Grid-based Distributed Data Mining Systems, Algorithms and Services [*]

Domenico Talia[†]

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 Grid-based distributed data mining.

1 Introduction.

Distributed data mining in distributed environments like virtual organization networks, the Internet, corporate intranets, sensor networks, and other decentralized infrastructures questions the suitability of centralized KDD architectures for large-scale knowledge in a networked environment. Distributed data mining works by analyzing data in a distributed fashion and pays careful attention to the trade-off between centralized collection and distributed analysis of data [7]. When the data sets are large scaling up the speed of the data mining task is crucial. Parallel knowledge discovery techniques address this problem by using high performance multicomputer machines. The increasing availability of such machines calls for extensive development of data analysis algorithms that can scale up as we attempt to analyze data sets measured in terabytes and petabytes on parallel machines with hundreds or thousands of processors. This technology is particularly suitable for applications that typically deal with very large amount of data (e.g., transaction data, scientific simulation and telecom data) that cannot be analyzed on traditional machines in acceptable times. Grid technology integrates both distributed computing and parallel computing, thus it represent a critical infrastructure for high-performance distributed knowledge discovery.

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

[*] Research described in this paper is partially supported by CoreGRID, GRID.IT, and GeoPKDD.it.

[†] DEIS, University of Calabria, Italy.

privacy considerations.

A few research framework currently exists for deploying distributed data mining applications in grids [3]. Some of them are general environments supporting execution of data data mining tasks on machines that belong to a Grid, others are single mining tasks for specific applications that have been "gridfied", and some others are implementations of single data mining algorithms. As the Grid is becoming a well accepted computing infrastructure in science and industry, it is necessary to provide general data mining services, algorithms, and applications that help analysts, scientists, organizations, and professionals to leverage Grid capacity in supporting high-performance distributed computing for solving their data mining problem in a distributed way.

This paper discusses some approaches for exploting Grid computing to support distributed data mining by using Grids as decentralized high-performance platforms where to execute data mining tasks and knowledge discovery algorithms and applications.

2 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 [1], 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 [2] that we shortly discuss in the next section, 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.

In the implementation of data mining systems, algorithms and applications over computational Grids a main issue is the integration of two main demands: synthesizing useful and usable knowledge from data and performing sophisticated large-scale computations leveraging the Grid infrastructure. Such integration must pass through a clear representation of the knowledge base used to translate moderately abstract domain-specific queries into computations and data analysis operations able to answer such queries by operating on the underlying systems. The systems discussed here provide different approaches for supporting knowledge discovery on Grids. Research projects such as the TeraGrid project and the GridDataMining project aim at developing data mining services on Grids, whereas systems like the Knowledge Grid, Discovery Net, and GridMiner developed KDD systems for designing complete distributed knowledge discovery processes on grids.

On the basis of this previous experiences it is necessary to design and implement Grid-based distributed data mining services that leveraging the OGSA and WSRF standards will provide a distributed data mining open service midddleware by which users can design higher level distributed data mining services that cover the main steps of the KDD process and offer typical distributed data mining patterns such as collective learning, ensemble learning, meta-learning, and other concurrent models for composing data mining applications.

3 Grid Services for Distributed Data Mining.

The *Service Oriented Architecture* (*SOA*) is essentially a programming model for building flexible, modular, and interoperable software applications. SOA enables the assembly of applications through parts regardless of their implementation details, deployment location, and initial objective of their development. Another principle of service oriented architectures is, in fact the reuse of software within different applications and processes.

The Grid community adopted the *Open Grid Services Architecture* (*OGSA*) as an implementation of the SOA model within the Grid context. OGSA provides a well-defined set of basic interfaces for the development of interoperable Grid systems and applications [6]. OGSA adopts Web Services as basic technology. Web Services are an important paradigm focusing on simple, Internet-based standards, such as the *Simple Object Access Protocol* (*SOAP*) and the *Web Services Description Language* (*WSDL*), to address heterogeneous distributed computing. Web services defines techniques for describing software components to be accessed, methods for accessing these components, and discovery mechanisms that enable the identification of relevant service providers.

