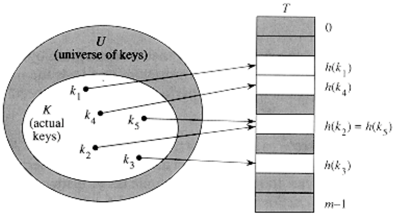
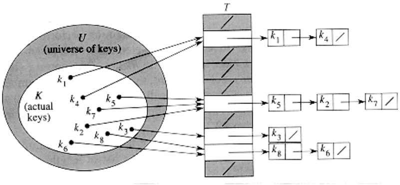
Hešovanie

kuko

20.10.2020

Vybrané partie z dátových štruktúr


T
0
U
(universe of keys)
K
(actual
keys)
k_1
k_4
k_5
k_2
k_3
$h(k_1)$
$h(k_4)$
$h(k_2) = h(k_5)$
$h(k_3)$
$m-1$


T
U
(universe of keys)
K
(actual
keys)
k1
k4
k5
k7
k2
k3
k8
k6

priemerná čas. zlož.

$\Longrightarrow$

očakávaná čas. zlož.

rodina hešovacích funkcií $\mathscr{H}$

$$h \in_R \mathscr{H}$$

$\mathscr{H}$ je univerzálna ak

$$\forall x_1 \neq x_2 : \Pr_{h \in_R \mathscr{H}}[h(x_1) = h(x_2)] \leq 1/n$$

$\mathscr{H}$ je 1-nezávislá, ak

$$\forall x, y : \Pr_{h \in_R \mathscr{H}}[h(x_1) = y] = 1/n$$

$\mathscr{H}$ je 1-nezávislá, ak

$$\forall x, y : \Pr_{h \in_R \mathscr{H}}[h(x_1) = y] = O(1/n)$$

$\mathscr{H}$ je k-nezávislá, ak

$$\forall \underbrace{x_1, \ldots, x_k}_{\text{rôzne}}, y_1, \ldots, y_k : \Pr_{h \in_R \mathscr{H}}[h(x_1) = y_1 \wedge \cdots \wedge h(x_k) = y_k] = O(1/n^k)$$

$\forall x, y : \Pr_h[h(x) = y] = O(1/n)$ a pre rôzne $x_1, \ldots, x_k$ sú náhodné premenné $h(x_1), \ldots, h(x_k)$ skoro nezávislé

$\mathscr{H}$ je k-nezávislá, ak

$$\forall \underbrace{x_1, \ldots, x_k}_{\text{rôzne}}, y_1, \ldots, y_k : \Pr_{h \in_R \mathscr{H}}[h(x_1) = y_1 \wedge \cdots \wedge h(x_k) = y_k] = O(1/n^k)$$

$\forall x, y : \Pr_h[h(x) = y] = O(1/n)$ a pre rôzne $x_1, \ldots, x_k$ sú náhodné premenné $h(x_1), \ldots, h(x_k)$ skoro nezávislé

- $x \mapsto (ax \bmod p) \bmod n$ je univerzálna
- $x \mapsto (ax) >> (w-k)$ (ak $n = 2^k$ a register má w bitov) je univerzálna
- $x \mapsto ((ax+b) \bmod p) \bmod n$ sú 2-nezávislé $(a, b \in \mathbb{Z}_p,\ a \neq 0)$
- všeobecne $x \mapsto ((a_k x^k + \cdots + a_1 x + a_0) \bmod p) \bmod n$ sú k-nezávislé
- jednoduché tabulačné hešovanie je 3-nezávislé
  - vygenerujeme si tabuľky $T_1, \ldots, T_c$ veľkosti $u^{1/c}$ s úplne náhodnou hešovacou fn.;
  - na $x \in U$ sa pozeráme ako na vektor $x = x_1 \cdots x_c$ a $h(x) = T_1(x_1) \oplus \cdots \oplus T_c(x_c)$

