A Study of Time-decayed Aggregation of Distributed Streaming Data

Ph.D. Dissertation Research, Bojian Xu (bojianxu@iastate.edu)
Department of Electrical & Computer Engineering, Iowa State University

Introduction

Many real world data naturally arise as streams (see examples below). These streams of data need to be monitored to collect traffic statistics, detect trends and anomalies, tune system performance and help make the business policies. However, due to the large size of such streaming data, conventional data processing methods are not feasible. My Ph.D. dissertation research studies fundamental problems in the processing of such streaming data in a time and space efficient manner, with applications to network and database management.

Example Data Streams

- IP Packet Stream
- Stock Exchange Records
- IP Packet Flow
- Database Accessing
- Traffic Monitoring Data

Challenges in Data Stream Processing

- One pass processing
- Fast processing of each streaming data
- Working space much smaller than stream size
- Continuous Queries

Distributed Asynchronous Data Streams Processing

Why Asynchronous Streams?

- Synchronous stream
- Asynchronous stream

Synchronous stream:

- Timestamp sliding window
- Network delay & multi-path routing

Asynchronous stream:

- Backward Decay is costly
- Hardly track all the ages in small space

Distributed Data Stream Phenomenon

- IP packet flow through network links
- Sensor observations in a sensor network
- Access sequence of a distributed database
- Trading log at stock exchanging markets
- Sequences of search key words at Google
- web servers

Example Applications

- Online network monitoring
- Energy-efficient sensor network computing
- Database access traffic analysis
- Online business data analysis for decision makers

Data Stream Model

\[ \text{Stream: } R = \{v_0, v_1, v_2, \ldots \} \]

- \( v_i \) - value
- \( w_i \) - initial weight
- \( \tau_i \) - unique id
- \( t_i \) - timestamp

Lower Bound 1: Sublinear space upper bound for reliable error guaranteed estimate

Upper Bound 1: Sublinear space upper bound for adversarial error guaranteed estimate

Visualizations of the Contributions from Prior Works and Our Work

(1) J. Gehrke, F. Korn and D. Srivastava. SIGMOD 2001

Main Collaborators

Srikanta Tirthapura (Thesis Advisor)
Associate Professor
Department of Electrical & Computer Engineering
Iowa State University

Graham Cormode (Internship mentor)
Principle Researcher
AT&T Labs – Research, Florham Park, NJ

Achievements

We handle: Asynchronous Streaming Data

- Order of data delivery is not the same as the order of data generation
- Inevitable in distributed context, such as networking data, where skewness is caused by network delay or multipath routing

We handle: Multi-dimensional Correlated Streaming Data

- Multi-dimensional data
- Data are correlated among different attributes
- Aggregate one dimension of the data where other dimensions satisfy a predicate given by users at query time

Examples:

- Average transmission delay over the IP packets whose payload is \( \leq 100 \) bytes
- Average latency to access a URL over the IP packets received in the last 24 hours

We handle: Time-decayed Streaming Data

- Newer data are more important
- Weight of each data is continuously being eliminated

Example:

- Linearly averaged latency to access a URL over the IP packets received in the last 24 hours

Selected Publications

   - Selected as one of the 7 best papers out of 231 submissions and awarded a special prize of $1,000 and a plaque in the Journal of the ACM!

   - Time-Decaying Sketches for Sensor Data Aggregation.

   - Journal paper (42 pages) under submission to Statistical Analysis and Data Mining.
   - Journal paper (42 pages) under submission to Journal of Parallel and Distributed Computing.

Summary

We proposed the new concept asynchronous stream, a more robust model for network data. It led to a new research direction investigating data decay by other researchers.

We designed the first general purpose small space data structure for distributed streaming data aggregation. 1) (a) surrogate data structure. 2) Supports time-decayed data based on any time decay model. 3) Handles asynchronous streams. 4) Can return accuracy guaranteed estimates for a large family of aggregates.

We proposed a practical time decay model for streaming systems to apply data aggregation and sampling over temporal streaming data. It nearly has no extra cost upon aggregation and sampling over temporal streaming data.

We conducted the first comprehensive study on the time-decayed correlated streaming data aggregation. Our results not only significantly improved the prior results, but also closed open problems proposed by previous work.