

9th SIAM International Conference on Data Mining 2009

Proceedings in Applied Mathematics 133

**Sparks, Nevada, USA
30 April - 2 May 2009**

Volume 1 of 3

ISBN: 978-1-61567-109-0

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© (2009) by SIAM: Society for Industrial and Applied Mathematics
All rights reserved.

Printed by Curran Associates, Inc. (2009)

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

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
Web: www.proceedings.com

TABLE OF CONTENTS

VOLUME 1

SESSION S1: CLUSTERING

GAD: General Activity Detection for Fast Clustering on Large Data	1
<i>Xin Jin, Sangkyum Kim, Jiawei Han, Liangliang Cao, Zhijun Yin</i>	
CORE: Nonparametric Clustering of Large Numeric Databases	13
<i>Andrej Taliun, Michael H. Böhlen, Arturas Mazeika</i>	
Constraint-Based Subspace Clustering	25
<i>Elisa Fromont, Adriana Prado, Céline Robardet</i>	
Integrated KL (K-means – Laplacian) Clustering: A New Clustering Approach by Combining Attribute Data and Pairwise Relations	37
<i>Fei Wang, Chris Ding, Tao Li</i>	
Hybrid Clustering of Text Mining and Bibliometrics Applied to Journal Sets	48
<i>Xinhai Liu, Shi Yu, Yves Moreau, Bart De Moor, Wolfgang Glänzel, Frizo Janssens</i>	

SESSION S2: TIME SERIES

Event Discovery in Time Series	60
<i>Dan Preston, Pavlos Protopapas, Carla Brodley</i>	
FuncICA for Time Series Pattern Discovery	72
<i>Nishant Mehta, Alexander Gray</i>	
Autocannibalistic and Anyspace Indexing Algorithms with Application to Sensor Data Mining	84
<i>Lexiang Ye, Xiaoyue Wang, Eamonn Keogh, Agenor Mafra-Neto</i>	
Proximity-Based Anomaly Detection Using Sparse Structure Learning	96
<i>Tsuyoshi Idé, Aurelie C. Lozano, Naoki Abe, Yan Liu</i>	
Optimal Distance Bounds on Time-Series Data	108
<i>Michail Vlachos, Suleyman S. Kozat, Philip S. Yu</i>	

SESSION S3: STATISTICAL METHODS AND APPLICATIONS

Application of Bayesian Partition Models in Warranty Data Analysis	120
<i>Markus Mueller, Christoph Schlieder, Axel Blumenstock</i>	
Learning Random-Walk Kernels for Protein Remote Homology Identification and Motif Discovery	132
<i>Renqiang Min, Rui Kuang, Anthony Bonner, Zhaolei Zhang</i>	
Outlier Detection with Globally Optimal Exemplar-Based GMM	144
<i>Xingwei Yang, Longin Jan Latecki, Dragoljub Pokrajac</i>	
Prior-Free Rare Category Detection	154
<i>Jingrui He, Jaime Carbonel</i>	
A Family of Large Margin Linear Classifiers and Its Application in Dynamic Environments	163
<i>Jianqiang Shen, Thomas G. Dietterich</i>	

SESSION S4: UNSUPERVISED LEARNING AND CLUSTERING

DensEst: Density Estimation for Data Mining in High Dimensional Spaces	172
<i>Emmanuel Müller, Ira Assent, Ralph Krieger, Stephan Günemann, Thomas Seidl</i>	
A Framework for Exploring Categorical Data	184
<i>Varun Chandola, Shyam Boriah, Vipin Kumar</i>	
Discovering Substantial Distinctions among Incremental Bi-Clusters	196
<i>Faris Alqadah, Raj Bhatnagar</i>	
Bayesian Cluster Ensembles	208
<i>Hongjun Wang, Hanhuai Shan, Arindam Banerjee</i>	

Agglomerative Mean-Shift Clustering via Query Set Compression	220
<i>Xiao-Tong Yuan, Bao-Gang Hu, Ran He</i>	

SESSION S5: SATA STREAM MINING

Adaptive Concept Drift Detection	232
<i>Anton Dries, Ulrich Rückert</i>	
Scalable Distributed Change Detection from Astronomy Data Streams Using Local, Asynchronous Eigen Monitoring Algorithms	244
<i>Kamalika Das, Kanishka Bhaduri, Sugandha Arora, Wesley Griffin, Kirk Borne, Chris Giannella, Hillol Kargupta</i>	
Positive Unlabeled Learning for Data Stream Classification	256
<i>Xiao-Li Li, Philip S. Yu, Bing Liu, See-Kiong Ng</i>	
Time-Decayed Correlated Aggregates over Data Streams	268
<i>Graham Cormode, Srikanth Tirthapura, Bojian Xu</i>	
Multi-Modal Hierarchical Dirichlet Process Model for Predicting Image Annotation and Image-Object Label Correspondence	280
<i>Oksana Yakhnenko, Vasant Honavar</i>	

