National Aeronautics and Space Administration
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Improving search relevancy for oceanographic data discovery

Ed Armstrong[1], Chaowei Yang[2], David Moroni[1], Thomas Huang[1], Lewis Mcggibney[1], Frank Greguska[1], Yongyao Jiang[2], Yun Li[2], Christopher Finch[1]

[1]Physical Oceanographic DAAC
NASA Jet Propulsion Laboratory, Pasadena, CA
[2]George Mason University, Fairfax, VA

19th GHRSST Science Team Meeting
Darmstadt, Germany
7 June 2018

Background

- Dataset Ranking is a long-standing problem in geospatial data discovery…data diversity and heterogeneity, user search intent
- Determining best and more relevant dataset is important
  - Saves time and less dataset exploration
  - Improve research results
  - Less need to leverage human resources
  - Improve machine to machine interfaces, ontology performance
- Historically it has been driven by community word of mouth, publication references, and even trial and error

PO.DAAC Free Text Search Capabilities

- Driven by Solr/Lucene index of PO.DAAC metadata
  - Limited search factors: term frequency (pre-defined keyword list), inverse document frequency, and dataset popularity
  - Implements a default “OR” between keywords
  - Suffers from low relevancy search precision
    - E.g., the search will often return good number of datasets (reasonable recall) but a low number of relevant datasets (precision is poor)
      - “OR” syntax often returns unrelated datasets
      - Incomplete indexing. Newer versions, release date, processing levels etc. not considered
      - User popularity (unique users) is an imperfect factor

The MUDROD Project

- **Mining and Utilizing Dataset Relevancy from Oceanographic Data (MUDROD)**
  - 2014 funded NASA Advanced Information Systems Technology (AIST) project
    - Technology Readiness Level development from an approximately Level 4 to Level 6-7
  - Specifically targeted to improve search relevance for earth science data in the PO.DAAC
  - Built on services previously implemented for the hydrology community

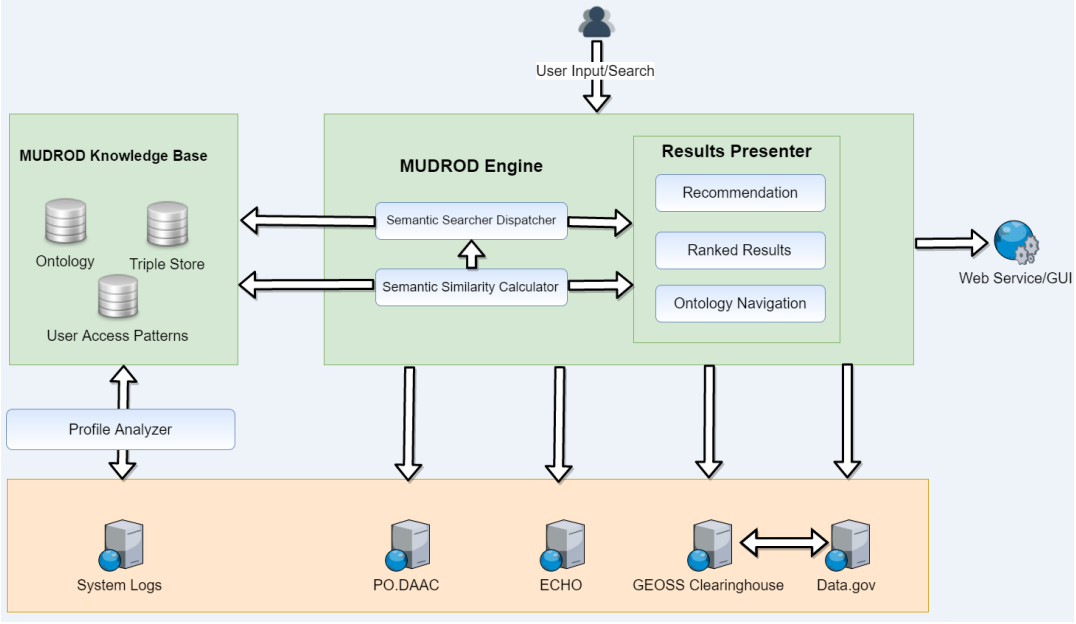
Use Cases for MUDROD Development

- Rank most recent versions of datasets higher
- Rank new mission dataset higher
- Allow user choice of “AND” vs “OR” or phrase keyword syntax
- Improve search across different ocean variables
- Find (and rank) related PO.DAAC datasets
- Prioritize datasets that have been vetted by ”domain experts”
- Consider user search intent, e.g.
  - Climate users vs real time applications users
  - High spatial resolution vs low spatial resolution

