Journal of Advances in Science and Technology

Vol. IV, Issue No. VII, November-2012, ISSN 2230-9659

AN ANALYSIS AROUND THE STUDY OF DISTRIBUTED DATA MINING METHOD IN THE GRID ENVIRONMENT: TECHNIQUE, ALGORITHMS AND SERVICES

AN INTERNATIONALLY INDEXED PEER REVIEWED & REFEREED JOURNAL

www.ignited.in

Journal of Advances in Science and Technology
Vol. IV, Issue No. VII, November-2012, ISSN 2230-9659

An Analysis around the Study of Distributed Data Mining Method in the Grid Environment: Technique, Algorithms and Services

Shoban Babu Sriramoju[1] Dr. Atul Kumar[2]

[1]Research Scholar, CMJ University, Shillong, Meghalaya

[2]Prof. CMJ University, Shillong, Meghalaya

Abstract – Distribution of data and computation allows for solving larger problems and execute applications that are distributed in nature. The Grid is a distributed computing infrastructure that enables coordinated resource sharing within dynamic organizations consisting of individuals, institutions, and resources. The Grid extends the distributed and parallel computing paradigms allowing resource negotiation and dynamical allocation, heterogeneity, open protocols and services. Grid environments can be used both for compute intensive tasks and data intensive applications as they offer resources, services, and data access mechanisms.

Data mining algorithms and knowledge discovery processes are both compute and data intensive, therefore the Grid can offers a computing and data management infrastructure for supporting decentralized and parallel data analysis. This paper discusses how Grid computing can be used to support distributed data mining. Grid-based data mining uses Grids as decentralized high-performance platforms where to execute data mining tasks and knowledge discovery algorithms and applications. Here we outline some research activities in Grid-based data mining, some challenges in this area and sketch some promising future directions for developing Gridbased distributed data mining.

Data mining algorithms are widely used today for the analysis of large corporate and scientific datasets stored in databases and data archives. Industry, science, and commerce fields often need to analyze very large datasets maintained over geographically distributed sites by using the computational power of distributed and parallel systems. The grid can play a significant role in providing an effective computational support for distributed knowledge discovery applications. For the development of data mining applications on grids we designed a system called KNOWLEDGE GRID. This paper describes the KNOWLEDGE GRID framework and presents the toolset provided by the KNOWLEDGE GRID for implementing distributed knowledge discovery.

In many industrial, scientific and commercial applications, it is often necessary to analyze large data sets, maintained over geographically distributed sites, by using the computational power of distributed and parallel systems. The grid can play a significant role in providing an effective computational support for knowledge discovery applications. We describe software architecture for geographically distributed high-performance knowledge discovery applications called Knowledge Grid, which is designed on top of computational grid mechanisms, provided by grid environments such as Globus. The Knowledge Grid uses the basic grid services such as communication, authentication, information, and resource management to build more specific parallel and distributed knowledge discovery tools and services.

Grid computing has emerged as an important new branch of distributed computing focused on large-scale resource sharing and high-performance orientation. In many applications, it is necessary to perform the analysis of very large data sets. The data are often large, geographically distributed and it's complexity is increasing. In these area grid technologies provides effective computational support for applications such as knowledge discovery. This paper is an introduction to Grid infrastructure and its potential for machine learning tasks.

----------♦----------

INTRODUCTION

Today large amounts of data are collected and warehoused. Data sets are generated and stored at enormous speed in local databases, from remote sources or from the sky. At the same time, scientific simulations generating terabytes of data are performed in many laboratories. E-commerce and e-

business applications store and manage huge databases about products, clients and transactions.

Unfortunately, we are much better at storing data than extracting knowledge from it. Large datasets are hard to understand and traditional techniques are infeasible for raw data.

Data mining helps scientists in hypothesis formation in biology, medicine, physics, and engineering. Companies use data mining techniques to provide better, customized services and support decision making. In all these different areas, massive data collections of terabyte and petabyte scale need to be used and analyzed. Moreover, in many cases datasets must be shared by large communities of users that pool their resources from different sites belonging to a single company, or from a large number of laboratories, plants, or public organizations.

