Data Mining in the Grid Environment

Martin Sarnovský

Katedra Kybernetiky a Umelej Inteligencie, Technická Univerzita, Letná 9, 040 01 Košice, Slovakia

Martin.Sarnovsky@tuke.sk

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

Keywords:Grid Computing, Knowledge Grid, Data mining, Distributed data mining

1 Introduction

Grid is a technology, that allows from geographically distributed computational and memory resources create an ***universal computing system*** with extreme performance and capacity. System has the features of global worldwide computer, where all of the the componenets are connected together via Internet. For the user it appears like a common workstation, but some segments of a solvedj task are computed in different parts of system. The main advantage of the grid is high efficiency of using technological capacity.

The term „the Grid“ was first used in the 1990 to denote a proposed distributed computing infrastructure for advanced science and engineering. The Grid means coordinated resource sharing and problem sloving in dynamic, multi-institutional virtual organizations. The sharing means direct access to computers, software, data and resources. Another definition of the Grid is „*Grid is the hardware and software infrastructure that provides dependable, consistent, pervasive and inexpensive access to high-end computational capabilities*“ [1]. Sharing is highly controlled, clearly defining what is shared, who is allowed to share, and the conditions under which sharing occurs. Individuals or institutions defined by sharing rules form are *virtual organizations*.

The main building block of the Grid is a network. Geographically distributed resources are linked together via networks. Networks also allow them to be used collectively. Networks connect the resources on the Grid, the most prevalent of which are computers with data storage. Computational elements can be on any level of power and capability. Some of Grids involve nodes that are high-performance machines or clusters. These Grid nodes provides major resources for simulation, analysis, datamining and another activities.

2 The Evolution of the Grids

During the 1980s and 1990s researchers from multiple disciplines began to come together to solve problems for which large-scale computational infrastructure is a key tool to achieve results. The problems inherent in conducting multidisciplinary and often geographically dipersed collaborations provided researchers experience with *coordination* and *distribution* – two fundamental concepts of Grid computing [2].

During the middle 90s early Grids started as the projects linking supercomputing sites to provide resources to a range of high-performance applications. For example FAFNER (Factoring via Network-Enabled Recursion) was developed for factoring large prime numbers [3]. Contributors supported the project by providing computational power of their web servers. The calculations were done by a CGI script at the respective server. This project was a precursor of projects like SETI(at)home.

The first modern Grid is generally considered to be the I-WAY [4]. The I-WAY project started in 1995 was experimental high-performance network linking numerous high-end computers and advanced visualization environments. The main idea was to unify resources at supercomputing centers. Over 17 sites were networked together, and over 60 applications were developed and deployed on the I-WAY as well as a Grid infrastructure to provide access, security, coordinate resources and other activities. I-WAY is important because it was precursor of Globus project which is at present standard for developing of the grid applications. The I-WAY project linked a numbers of US national supercomputing sites. But now the grid infrastucture allows to couple more than just a few specialized supercomputing centers. Grid is now more ubiquitious because of using the definitions and standards.

In the late 1990s, Grid researchers came together in the Grid Forum, subsequently expanding into the Global Grid Forum, where much of research is now evolving into the standards for future Grids like Open Grids Services Architecture, which integrates Globus and Web services approaches.

2.1 Open Grid Service Architecture

Open Grid Service Architecture (OGSA) defines the blueprint of how a Grid should look appear, including the infrastructure. It also defines the programming model of Grid services. It provides instructions onn how to build a Grid service by outlining the required components beeded to build and deliver Grid solution.

OGSA itself is described by the GGF as follows: “Successful realization of the OGSA vision of a broadly applicable and adopted framework for distributed system integration requires the early definition of a core set of interfaces, behaviors, resource models, bindings and so forth: what we call the OGSA Platform” [5].

2.2 Open Grid Services Infrastructure (OGSI)

OGSI is the specification of OGSA infrastructure. It is the middleware for Grid services. It defines how to build the Grid service, outlining the mechanisms for creating, managing, and exchanging information for Grid services.

