KDD-99: The Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining

Surajit Chaudhuri
Microsoft Research
One Microsoft Way
Redmond, WA 98052
surajitc@microsoft.com

David Madigan
AT&T Labs
180 Park Avenue
Florham Park, NJ 07932
madigan@research.att.com

Usama Fayyad
Microsoft Research
One Microsoft Way
Redmond, WA 98052
fayyad@acm.org

ABSTRACT

KDD-99 was the fifth conference in the KDD series attracting over 200 high quality submissions and almost 600 attendees. Here we describe some of the highlights of the technical program.

Keywords

KDD Conference overview, ACM SIGKDD.

1. INTRODUCTION

The demand for effective tools for knowledge discovery and data reduction and mining is large and growing rapidly, especially as a crucial component of data-warehouses. Meeting this demand requires input from multiple disciplines including databases, machine learning, pattern recognition, and statistics. Since 1995, the annual KDD conferences have served as a high-quality technical venue where these disciplines can interact and present the latest advances in algorithms and applications.

For its first four years as a formal conference, KDD was organized by AAAI and collocated with other major conferences in AI, statistics, and databases. In recognition of the growth of the community and to better serve its distinct needs, the ACM Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD) was formed. Starting with 1999, the KDD conference has become an ACM conference and is the official annual conference of SIGKDD. This is an important step in ensuring the successful continuation of the conference series and provides access to the extensive resources of the ACM. It also better reflects the breadth of the field and its overlap with many aspects of computing. The formation of SIGKDD represents a critical milestone in the growth of this nascent field as an area of research in its own right. We are grateful to AAAI for their early support and for their continued sponsorship of the conference.

2. CONFERENCE PROGRAM

The program for this year's conference certainly demonstrated continued extraordinary breadth. Unlike previous years, this year submissions were allowed to be significantly longer (20 pages) in order to enable a more careful review. There was no separate submission format for poster papers. The program committee accepted as poster papers submissions that were judged to have elements of novelty and promise but that were not judged to be quite as mature as the full-length papers. Each poster paper received an associated 2-minute presentation in the plenary program to insure sufficient exposure of the work, in addition to being presented in the 3 hour poster session. The poster session was very well-attended and gave plenty of opportunities for discussion: allowing authors to elaborate at length, and attendees to ask more detailed and focused questions than they would be able to ask in a regular presentation session.

The call for papers attracted 237 full-length research papers from 32 countries on 6 continents. Approximately half of all the submissions came from North America. A significant majority of the submissions (63%) classified their contribution as belonging to KDD Techniques[1]. The remaining submissions were distributed among:

- Enterprise Data (13%),
- Implementation Issues (14%),
- and Human Interaction and KDD Processes (10%).

Relatively few submissions considered topics such Data Cleaning, Visualization, KDD Benchmarks, KDD Theory, and Vertical Applications.

Of the 237 submissions, 27 were accepted for full presentation and 25 were accepted for poster presentation. While the program committee's decisions were made strictly based on quality, papers submitted on scalable mining algorithms, interaction between mining & database querying, and case studies enjoyed better than average acceptance rates. In contrast, the program committee was able to recommend acceptance for a smaller fraction of submitted papers on topics such as mining the web and incremental mining. While the variance in acceptance rates can hardly be considered to have any great significance, we thought that these facts might make this report a little more interesting!

As in previous years, a Best Paper Awards Chair, aided by the program co-chairs and members of the awards committee, identified outstanding papers in two categories: fundamental research and applications/applied research. Research papers were judged by the theoretical significance and originality of their contribution. Application papers were judged by the practical significance and current or potential usefulness of the work.

- The winner in the fundamental research category was Pedro Domingos for his paper "MetaCost: A general method for making classifiers cost-sensitive." Honorable mention went

[1] With hindsight, it was a mistake on our part to offer a catch-all phrase such as *KDD Techniques* as one of the subject areas.

to Bill DuMouchel, Chris Volinsky, Ted Johnson, Corinna Cortes, and Daryl Pregibon for their paper "Squashing flat files flatter."

- The winners in the applications and applied research category were Wenke Lee, Sal Stolfo, and Kui Mok for their paper "Mining in a Data-flow Environment: Experience in Network Intrusion Detection." Honorable mentions went to Tom Brijs, Gilbert Swinnen, Koen Vanhoof, and Geert Wets for "Using Association Rules for Product Assortment Decisions: A Case Study" and to Seth Rogers, Pat Langley, and Christopher Wilson for "Mining GPS Data to Augment Road Models."

We are grateful to Gregory Piatesky-Shapiro for serving as this year's Awards Chair. We also thank the members of the awards committee for their help.

3. INDUSTRIAL TRACK

This year, for the first time, the conference featured an industrial track. The industrial track provides a much-needed forum that allows practitioners in the data mining industry to share thoughts and problems with researchers. The track, co-chaired by Jim Gray and Ron Kohavi included nine talks covering a broad spectrum of topics that emphasized application and practice issues in KDD. The talks were selected from 28 submissions. Submissions consisted of extended abstracts (up to five pages in length). We are grateful to Gregory Piatesky-Shapiro for helping the industrial track chairs with the evaluation of industrial track submissions.

