The Second International Conference on Knowledge Discovery and Data Mining

Sponsored by the American Association for Artificial Intelligence
Portland, Oregon, USA, August 2-4, 1996

Program and Schedule

Contents

Events on Friday August 2
Plenary Session
Invited Talk
Plenary Session
Technology Spotlight T1 (Posters)
Paper Session 1: Scalable Data Mining Systems
Technology Spotlight T2 (Posters)
Paper Session 2A: Scalability and Extensibility
Paper Session 2B: Applications I
Plenary Session
Invited Talk
Plenary Session
Technology Spotlight T3 (Posters)
Paper Session 3: Spatial and Text Data Mining
Technology Spotlight T4 (Posters)
Paper Session 4A: Decision-Tree and Rule Induction
Special Paper Session 4B: Systems for Mining Large Databases
Paper Session 5A: Mining with Noise and Missing Data
Session 5B: Panel Discussion: Systems for Mining Large Databases
Opening Reception & Poster and Demonstration Session
Demonstrations

Events on Saturday August 3
Plenary Session
Invited Talk
Plenary Session
Technology Spotlight T5 (Posters)
Paper Session 6: Prediction and Deviation
Technology Spotlight T6 (Posters)
Paper Session 7A: Prediction
Paper Session 7B: Applications II
Plenary Session
Invited Talk
Plenary Session
Paper Session 8:Combining Data Mining and Machine Learning

Poster Session II
Paper Session 9A: Approaches to Numeric Data
Special Paper Session 9B: Scalable and Distributed Applications of KDD
Paper Session 10A: Pattern-Oriented Data Mining
Session 10B: Panel Discussion
KDD-96 Conference Banquet
Invited Talk

Events on Sunday August 4
Joint UAI-96/KDD-96 Plenary Sessions
Introductory Remarks
Session 11: Learning, Probability, and Graphical Models I
Session 12: Learning, Probability, and Graphical Models II
Summary Panel and Closing Remarks
KDD Wrap-up Business Meeting

Sessions
1: Scalable Data Mining Systems
2A: Scalability and Extensibility
2B: Applications I
3: Spatial and Text Data Mining
4A: Decision-Tree and Rule Induction
4B: Systems for Mining Large Databases
5A: Mining with Noise and Missing Data
5B: Panel Discussion: Systems for Mining Large Databases
6: Prediction and Deviation
7A: Prediction
7B: Applications II
8:Combining Data Mining and Machine Learning
9A: Approaches to Numeric Data
9B: Scalable and Distributed Applications of KDD
10A: Pattern-Oriented Data Mining
10B: Panel Discussion

Technology Spotlight
Technology Spotlight T1 (Posters)
Technology Spotlight T2 (Posters)
Technology Spotlight T3 (Posters)
Technology Spotlight T4 (Posters)
Technology Spotlight T5 (Posters)
Technology Spotlight T6 (Posters)

Friday August 2

8:30 - 9:45 AM

Plenary Session

Room B113-116, Oregon Convention Center

Welcome and Introduction
Evangelos Simoudis, KDD-96 Program Cochair

Invited Talk

Harnessing the Human in Knowledge Discovery
Georges G. Grinstein, University of Massachusetts at Lowell and The MITRE Corporation

Knowledge, the primary goal of data analysis and exploration, is most often discovered by generating information (structure) from data, and then abstracting non-trivial patterns (rules or associations for example) from the information. The discovery process can be done using visualization, data mining, statistics, neural networks, or mathematical modeling and simulation. Visualization is different from the rest, however, in that it is also the actual mechanism by which the analyses and their results can be presented to the user. We will present a brief history of alternative visualizations and how they have been applied to various data visualization problems. The emphasis will be on how exploratory visualization can support the knowledge discovery process, including concept development for database management, database visualizations, and minimally structured dataset visualizations.

9:45 - 10:00 AM

Coffee Break

10:00 - 11:00 AM

Plenary Session

Room B113-116, Oregon Convention Center

10:00 - 10:10 AM

Technology Spotlight T1 (Posters)

Mining Associations in Text in the Presence of Background Knowledge
Ronen Feldman, Bar-Ilan University, Israel and Haym Hirsh, Rutgers University

Undiscovered Public Knowledge: A Ten-Year Update
Don R. Swanson and Neil R. Smalheiser, University of Chicago

Developing Tightly-Coupled Data Mining Applications on a Relational Database System
Rakesh Agrawal and Kyuseok Shim, IBM Almaden Research Center

Mining Entity-Identification Rules for Database Integration
M. Ganesh and Jaideep Srivastava, University of Minnesota; Travis Richardson, Apertus Technologies, Inc.

