Lecture Notes in Computer Science 9670

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

Editorial Board

David Hutchison
Lancaster University, Lancaster, UK
Takeo Kanade
Carnegie Mellon University, Pittsburgh, PA, USA
Josef Kittler
University of Surrey, Guildford, UK
Jon M. Kleinberg
Cornell University, Ithaca, NY, USA
Friedemann Mattern
ETH Zurich, Zürich, Switzerland
John C. Mitchell
Stanford University, Stanford, CA, USA
Moni Naor
Weizmann Institute of Science, Rehovot, Israel
C. Pandu Rangan
Indian Institute of Technology, Madras, India
Bernhard Steffen
TU Dortmund University, Dortmund, Germany
Demetri Terzopoulos
University of California, Los Angeles, CA, USA
Doug Tygar
University of California, Berkeley, CA, USA
Gerhard Weikum
Max Planck Institute for Informatics, Saarbrücken, Germany

More information about this series at http://www.springer.com/series/8637

Abdelkader Hameurlain · Josef Küng
Roland Wagner · Ladjel Bellatreche
Mukesh Mohania (Eds.)

Transactions on Large-Scale Data- and Knowledge-Centered Systems XXVI

Special Issue on Data Warehousing and Knowledge Discovery

Springer

Editors-in-Chief

Abdelkader Hameurlain
IRIT, Paul Sabatier University
Toulouse
France

Josef Küng
FAW, University of Linz
Linz
Austria

Roland Wagner
FAW, University of Linz
Linz
Austria

Guest Editors

Ladjel Bellatreche
LIAS/ISAE-ENSMA
Chasseneuil
France

Mukesh Mohania
IBM India Research Lab
New Delhi
India

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Computer Science
ISBN 978-3-662-49783-8 ISBN 978-3-662-49784-5 (eBook)
DOI 10.1007/978-3-662-49784-5

Library of Congress Control Number: 2016932745

© Springer-Verlag Berlin Heidelberg 2016
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer-Verlag GmbH Berlin Heidelberg

Special Issue of DaWak 2014

We welcome you to this special issue dedicated to the best papers presented at the 16th International Conference on Data Warehousing and Knowledge Discovery (DaWaK), held in Munich, Germany, September 1–5, 2014. The subject of data warehousing and knowledge discovery has been widely accepted as a key technology for enterprises and organizations to improve 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 considered become more and more complex in both structure and semantic. New developments such as cloud computing and big data add to the challenges with massive scaling, a new computing infrastructure, and new types of data. Consequently, the process of retrieval and knowledge discovery from this deluge of heterogeneous complex data represents the litmus test for the research in the area.

The DaWaK conference has become one of the most important international scientific events bringing together researchers, developers, and practitioners to discuss the latest research issues and experiences in developing and deploying data warehousing and knowledge discovery systems, applications, and solutions. DaWaK is in the top 20 of the Google Scholar ranking related to data mining and analysis (http://scholar.google.com/citations?view_op=top_venues&hl=fr&vq=eng_dataminingанalysis). This year's DaWaK conference built on this tradition of facilitating the cross-disciplinary exchange of ideas, experience, and potential research directions. DaWaK 2014 sought to introduce innovative principles, methods, models, algorithms and solutions, industrial products, and experiences to challenging problems faced in the development of data warehousing, knowledge discovery, data mining applications, and the emerging area of high-performance computing (HPC).

The DaWaK 2014 call for papers attracted 109 papers and the Program Committee finally selected 34 full papers and eight short papers, yielding an acceptance rate of 31%. The accepted papers cover a number of broad research areas on both theoretical and practical aspects of data warehouse and knowledge discovery. In the area of data warehousing, the topics covered included modeling and ETL, ontologies, real-time data warehouses, query optimization, the MapReduce paradigm, storage models, scalability, distributed and parallel processing and data warehouses and data mining applications integration, recommendation and personalization, multidimensional analysis of text documents, and data warehousing for real-world applications such as health, bio-informatics, telecommunication, etc. In the areas of data mining and knowledge discovery, the topics included stream data analysis and mining, traditional data mining techniques, topics such as frequent item sets, clustering, association, classification ranking and application of data mining technologies to real-world problems, and fuzzy mining, skyline, etc. It is especially notable to see that some papers covered emerging real-world applications such as bioinformatics, social networks, telecommunication, brain analysis, etc.

This year we had three special issues for the following well-known journals: *Knowledge and Information Systems: An International Journal*, Springer, *Journal of Concurrency and Computation: Practice and Experience*, Wiley, and *Transactions on Large-Scale Data- and Knowledge-Centered Systems - TLDKS*, Springer.

Out of the 34 full papers, we invited the authors of seven papers to be included in the special issue of the LNCS journal *Transactions on Large-Scale Data- and Knowledge-Centered Systems* and after a second round of reviews, we finally accepted four papers. Thus, the relative acceptance rate for the papers included in this special issue is competitive. Needless to say, these four papers represent innovative and high-quality research. Incidentally, they uniformly cover the two major topics of the DaWaK conference: data warehousing (data cube computation and the process of the construction and analysis of a data warehouse in the context of cancer epidemiology) and knowledge discovery (pattern mining algorithms and frequent item-set border approximation)

We congratulate the authors of these four papers and thank all authors who submitted articles to DaWaK.

