

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, Zurich, 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, Chennai, 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/7409>

Carlos Ordonez · Ladjel Bellatreche (Eds.)

Big Data Analytics and Knowledge Discovery

20th International Conference, DaWaK 2018
Regensburg, Germany, September 3–6, 2018
Proceedings

Editors

Carlos Ordonez
University of Houston
Houston, TX
USA

Ladjel Bellatreche
LIAS/ISAE-ENSMA
Chasseneuil-du-Poitou
France

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Computer Science
ISBN 978-3-319-98538-1 ISBN 978-3-319-98539-8 (eBook)
<https://doi.org/10.1007/978-3-319-98539-8>

Library of Congress Control Number: 2018950524

LNCS Sublibrary: SL3 – Information Systems and Applications, incl. Internet/Web, and HCI

© Springer Nature Switzerland AG 2018

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. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Big data analytics and knowledge discovery remain hot research areas for both academia and the software industry, further fueled by advances in hardware and software. Important research topics associated with these major themes include data lakes (schema-free repositories), database design (ER modeling, prototyping), data integration (especially linking structured and semistructured data sources), big data management (mixing relational tables, text and any files), query languages (SQL and beyond), scalable analytic algorithms, parallel systems (cloud, parallel database systems, Spark, MapReduce, HDFS), theoretical foundations, and practical applications.

With a track record of 20 editions, the International Conference on Big Data Analytics and Knowledge Discovery (DaWaK) has established itself as a high-quality forum for researchers, practitioners and developers in the field of big data analytics. This year's conference (DaWaK 2018) built on this tradition, facilitating the interdisciplinary exchange of ideas, theory, techniques, experiences, and future research directions. DaWaK 2018 aimed to introduce innovative principles, methods, models, algorithms, industrial products, and experiences to solve challenging problems faced in the development of new-generation data management and analytic systems in the big data era.

Our call for papers attracted 76 papers, from which the Program Committee finally selected 13 full papers and 16 short papers, yielding an acceptance rate of 17% for full papers and 38% overall. Each paper was reviewed by at least three reviewers and in some cases up to five. Accepted papers cover a number of broad research areas on both theoretical and practical aspects. Some trends found in accepted papers include new generations of data warehouses, data lakes, data pre-processing, data mining, cloud computing, query processing, sequences, graph analytics, privacy-preserving data mining, and parallel processing. On the other hand, the program featured interesting case studies on social networks, Twitter sentiment analysis, understanding ground transportation modes, and E-commerce, among others.

For this 20th edition of DaWaK, we were pleased to have Prof. Il-Yeol Song from Drexel University (USA) as keynote speaker, giving an intriguing talk entitled: "Smart Aging: Topics, Applications, Technologies, and Agenda." Il-Yeol is an ACM Distinguished Scientist, an ER Fellow, and recipient of the 2015 Peter P. Chen Award in Conceptual Modeling.

Thanks to the history and reputation of DaWaK, editors of well-known journals agreed to receive extended versions of best papers selected from our program. This year, we were pleased to have two special issues in: *Data and Knowledge Engineering* (DKE, Elsevier) and *Transactions on Large-Scale Data- and Knowledge-Centered Systems* (TLDKS, Springer).

We would like to thank all authors for submitting their papers to DaWaK 2018 and we hope they submit again in the future. We express our gratitude to all the Program Committee members who provided high-quality reviews. We appreciate the great

efforts of Amin Anjomshoaa for helping extend the confdriver system with several innovations to improve paper reviews, to help in deciding between full and short length, to manage a conference-to-journal long-term review process, and to create an interesting, packed, thought-provoking conference program. Finally, we would like to especially thank Gabriela Wagner for her endless help and patience.

For conference attendees, we hope they enjoyed the technical program, informal meetings, and interaction with colleagues from all over the world; and of course, we are confident they liked the picturesque city of Regensburg, Germany. For the readers of these proceedings, we hope these papers are interesting and they give you ideas for future research.

