How Distributed Data Mining Tasks can Thrive as Services on Grids

Domenico Talia and Paolo Trunfio
DEIS, Università della Calabria, Italy
{talia, trunfio}@deis.unical.it

Abstract

Through a service-based approach it is possible to define services for supporting distributed and pervasive business intelligence applications in Grids. Those services can address all the tasks needed in data mining and in knowledge discovery processes starting from data selection and transport, to data analysis, knowledge models representation and visualization. By exploiting the Grid services features it is possible to develop data mining services accessible every time and everywhere by providing a sort of knowledge discovery ecosystem formed of a large numbers of decentralized data analysis services. We worked in this direction by providing Grid-based architectures and services for distributed and pervasive knowledge discovery. This paper discusses how Grid frameworks can be developed as a collection of Grid services and how they can be used to develop distributed data analysis tasks and knowledge discovery processes using the SOA model.

1. Introduction

Computer science applications are becoming more and more network centric, ubiquitous, knowledge intensive, and computing demanding. This trend will result soon in an ecosystem of pervasive applications and services that professionals and end-users can exploit everywhere. A long term perspective can be envisioned where a collection of services and applications will be accessed and used as public utilities, like water, gas and electricity are used today.

Key technologies for implementing that perspective are SOA and Web services, semantic Web and ontologies, pervasive computing, P2P systems, Grid computing, ambient intelligence architectures, data mining and knowledge discovery tools, Web 2.0 facilities, mashup tools, and decentralized programming models. In fact, it is mandatory to develop solutions that integrate some or many of those technologies to provide future knowledge-intensive software utilities. The Grid can represent a future cyber infrastructure for efficiently supporting that model.

In the area of Grid computing a proposed approach in accordance with the trend outlined above is the Service-Oriented Knowledge Utilities (SOKU) model [1] that envisions the integrated use of a set of technologies that are considered as a solution to information, knowledge and communication needs of many knowledge-based industrial and business applications. The SOKU approach stems from the necessity of providing knowledge and processing capabilities to everybody, thus supporting the advent of a competitive knowledge-based economy. Although the SOKU model is not yet implemented, Grids are increasingly equipped with data management tools, semantic technologies, complex workflows, data mining features and other Web intelligence approaches. These technologies can facilitate the process of having Grids as a strategic component for pervasive knowledge intensive applications and utilities.

Grids were originally designed for dealing with problems involving large amounts of data and/or compute-intensive applications. Today, however, Grids enlarged their horizon as they are going to run business applications supporting consumers and end users [2]. To face those new challenges, Grid environments must support adaptive data management and data analysis applications by offering resources, services, and decentralized data access mechanisms. In particular, according to the service oriented architecture (SOA) model, data mining tasks and knowledge discovery processes can be delivered as services in Grid-based infrastructures.

Through a service-based approach we can define integrated services for supporting distributed business intelligence tasks in Grids. Those services can address all the aspects that must be considered in data mining and in knowledge discovery processes such as data selection and transport, data analysis, knowledge models representation and visualization. We worked in this direction for providing Grid based architectures and services for distributed knowledge discovery such as the *Knowledge Grid* [3, 4], the *Weka4WS* toolkit [5], and mobile Grid services for data mining [6].

This paper discusses how Grid frameworks such those mentioned above can be developed as a collection of Grid services and how they can be used to develop distributed data analysis tasks and knowledge discovery processes using the SOA model.

The rest of the chapter is organized as follows. Section 2 introduces Grids as new infrastructures for running data mining applications. Section 3 discusses a strategy based on the use of Grid services for the design of distributed

knowledge discovery services. Section 4 shortly describes the services implemented in the Knowledge Grid, the Weka4WS and the mobile Grid service frameworks. Finally, Section 5 concludes the paper.

2. On Grids and data mining

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 to increase the efficiency 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 applications demand for both compute and data management facilities. Therefore the Grid is a good candidate offering 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. For example, 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 is stored. This is in contrast to the practice of selecting data and transferring 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.

A few research frameworks currently exist for deploying distributed data mining applications in Grids [7]. Some of them are general environments supporting execution of 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.

