PARALLEL DATA MINING WITH DBMS FACILITIES.

This Chapter continues to follow the spirit of integrating data mining and parallel processing, as presented in the previous Chapter, but it goes further in that it adds DBMS facilities to the picture. The basic motivation is that most large data warehouses are already stored on parallel database servers, offering high-performance DBMS facilities (Chapter 8). Hence, it seems natural that these servers should be used to realize a cost-effective, holistic framework for parallel data mining and for the entire knowledge discovery process.

We start by presenting an overview of integrated data mining/data warehouse frameworks and discussing their advantages over a conventional, non-integrated (client-based) framework. Two basic integrated frameworks are discussed in this Chapter. One of them is server-based, i.e. the data mining algorithm is run on the parallel database server. The other is hybrid client/server-based, i.e. some procedures of the data mining algorithm are performed on the client, but other procedures (the data-intensive, time-consuming ones) are performed on the parallel database server.

In this Chapter we also present generic, set-oriented primitives for the hybrid client/server-based framework. In particular, we present two such primitives, one developed for the rule induction paradigm and the other developed for the instance-based learning paradigm. Finally, for the sake of completeness, we also briefly discuss the topic of parallel data mining using specialized-hardware parallel database servers.

11.1 AN OVERVIEW OF INTEGRATED DATA MINING/DATA WAREHOUSE FRAMEWORKS.

In this Chapter we are interested in integrated data mining/data warehouse frameworks for Knowledge Discovery in Databases (KDD). Although our discussion will focus on the data mining step of the KDD process, we stress that this integration is also beneficial for other steps of the KDD process (Chapter 3). The motivation for integrating KDD algorithms and data warehouses will be discussed in Section 11.2. First, however, it is necessary to elaborate the concept of integrated data mining/data warehouse frameworks.

We assume that the data warehouse system has a client-server architecture. In this architecture the data warehouse is stored on a parallel database server, while users
access the data warehouse through client machines, typically PCs and/or workstations. We observe that this is the basic architecture of most large data warehouses [IBC 95], which are stored on parallel database servers for reasons of efficiency and scalability.

In this Section we discuss two integrated data mining/data warehouse frameworks. In one of them the data mining algorithm is fully embedded on the parallel database server, while in the other the data mining algorithm is partially embedded on the parallel database server. We will refer to these frameworks as server-based and hybrid client/server-based KDD frameworks, for short.

Before we discuss these two frameworks, it is important to clarify how they differ from a "conventional", client-based framework, where there is no integration between the data mining algorithm and the data warehouse. This difference is illustrated in Figures 11-1 through 11-3, which show, respectively, the (conventional) client-based, the server-based and the hybrid client/server-based KDD frameworks. Let us discuss each of these frameworks in turn. Throughout this discussion we will refer to the symbolic convention used in Figures 11-1 through 11-3, namely: the relative size of the squares, circles and triangles, representing respectively data, data mining (DM) algorithm and discovered knowledge, roughly indicate the size of the corresponding object.

**Figure 11-1:** "Conventional", client-based KDD framework.

**Figure 11-2:** Server-based KDD framework.