Datové struktury I

6. přednáška: Výběr hešovací funkce

Jirka Fink

https://ktiml.mff.cuni.cz/~fink/

Katedra teoretické informatiky a matematické logiky
Matematicko-fyzikální fakulta
Univerzita Karlova v Praze

Zimní semestr 2024/25

Základní pojmy

- Značení: $[n] = \{0, \dots, n-1\}$
- Máme univerzum U všech prvků
- Chceme uložit podmnožinu $S \subseteq U$ velikosti n
- Uložíme S do pole velikosti m pomocí hešovací funkce $h : U \rightarrow M$, kde $M = [m]$
- Hešovacím systémem $\mathcal{H}$ rozumíme libovolnou množinu hešovacích funkcí
- Dva prvky $x, y \in S$ kolidují, jestliže $h(x) = h(y)$
- Uvažujeme universum $U = [u]$ pro libovolné $u \in \mathbb{N}$, pokud není uvedeno jinak

c-universální hešovací systém

c-universální systém (ekvivalentní definice)

Systém hešovacích funkcí $\mathcal{H}$ je c-universální, jestliže pro všechna různá $x, y \in U$ ①

- počet hešovacích funkcí $h \in \mathcal{H}$ splňujících $h(x) = h(y)$ je nejvýše $\frac{c|\mathcal{H}|}{m}$
- náhodně zvolená $h \in \mathcal{H}$ splňuje $P[h(x) = h(y)] \leq \frac{c}{m}$ ② ③

Příklad c-universálního hešovacího systému

- Parametry: p a m, kde $p > u$ je prvočíslo
- Hešovací funkce $h_a(x) = (ax \mod p) \mod m$
- Hešovací systém $\mathcal{H} = \{h_a;\ a \in [p]\}$ je c-universální
- Hešovací funkce ze systému $\mathcal{H}$ je určena hodnotou a
- Tedy náhodný výběr hešovací funkce z $\mathcal{H}$ je náhodné vygenerování $a \in [p]$

1. Navíc obvykle vyžadujeme, aby hešovací funkci šlo spočítat v čase $\mathcal{O}(1)$ a aby funkci bylo možné popsat $\mathcal{O}(1)$ parametry.
2. Náhodný výběr hešovací funkce má vždy rovnoměrné rozdělení na celém systému.
3. Úplně náhodný hešovací systém je 1-universální, protože $h(x)$ padne do nějaké přihrádky a $h(y)$ má uniformní distribuci nezávislou na $h(x)$, a proto $P[h(x) = h(y)] = \frac{1}{m}$.

(k,c)-nezávislý hešovací systém

(2,c)-nezávislý systém hešovacích funkcí (ekvivalentní definice)

Systém hešovacích funkcí $\mathcal{H}$ je $(2, c)$-nezávislý, pokud pro každé $x_1, x_2 \in U$ a $x_1 \neq x_2$ a $z_1, z_2 \in M$

- počet $h \in \mathcal{H}$ splňujících $h(x_1) = z_1$ a $h(x_2) = z_2$ je nejvýše $\frac{c|\mathcal{H}|}{m^2}$
- náhodně zvolená $h \in \mathcal{H}$ splňuje $P[h(x_1) = z_1 \text{ a } h(x_2) = z_2] \leq \frac{c}{m^2}$

(k, c)-nezávislý systém hešovacích funkcí

Nechť $k \in \mathbb{N}$, $K = \{1, \dots, k\}$ a $c \geq 1$.
Systém hešovacích funkcí $\mathcal{H}$ je (k, c)-nezávislý, pokud náhodně zvolená $h \in \mathcal{H}$ splňuje

$$P[h(x_i) = z_i \, \forall i \in K] \leq \frac{c}{m^k}$$

pro všechna po dvou různá $x_1, \dots, x_k \in U$ a všechna $z_1, \dots, z_k \in M$.

k-nezávislý systém hešovacích funkcí

- Systém $\mathcal{H}$ je k-nezávislý, pokud je (k, c)-nezávislý pro nějaké $c \geq 1$.
- Systém $\mathcal{H}$ je silně k-nezávislý, pokud je $(k, 1)$-nezávislý.

