Hešování řetězců

$\mathcal{S} = \{ h_t \mid t \in \mathbb{Z}_p^d \}$ 1-univ. $\qquad h_t(x_0, \dots, x_{d-1}) = \sum_{i=0}^{d-1} t_i x_i \pmod p$ $\quad$ p prvočíslo, $d \geq 1$

Tvr. $\mathcal{R} = \{ h_a \mid a \in \mathbb{Z}_p \}$ d-univ. $\qquad h_a(x_0, \dots, x_{d-1}) = \sum_{i=0}^{d-1} x_i a^i \pmod p$

Dk: $x \neq y$ $\quad \Pr_a[h_a(x) = h_a(y)] = \Pr_a[\sum_i (x_i - y_i) a^i = 0] \leq \frac{d}{p}$ $\qquad h_a : \mathbb{Z}_p^d \to \mathbb{Z}_p$

□

Lem: $\mathcal{F}$ c-univ. fce $f : \mathcal{U} \to [r]$, $\mathcal{G}$ $(2,d)$-nez. fce $g : [r] \to [m]$ $(r \geq m)$. Pak

$\mathcal{H} = \mathcal{F} \circ \mathcal{G} = \{ f \circ g \mid f \in \mathcal{F}, g \in \mathcal{G} \}$ je $(2, c')$-nez., kde $c' = (\frac{cm}{r} + 1) d$.

Dk: $x_1 \neq x_2$, $i_1, i_2 \in [m]$

M: (match) $\quad h(x_1) = i_1$ a $h(x_2) = i_2$

$\qquad h(x_1) = g(f(x_1))$, $\quad h(x_2) = g(f(x_2))$

C: (kolize) $\quad f(x_1) = f(x_2)$

$\Pr[M] = \Pr[M \wedge \neg C] + \Pr[M \wedge C] =$

$= \Pr[M \mid \neg C] \Pr[\neg C] + \Pr[M \mid C] \Pr[C]$

$\leq \frac{d}{m^2} \cdot 1 + \frac{d}{m} \frac{c}{r} = \frac{c'}{m^2}$

□

Př: $\mathcal{L}' = \{ h_{a,b} \mid a,b \in \mathbb{Z}_p \}$ $h_{a,b}(x) = ((ax+b) \bmod p) \bmod m$

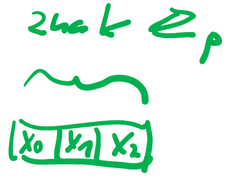
je (2,4)-nez.

$\mathcal{S} \circ \mathcal{L}'$ je (2,8)-nez., pro $p \geq 4m$ (2,5)-nez.

$\mathcal{R} \circ \mathcal{L}'$ je (2,5)-nez. pro $p \geq 4dm$.

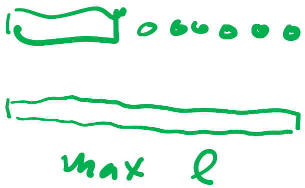
čas $O(d)$

kombinace: při $p >> k$ (vel. abecedy)



$x_0 + kx_1 + k^2 x_2 < p$

řetězce s lib. délkou:

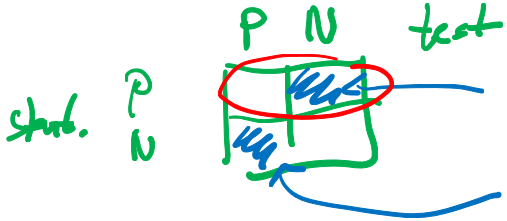


← pomocný znak ␣

kódovaný 0 ⇒ urychlí výpočet.

