Content-Based Metadata System: A Workbench to Prototype Data Mining Concepts

Ramachandran Suresh, NASA/MTECH
Joel Sachs, UMBC GEST
Robin Pfister, NASA Goddard Space Flight Center
Jeanne Behnke, NASA Goddard Space Flight Center

Abstract

The purpose of this prototype is to identify appropriate higher-level (levels 3 and 4) datasets for an on-line Content-Based Metadata (CBM) "warehouse". Higher-level data products contain processed, geophysical parameter data on a space-time grid, and were derived based on widely accepted algorithms. If brought to an on-line environment, they could serve as content-based metadata for searching other data held in NASA archives. Such a warehouse would also serve as a useful workbench for data mining tools, allowing researchers to explore what is and isn't useful to the science community so that true and useful data mining algorithms can be developed and implemented as part of NASA's data systems. As well, with an appropriate user interface, such an archive would be a very useful research-planning tool that does not exist today.

We describe a 3-phase prototype that will provide the above utilities to scientists. Phase I involves identifying appropriate higher-level datasets, and making them usefully available on line. We have analyzed and identified level-3 data products from EOS Terra, TRMM and other missions. A survey of science users and literature review were conducted to identify the geophysical parameters of importance. A variety of database, mark-up, and COTS technologies are being evaluated. This paper discusses the current status of the project and potential technologies that will be used.

1. Introduction

The main challenges in data retrieval and dissemination is to accurately and precisely provide the user with only those data that will meet his/her needs. This requires content-based search capabilities. A user is unable, today, to issue queries like “Show me areas where ground level ozone concentrations have increased by 5 % over the last decade.” This is because traditional metadata parameters (date, coordinates, instrument, etc.) do not tell us about the phenomena actually demonstrated by the data. Higher level data products (levels 3 and 4) can serve as content-based metadata for searching lower level products. Prototyping this capability is a goal of this project.

Content-based search and data mining are highly related. Either one can be used as the basis for the other. For example, data mining software (specifically, classification tools) can be used to identify hurricanes in historical weather data. This would enable a response to the content-based query “Show me all the hurricanes in the past 7 years”. Working the other way, a warehouse supporting content-based search could be used as a mining base. Achieving this second scenario is another goal of this project.

Another goal is to prototype the archive as a research-planning tool. Section 2 describes these 3 goals; section 3 summarizes the results of a questionnaire we used to determine use cases and requirements; section 4 describes the data sets that have been selected; section 5 briefly discusses the technologies we are evaluating; and section 6 concludes with progress to date, next steps, and future work. The remainder of section 1 is devoted to provide a quick introduction to the different levels of Earth science data products.

Earth Science data products begin as unprocessed instrument/payload data at full resolution (Level 0). Level 1 data is still unprocessed, but has been time-referenced, and annotated with ancillary information, such as geo-location, and sensor calibration information. Level 2 data has been processed into a derived geophysical product, and re-sampled into a selected map projection. Level 3 data comprises variables mapped on uniform space-time grid scales, often with global coverage. Level 3 products include geophysical parameters derived from multiple instruments. Level 4 data products are typically model output or results from analysis of lower level data.

2. Goals

Higher-level products contain derived geophysical parameters. Because they are of relatively small volume it is possible to put them into an on-line environment for a variety of uses. We are prototyping a multi-purpose on-line archive of higher-level (mostly level 3) data products.

i. Content-Based Metadata for Searching Other Products

Pre-computation of selected statistics is a common approach to providing content-based metadata. A vexing question is always which statistics to pre-compute. Level 3 holdings can be viewed as pre-computed summaries of lower level data sets. What's more, level 3 datasets have been constructed by scientists, and so we are guaranteed of their scientific relevance. It seems natural to us to use level 3 data as metadata for the lower level products. We envision scientists being able to use the values of the CBM Workbench products to help constrain query criteria against data held in EOSDIS. For example, today geophysical parameter-based queries are limited to geophysical parameter name. A specific query today might be "find all data that contains temperature or humidity values". With the CBM Workbench, this query could become "find all data where temperature values are between 25 and 30 C and Humidity values are between 80 and 100 percent".

ii. Research Planning

With level 3 data holdings on-line, a data visualization-based query system would allow a scientist to plot various geophysical parameters and to visualize features, anomalies and trends that would be interesting for further research. In addition, the system would allow the user to draw a box around the feature and issue a query to EOSDIS to find all data (or all data of a certain type) associated with that anomaly.

iii. Data Mining Test-bed.

Today in the retail industry, data mining is used for market basket analysis (e.g., to discover rules like "People who buy diapers on Thursday buy beer at the same time."). In the credit industry, data mining is used to establish credit worthiness (e.g., to discover rules like "People with no job and no bank account often default on their loans."). In the insurance industry, data mining is used to detect common patterns in claims known to be fraudulent.