POSTER SPOTLIGHTS

A Bayesian Approach to Graphy Regression with Relevant Subgraph Selection	291
<i>Silvia Chiappa, Hiroto Saigo, Koji Tsuda</i>	
A Hybrid Data Mining Metaheuristic for the p-Median Problem	301
<i>Alexandre Plastino, Erick R. Fonseca, Richard Fuchshuber, Simone de L. Martins, Alex A. Freitas, Martino Luis, Said Salhi</i>	
A New Constraint for Mining Sets in Sequences	313
<i>Boris Cule, Bart Goethals, Céline Robardet</i>	
A Re-evaluation of the Over-Searching Phenomenon in Inductive Rule Learning	325
<i>Frederik Janssen, Johannes Fürnkranz</i>	
A Semi-Supervised Framework for Feature Mapping and Multiclass Classification	337
<i>Bo Chen, Wai Lam, Ivor Tsang, Tak-Lam Wong</i>	
Aligned Graph Classification with Regularized Logistic Regression	349
<i>Brian Quanz, Jun Huan</i>	
An Entity Based Model for Coreference Resolution	361
<i>Michael Wick, Aron Culitta, Khashayar Rohanimanesh, Andrew McCallum</i>	
Analyses for Service Interaction Networks with Applications to Service Delivery	373
<i>S. Kameshwaran, Sameep Mehta, Vinayaka Pandit, Gyana Parija, Sudhanshu Singh, N. Viswanadham</i>	
Change-Point Detection in Time-Series Data by Direct Density-Ratio Estimation	385
<i>Yoshinobu Kawahara, Masashi Sugiyama</i>	
Context Aware Trace Clustering: Towards Improving Process Mining Results	397
<i>R. P. Jagadeesh Chandra Bose, Wil M. P. van der Aalst</i>	
Detection and Characterization of Anomalies in Multivariate Time Series	409
<i>Haibin Cheng, Pang-Ning Tan, Christopher Potter, Steven Klooster</i>	
Discovery of Geospatial Discriminating Patterns from Remote Sensing Datasets	421
<i>Wei Ding, Tomasz Stepinski, Josue Salazar</i>	
Diversity-Based Weighting Schemes for Clustering Ensembles	433
<i>Francesco Gullo, Andrea Tagarelli, Sergio Greco</i>	
Divide and Conquer Strategies for Effective Information Retrieval	445
<i>Jie Chen, Yousef Saad</i>	
Speeding Up Secure Computations via Embedded Caching	457
<i>K. Zhai, W. K. Ng, A. R. Herianto, S. Han</i>	
Exact Discovery of Time Series Motifs	469
<i>Abdullah Mueen, Eamonn Keogh, Qiang Zhu, Sydney Cash, Brandon Westover</i>	
Exploiting Semantic Constraints for Estimating Supersenses with CRFs	481
<i>Gerhard Paa?, Frank Reichartz</i>	
Feature Weighted SVMs Using Receiver Operating Characteristics	493
<i>Shaoyi Zhang, M. Maruf Hossain, Md. Rafiul Hassan, James Bailey, Kotagiri Ramamohanarao</i>	
FEDRA: A Fast and Efficient Dimensionality Reduction Algorithm	505
<i>Panagis Magdalinos, Christos Doulkeridis, Michalis Vazirgiannis</i>	
Finding Representative Association Rules from Large Rule Collections	517
<i>Warren L. Davis IV, Peter Schwarz, Evimaria Terzi</i>	

FutureRank: Ranking Scientific Articles by Predicting their Future PageRank	529
<i>Hassan Sayyadi, Lise Getoor</i>	