4. INVITED TALKS

The conference featured three invited talks. In the first keynote address, "Data Mining: Crossing the Chasm", Rakesh Agrawal (IBM Almaden Research Center), outlined the challenges facing data mining in transitioning from a niche to a mainstream technology. On the second day, Richard Hackathorn (WebFarming.com) emphasized the importance of the convergence of data-warehousing, web information discovery, and information visualization. In the final keynote address, Daryl Pregibon (AT&T Laboratories), reviewed the accomplishments, lessons and future directions of the InfoLab at AT&T Laboratories.

5. PANELS, TUTORIALS, AND EXHIBITS

Three panels were included in the conference program to discuss topics of interest and controversy. Integrating Data Mining into Vertical Solutions panel considered the importance of specialization and embedding of data mining in complete business solutions. Another panel covered the potential dark sides of data mining associated with data snooping and the possibilities of mistaking random noice or chance correlations for "knowledge". This being the tenth anniversary of KDD, the final panel brought together a few of the researchers who have been in the field since its foundation to reflect upon the progress, lack of it, and the challenges remaining.

The tutorial program consisted of six tutorials that covered many topics including: scalable algorithms for classification, clustering, and pattern detection (3 tutorials), mining unstructured data, methods for combining estimators (ensemble methods for prediction), and finally te integration of data mining solutions in the business process.

KDD-99 also included two workshops on topics of interest to the field, a demo session of five major demos which was part of the exhibits program. There exhibits programs was very active and featured 27 exhibitors which represented only a sampling of the rapidly growing space of companies related to or interested in data mining.

Our sincere thanks to functional chairs Padhraic Smyth, Jiawei Han, Rakesh Agrawal and Ismail Parsa for the panel, tutorial, workshop and demos/exhibit program, respectively.

6. PEOPLE

Assembling a conference of this magnitude requires a significant team effort and we have many to thank for their help. We are grateful to the program committee for their high quality work. Reviewing can be a thankless job but the strength of the conference program depends on quality reviews and this year's committee did an outstanding job.

The Microsoft Research Web Support team did a terrific job of putting together and maintaining the home page for the conference as well as providing service using the conference management software, built at Microsoft Research. The software enabled electronic as well as hardcopy submission of papers by authors. The electronically submitted papers were made available for downloading for evaluation. The program committee members were able to submit reviews and to view rankings of papers electronically. We would like to specially thank Jay Grieves and Jonathan Simon of the Web support team at Microsoft Research.

Won Kim, the SIGKDD Chair, was generous with his advice and time in helping us with KDD-99. He also did much of the work in coordinating hotel arrangements, the budget, and other organizational issues. We are grateful to Jenny Zhang, the local arrangements chair and Kyuseok Shim, the proceedings editor for the long hours they spent. We would also like to thank the ACM organizers, primarily Alisa Rivkin and her staff for help with organization, registration, and printing of proceedings and conference program. Special thanks to Donna Baglio at ACM. Eui Kyeong Hong did a lot of work as registration chair, conference treasurer, and coordinator of student volunteers. Foster Provost did an excellent job publicizing the conference and coordinating the design of the KDD Poster. We thank James Gary of Bucket Arts for the proceedings cover and poster design and graphics work. This year's KDD Cup competition was organized and administered by Charles Elkan. We thank R. Uthurusamy, our sponsorship chair, and especially the sponsors who worked with him for their generosity in supporting KDD-99 and SIGKDD. KDD-99 attracted a record $60K in sponsorship support this year.

Finally, we would like to thank the authors and all other contributors for their time and effort in creating such a distinguished program. We hope that the SIGKDD annual conferences continue to be the highest quality technical forum for presenting research and applications in knowledge discovery and data mining.

About the authors:

Surajit Chaudhuri (http://research.microsoft.com/~surajitc) is a senior researcher in the database group at Microsoft Research. His current research interests are in integration of data mining with

database technology, OLAP & data-warehousing, and self-tuning database systems. He is on the editorial board of the Journal of Data Mining and Knowledge Discovery. Surajit served as a Program co-chair of the KDD-99 conference.

David Madigan (http://www.research.att.com/~madigan) works in the Statistics research group at AT&T Labs. Previously he was on the faculty at the Department of Statistics at the University of Washington. His current research interests include collaborative filtering and large-scale statistical modeling. He is the founder of Talaria, Inc., a Seattle-based research company. David served as a Program co-chair of the KDD-99 conference.

Usama Fayyad (http://research.microsoft.com/~fayyad) is a Senior Researcher at Microsoft Research where he heads the Data Mining & Exploration (DMX) Group. His work with Microsoft product groups includes the development of mining components that ship with Microsoft Site Server and on developing scalable algorithms for mining large databases and architecting their fit with server products such as Microsoft SQL Server and OLAP Services. He also leads the core development of data mining algorithms that ship in SQL Server 2000. Prior to joining Microsoft he was at the Jet Propulsion Laboratory, California Institute of Technology where he headed the Machine Learning Systems Group. He is a co-editor of *Advances in Knowledge Discovery and Data Mining* (MIT Press, 1996) and is an Editor-in-Chief of the journal: *Data Mining and Knowledge Discovery*. He served as program co-chair of KDD-94 and KDD-95 and as general chair of KDD-96 and KDD-99.