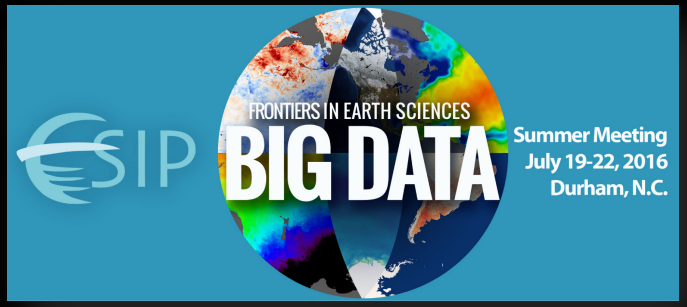
Challenges and Progress in Search Relevancy

Edward Armstrong, Lewis John McGibbney, Kim Whitehall

ESIP Summer Meeting

July 20, 2016



National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Introduction

The ESDSWG Search Relevance Working Group exists with the primary aim of improving search result relevance for EOSDIS data.
The main stakeholders are end users, especially those who *do not already know what data are available.*
We intend for our WG efforts to be relevant and benefit both ESDIS and DAACs by providing a better search experience, with more users finding the data they are looking for.

Technical Chairs:

- Edward Armstrong, Senior Data Engineer NASA JPL
- Lewis John McGibbney, Data Scientist II NASA JPL

ESDIS PoC:

- Christopher Lynnes

In a Nutshell:

- Formed in mid May 2015
- WG roster currently includes around 20 individuals
- 2016-17 Activity includes 5 active sub groups… more to come on them!

2015 Working Group Action Plan

Mission Statement

Improve search results relevance for EOSDIS data.

Stakeholders

- End Users (esp. those that do not know what data re available)
- EOSDIS and DAACs by providing a better search experience

Approach

Subgroups will focus on specific topics:

- Dataset Relationships
- Spatial Relevance
- Temporal Relevance
- Federated Search
- Dataset Relevance Heuristics

- 6 inactive Subgroups (2016)

Outcomes, Deliverables

- Recommendation for Essential Metrics
- Benchmarking study of dataset rankings for canonical studies
- Recommendations for enhancement to Common Metadata Repository

Dataset Relationships Subgroup

Aim:
To provide a common framework for identifying relationships across datasets with the purpose of lowering the barrier to obtaining similar datasets for a given user query.

Objectives:
We propose this in the following ways:

- **Modeling User History/Behaviors**: just as eCommerce platforms such as eBay, Amazon, etc. offer supplementary information to users that selected *X* also selected *Y and Z.*
- **Feature Co-occurrence present within Scientific Data Sources**: utilizing external academic literature (for example) where relationships are built based upon the presence of features within the same academic literature.
- **Utilization of Linked and Semantic Data Vocabularies within Dataset Home Pages**: relevant vocabularies such as Sweet, Schema.org, etc. should be further evaluated and proposed as mandatory for building relationships and an online presence on commercial search engines.

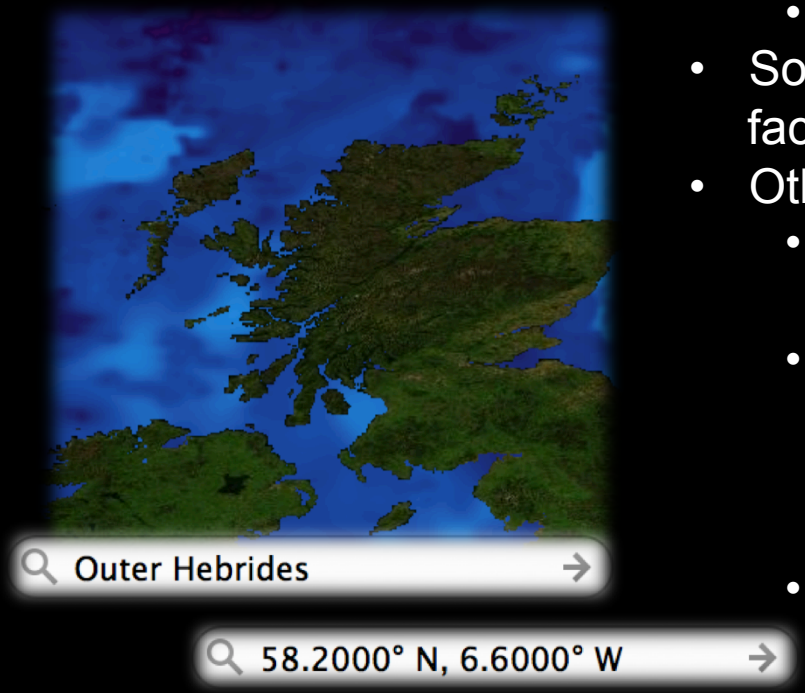
Spatial Relevance Subgroup

Aim:
Improve relevance ranking based on dataset spatial characteristics.

