Introduction to Data Stream Processing

Sumit Ganguly
Indian Institute of Technology Kanpur

Abstract

Data Streaming is a model of computation where data arrives continuously and at very high speeds, for example, network router switch data transactional data, sensor network data (bridge/building monitoring, GPS sensor data) etc.. Data streaming applications process data at very high speeds, while computing data summaries and prominent data statistics using (significantly) sub-linear space and very fast update times (O(1) or poly-logarithmic), while allowing approximation factors typically $1 \pm \epsilon$ and error probability at most δ.

The general streaming model defines a stream as a sequence of updates of the form $\mathcal{S} = (i_1, v_1), (i_2, v_2), \ldots, (i_t, v_t), \ldots$, where, $i_1, i_2, \ldots, i_n, \ldots$ are items (e.g., (src-IP, dest-IP) pairs, URLs etc.) that take values from a large domain $\{1, \ldots, N\}$. An N-dimensional frequency vector $f = (f_1, \ldots, f_N)$ is conceptually maintained that is initialized to the zero vector and is updated by each update (i_t, v_t) as follows: $f_{i_t} \leftarrow f_{i_t} + v_t$. The v_t's are the updates made to the *frequency* of the item f_{i_t}, where, $|v_t| \leq M$, for some positive upper bound M. This model is the *general update* model of data streaming. Several restricted models have been popularly studied, (1) the *unit insert-only* model, where, all the v_t's are equal to 1, (2) the *general insertion* model, where, $1 \leq v_t \leq M$.

Among the most popular problems studied in this model are, (1) estimating the number of distinct items in the general insertion model: $F_0 = \{i \mid f_i > 0\}$ and its corresponding ℓ_0 measure for general update streams $\ell_0(t) = \{i \mid f_i \neq 0\}$, (2) the ℓ_p/F_p frequency moment problem: $\ell_p = \left(\sum_{i \in [N]} |f_i|^p\right)^{1/p}$. It is well-known that the ℓ_p/F_p problem can be estimated within error factors $1 \pm \epsilon$ and with confidence $1 - \delta$ using space $O(\epsilon^{-2} \log(1/\delta) \log N)$ bits for $0 < p < 2$, and requires $\Omega(n^{1-2/p})$ bits for $p > 2$. A closely related classical result is the Johnson-Lindenstrauss Lemma for ℓ_2 estimation, with applications to a wide variety of problems, including a fast approximation to the numerical linear regression problem.