Reprezentace množiny řetězců nad abecedou Σ — např. $\{0,1\}$ $\{0,\_,9\}$ $\{a,\_,z\}$

$\uparrow \{x_1 \_ x_n\}$

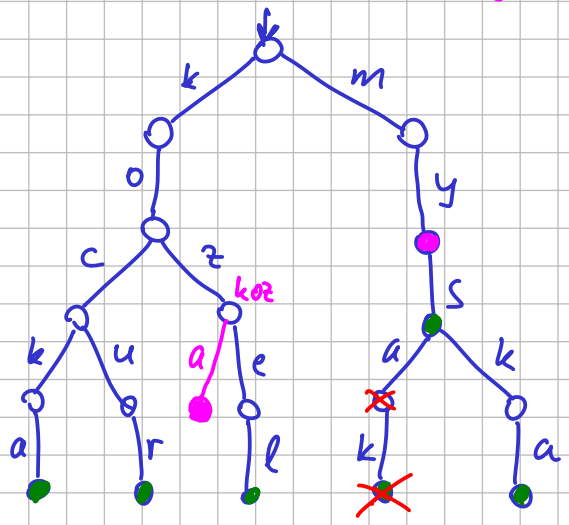
Pokud použijeme BVS: $\mathcal{O}(\log n \cdot$ délka řetězce$)$ $\leq |y|$

Find(y)

Písmenkový strom (trie ... kombinace slov tree, retrieval)

{ kocka, kocur, kozel, mys, mysak, myska }

kozka myš



vrcholy ~ prefixy slov v množině

D.S.: strom (ve vrcholu pole ukazatelů index. abecedou) + značky ve vrcholech příp. hodnota přísl. slovu

Member(y) $\mathcal{O}(|y|)$

Insert(y) $\mathcal{O}(|y|)$

Delete(y) $\mathcal{O}(|y|)$

paměť: $\mathcal{O}(|\Sigma| \sum_i |x_i|)$

Co když je abeceda velká? Pole nahradíme vyhl. stromem → paměť $\mathcal{O}(\sum_i |x_i|)$, čas $\mathcal{O}(|y| \cdot \log |\Sigma|)$

Číslicový strom (radix trie) → zápis čísla v soustavě o základu z uložíme jako řetězec $\mathcal{O}(|y| + \log |\Sigma|)$

základ z, n čísel z rozsahu $\{0 \_ L-1\}$ → # číslic $\sim \log_z L$

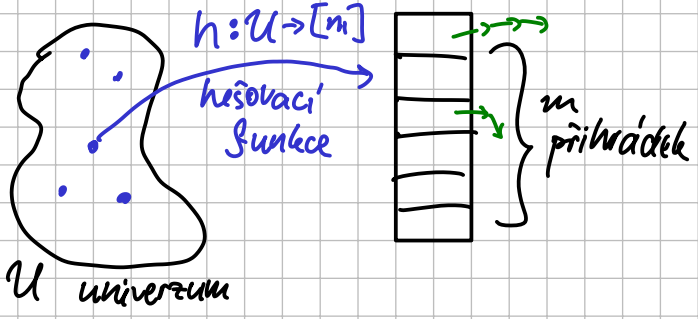
→ čas na operaci je $\mathcal{O}(\log_z L)$

pokud jsou čísla různá, pak $L \geq n \Rightarrow \log L \geq \log n$

HEŠOVÁNÍ

$\{0 \_ m-1\}$

$h: U \to [m]$ hešovací funkce



kolize ≡ více prvků v přihrádce

v přihrádce seznam [tabulka = pole ukazatelů]

Příklad: 1212 935 1918 1948 1968 1989 2021

$h(x) := x \bmod 10$

0	1	2	3	4	5	6	7	8	9
	2021	1212			935			1918 1948 1968	1989

představa: rovnoměrné rozložení n prvků do přihrádek

↳ typicky n/m prvků v přihrádce

Find, Insert, Delete pracují typicky v čase $\mathcal{O}(1)$ ↳ konstanta pro $m \in \Omega(n)$

