

Big Sequence Data Analysis: From Stream Processing Technology to Applications in Sleep Medicine

Hiroyuki Kitagawa

International Institute for Integrative Sleep Medicine

University of Tsukuba

Tsukuba, Japan

Big data has been bringing a huge impact on every aspect of human activities and human society. Big data technologies are major driving forces which lead us to the next generation society and industry. Many research and development projects have been launched to advance big data technologies and their applications. Japan is not an exception. My research group has been involved in several projects on big data, including "Research and Development on Real World Big Data Integration and Analysis" supported by MEXT, Japan. This talk will give an overview of the project and digest technologies we have developed through such research projects. Actually, big data covers diverse perspectives. Sequence data such as event streams, moving object trajectories, and biological signals plays a more and more important role in big data analysis. I will focus on our recent research activities on big sequence data analysis. They include stream OLAP framework, traceability and lineage tracking in stream analysis, and biological sequence data analysis for sleep medicine.

As for stream OLAP framework, I will introduce our system named StreamingCube, which provides OLAP facilities integrated with data stream processing, and discuss optimization schemes for efficient OLAP analysis. Traceability/trust is a key issue in big data analysis, and data lineage has been studied to enhance the traceability of data analysis in the database community. Data lineage allows information to be traced to its origin by showing how the results were derived. However, data analysis is becoming increasingly complex both with regard to the target (e.g., images, videos, and texts) and technology (e.g., AI and machine learning). Thus, simply showing the source data will not be enough to guarantee traceability. To address this, we propose the concept of Augmented Lineage, which is an extended lineage and enhances the traceability in complex data analyses. The talk explains the basic concept of Augmented Lineage and its application to stream processing. Finally, we discuss another type of sequence data analysis, namely human biological sequence data analysis. Especially, I will digest our recent research on automated sleep stage scoring using human EEG, EOG, EMG sequence data.