q \cdot clicks(q, d)$$

RankSVM

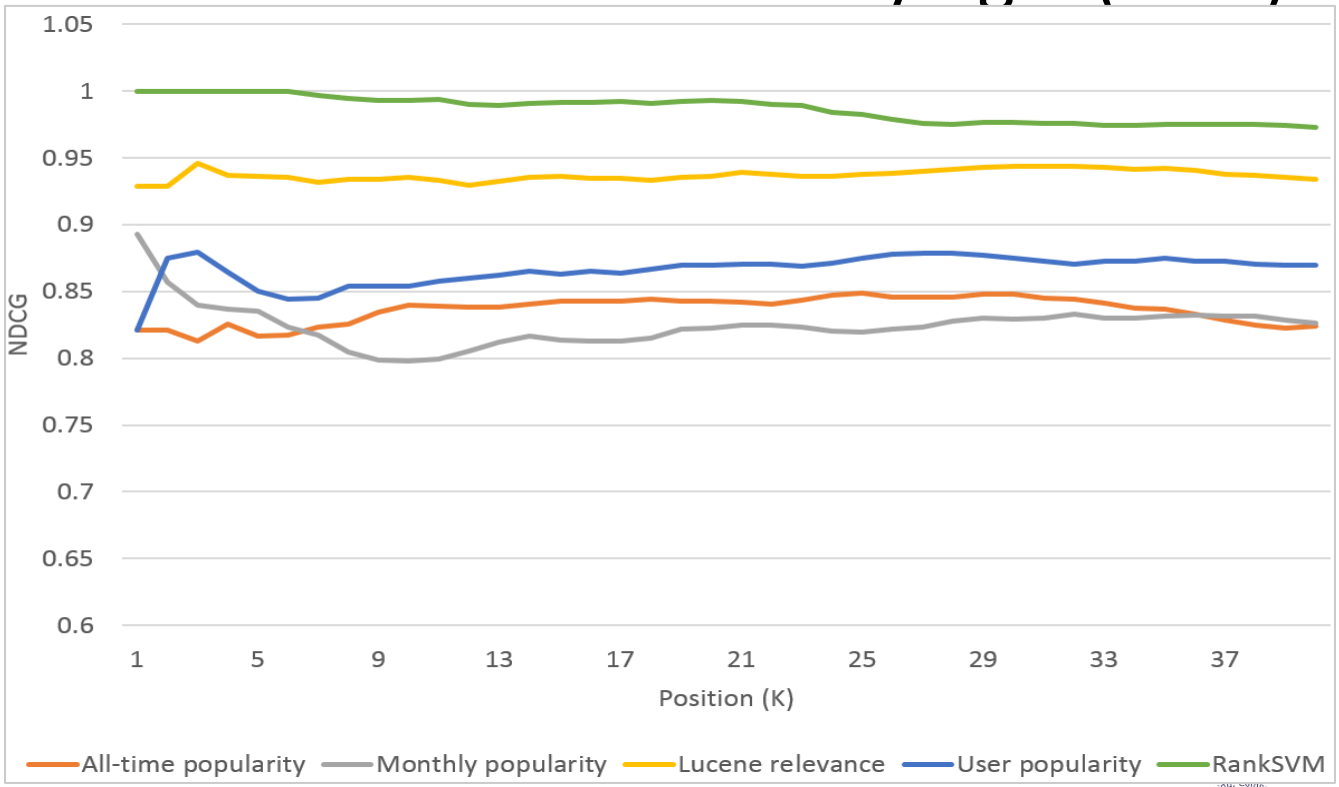
- One of the well-recognized ML ranking algorithm
- Convert a ranking problem into a classification problem that a regular SVM algorithm can solve
- 3 main steps
  - 1) Standardize: mean = 0, std = 1
    - SVM is not scale invariant
    - Over-optimized
    - Longer to train
  - 2) For any pair of training data, calculate the difference
  - 3) A ranking problem becomes a binary classification problem, where SVM is applied to find the optimal decision boundary

Architecture



- All of these (except for training) can be finished within 2 seconds
- None of the open source mainstream ML library provide any ranking algorithm
- Implemented it by ourselves with the aid of Spark MLlib

NDCG (K) for five different ranking methods at varying K (1-40)



Jiang, Y., Y. Li, C. Yang, K. Liu, E. M. Armstrong, T. Huang, D. Moroni & L. J. McGibbney (2017) Towards intelligent geospatial discovery: a machine learning ranking framework. International Journal of Digital Earth (minor revision)

Data recommendation

How to recommend geospatial data?