1. Systém $\mathcal{H} = \{h_a(x) = a;\ a \in M\}$ je 1-nezávislý, ale nepoužitelný ①
2. V definici k-nezávislého systému nemůžeme vyžadovat, aby přihrádky $z_1, \dots, z_k$ byly po dvou různé ②
3. (k, c)-nezávislý systém hešovacích funkcí je $(k-1, c)$-nezávislý ③
4. $(2, c)$-nezávislý systém hešovacích funkcí je c-universální ④
5. Existuje 1-universální systém, který není 2-nezávislý ⑤
6. Existuje silně k-nezávislý systém, který není $(k+1)$-nezávislý ⑥
7. Pro každý hešovací systém $\mathcal{H}$ a pro všechna po dvou různá $x_1, \dots, x_k \in U$ existují $z_1, \dots, z_k \in M$ taková, že $P[h(x_i) = z_i\, \forall i \in K] \geq \frac{1}{m^k}$ ⑦
8. Jestliže $\mathcal{H}$ je silně k-nezávislý, pak pro po dvou různá $x_1, \dots, x_k \in U$ a pro $z_1, \dots, z_k \in M$
    - $P[h(x_i) = z_i\, \forall i \in K] = \frac{1}{m^k}$
    - $P[h(x_k) = z_k | h(x_i) = z_i\, \forall i = 1, \dots, k-1] = \frac{1}{m}$
9. Jestliže $\mathcal{H}$ je (k, c)-nezávislý, pak $|\mathcal{H}| \geq \frac{m^k}{c}$ a na identifikaci funkce z $|\mathcal{H}|$ potřebujeme alespoň $k \log m - \log c$ bitů ⑧

1. Z tohoto důvodu jsme definovali k-nezávislost pro $k \geq 2$,
2. Podle takto upravené definice by systém $\mathcal{H} = \{h_a(x) = a;\ a \in M\}$ byl k-nezávislý pro libovolné k,
3. $P[h(x_i) = z_i \, \forall i = 1, \dots, k-1] = P[\exists z_k \in M :\ h(x_i) = z_i \, \forall i \in K] \leq \sum_{z_m \in M} P[h(x_i) = z_i \, \forall i \in K] \leq m \frac{c}{m^k} = \frac{c}{m^{k-1}}$
4. $P[h(x) = h(y)] = P[\exists z \in M :\ h(x) = z \text{ a } h(y) = z] \leq \sum_{z \in M} P[h(x) = z \text{ a } h(y) = z] \leq m \frac{c}{m^2} = \frac{c}{m}$
5. Uvažujme systém $\mathcal{H}$ všech funkcí $h : U \to M$ takových, že $h(0) = 0$ a $h(1) = 1$, t.j. dva prvky mají pevné přihrádky a ostatní prvky náhodné přihrádky. Pak $P[h(x) = h(y)] \leq \frac{1}{m}$, ale $P[h(0) = 0 \text{ a } h(1) = 1] = 1$.
6. $\mathcal{H} = \{h : U \to M;\ h(0) + h(1) + \cdots + h(k) = 0 \mod m\}$
7. Kdyby $P[h(x_i) = z_i \, \forall i \in K] < \frac{1}{m^k}$ pro všechna $z_1, \dots, z_k \in M$, pak $1 = P[\exists z_1, \dots, z_k \in M :\ h(x_i) = z_i \, \forall i \in K] \leq \sum_{z_1, \dots, z_k \in M} P[h(x_i) = z_i \, \forall i \in K] < m^k \frac{1}{m^k} = 1$.
8. Zvolme $h' \in \mathcal{H}$ a $x_1, \dots, x_n \in U$ různé. Nechť $z_i = h'(x_i)$ a β značí počet $h \in \mathcal{H}$ splňujících $h(x_i) = z_i$ pro všechna $i \in K$. Zřejmě $\beta \geq 1$. Z $P[h(x_i) = z_i \, \forall i \in K] = \frac{\beta}{|\mathcal{H}|} \leq \frac{c}{m^k}$ plyne $|\mathcal{H}| \geq \frac{m^k}{c}$.