Static Versus Dynamic Sampling for Data Mining
George H. John and Pat Langley, Stanford University

10:10 - 10:50 AM

Paper Session 1: Scalable Data Mining Systems

An Overview of Issues in Developing Industrial Data Mining and Knowledge Discovery Applications
Gregory Piatetsky-Shapiro, GTE Laboratories; Ron Brachman, AT&T Research; Tom Khabaza, ISL, United Kingdom; Willi Kloesgen, GMD, Germany; and Evangelos Simoudis, IBM Almaden Research Center

Quakefinder: A Scalable Data Mining System for Detecting Earthquakes from Space
Paul Stolorz and Christopher Dean, Jet Propulsion Laboratory, California Institute of Technology

10:50 - 11:00 AM

Technology Spotlight T2 (Posters)

Induction of Condensed Determinations
Pat Langley, Stanford University

Data Mining with Sparse and Simplified Interaction Selection
Gerald Fahner, International Computer Science Institute

Extraction of Spatial Proximity Patterns by Concept Generalization
Edwin M. Knorr and Raymond T. Ng, University of British Columbia, Canada

Pattern Discovery in Temporal Databases: A Temporal Logic Approach
Balaji Padmanabhan and Alexander Tuzhilin, New York University

Reverse Engineering Databases for Knowledge Discovery
Stephen Mc Kearney, Bournemouth University and Huw Roberts, BT Laboratories, United Kingdom

11:00 AM - 12:00 PM

Two Parallel Sessions

Paper Session 2A: Scalability and Extensibility

Room B113-116, Oregon Convention Center

Extensibility in Data Mining Systems
Stefan Wrobel, Dietrich Wettschereck, Edgar Sommer, and Werner Emde, GMD, FIT.KI, Germany

Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid
Ron Kohavi, Silicon Graphics, Inc.

Data Mining and Model Simplicity: A Case Study in Diagnosis
Gregory M. Provan, Rockwell Science Center and Moninder Singh, University of Pennsylvania

Paper Session 2B: Applications I

Room A105-106, Oregon Convention Center

Automated Discovery of Active Motifs in Multiple RNA Secondary Structures
Jason T. L. Wang, New Jersey Institute of Technology; Bruce A. Shapiro, National Institutes of Health; Dennis Shasha, New York University; Kaizhong Zhang, The University of Western Ontario, Canada; and Chia-Yo Chang, New Jersey Institute of Technology

Using a Hybrid Neural/Expert System for Data Base Mining in Market Survey Data
Victor Ciesielski and Gregory Palstra, Royal Melbourne Institute of Technology, Australia

Automated Discovery of Medical Expert System Rules from Clinical Databases Based on Rough Sets
Shusaku Tsumoto and Hiroshi Tanaka, Tokyo Medical and Dental University, Japan

12:00 - 1:30 PM

Lunch

1:30 - 2:30 PM

Plenary Session

Room B113-116, Oregon Convention Center

Invited Talk

Efficient Implementation of Data Cubes Via Materialized Views
Jeffrey D. Ullman, Stanford University

Data warehouses are collections of materialized views of source data. The optimal set of views to materialize depends on the assumed distribution of queries that will be posed about the data. Given a query distribution, a "greedy" approach to selecting materialized views picks a sequence of views, each of which provides the maximum "benefit" (reduction in average query cost), given the set of views previously chosen for materialization. Under a variety of assumptions about the way possible views relate to one another, greedy approaches are guaranteed to come within 63% of the optimum benefit. In fact, in some of these cases, such as the important case of a "data cube" storing multidimensional data, it can be shown that no polynomial algorithm can be guaranteed to come closer than 63%.

