

2009 IEEE International Conference on Data Mining Workshops

(ICDMW 2009)

**Miami, Florida, USA
6 December 2009**



**IEEE Catalog Number: CFP0956B-PRT
ISBN: 978-1-4244-5384-9**

2009 IEEE International Conference on Data Mining Workshops

ICDMW 2009

Table of Contents

Message from General Co-chairs
Message from Workshop Co-chairs
Workshop Organizing Committees
Workshop Program Committees

Domain Driven Data Mining

SISC: A Text Classification Approach Using Semi Supervised Subspace Clustering	1
<i>Mohammad Salim Ahmed and Latifur Khan</i>	
High Quality True-Positive Prediction for Fiscal Fraud Detection	7
<i>Stefano Basta, Fabio Fassetti, Massimo Guarascio, Giuseppe Manco, Fosca Giannotti, Dino Pedreschi, Laura Spinsanti, Gianfilippo Papi, and Stefano Pisani</i>	
Building Classifiers with Independency Constraints	13
<i>Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy</i>	
Discovering Domain Specific Concepts within User-Generated Taxonomies	19
<i>Jonathan Klinginsmith, Malika Mahoui, Yuqing Wu, and Josette Jones</i>	
A Game Theoretical Model for Adversarial Learning	25
<i>Wei Liu and Sanjay Chawla</i>	
Discovery of Quantitative Sequential Patterns from Event Sequences	31
<i>Fumiya Nakagaito, Tomonobu Ozaki, and Takenao Ohkawa</i>	
Information Extraction for Clinical Data Mining: A Mammography Case Study	37
<i>Houssam Nassif, Ryan Woods, Elizabeth Burnside, Mehmet Ayvaci, Jude Shavlik, and David Page</i>	

A New Minimally Supervised Learning Method for Semantic Term Classification - Experimental Results on Classifying Ratable Aspects Discussed in Customer Reviews	43
<i>Thao Pham Thanh Nguyen, Takahiro Hayashi, Rikio Onai, Yuhei Nishioka, Takamasa Takenaka, and Masaya Mori</i>	
A Paradigm Shift: Combined Literature and Ontology-Driven Data Mining for Discovering Novel Relations in Biomedical Domain	51
<i>Yakub Sebastian, Brian Chung Shiong Loh, and Patrick Hang Hui Then</i>	
Weighted Frequent Subgraph Mining in Weighted Graph Databases	58
<i>Masaki Shinoda, Tomonobu Ozaki, and Takenao Ohkawa</i>	
Detecting and Interpreting Variable Interactions in Observational Ornithology Data	64
<i>Daria Sorokina, Rich Caruana, Mirek Riedewald, Wesley Hochachka, and Steve Kelling</i>	
Tree-Based Approach to Missing Data Imputation	70
<i>Peerapon Vateekul and Kanoksri Sarinnapakorn</i>	
Theoretical Study of the Relationship between Diversity and Single-Class Measures for Class Imbalance Learning	76
<i>Shuo Wang and Xin Yao</i>	
Multi-sphere Support Vector Data Description for Outliers Detection on Multi-distribution Data	82
<i>Yanshan Xiao, Bo Liu, Longbing Cao, Xindong Wu, Chengqi Zhang, Zhifeng Hao, Fengzhao Yang, and Jie Cao</i>	
Integrating Knowledge in Search of Biologically Relevant Genes	88
<i>Zheng Zhao, Shashvata Sharma, Nitin Agarwal, Huan Liu, Jiangxin Wang, and Yung Chang</i>	
Towards Context Aware Food Sales Prediction	94
<i>Indre Žliobaite, Jorn Bakker, and Mykola Pechenizkiy</i>	
Privacy Aspects of Data Mining	
Privacy Preserving Classification with Emerging Patterns	100
<i>Piotr Andruszkiewicz</i>	
Efficient Anonymizations with Enhanced Utility	106
<i>Jacob Goldberger and Tamir Tassa</i>	
A Practical Differentially Private Random Decision Tree Classifier	114
<i>Geetha Jagannathan, Krishnan Pillaipakkamnatt, and Rebecca N. Wright</i>	
A Differentially Private Graph Estimator	122
<i>Darakhshan J. Mir and Rebecca N. Wright</i>	
An Attack on the Privacy of Sanitized Data that Fuses the Outputs of Multiple Data Miners	130
<i>Michal Sramka, Reihaneh Safavi-Naini, and Jörg Denzinger</i>	

