# Lecture Notes in Computer Science 12393


## Founding Editors

Gerhard Goos
*Karlsruhe Institute of Technology, Karlsruhe, Germany*
Juris Hartmanis
*Cornell University, Ithaca, NY, USA*

## Editorial Board Members

Elisa Bertino
*Purdue University, West Lafayette, IN, USA*
Wen Gao
*Peking University, Beijing, China*
Bernhard Steffen
*TU Dortmund University, Dortmund, Germany*
Gerhard Woeginger
*RWTH Aachen, Aachen, Germany*
Moti Yung
*Columbia University, New York, NY, USA*

More information about this series at http://www.springer.com/series/7409

Min Song · Il-Yeol Song · Gabriele Kotsis ·
A Min Tjoa · Ismail Khalil (Eds.)

# Big Data Analytics and Knowledge Discovery

22nd International Conference, DaWaK 2020
Bratislava, Slovakia, September 14–17, 2020
Proceedings

Springer

*Editors*
Min Song
Department of Library
and Information
Yonsei University
Seoul, Korea (Republic of)

Gabriele Kotsis
Johannes Kepler University
of Linz
Linz, Austria

Ismail Khalil
Johannes Kepler University
of Linz
Linz, Oberösterreich, Austria

Il-Yeol Song
Drexel University
Philadelphia, PA, USA

A Min Tjoa
Software Competence
Center Hagenberg (Au)
Vienna, Wien, Austria

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Computer Science
ISBN 978-3-030-59064-2 ISBN 978-3-030-59065-9 (eBook)
https://doi.org/10.1007/978-3-030-59065-9

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

© Springer Nature Switzerland AG 2020
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, expressed 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

During the past 21 years, the International Conference on Big Data Analytics and Knowledge Discovery (DaWaK) has been a prominent international forum in data warehousing, big data management, and analytics areas. DaWaK has brought researchers and practitioners together to exchange their experience and knowledge in developing and managing the systems, applications, and solutions for data warehousing, big data analytics, and knowledge discovery. With a track record of 22 editions, 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 2020) also builds on this tradition, facilitating the interdisciplinary exchange of ideas, theories, techniques, experiences, and future research directions. The DaWaK submission covered broad research areas on both theoretical and practical aspects of big data analytics and knowledge discovery. In the areas of big data analytics, the topics covered by submitted papers included data lakes, meta-data management, data engineering for data science, ETL processing, database encryption, text analytics, data integration, data-driven paradigms, NoSQL databases, and big data and cloud computing. In the areas of knowledge discovery, the topics covered by submitted papers included not only traditional data mining papers on classification, regression, clustering, anomaly detection, association rules, recommender systems, ensemble techniques, but also recent hot topics on machine learning algorithms and optimization, deep learning, graph analytics, stream data analytics, text analytics, and privacy-preserving data analysis. In the areas of application-oriented areas, the topics covered by submitted papers included energy-related data, smart meter stream data, financial time-series data, credit scoring, smart farming, tour data analysis, and biomedical data analysis. It was notable to see that some innovative papers covered emerging real-world applications such as smart farming, smart building, and stream data analytics. All these diverse papers show that DaWaK has become a key forum for innovative developments and real-world applications in the areas of big data analytics and knowledge discovery technologies.

Dawak 2020's call for papers attracted 77 papers from 25 countries, from which the Program Committee selected 15 full papers and 14 short papers, yielding an acceptance rate of 20% for full paper category and of 38% for both categories. Each paper was reviewed by at least three reviewers and in some cases up to four. Accepted papers cover a variety of research areas on both theoretical and practical aspects. Some trends found in accepted papers include data lakes, as a new generation of big data repository, data pre-processing, data mining, text mining, sequences, graph mining, and parallel processing. Thanks to the reputation of DaWaK, selected best papers of DaWaK 2020 will be invited for a special issue in *Data & Knowledge Engineering* (DKE, Elsevier) journal. We would like to thank the DKE Editor-in-Chief, Prof. Peter P. Chen, who approved the special issue.

