Výběr hešovací funkce

25 November, 2021 10:06

- pokud jsem data distribuována náhodně (uniformně a nezávisle v U) pak libovolná rozumná fce $h: U \to \{1, \dots, m\}$ bude fungovat dobře
  - tj. $h^{-1}(a)$ je (± 1) stejně velké pro všechna $a \in \{1, \dots, m\}$

ALE:

- data jsou málokdy distribuována náhodně
  - → nutno brát náhodně hašovací fce.

př.: vezmeme pevné $h: U \to \{1, \dots, m\}$ $|U| \geq m \cdot n$

$\Rightarrow \exists a$ t.ž. $|h^{-1}(a)| \geq n \Rightarrow \exists S \subseteq h^{-1}(a)$ kde $|S| = n$.

→ pro každou pevně zvolenou hašovací fci existuje špatná množina (→ DDOS attack)

- ideálně: $h: U \to [m]$ vybráno zcela náhodně
  - tj. $\forall x \in U$ $h(x)$ je zvoleno nezávisle a uniformně z $\{1, \dots, m\}$

problém: pro uložení h potřebujeme tabulku velikosti $|U| \cdot \log m$

→ jsme zpět na začátku.

→ chceme podmnožinu $\mathcal{H}$ všech hašovacích fcí, t.ž. náhodně zvolené $h \in \mathcal{H}$ se bude chovat dobře z hlediska použití pro hašování a $\mathcal{H}$ bude [illegible] malé. [illegible] $\log |\mathcal{H}|$ bitů.

→ popis hešek vyžaduje $\log n$ bitů

- **minimální požadavek na $\mathcal{H}$:**

pro libovolné prvky $x, y \in U$, $x \neq y$

$$\Pr_{h \in \mathcal{H}}[h(x) = h(y)] \leq \frac{1}{m} \qquad (*)$$

tedy pravděpodobnost kolize dvou prvků je jako nejhůře jako u náhodné fce.

$\mathcal{H}$, která splňuje $(*)$, je **univerzální** hašovací systém

- **silnější požadavek:**

pro $\forall$ prvky $x, y \in U$ $x \neq y$ a $\forall$ prvky $r, q \in \{1, \dots m\}$

$$\Pr_{h \in \mathcal{H}}[h(x) = r \ \& \ h(y) = q] = \frac{1}{m^2} \qquad (**)$$

$\mathcal{H}$, která splňuje $(**)$, je **2-nezávislý hašovací systém**

(občas též. **2-univerzální**)

- **obecněji:**

$\forall$ prvky $x_1, x_2, \dots x_k \in U$ a $\forall r_1, r_2, \dots r_k \in \{1, \dots m\}$

$$\Pr_{h \in \mathcal{H}}[h(x_1) = r_1 \ \& \ h(x_2) = r_2 \dots \ \& \ h(x_k) = r_k] = \frac{1}{m^k}$$

... **k-nezávislý hašovací systém**

Např:

- $\mathcal{H}$ je 5-nezávislý → lineární přidávání funguje dobře
- $\mathcal{H}$ je $\lg n$-nezávislý → kukačkové hašování funguje dobře.

kukačkové hašování funguje dobře také s tabulkovým hašováním (viz. dále)

Perfektní hašování

$\mathcal{H}$ ... univerzální hašovací systém $U \to \{0, \dots, m-1\}$

uvažme pevné $S \subseteq U$ $|S| = n$

- když hašuje S perfektně pokud
$$\forall (x,y) \in S^2 \quad x \neq y \quad h(x) \neq h(y).$$
- Pokud $m \geq kn^2$, pro $k \geq 1$, pak
$$\Pr_{h \in \mathcal{H}}[h \text{ hašuje } S \text{ perfektně}] \geq 1 - \frac{1}{k}$$