Grid computing has been proposed as a novel computational model, distinguished from conventional distributed computing by its focus on large-scale resource sharing, innovative applications, and, in some cases, high-performance orientation. Today grids can be used as effective infrastructures for distributed high-performance computing and data processing. A grid is a geographically distributed computation infrastructure composed of a set of heterogeneous machines that users can access via a single interface. Grids therefore, provide common resource-access technology and operational services across widely distributed virtual organizations composed of institutions or individuals that share resources.

In the last five years, toolkits and software environments for implementing grid applications have become available. These include Legion, Condor, and Unicore. In particular, Foster and Kesselman's Globus Toolkit is the most widely used middleware in scientific and data-intensive grid applications, and is becoming a de facto standard for implementing grid systems. The toolkit addresses security, information discovery, resource and data management, communication, fault detection, and portability issues. It does so through mechanisms, composed as bags of services, that execute operations in grid applications. Today, Globus and the other grid tools are used in many projects worldwide. Although most of these projects are in scientific and technical computing, there is a growing number of grid projects in education, industry, and commerce.

Together with the grid shift toward industry and business applications, a parallel shift toward the implementation of data grids has been registered. Data grids are designed to allow large data sets to be stored in repositories and moved with almost the same ease that small files can be moved. They represent an enhancement of computational grids, driven by the need to handle large data sets without repeated authentication, aiming to support the implementation of distributed data-intensive applications. Significant examples are the EU Data Grid, the Particle Physics Data Grid, the Japanese Grid Data Farm, and the Globus Data Grid project.

Data grid middleware is central for management of data movement and replication on grids. Furthermore, in many scientific and business areas it is necessary to use tools and environments for analysis, inference and discovery over the available data. Scientists and engineers can use those environments for implementing grid-based problem solving environments for doing "virtual" scientific experiments. Analysts can follow the same approach in mining large volumes of data to support decision making. Therefore, the evolution of data grids is represented by knowledge grids offering high-level tools and models for the distributed mining and extraction of knowledge from data repositories available on the grid. The development of such an infrastructure is the main goal of our research work, focused on the design and implementation of an environment for geographically distributed high-performance knowledge discovery applications called KNOWLEDGE GRID.

Grid computing represents the natural evolution of distributed computing and parallel processing technologies. The Grid is a distributed computing infrastructure that enables coordinated resource sharing within dynamic organizations consisting of individuals, institutions, and resources. The main aim of grid computing is to give organizations and application developers the ability to create distributed computing environments that can utilize computing resources on demand.

Grid computing can leverage the computing power of a large numbers of server computers, desktop PCs, clusters and other kind of hardware. Therefore, it can help increase efficiencies and reduce the cost of computing networks by decreasing data processing time and optimizing resources and distributing workloads, thereby allowing users to achieve much faster results on large operations and at lower costs.

Data mining algorithms and knowledge discovery processes are both compute and data intensive, therefore the Grid offers a computing and data management infrastructure for supporting decentralized and parallel data analysis. The opportunity of utilizing grid based data mining systems, algorithms and applications is interesting to users wanting to analyze data distributed across geographically dispersed heterogeneous hosts. Grid-based data mining would allow corporate companies to distribute compute-intensive data analysis among a large number of remote resources. At the same time, it can lead to new algorithms and techniques that would allow organizations to mine data where it are stored. This is in contrast to the practice of having to select data and transfer it into a centralized site for mining. As we know centralized analysis is difficult to perform because data is becoming increasingly

larger, geographically dispersed, and because of security and privacy considerations.

In many scientific and business areas, massive data collections of terabyte and petabyte scale need to be used and analyzed. These huge amounts of data represent a critical resource in several application areas. Moreover, in many cases these data sets must be shared by large communities of users that pool their resources from different sites of a single organization or from a large number of institutions.

Grids are geographically distributed platforms for computation, composed of a set of heterogeneous machines accessible to their users via a single interface. Grid computing has been proposed as an important computational model, distinguished from conventional distributed computing by its focus on large-scale resource sharing, innovative applications, and, in some cases, high-performance orientation.

