Iowa State University, Computer Science Department

Computer Science 535
Project 4

Jason Hu, Joo Seung Song

December 11, 2019

1 Count min sketch

1.1 Frequency estimation

We tried to empirically estimate $\Pr[\hat{f_x} \geq f_x + \epsilon N]$, but the probability is always zero, so we decide to measure error in a different way.

To see which frequencies are well approximated, we conduct experiments as follows. For every experiment, we fix the size of the stream to be $N = 2000000$, and we sample element e from interval [0,1234411] to be inserted for f_e times where f_e is uniformly sampled from interval [0, 1000]. Within the stream, we insert an element e' sampled from interval [1234411,1235411] to be inserted k times, where k variable unique to each experiment. We conduct 7 experiments in total for $k \in \{1, 10, 100, 1000, 10000, 100000, 1000000\}$. For every k, we repeatedly generate 100 streams to calculate the expected error rate. The error rate is calculated as $E = \hat{f_{e'}} - f_{e'}/N$. By algorithm design, we expected $E < \epsilon$, which we set to be $\epsilon = 0.2$ and $\delta = 0.5$ for all experiments. The results are presented in the following table.

$f_{e'}$	1	10	100	1000	10000	100000	1000000
E	0.0849	0.0861	0.0841	0.0866	0.0833	0.0796	0.0419

As the frequency increases, the estimation error reduces. The CMS has better estimation for large frequency items.

1.2 Heavy hitter

We implemented the heavy hitter correctly, so all q-heavy hitters are contained in the returned set (perfect recall). To measure the quality of the approximated heavy hitters, we calculate the precision as well as the probability that $y \in B$ for returned set B and y such that $f_y < r$. In all experiments, we have $q = 2\epsilon$ and $r = \epsilon$

All experiment settings are the same except the following. The universe size is $U = 1234411$. Frequency of an element is sampled from interval [0,1000]. We vary stream size N, error margin ϵ and error probability δ. The results are presented in the following table.

N	ϵ	δ	hh count	precision	$\hat{\Pr}[y \in B \mid f_y < r]$
10^6	10^{-3}	0.5	15	0	0
10^6	10^{-5}	0.5	1935	1	0
10^5	10^{-3}	0.5	160	1	0
10^5	10^{-4}	0.5	196	1	0
10^4	10^{-3}	0.5	22	1	0
10^4	10^{-4}	0.5	23	1	0

We see that the heavy hitter estimation is generally very robust. The first row of the table occurred because ϵ is so large that there was no heavy hitter in the stream. For all experiments, the heavy hitters were retrieved with perfect precision.

2 Count sketch

2.1 Frequency estimation

Like count min sketch, we repeat the same set of experiments for $k \in \{1, 10, 100, 1000, 10000, 100000, 1000000\}$. The results are presented in the following table.

$f_{e'}$	1	10	100	1000	10000	100000	1000000
E	0.01092	0.00984	0.01049	0.01090	0.00989	0.00930	0.00550

Similar to the count min sketch, the count sketch also performs better for estimating large frequency items. Note that we did not compute or estimate F_2 here, as we used the same error metrics $E = |\hat{f}_{e'} - f_{e'}|/N$. But because the count min sketch and count sketch use different memory, we will need more experiments to compare the two.

3 Same memory comparison

We select the parameters in a way that uses the almost the same amount of memory (5336 integers for CMS and 5340 integers for count sketch) and compare the performance of count min sketch and count sketch. CMS has $\epsilon = 0.0015$, $\delta = 0.046$. Count sketch has $\epsilon = 0.053$, and $\delta = 0.0184$. The stream is created with similar procedures as above, where $N = 2 * 10^6$ and sample frequency

$f_{e'}$	1	10	100	1000	10000	100000	1000000
CMS E	0.00012	0.00016	0.00011	0.00016	0.00014	0.00013	0.00002
Count Sketch E	0.00169	0.00186	0.00175	0.00178	0.00187	0.00170	0.00089

is from interval $[0, 1000]$. The error is measured by comparing with the stream's L0 norm as experiments above: $E = \hat{f}_{e'} - f_{e'}/N$. The results are presented in the following table.

As we see from the table, the count min sketch is more accurate than the count sketch with the same amount of memory.

4 Count sketch

The space dataset from PA2 was used to create the document term frequency matrix, where each document's term frequency vector was created and reduced by the AMS dimensionality reduction algorithm. For parameters, $\epsilon = 0.05$ and $\delta = 0.05$.

All pairs of documents d_i and d_j have original distance computed $L_2(d_i, d_j)$ as well as the reduced distance $L_2(r(d_i), r(d_j))$. We computed the mean error according to the formula

$$E = \frac{|L_2^2(d_i, d_j) - L_2^2(AMS(d_i), AMS(d_j))|}{L_2^2(d_i, d_j)}$$

, which is $E = 0.899$, far greater than ϵ. We also calculated the percentage of documents with $E > \epsilon$, which is 0.986, far greater than δ. We do not understand how the L2 distance reduction preserves the distance of document pairs.

To prove that our AMS reduction algorithm is implemented correctly, we tested if the original document vector's L2 norm is similar to the reduced vector's L2 norm. We computed E for every document d_i according to the formula given in the lecture

$$E = \frac{|L_2^2(d_i) - L_2^2(AMS(d_i))|}{L_2^2(d_i)}$$

The average E is 0.000947, smaller than ϵ. We calculated the percentage of documents with $E > \epsilon$, which is 0.000996, smaller than δ. This shows that the AMS algorithm is implemented correctly.