Differential Privacy for Clinical Trial Data: Preliminary Evaluations	138
<i>Duy Vu and Aleksandra Slavković</i>	

Internet Multimedia Mining

Video2Text: Learning to Annotate Video Content	144
<i>Hrishikesh Aradhye, George Toderici, and Jay Yagnik</i>	
Automatic Image Annotations by Mining Web Image Data	152
<i>Guiguang Ding, Jianmin Wang, Na Xu, and Lu Zhang</i>	
A Study of Language Model for Image Retrieval	158
<i>Bo Geng, Linjun Yang, and Chao Xu</i>	
MSRA-MM 2.0: A Large-Scale Web Multimedia Dataset	164
<i>Hao Li, Meng Wang, and Xian-Sheng Hua</i>	
Localized Content Based Image Retrieval with Self-Taught Multiple Instance Learning	170
<i>Qifeng Qiao and Peter A. Beiling</i>	
Mining Event Definitions from Queries for Video Retrieval on the Internet	176
<i>Kimiaki Shirahama, Chieri Sugihara, Kana Matsumura, Yuta Matsuoka, and Kuniaki Uehara</i>	
Cross-Domain Web Image Annotation	184
<i>Si Si, Dacheng Tao, and Kwok-Ping Chan</i>	
TubeTagger - YouTube-based Concept Detection	190
<i>Adrian Ulges, Markus Koch, Damian Borth, and Thomas M. Breuel</i>	
Semantic Linking between Video Ads and Web Services with Progressive Search	196
<i>Bo Wang, Jinqiao Wang, Shi Chen, Ling-Yu Duan, and Hanqing Lu</i>	
Mining Personal Image Collection for Social Group Suggestion	202
<i>Jie Yu, Xin Jin, Jiawei Han, and Jiebo Luo</i>	

Knowledge Discovery from Climate Data: Prediction, Extremes and Impacts

Induction of Mean Output Prediction Trees from Continuous Temporal Meteorological Data	208
<i>Dima Alberg, Mark Last, Roni Neuman, and Avi Sharon</i>	
Data Mining Geophysical Content from Satellites and Global Climate Models	214
<i>David Erickson, Jamison Daniel, Melissa Allen, Auroop Ganguly, Forrest Hoffman, Steven Pawson, Lesley Ott, and Eric Neilson</i>	
Understanding Climate Change Patterns with Multivariate Geovisualization	217
<i>Hai Jin and Diansheng Guo</i>	
Motivating Complex Dependence Structures in Data Mining: A Case Study with Anomaly Detection in Climate	223
<i>Shih-Chieh Kao, Auroop R. Ganguly, and Karsten Steinhaeuser</i>	