The main original application area was advanced science and engineering. Recently, grid computing is emerging as an effective paradigm for coordinated resource sharing and problem solving in dynamic, multi-institutional virtual organizations operating in the industry and business arena. Thus, today grids can be used as effective infrastructures for distributed high-performance computing and data processing.

Together with the grid shift towards industry and business applications, a parallel shift toward the implementation of data grids has been registered. Data grids are designed in order to allow large data sets to be stored in repositories and moved about with the same ease that small files can be moved. They represent an enhancement of computational grids, driven by the need to handle large data sets without constant, repeated authentication, aiming to support the implementation of distributed data-intensive applications. Data grids seem to be present largely motivated by the data handling needs of next-generation particle accelerators.

DISTRIBUTED DATA MINING AND GRIDS

Today many organizations, companies, and scientific centers produce and manage large amounts of complex data and information. Climate data, astronomic data and company transaction data are just some examples of massive amounts of digital data repositories that today must be stored and analyzed to find useful knowledge in them. This data and information patrimony can be effectively exploited if it is used as a source to produce knowledge necessary to support decision making. This process is both computationally intensive and collaborative and distributed in nature. Unfortunately, high-level products to support the knowledge discovery and management in distributed environments are lacking. This is particularly true in Grid-based knowledge discovery, although some research and development projects and activities in this area are going to be activated mainly in Europe and USA, such as the Knowledge Grid, the Discovery Net, and the AdAM project. Workflows are mapped on a Grid, assigning its nodes to the Grid hosts and using interconnections for communication among the workflow components (nodes). In the latest years, through the Open Grid Services Architecture (OGSA), the Grid community defined Grid services as an extension of Web services for providing a standard model for using the Grid resources and composing distributed applications as composed of several Grid services. OGSA provides an extensible set of services that virtual organizations can aggregate in various ways defines uniform exposed-service semantics, the so-called Grid service, based on concepts and technologies from both the Grid computing and Web services communities. Recently the Web Service Resource Framework (WSRF) was defined as a standard specification of Grid services for providing interoperability with standard Web services so building a bridge between the Grid and the Web. The creation of Knowledge Grids on top of data and computational Grids is the enabling condition for developing high-performance data mining tasks and knowledge discovery processes and meeting the challenges posed by the increasing demand for power and abstractness coming from complex data mining scenarios in science and engineering. Research projects such as the TeraGrid project and the Grid Data Mining project aim at developing data mining services on Grids, whereas systems like the Knowledge Grid, Discovery Net, and Grid- Miner developed KDD systems for designing complete distributed knowledge discovery processes on grids.

This is particularly true in Grid-based knowledge discovery, although some research and development projects and activities in this area are going to be activated mainly in Europe and USA, such as the Knowledge Grid, the Discovery Net, and the AdAM project. In particular, the Knowledge Grid provides a middleware for knowledge discovery services for a wide range of high performance distributed applications. Examples of large and distributed data sets available today include gene and protein databases, network access and intrusion data, drug features and effects data repositories, astronomy data files, and data about web usage, content, and structure. Knowledge discovery procedures in all these application areas typically require the creation and management of complex, dynamic, multi-step workflows. At each step, data from various sources can be moved, filtered, and integrated and fed into a data mining tool. Based on the output results, the analyst chooses which other data sets and mining components can be integrated in the workflow or how to iterate the process to get a knowledge model.

Workflows are mapped on a Grid assigning its nodes to the Grid hosts and using interconnections for communication among the workflow components (nodes).

In the latest years, through the Open Grid Services Architecture (OGSA), the Grid community defined Grid services as an extension of Web services for providing a standard model for using the Grid resources and composing distributed applications as composed of several Grid services. OGSA provides an extensible set of services that virtual organizations can aggregate in various ways defines a uniform exposed-service semantics, the so-called Grid service, based on concepts and technologies from both the Grid computing and Web services communities. Recently the Web Service Resource Framework (WSRF) was defines as a standard specification of Grid services for providing interoperability with standard Web services so building a bridge between the Grid and the Web.

