DATA 2020

9th International Conference on Data Science, Technology and Applications

Final Program and Book of Abstracts

7 - 9 July, 2020

http://www.dataconference.org

SPONSORED BY

INSTICC

PAPERS AVAILABLE AT

SCITEPRESS DIGITAL LIBRARY

DATA 2020
Final Program and Book of Abstracts

9th International Conference on Data Science, Technology and Applications

Online Streaming
July 7 - 9, 2020

Sponsored by
INSTICC - Institute for Systems and Technologies of Information, Control and Communication

ACM In Cooperation
ACM SIGMIS - ACM Special Interest Group on Management Information Systems

74: Learning Interpretable and Statistically Significant Knowledge from Unlabeled Corpora of Social Text Messages: A Novel Methodology of Descriptive Text Mining, *by Giacomo Frisoni, Gianluca Moro and Antonella Carbonaro* . 24
75: Sentiment Polarity Classification of Corporate Review Data with a Bidirectional Long-Short Term Memory (biLSTM) Neural Network Architecture, *by R. Loke and O. Kachaniuk* 24

Session 6 (14:45 - 16:30)
Room 3: *Data Science* 24
33: Classification of Products in Retail using Partially Abbreviated Product Names Only, *by Oliver Allweyer, Christian Schorr, Rolf Krieger and Andreas Mohr* 24
24: Automatic Detection of Gait Asymmetry, *by Maciej Cwierlikowski and Mercedes Torres* 24
29: Improving Statistical Reporting Data Explainability via Principal Component Analysis, *by Shengkun Xie and Clare Chua-Chow* 25
48: iTLM: A Privacy Friendly Crowdsourcing Architecture for Intelligent Traffic Light Management, *by Christian Roth, Mirja Nitschke, Matthias Hörmann, Doğan Kesdoğan and Doğan Kesdoğan* 25
67: Predicting the Environment of a Neighborhood: A Use Case for France, *by Nelly Barret, Fabien Duchateau, Franck Favetta and Loïc Bonneval* 25

Doctoral Consortium on Data Management Technologies, Applications and Software Technologies (DCDATASOFT) (14:45 - 17:15)
Room 2 25
1: Social Modeling based on Event Detection: Research Outline, *by Aigerim Mussina and Sanzhar Aubakirov* 25
2: An Approach to Manage Software Documentation in Scrum Projects, *by Anbreen Javed* 26
3: Deep Reinforcement Learning Applied to a University Environment, *by Gulim Moldash* 26
4: Data Quality Evaluation using Probability Models, *by Allen O'Neill* 26
5: Industrial Relevance of Software Testing Education: Improving Software Testing Curricula, *by Bushra Hamid* 26
6: Big Data Analysis Systems in IoT Environments for Managing Privacy and Digital Identity: Pseudonymity, De-anonymization and the Right to Be Forgotten, *by Emanuela Podda* 26
7: Improving Monitoring and Evaluation of Software Engineering Curriculum: A Concept Map based Approach, *by Farkhanda Qamar* 27

Thursday Sessions: July 9

Session 7 (09:45 - 11:15)
Room 3: *Databases* 31
23: A Framework for Creating Policy-agnostic Programming Languages, *by Fabian Bruckner, Julia Pampus and Falk Howar* 31
30: Integrating Lightweight Compression Capabilities into Apache Arrow, *by Juliana Hildebrandt, Dirk Habich and Wolfgang Lehner* 31
25: Hybrid Multi-model Multi-platform (HM3P) Databases, *by Sven Groppe and Jinghua Groppe* 31

Poster Session 1 (11:15 - 12:15)
Room Posters DATA - DATA 31
11: Trading Desk Behavior Modeling via LSTM for Rogue Trading Fraud Detection, *by Marine Neyret, Jaouad Ouaggag and Cédric Allain* 31
31: Toward a New Quality Measurement Model for Big Data, *by Mandana Omidbakhsh and Olga Ormandjieva* 32
37: Reference Data Abstraction and Causal Relation based on Algebraic Expressions, *by Susumu Yamasaki and Mariko Sasakura* 32
42: A Conceptual Framework for a Flexible Data Analytics Network, *by Daniel Tebernum and Dustin Chabrowski* 32
44: Identification of Social Influence on Social Networks and Its Use in Recommender Systems: A Systematic Review, *by Lesly Camacho and Solange Alves-Souza* 32
46: Trust Profile based Trust Negotiation for the FHIR Standard, *by Eugene Sanzi and Steven Demurjian* 33
58: Initializing k-means Clustering, *by Christian Borgelt and Olha Yarikova* 33
71: Context-aware Retrieval and Classification: Design and Benefits, *by Kurt Englmeier* 33
Room Posters DATA - Open Communication 33
1: How to Develop Digital Products for Industrial Environments: The Data Science & Engineering Process in PLM, *by Peter Louis and Ralf Russ* 33