2:30 - 3:30 PM

Plenary Session

Room B113-116, Oregon Convention Center

2:30 - 2:40 PM

Technology Spotlight T3 (Posters)

Exploiting Background Knowledge in Automated Discovery
John M. Aronis, University of Pittsburgh; Foster J. Provost, NYNEX Science & Technology; and Bruce G. Buchanan, University of Pittsburgh

Maintenance of Discovered Knowledge: A Case in Multi-Level Association Rules
David W. Cheung, University of Hong Kong; Vincent T. Ng, Hong Kong Polytechnic University; and Benjamin W. Tam, The University of Hong Kong

Analysing Binary Associations
Arno J. Knobbe and Pieter W. Adriaans, Syllogic, The Netherlands

Evaluating the Interestingness of Characteristic Rules
Micheline Kamber, Simon Fraser University and Rajjan Shinghal, Concordia University, Canada

Exceptional Knowledge Discovery in Databases Based on Information Theory
Einoshin Suzuki, Yokohama National University and Masamichi Shimura, Tokyo Institute of Technology, Japan

2:40 - 3:20 PM

Paper Session 3: Spatial and Text Data Mining

A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise
Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu, University of Munich, Germany

Self-Organizing Maps of Document Collections: A New Approach to Interactive Exploration
Krista Lagus, Timo Honkela, Samuel Kaski, and Teuvo Kohonen, Helsinki University of Technology, Finland

3:20 - 3:30 PM

Technology Spotlight T4 (Posters)

RITIO - Rule Induction Two In One
David Urpani, CSIRO; Xindong Wu, Monash University; and Jim Sykes, Swinburne University of Technology, Australia

Growing Simpler Decision Trees to Facilitate Knowledge Discovery
Kevin J. Cherkauer and Jude W. Shavlik, University of Wisconsin

Data Mining and Tree-Based Optimization
Robert Grossman, Magnify, Inc. and University of Illinois; Haim Bodek and Dave Northcutt, Magnify, Inc.; Vince Poor, Princeton University

SE-Trees Outperform Decision Trees in Noisy Domains
Ron Rymon, University of Pittsburgh

Efficient Specific-to-General Rule Induction
Pedro Domingos, University of California, Irvine

3:30 - 3:50 PM

Coffee Break

3:50 - 4:50 PM

Two Parallel Sessions

Paper Session 4A: Decision-Tree and Rule Induction

Room B113-116, Oregon Convention Center

Error-Based and Entropy-Based Discretization of Continuous Features
Ron Kohavi, Silicon Graphics, Inc. and Mehran Sahami, Stanford University

Discovery of Relevant New Features by Generating Non-Linear Decision Trees
Andreas Ittner, Chemnitz University of Technology and Michael Schlosser, Fachhochschule Koblenz, Germany

Linear-Time Rule Induction
Pedro Domingos, University of California, Irvine

Special Paper Session 4B: Systems for Mining Large Databases

A105-106, Oregon Convention Center

The Quest Data Mining System
Rakesh Agrawal, Manish Mehta, John Shafer, and Ramakrishnan Srikant, IBM Almaden Research Center; Andreas Arning and Toni Bollinger, IBM German Software Development Laboratory, Germany

DataMine: Application Programming Interface and Query Language for Database Mining
Tomasz Imielinski, Aashu Virmani, and Amin Abdulghani, Rutgers University

DBMiner: A System for Mining Knowledge in Large Relational Databases
Jiawei Han, Yongjian Fu, Wei Wang, Jenny Chiang, Wan Gong, Krzystof Koperski, Deyi Li, Yijun Lu, Amynmohamed Rajan, Nebojsa Stefanovic, Betty Xia, and Osmar R. Zaiane, Simon Fraser University, Canada

4:50 - 5:30 PM

Two Parallel Sessions

Paper Session 5A: Mining with Noise and Missing Data

Room B113-116, Oregon Convention Center

Imputation of Missing Data Using Machine Learning Techniques
Kamakshi Lakshminarayan, Steven A. Harp, Robert Goldman, and Tariq Samad, Honeywell Technology Center

Discovering Generalized Episodes Using Minimal Occurrences
Heikki Mannila and Hannu Toivonen, University of Helsinki, Finland