Objectives:
Improvements to DAAC and CMR search methods for spatial search

Metrics considered:

- Spatial overlap to request ranked with highest priority (collection or granule search).
  - Sorting by regional area or bounding box
  - Higher percent overlap ranked higher
- Sorting and ranking by spatial resolution: both facet and free text searches
- Other metrics considered or lower priority
  - Spatial proximity to an "event" (e.g., flood or hurricane)
  - Sorting by "true" spatial resolution of L3/ L4 datasets: i.e., feature detection (This requires complex spectral analysis, there are different ways to do this)
  - Search and rank by spatial projection

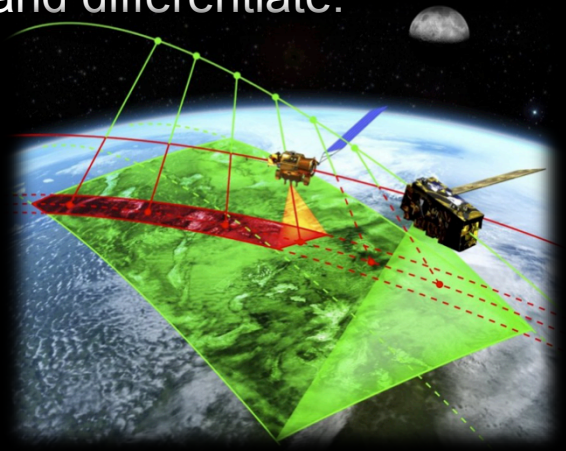


Temporal Relevance Subgroup

Aim:
Improve relevancy ranking based on dataset temporal characteristics.

Objectives:
Improvements to DAAC and CMR search engines for temporal search
With the aim of ranking and sorting datasets, investigate and differentiate:

- **time resolution**
- **time coverage**
- **repeat coverage**
- **an "event"**



Metrics considered:

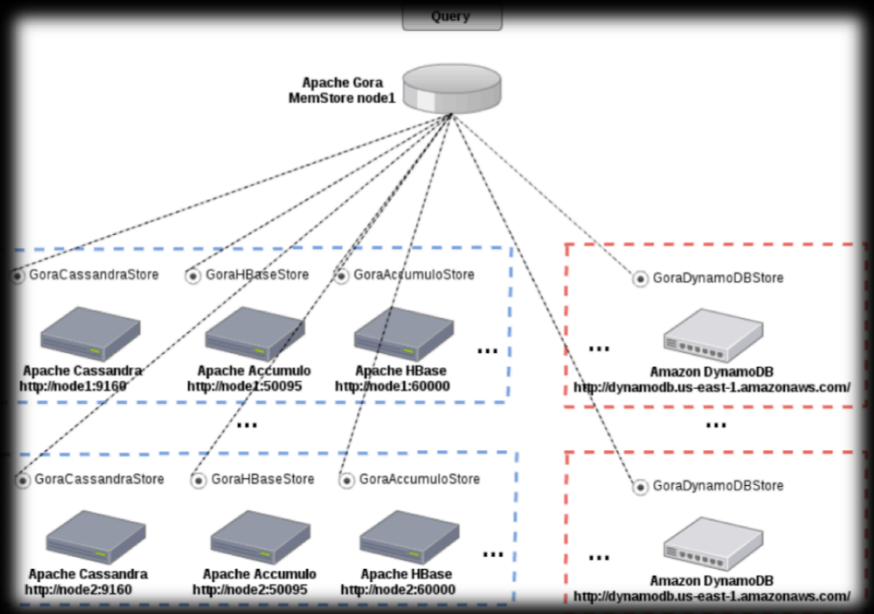
- temporal resolution: hourly, daily, weekly etc.
- orbital repeat and sampling
- time series length
- Other metrics considered or lower priority
  - temporal proximity to an "event" (e.g., flood, earthquake, hurricane)

Federated Search Subgroup

Aim:
Provide substantiated metrics and guidance on improving Information Retrieval practices within a Federated Search context.

Objectives:
Develop retrieval metrics that allow the simultaneous search of multiple searchable resources where users make a single query request which is distributed to search engines/indexes/resources participating in the federation.

- **resource identification**: from a number of available information resources, which ones are selected, and
- **results merging**: based on the execution of queries, the mechanism by which results are ranked and returned to the user as a singly ranked list of results.



