

2008 IEEE International Conference on Data Mining (ICDM)

**Pisa, Italy
15 – 19 December 2008**

Pages 1-557



**IEEE Catalog Number: CFP08278-PRT
ISBN: 978-1-4244-4101-3**

TABLE OF CONTENTS

Regular papers

On-line LDA: Adaptive Topic Models for Mining Text Streams with Applications to Topic Detection and Tracking	1
<i>Loulwah AlSumait, Daniel Barbará, Carlotta Domeniconi</i>	
Unsupervised Cross-Domain Learning by Interaction Information Co-clustering	11
<i>Shin Ando, Einoshin Suzuki</i>	
Paired Learners for Concept Drift	21
<i>Stephen H. Bach, Marcus A. Maloof</i>	
Predicting Future Decision Trees from Evolving Data	31
<i>Mirko Böttcher, Martin Spott, Rudolf Kruse</i>	
A Randomized Approach for Approximating the Number of Frequent Sets	41
<i>Mario Boley, Henrik Grosskreutz</i>	
A Non-parametric Semi-supervised Discretization Method	51
<i>A. Bondu, M. Boulle, V. Lemaire, S. Loiseau, B. Duval</i>	
Non-negative Matrix Factorization on Manifold	61
<i>Deng Cai, Xiaofei He, Xiaoyun Wu, Jiawei Han</i>	
Anti-monotonic Overlap-Graph Support Measures	71
<i>Toon Calders, Jan Ramon, Dries Van Dyck</i>	
SeqStream: Mining Closed Sequential Patterns over Stream Sliding Windows	81
<i>Lei Chang, Tengjiao Wang, Dongqing Yang, Hua Luan</i>	
SPARCL: Efficient and Effective Shape-Based Clustering	91
<i>Vineet Chaoji, Mohammad Al Hasan, Saeed Salem, Mohammed J. Zaki</i>	
Graph OLAP: Towards Online Analytical Processing on Graphs	101
<i>Chen Chen, Xifeng Yan, Feida Zhu, Jiawei Han, Philip S. Yu</i>	
Exploiting Local and Global Invariants for the Management of Large Scale Information Systems	111
<i>Haifeng Chen, Haibin Cheng, Guofei Jiang, Kenji Yoshihira</i>	
DECK: Detecting Events from Web Click-Through Data	121
<i>Ling Chen, Yiqun Hu, Wolfgang Nejdl</i>	
Mining Order-Preserving Submatrices from Data with Repeated Measurements	131
<i>Chun Kit Chui, Ben Kao, Kevin Y. Yip, Sau Dan Lee</i>	
Start Globally, Optimize Locally, Predict Globally: Improving Performance on Imbalanced Data	141
<i>David A. Cieslak, Nitesh V. Chawla</i>	
Generalized Framework for Syntax-Based Relation Mining	151
<i>Bonaventura Coppola, Alessandro Moschitti, Daniele Pighin</i>	
Formal Models for Expert Finding on DBLP Bibliography Data	161
<i>Hongbo Deng, Irwin King, Michael R. Lyu</i>	
ReDSOM: Relative Density Visualization of Temporal Changes in Cluster Structures Using Self-Organizing Maps	171
<i>Denny, Graham J. Williams, Peter Christen</i>	