Although COVID-19 struck all aspects of daily lives, and holding an international conference became extremely difficult, DaWak 2020 succeeded in terms of the number of submissions and the quality of submitted papers. We would like to thank all the authors who submitted their papers to DaWaK 2020, as their contributions formed the basis of an excellent program for DaWaK 2020. We would also like to express our sincere gratitude to all Program Committee members and the external reviewers who reviewed the papers profoundly and in a timely manner. Finally, we would like to thank the support and guidance from the DEXA conference organizers, especially, Prof. Ismail Khalil for having provided a great deal of assistance in planning and administering DaWaK 2020.

For virtual conference attendants, we hope they enjoy the technical program and interaction with colleagues from all over the world. For readers of these proceedings, we hope these papers are interesting and they provide insights and ideas for you to conduct future research.

September 2020

Min Song
Il-Yeol Song

# Organization

## General Chair

| | |
|---|---|
| Bernhard Moser | Software Competence Center Hagenberg (SCCH), Austria |

## Program Committee Chairs

| | |
|---|---|
| Min Song | Yonsei University, South Korea |
| Il-Yeol Song | Drexel University, USA |

## Steering Committee

| | |
|---|---|
| Gabriele Kotsis | Johannes Kepler University Linz, Austria |
| A Min Tjoa | Vienna University of Technology, Austria |
| Ismail Khalil | Johannes Kepler University Linz, Austria |

## Program Committee and Reviewer

| | |
|---|---|
| Alberto Abello | Universitat Politècnica de Catalunya, Spain |
| Julien Aligon | Université de Toulouse, France |
| Toshiyuki Amagasa | University of Tsukuba, Japan |
| Amin Beheshti | Macquarie University, Australia |
| Ladjel Bellatreche | LIAS, ENSMA, France |
| Sadok Ben Yahia | Université de Tunis El Manar, Tunis |
| Fadila Bentayeb | ERIC, Université Lumière Lyon 2, France |
| Rafael Berlanga | Universitat Jaume I, Spain |
| Jorge Bernardino | ISEC, Polytechnic Institute of Coimbra, Portugal |
| Besim Bilalli | Universitat Politècnica de Catalunya, Spain |
| Sandro Bimonte | Irstea, France |
| Kamel Boukhalfa | University of Science and Technology Houari Boumediene, Algeria |
| Stephane Bressan | National University of Singapore, Singapore |
| Tania Cerquitelli | Politecnico di Torino, Italy |
| Sharma Chakravarthy | The University of Texas at Arlington, USA |
| Isabelle Comyn-Wattiau | ESSEC Business School, France |
| Alfredo Cuzzocrea | ICAR-CNR, University of Calabria, Italy |
| Laurent D'Orazio | University of Rennes, CNRS-IRISA, France |
| Jérôme Darmont | Université de Lyon, France |
| Karen Davis | Miami University, USA |
| Ibrahim Dellal | LIAS, ENSMA, France |
| Curtis Dyreson | Utah State University, USA |

