

# **10th SIAM International Conference on Data Mining 2010**

**Proceedings in Applied Mathematics 136**

**Columbus, Ohio, USA  
29 April-1 May 2010**

**Volume 1 of 2**

**Editors:**

**Srinivasan Parthasarathy  
Bing Liu  
Bart Goethals**

**Jian Pei  
Chandrika Kamath**

**ISBN: 978-1-61738-652-7**

**Printed from e-media with permission by:**

Curran Associates, Inc.  
57 Morehouse Lane  
Red Hook, NY 12571



**Some format issues inherent in the e-media version may also appear in this print version.**

Copyright© (2010) by SIAM: Society for Industrial and Applied Mathematics  
All rights reserved.

Printed by Curran Associates, Inc. (2010)

For permission requests, please contact SIAM: Society for Industrial and Applied Mathematics  
at the address below.

SIAM  
3600 Market Street, 6th Floor  
Philadelphia, PA 19104-2688 USA

Phone: (215) 382-9800  
Fax: (215) 386-7999

[siambooks@siam.org](mailto:siambooks@siam.org)

**Additional copies of this publication are available from:**

Curran Associates, Inc.  
57 Morehouse Lane  
Red Hook, NY 12571 USA  
Phone: 845-758-0400  
Fax: 845-758-2634  
Email: [curran@proceedings.com](mailto:curran@proceedings.com)  
Web: [www.proceedings.com](http://www.proceedings.com)

# TABLE OF CONTENTS

Volume 1

## **SESSION S1: TEXT MINING**

<b>Text Categorization Using Word Similarities Based on Higher Order Co-Occurrences</b> .....	1
<i>Syed Fawad Hussain, Gilles Bisson</i>	
<b>Exploiting Associations between Word Clusters and Document Classes for Cross-Domain Text Categorization</b> .....	13
<i>Fuzhen Zhuang, Ping Luo, Hui Xiong, Qing He, Yuhong Xiong, Zhongzhi Shi</i>	
<b>Semi-Supervised Bio-Named Entity Recognition with Word-Codebook Learning</b> .....	25
<i>Pavel P. Kuksa, Yanjun Qi</i>	
<b>Improving Accessibility of Transaction-centric Web Objects</b> .....	37
<i>Muhammad Asiful Islam, Faisal Ahmed, Yevgen Borodin, Jalal Mahmud, I. V. Ramakrishnan</i>	

## **SESSION S2: PRIVACY AND TRUST**

<b>Reconstructing Randomized Social Networks</b> .....	49
<i>Niko Vuokko, Evimaria Terzi</i>	
<b>Reconstruction from Randomized Graph via Low Rank Approximation</b> .....	60
<i>Leting Wu, Xiaowei Ying, Xintao Wu</i>	
<b>Do You Trust to Get Trust? A Study of Trust Reciprocity Behaviors and Reciprocal Trust Prediction</b> .....	72
<i>Viet-An Nguyen, Ee-Peng Lim, Hwee-Hoon Tan, Jing Jiang, Aixin Sun</i>	
<b>Publishing Skewed Sensitive Microdata</b> .....	84
<i>Yabo Xu, Ke Wang, Ada Wai-Chee Fu, Raymond Chi-Wing Wong</i>	

## **SESSION S3: CLUSTERING**

<b>A SAT-Based Framework for Efficient Constrained Clustering</b> .....	94
<i>Ian Davidson, S. S. Ravi, Leonid Shamis</i>	
<b>Spectral and Semidefinite Relaxation of the CLUHSIC Algorithm</b> .....	106
<i>Wen-Yun Yang, James T. Kwok, Bao-Liang Lu</i>	
<b>Generation of Alternative Clusterings Using the CAMI Approach</b> .....	118
<i>Xuan Hong Dang, James Bailey</i>	
<b>Making k-Means Even Faster</b> .....	130
<i>Greg Hamerly</i>	

