Mining as a Service

Rahul Ramachandran*, University of Alabama Huntsville

*rramachandran@itsc.uah.edu

Sunil Movva, University of Alabama Huntsville

U.S Nair, University of Alabama Huntsville

Chris Lynnes, NASA/GSFC

Peter Fox, RPI

Era of Data Intensive Science

There are two main problem areas that drive the need for providing data mining functionality to the science community. First, since the Earth functions as a complex system existing information system infrastructure and science practices are inadequate to address the many interrelated science problems. Solving these complex problems requires new approaches, such as data mining, for analyzing science datasets. Second, the volume of raw data being collected and stored in different data archives today defies even partial manual examination. Data mining, however, provides an automated method for data intensive analysis. Data mining addresses both of these problems by bringing together techniques and algorithms from a multitude of disciplines to analyze and explore these massive data sets [Fayyad et al., 1996]. These techniques are borrowed from the diverse fields of machine learning, statistics, databases, expert systems, pattern recognition and data visualization [Ramachandran et al., 2006]. Data mining is an iterative, multi-step process which includes data preparation, cleaning, preprocessing, feature extraction, selection and application of the mining algorithm(s), and, finally, analysis and interpretation of the result or the discovered patterns [Dunham, 2003].

Service-Oriented Architecture

Web services are well-defined, self-contained functions that can be invoked via the Internet. The consumers of web services are either other computer applications or users interfacing with thin web clients that communicate, usually over HTTP, using XML-based web services specifications including SOAP, WSDL, and UDDI. The web services paradigm creates interoperable computing assets by providing a standards-based way to describe them. This allows not only integration of disparate systems with reduced cost and complexity but also the creation of new applications with minimal effort. Web services form the basis for the concept of a Service Oriented Architecture (SOA), providing a path to exploit existing computing assets and transform them into more agile services. The underlying principle of an SOA is to

minimize the need for writing new code each time a new processing capability is required, and to provide optimal means to reuse existing computing assets. In the context of science data processing and analysis, this architecture can provide a suite of data processing and analysis functions to scientists and end-users in a flexible processing system. In addition, different user algorithms packaged as web services can easily be integrated with other processing and analysis services for quick testing and comparison.

Mining in the SOA paradigm

The SOA paradigm allows mining algorithms to be made available as web services along with backend computational resources [Ramachandran et al., 2006]. Scientists can use their browser as a client to create mining workflows, access data from distributed locations and execute the workflow at a distributed computational resource. This paper will describe in detail the architecture (shown in Fig 1) for providing data mining or any analysis algorithm as a service. The architecture has two reusable components. The first is a standards-based backend service infrastructure. The details of the service infrastructure will be discussed in this paper. The second component is the portal interface which is critical as its used by scientist to not only create the workflows but also to document and share their workflows, analysis and results with others. Examples of actual use of these mining services by scientists will also be shared in this paper.

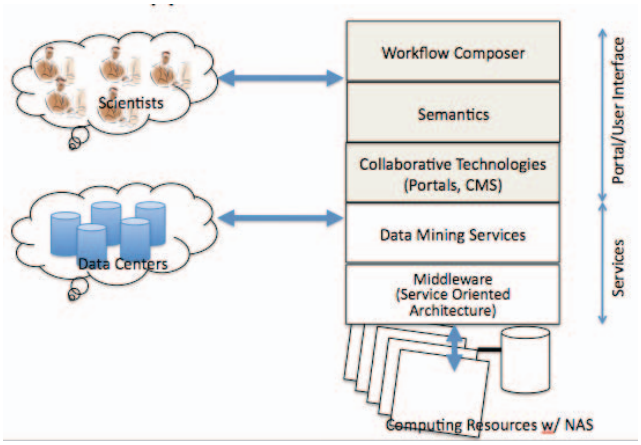

Fig 1: Conceptual architecture for providing Mining as a Service to the science community

References

M. H. Dunham, Data Mining: Introduction and Advanced Topics, Pearson Education, 2003.

U. M. Fayyad, G. Piatetsky-Shapiro and P. Smyth, eds., From Data Mining to Knowledge Discovery: An Overview, The MIT Press, 1996.

R. Ramachandran, S. Graves, J. Rushing, K. Keiser, M. Maskey, H. Lin and H. Conover, ADaM Services: Scientific Data Mining in the Service Oriented Architecture Paradigm, in W. Dubitsky, ed., Data Mining Techniques in Grid Computing Environments, Wiley Publication, 2008.

R. Ramachandran, J. Rushing, X. Li, C. Kamath, H. Conover and S. Graves, Bird's Eye View of Data Mining in Geosciences, in A.K.Sinha, ed., Geoinformatics: Data to Knowledge, Geological Society of America Special Paper 397, 2006.