

# Lecture Notes in Artificial Intelligence 4819

Edited by J. G. Carbonell and J. Siekmann

Subseries of Lecture Notes in Computer Science

Takashi Washio Zhi-Hua Zhou  
Joshua Zhexue Huang Xiaohua Hu  
Jinyan Li Chao Xie Jieyue He  
Deqing Zou Kuan-Ching Li  
Mário M. Freire (Eds.)

# Emerging Technologies in Knowledge Discovery and Data Mining

PAKDD 2007 International Workshops  
Nanjing, China, May 22, 2007  
Revised Selected Papers



Springer

## Series Editors

Jaime G. Carbonell, Carnegie Mellon University, Pittsburgh, PA, USA  
Jörg Siekmann, University of Saarland, Saarbrücken, Germany

## Volume Editors

Takashi Washio, Osaka University, Japan (washio@ar.sanken.osaka-u.ac.jp)

Zhi-Hua Zhou, Nanjing University, China (zhouzh@nju.edu.cn)

Joshua Zhexue Huang, The University of Hong Kong, China (jhuang@eti.hku.hk)

Xiaohua Hu, Drexel University, USA (thu@cis.drexel.edu)

Jinyan Li, Nanyang Technological University, Singapore (jyli@ntu.edu.sg)

Chao Xie, Georgia State University, USA (cxie@cs.gsu.edu)

Jieyue He, Southeast University, China (jieyuehe@seu.edu.cn)

Deqing Zou, Huazhong University of Science and Technology, China  
(deqingzou@hust.edu.cn)

Kuan-Ching Li, Providence University, Taiwan (kuaneli@pu.edu.tw)

Mário M. Freire, University of Beira Interior, Portugal (mario@di.ubi.pt)

Library of Congress Control Number: 2007941074

CR Subject Classification (1998): I.2, H.2.7-8, H.3, H.5.1, G.3, J.3

LNCS Sublibrary: SL 7 – Artificial Intelligence

ISSN	0302-9743
ISBN-10	3-540-77016-X Springer Berlin Heidelberg New York
ISBN-13	978-3-540-77016-9 Springer Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

Springer is a part of Springer Science+Business Media  
springer.com

© Springer-Verlag Berlin Heidelberg 2007  
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India  
Printed on acid-free paper SPIN: 12197576 06/3180 5 4 3 2 1 0

# Preface

The techniques of knowledge discovery and data mining (KDD) have rapidly developed along the significant progress of the computer and its network technologies in the last two decades. The attention and the number of researchers in this domain continue to grow in both international academia and industry. The Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD) is a worldwide representative international conference on the research areas of KDD. Under the current spread of KDD techniques in our society, PAKDD 2007 invited the organizers to an industrial track on KDD techniques and, moreover, called for enterprising proposals of workshops on novel and emerging topics in KDD studies. The PAKDD industrial track and the PAKDD workshops both attracted more than 350 paper submissions from 22 countries in total, of which only 81 high-quality papers were accepted after strict reviews. The conference provided many informal and vibrant opportunities for researchers and industry practitioners to share their research positions, original research results and practical development experiences on specific new challenges, emerging issues, new technology trends and solutions to real-world problems.

The objective of this volume is to offer the excellent presentations to the public, and to promote the study exchange among researchers worldwide. Sixty-two outstanding papers in the industrial track and the workshops were further selected, among the 81 paper presentations, under even more rigorous reviews by the organizers, and these papers were carefully brushed up and included in this post-proceedings volume.

The first part of this volume contains ten outstanding papers presented in the industrial track. This track was organized to attract papers on new technology trends and real-world solutions in different industry sectors. The succeeding chapters include papers selected from the three workshops of BioDM 2007, HPDMA 2007 and SSDU 2007. The 2nd BioDM Workshop on Data Mining for Biomedical Applications (BioDM 2007) aimed at attracting top researchers, practitioners and students from around the world to discuss data mining applications in the field of bioinformatics. There are many computationally challenging problems in the analysis of the diverse and voluminous data provided in the field, and the readers will find 11 challenging papers on these problems. The 2007 International Workshop on High-Performance Data Mining and Applications (HPMDA 2007) addressed high-performance data mining methods and applications from both algorithmic and system perspectives. It includes 25 papers on how to achieve efficient mining of useful information from the data available, as well as on the topics of parallel hardware platforms, clusters and large-scale distributed computing infrastructures. The 2007 International Workshop on Service, Security and Its Data Management for Ubiquitous Computing (SSDU 2007) was held to foster research in the areas of security and intelligence integration

into ubiquitous computing and data management technology. Sixteen papers discussing the many security risks and problems in the ubiquitous computing environment are included in the volume.

