Lecture 3 (Advanced Algos) Jan 25, 2022

Hashing

Problem: {Dictionary} • preprocess set $S \subset U$, $|S| = n$.
("universe")

• query: given x, report if "$x \in S$"

Goal: fast query, small space.

U = all IPs. 2^{32}

$S \subseteq U$.

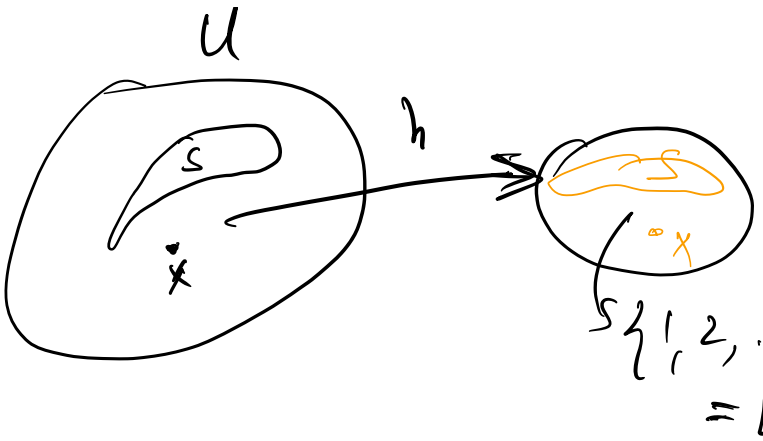
Solutions: 1) store S
query: scan S → $O(n)$ time

2) Binary Search Tree on S
space: $O(n)$
query: $O(\lg n)$.

3) store index: table $T[1..U]$
$T[i] = 1$ iff $i \in S$.
query: $O(1)$.
space: $O(U)$ bits.

Goal: obtain $O(1)$ q.t. & $O(n)$ space

Hashing:



No Collision Prop (NCP): $\forall x \in U$, $\forall y \in S$, $x \neq y$:

$h(x) \neq h(y)$.

Sol 4 (NCP): store table T s.t.

$T[h(i)] = 1$ iff $i \in S$.

$T[h(x)] = 1$ iff $x \in S$ (NCP).

Space: $\Theta(m)$.

$m = 10n$.

Fact: $\exists h$ that satisf NCP $m = n+1$.

$S = \{i_1, i_2, .. i_n\}$ $h(i_1) = 1$

$h(i_2) = 2$.

$h(i_n) = n$.

$h(x) = n+1$ $\forall x \notin S$.

Issues: computing $h(x)$?

as hard as Dict. problem.

Sol 5: pick h indep. of S at random.

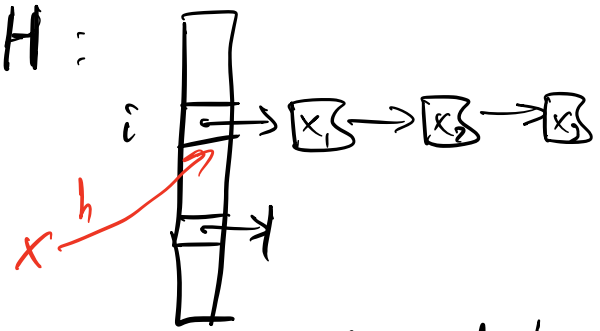
$\mathcal{H} = \{$all hash functions $h: U \to [m]\}$.

$|\mathcal{H}| = m^{|U|}$.

$h \in_r \mathcal{H}$.

Data structure: UCP does not hold (with large prob.)

app: use chaining.



$H[i]$ = stores a linked list of all y's s.t. $h(y) = i$.

query: look up L.L. of $h(x)$, iterate and check if $\ell = x$.

query time: prop to size of bucket $h(x)$

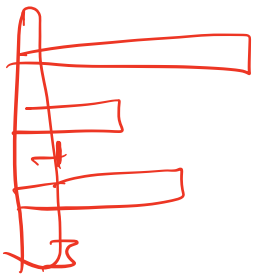
$C_x \triangleq$ # elements y in S, different from x, s.t. $h(x) = h(y)$.

Q. t. $= O(C_x)$.

fix x : $\mathop{\mathbb{E}}_h [C_x] = \mathop{\mathbb{E}}_h \left[\sum_{\substack{y \in S \\ y \neq x}} \mathbb{1}[h(x)=h(y)] \right]$

indicator r.v. event $h(x)=h(y)$.

