New Data Discovery Paradigms for the EarthCube

Rahul Ramachandran, Sara Graves, Helen Conover
Information Technology and Systems Center
University of Alabama Huntsville
Huntsville, AL 35899
rahul.ramachandran@uah.edu

Two new Data Discovery paradigms as *core capabilities* are envisioned for EarthCube. These new capabilities will expedite data discovery and use, and accelerate scientific research within the geosciences. The first data discovery capabilities will be based on the data content rather than just metadata. The second data discovery will support interdisciplinary research within EarthCube and utilize ontology driven semantics for data, service and information aggregation.

1. Content Based Data Discovery

The volume of raw data being collected and stored by different science instruments (or instrumentation) today defies even the partial manual examination by scientists. Data mining addresses this data glut problem by bringing to bear techniques and algorithms from a multitude of disciplines to analyze and explore these massive data sets. Data mining is a multidisciplinary domain borrowing techniques from the fields of machine learning, statistics, databases, expert systems, pattern recognition and data visualization. However, Data mining is an iterative, complex multi-step process. To fully exploit data mining process requires detailed knowledge of both the data, and the mining algorithms that can be brought to bear on it. Data mining tends to be too difficult and intricate an approach for a casual user to employ. A geoscientist typically requires the help of a data mining expert in order to employ data mining tools with confidence. Consequently, despite the availability of scientific mining tools [7,8,9], the use of data mining remains a niche activity restricted to "knowledge discovery". It still remains too specialized and is not an integral part of the analysis arsenal of most researchers.

We envision Earth Cube to expand the role of data mining and utilize data mining techniques to develop Content Based Data Discovery (CBDD) tools to bridge the existing gap between metadata based data discovery and the knowledge discovery. The CBDD tools will index data files in the Earth Cube using features that represent scientific concepts. Using these CBDD tools, researchers will be able to easily sift through large volumes of data to “discover” data files based on prescribed targets of interest. These CBDD systems will leverage the research developments from Content Based Image Retrieval Systems (CBIR) [2] and Information Mining [1,5].

2. Information Aggregation to build customized Data Albums

As described earlier, one of the largest continuing challenges in any geo science investigation is the discovery and access of useful science content from the increasingly large volumes of Earth science data and related information available. For example, in hurricane science, many hurricane researchers are familiar with limited, but specific datasets, but often are unaware of or unfamiliar with a large quantity of other resources. Finding airborne or satellite data relevant to a given storm often requires a time consuming search through web pages and data archives. Background information related to damages, deaths, and injuries requires extensive online searches for news reports and official storm summaries. This search process could be made much more efficient and productive if a system could provide not just links to airborne and satellite web sites, but links to specific instrument data and satellite data relevant to the storm(s), as well as all related reports, summaries, news stories, and images. Not only would such a system provide a richer resource of information to the researcher, it would also increase the exposure of scientific datasets to a community that may be unfamiliar with or hesitant to use the data because of difficulties in tracking down the individual satellite overpasses or airborne flights relevant to a storm.

Similarly, the search for individual convective events needed to develop case studies for weather forecast validation can be a challenge. Such case studies are important for improving the initial conditions of numerical weather prediction forecasts in order to enable improved prediction of challenging weather phenomena, such as convective thunderstorms that can produce damaging winds, hail, lightning, and tornadoes. While information is readily available about highly publicized cases such as the recent tornado outbreaks, less visible severe weather events that may make meaningful case studies might not be as well known to researchers. A tool that scours the Internet for textual forecasts leading up to an event, information about storm reports and news, and the necessary datasets for running model forecasts related to that given case would not only provide researchers with an opportunity to better use earth science data but also save them time in searching multiple data sources for all of this information.

The problem of locating data and information for geoscience case studies is very similar to information overload faced by web users today. Ever increasing online content has overwhelmed the traditional approach of browsing the web for information. A technological response to this problem is “aggregators” for gathering information from different sources. Aggregators reduce time spent visiting individual websites by automatically pulling content for users. They also support personalization, with the selection of content sources based on user preferences. Furthermore, advances in aggregator technology provide for content “curation”, where in addition to gathering information, the tool organizes, categorizes and ranks content by relevance.

We envison the EarthCube with capabilities such as ontology driven content aggregation to improve discovery and use of multi-instrument geoscience data, tools and related information meeting user-customizable criteria. These capabilities can be built on existing smart search technology [3,4,5] by adapting web topic monitoring, ontologies

and information aggregation concepts to geocience in order to create curated "Data Albums". These data albums will serve as compiled collections of information related to a specific science topic or event, containing links to relevant data files from different instruments; tools and services for visualization and analysis; information about the event contained in news reports, images or videos to supplement research analysis; and literature references.

3. References

[1] Datcu, M.;, H. Daschiel, A.; Pelizzari, M. Quartulli, A. Galoppo, A. Colapicchioni, M. Pastori, K. Seidel, P.G. Marchetti, and S.; D'Elia. 2003. Information mining in remote sensing image archives: system concepts IEEE Transactions on Geoscience and Remote Sensing 41 (12).

[2] Datta, Ritendra, Dhiraj Joshi, Jia Li, and James Z. Wang. 2008. Image Retrieval: Ideas, Influences, and Trends of the New Age. ACM Computing Surveys 40 (2):1-60.

[3] Movva, Sunil, Rahul Ramachandran, Sara Graves, and Helen Conover. 2008. Customizable Search Engine with Semantic and Resource Aggregation Capability. Paper read at The Semantic Web meets the Deep Web Workshop, IEEE Joint Conference on E-Commerce Technology and Enterprise Computing, E-Commerce and E-Services July 23, 2008, at Washington DC.

[4] Movva, Sunil, Rahul Ramachandran, Xiang Li, Phani Cherukuri, and Sara Graves. 2007. Customizing a Semantic Search Engine and Resource Aggregator for different Science Domains. Paper read at Geoinformatics, at San Diego, CA.

[5] Munoz, Ines Maria Gomez, and Mihai Datcu. 2010. Image Information Mining Systems, Geoscience and Remote Sensing New Achievements. In Geoscience and Remote Sensing New Achievements, edited by P. I. a. D. R. (Ed.): InTech.

[6] Perez, Sarah. A Productive Application of Semantic Search 2009 [cited. Available from http://www.readwriteweb.com/archives/a_productive_application_of_semantic_search.php.

[7] Ramachandran, Rahul, Sara Graves, Todd Berendes, Manil Maskey, C. Chidambaram, S. Christopher, Patrick Hogan, and Tom Gaskins. 2009. GLIDER: A comprehensive software tool to visualize, analyze and mine satellite imagery. Paper read at IEEE International Geoscience & Remote Sensing Symposium, at Cape Town, South Africa.

[8] Ramachandran, Rahul, John Rushing, Xiang Li, Chandrika Kamath, Helen Conover, and Sara Graves. 2006. Bird's Eye View of Data Mining in Geosciences. In Geoinformatics: Data to Knowledge, edited by A.K.Sinha: Geological Society of America Special Paper 397.

[9] Rushing, John, Rahul Ramachandran, Udaysankar Nair, Sara Graves, Ron Welch, and Amy Lin. 2005. ADaM: A Data Mining Toolkit for Scientists and Engineers. Computers & Geosciences 31 (5):607-618.