# Table of Contents

Welcome Message from the Conference Chairs ................................................................. xiii
Preface ............................................................................................................................... xiv
Workshops and Organizations ........................................................................................... xvi

## Data Mining in Web 2.0 Environments

Ask the Crowd to Find out What’s Important ........................................................................ 3  
*Sisay Fissaha Adafre and Maarten de Rijke*

Aspect Summarization from Blogosphere for Social Study .................................................. 9  
*Chia-Hui Chang and Kun-Chang Tsai*

FiVaTech: Page-Level Web Data Extraction from Template Pages .................................... 15  
*Mohammed Kayed, Chia-Hui Chang, Khaled Shaalan, and Moheb Ramzy Girgis*

SOPS: Stock Prediction Using Web Sentiment ..................................................................... 21  
*Vivek Sehgal and Charles Song*

HSN-PAM: Finding Hierarchical Probabilistic Groups from Large-Scale Networks ............ 27  
*Haizheng Zhang, Wei Li, Xuerui Wang, C. Lee Giles, Henry C. Foley, and John Yen*

Extracting Author Meta-Data from Web Using Visual Features ......................................... 33  
*Shuyi Zheng, Ding Zhou, Jia Li, and C. Lee Giles*

## Knowledge Discovery from Multimedia Data and Multimedia Applications

Automatic Generation of Traditional Style Painting by Using Density-Based Color Clustering ......................................................................................................................... 41  
*Guangyan Huang, Zhiming Ding, and Jing He*

Semi-Automatic Semantic Annotation of Images .................................................................... 45  
*Suzanne Little, Ovidio Salvetti, and Petra Perner*
Bit Sequences and Biclustering of Text Documents .......................................................... 51
Selim Mimaroglu and Dan A. Simovici

Video Data Mining: Discovering Topics by Burst Detection in Video Streams .................. 57
Kimiaki Shirahama and Kuniaki Uehara

Tensor Space Learning for Analyzing Activity Patterns from Video Sequences .................. 63
Liang Wang, Christopher Leckie, Xiaozhe Wang, Ramamohanarao Kotagiri, and Jim Bezdek

Adapting SVM Classifiers to Data with Shifted Distributions ........................................... 69
Jun Yang, Rong Yan and Alexander G. Hauptmann

**Mining and Management of Biological Data**

A Comparative Study of Methods for Transductive Transfer Learning .............................. 77
Andrew Arnold, Ramesh Nallapati, and William W. Cohen

Assessing Reliability of Protein-Protein Interactions by Semantic Data Integration ............. 83
Young-Rae Cho, Woochang Hwang, and Aidong Zhang

Characterizing RNA Secondary-Structure Features and Their Effects on Splice-Site Prediction ......................................................................................................................... 89
Rezarta Islamaj Dogan, Lise Getoor, and W. John Wilbur

Statistical Approaches to Identifying Androgen Response Elements ............................... 95
Li Li, Steffen Heber, Qiang Zhang, and Melvin E. Andersen

Mapping Gene/Protein Names in Free Text to Biomedical Databases ............................ 101
Hongfang Liu, Manabu Torii, Zhang-zhi Hu, and Cathy Wu

A Content Based Pattern Analysis System for a Biological Specimen Collection ............... 107
Joyita Mallik, Ashok Samal, and Scott L. Gardner

Discovering Gene Expression Data from the Tables of Full Text Publications ..................... 113
Brigitte Mathiak, Andreas Kupfer, Carolina Rio Bartulos, Tatjana Scope, Johann Weiland, and Silke Eckstein

Modeling and Management of Signal Transduction Pathways with Live Sequence Charts .... 119
Claudia Täubner, Brigitte Mathiak, and Silke Eckstein

**Data Mining in Medicine**

Developing an Integrated Time-Series Data Mining Environment for Medical Data Mining ............................................................................................................................................. 127
Hidenao Abe, Hideto Yokoi, Miho Ohsaki, and Takahira Yamaguchi

Time-Annotated Sequences for Medical Data Mining .......................................................... 133
Michele Berlingerio, Francesco Bonchi, Fosca Giannotti, and Franco Turini

Automatically Finding Images for Clinical Decision Support ............................................ 139
Dina Demner-Fushman, Sameer Antani, and George R. Thoma
Content-Based Support Environment (C-BASE): Data Preparation and Similarity Measurement .......... 145
Ameen Eetemadi, Mohammad-Reza Siadat, Hamid Soltanian-Zadeh, Farshad Fotouhi, and Kost Elisevich

