Lecture Notes in Computer Science 5182

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

David Hutchison
Lancaster University, UK

Takeo Kanade
Carnegie Mellon University, Pittsburgh, PA, USA

Josef Kittler
University of Surrey, Guildford, UK

Jon M. Kleinberg
Cornell University, Ithaca, NY, USA

Alfred Kobsa
University of California, Irvine, CA, USA

Friedemann Mattern
ETH Zurich, Switzerland

John C. Mitchell
Stanford University, CA, USA

Moni Naor
Weizmann Institute of Science, Rehovot, Israel

Oscar Nierstrasz
University of Bern, Switzerland

C. Pandu Rangan
Indian Institute of Technology, Madras, India

Bernhard Steffen
University of Dortmund, Germany

Madhu Sudan
Massachusetts Institute of Technology, MA, USA

Demetri Terzopoulos
University of California, Los Angeles, CA, USA

Doug Tygar
University of California, Berkeley, CA, USA

Gerhard Weikum
Max-Planck Institute of Computer Science, Saarbruecken, Germany

Il-Yeol Song Johann Eder
Tho Manh Nguyen (Eds.)

Data Warehousing and Knowledge Discovery

10th International Conference, DaWaK 2008
Turin, Italy, September 2-5, 2008
Proceedings

Springer

Volume Editors

Il-Yeol Song
Drexel University
College of Information Science and Technology
Philadelphia, PA 19104, USA
E-mail: song@drexel.edu

Johann Eder
University of Klagenfurt
Information and Communication Systems
Universitätsstraße 65-67, 9020 Klagenfurt, Austria
E-mail: eder@isys.uni-klu.ac.at

Tho Manh Nguyen
Vienna University of Technology
Institute of Software Technology and Interactive Systems
Favoritenstraße 9-11/188, 1040 Wien, Austria
E-mail: tho@ifs.tuwien.ac.at

Library of Congress Control Number: 2008933459

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

LNCS Sublibrary: SL 3 – Information Systems and Application, incl. Internet/Web and HCI

ISSN 0302-9743
ISBN-10 3-540-85835-0 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-85835-5 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 2008
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12466535 06/3180 5 4 3 2 1 0

Preface

Data Warehousing and Knowledge Discovery have been widely accepted as key technologies for enterprises and organizations as a means of improving their abilities in data analysis, decision support, and the automatic extraction of knowledge from data. With the exponentially growing amount of information to be included in the decision making process, the data to be processed is becoming more and more complex in both structure and semantics. Consequently, the process of retrieval and knowledge discovery from this huge amount of heterogeneous complex data constitutes the reality check for research in the area.

During the past few years, the International Conference on Data Warehousing and Knowledge Discovery (DaWaK) has become one of the most important international scientific events to bring together researchers, developers and practitioners. The DaWaK conferences serve as a prominent forum for discussing the latest research issues and experiences in developing and deploying data warehousing and knowledge discovery systems, applications, and solutions. This year's conference, the 10th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2008), continued the tradition of facilitating the cross-disciplinary exchange of ideas, experience and potential research directions. DaWaK 2008 sought to disseminate innovative principles, methods, algorithms and solutions to challenging problems faced in the development of data warehousing, knowledge discovery and data mining applications.

The papers presented covered a number of broad research areas on both theoretical and practical aspects of data warehousing and knowledge discovery. In the areas of data warehousing, the topics covered included advanced techniques in OLAP and multidimensional designing and modeling, advanced OLAP and Cube Processing, innovation of ETL processes and data integration problems, materialized view optimization, distributed and parallel processing in data warehousing, data warehouses and data mining applications integration, multidimensional analysis of text documents, and data warehousing for real-world applications such as medical applications, spatial applications, and bioinformatics data warehouses. In the areas of data mining and knowledge discovery, the topics covered included text mining and taxonomy, web information discovery, stream data analysis and mining, ontology-based data and non-standard data mining techniques, machine learning, constraint-based mining, and traditional data mining topics such as mining frequent item sets, clustering, association, classification, ranking, and applications of data mining technologies to real-world problems. It is especially notable to see that some papers covered emerging real-world applications such as bioinformatics, geophysics, and terrorist networks, as well as integration of multiple technologies such as conceptual modeling of knowledge discovery process and results, integration of semantic web into data warehousing and OLAP technologies, OLAP mining, and imprecise or fuzzy OLAP. All these papers show that data warehousing and knowledge discovery technologies are maturing and making an impact on real-world applications.

