


Lecture Notes in Business Information Processing

390

Series Editors

Wil van der Aalst 

RWTH Aachen University, Aachen, Germany

John Mylopoulos 

University of Trento, Trento, Italy

Michael Rosemann 

Queensland University of Technology, Brisbane, QLD, Australia

Michael J. Shaw

University of Illinois, Urbana-Champaign, IL, USA

Clemens Szyperski

Microsoft Research, Redmond, WA, USA

More information about this series at <http://www.springer.com/series/7911>


Ralf-Detlef Kutsche · Esteban Zimányi (Eds.)

Big Data Management and Analytics

9th European Summer School, eBISS 2019
Berlin, Germany, June 30 – July 5, 2019
Revised Selected Papers

Editors

Ralf-Detlef Kutsche
Technische Universität Berlin
Berlin, Germany

Esteban Zimányi 
Université libre de Bruxelles
Brussels, Belgium

ISSN 1865-1348 ISSN 1865-1356 (electronic)
Lecture Notes in Business Information Processing
ISBN 978-3-030-61626-7 ISBN 978-3-030-61627-4 (eBook)
<https://doi.org/10.1007/978-3-030-61627-4>

© 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

The 9th European Big Data Management and Analytics Summer School (eBISS 2019¹) took place in Berlin, Germany, in July 2019. Tutorials were given by renowned experts and covered advanced aspects of analytics and big data. This volume contains the lecture notes of the summer school.

The first chapter is devoted to actionable conformance checking. In the context of business processes, conformance checking aims at comparing a process model with an event log of the same process in order to assess whether the actual execution of a business process conforms to the model and vice versa. Although conformance checking has been receiving increasing attention in the last years, making the output of a conformance checking process actionable is still a real challenge. This chapter provides an introductory overview of the main techniques of the conformance checking field. In order to make it actionable, simple Python code snippets are provided to illustrate how an organization can start a conformance checking project on its own data. The chapter also provides pointers to open-source scripting libraries that can be used to make conformance checking and process mining actionable.

The second chapter provides an introduction to text analytics. It starts by presenting sources of textual data and the main challenges in text analysis. The chapter then surveys the various steps and methods involved in a typical processing pipeline. Since the steps to be realized heavily depend on the analytical task that is to be achieved, it is therefore necessary to identify the problem at hand and align the process accordingly. The chapter provides illustrative examples in each of the steps of the process and concludes by describing potential applications of text analytics, including sentiment analysis and automatic generation of content.

The third chapter is devoted to automated machine learning. Nowadays, machine learning techniques and algorithms are employed in almost every application domain to extract valuable knowledge from the massive amounts of data produced every day in our digital world. However, building a high-quality machine learning model is an iterative, complex, and time-consuming process that requires knowledge and experience. Given the continuous increase of the amount of digital data produced, it has been acknowledged that the number of data scientists cannot scale to address these challenges. The chapter gives an overview of the state-of-the-art tools and frameworks that have been proposed for tackling the challenges of machine learning automation. It concludes by discussing some research directions and open challenges required to achieve the vision and goals of automated machine learning.

The fourth chapter addresses the problem of determining how travel time can be computed from GPS data. The volume of GPS data collected from moving vehicles has increased significantly over the last years. Nowadays, it is possible to analyze the traffic on most of the road networks without installing roadside equipment. The chapter

¹ <http://cs.ulb.ac.be/conferences/ebiss2019/>.

presents a generic data model for travel time prediction that has a global scope and is applicable when GPS data and a road network graph is present. It defines several weather classes (dry, fog, rain, and snow) and shows their impact on travel time in various road categories (motorway, secondary, tertiary, and residential). The paper also analyzes other weather characteristics such as outside temperature and wind as well as regional differences. These results are presented in the context of a large-scale nationwide study performed in Denmark, where GPS data collected from 10,560 vehicles over five years is integrated with OpenStreetMap data and detailed weather information from the NOAA.

