Longbing Cao Yong Feng Jiang Zhong (Eds.)

Advanced Data Mining and Applications

6th International Conference, ADMA 2010 Chongqing, China, November 2010 Proceedings, Part I

Part I

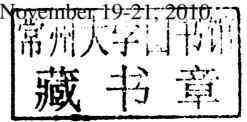


Longbing Cao Yong Feng Jiang Zhong (Eds.)

Advanced Data Mining and Applications

6th International Conference, ADMA 2010

Chongqing, China, November 19-21, Proceedings, Part I





Series Editors

Randy Goebel, University of Alberta, Edmonton, Canada Jörg Siekmann, University of Saarland, Saarbrücken, Germany Wolfgang Wahlster, DFKI and University of Saarland, Saarbrücken, Germany

Volume Editors

Longbing Cao
University of Technology Sydney
Faculty of Engineering and Information Technology
Sydney, NSW 2007, Australia
E-mail: longbing.cao-1@uts.edu.au

Yong Feng Chongqing University College of Computer Science Chongqing, 400030, China E-mail: fengphd@msn.com

Jiang Zhong Chongqing University College of Computer Science Chongqing, 400030, China E-mail: zhongjiang@cqu.edu.cn

Library of Congress Control Number: 2010939048

CR Subject Classification (1998): I.2, H.3, H.4, H.2.8, J.1, H.5

LNCS Sublibrary: SL 7 – Artificial Intelligence

ISSN 0302-9743

ISBN-10 3-642-17315-2 Springer Berlin Heidelberg New York

ISBN-13 978-3-642-17315-8 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.com

© Springer-Verlag Berlin Heidelberg 2010 Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India Printed on acid-free paper 06/3180

Preface

With the ever-growing power of generating, transmitting, and collecting huge amounts of data, information overload is now an imminent problem to mankind. The overwhelming demand for information processing is not just about a better understanding of data, but also a better usage of data in a timely fashion. Data mining, or knowledge discovery from databases, is proposed to gain insight into aspects of data and to help people make informed, sensible, and better decisions. At present, growing attention has been paid to the study, development, and application of data mining. As a result there is an urgent need for sophisticated techniques and tools that can handle new fields of data mining, e.g., spatial data mining, biomedical data mining, and mining on high-speed and time-variant data streams. The knowledge of data mining should also be expanded to new applications.

The 6th International Conference on Advanced Data Mining and Applications (ADMA 2010) aimed to bring together the experts on data mining throughout the world. It provided a leading international forum for the dissemination of original research results in advanced data mining techniques, applications, algorithms, software and systems, and different applied disciplines. The conference attracted 361 online submissions from 34 different countries and areas. All full papers were peer reviewed by at least three members of the Program Committee composed of international experts in data mining fields. A total number of 118 papers were accepted for the conference. Amongst them, 63 papers were selected as regular papers and 55 papers were selected as short papers. The Program Committee worked very hard to select these papers through a rigorous review process and extensive discussion, and finally composed a diverse and exciting program for ADMA 2010. The ADMA 2010 program was highlighted by three keynote speeches from outstanding researchers in advanced data mining and application areas: Kotagiri Ramamohanarao, Chengqi Zhang, and Vladimir Brusic.

September 2010

Longbing Cao Yong Feng Jiang Zhong

Organization

ADMA 2010 was organized by Chongqing University, China, and the School of Information Technology and Electrical Engineering, the University of Queensland, Australia, sponsored by the National Natural Science Foundation of China, Chongqing Science and Technology Commission, Chongqing Academy of Science and Technology, and technically co-sponsored by IEEE Queensland Section.

Organizing Committee

Steering Committee Chair

Xue Li

University of Queensland, Australia

Keynote Speakers

Kotagiri Ramamohanarao

Chengqi Zhang Vladimir Brusic University of Melbourne, Australia

University of Technology Sydney, Australia

Dana Farber Cancer Institute, USA

General Co-chairs

Charles Ling

Shangbo Zhou

Jie Xu

The University of Western Ontario, Canada

Chongqing University, China

Leeds University, UK

Program Co-chairs

Jinyan Li Longbing Cao

Zhongyang Xiong

Nanyang Technological University, Singapore University of Technology Sydney, Australia