Session 5B: Panel Discussion: Systems for Mining Large Databases

A105-106, Oregon Convention Center

6:00 - 8:00 PM

Opening Reception & Poster and Demonstration Session

Room C123-124, Oregon Convention Center

Demonstrations

Product Demonstration Program

Extensibility in Data Mining Systems
Stefan Wrobel, Dietrich Wettschereck, Edgar Sommer, and Werner Emde, GMD, FIT.KI, Germany

DBMiner: A System for Mining Knowledge in Large Relational Databases
Jiawei Han, Yongjian Fu, Wei Wang, Jenny Chiang, Wan Gong, Krzystof Koperski, Deyi Li, Yijun Lu, Amynmohamed Rajan, Nebojsa Stefanovic, Betty Xia, and Osmar R. Zaiane, Simon Fraser University, Canada

Webfind: Mining External Sources To Guide WWW Discovery.
Alvaro E. Monge and Charles P. Elkan, University of California, San Diego

MM - Mining with Maps
Raymond T. Ng, University of British Columbia, Canada

Decisionhouse
Nicholas J. Radcliffe, Quadstone Ltd., United Kingdom

STARC - A New Data Mining Tool
Damir Gainanow, Andre Matweew, and Michael Thess, DATA-Center Ltd., Russia and Scholz & Thess Software GbR, Germany

D-SIDE: A Probabilistic DeciSIon enDorsement Environment
Petri Kontkanen, Petri Myllymaki, and Henry Tirri, University of Helsinki, Finland

MineSet
Steven Reiss and Mario Schkolnick, Silicon Graphics, Inc.

Optimization Related Data Mining Using the PATTERN System
H. Bodek, R. L. Grossman, D. Northcutt, and H. V. Poor, Magnify, Inc. and Princeton University

Management Discovery Tool
Ken O'Flaherty, NCR Corporation

Clementine Data Mining System
Colin Shearer, Integral Solutrions Ltd., United Kingdom

Mining Associations in Text in the Presence of Background Knowledge
Ronen Feldman, Bar-Ilan Univesity, Israel and Haym Hirsh, Rutgers University

DataMine: Ad Hoc KDD Querying
Tomasz Imielinski and Aashu Virmani, Rutgers University

Saturday August 3

8:30 - 9:30 AM

Plenary Session

Room B113-116, Oregon Convention Center

Invited Talk

Small Sample Size Paradigm in Statistical Inference
Vladimir Vapnik, AT&T Research Laboratories

Vladimir Vapnik will describe (from both the theoretical and the applied point of view) a new approach to statistical inference that is based on the minimization of the guaranteed risk for a fixed sample size, which provides a high level of generalization ability and in many cases contradicts the existing classical paradigms.

9:30 - 9:45 AM

Coffee Break

9:45 - 11:05 AM

Plenary Session

Room B113-116, Oregon Convention Center

9:45 - 9:55 AM

Technology Spotlight T5 (Posters)

A Genetic Algorithm-Based Approach to Data Mining
Ian W. Flockhart, Quadstone Ltd. and Nicholas J. Radcliffe, Quadstone Ltd. and University of Edinburgh, United Kingdom

Deriving Queries from Results Using Genetic Programming
Tae-Wan Ryu and Christoph F. Eick, University of Houston

Discovering Classification Knowledge in Databases Using Rough Sets
Ning Shan, Wojciech Ziarko, Howard J. Hamilton, and Nick Cercone, University of Regina, Canada

Representing Discovered Patterns Using Attributed Hypergraph
Yang Wang and Andrew K.C.Wong, University of Waterloo, Canada

Interactive Knowledge Discovery from Marketing Questionnaire Using Simulated Breeding and Inductive Learning Methods
Takao Terano, The University Tsukuba, Tokyo and Yoko Ishino, The University of Tokyo, Japan

9:55 - 10:55 AM

Paper Session 6: Prediction and Deviation

A Comparison of Approaches for Maximizing the Business Payoff of Prediction Models
Brij Masand and Gregory Piatetsky-Shapiro, GTE Laboratories

A Linear Method for Deviation Detection in Large Databases
Andreas Arning, IBM German Software Development Laboratory, Germany; Rakesh Agrawal and Prabhakar Raghavan, IBM Almaden Research Center

Multiple Uses of Frequent Sets and Condensed Representations: Extended Abstract
Heikki Mannila and Hannu Toivonen, University of Helsinki, Finland