From 143 submitted abstracts, we received 121 papers from 37 countries. The Program Committee finally selected 40 papers, making an acceptance rate of 33.05% of submitted papers.

We would like to express our gratitude to all the Program Committee members and the external reviewers, who reviewed the papers very profoundly and in a timely manner. Due to the high number of submissions and the high quality of the submitted papers, the reviewing, voting and discussion process was an extraordinarily challenging task. We would also like to thank all the authors who submitted their papers to DaWaK 2008, as their contributions formed the basis of this year's excellent technical program.

Many thanks go to Gabriela Wagner for providing a great deal of assistance in administering the DaWaK management issues as well as to Mr. Amin Andjomshoaa for the conference management software.

September 2008

Il Yeol Song
Johann Eder
Tho Manh Nguyen

Organization

Conference Program Chairpersons

Il-Yeol Song	Drexel University, USA
Johann Eder	University of Klagenfurt, Austria
Tho Manh Nguyen	Vienna University of Technology, Austria

Program Committee Members

Alberto Abello	Universitat Politecnica de Catalunya, Spain
Jose Balcazar	Universitat Politecnica de Catalunya, Spain
Elena Baralis	Politecnico di Torino, Italy
Ladjel Bellatreche	LISI/ENSMA, France
Bettina Berendt	Humboldt University Berlin, Germany
Petr Berka	University of Economics Prage, Czech Republic
Jorge Bernardino	ISEC, Polytechnic Institute of Coimbra, Portugal
Sourav Bhowmick	Nanyang Technological University, Singapore
Francesco Bonchi	ISTI - C.N.R., Italy
Henrik Bostrom	University of Skövde, Sweden
Jean Francois Boulicaut	INSA Lyon, France
Stephane Bressan	National University of Singapore, Singapore
Peter Brezany	University of Vienna, Austria
Luca Cabibbo	Università Roma Tre, Italy
Tru Hoang Cao	Ho Chi Minh City University of Technology, Vietnam
Jesus Cerquides	University of Barcelona, Spain
Rada Chirkova	NC State University, USA
Sunil Choenni	University of Twente and Dutch Ministry of Justice, Netherlands
Ezeife Christie	University of Windsor, Canada
Frans Coenen	The University of Liverpool, UK
Bruno Cremilleux	Université de Caen, France
Alfredo Cuzzocrea	University of Calabria, Italy
Agnieszka Dardzinska	Bialystok Technical University, Poland
Karen Davis	University of Cincinnati, USA
Kevin Desouza	University of Washington, USA
Draheim Dirk	Software Competence Center Hagenberg, Austria
Johann Eder	University of Klagenfurt, Austria
Tapio Elomaa	Tampere University of Technology, Finland
Roberto Esposito	University of Turin, Italy
Vladimir Estivill-Castro	Griffith University, Australia
Ling Feng	Tsinghua University, China

