University of Pisa

Department of computer science

Algorithm design

Eighth hands-on: Count-min sketch: range queries

Domenico Erriquez

1. PROBLEM

Consider the counters$F[i]$ for 1≤ $i \leq n$, where n is the number of items in the stream of any length. At any time, we know that $||F||$ is the total number of items (with repetitions) seen so far, where each $F[i]$ contains how many times item i has been so far. We saw that CM-sketches provide a FPTAS $F'[i]$ such that $F[i] \leq F'[i] \leq F[i] + \varepsilon\ ||F||$, where the latter inequality holds with probability at least $1 - \delta$.
Consider now a range query (a,b), where we want $F_{ab} = \sum_{a \leq i \leq b} F[i]$. Show how to adapt CM-sketch so that a FPTAS F'_{ab} is provided:

- Baseline is $\sum_{a \leq i \leq b} F'[i]$, but this has drawbacks as both time and error grows with $b - a + 1$.
- Consider how to maintain counters for just the sums when $b - a + 1$ is any power of 2 (less or equal to n):
  - Can we now answer quickly also when $b - a + 1$ is not a power of two?
  - Can we reduce the number of these power-of-2 intervals from $n \log n\ to\ 2n$?
  - Can we bound the error with a certain probability? Suggestion: it does not suffices to say that it is at most δ the probability of error of each individual counter; while each counter is still the actual wanted value plus the residual as before, it is better to consider the sum V of these wanted values and the sum X of these residuals, and apply Markov's inequality to V and X rather than on the individual counters

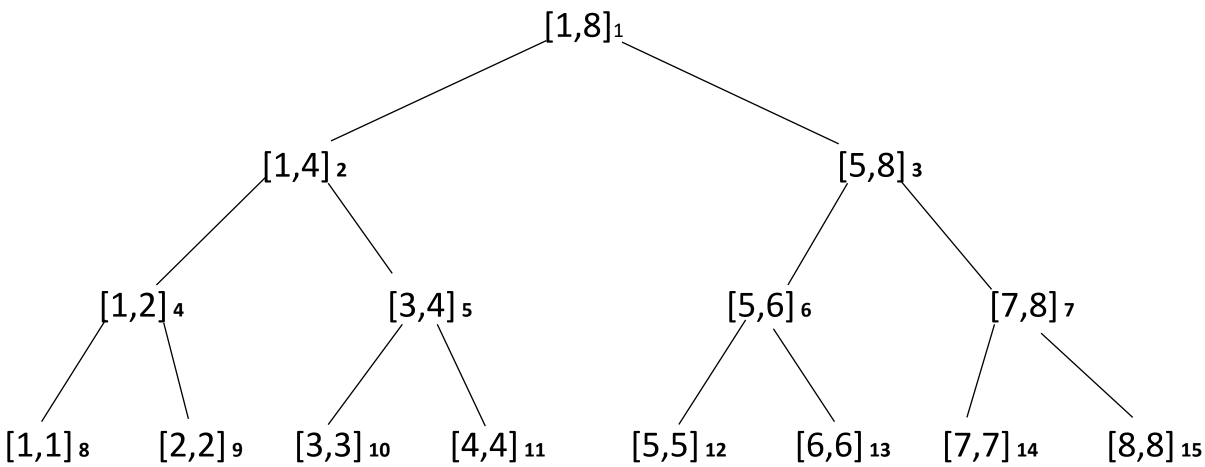
2. SOLUTION

The baseline solution consists of just sum all the counters $F'[i]\ a \leq i \leq b$, but since each counter has some error, if we sum up all the counters, then also the final error will grow linearly as large the range that is $b - a + 1$.

To assure that the error does not increase linearly to the size of the range but logarithmically, we can maintain counters for the sums of ranges which length is a power of 2. The idea is to make estimations for ranges and not for each single item. A way to build the range is the following: for each $i \leq n$, we have all the ranges $[i,\ i + 2^y - 1]$ such that $i + 2^y - 1 \leq n$ for $y > 0$. In other words, for each starting position i we have all the possible ranges with length of power of 2, up to the end. The number of these possible ranges is at most $\log n$, so the total number of ranges is indeed $O(n \log n)$.

Dyadic ranges

To use a smaller number of ranges we can make use of dyadic ranges. A range $[a, b]$ can be split into ranges of length of power of 2 called dyadic ranges. An illustration with $n = 8$ is the following:



The set of dyadic ranges is $\sim 2n$.

The idea is to use $\log n$ count_min sketch, one for each level of the tree above. With this approach we still have one count-min sketch to estimate the number of times each element occurs, but we have also count-min sketches to estimate the number of elements in ranges.

UPDATE AND QUERY OPERATIONS

If we want to update the value of some $i \in [1, n]$ we can no longer do simply $F[i] + +$ as before, but we need to traverse the tree updating the counters in the ranges where the element i appears. For example, if we need to update the element 5 in the tree above, the ranges to be updated are: $(1,8), (5,8), (5,6), (5,5)$. The update operation then it takes $O(\log n)$ time.

