COMPUTER ENGINEERING SERIES

METAHEURISTICS SET

Volume 5

Metaheuristics for Big Data

Clarisse Dhaenens

Laetitia Jourdan

ISTE

WILEY

Metaheuristics Set

coordinated by
Nicolas Monmarché and Patrick Siarry

Volume 5

Metaheuristics for Big Data

Clarisse Dhaenens
Laetitia Jourdan

ISTE

WILEY

First published 2016 in Great Britain and the United States by ISTE Ltd and John Wiley & Sons, Inc.

ISTE Ltd
27-37 St George's Road
London SW19 4EU
UK

www.iste.co.uk

John Wiley & Sons, Inc.
111 River Street
Hoboken, NJ 07030
USA

www.wiley.com

Library of Congress Control Number: 2016944993

British Library Cataloguing-in-Publication Data
A CIP record for this book is available from the British Library
ISBN 978-1-84821-806-2

Metaheuristics for Big Data

Acknowledgments

This book is an overview of metaheuristics for Big Data. Hence it is based on a large literature review conducted by the authors in the Laboratory CRIStAL (Research Center in Computer Science, Signal and Automatics), University of Lille and CNRS, France and in the Lille Nord Europe Research Center of INRIA (French National Institute for Computer Science and Applied Mathematics) between 2000 and the present. We are grateful to our former and current PhD students and colleagues for all the work they have done together with us that has led to this book.

We are particularly grateful to Aymeric Blot, Fanny Dufossé, Lucien Mousin and Maxence Vandromme who read and corrected the first versions of this book. A special word of gratitude to Marie-Elénore Marmion who read carefully and commented on several chapters.

We would like to thank Nicolas Monmarché and Patrick Siarry for their proposal to write this book and for their patience! Sorry for the time we took.

Finally, we would like to thank our families for their support and love.

Clarisse DHAENENS and Laetitia JOURDAN

Introduction

Big Data: a buzzword or a real challenge?

Both answers are suitable. On the one hand, the term *Big Data* has not yet been well defined, although several attempts have been made to give it a definition. Indeed, the term *Big Data* does not have the same meaning according to the person who uses it. It could be seen as a buzzword: *everyone talks about Big Data but no one really manipulates it.*

On the other hand, the characteristics of *Big Data*, often reduced to the three "Vs" – volume, variety and velocity – introduce plenty of new technological challenges at different phases of the Big Data process. These phases are presented in a very simple way in Figure I.1.

Starting from the generation of data, its storage and management, analyses can be made to help decision-making. This process may be reiterated if additional information is required. At each phase, some important challenges arise.

Indeed, during the generation and capture of data, some challenges may be related to technological aspects that are linked to the acquisition of real-time data, for example. However, at this phase, challenges are also related to the identification of meaningful data.

The storage and management phase leads to two critical challenges: first, the infrastructures for the storage of data and its transportation; second, conceptual models to provide well-formed available data that may be used for analysis.

Figure I.1. *Main phases of a Big Data process*

Then, the analysis phase has its own challenges, with the manipulation of heterogeneous massive data. In particular, when considering the knowledge extraction, in which unknown patterns have to be discovered, analysis may be very complex due to the nature of data manipulated. This is at the heart of data mining. A way to address data mining problems is to model them as optimization problems. In the context of Big Data, most of these problems are large-scale ones. Hence metaheuristics seem to be good candidates to tackle them. However, as we will see in the following, metaheuristics are suitable not only to address the large size of the problem, but also to deal with other aspects of Big Data, such as variety and velocity.

The aim of this book is to present how *metaheuristics can provide answers to some of the challenges induced by the Big Data context* and particularly within the data analytics phase.

This book is composed of three parts. The first part is an introductory part consisting of three chapters. The aim of this part is to provide the reader with elements to understand the following aspects.