- Use geospatial metadata for content-based recommendation
  - -Metadata spatiotemporal similarity
  - -Metadata attribute similarity
  - -Metadata content similarity

- Leverage user behaviors data for CF recommendation
  - -Session-based co-occurrence of data

Geographic metadata

Attribute type	Attribute name	Attribute description
Spatiotemporal attributes	DatasetCoverage-EastLon	The East longitude of the bounding rectangle
	DatasetCoverage-WestLon	The West longitude of the bounding rectangle
	DatasetCoverage-NorthLat	The North latitude of the bounding rectangle
	DatasetCoverage-SouthLat	The South latitude of the bounding rectangle
	DatasetCoverage-StartTimeLong	The start time of the data
	DatasetCoverage-StopTimeLong	The end time of the data
Categorical geographic attributes	DatasetRegion-Region	Region of dataset. Such as global, Atlantic
	Dataset-ProjectionType	Project type like cylindrical lat-lon
	Dataset-ProcessingLevel	Data processing level
	DatasetPolicy-DataFormat	Data format e.g. HDF, NetCDF
	DatasetSource-Sensor-ShortName	Short name of sensor
Ordinal geographic attributes	Dataset-TemporalResolution	Temporal resolution of dataset
	Dataset-TemporalRepeat	Temporal resolution of dataset
	Dataset-SpatialResolution	Spatial resolution of dataset
Descriptive attributes	Dataset-description	Describe the content of the dataset

Spatiotemporal similarity

- Spatial variables: NorthLat, SouthLat, WestLon, EastLon
- Temporal variables: DatasetCoverage-StartTimeLong, StopTimeLong
- Use volume overlap ratio to calculate similarity

$$spatiotemporal{\downarrow}sim(r{\downarrow}i, r{\downarrow}j) = (volume(r{\downarrow}i \cap r{\downarrow}j)/volume(r{\downarrow}i) + volume(r{\downarrow}i \cap r{\downarrow}j)/volume(r{\downarrow}j)) * 0.5$$

$$volume(r) = |eastlon - westlon| * |southlat - northlat| * |endtime - startime|$$

Categorical similarity

- Fixed number of values
- No intrinsic ordering
- sensor-name: "AMSR-E", "MODIS", "AVHRR-3" and "WindSat"

$$categorical\_var\_sim(v_i, v_j) = v_i \cap v_j / v_i \cup v_j$$

Ordinal similarity

Ordinal attribute is similar to categorical attribute but its values has a clear order, e.g. spatial resolution

- Converted into rank from 1 to R
- Nominalize these ranks for similarity calculation

$$norm\_rank(v\downarrow i)=Rankv\downarrow i+1/R+1$$

$$ordinal\downarrow var\downarrow sim(v\downarrow i,v\downarrow j)=1-|norm\downarrow rank(v\downarrow i)-norm\downarrow rank(v\downarrow j)|$$

Descriptive similarity

Step 1: Phrase extraction

1. Extract term candidates from metadata description with POS (part of speech) Tagging
2. Introduce “occurrence” and “strength” to filter out terms from candidates.
   - “occurrence”: occurrences number of terms
   - “strength”: the number of words in a term