The Grid community has adopted the *Open Grid Services Architecture (OGSA)* as an implementation of the SOA model within the Grid context. In OGSA every resource is represented as a Web service that conforms to a set of conventions and supports standard interfaces. OGSA provides a well defined set of Web service interfaces for the development of interoperable Grid systems and applications. Recently the WS-Resource Framework (*WSRF*) has been adopted as an evolution of early OGSA implementations. WSRF defines a family of technical specifications for accessing and managing stateful resources using Web services. The composition of a Web service and a stateful resource is termed as *WS-Resource*. The possibility to define a state associated to a service is the most important difference between WSRF compliant Web services, and pre-WSRF ones. This is a key feature in designing Grid applications, since WS-Resources provide a way to represent, advertise, and access properties related to both computational resources and applications.

3. Data mining Grid services

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 knowledge discovery processes from data selection and transport, to data analysis, knowledge model representation and visualization. This can be done by designing services corresponding to

- ***single steps*** that compose a KDD process such as pre-processing, 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 can 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 Processing* (*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 datasets.

This approach also preserves privacy and prevents disclosure of data beyond the original sources. Finally, basic Grid mechanisms for handling security, trustiness, monitoring, and scheduling distributed tasks can be used to provide efficient implementation of high-performance distributed data analysis.

4. Grid Service-based data mining frameworks

After introducing the service-oriented approach for the implementation of distributed data mining on Grids, in the rest of the paper, we shortly describe three systems that we developed according to that service-based model. Those systems show the feasibility of the proposed approach.

4.1. Weka4WS

Weka4WS is a framework that extends the widely used open source Weka toolkit 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 a Grid 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*) [8]. 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 is automatically uploaded by means of the GT4 data management services. 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. Through the GUI, a user can start the execution either locally or on a (generic or specific) remote Grid node (see Figure 1). 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 computing nodes, 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 a pane of the GUI.

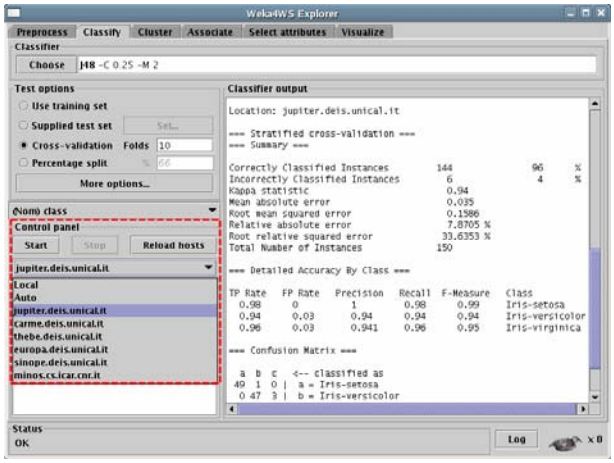

Figure 1. The Weka4WS GUI. The red box shows the panel that allows selecting a Grid node where to run the data mining task.

A recent paper [9] presents a performance analysis of the execution mechanisms described above. 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 datasets, and confirms the efficiency of the WSRF framework as a means for executing data mining tasks on remote resources. By exploiting such mechanisms, Weka4WS is an open source system that provides an effective way to perform compute-intensive distributed data analysis on large-scale Grid environments (it can be downloaded from grid.deis.unical.it/weka4ws).

4.2. The KNOWLEDGE GRID

The *Knowledge Grid* [3] is a Grid services-based environment providing knowledge discovery services for a wide range of high performance distributed applications. It offers users high-level abstractions and a set of services by which they can integrate Grid resources to support all the phases of the knowledge discovery process.

The Knowledge Grid supports such activities by providing mechanisms and higher level services for searching resources (data, algorithms, etc.), representing, creating, and managing knowledge discovery processes, and for composing existing data services and data mining services in a structured manner, allowing designers to plan, store, document, verify, share and re-execute their workflows as well as manage their output results.
The Knowledge Grid architecture is composed of a set of services divided in two layers: the *Core K-Grid layer* and the *High-level K-Grid layer*. The first layer interfaces the basic and generic Grid middleware services, while the second layer interfaces the user by offering a set of services for the design and execution of knowledge discovery applications. Both layers make use of repositories that provide information about resource metadata, execution plans, and knowledge obtained as result of knowledge discovery applications.
In the Knowledge Grid environment, discovery processes are represented as workflows that a user may compose using both concrete and abstract Grid resources. Knowledge discovery workflows are defined using a visual interface that shows resources (data, tools, and hosts) to the user and offers mechanisms for integrating them in a workflow. Information about single resources and workflows are stored using an XML-based notation that represents a workflow (called execution plan in the Knowledge Grid terminology) as a data-flow graph of nodes, each one representing either a data mining service or a data transfer service. The XML representation allows the workflows for discovery processes to be easily validated, shared, translated into executable scripts, and stored for future executions.