Nonnegative Matrix Factorization for Combinatorial Optimization: Spectral Clustering, Graph Matching, and Clique Finding	181
<i>Chris Ding, Tao Li, Michael I. Jordan</i>	
Space Efficient String Mining under Frequency Constraints	191
<i>Johannes Fischer, Veli Mäkinen, Niki Välimäki</i>	
Efficient Discovery of Statistically Significant Association Rules	201
<i>Wilhelmiina Hämmäläinen, Matti Nykänen</i>	
Interpreting PET Scans by Structured Patient Data: A Data Mining Case Study in Dementia Research	211
<i>Andreas Hapfelmeier, Jana Schmidt, Marianne Mueller, Stefan Kramer, Robert Perneczky, Alexander Kurz, Alexander Drzezga</i>	
Inlier-Based Outlier Detection via Direct Density Ratio Estimation	221
<i>Shohei Hido, Yuta Tsuboi, Hisashi Kashima, Masashi Sugiyama, Takafumi Kanamori</i>	
Supervised Inductive Learning with Lotka-Volterra Derived Models	231
<i>Karen Hovsepian, Peter Anselmo, Subhasish Mazumdar</i>	
A Novel Language-Model-Based Approach for Image Object Mining and Re-ranking	241
<i>Jen-Hao Hsiao, Chu-Song Chen, Ming-Syan Chen</i>	
Maximum Margin Clustering with Pairwise Constraints	251
<i>Yang Hu, Jingdong Wang, Nenghai Yu, Xian-Sheng Hua</i>	
Collaborative Filtering for Implicit Feedback Datasets	261
<i>Yifan Hu, Yehuda Koren, Chris Volinsky</i>	
Semi-supervised Learning from General Unlabeled Data	271
<i>Kaizhu Huang, Zenglin Xu, Irwin King, Michael R. Lyu</i>	
Metropolis Algorithms for Representative Subgraph Sampling	281
<i>Christian Hübler, Hans-Peter Kriegel, Karsten Borgwardt, Zoubin Ghahramani</i>	
Learning on Weighted Hypergraphs to Integrate Protein Interactions and Gene Expressions for Cancer Outcome Prediction	291
<i>TaeHyun Hwang, Ze Tian, Rui Kuangy, Jean-Pierre Kocher</i>	
A Fast Method to Mine Frequent Subsequences from Graph Sequence Data	301
<i>Akihiro Inokuchi, Takashi Washio</i>	
Overlapping Matrix Pattern Visualization: A Hypergraph Approach	311
<i>Ruoming Jin, Yang Xiang, David Fuhry, Feodor F. Dragan</i>	
A Robust Discriminative Term Weighting Based Linear Discriminant Method for Text Classification	321
<i>Khurum Nazir Junejo, Asim Karim</i>	
Clustering Uncertain Data Using Voronoi Diagrams	331
<i>Ben Kao, Sau Dan Lee, David W. Cheung, Wai-Shing Ho, K. F. Chan</i>	
SCS: A New Similarity Measure for Categorical Sequences	341
<i>Abdellali Kelil, Shengrui Wang</i>	
Toward Faster Nonnegative Matrix Factorization: A New Algorithm and Comparisons	351
<i>Jingu Kim, Haesun Park</i>	
Scalable Tensor Decompositions for Multi-aspect Data Mining	361
<i>Tamara G. Kolda, Jimeng Sun</i>	
Mining Periodic Behavior in Dynamic Social Networks	371
<i>Mayank Lahiri, Tanya Y. Berger-Wolf</i>	