Chapter 1, *Optimization and Big Data*, provides elements to understand the main issues led by the *Big Data* context. It then reveals what characterizes *Big Data* and focuses on the analysis phase and, more precisely, on the data

mining task. This chapter indicates how data mining problems may be seen as combinatorial optimization problems and justifies the use of metaheuristics to address some of these problems. A section is also dedicated to the performance evaluation of algorithms, as in data mining, a specific protocol has to be followed.

Chapter 2 presents an *introduction to metaheuristics*, to make this book self-contained. First, common concepts of metaheuristics are presented and then the most widely known metaheuristics are described with a distinction between single solution-based and population-based methods. A section is also dedicated to multi-objective metaheuristics, as many of them have been proposed to deal with data mining problems.

Chapter 3 provides indications on *parallel optimization* and the way metaheuristics may be parallelized to tackle very large size problems. As it will be revealed, the parallelization is considered not only to deal with large problems, but also to provide better quality solutions.

The second part, composed of the following four chapters, is the heart of the book. Each of these chapters details a data mining task and indicates how metaheuristics can be used to deal with it.

Chapter 4 begins the second part of the book and is dedicated to *clustering*. This chapter first presents the clustering task that aims to group similar objects and some of the classical approaches to solve it. Then, the chapter provides indications on the modeling of the clustering task as an optimization problem and focuses on the quality measures that are commonly used, on the interest of a multi-objective resolution approach and on the representation of a solution in metaheuristics. An overview of multi-objective methods is then proposed. The chapter ends with a specific and difficult point in the clustering task: how the estimation of the quality of a clustering solution and its validation can be done.

Chapter 5 deals with *association rules*. It first describes the corresponding data mining task and the classical approach: the *a priori* algorithm. Then, the chapter indicates how this task may be modeled as an optimization task and then focuses on metaheuristics proposed to deal with this task. It differentiates the metaheuristics according to the type of rules that are considered: categorical association rules, quantitative association rules or

fuzzy association rules. A general table summarizes the most important works of the literature.

Chapter 6 is dedicated to *supervised classification*. Data mining is of great importance as it allows the prediction of the class of a new observation regarding information from observations whose classes are known. The chapter first gives a description of the classification task and briefly presents standard classification methods. Then, an optimization perspective of some of these standard methods is presented as well as the use of metaheuristics to optimize some of them. The last part of the chapter is dedicated to the use of metaheuristics for the search of classification rules, viewed as a special case of association rules.

Chapter 7 deals with *feature selection for classification* that aims to reduce the number of attributes and to improve the classification performance. The chapter uses several notions that are presented in Chapter 6 on classification. After a presentation of generalities on feature selection, the chapter gives its modeling as an optimization problem. Different representations of solutions and their associated search mechanisms are then presented. An overview of metaheuristics for feature selection is finally proposed.

Finally, *the last part* is composed of a single chapter (Chapter 8) which presents *frameworks* dedicated to data mining and/or metaheuristics. A short comparative survey is provided for each kind of framework.

Browsing the different chapters, the reader will have an overview of the way metaheuristics have been applied so far to tackle problems that are present in the Big Data context, with a focus on the data mining part, which provides the optimization community with many challenging opportunities of applications.

Contents

1

Optimization and Big Data

The term *Big Data* refers to vast amounts of information that come from different sources. Hence *Big Data* refers not only to this huge data volume but also to the diversity of data types, delivered at various speeds and frequencies. This chapter attempts to provide definitions of *Big Data*, the main challenges induced by this context, and focuses on Big Data analytics.

1.1. Context of Big Data

As depicted in Figure 1.1, the evolution of Google requests on the term "Big Data" has grown exponentially since 2011.

