9th International Conference on Data Science, Technology and Applications (DATA 2020)

Online
7 – 9 July 2020

Editors:

Slimane Hammoudi
Christoph Quix
Jorge Bernardino

ISBN: 978-1-7138-4058-9

Printed from e-media with permission by:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571

Some format issues inherent in the e-media version may also appear in this print version.

Copyright© (2020) by SCITEPRESS – Science and Technology Publications, Lda.
All rights reserved.

Printed with permission by Curran Associates, Inc. (2021)

For permission requests, please contact SCITEPRESS – Science and Technology Publications, Lda.
at the address below.

SCITEPRESS – Science and Technology Publications, Lda.
Avenida de S. Francisco Xavier, Lote 7 Cv. C,
2900-616 Setúbal, Portugal

Phone: +351 265 520 185
Fax: +351 265520 186

info@scitepress.org

Additional copies of this publication are available from:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: 845-758-0400
Fax: 845-758-2633
Email: curran@proceedings.com
Web: www.proceedings.com

CONTENTS

INVITED SPEAKERS

KEYNOTE SPEAKERS

On the Pareto Principle in Process Mining, Task Mining, and Robotic Process Automation 5
Wil M. P. van der Aalst

From Data to the Press: Data Management for Journalism and Fact-Checking 13
Ioana Manolescu

PAPERS

FULL PAPERS

Auxiliary Decision-making for Controlled Experiments based on Mid-term Treatment Effect Prediction: Applications in Ant Financial's Offline-payment Business 19
Gang Li and Huizhi Xie

A Framework for Creating Policy-agnostic Programming Languages 31
Fabian Bruckner, Julia Pampus and Falk Howar

Capability-based Scheduling of Scientific Workflows in the Cloud 43
Michel Krämer

Integrating Lightweight Compression Capabilities into Apache Arrow 55
Juliana Hildebrandt, Dirk Habich and Wolfgang Lehner

Classification of Products in Retail using Partially Abbreviated Product Names Only 67
Oliver Allweyer, Christian Schorr, Rolf Krieger and Andreas Mohr

On Generating Efficient Data Summaries for Logistic Regression: A Coreset-based Approach 78
Nery Riquelme-Granada, Khuong An Nguyen and Zhiyuan Luo

Catalog Integration of Low-quality Product Data by Attribute Label Ranking 90
Oliver Schmidts, Bodo Kraft, Marvin Winkens and Albert Zündorf

A Gradient Descent based Heuristic for Solving Regression Clustering Problems 102
Enis Kayış

Is Open Data Ready for Use by Enterprises? Learnings from Corporate Registers 109
Pavel Krasikov, Timo Obrecht, Christine Legner and Markus Eurich

Learning Interpretable and Statistically Significant Knowledge from Unlabeled Corpora of Social Text Messages: A Novel Methodology of Descriptive Text Mining 121
Giacomo Frisoni, Gianluca Moro and Antonella Carbonaro

SHORT PAPERS

Applied Data Science: An Approach to Explain a Complex Team Ball Game
Friedemann Schwenkreis and Eckard Nothdurft 135

Trading Desk Behavior Modeling via LSTM for Rogue Trading Fraud Detection
Marine Neyret, Jaouad Ouaggag and Cédric Allain 143

Leveraging Machine Learning for Fake News Detection
Elio Masciari, Vincenzo Moscato, Antonio Picariello and Giancarlo Sperlì 151

Multi-view Clustering Analyses for District Heating Substations
Shahrooz Abghari, Veselka Boeva, Jens Brage and Håkan Grahn 158

Automatic Detection of Gait Asymmetry
Maciej Cwierlikowski and Mercedes Torres Torres 169

Hybrid Multi-model Multi-platform (HM3P) Databases
Sven Groppe and Jinghua Groppe 177

Improving Statistical Reporting Data Explainability via Principal Component Analysis
Shengkun Xie and Clare Chua-Chow 185

Toward a New Quality Measurement Model for Big Data
Mandana Omidbakhsh and Olga Ormandjieva 193

Improving Public Sector Efficiency using Advanced Text Mining in the Procurement Process
Nikola Modrušan, Kornelije Rabuzin and Leo Mršić 200

Reference Data Abstraction and Causal Relation based on Algebraic Expressions
Susumu Yamasaki and Mariko Sasakura 207

Real-time Visualization of Sensor Data in Smart Manufacturing using Lambda Architecture
Nadeem Iftikhar, Bartosz Piotr Lachowicz, Akos Madarasz, Finn Ebertsen Nordbjerg, Thorkil Baattrup-Andersen and Karsten Jeppesen 215

A Conceptual Framework for a Flexible Data Analytics Network
Daniel Tebernum and Dustin Chabrowski 223

Identification of Social Influence on Social Networks and Its Use in Recommender Systems: A Systematic Review
Lesly Alejandra Gonzalez Camacho and Solange Nice Alves-Souza 234

Trust Profile based Trust Negotiation for the FHIR Standard
Eugene Sanzi and Steven A. Demurjian 242

iTLM: A Privacy Friendly Crowdsourcing Architecture for Intelligent Traffic Light Management
Christian Roth, Mirja Nitschke, Matthias Hörmann and Doğan Kesdoğan 252

Initializing k-means Clustering
Christian Borgelt and Olha Yarikova 260

The Need for an Enterprise Risk Management Framework for Big Data Science Projects
Jeffrey Saltz and Sucheta Lahiri 268

Towards Large-scale Gaussian Process Models for Efficient Bayesian Machine Learning 275
Fabian Berns and Christian Beecks

Ensemble Clustering based Semi-supervised Learning for Revenue Accounting Workflow Management 283
Tianshu Yang, Nicolas Pasquier and Frederic Precioso

Predicting the Environment of a Neighborhood: A Use Case for France 294
Nelly Barret, Fabien Duchateau, Franck Favetta and Loïc Bonneval

Context-aware Retrieval and Classification: Design and Benefits 302
Kurt Englmeier

Sentiment Polarity Classification of Corporate Review Data with a Bidirectional Long-Short Term Memory (biLSTM) Neural Network Architecture 310
R. E. Loke and O. Kachaniuk

Towards Self-adaptive Defect Classification in Industrial Monitoring 318
Andreas Margraf, Jörg Hähner, Philipp Braml and Steffen Geinitz

A CART-based Genetic Algorithm for Constructing Higher Accuracy Decision Trees 328
Elif Ersoy, Erinç Albey and Enis Kayış

AUTHOR INDEX 339