Unsupervised Face Annotation by Mining the Web	381
<i>Duy-Dinh Le, Shin'ichi Satoh</i>	
Border Sampling through Coupling Markov Chain Monte Carlo	391
<i>Guichong Li, Nathalie Japkowicz, Trevor J. Stocki, R. Kurt Ungar</i>	
Computationally Efficient Estimators for Dimension Reductions Using Stable Random Projections	401
<i>Ping Li</i>	
Isolation Forest	411
<i>Fei Tony Liu, Kai Ming Ting, Zhi-Hua Zhou</i>	
TEFE: A Time-Efficient Approach to Feature Extraction	421
<i>Li-Ping Liu, Yang Yu, Yuan Jiang, Zhi-Hua Zhou</i>	
Transductive Component Analysis	431
<i>Wei Liu, Dacheng Tao, Jianzhuang Liu</i>	
Modeling and Predicting the Helpfulness of Online Reviews	441
<i>Yang Liu, Xiangji Huang, Aijun An, Xiaohui Yu</i>	
LBF: A Labeled-Based Forecasting Algorithm and Its Application to Electricity Price Time Series	451
<i>Francisco Martínez-Álvarez, Alicia Troncoso, José C. Riquelme, Jesús S. Aguilar-Ruiz</i>	
Enhancing the Stability of Spectral Ordering with Sparsification and Partial Supervision: Application to Paleontological Data	460
<i>Dimitrios Mavroeidis, Ella Bingham</i>	
Scaling up Classifiers to Cloud Computers	470
<i>Christopher Moretti, Karsten Steinhaeuser, Douglas Thain, Nitesh V. Chawla</i>	
What Sperner Family Concept Class is Easy to Be Enumerated?	480
<i>Atsuyoshi Nakamura, Mineichi Kudo</i>	
Learning by Propagability	490
<i>Bingbing Ni, Shuicheng Yan, Ashraf Kassim, Loong Fah Cheong</i>	
One-Class Collaborative Filtering	500
<i>Rong Pan, Yunhong Zhou, Bin Cao, Nathan N. Liu, Rajan Lukose, Martin Scholz, Qiang Yang</i>	
DisCo: Distributed Co-clustering with Map-Reduce: A Case Study towards Petabyte-Scale End-to-End Mining	510
<i>Spiros Papadimitriou, Jimeng Sun</i>	
Learning Bayesian Networks: A MAP Criterion for Joint Selection of Model Structure and Parameter	520
<i>Carsten Riggelsen</i>	
Bayesian Co-clustering	528
<i>Hanhuai Shan, Arindam Banerjee</i>	
Temporal-Relational Classifiers for Prediction in Evolving Domains	538
<i>Umang Sharan, Jennifer Neville</i>	
xCrawl: A High-Recall Crawling Method for Web Mining	548
<i>Kostyantyn Shchekotykhin, Dietmar Jannach, Gerhard Friedrich</i>	
Comparison of Cluster Representations from Partial Second- to Full Fourth-Order Cross Moments for Data Stream Clustering	558
<i>Mingzhou (Joe) Song, Lin Zhang</i>	
Web Mining for Understanding Stories through Graph Visualisation	568
<i>Ilija Subašić, Bettina Berendt</i>	

Balancing Spectral Clustering for Segmenting Spatio-temporal Observations of Multi-agent Systems	578
<i>Bálint Takács, Yiannis Demiris</i>	
Finding Good Itemsets by Packing Data	586
<i>Nikolaj Tatti, Jilles Vreeken</i>	
Measuring Proximity on Graphs with Side Information	596
<i>Hanghang Tong, Huiming Qu, Hani Jamjoom</i>	
Fast Counting of Triangles in Large Real Networks without Counting: Algorithms and Laws	606
<i>Charalampos E. Tsourakakis</i>	
Improving Collaborative Filtering Recommendations Using External Data	616
<i>Akhmed Umyarov, Alexander Tuzhilin</i>	
A Generative Probabilistic Model for Multi-label Classification	626
<i>Hongning Wang, Minlie Huang, Xiaoyan Zhu</i>	
SpecVAT: Enhanced Visual Cluster Analysis	636
<i>Liang Wang, Xin Geng, James Bezdek, Christopher Leckie, Ramamohanarao Kotagiri</i>	
Dirichlet Process Based Evolutionary Clustering	646
<i>Tianbing Xu, Zhongfei (Mark) Zhang, Philip S. Yu, Bo Long</i>	
Evolutionary Clustering by Hierarchical Dirichlet Process with Hidden Markov State	656
<i>Tianbing Xu, Zhongfei (Mark) Zhang, Philip S. Yu, Bo Long</i>	
TOFA: Trace Oriented Feature Analysis in Text Categorization	666
<i>Jun Yan, Ning Liu, Qiang Yang, Weiguo Fan, Zheng Chen</i>	
Clustering Distributed Time Series in Sensor Networks	676
<i>Jie Yin, Mohamed Medhat Gaber</i>	
M3MIML: A Maximum Margin Method for Multi-instance Multi-label Learning	686
<i>Min-Ling Zhang, Zhi-Hua Zhou</i>	
<u>Short Papers</u>	
RTM: Laws and a Recursive Generator for Weighted Time-Evolving Graphs	696
<i>Leman Akoglu, Mary McGlohon, Christos Faloutsos</i>	
A Shrinkage Approach for Modeling Non-stationary Relational Autocorrelation	702
<i>Pelin Angin, Jennifer Neville</i>	
Latent Dirichlet Allocation and Singular Value Decomposition Based Multi-document Summarization	708
<i>Rachit Arora, Balaraman Ravindran</i>	
INSCY: Indexing Subspace Clusters with In-Process-Removal of Redundancy	714
<i>Ira Assent, Ralph Krieger, Emmanuel Müller, Thomas Seidl</i>	
A Conservative Feature Subset Selection Algorithm with Missing Data	720
<i>Alex Aussem, Sergio Rodrigues de Morais</i>	
Nonparametric Monotone Classification with MOCA	726
<i>Nicola Barile, Ad Feelders</i>	
Mining Large Networks with Subgraph Counting	732
<i>Ilaria Bordino, Debora Donato, Aristides Gionis, Stefano Leonardi</i>	
Comparative Evaluation of Anomaly Detection Techniques for Sequence Data	738
<i>Varun Chandola, Varun Mithal, Vipin Kumar</i>	