10:55 - 11:05 AM

Technology Spotlight T6 (Posters)

Learning Limited Dependence Bayesian Classifiers
Mehran Sahami, Stanford University

The Field Matching Problem: Algorithms and Applications
Alvaro E. Monge and Charles P. Elkan, University of California, San Diego

Performing Effective Feature Selection by Investigating the Deep Structure of the Data
Marco Richeldi and Pier Luca Lanzi, CSELT, Italy

Inferring Hierarchical Clustering Structures by Deterministic Annealing
Thomas Hofmann and Joachim M. Buhmann, Rheinische Friedrich-Wilhelms-Universität, Germany

Efficient Search for Strong Partial Determinations
Stefan Kramer and Bernhard Pfahringer, Austrian Research Institute for Artificial Intelligence, Austria

11:05 AM - 12:05 PM

Two Parallel Sessions

Paper Session 7A: Prediction

B113-116, Oregon Convention Center

Predictive Data Mining with Finite Mixtures
Petri Kontkanen, Petri Myllymäki, and Henry Tirri, University of Helsinki, Finland

An Empirical Test of the Weighted Effect Approach to Generalized Prediction Using Recursive Neural Nets
Rense Lange, University of Illinois at Springfield

Planning Tasks for Knowledge Discovery in Databases; Performing Task-Oriented User-Guidance
Robert Engels, University of Karlsruhe, Germany

Paper Session 7B: Applications II

Room A105-106, Oregon Convention Center

KDD for Science Data Analysis: Issues and Examples
Usama Fayyad, Microsoft Research; David Haussler, University of California, Santa Cruz; and Paul Stolorz, Jet Propulsion Laboratory, California Institute of Technology

Detecting Early Indicator Cars in an Automotive Database: A Multi-Strategy Approach
Ruediger Wirth and Thomas P. Reinartz, Daimler-Benz AG, Germany

Discovering Knowledge in Commercial Databases Using Modern Heuristic Techniques
B. de la Iglesia, J. C. W. Debuse, and V. J. Rayward-Smith, University of East Anglia, United Kingdom

12:05 - 1:30 PM

Lunch

1:30 - 2:30 PM

Plenary Session

B113-116, Oregon Convention Center

Invited Talk

Data Integration and Analysis in a Client Server Environment: The Sara Lee Meat Experience
Perry K. Youngs, Sara Lee Corporation

The role of marketing research is currently going through dramatic changes in the United

States as census based syndicated scanner data is becoming available to retailers and manufacturers. This change is being led by ECR and Category Management initiatives that are removing costs from distribution channels. In an attempt to manage the ever increasing amounts of information needed for this endeavor, client server based information systems are being developed with new data warehousing technology.

Sara Lee Meats has just successfully implemented the conversion of a main frame based system to a client server based system using a three tier object technology from Information Advantage, Incorporated and data warehousing technology from Red Brick Systems, Incorporated. Youngs will discuss Sara Lee Meat's experiences relating to data integration

and analysis in a client server environment.

2:30 - 3:30 PM

Plenary Session

B113-116, Oregon Convention Center

Paper Session 8:Combining Data Mining and Machine Learning

Combining Data Mining and Machine Learning for Effective User Profiling
Tom Fawcett and Foster Provost, NYNEX Science and Technology

Sharing Learned Models among Remote Database Partitions by Local Meta-Learning
Philip K. Chan, Florida Institute of Technology and Salvatore J. Stolfo, Columbia University

Local Induction of Decision Trees: Towards Interactive Data Mining
Truxton Fulton, Simon Kasif, and Steven Salzberg, Johns Hopkins University; David Waltz, NEC Research Institute

3:30 - 3:50 PM

Coffee Break

3:50 - 4:50 PM

Three Parallel Sessions

Poster Session II

Room C123, Oregon Convention Center

Paper Session 9A: Approaches to Numeric Data

Room B113-116, Oregon Convention Center

Mining Knowledge in Noisy Audio Data
Andrzej Czyzewski, Technical University of Gdansk, Poland

A Method for Reasoning with Structured and Continuous Attributes in the INLEN-2 Multistrategy Knowledge Discovery System
Kenneth A. Kaufman, George Mason University and Ryszard S. Michalski, George Mason University and Polish Academy of Sciences, Poland