MUDROD search relevance methodology and technical objectives

Methodology

- Analyze web/ftp logs and metadata to discover user knowledge (query and data relationships)
- Construct knowledge base by combining semantics and profile analyzer
- Implement Machine Learning on a large training set of factors
- Improve data discovery by 1) better ranking; 2) recommendation; 3) ontology navigation



MUDROD modules

Technology (four technological modules)

- PO.DAAC FTP and web log processing and session construction
- Semantic analysis of user queries & navigation, and metadata records
- Machine learning applied to search ranking training set
- Dataset recommendation engine

Objectives and algorithm factors

- Put the most desired dataset 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 (factors) by considering user behavior, query-text match and examining common geospatial metadata attributes.
  - Geospatial metadata attributes (next slide)
  - Query – metadata content overlap (spatial similarity)
  - User behavior modeling from FTP/web logs (popularity and semantic similarity)

Ranking features – Metadata attributes

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

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

Tying it all together – Machine Learning, the Rank SupportVectorMachine (RankSVM)

- One of the well-recognized Machine Learning ranking algorithms
- Convert a ranking problem into a classification problem that a regular SVM algorithm can solve
  - A classifier trained to predict the ranking order of data pairs
- A ranking problem becomes a binary classification problem, where SVM is applied to find the optimal decision boundary
- Has the best NCDG

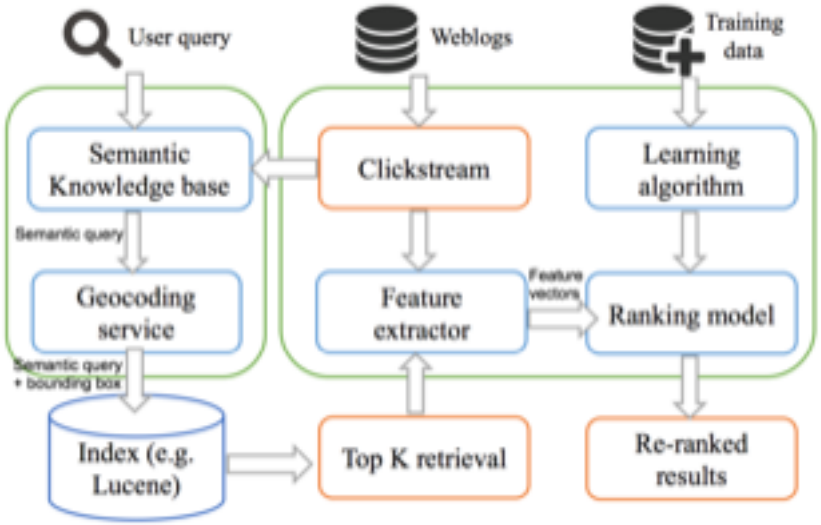


Figure 1. System workflow and architecture

National Aeronautics and Space Administration
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

MUDROD vs PO.DAAC search animation

National Aeronautics and Space Administration
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Comparison of "VIIRS (or) SST" search results

PO.DAAC (Solr)

MUDROD

Not Relevant !! (AVHRR SST datasets)

- MUDROD results:
  - All relevant VIIRS
  - Ordered by version
  - Improved precision

National Aeronautics and
Space Administration
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