Chongqing University, China

Publicity Chair

Xueming Li

Chongqing University, China

Regional Organization Co-chairs

Jiang Zhong

Yong Feng

Chongqing University, China

Chongqing University, China

Finance Chair

Yong Feng

Chongqing University, China

China Registration Chair

Li Wan Chongqing University, China

China Web Master

Quanji Qiu Chongqing University, China Xiaoran Lin Chongqing University, China

China Secretariat

Min Zheng Chongqing University, China Yanyan Zou Chongqing University, China Haoyang Ren Chongqing University, China Junhui Wang Chongqing University, China

Program Committee

Hua Li, Canada Arlindo Oliveira, Portugal Dragan Gamberger, Croatia Andre Ponce Leao, Brazil Andrew Kusiak, America Wang Shuliang, China Christophe Giraud-Carrier, USA Daniel Neagu, UK Liu Zhen, Japan Daniel Sanchez, Spain Dianhui Wang, Australia Fernando Berzal, Spain Gang Li, Australia Jan Rauch, Czech Republic Jean-Gabriel Ganascia, France Joseph Roure, UK Juho Rousu, USA Junbin Gao, Australia Paul Vitanyi, USA Petr Berka, Czech Republic Rui Camacho, Portugal Wanquan Liu, Australia Christophe Rigotti, India Xiaochun Cheng, UK Yonghong Peng, UK Zhaoli Zhu, China

Peng Han, China

Cai Yueping, Japan

Yu Qiao, China Guang Chen, China Xinyang Ying, China Guobin Zhou, China Yun Li. China Jun Zhao, China Hong Tang, China Hao Wang, China Hong Yu, China Li Li, China Ling Ou, China Zili Zhang, China Xingang Zhang, China Xiaofeng Liao, China Kaigui Wu, China Yufang Zhang, China Hua Li, China Xiaofan Yang, China Jiang Zhong, China Yong Feng, China Ji Li. China Li Wan, China Chengliang Wang, China Chunxiao Ye, China Huiping Cao, USA Yan Qi, USA Yanchang Zhao, Australia Sumon Shahriar, Australia Senthil Kumar, India Alfredo Cuzzocrea, Italy Rong Xie, China Ji Liu, China Changze Wu, China Yixiong Chen, China Qi Xie, China Pan He, China BaoHua Qiang, China Zhixing Li, China Fenghua Tu, China Xiaohong Zhang, China Chuan Li, China

Table of Contents – Part I

Ι	Data	Mining	Foundations
---	------	--------	--------------------

Cost Sensitive Classification in Data Mining	1
Web Users Access Paths Clustering Based on Possibilistic and Fuzzy Sets Theory	12
Discriminative Markov Logic Network Structure Learning Based on Propositionalization and χ^2 -Test	24
EWGen: Automatic Generation of Item Weights for Weighted Association Rule Mining	36
Best Clustering Configuration Metrics: Towards Multiagent Based Clustering	48
On Probabilistic Models for Uncertain Sequential Pattern Mining Muhammad Muzammal and Rajeev Raman	60
Cube Based Summaries of Large Association Rule Sets	73
A Perceptron-Like Linear Supervised Algorithm for Text Classification	86
Research on Time Series Forecasting Model Based on Moore Automata	98
A Clustering Algorithm FCM-ACO for Supplier Base Management Weining Liu and Lei Jiang	106
Nearest Neighbour Distance Matrix Classification	114
Classification Inductive Rule Learning with Negated Features Stephanie Chua, Frans Coenen, and Grant Malcolm	125

Fast Retrieval of Time Series Using a Multi-resolution Filter with Multiple Reduced Spaces Muhammad Marwan Muhammad Fuad and Pierre-François Marteau	1
DHPTID-HYBRID Algorithm: A Hybrid Algorithm for Association Rule Mining	1
An Improved Rough Clustering Using Discernibility Based Initial Seed Computation	1
Fixing the Threshold for Effective Detection of Near Duplicate Web Documents in Web Crawling	1
Topic-Constrained Hierarchical Clustering for Document Datasets Ying Zhao	1
Discretization of Time Series Dataset Using Relative Frequency and K-Nearest Neighbor Approach	1
MSDBSCAN: Multi-density Scale-Independent Clustering Algorithm Based on DBSCAN	2
An Efficient Algorithm for Mining Erasable Itemsets	2
Discord Region Based Analysis to Improve Data Utility of Privately Published Time Series	2
Deep Web Sources Classifier Based on DSOM-EACO Clustering Model	2
Kernel Based K-Medoids for Clustering Data with Uncertainty	2
Frequent Pattern Mining Using Modified CP-Tree for Knowledge Discovery	2
Spatial Neighborhood Clustering Based on Data Field	20

