L11: Streaming : Frequent Items and Quantiles

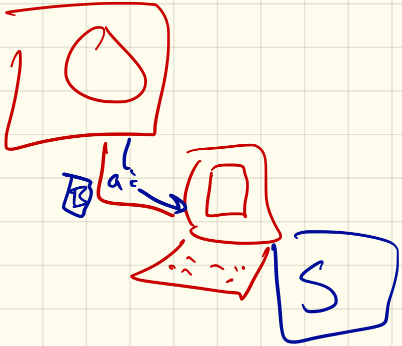
Jeff M. Phillips

February 20, 2019

Big Data

- Sampling
- Streaming: One Pass
- MapReduce

  Distributed



Streaming

m, n too big

Data: $A = \langle a_1, a_2, \dots, a_i, \dots, a_m \rangle$

$a_i \in [n]$ Domain

$[n]$ = IP address

= words in dictionary

k-gram

or size

counter = $O(\log m)$ bits

label = $O(\log n)$ bits

hash tables (hash function)

$O(\log n + \log m)$

Streaming Model

$$A = \langle a_1, a_2, \ldots, a_j, \ldots a_m \rangle$$

$a_j \in [n] \leftarrow$ Domain

m, n very large

Space: $(\log m + \log n)$ counter label

frequencies

$$f_j = |\{a_i \in A \mid a_i = j\}|$$

$$F_0 = \sum_i f_j^0 = \#\text{ distinct elements}$$

$$F_1 = \sum_j f_j^1 = m = \#\text{ elements}$$

Hyperloglog

$$F_2 = \sqrt{\sum_j f_j^2} = \text{join size}$$

MAJORITY

Is one IP address on more than half of all packets?

Is some $f_j > m/2$?

If so, which one?

Then report j s.t. $f_j > m/2$.

If not, guess.

If not, return any $j \in [n]$.

Majority

counter

label

Majority(A)

Set $c = 0$ and $\ell = \emptyset$

for $i = 1$ **to** m **do**

  if $(a_i = \ell)$ **then**

    $c = c + 1$ increment

  else

    $c = c - 1$ decrement

  if $(c < 0)$ **then**

    $c = 1$, $\ell = a_i$

return ℓ

Heavy Hitters / Frequent Items

Report all $f_i \geq m/k$

Freq Item

For all $i \in [n]$ have $\hat{f}_j$ $\leftarrow$ aprox

$$\underbrace{\overbrace{f_j - \frac{m}{k} \leq \hat{f}_j \leq f_j + \frac{m}{k}}^{\text{CMS}}}_{\text{MG}}$$

$k = \frac{1}{\varepsilon}$

$\hookrightarrow \frac{m}{k} = \varepsilon m$

$\varepsilon = \%$ error

$\varepsilon = 0.01 \Rightarrow 1\%$ error

$\hookrightarrow k = 100$

Misra – Gries Algo

Cache: $k-1$ counters
$k-1$ labels ← k guesses

$C[1] \quad C[2] \quad \cdots \quad C[k]$

$L[1] \quad L[2] \quad \cdots \quad L[k]$

- if $a_j = L[j']$ $\quad C[j']{+}{+}$
- else and some $C[j'] = 0$ ← empty counter
  then $L[j'] = a_j \quad C[j'] = 1$
- else Decrement all counters
  $\forall j' \in [k] \quad C[j'] \leftarrow -$

Total # Decrements $\leq \frac{m}{k}$

Misra-Gries

counter array C : $C[1], C[2], \ldots, C[k-1]$
location array L : $L[1], L[2], \ldots, L[k-1]$

<u>Misra-Gries(A)</u>

Set all $C[i] = 0$ and all $L[i] = \emptyset$
for $i = 1$ **to** m **do**
  if $(a_i = L[j])$ **then**
    $C[j] = C[j] + 1$
  else
    if (some $C[j] = 0$) **then**
      Set $L[j] = a_i$ & $C[j] = 1$
    else
      for $j \in [k-1]$ **do** $C[j] = C[j] - 1$
return C, L

$f_j - \frac{m}{k} \leq \hat{f}_j \leq f_j$

Total # Decrements

$m = 1000$ $k = 100$

$\hat{f}_j = \cancel{150}$ 140

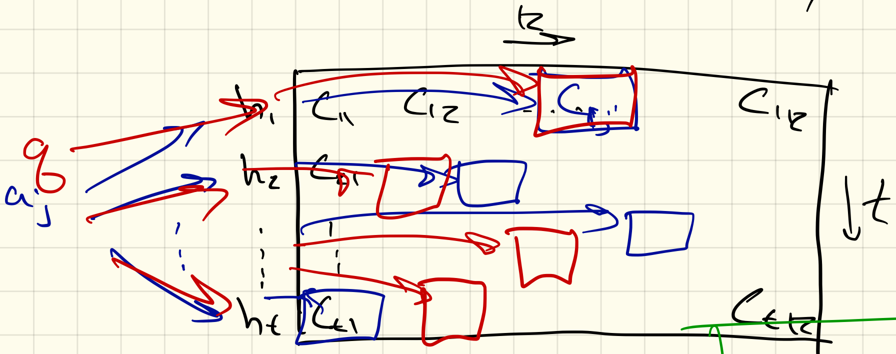
Q. is $f_j > 0.15 \cdot m$?

