Accepted Manuscript

Towards modeling and optimization of features selection in Big Data based social Internet of Things

Awais Ahmad, Murad Khan, Anand Paul, Sadia Din, M. Mazhar Rathore, Gwanggil Jeon, Gyu Sang Chio

Please cite this article as: A. Ahmad, M. Khan, A. Paul, S. Din, M.M. Rathore, G. Jeon, G.S. Chio, Towards modeling and optimization of features selection in Big Data based social Internet of Things, *Future Generation Computer Systems* (2017), https://doi.org/10.1016/j.future.2017.09.028

Towards Modeling and Optimization of Features Selection in Big Data based Social Internet of Things

Awais Ahmad, Murad Khan, Anand Paul, Sadia Din, M. Mazhar Rathore, Gwanggil Jeon, Gyu Sang Chio

Abstract— **The growing gap between users and the Big Data analytics requires innovative tools that address the challenges faced by big data volume, variety, and velocity. Therefore, it becomes computationally inefficient to analyze and select features from such massive volume of data. Moreover, advancements in the field of Big Data application and data science poses additional challenges, where a selection of appropriate features and High-Performance Computing (HPC) solution has become a key issue and has attracted attention in recent years. Therefore, keeping in view the needs above, there is a requirement for a system that can efficiently select features and analyze a stream of Big Data within their requirements. Hence, this paper presents a system architecture that selects features by using Artificial Bee Colony (ABC). Moreover, a Kalman filter is used in Hadoop ecosystem that is used for removal of noise. Furthermore, traditional MapReduce with ABC is used that enhance the processing efficiency. Moreover, a complete four-tier architecture is also proposed that efficiently aggregate the data, eliminate unnecessary data, and analyze the data by the proposed Hadoop-based ABC algorithm. To check the efficiency of the proposed algorithms exploited in the proposed system architecture, we have implemented our proposed system using Hadoop and MapReduce with the ABC algorithm. ABC algorithm is used to select features, whereas, MapReduce is supported by a parallel algorithm that efficiently processes a huge volume of data sets. The system is implemented using MapReduce tool at the top of the Hadoop parallel nodes with near real-time.**

Awais Ahmad is with the Department of Information and Communication Engineering, Yeungnam University, Korea (e-mail: aahmad.marwat@gmail.com).

Murad Khan is with the Department of Computer Science, Sarhad University of Science and Information Technology, Peshawar, Pakistan (e-mail: muradkhan23@gmail.com).

Anand Paul is with the School Of Computer Science and Engineering, Kyungpook National University, Daegu, 702-701, Korea (e-mail: paul.editor@gmail.com).

Sadia Din is with the School Of Computer Science and Engineering, Kyungpook National University, Daegu, 702-701, Korea (e-mail: research.2486@gmail.com).

M. Mazhar Rathore is with the School Of Computer Science and Engineering, Kyungpook National University, Daegu, 702-701, Korea (e-mail: rathoremazhar@gmail.com).

Gwanggil Jeon is with Department of Embedded Systems Engineering, Incheon National University, Incheon, Korea (email: gjeon@inu.ac.kr)

Gyu Sang Chio is with the Department of Information and Communication Engineering, Yeungnam University, Korea (e-mail: castchoi@ynu.ac.kr).

Moreover, the proposed system is compared with Swarm approaches and is evaluated regarding efficiency, accuracy and throughput by using ten different data sets. The results show that the proposed system is more scalable and efficient in selecting features.

Index Terms— **SIoT, Big Data, ABC algorithm, feature selection.**

I. INTRODUCTION

THE most important paradigm that fills the gap between physical world and cyber world is the Social Internet of Things (SIoT). Recent advancement in the field of IoT has led to the digitization of the physical world where the configuration of novel applications and services are high in demand. For such advancements, a variety of things is grouped together to share information with the help of Internet. These things include radio frequency identification (RFID) tags, sensors, actuators, mobile equipment, computers, medical sensors, etc. and are connected to each other through wires or wirelessly. The evolved things in the SIoT can sense the physical environment, collect data, transfer or disseminate data, process data for appropriate applications, and communicate with other things. Hence, SIoT came up with a power technology that helps in understanding the physical world and to response to outer stimuli. Hence, an ultimate solution that gives us insight to the real-world in real-time.

Apparently, advancements in the field of SIoT pose new challenges when it comes to implementation [1]. Since SIoT is a mixture of heterogeneous things, making it quite different from traditional networks, thus it becomes more complex due to its scalable property [2]. As a result, the evolved things cannot directly apply to SIoT. Moreover, given the complex, heterogeneous nature of SIoT, various things communicating with each other under its umbrella consume a high volume of memory, processing power, and high bandwidth. Thus, IoT tends to generate a huge volume of data, referred to Big Data. The term Big Data is categorized into some specific types of data sets, which comprises formless data, which resides in the data layer of the technical computation applications as well as the web. Usually, Big Data comprises 3 V's, such as Volume (referred to as the size of the data sets), Velocity (referred to as high-speed processing and analyzing) and variety (referred to as different data sources, i.e., heterogeneous networks). To