- $x \mapsto (ax \bmod p) \bmod n$ je univerzálna
- $x \mapsto (ax) >> (w-k)$ (ak $n = 2^k$ a register má w bitov) je univerzálna
- $x \mapsto ((ax+b) \bmod p) \bmod n$ sú 2-nezávislé $(a, b \in \mathbb{Z}_p,\ a \neq 0)$
- všeobecne $x \mapsto ((a_k x^k + \cdots + a_1 x + a_0) \bmod p) \bmod n$ sú k-nezávislé
- jednoduché tabulačné hešovanie je 3-nezávislé
  - vygenerujeme si tabuľky $T_1, \ldots, T_c$ veľkosti $u^{1/c}$ s úplne náhodnou hešovacou fn.;
  - na $x \in U$ sa pozeráme ako na vektor $x = x_1 \cdots x_c$ a $h(x) = T_1(x_1) \oplus \cdots \oplus T_c(x_c)$

- $x \mapsto (ax \bmod p) \bmod n$ je univerzálna
- $x \mapsto (ax) >> (w-k)$ (ak $n = 2^k$ a register má w bitov) je univerzálna
- $x \mapsto ((ax+b) \bmod p) \bmod n$ sú 2-nezávislé $(a, b \in \mathbb{Z}_p,\ a \neq 0)$
- všeobecne $x \mapsto ((a_k x^k + \cdots + a_1 x + a_0) \bmod p) \bmod n$ sú k-nezávislé
- jednoduché tabulačné hešovanie je 3-nezávislé
  - vygenerujeme si tabuľky $T_1, \ldots, T_c$ veľkosti $u^{1/c}$ s úplne náhodnou hešovacou fn.;
  - na $x \in U$ sa pozeráme ako na vektor $x = x_1 \cdots x_c$ a $h(x) = T_1(x_1) \oplus \cdots \oplus T_c(x_c)$

- $x \mapsto (ax \bmod p) \bmod n$ je univerzálna
- $x \mapsto (ax) >> (w-k)$ (ak $n = 2^k$ a register má w bitov) je univerzálna
- $x \mapsto ((ax+b) \bmod p) \bmod n$ sú 2-nezávislé $(a, b \in \mathbb{Z}_p,\ a \neq 0)$
- všeobecne $x \mapsto ((a_k x^k + \cdots + a_1 x + a_0) \bmod p) \bmod n$ sú k-nezávislé
- jednoduché tabulačné hešovanie je 3-nezávislé
  - vygenerujeme si tabuľky $T_1, \ldots, T_c$ veľkosti $u^{1/c}$ s úplne náhodnou hešovacou fn.;
  - na $x \in U$ sa pozeráme ako na vektor $x = x_1 \cdots x_c$ a $h(x) = T_1(x_1) \oplus \cdots \oplus T_c(x_c)$

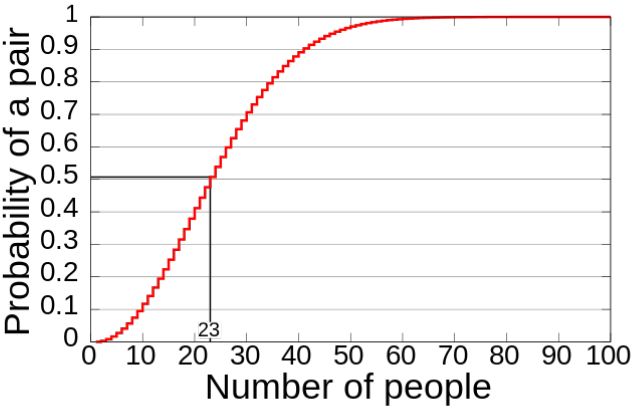
- $x \mapsto (ax \bmod p) \bmod n$ je univerzálna
- $x \mapsto (ax) >> (w-k)$ (ak $n = 2^k$ a register má w bitov) je univerzálna
- $x \mapsto ((ax+b) \bmod p) \bmod n$ sú 2-nezávislé ($a, b \in \mathbb{Z}_p$, $a \neq 0$)
- všeobecne $x \mapsto ((a_k x^k + \cdots + a_1 x + a_0) \bmod p) \bmod n$ sú k-nezávislé
- jednoduché tabulačné hešovanie je 3-nezávislé
  - vygenerujeme si tabuľky $T_1, \ldots, T_c$ veľkosti $u^{1/c}$ s úplne náhodnou hešovacou fn.;
  - na $x \in U$ sa pozeráme ako na vektor $x = x_1 \cdots x_c$ a $h(x) = T_1(x_1) \oplus \cdots \oplus T_c(x_c)$

hešovanie so zreťazením	$O(1)$	univerzálna rodina $\mathscr{H}$
lineárne sondovanie	$O(1/\varepsilon^2)$	$n \geq (1+\varepsilon)m$
	$O(1/\varepsilon^2)$	5-nezávislá rodina $\mathscr{H}$
	$O(1/\varepsilon^2)$	tabulačné hešovanie

dá sa vyhľadávanie v $O(1)$ v najhoršom prípade?
čo keby sme vedeli všetky prvky dopredu?