Identifying Exacerbating Cases in Chronic Diseases Based on the Cluster Analysis of Trajectory Data on Laboratory Examinations.......................................................... 151
Shoji Hirano and Shusaku Tsumoto

Predictive Data Mining for Lung Nodule Interpretation ................................................................. 157
William Horsthemke, Ekarin Varutbangkul, Daniela Raicu, and Jacob Furst

Predictive Data Mining to Learn Health Vitals of a Resident in a Smart Home .............................. 163
Vikramaditya Jakkula

Utilization of Data-Mining Techniques for Evaluation of Patterns of Asthma Drugs Use by Ambulatory Patients in a Large Health Maintenance Organization......................................... 169
Mark Last, Rafael Carel, and Dotan Barak

Analysis of Relationship between Blood Stream Infection and Clinical Background in Patients’ Lactobacillus Therapy by Data Mining .............................................................. 175
Kimiko Matsuoka, Shigeki Yokoyama, Kumitomo Watanabe, and Shusaku Tsumoto

Data Modeling for Content-Based Support Environment (C-BASE): Application on Epilepsy Data Mining ............................................................................................................... 181
Mohammad-Reza Siadat, Hamid Soltanian-Zadeh, Farshad Fotouhi, Ameen Eetemadi, and Kost Elisevich

Optimization-Based Data Mining Techniques with Applications

Data Clustering with a Relational Push-Pull Model ........................................................................ 189
Adam Anthony and Marie desJardins

Learning What Makes a Society Tick................................................................................................ 195
Hung-Ching (Justin) Chen, Mark Goldberg, Malik Magdon-Ismail, and William A. Wallace

Distance Metric Learning through Optimization of Ranking.......................................................... 201
Kreshna Gopal and Thomas R. Ioerger

Using Data Mining to Estimate Missing Sensor Data ...................................................................... 207
Le Gruenwald, Hamed Chok, and Mazen Aboukhamis

Cluster Analysis and Optimization in Color-Based Clustering for Image Abstract .......................... 213
Jing He, Guangyan Huang, Yanchun Zhang, and Yong Shi

Predicting and Optimizing Classifier Utility with the Power Law .................................................. 219
Mark Last

Mining Distance-Based Outliers from Categorical Data ................................................................. 225
Shuxin Li, Robert Lee, and Sheau-Dong Lang

Feature Selection for Nonlinear Kernel Support Vector Machines ................................................. 231
Olvi L. Mangasarian and Edward W. Wild
Categorical Attribute Transformation Technique for Multiple Criteria Quadratic Programming Classification Model ................................................................. 237
   Yi Peng and Gang Kou

Physical Analysis of Precipitation Factors Based on SVM Method ................................................................. 243
   Xiong Qiufen, Gao Jie, Liu Huanzhu, and Shao Mingxuan

An Efficient Fitness Assignment Based on Dominating Tree ................................................................. 247
   Chuan Shi, Zhongzhi Shi, and Bin Wu

A Regularized Multiple Criteria Linear Program for Classification ............................................................... 253
   Yong Shi, Yingjie Tian, Xiaojun Chen, and Peng Zhang

Dual Fuzzy-Possibilistic Co-clustering for Document Categorization ............................................................... 259
   William-Chandra Tjhi and Lihui Chen

Generalized Additive Models from a Neural Network Perspective ............................................................... 265
   D. A. de Waal and J. V. du Toit

A Novel Rule Weighting Approach in Classification Association Rule Mining ............................................................... 271
   Yanbo J. Wang, Qin Xin, and Frans Coenen

Semi-supervised Kernel Logistic Regression and Its Extension to Active Learning Based on A-Optimality ............................................................... 277
   Yasutoshi Yajima, and Teppei Sato

Classification with Choquet Integral with Respect to Signed Non-additive Measure ............................................................... 283
   Nian Yan, Zhenyuan Wang, and Zhengxin Chen

Multiple-Criteria Linear Programming for VIP E-Mail Behavior Analysis ............................................................... 289
   Peng Zhang and Jingran Dai

High Performance Data Mining

A Novel Parallel Boolean Approach for Discovering Frequent Itemsets ............................................................... 297
   Wassim Ayadi and Khedija Arour

Collaborative Filtering Using Orthogonal Nonnegative Matrix Tri-factorization ............................................................... 303
   Gang Chen, Fei Wang, and Changshui Zhang

