DISTRIBUTED AND PARALLEL DATA MINING WITH GRID COMPUTING ENVIRONMENT

AMIT KAPOOR[1]

Abstract

Advance in computing and communication over wired and wireless network have resulted in many distributive computing environments. Many of these environments have different distributed sources of voluminous data and multiple compute nodes. The Various science applications are typically at the forefront of large scale computing problems.

Fundamental scientific problems currently being explored generate gradually more complex data, require more realistic simulations of the processes under study and demand greater and more difficult visualizations of the results. These problems often require numerous complex calculations and collaboration among people with multiple disciplines and geographic locations. Examples of scientific grand challenge problems include multi-scale environmental modeling and ecosystem simulations, biomedical imaging and biomechanics, nuclear power and weapons simulations, fluid dynamics and fundamental computational science.

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. The grid can play 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.

[1] Assistant Professor, Maharaja Agrasen Institute of Management and Technology, Jagadhri, Haryana, India

Data mining techniques can be efficiently deployed in a grid environment and operational grids can be mined for patterns that may help to optimize the effectiveness and efficiency of the grid computing infrastructure. 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.

***Keywords**: Grid Computing, Distributed Data Mining, Data mining algorithms, Knowledge Grid.*

1. Introduction

Grid is the notion of providing computing power transparently in an analogy with electrical power. It aims to aggregate distributed computing resources, hide their specifications and present a homogeneous inter-face to end users for high performance or high throughput computation 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.

Basically grid can term as **resource sharing +problem solving.**

Resource sharing can define direct access to computer, various software, hardware, and data. . Sharing is highly controlled, clearly defining what is shared, who is allowed to share, and the conditions under where and which sharing occurs.

Problem can be solved through a network because all resources are linked with a network. Network connects all the resources with the grid and allows them to be used collectively.

In recent developments we have seen an unprecedented growth of data and information in a wide range of knowledge sectors. The term information explosion describes the rapidly increasing amount of published information and its effects on society. It has been estimated that the amount of new information produced in the world increases by 30 per cent each year. The Population Reference Bureau 1 estimates that 800 MB of recorded information are produced per person each year (assuming a world population of 6.3 billion). Many organizations, companies and scientific centres produce and store large amounts of complex data and information. Examples include climate and astronomy data, economic and financial transactions and data from many scientific disciplines. To justify their existence and maximize their use, these data need to be stored and analysed. The larger and the more complex these data, the more time consuming and costly is their storage and analysis. Data mining has been developed to address the information needs in modern knowledge sectors. Data mining refers to the non-trivial process of identifying valid, novel, potentially useful and understandable patterns in large volumes of data. Because of the information explosion phenomenon, data mining has become one of the most important areas of research and development in computer science.

2. Distributed Data Mining

Data Mining evolution has outlined the development of new contributions in any of the following two lines:

(i) New algorithms, theoretical models or data mining techniques

(ii) Technological and design research for new data mining systems and architectures

The same can be asserted for 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 conventional 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. A primary motivation for Distributed Data Mining (DDM) is that a lot of data is inherently distributed. Merging of remote data at a central site to perform data mining will result in unnecessary communication overhead and algorithmic complexities. For example, consider the NASA Earth Observing System Data and Information System (EOSDIS) which manages data from earth science research satellites and field measurement programs. It provides data archiving, distribution, and information management services and holds more than 1450 datasets that are stored and managed at many sites throughout the United States. It manages extraordinary rates and volumes of scientific data. A centralized data mining system may not be adequate in such a dynamic, distributed environment. Indeed, the resources required to transfer and merge the data on a centralized site may become implausible at such a rapid rate of data arrival. Data mining techniques that minimize communication between sites are quite valuable. Some examples are distributed or parallel algorithms for association rules, classification rules, sequence patterns or clustering algorithm.

3. Grid & Distributed Data Mining

Grid computing represents the natural evolution of distributed computing and parallel-processing technologies. Basically, grid computing employs groups of locally or remotely networked machines to work together on specific computational tasks to harness the power of many computers in a network. The primary aim of grid computing is to give IT organizations and application developers the ability to create distributed computing environments that can utilize computing resources on demand. In practice, grid computing can leverage the processing capacity of hundreds, or even thousands, of computers. Thus 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.