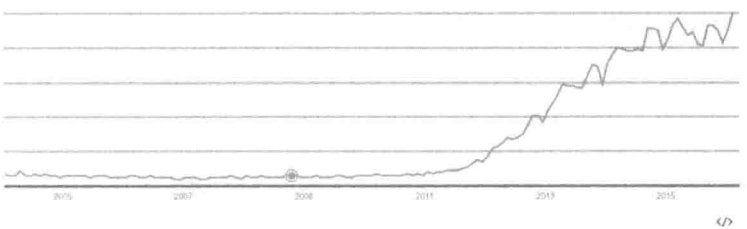

Figure 1.1. *Evolution of Google requests for "Big Data" (Google source)*

How can we explain the increasing interest in this subject? Some responses may be formulated, when we know that everyday 2.5 quintillion bytes of data are generated – such that 90% of the data in the world today

have been created in the last two years. These data come from everywhere, depending on the industry and organization: sensors are used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records and cellphone GPS signals, to name but a few [IBM 16b]. Such data are recorded, stored and analyzed.

1.1.1. *Examples of situations*

Big Data appears in a lot of situations where large amounts of complex data are generated. Each situation presents challenges to handle. We may cite some examples of such situations:

– *Social networks*: the quantity of data generated in social networks is huge. Indeed, monthly estimations indicate that 12 billion tweets are sent by about 200 million active users, 4 billion hours of video are watched on YouTube and 30 billion pieces of content are shared on Facebook [IBM 16a]. Moreover, such data are of different formats/types.

– *Traffic management*: in the context of creation of smart cities, the traffic within cities is an important issue. This becomes feasible, as the widespread adoption in recent years of technologies such as smartphones, smartcards and various sensors has made it possible to collect, store and visualize information on urban activities such as people and traffic flows. However, this also represents a huge amount of data collected that need to be managed.

– *Healthcare:* in 2011, the global size of data in healthcare was estimated as 150 exabytes. Such data are unique and difficult to deal with because: 1) data are in multiple places (different source systems in different formats including text as well as images); 2) data are structured and unstructured; 3) data may be inconsistent (they may have different definitions according to the person in charge of filling data); 4) data are complex (it is difficult to identify standard processes); 5) data are subject to regulatory requirement changes [LES 16].

– *Genomic studies:* with the rapid progress of DNA sequencing techniques that now allows us to identify more than 1 million SNPs (genetic variations), large-scale genome-wide association studies (GWAS) have become practical. The aim is to track genetic variations that may, for example, explain genetic susceptibility for a disease. In their analysis on the new challenges induced by these new massive data, Moore *et al.* first indicate the necessity of the development on new biostatistical methods for quality control, imputation and

analysis issues [MOO 10]. They also indicate the challenge of recognizing the complexity of the genotype–phenotype relationship that is characterized by significant heterogeneity.

In all these contexts, the term *Big Data* is now become a widely used term. Thus, this term needs to be defined clearly. Hence, some definitions are proposed.

1.1.2. *Definitions*

Many definitions of the term *Big Data* have been proposed. Ward and Baker propose a survey on these definitions [WAR 13]. As a common aspect, all these definitions indicate that size is not the only characteristic.

A historical definition was given by Laney from Meta Group in 2001 [LAN 01]. Indeed, even if he did not mention the term "Big Data", he identified, mostly for the context of e-commerce, new data management challenges along three dimensions – the three "Vs": volume, velocity and variety:

– *Data volume:* as illustrated earlier, the number of data created and collected is huge and the growth of information size is exponential. It is estimated that 40 zettabytes (40 trillion gigabytes) will be created by 2020.

– *Data velocity:* data collected from connected devices, websites and sensors require specific data management not only because of real-time analytics needs (analysis of streaming data) but also to deal with data obtained at different speeds.

– *Data variety:* there is a variety of data coming from several types of sources. Dealing simultaneously with such different data is also a difficult challenge.

The former definition has been extended. First, a fourth "V" has been proposed: *veracity*. Indeed, another important challenge is the uncertainty of data. Hence around 1 out of 3 business leaders do not trust the information they use to make decisions [IBM 16a]. In addition, a fifth "V" may also be associated with Big Data: *value*, in a sense that the main interest to deal with data is to produce additional value from information collected [NUN 14].