In OGSA every resource (e.g., computer, storage, program) is represented as a *Grid Service*: a Web Service that conforms to a set of conventions and supports standard interfaces. This service-oriented view addresses the need for standard interface definition mechanisms, local and remote transparency, adaptation to local OS services, and uniform service semantics.

OGSA defines standard mechanisms for creating, naming, and discovering transient Grid Service instances; provides location transparency and multiple protocol bindings for service instances; and supports integration with underlying native platform facilities. OGSA also defines mechanisms required for creating and composing sophisticated distributed systems, including lifetime management, change management, and notification. The *WS-Resource Framework* (*WSRF*) was recently proposed as a refactoring and evolution of Grid Services aimed at exploiting new Web Services standards, and at evolving OGSI based on early implementation and application experiences [5].

WSRF provides the means to express state as stateful resources and codifies the relationship between Web Services and stateful resources in terms of the *implied resource pattern*, which is a set of conventions on Web Services technologies, in particular XML, WSDL, and *WS-Addressing.* A stateful resource that participates in the implied resource pattern is termed a *WS-Resource.* The framework describes the WS-Resource definition and association with the description of a Web Service interface, and describes how to make the properties of a WS-Resource accessible through a Web Service interface.

Through WSRF is possible to define basic services for supporting distributed data mining tasks in Grids. Those services can address all the aspects that must be considered in data mining and in knowledge discovery processes from data selection and transport to data analysis, knowledge models representation and visualization. To do this it is necessary to define services corresponding to

- single steps that compose a KDD process such as preprocessing, filtering, and visualization;
- single data mining tasks such as classification, clustering, and rule discovery;
- distributed data mining patterns such as collective learning, parallel classification and meta-learning models;
- data mining applications including all or some of the previous tasks expressed through a multi-step scientific workflows.

This collection of data mining services can constitute an *Open Service Framework for Grid-based Data Mining.* This framework might allow developers to design distributed KDD processes as a composition of single services that are available over a Grid. At the same time, those services should exploit other basic Grid services for data transfer and management such as Reliable File Transfer (RFT), Replica Location Service (RLS), Data Access and Integration (OGSA-DAI) and Distributed Query procesing (OGSA-DQP). Moreover, distributed data mining algorithms can optimize the exchange of data needed to develop global knowledge models based on concurrent mining of remote data sets. This approach also preserves privacy and prevent disclosure of data beyond the original sources. Finally, Grid basic mechanisms for handling security, monitoring, and scheduling distributed tasks can be used to provide efficient implementation of high-performance distributed data analysis.

In the following we describes two systems that have been developed according this service-based approach to develop distributed data mining in grids.

3.1 The Knowledge Grid framework. The Knowledge Grid framework is a system implemented to support the development of distributed KDD processes in a Grid [2]. It uses basic Grid mechanisms to build specific knowledge discovery services. These services can be developed in different ways using the available Grid environments. This approach benefits from "standard" Grid services that are more and more utilized and offers an open distributed knowledge discovery architecture that can be configured on top of Grid middleware in a simple way.

The Knowledge Grid provides users with high-level abstractions and a set of services by which is possible to integrate Grid resources to support all the phases of the knowledge discovery process, as well as basic, related tasks like data management, data mining, and knowledge representation. Therefore, it allows end-users to concentrate on the knowledge discovery process they must develop, without worrying about the Grid infrastructure and its low-level details. The framework supports distributed data mining on the Grid by providing mechanisms and higher level services for searching resources, representing, creating, and managing knowledge discovery processes, and for composing existing data services and data mining services as structured, compound services, so as to allow users to plan, store, document, verify, share and (re-)execute their applications, as well as manage their output results.

We are developing the Knowledge Grid in terms of the OGSA model. In this implementation, each Knowledge Grid service (*K-Grid service*) is exposed as a Web Service that exports one or more operations (*OPs*), by using the WSRF conventions and mechanisms. The operations exported by high-level K-Grid services (data access services (DAS), tools and algorithms access services (TAAS), execution plan management services (EPMS), and result presentation services (RPS)) are designed to be invoked by user-level applications, whereas operations provided by core K-Grid services (knowledge directory services (KDS) and resource access and execution services (RAEMS)) are thought to be invoked by high-level and core K-Grid services (see figure 1).