Eduardo Fernandez-Medina	University of Castilla-La Mancha, Spain
Johannes Fernkranz	TU Darmstadt, Germany
Dragan Gamberger	Rudjer Boskovic Institute, Croatia
Chris Giannella	Goucher College, USA
Matteo Golfarelli	DEIS - University of Bologna, Italy
Gunter Grieser	Technical University Darmstadt, Germany
Mirsad Hadzikadic	College of Information Technology, UNC Charlotte, USA
Eui-Hong Han	iXmatch Inc., USA
Wook-Shin Han	Kyungpook National University, Korea
Eyke Hermeier	University of Magdeburg, Germany
Se June Hong	IBM TJ Watson Research Center, USA
Frank Hoppner	University of Applied Sciences Braunschweig/Wolfenbuettel, Germany
Andreas Hotho	University of Kassel, Germany
Jimmy Huang	York University, Canada
Yong Hwan-Seung	Ewha Womans University, Korea
Ryutaro Ichise	National Institute of Informatics, Japan
Yoo Illhoi	University of Missouri, Columbia, USA
Hollmen Jaakko	Helsinki University of Technology, Finland
Fan Jianping	University of North Carolina, USA
Alipio M. Jorge	FEP/LIACC, University of Porto, Portugal
Panos Kalnis	National University of Singapore, Singapore
Murat Kantarcioglu	University of Texas at Dallas, USA
Mehmet Kaya	Firat University, Turkey
Eamonn Keogh	University of California - Riverside, USA
Jinho Kim	Kangwon National University, Korea
Sang-Wook Kim	Han Yang University, Korea
Jörg Kindermann	Fraunhofer Institute Intelligent Analysis and Information Systems (IAIS), Germany
Arno Knobbe	Universiteit Utrecht, Netherlands
Igor Kononenko	University of Ljubljana, Slovenia
Stefan Kramer	Technische Universität München, Germany
Michihiro Kuramochi	Google Inc., USA
Christine Largeron	Laboratoire Hubert Curien University Jean Monnet Saint-Etienne, France
Jens Lechtenbörger	Universität Münster, Germany
Guanling Lee	National Dong Hwa University, Taiwan
Jae-Gil Lee	UIUC, USA
Young-Koo Lee	Kyung Hee University, Korea
Yue-Shi Lee	Ming Chuan University, Taiwan
Jinyan Li	Institute for Infocomm Research, Singapore
Xuemin Lin	UNSW, Australia
Woong-Kee Loh	Sungkyul University, South Korea
Sergio Lujan-Mora	University of Alicante, Spain
Donato Malerba	Università degli Studi di Bari, Italy

Nikos Mamoulis	University of Hong Kong, Hong Kong
Giuseppe Manco	ICAR-CNR, National Research Council, Italy
Damiani Maria-Luisa	Università degli Studi di Milano, Italy
Michael May	Fraunhofer Institut für Autonome Intelligente Systeme, Germany
Rosa Meo	University of Turin, Italy
Mukesh Mohania	I.B.M. India Research Lab, India
Yang-Sae Moon	Dept. of Computer Science, Kangwon National University, Korea
Eduardo Morales	INAOE, Mexico
Alexandros Nanopoulos	Aristotle University of Thessaloniki, Greece
Wee-Keong Ng	Nanyang Tech. University, Singapore
Tho Manh Nguyen	Vienna University of Technology, Austria
Richard Nock	Université des Antilles-Guyane, Martinique, France
Arlindo Oliveira	IST/INESC-ID, Portugal
Themis Palpanas	University of Trento, Italy
Byung-Kwon Park	Dong-A University, Korea
Torben Bach Pedersen	Aalborg University, Denmark
Dino Pedreschi	University of Pisa, Italy
Jaakko Peltonen	Helsinki University of Technology, Finland
Clara Pizzuti	ICAR-CNR, Italy
Lubomir Popelinsky	Masaryk University, Czech Republic
David Powers	The Flinders University of South Australia, Australia
Jan Ramon	Katholieke Universiteit Leuven, Belgium
Zbigniew Ras	University of North Carolina, Charlotte, USA
Andrew Rau-Chaplin	Dalhousie University, Canada
Mirek Riedewald	Cornell University, USA
Gilbert Ritschard	University of Geneva, Switzerland
Stefano Rizzi	University of Bologna, Italy
Domenico Sacca	University of Calabria & ICAR-CNR, Italy
Monica Scannapieco	University of Rome, Italy
Josef Schiefer	Vienna University of Technology, Austria
Markus Schneider	University of Florida, USA
Michael Schrefl	University of Linz, Austria
Giovanni Semeraro	University of Bari, Italy
Alkis Simitsis	Stanford University, USA
Dan Simovici	University of Massachusetts at Boston, USA
Andrzej Skowron	Warsaw University Banacha 2, Poland
Carlos Soares	University of Porto, Portugal
Il-Yeol Song	Drexel University, USA
Min Song	New Jersey Institute of Technology, USA
Jerzy Stefanowski	Poznan University of Technology, Poland
Olga Stepankova	Czech Technical University, Czech Republic
Reinhard Stolle	BMW Car IT, Germany
Jan Struyf	Katholieke Universiteit Leuven, Belgium
Mika Sulkava	Helsinki University of Technology, Finland
Ah-Hwee Tan	Nanyang Technological University, Singapore