We hope this book contributes to the growth of the worldwide community of KDD researchers.

September 2007

Takashi Washio  
Zhi-Hua Zhou

# Organization and Editorial Board

The paper selection of the industrial track and the workshops was made by the Program Committee of each respective organization. After the paper selection, the book was edited and managed by the volume editors.

## Volume Editors

Takashi Washio (Osaka University, Japan)

Zhi-Hua Zhou (Nanjing University, China)

Joshua Z. Huang (The University of Hong Kong, Hong Kong)

Xiaohua Hu (Drexel University, USA)

Jinyan Li (Institute for Infocomm Research, Singapore)

Chao Xie (Georgia State University, USA)

Jieyue He (Southeast University, China)

Deqing Zou (Huazhong University of Science and Technology, China)

Kuan-Ching Li (Providence University, Taiwan)

Mario Freire (University of Beira Interior, Portugal)

# Table of Contents

## PAKDD Industrial Track and Workshops 2007

### Industrial Track

PAKDD 2007 Industrial Track Workshop .....	1
<i>Joshua Zhexue Huang and Yunming Ye</i>	
A Survey of Open Source Data Mining Systems .....	3
<i>Xiaojun Chen, Yunming Ye, Graham Williams, and Xiaofei Xu</i>	
Predicting the Short-Term Market Reaction to Asset Specific News: Is Time Against Us? .....	15
<i>Calum Robertson, Shlomo Geva, and Rodney Wolff</i>	
Frequency-Weighted Fuzzy Time-Series Based on Fibonacci Sequence for TAIEX Forecasting .....	27
<i>Hia Jong Teoh, Tai-Liang Chen, and Ching-Hsue Cheng</i>	
Probabilistic Techniques for Corporate Blog Mining .....	35
<i>Flora S. Tsai, Yun Chen, and Kap Luk Chan</i>	
Mining Chat Conversations for Sex Identification .....	45
<i>Cemal Köse, Özcan Özyurt, and Guychmyrat Amanmyradov</i>	
Mining High Impact Exceptional Behavior Patterns .....	56
<i>Longbing Cao, Yanchang Zhao, Fernando Figueiredo, Yuming Ou, and Dan Luo</i>	
Practical Issues on Privacy-Preserving Health Data Mining .....	64
<i>Huidong (Warren) Jin</i>	
Data Mining for Intelligent Structure Form Selection Based on Association Rules from a High Rise Case Base .....	76
<i>Shihai Zhang, Shujun Liu, and Jinping Ou</i>	
CommonKADS Methodology for Developing Power Grid Switching Orders Systems .....	87
<i>Ming Zhou, Jianwen Ren, Jianxun Qi, Dongxiao Niu, and Gengyin Li</i>	
Discovering Prediction Model for Environmental Distribution Maps ....	99
<i>Ke Zhang, Huidong Jin, Nianjun Liu, Rob Lesslie, Lei Wang, Zhouyu Fu, and Terry Caelli</i>	

# Workshop of BioDM 2007

Workshop BioDM'07—An Overview . . . . .	110
<i>Jinyan Li and Xiaohua Hu</i>	
Extracting Features from Gene Ontology for the Identification of Protein Subcellular Location by Semantic Similarity Measurement . . . . .	112
<i>Guoqi Li and Huanye Sheng</i>	
Detecting Community Structure in Complex Networks by Optimal Rearrangement Clustering . . . . .	119
<i>Rui-Sheng Wang, Yong Wang, Xiang-Sun Zhang, and Luonan Chen</i>	
The HIV Data Mining Tool for Government Decision-Making Support . . . . .	131
<i>Huijun Liu, Qunying Xiao, and Zhengwei Zhu</i>	
Negative Localized Relationship Among p70S6 with Smad1, 2, 3 and p38 in Three Treated Human Cancer Cell Lines . . . . .	142
<i>Lin Wang, Minghu Jiang, Stefan Wolf, and Yinghua Lu</i>	
Cancer Identification Based on DNA Microarray Data . . . . .	153
<i>Yihui Liu</i>	
Incorporating Dictionary Features into Conditional Random Fields for Gene/Protein Named Entity Recognition . . . . .	162
<i>Hongfei Lin, Yanpeng Li, and Zhihao Yang</i>	
Translation and Rotation Invariant Mining of Frequent Trajectories: Application to Protein Unfolding Pathways . . . . .	174
<i>Alexander Andreopoulos, Bill Andreopoulos, Aijun An, and Xiaogang Wang</i>	
Genetic-Annealing Algorithm for 3D Off-lattice Protein Folding Model . . . . .	186
<i>Xiaolong Zhang, Xiaoli Lin, Chengpeng Wan, and Tingting Li</i>	
Biclustering of Microarray Data Based on Singular Value Decomposition . . . . .	194
<i>Wen-Hui Yang, Dao-Qing Dai, and Hong Yan</i>	
On the Number of Partial Least Squares Components in Dimension Reduction for Tumor Classification . . . . .	206
<i>Xue-Qiang Zeng, Guo-Zheng Li, Geng-Feng Wu, and Hua-Xing Zou</i>	
Mining Biosignal Data: Coronary Artery Disease Diagnosis Using Linear and Nonlinear Features of HRV . . . . .	218
<i>Heon Gyu Lee, Ki Yong Noh, and Keun Ho Ryu</i>	



