Topic 22
High Performance Data Mining and Knowledge Discovery

David Skillicorn and Domenico Talia

Co-chairmen

Many, perhaps most, organizations use computers when they interact with their customers. As a result, and almost by accident, many organizations have accumulated huge amounts of data about such interactions. Over the past five to ten years, they have increasingly tried to use this data for commercial advantage. This process began by accumulating transaction data into *data warehouses*, where it could be made available for decision support and retrospective analysis. The effectiveness of such analysis largely depends on the ability of individuals to induce queries that will reveal key facts about the organization and its customers.

Increasingly, both the volume of data and its complexity have taken the problem beyond the ability of any individual to analyze. *Data mining* is the automated analysis of large volumes of data, looking for the relationships and knowledge that are implicit in large volumes of data and are 'interesting' in the sense of impacting an organization's practice. Research and development work in the area of knowledge discovery and data mining concerns the study and definition of techniques, methods, and tools for the extraction of novel, useful, and implicit patterns from data. It builds on machine learning, database technology, and statistics, but is distinguished by problems of scale: the data involved is so large that most applications tend to use conceptually straightforward, but carefully optimized, algorithms.

There is a natural confluence between parallel computation and data mining. For researchers in parallel computation, data mining is an application area that is growing in importance, and that introduces interesting new problems (irregularity, data representation and storage, multiple parallelization strategies, symbolic computation) that have not been so critical in scientific and numerical computing. For organizations who want to use data mining in their day to day work, parallel computation offers increased performance, which in turn may translate into commercial advantage. When data mining tools are implemented on high-performance parallel computers, they can analyze massive databases in a reasonable time. Faster processing also means that users can experiment with more models to understand complex data. High performance makes it practical for users to analyze greater quantities of data. Larger databases, in turn, yield improved predictions.

Data mining, even sequentially, is not yet mature, and many of the existing applications are relatively unsophisticated. Nevertheless, it seemed useful to explore the fledgling projects that are looking at the connections between parallel

P. Amestoy et al. (Eds.): Euro-Par'99, LNCS 1685, pp. 1439–1440, 1999.

computing and data mining. This track has assembled a small number of papers describe such research experiences.

The first paper "Mining of Association Rules in Very Large Databases: A Structured Parallel Approach" by Becuzzi, Coppola, and Vanneschi, presents a case study implementing the Apriori parallel association rule algorithm using the skeleton-based language SkIE. The paper is as much about parallel software engineering as it is about data mining. It demonstrates the effectiveness of the skeleton approach as a software construction technique for a real problem. The development of a parallel program from a a sequential one, and its subsequent tuning, are shown to be straightforward and inexpensive. The second paper "Parallel k-means Clustering for Large Data Sets" by Stoffel and Belkoniene presents a parallel version of the k-means clustering algorithm where data points are distributed across processors. Experiments on a cluster of 32 PCs are discussed. This implementation shows interesting scalability properties of the h-means version of the algorithm on distributed-memory parallel computers. The third paper "Performance Analysis for Parallel Generalized Association Rule Mining on a Large Scale PC Cluster", by Shintani, Oguchi and Kitsuregawa, addresses the parallel computation of association rules with an item hierarchy. The authors present a performance evaluation, on a large scale PC cluster, of algorithms for mining generalized association rules. The performance results show that the algorithms are effective for handling skew on highly parallel computers. The last paper "Inducing Load Balancing and Efficient Data Distribution Prior to Association Rule Discovery in a Parallel Environment" by Manning and Keane, proposes a statistical method to achieve efficient data partitioning in parallel data mining. In particular, the authors suggest a database redistribution based on principal component analysis (PCA) to achieve load balancing and reduction in candidate set duplication in a parallel algorithm for mining association rules.

The four papers presented in this session give some idea of the issues and approaches to using parallel computation to implement scalable data mining algorithms. Parallel data mining is a long way from mature status. We hope the results presented here will stimulate interest in the confluence of parallel computing and data mining, and will lead to more work in this important area.

We would like to thank the two Vice-Chairs, Vipin Kumar and Hannu Toivonen, who organized the track and reviewed papers with us. We would also like to thank the external referees who helped us review all of the submissions. Their work tends to be unnoticed, but is not unappreciated.