Enterprise Big Data Engineering, Analytics, and Management

Martin Atzmueller
University of Kassel, Germany

Samia Oussena
University of West London, UK

Thomas Roth-Berghofer
University of West London, UK

A volume in the Advances in Business Information
Systems and Analytics (ABISA) Book Series

BUSINESS SCIENCE
Reference
An Imprint of IGI Global

Table of Contents

Foreword

Preface

Section 1
Foundational Issues

Chapter 1
How Big Does Big Data Need to Be?
Martin Stange, Leuphana University, Germany
Burkhardt Funk, Leuphana University, Germany

Chapter 2
Strategic Management of Data and Challenges for Organizations: Strategy Development and Business Value
Stephen Andrew Roberts, University of West London, UK
Bruce Laurie, University of West London, UK

Chapter 3
Data Stream Mining of Event and Complex Event Streams: A Survey of Existing and Future Technologies and Applications in Big Data
Chris Wrench, University of Reading, UK
Frederic Stahl, University of Reading, UK
Giuseppe Di Fatta, University of Reading, UK
Vidhyalakshmi Karthikeyan, BT, UK
DetlefD. Nauck, BT, UK

Chapter 4
Research Challenges in Big Data Analytics
Sivamathi Chokkalingam, Bharathiar University, India
Vijayarani S., Bharathiar University, India

Section 2
Tools and Methods

Chapter 5
Descriptive and Predictive Analytical Methods for Big Data 66
Sema A. Kalaian, Eastern Michigan University, USA
Rafa M. Kasim, Indiana Tech University, USA
Nabeel R. Kasim, University of Michigan, USA

Chapter 6
A Framework to Analyze Enterprise Social Network Data 84
Janine Viol Hacker, University of Erlangen – Nuremberg, Germany
Freimut Bodendorf University of Erlangen – Nuremberg, Germany
Pascal Lorenz, University of Haute Alsace, France

Chapter 7
Big Data Analytics Using Local Exceptionality Detection 108
Martin Atzmueller, University ofKassel, Germany
Dennis Mollenhauer, University ofKassel, Germany
Andreas Schmidt, University ofKassel, Germany

Chapter 8
Statistical Features for Extractive Automatic Text Summarization 126
Yogesh Kumar Meena, MNIT Jaipur, India
Dinesh Gopalani, MNIT Jaipur, India

Section 3
Case Studies and Application Areas

Chapter 9
Data Modeling and Knowledge Discovery in Process Industries 146
Benjamin Klopper, ABB Corporate Research Center, Germany
Marcel Dix, ABB Corporate Research Center; Germany
David Arnu, RapidMiner GmbH, Germany
Dikshith Siddapura, ABB Corporate Research Center, Germany

Chapter 10
Data Preparation for Big Data Analytics: Methods and Experiences 157
Andreas Schmidt, University ofKassel, Germany
Martin Atzmueller, University ofKassel, Germany
Martin Hollender, ABB Corporate Research Center, Germany

Chapter 11
Semantification of Large Corpora of Technical Documentation 171
Sebastian Furth, denkbares GmbH, Germany
Joachim Baumeister, denkbares GmbH, Germany & University of Wiirzburg, Germany

Chapter 12
Application of Complex Event Processing Techniques to Big Data Related to Healthcare: A Systematic Literature Review of Case Studies 201
Fehmida Mohamedali, University of West London, UK
Samia Oussena, University of West London, UK

Chapter 13
Using Big Data in Collaborative Learning 221
Liz Sokolowski, University of West London, UK
Samia Oussena, University of West London, UK

Compilation of References *238*

About the Contributors *264*

Index ***270***

Detailed Table of Contents

Foreword **xiv**

Preface xvi

Section 1
Foundational Issues

Chapter 1
How Big Does Big Data Need to Be? 1
Martin Stange, Leuphana University, Germany
Burkhardt Funk, Leuphana University, Germany

Collecting and storingof as manydata aspossible iscommon practice in many companies these days.To reduce costs of collecting and storing data that is not relevant, it is important to define which analytical questions are to be answered and how much data is needed to answer these questions. In this chapter, a process to define an optimal sampling size is proposed. Based on benefit/cost considerations, the authors show how to find the sample size that maximizes the utility of predictive analytics. By applying the proposed process to a case study is shown that only a very small fraction of the available data set is needed to make accurate predictions.

Chapter 2
Strategic Management of Data and Challenges for Organizations: Strategy Development and Business Value 13
Stephen Andrew Roberts, University of West London, UK
Bruce Laurie, University of West London, UK

Public, organizational and personal data has never been so much in the forefront of discussion and attention as at the present time. The term 'Big Data' (BD) has become part of public discourse, in the press, broadcast media and on the web. Most people in the wider public have very little idea of what it is and what it means but anyone who gives it a thought will see it as contemporary and relevant to life as much as to business.This paper is directed towards the perspectivesof people working in,managing and developingorganizations which arededicated to fulfilling theirrespective purposes. All organizationsneed to understand their strategic purpose and to develop strategies and tactical responses accordingly. The organizations' purpose and the frameworks and resources adopted are part of its quest for achievement which creates value and worth. BD is a potential and actual source of value.

