9. Hash functions

Friday, September 24, 2021 2:41 AM

Today: • Hash functions

Intuitively, it's helpful to sometimes "randomly" map numbers

Given a domain U, and a range $[m] = \{0, \ldots, m-1\}$, a

truly random hash function is a map

$h : U \to [m]$ where each $h(x)$ is an i.i.d. uniform r.v. in $[m]$

h is a $|U|$-dim r.v. chosen uniformly at random from $U \to [m]$ $\left([m]^U\right)$

$|U| \log_2 m$ bits space complexity

- $|U|$: from each element in U
- $\log_2 m$: store what it maps to in $[m]$

Impractical.

Def. A hash function $h : U \to [m]$ is a r.v. in the class of all functions $U \to [m]$.
(but not necessarily uniformly distributed)

Ex (trivial) Let $h : [m] \to [m]$ always be the identity function

prime field Let $h : [p] \to [p]$ for a prime p given by

$h(x) = (ax + b) \bmod p$, where a, b are uniformly randomly chose in $[p]$

Def. A family $\mathcal{H}$ of hash functions $h : U \to [m]$ is universal if
$\forall\, x, y \in U,\ x \neq y,$

$$\Pr_{h \in \mathcal{H}}[h(x) = h(y)] \leq \frac{1}{m}$$ (Note prob $= \frac{1}{m}$ for truly random)

Def. c-approximately universal is $\Pr_h[h(x) = h(y)] \leq \frac{c}{m}$, $c = O(1)$

If $u \leq m$, is the identity function $f(x) = x$ a universal hash function on $[u] \to [m]$?

If $u \geq m$, is the mod function $f(y) = x \bmod m$ universal on $[u] \to [m]$?

No, because 1 and $1+m$ collide with prob. 1.

Application to hash tables with chaining

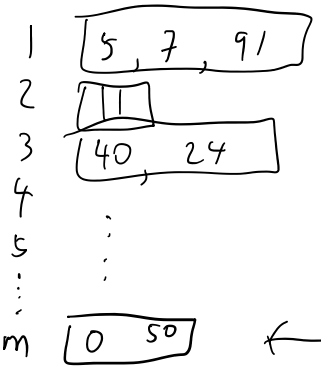
Suppose we have a set S of length n of integers in $[u]$.

How long does it take to check if a particular integer $x \in S$?

Depends on how store S.

	time	space	
• Unsorted list:	$O(n)$	$O(n)$	
• Sorted list:	$O(\log n)$	$O(n)$	via binary search
• Bitmap:	$O(1)$	$O(u)$	

• Hash table of size $m \geq n$ where we have an array L of m buckets



In each bucket i, we have an unordered list of all items $x \in S$ s.t. $h(x) = i$.

If h is universal and our query $x \notin S$.

Let $I(y)$ be the indicator variable for the event $h(x) = h(y)$.

Then $\mathop{\mathbb{E}}_h \left| L[h(x)] \right| = \mathop{\mathbb{E}}_h \left[\sum_{y \in S} I(y) \right] = \sum_{y \in S} \mathop{\mathbb{E}}_h I(y) = \sum_{y \in S} \Pr_h [h(y) = h(x)] \leq \frac{n}{m} \leq 1$

In expectation, we only need to check a list of size 1.

So queries are average $O(1)$ time, and space Δ $O(n)$.