A Distributed Computing Infrastructure for the Evaluation of Climate Models Using NASA Observational Data	231
<i>Chris A. Mattmann, Daniel J. Crichton, Amy Braverman, Dean Williams, Michael Gunson, David Woollard, Sean Kelly, and Michael Cayanan</i>	
Ensemble-Vis: A Framework for the Statistical Visualization of Ensemble Data	233
<i>Kristin Potter, Andrew Wilson, Peer-Timo Bremer, Dean Williams, Charles Doutriaux, Valerio Pascucci, and Chris R. Johnson</i>	
Uncertainty Quantification in the Presence of Limited Climate Model Data with Discontinuities	241
<i>Khachik Sargsyan, Cosmin Safta, Bert Deusschere, and Habib Najm</i>	
Change Detection in Climate Data over the Iberian Peninsula	248
<i>Raquel Sebastião, Pedro Pereira Rodrigues, and João Gama</i>	
The Flexible Climate Data Analysis Tools (CDAT) for Multi-model Climate Simulation Data	254
<i>Dean N. Williams, Charles M. Doutriaux, Robert S. Drach, and Renata B. McCoy</i>	
Large-Scale Data Mining: Theory and Applications - Long Presentations	
Link Prediction on Evolving Data Using Matrix and Tensor Factorizations	262
<i>Evrin Acar, Daniel M. Dunlavy, and Tamara G. Kolda</i>	
SIMPLE: A Strategic Information Mining Platform for Licensing and Execution	270
<i>Ying Chen, Scott Spangler, Jeffrey Kreulen, Stephen Boyer, Thomas D. Griffin, Alfredo Alba, Amit Behal, Bin He, Linda Kato, Ana Lelescu, Cheryl Kieliszewski, Xian Wu, and Li Zhang</i>	
HOCT: A Highly Scalable Algorithm for Training Linear CRF on Modern Hardware	276
<i>Tianyuan Chen, Lei Chang, Jianqing Ma, Wei Zhang, and Feng Gao</i>	
Why Naive Ensembles Do Not Work in Cloud Computing	282
<i>Wenxuan Gao, Robert Grossman, Philip S. Yu, and Yunhong Gu</i>	
EigenSpokes: Surprising Patterns and Scalable Community Chipping in Large Graphs	290
<i>B. Aditya Prakash, Mukund Seshadri, Ashwin Sridharan, Sridhar Machiraju, and Christos Faloutsos</i>	
Toolkit-Based High-Performance Data Mining of Large Data on MapReduce Clusters	296
<i>Dennis Wegener, Michael Mock, Deyaa Adranale, and Stefan Wrobel</i>	
Scalable Attribute-Value Extraction from Semi-structured Text	302
<i>Yuk Wah Wong, Dominic Widdows, Tom Lokovic, and Kamal Nigam</i>	

Large-Scale Data Mining: Theory and Applications - Short Presentations

Estimating the Parameters of Randomly Interleaved Markov Models	308
<i>Daniel Gillblad, Rebecca Steinert, and Diogo R. Ferreira</i>	
Sparse Least-Squares Methods in the Parallel Machine Learning (PML) Framework	314
<i>Ramesh Natarajan, Vikas Sindhwani, and Shirish Tatikonda</i>	
Fast Induction of Multiple Decision Trees in Text Categorization from Large Scale, Imbalanced, and Multi-label Data	320
<i>Peerapon Vateekul and Miroslav Kubat</i>	
Greedy is not Enough: An Efficient Batch Mode Active Learning Algorithm	326
<i>Zuobing Xu, Christopher Hogan, and Robert Bauer</i>	
Efficient Dense Structure Mining Using MapReduce	332
<i>Shengqi Yang, Bai Wang, Haizhou Zhao, and Bin Wu</i>	
Efficient Incremental Mining of Qualified Web Traversal Patterns without Scanning Original Databases	338
<i>Jia-Ching Ying, Vincent S. Tseng, and Philip S. Yu</i>	
Compressed Spectral Clustering	344
<i>Bin Zhao and Changshui Zhang</i>	

Optimization-Based Methods for Emerging Data Mining Problems

Mining of Attribute Interactions Using Information Theoretic Metrics	350
<i>Pritam Chanda, Young-Rae Cho, Aidong Zhang, and Murali Ramanathan</i>	
Kernel K-means Based Framework for Aggregate Outputs Classification	356
<i>Shuo Chen, Bin Liu, Mingjie Qian, and Changshui Zhang</i>	
GLSVM: Integrating Structured Feature Selection and Large Margin Classification	362
<i>Hongliang Fei, Brian Quanz, and Jun Huan</i>	
An Effective Network Partitioning Algorithm Based on Two-Point Diffusing Strategy	368
<i>Chengying Mao</i>	
Nonsmooth Bilevel Programming for Hyperparameter Selection	374
<i>Gregory M. Moore, Charles Bergeron, and Kristin P. Bennett</i>	
A New Measure of Feature Selection Algorithms' Stability	382
<i>Jana Novovicová, Petr Somol, and Pavel Pudil</i>	
Improved Multi Label Classification in Hierarchical Taxonomies	388
<i>Kunal Punera and Suju Rajan</i>	
Probabilistic Labeled Semi-supervised SVM	394
<i>Mingjie Qian, Feiping Nie, and Changshui Zhang</i>	
Feature Selection for Maximizing the Area Under the ROC Curve	400
<i>Rui Wang and Ke Tang</i>	