Metrics considered:
Normalized Discounted Cumulative Gain: measure topical relevance as the main metric.

Case Study(ies):
NSIDC Federated Arctic Data Explorer

Dataset Relevance Heuristics Subgroup

Aim:

Relevance Heuristics seeks to leverage the years of experience that EOSDIS has in serving end users to establish rules of thumb for returning the most relevant search results. We aim to implement at least the top 2 or 3 heuristics in CMR and other search engines. This effort is therefore a conglomerate of all other WG efforts.

Objectives:

We propose this in the following ways:

- Obtain WG agreement on useful heuristics for sorting or scoring results in order of relevance.
- Identify search-support metrics to be collected (e.g., dataset popularity).
- Identify performance-criteria metrics to be used in evaluating relevancy results (e.g., AP Correlation, Precision and Recall).

Metrics and Case Studies:

- Golden datasets from the DAACs: dataset collections for various canonical queries, ranked manually by DAAC subject matter experts.
- Click event logging of the dataset results for CMR and/or the Earthdata Search Client
- Search Precision and Recall are key accuracy heuristics

Summary of current recommendations

- **Spatial/Temporal Relevance**
  - Spatial and temporal overlap, and intrinsic spatial and temporal resolution are key factors to be used in search result ranking
  - Ranked results from targeted keyword driven regional spatial searches should be improved
- **Dataset Heuristics**
  - Accuracy heuristics using search recall and precision should be implemented.
  - Keyword searches should weighted to favor collections containing controlled science keywords (e.g., GCMD keywords)
  - Search clients should record click through of dataset results
- **Federated Search**
  - Federated searched should utilize normalized discounted cumulative gain (nDCG) to measure topical relevance as the primary metric when merging results from the federated queries
- **Dataset Relationships**
  - Scientific literature should be mined for data and dataset semantic relations and keyword associations. This information can be used to quantify relationships via ontologies and databases holding triple stores that search engines can eventually leverage.

Full oral report presented at ’16 ESDSWG meeting. Written report in final form submitted shortly thereafter and available at http://bit.ly/29U9sBc

2016 Working Group Action Plan

Mission Statement

Improve search results relevance for EOSDIS data.

Stakeholders

- **EOSDIS and DAACs by providing a better search experience**
- **End Users (experience and novices)**

Approach

Subgroups will focus on specific topics:

- **Content-based Optimization for Commercial Search Engines**
- **Dataset Relevance Heuristics**
- **Granule-level Relevance**
- **Semantic Dataset Relationships**
- **User Characterization**

Engage ESIP Discovery and Semantic Technologies Groups

- ESIP search relevance hackathon
- ESIP search relevance breakout

Outcomes, Deliverables

1. Recommendations for Enhancements to Common Metadata Repository (CMR)
2. Evaluate the results from ESIP hackathon in CMR
3. Mid-term report on commercial search engine sub group findings
4. Empirical study of user characteristics based on URS profiles
5. Recommendations for DAAC weighted CMR scoring

Guest Presentations

A number of subject matter experts have come to present to the WG at our Telecons.

- Prof. Grace Hui Yang, Assistant Professor, Dept. Computer Science, Georgetown University - Dynamic Information Retrieval Modeling
- Lindsey Spratt, Highfleet Inc. - HIGHFLEET Semantic Federation Integrated with an Ontologically-Driven Deductive Database System.
- Dr. Djoerd Hiemstra, Associate Professor Database and Search Engine Technology at the University of Twente, Netherlands – Federated Search
- …many more have taken place and even more to come in upcoming meetings!

https://wiki.earthdata.nasa.gov/display/ESDSWG/Guest+Presentations

Working Group Resources

The WG maintains a growing active archive of resources including free courses, books and relevant literature as well as a growing information resources for desk studies e.g. Goddard DISC Mirador search queries

https://wiki.earthdata.nasa.gov/display/ESDSWG/Search+Relevance+Resources

Community Outreach/Collaborations

See us at 2016 Fall AGU Session IN030: ***New Approaches to Data Discovery Across Geoscience Domains***

- We would appreciated more DAAC representatives (and others) on our WG.
- More input from the needs and experiences of data centers
- Develop more Use Cases for development of search relevancy metrics
- Improve our progress and help address WG commitments and plans for 2016

Relevant/Honorable Mentions

Deep Insights – Search Analytics for the Domain Sciences
http://sched.co/6uHS
Contact: Chris Mattmann

Search Relevancy 101
http://sched.co/7X6Q
Contact: Chris Lynnes, Doug Newman

Thank you all… very much

Questions?

Find us on our mailing list

esdswg-search@lists.nasa.gov