Perfektné hešovanie

Narodeninový paradox



$$1 \times \left(1 - \frac{1}{365}\right) \times \left(1 - \frac{2}{365}\right) \times \cdots \times \left(1 - \frac{n-1}{365}\right)$$

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

$$\approx 1 \cdot e^{-1/365} \cdot e^{-2/365} \cdots e^{-(n-1)/365} = e^{-(n(n-1)/2)/365}$$

$$1 \times \left(1 - \frac{1}{365}\right) \times \left(1 - \frac{2}{365}\right) \times \cdots \times \left(1 - \frac{n-1}{365}\right)$$

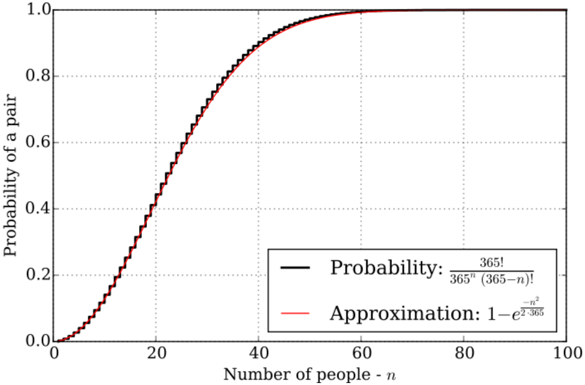
$$e^x \approx 1 + x$$

$$\approx 1 \cdot e^{-1/365} \cdot e^{-2/365} \cdots e^{-(n-1)/365} = e^{-(n(n-1)/2)/365}$$

$$1 \times \left(1 - \frac{1}{365}\right) \times \left(1 - \frac{2}{365}\right) \times \cdots \times \left(1 - \frac{n-1}{365}\right)$$

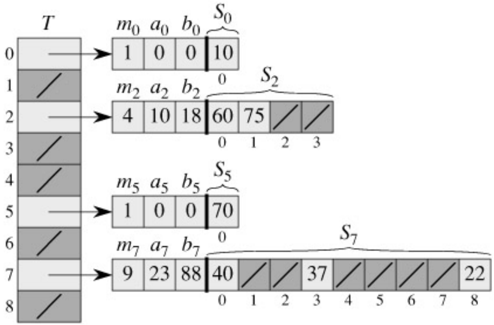
$$e^x \approx 1 + x$$

$$\approx 1 \cdot e^{-1/365} \cdot e^{-2/365} \cdots e^{-(n-1)/365} = e^{-(n(n-1)/2)/365}$$



Probability of a pair
Number of people - n
Probability: $\frac{365!}{365^n\,(365-n)!}$
Approximation: $1-e^{\frac{-n^2}{2\cdot 365}}$
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
100

- (FKS'84) dve úrovne hešovania, namiesto zreťazenia tabuľku kvadratickej veľkosti



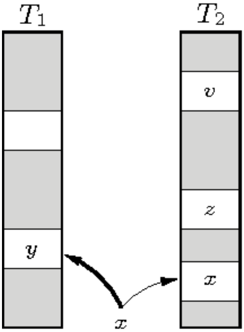
- $\mathrm{E}[\#\text{kolízií}] = m^2 \cdot O(1/n) \leq 1/2$ pre dosť veľké $n = \Theta(m^2)$
- žiadna kolízia s pp. aspoň 1/2 (Markov: $\Pr[X \geq k\mu] \leq 1/k$)
- $\mathrm{E}[\sum_t C_t^2] = \sum_t \mathrm{E}[C_t^2] = O(\mathrm{E}[\#\text{kolízií}]) + O(n)$
- $= m^2 \cdot O(1/n) = O(n)$ pre $m = \Theta(n)$

