

2007 IEEE 7th International Conference on Data Mining (ICDM 2007)

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



**IEEE Catalog Number: CFP07278-PRT
ISBN: 978-1-4244-3031-4**



Proceedings of ICDM 2007

Table of Contents

Welcome Message from the Conference Chairs

Preface

Conference Organization

Program Committee

Non-PC Reviewers

Corporate Sponsors

Invited Speakers and Their Talk Descriptions

Tutorials and Their Descriptions

REGULAR PAPERS

How Much Noise Is Too Much: A Study in Automatic Text Classification	3
<i>Sumeet Agarwal, Shantanu Godbole, Diwakar Punjani, and Shourya Roy</i>	
Clustering Needles in a Haystack: An Information Theoretic Analysis of Minority and Outlier Detection	13
<i>Shin Ando</i>	
Efficient Data Sampling in Heterogeneous Peer-to-Peer Networks	23
<i>Benjamin Arai, Song Lin, and Dimitrios Gunopoulos</i>	
Temporal Analysis of Semantic Graphs Using ASALSAN	33
<i>Brett W. Bader, Richard A. Harshman, and Tamara G. Kolda</i>	

Scalable Collaborative Filtering with Jointly Derived Neighborhood Interpolation Weights	43
<i>Robert M. Bell and Yehuda Koren</i>	
Rule Cubes for Causal Investigations	53
<i>Axel Blumenstock, Franz Schweiggert, and Markus Müller</i>	
The Chosen Few: On Identifying Valuable Patterns	63
<i>Björn Bringmann and Albrecht Zimmermann</i>	
Spectral Regression: A Unified Approach for Sparse Subspace Learning	73
<i>Deng Cai, Xiaofei He, and Jiawei Han</i>	
Mining Frequent Itemsets in a Stream	83
<i>Toon Calders, Nele Dexters, and Bart Goethals</i>	
A Cascaded Approach to Biomedical Named Entity Recognition Using a Unified Model	93
<i>Shing-Kit Chan, Wai Lam, and Xiaofeng Yu</i>	
Incorporating User Provided Constraints into Document Clustering	103
<i>Yanhua Chen, Manjeet Rege, Ming Dong, and Jing Hua</i>	
Depth-Based Novelty Detection and Its Application to Taxonomic Research	113
<i>Yixin Chen, Henry L. Bart Jr., Xin Dang, and Hanxiang Peng</i>	
Detecting Fractures in Classifier Performance	123
<i>David A. Cieslak and Nitesh V. Chawla</i>	
Non-redundant Multi-view Clustering via Orthogonalization	133
<i>Ying Cui, Xiaoli Z. Fern, and Jennifer G. Dy</i>	
On Appropriate Assumptions to Mine Data Streams: Analysis and Practice	143
<i>Jing Gao, Wei Fan, and Jiawei Han</i>	
ORIGAMI: Mining Representative Orthogonal Graph Patterns	153
<i>Mohammad Al Hasan, Vineet Chaoji, Saeed Salem, Jeremy Besson, and Mohammed J. Zaki</i>	
Efficient Algorithms for Mining Significant Substructures in Graphs with Quality Guarantees	163
<i>Huahai He and Ambuj K. Singh</i>	
Dynamic Micro Targeting: Fitness-Based Approach to Predicting Individual Preferences	173
<i>Tianyi Jiang and Alexander Tuzhilin</i>	
Data Discretization Unification	183
<i>Ruoming Jin, Yuri Breitbart, and Chibuike Muoh</i>	
Improving Knowledge Discovery in Document Collections through Combining Text Retrieval and Link Analysis Techniques	193
<i>Wei Jin, Rohini K. Srihari, Hung Hay Ho, and Xin Wu</i>	