September 2018

Carlos Ordonez
Ladjel Bellatreche

Organization

Program Committee Co-chairs

Carlos Ordonez	University of Houston, USA
Ladjel Bellatreche	ISAE-ENSMA, France

Program Committee

Alberto Abelló	Universitat Politècnica de Catalunya, Spain
Toshiyuki Amagasa	University of Tsukuba, Japan
Elena Baralis	Politecnico di Torino, Italy
Ladjel Bellatreche	ENSMA, France
Sadok Ben Yahia	Faculty of Sciences of Tunis, Tunisia
Jorge Bernardino	ISEC - Polytechnic Institute of Coimbra, Portugal
Vasudha Bhatnagar	Delhi University, India
Omar Boussaid	University of Lyon/Lyon 2, France
Stephane Bressan	National University of Singapore, Singapore
Wellington Cabrera	Teradata, USA
Sharma Chakravarthy	The University of Texas at Arlington, USA
Isabelle Comyn-Wattiau	ESSEC Business School, Paris, France
Alfredo Cuzzocrea	University of Trieste, Italy
Laurent d'Orazio	University of Rennes 1, France
Soumyava Das	Teradata, USA
Karen Davis	Miami University, USA
Claudia Diamantini	Università Politecnica delle Marche, Italy
Josep Domingo-Ferrer	Universitat Rovira i Virgili, Spain
Dejing Dou	University of Oregon, USA
Curtis Dyreson	Utah State University, USA
Markus Endres	University of Augsburg, Germany
Leonidas Fegaras	The University of Texas at Arlington, USA
Filippo Furfaro	DIMES, University of Calabria, Italy
Pedro Furtado	Universidade de Coimbra, Portugal
Carlos Garcia-Alvarado	Autonomic LLC, USA
Javier Garcia-Garcia	C3 UNAM University, Mexico
Kazuo Goda	The University of Tokyo, Japan
Matteo Golfarelli	DISI, University of Bologna, Italy
Sergio Greco	University of Calabria, Italy
Abdelkader Hameurlain	Paul Sabatier University, France
Takahiro Hara	Osaka University, Japan
Frank Hoppner	Ostfalia University of Applied Sciences, Germany
Yoshiharu Ishikawa	Nagoya University, Japan
Stéphane Jean	LIAS/ISAE-ENSMA and University of Poitiers, France

Lili Jiang	Umeå University, Sweden
Selma Khouri	LCSI/ESI (Algeria) and LIAS/ISAE-ENSMA (France), Algeria
Uday Kiran Rage	University of Tokyo, Japan
Nhan Le-Thanh	Nice Sophia Antipolis University, France
Jens Lechtenboerger	Westfälische Wilhelms–Universität Münster, Germany
Carson Leung	University of Manitoba, Canada
Sofian Maabout	University of Bordeaux, France
Yannis Manolopoulos	Aristotle University of Thessaloniki, Greece
Patrick Marcel	Université François Rabelais Tours, France
Jun Miyazaki	Tokyo Institute of Technology, Japan
Anirban Mondal	Ashoka University, India
Yasuhiko Morimoto	Hiroshima University, Japan
Makoto Onizuka	Osaka University, Japan
Carlos Ordonez	University of Houston, USA
Alex Poulouvassilis	Birkbeck, University of London, UK
Praveen Rao	University of Missouri-Kansas City, USA
Goce Ristanoski	Data61, CSIRO, Australia
Oscar Romero	Universitat Politècnica de Catalunya, Spain
Laura Rusu	IBM, Australia
Ilya Safro	Clemson University, USA
Sherif Sakr	Software Systems Research Group, NICTA, University of New South Wales, Australia
Abhishek Santra	University of Texas at Arlington, USA
Alkis Simitsis	HP Labs, USA
Emanuele Storti	Università Politecnica delle Marche, Italy
Aditya Telang	University of Texas at Arlington, USA
Olivier Teste	IRIT, University of Toulouse, France
Dimitri Theodoratos	New Jersey Institute of Technology, USA
Predrag Tosic	Washington State University, USA
Panos Vassiliadis	University of Ioannina, Greece
Robert Wrembel	Poznan University of Technology, Poland
Yinghui Wu	Washington State University, USA
Wai GenYee	Grubhub, USA
Yinuo Zhang	Teradata, USA
Yiqun Zhang	VoltDB Inc., USA