2.3 Globus Toolkit

Globus provides software architecture that enables to handle and manage distributed and heterogeneous resources. It is an open source product that provides the services and also C/C++ libraries [6] is designed to develop the Grid applications. The toolkit includes software services and libraries for resource monitoring, discovery, and management, plus security and file management. Globus provides a layered architecture and is de facto standard for developing grid applications.

- Grid Resource Allocation Manager (GRAM) is used for allocation, monitoring and control of computational resources
- GridFTP is extansion of FTP, it contains features like management for paralelism for high-speed transfers etc.
- Grid Service Infrastructure – supplies authentification and security services
- Lightweight Directory Access Protocol – for distributed access to structure and state information
- Global Access to Secondary Storage – remote access via parallel interfaces
- Globus Executable Management – construction, location and caching executables
- Globus Advanced Reservation and Allocation – reserves and allocates resources

3 Data Mining

The knowlegde discovery in databases (KDD) is a process of extraction of new, previously unknown information from lagre datasets. Datamining is one step of the KDD process, where machine learning methods are applied on data in order to extract data patterns. Datamining is a process controlled by the user. This process is interactive, so user has to know how to control the process by setting the parameters.

3.1 Distributed Data Mining

Nowadays, the information overload means big problem, so data mining algorithms working on very large data sets take very long times on comventional computers to get results. One approach to solve this problem is parallel computing – parallel data mining algorithms can offer an effective way to mine very large data sets.

Parallel and distributed knowledge discovery is based on the use of networks for the mining of data in a distributed and parallel fashion. It is possible to manage and analyze data, that are geographically distributed in different data warehouses.

The main idea in this area is to use the Grid computing performance for machine learning tasks.

Nowadays, the process of data mining is one of the most important topics in scientific and business problems. There is a huge amount of data that can help to solve many of these problems. However, data is geographically distributed in various locations and belongs to several organizations. Furthermore, it is stored in different kind of systems and it is represented in many formats.

Grids provide access to distributed computing and data resources, allowing dataintensive applications to improve significantly data access, management and analysis. Grid systems responsible for tackling and managing large amounts of data in geographically distributed environments are usually named data grids. Generic data management systems, and particularly data mining grids, involve a great number of challenges. There are a couple of existing applications that use Grid to solve these problems.

3.2 GridMiner

GridMiner is a novel service-oriented grid-aware application, which supports the single steps of the knowledge discovery process by a set of OGSA services covering data integration, data preprocessing, and data mining methods [7]. GridMiner provides both centralized and distributed datamining methods. In the centralized scenario the user wants to perform some data preprocessing first and

analyze the resulting dataset via data mining algorithms in a centralized fashion – on one host. In the distributed data mining model two or more Grid nodes simultaneously participate in the data mining task. This distributed model uses parallel data mining algorithms that are extensions of the algorithms developed for distributed-memory parallel architectures and more complex workflow. It is derived by typical discovery process, that is orchestrated by a workflow engine.

Conclusions

In this paper I gave a short overview of evolution, application of the Grid technology to knowledge discovery in databases process. Grid computing is now an important new technology with many commercial and industrial enterprises. It is possible to extend the Grid technology to category of applications such as data mining. Data mining in the Grid environment can provide highly efficient and powerful data analysis and knowledge discovery solution.

References

[1] I. Foster, C Kesselman, *Computational Grids*, The Grid – Blueprint for a new Computing Infrastructure, Morgan Kaufmann, 1999

[2] F. Berman, A. J. G. Hey, G. C. Fox, *Grid Computing-making the global infrastructure a reality,* Wiley, 2003

[3] FAFNER: Factoring via network enabled recursion. http://cs-www.bu.edu/cgi-bin/FAFNER/factor.pl

[4] I. Foster, J Geisler, W Nickles, W. Smith and S. Tuecke: Software infrastructure for the I-WAY high performance distributed computing experiment. In *Proc. 5th IEEE Symposium on High Performance Distributed Computing*

[5] I. Foster, C. Kesselman. Globus: A metacomputing infrastructure toolkit. *International Journal of Supercomputer Applications,* 1997

[6] The Globus Project, Globus toolkit 3.0. http://www.globus.org

[7] P. Brezany, J. Hofer, A. M. Tjoa, A. Wohrer. GridMiner: An infrastructure for data mining on computational grids, 2003.