Future Generation Computer Systems 67 (2017) 286–296

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Geometrical and topological approaches to Big Data

Václav Snášel [a], Jana Nowaková [a,*], Fatos Xhafa [b], Leonard Barolli [c]

[a] Department of Computer Science, Faculty of Electrical Engineering and Computer Science, VŠB - Technical University of Ostrava, 17. listopadu 15/2172, 708 33 Ostrava - Poruba, Czech Republic

[b] Department of Computer Science, Technical University of Catalonia, C/Nord, Omega Bld, C/Jordi Girona 1-3, 08034 Barcelona, Spain

[c] Department of Information and Communication Engineering, Faculty of Information Engineering, Fukuoka Institute of Technology (FIT), 3-30-1 Wajiro-higashi, Higashi-ku, Fukuoka 811-0295, Japan

HIGHLIGHTS

- An overview of state-of-the-art in geometrical and topological approach to Big Data.
- Trends in geometrical and topological approach to Big Data.
- Big Data visualization.
- Discussion of current techniques and future trends to address the applications.

ARTICLE INFO

Article history:
Received 6 March 2016
Received in revised form
25 May 2016
Accepted 6 June 2016
Available online 29 June 2016

Keywords:
Big Data
Industry 4.0
Topological data analysis
Persistent homology
Dimensionality reduction
Big Data visualization

ABSTRACT

Modern data science uses topological methods to find the structural features of data sets before further supervised or unsupervised analysis. Geometry and topology are very natural tools for analysing massive amounts of data since geometry can be regarded as the study of distance functions. Mathematical formalism, which has been developed for incorporating geometric and topological techniques, deals with point cloud data sets, i.e. finite sets of points. It then adapts tools from the various branches of geometry and topology for the study of point cloud data sets. The point clouds are finite samples taken from a geometric object, perhaps with noise. Topology provides a formal language for qualitative mathematics, whereas geometry is mainly quantitative. Thus, in topology, we study the relationships of proximity or nearness, without using distances. A map between topological spaces is called continuous if it preserves the nearness structures. Geometrical and topological methods are tools allowing us to analyse highly complex data. These methods create a summary or compressed representation of all of the data features to help to rapidly uncover particular patterns and relationships in data. The idea of constructing summaries of entire domains of attributes involves understanding the relationship between topological and geometric objects constructed from data using various features.

A common thread in various approaches for noise removal, model reduction, feasibility reconstruction, and blind source separation, is to replace the original data with a lower dimensional approximate representation obtained via a matrix or multi-directional array factorization or decomposition. Besides those transformations, a significant challenge of feature summarization or subset selection methods for Big Data will be considered by focusing on scalable feature selection. Lower dimensional approximate representation is used for Big Data visualization.

The cross-field between topology and Big Data will bring huge opportunities, as well as challenges, to Big Data communities. This survey aims at bringing together state-of-the-art research results on geometrical and topological methods for Big Data.

1. Introduction

Big Data is everywhere as high volumes of varieties of valuable precise and uncertain data can be easily collected or generated at high velocity in various real-life applications. The explosive growth in web-based storage, management, processing, and accessibility

* Corresponding author.
E-mail addresses: vaclav.snasel@vsb.cz (V. Snášel), jana.nowakova@vsb.cz (J. Nowaková), fatos@cs.upc.edu (F. Xhafa), barolli@fit.ac.jp (L. Barolli).

http://dx.doi.org/10.1016/j.future.2016.06.005

of social, medical, scientific and engineering data has been driven by our need for fundamental understanding of the processes which produce this data. It is predicted that volume of the produced data could reach 44 zettabytes in 2020 [1]. The enormous volume and complexity of this data propel technological advancements realized as exponential increases in storage capability, processing power, bandwidth capacity and transfer velocity. This is, partly, because of new experimental methods, and in part because of the increase in the availability of high-powered computing technology. Massive amounts of data (Big Data) are too complex to be managed by traditional processing applications. Nowadays, it includes the huge, complex, and abundant structured and unstructured data that is generated and gathered from several fields and resources. The challenges of managing massive amounts of data include extracting, analysing, visualizing, sharing, storing, transferring and searching such data. Currently, traditional data processing tools and their applications are not capable of managing Big Data. Therefore, there is a critical need to develop effective and efficient Big Data processing techniques. Big Data has five characteristics: volume, velocity, variety, veracity and value [2]. Volume refers to the size of the data for processing and analysis. Velocity relates to the rate of data growth and usage. Variety means the different types and formats of the data used for processing and analysis. Veracity concerns the accuracy of results and analysis of the data. Value is the added value and contribution offered by data processing and analysis.