Lineární kongruence

Lemma

Pro libovolná různá $x_1, x_2 \in [p]$ rovnice

$$y_1 = ax_1 + b \mod p$$
$$y_2 = ax_2 + b \mod p$$

definují bijekci mezi $(a, b) \in [p]^2$ a $(y_1, y_2) \in [p]^2$, kde p je prvočíslo.

Důkaz

Pro danou dvojici (y_1, y_2) existuje jediná dvojice (a, b) splňující rovnice

- Odečtením dostáváme $a(x_1 - x_2) \equiv_p y_1 - y_2$ ①
- V tělese $GF(p) = \mathbb{Z}_p$ dostáváme $a = (y_1 - y_2)(x_1 - x_2)^{-1}$, $b = y_1 - ax_1$

Pozorování

- Nechť $p \geq |U|$ je prvočíslo, kde $U = [u]$
- Uvažujme hešovací funkci $h_{a,b}(x) = ax + b \mod p$
- Pak systém $\mathcal{H} = \{h_{a,b};\ a, b \in [p]\}$ je $(2, 1)$-nezávislý ②

1. $\equiv_n$ značí rovnost modulo n
2. Důkaz plyne z předcházejícího lemmatu, protože bijekce zaručuje, že pro různá $x_1, x_2 \in [p]$ a $y_1, y_2 \in [p]$ existuje právě jedna $h \in \mathcal{H}$ taková, že $h(x_1) = y_1$ a $h(x_2) = y_2$.

Skládání modulo m

Lemma

- Nechť systém $\mathcal{H}$ je $(2, c)$-nezávislý z U do $[r]$ a $m \leq r$
- Pak $\mathcal{H} \bmod m = \{x \to h(x) \bmod m;\ h \in \mathcal{H}\}$ je $2c$-universální a $(2, 4c)$-nezávislý

Důkaz

- Zvolme různá $x_1, x_2 \in U$ a označme $y_1 = h(x_1)$ a $y_2 = h(x_2)$
- $2c$-universálnost:
  - $P[h(x_1) \equiv_m h(x_2)] = \sum_{y_1 \equiv_m y_2} P[h(x_1) = y_1 \text{ a } h(x_2) = y_2]$
  - $P[h(x_1) = y_1 \text{ a } h(x_2) = y_2] \leq c/r^2$
  - Sumu sčítáme přes r hodnot y_1 a $\lceil r/m \rceil$ hodnot y_2
  - $\lceil \frac{r}{m} \rceil \leq \frac{r+m-1}{m} \leq \frac{2r}{m}$
  - Celkem $P[h(x_1) \equiv_m h(x_2)] \leq \frac{c}{r^2} \cdot r \cdot \frac{2r}{m} = \frac{2c}{m}$
- $(2, 4c)$-nezávislost
  - $P[h(x_1) = z_1 \text{ a } h(x_2) = z_2] = \sum_{y_1 \equiv_m z_1 \text{ a } y_2 \equiv_m z_2} P[h(x_1) = y_1 \text{ a } h(x_2) = y_2]$
  - Sumu sčítáme přes nejvýše $\lceil r/m \rceil$ hodnot y_1 a y_2
  - Celkem $P[h(x_1) = z_1 \text{ a } h(x_2) = z_2] \leq \frac{c}{r^2} \cdot \left(\frac{2r}{m}\right)^2 = \frac{4c}{m^2}$

Multiply-mod-prime

Pozorování

- Nechť $p \geq |U|$ je prvočíslo, kde $U = [u]$
- Uvažujme hešovací funkci $h_{a,b}(x) = ax + b \bmod p$
- Pak systém $\mathcal{H} = \{h_{a,b};\ a, b \in [p]\}$ je $(2, 1)$-nezávislý ①

Lemma

- Nechť systém $\mathcal{H}$ je $(2, c)$-nezávislý z U do $[r]$ a $r \geq m$
- Pak $\mathcal{H} \bmod p = \{x \to h(x) \bmod p;\ h \in \mathcal{H}\}$ je $2c$-universální a $(2, 4c)$-nezávislý