Learning from Biased Data Using Mixture Models
A.J. Feelders, Data Distilleries Ltd., The Netherlands

Special Paper Session 9B: Scalable and Distributed Applications of KDD

Room A105-106, Oregon Convention Center

Parallel Halo Finding in *N*-body Cosmology Simulations
David W. Pfitzner, Mount Stromlo Observatory, Australia and John K. Salmon, California Institute of Technology

Scalable Exploratory Data Mining of Distributed Geoscientific Data
Eddie C. Shek, University of California, Los Angeles and Hughes Research Laboratories; Richard R. Muntz, Edmond Mesrobian, and Kenneth Ng, University of California, Los Angeles

Knowledge Discovery in RNA Sequence Families of HIV Using Scalable Computers
Ivo L. Hofacker, University of Illinois; Martijn A. Huynen, Los Alamos National Laboratory and Santa Fe Institute; Peter F. Stadler, University of Vienna and Santa Fe Institute; Paul E. Stolorz, Jet Propulsion Laboratory, California Institute of Technology

4:50 - 5:30 PM

Two Parallel Sessions

Paper Session 10A: Pattern-Oriented Data Mining

Room B113-116, Oregon Convention Center

Metapattern Generation for Integrated Data Mining
Wei-Min Shen, University of Southern California and Bing Leng, Inference Corporation

Automated Pattern Mining with a Scale Dimension
Jan M. Zytkow, Wichita State University and Polish Academy of Sciences, Poland; Robert Zembowicz, Wichita State University

Session 10B: Panel Discussion

Room A105-106, Oregon Convention Center

Scalable and Distributed Applications of KDD

The Promise and Challenge of Data Mining with High Performance Computers

7:00 PM

KDD-96 Conference Banquet

Benson Hotel

Invited Talk

Advanced Scout: Data Mining and Knowledge Discovery in NBA Data
Inderpal Bhandari, IBM, T.J. Watson Research Center

Sunday August 4

All sessions will be held in Room B114-116, Oregon Convention Center

8:30 - 11:35 PM

Joint UAI-96/KDD-96 Plenary Sessions

Selected talks on learning graphical models from the UAI-96 and KDD-96 proceedings. UAI badges will be honored at the Portland Convention Center for the joint session.

8:30 - 8:40 AM

Introductory Remarks

UAI Meets KDD
Usama Fayyad and Eric Horvitz, Microsoft Research

8:40 - 10:00 AM

Session 11: Learning, Probability, and Graphical Models I

KDD-96: Knowledge Discovery and Data Mining: Towards a Unifying Framework
Usama Fayyad, Microsoft Research; Gregory Piatetsky-Shapiro, GTE Laboratories; and Padhraic Smyth, University of California, Irvine

UAI-96: Efficient Approximations for the Marginal Likelihood of Incomplete Data Given a Bayesian Network
D. Chickering, University of California, Los Angeles and D. Heckerman, Microsoft Research

KDD-96: Clustering Using Monte Carlo Cross-Validation
Padhraic Smyth, University of California, Irvine

UAI-96: Learning Equivalence Classes of Bayesian Network Structures
D. Chickering, University of California, Los Angeles

10:00 - 10:15 AM

Coffee Break

10:15 - 11:35 AM

Session 12: Learning, Probability, and Graphical Models II

UAI-96: Learning Bayesian Networks with Local Structure
N. Friedman, Stanford University and M. Goldszmidt, SRI International

KDD-96: Rethinking the Learning of Belief Network Probabilities
Ron Musick, Lawrence Livermore National Laboratory

UAI-96: Bayesian Learning of Loglinear Models for Neural Connectivity
K. Laskey and L. Martignon

KDD-96: Harnessing Graphical Structure in Markov Chain Monte Carlo Learning
Paul E. Stolorz, Jet Propulsion Laboratory, California Institute of Technology and Philip C. Chew, University of Pennsylvania

11:35 AM - 12:30 PM

Lunch

(Box lunch will be served--May overlap with Summary Panel Session)

12:30 - 1:20 PM

Summary Panel and Closing Remarks

"What Have We Discovered?"

1:20 - 2:00 PM

KDD Wrap-up Business Meeting

2:00 PM

Conference Adjourns