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.

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.

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 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.

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.

4. Distributed Vs Centralized Data Mining

The goal of distributed data mining is to get global knowledge from the local data at distributed sites (N. Zhang, et al, 2009). Recently, many companies, organizations and research centers have been generating and manipulating huge amounts of digital data and information. The digital data are stored in distributed repositories for more reliable and fast access of information. Basically two approaches can be used for mining data from the distributed database: one is Distributed Data Mining (DDM) and the other is Centralized Data Mining (CDM) (M. F. Santos, et al, 2010,

C. Clifton et al, 2002). Centralized Data Mining is also known as warehousing method (N. Zhang, et al, 2009). In CDM, data is stored in the different local databases, but for mining purpose, all data has to be transferred from local databases to the centralized data repository. There are many excising applications that are used this principle of collecting data at centralized site and running an algorithm on that data. Figure 1 depicts the centralized method when considering three geographically distributed databases. These three databases may be the parts of one organization, while the execution of distributed data mining, the data which is stored in the local database has to send to the central repository. The data mining algorithm is applied on that collected data, which is in the central repository and generate global model.

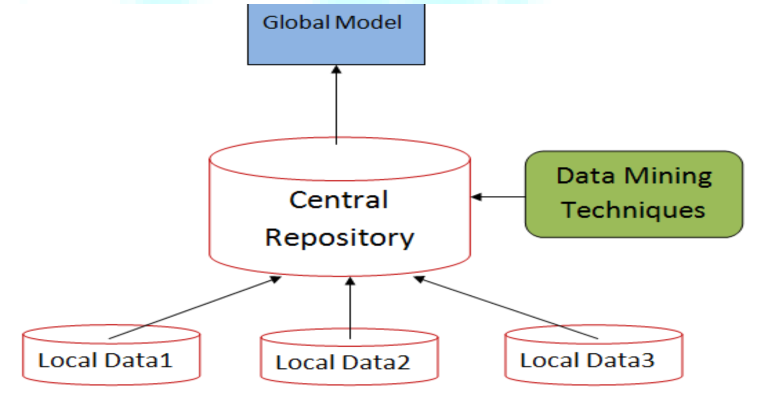

Figure 1: Centralized Data Mining (CDM) Approach

The size and security of data are two main concerns in the centralized mining method. The large size of the data will increase the communication and computational cost of mining process. The size of the local data may vary from one site to another and it is not controlled by the user. Algorithm can retrieve local data depending on the requirement of the user, but the size is not predictable. High bandwidth communication channel is required for sending large sized data to a central site; otherwise, it will take longer to send data to central site. After receiving all data from different nodes, mining algorithm would be applied to the centralized data. Because of the large size of the data, generating the centralized model would be slower, resulting in higher computational cost. The privacy is another main issue of sending data to a central repository. For example, an insurance company with different branches may be unwilling to transmit large amounts of data across a network (C. Clifton et al, 2002).