On Locally Linear Classification by Pairwise Coupling	744
<i>Feng Chen, Chang-Tien Lu, Arnold P. Boedihardjo</i>	
A Probability Model for Projective Clustering on High Dimensional Data	750
<i>Lifei Chen, Qingshan Jiang, Shengrui Wang</i>	
Estimating Aggregates over Multiple Sets	756
<i>Edith Cohen, Haim Kaplan</i>	
A Joint Matrix Factorization Approach to Unsupervised Action Categorization	762
<i>Peng Cui, Fei Wang, Li-Feng Sun, Shi-Qiang Yang</i>	
Finding Alternative Clusterings Using Constraints	768
<i>Ian Davidson, Zijie Qi</i>	
Efficient Feature Selection in the Presence of Multiple Feature Classes	774
<i>Paramveer S. Dhillon, Dean Foster, Lyle H. Ungar</i>	
Why Stacked Models Perform Effective Collective Classification	780
<i>Andrew Fast, David Jensen</i>	
Multiplicative Mixture Models for Overlapping Clustering	786
<i>Qiang Fu, Arindam Banerjee</i>	
Anomaly Detection Support Vector Machine and Its Application to Fault Diagnosis	792
<i>Ryohei Fujimaki</i>	
A Recommendation System for Preconditioned Iterative Solvers	798
<i>Thomas George, Anshul Gupta, Vivek Sarin</i>	
Cost-Sensitive Parsimonious Linear Regression	804
<i>Robby Goetschalckx, Kurt Driessens, Scott Sanner</i>	
Text Mining in Radiology Reports	810
<i>Tianxia Gong, Chew Lim Tan, Tze Yun Leong, Cheng Kiang Lee, Boon Chuan Pang, C. C. Tchoyoson Lim, Qi Tian, Suisheng Tang, Zhuo Zhang</i>	
A Hierarchical Algorithm for Clustering Uncertain Data via an Information-Theoretic Approach	816
<i>Francesco Gullo, Giovanni Ponti, Andrea Tagarelli, Sergio Greco</i>	
Discovering Significant Patterns in Multi-stream Sequences	822
<i>Robert Gwadera, Fabio Crestani</i>	
Graph-Based Rare Category Detection	828
<i>Jingrui He, Yan Liu, Richard Lawrence</i>	
Clustering Documents with Active Learning Using Wikipedia	834
<i>Anna Huang, David Milne, Eibe Frank, Ian H. Witten</i>	
Direct Zero-Norm Optimization for Feature Selection	840
<i>Kaizhu Huang, Irwin King, Michael R. Lyu</i>	
Discovering Flow Anomalies: A SWEET Approach	846
<i>James M. Kang, Shashi Shekhar, Christine Wennen, Paige Novak</i>	
Boosting Relational Sequence Alignments	852
<i>Andreas Karwath, Kristian Kersting, Niels Landwehr</i>	
Support Vector Regression for Censored Data (SVRC): A Novel Tool for Survival Analysis	858
<i>Faisal M. Khan, Valentina Bayer Zubek</i>	
Nearest Neighbour Classifiers for Streaming Data with Delayed Labelling	864
<i>Ludmila I. Kuncheva, J. Salvador Sánchez</i>	