The last chapter introduces the Laplacian matrix as an efficient tool for addressing various tasks in machine learning. Many machine learning problems can be expressed by means of a graph with nodes representing training samples and edges representing the relationship between samples in terms of similarity, temporal proximity, or label information. As graphs can be represented by matrices, the chapter advocates the use of a Laplacian matrix, which allows us to assign each node a value that varies only slightly between strongly connected nodes and more between distant nodes. Such an assignment can be used to extract a useful feature representation, find a good embedding of data in a low dimensional space, or perform clustering on the original samples. The chapter starts by introducing the Laplacian matrix and then presents several algorithms designed around it for data visualization and feature extraction.

In addition to the lectures corresponding to the chapters described above, there were four additional lectures, as follows:

- Ralf-Detlef Kutsche from Technische Universität Berlin, Germany: Science Methodology
- Begüm Demir from Technische Universität Berlin, Germany: Deep Earth Query, Advances in Remote Sensing Image Characterization and Indexing from Massive Archives
- Aymen Cherif from Eura Nova, Belgium: Deep Learning, Current Applications and Future Trends
- Albert Bifet from Télécom ParisTech, France: Machine Learning for Data Streams

These lectures have no associated chapter in this volume.

As for the previous editions, eBISS joined forces with the Erasmus Mundus IT4BI-DC consortium and hosted its doctoral colloquium aiming at community building and promoting a corporate spirit among PhD candidates, advisors, and researchers of different organizations. The corresponding two sessions, each organized in two parallel tracks, included the following presentations:

- Judith Awiti, Evolving ETL workflows in a big data environment
- Jam Jahanzeb Behan, Statistical multidimensional data modeling based on Linked Open Data
- Moditha Hewasinghage, Physical design in document stores
- Mohsin Iqbal, Spatio-textual analytics
- Suela Isaj, Multi-source spatial entity linkage
- Nusrat Jahan Lisa, Database operations on top of complex system design
- Rediana Koci, A data-driven approach to prescribe Web API evolution

- Subba Lawan, Bitmap indexing for big data
- Shumet Tadesse Nigatu, Semi-automatic generation of data intensive APIs
- Olga Rybnytska, Prescriptive analytics for physical systems models

We would like to thank the attendants of the summer school for their active participation, as well as the speakers and their co-authors for the high quality of their contribution in a constant evolving and highly competitive domain. Finally, we would like to thank the external reviewers for their careful evaluation of the chapters.

June 2020

Ralf-Detlef Kutsche
Esteban Zimányi

Organization

The 9th European Big Data Management and Analytics Summer School (eBISS 2019) was organized by the Technische Universität Berlin, Germany, and the Department of Computer and Decision Engineering (CoDE) of the Université libre de Bruxelles, Belgium.

Program Committee

Alberto Abelló	Universitat Politècnica de Catalunya, BarcelonaTech, Spain
Ralf-Detlef Kutsche	Technische Universität Berlin, Germany
Boudewijn van Dongen	Technische Universiteit Eindhoven, The Netherlands
Nacéra Bennacer	CentraleSupélec, France
Esteban Zimányi	Université libre de Bruxelles, Belgium

External Referees

Judith Awiti	Université libre de Bruxelles, Belgium
Oscar Romero	Universitat Politècnica de Catalunya, BarcelonaTech, Spain
Mahmoud Sakr	Université libre de Bruxelles, Belgium
Alejandro Vaisman	Instituto Tecnológica de Buenos Aires, Argentina
Stijn Vansummen	Université libre de Bruxelles, Belgium
Robert Wrembel	Poznan University of Technology, Poland
Jianqiu Xu	Nanjing University of Aeronautics and Astronautics, China

Sponsorship and Support

Education, Audiovisual and Culture Executive Agency (EACEA)

Contents

Actionable Conformance Checking: From Intuitions to Code	1
<i>Josep Carmona, Matthias Weidlich, and Boudewijn van Dongen</i>	
Introduction to Text Analytics	25
<i>Agata Filipowska and Dominik Filipiak</i>	
Automated Machine Learning: Techniques and Frameworks	40
<i>Radwa Elshawy and Sherif Sakr</i>	
Travel-Time Computation Based on GPS Data	70
<i>Kristian Torp, Ove Andersen, and Christian Thomsen</i>	
Laplacian Matrix for Dimensionality Reduction and Clustering	93
<i>Laurenz Wiskott and Fabian Schönfeld</i>	
Author Index	121