Chapter 1.10
Parallel, Distributed, and Grid-Based Data Mining: Algorithms, Systems, and Applications

Moez Ben HajHmida
Faculty of Sciences of Tunis, Tunisia

Antonio Congiusta
University of Calabria, Italy & University of Salerno, Italy

ABSTRACT

Knowledge discovery has become a necessary task in scientific, life sciences, and business fields, both for the growing amount of data being collected and for the complexity of the analysis that need to be performed on it. Classic data mining techniques, developed for centralized sites, often reveal themselves inadequate, due to some unique characteristics of today's data sources. In such cases, sequential approaches to data mining cannot provide for scalability, in terms of the data dimensionality, size, and runtime performance. Moreover, the increasing trend towards decentralized business organizations, distribution of users, software, and hardware systems magnifies the need for more advanced and flexible approaches and solutions. Life science is one of the application areas that best resemble such scenario. This chapter presents the state of the art about the major data mining techniques, systems and approaches. A detailed taxonomy is drawn by analyzing and comparing parallel, distributed and Grid-based data mining methods, with a particular focus on the exploitation of large and remotely dispersed datasets and/or high-performance computers.

DOI: 10.4018/978-1-4666-0879-5.ch1.10

INTRODUCTION

Data mining aims at extracting hidden information from large data repositories (e.g., databases, file archives, digital libraries) for building valuable knowledge patterns and predictive models. The main data mining tasks are association rules discovery, classification and clustering.

Data mining is a massive computing task that deals with memory resident data. With the huge amount of stored data in centralized or distributed systems, traditional data mining techniques encounter limitations and shortcomings that often lead to inefficiencies. The need for parallel and distributed computing becomes inevitable to deal with large-scale data mining (Freitas, 1998; Kargupta, 2000), and for addressing complex needs and scenarios encountered in business as well as research organizations.

Parallel Data Mining (PDM) is targeted to tightly-coupled systems, like shared or distributed memory machines, and clusters based on fast networks. Distributed Data Mining (DDM) deals with loosely-coupled systems: clusters with average-fast or slow networks and geographically distributed computing nodes. The main differences between PDM and DDM are the number of involved computing nodes, the communication costs, and the degree of data distribution.

The advances in network technologies have produced huge amount of data stored on geographically distributed databases and repositories. When these amounts of data are owned by different organizations and hosted by non-dedicated computing resources, parallel and distributed data mining techniques start showing their limits. The Grid is the computing architecture that provides means for utilizing geographically distributed resources as a single meta-system. The emergence of such new infrastructure is highly beneficial to large-scale and compute-intensive data mining, as it offers new opportunities to optimize and speed-up mining processes.

Unlike previous parallel and distributed data mining surveys (Kargupta, 2000; Zaki, 2000), this chapter differentiates between:

- PDM, where learning methods are often platform dependent, databases are centralized (for example in a cluster or a supercomputer), and network connections are reliable and fast;
- DDM, in which learning methods are based on sharing nothing machines with a much slower network and databases are naturally distributed;
- and, in addition, the Grid Data Mining (GDM) category of algorithms and systems, which is deeply explored.

Although GDM shares many commonalities with PDM and DDM, there are platform peculiarities and requirements implying that efforts and obtained results in such area cannot be compared (in a homogeneous way) with those achieved by PDM and DDM.

The remainder of this chapter is organized as follows: Section 2 contains a background on classical data mining techniques; Section 3 presents parallel systems and the related programming paradigms, it details the techniques used in parallel data mining and classify them on the basis of the employed method; Section 4 presents the main differences between parallel and distributed data mining techniques, then it discusses the major distribution methods and their transition from parallel to distributed systems; Section 5 contains a description of the knowledge discovery process in Grid environments, it focuses on Grid infrastructures and frameworks designed for such purpose; finally Section 6 draws conclusions and highlights some future trends.