$\left[\max_x \mathop{\mathbb{E}}_h [C_x] \right]$



$= \sum_{\substack{y \in S \\ y \neq x}} \left[\mathop{\mathbb{E}}_h \mathbb{1}[h(x)=h(y)] \right]$

$= \sum_{y \in S, y \neq x} \Pr[h(x)=h(y)]$

$= \sum_{y \in S, y \neq x} \frac{1}{m} \leq n/m$.

Conclusion: expected q. time : $O(n/m+1)$.

$\Rightarrow O(1)$ if $m = 2n$.

Remark: what is $\mathop{\mathbb{E}}_h [\max_x C_x] = ?$

what is the size of largest bucket / linked list?

answer: $\Theta\left(\frac{\lg n}{\lg \lg n}\right)$ $m = \Theta(n)$.

Issues 1: how to choose/store $h \in \mathcal{H}$?

– " 2: worst-case runtime is $\gg$ const.

$\sim$ logarithmic.

Sol 6: (no more assumptions)

how to choose $h \in \mathcal{H}$. fully random.

Idea: use smaller / more structured families $\mathcal{H}$.

Def: $\mathcal{H}$ is universal iff $\forall x \neq y$:

$$\Pr_{h \in \mathcal{H}}[h(x) = h(y)] = 1/m.$$

Remark: our analysis from Sol 5 used only fact that $\mathcal{H}$ is universal.

Def: $\mathcal{H}$ is k-wise independent iff.

$\forall$ distinct $x_1, \ldots x_k \in U$, any vals $v_1 \ldots v_k \in [m]$:

$$\Pr_{h \in \mathcal{H}}\left[h(x_i) = v_i,\ i = 1 \ldots k\right] = \frac{1}{m^k}.$$

Fact: if $\mathcal{H}$ is 2-wise indep $\Rightarrow$ universal.

$\forall x \neq y$ $\Pr[h(x)=h(y)] = \sum_v \Pr[h(x)=v \wedge h(y)=v] = m \cdot \frac{1}{m^2} = 1/m.$

Fact: $\forall k$, there exist k-wise indep h.f. $\mathcal{H}$ with $\lg|\mathcal{H}| = O(k \cdot \lg U)$.

and time to compute $h(x)$ is $\mathcal{B}$.

2-wise: $h_{a,b}(x) = (a + bx) \% m$.

Def: $\mathcal{H}$ is α-universal, $\alpha \geq 1$, iff

$\forall x \neq y$: $\Pr_h[h(x) = h(y)] \leq \frac{\alpha}{m}$.

Remark: $\mathbb{E}[C_x] \leq \alpha \cdot \frac{n}{m}$ ($h \in \mathcal{H}$ α-univ.)

Example of 2-universal $\mathcal{H}$:

$a \in [U]$ at random, odd.

$$h_a(x) = \left\lfloor ((a \cdot x) \% |U|) \cdot \frac{m}{|U|} \right\rfloor$$

$h_a : U \to \{0, 1, \ldots m-1\}$.

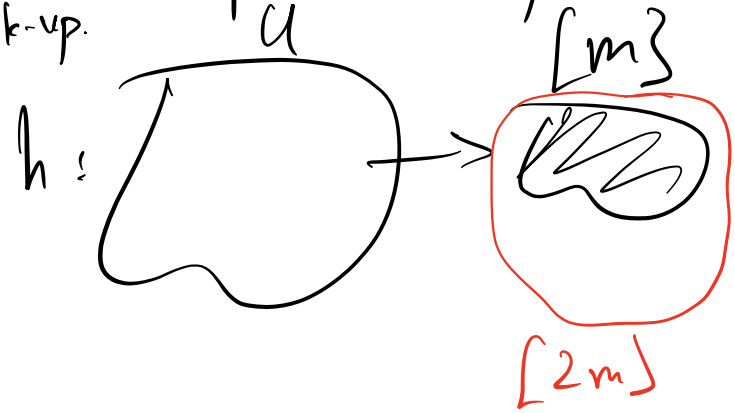
$\mathcal{H} = \{h_a$ for $a \in [U]$ and odd$\}$.

Conclusion (Sol 6): set $m = n$.

pick $h \in \mathcal{H}$ 2-univ.

space: $O(\lg U + n \cdot \lg U)$ bits to store
h.f. hash table hash table.

q.t: $O(1) + O(1)$ expected (word operations)
h.f. eval. look-up.



Sol 7: Perfect Hashing.

Goal: get $O(1)$ q.t. deterministically.

idea: trade-off hash func construct/description to depend (a little) on S

vs # collisions.