Seoul National University

M1522.001400-001 Introduction to Data Mining

Homework 2: Mining Data Streams (Chapter 4)

2017-18538 Hwang SunYoung

1.

(1) We can assume each tuple has different key.

The probability that include tuple 'f' in 20% of samples is 1- (The probability that does not include tuple 'f' in 20% of samples). The number of tuple 'f' is two. The probability that does not include 'one' 'f' is $\frac{8}{10}$.

So the probability that include tuple 'f' in 20% of samples is

$$1-\left(\frac{8}{10}\right)\left(\frac{8}{10}\right)=0.36$$

(2) a, b, e are in the three buckets at t2.

(3) 3 buckets = s, 6 data = n → $\frac{s}{n}=\frac{3}{6}$. this probability is that we keep 6th element(t5)

The probability that replace 'a' in the sample is $\frac{1}{3}$ (there are 3 buckets to store the sampled data and sample data has one 'a')

$(\frac{3}{6})(\frac{1}{3})=\frac{1}{6}$

The probability that the tuple 'a' is replaced at t5 is $\frac{1}{6}$

2.

(1) To calculate the estimated number of 1s in the last 15 bits, sum the size of all buckets but the last and add half the size of the last bucket.

1+1+2+2+$\frac{4}{2}$=8

(2) To calculate the estimated number of 0s in the last 15 bits,

15- (the estimated number of 1s in the last 15 bits) = 15-8 = 7

(3)

...0 0 1 0 0 1 0 0 1 0 0 1 / 0 1 1 1 1 0 0 0 0 1 0 1 0 1 0

$1+1+2+2+\frac{4}{2}=8$

The estimated number of 1's in the last 15 bits is 8.

(4)

...0 0 1 0 0 1 0 0 1 0 0 1 / 0 1 1 1 1 0 0 0 0 1 0 1 0 1 1

...0 0 1 0 0 1 0 0 1 0 0 1 / 0 1 1 1 1 0 0 0 0 1 0 1 0 1 1 ←merge

We do not merge the leftmost bucket because the size of leftmost bucket is 8.

$1+2+4+\frac{4}{2}=9$

The estimated number of 1's in the last 15 bits is 9.

3.

(1) Because each n URL hash k times, the time complexity of checking n URLs is n*k*O(1)=O(nk).

(2) |S|=5 billion = m, |B| = 8 billion = n.

Optimum number of hash functions is $\frac{8}{5}ln2$ = 1.109 ≒ 1.

(3) |S|=1 billion = m, |B| = 8 billion = n.

(False positive probability) = $(1-e^{-km/n})^k$ k is the number of hash functions.

$(1-e^{-k/8})^k \leq 0.05$

If k=1, $(1-e^{-1/8})^1 = 0.118 > 0.05$

If k=2, $(1-e^{-1/4})^2 = 0.049 < 0.05$

The minimum number of hash functions is 2.

4.

(1) the number of a: 4, the number of b: 3, the number of c: 2, the number of d: 4

S= 4^2+3^2+2^2+4^2 = 45

The second moment(surprise number) of the stream is 45.

(2) n=13, c=X.val

4^{th} X.el = d, X.val = 4

f(X)=13(3*4^2-3*4+1) = 481

5^{th} X.el = a, X.val = 3

f(X)=13(3*3^2-3*3+1) = 247

10^{th} X.el = c, X.val = 1

f(X)=13(3*1^2-3*1+1) = 13

E[f(X)] = (481+247+13)/3=247

The third moment of the stream using AMS method is 247.