Topic 5
Parallel and Distributed Databases

Marta Patiño-Martinez, Genoveva Vargas-Solar, Elena Baralis, and Bettina Kemme

Topic Chairs

Advances in data exploitation (access, query, retrieval, analysis, mining) are inherent to current and future information systems. Today, accessing great volumes of information is reality; tomorrow data intensive management systems will enable huge user communities to transparently access multiple pre-existing autonomous, distributed and heterogeneous resources (data, documents, services). Existing data management solutions do not provide efficient techniques for exploiting and mining tera-datasets available in clusters, peer to peer and grid architectures. Parallel and distributed databases are a key element for achieving scalable, efficient systems that will both cost-effectively manage and extract knowledge from huge amounts of highly distributed and heterogeneous digital data repositories.

The track received 16 submissions from which 6 papers were accepted for publication.

The first paper is devoted to P2P data management. "Topology-Aware Approach for Distributed Data Reconciliation in P2P Networks" presents an algorithm for data reconciliation that exploits knowledge about the topology of the underlying P2P network to reduce the communication cost. Two other papers are devoted to parallel and distributed algorithms for data mining. "Efficient Distributed Data Condensation for Nearest Neighbor Classification" presents a distributed algorithm for computing a consistent subset of the training datasets for the nearest neighbor classification whilst "Parallel Nearest Neighbour algorithms for Text Categorization" explores the parallelization of the nearest neighbour problem.

"Handling Request Variability for QoS-max Measures" focuses on how optimizing the multi-programming level of the database to maximize the throughput.

"Search Engine Accepting On-Line Updates" discusses how to address updates in parallel search web engines. Finally, "A Multi-layer Collaborative Cache for Question Answering" addresses how to distribute a cache for answering queries efficiently.