Distributed and Parallel Databases, 11, 123–127, 2002

Introduction: Recent Developments in Parallel and Distributed Data Mining

MOHAMMED J. ZAKI zaki@cs.rpi.edu
Computer Science Department, Rensselaer Polytechnic Institute, Troy, NY 12180, USA

YI PAN pan@cs.gsu.edu
Department of Computer Science, Georgia State University, Atlanta, GA 30303, USA

Introduction

Data Mining and Knowledge Discovery in Databases (KDD) is an interdisciplinary field merging ideas from statistics, machine learning, databases, and parallel and distributed computing. It has been engendered by the phenomenal growth of data in all spheres of human endeavor, and the economic and scientific need to extract useful information from the collected data. The key challenge in data mining is the extraction of knowledge and insight from massive databases.

KDD refers to the overall process of discovering new patterns or building models from a given dataset. There are many steps involved in the KDD enterprise which include data selection, data cleaning and preprocessing, data transformation and reduction, data-mining task and algorithm selection, and finally post-processing and interpretation of discovered knowledge. This KDD process tends to be highly iterative and interactive.

Typically data mining has the two high level goals of *prediction* and *description*. In prediction, we are interested in building a model that will predict unknown or future values of attributes of interest, based on known values of some attributes in the database. In KDD applications, the description of the data in human-understandable terms is equally if not more important than prediction. Two main forms of data mining can be identified. In *verification-driven* data mining the user postulates a hypothesis, and the system tries to validate it. The common verification-driven operations include query and reporting, multidimensional analysis or On-Line Analytical Processing (OLAP), and statistical analysis. *Discovery-driven* mining, on the other hand, automatically extracts new information from data. The typical discovery-driven tasks include association rules, sequential patterns, classification and regression, clustering, similarity search, deviation detection, etc.

While data mining has its roots in the traditional fields of machine learning and statistics, the sheer volume of data today poses the most serious problem. For example, many companies already have data warehouses in the terabyte range (e.g., FedEx, UPS, Walmart). In addition to business oriented data mining, data mining and domain knowledge plays a significant role in knowledge discovery and refinement in engineering, scientific, and medical databases, which are reaching gigantic proportions (e.g., NASA space missions, Human Genome Project) and require both large memory and disk space and high speed computing.

Traditional methods typically made the assumption that the data is memory resident. This assumption is no longer tenable. Implementation of data mining ideas in high-performance parallel and distributed computing environments is thus becoming crucial for ensuring system scalability and interactivity as data continues to grow inexorably in size and complexity.

This special issue of Distributed and Parallel Databases provides a forum for the sharing of original research results and practical development experiences among researchers and application developers from different areas related to parallel and distributed data mining. Papers for this special issue were selected to address data mining methods and processes from both an algorithmic and systems perspective in parallel and distributed environments.

The algorithmic aspects involve the design of efficient, scalable, disk-based, parallel and distributed algorithms for large-scale data mining tasks. The challenge is to develop methods that scale to thousands of attributes and millions of transactions. The techniques of interest span all major classes of data mining methods such as association rules, sequences, classification, clustering, deviation detection, as well as various pre-processing and post-processing operations like sampling, feature selection, data reduction and transformation, rule grouping and pruning, exploratory and interactive browsing, meta-level mining, etc.

The systems issues focus on actual implementation of the algorithms on a variety of parallel hardware platforms, including shared-memory systems (SMPs), distributed-memory systems, network of workstations, hybrid systems consisting of a cluster of SMPs, geographically distributed systems, etc. The key challenges include improving the load balancing, improving locality, eliminating false sharing on SMPs, minimizing synchronization, minimizing communication, maximizing accuracy of distributed models, integrating heterogeneous sources, and finding appropriate data layouts. Papers dealing with integration of mining with databases and data-warehousing, as well as successful applications, were also sought.

