# Workshop Report: Large-Scale Parallel KDD Systems

Mohammed J. Zaki
Computer Science Department
Rensselaer Polytechnic Institute, Troy, NY 12180
zaki@cs.rpi.edu,
http://www.cs.rpi.edu/~zaki

Ching-Tien (Howard) Ho
IBM Almaden Research Center
650 Harry Rd., San Jose, CA 95120
ho@almaden.ibm.com

## 1. INTRODUCTION

With the unprecedented rate at which data is being collected today in almost all fields of human endeavor, there is an emerging economic and scientific need to extract useful information from it. For example, many companies already have data-warehouses in the terabyte range (e.g., FedEx, Walmart). The World Wide Web has an estimated 800 million web-pages. Similarly, scientific data is reaching gigantic proportions (e.g., NASA space missions, Human Genome Project). High-performance, scalable, parallel and distributed computing is crucial for ensuring system scalability and interactivity as datasets continue to grow in size and complexity.

To address this need the authors organized the workshop on Large-Scale Parallel KDD Systems, which was held in conjunction with the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, on August 15th, 1999, San Diego, California.

The goal of this workshop was to bring researchers and practitioners together in a setting where they could discuss the design, implementation, and deployment of large-scale parallel knowledge discovery (PKD) systems, which can manipulate data taken from very large enterprise or scientific databases, regardless of whether the data is located centrally or is globally distributed. Relevant topics identified for the workshop included:

- How to develop a rapid-response, scalable, and parallel knowledge discovery system that supports global organizations with terabytes of data.
- How to address some of the challenges facing current state-of-the-art data mining tools. These challenges include relieving the user from time and volume constrained tool-sets, evolving knowledge stores with new knowledge effectively, acquiring data elements from heterogeneous sources such as the Web or other repositories, and enhancing the PKD process by incrementally updating the knowledge stores.
- How to leverage high performance parallel and distributed techniques in all the phases of KDD, such as initial data selection, cleaning and preprocessing, transformation, data-mining task and algorithm selection and its application, pattern evaluation, management of discovered knowledge, and providing tight coupling between the mining engine and database/file server.
- How to facilitate user interaction and usability, allowing the representation of domain knowledge, and to maximize understanding during and after the process. That is, how to build an adaptable knowledge engine which supports business decisions, product creation and evolution, and leverages information into usable or actionable knowledge.

The workshop attracted around 50 participants, from academia, industry and government labs., underscoring the active interest in this topic from a variety of perspectives.

The program of the workshop included eight contributed papers, four invited talks and a closing panel on the future of large-scale KDD systems. The online proceedings for the contributed papers and invited talks is available at the workshop URL at http://www.cs.rpi.edu/~zaki/WKDD99. A hard copy is also available upon request [13].

The workshop presentations were organized into four sessions. Each session began with an invited talk, followed by two papers on a common theme. We will summarize the workshop contents in the rest of the report.

## 2. INVITED TALKS

The first talk was given by Reagan Moore, from San Diego Supercomputer Center. Reagan talked about the implementation of data grids [7], which are inherently distributed and heterogeneous systems that link massive data and computational resources. Such data grids provide support for information discovery, remote processing capability, and data analysis and mining. Functionalities provided by data grids include name transparency (attribute based data access), location transparency (local or remote datasets), protocol transparency (file systems, databases, or archives) and time transparency (optimized retrieval time via data caching, replication, aggregation, filtering and parallel I/O).

Umeshwar Dayal from Hewlett-Packard Labs. summarized the requirements of business intelligence applications, and described a data warehouse based architecture for meeting those requirements [2]. He stressed the need to focus on the entire end-to-end process of KDD, with parallel algorithms as just one step, and he noted the fact that due to rapid data flow rates and high data volume, today's business intelligence applications require continuous analysis and mining. On the architectural side he focused on fast data cube construction, retirement policies for moving data from the warehouse on to an archive, data mining using cubes and OLAP tools, and distributed warehousing and mining.

The third talk was by Graham Williams from the Cooperative Research Center for Advanced Computational Systems, Australia. He presented Data Miner's Arcade [11], a java-based platform-independent system for integrating multiple analysis and mining tools, using a common API, and providing seamless data access across multiple systems. Components of the DM Arcade include parallel algorithms (e.g., BMARS - multiple adaptive regression B-splines), virtual environments for data visualization, and data management for mining.

