Relevancy Ranking of Satellite Dataset Search Results

WGISS 2017

Christopher Lynnes (NASA ESDIS)
Patrick Quinn (Element 84)
James Norton (Element 84)
The Variety problem in Big Data from Satellites
Variety = Choice

Choice = Good

(Right?)
The Earth Observing System Data and Information System (EOSDIS)

data downlink

capture and clean

capture

archive

distribute

Research

Applications

Education

Users
The Variety problem in Big Earth Data from Satellites

Distinct Science Products Distributed

- Years: 1998 to 2016
- Y-axis: 0 to 4,000
- Data points show a steady increase from 1998 to 2016.
Earthdata Search

Discover Earth Science Data

Search NASA Earth Science data by keyword and filter by time or space.

- ozone

Browse All Data

See featured collections or use categories to narrow your results.
Too many datasets to sift manually
Where does Variety come from?

Instruments
  Fundamental differences: sounders, limb sounders, imagers...
  Incremental evolution in instrument design

Satellites
  “Same” instrument on different satellites

Processing Level
  Calibrated -> Swath -> Grid -> Model

Processing Algorithm
  Different basic principles
  Incremental evolution in algorithm development

Temporal Resolution
  daily, 5-day, 8-day, monthly, yearly

Spatial Resolution...
Example: Time Aggregation

Aerosol Optical Depth at 555 nm from Multi-angle Imaging Spectro-Radiometer

- Daily
- Monthly

- Selected data range was 2003-08-21 - 2006-09-21. Title reflects the date range of the granules that went into making this result.
What to do?

Emulate the best search engines: return the most relevant results at the top of the list
“how well a retrieved document or set of documents meets the information need of the user”
HOW?
Relevancy Ranking Heuristics

Heuristic = “rule of thumb”
Basis is 20+ years of serving satellite data to researchers
The Content Heuristic*

Got ozone?
“New-and-improved” Heuristics
New-and-Improved Processing Version
New processing version is also more likely to be up to date
Newer instrument is usually better than previous instruments
Region of Interest Overlap
## Time Range Heuristic

Datasets covering the user’s full time range are better than those covering just part of it.

<table>
<thead>
<tr>
<th>Time range of interest</th>
<th>2005</th>
<th>2006</th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOMS-Earth Probe</td>
<td>Red</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ozone Monitoring Inst.</td>
<td>Green</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Meh. Yeah!
Spatial Heuristic

Data covering the user’s full area are better than those covering just part of it. This is not as good as...
Spatial Heuristic

This...
User-centric Heuristics
Community Usage Heuristic

The dataset most often used by the community is more likely to be useful

<table>
<thead>
<tr>
<th>Data Product</th>
<th>Users**</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aqua AIRS Level 3 Daily Standard Physical Retrieval (AIRS only)*</td>
<td>164</td>
</tr>
<tr>
<td>Aqua AIRS Level 3 Daily Standard Physical Retrieval (AIRS+AMSU)*</td>
<td>714</td>
</tr>
</tbody>
</table>

*Version 6
** Jan 1, 2016 - June 20, 2016
# User Intent Heuristics

<table>
<thead>
<tr>
<th>User type or intent*</th>
<th>The most relevant datasets are...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Applications users</td>
<td>High spatial resolution, near-real-time</td>
</tr>
<tr>
<td>Students</td>
<td>Easier to use data <em>e.g., L3 grids in netCDF</em></td>
</tr>
<tr>
<td>Climate Modeler</td>
<td>Datasets on Climate Model Grid</td>
</tr>
</tbody>
</table>