Pozorování: Systém Multiply-mod-prime

- Nechť $p \geq |U| \geq m$ je prvočíslo, kde $U = [u]$
- $h_{a,b}(x) = (ax + b \bmod p) \bmod m$
- $\mathcal{H} = \{h_{a,b};\ a, b \in [p]\}$
- Systém $\mathcal{H}$ je 2-universální a (2,4)-nezávislý, ale není 3-nezávislý

1. Důkaz plyne z následujícího lemmatu, protože bijekce zaručuje, že pro různá $x_1, x_2 \in [p]$ a $y_1, y_2 \in [p]$ existuje právě jedna $h \in \mathcal{H}$ taková, že $h(x_1) = y_1$ a $h(x_2) = y_2$.

Zobecněné modulo m

Lemma

- Nechť systém $\mathcal{H}$ je (k, c)-nezávislý z U do $[r]$ a $r \geq 2km$
- Pak $\mathcal{H} \mod m = \{x \to h(x) \mod m;\ h \in \mathcal{H}\}$ je $(k, 2c)$-nezávislý

Důkaz

- Zvolme různá $x_1, \dots, x_k \in U$, $z_1, \dots, z_k \in M$ a označme $y_i = h(x_i)$
- $P[h(x_i) \mod m = z_i\ \forall i \in K] = \sum P[y_i \equiv_m z_i\ \forall i \in K]$
- Sumu sčítáme přes nejvýše $\lceil \frac{r}{m} \rceil \leq \frac{r+m-1}{m}$ hodnot $y_1, \dots, y_k$
- Z odhadu $1 + x \leq e^x$ plyne
  $P[h(x_i) \mod m = z_i\ \forall i \in K] \leq \frac{c}{r^k} \cdot \left(\frac{r+m-1}{m}\right)^k = \frac{c}{m^k} \cdot \left(\frac{r+m-1}{r}\right)^k \leq$
  $\frac{c}{m^k} \cdot \left(1 + \frac{m}{r}\right)^k = \frac{c}{m^k} \cdot e^{km/r} \leq \frac{c}{m^k} \cdot e^{1/2} \leq \frac{2c}{m^k}$

Poly-mod-prime

Lemma

- Nechť systém $\mathcal{H}$ je (k, c)-nezávislý z U do $[r]$ a $r \geq 2km$
- Pak $\mathcal{H} \mod m = \{x \to h(x) \mod m;\ h \in \mathcal{H}\}$ je $(k, 2c)$-nezávislý

Věta z algebry: Jednoznačnost interpolace polynomem

Pro každé těleso T, $k > 1$ celočíselné, po dvou různá $x_1, \dots, x_k \in T$ a $y_1, \dots, y_k \in T$ existuje právě jeden polynom $p_a(x) = \sum_{i=0}^{k-1} a_i x^i$ stupně $k - 1$ s koeficienty $a_0, \dots, a_{k-1} \in T$ takový, že $p(x_i) = y_i$ pro všechna $i \in K$.

Pozorování: Systém Poly-mod-prime

- Nechť p je prvočíslo, $a \in \mathbb{Z}_p^k$, $x \in \mathbb{Z}_p$
- $h_a(x) = \sum_{i=0}^{k-1} a_i x^i$
- Systém $P_k = \left\{h_a;\ a \in \mathbb{Z}_p^k\right\}$ je $(k, 1)$-nezávislý ①
- Systém $P_k \mod m$ je $(k, 2)$-nezávislý pro $p \geq 2km$ ②

1. Důkaz přímo plyne z jednoznačnosti polynomu
2. Jestliže p je prvočíslo, tak hešovací funkci lze zapsat jako $h_a(x) = (\sum_{i=0}^{k-1} a_i x^i \mod p) \mod m$, kde aritmetické operace jsou nad celými čísly.