WiFiViz: Effective Visualization of Frequent Itemsets	870
<i>Carson Kai-Sang Leung, Pourang P. Irani, Christopher L. Carmichael</i>	
Fast and Memory Efficient Mining of High Utility Itemsets in Data Streams	876
<i>Hua-Fu Li, Hsin-Yun Huang, Yi-Cheng Chen, Yu-Jiun Liu, Suh-Yin Lee</i>	
HIREL: An Incremental Clustering Algorithm for Relational Datasets	882
<i>Tao Li, Sarabjot S. Anand</i>	
Time Sensitive Ranking with Application to Publication Search	888
<i>Xin Li, Bing Liu, Philip Yu</i>	
Releasing the SVM Classifier with Privacy-Preservation	894
<i>Keng-Pei Lin, Ming-Syan Chen</i>	
Text Cube: Computing IR Measures for Multidimensional Text Database Analysis	900
<i>Cindy Xide Lin, Bolin Ding, Jiawei Han, Feida Zhu, Bo Zhao</i>	
Multi-Space-Mapped SVMs for Multi-class Classification	906
<i>Bo Liu, Longbing Cao, Philip S. Yu, Chengqi Zhang</i>	
Spotting Significant Changing Subgraphs in Evolving Graphs	912
<i>Zheng Liu, Jeffrey Xu Yu, Yiping Ke, Xuemin Lin, Lei Chen</i>	
Classifying High-Dimensional Text and Web Data Using Very Short Patterns	918
<i>Hassan H. Malik, John R. Kender</i>	
A Practical Approach to Classify Evolving Data Streams: Training with Limited Amount of Labeled Data	924
<i>Mohammad M. Masud, Jing Gao, Latifur Khan, Jiawei Han, Bhavani Thuraisingham</i>	
Spatiotemporal Relational Probability Trees: An Introduction	930
<i>Amy McGovern, Nathan C. Hiers, Matthew Collier, David J. Gagne_II, Rodger A. Brown</i>	
Stream Sequential Pattern Mining with Precise Error Bounds	936
<i>Luiz F. Mendes, Bolin Ding, Jiawei Han</i>	
Organic Pie Charts	942
<i>Fabian Moerchen</i>	
Frequent Subgraph Retrieval in Geometric Graph Databases	948
<i>Sebastian Nowozin, Koji Tsuda</i>	
Alert Detection in System Logs	954
<i>Adam J. Oliner, Alex Aiken, Jon Stearley</i>	
Variance Minimization Least Squares Support Vector Machines for Time Series Analysis	960
<i>Róbert Ormándi</i>	
Quantitative Association Analysis Using Tree Hierarchies	966
<i>Feng Pan, Lynda Yang, Leonard McMillan, Fernando Pardo Manuel de Villena, David Threadgill, Wei Wang</i>	
Sparse Maximum Margin Logistic Regression for Credit Scoring	972
<i>Sabyasachi Patra, Kripa Shanker, Debasis Kundu</i>	
Similarity Learning for Nearest Neighbor Classification	978
<i>Ali Mustafa Qamar, Eric Gaussier, Jean-Pierre Chevallet, Joo Hwee Lim</i>	
RBNBC: Repeat Based Naive Bayes Classifier for Biological Sequences	984
<i>Pratibha Rani, Vikram Pudi</i>	
Multi-label Classification Using Ensembles of Pruned Sets	990
<i>Jesse Read, Bernhard Pfahringer, Geoff Holmes</i>	