Chapter 3
Data Stream Mining of Event and Complex Event Streams: A Survey of Existing and Future Technologies and Applications in Big Data 24
Chris Wrench, University of Reading, UK
Frederic Stahl, University of Reading, UK
Giuseppe Di Fatta, University of Reading, UK
Vidhyalakshmi Karthikeyan, BT, UK
DetlefD. Nauck, BT, UK

Complex Event Processing has been a growing field for the last ten years. It has seen the development of a number of methods and tools to aid in the processing of event streams and clouds though it has also been troubled by the lack of a cohesive definition. This paper aims to layout the technologies surrounding CEP and to distinguish it from the closely related field of Event Stream Processing. It also aims to explore the work done to apply Data Mining Techniques to both of these fields. An outline of stream processing technologies is laid out including the Data Stream Mining techniques that have been adapted for CEP.

Chapter 4
Research Challenges in Big Data Analytics 48
Sivamathi Chokkalingam, Bharathiar University, India
Vijayarani S,, Bharathiar University, India

The term Big Data refers to large-scale information management and analysis technologies that exceed the capability of traditional data processing technologies. Big Data is differentiated from traditional technologies in three ways: volume, velocity and variety of data. Big data analytics is the process of analyzing large data sets which contains a variety of data types to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful business information. Since Big Data is new emerging field, there is a need for development of new technologies and algorithms for handling big data. The main objective of this paper is to provide knowledge about various research challenges of Big Data analytics. A brief overview of various types of Big Data analytics is discussed in this paper. For each analytics, the paper describes process steps and tools. A banking application is given for each analytics. Some of research challenges and possible solutions for those challenges of big data analytics are also discussed.

Section 2
Tools and Methods

Chapter 5
Descriptive and Predictive Analytical Methods for Big Data 66
Sema A. Kalaian, Eastern Michigan University, USA
Rafa M. Kasim, Indiana Tech University, USA
Nabeel R. Kasim, University of Michigan, USA

Data analytics and modeling are powerful analytical tools for knowledge discovery through examining and capturing the complex and hidden relationships and patterns among the quantitative variables in the existing massive structured Big Data in efforts to predict future enterprise performance. The main purpose of this chapter is to present a conceptual and practical overview of some of the basic and advanced analytical tools for analyzing structured Big Data. The chapter covers descriptive and predictive analytical

methods. Descriptive analytical tools such as mean, median, mode, variance, standard deviation, and data visualization methods (e.g., histograms, line charts) are covered. Predictive analytical tools for analyzing Big Data such as correlation, simple- and multiple- linear regression are also covered in the chapter.

Chapter 6
A Framework to Analyze Enterprise Social Network Data 84
Janine Viol Hacker, University of Erlangen – Nuremberg, Germany
Freimut Bodendorf University of Erlangen – Nuremberg, Germany
Pascal Lorenz, University of Haute Alsace, France

Enterprise Social Networks have a similar set of functionalities as social networking sites but are run as closed applications within a company's intranet. Interacting and communicating on the Enterprise Social Networks, the users, i.e. a company's employees, leave digital traces. The resulting digital record stored in the platform's back end bears great potential for enterprise big data engineering, analytics, and management. This book chapter provides an overview of research in the area of Enterprise Social Networks and categorizes Enterprise Social Network data based on typical functionalities of these platforms. It introduces exemplary metrics as well as a process for the analysis of ESN data. The resulting framework for the analysis of Enterprise Social Network data can serve as a guideline for researchers in the area of Enterprise Social Network analytics and companies interested in analyzing the data stored in the application's back end.

Chapter 7
Big Data Analytics Using Local Exceptionality Detection 108
Martin Atzmueller, University of Kassel, Germany
Dennis Mollenhauer, University ofKassel, Germany
Andreas Schmidt, University ofKassel, Germany

Large-scale data processing is one of the key challenges concerning many application domains, especially considering ubiquitous and big data. In these contexts, subgroup discovery provides both a flexible data analysis and knowledge discovery method. Subgroup discovery and pattern mining are important descriptive data mining tasks. They can be applied, for example, in order to obtain an overview on the relations in the data, for automatic hypotheses generation, and for a number of knowledge discovery applications. This chapter presents the novel SD-MapR algorithmic framework for large-scale local exceptionality detection implemented using subgroup discovery on the Map/Reduce framework. We describe the basic algorithm in detail and provide an experimental evaluation using several real-world datasets. We tackle two algorithmic variants focusing on simple and more complex target concepts, i.e., presenting an implementation of exceptional model mining on large attributed graphs. The results of our evaluation show the scalability of the presented approach for large data sets.

