An efficient algorithm for reducing the flow of real-time data stream with least sampling error

Devesh Kumar Lal* and Ugrasen Suman

School of Computer Science and IT,
Devi Ahilya University,
Indore, India
Email: devesh2222@gmail.com
Email: ugrasen123@yahoo.com
*Corresponding author

Abstract: Nature of data stream is determined after complete scanning of whole data sets during real-time data processing. However, it becomes inconvenient to process entire data stream at once in real-time data stream processing. Thus, a sheer sized fixed window of data streams is processed at a particular time. The intensification of sheer sized fixed window at processing node is mitigated by reducing the flowing rate of data stream. Heuristic clustering windowing (HCW) approach and partial blind window (PBW) algorithms are proposed for reducing the flow of data stream with least sampling error. These approaches consist of the combination of systematic sampling and clustering mechanism. A clustering approach is applied on one fraction of data streams whereas systematic sampling handles other portion of streams. These approaches are helpful in reducing flow of data streams in minimum latency.

Keywords: clustering approach; data stream; data processing; data sampling; real-time big data; systematic sampling.

Reference to this paper should be made as follows: Lal, D.K. and Suman, U. (2020) 'An efficient algorithm for reducing the flow of real-time data stream with least sampling error', *Int. J. Big Data Intelligence*, Vol. 7, No. 4, pp.186–193.

Biographical notes: Devesh Kumar Lal is pursuing his PhD in Computer Science from Devi Ahilya University Indore, and did his Master of Engineering from Institute of Engineering and Technology Indore India. He has completed his Bachelor in Engineering from KCB Technical Academy RGTU Bhopal India. He has more than two years of research experience in real time big data processing and currently working as Senior Research Fellow (SRF) under Council of Scientific and Industrial Research New Delhi India.

Ugrasen Suman is working as a Professor at School of Computer Science and Information Technology, Devi Ahilya University, Indore, India. He has more than 17 years of teaching and research experience. His areas of research are software engineering, information system, software reuse, software maintenance, and reengineering, agile methodologies, software architectures, service-oriented computing, knowledge management and mining. He has authored three books *Software Engineering: Concepts & Practices* (Cengage Learning, 2013), *Software Engineering* (Cengage Learning, 2018), *Object Oriented Analysis & Design with UML*. He has published three book chapters and 88 research papers in national and international journals/ conferences.

1 Introduction

Data centric research has been widely explored by researchers, which incorporates various domains such as, IOT, data stream analytics, learning, etc. The unbounded generated data streams are processed by stream processing engines (SPE) with deployment either in cloud or localised. The large volume of data streams enhances the processing cost in cloud. Whereas in static localise processing unit unable to scale up with respect to the growth of data stream. The sampling of population data streams provides a solution to these challenges by reducing streams.

The processing of exponential growth of data streams, which process in stable hardware configurations for perpetual span of time is anomalous. Adding pre-processing of data streams such as, noise removal, anomaly detection, feature extraction, outlier detection, filtering, etc., is critical for real-time applications (Rehman et al., 2016). The processing of such data streams reduces efficiencies with increased latency, storage, and computational unit. The sustainable increased rate of data stream generation is directly affects its processing cost. The reduction in data streams acquisition by factor of two with minimal semantics alteration may support for providing better efficiency. In real-time stream processing, data acquisition plays a significant role. At the time of data stream acquisition from multifarious sources, the characteristics such as, variety, volume, veracity, and velocity has to consider in real-time.

Copyright © 2020 Inderscience Enterprises Ltd.