Topic 5
Parallel and Distributed Databases, Data Mining and Knowledge Discovery

Harald Kosch, David Skilicorn, and Domenico Talia

Topic Chairpersons

We would like to welcome you to Paderborn and to the Europar 2002 topic on Parallel and Distributed Databases, Data Mining and Knowledge Discovery.

Current research and applications in parallel and distributed database systems are stimulated both by advances in technology (e.g., parallel architectures, high performance networks) and by the requirements of new applications (e.g., multimedia, medicine). These applications are mostly data-hungry and are running on very large databases with the goal of extracting information diamonds. Data mining is one of the key techniques here. However, these intensive data-consuming applications suffer from performance problems and and bottlenecks caused by single sources of data. Introducing data distribution and parallel processing helps to overcome these resource problems and to achieve guaranteed throughput, quality of service, and system scalability.

This year, 13 papers were submitted, more than in previous years. The range and quality of the submitted papers was impressive and reflects advances in technology, as well as describing solutions for application-specific problems. Each paper was reviewed by at least three reviewers and, in the end, we were able to select 6 regular papers and 3 short ones. This shows the high-quality of the submissions. From this, three full sessions have been scheduled. The first one is dedicated to aspects of parallel and distributed database systems, the second one deals with strategies for parallel data mining, and the final third one contains interesting papers on data-grid applications and distribution aspects in database systems.

Parallel databases are currently making a shift from relational query processing and traditional transaction management towards the handling of new data-hungry applications, like multimedia, introducing new operators to query processing. The first paper in the session, "Dynamic Query Scheduling in Parallel Data Warehouses", by H. Märtens, E. Rahm and T. Stöhr introduces a new skew-aware query scheduling strategy for shared-disk parallel systems which considers both disks and processors. The tested queries stem from data warehouse applications. The second paper, "Speeding Up Navigational Requests in a Parallel Object Database System" by J. Smith, P. Watson, S. Sampaio and N. Paton is dedicated to query processing in object-oriented parallel databases. This is, in our opinion, the first framework which completely covers parallelization of so-called path queries including functions. Finally, the third paper, "Retrieval of Multispectral Satellite Imagery on Cluster Architectures", by T. Bretschneider and O. Kao falls into parallel image retrieval. It reflects the need for parallel

B. Monien and R. Feldmann (Eds.): Euro-Par 2002, LNCS 2400, pp. 319–320.

query processing in images, and multimedia databases in general, and parallelizes efficiently similarity-based operators on the image database systems.

The need to analyze large amounts of data is the main motivation for the implementation of parallel data mining algorithms. This trend also benefits from the wider availability of cost-effective parallel machines such as clusters and SMPs. Session 2 includes three papers that discuss parallel data mining algorithms and systems. The first paper "Shared Memory Parallelization of Decision Tree Construction Using a General Data Mining Middleware" by Jin and Agrawal, describes the use of a data-mining framework for developing shared-memory parallel implementations of classifiers based on decision trees. Experiments presented in the paper show that applying a set of techniques results in good performance. The second paper "Characterizing the Scalability of Decision-Support Workloads on Clusters and SMP Systems" by Zhang, Sivasubramaniam, Nagar, and Franke, discusses the scalability of the TPC-H decision support benchmark on a cluster and on an SMP machine. The paper focus is on the impact of various hardware parameters such as CPU, memory, disk and network. The evaluation results show that for both the cluster and the SMP environments, the CPU and memory resources are not major contributors to performance beyond a point, but the I/O parallelism. The last paper "A Parallel Learning Algorithm for Text Classification on PIRUN Beowulf Cluster" by Kruengkrai and Jaruskulch, presents a parallel learning algorithm for text classification based on the combination of the expectation-maximization algorithm and the naive Bayes. The preliminary experimental results discussed in the paper show that the proposed parallel implementation in the mining of up to 10000 documents has reasonable speedup.

The growing size of online data, and the fact that it is often distributed, create new challenges for data-intensive applications. The application papers in Session 3 illustrate this trend. The paper by Orlando *et al.* "Scheduling High Performance Data Mining Tasks on a Data Grid Environment" explores issues of scheduling distributed data mining algorithms using cost information. Unlike conventional grid resource discovery, this paper shows how to use samples to determine the likely behavior of computing resources. The second paper, by Kwok *et al.*, "Parallel Fuzzy c-Means Clustering for Large Data Sets" illustrates the use of parallelism in data mining. Clustering data is a major application. This paper goes beyond the standard k-means approach to clustering to parallelize the fuzzy c-means algorithm using MPI. Finally, the paper by Boukerche and Tuck "A delayed-Initiation Risk-Free Multiversion temporally correct algorithms" present a general algorithm for risk-free multiversion concurrency, extending work presented at Europar in 2001.

In closing, we would like to thank the authors who submitted a contribution, as well as the Europar Organizing Committee, and the referees with there highly useful comments, whose efforts have made this conference, and Topic 05 possible.