

Proceedings

ICDM Workshops 2007

*Seventh IEEE International Conference
on Data Mining – Workshops*

*28-31 October 2007
Omaha, Nebraska, USA*



Los Alamitos, California
Washington • Tokyo



TABLE OF CONTENTS

DATA MINING IN WEB 2.0 ENVIROMENTS

Ask the Crowd to Find out What's Important	1
<i>Sisay Fissaha Adafre; Maarten de Rijke</i>	
Aspect Summarization from Blogosphere for Social Study	7
<i>Chia-Hui Chang; Kun-Chang Tsai</i>	
FiVaTech: Page-Level Web Data Extraction from Template Pages	13
<i>Mohammed Kayed; Chia-Hui Chang; Khaled Shaalan; Moheb Ramzy Girgis</i>	
SOPS: Stock Prediction Using Web Sentiment	19
<i>Vivek Sehgal; Charles Song</i>	
HSN-PAM: Finding Hierarchical Probabilistic Groups from Large-Scale Networks	25
<i>Haizheng Zhang; Wei Li; Xuerui Wang; C. Lee Giles; Henry C. Foley; John Yen</i>	
Extracting Author Meta-Data from Web Using Visual Features	31
<i>Shuyi Zheng; Ding Zhou; Jia Li; C. Lee Giles</i>	

KNOWLEDGE DISCOVERY FROM MULTIMEDIA DATA AND MULTIMEDIA APPLICATIONS

Automatic Generation of Traditional Style Painting by Using Density-Based Color Clustering	37
<i>Guangyan Huang; Zhiming Ding; Jing He</i>	
Semi-Automatic Semantic Annotation of Images	41
<i>Suzanne Little; Ovidio Salvetti; Petra Perner</i>	
Bit Sequences and Biclustering of Text Documents	47
<i>Selim Mimaroglu; Dan A. Simovici</i>	
Tensor Space Learning for Analyzing Activity Patterns from Video Sequences	53
<i>Liang Wang; Christopher Leckie; Xiaozhe Wang; Ramamohanarao Kotagiri; Jim Bezdek</i>	
Adapting SVM Classifiers to Data with Shifted Distributions	59
<i>Jun Yang; Rong Yan; Alexander G. Hauptmann</i>	

MINING AND MANAGEMENT OF BIOLOGICAL DATA

A Comparative Study of Methods for Transductive Transfer Learning	65
<i>Andrew Arnold; Ramesh Nallapati; William W. Cohen</i>	
Assessing Reliability of Protein-Protein Interactions by Semantic Data Integration	71
<i>Young-Rae Cho; Woochang Hwang; Aidong Zhang</i>	
Characterizing RNA Secondary-Structure Features and Their Effects on Splice-Site Prediction	77
<i>Rezarta Islamaj Dogan; Lise Getoor; W. John Wilbur</i>	
Statistical Approaches to Identifying Androgen Response Elements	83
<i>Li Li; Steffen Heber; Qiang Zhang; Melvin E. Andersen</i>	

Mapping Gene/Protein Names in Free Text to Biomedical Databases	89
<i>Hongfang Liu; Manabu Torii; Zhang-zhi Hu; Cathy Wu</i>	
A Content Based Pattern Analysis System for a Biological Specimen Collection	95
<i>Joyita Mallik; Ashok Samal; Scott L. Gardner</i>	
Discovering Gene Expression Data from the Tables of Full Text Publications	101
<i>Brigitte Mathiak; Andreas Kupfer; Carolina Rio Bartulos; Tatjana Scope; Johann Weiland; Silke Eckstein</i>	
Modeling and Management of Signal Transduction Pathways with Live Sequence Charts	107
<i>Claudia Täubner; Brigitte Mathiak; Silke Eckstein</i>	

DATA MINING IN MEDICINE

Developing an Integrated Time-Series Data Mining Environment for Medical Data Mining	113
<i>Hidenao Abe; Hideto Yokoi; Miho Ohsaki; Takahira Yamaguchi</i>	
Time-Annotated Sequences for Medical Data Mining	119
<i>Michele Berlingerio; Francesco Bonchi; Fosca Giannotti; Franco Turini</i>	
Automatically Finding Images for Clinical Decision Support	125
<i>Dina Demner-Fushman; Sameer Antani; George R. Thoma</i>	
Identifying Exacerbating Cases in Chronic Diseases Based on the Cluster Analysis of Trajectory Data on Laboratory Examinations	131
<i>Shoji Hirano; Shusaku Tsumoto</i>	
Predictive Data Mining for Lung Nodule Interpretation	137
<i>William Horsthemke; Ekarin Varutbangkul; Daniela Raicu; Jacob Furst</i>	
Predictive Data Mining to Learn Health Vitals of a Resident in a Smart Home	143
<i>Vikramaditya Jakkula</i>	
Utilization of Data-Mining Techniques for Evaluation of Patterns of Asthma Drugs Use by Ambulatory Patients in a Large Health Maintenance Organization	149
<i>Mark Last; Rafael Carel; Dotan Barak</i>	
Analysis of Relationship between Blood Stream Infection and Clinical Background n Patients' Lactobacillus Therapy by Data Mining	155
<i>Kimiko Matsuoka; Shigeki Yokoyama; Kunitomo Watanabe; and Shusaku Tsumoto</i>	
Data Modeling for Content-Based Support Environment (C-BASE): Application on Epilepsy Data Mining	161
<i>Mohammad-Reza Siadat; Hamid Soltanian-Zadeh; Farshad Fotouhi; Ameen Eetemadi; Kost Elisevich</i>	

