Spatial and temporal sampling of distributed data streams

Raja CHIKY and Georges HEBRAIL
Télécom ParisTech, UMR CNRS 5141
46 rue barrault 75013, Paris
first.last@telecom-paristech.fr

Abstract

Remote sensing enables data collection from multiple sources in a unique central server for a large wide of applications which are typically monitoring and system supervision, transmitting alarms in real time and more generally producing synthesis to help in business decision. The volume of data collected is generally too big to be stored entirely. Data Stream Management Systems are generic tools for processing data streams: data elements arrive on-line and stay only for a limited time period in memory. For many stream processing applications, one may not wish to completely lose the entire stream data for analyzing 'past' and 'present'data. Thus a stream processing system must also provide a stored historical data.

In this paper, we consider a distributed computing environment, describing a collection of \( N \) remote sensors that feed a unique central server with numeric and uni-dimensional data streams (also called curves). The central server has a limited memory but should be able to compute aggregated value of any subset of the stream sources from a large time horizon including old and new data streams. Two approaches are studied to reduce the size of data: (1) spatial sampling [2] only consider a random sample of the sources observed at every instant; (2) temporal sampling [1] consider all sources but samples the instants to be stored. The computation of aggregates involves statistical inference in the case of spatial sampling and interpolation in the case of temporal sampling. To the best of our knowledge, there is no method for estimating interpolation errors at each timestamp that would take into account some curve features such as the knowledge of the integral of the curve during the period. We propose two approaches: one uses a stochastic process for interpolation (stochastic approach) and the other uses the past of the data curve (naive approach).

Références