Topic 05
Parallel and Distributed Databases, Data Mining and Knowledge Discovery

Harald Kosch, Pedro R. Falcone Sampaio, Abdelkader Hameurlain, and Lionel Brunie

Topic Chairpersons

We would like to welcome you to the Euro-Par 2001 topic on Parallel and Distributed Databases and Data Mining. It is the first time that the parallel and distributed processing in databases topic at the Euro-Par Conference Series has expanded to incorporate contributions coming from the Data Mining area. From the scope and nature of the submissions received, we were quick to recognize the fruitfulness of this symbiosis, stemming from the fact that many technical challenges and solutions can be shared (for instance, the problem of parallel access to multidimensional data structures). As a result of the topic expansion, we are proud to present two sessions with 6 paper presentations for which we expect some controversy, as well as instructive insights originating from the research discussions. We also hope that you enjoy your visit to Manchester.

Today, the World Wide Web is regarded as a distributed global information resource, which is not only important to individual users, but also to business organizations. New intensive data consuming applications (e.g. Multimedia Content Management, Data Mining, Decision Support, E-Commerce) emerged in this environment. They often suffer from performance problems and single database source limitations. Introducing data distribution and parallel processing helps to overcome resource bottlenecks and to achieve guaranteed throughput, quality of service, and system scalability. Recent developments in parallel and distributed architectures supported by high performance networks and intelligent middleware offer parallel and distributed databases a great opportunity to make a step from being highly specialized systems to supporting cost-effective every day applications.

This year 10 papers were submitted, a considerable increase with regard to the previous edition of the topic at Euro-Par. The quality and range of the 10 submitted papers was remarkable, and we needed intensive peer discussions to select the best ones. The growing number of submissions shows the popularity of research on parallel and distributed databases, as well as the strong link with its ‘sister’ field of data mining. All papers were reviewed by four reviewers, where, besides the PC members, 19 external reviewers supported the process. The feedback to the authors were thorough and detailed, and many reviewers took additional efforts to point out further possible research directions. Using these referee’s reports as guidelines, the program committee chose six papers for publication and presentation at the conference. Four were selected as regular papers, and two as research notes. These were scheduled to be presented in

R. Sakellariou et al. (Eds.): Euro-Par 2001, LNCS 2150, pp. 278–279, 2001.

two sessions, on Wednesday 29th of August during the afternoon. One session is dedicated to aspects focusing on performance and implementation in parallel and distributed databases and the other session on parallel data-mining oriented aspects.

The first full session focuses on performance and implementation techniques in parallel and distributed databases. The three papers chosen for this session bring new results in classical areas of parallel and distributed database research.

Highlighting the consolidation of the ODMG as a step forward in providing standards for object database systems, and the increasing importance of efficient parallel implementations of query processing operators in ODMG compliant object databases, the first paper by Sandra de F. Mendes Sampaio, Jim Smith, Norman W. Paton and Paul Watson presents experimental results of several join algorithms implemented in the Polar ODMG compliant parallel object database system. The second paper, by Holger Märtens, brings new insights on the relation between different skew types and load balancing methods in parallel database systems, suggesting some anti-skew measures to alleviate skew effects. The third paper, by Azzedine Boukerche and Terry Tuck, addresses the problem of improving concurrency control in distributed databases, by using implementation techniques based on predeclaring tables that will be updated when a transaction begins.

The second full session focuses on parallel processing in data mining. Reflecting the increasing importance of data mining techniques in industrial applications, all three papers contribute to a more efficient processing of these complex and resource-hungry techniques.

In the first paper, Valerie Guralnik, Nivea Garg and George Karypis present two parallel formulations (using once a data and once a task parallelisation) of a serial sequential pattern discovery algorithm based on tree projection that are well suited for distributed memory parallel computers. The discovery of sequential patterns is becoming increasingly useful and essential in many commercial applications and an acceptable mining performance is critical to these applications. In the second paper Attila Gürsoy and İlker Cengiz have developed and evaluated two parallelisation schemes for a tree-based k-means method on shared memory architectures. They use a data-parallelization paradigm and study data pattern composition necessary for efficient processing. This paper is highly related to the first one and we expect some interesting comparative discussions in this session. In the third paper, Domenica Arlia and Massimo Coppola have developed and implemented a parallelisation of DBSCAN, a broadly used Data Mining algorithm for density-based spatial clustering. Interesting, here, is the applicability of the developed methods to multimedia, or spatial parallel databases and makes therefore a link to the first session.

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