Modern data science uses so-called topological methods to find the structural features of data sets before further supervised or unsupervised analysis. Geometry and topology are very natural tools for analysing massive amounts of data since geometry can be regarded as the study of distance functions. Besides the heterogeneity of distance functions, another issue is related to distance functions on large finite sets of data. The mathematical formalism which has been developed for incorporating geometric and topological techniques deals with point clouds, i.e. finite sets of points equipped with proximity or nearness or distance functions [3,4]. It then adapts tools from the various branches of geometry and topology for the study of point clouds [5]. The point clouds are finite samples taken from a geometric object, perhaps with noise.

Geometrical and Topological methods are tools for analysing highly complex data [3]. These methods create a summary or a compressed representation of all of the data features to help rapidly uncover patterns and relationships in data. The idea of constructing summaries of entire domains of parameter values involves understanding the relationship between geometric objects constructed from data using various parameter values, e.g. [8].

One problem with Big Data analysis, which is very actual, is that the currently used methods based on model creation, simulation of the created model and then assessment, whether the original data corresponds to data obtained using the created model–model verification cannot be applied. The described process is useful and appropriate for solving classic problems such as physical problems, because the theoretical background of these problems has been researched and understood enough, so it could be reconstructed to fit the model. For Big Data processing, the first problem is that we are not able to define the concrete hypothesis of the data feature which could be tested. Due to this, for the Big Data problem, the same approach as with the classic physical problem cannot be used. Therefore, the main aim of the research is not to define a model, but to be able to mine accurately and automatically interesting features of Big Data sets. In many cases, the data to be examined is often based on shapes that are not easy to capture using traditional methods [9].

A common thread in various approaches for noise removal, model reduction, feasibility reconstruction, and blind source separation, is to replace the original data with a lower dimensional approximate representation obtained via a matrix or multi-directional array factorization or decomposition. Besides those transformations, a significant challenge of feature summarization or subset selection methods for Big Data will be considered by focusing on scalable feature selection. Lower dimensional approximate representation is used for Big Data visualization to be able to visualize data in the understandable form. This approach—dimensionality reduction can be also understood as a method for feature compression, see Fig. 6.

The whole paper is organized as follows: in Section 2, a brief introduction to Big Data technologies is given. In the next Section 3, a brief motivational example is presented. A short mathematical background is introduced in Section 4. This part contains a brief review of topology, metric space, homology and persistent homology theory, manifolds and Morse theory. In the following Section 5, a brief review of homology and persistent homology theory is introduced. Various applications of geometrical and topological methods are presented in Section 6. Big Data visualization is discussed in Section 7. In this section, we discuss methods to create a summary or compressed representation of all of the data features to help visualize hidden relationships in data. This is followed by the section described and introduced new, perspective Big Data challenges. This paper ends with conclusions in Section 9.

2. Big Data technologies during time

The manner in which data is stored, transmitted, analysed and visualized has varied over time; the rise of all fields of human activities is always connected with an increase in technological possibilities, as with the political situation, development of the socio-economical arrangement and industry. In 1936, Franklin D. Roosevelt's administration in the USA, after Social Security became law, ordered from IBM the development of the punch card-reading machine to be able to collect data from all Americans and employers. This biggest accounting operation of all time, as it was called at that time, can be considered as the first major data project [10–13]. As already mentioned, the political situation always has a big influence on the rise of technology and the main mover of its development has always been war and money. During World War II, the British invented, in 1943, a machine *Colossus* to decipher German codes. The device, which searched for patterns in encrypted messages at a rate of 5000 characters per second, is known as the first data-processing machine [14,10]. Big Data as a term has been one of the biggest trends in recent years, leading to an increase in research, as well as industry and government applications [15–17]. The continued improvements in high-performance computing and high resolution sensing capabilities have resulted in data of unprecedented size and complexity. Data is deemed a powerful raw material that can impact multidisciplinary research.

2.1. Data storage

We face a wave of data; the amount of data is so big that a lot of information is never looked at by anybody [18]. The next problematic aspect of data is that a big part of it is redundant, e.g. one video due to many existing video formats, its resolution and subtitles in many languages [19] takes up lots of space, which is necessary from an informational point of view but generally it does not bring anything new. The manner in which data is stored has changed: what was sufficient in 1965, when the US Government decided to found the first data centre to store 175 million sets of fingerprints and 742 million tax returns and store data onto magnetic computer tape [20], is nowadays unusable. Traditionally, persistent data is still stored using hard