A STORY OF DISTINCT ELEMENTS

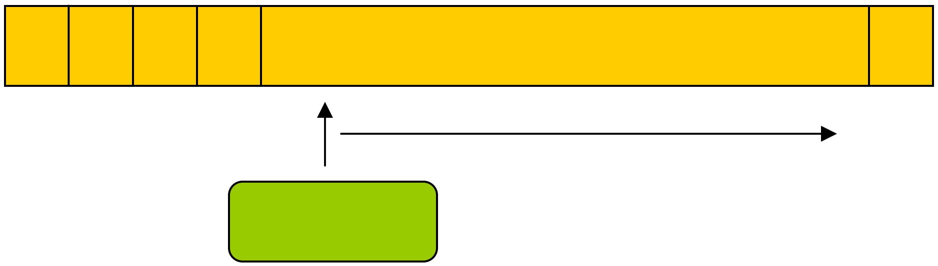
Ravi Kumar
Yahoo! Research
Sunnyvale, CA
ravikumar@yahoo-inc.com

ε results about F_0

(This represents joint works with Bar-Yossef, Jayram, Sivakumar, Trevisan)

Data stream model

Modeling efficient computation on massive data



Compute a function of inputs $X = x_1, \ldots, x_n$

Approximate, randomize, and be space-efficient!

Finding distinct elements

- Given $X = x_1, \ldots, x_n$ compute $F_0(X)$, the number of distinct elements in X, in the data stream model

  Assume $x_i \in [m]$
- (ε,δ)-approximation: Output $F'_0(X)$ such that with probability at least $1 - \delta$, $F'_0(X) = (1 \pm \varepsilon) F_0(X)$
- Zeroth frequency moment
- Assume $\log m = O(\log n)$; otherwise hash input
- Sampling needs lots of space
- Without randomization and approximation, this problem is uninteresting

Some applications

- Web analysis
  - How many different queries were processed by the search engine in the last 48 hours?
  - How many non-duplicate pages have been crawled from a given web site?
  - How many unique ads has the user clicked on (or) how many unique users ever clicked a given ad?
- Databases
  - Query selectivity
  - Query planning and execution
- Networks
  - Smart traffic routing

Some previous work

- [Flajolet, Martin]: Assumed ideal hash functions
- [Alon, Matias, Szegedy]: Pairwise independent hashing

  $(2+\varepsilon)$-approximation using $O(\log m)$ space
- [Cohen]: Similar to FM, AMS
- [Gibbons, Tirthapura]: Hashing-based

  ε-approximation using $O(1/\varepsilon^2 \log m)$ space
- [Bar-Yossef, Kumar, Sivakumar]: Hashing-based, range-summable

  ε-approximation using $O(1/\varepsilon^3 \log m)$ space
- [Cormode, Datar, Indyk, Muthukrishnan]: Stable distributions

  ε-approximation using $O(1/\varepsilon^2 \log m)$ space

The rest of the talk

- Upper bounds
- Lower bounds

Upper bounds

What is the goal beyond $O(1/\varepsilon^2 \log m)$ space?

Can we get upper bounds of the form

$$\tilde{O}(1/\varepsilon^2 + \log m)$$

where $\tilde{O}$ hides factors of the form $\log 1/\varepsilon$ and $\log \log m$?

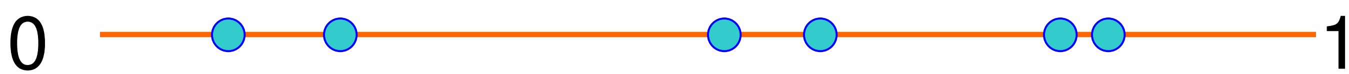
Three algorithms with improved upper bounds

Summary of the bounds

- ALG I: Space $O(1/\varepsilon^2 \log m)$ and time $\tilde{O}(\log m)$ per element
- ALG II: Space $\tilde{O}(1/\varepsilon^2 + \log m)$ and time $\tilde{O}(1/\varepsilon^2 \log m)$ per element
- ALG III: Space $\tilde{O}(1/\varepsilon^2 + \log m)$ and time $\tilde{O}(\log m)$ amortized per element