Count – Min Sketch

turnstyle model

$k \cdot t$ counters , t hash functions

$h : [n] \to [k]$



$\tilde{O}(\log n + \log m)$ space

$k = \frac{2}{\varepsilon}$ $\quad t = \log \frac{1}{\delta}$

prob. failure

for $a_j \in A$

for i in $1 \dots t$

$C_{i,h_i(j)}$ ++

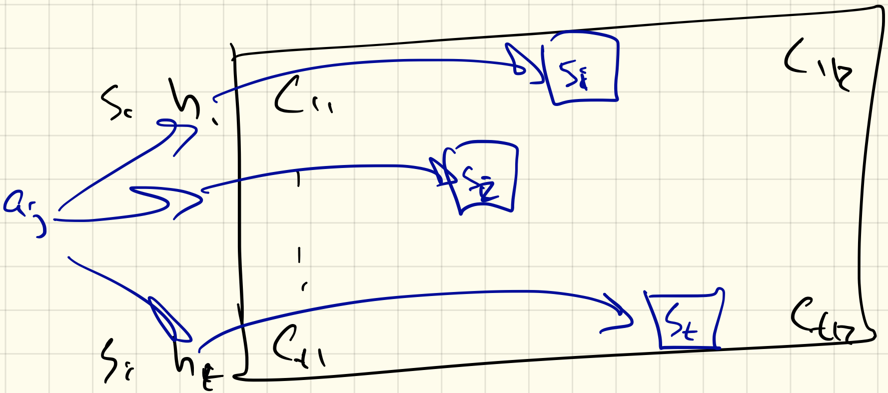
linear sketch

Query $q \in [n]$

$$\hat{f}_q = \min_i C_{i,h_i(q)}$$

$$f_q \le \hat{f}_q \le f_q + \varepsilon m \quad \text{w.p.} > 1-\delta$$

Count Sketch



$t \cdot k$ counters

t hash fxns

t random sign hashs

$s_i : [n] \to \{-1, +1\}$

Query q

$$\hat{f}_q = \underset{i}{\text{median}}\left(s_i(q) \cdot C_{i,h_i(q)}\right)$$

$k = O(1/\varepsilon^2)$

$t = \log 1/\delta$

$$E[\hat{f}_q] = f_q$$

$$|\hat{f}_q - f_q| \le \varepsilon \cdot F_2$$

for $a_j \in A$

  for i in $1 \ldots t$

$$C_{i,h_i(j)} = C_{i,h_i(j)} + s_i(j)$$

Frugal Median

Frugal Median(A)

Set $\ell = 0$.
for $i = 1$ **to** m **do**
  if $(a_i > \ell)$ **then**
    $\ell \leftarrow \ell + 1$.
  if $(a_i < \ell)$ **then**
    $\ell \leftarrow \ell - 1$.
return ℓ.

Frugal Quantile

Frugal Quantile(A, ϕ) e.g. $\phi = 0.75$

Set $\ell = 0$.
for $i = 1$ **to** m **do**
  $r = \text{Unif}(0, 1)$ (at random)
  if $(a_i > \ell$ **and** $r > 1 - \phi)$ **then**
    $\ell \leftarrow \ell + 1$.
  if $(a_i < \ell$ **and** $r > \phi)$ **then**
    $\ell \leftarrow \ell - 1$.
return ℓ.

Frequent Itemsets : Apriori

$A = \{T_1, T_2, \ldots T_m\}$

$$T_1 = \{1, 2, 3, 4, 5\}$$
$$T_2 = \{2, 6, 7, 9\}$$
$$T_3 = \{1, 3, 5, 6\}$$
$$T_4 = \{2, 6, 9\}$$
$$T_5 = \{7, 8\}$$
$$T_6 = \{1, 2, 6\}$$
$$T_7 = \{0, 3, 5, 6\}$$
$$T_8 = \{0, 2, 4\}$$
$$T_9 = \{2, 4\}$$
$$T_{10} = \{6, 7, 9\}$$
$$T_{11} = \{3, 6, 9\}$$
$$T_{12} = \{6, 7, 8\}$$