Hashing 3

Feb 25 2021

Construction 3 (using finite fields)

$GF(2^u)$ each element u-bit vector

pick two random elements $a, b \in GF(2^u)$ (uniformly at random)

$x \in U$

$h(x) = ax + b$

Q: 2-wise indp ?

$$ax_1 + b = \alpha_1$$

$$ax_2 + b = \alpha_2$$

$\Rightarrow$ $a =$

$b =$

$[u] \longrightarrow [u]$

We have $[U] \to [U]$ ⟶ vector of length u

We want $[U] \to [M] \to$ vector of len. m $(m < u)$

$|U| = 2^u$ $\quad$ $|M| = 2^m$

$|M| << |U|$

$m << u$

Truncate to m bits! (Ex)

---

k-wise indep.

$\mathbb{F}(2^u)$

linear fn ⟶ polynomial fn!

Pick uniformly at random : $a_0 \dots a_{k-1} \in GF(2^u)$

$$h(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_{k-1} x^{k-1}$$

To prove : k-wise indep.

Other example : Tabulation

Hash tables : closed addressing
Open addressing

Open addressing:

No separate DS

All keys stored in one array.

Linear probing:

$h(x)$

go $(h(x)+1) \bmod M$

Can have constant time look-ups.

But universal hashing not sufficient

5-wise indep is necessary

other probe sequences: step-size, quadratic etc.

Cuckoo Hashing;

Open addressing

by Pagh & Rodler (2004)

Take two tables $T1$ & $T2$

Both of size M_2 $O(N)$

Two hash fns $h_1, h_2 : U \longrightarrow [M]$

$h_1, h_2 \in H$ ← hash fn. family

Assume H is fully-random ($O(\log N)$-wise indep)

Insertion of an element x:

$T1[h_1(x)]$ or $T2[h_2(x)]$

Bumping out process ....

If more than $6 * \log N$ bumps rehash everything.

Theorem: Expected time for insertion is $O(1)$ if $M \geq 4N$ [5]

M ← table size

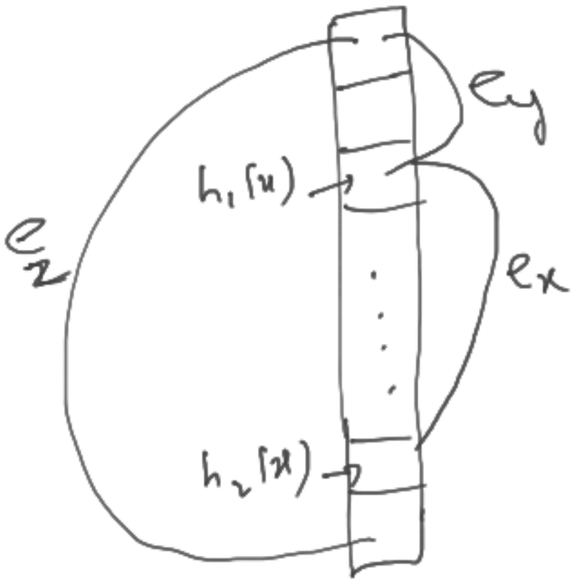
Proof sketch:

Cuckoo graph (G_1):

M vertices

Edges: elements (to be) inserted

$e_x = (h_1(x), h_2(x))$



Bucket of x

$B(x)$ = set of nodes of G_1 reachable from $h_1(x)$ or $h_2(x)$

E[insertion time for x] = $E[|B(x)|]$

↖ - G is random
- so is $B(x)$

Goal: $E[|B(x)|] \leq O(1)$

$$E[|B(x)|] = \sum_{\substack{y \in S \\ y \neq x}} P[e_y \in B(x)]$$

$$\leq N \, P[e_y \in B(x)]$$

Sufficient to show $P[e_y \in B(x)] \leq O(1/M)$

Lemma: For any i, j in $[M]$

$$P[\text{there exists a path of length } \ell \text{ between } i \& j \text{ in the cuckoo graph } G_t] \leq \frac{1}{2^\ell M}$$

Proof: Via induction.

$\ell = 1$: $P(\text{edge between } i \& j)$

$= P(\exists y \text{ s.t } e_y \text{ exists in } G_t)$

$\leq N \cdot \frac{2}{M^2}$ ← $P[(h_1(y) = i \cap h_2(y) = j) \cup (h_1(y) = j \cap h_2(y) = i)]$

$\leq \frac{1}{2 \cdot M}$

Then induction on ℓ.

(Ex)

To show: $P[e_y \in B(x)] \leq O(1/M)$

using the lemma,

$$P[e_y \in B(x)] \leq \sum_{l \geq 1} \frac{1}{2^l M}$$

$$= O(1/M).$$

$M \geq 4N \Rightarrow$ 25% space efficiency

For $d = 3$, experimentally >90% space efficiency

Bloom Filters:

Limited ops: membership query

Allows for mistakes:

only false positives; no false negatives.

- Array T of M bits
  - initialized to 0
- k hash funcs. $h_1, h_2 \ldots h_k : U \to [M]$
  - assume fully random
- Adding: $x \in S$

  set bits $T[h_1(x)]\ T[h_2(x)] \ldots T[h_k(x)]$ to 1

Membership query :