The development of data mining software for Grids will offer tools and environments to support the process of analysis, inference, and discovery over distributed data available in many scientific and business areas. The creation of Knowledge Grids on top of data and computational Grids is the enabling condition for developing high-performance data mining tasks and knowledge discovery processes and meeting the challenges posed by the increasing demand for power and abstractness coming from complex data mining scenarios in science and engineering. The same can occur in industry and commerce, where analysts need to be able to mine the large volumes of information that can be distributed over different sites to support corporate decision making. The design of distributed data mining in Grids can benefit from the layered Grid architecture, with lower levels providing middleware support for higher level application-specific services.

GRID COMPUTING AS A TECHNIQUE FOR DISTRIBUTED SCENARIO

Today amounts of data are collected and warehoused. Data sets are generated and stored at enormous speed in local databases, from remote sources or from the sky. At the same time, scientific simulations generating terabytes of data are performed in many laboratories. E-commerce and e-business applications store and manage huge databases about products, clients and transactions. Unfortunately, we are much better at storing data than extracting knowledge from it. Large datasets are hard to understand and traditional techniques are infeasible for raw data. Data mining helps scientists in hypothesis formation in biology, medicine, physics, and engineering. Companies use data mining techniques to provide better, customized services and support decision making. In all these different areas, massive data collections of terabyte and petabyte scale need to be used and analyzed. Moreover, in many cases datasets must be shared by large communities of users that pool their resources different sites belonging to a single company, or from a large number of laboratories, plants, or public organizations. Grid computing has been proposed as a novel computational model, distinguished from conventional distributed computing by its focus on large-scale resource sharing, innovative applications, and, in some cases, high-performance orientation. Today grids can be used as effective infrastructures for distributed high-performance computing and data processing. A grid is a geographically distributed computation infrastructure composed of a set of heterogeneous machines that users can access via a single interface. Grids therefore, provide common resource-access technology and operational services across widely distributed virtual organizations composed of institutions or individuals that share resources. Although originally intended for advanced science and engineering applications, grid computing has emerged as a paradigm for coordinated resource sharing and problem solving in dynamic, multi-institutional virtual organizations in industry and business. Grid applications include the following:

• Intensive simulations on remote supercomputers;

• Cooperative visualization of very large scientific data sets;

• distributed processing for computationally demanding data analysis;

• coupling of scientific instruments with remote computers and data archives.

In the last five years, toolkits and software environments for implementing grid applications have become available. These include Legion, Condor, and Unicore. In particular, Foster and Kesselman's Globus Toolkit is the most widely used middleware in scientific and data-intensive grid applications, and is becoming a de facto standard for implementing grid systems. The toolkit addresses security, information discovery, resource and data management, communication, fault detection, and portability issues. It does so through mechanisms, composed as bags of services that execute operations in grid applications. Today, Globus and the other grid tools are used in many projects worldwide. Although most of these projects are in scientific and technical computing, there is a growing number of grid projects in education, industry, and commerce. Together with the grid shift toward industry and business applications, a parallel shift toward the implementation of data grids has been registered.

Data grids are designed to allow large data sets to be stored in repositories and moved with almost the same ease that small files can be moved. They represent an enhancement of computational grids, driven by the need to handle large data sets without repeated authentication, aiming to support the implementation of distributed data-intensive applications. Significant examples are the EU Data

Grid, the Particle Physics Data Grid, the Japanese Grid Data Farm, and the Globus Data Grid project. Data grid middleware is central for management of data movement and replication on grids. Furthermore, in many scientific and business areas it is necessary to use tools and environments for analysis, inference and discovery over the available data. Scientists and engineers can use those environments for implementing grid-based problem solving environments for doing "virtual" scientific experiments. Analysts can follow the same approach in mining large volumes of data to support decision making. Therefore, the evolution of data grids is represented by knowledge grids offering high-level tools and models for the distributed mining and extraction of knowledge from data repositories available on the grid.

THE KNOWLEDGE GRID STRUCTURE

The KNOWLEDGE GRID architecture uses basic grid mechanisms to build specific knowledge discovery services on top of grid toolkits and services. These services can be developed in different ways using the available grid environments. The current implementation is based on the Globus Toolkit. Like Globus, the KNOWLEDGE GRID offers global services based on the cooperation and combination of local services. We designed the KNOWLEDGE GRID architecture so that more specialized data mining tools are compatible with lower-level grid mechanisms and data grid services. This approach benefits from "standard" Grid services that are more and more utilized and offers an open parallel and distributed knowledge discovery architecture that can be configured on top of grid middleware in a simple way.

