Distributed Data Mining System with Gateway for Virtual Observatories

N. Balac[1], Olschanowsky[1], H. Karimabadi[2], T. Sipes[2], S. Ferenci[3], R. Chandran[3], R. Fujimoto[3], A. Roberts[4]

[1]SDSC, [2]SciberQuest, Inc., [3]Georgia Tech, [4]Goddard Space Flight Center

natashab@sdsc.edu, roman2u@sdsc.edu, homak@sciberquest.com, tsipes@sciberquest.com, ferenci@cc.gatech.edu, fujimoto@cc.gatech.edu,chandranramesh@gmail.com, aaron.roberts@nasa.gov

Abstract

A common problem facing diverse fields of science and technology is analysis of large, distributed data sets. RemoteMiner is a distributed and parallel data mining software that includes modules for data access, data preparation and mining operation. An overview of RemoteMiner is presented and its application as a tool to extend the capability of virtual space physics observatory from data portal to a science analysis center is discussed.

1. Introduction

With the information age has come a dramatic increase in our ability to generate and collect data while the traditional techniques to analyze this tsunami of data have proven woefully inadequate. It has been estimated that the amount of information in the world doubles every 20 months and the size and number of databases are increasing even faster. Equally important is the growing need to access and analyze data that is on distributed and heterogeneous locations and parties. At the same time, computing and data Grids are emerging as the choice infrastructure for handling large distributed data sets. Here we present an overview of RemoteMiner which was developed at SciberQuest, Inc. in collaboration with Georgia Tech and San Diego Supercomputer Center. RemoteMiner is a general purpose parallel and distributed data mining software. It is currently been adapted for use with virtual observatories in heliophysics. The paper is organized as follows. We find it useful to first discuss the user workflow as a way to illustrate the functionality of RemoteMiner. We then describe the architecture and different elements of RemoteMiner to enable the specified functionality.

2. User Workflow

The following workflow describes the various necessary stages from data access and preparation to performing of data mining and subsequent analysis and saving of workspace. The first step is launching of user interface where the user is prompted to log in with a username and password. Then the user will be able to search for available data sets from local file system, SRB [2] file system (see below), or other external data sources. In case of virtual observatories (VOs), a gateway is provided to pass queries to VOs, and download a selected data set. Next, the user can perform simple analysis of the data to choose the segments of data of interest for further exploration. At this point, any previous results from the analysis of the chosen data set, by either the same user or by others, will be displayed in form of a content-based meta data. Upon completion of the data selection process, the user is ready to perform data mining tasks which can be run on backend computing resources and does not require the user to remain logged in. When the task completes the user is notified. The workspace and data mining work flow are saved and can be made public or sharable to a select group based on user's discretion and permissions set. At any given step, the user can go back to previous steps.

3. Underlying Architecture

Figure 1 illustrates the architecture of RemoteMiner [4]. The user interface (UI) is Java based. Initially we considered browser/cgi-script based interface but chose Java due to its ubiquitous nature and quicker development time. The RemoteMiner server hosts all services including Data Mining Manager (DMM), StreamCache (SC), Resource Manager (PM), External Gateways, and System State (SS). DMM maps data mining tasks to available computing resources, schedules and monitors the execution of those activities. StreamCache efficiently moves data to be mined from storage location to compute nodes using SRB services based on the mapping created by the DMM. Resource Manager monitors the compute nodes and collects information such as load, progress of a task, and network throughput. The information collected

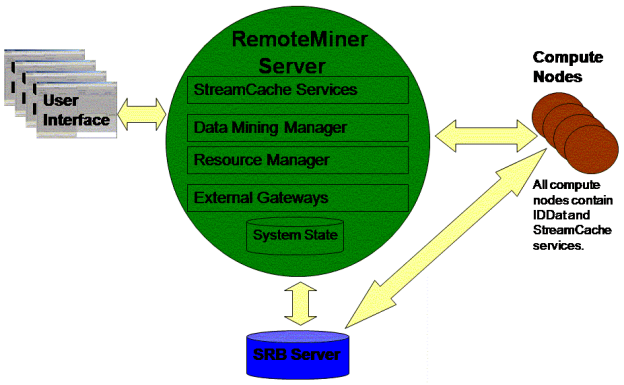

Figure 1. RemoteMiner Architecture

is stored in System State (SS) and is used by the DMM. External Gateways provide access to data sources outside the SRB file system. For example, a gateway can provide access to data stored at a Virtual Observatory (VO). System State stores information of the system.

4. SRB

We have leveraged SRB in the development of RemoteMiner. SRB is a state of the art Data Grid Management System (DGMS) or simply a logical distributed file system based on a client-server architecture which presents the user with a single global logical namespace or file hierarchy. The SRB DGMS has features to support collaborative management of distributed data including: controlled sharing, publication, replication, and transfer, attribute based organization, metadata, data discovery, and preservation of distributed data. The SRB has become a default DGMS for collaborative data management in multiple academic data centers around the world including: US, UK, Taiwan, Australia, Japan, and more countries. It is estimated that SRB brokers more than 1.5 Petabytes of data worldwide.

This client-server or server-server middleware (or DGMS) provides a uniform interface to heterogeneous data resources (UNIX FS, HPSS, UniTree, NTFS, SAM-QFS, DBMS, Disk, etc.) distributed on multiple heterogeneous hosts (UNIX, Linux, AIX, OSX, Windows, etc.). This DGMS The can be used to enable Distributed Logical File Systems, Distributed Digital Libraries, Distributed Persistent Archives, semantic Web and Virtual Object Ring Buffers. It provides its clients with a single logical file hierarchy for easily managing massive distributed data sets. The SRB has scalable library functions that can be utilized by higher-level software. However, it is more complete than many middleware software systems as it implements a comprehensive distributed data management environment, including end-user client applications such as load libraries (Windows, Perl, Python), browsers (web, Windows), digital libraries (DSpace, Fedora, OAI-PMH), workflows (WSDL, Kepler actors, Matrix), and Unix style shell commands.