Multiple Instance Transfer Learning	406
<i>Dan Zhang and Luo Si</i>	

Transfer Mining

Detecting Similarity of Transferring Datasets Based on Features of Classification Rules	412
<i>Hidenao Abe and Shusaku Tsumoto</i>	
Transferred Feature Selection	416
<i>Wei Bi, Yuan Shi, and Zhenzhong Lan</i>	
Set-Based Boosting for Instance-Level Transfer	422
<i>Eric Eaton and Marie desJardins</i>	
Knowledge Transfer among Heterogeneous Information Networks	429
<i>Evan Wei Xiang, Nathan N. Liu, Sinno Jialin Pan, and Qiang Yang</i>	
Towards a Universal Text Classifier: Transfer Learning Using Encyclopedic Knowledge	435
<i>Pu Wang and Carlotta Domeniconi</i>	
Knowledge Transformation by Cross-Domain Belief Propagation	441
<i>Fei Wang and Tao Li</i>	
Determining the Training Window for Small Sample Size Classification with Concept Drift	447
<i>Indre Žliobaite and Ludmila I. Kuncheva</i>	

Semantic Aspects in Data Mining

Pattern Mining over Star Schemas in the Onto4AR Framework	453
<i>Cláudia Antunes</i>	
Document Clustering Using Semantic Kernels Based on Term-Term Correlations	459
<i>Ahmed K. Farahat and Mohamed S. Kamel</i>	
Semantic-Rich Markov Models for Web Prefetching	465
<i>Nizar R. Mabroukeh and Christie I. Ezeife</i>	
Parameterized Contrast in Second Order Soft Co-occurrences: A Novel Text Representation Technique in Text Mining and Knowledge Extraction	471
<i>Amir Hossein Razavi, Stan Matwin, Diana Inkpen, and Alexandre Kouznetsov</i>	
A WordNet-Based Semantic Model for Enhancing Text Clustering	477
<i>Shady Shehata</i>	
A Data Mining Method to Extract and Rank Papers Describing Coexpression Predicates Semantically	483
<i>Chengcui Zhang, Richa Tiwari, and Wei-Bang Chen</i>	

Mining Multiple Information Sources

LNBC: A Link-Based Naive Bayes Classifier	489
<i>Bahareh Bina, Oliver Schulte, and Hassan Khosravi</i>	
TagLearner: A P2P Classifier Learning System from Collaboratively Tagged Text Documents	495
<i>Haimonti Dutta, Xianshu Zhu, Tushar Mahule, Hillol Kargupta, Kirk Borne, Codrina Lauth, Florian Holz, and Gerhard Heyer</i>	
Mining Multiple Satellite Sensor Data Using Collaborative Clustering	501
<i>Germain Forestier, Cédric Wemmert, Pierre Gancarski, and Jordi Inglada</i>	
Feature Selection with High-Dimensional Imbalanced Data	507
<i>Jason Van Hulse, Taghi M. Khoshgoftaar, Amri Napolitano, and Randall Wald</i>	
Theoretically Optimal Distributed Anomaly Detection	515
<i>Aleksandar Lazarevic, Nisheeth Srivastava, Ashutosh Tiwari, Josh Isom, Nikunj Oza, and Jaideep Srivastava</i>	
Hybrid Clustering by Integrating Text and Citation Based Graphs in Journal Database Analysis	521
<i>Xinhai Liu, Shi Yu, Yves Moreau, Frizo Janssens, Bart De Moor, and Wolfgang Glänzel</i>	
Frequent Pattern Discovery from a Single Graph with Quantitative Itemsets	527
<i>Yuuki Miyoshi, Tomonobu Ozaki, and Takenao Ohkawa</i>	
Efficient Discovery of Closed Hyperclique Patterns in Multidimensional Structured Databases	533
<i>Tomonobu Ozaki and Takenao Ohkawa</i>	
Bucket Learning: Improving Model Quality through Enhancing Local Patterns	539
<i>Guangzhi Qu and Hui Wu</i>	
Improving Similarity Join Algorithms Using Fuzzy Clustering Technique	545
<i>Lisa Tan, Farshad Fotouhi, William Grosky, Horia F. Pop, and Nouredine Mouaddib</i>	
Mining Data from Multiple Software Development Projects	551
<i>H. Wang, T.M. Khoshgoftaar, K. Gao, and N. Seliya</i>	
Mining for Core Patterns in Stock Market Data	558
<i>Jianfei Wu, Anne Denton, Omar Elariss, and Dianxiang Xu</i>	

