Chapter XVII
Using Grids for Distributed Knowledge Discovery

Antonio Congiusta
DEIS—University of Calabria, Italy

Domenico Talia
DEIS—University of Calabria, Italy

Paolo Trunfio
DEIS—University of Calabria, Italy

ABSTRACT

Knowledge discovery is a compute- and data-intensive process that allows for finding patterns, trends, and models in large datasets. The grid can be effectively exploited for deploying knowledge discovery applications because of the high performance it can offer and its distributed infrastructure. For effective use of grids in knowledge discovery, the development of middleware is critical to support data management, data transfer, data mining and knowledge representation. To such purpose, we designed the Knowledge Grid, a high-level environment providing for grid-based knowledge discovery tools and services. Such services allow users to create and manage complex knowledge discovery applications, composed as workflows that integrate data sources and data-mining tools provided as distributed grid services. This chapter describes the Knowledge Grid architecture and describes how its components can be used to design and implement distributed knowledge discovery applications. Then, the chapter describes how the Knowledge Grid services can be made accessible using the open grid services architecture (OGSA) model.

INTRODUCTION

Knowledge discovery in databases (KDD) is often both a compute- and data-intensive process. When large datasets are coupled with geographic distribution of data, users, and systems, a variety of technologies must be combined for implementing high-performance distributed knowledge dis-

covery systems. Most of the current off-the-shelf KDD environments require central aggregation of data that, in many cases, is distributed. Data storage in a single site may not always be feasible because of limited network bandwidth, security concerns, scalability problems, and other practical issues.

Data mining in large settings like virtual organization networks, the Internet, corporate intranets, sensor networks, and the emerging world of ubiquitous computing, questions the suitability of centralized KDD architectures for large-scale knowledge discovery in a networked environment. The field of distributed KDD offers an alternative approach. It works by analyzing data in a distributed fashion, and pays particular attention to the trade-off between centralized collection and distributed analysis of data.

When the datasets are large, scaling up the speed of the KDD process is a crucial issue. Distributed knowledge discovery techniques address this problem by using high-performance multicomputer machines and a decentralized approach for mining large datasets that can be used when several interconnected machines are available for running distributed data-mining models. The increasing availability of such machines and networks calls for extensive development of data-analysis algorithms able to scale with datasets, measured in terabytes and petabytes, on distributed and parallel machines with hundreds or thousands of processors. Knowledge discovery is speeded up by executing, in a distributed way, a number of data mining processes on different data subsets, and then combining the results through metalearning. This technology is particularly suitable for applications that typically deal with very large amounts of data (e.g., transaction data, scientific simulation, and telecommunication data) that cannot be analyzed in a single site on traditional machines in acceptable times. Moreover, parallel data-mining algorithms can be a component of distributed data-mining applications that can exploit either parallelism or data distribution.

Grid technology integrates both distributed and parallel computing; thus, it represents a critical infrastructure for high-performance distributed knowledge discovery. Grid computing is receiving increasing attention both from the research community and from industry and governments, looking to this new computing infrastructure as a key technology for solving complex problems and implementing distributed high-performance applications (Foster, Kesselman, Nick, & Tuecke, 2002). Today there is a large number and variety of grid tools and middleware that allow the user community to use grids for implementing a larger set of applications, with respect to 1 or 2 years ago.

The term "grid" defines a global distributed computing platform through which—like in a power grid—users gain ubiquitous access to a range of services, computing, and data resources. The driving grid applications are traditional high-performance applications, such as high-energy particle physics, and astronomy and environmental modeling, in which experimental devices create large quantities of data that require scientific analysis.

Grid computing differs from conventional distributed computing because it focuses on large-scale resource sharing, offers innovative applications, and, in some cases, it is geared toward high-performance systems. 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. Therefore, today's grids can be used as effective infrastructures for distributed high-performance computing and data processing. Grid applications include:

- Intensive simulations on remote supercomputers.