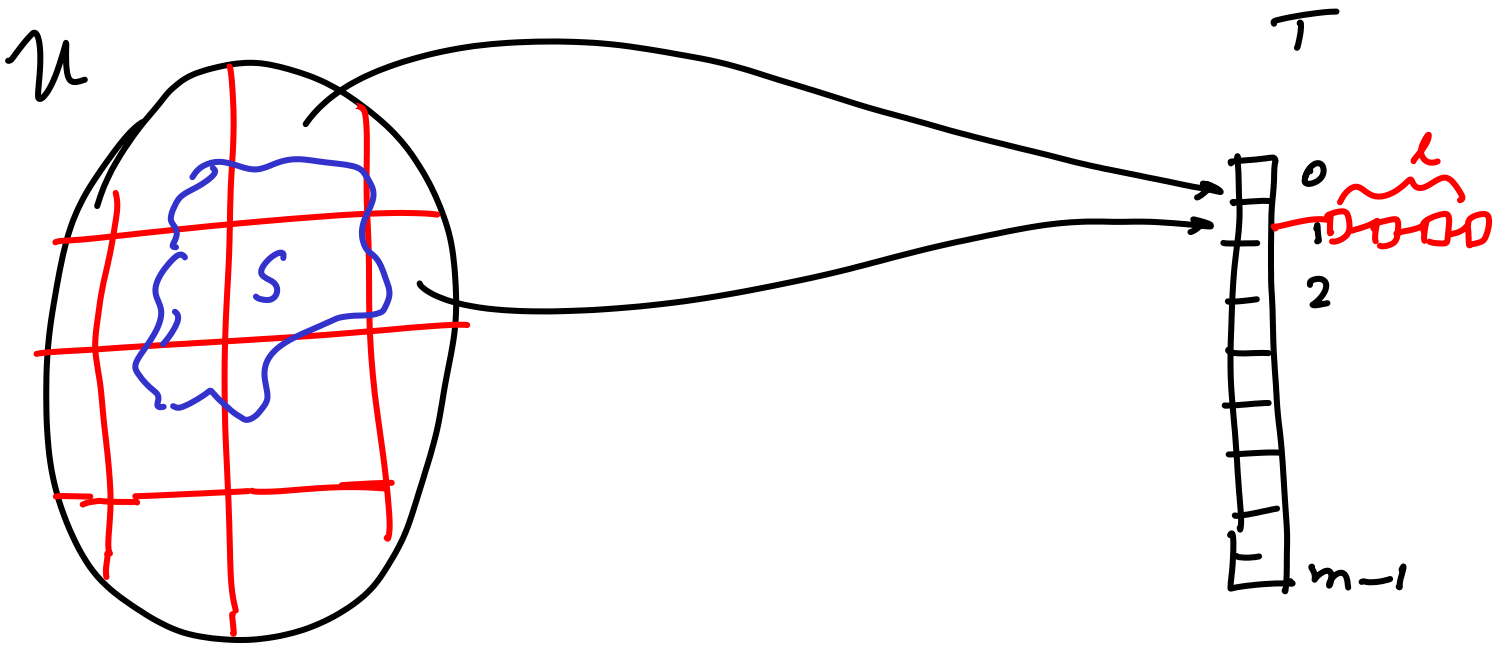
COL 702 Lecture 19 Oct 12

$\mathcal{U}$: universe of elements $|\mathcal{U}| = N$

S is a set of n elements $S \subset \mathcal{U}$ that is hashed into a table of size m

$N \gg n$ and n and m are approx same

$h : \mathcal{U} \to m$ h is a hash function



We say $x \neq y$ collide if $h(x) = h(y)$

Collisions could lead to long chains that cause high search time in hash tables

Special case: All the n elements of S are chosen uniformly at random from U

What is the <u>expected</u> length of any chain.

Consider location 0: what is expected # of elements that get mapped to 0

Let $X_i = \begin{cases} 1 & \text{if the } i\text{th element is mapped to location 0} \\ 0 & \text{otherwise} \end{cases}$

$$E\left[\text{Length of chain in location 0}\right] = E\left[\sum X_i\right]$$

$$= \sum E[X_i] \quad \text{using linearity of expectation}$$

$$= \sum p_i \quad \text{where } p_i \text{ is the prob that } X_i = 1$$

$$= \sum_{i=1}^{n} \frac{1}{m} = \frac{n}{m}$$

Why is $p_i = \frac{1}{m}$ only under the assumption that U is equally partitioned on the basis of inverse mapping from T, i.e. $h^{-1}(0) : h^{-1}(1) = \ldots$

The basic strategy is to use multiple hash functions, i.e. consider a set H of hash functions and choose one of them randomly and hope that the set S is "scattered" well w.r.t the chosen function

A set H of hash functions is called c-universal if for any pair $x, y \in U$ $x \neq y$

$$\sum_{h \in H} \delta_h(x, y) \leq \frac{c \cdot |H|}{m}$$

where $\delta_h(x, y) = 1$ if $h(x) = h(y)$
$= 0$ otherwise

$\delta_h(x, y)$ is called the collision function

It implies that if a hash function h_1 is chosen at random from H then the probability that $h_1(x) = h_1(y)$ is $\sim \frac{1}{m}$

For any element $x \in S$, let

$$\delta_h(x,S) = \sum_{y \in S} \delta_h(x,y).$$

Then $$\sum_{h \in H} \delta_h(x,S) = \sum_{h \in H} \sum_{\substack{y \in S \\ y \neq x}} \delta_h(x,y)$$

$$= \sum_{\substack{y \in S \\ y \neq x}} \left[\sum_{h \in H} \delta_h(x,y) \right]$$

$$\leq \sum_{y \in S} c \frac{|H|}{m} = c \cdot \frac{|H| n}{m}$$

$|S| = n$

Then it follows that the expected no. of elements in S that collide with x for a randomly chosen hash function $h \leq \frac{cn}{m}$

Question: Can we construct universal hash function?

Example 1: Suppose

$$h_a(x) = \left((x+a) \bmod N\right) \bmod m$$

where $a \in \{0, 1, 2, \ldots N-1\}$

Show that this is not universal

Example 2 : $h_a(x) = \left(xa \bmod N\right) \bmod m$

$a = \{1, 2, \ldots N-1\}$

Assume that N is prime

(There exists a prime between ℓ and 2ℓ for any integer ℓ : Bertrand's postulate)

Is $h_a(x)$ universal ?

$|H| = N-1$ $\quad a \neq 0$

For a pair x, y we want to count for how many a's

$h_a(x) = h_a(y)$

$\Rightarrow \left(ax \bmod N\right) \bmod m = \left(ay \bmod N\right) \bmod m$

$$\Rightarrow \left(\left(ax - ay \right) \bmod N \right) \bmod m = 0$$

$$\Rightarrow \quad a(x-y) \bmod N = km$$

for $k = \pm 1, \pm 2, \pm 3, \ldots \pm \left\lfloor \frac{N}{m} \right\rfloor$

For any fixed k

$a(x-y) = km$ has a unique soln for a

i.e. $a = (x-y)^{-1} \cdot km$

where the inverse is w.r.t the multiplicative group modulo N

So total # of a's for which $h_a(x) = h_a(y) \leq 2 \cdot \frac{N}{m}$

$|H| = N-1$

value is $2 \cdot \frac{N}{m}$