ALG I: Basic idea

Suppose h:[m] → (0, 1) is truly random



Then min $(h(x_i))$ is roughly ~ $1/F_0(X)$
Reciprocal of this value is $F_0(X)$ [FM, AMS]

More robust: Keep the t-th smallest value v_t
v_t is roughly ~ t/F_0
A good estimator of F_0 is t/v_t

ALG I: Details

$t = 1/\varepsilon^2$; $h:[m] \rightarrow h[m^3]$, pairwise indep.; $T = \emptyset$

for i = 1, …, n do

  $T \leftarrow$ t smallest values in $T \cup h(x_i)$

v_t = t-th smallest value in T

Output $tm^3/v_t = F'_0(X)$

- Space: O(log m) for h and $O(1/\varepsilon^2 \log m)$ for T
- Time: Balanced binary search tree for T

ALG I: Analysis

h is pairwise independent, injective whp

$Y = \{ y_1, \ldots, y_k \}$ be distinct values, $F_0 = k$

Suppose $F'_0 > (1+\varepsilon) F_0$

  means $h(y_1), \ldots, h(y_k)$ has t values smaller than $tm^3/(F_0(1+\varepsilon))$

  Pr[this event] < 1/6 by Chebyshev

Similar analysis for $F'_0 < (1-\varepsilon) F_0$

ALG II: Basic idea

Suppose we know rough value of F_0, say R

Suppose h:[m] → [R] is truly random

Define r = Pr_h[h maps some x_i to 0]

$$r = 1 - \left(1 - \frac{1}{R}\right)^{F_0}$$

If R and F_0 are close, then r is all we need

Estimate R using [AMS]

$$r = \sum_{i=1}^{F_0} (-1)^{i+1} \binom{F_0}{i} R^{-i}$$

Estimate r using sufficiently indep. hash functions

ALG II: Some details

H be (log1/ε)-wise independent hash family

Estimator $p = Pr_{h \in H}$[h maps some x_i to 0]

p matches first log1/ε terms in expansion of r

Chebyshev inequality, inclusion-exclusion

p and r will be close if $1/\varepsilon^2$ estimators (hash functions) are deployed

Create these hash functions from a master hash

ALG III: Basic idea

Overview of algorithm of [GT] and [BKS]

Suppose h: [m] → [m] is pairwise indep.

Let h_t = projection of h onto last t bits

Find min t for which $r = \#\{x_i \mid h_t(x_i) = 0\} < 1/\varepsilon^2$

Output $r\, 2^t$

Can do space-efficiently since if $h_{t+1}(x_i) = 0$ then $h_t(x_i) = 0$ and so can filter

ALG III: Some details

- Space = $1/\varepsilon^2 \log m$
- Obs: Need not store elements explicitly
- Use a secondary hash function g
  - g succinct, injective
  - g suffices to store trailing zeros

- Space: $\log m + 1/\varepsilon^2 (\log 1/\varepsilon + \log \log m)$
- Amortized time: $\tilde{O}(\log m + \log 1/\varepsilon)$

Lower bounds

The general paradigm

- Consider communication complexity of a certain problem
  - One-way
  - Multi-round
- Reduce it to that of computing F_0 in the data stream model
- Obtain one-pass or multi-pass space lower bound

$\Omega(\log m)$ lower bound [AMS]

Reduction from set equality problem

Alice given X, Bob given Y, both m-bit vectors, and the question is if X = Y

- Randomized space bound of $\Omega(\log m)$

X' = φ(X), Y' = φ(Y) where φ is error-correcting code

- YES case: if X = Y, then F_0(X' U Y') = n'
- NO case: if X ≠ Y, then F_0(X' U Y') ~ 2n'

One-pass $\Omega(1/\varepsilon)$ lower bound

Reduction from set disjointness with special instances

Alice has bit vector X with $|X| = m/2$, Bob has bit vector Y with $|Y| = \varepsilon m$

- Treated as sets

  YES instance: X contains Y

  NO instance: $X \cap Y = \varnothing$

- One-pass lower bound [BJKS]: $\Omega(1/\varepsilon)$

$Z = (1, x_1) \ldots (m, x_m)\ (1, y_1) \ldots (m, y_m)$

- YES case: If X contains Y, then $F_0(Z) = m/2$
- NO case: If X and Y are disjoint, $F_0(Z) = m/2 + \varepsilon\, m = m/2(1 + 2\,\varepsilon)$