Bloomovy filtry

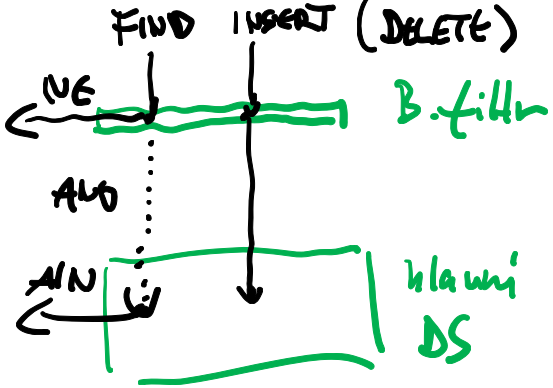
filtr: test s jednostranou chybou (bez falešně negativních)



falešně negativní (ideálně 0, senzitivita = 100 %)

falešně pozitivní (co nejméně falešně poz.)

B. filtr: DS přibližně reprezentující množinu bez falešně negativních



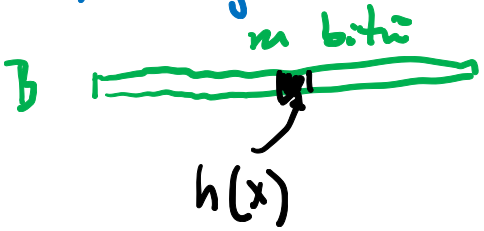
⇒ odpověď NE ... určitě NE

ANO ... možná ANO

cíl: paměťově efektivní

co nejmenší počet falešně pozitivních.

1-násobný B. filtr



$h: U \to [m]$ z univ. systému

$\text{FIND}(x) = \text{ANO} \iff B[h(x)] = 1$

⇒ bez falešně negativních

$x_1, \dots, x_n$ vložené, y ne (skutečně negativní)

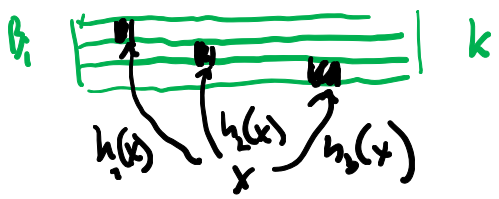
$$\Pr[y \text{ falešně pozitivní}] = \Pr[h(y) = h(x_i) \text{ pro nějaké } i] \le \sum_{i=1}^{n} \Pr[h(y) = h(x_i)] \le \frac{cn}{m} < \varepsilon$$

Pro dané n, cílová pravd. falešně pozitivní $\varepsilon > 0$:

$\Rightarrow m = \lceil n/\varepsilon \rceil$ (lin. vzhledem $1/\varepsilon$)

Př.: $n = 10^6$, $\varepsilon = 0.01$ $\Rightarrow$ $m = 100$ Mb (pro $U = 2^{32}$ lze mít h. tabulku s $\alpha < 1/3$)

Vícepásmový filtr



$h_i : U \to [m]$ nezávisle zvolené 2 c-univ. systém

INSERT(x): $B_i[h_i(x)] := 1$ $\forall i := 1 \dots k$

FIND(x) = ANO $\Leftrightarrow$ $B_i[h_i(x)] = 1$ $\forall i$

$\Rightarrow$ bez falešně neg., čas $O(k)$

Věta: Nechť $\varepsilon > 0$, n max. počet prvků. Pro $k = \lceil \log 1/\varepsilon \rceil$ a $m = 2n$ má B. filtr má pravděp. falešně pozitivních $< \varepsilon$.

Dk: $x_1 \dots x_n$ vložené, y různé (členství neg.)

$$\Pr[y \text{ falešně poz.}] = \Pr[\forall i\ \exists j\ h_i(y) = h_i(x_j)] = \prod_{i=1}^{k} \Pr[h_i(y) = h_i(x_j) \text{ pro nějaké } j]$$

$$\leq \prod_{i=1}^{k} \frac{cn}{m} = 2^{-k} \leq \varepsilon \quad \square$$

(pro $c = 1$)