VOLUME 2

Highlighting Diverse Concepts in Documents	541
<i>Kun Liu, Evimaria Terzi, Tyrone Grandison</i>	
Identifying Information-Rich Subspace Trends in High-Dimensional Data	553
<i>Snehal Pokharkar, Chandan K. Reddy</i>	
Low-Entropy Set Selection	565
<i>Hannes Heikinheimo, Jilles Vreeken, Arno Siebes, Heikki Mannila</i>	
Measuring Discrimination in Socially-Sensitive Decision Records	577
<i>Dino Pedreschi, Salvatore Ruggieri, Franco Turini</i>	
Mining Cohesive Patterns from Graphs with Feature Vectors	589
<i>Flavia Moser, Recep Colak, Arash Rafiey, Martin Ester</i>	
Mining Complex Spatio-Temporal Sequence Patterns	601
<i>Florian Verhein</i>	
Mining for Surprise Events Within Text Streams	613
<i>Paul Whitney, Dave Engel, Nick Cramer</i>	
Multi-field Correlated Topic Modeling	624
<i>Konstantin Salomatin, Yiming Yang, Abhimanyu Lad</i>	
Multiple Kernel Clustering	634
<i>Bin Zhao, James T. Kwok, Changshui Zhang</i>	
MUSK: Uniform Sampling of k Maximal Patterns	646
<i>Mohammad Al Hasan, Mohammed Zaki</i>	
Noise Robust Classification Based on Spread Spectrum	658
<i>Joern David</i>	
Non-negative Matrix Factorization, Convexity and Isometry	669
<i>Nikolaos Vasiloglou, Alexander G. Gray, David V. Anderson</i>	
Non-parametric Information-Theoretic Measures of One-Dimensional Distribution Functions from Continuous Time Series	681
<i>Paolo D'Alberto, Ali Dasdan</i>	
On Maximum Coverage in the Streaming Model & Application to Multi-topic Blog-Watch	693
<i>Barna Saha, Lise Getoor</i>	
On Randomness Measures for Social Networks	705
<i>Xiaowei Ying, Xintao Wu</i>	
On Segment-Based Stream Modeling and Its Applications	717
<i>Charu C. Aggarwal</i>	
On the Comparison of Relative Clustering Validity Criteria	729
<i>Lucas Vendramin, Ricardo J. G. B. Campello, Eduardo R. Hruschka</i>	
Parallel Pairwise Clustering	741
<i>Elad Yom-Tov, Noam Slonim</i>	
PICC Counting: Who Needs Joins When You Can Propagate Efficiently?	752
<i>Jong Wook Kim, K. Selçuk Candan</i>	
Providing Privacy through Plausibly Deniable Search	764
<i>Mummooorthy Murugesan, Chris Clifton</i>	
Randomization Techniques for Graphs	776
<i>Sami Hanhijärvi, Gemma C. Garriga, Kai Puolamäki</i>	
Semi-supervised Learning by Sparse Representation	788
<i>Shuicheng Yan, Huan Wang</i>	
ShatterPlots: Fast Tools for Mining Large Graphs	798
<i>Ana Paula Appel, Deepayan Chakrabarti, Christos Faloutsos, Ravi Kumar, Jure Leskovec, Andrew Tomkins</i>	
Spatially Cost-Sensitive Active Learning	810
<i>Alexander Liu, Goo Jun, Joydeep Ghosh</i>	
Structure and Dynamics of Research Collaboration in Computer Science	822
<i>Christian Bird, Earl Barr, Andre Nash</i>	
Text Categorization with All Substring Features	834
<i>Daisuke Okanohara, Jun'ichi Tsujii</i>	
The Set Classification Problem and Solution Methods	843
<i>Xia Ning, George Karypis</i>	
Topic Evolution in a Stream of Documents	855
<i>André Gohr, Alexander Hinneburg, René Schult, Myra Spiliopoulou</i>	

Tracking User Mobility to Detect Suspicious Behavior	867
<i>Gaurav Tandon, Philip K. Chan</i>	

SESSION S6: SUPERVISED LEARNING

Toward Optimal Ordering of Prediction Tasks	879
<i>Abhimanyu Lad, Yiming Yang, Rayid Ghani, Bryan Kisiel</i>	
Hierarchical Linear Discriminant Analysis for Beamforming	889
<i>Jaegul Choo, Barry L. Drake, Haesun Park</i>	
Twin Vector Machines for Online Learning on a Budget	901
<i>Zhuang Wang, Slobodan Vucetic</i>	
The Metric Dilemma: Competence-Conscious Associative Classification	913
<i>Adriano Veloso, Mohammed Zaki, Wagner Meira Jr., Marcos Gonçalves</i>	