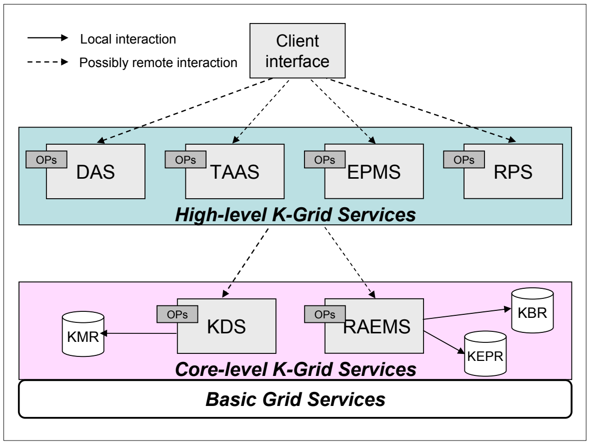

Figure 1: Interactions between a client and the Knowledge Grid environment.

In the WSRF-based implementation of the Knowledge Grid each service is exposed as a Web Service that exports one or more operations (OPs), by using the WSRF conventions and mechanisms. The operations exported by High-level K-Grid services are designed to be invoked by user-level applications only, whereas the operations provided by Core K-Grid services are thought to be invoked by High-level as well as Core K-Grid services.

Users can access the Knowledge Grid functionalities by using a client interface located on their machine. The client interface can be an integrated visual environment that allows for performing basic tasks (e.g., searching of data and software, data transfers, simple job executions), as well as for composing distributed data mining applications described by arbitrarily complex execution plans. The client interface performs its tasks by invoking the appropriate operations provided by the different High-level K-Grid services. Those services may be in general executed on a different Grid node; therefore the interactions between the client interface and High-level K-Grid services are possibly remote.

3.2 Weka4WS. Weka4WS is a framework that extends the widely used open source Weka toolkit [9] for supporting distributed data mining on WSRF-enabled Grids. Weka4WS adopts the WSRF technology for running remote data mining algorithms and managing distributed computations. The Weka4WS user interface supports the execution of both local and remote data mining tasks. On every computing node, a WSRF-compliant Web Service is used to expose all the data mining algorithms provided by the Weka library.

The Weka4WS software prototype has been developed by using the Java WSRF library provided by

Globus Toolkit (GT4). All involved Grid nodes in Weka4WS applications use the GT4 services for standard Grid functionality, such as security, data management, and so on. We distinguish those nodes in two categories on the basis of the available Weka4WS components: *user nodes* that are the local machines providing the Weka4WS client software; and *computing nodes* that provide the Weka4WS Web Services allowing for the execution of remote data mining tasks. Data can be located on computing nodes, user nodes, or third-party nodes (e.g., shared data repositories). If the dataset to be mined is not available on a computing node, it can be uploaded by means of the GT4 data management services.

Figure 2 shows the software components of user nodes and computing nodes in the Weka4WS framework. User nodes include three components: *Graphical User Interface* (*GUI*), *Client Module* (*CM*), and *Weka Library* (*WL*). The GUI is an extended Weka Explorer environment that supports the execution of both local and remote data mining tasks. Local tasks are executed by directly invoking the local WL, whereas remote tasks are executed through the CM, which operates as an intermediary between the GUI and Web Services on remote computing nodes.

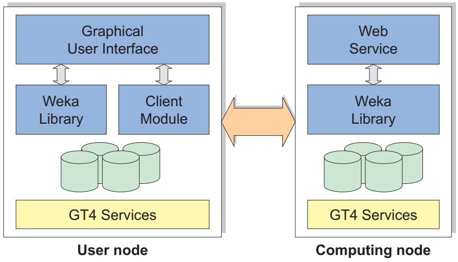

Figure 2: Software components of user nodes and computing nodes.

