Workshop Report: 2000 ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery

Dimitrios Gunopulos
University of California
Riverside, CA
dg@cs.ucr.edu

Rajeev Rastogi
Bell Laboratories
Murray Hill, NJ 07974
rastogi@research.bell-labs.com

The 2000 Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD) was held on May 14, 2000 in Dallas in conjunction with the ACM SIGMOD 2000 conference. In the call for papers, in addition to novel data mining algorithms and experiences in deploying data mining systems and applications, we solicited papers on the following three important topics:

- **Foundations of data mining.** There are numerous techniques that fall under the broad umbrella of "data mining" algorithms (e.g., association rules, classification, clustering). A framework or model that unifies these disparate techniques would be a major contribution to the field (e.g., is data mining simply the process of generating "good" summaries of the data?).
- **Data mining techniques for vertical applications.** These are techniques that focus on specific applications and exploit domain-specific knowledge to mine the data more effectively. Examples of such applications include E-commerce, fraud detection (telecom, credit-card), network management, customer relationship management, etc.
- **Novel Data Applications.** These are data mining techniques and applications that focus on exploring novel kinds of data, such as geographical data, environmental data, or medical data.

We received 19 submissions, of which we selected 11 papers for inclusion in the workshop proceedings and presentation at the workshop. The workshop program also included 2 invited talks by H. V. Jagadish (University of Michigan, Ann Arbor) and Diane Lambert (Bell Labs, Lucent) and a panel moderated by Umeshwar Dayal (HP Labs).

Invited Talks.

The first invited talk by H. V. Jagadish was titled "Incompleteness in Data Mining". The main claim made by Jagadish was that techniques that can find SOME interesting patterns cheaply are much more valuable than an exhaustive enumeration of ALL patterns, which can be extremely expensive. Further, knowledge discovery is most effective when the human analyst is involved in the endeavor. For this, data mining techniques need to be interactive, delivering real-time responses and feedback, which is only possible with incomplete and approximate answers. The case for incompleteness was made using the notion of *fascicles*, which are subsets having very similar values for many attributes. Jagadish's experience was that randomized algorithms for finding fascicles were as effective as exhaustive algorithms. Jagadish concluded his talk by listing the following four desiderata for frameworks that permit incompleteness to be exploited: (1) *Tunability* or the ability to specify the degree of incompleteness, (2) *Incrementality* or the ability to exploit previous computations in subsequent iterations, (3) *Focusing* or the ability to incorporate user-specified constraints into the computation, and (4) *Quality loss guarantees* or the ability to quantify analytically the loss of quality corresponding to a chosen degree of incompleteness.

The second invited talk by Diane Lambert focused on the role of statistics in data mining. Diane described statistics as a means to extract information from data that are noisy or uncertain. She discussed a successful application of statistics to detecting telecommunications fraud as it happens. The basic approach is to use statistical methods for predicting legitimate calling behavior for each customer and score the customer's calls against that distribution. According to Diane, massive data raise many questions beyond analysis that could be much better addressed by statisticians and computer scientists together than by either group working alone. As an example, while privacy and confidentiality issues have been discussed by statisticians for decades, most proposals are not feasible for large sets of data. Diane finally cited other branches of mathematics that are fundamental to understanding massive data sets: probability, bayesian networks, coding, compression, approximation theory, functional analysis and algebra.

Papers.

The papers were presented in two sessions: the first session focused on novel algorithms for data mining, while the second focused on data mining applications.

Novel Algorithms. Wang and Zaniolo propose a unified solution for data mining, complex OLAP queries and special data types such as time series. The solution is based on User-Defined Aggregates (UDAs) expressed in a new SQL-like language called AXL. Pei, Han and Mao focus on using closed itemsets to substantially reduce the number of association rules, and propose an efficient algorithm for mining closed itemsets. Jermaine and Miller present a new data reduction method for efficiently constructing a high-dimensional, joint probability distribution that is subsequently used to find approximate answers to high-dimensional data cube queries. Wang and Wang propose a multilevel filtering scheme for high-dimensional nearest neighbor search in which successively better approximations are used to refine the candidate set. Finally, Stanoi, Agrawal and El Abbadi exploit results from computational geometry for answering *reverse nearest neighbor* (RNN) queries which is the problem of finding the set of data points that have the query point as the nearest neighbor.