Finding Cohesive Clusters for Analyzing Knowledge Communities	203
<i>Vasileios Kandylas, S. Phineas Upham, and Lyle H. Ungar</i>	
Succinct Matrix Approximation and Efficient k -NN Classification	213
<i>Rong Liu and Yong Shi</i>	
A Pairwise Covariance-Preserving Projection Method for Dimension Reduction	223
<i>Xiaoming Liu, Zhaohui Wang, Zhilin Feng, and Jinshan Tang</i>	
Community Learning by Graph Approximation.....	232
<i>Bo Long, Xiaoyun Xu, Zhongfei Zhang, and Philip S. Yu</i>	
Parallel Mining of Frequent Closed Patterns: Harnessing Modern Computer Architectures.....	242
<i>Claudio Lucchese, Salvatore Orlando, and Raffaele Perego</i>	
Supervised Learning by Training on Aggregate Outputs	252
<i>David R. Musicant, Janara M. Christensen, and Jamie F. Olson</i>	
Sample Selection for Maximal Diversity.....	262
<i>Feng Pan, Adam Roberts, Leonard McMillan, Fernando Pardo Manuel de Villena, David Threadgill, and Wei Wang</i>	
Mining Statistical Information of Frequent Fault-Tolerant Patterns in Transactional Databases	272
<i>Ardian Kristanto Poernomo and Vivekanand Gopalkrishnan</i>	
Lightweight Distributed Trust Propagation	282
<i>Daniele Quercia, Stephen Hailes, and Licia Capra</i>	
Social Network Extraction of Academic Researchers	292
<i>Jie Tang, Duo Zhang, and Limin Yao</i>	
General Averaged Divergence Analysis.....	302
<i>Dacheng Tao, Xuelong Li, Xindong Wu, and Stephen J. Maybank</i>	
Maximum Entropy Based Significance of Itemsets.....	312
<i>Nikolaj Tatti</i>	
Local Probabilistic Models for Link Prediction.....	322
<i>Chao Wang, Venu Satuluri, and Srinivasan Parthasarathy</i>	
Improving Text Classification by Using Encyclopedia Knowledge.....	332
<i>Pu Wang, Jian Hu, Hua-Jun Zeng, Lijun Chen, and Zheng Chen</i>	
Language-Independent Set Expansion of Named Entities Using the Web.....	342
<i>Richard C. Wang and William W. Cohen</i>	
Structure-Based Statistical Features and Multivariate Time Series Clustering	351
<i>Xiaozhe Wang, Anthony Wirth, and Liang Wang</i>	
A Generalization of Proximity Functions for K-Means.....	361
<i>Junjie Wu, Hui Xiong, Jian Chen, and Wenjun Zhou</i>	

Multilevel Belief Propagation for Fast Inference on Markov Random Fields.....	371
<i>Liang Xiong, Fei Wang, and Changshui Zhang</i>	
Disk Aware Discord Discovery: Finding Unusual Time Series in Terabyte Sized Datasets.....	381
<i>Dragomir Yankov, Eamonn Keogh, and Umaa Rebbapragada</i>	
Binary Matrix Factorization with Applications	391
<i>Zhongyuan Zhang, Tao Li, Chris Ding, and Xiangsun Zhang</i>	

SHORT PAPERS

A Semantic Kernel for Semi-structured Documents.....	403
<i>Sujeewan Aseervatham, Emmanuel Viennet, and Younès Bennani</i>	
DUSC: Dimensionality Unbiased Subspace Clustering	409
<i>Ira Assent, Ralph Krieger, Emmanuel Müller, and Thomas Seidl</i>	
Finding Predictive Runs with LAPS.....	415
<i>Suhrid Balakrishnan and David Madigan</i>	
Latent Dirichlet Conditional Naive-Bayes Models.....	421
<i>Arindam Banerjee and Hanhuai Shan</i>	
Efficient Kernel Discriminant Analysis via Spectral Regression	427
<i>Deng Cai, Xiaofei He, and Jiawei Han</i>	
Zonal Co-location Pattern Discovery with Dynamic Parameters	433
<i>Mete Celik, James M. Kang, and Shashi Shekhar</i>	
Predicting Blogging Behavior Using Temporal and Social Networks	439
<i>Bi Chen, Qiankun Zhao, Bingjun Sun, and Prasenjit Mitra</i>	
gApprox: Mining Frequent Approximate Patterns from a Massive Network.....	445
<i>Chen Chen, Xifeng Yan, Feida Zhu, and Jiawei Han</i>	
Document Transformation for Multi-label Feature Selection in Text Categorization.....	451
<i>Weizhu Chen, Jun Yan, Benyu Zhang, Zheng Chen, and Qiang Yang</i>	
Recommendation via Query Centered Random Walk on K-Partite Graph	457
<i>Haibin Cheng, Pang-Ning Tan, Jon Sticklen, and William F. Punch</i>	
Bandit-Based Algorithms for Budgeted Learning	463
<i>Kun Deng, Chris Bourke, Stephen Scott, Julie Sunderman, and Yaling Zheng</i>	
Extracting Product Comparisons from Discussion Boards.....	469
<i>Ronen Feldman, Moshe Fresco, Jacob Goldenberg, Oded Netzer, and Lyle Ungar</i>	
Mining Interpretable Human Strategies: A Case Study.....	475
<i>Xiaoli Z. Fern, Chaitanya Komireddy, and Margaret Burnett</i>	