VIIRS L2P

VIIRS_NPP-NAVO-L2P-v2.0

Long Name	GHRSST Level 2P 1 m Depth Global Sea Surface Temperature from the Visible Infrared Imaging Radiometer Suite (VIIRS) on the Suomi NPP satellite (GDS version 2)
Landing Page	https://podaac.jpl.nasa.gov/dataset/VIIRS_NPP-NAVO-L2P-v2.0
DOI	10.5067/GHVRS-2PN01
Measurement	Oceans > Ocean Temperature > Sea Surface Temperature > Skin Sea Surface Temperature
Version	2.0
Description	A global Group for High Resolution Sea Surface Temperature (GHRSST) Level 2P dataset based on retrievals from the Visible Infrared Imaging Radiometer Suite (VIIRS). This sensor resides on the Suomi National Polar-orbiting Operational Environmental Satellite System (NPOESS) Preparatory Project (NPP) satellite launched on 28 October 2011. The VIIRS instrument is a a 22-band, multi-spectral scanning radiometer with a 3040 km swath width that builds on the heritage of the MODIS , AVHRR and SeaWiFS sensors for sea surface temperature (SST) and ocean color. For the infrared bands for SST the effective pixel size is 740 meters at nadir and the pixel size variation across the swath is constrained to no more than 1600 meters at the edge of the swath. However, the processing of this dataset aggregates two pixels into one so the resolution is 1500 meters at nadir. This dataset adheres to the GHRSST Data Processing Specification (GDS) version 2 format specifications.
Processing Level	2P
Coverage	Region: Global Northernmost Latitude: 90 degrees Southernmost Latitude: 90 degrees Westernmost Longitude: -180 degrees Easternmost Longitude: 180 degrees Time Span: 2016-02-25 to Present
Spatial Resolution	0.007 degrees (latitude) x 0.007 degrees (longitude)
Temporal Repeat	12 Hour
Sensor	VIIRS
Project	GHRSST
Format	NETCDF
Data Access	http://data.nodc.noaa.gov/cgi-bin/nph-dods/ghrsst/L2P/VIIRS_NPP/NAVO/v2 ftp://ftp.nodc.noaa.gov/pub/data.nodc/ghrsst/GDS2/L2P/VIIRS_NPP/NAVO/v2 https://podaac-opendap.jpl.nasa.gov/opendap/allData/ghrsst/data/GDS2/L2P/VIIRS_NPP/NAVO/v2/ ftp://podaac-ftp.jpl.nasa.gov/allData/ghrsst/data/GDS2/L2P/VIIRS_NPP/NAVO/v2

Related Datasets

- VIIRS_NPP-NAVO-L2P-v1.0
- VIIRS_NPP-OSPO-L2P-v2.4
- VIIRS_NPP-OSPO-L3U-v2.4
- VIIRS_NPP-OSPO-L2P-v2.3
- VIIRS_SST_NPP_NAR-OSISAF-L3C-v1.0
- CMC0.1deg-CMC-L4-GLOB-v3.0
- Geo_Polar_Blended-OSPO-L4-GLOB-v1.0
- DMI_OI-DMI-L4-GLOB-v1.0
- JPL-L2P-MODIS_A
- MODIS_A-JPL-L2P-v2014.0

How does MUDROD find related Datasets?

Related Keyword Searches

- VIIR NPP JPL L2P V2016.0 (0.91)
- VIIR (0.8)
- IASI (0.79)
- SUOMI NPP (0.74)

Summary

- **Dataset heterogeneity and number, and understanding user intent still represent challenges for effective earth data search**
- MUDROD demonstrated tangible improvements in the search precision over the current default PO.DAAC Solr search result
  - Results vetted by oceanographic domain experts
- MUDROD key features:
  - Implemented 11 factors derived from log mining, query analysis and metadata attributes in a Machine Learning algorithm
  - A dataset recommendation algorithm implemented to improve 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
- Deployed at: https://podaac.jpl.nasa.gov/podaac_labs and https://mudrod.jpl.nasa.gov
  - Publications and technical related documentation can be accessed: at https://mudrod.github.io/
- Will be extended in the OceanWorks Project framework
  - Support near real-time data ingestion to dynamically update knowledge base
  - Develop improved query understanding module to better interpret user's search intent (e.g. "ocean wind level 3" -> "ocean wind" AND "level 3")
  - Event based tagging, e.g. return appropriate datasets and granules for hurricanes
  - Infuse into PO.DAAC keyword text search

Backups

National Aeronautics and Space Administration
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

PO.DAAC Web Portal

Dataset Discovery

- Faceted Browsing
- **Keyword search**
- Dataset Information Page/DOI Landing Pages
- Granule browsing through date tree

PO.DAAC User Working Group Recommendations

- Dataset Ranking is a long-standing problem in geospatial data discovery…data diversity and heterogeneity, user search intent
- UWG recommendations over past several years
- *…..Improve search and discovery of PO.DAAC dataset via free text (.e.g., keyword) and facets*
- *…..Develop advanced search capabilities*
- While faceted search provides a systematic approach to group data artifacts, facets are still static and rely on manual keywords tagging.
- Search relevance requires multi-dimensional dynamic ranking of data

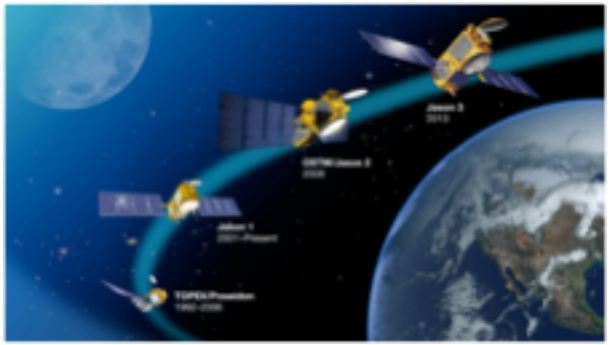
National Aeronautics and Space Administration
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

