Fixing Dataset Search

Chris Lynnes
GES DISC
An Experiment...

Search for “Ozone” data distributed by GES DISC
Relevant = dataset-actually-*has*-ozone

<table>
<thead>
<tr>
<th></th>
<th>Results</th>
<th>Precision</th>
<th>Recall</th>
<th>False Pos</th>
<th>False Neg</th>
</tr>
</thead>
<tbody>
<tr>
<td>GCMD</td>
<td>209</td>
<td>61%</td>
<td>98%</td>
<td>81</td>
<td>2</td>
</tr>
<tr>
<td>ECHO</td>
<td>179</td>
<td>65%</td>
<td>89%</td>
<td>63</td>
<td>14</td>
</tr>
<tr>
<td>Mirador</td>
<td>65</td>
<td>82%</td>
<td>41%</td>
<td>12</td>
<td>77</td>
</tr>
</tbody>
</table>

Precision = \[
\frac{\text{Relevant} \cap \text{Retrieved}}{\text{Retrieved}}
\]

Recall = \[
\frac{\text{Relevant} \cap \text{Retrieved}}{\text{Relevant}}
\]
Ouch.
Top 10: GCMD

1. MLS/Aura Level 2 Hydroperoxy (HO2) Mixing Ratio
2. MLS/Aura Level 2 Temperature
3. MLS/Aura Level 2 Hydroxyl (OH) Mixing Ratio
4. MLS/Aura Level 2 Nitric Acid (HNO3) Mixing Ratio
5. MLS/Aura Level 2 Water Vapor (H2O) Mixing Ratio
6. MLS/Aura Level 2 Hydrogen Chloride (HCl) Mixing Ratio
7. MLS/Aura Level 2 Nitrous Oxide (N2O) Mixing Ratio
8. MLS/Aura Level 2 Ozone (O3) Mixing Ratio
9. MLS/Aura Level 2 Chlorine Monoxide (ClO) Mixing Ratio
10. MLS/Aura Level 2 Carbon Monoxide (CO) Mixing Ratio
Top 10: ECHO

1. OMI/Aura Formaldehyde (HCHO) Total Column Global 0.25deg Lat/Lon Grid
2. Aqua AIRS Level 3 Daily Standard Physical Retrieval (AIRS-only)
3. OMI/Aura Zoom-in Ground Pixel Corners 1-Orbit L2 Swath 13x12km
4. MLS/Aura Near-Real-Time L2 Temperature
5. OMI/Aura NO2 Cloud-Screened Total and Tropospheric Column Daily L3 Global 0.25deg Lat/Lon Grid
6. UARS Improved Stratospheric and Mesospheric Sounder (ISAMS) Level 3AL
7. UARS Cryogenic Limb Array Etalon Spectrometer (CLAES) Level 3AL
8. SBUV2/NOAA-16 Ozone (O3) Nadir Profile and Total Column Daily L2
9. TOMS/Earth Probe UV Reflectivity Monthly L3 Global 1x1.25 deg Lat/Lon Grid
10. GLA DAILY GRIDS from NOAA-10
Top 10: Mirador

1. OMI/Aura Ozone (O3) Total Column 1-Orbit L2 Swath 13x24 km
2. OMI/Aura DOAS Total Column Ozone Zoomed 1-Orbit L2 Swath 13x12km
3. OMI/Aura Ozone (O3) Total Column Daily L2 Global 0.25 deg Lat/Lon Grid
4. SBUV/Nimbus-7 Ozone Profile, Ozone Total Column 1-Orbit L2 200x200 km
5. SBUV2/NOAA-09 Ozone Profile, Ozone Total Column 1-Orbit L2 200x200 km
6. SBUV2/NOAA-11 Ozone Profile, Ozone Total Column 1-Orbit L2 200x200 km
7. SBUV2/NOAA-16 Ozone Profile, Ozone Total Column 1-Orbit L2 200x200 km
8. OMI/Aura Ozone (O3) DOAS Total Column Daily L2 Global 0.25 deg Lat/Lon Grid
9. OMI/Aura Ozone (O3) DOAS Total Column 1-Orbit L2 Swath 13x24 km
10. GOZCARDS Source Data for Ozone Monthly Zonal Means on a Geodetic Latitude and Pressure Grid

Better, but still not that good:
- No L3
- L2G > L3
- Pre-EOS > EOS (#4-#7)
- Specialty product (Zoomed) > Global
So That’s Why Users Hate on Search
Implementing *good* relevancy ranking may be the single most important thing we could do to improve our search
How to Relevant

• It’s what made Google famous
• Ironically, Google’s PageRank does not work well for dataset “documents”
• OTOH, DAACs and EOSDIS have been helping users select data for > 20 years
• We should know by now what users want
Epiphany #2

Where We Slipped Up:

Just because users would rather
SEARCH WITH freetext keywords,
that doesn’t mean we have to
FIND BY freetext methods.
Relevancy Heuristics

• An attempt to return datasets in the most relevant order, given:
  – Keywords provided by user
  – Likely user intent, inferred from:
    • Empirical experience
    • Type of user
    • Referrer
    • Keyword

• N.B.: *No heuristic works for everybody*
  – Just trying to make as many users happy as possible
Data Content Heuristic

Match in measurement/parameter/GCMD specific keyword

*is more relevant than*

Match in any other field

*Example:

Ozone as a variable

*is more relevant than*

Ozone in “Ozone Monitoring Instrument”*
Data Content “Noun” Heuristic

“Nouns” in a Measurement Match

is more relevant than

Modifiers in a Measurement Match

Example:

“Ozone” in “Total Column Ozone”

is more relevant than

“Total” in “Total Column Ozone”
Data Version Heuristic

More Recent Version

is more relevant than

Less Recent Version

Example:

MLS Whatever V003

is more relevant than

MLS Whatever V002
Date Range Coverage Heuristic

Date Range Coverage Score

\[
\text{Date Range Coverage Score} = \frac{\% \text{ of Search Date Range covered by the dataset}}{8}
\]

Example:

User: want Oct 2007 to Oct 2014 (8 yr)

Coverage = 6/8 = 75%
Ease of Use Heuristic

Processing Level as a Proxy:

• L4 – model, ergo gridded + usually gap-free
• L3 – nicely gridded
• L2 – processed to physical variables but tricky to map and interpret quality flags
• L1 – Do-it-yourself
• L0 – Science teams only (or mostly)
“What’s New” Heuristic*

Newer datasets
are usually better than
Older datasets

*Resonates with people’s Recency Bias
Metrics-based Heuristics

• Data popularity over last N months
• Data citations over last N months
• Associative popularity...not so much
  – “Customers Who Bought This Item Also Bought”
  – “What Other Items Do Customers Buy After Viewing This Item?”
  – N.B.: You have to get to and select a relevant item FIRST
User Type and Intent Modeling

*Use it to weight or otherwise modify previous heuristics*

- User interests
  - URS
  - Other search words (“Rain” + “Landslides”)
- Referrals from Portals or Tools
  - Applications* portals/tools
  - Education portals/tools
- Use of Jargon or Technical Terms
  - AOD, SST, L2, OMI, SNPP...

*Applications like “Landslide Prediction”, not like “MS Office”*
What Do We Need?

• The ability to rank CMR dataset results according to heuristics
  – Ideally, client-selectable heuristics and scoring
• The ability to experiment with ranking schemes against CMR metadata
  – Lure researchers to develop ranking schemes
• Discussion forum for developing additional methods for ranking and modeling user intent
Discuss...

For White Paper, see:

https://wiki.earthdata.nasa.gov/display/CMR/Relevancy+Ranking+of+Data+Collections