Table of Contents – Part 1	AΠ
Surrounding Influenced K-Nearest Neighbors: A New Distance Based Classifier	270
A Centroid k -Nearest Neighbor Method	278
Mining Spatial Association Rules with Multi-relational Approach Min Qian, Li-Jie Pu, Rong Fu, and Ming Zhu	286
An Unsupervised Classification Method of Remote Sensing Images Based on Ant Colony Optimization Algorithm	294
A Novel Clustering Algorithm Based on Gravity and Cluster Merging	302
II Data Mining in Specific Areas	
Evolution Analysis of a Mobile Social Network	310
Distance Distribution and Average Shortest Path Length Estimation in Real-World Networks	322
Self-adaptive Change Detection in Streaming Data with Non-stationary Distribution	334
Anchor Points Seeking of Large Urban Crowd Based on the Mobile Billing Data	346
Frequent Pattern Trend Analysis in Social Networks Puteri N.E. Nohuddin, Rob Christley, Frans Coenen, Yogesh Patel, Christian Setzkorn, and Shane Williams	358
Efficient Privacy-Preserving Data Mining in Malicious Model	370
Analyze the Wild Birds' Migration Tracks by MPI-Based Parallel Clustering Algorithm	383

Formal Concept Analysis Based Clustering for Blog Network Visualization	394
Finding Frequent Subgraphs in Longitudinal Social Network Data Using a Weighted Graph Mining Approach	405
Weigted-FP-Tree Based XML Query Pattern Mining	417
Privacy-Preserving Data Mining in Presence of Covert Adversaries Atsuko Miyaji and Mohammad Shahriar Rahman	429
Multiple Level Views on the Adherent Cohesive Subgraphs in Massive Temporal Call Graphs	441
Combating Link Spam by Noisy Link Analysis	453
High Dimensional Image Categorization François Poulet and Nguyen-Khang Pham	465
Efficiently Mining Co-Location Rules on Interval Data Lizhen Wang, Hongmei Chen, Lihong Zhao, and Lihua Zhou	477
Multiple Attribute Frequent Mining-Based for Dengue Outbreak Zalizah Awang Long, Azuraliza Abu Bakar, Abdul Razak Hamdan, and Mazrura Sahani	489
A Top-Down Approach for Hierarchical Cluster Exploration by Visualization	497
Distributed Frequent Items Detection on Uncertain Data	509
Mining Uncertain Sentences with Multiple Instance Learning Feng Ji, Xipeng Qiu, and Xuanjing Huang	521
WeightLOFCC: A Heuristic Weight-Setting Strategy of LOF Applied to Outlier Detection in Time Series Data	529
TGP: Mining Top-K Frequent Closed Graph Pattern without Minimum Support	537

625

Table of Contents – Part II

III Data Mining Methodologies and Processes	
Incremental Learning by Heterogeneous Bagging Ensemble	1
CPLDP: An Efficient Large Dataset Processing System Built on Cloud Platform	13
A General Multi-relational Classification Approach Using Feature Generation and Selection	21
A Unified Approach to the Extraction of Rules from Artificial Neural Networks and Support Vector Machines	34
A Clustering-Based Data Reduction for Very Large Spatio-Temporal Datasets	43
Change a Sequence into a Fuzzy Number	55
Multiple Kernel Learning Improved by MMD	63
A Refinement Approach to Handling Model Misfit in Semi-supervised Learning	75
Soft Set Approach for Selecting Decision Attribute in Data Clustering	87
Comparison of BEKK GARCH and DCC GARCH Models: An Empirical Study	99
Adapt the mRMR Criterion for Unsupervised Feature Selection $Junling \ Xu$	111