1. KNOWLEDGE GRID services

The Knowledge Grid services are organized in two hierarchic levels:

• the Core K-grid layer;

• the High level K-grid layer.

The former refers to services directly implemented on the top of generic grid services, the latter is used to describe, develop and execute distributed knowledge discovery computations over the Knowledge Grid. The Knowledge Grid layers are depicted in Fig. 1. The figure shows layers as implemented on the top of Globus services; moreover, the Knowledge Grid data and metadata repositories are also shown. In the following the term K-grid node will denote a Globus node implementing the Knowledge Grid services.

Core K-grid layer- The Core K-grid layer offers the basic services for the definition, composition and execution of a distributed knowledge discovery computation over the grid. Its main goals are the management of all metadata describing features of data sources, third party data mining tools, data management, and data visualization tools and algorithms. Moreover, this layer coordinates the application execution by attempting to fulfill the application requirements and the available grid resources. This layer comprises two main services.

KNOWLEDGE directory service (KDS). This service extends the basic Globus MDS service and it is responsible for maintaining a description of all the data and tools used in the Knowledge Grid.

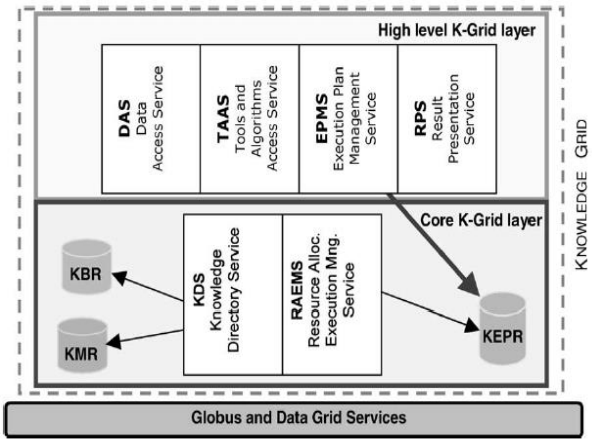

Fig. 1. The Knowledge Grid architecture.

High level K-grid layer - The High level K-grid layer includes services used to compose, validate, and execute a parallel and distributed knowledge discovery computation. Moreover, the layer offers services to store and analyze the discovered knowledge.

2. Globus Toolkit Services

The main services offered by Globus Toolkit 2 are the following:

• Grid security infrastructure (GSI). Enables secure authentication and communication over an open network providing a number of services, including mutual authentication and single sign-on run-anywhere authentication, with support for local control over access rights and mapping from global to local user identities. GSI is based on public key encryption, X.509 certificates, and the secure sockets layer (SSL) communication protocol.

• Monitoring and discovery service (MDS). Provides a framework for publishing and accessing information about grid resources by using the lightweight directory access protocol (LDAP) as a uniform interface to such information. MDS provides two types of directory services: the grid resource information service (GRIS) and the grid index

information service (GIIS). A GRIS can answer queries about the resources of a particular grid node; examples of information provided include host identity (e.g., operating systems and versions), as well as more dynamic information such as current CPU load and memory availability. A GIIS combines the information provided by a set of GRIS services managed by an organization, giving a coherent system image that can be explored or searched by grid applications.

• Globus resource allocation manager (GRAM). Provides facilities for resource allocation and process creation, monitoring, and management. GRAM simplifies the use of remote systems by providing a single standard interface for requesting and using remote system resources for the execution of jobs. The most common use of GRAM is remote job submission and control, to support distributed computing applications.

• Dynamically-updated resource online co-allocator (DUROC). Manages multirequests of resources, delivers requests to different GRAMs and provides time-barrier mechanisms among jobs. In Globus, a GRAM provides an interface to submit jobs on a particular set of physical resources, whereas the DUROC is used to coordinate transactions with independent GRAMs.

