DATA SCIENCE WITH SEMANTIC TECHNOLOGIES

THEORY, PRACTICE AND APPLICATION

Edited by
Archana Patel
Narayan C. Debnath
Bharat Bhusan

Scrivener Publishing

WILEY

Table of Contents

List of Tables

List of Illustrations

Chapter 5

Chapter 6

Scrivener Publishing
100 Cummings Center, Suite 541J
Beverly, MA 01915-6106
Publishers at Scrivener
Martin Scrivener (martin@scrivenerpublishing.com)
Phillip Carmical (pcarmical@scrivenerpublishing.com)

Data Science with Semantic Technologies

Theory, Practice, and Application

Edited by

Archana Patel
Department of Software Engineering, School of Computing and Information Technology, Eastern International University, Vietnam

Narayan C. Debnath
School of Computing and Department of Computer Science and Engineering, School of Engineering Vietnam

and

Bharat Bhusan
Technology, Sharda University Information Technology, India

WILEY

This edition first published 2022 by John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA and Scrivener Publishing LLC, 100 Cummings Center, Suite 541J, Beverly, MA 01915, USA © 2022 Scrivener Publishing LLC
For more information about Scrivener publications please visit www.scrivenerpublishing.com.

Wiley Global Headquarters
111 River Street, Hoboken, NJ 07030, USA

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Limit of Liability/Disclaimer of Warranty
While the publisher and authors have used their best efforts in preparing this work, they make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation any implied warranties of merchant-ability or fitness for a particular purpose. No warranty may be created or extended by sales representatives, written sales materials, or promotional statements for this work. The fact that an organization, website, or product is referred to in this work as a citation and/or potential source of further information does not mean that the publisher and authors endorse the information or services the organization, website, or product may provide or recommendations it may make. This work is sold with the understanding that the publisher is not engaged in rendering professional services. The advice and strategies contained herein may not be suitable for your situation. You should consult with a specialist where appropriate. Neither the publisher nor authors shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages. Further, readers should be aware that websites listed in this work may have changed or disappeared between when this work was written and when it is read.

Library of Congress Cataloging-in-Publication Data

ISBN 9781119864981

Cover image: PixaBay.Com
Cover design by Russell Richardson

Set in size of 11pt and Minion Pro by Manila Typesetting Company, Makati, Philippines

Printed in the USA

10 9 8 7 6 5 4 3 2 1

Preface

Data Science is an invaluable resource that deals with vast volumes of data using modern tools and techniques to find unseen patterns, derive meaningful information, and make business decisions. To create intelligence in data science, it becomes necessary to utilize the semantic technologies which allow machine-readable representation of data. This intelligence uniquely identifies and connects data with common business terms, and also enables users to communicate with data. Instead of structuring the data, semantic technologies help users to understand the meaning of the data by using the concepts of semantics, ontology, OWL, linked data, and knowledge graphs. These technologies assist organizations in understanding all of the stored data, adding value to it, and enabling insights that were not available before. Organizations are also using semantic technologies to unearth precious nuggets of information from vast volumes of data and to enable more flexible use of data. These technologies can deal with the existing problems of data scientists and help them in making better decisions for any organization. All of these needs are part of a focused shift towards utilization of semantic technologies in data science that provide knowledge along with the ability to understand, reason, plan, and learn with existing and new data sets. These technologies also generate expected, reproducible, user-desired results.

This book aims to provide a roadmap for the deployment of semantic technologies in the field of data science. Moreover, it highlights how data science enables the user to create intelligence through these technologies by exploring the opportunities and eradicating the challenges

in current and future time frames. It can serve as an important guide to applications of data science with semantic technologies for the upcoming generation and thus is a unique resource for scholars, researchers, professionals and practitioners in this field. Following is a brief description of the subjects covered in the 15 chapters of the book.

– Chapter 1 provides a brief introduction to data science. It addresses various aspects of data science such as what a data scientist does and why data science is in demand; the history of data science and how it differs from business intelligence; the life cycle of data science and data science components; why data science is important; the challenges of data science; the tools used for data science; and the benefits and applications of data science.

