Big Data Visualization: Methods and Software

Masterarbeit

zur Erlangung des akademischen Grades „Master of Science (M. Sc.)“ im Studiengang Wirtschaftswissenschaft der Wirtschaftswissenschaftlichen Fakultät der Leibniz Universität Hannover

vorgelegt von

Name: Schmitz [redacted] Vorname: Jan-Niklas [redacted] [redacted]

Prüfer: Prof. Dr. M. H. Breitner

Ort, den: Hannover, 01.04.2019

TABLE OF CONTENTS

1 INTRODUCTION

"We are drowning in information
but starved for knowledge."

- Naisbitt (1982)[1]

We are currently living in a data-driven world that faces tremendous challenges and opportunities concerning the efficient and effective exploitation of data. As of 2012, Data is created at a staggering speed with roughly 2.5 exabytes doubling every 40 months.[2] Steady improvements in storage capabilities and means of data collection possibilities profoundly influenced the way data is dealt with these days. The possibilities to generate and store data is developing at a much faster pace as the feasibility to make use of it in decision-making processes. The plain acquisition of data is no longer the impelling problem but making sense of the available data, generate knowledge from it using appropriate methods and models, and incorporate and leverage this knowledge in future decision-making processes is posing a formidable challenge. This gap of available information and its inherent knowledge that is not extracted from it is entirely in line with the citation of Naisbitt (1982) from almost four decades ago.

Especially for businesses and decision-makers, it is of crucial importance to make timely decisions that are grounded in the fundamental analysis of related data. However, conventional analysis methods are stretched to its limits when dealing with the diversity and amount of the data nowadays. Companies are facing an abundance of data nowadays providing an inherent value that can be accessed through the utilization of sophisticated software and algorithms that are mainly justified for handling big data. One primary target in big data analysis is to detect and discover patterns inside the vast amounts of data. Visualizing the enormous amount of data make them more accessible for human beings using the perceptive and cognitive abilities. The interdisciplinary approach of visual analytics (VA) addresses the dynamic interplay between the domain expert's knowledge, the perception and cognitive abilities of humans, and the machine including the computational power. Visual analytics comprises various highly related fields of research such as visualization, data mining, data management, and statistics. The core idea of VA is to combine these research fields into one. Available commercial software solutions target this information overload

[1] Naisbitt, 1982: 24

[2] McAfee et al., 2012: 64

problem by providing state-of-the-art visualization techniques and enabling users without knowledge of any programming language to efficiently explore and visualize the respective data.

1.1 Research Question

This thesis addresses different visualization approaches for big and high-dimensional data sets and decision support systems that forge a hybrid intelligence of the domain expert and the machine. In particular, this thesis is devoted to the question of how the tight integration of human and machine forming a hybrid intelligence can amplify the decision-making process through the utilization of a VA system?

1.2 Thesis Outline

The remainder of this thesis is structured as follows: Chapter 2 provides a fundamental literature review of the visualization approaches, decision-making processes, VA framework, and corresponding model extensions. Chapter 3 distinguishes open source VA toolkits and selected self-service business intelligence application devoted to the visualization of large data sets. Chapter 4 provides a review of statistical methods addressing the visualization of big and in particular high-dimensional data. Particularly dimensionality reduction techniques and a clustering algorithm are explained. Chapter 5 describes the applied methodology for the development of a web-based hybrid intelligent VA application dedicated to amplifying decision support by incorporating the domain expert into the analysis. Chapter 6 provides an application of the developed solution to three generic case studies. Chapter 7 depicts the limitations of the application and the further outlook for research and practitioners. Ultimately, chapter 8 concludes this thesis.

8 CONCLUSION

Living in a data-driven world provides numerous challenges for businesses and research in efficiently exploiting the underlying insights and knowledge in order to incorporate it into the decision-making process. However, when performed in an orderly manner the competitive advantages and possibilities for harnessing this knowledge exceed the challenges by far. This thesis focused primarily on the possibilities of integrating VA into a data-driven decision-making process shaping a hybrid intelligent system that integrates the prime advantaged of both, the human and computer side. Current state-of-the-art visualization and DSS approaches provided the basis for further research. Initiated by the VA process model further advancements with respect to the tight integration of domain-experts and computers were examined. This progress provided a more detailed view of how exactly the domain-expert is effectively integrated through various interaction taxonomies. Leading software applications already address the topic of integrating users without the capabilities of an ordinary statistical programming language to efficiently and interactively engage with the data. Two of the leading software were described regarding their specific features, advantages and particular drawbacks in statistical computation possibilities. Specific statistical algorithms and techniques that are devoted to handling high-dimensional data sets were evaluated and presented subsequently.

In order to address the missing capabilities of commercial VA systems to implement complex statistical algorithms, this thesis presents CIVEDA, a specially developed VA system. This system encompasses current state-of-the-art interactive visualizations and statistical approaches towards the challenges that emerge when dealing with big and high-dimensional data sets. CIVEDA is a hybrid intelligence system combining the strengths of human cognition and perception and the computational efficiency of computers. It serves as an initial model for a VA system developed to amplify data-based decision-making. The primary components of CIVEDA the visual exploration of the data sets in an exploratory data analytics demeanor, variable modifications through preprocessing techniques, implementation of dimensionality reduction techniques, and execution of machine learning algorithms. The principal contribution of the presented thesis is to provide a practical approach to integrate VA into data-driven decision-making system. The two presented case studies prove the valuable approach to interactively engage the domain expert with the analysis system and facilitate to amplify the knowledge generation process.

VA systems provide an excellent opportunity to integrate domain-expert’s knowledge throughout the entire analytics process and make data-driven decision-making more independently. As a result of the ever-increasing speed at which data is generated and stored

future analysis, and exploration tasks can neither be conducted solely by humans or the machine. Thus, engaging with complex problems and topics requires tight integration of the computational efficiency and storage capabilities of machines with the human intelligence and cognitive abilities.