| | |
|---|---|
| Markus Endres | University of Passau, Germany |
| Leonidas Fegaras | The University of Texas at Arlington, USA |
| Philippe Fournier-Viger | Harbin Institute of Technology, China |
| Filippo Furfaro | DIMES, University of Calabria, Italy |
| Pedro Furtado | Universidade de Coimbra, CISUC, Portugal |
| Carlos Garcia-Alvarado | Pivotal Inc., USA |
| Kazuo Goda | The University of Tokyo, Japan |
| Matteo Golfarelli | University of Bologna, Italy |
| Marcin Gorawski | Silesian University of Technology and Wroclaw University of Technology, Poland |
| Hyoil Han | Illinois State University, USA |
| Frank Höppner | Ostfalia University of Applied Sciences, Germany |
| Stephane Jean | LISI, ENSMA, University of Poitiers, France |
| Petar Jovanovic | Universitat Politècnica de Catalunya, Barcelona Tech, Spain |
| Vana Kalogeraki | Athens University of Economics and Business, Greece |
| Min-Soo Kim | DGIST, South Korea |
| Uday Kiran | The University of Tokyo, Japan |
| Nicolas Labroche | Université François-Rabelais Tours, France |
| Jens Lechtenbörger | University of Muenster, Germany |
| Young-Koo Lee | Kyung Hee University, South Korea |
| Jae-Gil Lee | Korea Advanced Institute of Science and Technology, South Korea |
| Mong Li Lee | National University of Singapore, Singapore |
| Wolfgang Lehner | TU Dresden, Germany |
| Daniel Lemire | LICEF Research Center, Université du Québec, Canada |
| Sebastian Link | The University of Auckland, New Zealand |
| Xiufeng Liu | Danish Technical University, Denmark |
| Sofian Maabout | LaBRI, University of Bordeaux, France |
| Sanjay Madria | Missouri S & T, USA |
| Patrick Marcel | Université François Rabelais Tours, France |
| Adriana Marotta | Universidad de la Republica, Uruguay |
| Alejandro Maté | University of Alicante, Spain |
| Jose-Norberto Mazon | University of Alicante, Spain |
| Amin Mesmoudi | LIAS, Université de Poitiers, France |
| Jun Miyazaki | Tokyo Institute of Technology, Japan |
| Anirban Mondal | The University of Tokyo, Japan |
| Yang-Sae Moon | Kangwon National University, South Korea |
| Yasuhiko Morimoto | Hiroshima University, Japan |
| Sergi Nadal | Universitat Politècnica de Catalunya, Spain |
| Bernd Neumayr | Johannes Kepler University Linz, Austria |
| Makoto Onizuka | Osaka University, Japan |
| Carlos Ordonez | University of Houston, USA |
| Torben Bach Pedersen | Aalborg University, Denmark |
| Verónika Peralta | University of Tours, France |
| Praveen Rao | University of Missouri, USA |

| | |
|---|---|
| Stefano Rizzi | University of Bologna, Italy |
| Oscar Romero | Universitat Politècnica de Catalunya, Spain |
| Keun Ho Ryu | Chungbuk National University, South Korea |
| Alkis Simitsis | HP Labs, USA |
| Benkrid Soumia | ESI, Algeria |
| Dimitri Theodoratos | New Jersey Institute of Technology, USA |
| Maik Thiele | TU Dresden, Germany |
| Christian Thomsen | Aalborg University, Denmark |
| Predrag Tosic | Washington State University, USA |
| Juan Trujillo | University of Alicante, Spain |
| Panos Vassiliadis | University of Ioannina, Greece |
| Robert Wrembel | Poznan University of Technology, Poland |
| Hwan-Seung Yong | Ewha Womans University, South Korea |
| Yongjun Zhu | Sungkyunkwan University, South Korea |

## Organizers

# Contents


**Position Paper**

Analyzing the Research Landscape of DaWaK Papers from 1999 to 2019 . . . 3
*Tatsawan Timakum, Soobin Lee, Il-Yeol Song, and Min Song*

**Applications**

DHE$^2$: Distributed Hybrid Evolution Engine for Performance Optimizations of Computationally Intensive Applications . . . . . 17
*Oana Stroie, Elena-Simona Apostol, and Ciprian-Octavian Truică*

Grand Reports: A Tool for Generalizing Association Rule Mining to Numeric Target Values . . . . . 28
*Sijo Arakkal Peious, Rahul Sharma, Minakshi Kaushik, Syed Attique Shah, and Sadok Ben Yahia*

Expected vs. Unexpected: Selecting Right Measures of Interestingness . . . . . 38
*Rahul Sharma, Minakshi Kaushik, Sijo Arakkal Peious, Sadok Ben Yahia, and Dirk Draheim*

SONDER: A Data-Driven Methodology for Designing Net-Zero Energy Public Buildings . . . . . 48
*Ladjel Bellatreche, Felix Garcia, Don Nguyen Pham, and Pedro Quintero Jiménez*

Reverse Engineering Approach for NoSQL Databases . . . . . 60
*Fatma Abdelhedi, Amal Ait Brahim, Rabah Tighilt Ferhat, and Gilles Zurfluh*

**Big Data/Data Lake**

HANDLE - A Generic Metadata Model for Data Lakes . . . . . 73
*Rebecca Eichler, Corinna Giebler, Christoph Gröger, Holger Schwarz, and Bernhard Mitschang*

