Weighted distinct elements

April 7, 2020

1 Introduction

The purpose of this writeup is to present solutions to the weighted version of the streaming distinct elements problem.

We are given a stream of elements from a universe $\mathcal{U}$ of size N. Every element $i \in \mathcal{U}$ has weight $w_i > 0$. Let f_i denote the number of occurences of i on the stream. Our goal is to esimate $\sum_{i:f_i>0} w_i$ up to a $(1 \pm \epsilon)$ approximation, with high probability (say 0.99). The special case $\forall_i w_i = 1$ is the usual (unweighted) distinct elements problem.

We will assume the weights are integers in $\{1, \ldots, W\}$. Note that if the weights are real-valued, as long as we have a lower bound W_{min} on the weights, we can transition to the integer weight setting by discretization.

2 Unweighted distinct elements

We state the classic result of Flajolet-Martin (FM) for the unweighted case, since the weighted solutions below will largely be by reduction to it.

Theorem 2.1 (FM)**.** *For the unweighted distinct elements problem, there is a solution that uses* $O(\epsilon^{-2} \log N)$ *space, and* $O(\log N)$ *time per stream element.*

The HyperLogLog algorithm partitions the stream at random into several sub-streams, runs FM on each sub-stream, and returns the harmonic average of the sub-stream FM estimates.

3 Solution I: $O(\epsilon^{-3} \log N \cdot \log W)$ space

One solution for the weighted case is to round each weight w_i to its nearest power of $(1+\epsilon)$. There are only $L = \lceil \log_{1+\epsilon} W \rceil = O(\epsilon^{-1} \log W)$ rounded weights. For every $k = 0, \ldots, L-1$, we run a copy FM_k of FM and feed into it the elements whose rounded weight is $(1+\epsilon)^k$. Let E_k be the output estimate of FM_k. We return $\sum_{k=0}^{L-1} (1+\epsilon)^k E_k$. The correctness of the estimate follows immediately from that of FM.

The total space is L times the space of FM. Next we will consider solutions with better space usage (in particular, comparable maybe up to log factors to unweighted FM).

4 Solution II: $O(\epsilon^{-2} \log(NW))$ space, $O(W \log N)$ time

Another solution for the weighted case is to turn every element i into w_i unweighted elements, (i, j) for $j = 1, \ldots, w_i$. We thus generate from our input stream a new stream whose universe is

$\mathcal{U} \times \{1, \ldots, W\}$. Observe that the number of distinct elements in the new stream is equal to the solution of weighted distinct elements on the original stream. Hence we can simply run FM on the new stream and return its output.

The space usage is the same as FM on a universe of size NW, thus $O(\epsilon^{-2} \log(NW))$. However, for each input element we might need to insert as many as W elements into FM, so the time per element is $O(W \log N)$. This is exponential is the description size of the weights, so we would like to improve on it.