数据科学与工程算法基础 习题5

10195501436 龚敬洋

3

输入	操作	结果
b	插入	$F=\{(b,1)\}$
a	插入	$F=\{(b,1),(a,1)\}$
c	插入，删除	$F=\{(b,1),(a,1),(c,1)\}$，$F=\{\}$
a	插入	$F=\{(a,1)\}$
d	插入	$F=\{(a,1),(d,1)\}$
e	插入，删除	$F=\{(a,1),(d,1),(e,1)\}$，$F=\{\}$
a	插入	$F=\{(a,1)\}$
f	插入	$F=\{(a,1),(f,1)\}$
a	更新	$F=\{(a,2),(f,1)\}$
d	插入，删除	$F=\{(a,2),(f,1),(d,1)\}$，$F=\{(a,1)\}$

4

易知

$$E(\hat{f}_{a_j})=f_a, Var(\hat{f}_{a_j})=\frac{||f_a||_2^2}{k}$$

故由Chebyshev不等式可知

$$P\left(|\hat{f}_{a_j}-f_a|>\varepsilon f_a\right)\le\frac{Var(\hat{f}_{a_j})}{\varepsilon^2 f_a^2}=\frac{1}{k\varepsilon^2}(j=1,2,\cdots,k)$$

其中 $\hat{f}_{a_j}$ 为单次Basic Count Sketch算法的输出值，f_a 为真实值

由上式易见当 $k=\mathcal{O}(1/(\varepsilon^2\delta))$ 时，$\hat{f}_{a_j}$ 偏离 εf_a 概率小于 $1/3$

定义

$$Y_i=\begin{cases}1, \left|\dfrac{1}{k}\displaystyle\sum_{j=1}^{k}\hat{f}_{a_j}-f_a\right|>\varepsilon f_a\\ 0, otherwise\end{cases}$$

则 $E(Y_i)=P(Y_i=1)<\frac{1}{3}$

若运行 t 次Basic Count Sketch算法，其失败次数的期望不会超过 $t/3$。进一步的，若Count Sketch算法失败，则中位数左边或右边的都失败，也即至少有一半的Basic Count Sketch失败。

由Chernoff不等式可知

$$
\begin{aligned}
P\left(\sum_{i=1}^{t} Y_i > \frac{t}{2}\right) &= P\left(\sum_{i=1}^{t} Y_i > \left(1+\frac{1}{2}\right)\frac{t}{3}\right) \\
&\leq P\left(\sum_{i=1}^{t} Y_i > \left(1+\frac{1}{2}\right)\mu\right) \\
&\leq \exp\left\{-\frac{1}{4}\cdot\mu\cdot\left(\frac{1}{2}\right)^2\right\} \\
&< \delta
\end{aligned}
$$

因此

$$
\frac{t}{3} \leq \mu \leq 16\ln\frac{1}{\delta}
$$

也即 $t = \mathcal{O}(\log(1/\delta))$

5

此时

$$
E(\hat{f}_a) = \frac{1}{t}\sum_{i=1}^{t} E(\hat{f}_{a_i}) = \frac{1}{t}\cdot t\cdot\hat{f}_a = \hat{f}_a
$$

$$
Var(\hat{f}_a) = \frac{1}{t^2}\sum_{i=1}^{t} Var(\hat{f}_{a_i}) = \frac{1}{t^2}\cdot t\cdot\frac{||f_{-a}||_2^2}{k} = \frac{||f_{-a}||_2^2}{tk}
$$

故由Chebyshev不等式可知

$$
\begin{aligned}
P\left(|\hat{f}_a - f_a| \geq \varepsilon||f||_2\right) &\leq P\left(|\hat{f}_a - f_a| \geq \varepsilon||f_{-a}||_2\right) \\
&\leq \frac{Var(\hat{f}_a)}{\varepsilon^2||f_{-a}||_2^2} \\
&= \frac{1}{tk\varepsilon^2} \\
&< \delta
\end{aligned}
$$

又 $t = \log(1/\delta)$

因此

$$
k = \mathcal{O}\left(\frac{1}{\delta\varepsilon^2}\right)
$$

6

(1)

数据流	4	1	3	5	1	3	2	6	7	0	9
$h_1(x) = (3x+2) \mod 8$	6	5	3	1	5	3	0	4	7	2	5
$h_2(x) = (7x+5) \mod 8$	1	4	2	0	4	2	3	7	6	5	4
$h_3(x) = (5x+3) \mod 8$	7	0	2	4	0	2	5	1	6	3	0

因此CM Sketch矩阵为

$$\begin{pmatrix} 1 & 1 & 1 & 2 & 1 & 3 & 1 & 1 \\ 1 & 1 & 2 & 1 & 3 & 1 & 1 & 1 \\ 3 & 1 & 2 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

可知

$$\begin{aligned} \hat{f}_0 &= \min(1,1,1) = 1 \\ \hat{f}_1 &= \min(3,3,3) = 3 \\ \hat{f}_2 &= \min(1,1,1) = 1 \\ \hat{f}_3 &= \min(2,2,2) = 2 \\ \hat{f}_4 &= \min(1,1,1) = 1 \\ \hat{f}_5 &= \min(1,1,1) = 1 \\ \hat{f}_6 &= \min(1,1,1) = 1 \\ \hat{f}_7 &= \min(1,1,1) = 1 \\ \hat{f}_9 &= \min(3,3,3) = 3 \end{aligned}$$

故CM Sketch估计的频繁项为 1 和 9

(2) 由（1）可知算法对元素 $0,2,3,4,5,6,7$ 的计数时准确的，但由于 $h(1) = h(9) \mod 8$，导致计数器的每一行 1 和 9 都产生冲突，因此对元素 1 和 9 的计数偏大

(3) 将哈希函数的个数增加至 $\lceil \log(1/\delta) \rceil$，哈希表的大小增加为 $2/\varepsilon$

现希望有 $1-\delta$ 的概率使得 $\hat{f}_a - f_a \le \varepsilon n$

例如取 $\delta = 0.05, \varepsilon = \frac{2}{11}$，则

$$\begin{aligned} w &= \left\lceil \frac{2}{\varepsilon} \right\rceil = 11 \\ d &= \left\lceil \ln \frac{1}{0.05} \right\rceil = 3 \end{aligned}$$

因此可将哈希函数修改为

$$\begin{aligned} h(x) &= (3x+2) \mod 11 \\ h(x) &= (7x+5) \mod 11 \\ h(x) &= (5x+3) \mod 11 \end{aligned}$$