Online Analysis of Data Streams

by
Jin Cao
Bell Labs
cao@research.bell-labs.com

Abstract

Massive data streams are becoming increasingly commonplace in many areas of science and technology. Example of such are network packet traces, or business transaction records. With massive data, and limited storage and computation power, analysis of such data becomes difficult even for very simple task such as simple counting of the number of distinct values. In this talk, I shall give a statistician's perspective for analyzing such data streams using online methods, which means the data cannot be stored and will be seen only once. I will survey our recent work on data streaming problems arising in analyzing network packet traces, and give a more detailed account on the problem of cardinality counting and quantile estimation.