PO.DAAC Data: 500+ datasets

NASA Missions & Projects

Seasat, TOPEX/Poseidon, Jason-1, NSCAT, SeaWinds on ADEOS-II, QuikSCAT, ISS-RapidSCAT, GRACE, GHRSST, SPURS, MEaSUREs, Aquarius, CYGNSS, GRACE-FO (2017)

Upcoming: COWVR, AirSWOT, SWOT, GRACE-2

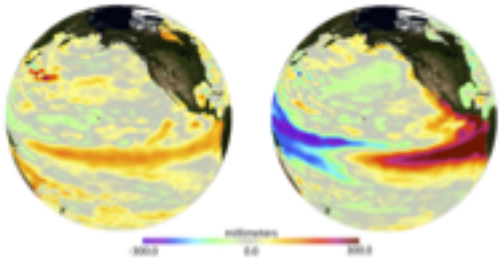


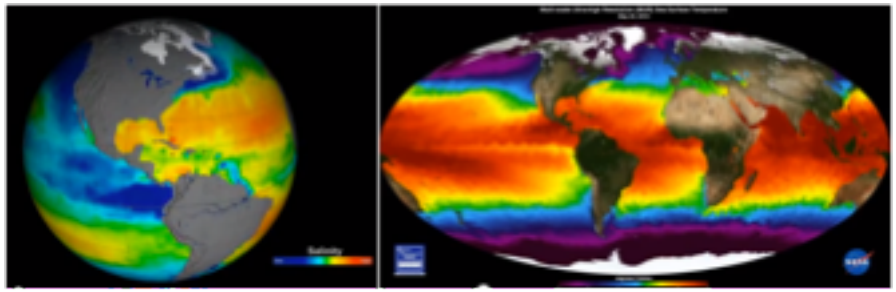
Ocean & Climate Community Driven

Value-added datasets in support of NASA programs

Gravity
Ocean Circulation & Currents
Ocean Surface Salinity
Ocean Surface Topography
Ocean Vector Winds
Sea Surface Temperature

Hydrology
Ocean Color
Sea Ice





Additional potential future improvements

- 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 near real-time data ingestion to dynamically update knowledge base
- Leverage advanced computing techniques to speed up the process

Relevancy

- Dataset relevancy and selection factors
  - Data quality, accuracy
  - Documentation, interoperability
  - Time series length, latency
  - Images
  - Resolution
  - Application
- PO.DAAC facets address some of these but there is lots to improve

Objectives

- Objectives:
  - Dataset relevancy mining, Dataset similarity calculation, Recommendations based on metadata attributes and user workflow patterns
- JPL contributions:
  - Developing use cases for search, discovery and utilization of earth science data focusing on datasets residing the JPL PO.DAAC
  - Ontology development for improving semantic searching of MUDROD
  - Implementation in PO.DAAC labs
  - Leverage ESDSWG Data Quality recommendations on distinguishability

• MUDROD

- *(Mining and Utilizing Dataset Relevancy from Oceanographic Dataset)*

- MUDROD funded by NASA AIST project intends to improve search relevancy and ranking for NASA earth science products from a user perspective with the Metadata, Usage Metrics, and User Feedback information.