OPTIMIZATION – BASED DATA MINING TECHNIQUES WITH APPLICATIONS

Data Clustering with a Relational Push-Pull Model	167
<i>Adam Anthony; Marie desJardins</i>	
Learning What Makes a Society Tick	173
<i>Hung-Ching (Justin) Chen; Mark Goldberg; Malik Magdon-Ismael; William A. Wallace</i>	
Distance Metric Learning through Optimization of Ranking	179
<i>Kreshna Gopal; Thomas R. Ioerger</i>	
Using Data Mining to Estimate Missing Sensor Data	185
<i>Le Gruenwald; Hamed Chok; Mazen Aboukhamis</i>	

Cluster Analysis and Optimization in Color-Based Clustering for Image Abstract	191
<i>Jing He; Guangyan Huang; Yanchun Zhang; Yong Sh</i>	
Predicting and Optimizing Classifier Utility with the Power Law	197
<i>Mark Last</i>	
Mining Distance-Based Outliers from Categorical Data	203
<i>Shuxin Li; Robert Lee; Sheau-Dong Lang</i>	
Feature Selection for Nonlinear Kernel Support Vector Machines	209
<i>Olvi L. Mangasarian; Edward W. Wild</i>	
Physical Analysis of Precipitation Factors Based on SVM Method	215
<i>Xiong Qiufen; Gao Jie; Liu Huanzhu; Shao Mingxuan</i>	
An Efficient Fitness Assignment Based on Dominating Tree	219
<i>Chuan Shi; Zhongzhi Shi; Bin Wu</i>	
A Regularized Multiple Criteria Linear Program for Classification	225
<i>Yong Shi; Yingjie Tian; Xiaojun Chen; Peng Zhang</i>	
Dual Fuzzy-Possibilistic Co-clustering for Document Categorization	231
<i>William-Chandra Tjhi; Lihui Chen</i>	
Generalized Additive Models from a Neural Network Perspective	237
<i>D. A. de Waal; J. V. du Toit</i>	
A Novel Rule Weighting Approach in Classification Association Rule Mining	243
<i>Yanbo J. Wang; Qin Xin; Frans Coenen</i>	
Semi-supervised Kernel Logistic Regression and Its Extension to Active Learning Based on A-Optimality	249
<i>Yasutoshi Yajima; Teppei Sato</i>	
Classification with Choquet Integral with Respect to Signed Non-additive Measure	255
<i>Nian Yan; Zhenyuan Wang; Zhengxin Chen</i>	
Multiple-Criteria Linear Programming for VIP E-Mail Behavior Analysis	261
<i>Peng Zhang; Jingran Dai</i>	
 <u>HIGH PERFORMANCE DATA MINING</u>	
A Novel Parallel Boolean Approach for Discovering Frequent Itemsets	266
<i>Wassim Ayadi; Khedija Arour</i>	
Collaborative Filtering Using Orthogonal Nonnegative Matrix Tri-factorization	272
<i>Gang Chen; Fei Wang; Changshui Zhang</i>	
Extracting Knowledge to Predict TSP Asymptotic Time Complexity	278
<i>Paula Cecilia Fritzsche; Dolores Rexachs; Emilio Luque</i>	
Twin Kernel Embedding with Back Constraints	288
<i>Yi Guo; Paul W. Kwan; Junbin Gao</i>	
An Efficient Technique for Mining Approximately Frequent Substring Patterns	294
<i>Xiaonan Ji; James Bailey</i>	
Robust Unsupervised and Semisupervised Bounded C-Support Vector Machines	300
<i>Zhao Kun; Tian Ying-jie; Deng Nai-yang</i>	
Sparse Word Graphs: A Scalable Algorithm for Capturing Word Correlations in Topic Models	306
<i>Ramesh Nallapati; Amr Ahmed; William Cohen; Eric Xing</i>	

