Workshop Report: 1999 ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery

Kyuseok Shim
Lucent Bell Laboratories
600 Mountain Ave, Murray Hill, NJ 20974
shim@research.bell-labs.com

Ramakrishnan Srikant
IBM Almaden Research Center
650 Harry Road, San Jose, CA 95120
srikant@us.ibm.com

The Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD) was started three years ago as a forum for database researchers to discuss and investigate data mining research issues. The first two workshops primarily focused on advances in data mining algorithms. In 1998, the workshop additionally emphasized data mining systems, applications and solutions. This year, as the field matured, we placed an additional emphasis on future directions.

After careful review by the program committee, 12 papers were chosen for presentation and inclusion in the workshop proceedings out of 21 submissions. In addition, the program included two invited talks by Heikki Mannila of Microsoft Research and George John of E.piphany Inc., and a round-table on "lessons learnt and future directions" moderated by Umeshwar Dayal. There were around 140 attendees at the workshop.

Invited Talks

The first invited talk, "Theoretical Frameworks for Data Mining", was presented by Heikki Manila. Requirements for a good theory, apart from being nice, clean, useful, etc., include recognizing the special features of data mining: data mining is a process; the user is part of the loop; there is not a single criterion of what is interesting; background knowledge; probabilistic nature of the patterns; different tasks of data mining; and different forms of data. He discussed the pros and cons of five candidates for "a theory of data mining":

- The pattern discovery approach yields useful algorithmic insights such as level-wise search and monotonicity.
- Probabilistic approaches (e.g. hierarchical Bayesian models) consider data mining as finding the joint distribution of the random variables.
- With the data compression/MDL approach, a good representation is one that makes it possible to code the data using few bits.
- The microeconomic view of data mining can formalize data mining problems such as pattern discovery and clustering as a nonlinear optimization problem.
- Inductive databases suggest an architecture for data mining systems, but seem difficult to implement.

He concluded his talk by arguing that while the practice of data mining seems to be proceeding nicely without theory, there is really no coherent whole. Hence we need a theory to guide the development of practice and connect existing areas.

The second invited talk, "Data Mining in the Real World", was presented by George H. John. He illustrated with examples that average citizens are frequently exposed to data mining techniques without noticing them in their daily life. For instance, sophisticated fraud detection systems keep an eye on hundreds of millions of accounts worldwide looking for unusual buying behavior. He classified data mining into three varieties. Ad-hoc data mining is descriptive and supports decisions indirectly. In prescriptive mining projects, the results of mining directly suggest actions. Embedded data mining is invisible, e.g. campaign management software automatically tunes real campaigns based on test results. For widespread use of data mining in businesses, he suggested that data mining software should move from the "ad-hoc" model to the "embedded" model.

Papers

There were 12 accepted papers in 4 categories: applications/experiences, association rules, algorithms, and future directions. The papers can be downloaded from http://www.almaden.ibm.com/cs/dmkd/.

Applications/Experiences. S. Morishita et al. illustrated the use of association rules and clustering technologies to derive biological facts efficiently in gene expression databases. F. Giannotti et al. presented two case studies that show how a suitable integration of deductive reasoning and inductive reasoning provides a powerful formalism that allows a high degree of expressiveness in specifying expert rules, or business rules, the ability to formalize the overall KDD process and the separation of concerns between the specification level and the mapping to the underlying databases

and data mining tools.

Association Rules. D. Meretakis et al. showed that various classifiers (decision-tree, Naive Bayes, Bayesian networks, nearest neighbor) can be seen as the mining and use of frequent itemsets. B. Goethals et al. discussed the tradeoffs between a priori versus a posteriori filtering of association rules. D. Shah et al. formalized principles that can be used to eliminate redundant rules.

Algorithms. M. Dash et al. showed that entropy is a good measure for dimensionality reduction in unsupervised learning. G. Grahne et al. examined the problem of efficiently computing correlated sets satisfying given constraints. J. Han et al. presented a method for mining significant patterns and successful actions in a large planbase using a divide-and-conquer strategy.

Future Directions. T. Johnson et al. proposed a toolkit for data analysis and mining. Their intuition was that data space partitioning can be used to unify many aspects of data mining and analysis. D. Barbara suggested using the fractal dimension to uncover anomalous patterns in datasets. V. Raman et al. showed a scalable spreadsheet for interactive data analysis that combines exploration, grouping and aggregation. H. Wang et al. showed how user-defined aggregates can be used for data mining algorithms.

Round-Table

The round-table, on "Lessons Learnt and Future Directions" was organized and moderated by Umeshwar Dayal. The participants included Alon Levy, H.V. Jagadish, Dennis Shasha, Kyuseok Shim and Ramakrishnan Srikant.

- Alon Levy made the distinction between mining data on the web and mining processes. He argued that while everyone knows about mining data, mining processes is very new and yet relatively unexplored. Examples include mining what people are looking for on the web, where are the hotspots on the web, and where people are turning to when different events happen.
- H.V. Jagadish pointed out that much of the data to be mined does not reside in databases. Hence more research needs to be done to support data cleaning, as well as interactive exploration of data.
- Dennis Shasha argued that data mining research continues to generate many new techniques and products. This is good, because some of these will be widely useful. This is also bad, because potential users have difficulty finding out which technology will work best for them. The community needs a service: a database of case studies, rich enough to guide users to the best technology for their needs.
- Kyuseok Shim presented the following future research directions: (1) scaling up existing algorithms from AI, ML and IR communities, (2) new methodologies and applications, (3) incorporation of constraints into existing data mining algorithms, (4) tight-coupling with DBMS and (5) Web mining.
- Srikant used examples to show that the "bar" for success in customer engagements is very high, often requiring a champion of the process and/or innovative application of techniques. Three approaches to crossing the chasm from the "early adopter" stage into the mainstream might be partnerships with application vendors, piggy-backing on the success of database, and the web. In particular, the web enables mining applications that were previously infeasible.

About the authors:

Kyuseok Shim is a Member of Technical Staff (MTS) at Bell Laboratories. He is currently involved with the Serendip data mining project. Before that, he worked for the Quest data mining project at IBM Almaden Research Center. He has served as a program committee member on ICDE '97, KDD '98, SIGMOD '99, SIGKDD '99, ICDE '00 and VLDB '00 conferences. He gave a data mining tutorial at CIKM '98, ICDE '99 and SIGKDD '99 conferences. He has been working in the area of databases focusing on data mining, data warehousing, query processing and query optimization.

Ramakrishnan Srikant is a Research Staff Member at IBM Almaden Research Center. He has received several Outstanding Technical Achievement Awards and Invention Achievement Awards from IBM, and was named a Master Inventor in 1999. Srikant was co-chair of the 1999 ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. He has served on the program committees of KDD '96, KDD '97, KDD '98 and SIGKDD '99, as well as several workshops. His current interests include data mining, text and web mining, and electronic commerce.