Chapter 14
Data Mining for High Performance Computing

Shen Lu
Soft Challenge LLC, USA

ABSTRACT

With the development of information technology, the size of the dataset becomes larger and larger. Distributed data processing can be used to solve the problem of data analysis on large datasets. It partitions the dataset into a large number of subsets and uses different processors to store, manage, broadcast, and synchronize the data analysis. However, distributed computing gives rise to new problems such as the impracticality of global communication, global synchronization, dynamic topology changes of the network, on-the-fly data updates, the needs to share resources with other applications, frequent failures, and recovery of resource. In this chapter, the concepts of distributed computing are introduced, the latest research are presented, the advantage and disadvantage of different technologies and systems are analyzed, and the future trends of the distributed computing are summarized.

INTRODUCTION

Nowadays, with the size of customer's datasets become larger and larger, distributing datasets into different machines provides an effective solution for data processing. However, distributed data processing [Corbett et al. 2012] introduces the issues of global synchronization and communication among machines. Nodes are required to work independently and also cooperate together where nodes generate the data locally and update the data whenever new data arrive. This means that for all practical purpose the nodes should receive the data from its immediate neighbors and compute the results through local negotiation. Therefore, a global database is constructed through each node of data exchange from its immediate neighbors.

With the advance of the technology, a database can now be partitioned into a large number of computers, such as grid computing platforms (Villegas, et. al. 2010. Talia, 2006), federal database systems (McLeod & Heimbigner, 2009), and peer-to-peer computing environments (Datta et al., 2008). However, parallel processing assumes the availability of parallel processors, even though data mining algorithms are stand-alone processes and do not require parallel processors. Distributed data

DOI: 10.4018/978-1-4666-7461-5.ch014

processing divide the transactional dataset D into N non-overlapping partitions, D_1, D_2, D_3, ..., D_n.

Many data analysis and mining algorithms have been proposed that focus on improving the efficiency of the algorithms via parallelism, which uses hash-based technology, transaction reduction, partitioning, and sampling. Partitions are distributed to processors. Each processor finishes data analysis on a partition independently and creates its local result against its own dataset partition. After that, processors exchange their local dataset partitions and results for the global synchronization. The master processor is different from other processors. It plays a central role in distributed computing, which can not only work as a standalone processor, but also manage, broadcast, and synchronize other processors.

However, distributed computing gives rise to new problems such as the impracticality of global communication, global synchronization, dynamic topology changes of the network, on-the-fly data updates, the needs to share resources with other applications, frequent failures, and recovery of resource. Distributed data mining algorithms have been provided for this purpose. For example, distributed system can be considered as the combination of several individual processors in which every node in the system can reach the exact solution, impose very little communication overhead, transparently tolerate network topology changes and node failures, and quickly adjust to changes in the data as they occur. Another example is the distributed system with complicated master processor and simple distributed nodes, in which the master processor can be used to manage, broadcast and synchronize and other processors can only perform heavy-duty and time-consuming computation. Different architectures use different strategies to eventually finish global communication, global synchronization, dynamic topology changes of the network, on-the-fly data updates.

In this chapter, we introduce databases for high performance computing (HPC), data mining for high performance computing, visualization for modeling and simulation, and input/output performance tuning.

DATA MINING

Data Mining as the Evolution of Information Technology

Data mining can be used to integrate, manage, analyze and predict information. With the development of World Wide Web, data storage devices, and data collecting machines, a vast amount of information are collected each day from business, science, medicine and almost every aspect of daily life. The fast-growing, tremendous amount of data, collected and stored in large and numerous data repositories, has far exceeded our human ability for comprehension without powerful tools. We need to understand data, use data to help make decisions, find interesting knowledge from data and so on. Data mining is the process of discovering interesting patterns and knowledge from large amounts of data. However, the process of knowledge discovery includes several steps, such as data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation, and knowledge presentation.

What Kinds of Data can be Mined?

Regarding temporal data, we can mine banking data for changing trends, which may aid in the scheduling of bank tellers according to the volume of customer traffic. Stock exchange data can be mined to uncover trends that could help you plan investment strategies. We could mine computer network data streams to detect intrusions based on the anomaly of message flows (Lau et. al, 2013), which may be discovered by clustering, dynamic construction of stream models or by comparing the current frequent patterns with those at a previous time. With spatial data, we may look for patterns that describe changes in metropolitan