## Workshop of HPDMA 2007

High Performance Data Mining and Applications Overview . . . . .	229
<i>Chao Xie and Jieyue He</i>	
Approximately Mining Recently Representative Patterns on Data Streams . . . . .	231
<i>Jia-Ling Koh and Yuan-Bin Don</i>	
Finding Frequent Items in Data Streams Using ESBF . . . . .	244
<i>ShuYun Wang, XiuLan Hao, HeXiang Xu, and YunFa Hu</i>	
A New Decision Tree Classification Method for Mining High-Speed Data Streams Based on Threaded Binary Search Trees . . . . .	256
<i>Tao Wang, Zhoujun Li, Xiaohua Hu, Yuejin Yan, and Huowang Chen</i>	
Progressive Subspace Skyline Clusters Mining on High Dimensional Data . . . . .	268
<i>Rong Hu, Yansheng Lu, Lei Zou, and Chong Zhou</i>	
Efficient Privacy Preserving Distributed Clustering Based on Secret Sharing . . . . .	280
<i>Selim V. Kaya, Thomas B. Pedersen, Erkay Savaş, and Yücel Saygın</i>	
SePMa: An Algorithm That Mining Sequential Processes from Hybrid Log . . . . .	292
<i>Xiaoyu Huang, Huiling Zhong, and Wenxue Cai</i>	
Evaluate Structure Similarity in XML Documents with Merge-Edit-Distance . . . . .	301
<i>Chong Zhou, Yansheng Lu, Lei Zou, and Rong Hu</i>	
Ensemble Learning Based Distributed Clustering . . . . .	312
<i>Genlin Ji and Xiaohan Ling</i>	
Deploying Mobile Agents in Distributed Data Mining . . . . .	322
<i>Xining Li and JingBo Ni</i>	
ODDC: Outlier Detection Using Distance Distribution Clustering . . . . .	332
<i>Kun Niu, Chong Huang, Shubo Zhang, and Junliang Chen</i>	
Spatial Clustering with Obstacles Constraints Using Ant Colony and Particle Swarm Optimization . . . . .	344
<i>Xueping Zhang, Jiayao Wang, Zhongshan Fan, and Bin Li</i>	
A High Performance Hierarchical Cubing Algorithm and Efficient OLAP in High-Dimensional Data Warehouse . . . . .	357
<i>Kongfa Hu, Zhenzhi Gong, Qingli Da, and Ling Chen</i>	