Through the GUI a user can either: *i*) start the execution locally by using the *Local* pane; *ii*) start the execution remotely by using the *Remote* pane. Each task in the GUI is managed by an independent thread. Therefore, a user can start multiple distributed data mining tasks in parallel on different Web Services, this way taking full advantage of the distributed Grid environment. Whenever the output of a data mining task has been received from a remote computing node, it is visualized in the standard *Output* pane. A recent paper [8] presents the architecture, details of user interface, and performance analysis of Weka4WS in executing a distributed data mining task in different network scenarios.

The experimental results demonstrate the low overhead of the WSRF Web service invocation mechanisms with respect to the execution time of data mining algorithms on large data sets and the efficiency of the WSRF framework as a means for executing data mining tasks on remote resources. By exploiting such mechanisms, Weka4WS provides an effective way to perform compute-intensive distributed data analysis on large-scale Grid environments. Weka4WS can be downloaded from http://grid.deis.unical.it/weka4ws.

4 Some Concluding Remarks.

Many experts in IT, science, finance and commerce are recognizing the importance of scalable data mining solutions in their business. From the Bill Gates' keynote speech at Supercomputing 05: *... And so now we need to take the techniques that have been developed for things like business intelligence and data mining that goes on around that and think how we can apply those in these realms as well, how we can take every step of the process and have it be very visual and only require as much software understanding as is absolutely necessary.* we can conclude that the importance of high-performance data mining is going to be considered a real added value.

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 end-users to concentrate on the knowledge discovery process without worrying about Grid infrastructure details.

In the next years the Grid will be used as a platform for implementing and deploying geographically distributed knowledge discovery and knowledge management services and applications. Some ongoing efforts in this direction have been recently started. Example of systems such as the Discovery Net, the AdAM system, and the Knowledge Grid discussed here show the feasibility of the approach and can represent the first generation of knowledge-based pervasive Grids. The future use of the Grid is mainly related to its ability embody many of those properties and to manage world-wide complex distributed applications. Among those, knowledge-based applications are a major goal. To reach this goal, the Grid needs to evolve towards an open decentralized infrastructure based on interoperable high-level services that make use of knowledge both in providing resources and in giving results to end users [4]. Software technologies for the implementation and deployment of knowledge Grids as we discussed in this paper will provide important elements to build up knowledge-based applications on a local Grids or on a World Wide Grid. These models, techniques, and tools can provide the basic components for developing Grid-based complex systems such as distributed knowledge management systems providing pervasive access, adaptivity, and high performance for virtual organizations in science, engineering and industry that need to produce knowledge-based applications.

5 Acknowledgements

I would like to thanks Antonio Congiusta and Paolo Trunfio for their work on some systems cited in this paper.

References

[1] F. Berman. *From TeraGrid to Knowledge Grid*, Communications of the ACM, 44(11), pp. 27–28, 2001.

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

[3] M. Cannataro, A. Congiusta, C. Mastroianni, A. Pugliese, D. Talia, P. Trunfio, *Grid-Based Data Mining and Knowledge Discovery*, In: Intelligent Technologies for Information Analysis, N. Zhong and J. Liu (eds.), Springer-Verlag, chapt. 2 (2004), pp. 19–45.

[4] M. Cannataro, D. Talia, *Semantics and Knowledge Grids: Building the Next-Generation Grid*, IEEE Intelligent Systems, 19(1), (2004), pp. 56–63.

[5] K. Czajkowski et al., The WS-Resource Framework Version 1.0. http://www-106.ibm.com/developerworks/library/ws-resource/ws-wsrf.pdf.

[6] I. Foster, C. Kesselman, J. Nick, and S. Tuecke, *The Physiology of the Grid*, In: F. Berman, G. Fox, and A. Hey (eds.), Grid Computing: Making the Global Infrastructure a Reality, Wiley, pp. 217–249, (2003).

[7] 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).

[8] D. Talia, P. Trunfio, O. Verta. *Weka4WS: a WSRF-enabled Weka Toolkit for Distributed Data Mining on Grids*. Proc. PKDD 2005), Porto, Portugal, October 2005, LNAI vol. 3721, pp. 309–320, Springer-Verlag, 2005.

[9] H. Witten and E. Frank. *Data Mining: Practical machine learning tools with Java implementations*. Morgan Kaufmann, 2000.