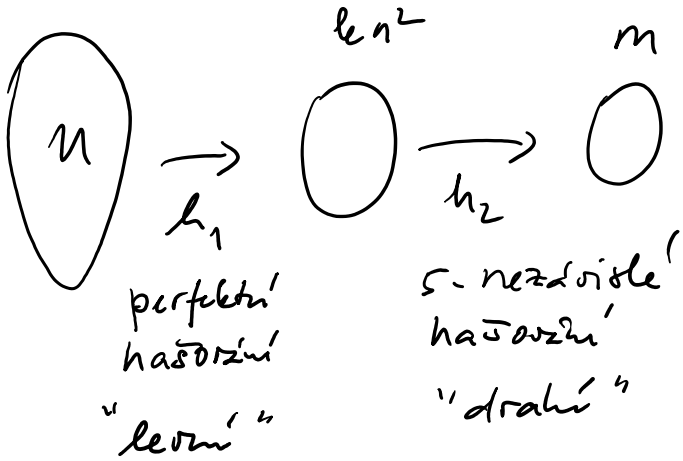
Dk:

$$\Pr_{h \in \mathcal{H}}[\overbrace{\exists x \neq y \in S,\ h(x) = h(y)}^{\text{existuje kolize}}] \leq \sum_{(x,y) \in S^2} \Pr_{h \in \mathcal{H}}[h(x) = h(y)]$$
$$\leq n^2 \cdot \frac{1}{m} \leq n^2 \cdot \frac{1}{kn^2} = \frac{1}{k}$$

$$\Pr_{h \in \mathcal{H}}[h \text{ hašuje } S \text{ perfektně}] = \Pr_{h \in \mathcal{H}}[\text{nenastává kolize prvků v } S]$$
$$= 1 - \Pr_h[\text{nastává kolize}]$$
$$\geq 1 - \frac{1}{k} \quad \square$$

Perfektní hašování lze použít jako mezikrok při hašování

U ... obrovská velikost, např. řetězce



$h : U \to \{0, \dots, m-1\}$ $\quad h(x) = h_2(h_1(x))$

Pokud h_1 je perfektní na S, vlastnosti určují h_2

↳ stačí u s pstí $\geq 1 - \frac{1}{k}$

lze volit $k \gg n$.

Tabulkové hašování

- náhodné fce vyžadují $|U|$ bitů paměti
- kompromis nabízí tabulkové hašování:

obecně $x_1, x_2, \dots x_d \in \{0, \dots, w-1\}$

náhodné tabulky $T_1, T_2, \dots T_d : \{0, \dots w-1\} \to \{0,1\}^k$

$$h(x_1, \dots x_d) = T[x_1] \oplus T[x_2] \oplus \dots T[x_d]$$

↑ XOR po bitech

$$U = [w]^d$$

Př.: $U = \{0,1\}^{32}$ $\quad x \in U$ interpretuji jako $x_1, \dots x_4 \in \{0,1\}^8$

$h(x) = T_1[x_1] \oplus T_2[x_2] \oplus T_3[x_3] \oplus T_4[x_4]$.

$T_1, \dots T_4 : \{0,1\}^8 \to \{0,1\}^{16}$

$m = 2^{16}$

místo $16 \cdot 2^{32}$ bitů (≈ 8 GB) stačí

$4 \cdot 16 \cdot 2^8$ bitů $= 2$ kB.

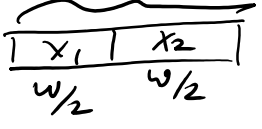
... lze použít pro binkačkové hašování [Patrascu-Thorup]

<u>speciální případ</u> – S-nezávislé:

$x \in \{0, \dots |U|-1\}$

$x_1 = \frac{x}{\sqrt{|U|}}$ $\quad x_2 = x \bmod \sqrt{|U|}$

$|U| = 2^w$



$T_1, T_2 : \{0, \dots \sqrt{|U|}-1\} \to \{0,1\}^k$

$T_3 : \{0, \dots 2\sqrt{|U|}\} \to \{0,1\}^k$

$h(x) = T_1[x_1] \oplus T_2[x_2] \oplus T_3[x_1 + x_2]$.