Volba hešovací fce prakticky

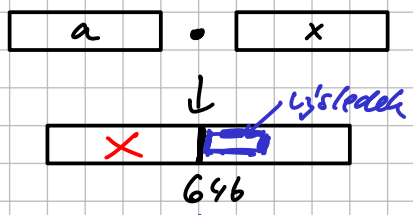
- $x \mapsto (ax) \bmod m$ — často prvočíslo
  - a nesoudělné s m
  - typ. $a \doteq 0.618\,m$

lineární kongruence

- pro $\mathcal{U} = [2^w]$, $m = 2^k$
  $x \mapsto (ax \bmod 2^w) >> w-k$

multiply-shift

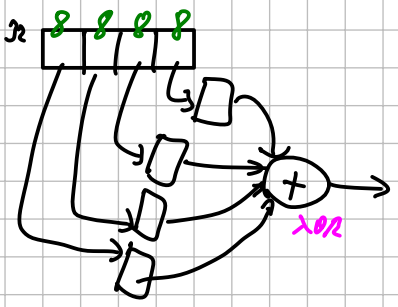
32-bit počítač



multiply-shift

- $x_0 \dots x_{d-1} \mapsto \left(\sum_i a_i x_i\right) \bmod m$ — skalární součin
  parametry $a_0 - a_{d-1}$
- $x_0 \dots x_{d-1} \mapsto \left(\sum_i a^i x_i\right) \bmod m$ — polynom

tabelační hešování



"Naplňovací heš. tabulka"

- sledujeme $\alpha := \frac{n}{m}$ — hustota / faktor naplnění

chceme $\alpha \le$ konst.

- naroste-li α příliš, přehešujeme do více přihrádek

zvolíme $m \to 2m$

1, 2, 4, 8, 16, ...

přehešovat z 2^i do 2^{i+1} trvá $\Theta(n + 2^{i+1}) = \Theta(n)$

$\frac{n}{2^i} >$ konst. $\Rightarrow 2^i < \frac{n}{\text{konst.}}$ $\Rightarrow$ 2^i je $\Theta(n)$

$\frac{n-1}{2^i} <$ konst. $> \frac{n-1}{\text{konst.}}$

zjednodušení: $\alpha \le 1$

počty přihrádek jsou 1, 2, 4, 8, 16, ...

$2^{i-1} \to 2^i \to 2^{i+1}$

$2^{i-1}+1$ $\quad$ 2^i+1

trvá $\Theta(2^i) \to O(1)$ na přidaný prvek [amortizace]

přibylo 2^{i-1} prvků

Teorie

- volíme hešovací funkci náhodně z nějakého systému funkcí

parametrizované např.

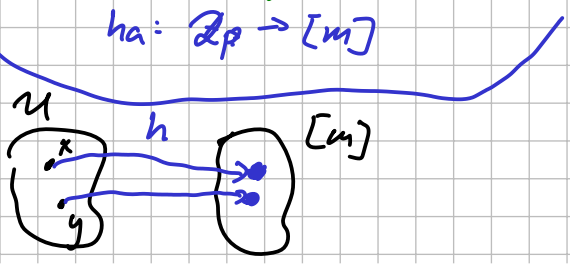
$\mathcal{H} := \{h_a \mid a \in \mathbb{Z}_p\}$

$h_a(x) := (ax) \bmod m$

$h_a : \mathbb{Z}_p \to [m]$

Def: Systém funkcí $\mathcal{H}$ z $\mathcal{U}$ do $[m]$ je c-univerzální pro $c > 0$

$\equiv \forall x, y \in \mathcal{U},\ x \ne y : \Pr_{h \in \mathcal{H}}[h(x) = h(y)] \le \frac{c}{m}$.



Příklad: Skalární součin nad tělesem $\mathbb{Z}_p$.

$U = \mathbb{Z}_p^d$, $a \in \mathbb{Z}_p^d$, přihrádky: $\mathbb{Z}_p$

$\mathcal{H} = \{ h_a \mid a \in \mathbb{Z}_p^d \}$

$h_a(x) = a \cdot x$ (v tělese) $\left(\sum_i a_i x_i\right) \bmod p$

Věta: Tento systém je 1-univerzální.

Dk: pro $x \neq y$ $\Pr_{h \in \mathcal{H}}[h(x) = h(y)]$

$= \Pr_a [a \cdot x = a \cdot y]$

$a \cdot (x - y) = 0$

$\sum_{i=1}^{d} a_i (x_i - y_i) = 0$

$a \cdot z$ rovnoměrné $\sim \mathbb{Z}_p$ ($z \neq 0$)

Dáno $x_1 \neq y_1$

Nechť zafixuji $a_2 - a_d$

$a_1 (x_1 - y_1) + \sum_{i=2}^{d} a_i (x_i - y_i) = 0$

a_1: neznámá; $(x_1 - y_1)$: nenulová konst.; $\sum_{i=2}^{d} a_i (x_i - y_i)$: konst.

Lineární rovnice $\Rightarrow$ ∃! řešení a_1

$\Pr_a[a_1 \text{ je řešení}] = \frac{1}{p}$ ∎