## **SESSION S4: PATTERN MAKING**

<b>On Mining Statistically Significant Attribute Association Information</b> .....	141
<i>Pritam Chanda, Jianmei Yang, Aidong Zhang, Murali Ramanathan</i>	
<b>An Information-Theoretic Approach to Finding Informative Noisy Tiles in Binary Databases</b> .....	153
<i>Kleanthis-Nikolaos Kontonasis, Tijl De Bie</i>	
<b>Mining Top-K Patterns from Binary Datasets in Presence of Noise</b> .....	165
<i>Claudio Lucchese, Salvatore Orlando, Raffaele Perego</i>	
<b>Formal Concept Sampling for Counting and Threshold-Free Local Pattern Mining</b> .....	177
<i>Mario Boley, Thomas Gärtner, Henrik Grosskreutz</i>	

## **SESSION S5: RECOMMENDATION**

<b>Alleviating the Sparsity Problem in Collaborative Filtering by Using an Adapted Distance and a Graph-Based Method</b> .....	189
<i>Beau Piccart, Jan Struyf, Hendrik Blockeel</i>	

<b>Collaborative Filtering: Weighted Nonnegative Matrix Factorization Incorporating User and Item Graphs</b> .....	199
<i>Quanquan Gu, Jie Zhou, Chris Ding</i>	
<b>Temporal Collaborative Filtering with Bayesian Probabilistic Tensor Factorization</b> .....	211
<i>Liang Xiong, Xi Chen, Tzu-Kuo Huang, Jeff Schneider, Jaime Carbonell</i>	
<b>Residual Bayesian Co-clustering for Matrix Approximation</b> .....	223
<i>Hanhuai Shan, Arindam Banerjee</i>	

## **SESSION S6: SUPPORT VECTOR MACHINES**

<b>Two-View Transductive Support Vector Machines</b> .....	235
<i>Guangxia Li, Steven C. H. Hoi, Kuiyu Chang</i>	
<b>Fast Stochastic Frank–Wolfe Algorithms for Nonlinear SVMs</b> .....	245
<i>Hua Ouyang, Alexander Gray</i>	
<b>Single-Pass Distributed Learning of Multi-class SVMs Using Core-Sets</b> .....	257
<i>Stefano Lodi, Ricardo Nanculef, Claudio Sartori</i>	
<b>Nonnegative Principal Component Analysis for Proteomic Tumor Profiles</b> .....	269
<i>Xiaoxu Han</i>	

## **SESSION S7: SUPERVISED LEARNING**

<b>Efficient Nonnegative Matrix Factorization with Random Projections</b> .....	281
<i>Fei Wang, Ping Li</i>	
<b>Bridging Domains with Words: Opinion Analysis with Matrix Tri-Factorizations</b> .....	293
<i>Tao Li, Vikas Sindhwani, Chris Ding, Yi Zhang</i>	
<b>Exact Passive-Aggressive Algorithm for Multiclass Classification Using Support Class</b> .....	303
<i>Shin Matsushima, Nobuyuki Shimizu, Kazuhiro Yoshida, Takashi Ninomiya, Hiroshi Nakagawa</i>	

## **SESSION S8: SPATIAL-TEMPORAL PATTERN MINING**

<b>Robust Mining of Time Intervals with Semi-Interval Partial Order Patterns</b> .....	315
<i>Fabian Moerchen, Dmitriy Fradkin</i>	
<b>Cascading Spatio-Temporal Pattern Discovery: A Summary of Results</b> .....	327
<i>Pradeep Mohan, Shashi Shekhar, James A. Shine, James P. Rogers</i>	
<b>Frequentness-Transition Queries for Distinctive Pattern Mining from Time-Segmented Databases</b> .....	339
<i>Shin-Ichi Minato, Takeaki Uno</i>	
<b>Consecutive Ones Property and Spectral Ordering</b> .....	350
<i>Niko Vuokko</i>	

## **SESSION S9: UNCERTAINTY IN DATA MINING**

<b>Naive Bayes Classifier for Positive Unlabeled Learning with Uncertainty</b> .....	361
<i>Jiazhen He, Yang Zhang, Xue Li, Yong Wang</i>	
<b>On Multidimensional Sharpening of Uncertain Data</b> .....	373
<i>Charu C. Aggarwal</i>	
<b>Subspace Clustering for Uncertain Data</b> .....	385
<i>Stephan Günnemann, Hardy Kremer, Thomas Seidl</i>	
<b>On the Use of Combining Rules in Relational Probability Trees</b> .....	397
<i>Daan Fierens</i>	