2-nezávislé hašovací systémy

PŘ:

1) p ... prvočíslo $\quad U \subseteq \{0, \dots p-1\}$

$\mathcal{H} = \{ h_{a,b} : U \to \{0, \dots p-1\} ; \ a, b \in \{0, \dots, p-1\} \}$

kde $h_{a,b}(x) = ax + b \bmod p$

vybrat náhodné $h \in \mathcal{H}$ znamená vybrat náhodně

<u>a</u> a <u>b</u> $\in \{0, \dots, p-1\}$

(právě jedno řešení: $ax + b = r \bmod p$, $ay + b = s \bmod p$) $\to 2 \log p$ bitů na popis <u>h</u>.

(prvočíslo: $ax+b=r \bmod p$; jednoznačné řešení: $ay+b=s \bmod p$) $\rightarrow$ $2\log p$ bitů na popis h.

Chci $h: U \rightarrow \{0,\dots,m-1\}$ $\quad m \le |U| \le p$ (prvočíslo)

$$h_{a,b}(x) = ((ax+b) \bmod p) \bmod m$$

kde a,b jsou volby náhodně z $\{0,\dots,p-1\}$

Platí: $\forall x \neq y \in U \quad \forall r,q \in \{0,\dots m-1\}$

$$\frac{1}{4m^2} \le \Pr_{a,b}\left[h_{a,b}(x)=r \;\&\; h_{a,b}(y)=q\right] \le \frac{4}{m^2}$$

(nevýhoda – vyžaduje počítání $\bmod p$ … pomalé)

2) $w, k \ge 1$ celá čísla

$$h: \{0,1\}^w \rightarrow \{0,1\}^k$$

$$\mathcal{H} = \{ h_{A,b} : \{0,1\}^w \rightarrow \{0,1\}^k, \; A \in \{0,1\}^{k \times w}, \; b \in \{0,1\}^k \}$$

kde $h_{A,b}(x) = Ax + b$.
↳ maticové násobení $\bmod GF[2]$.

3) konvoluce

$w, k \ge 1$ celá čísla

$$h: \{0,1\}^w \rightarrow \{0,1\}^k$$

$$\mathcal{H} = \{ h_{a,b} \} \quad a \in \{0,1\}^{w+k-1} \;\&\; b \in \{0,1\}^k \}$$

$$\left(h_{a,b}(x)\right)_{\substack{j\text{-tý} \\ \text{bit}}} = b_j + \sum_{i=1}^{w} a_{i+j-1} \cdot x_i \qquad j = 1,\dots,k$$

4) multiply-shift (Dietzfelbinger '92)

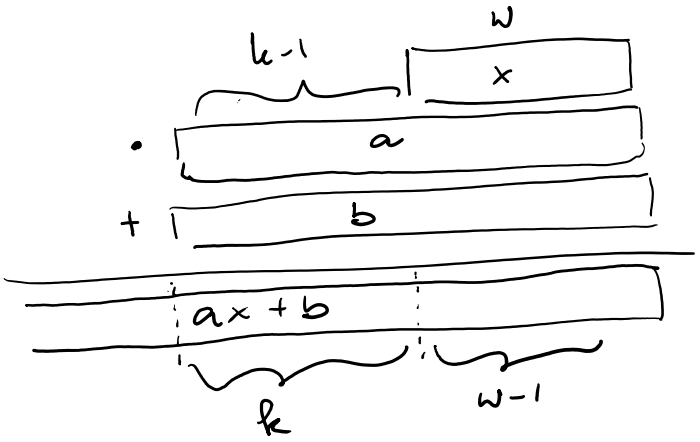
$w, k \geq 1$ celá čísla

$h : \{0,1\}^w \to \{0,1\}^k$

$\mathcal{H} = \{ h_{a,b} : a,b \in \{0,1\}^{w+k-1} \}$

$h_{a,b}(x) = \left[(ax+b) >> (w-1) \right]_{1..k}$

↑ nejvyšší bity 1...k



např. $w = 32$
$k = 16$
stačí 64-bitová aritmetika
– jedno násobení, >>, &

lze také zrobit podmnožinu
bitů 1..k → pro $k' \leq k$ 1..k'

---

2), 3) nepraktické

4) rychlé, snadné, praktické, nepotřebuje dělení, pouze jedno násobení

1) často používané, potřebuje dělení.