Cross-Mining Binary and Numerical Attributes.....	481
<i>Gemma C. Garriga, Hannes Heikinheimo, and Jouni K. Seppänen</i>	
PRISM: A Primal-Encoding Approach for Frequent Sequence Mining.....	487
<i>Karam Gouda, Mosab Hassaan, and Mohammed J. Zaki</i>	
Using Burstiness to Improve Clustering of Topics in News Streams.....	493
<i>Qi He, Kuiyu Chang, and Ee-Peng Lim</i>	
Bayesian Folding-In with Dirichlet Kernels for PLSI.....	499
<i>Alexander Hinneburg, Hans-Henning Gabriel, and André Gohr</i>	
Confident Identification of Relevant Objects Based on Nonlinear Rescaling Method and Transductive Inference.....	505
<i>Shen-Shyang Ho and Roman Polyak</i>	
Training Conditional Random Fields by Periodic Step Size Adaptation for Large-Scale Text Mining.....	511
<i>Han-Shen Huang, Yu-Ming Chang, and Chun-Nan Hsu</i>	
Semi-supervised Document Clustering via Active Learning with Pairwise Constraints.....	517
<i>Ruizhang Huang and Wai Lam</i>	
Computing Correlation Anomaly Scores Using Stochastic Nearest Neighbors.....	523
<i>Tsuyoshi Idé, Spiros Papadimitriou, and Michail Vlachos</i>	
On Meta-Learning Rule Learning Heuristics.....	529
<i>Frederik Janssen and Johannes Fürnkranz</i>	
Web Site Recommendation Using HTTP Traffic.....	535
<i>Ming Jia, Shaozhi Ye, Xing Li, and Julie Dickerson</i>	
Trend Motif: A Graph Mining Approach for Analysis of Dynamic Complex Networks.....	541
<i>Ruoming Jin, Scott McCallen, and Eivind Almaas</i>	
Analyzing and Detecting Review Spam.....	547
<i>Nitin Jindal and Bing Liu</i>	
A Computational Approach to Style in American Poetry.....	553
<i>David M. Kaplan and David M. Blei</i>	
Change-Point Detection in Time-Series Data Based on Subspace Identification.....	559
<i>Yoshinobu Kawahara, Takehisa Yairi, and Kazuo Machida</i>	
Optimal Subsequence Bijection.....	565
<i>Longin Jan Latecki, Qiang Wang, Suzan Koknar-Tezel, and Vasileios Megalooikonomou</i>	
Connections between Mining Frequent Itemsets and Learning Generative Models.....	571
<i>Srivatsan Laxman, Prasad Naldurg, Raja Sripada, and Ramarathnam Venkatesan</i>	

