Editorial: Online, Interactive, and Anytime Data Mining

Mohammed J. Zaki
Computer Science Department
Rensselaer Polytechnic Institute
Troy NY 12180 USA
zaki@cs.rpi.edu

1. Introduction

Discovering interesting nuggets in large databases can be very compute and I/O intensive. This high computational cost may be acceptable when the database is static since the discovery is done off-line, and several approaches to this problem have been presented in the literature. However, many domains with streaming data, such as electronic commerce, web mining, stock analysis, intrusion detection, fault monitoring, etc., impose time and memory constraints on the mining process. In such domains, where the databases are updated continuously and user interactions modify the search parameters, running the discovery program from scratch is unfeasible, and having the user wait inordinately long is unacceptable. Hence, there is a need for techniques that can effectively mine or handle:

- Streaming data (online updations/deletions)
- User interactions (modifying/constraining the search space)
- Anytime response (partial/approximate results)

This special issue presents some of the current work addressing the above problems. It includes one invited paper and four contributed papers. A brief overview of these articles follows.

2. Special Issue Articles

In their invited paper, "Mining Data Stream under Block Evolution," Venkatesh Ganti, Johannes Gehrke and Raghu Ramakrishnan present incremental mining techniques for the block evolution model, i.e., where a dataset is updated periodically through additions or deletions of a block of transactions; this is common practice in today's data warehouses. They describe their work on GEMM and FOCUS frameworks for data mining model maintenance (frequent sets and decision trees) and change detection under block evolution model, respectively.

In the first contributed paper entitled, "Towards Effective and Interpretable Data Mining by Visual Interaction," Charu Aggarwal discusses the limitation of purely automated mining techniques in finding interesting or useful nuggets, since it is often difficult to objectively specify such constraints. He goes on to show how visual feedback in the intermediate mining steps can help combine the computational power of machines with the intuitive and subjective abilities of human users. The paper presents visual methods for the clustering and nearest neighbor problems in high-dimensional spaces, and discusses extensions to other problems like classification and associations.

In the paper, "Requirements for Clustering Data Streams," Daniel Barbará first discusses the desiderata for clustering and outlier detection methods when dealing with streaming data. The paper goes on to discuss existing algorithms on incremental clustering and discusses whether/how they meet these requirements. It shows how the new fractal clustering method meets most of the challenges in data stream clustering.

In the third paper, "Interactive Mining and Knowledge Reuse for the Closed Itemset Incremental Mining Problem," Luminita Dumitriu describes a user-driven approach to mining association rules. The paper uses frequent closed itemsets and the associated concept lattice to incrementally mine rules pertaining to newly added items, reusing previous knowledge where possible. It also allows contractions where certain items are removed from consideration. The paper thus addresses issues of incrementality as well as interactivity.

In the final paper entitled, "MobiMine: Monitoring the Stock Market from a PDA," Hillol Kargupta, Byung-Hoon Park, Sweta Pittie, Lei Liu, Deepali Kushraj, and Kakali Sarkar present their work on mining data streams in a mobile environment. The paper describes a prototype system that allows mobile users to monitor time-critical financial data on a PDA via personalization. It uses the Fourier analysis of decision trees to compress and communicate them to mobile users over limited bandwidth networks.

3. Reports from KDD-2001

In addition to the special issue articles, this volume of SIGKDD Explorations also includes reports from KDD Cup 2001, and the panel and workshops from SIGKDD Conference, 2001. Cheng et al. describe the KDD Cup competition.

Next come reports from the workshops associated with SIGKDD 2001 Conference. Zaïane and Simoff present a summary of the 2nd Multimedia Data Mining workshop. Kamath describes 4th Workshop on Mining Scientific Datasets. Eick and Keim discuss the workshop on Visual Data Mining. Zaki, Wang and Toivonen present the report from the Data Mining in Bioinformatics workshop.

Domingos summarizes the SIGKDD01 panel discussion on "When and How to Subsample." Gehrke reports on the panel on New Research Directions in KDD.

4. Other Reports

We are also happy to have a report by Simoff on the International Workshop on Visual Data Mining, which was held in conjunction with the 12th European Conference on Machine Learning (ECML'01) and the 5th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'01).

About the Guest Editor:

Mohammed J. Zaki is an Assistant Professor in the Computer Sciences Department at Rensselaer Polytechnic Institute. He

received his M.S. ('95) and Ph.D. ('98) degrees in computer science from the University of Rochester. He is leading the SPIDER (Scalable, Parallel and Interactive Data mining and Exploration at Rensselaer) data mining project at RPI. He is especially interested developing novel data mining techniques for applications like bioinformatics, web mining, material informatics, etc. He recently received a CAREER Award from the National Science Foundation for his research on "Application-Oriented Large-Scale Parallel Data Mining." He has published over 60 papers on data mining, has co-edited 4 books (including "Large-scale Parallel Data Mining," Springer-Verlag, 2000), has co-chaired/guest-edited several journal special issues and workshops on high performance data mining and bioinformatics, and has served as program committee member on many international conferences on data mining.