

# Contents

## New Generation Data Warehouses Design

Evaluation of Data Warehouse Design Methodologies in the Context of Big Data . . . . .	3
<i>Francesco Di Tria, Ezio Lefons, and Filippo Tangorra</i>	
Optimal Task Ordering in Chain Data Flows: Exploring the Practicality of Non-scalable Solutions. . . . .	19
<i>Georgia Kougka and Anastasios Gounaris</i>	
Exploiting Mathematical Structures of Statistical Measures for Comparison of RDF Data Cubes. . . . .	33
<i>Claudia Diamantini, Domenico Potena, and Emanuele Storti</i>	
S2D: Shared Distributed Datasets, Storing Shared Data for Multiple and Massive Queries Optimization in a Distributed Data Warehouse . . . . .	42
<i>Rado Ratsimbazafy, Omar Boussaid, and Fadila Bentayeb</i>	

## Cloud and NoSQL Databases

Enforcing Privacy in Cloud Databases . . . . .	53
<i>Somayeh Sobati Moghadam, Jérôme Darmont, and Gérald Gavin</i>	
TARDIS: Optimal Execution of Scientific Workflows in Apache Spark . . . . .	74
<i>Daniel Gaspar, Fabio Porto, Reza Akbarinia, and Esther Pacitti</i>	
MDA-Based Approach for NoSQL Databases Modelling . . . . .	88
<i>Fatma Abdelhedi, Amal Ait Brahim, Faten Atigui, and Gilles Zurfluh</i>	

## Advanced Programming Paradigms

MiSeRe-Hadoop: A Large-Scale Robust Sequential Classification Rules Mining Framework. . . . .	105
<i>Elias Egho, Dominique Gay, Romain Trinquart, Marc Boullé, Nicolas Voisine, and Fabrice Clérot</i>	
An Efficient Map-Reduce Framework to Mine Periodic Frequent Patterns . . .	120
<i>Alampally Anirudh, R. Uday Kiran, P. Krishna Reddy, M. Toyoda, and Masaru Kitsuregawa</i>	

MapReduce-Based Complex Big Data Analytics over Uncertain and Imprecise Social Networks . . . . .	130
<i>Peter Braun, Alfredo Cuzzocrea, Fan Jiang, Carson Kai-Sang Leung, and Adam G.M. Pazdor</i>	

**Non-functional Requirements Satisfaction**

A Case for Abstract Cost Models for Distributed Execution of Analytics Operators . . . . .	149
<i>Rundong Li, Ningfang Mi, Mirek Riedewald, Yizhou Sun, and Yi Yao</i>	

Pre-processing and Indexing Techniques for Constellation Queries in Big Data. . . . .	164
<i>Amir Khatibi, Fabio Porto, Joao Guilherme Rittmeyer, Eduardo Ogasawara, Patrick Valduriez, and Dennis Shasha</i>	

A Lightweight Elastic Queue Middleware for Distributed Streaming Pipeline . . . . .	173
<i>Weiping Qu and Stefan Dessloch</i>	

Modeling Data Flow Execution in a Parallel Environment . . . . .	183
<i>Georgia Kougka, Anastasios Gounaris, and Ulf Leser</i>	

**Machine Learning**

Accelerating K-Means by Grouping Points Automatically . . . . .	199
<i>Qiao Yu and Bi-Ru Dai</i>	

A Machine Learning Trainable Model to Assess the Accuracy of Probabilistic Record Linkage . . . . .	214
<i>Robespierre Pita, Everton Mendonça, Sandra Reis, Marcos Barreto, and Spiros Denaxas</i>	

An Efficient Approach for Instance Selection . . . . .	228
<i>Joel Luís Carbonera</i>	

Search Result Personalization in Twitter Using Neural Word Embeddings . . .	244
<i>Sameendra Samarawickrama, Shanika Karunasekera, Aaron Harwood, and Ramamohanarao Kotagiri</i>	

Diverse Selection of Feature Subsets for Ensemble Regression . . . . .	259
<i>Arvind Kumar Shekar, Patricia Iglesias Sánchez, and Emmanuel Müller</i>	

K-Means Clustering Using Homomorphic Encryption and an Updatable Distance Matrix: Secure Third Party Data Clustering with Limited Data Owner Interaction . . . . .	274
<i>Nawal Almutairi, Frans Coenen, and Keith Dures</i>	

Reweighting Forest for Extreme Multi-label Classification . . . . .	286
<i>Zhun-Zheng Lin and Bi-Ru Dai</i>	

### **Social Media and Twitter Analysis**

A Relativistic Opinion Mining Approach to Detect Factual or Opinionated News Sources . . . . .	303
<i>Erhan Sezerer and Selma Tekir</i>	

A Reliability-Based Approach for Influence Maximization Using the Evidence Theory . . . . .	313
<i>Siwar Jendoubi and Arnaud Martin</i>	

Sentiment Analysis on Twitter to Improve Time Series Contextual Anomaly Detection for Detecting Stock Market Manipulation . . . . .	327
<i>Koosha Golmohammadi and Osmar R. Zaiane</i>	

Automatic Segmentation of Big Data of Patent Texts . . . . .	343
<i>Mustafa Sofean</i>	

### **Sentiment Analysis and User Influence**

Tag Me a Label with Multi-arm: Active Learning for Telugu Sentiment Analysis . . . . .	355
<i>Sandeep Sricharan Mukku, Subba Reddy Oota, and Radhika Mamidi</i>	

Belief Temporal Analysis of Expert Users: Case Study Stack Overflow . . . . .	368
<i>Dorra Attiaoui, Arnaud Martin, and Boutheina Ben Yaghlane</i>	

Leveraging Hierarchy and Community Structure for Determining Influencers in Networks . . . . .	383
<i>Sharanjit Kaur, Rakhi Saxena, and Vasudha Bhatnagar</i>	

Using Social Media for Word-of-Mouth Marketing . . . . .	391
<i>Nagendra Kumar, Yash Chandarana, Konjengbam Anand, and Manish Singh</i>	

### **Knowledge Discovery**

Knowledge Discovery of Complex Data Using Gaussian Mixture Models . . . . .	409
<i>Linfei Zhou, Wei Ye, Claudia Plant, and Christian Böhm</i>	

Optimized Mining of Potential Positive and Negative Association Rules . . . . .	424
<i>Parfait Bemarisika and André Totohasina</i>	

Extracting Non-redundant Correlated Purchase Behaviors  
by Utility Measure . . . . . 433  
*Wensheng Gan, Jerry Chun-Wei Lin, Philippe Fournier-Viger,  
and Han-Chieh Chao*

**Data Flow Management and Optimization**

Detecting Feature Interactions in Agricultural Trade Data  
Using a Deep Neural Network . . . . . 449  
*Jim O'Donoghue, Mark Roantree, and Andrew McCarren*

Air Quality Monitoring System and Benchmarking . . . . . 459  
*Xiufeng Liu and Per Sieverts Nielsen*

Electric Vehicle Charging Station Deployment for Minimizing  
Construction Cost . . . . . 471  
*Kai Li and Shuai Wang*

**Author Index** . . . . . 487

Big Data Analytics and Knowledge Discovery  
19th International Conference, DaWaK 2017, Lyon,  
France, August 28–31, 2017, Proceedings  
Bellatreche, L.; Chakravarthy, S. (Eds.)  
2017, XIV, 488 p. 137 illus., Softcover  
ISBN: 978-3-319-64282-6