Applications. Mining Web access logs to generate user profiles is one possible approach to Web personalization. Joshi and Krishnapuram analyze Web logs to discover user sessions and then use a robust fuzzy clustering algorithm to extract user profiles. Chawla et. al. present spatial statistical techniques which can effectively model the notion of spatial-autocorrelation and apply it to the problem of predicting bird nest locations in a wetland. Ordonez, Santana and Braal use association rules to discover patterns in medical data which are typically small, have high dimensionsionality, and numerical, categorical and image attributes. Agichtein, Eskin and Gravano describe the Snowball system that extracts structured relations from unstructured text collections, that can subsequently be queried or mined. Wong and Fu address the problem of finding the structure of Web documents and present a hierarchical structure to represent the relation among text data in Web documents. Finally, Hogl, Stoyan and Muller present the Knowledge Discovery Assistant (KDA) that enables business users to interact with the data mining system using a high-level language.

Panel.

The panel included Umesh Dayal (Moderator), Jawei Han, Raymond Ng, Dimitrios Gunopulos and Rajeev Rastogi. One of the topics that panelists discussed was possible new challenging and unusual applications in Data Mining, and whether these applications raise new architectural issues that need to be addressed.

Han, after providing a brief overview of the history of the workshop, offered suggestions for making data mining more successful in the commercial marketplace – these included integrating it closely with the database system, incorporating domain-specific knowledge into the data mining tools (generic tools are too simple for sophisticated applications), and building mining functions into information services like Web search engines. Ng presented the outline of a unified framework to support many major mining and analysis tools. He also presented a related algebra that can be used to describe tasks and operations in this general framework.

Gunopulos gave three examples of interesting new applications. Efficient interactive mining of large datasets will allow more focused, "man in the loop" mining. Faster (online or approximation) algorithms are needed, as well as new data approximation techniques. Integrating mining operations with the database engines will allow easier deployment of models and speed up execution by taking advantage of free parallelism. Possible approaches are finding fundamental operations to push into SQL, and expand SQL to build or describe models. Interdisciplinary problems provide important motivation and validation of data mining research. One such problem is mining spatio-temporal data that describe wildlife fires.

Rastogi discussed applications of data mining in the internet. These applications fall into three categories: E-commerce, information retrieval (search) and network management. Examples of E-commerce applications are generation of user profiles and targetted Web advertizing based on user access patterns that can be extracted from Web logs. Information retrieval applications include automatic construction topic hierarchies, and identification of hubs and authorities for specific topics on the Web. Finally, network management applications include content delivery, identification of the best (least congested) server and path along which to deliver content and quickly identifying failure of links and nodes in the network. One of the audience questions related to the difference between Web mining and traditional data mining – Rastogi pointed to the presence of hyperlinks in Web documents as one major difference. Another concern expressed by the audience related to the issue of privacy in Web mining.

The panel concluded with suggestions on how the workshop could be improved. Audience suggestions included fewer paper presentations, and more panels, tutorials (on new and interesting topics) and discussion sessions for future workshops to improve the cross-fertilization between areas like AI, statistics and databases.

About the Authors.

Dimitrios Gunopulos received a Ph.D. in Computer Science from Princeton University in 1995. Prior to that he received a M.A. in Computer Science from Princeton and a Diploma in Computer Engineering from the University of Patras. Dr. Gunopulos is currently an Assistant Professor in the Department of Computer Science and Engineering at the University of California, Riverside. He has also held positions at IBM Almaden and the Max-Plank-Institut for Informatik. His research interests include data mining, databases, algorithms, and computational geometry. His current research focuses on techniques for approximating range queries, on applying data mining techniques to geospatial data, and on database indexing techniques.

Rajeev Rastogi is the Director of the Internet Management Research Department at Bell Laboratories, Lucent Technologies. He received the B. Tech degree in Computer Science from the Indian Institute of Technology, Bombay in 1988, and the masters and Ph.D. degrees in Computer Science from the University of Texas, Austin, in 1990 and 1993, respectively. He joined Bell Laboratories in Murray Hill, New Jersey, in 1993 and became a Distinguished Member of Technical Staff (DMTS) in 1998.

Rajeev Rastogi is active in the field of databases and has served as a program committee member for several conferences in the area. His writings have appeared in a number of ACM and IEEE publications and other professional conferences and journals. His research interests include database systems, storage systems, knowledge discovery and network management. His most recent research has focused on the areas of network management, data mining, high-performance transaction systems, continuous-media storage servers, tertiary storage systems, and multidatabase transaction management.