Parallelized Variational EM for Latent Dirichlet Allocation: An Experimental Evaluation of Speed and Scalability	312
<i>Ramesh Nallapati; William Cohen; John Lafferty</i>	
WC-Clustering: Hierarchical Clustering Using the Weighted Confidence Affinity Measure	318
<i>Baoying Wang; Imad Rahal</i>	
Semi-supervised Clustering Using Bayesian Regularization	324
<i>Zuobing Xu; Ram Akella; Mike Ching; Renjie Tang</i>	
Experimental Comparison of Feature Subset Selection Methods	330
<i>Chulmin Yun and Jihoon Yang</i>	
Utility-Based Web Path Traversal Pattern Mining	336
<i>Lin Zhou; Ying Liu; Jing Wang; Yong Shi</i>	

MINING GRAPHS AND COMPLEX STRUCTURES

Combining Collective Classification and Link Prediction	342
<i>Mustafa Bilgic; Galileo Mark Namata; Lise Getoor</i>	
Simultaneous Heterogeneous Data Clustering Based on Higher Order Relationships	348
<i>Shouchun Chen; Fei Wang; Changshui Zhang</i>	
Discovering Structural Anomalies in Graph-Based Data	354
<i>William Eberle; Lawrence Holder</i>	
Subgraph Support in a Single Large Graph	360
<i>Mathias Fiedler; Christian Borgel</i>	
Tree Planar Languages	366
<i>Christophe Costa Florêncio</i>	
An Examination of Experimental Methodology for Classifiers of Relational Data	372
<i>Brian Gallagher; Tina Eliassi-Rad</i>	
GDClust: A Graph-Based Document Clustering Technique	378
<i>M. Shahriar Hossain; Rafal A. Angryk</i>	
Learning Term Dependency Links Using Information Theoretic Inclusion Measure	384
<i>Masoud Makrehchi; Mohamed S. Kamel</i>	
Exploiting Network Structure for Active Inference in Collective Classification	390
<i>Matthew J. Rattigan; Marc Maier; David Jensen</i>	
Resume Mining of Communities in Social Network	396
<i>Bin Wu; Xin Pei; JianBin Tan; Yi Wang</i>	
A Divisive Hierarchical Structural Clustering Algorithm for Networks	402
<i>Nurcan Yuruk; Mutlu Mete; Xiaowei Xu; Thomas A. J. Schweiger</i>	

DATA MINING ON UNCERTAIN DATA

Counterpropagation Neural Network for Stochastic Conditional Simulation An Application with Berea Sandstone	408
<i>Lance E. Besaw; Donna M. Rizzo</i>	
Genre Categorization of Web Pages	414
<i>Jebari Chaker; Ounelli Habib</i>	
Segmenting Multi-attribute Sequences Using Dynamic Bayesian Networks	424
<i>Robert Gwadera; Janne Toivola; Jaakko Hollmén</i>	

Error-Aware Density-Based Clustering of Imprecise Measurement Values	430
<i>Dirk Habich; Peter B. Volk; Wolfgang Lehner; Ralf Dittmann; Clemens Utzny</i>	
Skewed Class Distributions and Mislabeled Examples	436
<i>Jason Van Hulse; Taghi M. Khoshgoftaar; Amri Napolitano</i>	
Reducing UK-Means to K-Means	442
<i>S. D. Lee; Ben Kao; and Reynold Cheng</i>	
Efficient Mining of Frequent Patterns from Uncertain Data	448
<i>Carson Kai-Sang Leung; Christopher L. Carmichael; Boyu Hao</i>	
A Novel Ordering-Based Greedy Bayesian Network Learning Algorithm on Limited Data	454
<i>Feng Liu; Fengzhan Tian; Qiliang Zhu</i>	
Granularity Conscious Modeling for Probabilistic Databases	460
<i>Eirinaios Michelakis; Daisy Zhe Wang; Minos Garofalakis; Joseph M. Hellerstein</i>	
Representing Tuple and Attribute Uncertainty in Probabilistic Databases	466
<i>Prithviraj Sen; Amol Deshpande; Lise Getoor</i>	
Targeting Input Data for Acoustic Bird Species Recognition Using Data Mining	472
<i>Erika Vilches; Iván A. Escobar; Edgar E. Vallejo; Charles E. Taylor</i>	
Incremental Integration of Probabilistic Models Learned from Data	478
<i>Jian Xu; Pedrito Maynard-Zhang; Jianhua Chen</i>	