Multiply-shift

Multiply-shift

- Předpokládáme, že $|U| = 2^w$ a $m = 2^l$
- $h_a(x) = (ax \bmod 2^w) >> (w - l)$
- $\mathcal{H} = \{h_a;\ a$ je liché w-bitové číslo $\}$

Implementace v C

```
uint64_t hash(uint64_t x, uint64_t l, uint64_t a)
{ return (a*x) >> (64-l); }
```

Vlastnosti systému multiply-shift

- 2-universální
- Velmi rychlý na reálných počítačích
- V praxi často používaný
- Celý výpočet musí být proveden v neznaménkových celočíselných typech, protože ze součinu *ax* potřebujeme získat posledních *w* bitů

Tabulkové hešování

- $\oplus$ značí bitový XOR
- Předpokládáme, že $u = 2^w$ a $m = 2^l$ a w je násobek $d \geq 2$
- Bitový zápis čísla $x \in U$ rozdělíme na d částí $x^1, \dots, x^d$ po $\frac{w}{d}$ bitech
- Pro každé $i = 1, \dots, d$ vybereme náhodnou hešovací funkci $T_i : [2^{w/d}] \to M$
- Hešovací funkce je $h(x) = T_1(x^1) \oplus \cdots \oplus T_d(x^d)$
- K vygenerování h potřebujeme $d \cdot 2^{w/d}$ náhodných čísel z rozsahu $M = [2^l]$

Ilustrativní příklad

- Uvažujme $w = 12$ a $d = 3$
- Nejprve vygeneruje náhodné funkce $T_1, T_2, T_3 : [2^4] \to M$
- Číslo $x = 101100111001$ rozdělíme na $x^1 = 1011$, $x^2 = 0011$ a $x^3 = 1001$
- Výpočet hešovací funkce je
  $h(x) = T_1(x^1) \oplus T_2(x^2) \oplus T_3(x^3) = T_1(1011) \oplus T_2(0011) \oplus T_3(1001)$

Univerzalita

Tabulkové hešování je silně 3-nezávislé, ale není 4-nezávislé.

Tabulkové hešování

- Předpokládáme, že $u = 2^w$ a $m = 2^l$ a w je násobek d
- Bitový zápis čísla $x \in U$ rozdělíme na d částí $x^1, \dots, x^d$ po $\frac{w}{d}$ bitech
- Pro každé $i = 1, \dots, d$ vybereme náhodnou hešovací funkci $T_i : [2^{w/d}] \to M$
- Hešovací funkce je $h(x) = T_1(x^1) \oplus \cdots \oplus T_d(x^d)$

Univerzalita

Tabulkové hešování je 3-nezávislé, ale není 4-nezávislé.

Důkaz 2-nezávislosti (3-nezávislost je ponechána na cvičení)

- Mějme dva prvky x_1 a x_2 lišící se v i-tých částech
- Nechť $h_i(x) = T_1(x^1) \oplus \cdots \oplus T_{i-1}(x^{i-1}) \oplus T_{i+1}(x^{i+1}) \oplus \cdots \oplus T_d(x^d)$
- $P[h(x_1) = z_1] = P[h_i(x_1) \oplus T_i(x_1^i) = z_1] = P[T_i(x_1^i) = z_1 \oplus h_i(x_1)] = \frac{1}{m}$ 1
- Náhodné jevy $h(x_1) = z_1$ a $h(x_2) = z_2$ jsou nezávislé
  - Náhodné proměnné $T_i(x_1^i)$ a $T_i(x_2^i)$ jsou nezávislé
  - Náhodné jevy $T_i(x_1^i) = z_1 \oplus h_i(x_1)$ a $T_i(x_2^i) = z_2 \oplus h_i(x_2)$ jsou nezávislé
- $P[h(x_1) = z_1 \text{ a } h(x_2) = z_2] = P[h(x_1) = z_1]P[h(x_2) = z_2] = \frac{1}{m^2}$

1. $T_i(x_1^i)$ nabývá všech hodnot z M se stejnou pravděpodobností $\frac{1}{m}$ a náhodné proměnné $T_i(x_1^i)$ a $z_1 \oplus h_i(x_1)$ jsou nezávislé.