5. Additional Details

RemoteMiner uses SRB to store data files, meta-data, workspace information, and utilize SRB's file transfer services. Note that SRB is optimized for managing and moving large files (parallel I/O) [1]. Our compute node selection policies will attempt to pair a data mining task with a compute node that is already caching the necessary data. RemoteMiner considers the cost of moving data to the compute node and will attempt to select a node that reduces this cost. RM records metrics regarding previous operations to be used for making future decisions. ScreamCache policies are based on the fact that compute nodes are capable of caching files for future processing. StreamCache services ensure that cache entries are current. Policies determine how files are managed at the compute nodes. Currently a First-In-First-Out FIFO policy is implemented but other policies will be explored. The communication between UI and RemoteMiner server is over socket based TCP/IP connections, i.e. UI connects to open ports on the server.

6. Data Mining Algorithms

Given that some of the data mining operations can be time consuming, especially for high dimensional data, the use of parallel machines can significantly reduce the time it takes to obtain results. RemoteMiner [4] is designed to incorporate a variety of data mining algorithms. In its first release, we used the open source Weka-Parallel [3] (http://weka-parallel.sourceforge.net/) which is the parallel version of the widely used Weka [5] toolkit. Weka which is written in Java includes a large number of machine learning and data mining algorithms originally developed at the University of Waikato in New Zealand. We are also in the process of incorporating a set of new data mining algorithms with analytical capabilities called MineTool [4]. Other techniques will be added in time.

7. Application to Virtual Observatories

Virtual observatories in heliophysics have been primarily designed to address the need to access geographically distributed data with different formats

collected from various instruments. A brief description of VSPO taken from their site (http://vspo.gsfc.nasa.gov) is given here. “Current plans at both at NASA and in the Sun-Earth connection community include making data easily available from all missions relevant to the global problem of the effects of solar particles and fields on the Earth and the solar system (a "Heliophysics Great Observatory"). The VSPO is an evolving system for accomplishing this task. The basic philosophy, shared with the Virtual Solar Observatory and many other such projects, is to register data products from disparate repositories using a common language that allows searching across datasets, retrieving data, and performing analysis and visualization in a uniform way”.

Our goal has been to augment the functionality of VOs by providing data mining capabilities. Accordingly, we have developed a gateway that facilitates performing queries of data on VSPO server. Back-end services and extensions to the GUI have been added to support searching through data located at VSPO. In addition, data access is supported from CDAWeb.

8. GUI

The GUI is written entirely in Java. It contains several tabs (Fig. 2) that enable the user to authenticate themselves, select files from local, SRB, or remote sites, prepare the files for data mining operation, select a particular data mining suite, visualize the results, and save the workspace.

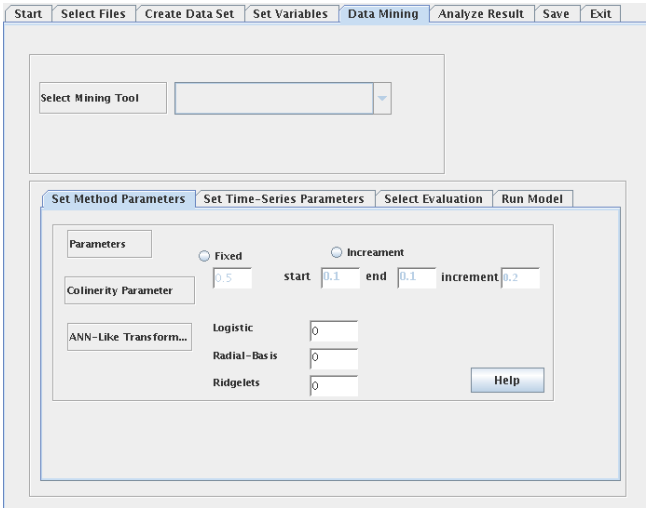

Figure 2. RemoteMiner’s GUI

9. Conclusion

As a result of the dramatic increase of the availability of data from various sources, many scientific disciplines are facing the challenge to collect, manage, share and analyze massive data sets in a networked environment. The existing Cyber-infrastructure is currently not supporting the large data needs of many communities adequately. To address this issue, we have presented RemoteMiner, a next generation data mining and cyber enabled discovery tool. This high-performance, distributed and parallel data mining architecture addresses these challenges in an effective, efficient, collaborative and user friendly manner. An example of this new system’s capacity to extend the capability of virtual space physics observatory from data portal to a science analysis center is presented.

10. References

[1] N. Ali, M. Lauria, “SEMPLAR: high-performance remote parallel I/O over SRB.” CCGRID 2005: 366-373.

[2] C. Baru, R. Moore, A. Rajasekar, M. Wan, "*The SDSC Storage Resource Broker*," Proc. CASCON'98 Conference, Nov.30-Dec.3, 1998, Toronto, Canada.

[3] S. Celis, D.R. Musicant “Weka-Parallel: Machine Learning in Parallel.” Report at http://sourceforge.net/projects/weka-parallel.

[4] H. Karimabadi, T. B. Sipes, H. White, M. Marinucci, A. Dmitriev, J.K. Chao, J. Driscoll, and N. Balac, “Data Mining in Space Physics: 1. The MineTool Algorithm”, in press, J. Geophys. Res., 2007

[5] I. H. Witten, E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, Location, June 2005.