Extracting Knowledge to Predict TSP Asymptotic Time Complexity ............................................................... 309
   Paula Cecilia Fritzsche, Dolores Rexachs, and Emilio Luque

Twin Kernel Embedding with Back Constraints ......................................................................................... 319
   Yi Guo, Paul W. Kwan, and Junbin Gao

An Efficient Technique for Mining Approximately Frequent Substring Patterns ............................................................... 325
   Xiaonan Ji and James Bailey

Robust Unsupervised and Semisupervised Bounded C-Support Vector Machines ............................................................... 331
   Zhao Kun, Tian Ying-jie, and Deng Nai-yang

Large PSD Matrix Estimation from Partial Elements ......................................................................................... 337
   Hiroshi Kuwajima and Takashi Washio
Sparse Word Graphs: A Scalable Algorithm for Capturing Word Correlations in Topic Models ................................................................. 343
  Ramesh Nallapati, Amr Ahmed, William Cohen, and Eric Xing

Parallelized Variational EM for Latent Dirichlet Allocation: An Experimental Evaluation of Speed and Scalability ................................................................. 349
  Ramesh Nallapati, William Cohen, and John Lafferty

WC-Clustering: Hierarchical Clustering Using the Weighted Confidence Affinity Measure ................................. 355
  Baoying Wang and Imad Rahal

Semi-supervised Clustering Using Bayesian Regularization ................................................................. 361
  Zuobing Xu, Ram Akella, Mike Ching, and Renjie Tang

Experimental Comparison of Feature Subset Selection Methods ................................................................. 367
  Chulmin Yun and Jihoon Yang

Utility-Based Web Path Traversal Pattern Mining ..................................................................................... 373
  Lin Zhou, Ying Liu, Jing Wang, and Yong Shi

Mining Graphs and Complex Structures

Combining Collective Classification and Link Prediction ..................................................................................... 381
  Mustafa Bilgic, Galileo Mark Namata, and Lise Getoor

Simultaneous Heterogeneous Data Clustering Based on Higher Order Relationships ................................................................. 387
  Shouchun Chen, Fei Wang, and Changshui Zhang

Discovering Structural Anomalies in Graph-Based Data ..................................................................................... 393
  William Eberle and Lawrence Holder

Subgraph Support in a Single Large Graph .............................................................................................. 399
  Mathias Fiedler and Christian Borgelt

Tree Planar Languages ..................................................................................................................................... 405
  Christophe Costa Florêncio

An Examination of Experimental Methodology for Classifiers of Relational Data ......................................... 411
  Brian Gallagher and Tina Eliassi-Rad

GDClust: A Graph-Based Document Clustering Technique ..................................................................................... 417
  M. Shahriar Hossain and Rafal A. Angryk

Learning Term Dependency Links Using Information Theoretic Inclusion Measure ................................................................. 423
  Masoud Makrehchi and Mohamed S. Kamel

Exploiting Network Structure for Active Inference in Collective Classification ................................................................. 429
  Matthew J. Rattigan, Marc Maier, and David Jensen

Resume Mining of Communities in Social Network ..................................................................................... 435
  Bin Wu, Xin Pei, JianBin Tan, and Yi Wang

A Divisive Hierarchical Structural Clustering Algorithm for Networks ................................................................. 441
  Nurcan Yuruk, Mutlu Mete, Xiaowei Xu, and Thomas A. J. Schweiger
Data Mining on Uncertain Data

Counterpropagation Neural Network for Stochastic Conditional Simulation:
An Application with Berea Sandstone ........................................................................................................ 449
   Lance E. Besaw and Donna M. Rizzo

Genre Categorization of Web Pages ........................................................................................................ 455
   Jebari Chaker and Ounelli Habib

Segmenting Multi-attribute Sequences Using Dynamic Bayesian Networks ............................................ 465
   Robert Gwadera, Janne Toivola, and Jaakko Hollmén

Error-Aware Density-Based Clustering of Imprecise Measurement Values ........................................... 471
   Dirk Habich, Peter B. Volk, Wolfgang Lehner, Ralf Dittmann, and Clemens Utzny

Skewed Class Distributions and Mislabeled Examples ............................................................................ 477
   Jason Van Hulse, Taghi M. Khoshgoftaar, and Amri Napolitano

Reducing UK-Means to K-Means ............................................................................................................. 483
   S. D. Lee, Ben Kao, and Reynold Cheng