Solving Consensus and Semi-supervised Clustering Problems Using Nonnegative Matrix Factorization	577
<i>Tao Li, Chris Ding, and Michael I. Jordan</i>	
Failure Prediction in IBM BlueGene/L Event Logs	583
<i>Yinglung Liang, Yanyong Zhang, Hui Xiong, and Ramendra Sahoo</i>	
A Text Classification Framework with a Local Feature Ranking for Learning Social Networks.....	589
<i>Masoud Makrehchi and Mohamed S. Kamel</i>	
Optimizing Frequency Queries for Data Mining Applications.....	595
<i>Hassan H. Malik and John R. Kender</i>	
Detecting Subdimensional Motifs: An Efficient Algorithm for Generalized Multivariate Pattern Discovery.....	601
<i>David Minnen, Charles Isbell, Irfan Essa, and Thad Starner</i>	
Consensus Clusterings.....	607
<i>Nam Nguyen and Rich Caruana</i>	
High-Speed Function Approximation.....	613
<i>Biswanath Panda, Mirek Riedewald, Johannes Gehrke, and Stephen B. Pope</i>	
Weighted Additive Criterion for Linear Dimension Reduction.....	619
<i>Jing Peng and Stefan Robila</i>	
Local Word Bag Model for Text Categorization.....	625
<i>Wen Pu, Ning Liu, Shuicheng Yan, Jun Yan, Kunqing Xie, and Zheng Chen</i>	
Sampling for Sequential Pattern Mining: From Static Databases to Data Streams	631
<i>Chedy Raïssi and Pascal Poncelet</i>	
Can the Content of Public News Be Used to Forecast Abnormal Stock Market Behaviour?	637
<i>Calum Robertson, Shlomo Geva, and Rodney C. Wolff</i>	
An Efficient Spectral Algorithm for Network Community Discovery and Its Applications to Biological and Social Networks	643
<i>Jianhua Ruan and Weixiong Zhang</i>	
Exploration of Link Structure and Community-Based Node Roles in Network Analysis.....	649
<i>Jerry Scripps, Pang-Ning Tan, and Abdol-Hossein Esfahanian</i>	
A Support Vector Approach to Censored Targets	655
<i>Pannagadatta K. Shivaswamy, Wei Chu, and Martin Jansche</i>	
Understanding Discrete Classifiers with a Case Study in Gene Prediction	661
<i>Muhammad Subianto and Arno Siebes</i>	
Statistical Learning Algorithm for Tree Similarity.....	667
<i>Atsuhiko Takasu, Daiji Fukagawa, and Tatsuya Akutsu</i>	

A Novel Criterion for Onset Detection: Differential Information Redundancy with Application to Human Movement Initiation	673
<i>Gert Van Dijck, Marc M. Van Hulle, and Jo Van Vaerenbergh</i>	
Using Significant, Positively Associated and Relatively Class Correlated Rules for Associative Classification of Imbalanced Datasets	679
<i>Florian Verhein and Sanjay Chawla</i>	
Preserving Privacy through Data Generation	685
<i>Jilles Vreeken, Matthijs van Leeuwen, and Arno Siebes</i>	
Transitional Patterns and Their Significant Milestones	691
<i>Qian Wan and Aijun An</i>	
Topical N-Grams: Phrase and Topic Discovery, with an Application to Information Retrieval	697
<i>Xuerui Wang, Andrew McCallum, and Xing Wei</i>	
Mechanism Design for Clustering Aggregation by Selfish Systems	703
<i>Pinata Winoto, Yiu-ming Cheung, and Jiming Liu</i>	
estMax: Tracing Maximal Frequent Itemsets over Online Data Streams	709
<i>Ho Jin Woo and Won Suk Lee</i>	
Locally Constrained Support Vector Clustering.....	715
<i>Dragomir Yankov, Eamonn Keogh, and Kin Fai Kan</i>	
Cocktail Ensemble for Regression.....	721
<i>Yang Yu, Zhi-Hua Zhou, and Kai Ming Ting</i>	
Incremental Subspace Clustering over Multiple Data Streams.....	727
<i>Qi Zhang, Jinze Liu, and Wei Wang</i>	
Noise Modeling with Associative Corruption Rules	733
<i>Yan Zhang and Xindong Wu</i>	
Co-ranking Authors and Documents in a Heterogeneous Network.....	739
<i>Ding Zhou, Sergey A. Orshanskiy, Hongyuan Zha, and C. Lee Giles</i>	
Discovering Temporal Communities from Social Network Documents	745
<i>Ding Zhou, Isaac Councill, Hongyuan Zha, and C. Lee Giles</i>	
Efficient Discovery of Frequent Approximate Sequential Patterns	751
<i>Feida Zhu, Xifeng Yan, Jiawei Han, and Philip S. Yu</i>	
Active Learning from Data Streams	757
<i>Xingquan Zhu, Peng Zhang, Xiaodong Lin, and Yong Shi</i>	
Lazy Bagging for Classifying Imbalanced Data.....	763
<i>Xingquan Zhu</i>	

Author Index