For the query operation we need to find the minimum set of non-overlapping dyadic ranges that covers the range of the query and sum the values of their estimation. For example, for the query(3,6) the ranges $(3,4), (5,6)$ are taken and their estimation is summed up. The query operation takes $O(\log n)$ time.

The error grows logarithmically instead than liner(case of baseline), this is because each interval of size n can be represented using at most $2 \log n$ non-overlapping dyadic ranges, this is because at each level at most two dyadic ranges can be taken.

ERROR ANALYSIS

Let's first declare D as the set of all the dyadic ranges, and $D'_{[a,b]} \subset D$ as the minimal set of non-overlapping ranges that cover the range $[a, b]$.

Let's have a fixed hash function j of the count-min sketch which estimates the range $[a, b] \in D$. We define an indicator variable $I_{j,ab,cd}$:

$$I_{j,ab,cd} = \begin{cases} 1, & if\ h_j([a,b]) = h_j([c,d]) \wedge\ [a,b] \neq [c,d] \\ 0, & otherwise \end{cases}$$

So, the indicator variable is equal to 1, if the hash function has a collision between two different ranges.

With the indicator variable just defined we can represent with $Y_{j,ab}$ the residual (i.e the error for $[a,b]$ because of the collision) for a dyadic range $[a,b]$ and a hash function j:

$$Y_{j,ab} = \sum_{\substack{[c,d]\in D\ \wedge\ [a,b]\neq[c,d] \\ |a-b|=|c-d|}} I_{j,ab,cd} F_{[c,d]}$$

We can analyze now the expected error of a dyadic range $[a,b]$ using the expression of $Y_{j,ab}$.

$$E[\,Y_{j,ab}] = E\left[\sum_{\substack{[c,d]\in D\ \wedge\ [a,b]\neq[c,d] \\ |a-b|=|c-d|}} I_{j,ab,cd} F_{[c,d]}\right] = \sum_{\substack{[c,d]\in D\ \wedge\ [a,b]\neq[c,d] \\ |a-b|=|c-d|}} E\big[I_{j,ab,cd} F_{[c,d]}\big]$$

$$= \sum_{\substack{[c,d]\in D\ \wedge\ [a,b]\neq[c,d] \\ |a-b|=|c-d|}} \Pr\,[\,I_{j,ab,cd} = 1] * F_{[c,d]} \ \leq \sum_{\substack{[c,d]\in D\ \wedge\ [a,b]\neq[c,d] \\ |a-b|=|c-d|}} \frac{\varepsilon}{e} * F_{[c,d]}$$

$$= \frac{\varepsilon}{e} \sum_{\substack{[c,d]\in D\ \wedge\ [a,b]\neq[c,d] \\ |a-b|=|c-d|}} F_{[c,d]} \leq \frac{\varepsilon}{e} \|F\|$$

In this way we computed the expected error for a dyadic range. Now the expected error for a generic range $[a,b]$ which can be expressed using at most $2\log n$ dyadic ranges, can be defined as follows:

$$X_{jab} = \sum_{[c,d]\in D'_{[a,b]}} Y_{jcd}$$

$$E = [X_{jab}] = E\left[\sum_{[c,d]\in D'_{[a,b]}} Y_{jcd}\right] = \sum_{[c,d]\in D'_{[a,b]}} E[Y_{jcd}] \leq \sum_{[c,d]\in D'_{[a,b]}} \frac{\varepsilon}{e}\|F\| = 2\log n\ \frac{\varepsilon}{e}\|F\|$$

The estimated counter for the query $[a,b]$ given a fixed hash function j is:

$$F'_{[a,b]} = F_{[a,b]} + X_{jab}$$

Finally, we can use Markov's inequality to bound the probability.

$$
\begin{aligned}
\Pr\left[\ [\forall j\ \in [r]: F'_{[a,b]} \geq\ F_{[a,b]} +\ 2\log n\,\varepsilon\|F\|\right] &= \prod_{j=1}^{r} Pr\left[F_{[a,b]} +\ X_{jab} \geq\ F_{[a,b]} +\ 2\log n\,\varepsilon\|F\|\right] \\
&= \prod_{j=1}^{r} Pr\left[X_{jab} \geq\ 2\log n\,\varepsilon\|F\|\right] \\
&\leq \prod_{j=1}^{r} \frac{E\left[X_{jab}\right]}{2\log n\,\varepsilon\|F\|} \\
&\leq \frac{2\log n\ \frac{\varepsilon}{e}\|F\|}{2\log n\,\varepsilon\|F\|} \\
&= \prod_{j=1}^{r} \frac{1}{e} = \left(\frac{1}{e}\right)^{r}
\end{aligned}
$$