Additional Reviewers

Mohammed El Malki	Toulouse University/IRIT-UMR5505, France
Sergio Martínez	Universitat Rovira i Virgili, Catalonia
Luis del Vasto	Universitat Rovira i Virgili, Catalonia
Michael Bamiloshin	Universitat Rovira i Virgili, Catalonia
Calvin S. H. Hoi	University of Manitoba, Canada
Fan Jiang	University of Northern British Columbia, Canada
Adam G. M. Pazdor	University of Manitoba, Canada

Oluwafemi A. Sarumi	Federal University of Technology – Akure, Nigeria
Neelabh Pant	Walmart, Plano, Texas, USA
Abhishek Santra	University of Arlington, Texas, USA
Soumyava Das	Teradata, CA, USA
Aggeliki Dimitriou	National Technical University of Athens, Greece
Michael Lan	New Jersey Institute of Technology, USA
Xiaoying Wu	Wuhan University, China
Daniele Apiletti	Politecnico di Torino, Italy
Andrea Pasini	Politecnico di Torino, Italy
Eliana Pastor	Politecnico di Torino, Italy
Anastasios Gounaris	Aristotle University of Thessaloniki, Greece
Irina Trubitsyna	University of Calabria, Italy
Gianvincenzo Alfano	University of Calabria, Italy
Francesco Parisi	University of Calabria, Italy
Sharanjit Kaur	University of Delhi, India
Daniel Ernesto Lopez Barron	University of Missouri-Kansas City, USA
Amira Mouakher	Faculty of Sciences of Tunis, Tunisia

Smart Aging: Topics, Applications, Technologies, and Agenda (Abstract of Keynote Speaker)

Il-Yeol Song

College of Computing and Informatics, Drexel University,
Philadelphia, PA, USA
songiy@drexel.edu

Abstract. Aging is a rapidly growing social problem in the developed world. It is critically important to mitigate the effects of aging, improve elderly people's life, and improve overall quality of healthcare environments. Smart aging addresses those challenges by intelligently utilizing modern biomedical, digital healthcare, big data computing & analytics, IOT, and communication technologies. In this talk, I will first review several innovative R&D projects and services for smart aging. I will then present a comprehensive review of various research activities on smart aging from the content analysis of public web pages and the web pages of NIH funded research projects related to smart aging. I will then cover recent developments in big data technologies for smart health, including healthcare data warehouses, data lakes and big data analytics. I will conclude my talk with a summary of suggestions on smart aging projects and research topics in smart aging.

Keywords: Smart aging · Big data technologies · Smart health
Healthcare data warehouses · Healthcare data lake

Contents

Graph Analytics

Graph BI & Analytics: Current State and Future Challenges.	3
<i>Amine Ghrab, Oscar Romero, Salim Jouili, and Sabri Skhiri</i>	
Community Detection in Who-calls-Whom Social Networks.	19
<i>Ciprian-Octavian Truică, Olivera Novović, Sanja Brdar, and Apostolos N. Papadopoulos</i>	
FedS: Towards Traversing Federated RDF Graphs	34
<i>Qaiser Mehmood, Alok Kumar Jha, Dietrich Rebholz-Schuhmann, and Ratnesh Sahay</i>	

Case Studies

Adversarial Spiral Learning Approach to Strain Analysis for Bridge Damage Detection.	49
<i>Takaya Kawakatsu, Akira Kinoshita, Kenro Aihara, Atsuhiko Takasu, and Jun Adachi</i>	
CoRe: Generating a Computationally Representative Road Skeleton - Integrating AADT with Road Structure	59
<i>Rohith Reddy Sankepally and K. S. Rajan</i>	
E-Commerce Product Recommendation Using Historical Purchases and Clickstream Data.	70
<i>Ying Xiao and C. I. Ezeife</i>	
Effective Classification of Ground Transportation Modes for Urban Data Mining in Smart Cities.	83
<i>Carson K. Leung, Peter Braun, and Adam G. M. Pazdor</i>	
Location Prediction Using Sentiments of Twitter Users	98
<i>Ritu Singh and Durga Toshniwal</i>	

