sample < fixed portion / fixed size. Filtering (Bloom). Counting < distinct element / distinct moments sliding window

一、sampling

1. fixed portion sampling.

i) naive solution. s queries once, d queries twice, truly fraction of query with duplicates = $\frac{d}{d+s}$

however. we have $\frac{\frac{d}{100}}{\frac{s}{10}+\frac{d}{100}+\frac{19d}{100}} = \frac{d}{10s+19d} \neq \frac{d}{d+s}$

ii) solution: hash key and select 1/e

2. fixed size. sampling

sample size: s.

i ≤ s 时 都进 sample

i > s 时. ith 每个 S_i 的 prob be sample → if keep, kick off one with prob $\frac{1}{s}$

new arrive ele x already is S. prob(keep) $= (1-\frac{s}{n+1}) + \frac{s}{n+1}\cdot\frac{s-1}{s} = \frac{n}{n+1}$

二、bloom Filtering

1. check whether have seen new arrival before → no false negative
   is BF says no, new item not in set.
   but many false positive, if say yes, 不定 in set.

2. component. { A: array of n bits [0, 1, ..., n-1]
   set of hash functions. → input element, output 0, 1, ..., n-1.
   set of objects }

3. construct:
   用 hash function h_j 对 S 中的 objective 完成 A 的 n 个 buckets 中. 则有 A[i] = 1.
   对新 objective, 用 $h_1, \dots, h_j$ 这些 hash 若有一个 $h_i(0) = i$, 且 A[i] = 0. 则 0 不在 S 中.

4. false positive. x not seen before, but identified as in S.

   fpr. upper bound. n bits array. k hash functions. m elements. inserted.
   f: fraction of 1s in array

   FPR = f^k. 其中 $f \le \frac{mk}{n}$.

5. estimation of f.

   $\lim_{n\to\infty}(1+\frac{1}{n})^n = e \Rightarrow (1-\frac{1}{t})^t = e^{-1} \Rightarrow (1-\frac{1}{t})^{t\cdot\frac{d}{t}} = e^{-\frac{d}{t}} \Rightarrow f = 1-e^{-\frac{km}{n}}$

   $fpr = f^k = (1-e^{-\frac{km}{n}})^k$.

   optimal $k = \frac{n}{m}\ln 2$.

三、counting = distinct element + distinct moments

1. distinct element.

   For 1 elg. hash element to a long str. ⇒ set R = longest length of trailing zero
   ⇒ estimation $= 2^R$.

   $1-e^{-\frac{m}{2^r}} \sim \frac{m}{2^r}$ if $2^r \gg m$

h(a) for element a has at least r trailing 0's : $p = 1 - e^{-\frac{m}{2^r}}$
no element a has 〃 : $p' = e^{-\frac{m}{2^r}}$

if $2^r \gg m$. $p \to 0, p' = 1 \Rightarrow$ r 不会太大
if $2^r \ll m$. $p \to 1, p' \to 0 \Rightarrow$ r 不会太小 } 2^R is around m.

2. distinct moments.

(1) k^{th}-moments of stream $S = \sum_{i=1}^{n} (m_i)^k$. → n = # of distinct value in S.
m_i = # of occurrences of v_i in S.

(2). 0-th-moments : # of distinct element
1st -moments : length of S.
2nd 〃 : surprise number, 越大越 unbalance.

(3) AMS: estimate 2nd-moments of S.
随机选取下标.
X.value = X在被选中后的位置的occurrences.
estimation $= \frac{n}{k} \sum_{i=1}^{k} (2X_i.\text{value} - 1)$.

(4) estimation of 3rd-moments $= \frac{n}{k} \sum (3X_k.v^2 - 3X_k.v + 1)$.

四. sliding window : counting # of 1's in window