SESSION S7: PRIVACY AND SOCIAL NETWORKS

AMORI: A Metric-Based One Rule Inducer	925
<i>Niklas Lavesson, Paul Davidsson</i>	
Identifying Unsafe Routes for Network-Based Trajectory Privacy	937
<i>Aris Gkoulalas-Divanis, Vassilios S. Verykios, Mohamed F. Mokbel</i>	
Privacy Preservation in Social Networks with Sensitive Edge Weights	949
<i>Lian Liu, Jie Wang, Jinze Liu, Jun Zhang</i>	
Graph Generation with Prescribed Feature Constraints	961
<i>Xiaowei Ying, Xintao Wu</i>	
Detecting Communities in Social Networks Using Max-Min Modularity	973
<i>Jiyang Chen, Osmar R. Zaiane, Randy Goebel</i>	
A Bayesian Approach Toward Finding Communities and Their Evolutions in Dynamic Social Networks	985
<i>Tianbao Yang, Yun Chi, Shenghuo Zhu, Yihong Gong, Rong Jin</i>	
Efficient Discovery of Interesting Patterns Based on Strong Closedness	997
<i>Mario Boley, Tamás Horváth, Stefan Wrobel</i>	

SESSION S8: RELATIONAL MINING AND HIGH PERFORMANCE LEARNING

Efficient Computation of Partial-Support for Mining Interesting Itemsets	1009
<i>Ardian Kristanto Poernomo, Vivekanand Gopalkrishnan</i>	
Grammar Mining	1021
<i>Siegfried Nijssen, Luc De Raedt</i>	
Top-k Correlative Graph Mining	1033
<i>Yiping Ke, James Cheng, Jeffrey Xu Yu</i>	
High Performance Parallel/Distributed Biclustering Using Barycenter Heuristic	1045
<i>Arija Nisar, Waseem Ahmad, Wei-keng Liao, Alok Choudhary</i>	

SESSION S9: MINING GRAPHS AND SEMI STRUCTURED DATA

MultiVis: Content-Based Social Network Exploration through Multi-way Visual Analysis	1057
<i>Jimeng Sun, Spiros Papadimitriou, Ching-Yung Lin, Nan Cao, Shixia Liu, Weihong Qian</i>	
Near-optimal Supervised Feature Selection among Frequent Subgraphs	1069
<i>Marisa Thoma, Hong Cheng, Arthur Gretton, Jiawei Han, Hans-Peter Kriegel, Alex Smola, Le Song, Philip S. Yu, Xifeng Yan, Karsten Borgwardt</i>	
Polynomial-Delay and Polynomial-Space Algorithms for Mining Closed Sequences, Graphs, and Pictures in Accessible Set Systems	1081
<i>Hiroki Arimura, Takeaki Uno</i>	

VOLUME 3

Link Propagation: A Fast Semi-supervised Learning Algorithm for Link Prediction	1093
<i>Hisashi Kashima, Tsuyoshi Kato, Yoshihiro Yamanishi, Masashi Sugiyama, Koji Tsuda</i>	

Understanding Importance of Collaborations in Co-authorship Networks: A Supportiveness Analysis Approach	1105
<i>Yi Han, Bin Zhou, Jian Pei, Yan Jia</i>	

SESSION S10: TEXT MINING DATA REDUCTION

Topic Cube: Topic Modeling for OLAP on Multidimensional Text Databases	1117
<i>Duo Zhang, Chengxiang Zhai, Jiawei Han</i>	
Local Relevance Weighted Maximum Margin Criterion for Text Classification	1129
<i>Quanquan Gu, Jie Zhou</i>	
Multi-topic Based Query-Oriented Summarization	1141
<i>Jie Tang, Limin Yao, Dewei Chen</i>	
Straightforward Feature Selection for Scalable Latent Semantic Indexing	1153
<i>Jun Yan, Shuicheng Yan, Ning Liu, Zheng Chen</i>	
Parallel Large Scale Feature Selection for Logistic Regression	1165
<i>Sameer Singh, Jeremy Kubica, Scott Larsen, Daria Sorokina</i>	

SESSION S11: MINING SPATIO-TEMPORAL DATA AND EFFICIENT LEARNING

Travel-Time Prediction Using Gaussian Process Regression: A Trajectory-Based Approach	1177
<i>Tsuyoshi Idé, Sei Kato</i>	
Discretized Spatio-Temporal Scan Window	1189
<i>Seyed H. Mohammadi, Vandana P. Janeja, Aryya Gangopadhyay</i>	
Finding Links and Initiators: A Graph-Reconstruction Problem	1201
<i>Heikki Mannila, Evinaria Terzi</i>	
Efficient Multiplicative Updates for Support Vector Machines	1212
<i>Vamsi K. Potluru, Sergey M. Plis, Morten Mørup, Vincent D. Calhoun, Terran Lane</i>	
Efficient Active Learning with Boosting	1224
<i>Zheng Wang, Yangqiu Song, Changshui Zhang</i>	

