Design and Analysis of Algorithms I

Data Structures

Universal Hash Functions: Definition and Example

Overview of Universal Hashing

Next: details on randomized solution (in 3 parts).

Part I: proposed definition of a "good random hash function". ("universal family of hash functions")

Part II: Concrete example of simple + practical such functions

Part III: justification of definition: "good functions" lead to "good performance".

Universal Hash Functions

Definition: Let $\mathcal{H}$ be a set of hash functions from U to $\{0,1,2,\ldots,n-1\}$.

$\mathcal{H}$ is **universal** if and only if:

for all $x,y \in U$ (with $x \neq y$)

$$\Pr_{h \in \mathcal{H}}\left[\begin{array}{c} x,y \text{ collide} \\ (h(x)=h(y)) \end{array}\right] \leq \frac{1}{n}$$

(n = # of buckets)

when h is chosen uniformly at random from $\mathcal{H}$.

(i.e., collision probability as small as with "gold standard" of perfectly random hashing)

Tim Roughgarden

Consider a hash function family H, where each hash function of H maps elements from a universe U to one of n buckets. Suppose H has the following property: for every bucket i and key k, a $1/n$ fraction of the hash functions in H map k to i. Is H universal?

Yes: Take H = all functions from U to $\{0,1,2,\ldots,n-1\}$.

- Yes, always.
- No, never.
- Maybe yes, maybe no (depends on the H).
- Only if the hash table is implemented using chaining.

No: Take H = the set of n different constant functions.

Example: Hashing IP Addresses

Let U = IP addresses (of the form (x_1, x_2, x_3, x_4), with each $x_i \in \{0,1,2,\ldots,255\}$)

Let n = a prime (e.g., small multiple of # of objects in HT)

Construction: Define one hash function h_a per 4-tuple $a = (a_1, a_2, a_3, a_4)$ with each $a_i \in \{0,1,2,3,\ldots,n-1\}$.

(n^4 such functions)

Define: h_a: IP addrs $\rightarrow$ buckets

by $h_a(x_1, x_2, x_3, x_4) = \begin{pmatrix} a_1x_1 + a_2x_2 + \\ a_3x_3 + a_4x_4 \end{pmatrix} \bmod n$

Tim Roughgarden

A Universal Hash Function

Define: $\mathcal{H} = \{ h_a \mid a_1, a_2, a_3, a_4 \in \{0, 1, 2, \ldots, n-1\} \}$.

$h_a(x_1, x_2, x_3, x_4) = (a_1 x_1 + a_2 x_2 + a_3 x_3 + a_4 x_4) \bmod n$

Theorem: This family is universal.

Tim Roughgarden

Proof (Part I)

Consider distinct IP addresses (x_1, x_2, x_3, x_4), (y_1, y_2, y_3, y_4).

Assume: $x_4 \neq y_4$. **Question:** collision probability?

(i.e., $\Pr_{h_a \in \mathcal{H}}[h_a(x_1, \ldots, x_4) = h_a(y_1, \ldots, y_4)]$)

Note: collision $\Leftrightarrow$ $\underset{(\mathrm{mod}\ n)}{a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4} = \underset{(\mathrm{mod}\ n)}{a_1y_1 + a_2y_2 + a_3y_3 + a_4y_4}$

$\Leftrightarrow a_4(x_4 - y_4) \bmod n = \sum_{i=1}^{3} a_i(y_i - x_i) \bmod n$

Next: Condition on random choices of a_1, a_2, a_3.

(a_4 still random)

Tim Roughgarden

Proof (Part II)

The Story So Far: with a_1, a_2, a_3 fixed arbitrarily, how many choices of a_4 satisfy $\underbrace{a_4(x_4-y_4) \bmod n = \sum_{i=1}^{3} a_i(y_i-x_i) \bmod n}_{\Leftrightarrow\ x,y \text{ collide under } h_a}$?

still random

Some fixed number in $\{0,1,2,\ldots,n-1\}$

Key Claim: left-hand side equally likely to be any of $\{0,1,2,\ldots,n-1\}$.

[addendum: make sure n bigger than the maximum value of an a_i]

Reason: $x_4 \neq y_4$ ($x_4 - y_4 \neq 0 \bmod n$), n is prime, a_4 uniform at random.

$\rightarrow$ implies $\Pr_a[h_a(x)=h_a(y)] = \frac{1}{n}$

"Proof" by example: $n=7$, $x_4 - y_4 = 2$ or $3 \bmod n$

QED!