Grid-Based Clustering Algorithm Based on Intersecting Partition and Density Estimation.....	368
<i>Bao-Zhi Qiu, Xiang-Li Li, and Jun-Yi Shen</i>	
Depth First Generation of Frequent Patterns Without Candidate Generation.....	378
<i>Qunxiong Zhu and Xiaoyong Lin</i>	
Efficient Time Series Data Classification and Compression in Distributed Monitoring.....	389
<i>Sheng Di, Hai Jin, Shengli Li, Jing Tie, and Ling Chen</i>	
Best-Match Method Used in Co-training Algorithm.....	401
<i>Hui Wang, Liping Ji, and Wanli Zuo</i>	
A General Method of Mining Chinese Web Documents Based on GA&SA and Position-Factors.....	410
<i>Xi Bai, Jigui Sun, Haiyan Che, and Jin Wang</i>	
Data Management Services in ChinaGrid for Data Mining Applications.....	421
<i>Song Wu, Wei Wang, Muzhou Xiong, and Hai Jin</i>	
Two-Phase Algorithms for a Novel Utility-Frequent Mining Model.....	433
<i>Jieh-Shan Yeh, Yu-Chiang Li, and Chin-Chen Chang</i>	
Top-Down and Bottom-Up Strategies for Incremental Maintenance of Frequent Patterns.....	445
<i>Qunxiong Zhu and Xiaoyong Lin</i>	
GC-Tree: A Fast Online Algorithm for Mining Frequent Closed Itemsets.....	457
<i>Junbo Chen and ShanPing Li</i>	
Integration of Distributed Biological Data Using Modified K-Means Algorithm.....	469
<i>Jongil Jeong, Byunggul Ryu, Dongil Shin, and Dongkyoo Shin</i>	
A Parallel Algorithm for Enumerating All the Maximal $k$ -Plexes.....	476
<i>Bin Wu and Xin Pei</i>	
A Multi-dependency Language Modeling Approach to Information Retrieval.....	484
<i>Keke Cai, Chun Chen, Jiajun Bu, Guang Qiu, and Peng Huang</i>	
Factoid Mining Based Content Trust Model for Information Retrieval...	492
<i>Wei Wang, Guosun Zeng, Mingjun Sun, Huanan Gu, and Quan Zhang</i>	

## Workshop of SSDU 2007

Service, Security and Its Data Management for Ubiquitous Computing – Overview .....	500
<i>Jong Hyuk Park and Deqing Zou</i>	
Study on Trust Inference and Emergence of Economical Small-World Phenomena in P2P Environment .....	502
<i>Yufeng Wang, Yoshiaki Hori, and Kouichi Sakurai</i>	
A Secure Time Synchronization Protocol for Sensor Network .....	515
<i>Hui Li, Yanfei Zheng, Mi Wen, and Kefei Chen</i>	
On Grid-Based Key Pre-distribution: Toward a Better Connectivity in Wireless Sensor Network .....	527
<i>Abedelaziz Mohaisen, YoungJae Maeng, and DaeHun Nyang</i>	
A Distributed and Cooperative Black Hole Node Detection and Elimination Mechanism for Ad Hoc Networks .....	538
<i>Chang Wu Yu, Tung-Kuang Wu, Rei Heng Cheng, and Shun Chao Chang</i>	
A Novel Adaptive and Safe Framework for Ubicomp .....	550
<i>Xuanhua Shi and Jean-Louis Pizat</i>	
Reducing Inter-piconet Delay for Large-Scale Bluetooth Scatternets ....	562
<i>Chang Wu Yu, Kun-Ming Yu, and Shu Ling Lin</i>	
Security Analysis and Enhancement of One-Way Hash Based Low-Cost Authentication Protocol (OHLCAP) .....	574
<i>JeaCheol Ha, SangJae Moon, Juan Manuel Gonzalez Nieto, and Colin Boyd</i>	
An Effective Design of an Active RFID Reader Using a Cache of Tag Memory Data .....	584
<i>Seok-Young Jang, Sang-Hwa Chung, Won-Ju Yoon, and Seong-Joon Lee</i>	
Privacy Protection Scheme of RFID Using Random Number .....	596
<i>Soo-Young Kang and Im-Yeong Lee</i>	
A Hierarchical Composition of LU Matrix-Based Key Distribution Scheme for Sensor Networks .....	608
<i>Mi Wen, Yanfei Zheng, Hui Li, and Kefei Chen</i>	
Security Framework for Home Network: Authentication, Authorization, and Security Policy .....	621
<i>Geon Woo Kim, Deok Gyu Lee, Jong Wook Han, Sang Choon Kim, and Sang Wook Kim</i>	

Bogus Data Filtering in Sensor Networks.....	629
<i>Yong Ho Kim, Jong Hwan Park, Dong Hoon Lee, and Jongin Lim</i>	
Streaming Media Securely over Multipath Multihop Wireless Network .....	636
<i>Binod Vaidya, SangDuck Lee, Eung-Kon Kim, and SeungJo Han</i>	
Flexible Selection of Wavelet Coefficients Based on the Estimation Error of Predefined Queries .....	644
<i>Jaehoon Kim and Seog Park</i>	
Secured Web Services Based on Extended Usage Control .....	656
<i>Woochul Shin and Sang Bong Yoo</i>	
A Digital Rights Management Architecture for Multimedia in P2P .....	664
<i>Cheng Yang, Jianbo Liu, Aina Sui, and Yongbin Wang</i>	
<b>Author Index</b> .....	673