The gap-hamming problem [IW]

Alice given X, Bob given Y, both m-bit vectors

- Promise
  - YES instance: $h(X, Y) \geq m/2$
  - NO instance: $h(X, Y) \leq m/2 - \sqrt{m}$

Gap-hamming problem: distinguish the two cases in one-pass or multi-round communication model

Gap-hamming captures F0

- $Z = (1, x_1) \ldots (m, x_m)\ (1, y_1) \ldots (m, y_m)$
- $F_0(Z) = 2h(X,Y) + (m - h(X, Y)) = m + h(X,Y)$

- YES case: if $h(X, Y) \geq m/2$ then $F_0(Z) \geq 3m/2$
- NO case: if $h(X, Y) \leq m/2 - \sqrt{m}$ then $F_0(Z) \leq 3m/2 - \sqrt{m} = 3m/2(1 – 1/\sqrt{m})$

Can be shown that $\Omega((\sqrt{m})^c)$ lower bound for gap-hamming leads to $\Omega(1/\varepsilon^c)$ lower bound for F_0

Easy Ω(√m) lower bound for gap-hamming

Reduce from set disjointness of √m size

Alice given X, Bob given Y, both √m-bit vectors, and the question is if $X \cap Y = \emptyset$

- Randomized space bound of Ω(√m) [KS, R]

Each bit in X, Y is expanded to √m bit block so that if $x_i \neq y_i$ then this block has hamming distance √m/2 and if $x_i = y_i$ then has hamming distance 0

- YES case: if $X \cap Y = \emptyset$, then $h(X',Y') = m/2$
- NO case: if $X \cap Y \neq \emptyset$ then $h(X',Y') < m/2 - \sqrt{m}/2$

One-pass $\Omega(m)$ lower bound for gap-hamming [IW, W]

- Indyk and Woodruff, Woodruff showed $\Omega(m)$ lower bound in the one-way model
  - Using VC-dimension and embedding
  - We will show a simpler proof of this result

Reduction from indexing [JKS]

- Alice has n-bit vector T with $|T| = n/2$ and Bob has index i; assume $n/2$ is odd
- Using public randomness, Alice and Bob pick a random n-bit ± 1 vector r
- Alice computes $x = \text{sign}(\langle T, r \rangle)$
- Bob computes $y = \text{sign}(r_i)$
- Now look at the correlation between random variables x and y

Analyzing the correlation

Let $s = \sum_{i \in T} r_i$

n/2 odd implies $\Pr[s < 0] = \Pr[s > 0] = 1/2$

- NO case: If $i \notin T$, then x is independent of y

  so $\Pr[x = y] = \Pr[\text{sign}(s) = \text{sign}(r_i)] = 1/2$

- YES case: If $i \in T$, then let $s = s' + r_i$

  $\Pr[s' = 0] = \eta = c/\sqrt{n}$

  $\Pr[s' < 0] = \Pr[s' > 0] = (1 - \eta)/2$

  $\Pr[x = y] = \Pr[s' = 0] + \Pr[\text{sign}(s') = \text{sign}(r_i) \mid s' \neq 0]$

  $= \eta + (1 - \eta)/2 = (1 + c/\sqrt{n})/2$

Amplifying the gap

- We have random variables x and y with the property that
  - NO case: $\Pr[x = y] = 1/2$
  - YES case: $\Pr[x = y] = 1/2 + c'/\sqrt{n}$
- Repeat with different independent random vectors $r^1, r^2, \ldots, r^t$ to get t-bit vectors X and Y
  - Chernoff shows that if $t = O(n)$ then whp we have
    - NO case: $h(X, Y) \geq (1/2 - c_1)n$
    - YES case: $h(X, Y) \leq (1/2 - c_1)n - c_2\sqrt{n}$

Open problem

- Close the gap between the upper and lower bounds for F_0 for multi-pass algorithms
  - One-pass algorithm with space $O(1/\varepsilon^2)$
  - One-pass lower bound of $\Omega(1/\varepsilon^2)$

- Conjecture: the multi-pass space complexity of F_0 is $\Omega(1/\varepsilon^2)$

thank you!

ravikumar@yahoo-inc.com