- **Key features:**

❑ *Goals: (upper figure)*

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

❑ *Architecture (bottom figure)*

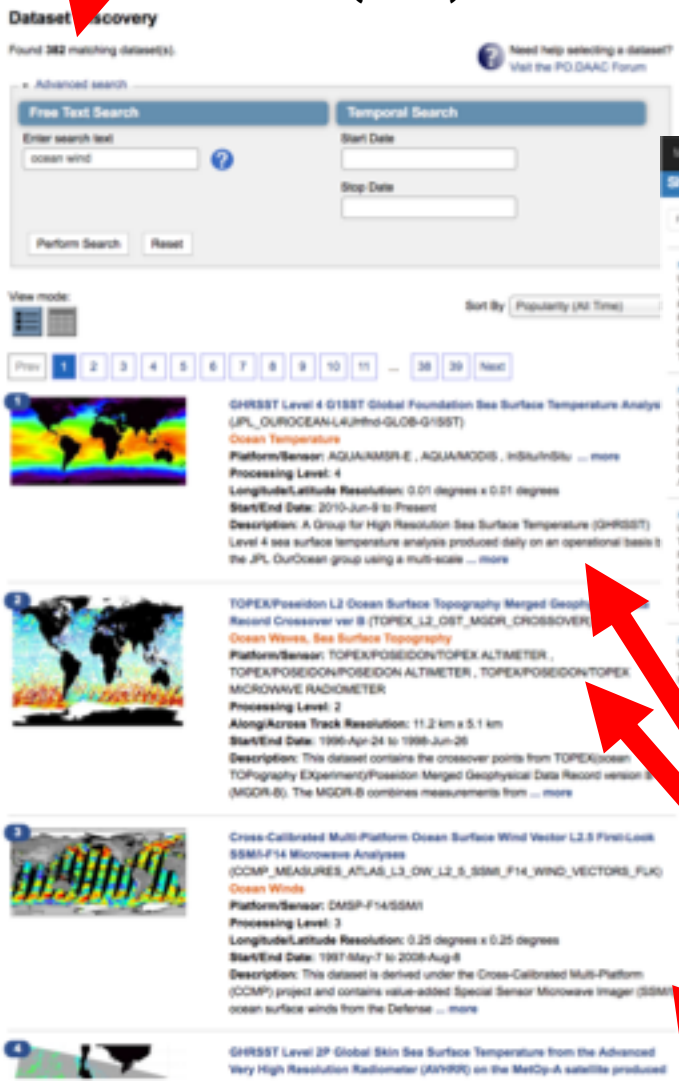
- The flowchart of constructing MUDROD smart search engine from user query to the ranked results, indicating a complicate process of user driven data search.

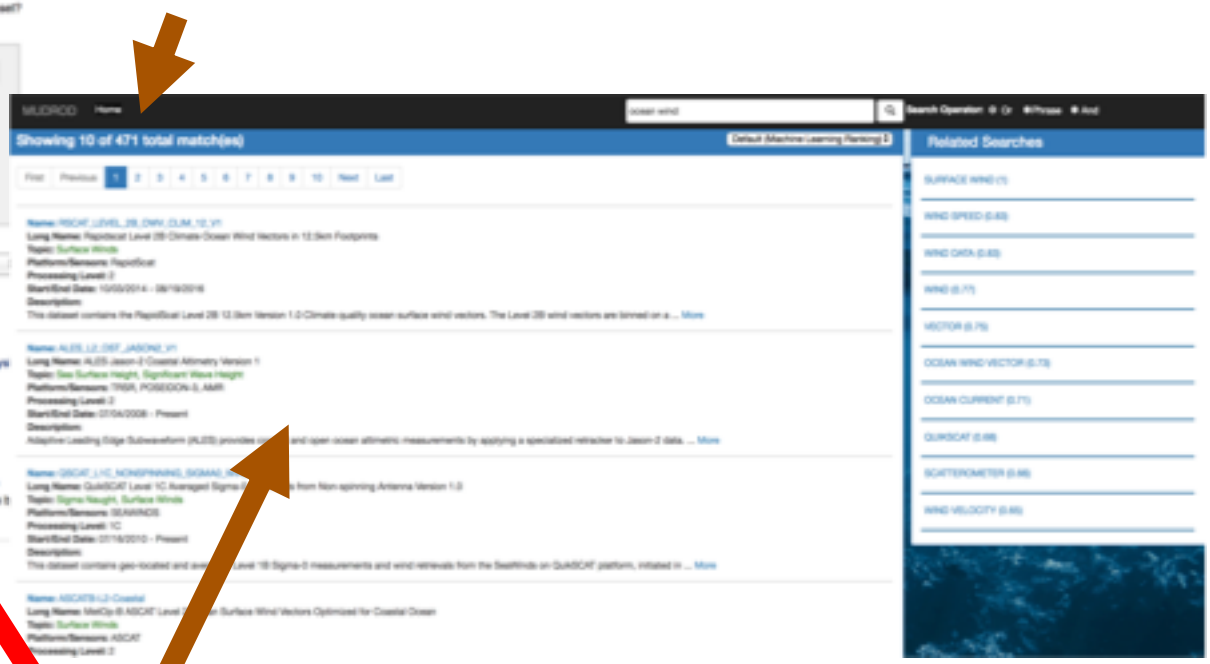
National Aeronautics and Space Administration
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Comparison of "ocean OR wind" search results

PO.DAAC (Solr)

MUDROD





Not Relevant !! (SST or SSH altimeter datasets)

- MUDROD results:
  - Recall similar
  - Precision improved !