In Earth Science, the story is a little bit different. In each industry example above, we start with a database with a high-level semantics. That is, the database deals with things such as "people", "diapers", "employment status", etc. In Earth Science, we begin with pixel data in binary form. A lot of processing must be done to extract semantics from the raw data. A lot of this processing involves classification and clustering, and so can be called data mining. Thus there is an equivalence between the content-based retrieval query "show me all images containing the spotted Owl's habitat", and the "data mining" problem of building a classifier that can spot habitat characteristics, based on a training set of the pre-identified habitat.

But once we've established the semantics of our images, we can do "industrial style" data mining. For example, suppose we have a bunch of geo-registered images. In one image, pixel values represent ground level ozone concentrations; while in another, pixel values represent ground surface temperature. Maybe we have 100, or 1000 of these images. We should be able to use commercial data mining software to find patterns, or correlations, that we wouldn't otherwise have looked for.

Level 3 data products have a high-level semantics, and should be amenable to this sort of general-purpose data mining.

The above goals are to be realized in 3 phases. The first phase is to identify appropriate higher-level datasets, and to bring them on-line. The remainder of the paper is devoted to describing our completed Phase 1 activities.

3. Survey

The first step was to build use cases, and to design requirements around them. To do this, we developed a questionnaire, and used it to survey researchers and data system developers. The questionnaire asked respondents for information about how they order data, how they use data, the types of queries that they issue, and the types of queries they would like to be able to issue. Those surveyed represented a mixture of academic and government research institutions involved in various aspects of Earth science research and data system development.

There are many similarities in the way users search and locate data, with most users relying on web based systems and tools. It is difficult for a user to order precisely the data s/he needs. After receiving data from an archive and other sources, users always have to extract data of their interest from the data they received. A combination of custom developed software, public domain utilities and COTS packages (ERDAS, IDL, Matlab, Intergraph, PC Image) are used to perform this extraction.

Although survey respondents typically had no trouble in using on-line ordering mechanisms to obtain the data they were looking for, a majority found the current data search and access capabilities for Earth science data limiting. They were of the opinion that it is complex and not user friendly. They also said that the inter-comparison of geophysical parameters is not easy today, and takes time since they have to order multiple data sets and extract the information in their own computing environment. Users said that a system that will enable on-line inter-comparison of multiple geophysical parameter values would be useful. Some of the parameters users want to compare are SST and Aerosol, SST and NDVI from multiple instruments, and Ozone and trace gases.

We also asked users to envision their "dream" data systems. Some responses were for Star Trek type capabilities, for example, "Computer, please identify all data relevant to my underlying scientific question, and do all necessary analysis to provide me with an answer." Others were for more modest data systems, and directly in line with the goals of this project (e.g. on-line, content-based search). Most respondents have not used data mining software, suggesting that current packages do not meet the needs of the Earth Science user community, and highlighting the importance of an Earth Science data mining test-bed.

The highlights of the survey are as follows:

1. Users like current on-line catalog and ordering features.
2. Users like simple user interfaces that will help them to locate the data of their interest.
3. Users desire seamless on-line data comparison from multiple instruments and satellites.
4. Data mining is not heavily used by the community.

5. When asked specifically, respondents said that an on-line level 3 data products based system would be useful, and would address many of their needs.

4. Data set analysis and selection

A list of data products from several sources including the EOS Terra site (http://grid2.gsfc.nasa.gov/~todirita/terra/terra_dataprod.html) and the SPSO database (http://spsosun.gsfc.nasa.gov/eosdata.html) and EOS interdisciplinary team were reviewed to identify suitable Level 3 data products. We found out that not all EOS Terra Level 3 data products are available yet. The bulk of the Level 3 data products available come from MODIS and CERES instruments. ASTER Digital Elevation Model (DEM) data is also available. No Level 3 data products were available from MISR and MOPITT instruments. Table 1 describes the initial list of data products selected for the study. The list of products will change as new products become available. These data products are currently viewable from the CBMW web site. Scales, projections and resolutions of the data products are being reviewed in order to select the appropriate list of parameters for the Phase 1 effort.