• Heartbeat monitor (HBM). Provides a mechanism for monitoring the state of processes. The HBM is designed to detect and report the failure of processes that have identified themselves to the HBM. It allows simultaneous monitoring of both Globus system processes and application processes associated with user computations. The HBM also provides notification of process status exception events, so that recovery actions can be taken.

• Grid FTP. Implements a high-performance, secure data transfer mechanism based on an extension of the FTP protocol that allows parallel data transfer, partial file transfer, and third-party (server-to-server) data transfer, using GSI for authentication. This allows grid applications to have ubiquitous, high-performance access to data in a way that is compatible with the most popular file transfer protocol in use today.

• Replica catalog and replica management. Provide facilities for managing data replicas, i.e., multiple copies of data stored in different systems to improve access across geographically-distributed grids. The replica catalog provides mappings between logical names for files and one or more copies of the files on physical storage systems; it is accessible via an associated library and a command-line tool. The replica management combines the replica catalog (for keeping track of replicated files) and Grid FTP (for moving data) to manage data replication.

3. DDM Using Grid Architecture

A possible infrastructure for a virtual organization, implemented using Grid technologies. The company has a central branch and several local branches (LB). Each branch is composed of a number of Grid nodes (GN) interconnected in a Grid infrastructure. In the case study, the data mining task is the discovery of the association rules in the local branch databases, and the implementation of the Grid infrastructure is based on the Globus toolkit. In the OGSA context, the association rules discovery task is exposed in the form of Grid services.

The mining service has several components specific to a Grid service: service data access, service data element, and service implementation. The association rules discovery service is interacting with the rest of the grid services: service registry, service creation, authorization, notification, manageability and concurrency. There are two types of grid services they are Apriori and Predictive Apriori Algorithms in which the Apriori Grid Service must comply with OGSA rules, constraints, standard interfaces and behavior.

EXECUTION OF GRID DATA MINING APPLICATIONS

The execution plan optimization and translation is performed by means of the RAEMS, whose basic functionalities are provided by the VEGA components and by the scheduler.

Currently, VEGA integrates an RSL generator module, which produces an Resource Specification Language (RSL) script that can be directly submitted to the Globus resource allocation manager (GRAM) of a grid node running the Globus Toolkit. In opposition with the XML execution plan, the RSL script entirely describes an instance of the designed computation, i.e., it specifies all the physical information needed for the execution.

The execution of the computation is performed by means of the VEGA execution manager module. The execution manager allows the system to authenticate a user to the grid, by using the Globus grid security infrastructure (GSI) services, and submits the RSL script to the Globus GRAM for its execution. The execution manager is also responsible of the monitoring of the jobs that compose the overall data mining computation during their life cycle. Finally, the execution manager collects the results of the distributed data mining computation and passes them to the RPS that, in turn, presents them to the user.

At the same time, ad hoc protocols and tools are crucial to perform efficient data transfer along heterogeneous networks having different latencies and bandwidths. To this end, in we proposed a system to enhance the use of the Grid FTP protocol for efficient data transfer on the grid. Such system is based on an algorithm that, on the basis of historical file transfer data, selects the appropriate Grid FTP

parameters for a required transfer session. Among other projects for distributed data analysis, Data Space proposes a significant system to address efficient data access and transfer over the grid.

Data Space is a Web services based infrastructure for exploring, analyzing, and mining remote and distributed data. Data Space applications employ a protocol for working with remote and distributed data called Data Space transfer protocol (DSTP).

DSTP simplifies working with data by providing direct support for common operations, such as working with attributes, keys and metadata. The DSTP protocol can be layered over specialized high performance transport protocols such as SABUL, that allows Data Space applications to effectively work on wide-area high-performance networks.

CONCLUSION

The development of practical grid computing techniques will have a profound impact on the way data is analyzed. In particular, the possibility of utilizing grid based data mining applications is very appealing to organizations wanting to analyze data distributed across geographically dispersed heterogeneous platforms. Grid based data mining would allow companies to distribute compute intensive analytic processing among different resources. Moreover, it might eventually lead to new integration and automated analysis techniques that would allow companies to mine data where it resides. This is in contrast to the current practice of having to extract and move data into a centralized location for mining processes that are becoming more difficult to conduct due to the fact that data is becoming increasingly geographically dispersed, and because of security and privacy considerations.