Articles in this special issue

Through rigorous reviews involving 36 referees, four papers were chosen from a pool of 12 papers submitted to this special issue. This is reflected in the high quality of the papers accepted.

In the first paper entitled, "Shared State for Distributed Interactive Data Mining Applications," Parthasarathy and Dwarkadas present and evaluate a distributed interactive data mining system, called InterAct, which supports data sharing efficiently by allowing caching, by communicating only the modified data, and by allowing relaxed coherence requirement specification for reduced communication overhead.

A typical clustering algorithm requires bringing all the data in a centralized warehouse, and involves large transmission cost. In the second paper entitled, "RACHET: An Efficient Cover-Based Merging of Clustering Hierarchies from Distributed Datasets," Samatova et al. present a hierarchical clustering method named RACHET for analyzing multi-dimensional distributed data. RACHET runs with at most linear time, space, and communication costs to build a global hierarchy of comparable clustering quality by merging locally generated clustering hierarchies.

In the paper, “Parallelizing the Data Cube,” Dehne et al. present a general methodology for the efficient parallelization of existing data cube construction algorithms. The methods used reduce inter-processor communication overhead by partitioning the load in advance and enable code reuse by permitting the use of existing sequential data cube algorithms for the sub-cube computations on each processor.

Boosting is a popular technique for constructing highly accurate classifier ensembles. In the last paper, “Boosting Algorithms for Parallel and Distributed Learning,” Lazarevic and Obradovic present new parallel and distributed boosting algorithms. They have applied their proposed methods to several data sets and the results show that parallel boosting can achieve the same or even better prediction accuracy, yet is much faster, than the standard sequential boosting.

In our opinion, the selected papers cover as broad as a range of topics possible within the area of parallel and distributed data mining. We hope the research community finds this special issue of use and interest. We thank all those who submitted papers to this special issue. We are also thankful to Professor Ahmed K. Elmagarmid, the editor-in-chief of DPD, for his guidance and support in integrating this issue. Many thanks also go to the reviewers for their prompt responses and helpful comments.

Resources on parallel and distributed KDD

There is a wealth of resources available for further exploration of parallel and distributed KDD, such as books, journal special issues and special topics workshops, as listed below.

Books on parallel and distributed data mining

1. A. Freitas and S. Lavington, Mining Very Large Databases with Parallel Processing, Kluwer Academic: Boston, MA, 1998.
2. M.J. Zaki and C.-T. Ho (Eds.), Large-Scale Parallel Data Mining, LNAI State-of-the-Art Survey, Vol. 1759, Springer-Verlag: Berlin, 2000.
3. H. Kargupta and P. Chan (Eds.), Advance in Distributed and Parallel Knowledge Discovery, AAAI Press, 2000.

Journal special issues

1. H. Kargupta, J. Ghosh, V. Kumar, and Z. Obradovic (Eds.), “Distributed and Parallel Knowledge Discovery,” Knowledge and Information Systems, vol. 3, no. 4, 2001.
2. V. Kumar, S. Ranka, and V. Singh (Eds.), “High performance data mining,” Journal of Parallel and Distributed Computing, vol. 61, no. 3, 2001.
3. A. Zomaya, T. El-Ghazawi, and O. Frieder (Eds.), “Parallel and Distributed Computing for Data Mining,” IEEE Concurrency, vol. 7, no. 4, 1999.
4. Y. Guo and R. Grossman (Eds.), “Scalable Parallel and Distributed Data Mining,” Data Mining and Knowledge Discovery: An International Journal, vol. 3, no. 3, 1999.
5. P. Stolorz and R. Musick (Eds.), “Scalable High-Performance Computing for KDD,” Data Mining and Knowledge Discovery: An International Journal, vol. 1, no. 4, 1997.