Př: $n = 10^6$, $\varepsilon = 0.01 \Rightarrow k = 7$, $m = 2 \cdot 10^6$

14 Mb

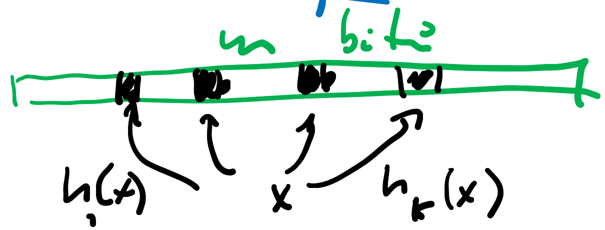
$\varepsilon = 0.001 \Rightarrow k = 10$, $m = 2 \cdot 10^6$, 20 Mb

celkem $M = 2n \lceil \log 1/\varepsilon \rceil$

Pozn: víme, že je třeba aspoň $n \log \lceil 1/\varepsilon \rceil$ bitů

dá se ukázat (h_i zcela náh. fce), pro $m \doteq 1.44\, n$ opět $\Pr \leq \varepsilon$.

jednoduchý Bloomův filtr



$h_i : \mathcal{U} \to [m]$ nez. zvolené, h_i zcela náhodné fce, $i = 1..k$

$$\text{FIND}(x) = \text{ANO} \iff B[h_i(x)] = 1 \ \forall i$$

$x_1 \dots x_n$ uložené, y různé $\Rightarrow$ nejvýše kn bitů 1

$$p = \Pr[B[i] = 0] = (1 - 1/m)^{nk} \approx e^{-nk/m} \Rightarrow k \approx -m/n \ln p$$

$[1 + x \approx e^x]$

$$\Pr[y \text{ falešně poz.}] = \Pr[B[h_i(y)] = 1 \text{ pro každé } i] \approx (1-p)^k = 2^{-k} \le \varepsilon$$

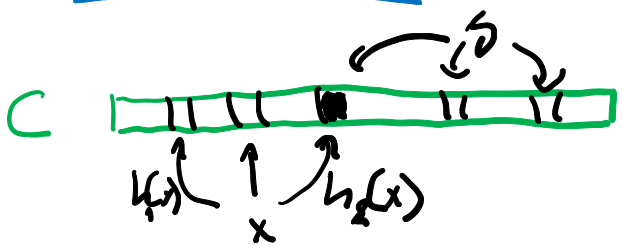
(↑ nez.)

$$\approx (1-p)^{-m/n \ln p} = e^{-m/n \ln p \ln(1-p)}$$

min. pro $p = 1/2$

$$\Rightarrow k = \lceil \log 1/\varepsilon \rceil, \quad m = 1.44\, n \lceil \log 1/\varepsilon \rceil$$

počítací filtr



$h_i : U \to [m]$

DELETE, INSERT: odečítat/přičítat 1 na $C[h_i(x)]$ pro i

FIND(x) = ANO $\Leftrightarrow$ $C[h_i(x)] > 0$

čítač b bitů $\Rightarrow$ $C[i] = 2^b - 1$ čítač "zamrzne"

$\Rightarrow$ nový typ falešně pozitivních, n prvků, h_i zcela náh. fce

$$Pr[C[i] \geq t] \leq \binom{n}{t}\left(\frac{1}{m}\right)^t \leq \left(\frac{ne}{t}\right)^t \left(\frac{1}{m}\right)^t \approx \left(\frac{e \ln 2}{t}\right)^t$$

$m \approx n / \ln 2 = 1.44\, n$

$$Pr[\text{nějaký čítač} \geq t] \leq m \cdot \left(\frac{e \ln 2}{b}\right)^t$$

Při $b = 4$ $\Rightarrow$ $t = 15$ $\quad Pr[C[i] \geq 16] \leq 3.06 \cdot 10^{-14}$

$m = 10^9$ $\Rightarrow$ $Pr[\text{nějaký čítač zamrzne}] \leq 3.06 \cdot 10^{-5}$ $\Rightarrow$ "4-bitový stačí"