Chapter 8
Statistical Features for Extractive Automatic Text Summarization 126
Yogesh Kumar Meena, MNIT Jaipur, India
Dinesh Gopalani, MNIT Jaipur, India

Automatic Text Summarization (ATS) enables users to save their precious time to retrieve their relevant information need while searching voluminous big data. Text summaries are sensitive to scoring methods, as most of the methods requires to weight features for sentence scoring. In this chapter, various statistical

features proposed by researchers for extractive automatic text summarization are explored. Features that perform well are termed as best features using ROUGE evaluation measures and used for creating feature combinations. After that, best performing feature combinations are identified. Performance evaluation of best performing feature combinations on short, medium and large size documents is also conducted using same ROUGE performance measures.

Section 3
Case Studies and Application Areas

Chapter 9
Data Modeling and Knowledge Discovery in Process Industries 146
Benjamin Klopper, ABB Corporate Research Center, Germany
Marcel Dix, ABB Corporate Research Center, Germany
David Amu, RapidMiner GmbH, Germany
Dikshith Siddapura, ABB Corporate Research Center, Germany

Dispersed data sources, incompatible data formats and a lack of non-ambiguous and machine readable meta-data is a major obstacle in data analytics and data mining projects in process industries. Often, meta-information is only available in unstructured format optimized for human consumption. This contribution outlines a feasible methodology for organizing historical datasets extracted from process plants in a big data platform for the purpose of analytics and machine learning model building in an industrial big data analytics project.

Chapter 10
Data Preparation for Big Data Analytics: Methods and Experiences 157
Andreas Schmidt, University ofKassel, Germany
Martin Atzmueller, University of Kassel, Germany
Martin Hollender, ABB Corporate Research Center, Germany

This chapter provides an overview of methods for preprocessing structured and unstructured data in the scope of Big Data. Specifically, this chapter summarizes according methods in the context of a real-world dataset in a petro-chemical production setting. The chapter describes state-of-the-art methods for data preparation for Big Data Analytics. Furthermore, the chapter discusses experiences and first insights in a specific project setting with respect to a real-world case study. Furthermore, interesting directions for future research are outlined.

Chapter 11
Semantification of Large Corpora of Technical Documentation 171
Sebastian Furth, denkbares GmbH, Germany
Joachim Baumeister, denkbares GmbH, Germany & University of Wiirzburg, Germany

The complexity of machines has grown dramatically in the past years. Today, they are built as a complex functional network of mechanics, electronics, and hydraulics. Thus, the technical documentation became a fundamental source for service technicians in their daily work. The technicians need fast and focused access methods to handle the massive volumes of documentation. For this reason, semantic search emerged as the new system paradigm for the presentation of technical documentation. However, the existent large corpora of legacy documentation are usually not semantically prepared. This fact creates

an invincible gap between new technological opportunities and the actual data quality at companies. This chapter presents a novel and comprehensive approach for the semantification of large volumes of legacy technical documents. The approach espescially tackles the veracity and variety existent in technical documentation and makes explicit use of their typical characteristics. The experiences with the implementation and the learned benefits are discussed in industrial case studies.

Chapter 12
Application of Complex Event Processing Techniques to Big Data Related to Healthcare: A Systematic Literature Review of Case Studies 201
Fehmida Mohamedali, University of West London, UK
Samia Oussena, University of West London, UK

Healthcare is a growth area for event processing applications. Computers and information systems have been used for collecting patient data in health care for over fifty years. However, progress towards a unified health care delivery system in the UK has been slow. Big Data, the Internet of Things (IoT) and Complex Event Processing (CEP) have the potential not only to deal with treatment areas of healthcare domain but also to redefine healthcare services. This study is intended to provide a broad overview of where in the health sector, the application of CEP is most used, the data sources that contribute to it and the types of event processing languages and techniques implemented. By systematic review of existing literature on the application of CEP techniques in Healthcare, a number of use cases have been identified to provide a detailed analysis of the most common used case(s), common data sources in use and highlight CEP query language types and techniques that have been considered.

Chapter 13
Using Big Data in Collaborative Learning 221
Liz Sokolowski, University of West London, UK
Samia Oussena, University of West London, UK

Big data emerged as a dominant trend for predictive analytics in many areas of industry and commerce. The study aimed to explore whether similar trends and benefits have been observed in the area of collaborative learning. The study looked at the domains in which the collaborative learning was undertaken. The results of the review found that the majority of the studies were undertaken in the Computing and Engineering or Social Science domains, primarily at undergraduate level. The results indicate that the data collection focus is on interaction data to describe the process of the collaboration itself, rather than on the end product of the collaboration. The student interaction data came from various sources, but with a notable concentration on data obtained from discussion forums and virtual learning environment logs. The review highlighted some challenges; the noisy nature of this data and the need for manual pre-processing of textual data currently renders much of it unsuitable for automated 'big data' analytical approaches.

Compilation of References *238*

About the Contributors *264*

Index *270*