Count-min sketch: range queries

Show and analyse the application of count-min sketch to range queries (i, j) for computing $\sum_{k=i}^{j} F[k]$. Hint: reduce the latter query to the estimate of just $t \leq 2\ log\ n$ counters $c_1, c_2, ..., c_t$. Note that in order to obtain a probability at most δ of error (i.e. that $\sum_{l=1}^{t} c_l > \sum_{k=i}^{j} F[k] + 2\epsilon\ log\ n||F||$), it does not suffices to say that it is at most δ the probability of error of each counter c_l: while each counter is still the actual wanted value plus the residual as before, it is better to consider the sum V of these t wanted values and the sum X of these residuals, and apply Markov's inequality to V and X rather than on the individual counters.

SOLUTION

We use the Dyadic Interval: a dyadic interval is a range of the form $[x2^y + 1...(x+1)2^y]$ for parameters x and y. Each point in the range $[1 \ldots n]$ is a member of $log_2 n$ dyadic interval, one for each y in the range $0 \ldots log_2(n) - 1$ (think it as tree, where we split the interval in 2 at each level).

A count-min sketch table is kept for each set of dyadic ranges of length 2^y, one for each level in the tree, thus we have $log_2 n$ vectors $\widetilde{F_i}$.

Witnessing a new element in the stream will therefore trigger an update to all the $log_2 n$ tables.

The idea for a range query is to partition the range into dyadic intervals and return as result the sum of the values stored in the CMS tables for the corresponding intervals [see this link for further reference].

It can be shown[1] that any range will be split at most into $2log_2 n$ dyadic intervals.

Therefore for each query we access $t \leq 2\ log\ n$ counters $c_1, c_2, ..., c_t$. We then, want to proof $Pr[\sum_{l=1}^{t} c_l > \sum_{k=i}^{j} F[k] + 2\epsilon\ log\ n||F||] < \delta$.

Firstly, we notice that each counter c_i represents an interval $[a, b]$ and it value is $\sum_{k=a}^{b}(F[k]) + X_i$, where X_i is the rubbish of that particular interval. Then, give an interval $[l, r]$, and the counters we have $\sum_{i=1}^{t} c_i = \sum_{k=l}^{r} F[k] + X$, where X is the total error (notice that this works because the interval are disjoint). Then we substitute to the previous equation and we have:

$$Pr[\sum_{l=1}^{t} c_l > \sum_{k=l}^{r} F[k] + 2\epsilon\ log\ n||F||] < \delta$$

$$Pr[\sum_{k=l}^{r} F[k] + X > \sum_{k=l}^{r} F[k] + 2\epsilon\ log\ n||F||] < \delta$$

$$Pr[X > 2\epsilon\ log\ n||F||] < \delta$$

Now, we apply Markov inequality and by the linearity of the expectation we have:

$$Pr[X > 2\epsilon\ log\ n||F||] \leq \frac{E[X]}{2\epsilon\ log\ n||F||} \leq \frac{\sum_{i}^{2log_2(n)} E[X_i]}{2\epsilon\ log\ n||F||}$$

Now, $\forall i. E[X_i] < \frac{\epsilon}{e}||F_i||_1$, but $\forall i, j. ||F_i||_1 = ||F_j||_1 = ||F||_1 \implies E[X_i] < \frac{\epsilon}{e}||F||_1$.
(by definition $||A||_1 = \sum_{h=1}^{n} |A[h]|$, that is in our case we sum the frequencies of symbols, which sum up to the same total amount no matter how we group them)

Thus we have

$$\frac{\sum_{i}^{2log_2(n)} E[X_i]}{2\epsilon\ log\ n||F||} \leq \frac{2\frac{\epsilon}{e} log\ n||F||}{2\epsilon\ log\ n||F||} = \frac{1}{e}$$

[1] First, note that:

- No three intervals of the same length can be contained in the partition of the same query, otherwise you could merge two of them;
- If there are two consecutive intervals of the same length, they must belong to two different larger "parent" intervals (otherwise you could replace them with their parent), i.e. their union cannot belong to the dyadic partition.

Thus, there are at most 2 intervals of each possible size in the partition. As the sizes are $log_2 n$ in all, the partition is made up of at most $2log_2 n$ intervals.

This is the probability of error (that the sum of garbage is more the $2\epsilon \; log \; n||F||$) in row j (the min, potentially distinct for each table).

Since we choose the row of each table that minimize the sum of the counter (by definition), then there must be an error in all the row r. Thus we have $\frac{1}{e^r} = \delta$ since $r = ln(\frac{1}{\delta})$.