Classification and Clustering

A Clustering Model for Uncertain Preferences Based on Belief Functions . . .	111
<i>Yiru Zhang, Tassadit Bouadi, and Arnaud Martin</i>	
A Novel Committee-Based Clustering Method	126
<i>Sonia Fiol-Gonzalez, Cassio Almeida, Simone Barbosa, and H�lio Lopes</i>	

KMN - Removing Noise from K-Means Clustering Results	137
<i>Benjamin Schelling and Claudia Plant</i>	
Subset Labeled LDA: A Topic Model for Extreme Multi-label Classification	152
<i>Yannis Papanikolaou and Grigorios Tsoumakas</i>	
Third Party Data Clustering Over Encrypted Data Without Data Owner Participation: Introducing the Encrypted Distance Matrix	163
<i>Nawal Almutairi, Frans Coenen, and Keith Dures</i>	
Pre-proccesing	
An Efficient Prototype Selection Algorithm Based on Spatial Abstraction . . .	177
<i>Joel Luís Carbonera and Mara Abel</i>	
Web Usage Data Cleaning: A Rule-Based Approach for Weblog Data Cleaning.	193
<i>Amine Ganibardi and Chérif Arab Ali</i>	
Anonymization of Multiple and Personalized Sensitive Attributes	204
<i>Jerry Chun-Wei Lin, Qiankun Liu, Philippe Fournier-Viger, Youcef Djenouri, and Ji Zhang</i>	
TRANS-AM: Discovery Method of Optimal Input Vectors Corresponding to Objective Variables	216
<i>Hiroaki Tanaka, Yu Suzuki, Koichiro Yoshino, and Satoshi Nakamura</i>	
Sequences	
Discovering Periodic Patterns Common to Multiple Sequences	231
<i>Philippe Fournier-Viger, Zhitian Li, Jerry Chun-Wei Lin, Rage Uday Kiran, and Hamido Fujita</i>	
Discovering Tight Space-Time Sequences	247
<i>Riccardo Campisano, Heraldo Borges, Fabio Porto, Fabio Perosi, Esther Pacitti, Florent Massegli, and Eduardo Ogasawara</i>	
Cloud and Database Systems	
CloudDBGuard: Enabling Sorting and Searching on Encrypted Data in NoSQL Cloud Databases	261
<i>Tim Waage and Lena Wiese</i>	
Query Processing on Large Graphs: Scalability Through Partitioning.	271
<i>Jay Bodra, Soumyava Das, Abhishek Santra, and Sharma Chakravarthy</i>	

Querying Heterogeneous Data in Graph-Oriented NoSQL Systems	289
<i>Mohammed El Malki, Hamdi Ben Hamadou, Max Chevalier, André Péninou, and Olivier Teste</i>	
Selection of Bitmap Join Index: Approach Based on Minimal Transversals	302
<i>Issam Ghabry, Sadok Ben Yahia, and M. Nidhal Jelassi</i>	
Scalable Random Sampling K-Prototypes Using Spark	317
<i>Mohamed Aymen Ben HajKacem, Chiheb-Eddine Ben N'cir, and Nadia Essoussi</i>	
Data Mining	
ERAPN, an Algorithm for Extraction Positive and Negative Association Rules in Big Data	329
<i>Parfait Bemarisika and André Totohasina</i>	
Multistep-ahead Prediction: A Comparison of Analytical and Algorithmic Approaches	345
<i>Fouad Bahrpeyma, Mark Roantree, and Andrew McCarren</i>	
Novel Data Segmentation Techniques for Efficient Discovery of Correlated Patterns Using Parallel Algorithms	355
<i>Amulya Kotni, R. Uday Kiran, Masashi Toyoda, P. Krishna Reddy, and Masaru Kitsuregawa</i>	
Time Series Distance Density Cluster with Statistical Preprocessing	371
<i>Ruizhe Ma, Soukaina Filali Boubrahimi, and Rafal Angryk</i>	
Debate Stance Classification Using Word Embeddings.	382
<i>Anand Konjengbam, Subrata Ghosh, Nagendra Kumar, and Manish Singh</i>	
Author Index	397