Robert Grossman and Yike Guo, in the final talk, discussed the issues and challenges in developing an open knowledge network for managing, mining, and modeling massive distributed datasets for wide-area distributed mining [5]. They noted that recent computing trends have made such a system within reach, since powerful PC clusters with fast networks can be built relatively cheaply, and the next-generation Internet makes it possible to mine distributed sites. They also presented the Terabyte Challenge Testbed, an international network of sites, for implementing and testing wide-area mining.

## 3. CONTRIBUTED PAPERS

The contributed papers were organized into four sessions spanning major data mining techniques from mining frameworks to associations, sequences, clustering and classification. Zaki chaired the two morning sessions, while Howard chaired the two afternoon sessions.

### 3.1 Mining Frameworks

The first paper of the session was by Bailey et al. [1]. They described the implementation of Osiris, a data server for wide-area distributed data mining, built upon clusters, meta-clusters (with commodity network like Internet) and super-clusters (with high-speed network). Osiris addresses three key issues: What data layout should be used on the server? What tradeoffs are there in moving data or predictive models between nodes? How data should be moved to minimize latency; what protocols should be used? Experiments were performed on a wide-area system linking Chicago and Washington via the NSF/MCI vBNS network.

Parthasarathy et al. [9] presented InterAct, an active mining framework for distributed mining. Active mining refers to methods that maintain valid mined patterns or models in the presence of user interaction and database updates. The framework uses mining summary structures that are maintained across updates or changes in user specifications. InterAct also allows effective client-server data and computation sharing. Active mining results were presented on a number of methods like discretization, associations, sequences, and similarity search.

### 3.2 Association Rules

Morishita and Nakaya [8] opened the second session with a novel parallel algorithm for mining correlated association rules. They mine rules based on the chi-squared metric that optimizes the statistical significance or correlation between the rule antecedent and consequent. A parallel branch-and-bound algorithm was proposed that uses a term rewriting technique to avoid explicitly maintaining list of open and closed nodes on each processor. Experiments on SMP platforms (with up to 128 processors) show very good speedups.

Shintani and Kitsuregawa [10] propose new load balancing strategies for generalized association rule mining using a gigabyte-sized database on a cluster of 100 PCs connected with an ATM network. In generalized associations the items are at the leaf levels in a hierarchy or taxonomy of items, and the goal is to discover rules involving concepts at multiple (and mixed) levels. They show that load balancing is crucial for performance on such large-scale clusters.

### 3.3 Clustering and Sequences

Dhillon and Modha [3] parallelized the K-Means clustering algorithm on a 16 node IBM SP2 distributed-memory system. They exploit the inherent data parallelism of the K-Means algorithm, by performing the point-to-centroid distance calculations in parallel. They demonstrated linear speedup on a 2GB dataset.

Zaki [12] presented pSPADE, a parallel algorithm for sequence mining. pSPADE divides the pattern search space into disjoint, independent sub-problems based on suffix-classes, each of which can be solved in parallel in an asynchronous manner. Task parallelism and dynamic inter- and intra-class load balancing is used for good performance. Results on a 12 processor SMP using up to a 1 GB dataset show good speedup and scaleup.

### 3.4 Classification

Goil and Choudhary [4] implemented a parallel decision tree classifier using the aggregates computed in multidimensional analysis or OLAP. They compute aggregates/counts per class along single dimensions, which can be used for computing the attribute split-points. Communication is minimized by coalescing messages and is done once per tree level. Experiments on a 16 node IBM SP2 were presented.

In the final paper, Hall et al. [6] describe distributed rule induction for learning a single model from disjoint datasets. They first learn local rules from a single site; these are merged to form a global rule set. It is shown that while this approach promises fast induction, accuracy tapers off (as compared to directly mining the whole database) as the number of sites increases. They suggested some heuristics to minimize this loss in accuracy.

## 4. LARGE-SCALE DATA MINING: WHERE IS IT HEADED?

This was the topic of the closing panel. The panelists were:

- Vipin Kumar, University of Minnesota
- Ron Musick, Lawrence Livermore National Labs.
- Foster Provost, New York University
- Mohammed Zaki, Rensselaer Polytechnic Institute

The panelists were asked to formulate their position on three main topics. The first dealt with questions like: Is large-scale necessary (can/should we mine terabyte datasets)? Is sampling enough for most cases? What is the role of parallelism/distributed computing in mining?

The second set of questions pertained to the nature of current research. Questions posed were: Shall we move beyond mining algorithms? Is there any interesting research in the other steps of KDD or is it just a matter of implementing it?

The final topic was on the future of large-scale data mining. Relevant questions were: What are the requirements of large-scale mining? How can these requirements be met? What is our vision for a complete large-scale mining system? Can we handle non-traditional datasets (beyond relational tables)? What are the challenges in such new domains?