Original text

Aquarius Level 3 sea surface salinity (SSS) standard mapped image data contains gridded 1 degree spatial resolution SSS averaged over daily, 7 day, monthly, and seasonal time scales. This particular data set is the seasonal climatology, Ascending sea surface salinity product for version 4.0 of the Aquarius data set, which is the official end of prime mission public data release from the AQUARIUS/SAC-D mission. Only retrieved values for Ascending passes have been used to create this product. The Aquarius instrument is onboard the AQUARIUS/SAC-D satellite, a collaborative effort between NASA and the Argentinian Space Agency Comision Nacional de Actividades Espaciales (CONAE). The instrument consists of three radiometers in push broom alignment at incidence angles of 29, 38, and 46 degrees incidence angles relative to the shadow side of the orbit. Footprints for the beams are: 76 km (along-track) x 94 km (cross-track), 84 km x 120 km and 96km x 156 km, yielding a total cross-track swath of 370 km. The radiometers measure brightness temperature at 1.413 GHz in their respective horizontal and vertical polarizations (TH and TV). A scatterometer operating at 1.26 GHz measures ocean backscatter in each footprint that is used for surface roughness corrections in the estimation of salinity. The scatterometer has an approximate 390km swath.

Extracted terms

Radiometers Measure Brightness Temperature,AQUARIUS/SAC Mission,Image Data,Broom Alignment,Resolution SSS,AQUARIUS/SAC Satellite,Scatterometer,Scatterometer,Aquarius Data,Argentinian Space Agency Comision Nacional,Incidence Angles,Time Scales,Actividades Espaciales,Cross-track Swath,Official End,Aquarius Instrument,Shadow Side,Ascending Sea Surface Salinity Product,Level,Level,Surface Roughness Corrections,Data Release,Salinity,Density, Salinity,Density, AQUARIUS,L3,SSS,SMIA,SEASONAL-CLIMATOLOGY,V4

Metadata abstract semantic similarity

Step 2: Represent metadata in the phrase vector space (The dimension lower than word feature space)

	Term 1	Term 2	Term 3	Term 4	Term 5	Term 6	Term 7	…	Term N
dataser1	1	0	1	0	0	0	0		1
dataser2	0	0	0	1	1	0	0		0
…									
dataset k	1	0	1	0	0	0	0		1

Step 3: Calculate cosine similarity

Session based recommendation

Calculate metadata similarity based on session level co-occurrence

	Session1	Session2	…	Session N
Data 1	1	0		1
Data 2	0	0		0
…				
Data k	1	0		1

$$Similarity\ (i,j) = N(i \frown j) / \sqrt{N(i)} * \sqrt{N(j)}$$

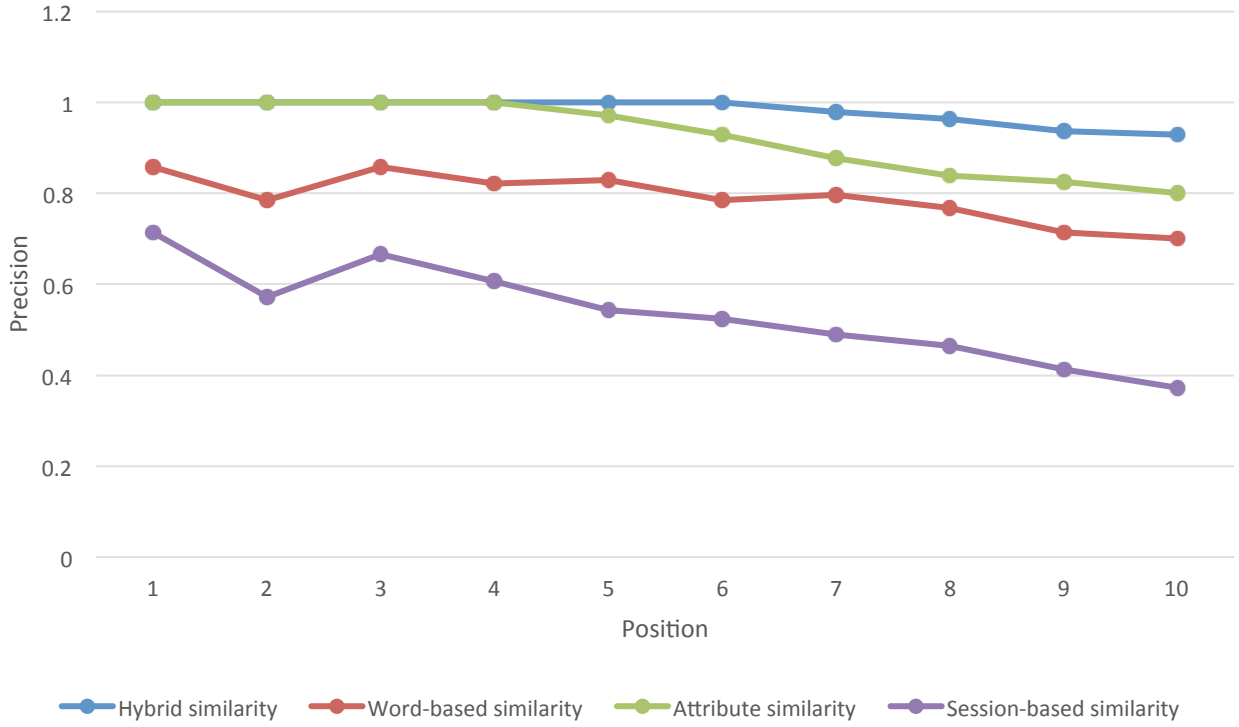
N(i): The number of sessions in which dataset i was viewed or download
N(j): The number of sessions in which dataset j was viewed or download
N($i \frown j$): The number of sessions in which both dataset i and j were viewed or download