Sl No	Data Product Id	Brief Description	Discipline	Total Granules
1	AST 14	AST08 Digital Elevation Models	Land	20
2	MOD 10	Snow mapping algorithm and the sea ice mapping algorithm	Land	20
3	MOD 11	Land Surface Temperature and Emissivity level 3 -1day	Land	10
4	MOD 11	Land Surface Temperature and Emissivity level 3 -8day	Land	20
5	MOD 12	Land Cover / Land Cover Change	Land	20
6	MOD 13	Gridded Vegetation Indices (Max NDVI and Integrated MVI) 1 Km ISIN Grid	Land	20
7	MOD 14	Thermal Anomalies and biomass burning level 3 - 1 day	Land	20
8	MOD 14	Thermal Anomalies and biomass burning level 3 - 8 day	Land	20
9	MOD 27	Annual Ocean Primary Productivity	Ocean	20
10	CER 03	CER_ES9_Terra-FM1_Edition1 / ERBE- like Monthly Regional Averages	Atmosphere	14
11	CER 03	CER_ES9_Terra-FM2_Edition1 / ERBE- like Monthly Regional Averages	Atmosphere	15
12	MOD 08	MODIS L3 8-day joint	Atmosphere	16
13	MOD 11	MODIS / Terra Land Surface Temperature / Emissivity daily L3 Global 1km ISIN	Atmosphere	10
14	AST 14	AST14 Digital Elevation Models	Land	10

Table 1. The initial list of data products selected for the study.

Science scenarios:

Recent work related to content based metadata search and data mining funded by NASA organizations was reviewed. Particularly, we concentrated on the SPIRE system developed by IBM, and the work done by University of Alabama, Huntsville and George Mason University. Several science scenarios were identified from the literature and investigated. In summary, the literature survey indicated that it is difficult to develop general-purpose algorithms for data mining and content-based metadata search. To our knowledge, all existing content-based search prototypes were built

for specific domains (e.g. hantavirus, fire ants, etc.). Our system will leverage existing higher-level products to support a much broader range of ad-hoc queries.

5. Technologies

In Phase 1, we are experimenting with a variety of technologies to determine the best architecture for bringing higher-level datasets on-line. We describe below a few of the technologies we are considering.

Database

We need to support queries that compare and aggregate data from multiple HDF-EOS files. An example of such a query is "show me areas where ground level ozone concentration has increased by more that 5% over the last decade". A natural way to do this would be to store the data in a relational database, taking advantage of the comparison and aggregation functions provided by the RDBMS. Note that we would not need to store all HDF objects in the database, but only the scientific datasets that contain the numeric information we need available for comparison.

Markup Languages

Our system will interoperate with ECHO. It will be capable of acting both as an ECHO client, and also as an ECHO service provider. This interoperability will be achieved by using the ECHO XML DTD for all communications with ECHO. As well, we intend to provide broader interoperability, with both other data portals and with search services (such as Google.) We will do this by employing the standards of the emerging *semantic web*. Specifically, our holdings and capabilities will be advertised with in an RDF based mark-up language. Thus, any web service complying with W3C semantic web standards will be able to make use of the CBMW. Specific languages being looked at include XDF, GML, DAML, and ESML.

Advanced Tools

We are considering a number of COTS packages. These include E-Cognition, an object oriented data mining tool capable of multi-scale image analysis, and a variety of GIS packages. Specifically, we are looking to the built-in ability of a GIS to overlay images of varying scale and resolution.

6. Conclusion and Future Work

We described our activities to date in building an online content-based metadata warehouse. At the end of Phase 1, we will have constructed the initial on-line archive, together with a CBM search tool, critical path methodology based on our requirements document and technology review to provide a data mining test-bed. We would also like to determine the feasibility of integrating such a system into the EOSDIS environment via the EOSDIS Clearing HOuse (ECHO).

Follow on phases will identify ways to manipulate and visualize the content-based metadata and will support users to search existing databases based on content-based metadata. In Phase 2, we will identify useful pattern recognition algorithms for feature identification. We will also identify useful visualization tools for prototyping new visualization features that will help scientists to visualize the content-based metadata. In Phase 3, we will prototype graphical data displays to observe interrelationships between diverse data, identify anomalies or other features of interest, and specify the feature as input criteria into a data system search, thus "mining" the archived data with the static, content-based metadata.

References

Content-based Search and Data Mining Cluster of ESIP, "Science Scenarios for Content-based Search and Data Mining", 2000. http://esipfed.org:8080/Clusters/Content-Based/Sci_scen.html

Z. Li, X. S. Wang, M. Kafatos, and. R. Yang, "A Pyramid Data Model for Supporting Content -based Browsing and

Knowledge Discovery", in Proceedings of the 10th International Conference on Scientific and Statistical Database Management (M.Rafanelli and M.Jarke, eds.) pp.170-179, IEEE, Computer Society, 1998.

R. Ramachandran, H. Conover, S. Graves and K. Keiser, "Algorithm Development and Mining (Adam) System for Earth Science Applications", Second Conference on Artificial Intelligence, 80th AMS Annual Meeting, Long Beach, January, 2000.

C.S. Li., L. Bergman, V. Castelli, J.R. Smith, J.J. Turek, A. Achuthan, Y. C. Chang, and M. Hill. IBM SPIRE System, Final Report NASA CAN NCC5-101, September, 1999.

E-Cognition. http://www.definiens-imaging.com/central/index.htm