DATA STREAMING MINING AND MANAGEMENT

Incremental Quantization for Aging Data Streams	484
<i>Fatih Altıparmak; David Chiu; and Hakan Ferhatosmanoglu</i>	
Hierarchical Classifier Combination and Its Application in Networks Intrusion Detection	490
<i>Morteza Analoui; Behrouz Minaei Bidgoli; Mohammad Hossein Rezvani</i>	
An Approach for Incremental Semi-supervised SVM	496
<i>Wael Emara; Mehmed Kantardzic</i>	
Sequential Change Detection on Data Streams	502
<i>S. Muthukrishnan; Eric van den Berg; Yihua Wu</i>	
Optimal Window Change Detection	508
<i>Jan Peter Patist</i>	
High-Speed Identification of Language and Script	514
<i>Alan Ratner; Ron Loui</i>	
Infrequent Item Mining in Multiple Data Streams	520
<i>Budhaditya Saha; Mihai Lazarescu; Svetha Venkatesh</i>	
Stream Event Detection: A Unified Framework for Mining Outlier, Change and Burst Simultaneously over Data Stream	526
<i>Zhijian Yuan; Kai Du; Yan Jia; Jiajia Miao</i>	
Toward Behavioral Modeling of a Grid System: Mining the Logging and Bookkeeping Files	532
<i>Xiangliang Zhang; Michèle Sebag; Cécile Germain</i>	

SPATIAL AND SPATIO-TEMPORAL DATA MINING

Space-Time Interpolation and Uncertainty Assessment of an Extreme Precipitation Index Using Geostatistical Cosimulation	538
<i>Ana Cristina Costa; Amílcar Soares</i>	
On Regional Association Rule Scoping	544
<i>Wei Ding; Christoph F. Eick; Xiaojing Yuan; Jing Wang; Jean-Philippe Nicot</i>	
A Compact Representation of Spatio-temporal Data	550
<i>Sigal Elnekave; Mark Last; Oded Maimon</i>	
Modeling Fundamental Geo-Raster Operations with Array Algebra	556
<i>Angelica Garcia Gutierrez; Peter Baumann</i>	
Pattern Mining as Abduction: From Snapshots to Spatio-temporal Sequential Patterns	562
<i>Shyamanta M. Hazarika</i>	
Query Expansion Using Topic and Location	568
<i>Shu Huang; Qiankun Zhao; Prasenjit Mitra; C. Lee Giles</i>	
Knowledge Discovery in Entity Based Smart Environment Resident Data Using Temporal Relation Based Data Mining	574
<i>Vikramaditya R. Jakkula; Aaron S. Crandall; Diane J. Cook</i>	
Space-Time Summarization of Multisensor Time Series. Case of Missing Data	580
<i>Marc Joliveau; Florian De Vuyst</i>	
Spatial Clustering Using the Likelihood Function	586
<i>April Kerby; David Marx; Ashok Samal; Viacheslav Adamchuck</i>	
Diagnosing Similarity of Oscillation Trends in Time Series	592
<i>Leonardo E. Mariote; Claudia Bauzer Medeiros; Ricardo da S. Torres</i>	
Formulating, Identifying and Analyzing Individual Spatial Knowledge	598
<i>Falko Schmid</i>	
Using Statistics and Spatial Data Mining to Study Land Cover in Wyoming Can We Predict Vegetation Types from Environmental Variables?	604
<i>ZongBo Shang; Jeffery D. Hamerlinck</i>	
The Vegetation Outlook (VegOut): A New Tool for Providing Outlooks of General Vegetation Conditions Using Data Mining Techniques	610
<i>Tsegaye Tadesse; Brian Wardlow</i>	
A Hybrid Classification Scheme for Mining Multisource Geospatial Data	616
<i>Ranga Raju Vatsavai; Budhendra Bhaduri</i>	
Fast Mining of Complex Spatial Co-location Patterns Using GLIMIT	622
<i>Florian Verhein; Ghazi Al-Naymat</i>	
Spatio-temporal Analysis of the Relationship between South American Precipitation Extremes and the El Niño Southern Oscillation	628
<i>Elizabeth Wu; Sanjay Chawla</i>	

PRIVACY ASPECTS OF DATA MINING

Hiding Sensitive Trajectory Patterns	634
<i>Osman Abul; Maurizio Atzori; Francesco Bonchi; Fosca Giannotti</i>	
Privacy-Preserving k-NN for Small and Large Data Sets	640
<i>Artak Amirbekyan; Vladimir Estivill-Castro</i>	

A Secure Clustering Algorithm for Distributed Data Streams	646
<i>Geetha Jagannathan; Krishnan Pillaipakkammatt; D. Umamo</i>	
Private Inference Control for Aggregate Database Queries.....	652
<i>Geetha Jagannathan; Rebecca N. Wright</i>	
Privacy-Preserving Data Mining Applications in the Malicious Model	658
<i>Murat Kantarcioglu; Onur Kardeş</i>	
"Secure" Logistic Regression of Horizontally and Vertically Partitioned Distributed Database.....	664
<i>Aleksandra B. Slavkovic; Yuval Nardi; Matthew M. Tibbits</i>	
Simultaneous Pattern and Data Hiding in Unsupervised Learning.....	670
<i>Jie Wang; Jun Zhang; Lian Liu; Dianwei Han</i>	
Author Index	