Hybrid recommendation

Recommendation method	Pros	Cons	Strategy
Descriptive similarity	1. Natural language processing methods can be adopted to find latent semantic relationship	1. Many datasets has nearly same abstract with few words/values changed 2. It is hard to extract detailed attributes from description	Used as the basic method of recommendation algorithm
Attribute similarity (spatiotemporal, ordinal, categorical)	1. As structured data, geographic metadata have many variables.	1. Variable values may be null or wrong 2. The quality depends on the weight assigned to every variable	As supplement to semantic similarity
Session concurrence	1. Reflect users' preference	1. Cold start problem: Newly published data don't have usage data	Fine-tune recommendation list

$$Recommend\ (i) = Wss * Descriptivec{\downarrow}Similarity(i) + \sum Wcv * Categorial{\downarrow}Similarity(i) + \sum Wov * Ordianl{\downarrow}Similarity(i) + Wstv * SpatioTemporalSimilarity + Wso * SessionSimilarity$$

Quantitative Evaluation



Hybrid similarity outperform other similarities since it integrates metadata attributes and user preference.

Y. Li, Jiang, Y., C. Yang, K. Liu, E. M. Armstrong, T. Huang, D. Moroni & L. J. McGibbney (2017) A Geospatial Data Recommender System based on Metadata and User Behaviour (revision)

Conclusion

- Log mining enables a data portal integrating implicit user preferences
- Word similarity retrieved by data mining tasks expands any given query to improve search recall and precision.
- The rich set of ranking features and the ML algorithm provide substantial advantages over using other ranking methods
- The recommendation algorithm can discover latent data relevancy
- The proposed architecture enables the loosely coupled software structure of a data portal and avoids the cost of replacing the existing system

Products

- Publications

- 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.
- Yang, C., et al., 2017. Big Data and cloud computing: innovation opportunities and challenges. *International Journal of Digital Earth*, *10*(1), pp.13-53. (the **2nd most read paper of IJDE** in it's decadal history)
- Jiang, Y., Y. Li, C. Yang, K. Liu, E. M. Armstrong, T. Huang & D. Moroni (2017) A Comprehensive Approach to Determining the Linkage Weights among Geospatial Vocabularies - An Example with Oceanographic Data Discovery. International Journal of Geographical Information Science (minor revision)
- Jiang, Y., Y. Li, C. Yang, K. Liu, E. M. Armstrong, T. Huang, D. Moroni & L. J. McGibbney (2017) Towards intelligent geospatial discovery: a machine learning ranking framework. International Journal of Digital Earth (minor revision)
- Y. Li, Jiang, Y., C. Yang, K. Liu, E. M. Armstrong, T. Huang, D. Moroni & L. J. McGibbney (2017) A Geospatial Data Recommender System based on Metadata and User Behaviour (revision)
- Jiang, Y., Y. Li, C. Yang, K. Liu, E. M. Armstrong, T. Huang, D. Moroni & L. J. McGibbney (2017) A smart web-based data discovery system for ocean sciences. (ongoing)