The four selected papers are summarized as follows:

The first paper titled "Banded Pattern Mining Algorithms in Multi-dimensional Zero-One Data," by Fatimah Abdullahi, Frans Coenen, and Russell Martin, studies the problem of banded pattern mining in high-dimensional binary data. A major novelty of the proposed algorithm is that it can deal with n-dimensional data, while traditional banded pattern mining methods are devised for 2D data. Unlike the competing methods, the proposed method does not rely on the generation of data permutation but on the much more efficient calculation of a "banding score." Two variations of the method are proposed, an approximate version (NDBPM_approx), which takes into account dimension pairs, and an exact version (NDBPM_exact), which considers the entire data space. The exact algorithm has two versions itself, which differ for the weighting scheme used (Euclidean distance or Manhattan distance). The efficiency and efficacy of the proposed algorithms are assessed first against competing methods in the 2D data space, and subsequently against each other in 3D and 5D data space. Experimental results are encouraging for both the approximate and for the exact version of the NDBPM algorithm, even if they are outperformed by competing algorithms in some cases.

The second paper titled "Frequent Itemset Border Approximation by Dualization," by Nicolas Durand and Mohamed Quafafou, presents a data mining approach, called FIBAD, for approximating frequent itemset borders. The aim is the reduction of the border size in order to make easier the exploitation of the contained itemsets. The proposed approach introduces an approximate dualization method considering both the maximal frequent itemsets (positive border) and the minimal infrequent itemsets (negative border). This method computes the approximate minimal hypergraph transversals using hypergraph reduction. Several experiments are conducted showing that the proposed approach outperforms the existing ones as it reduces effectively the size of the generated borders while remaining close to the exact solutions.

The third paper titled "Dynamic Materialization for Building Personalized Smart Cubes," by Daniel Antwi and Herna Viktor, addresses the issue of optimization of OLAP queries. It combines vertical partitioning, partial view materialization, and dynamic adaptation to the "user's interest," called personalization. An interesting overview of the state of the art is given, covering the issue of computing a data cube by considering two non-functional requirements: the reduction of the overall storage cost and the improvement of query performance. Considering vertical partitioning contributes toward managing dynamic data cube computation. Intensive experiments were conducted that show the efficiency of the proposal against the state-of-the-art studies.

The fourth paper entitled "Opening up Data Analysis for Medical Health Services: Data Integration and Analysis in Cancer Registries with CARESS," by David Korfkamp, Stefan Gudenkauf, Martin Rohde, Eunice Sirri, Joachim Kieschke, Kolja Blohm, Alexander Beck, Alexandr Puchkovskiy, and H.-Jürgen Appelrath, presents two software systems, CARESS (CARLOS Epidemiological and Statistical Data Exploration System) — an analytical information system for data warehouses deployed in epidemiological cancer registries in several German states — and CARELIS (CARLOS Record Linkage System), an upstream tool preparing data for CARESS compliant to restrictive German data privacy laws. CARESS addresses an issue in the German cancer epidemiology field where a new law demands the execution of comparable survival analyses. In order to be compliant with this new law, CARESS has been extended by a module that implements the demanded survival analysis methods and makes them easily accessible to a wider audience via the convenient CARESS user interface. The paper figures an important process in epidemiological cancer registries: the pass of a dataset into the registry and integration into a data warehouse using CARELIS and analyzing the data with CARESS.

February 2016

Ladjel Bellatreche
Mukesh Mohania

Organization

Editorial Board

Reza Akbarinia	Inria, France
Bernd Amann	LIP6 - UPMC, France
Dagmar Auer	FAW, Austria
Stéphane Bressan	National University of Singapore, Singapore
Francesco Buccafurri	Università Mediterranea di Reggio Calabria, Italy
Qiming Chen	HP-Lab, USA
Mirel Cosulschi	University of Craiova, Romania
Dirk Draheim	University of Innsbruck, Austria
Johann Eder	Alpen Adria University Klagenfurt, Austria
Georg Gottlob	Oxford University, UK
Anastasios Gounaris	Aristotle University of Thessaloniki, Greece
Theo Härder	Technical University of Kaiserslautern, Germany
Andreas Herzig	IRIT, Paul Sabatier University, France
Dieter Kranzlmüller	Ludwig-Maximilians-Universität München, Germany
Philippe Lamarre	INSA Lyon, France
Lenka Lhotská	Czech Technical University in Prague, Czech Republic
Vladimir Marik	Czech Technical University in Prague, Czech Republic
Franck Morvan	Paul Sabatier University, IRIT, France
Kjetil Nørvåg	Norwegian University of Science and Technology, Norway
Gultekin Ozsoyoglu	Case Western Reserve University, USA
Themis Palpanas	Paris Descartes University, France
Torben Bach Pedersen	Aalborg University, Denmark
Günther Pernul	University of Regensburg, Germany
Sherif Sakr	University of New South Wales, Australia
Klaus-Dieter Schewe	University of Linz, Austria
A Min Tjoa	Vienna University of Technology, Austria
Chao Wang	Oak Ridge National Laboratory, USA

External Reviewers

Mohammed Al-Kateb	Teradata, USA
Ladjel Bellatreche	LIAS/ISAE-ENSMA, France
Kamel Boukhalfa	USTHB, Algiers, Algeria
Markus Endres	University of Augsburg, Germany
Christie Ezeife	University of Windsor, Canada
Filippo Furfaro	University of Calabria, Italy

Carlos Garcia-Alvarado	Amazon Web Services, USA
Sergio Greco	University of Calabria, Italy
Selma Khouri	National High School of Computer Science, Algiers, Algeria
Sofian Maabout	University of Bordeaux, France
Mukesh Mohania	IBM India
Lu Qin	University of Technology, Sydney, Australia
Robert Wrembel	Poznan University of Technology, Poland
Karine Zeitouni	University of Versailles, Saint-Quentin-en-Yvelines, France

Contents