**Data Mining**

A SAT-Based Approach for Mining High Utility Itemsets from Transaction Databases . . . . . 91
*Amel Hidouri, Said Jabbour, Badran Raddaoui, and Boutheina Ben Yaghlane*

High-Utility Interval-Based Sequences . . . . . 107
*S. Mohammad Mirbagheri and Howard J. Hamilton*

Extreme-SAX: Extreme Points Based Symbolic Representation for Time Series Classification . . . . . 122
*Muhammad Marwan Muhammad Fuad*

Framework to Optimize Data Processing Pipelines Using Performance Metrics . . . . . 131
*Syed Muhammad Fawad Ali and Robert Wrembel*

A Scalable Randomized Algorithm for Triangle Enumeration on Graphs Based on SQL Queries . . . . . 141
*Abir Farouzi, Ladjel Bellatreche, Carlos Ordonez, Gopal Pandurangan, and Mimoun Malki*

Data Engineering for Data Science: Two Sides of the Same Coin . . . . . 157
*Oscar Romero and Robert Wrembel*

Mining Attribute Evolution Rules in Dynamic Attributed Graphs . . . . . 167
*Philippe Fournier-Viger, Ganghuan He, Jerry Chun-Wei Lin, and Heitor Murilo Gomes*

Sustainable Development Goal Relational Modelling: Introducing the SDG-CAP Methodology . . . . . 183
*Yassir Alharbi, Frans Coenen, and Daniel Arribas-Bel*

Mining Frequent Seasonal Gradual Patterns . . . . . 197
*Jerry Lonlac, Arnaud Doniec, Marin Lujak, and Stephane Lecoeuche*

Derivative, Regression and Time Series Analysis in SARS-CoV-2 . . . . . 208
*Pedro Furtado*

**Machine Learning and Deep Learning**

Building a Competitive Associative Classifier . . . . . 223
*Nitakshi Sood and Osmar Zaiane*

Contrastive Explanations for a Deep Learning Model on Time-Series Data . . . 235
*Jokin Labaien, Ekhi Zugasti, and Xabier De Carlos*

Cyberbullying Detection in Social Networks Using Deep Learning Based Models . . . . . 245
*Maral Dadvar and Kai Eckert*

Predicting Customer Churn for Insurance Data . . . . . 256
*Michael Scriney, Dongyun Nie, and Mark Roantree*

## Supervised Learning

Scalable Machine Learning on Popular Analytic Languages with Parallel Data Summarization . . . . . 269
*Sikder Tahsin Al-Amin and Carlos Ordonez*

Which Bills Are Lobbied? Predicting and Interpreting Lobbying Activity in the US . . . . . 285
*Ivan Slobozhan, Peter Ormosi, and Rajesh Sharma*

FIBS: A Generic Framework for Classifying Interval-Based Temporal Sequences . . . . . 301
*S. Mohammad Mirbagheri and Howard J. Hamilton*

Multivariate Time Series Classification: A Relational Way . . . . . 316
*Dominique Gay, Alexis Bondu, Vincent Lemaire, Marc Boullé, and Fabrice Clérot*

## Unsupervised Learning

Behave or Be Detected! Identifying Outlier Sequences by Their Group Cohesion . . . . . 333
*Martha Tatusch, Gerhard Klassen, and Stefan Conrad*

Detecting Anomalies in Production Quality Data Using a Method Based on the Chi-Square Test Statistic . . . . . 348
*Michael Mayr and Johannes Himmelbauer*

Learning from Past Observations: Meta-Learning for Efficient Clustering Analyses . . . . . 364
*Manuel Fritz, Dennis Tschechlov, and Holger Schwarz*

Parallel K-Prototypes Clustering with High Efficiency and Accuracy . . . . . 380
*Hiba Jridi, Mohamed Aymen Ben HajKacem, and Nadia Essoussi*

Self-Organizing Map for Multi-view Text Clustering . . . . . 396
*Maha Fraj, Mohamed Aymen Ben Hajkacem, and Nadia Essoussi*

**Author Index** . . . . . 409