Open-addressed hashing

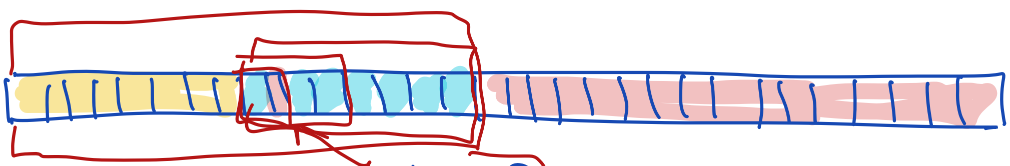
Insert(x):

for $i \leftarrow 0$ to $m-1$
  if $T[h_i(x)]$ is empty
    $T[h_i(x)] \leftarrow x$
    return
return FULL

$h_0(x), h_1(x), \ldots h_{m-1}(x)$ is a permutation of $0, 1, \ldots m-1$

Linear probing: $h_i(x) = (h_0(x) + i) \bmod m$

Binary probing: $h_i(x) = (h_0(x) \oplus i) \bmod m$ $(m = 2^\ell)$



$h_0(x) =$ 01010
$h_1(x) =$ 01011
$h_2(x) =$ 01000
$h_3(x) =$ 01001
$h_4(x) =$ 01110
⋮

FICTION: $(h_0(x), h_1(x) \ldots)$ is a random permutation

every perm has prob. $1/m!$

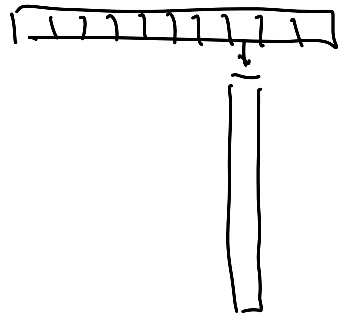
$T(n,m)$ = # probes need to insert nth item into a table of size m

$\boxed{m = 2n}$

$$E[T(n,m)] \le 1 + \frac{n}{m} E[T(n-1, m-1)]$$

Induction: $\le \frac{m}{m-n} = 2 = \frac{1}{1-\alpha}$

Chained:



universal hashing
$\Pr[h(x) = h(y)] \le 1/m$

$\Rightarrow O(1)$ lookup
$\Rightarrow O(1)$ amort exp. insertion / deletion

Loose binary probing: $B_k(x)$ = block of length 2^k containing $h_0(x)$

for $k \leftarrow 0$ to $\lg m$
  if $B_k(x)$ is not full
    put x into $B_k(x)$
    return

Time = O(size of largest full block containing $h_0(x)$)

$B_k(x)$ is full if all 2^k slots are occupied $\boxed{F_k = \Pr[B_k(x) \text{ is full}]}$

$B_k(x)$ is popular if $\#\{y \mid h_0(y) \in B_k(x)\} \geq 2^k$

popular $\Rightarrow$ full $P_k = \Pr[B_k(x) \text{ is popular}]$

B_k full $\Rightarrow$ B_k popular or sibling/uncle of B_k is popular

$$\boxed{F_k \leq P_k + \sum_{j \geq k} P_j}$$

We need to understand P_k (Fix k)

$Y = \#\{y \mid h_0(y) \in B_k(x)\}$ $\boxed{m = 2n}$

So $P_k = \Pr[Y \geq 2^k]$

$E[Y] = \frac{1}{2} \cdot 2^k$ assuming h_0 is uniform

$P_k = \Pr[Y \geq 2 \cdot E[Y]]$ $\circledast$

assume h_0 is pairwise-independent

$\Pr[Y \geq (1+\delta) E[Y]] < \frac{1}{\delta^2 E[Y]}$ Cheb's $\neq$

Set $\delta = 1$ $\Pr[Y \geq 2E[Y]] < \frac{1}{E[Y]} = \boxed{\frac{1}{2^{k-1}}}$

$\Rightarrow F_k = O(2^{-k})$

$$E[\text{Time}] = \sum_{k=0}^{\lg m} O(2^k) \cdot F_k = \sum_{k=0}^{\lg m} \underbrace{\cancel{O(2^k) \cdot O(2^{-k})}}_{O(1)} = O(\log m)$$

If hash values are 4-independent

Chebychebyshevshev's Inequality

$$\Pr[Y \geq (1+\delta)\mu] = O\left(\left(\frac{1}{\delta^2\mu}\right)^2\right)$$

$$\Pr[Y \geq 2 \cdot E[Y]] = O(4^{-k}) \Rightarrow F_k = O(4^{-k})$$

$$E[\text{Time}] = \sum_{k=0}^{\lg m} O(2^k) F_k = \sum_{k=0}^{\lg m} O(2^{-k}) = \boxed{O(1)}$$

We actually need 5-uniform hashing

For all distinct v, w, x, y, z

for all h, i, j, k, l

$$\Pr[h_0(v)=h \text{ and } \cdots \text{ and } h_0(z)=l] \leq \frac{1}{m^5}$$

Carter Wegman

$$h_0(x) = ((a+bx+cx^2+dx^3+ex^4) \bmod p) \bmod m$$

where $a,b,c,d,e \in_u [p]$

Thorup Zhang 2010:

Twisted tabulation:

$$h(x,y) = A[x] \oplus B[y] \oplus C[x+y]$$

5-uniform

Pătrașcu Thorup 2011:

Tabulation!

$$h(x_1 \ldots x_c) = A_1[x_1] \oplus A_2[x_2] \oplus \cdots \oplus A[x_c]$$

3-uniform

still get O(1) whp

depends on c