LINK ANALYSIS WORKSHOP

Defence and Security Applications of Network Technologies	1236
<i>I. Pestov</i>	
Anomaly Detection using Scan Statistics on Time Series Hypergraphs	1243
<i>Y. PArk, C.R. Priebe, D.J. Marchette, A. Youseef</i>	
Discovering Research Domains Using Distance Matrix and Co-Authorship Network	1252
<i>S. Hassan, R. Ichise</i>	
Privacy- Enhancing Distributed Higher-Order ARM	1258
<i>A. Nikolov, S. Li, W.M. Pottenger</i>	
Privacy Vulnerabilites with Background Information in Data Pertubation	1268
<i>L. Liu J. Wang, J. Zhang</i>	
PaCK: Scalable Parameter-Free Clustering on K-Partite Graphs	1278
<i>J. Carbonell, T. Eliassi-Rad, C. Faloutsos, J. He, S. Papadimtriou, H. Tong</i>	

TEXT MINING WORKSHOP

Text Mining 2009	1288
<i>M. Berry, J. Kogan</i>	
Mining for Emerging Technologies within Text Streams and Documents	1291
<i>D. Engel, P. Whitney, G. Calapristi, F. Brockman</i>	
Investigating the Role of Past Inferred Semantics in Generating Furuter Topic Models of Text Streams	1302
<i>L. AlSumait, D. Barbará</i>	
Threshold Setting and Performance Monitoring for Novel Text Mining	1310
<i>W. Tang, F. Tsai</i>	
FutureLens: Software for Text Visualization and Tracking	1320
<i>G.L. Shutt, A.A. Puretskiy, M.W. Berry</i>	
ChatCoder: Toward the Tracking and Categorization of Internet Predators	1327
<i>Dr. A. Kontostathis, Dr. L. Edwards, A. Leatherman</i>	

Content-Based Spam Filtering by Machine Learning Algorithms	1335
<i>E. Jiang</i>	
E-Mail Classification Based on NMF	1345
<i>A.G.K. Janecek, W.N. Gansterer</i>	
Constrained Clustering with k-means	1355
<i>Z. Su, J. Kogan, C. Nicholas</i>	

DYNAMIC NETWORKS WORKSHOP

ADN 09: 2nd International Workshop on Analysis of Dynamic Networks	1363
<i>T. Berger-Wolf, M. Magdon-Ismail, J. Saia</i>	
The Impact of Changes in Network Structure on Diffusion of Warnings	1364
<i>C. Hui, M. Magdon-Ismail, W.A. Wallace, M. Goldberg</i>	
Uncovering Cross-Dimension Group Structures in Multi-Dimensional Networks	1374
<i>L. Tang, H. Liu</i>	

MULTIMEDIA DATA MINING

Introduction to the Workshop on Multimedia Data Mining	1382
<i>J. Pan, M.L. Sapino</i>	
Multimedia Data Mining Theory and its Application in Semantic Imagery and Video Retrieval	1384
<i>Z. Zhang</i>	
Augmenting Points of Interest Recommendations with Music	1385
<i>M. Karminskas, F. Ricci</i>	
An Effective Video Retrieval System by Combining Visual and Textual Mining Techniques	1395
<i>J-H. Su, H-H. Yeh, V.S. Tseng</i>	
Employing Fractal Dimension to Analyze Climate and Remote Sensing Data Streams	1404
<i>L.A.S. Romani, M.X. Ribeiro, E.P.M. de Sousa, C. Traina Jr., A.J.M. Traina, J. Zullo Jr.</i>	
Multimedia Data Mining Workflows: Efficiency and Effectiveness	1416
<i>K.S. Candan</i>	

TUTORIALS

A Geometric Perspective on Dimensionality Reduction	1417
<i>D. Cai, X. He, J. Han</i>	
Mining When Classes are Imbalanced, Rare Events Matter More, and Errors Have Costs Attached	1476
<i>N. Chawla</i>	
Exploring Interaction Networks for Services Industry	1529
<i>S. Kameshwaran, V. Pandit, S. Mehta</i>	
Data Mining with Graphs and Matrices	1580
<i>F. Wang, T. Li, C. Ding</i>	

EXTRA ABSTRACTS

Automated Learning and Data Visualization	1637
<i>W.S. Cleveland</i>	
Applied Nonparametric Bayes	1638
<i>M.I. Jordan</i>	
A Geometric Perspective on Machine Learning and Data Mining	1639
<i>P. Niyogi</i>	
Semantics on the Web: How do we get there	1640
<i>R. Ramakrishnan</i>	
Author Index	