Foster Provost opened the panel session by challenging the need for terabyte-sized mining. He pointed that for many datasets sampling should be sufficient, but there remain cases where sampling is not adequate. These cases are not well understood. Thus, a better treatment of sampling is critically needed. He concluded with why distributed data mining has implications far beyond just scaling up (which he noted are not being addressed yet to any real extent). Distributed mining is clearly necessary for leveraging the vast amount of data and background knowledge distributed across a local network or across the Internet.

Ron Musick addressed the panel topics within the context of large-scale mining of scientific datasets. He showed the great disparity between scientific and business data. Scientific mesh data is very dense (millions or billions of zones), very large (petabytes), and almost no research has addressed mining such data. Important requirements include preparing data for mining, parallelism (which is a must), sampling, output uncertainty, model flux, etc. He also noted that there is little to no room for DBMS support, since current DBMS are ill-equipped to handle such data.

Vipin Kumar approached the topics from an algorithmic perspective. He pointed that memory limitations cause sequential algorithms to make many expensive data scans; parallel computing offers the promise of handling much larger data sets in a fraction of the time, which is crucial for interactive exploration. After discussing his experience parallelizing decision tree and associations methods, he pointed out opportunities for parallel computing in pre-processing steps such as feature selection, etc.

Zaki talked about the design and implementation of a large-scale high-performance parallel KDD system. Such a system should support mining over terabyte-sized datasets, centralized or distributed data, incremental changes, and heterogeneous sources. It should support all phases of KDD (pre/post-processing and mining) and should be modular and customizable. He pointed out open research issues in fast algorithms, parallelism and scalability, data locality and type, incremental and interactive methods, database integration, and understandability and usability.

Panelists' presentations were followed by a general Q & A session with active participation by the attendees. The need for large-scale mining systems was further highlighted in the ensuing discussion.

## 5. CONCLUSIONS

We conclude by observing that the need for large-scale KDD systems is real and immediate. Parallel and distributed computing is essential for providing scalable, incremental and interactive mining solutions. The field is in its infancy (as is most of data mining), and offers many interesting research directions to pursue. We hope that this workshop has been successful in bringing to surface the requirement and challenges in large-scale parallel KDD systems.

For latest developments in this area, look out for the new book, edited by the authors, entitled "Large-Scale Parallel Data Mining," to be published as Volume 1759 in Springer-Verlag's LNAI series, Feb. 2000. It contains revised versions of the workshop papers. With the addition of several invited chapters, it represents the state-of-the-art in parallel and distributed data mining methods. It should be useful for both researchers and practitioners interested in the design, implementation, and deployment of large-scale, parallel knowledge discovery systems.

More information on the workshop itself is available at http://www.cs.rpi.edu/∼zaki/WKDD99.

## Acknowledgements

We would like to thank all the invited speakers, authors and participants for contributing to the success of the workshop. Special thanks are due to the program committee for their support and help in reviewing the submissions.

## 6. REFERENCES

[1] S. Bailey, E. Creel, R. Grossman, S. Gutti, and H. Sivakumar. A high performance implementation of the data space transfer protocol (DSTP). In [13].

[2] U. Dayal. Large-scale data mining applications: Requirements and architectures. In [13].

[3] I. S. Dhillon and D. S. Modha. A data clustering algorithm on distributed memory machines. In [13].

[4] S. Goil and A. Choudhary. Efficient parallel classification using dimensional aggregates. In [13].

[5] R. Grossman and Y. Guo. Communicating data mining: Issues and challenges in wide area distributed data mining. In [13].

[6] L. O. Hall, N. Chawla, K. W. Bowyer, and W. P. Kegelmeyer. Learning rules from distributed data. In [13].

[7] R. Moore. Collection-based data management. In [13].

[8] S. Morishita and A. Nakaya. Parallel branch-and-bound graph search for correlated association rules. In [13].

[9] S. Parthasarathy, S. Dwarkadas, and M. Ogihara. Active data mining in a distributed setting. In [13].

[10] T. Shintani and M. Kitsuregawa. Parallel generalized association rule mining on large scale PC cluster. In [13].

[11] G. Williams. Integrated delivery of large-scale data mining systems. In [13].

[12] M. J. Zaki. Parallel sequence mining on SMP machines. In [13].

[13] M. J. Zaki and C.-T. Ho, editors. *Workshop on Large-Scale Parallel KDD Systems (with KDD99)*. Technical Report 99-8. Computer Science Dept., Rensselaer Polytechnic Institute, August 1999.