---

5) vektory $h : \{0,1\}^{dw} \to \{0,1\}^k$ $w' \geq w+k-1$

$\mathcal{H} = \{ h_{a_1, a_2, \ldots a_d, b} : a_1, \ldots a_d, b \in \{0,1\}^{w'} \}$

$h_{a_1 \ldots a_d, b}(x_1, \ldots x_d) = \left[\sum_{i=1}^{d} a_i x_i + b \right]_{w'-k+1 \ldots w'}$

$$h_{a_1 \dots a_d, b}(x_1, \dots x_d) = \left\lfloor \sum_{i=1}^{} a_i x_i + b \right\rfloor_{w'-k+1, \dots, w'}$$
$\in \{0,1\}^w$

$a_1, \dots a_d, b$ zvolím náhodně

• pro sudé d lze též použít:

$$h_{a_1 \dots a_d, b}(x_1, \dots x_d) = \left[\sum_{j=1}^{\frac{d}{2}} (x_{2j} + a_{2j})(x_{2j-1} + a_{2j-1}) + b \right]_{w'-k+1 \dots w'}$$

→ ušetří se polovina násobení

- vektory proměnné délky $d' < d$ se doplní 0 na d souřadnic

nebo pro d' sudé:

$$h_{a_1 \dots a_d}(x) = \left[\sum_{j=1}^{\frac{d'}{2}} (x_{2j} + a_{2j})(x_{2j-1} + a_{2j-1}) + a_{d'+1} \right]_{w'-k+1, \dots, w'}$$

Hašování řetězci:

$x_0, \dots x_{d-1} \in U$ prvočíslo $p \geq |U|$

$a \in \{0, \dots, p-1\}$

$$h_a(x_0, x_1, \dots x_{d-1}) = \sum_{i=0}^{d-1} x_i a^i \mod p$$

(lze počítat Hornerovým schématem odzadu:

$v_1 \leftarrow x_{d-1}$; $v_i \leftarrow v_{i-1} \cdot a + x_{d-i}$

$h_a(x_0, \dots x_{d-1}) = v_d$)

- $\forall x_1, \dots x_{d-1}, y_0, \dots y_{d-1} \in U \qquad \overline{x} \neq \overline{y}$

$$P_{r_a}\left[h_a(x_0, \dots x_{d-1}) = h_a(y_0, \dots y_{d-1})\right] \leq \frac{d}{p}$$

Dk: dva různé polynomy stupně $d-1$ se mohou shodovat v nejvýše $d-1$ bodech nad $GF[p]$. ☑

Př: $p = 2^{89} - 1$ ← Mersenovo prvočíslo $\qquad d \leq 2^{57} \rightarrow$ pravděpodobnost kolize $\leq \frac{1}{2^{32}}$

$h_a(\cdot\cdot)$ lze složit s náhodnou fcí $\{0, \dots, p-1\} \rightarrow \{0, \dots m-1\}$

$a, b, c \in \{0, \dots, p-1\}$

$$h_{a,b,c}(x_0, \dots x_{d-1}) = \left(\left(a\left(\sum_{i=0}^{d-1} x_i c^i\right) + b\right) \bmod p\right) \bmod m.$$

pokud $d \leq \frac{p}{m}$ pak je pst. kolize $\leq \frac{2}{m}$ ← dvou prvků.

Mersenova prvočísla: $2^{31}-1, 2^{61}-1, 2^{89}-1, 2^{107}-1$

$$p = 2^a - 1$$

$$y \bmod p \longrightarrow y = (y \,\&\, p) + (y >> a) \pmod p$$

↑ binární AND po bitech

---

k-nezávislé náhodní

$x \in \{0, \dots, p-1\}$ $\qquad p$ prvočíslo

$a_1, \dots a_k \in \{0, \dots, p-1\}$ náhodně

$$h_{a_1,\dots a_k}(x) = \sum_{i=1}^{k} a_i x^{i-1} \bmod p$$

$\mathcal{H} = \{ h_{a_1,\dots a_k} : \{0,\dots,p-1\} \to \{0,\dots p-1\} \}$ je k-nezávislé