A Tutorial Introduction to High Performance Data Mining

Robert Grossman

Magnify, Inc. and University of Illinois at Chicago

Abstract

The goal of the tutorial is to provide researchers, practitioners, and advanced students with an introduction to techniques from high performance computing and high performance data management which are applicable to data mining and to illustrate their use on a variety of practical data mining problems.

A fundamental problem in data mining is to develop data mining algorithms and systems which scale as the amount of data grows, as the dimension grows, and as the complexity of the data grows. Recently, 1) parallel and distributed variants of some standard data mining algorithms have been developed and 2) data mining systems have begun to develop higher performance access to databases and data warehouses. These and related topics will be covered in the tutorial. Finally, we will cover several case studies involving mining large data sets, from 10-500 Gigabytes in size.