## **SESSION S10: CLUSTERING AND OUTLIER DETECTION**

<b>The Application of Statistical Relational Learning to a Database of Criminal and Terrorist Activity</b> .....	409
<i>B. Delaney, A. Fast, W. Campbell, C. Weinstein, D. Jensen</i>	
<b>ContextTour: Contextual Contour Analysis on Dynamic Multi-Relational Clustering</b> .....	418
<i>Yu-Ru Lin, Jimeng Sun, Nan Cao, Shixia Liu</i>	
<b>Identifying Multi-Instance Outliers</b> .....	430
<i>Ou Wu, Jun Gao, Weiming Hu, Bing Li, Mingliang Zhu</i>	

<b>Mining Actionable Subspace Clusters in Sequential Data</b> .....	442
<i>Kelvin Sim, Ardian Kristanto Poernomo, Vivekanand Gopalkrishnan</i>	

### **SESSION S11: GRAPH MINING**

<b>GraSS: Graph Structure Summarization</b> .....	454
<i>Kristen Lefevre, Evimaria Terzi</i>	
<b>Mining Frequent Graph Sequence Patterns Induced by Vertices</b> .....	466
<i>Akihiro Inokuchi, Takashi Washio</i>	
<b>On Clustering Graph Streams</b> .....	478
<i>Charu C. Aggarwal, Yuchen Zhao, Philip S. Yu</i>	
<b>Inferring Probability Distributions of Graph Size and Node Degree from Stochastic Graph Grammars</b> .....	490
<i>Sourav Mukherjee, Tim Oates</i>	

### **SESSION S12: FEATURE LEARNING AND PREDICTION**

<b>p-ISOMAP: An Efficient Parametric Update for ISOMAP for Visual Analytics</b> .....	502
<i>Jaegul Choo, Chandan K. Reddy, Hanseung Lee, Haesun Park</i>	
<b>Confidence-Based Feature Acquisition to Minimize Training and Test Costs</b> .....	514
<i>Marie Desjardins, James Macglashan, Kiri L. Wagstaff</i>	
<b>Co-Selection of Features and Instances for Unsupervised Rare Category Analysis</b> .....	525
<i>Jingrui He, Jaime Carbonell</i>	
<b>Active Ordering of Interactive Prediction Tasks</b> .....	537
<i>Abhimanyu Lad, Yiming Yang</i>	
<b>Radius Plots for Mining Tera-Byte Scale Graphs: Algorithms, Patterns, and Observations</b> .....	548
<i>U Kang, Charalampos E. Tsourakakis, Ana Paula Appel, Christos Faloutsos, Jure Leskovec</i>	

### **SESSION S13: MINING LARGE GRAPHS**

<b>Spectral Analysis of Signed Graphs for Clustering, Prediction and Visualization</b> .....	559
<i>Jérôme Kunegis, Stephan Schmidt, Andreas Lommatzsch, Jürgen Lerner, Ernesto W. De Luca, Sahin Albayrak</i>	
<b>Fast Single-Pair SimRank Computation</b> .....	571
<i>Pei Li, Hongyan Liu, Jeffrey Xu Yu, Jun He, Xiaoyong Du</i>	
<b>A Heterogeneous Label Propagation Algorithm for Disease Gene Discovery</b> .....	583
<i>Taehyun Hwang, Rui Kuang</i>	
<b>Direct Density Ratio Estimation with Dimensionality Reduction</b> .....	595
<i>Masashi Sugiyama, Satoshi Hara, Paul Von Büna, Taiji Suzuki, Takafumi Kanamori, Motoaki Kawanabe</i>	

### **SESSION S14: FEATURE SELECTION**

<b>The Generalized Dimensionality Reduction Problem</b> .....	607
<i>Charu C. Aggarwal</i>	
<b>Convex Principal Feature Selection</b> .....	619
<i>Mahdokht Masaali, Yan Yan, Ying Cui, Glenn Fung, Jennifer G. Dy</i>	
<b>Generalized and Heuristic-Free Feature Construction for Improved Accuracy</b> .....	629
<i>Wei Fan, Erheng Zhong, Jing Peng, Olivier Verscheure, Kun Zhang, Jiangtao Ren, Rong Yan, Qiang Yang</i>	