Keynote Lecture (12:15 - 13:15)
Room Plenary 1 33
From Data to the Press: Data Management for Journalism and Fact-Checking, *by Ioana Manolescu* 33

Closing Session (13:15 - 13:30)
Room Plenary 1 34

of speeds on treadmill. Classification was conducted for each speed independently with several feature extraction techniques applied. Subjects elicited gait asymmetry, yet ground reaction forces were more discriminative than joint angles. Walking speed affected gait symmetry with larger discrepancies registered at slower speeds; the highest F1 scores were noted at the slowest condition (decision trees: 87.35%, k-NN: 91.46%, SVMs: 88.88%, ANNs: 87.22%). None of the existing research has yet addressed ML-assisted assessment of gait symmetry across a range of walking speeds using both, kinetic and kinematic information. The proposed methodology was sufficiently sensitive to discern subtle deviations in healthy subjects, hence could facilitate an early diagnosis when anomalies in gait patterns emerge.

Paper #29

Improving Statistical Reporting Data Explainability via Principal Component Analysis

Shengkun Xie and Clare Chua-Chow

Global Management Studies, Ted Rogers School of Management, Ryerson University, Toronto, Canada

Keywords: Explainable Data Analysis, Data Visualization, Principal Component Analysis, Size of Loss Frequency, Business Analytics.

Abstract: The study of high dimensional data for decision-making is rapidly growing since it often leads to more accurate information that is needed to make reliable decision. To better understand the natural variation and the pattern of statistical reporting data, visualization and interpretability of data have been an on-going challenging problem, mainly, in the area of complex statistical data analysis. In this work, we propose an approach of dimension reduction and feature extraction using principal component analysis, in a novel way, for analyzing the statistical reporting data of auto insurance. We investigate the functionality of loss relative frequency, to the size-of-loss as well as the pattern and variability of extracted features, for a better understanding of the nature of auto insurance loss data. The proposed method helps improve the data explainability and gives an in-depth analysis of the overall pattern of the size-of-loss relative frequency. The findings in our study will help the insurance regulators to make a better rate filling decision in the auto insurance that would benefit both the insurers and their clients. It is also applicable to similar data analysis problems in other business applications.

Paper #48

iTLM: A Privacy Friendly Crowdsourcing Architecture for Intelligent Traffic Light Management

Christian Roth, Mirja Nitschke, Matthias Hörmann, Doğan Kesdoğan and Doğan Kesdoğan

University of Regensburg, Regensburg, Germany

Keywords: Traffic Light, V2X, Privacy, Attribute-Based-Credentials, Privacy-ABC System, Reinforcement Learning, Privacy-by-design.

Abstract: Vehicle-to-everything (V2X) interconnects participants in vehicular environments to exchange information. This enables a broad range of new opportunities. We propose a self learning traffic light system which uses crowdsoured information from vehicles in a privacy friendly manner to optimize the overall traffic flow. Our simulation, based on real world data, shows that the information gain vastly decreases waiting time at traffic lights eventually reducing CO2 emissions. A privacy analysis shows that our approach provides a significant level of k-anonymity even in low traffic scenarios.

Paper #67

Predicting the Environment of a Neighborhood: A Use Case for France

Nelly Barret[1], Fabien Duchateau[1], Franck Favetta[1] and Loïc Bonneval[2]

[1] *LIRIS UMR5205, Université de Lyon, UCBL, Lyon, France*

[2] *Centre Max Weber, Université de Lyon, France*

Keywords: Data Science, Machine Learning, Data Integration, Environment Prediction, Neighbourhood Study.

Abstract: Notion of neighbourhoods is critical in many applications such as social studies, cultural heritage management, urban planning or environment impact on health. Two main challenges deal with the definition and representation of this spatial concept and with the gathering of descriptive data on a large area (country). In this paper, we present a use case in the context of real estate search for representing French neighbourhoods in a uniform manner, using a few environment variables (e.g., building type, social class). Since it is not possible to manually classify all neighbourhoods, our objective is to automatically predict this new information.

Doctoral Consortium **DCDATASOFT**
14:45 - 17:15 **Room 2**
Doctoral Consortium on Data Science Technologies, Applications and Software Technologies

Paper #1

Social Modeling based on Event Detection: Research Outline

Aigerim Mussina and Sanzhar Aubakirov

Department of Computer Science, al-Farabi Kazakh National University, Almaty, Kazakhstan

Keywords: Event Detection, Event Association, Online Social Network, Machine Learning.

Abstract: Social networks already play a significant role in human's daily life. We are used to sharing almost everything with other users. Therefore social networks have become an arena of enormous opportunities to perform data analysis. Social media analytics applies to digital marketing, social opinion analysis, political situation monitoring, natural disaster notification. Various commercial and government organizations want to track, manage and predict information threads flow in digital space. It is possible to detect events on which people are reacting in every moment in online social networks. Event detection is a powerful data analysing process useful for social modeling.