

Information Models for Large Table Trimming

Thursday 18 September 2025 11:55 (15 minutes)

Large datasets can rarely be presented or used in real time without significantly reducing their size. This paper discusses models of trimming timestamped event datasets while keeping the loss of information to a minimum. The presentation goes gradually from independent event models, where trimming of events does not change the order of the information contribution of the other events, to statistical models adjusted to incorporate cross-referencing of entries and memory effects into the information calculation. Based on the particular structure of the information function, various trimming strategies are discussed. Depending on the contents of the registered events, such models can be used to retain most of the information in the dataset, while significantly decreasing the computation time. This is particularly useful when dealing with frontends that can handle a limited amount of data, or with sampling the training data for machine learning models.

Author: MANOLACHE, Florin Bogdan (Carnegie Mellon University)

Co-author: Dr RUSU, Octavian

Presenter: MANOLACHE, Florin Bogdan (Carnegie Mellon University)

Session Classification: High Performance Computing in Science

Track Classification: Grid, Cloud & High Performance Computing in Science