### **SESSION S15: TIME SERIES**

<b>Unsupervised Discovery of Abnormal Activity Occurrences in Multi-Dimensional Time Series, with Applications in Wearable Systems</b> .....	641
<i>Alireza Vahdatpour, Majid Sarrafzadeh</i>	
<b>An Integrated Framework for Simultaneous Classification and Regression of Time-Series Data</b> .....	653
<i>Zubin Abraham, Pang-Ning Tan</i>	
<b>Multiresolution Motif Discovery in Time Series</b> .....	665
<i>Nuno Castro, Paulo Azevedo</i>	

<b>Time-Series Classification in Many Intrinsic Dimensions</b> .....	677
<i>Miloš Radovanovic, Alexandros Nanopoulos, Mirjana Ivanovic</i>	

### **SESSION S16: TENSORS**

<b>MACH: Fast Randomized Tensor Decompositions</b> .....	689
<i>Charalampos E. Tsourakakis</i>	
<b>Scalable Tensor Factorizations with Missing Data</b> .....	701
<i>Evrin Acar, Daniel M. Dunlavy, Tamara G. Kolda, Morten Mørup</i>	
<b>On Low-Rank Updates to the Singular Value and Tucker Decompositions</b> .....	713
<i>Michael J. O'Hara</i>	

### **SESSION S17: SOCIAL NETWORK MINING**

<b>Towards Finding Valuable Topics</b> .....	720
<i>Zhen Wen, Ching-Yung Lin</i>	
<b>Predicting Customer Churn in Mobile Networks through Analysis of Social Groups</b> .....	732
<i>Yossi Richter, Elad Yom-Tov, Noam Slonim</i>	
<b>Directed Network Community Detection: A Popularity and Productivity Link Model</b> .....	742
<i>Tianbao Yang, Yun Chi, Shenghuo Zhu, Yihong Gong, Rong Jin</i>	
<b>HCDF: A Hybrid Community Discovery Framework</b> .....	754
<i>Keith Henderson, Tina Eliassi-Rad, Spiros Papadimitriou, Christos Faloutsos</i>	

### **SESSION S18: CLASSIFICATION**

<b>A Robust Decision Tree Algorithm for Imbalanced Data Sets</b> .....	766
<i>Wei Liu, Sanjay Chawla, David A. Cieslak, Nitesh V. Chawla</i>	
<b>Multi-Label Classification without the Multi-Label Cost</b> .....	778
<i>Xiatian Zhang, Quan Yuan, Shiwan Zhao, Wei Fan, Wentao Zheng, Zhong Wang</i>	

## Volume 2

<b>Fast and Accurate Gene Prediction by Decision Tree Classification</b> .....	790
<i>Rong She, Jeffrey Shih-Chieh Chu, Ke Wang, Nansheng Chen</i>	
<b>On Classification of High-Cardinality Data Streams</b> .....	802
<i>Charu C. Aggarwal, Philip S. Yu</i>	

### **SESSION S19: CLASSIFICATION AND APPLICATIONS**

<b>Predictive Modeling with Heterogeneous Sources</b> .....	814
<i>Xiaoxiao Shi, Qi Liu, Wei Fan, Qiang Yang, Philip S. Yu</i>	
<b>A Probabilistic Framework to Learn from Multiple Annotators with Time-Varying Accuracy</b> .....	826
<i>Pinar Donmez, Jaime Carbonell, Jeff Schneider</i>	
<b>An Integrative Approach to Identifying Biologically Relevant Genes</b> .....	838
<i>Zheng Zhao, Jiangxin Wang, Shashvata Sharma, Nitin Agarwal, Huan Liu, Yung Chang</i>	
<b>A Compression Based Distance Measure for Texture</b> .....	850
<i>Bilson J. L. Campana, Eamonn J. Keogh</i>	