4.3. Mobile data mining services

The availability of client programs on mobile devices that can invoke the remote execution of data mining tasks and show the mining results is a significant added value for nomadic users and organizations that need to perform analysis of data stored in repositories far away from the site where users are working, allowing them to generate knowledge regardless of their physical location.
This section shortly discusses pervasive data mining of databases from mobile devices through the use of Grid Services. By implementing mobile Grid Services we allow
remote users to execute data mining tasks on a Grid from a mobile phone or a PDA and receive on those devices the results of a data analysis task.
The system is based on the client/server architecture shown in Figure 2. The architecture includes three types of components:

- *Data providers*: applications that generate the data to be mined.
- *Mobile clients*: the applications that require the execution of data mining computations on remote data.
- *Mining servers*: server nodes used for storing the data generated by data providers and for executing the data mining tasks submitted by mobile clients.

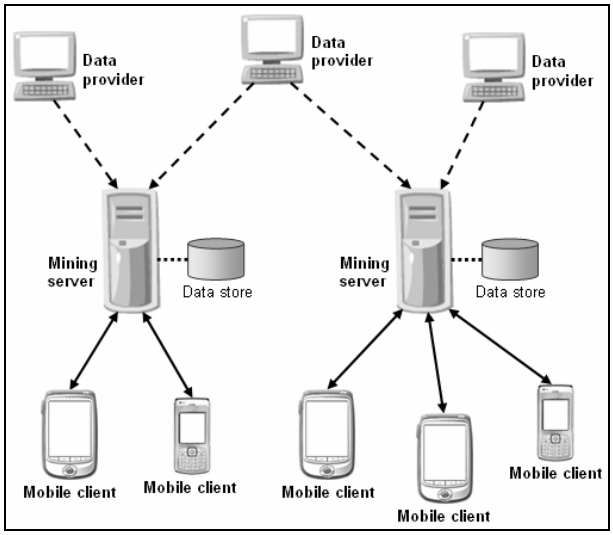

Figure 2. General architecture of the system.

Each Mining server exposes its functionalities through two Web services: the *Data Collection Service* (*DCS*) and the *Data Mining Service* (*DMS*). The DCS is invoked by data providers to store data on the server. The DMS is invoked by mobile clients to perform data mining tasks. Its interface defines a set of operations that allow to: obtaining the list of the available data sets and algorithms, submitting a data mining task, getting the current status of a computation, and getting the result of a given task.
The DMS can perform several data mining tasks from a subset of the algorithms provided by the Weka4WS systems. When a data mining task is submitted to the DMS, the appropriate algorithm of the Weka library is invoked on a Grid node to analyze the local data set specified by the mobile client.
The mobile client is composed by three components: the *MIDlet*, the *DMS Stub*, and the *Record Management System* (*RMS*). The MIDlet is a J2ME application allowing the user to perform data mining operations and visualize their results. The DMS Stub is a WSRF Service stub allowing the MIDlet to invoke the operations of a remote DMS. Even if the DMS Stub and the MIDlet are two logically separated components, they are distributed and installed as a single J2ME application. The RMS is a simple record-oriented database that allows J2ME applications to persistently store data across multiple invocations. In our system, the MIDlet uses the RMS to store the URLs of the remote DMSs that can be invoked

by the user. The list of URLs stored in the RMS can be updated by the user using a MIDlet functionality.
The small size of the screen is one of the main limitations of mobile device applications. In data mining tasks, in particular, a limited screen size can affect the appropriate visualization of complex results representing the discovered model. In our system we overcome this limitation by splitting the result in different parts and allowing a user to select which part to visualize at one time. Moreover, users can choose to visualize the mining model (e.g., a cluster assignment or a decision tree) either in textual form or as an image. In both cases, if the information does not fit the screen size, the user can scroll it by using the normal navigation facilities of the mobile device.
As an example, Figure 3 shows two screenshots of the mobile client taken from a test application. The screenshot on the left shows the menu for selecting which part of the result of a classification task must be visualized, while the screenshot on the right shows the result, in that case the pruned tree resulting from classification.