David Taniar	Monash University, Australia
Evimaria Terzi	IBM Almaden, USA
Dimitri Theodoratos	New Jersey's Science & Technology University, USA
A Min Tjoa	Vienna University of Technology, Austria
Juan Trujilo	University of Alicante, Spain
Jaideep Vaidya	Rutgers University, USA
Panos Vassiliadis	University of Ioannina, Greece
Roland Wagner	Johannes Kepler University Linz, Austria
Marek Wojciechowski	Poznan University of Technology, Poland
Wolfram Wöß	University of Linz, Austria
Hu Xiaohua	Drexel University, USA
An Yuan	Drexel University, USA
Mohammed Zaki	Rensselaer Polytechnic Institute, USA
Carlo Zaniolo	University of California, Los Angeles, USA
Xiuzhen Jenny Zhang	RMIT University, Australia
Shichao Zhang	Guangxi Normal University, China
Djamel Abdelkader Zighed	University of Lyon, France

Table of Contents

Conceptual Design and Modeling

UML-Based Modeling for What-If Analysis 1
Matteo Golfarelli and Stefano Rizzi

Model-Driven Metadata for OLAP Cubes from the Conceptual Modelling of Data Warehouses 13
Jesús Pardillo, Jose-Norberto Mazón, and Juan Trujillo

An MDA Approach for the Development of Spatial Data Warehouses . . . 23
Octavio Glorio and Juan Trujillo

OLAP and Cube Processing

Built-In Indicators to Discover Interesting Drill Paths in a Cube 33
Véronique Cariou, Jérôme Cubillé, Christian Derquenne, Sabine Goutier, Françoise Guisnel, and Henri Klajnmic

Upper Borders for Emerging Cubes 45
Sébastien Nedjar, Alain Casali, Rosine Cicchetti, and Lotfi Lakhal

Top_Keyword: An Aggregation Function for Textual Document OLAP 55
Franck Ravat, Olivier Teste, Ronan Tournier, and Gilles Zurfluh

Distributed Data Warehouse

Summarizing Distributed Data Streams for Storage in Data Warehouses 65
Raja Chiky and Georges Hébrail

Efficient Data Distribution for DWS 75
Raquel Almeida, Jorge Vieira, Marco Vieira, Henrique Madeira, and Jorge Bernardino

Data Partitioning in Data Warehouses: Hardness Study, Heuristics and ORACLE Validation 87
Ladjel Bellatreche, Kamel Boukhalfa, and Pascal Richard

Data Privacy in Data Warehouse

A Robust Sampling-Based Framework for Privacy Preserving OLAP 97
Alfredo Cuzzocrea, Vincenzo Russo, and Domenico Saccà

Generalization-Based Privacy-Preserving Data Collection 115
Lijie Zhang and Weining Zhang

Processing Aggregate Queries on Spatial OLAP Data 125
Kenneth Choi and Wo-Shun Luk

Data Warehouse and Data Mining

Efficient Incremental Maintenance of Derived Relations and BLAST Computations in Bioinformatics Data Warehouses 135
Gabriela Turcu, Svetlozar Nestorov, and Ian Foster

Mining Conditional Cardinality Patterns for Data Warehouse Query Optimization 146
Mikołaj Morzy and Marcin Krystek