Workshops

1. 4th IEEE IPDPS International Workshop on Parallel and Distributed Data Mining, 2001. http://www.cs.rpi.edu/~zaki/PDDM01
2. HiPC Special Session on Large-Scale Data Mining, 2000. http://www.cs.rpi.edu/~zaki/ LSDM/
3. ACM SIGKDD Workshop on Distributed Data Mining, 2000. http://www.eecs.wsu.edu/ ~hillol/DKD/dpkd2000.html
4. 3rd IEEE IPDPS Workshop on High Performance Data Mining, 2000. http://www.cs.rpi. edu/~zaki/HPDM/
5. ACM SIGKDD Workshop on Large-Scale Parallel KDD Systems, 1999. http://www.cs. rpi.edu/~zaki/WKDD99/
6. ACM SIGKDD Workshop on Distributed Data Mining, 1998. http://www.eecs.wsu.edu/ ~hillol/DDMWS/papers.html

List of reviewers

Charu C. Aggarwal, Gagan Agrawal, Daniel Barbara, Raj Bhatnagar, Christopher W. Clifton, Ayhan Demiriz, Sanjay Goil, Robert L. Grossman, Himanshu Gupta, Eui-Hong Han, Benjamin C.M. Kao, Hillol Kargupta, George Karypis, Bing Liu, Malik Magdon-Ismail, William A. Maniatty, Ron Musick, Salvatore Orlando, Srinivasan Parthasarathy, Jian Pei, Sakti Pramanik, Andreas Prodromidis, Naren Ramakrishnan, Rajeev Rastogi, Tobias Scheffer, Paola Sebastiani, David B. Skillicorn, Domenico Talia, Kathryn Thornton, Haixun Wang, Jason T.L. Wang, Graham Williams, Xindong Wu, and Osmar R. Zaiane.

Information about guest editors

Mohammed J. Zaki is currently an assistant professor of Computer Science at Rensselaer Polytechnic Institute. He received his M.S. and Ph.D. degrees in Computer Science from the University of Rochester in May 1995 and July 1998, respectively. His research interests include the design of efficient, scalable, and parallel algorithms for various data mining techniques. He is specially interested developing novel data mining techniques for applications like bioinformatics, web mining, and materials informatics.

Dr. Zaki received a National Science Foundation CAREER Award in 2001 for his work on parallel and distributed data mining; he has published over 50 papers in this area. He is an editor of the book, "Large-scale Parallel Data Mining," LNAI Vol. 1759, Springer-Verlag, 2000. In the past he has co-chaired several workshops in High-Performance, Parallel and Distributed Data Mining. He is a member of the ACM (SIGKDD, SIGMOD), and IEEE (IEEE Computer Society).

Yi Pan is an associate professor in the Department of Computer Science at Georgia State University. Previously, he was a faculty member in the Department of Computer Science at the University of Dayton. He received his B.Eng. degree in Computer Engineering from Tsinghua University, China, in 1982, and his Ph.D. degree in Computer Science from the University of Pittsburgh, USA, in 1991.

He has published more than 110 research papers. He has received many awards including Visiting Researcher Support Program Award from the International Information Science Foundation (2001), Outstanding Scholarship Award of the College of Arts and Sciences at University of Dayton (1999), the Japanese Society for the Promotion of Science Fellowship (1998), AFOSR Summer Faculty Fellowship (1997), NSF Research Opportunity Award (1994, 1996), Andrew Mellon Fellowship from Mellon Foundation (1990), the best paper award from PDPTA '96 (1996), and Summer Research Fellowship from the Research Council of the University of Dayton (1993). Dr. Pan is currently an area editor-in-chief of the Journal of Information, an associate editor of IEEE Transactions on Systems, Man, and Cybernetics, an editor of the Journal of Parallel and Distributed Computing Practices, an associate editor of the International Journal of Parallel and Distributed Systems and Networks, and on the editorial board of the Journal of Supercomputing. He has also served as a guest editor of special issues for several journals. Dr. Pan is a senior member of IEEE and a member of the IEEE Computer Society.