### **SESSION S20: MACHINE LEARNING**

<b>Fast Implementation of <math>\ell_1</math> Regularized Learning Algorithms Using Gradient Descent Methods</b> .....	862
<i>Yunpeng Cai, Yijun Sun, Yubo Cheng, Jian Li, Steve Goodison</i>	
<b>Learning Compressible Models</b> .....	872
<i>Yi Zhang, Jeff Schneider, Artur Dubrawski</i>	
<b>A Permutation Approach to Validation</b> .....	882
<i>Malik Magdon-Ismail, Konstantin Mertsalov</i>	

<b>Adaptive Informative Sampling for Active Learning</b> .....	894
<i>Zhenyu Lu, Xindong Wu, Josh Bongard</i>	

### **SESSION S21: POTPOURRI**

<b>Evaluating Query Result Significance in Databases via Randomizations</b> .....	906
<i>Markus Ojala, Gemma C. Garriga, Aristides Gionis, Heikki Mannila</i>	
<b>Cross-Selling Optimization for Customized Promotion</b> .....	918
<i>Nan Li, Yinghui Yang, Xifeng Yan</i>	
<b>A Generalized Tree Matching Algorithm Considering Nested Lists for Web Data Extraction</b> .....	930
<i>Nitin Jindal, Bing Liu</i>	
<b>Mining Maximally Banded Matrices in Binary Data</b> .....	942
<i>Faris Alqadah, Raj Bhatnagar, Anil Jegga</i>	

### **EIGHTH WORKSHOP ON TEXT MINING**

<b>Text Mining 2010</b> .....	954
<i>Michael W. Berry, Jacob Kogan</i>	
<b>Discovering Generalized Concepts from Documents Using a Category Graph</b> .....	958
<i>Tom Vacek, John Joseph, Daniel Boley</i>	
<b>A Hybrid Recommender Systems Based on Weighted Tags</b> .....	968
<i>Huizhi Liang, Yue Xu, Yuefeng Li, Richi Nayak, Gavin Shaw</i>	
<b>A Fast Algorithm for Nonnegative Tensor Factorization Using Block Coordinate Descent and an Active-Set-Type Method</b> .....	978
<i>Krishnakumar Balasubramanian, Jingu Kim, Andrey Pureskiy, Michael W. Berry, Haesun Park</i>	
<b>Orthogonal Nonnegative Matrix Factorization for Multi-Type Relational Clustering</b> .....	988
<i>Chengcheng Shen, Ying Liu</i>	
<b>EDLSI with PSVD Updating</b> .....	997
<i>April Kontostathis, Erin Moulding, Raymond J. Spiteri</i>	
<b>Large-Scale Text Mining</b> .....	1007
<i>Alan Ratner</i>	
<b>Improving Document Clustering Using a Hierarchical Ontology Extracted From Wikipedia</b> .....	1017
<i>Mostafa M. Hassan, Fakhreddine Karray, Mohamed S. Kamel</i>	
<b>Regularized Co-Clustering on Manifold</b> .....	1026
<i>Chengcheng Shen, Ying Liu</i>	
<b>Enhancing Document Clustering Using Hybrid Models for Semantic Similarity</b> .....	1036
<i>Ahmed K. Farahat, Mohamed S. Kamel</i>	
<b>Information Bottleneck Co-Clustering</b> .....	1046
<i>Pu Wang, Carlotta Domeniconi, Kathryn Blackmond Laskey</i>	

### **WORKSHOP ON HIGH PERFORMANCE ANALYTICS**

<b>Workshop on High Performance Analytics: Algorithms, Implementations, and Applications</b> .....	1057
<i>Amol Ghoting, Rong Yan, Xifeng Yan</i>	
<b>Mining Frequent Highly-Correlated Item-Pairs at Very Low Support Levels</b> .....	1059
<i>Ian Sandler, Alex Thoma</i>	
<b>Database Support for Bregman Co-Clustering</b> .....	1069
<i>Kuo-Wei Hsu, Jaideep Srivastava</i>	
<b>A Data Intensive Multi-Chunk Ensemble Technique to Classify Stream Data Using Map-Reduce Framework</b> .....	1079
<i>Tahseen Al-Khateeb, Mohammad Salim Ahmed, Mohammad Masud, Latifur Khan</i>	