Efficient Mining of Frequent Patterns from Uncertain Data .................................................................. 489
   Carson Kai-Sang Leung, Christopher L. Carmichael, and Boyu Hao

A Novel Ordering-Based Greedy Bayesian Network Learning Algorithm on Limited Data .................... 495
   Feng Liu, Fengzhan Tian, and Qiliang Zhu

Granularity Conscious Modeling for Probabilistic Databases ................................................................ 501
   Eirinaios Michelakis, Daisy Zhe Wang, Minos Garofalakis, and Joseph M. Hellerstein

Representing Tuple and Attribute Uncertainty in Probabilistic Databases ........................................... 507
   Prithviraj Sen, Amol Deshpande, and Lise Getoor

Targeting Input Data for Acoustic Bird Species Recognition Using Data Mining and HMMs .................. 513
   Erika Vilches, Iván A. Escobar, Edgar E. Vallejo, and Charles E. Taylor

Incremental Integration of Probabilistic Models Learned from Data ....................................................... 519
   Jian Xu, Pedrito Maynard-Zhang, and Jianhua Chen

Data Streaming Mining and Management

Incremental Quantization for Aging Data Streams ................................................................................... 527
   Fatih Altiparmak, David Chiu, and Hakan Ferhatosmanoglu

Hierarchical Classifier Combination and Its Application in Networks Intrusion Detection ..................... 533
   Morteza Analoui, Behrouz Minaei Bidgoli, and Mohammad Hossein Rezvani

An Approach for Incremental Semi-supervised SVM ........................................................................... 539
   Wael Emara and Mehmed Kantardzic

Quality of Service-Driven Stream Mining ................................................................................................. 545
   Marcel Karnstedt, Kai-Uwe Sattler, Dirk Habich, and Wolfgang Lehner
Sequential Change Detection on Data Streams ................................................................. 551
  S. Muthukrishnan, Eric van den Berg, and Yihua Wu

Optimal Window Change Detection .............................................................................. 557
  Jan Peter Patist

High-Speed Identification of Language and Script ...................................................... 563
  Alan Ratner and Ron Loui

Infrequent Item Mining in Multiple Data Streams ....................................................... 569
  Budhaditya Saha, Mihai Lazarescu, and Svetha Venkatesh

Stream Event Detection: A Unified Framework for Mining Outlier, Change
and Burst Simultaneously over Data Stream ................................................................. 575
  Zhijian Yuan, Kai Du, Yan Jia, and Jiajia Miao

Toward Behavioral Modeling of a Grid System: Mining the Logging and Bookkeeping Files .......... 581
  Xiangliang Zhang, Michèle Sebag, and Cécile Germain

Spatial and Spatio-temporal Data Mining

Space-Time Interpolation and Uncertainty Assessment of an Extreme Precipitation
Index Using Geostatistical Cosimulation........................................................................ 589
  Ana Cristina Costa and Amílcar Soares

On Regional Association Rule Scoping ........................................................................... 595
  Wei Ding, Christoph F. Eick, Xiaojing Yuan, Jing Wang, and Jean-Philippe Nicot

A Compact Representation of Spatio-temporal Data.................................................... 601
  Sigal Elnekave, Mark Last, and Oded Maimon

Modeling Fundamental Geo-Raster Operations with Array Algebra.............................. 607
  Angelica Garcia Gutierrez and Peter Baumann

Pattern Mining as Abduction: From Snapshots to Spatio-temporal Sequential Patterns .......... 613
  Shyamanta M. Hazarika

Query Expansion Using Topic and Location..................................................................... 619
  Shu Huang, Qiankun Zhao, Prasenjit Mitra, and C. Lee Giles

Knowledge Discovery in Entity Based Smart Environment Resident Data
Using Temporal Relation Based Data Mining................................................................. 625
  Vikramaditya R. Jakkula, Aaron S. Crandall, and Diane J. Cook

Space-Time Summarization of Multisensor Time Series. Case of Missing Data ................ 631
  Marc Joliveau and Florian De Vuyst

Spatial Clustering Using the Likelihood Function ............................................................ 637
  April Kerby, David Marx, Ashok Samal, and Viacheslav Adamchuck

Diagnosing Similarity of Oscillation Trends in Time Series ........................................... 643
  Leonardo E. Mariote, Claudia Bauzer Medeiros, and Ricardo da S. Torres