Tabulkové hešování

Tabulkové hešování není 4-nezávislé

1. Zvolíme prvky x_1, x_2, x_3 a x_4 takové, že
   - části x_1 splňují $x_1^1 = 0$, $x_1^2 = 0$, $x_1^i = 0$ pro $i \geq 3$
   - části x_2 splňují $x_2^1 = 1$, $x_2^2 = 0$, $x_2^i = 0$ pro $i \geq 3$
   - části x_3 splňují $x_3^1 = 0$, $x_3^2 = 1$, $x_3^i = 0$ pro $i \geq 3$
   - části x_4 splňují $x_4^1 = 1$, $x_4^2 = 1$, $x_4^i = 0$ pro $i \geq 3$
2. Platí $h(x_1) \oplus h(x_2) \oplus h(x_3) = h(x_4)$
3. Zvolme libovolná z_1, z_2, z_3 a nechť $z_4 = z_1 \oplus z_2 \oplus z_3$
4. Jestliže $h(x_1) = z_1$, $h(x_2) = z_2$ a $h(x_3) = z_3$, pak platí $h(x_4) = z_4$
5. $P[h(x_1) = z_1 \text{ a } h(x_2) = z_2 \text{ a } h(x_3) = z_3 \text{ a } h(x_4) = z_4] = \frac{1}{m^3} > \frac{c}{m^4}$

Hešování posloupností pevné délky

Scalar-mod-prime

- Chceme hešovat d-tici $x_1, \ldots, x_d \in \mathbb{Z}_p$, kde p je prvočíslo
- $\left\{x_1, \ldots, x_d \to \sum_{i=1}^d a_i x_i \mod p;\ a \in \mathbb{Z}_p^d\right\}$ je 1-universální
- $\left\{x_1, \ldots, x_d \to b + \sum_{i=1}^d a_i x_i \mod p;\ a \in \mathbb{Z}_p^d, b \in \mathbb{Z}_p\right\}$ je (2,1)-nezávislý
- $\left\{x_1, \ldots, x_d \to \left(b + \sum_{i=1}^d a_i x_i \mod p\right) \mod m;\ a \in \mathbb{Z}_p^d, b \in \mathbb{Z}_p\right\}$ je (2,4)-nezávislý

Důkaz 1-universálnosti ①

- Mějme různé $x, y \in \mathbb{Z}_p^d$ a BÚNO předpokládejme, že $x_1 \neq y_1$
- $P[a \cdot x \equiv_p a \cdot y] = P[a \cdot (x - y) \equiv_p 0] = P\left[a_1 \equiv_p \frac{\sum_{i=2}^d a_i(y_i - x_i)}{x_1 - y_1}\right] = 1/p$ ②

1. Pro 2-nezávislost stačí podobně nahlednout, že a_1, b jsou jednoznačně určené.
2. Náhodná proměnná a_1 musí nabývat jednu konkrétní hodnotu, což nastane s pravděpodobností $1/p$.

Hešování řetězců různých délek

Poly-mod-prime pro různě dlouhé řetězce I

- Chceme hešovat řetězec $x_1, \ldots, x_d \in \mathbb{Z}_p$, kde p je prvočíslo
- $\left\{x_1, \ldots, x_d \to \sum_{i=0}^{d-1} x_{i+1} a^i \mod p;\ a \in [p]\right\}$ je d-universální
- Dva různé polynomy stupně nejvýše $d - 1$ mají nejvýše d společných bodů, takže existuje nejvýše d kolidujících hodnot a.

Poly-mod-prime pro různě dlouhé řetězce II

- Chceme hešovat řetězec $x_1, \ldots, x_d \in U$ do M, kde $p \geq m$ je prvočíslo
- $h_{a,b,c}(x_1, \ldots, x_d) = \left(b + c\sum_{i=0}^{d-1} x_{i+1} a^i \mod p\right) \mod m$
- $\mathcal{H} = \{h_{a,b,c};\ a, b, c \in [p]\}$
- $P[h_{a,b,c}(x_1, \ldots, x_d) = h_{a,b,c}(x'_1, \ldots, x'_{d'})] \leq \frac{2}{m}$ pro různé řetězce délek $d, d' \leq \frac{p}{m}$.