Evaluating the Distance between Two Uncertain Categorical Objects Hongmei Chen, Lizhen Wang, Weiyi Liu, and Qing Xiao	122
Construction Cosine Radial Basic Function Neural Networks Based on Artificial Immune Networks	134
Spatial Filter Selection with LASSO for EEG Classification	142
Boolean Algebra and Compression Technique for Association Rule Mining Somboon Anekritmongkol and M.L. Kulthon Kasamsan	150
Cluster Based Symbolic Representation and Feature Selection for Text Classification	158
SimRate: Improve Collaborative Recommendation Based on Rating Graph for Sparsity	167
Logistic Regression for Transductive Transfer Learning from Multiple Sources	175
Double Table Switch: An Efficient Partitioning Algorithm for Bottom-Up Computation of Data Cubes	183
IV Data Mining Applications and Systems	
Tag Recommendation Based on Bayesian Principle Zhonghui Wang and Zhihong Deng	191
Comparison of Different Methods to Fuse Theos Images	202
Using Genetic K-Means Algorithm for PCA Regression Data in Customer Churn Prediction	210
Time-Constrained Test Selection for Regression Testing	221
Chinese New Word Detection from Query Logs	233

Table of Contents – Part II	XIX
Exploiting Concept Clumping for Efficient Incremental E-Mail Categorization	244
Topic-Based User Segmentation for Online Advertising with Latent Dirichlet Allocation	259
Applying Multi-Objective Evolutionary Algorithms to QoS-Aware Web Service Composition	270
Real-Time Hand Detection and Tracking Using LBP Features	282
Modeling DNS Activities Based on Probabilistic Latent Semantic Analysis	290
A New Statistical Approach to DNS Traffic Anomaly Detection	302
Managing Power Conservation in Wireless Networks	314
Using PCA to Predict Customer Churn in Telecommunication Dataset	326
Hierarchical Classification with Dynamic-Threshold SVM Ensemble for Gene Function Prediction	336
Personalized Tag Recommendation Based on User Preference and Content	348
Predicting Defect Priority Based on Neural Networks	356
Personalized Context-Aware QoS Prediction for Web Services Based on Collaborative Filtering	368
Hybrid Semantic Analysis System – ATIS Data Evaluation	376

Click Prediction for Product Search on C2C Web Sites $\dots \dots$

Xiangzhi Wang, Chunyang Liu, Guirong Xue, and Yong Yu

387

XX

Finding Potential Research Collaborators in Four Degrees of	200
Separation	399
Predicting Product Duration for Adaptive Advertisement Zhongqi Guo, Yongqiang Wang, Gui-rong Xue, and Yong Yu	411
An Algorithm for Available Bandwidth Estimation of IPv6 Network Quanjie Qiu, Zhiguo Li, and Zhongfu Wu	419
A Structure-Based XML Storage Method in YAFFS File System Ji Liu, Shuyu Chen, and Haozhang Liu	427
A Multi-dimensional Trustworthy Behavior Monitoring Method Based on Discriminant Locality Preserving Projections	435
NN-SA Based Dynamic Failure Detector for Services Composition in Distributed Environment	443
Two-Fold Spatiotemporal Regression Modeling in Wireless Sensor Networks	451
Generating Tags for Service Reviews	463
Developing Treatment Plan Support in Outpatient Health Care Delivery with Decision Trees Technique	475
Factor Analysis of E-business in Skill-Based Strategic Collaboration Daijiang Chen and Juanjuan Chen	483
Increasing the Meaningful Use of Electronic Medical Records: A Localized Health Level 7 Clinical Document Architecture System Jun Liang, Mei Fang Xu, Lan Juan Li, Sheng Li Yang, Bao Luo Li, De Ren Cheng, Ou Jin, Li Zhong Zhang, Long Wei Yang, and Jun Xiang Sun	491
Corpus-Based Analysis of the Co-occurrence of Chinese Antonym Pairs	500
Application of Decision-Tree Based on Prediction Model for Project Management	508

Table of Contents – Part II	XXI
Management Policies Analysis for Multi-core Shared Caches	514
Multi-core Architecture Cache Performance Analysis and Optimization Based on Distributed Method	522
The Research on the User Experience of E-Commercial Website Based on User Subdivision	529
An Ontology-Based Framework Model for Trustworthy Software Evolution	537
Multi-level Log-Based Relevance Feedback Scheme for Image Retrieval Huanchen Zhang, Weifeng Sun, Shichao Dong, Long Chen, and Chuang Lin	545
A Distributed Node Clustering Mechanism in P2P Networks	553
Exploratory Factor Analysis Approach for Understanding Consumer Behavior toward Using Chongqing City Card	561
Author Index	569