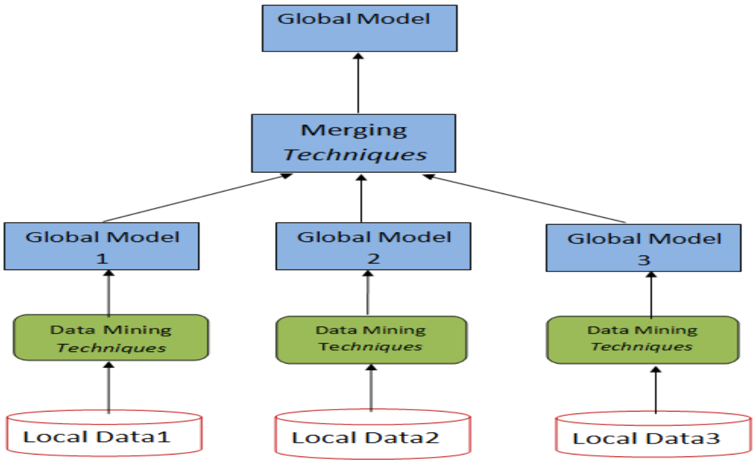

Figure 2: Distributed Data Mining (DDM) Approach

Distributed data mining means data mining in the distributed data sets (N. Zhang, et al, 2009). In DDM, data sets are stored in different local data sets, and hosted by local computers that are connected through a computer network (N. Zhang, et al, 2009). First, data mining is executed in all local environments, and then all these local models (results) from local nodes are combined. The combined model (result) is called Global Model (GM). Figure 3 depicts the DDM method with three geographically distributed databases. The first process of the DDM is to make three local models by applying mining algorithm at each site in parallel fashion. Central node will receive all these three local models and merge them to develop the GM. The DDM has several advantages compared to the CDM. First, the DDM system doesn't need to send original data to the central node; instead, system can send local trained models to the central site, making DDM more secure than CDM. Second, the local model has a fixed size, which is set by the user. Since the size of the local model would be less than the size of training data, DDM doesn't require high bandwidth for communicational channel thus reducing the communication cost. Third, the computational cost of the DDM is less because in DDM paradigm, data mining is executed in parallel fashion on small data sets from each node.

5. Conclusion

As the demand for automated analysis of large and distributed data grows, new data mining challenges in distributed computing environments emerge. The aim of DataMiningGrid project is to address some important requirements arising from modern data mining scenarios.

Major features of the DataMiningGrid technology include high performance, scalability, flexibility, ease of use, conceptual simplicity, compliance with emerging grid and data mining standards and use of mainstream grid and open technology. As a result, even larger-scale problems are envisaged and in many areas so-called grand challenge problems are being tackled. These problems put an even greater demand on the underlying computing resources. A growing class of applications that need large-scale resources is modern data mining applications in science, engineering and other areas. Grid technology is an answer to the increasing demand for affordable large-scale computing resources. The emergence of grid technology and the increasingly complex nature of data mining applications have led to a new synergy of data mining and grid. On one hand, the concept of a data mining grid is in the process of becoming a reality. A data mining grid facilitates novel data mining applications and provides a comprehensive solution for affordable high performance resources satisfying the needs of large-scale data mining problems. On the other hand, mining grid data is emerging as a new class of data mining application. Mining grid data could be understood as a methodology that could help to address the complex issues involved in running and maintaining large grid computing environments.

References

1. Cho,V. andW¨uthrich, B. (2002), ‘Distributedmining of classification rules’, *Knowledge and Information Systems* 4 (1), 1–30.
2. Kargupta, H., Huang,W., Sivakumar, K. and Johnson, E. (2001), ‘Distributed clustering using collective principal component analysis’, *Knowledge and Information Systems Journal 3*, 422–448.

3. Kargupta, H., Kamath, C. and Chan, P. (2000), Distributed and parallel data mining: Emergence, growth, and future directions, *in* ‘Advances in Distributed and Parallel Knowledge Discovery’, AAAI/MIT Press, pp. 409–416.
4. Stankovski, V., Swain, M., Kravtsov, V., Niessen, T., Wegener, D., Kindermann, J. and Dubitzky, W. (2008), ‘Grid-enabling data mining applications with DataMiningGrid: An architectural perspective’, *Future Generation Computer Systems* 24, 259–279.
5. Grid web site. URL: www.mygrid.org.uk
6. M. Cannataro, D. Talia, P. Trunfio, Distributed data mining on the grid, Future Generation Computer Systems 18 (8) (2002) 1101–1112
7. D. Hand, H. Mannila, P. Smyth, Principles of Data Mining, MIT Press, Cambridge, MA, 2001.
8. J. Trnkoczy, Z. Turk, V. Stankovski, A grid-based architecture ˆ for personalized federation of digital libraries, Library Collections, Acquisitions, and Technical Services 30 (2006) 139–153
9. Discovery Net. www.discovery-on-the.net