Up and Down: Mining Multidimensional Sequential Patterns Using Hierarchies 156
Marc Plantevit, Anne Laurent, and Maguelonne Teisseire

Clustering I

Efficient K-Means Clustering Using Accelerated Graphics Processors ... 166
S.A. Arul Shalom, Manoranjan Dash, and Minh Tue

Extracting Knowledge from Life Courses: Clustering and Visualization 176
Nicolas S. Müller, Alexis Gabadinho, Gilbert Ritschard, and Matthias Studer

A Hybrid Clustering Algorithm Based on Multi-swarm Constriction PSO and GRASP 186
Yannis Marinakis, Magdalene Marinaki, and Nikolaos Matsatsinis

Clustering II

Personalizing Navigation in Folksonomies Using Hierarchical Tag Clustering 196
Jonathan Gemmell, Andriy Shepitsen, Bamshad Mobasher, and Robin Burke

Clustered Dynamic Conditional Correlation Multivariate GARCH Model 206
Tu Zhou and Laiwan Chan

Document Clustering by Semantic Smoothing and Dynamic Growing Cell Structure (DynGCS) for Biomedical Literature 217
Min Song, Xiaohua Hu, Illhoi Yoo, and Eric Koppel

Mining Data Streams

Mining Serial Episode Rules with Time Lags over Multiple Data Streams 227
Tung-Ying Lee, En Tzu Wang, and Arbee L.P. Chen

Efficient Approximate Mining of Frequent Patterns over Transactional Data Streams 241
Willie Ng and Manoranjan Dash

Continuous Trend-Based Clustering in Data Streams 251
Maria Kontaki, Apostolos N. Papadopoulos, and Yannis Manolopoulos

Mining Multidimensional Sequential Patterns over Data Streams 263
Chedy Raïssi and Marc Plantevit

Classification

Towards a Model Independent Method for Explaining Classification for Individual Instances 273
Erik Štrumbelj and Igor Kononenko

Selective Pre-processing of Imbalanced Data for Improving Classification Performance 283
Jerzy Stefanowski and Szymon Wilk

A Parameter-Free Associative Classification Method 293
Loïc Cerf, Dominique Gay, Nazha Selmaoui, and Jean-François Boulicaut

Text Mining and Taxonomy I

The Evaluation of Sentence Similarity Measures 305
Palakorn Achananuparp, Xiaohua Hu, and Xiajiong Shen

Labeling Nodes of Automatically Generated Taxonomy for Multi-type Relational Datasets 317
Tao Li and Sarabjot S. Anand

Towards the Automatic Construction of Conceptual Taxonomies 327
Dino Ienco and Rosa Meo

Text Mining and Taxonomy II

Adapting LDA Model to Discover Author-Topic Relations for Email Analysis 337
Liqiang Geng, Hao Wang, Xin Wang, and Larry Korba

A New Semantic Representation for Short Texts 347
M.J. Martín-Bautista, S. Martínez-Folgoso, and M.A. Vila

Document-Base Extraction for Single-Label Text Classification 357
Yanbo J. Wang, Robert Sanderson, Frans Coenen, and Paul Leng

Machine Learning Techniques

How an Ensemble Method Can Compute a Comprehensible Model 368
José L. Triviño-Rodriguez, Amparo Ruiz-Sepúlveda, and Rafael Morales-Bueno

Empirical Analysis of Reliability Estimates for Individual Regression Predictions 379
Zoran Bosnić and Igor Kononenko

User Defined Partitioning - Group Data Based on Computation Model 389
Qiming Chen and Meichun Hsu

Data Mining Applications

Workload-Aware Histograms for Remote Applications 402
Tanu Malik and Randal Burns

Is a Voting Approach Accurate for Opinion Mining? 413
Michel Plantié, Mathieu Roche, Gérard Dray, and Pascal Poncelet

Mining Sequential Patterns with Negative Conclusions 423
Przemysław Kazienko

Author Index 433