– Chapter 2 provides an overview of the top 10 tools and applications that should be of interest to any data scientist. Its objective includes, but is not limited to, realizing the use of Python in developing solutions to data science tasks; recognizing the use of R language as an open-source data science provider; traveling around the SQL to provide structured models for data science projects; navigating through data analytics and statistics using Excel; and using D3.js scripting tools for data visualization. Also, practical examples/case studies are provided on data visualization, data analytics, regression, forecasting, and outlier detection.

– Chapter 3 presents the use of data modeling for data science, revealing the possibility of a new side of the data. The chapter covers different types of data (unstructured data, semi-structured data, structured data, hybrid (un/semi)-structured data and big data) and data model design.

– Chapter 4 shows data management by considering language based on the novelty view of data. The chapter focuses on data life cycle, data distribution and CAP theorem.

– Chapter 5 presents the role of data science in healthcare. There are several fields in the healthcare sector, such as predictive modeling, genetics, etc., which make use of data science for diagnosis and drug discovery, thereby increasing usability of precision medicine.

– Chapter 6 provides a new balanced binary search tree that generates two kinds of nodes: simple and class nodes. Two advantages make the new structure attractive. First, it subsumes the most popular data structures of AVL and Red-Black trees. Second, it proposes other unknown balanced binary search trees in which we can adjust the maximal height of paths between 1.44 lg(n) and 2 lg(n), where n is the number of nodes in the tree and lg the base-two logarithm.

– Chapter 7 shows the study of machine learning and deep learning algorithms with detailed and analytical comparisons, which help new and inexperienced medical professionals or researchers in the medical field. The proposed machine learning model has an accurate algorithm that works with rich healthcare data, a high-dimensional data handling system, and an intelligent framework that uses different data sources to predict heart disease. This chapter uses an ensemble-based deep learning model with optimal feature selection to improve accuracy.

– Chapter 8 presents an IoT-based automated fire control system in a mining area which will help to protect many valuable lives whenever an accident occurs due to fire. In the experimental application, different types of sensors for

temperature, moisture, and gas are used to sense the different environmental data.

– Chapter 9 offers an aspect identification method for sentiment sentences in review documents. The chapter describes two key tasks—one for extracting significant features from the reviews and another for identification of degrees of product reviews.

– Chapter 10 shows the research that paved the way for semantic technology. It then describes each of the semantic pillars with examples and explanations of the business value of each technology.

– Chapter 11 describes the ontology evaluation tools and then focuses on the evaluation of the security ontologies. The existing ontology evaluation tools are classified under two categories; namely, domain-dependent ontology evaluation tools and domain-independent ontology evaluation tools. The evaluation of security ontology assesses the quality of ontology among the available ontologies.

– Chapter 12 discusses the main concepts of health data, data science, health data science, examples of the application of health data science and related challenges. In addition, it also highlights the application of semantic technologies in health data science and the challenges that lie ahead of using these technologies.

– Chapter 13 proposes an original hybrid optimization approach based on two different mixed integer programming statements. The first statement is based on minimizing the sum of pairwise distances between all objects (PDC clustering), while the second statement is based on minimizing the total distance from objects to cluster centers (CC clustering). Computational experiments showed that the hybrid method developed for solving the

clustering problem combines the advantages of both approaches—the speed of the k-means method and the accuracy of PDC clustering—which makes it possible to get rid of the main drawback of the k-means, namely, the lack of guaranteed determining of the global optimum.

– Chapter 14 uses a model for the analysis of time series data which highly depend on the novel coronavirus 2019. This model predicts the future trend of confirmed, recovered, active, and death cases based on the available data from January 22, 2020 to May 29, 2021. The present model predicted the spread of COVID-19 for a future period of 30 days. The RMSE, MSE, MAE, and MdAPE metrics are used for the model evaluation.

– Chapter 15 focuses on systems that incorporated real-world data utilized by actual users. It first describes a new methodology for the survey and then covers the various domains where semantic technology can be applied and some of the most impressive systems developed in each domain.

Finally, the editors would like to sincerely thank all the authors and reviewers who contributed their time and expertise for the successful completion of this book. The editors also appreciate the support given by Scrivener Publishing, which allowed us to compile the first edition of this book.