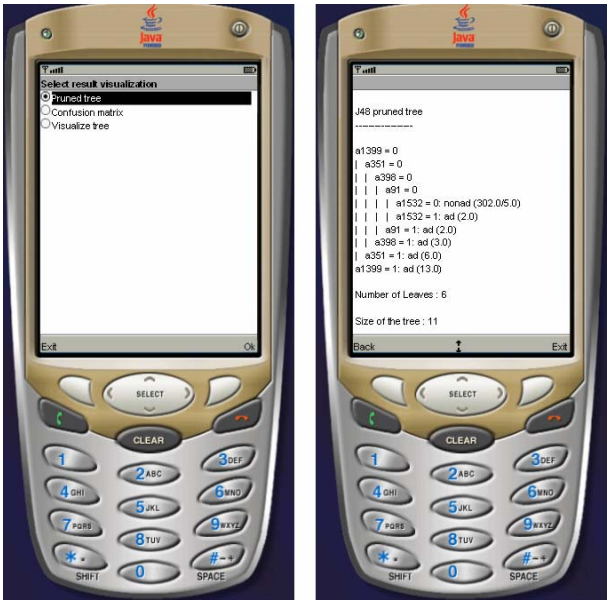

Figure 3. Two screenshots of the client applications running on the emulator of the Sun Java Wireless Toolkit.

Our early experiments show that the system performance depends almost entirely on the computing power of the server on which the data mining task is executed. On the contrary, the overhead due to the communication between MIDlet and Data Mining Service does not affect the execution time in a significantly way, since the amount of data exchanged between client and server is very small. In general, when the data mining task is relatively time consuming, the communication overhead is a negligible percentage of the overall execution time.

5. Final remarks

The main thesis of this paper is that the Grid can be used as an effective cyber infrastructure for implementing and deploying geographically distributed data mining and knowledge discovery services and applications. Future uses of the Grid are mainly related to the ability to utilize it as a knowledge-oriented platform able to run world-wide complex distributed applications. Among those, knowledge discovery applications are a major goal. To reach this goal, the Grid needs to evolve towards an open decentralized platform based on interoperable high-level services that make use of knowledge both in providing resources and in giving results to end users [10].
Software frameworks and technologies for the implementation and deployment of knowledge services, as those we discussed in this paper, provide key elements to build up data analysis applications on enterprise or campus Grids or on a World Wide Grid. Those models, techniques, and tools can be instrumented in Grids as decentralized and interoperable services that enable the development of complex systems such as distributed knowledge discovery suites and knowledge management systems offering pervasive access, adaptivity, and high performance to single users, professional teams, and virtual organizations in science, engineering and industry that need to create and use knowledge-based applications.

Acknowledgements

This research work is partially carried out under the FP6 Network of Excellence Core-GRID funded by the European Commission (Contract IST-2002-004265) and the TOCAI.IT project funded by MIUR.

6. References

[1] NGG3 Expert Group, “Strategic Future for European Grids: Next Generation GRIDs based on SOKU A new paradigm for service delivery and software infrastructure, Bruxelles, December 2005.

[2] Cannataro M., Talia D., “Semantics and Knowledge Grids: Building the NextGeneration Grid”, *IEEE Intelligent Systems*, 19(1), 2004, pp. 56–63.

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

[4] Cannataro M., Talia D., Trunfio P., “Design of Distributed Data Mining Applications on the Knowledge Grid”. In: *Data Mining: Next Generation Challenges and Future Directions*, H.

Kargupta, A. Joshi, K. Sivakumar, Y. Yesha (eds.), MIT Press, Menlo Park, California, 2004, pp. 67-88.

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

[6] Talia D., Trunfio P., "Mobile Data Mining on Small Devices Through Web Services". In: *Mobile Intelligence: When Computational Intelligence Meets Mobile Paradigm*, L. Yang, A. Waluyo, J. Ma, L. Tan, B. Srinivasan (eds.), John Wiley & Sons, 2007, to appear.

[7] Cannataro M., Congiusta A., Mastroianni C, Pugliese A., Talia D., Trunfio P., “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.

[8] Foster I., “Globus Toolkit Version 4: Software for Service-Oriented Systems”, *Proc. Conference on Network and Parallel Computing (NPC 2005)*, LNCS 3779, Springer-Verlag, 2005, pp. 2-13.

[9] Talia D., Trunfio P., Verta O., "WSRF Services for Composing Distributed Data Mining Applications on Grids: Functionality and Performance", *Proc. of the International Conference on Computational Science and its Applications (ICCSA 2006)*, Glasgow, UK, LNCS, vol. 3980, Springer-Verlag, May 2006, pp. 1080-1089.

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