### **SIAM SDM WORKSHOP ON DATA MINING FOR SMARTER INFRASTRUCTURE**

<b>SIAM SDM Workshop on Data Mining for Smarter Infrastructure</b> .....	1089
<i>Milind Naphade, Lucio Soibelman</i>	
<b>Discovering Flow Anomalies on a Smarter Water Infrastructure System</b> .....	1091
<i>James M. Kang, Shashi Shekhar</i>	

<b>A Data Mining Framework for Pipeline Monitoring Using Time Reversal</b> .....	1097
<i>Yujie Yinga, Lucio Soibelman, Joel Harley, Nicholas O'Donoghue, James H. Garrett Jr., Yuanwei Jin, José M. F. Moura, Irving J. Oppenheim</i>	
<b>Analyzing Cell-Phone Mobility and Social Events</b> .....	1103
<i>Francesco Calabrese, Giusy Di Lorenzo, Francisco Pereira, Liang Liu</i>	
<b>Traffic Analysis for Holidays and Planned Events</b> .....	1107
<i>Jing Dai, Chang-Tien Lu, Xiang Fei</i>	
<b>Market-Based Agents for Distributed Data Processing in Wireless Sensor and Actuator Networks</b> .....	1113
<i>Jerome P. Lynch, Andrew T. Zimmerman, Frank T. Ferrese</i>	
<b>Pattern Recognition Models for Smarter Infrastructure Sensing</b> .....	1119
<i>I. Brilakis, S. German</i>	

## **WORKSHOP ON MACHINE LEARNING AND DATA MINING FOR SUSTAINABLE DEVELOPMENT 2010**

<b>Workshop on Machine Learning and Data Mining for Sustainable Development 2010</b> .....	1125
<i>Tina Yu</i>	
<b>Probabilistic Conic Mixture Model and its Applications to Mining Spatial Ground Penetrating Radar Data</b> .....	1127
<i>Huanhuan Chen, Anthony G. Cohn</i>	
<b>A Simple Parallel Projection Optimization Algorithm Estimating a Large-Sized Input-Output Table for Environmental Impact Assessment</b> .....	1136
<i>Ting Yu, Julien Ugon, Regina Burachik, Manfred Lenzen</i>	

## **MINISYMPOSIUM**

<b>Towards Natural Language Based Data/Text Mining and Summarization Via Soft Approaches</b> .....	1143
<i>Janusz Kacprzyk, Slawomir Zadrozny</i>	
<b>A Perspective on the Role of Fuzzy Logic, Computational Linguistics and Natural Language Processing in Data Mining and Data Summarization</b> .....	1145
<i>Janusz Kacprzyk, Slawomir Zadrozny</i>	
<b>Some Features of Partitions Useful for Linguistic Data Mining</b> .....	1157
<i>Ronald R. Yager</i>	
<b>Estimation of Topic Cardinality in Document Collections</b> .....	1168
<i>Antoon Bronselaer, Saskia Debergh, Dirk Van Hyfte, Guy De Tré</i>	
<b>Importance and Ontology-Based Enhancement of Concepts in Structured Queries</b> .....	1177
<i>Zhan Li, Marek Reformat, Ronald R. Yager</i>	
<b>Evaluation of Abstraction-Based Data Models for Text via Supervised Learning Methods</b> .....	1190
<i>Richard A. McAllister, Rafal A. Angryk</i>	

## **TUTORIALS**

<b>Tutorial 1: Mining Sparse Representations: Theory, Algorithms, and Applications</b> .....	1199
<i>Jun Liu, Shuiwang Ji, Jieping Ye</i>	
<b>Tutorial 2: Outlier Detection Techniques</b> .....	1359
<i>Hans-Peter Kriegel, Peer Kröger, Arthur Zimek</i>	
<b>Tutorial 3: Ranking Methods in Machine Learning</b> .....	1432
<i>Shivani Agarwal</i>	
<b>Tutorial 3: On the Power of Ensemble: Supervised and Unsupervised Methods Reconciled</b> .....	1433
<i>Jing Gao, Wei Fan, Jiawei Han</i>	
<b>Author Index</b>	