The Editors
Archana Patel
Narayan C. Debnath
Bharat Bhusan
June 2022

1
A Brief Introduction and Importance of Data Science

Karthika N.[1]*, Sheela J.[1] and Janet B.[2]
[1]*Department of SCOPE, VIT-AP University, Amaravati, Andhra Pradesh, India*
[2]*Department of Computer Applications, National Institute of Technology, Tiruchirappalli, India*

Abstract

Data is very important component of any organization. According to International Data Corporation, by 2025, global data will reach to 175 zettabytes. They need data to help them make careful decisions in business. Data is worthless until it is transformed into valuable data. Data science plays a vital role in processing and interpreting data. It focuses on the analysis and management of data too. It is concerned with obtaining useful information from large datasets. It is frequently applied in a wide range of industries, including healthcare, marketing, banking, finance, policy work, and more. This enables companies to make informed decisions around growth, optimization, and performance. In this brief monograph, we address following questions.

What is data science and what does a data scientist do? Why data science is in demand? History of data science, how data science differs from business intelligence? The lifecycle of data science, data science components, why data science is important? Challenges of data science, tools used for data science, benefits and applications of data science.

Keywords: Data science, history, lifecycle, components, tools

1.1 What is Data Science? What Does a Data Scientist Do?

Data is very important component of any organization. According to International Data Corporation, by 2025, global data will reach to 175 zettabytes. They need data to help them make careful decisions in business. Data is worthless until it is transformed into valuable data. Data science plays a vital role in processing and interpreting

data. It focuses on the analysis and management of data too. It is concerned with obtaining useful information from large datasets. It is frequently applied in a wide range of industries, including healthcare, marketing, banking, finance, policy work, and more. This enables companies to make informed decisions around growth, optimization, and performance. In nutshell, Data science is an integrative strategy for deriving actionable insights from today's organizations' massive and ever-increasing data sets. Preparing data for analysis and processing, performing advanced data analysis, and presenting the findings to expose trends and allow stakeholders to make educated decisions are all part of data science [1, 2]. Data science experts are both well-known, data-driven individuals with advanced technical capabilities who can construct complicated quantitative algorithms to organize and interpret huge amounts of data in order to address questions and drive strategy in their company. This is combined with the communication and leadership skills required to provide tangible results to numerous stake-holders throughout a company or organization. Data scientists must be inquisitive and results-driven, with great industry-specific expertise and communication abilities that enable them to convey highly technical outcomes to non-technical colleagues. To create and analyze algorithms, they have a solid quantitative background in statistics and linear algebra, as well as programming experience with a focus on data warehousing, mining, and modeling [3].

1.2 Why Data Science is in Demand?

Data science is the branch of science concerned with the discovery, analysis, modeling, and extraction of useful information which has become a buzz in a lot of companies. Firms are increasingly aware that they have been sitting on

data treasure mines the priority with which this data must be analyzed, and ROI generated is obvious. We look at the most important reasons that data science professions are in high demand [4].

- **Data Organization**

  During the mid-2000s IT boom, the emphasis was on "lifting and shifting" offline business operations into automated computer systems. Digital content generation, transactional data processing, and data log streams have all been consistent throughout the last two decades. This indicates that every company now has a plethora of information that it believes can really be valuable but does not know how to use. This is apparent in Glassdoor's recent analysis, which identifies the 50 greatest jobs in modern era.

- **Scarcity of Trained Manpower**

  According to a McKinsey Global Institute study, by 2018, the United States will be short 190,000 data scientists, 1.5 million managers, including analysts who would properly comprehend and make judgments based on Big Data. The need is particularly great in India, where the tools and techniques are available but there are not enough qualified people. Data scientists, who can perform analytics, and analytics consultants, who can analyze and apply data, are two sorts of talent shortages, according to Srikanth Velamakanni, co-founder and CEO of Fractal Analytics. The supply of talent in these fields, particularly data scientists, is extremely limited, and the demand is enormous."

- **The Pay Is Outstanding**

  A data science position is currently one of the highest paying in the market. The national average income for a data scientist/analyst in the United States, according