- Source code: https://github.com/mudrod/mudrod
- PO.DAAC Labs: http://mudrod.jpl.nasa.gov/
- PD Leverage: http://pd.cloud.gmu.edu/

Component	Current TRL	Project end TRL	Description
Semantic search engine			
Search Dispatcher	7	7	Translating a user search query into a set of new semantic queries
Similarity calculator	7	7	Calculating the semantic similarity from weblogs, metadata, and ontology
Recommendation module	7	7	Recommending similar datasets to the clicked dataset
Ranking module	7	7	Re-ranking the search results based on RankSVM ML algorithm
Knowledge base			
Ontology	7	7	Extensions from SWEET ontology for earth science data
Triple Store	7	7	ESIP ontology repository
Vocabulary linkage discovery engine			
Profile analyzer	7	7	Extracting user browsing pattern from raw web logs
Web services/GUI			
Ranking service/presenter	7	7	Providing and presenting the ranked results
Recommendation service/presenter	7	7	Providing and presenting the related datasets
Ontology navigation service/presenter	7	7	Providing and presenting related searches

Next steps

- Add more features (e.g., temporal similarity)
- Create training data from web logs for RankSVM
- Develop a query understanding module to better interpret user's search intent (e.g. "ocean wind level 3" -> "ocean wind" AND "level 3")
- Support Solr
- Support near real-time data ingestion to dynamically update knowledge base
- Integration with DOMS and OceanXtremes for an ocean science analytics center
- Leverage advanced computing techniques to speed up the process

Presentations

- Yang C., Jiang Y., L Y., Armstrong E., Huang T., and Moroni D., 2015. “Utilizing Advanced IT Technologies to Support MUDROD to Advance Data Discovery and Access”, AGU, San Francisco, CA.
- Yang C., Jiang Y., L Y., Armstrong E., Huang T., and Moroni D., 2016. “Mining and Utilizing Dataset Relevancy from Oceanographic Dataset (MUDROD) Metadata, Usage Metrics, and User Feedback to Improve Data Discovery and Access”, ESIP winter meeting 2016, Washington D.C.
- Jiang Y., Yang C., L Y., Armstrong E., Huang T., and Moroni D., 2016. “A Comprehensive Approach to Determining the Linkage Weights among Geospatial Vocabularies - An Example with Oceanographic Data Discovery”, AAG 2016, San Francisco, CA.
- Yang C., Jiang Y., L Y., Armstrong E., Huang T., and Moroni D., 2016. “Mining and Utilizing Dataset Relevancy from Oceanographic Dataset (MUDROD) Metadata, Usage Metrics, and User Feedback to Improve Data Discovery and Access”, PO.DAAC UWG, Pasadena, CA.
- L Y., Yang C., Jiang Y., Armstrong E., Huang T., and Moroni D., 2016. “Leveraging cloud computing to speedup user access log mining”, Oceans16 MTS IEEE, Monterey, CA.
- Jiang Y., Yang C., L Y., Armstrong E., Huang T., and Moroni D., 2017. “Towards intelligent geospatial discovery: a machine learning ranking framework”, AAG 2017, Boston, MA.
- L Y., Yang C., Jiang Y., Armstrong E., Huang T., and Moroni D., 2017. “A geographic recommender system using metadata and user feedbacks”, AAG 2017, Boston, MA.

Acknowledgements

1. NASA AIST Program (NNX15AM85G)
2. PO.DAAC SWEET Ontology Team (Initially funded by ESTO)
3. Hydrology DAAC Rahul Ramachandran (providing the earlier version of NOESIS)
4. ESDIS for providing testing logs of CMR
5. All team members at JPL and GMU