In this scenario, the Grid can offer an effective infrastructure for deploying data mining and knowledge discovery applications. It can represent in a near future an effective infrastructure for managing very large data sources and providing high-level mechanisms for extracting valuable knowledge from them. To solve this class of tasks, advanced tools and services for knowledge discovery are vital. Here we discussed systems and services for implementing Grid-enabled knowledge discovery services by using dispersed resources connected through a Grid. These services allow professionals and scientists to create and manage complex knowledge discovery applications composed as workflows that integrate data sets and mining tools provided as distributed services on a Grid. They also allow users to store, share, and execute these knowledge discovery workflows as well as publish them as new components and services. As an example of this approach, we described how the Knowledge Grid and the Weka4WS systems provide a higher level of abstraction of the Grid resources for distributed knowledge discovery activities, thus allowing the endusers to concentrate on the knowledge discovery process without worrying about Grid infrastructure details.

Parallel and distributed data mining suites and computational grid systems are two critical elements of future high-performance computing environments for e-science (data-intensive experiments), e-business (distributed online services), and virtual organizations support (virtual teams, virtual enterprises).

The grid infrastructure is growing up very quickly and is going to be more and more complete and complex both in the number of tools and in the variety of supported applications. Along this direction, grid services are shifting from generic computation-oriented services to high-level information management and knowledge discovery services.

Knowledge grids will enable entirely new classes of advanced applications for dealing with the data deluge. Their integration is a challenge whose achievements could produce many benefits in several application areas. Grids are coupling compute-oriented services with data-oriented and high-level information management services. This trend enlarges the grid application scenario and offers opportunities for high-performance distributed knowledge-based systems and services such as data mining and knowledge discovery.

The KNOWLEDGE GRID system we discussed here is a significant component of this trend. It integrates and completes the data grid services by supporting distributed data analysis and knowledge discovery and knowledge management services [43].

REFERENCES

- A. Abraham, R. Buyya, and B. Nath, “Nature’s heuristics for scheduling jobs on computational grids,” in Proc. IEEE 8th Int. Conf. Advanced Computing Communications, 2000.
- Albert Y. Zomaya, Tarek El-Ghazawi, Ophir Frieder, “Parallel and Distributed Computing for Data Mining”, IEEE Concurrency, 1999.
- D. Arnold, H. Casanova, and J. Dongarra, “Innovation of the netsolve grid computing system,” Concur. Comput., 2002.
- F. Berman. From TeraGrid to Knowledge Grid, Communications of the ACM, 44(11), pp. 27–28, 2001.

- H. Dail, H. Casanova, and F. Berman, “AModular Scheduling Approach for Grid Application Development Environments,” UCSD, 2002.

- H. Kargupta and C. Kamath and P. Chan, Distributed and Parallel Data Mining: Emergence, Growth, and Future Directions, In: Advances in Distributed and Parallel Knowledge Discovery, AAAI/MIT Press, pp.409–416, (2000).

- I. Foster, C. Kesselman, S. Tuecke, The anatomy of the grid: enabling scalable virtual organizations, Intl. J. Supercomputer Appl. 15 (3) (2001).

- M. Cannataro, D. Talia, The Knowledge Grid, Communications of the ACM, 46(1), (2003), pp. 89–93.

- Michael D. Beynon, Tahsin Kurc, Alan Sussman, and Joel Saltz. Optimizing execution of component-based applications using group instances. In Proceedings of the Conference on Cluster Computing and the Grid (CCGRID), pp. 56–63, May 2001.

- S. Orlando, P. Palmerini, R. Perego, and F. Silvestri, “Scheduling high performance data mining tasks on a data grid environment,” Europar, 2002.

- Y. Morita et al., “Grid data farm for atlas simulation data challenges,” in Proc. Int. Conf. Computing High Energy Nuclear Physics, 2001, pp. 699–701.

- Zakim J, Pan Y, “Introduction: recent developments in parallel and distributed data mining,” Journal of Distributed Parallel Database, Vol. 11, pp. 123-127, 2002.