Spatial and Spatiotemporal Data Mining - Regular Papers

Spatio-temporal Multi-dimensional Relational Framework Trees	564
<i>Matthew Bodenhamer, Samuel Bleckley, Daniel Fennelly, Andrew H. Fagg, and Amy McGovern</i>	
Spatiotemporal Modeling and Monitoring of Atmospheric Hazardous Emissions Using Sensor Networks	571
<i>Guido Cervone, Anthony Stefanidis, Pasquale Franzese, and Peggy Agouris</i>	
Multi-granularity Visualization of Trajectory Clusters Using Sub-trajectory Clustering	577
<i>Cheng Chang and Baoyao Zhou</i>	

Deriving Low-Level Steering Behaviors from Trajectory Data	583
<i>Arie Croitoru</i>	
Greedy Optimization for Contiguity-Constrained Hierarchical Clustering	591
<i>Diansheng Guo</i>	
Spatially Adaptive Classification and Active Learning of Multispectral Data with Gaussian Processes	597
<i>Goo Jun, Ranga Raju Vatsavai, and Joydeep Ghosh</i>	
FARM : Feature-Assisted Aggregate Route Mining in Trajectory Data	604
<i>Shrikant Kashyap, Sujoy Roy, Mong Li lee, and Wynne Hsu</i>	
K-BestMatch Reconstruction and Comparison of Trajectory Data	610
<i>Mirco Nanni and Roberto Trasarti</i>	
Analyzing Abnormal Events from Spatio-temporal Trajectories	616
<i>Dhaval Patel, Chidansh Bhatt, Wynne Hsu, Mong Li Lee, and Mohan Kankanhalli</i>	
Multilayer Scene Similarity Assessment	622
<i>Anthony Stefanidis, Caixia Wang, Lu Xu, and Kevin M. Curtin</i>	
Spatiotemporal Relational Random Forests	630
<i>Timothy A. Supinie, Amy McGovern, John Williams, and Jennifer Abernathy</i>	
Accurate Discovery of Valid Convoys from Moving Object Trajectories	636
<i>Hyunjin Yoon and Cyrus Shahabi</i>	

Spatial and Spatiotemporal Data Mining - Short Papers

A Semi-supervised Framework for Simultaneous Classification and Regression of Zero-Inflated Time Series Data with Application to Precipitation Prediction	644
<i>Zubin Abraham and Pang-Ning Tan</i>	
Visualization and Classification of Power System Frequency Data Streams	650
<i>Jason N. Bank, Olufemi A. Omitaomu, Steven J. Fernandez, and Yilu Liu</i>	
Information Services and Middleware for the Coastal Sensor Web	656
<i>Surya S. Durbha, Roger L. King, Santhosh K. Amanchi, Shruthi Bheemireddy, and Nicolas H. Younan</i>	
Enhancing the DBSCAN and Agglomerative Clustering Algorithms to Solve Network Planning Problem	662
<i>Lamia Fattouh Ibrahim, Weam M. Minshawi, Isra Yosef Ekkab, Nehal Mahmoud Al-Jurf, Afnan Salem Babrahim, and Samar Faisl Al-Halees</i>	
Fast Visual Trajectory Analysis Using Spatial Bayesian Networks	668
<i>Thomas Liebig, Christine Körner, and Michael May</i>	

Author Index