Active Learning of Equivalence Relations by Minimizing the Expected Loss Using Constraint Inference	996
<i>Steffen Rendle, Lars Schmidt-Thieme</i>	
Iterative Subgraph Mining for Principal Component Analysis	1002
<i>Hiroto Saigo, Koji Tsuda</i>	
Clustering Geospatial Objects via Hidden Markov Random Fields	1008
<i>Makoto Sato, Shuuichiro Imahara</i>	
Collective Latent Dirichlet Allocation	1014
<i>Zhi-Yong Shen, Jun Sun, Yi-Dong Shen</i>	
Document-Word Co-regularization for Semi-supervised Sentiment Analysis	1020
<i>Vikas Sindhwani, Prem Melville</i>	
A Non-parametric Approach to Pair-Wise Dynamic Topic Correlation Detection	1026
<i>Yang Song, Lu Zhang, C. Lee Giles</i>	
Block-Iterative Algorithms for Non-negative Matrix Approximation	1032
<i>Suvrit Sra</i>	
A Novel Method of Combined Feature Extraction for Recognition	1038
<i>Tingkai Sun, Songcan Chen, Jingyu Yang, Pengfei Shi</i>	
Prediction of Skin Penetration Using Machine Learning Methods	1044
<i>Yi Sun, Gary P. Moss, Maria Prapopoulou, Rod Adams, Marc B. Brown, Neil Davey</i>	
A Topic Modeling Approach and Its Integration into the Random Walk Framework for Academic Search	1050
<i>Jie Tang, Ruoming Jin, Jing Zhang</i>	
Sequence Mining Automata: A New Technique for Mining Frequent Sequences under Regular Expressions	1056
<i>Roberto Trasarti, Francesco Bonchi, Bart Goethals</i>	
Filling in the Blanks - Krimp Minimisation for Missing Data	1062
<i>Jilles Vreeken, Arno Siebes</i>	
Computational Discovery of Motifs Using Hierarchical Clustering Techniques	1068
<i>Dianhui Wang, Nung Kion Lee</i>	
Inference Analysis in Privacy-Preserving Data Re-publishing	1074
<i>Guan Wang, Zutao Zhu, Wenliang Du, Zhouxuan Teng</i>	
Using Wikipedia for Co-clustering Based Cross-Domain Text Classification	1080
<i>Pu Wang, Carlotta Domeniconi, Jian Hu</i>	
Iterative Set Expansion of Named Entities Using the Web	1086
<i>Richard C. Wang, William W. Cohen</i>	
Experimental Evaluation of the Value of Structure: How to Efficiently Exploit Interdependencies in Sequence Labeling	1092
<i>Guillaume Wisniewski, Patrick Gallinari</i>	
Pseudolikelihood EM for Within-network Relational Learning	1098
<i>Rongjing Xiang, Jennifer Neville</i>	
Publishing Sensitive Transactions for Itemset Utility	1104
<i>Yabo Xu, Benjamin C. M. Fung, Ke Wang, Ada W. C. Fu, Jian Pei</i>	
Learning the Latent Semantic Space for Ranking in Text Retrieval	1110
<i>Jun Yan, Shuicheng Yan, Ning Liu, Zheng Chen</i>	

Robust Time-Referenced Segmentation of Moving Object Trajectories	1116
<i>Hyunjin Yoon, Cyrus Shahabi</i>	
Maximum Margin Embedding	1122
<i>Bin Zhao, Fei Wang, Changshui Zhang</i>	
Graph-Based Iterative Hybrid Feature Selection	1128
<i>ErHeng Zhong, Sihong Xie, Wei Fan, Jiangtao Ren, Jing Peng, Kun Zhang</i>	
Cleansing Noisy Data Streams	1134
<i>Xingquan Zhu, Peng Zhang, Xindong Wu, Dan He, Chengqi Zhang, Yong Shi</i>	
Author Index	