Algorithm Foundations of Data Science and Engineering
Welcome Tutorial :-)
Tutorial 3

GAO Ming

DaSE @ ECNU

11 Mar., 2019

1、 In count sketch for item frequency, the algorithm returns

$$\widehat{f_a} = \text{median}_{1 \le i \le t} g_i(a) C[i][h_i(a)]$$

for a query a. Please give reason for $t = O(\log(1/\delta))$.

解：易证，

$$E\widehat{(f_a)} = f_a, \mathrm{var}\widehat{(f_a)} = \frac{\|f_{-a}\|_2^2}{k}$$

由切比雪夫不等式，得

$$p(\left|\hat{f}_a - f_a\right| \ge \varepsilon \|f\|_2) \le p(\left|\hat{f}_a - f_a\right| \ge \varepsilon \|f_{-a}\|_2) \le \frac{\mathrm{var}\hat{(f_a)}}{\varepsilon^2 \|f_{-a}\|_2^2} = \frac{1}{k\varepsilon^2}$$

令 $\frac{1}{k\varepsilon^2} = \frac{1}{3}$，定义

$$Y_i = \begin{cases} 1, \left|\hat{f}_a - f_a\right| \ge \varepsilon \|f\|_2 \\ 0, otherwise \end{cases}$$

则有$P(Y_i=1)\le\frac{1}{3}$，记$\mu=E(\sum_{i=1}^{t}Y_i)\le\frac{t}{3}$ 由chernoff bound得，

$$p(\sum_{i=1}^{t}Y_i>\frac{t}{2})\le p(\sum_{i=1}^{t}Y_i>(1+\frac{1}{2})\mu)\le\exp(-\frac{\mu}{16})<\delta$$

又因为

$$\exp(-\frac{\mu}{16})\ge\exp(-\frac{t}{48})$$

所以，

$$\exp(-\frac{t}{48})<\delta$$

所以$t=O(\log\frac{1}{\delta})$，得证。

2、For the counting sketch algorithm, say the last line is changed from "On query a, report $\widehat{f_a}=\text{median}_{1\le i\le t}g_i(a)C[i][h_i(a)]$" to "On query a, report $\widehat{f_a}=\frac{\sum_{i=1}^{t}g_i(a)C[i][h_i(a)]}{t}$". The rest of the algorithm is kept as it is. Analyze the performance of this modified algorithm.

解：若$\widehat{f}_a = \frac{\sum_{i=1}^{t} g_i(a) C[i][h_i(a)]}{t}$ 则其方差为

$$\mathrm{var}(\hat{f}_a) = \frac{\|f_{-a}\|_2^2}{tk}$$

由切比雪夫不等式，得

$$p(\left|\hat{f}_a - f_a\right| \geq \varepsilon \|f\|_2) \leq p(\left|\hat{f}_a - f_a\right| \geq \varepsilon \|f_{-a}\|_2) \leq \frac{\mathrm{var}(\hat{f}_a)}{\varepsilon^2 \|f_{-a}\|_2^2} = \frac{1}{tk\varepsilon^2} < \delta$$

所以

$$t = O(\frac{1}{\delta\varepsilon^2})$$

3、Given the input streaming $b, a, c, a, d, e, a, f, a, d$, and $k = 3$, i.e., three counters. Please write down the executing process step by step and find the result of the Misra-Gries summary.

解：step1:input=b,opration=add,result为$F = \{(b, 1)\}$

step2:input=a,opration=add,result为$F = \{(b, 1), (a, 1)\}$

step3:input=c,opration=add+delete,result为$F=$

$\{(b,1),(a,1),(c,1)\}$->$F=\{\}$

step4:input=a,opration=add,result为$F=\{(a,1)\}$

step5:input=d,opration=add,result为$F=\{(a,1),(d,1)\}$

step6:input=e,opration=add+delete,result为$F=$

$\{(a,1),(d,1),(e,1)\}$->$F=\{\}$

step7:input=a,opration=add,result为$F=\{(a,1)\}$

step8:input=f,opration=add,result为$F=\{(a,1),(f,1)\}$

step9:input=a,opration=update,result为$F=\{(a,2),(f,1)\}$

step10:input=d,opration=add+delete,result为$F=$

$\{(a,2),(f,1),(d,1)\}$->$F=\{(a,1)\}$

the result of the Misra-Gries summary is $F=\{(a,1)\}$。

4、 From your opinion,

- Is the Misra-Gries summary mergable? That is, two summaries of different inputs of size k can be combined together to obtain a new summary of size k that summarizes the union of the two inputs.

解：能。不妨设两个输入流的数据量分别为n_1和n_2，

case 1：合并两个输入流的summaries。第一个输入流中出现次数大于$\frac{n_1}{k}$的项一定出现在summary，第二个输入流中出现次数大于$\frac{n_2}{k}$的项一定出现在summary，将两者的summary合并。

case 2：两个输入流的并的summaries。数据量为$n_1 + n_2$，其中出现次数大于$\frac{n_1+n_2}{k}$的项一定出现在summary。

易得，出现在case 2的summaries中的出现次数大于$\frac{n_1+n_2}{k}$的项一定出现在case 1的summaries，所以Misra-Gries summary是可合并的。

- Is the Misra-Gries summary suitable to be used in distributed and parallel environments?

  解：因为Misra-Gries summary是可合并的，所以能够在分布式和并行环境中使用。