- $\mathrm{E}[\#\text{kolízií}] = m^2 \cdot O(1/n) \leq 1/2$ pre dosť veľké $n = \Theta(m^2)$
- žiadna kolízia s pp. aspoň 1/2 (Markov: $\Pr[X \geq k\mu] \leq 1/k$)
- $\mathrm{E}[\sum_t C_t^2] = \sum_t \mathrm{E}[C_t^2] = O(\mathrm{E}[\#\text{kolízií}]) + O(n)$
- $= m^2 \cdot O(1/n) = O(n)$ pre $m = \Theta(n)$

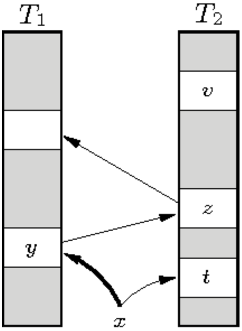
- $\mathrm{E}[\#\text{kolízií}] = m^2 \cdot O(1/n) \leq 1/2$ pre dosť veľké $n = \Theta(m^2)$
- žiadna kolízia s pp. aspoň $1/2$ (Markov: $\Pr[X \geq k\mu] \leq 1/k$)
- $\mathrm{E}[\sum_t C_t^2] = \sum_t \mathrm{E}[C_t^2] = O(\mathrm{E}[\#\text{kolízií}]) + O(n)$
- $= m^2 \cdot O(1/n) = O(n)$ pre $m = \Theta(n)$

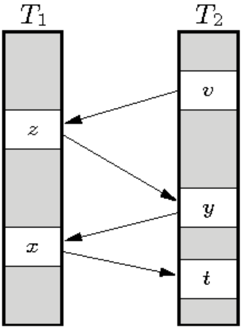
- $\mathrm{E}[\#\text{kolízií}] = m^2 \cdot O(1/n) \leq 1/2$ pre dosť veľké $n = \Theta(m^2)$
- žiadna kolízia s pp. aspoň 1/2 (Markov: $\Pr[X \geq k\mu] \leq 1/k$)
- $\mathrm{E}[\sum_t C_t^2] = \sum_t \mathrm{E}[C_t^2] = O(\mathrm{E}[\#\text{kolízií}]) + O(n)$
- $= m^2 \cdot O(1/n) = O(n)$ pre $m = \Theta(n)$

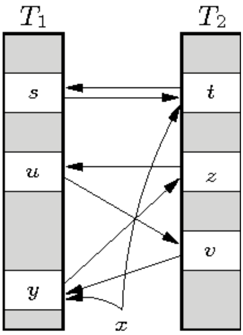
Kukučie hešovanie



- máme 2 tabuľky A, B dĺžky $2m$ a 2 hešovacie funkcie f, g
- hľadanie: prvok x je vždy v $A[f(x)]$ alebo $B[g(x)]$


T_1
T_2
v
z
y
t
x


T_1
T_2
v
z
y
x
t


T_1
T_2
s
t
u
z
v
y
x

Riešenie kolízií zreťazením

Koľko prvkov sa zahešuje na tú „najvyťaženejšiu“ pozíciu?

$$P_i = \#\text{prvkov, ktoré sa zahešujú na pozíciu } i$$

koľko je $\max P_i$?

- nech $p = 1/n = \Pr[\text{prvok zahešujeme na konkrétnu pozíciu}]$
- nech $q = 1 - p = \Pr[\text{prvok zahešuje inde}]$

$$\Pr[P_i = k] = \binom{n}{k} p^k q^{n-k}$$

- nech $p = 1/n = \Pr[\text{prvok zahešujeme na konkrétnu pozíciu}]$
- nech $q = 1 - p = \Pr[\text{prvok zahešuje inde}]$

$$\Pr[P_i = k] = \binom{n}{k} p^k q^{n-k}$$

- nech $p = 1/n = \Pr[\text{prvok zahešujeme na konkrétnu pozíciu}]$
- nech $q = 1 - p = \Pr[\text{prvok zahešuje inde}]$

$$\Pr[P_i = k] = \binom{n}{k} p^k q^{n-k}$$

$$\Pr[P_i = k] = \binom{n}{k} \left(\frac{1}{n}\right)^k \left(1 - \frac{1}{n}\right)^{n-k}$$

$$\left(\frac{n}{k}\right)^k \leq \binom{n}{k} \leq \left(\frac{ne}{k}\right)^k$$

$$1/e \leq \left(1 - \frac{1}{n}\right)^{n-k} \leq 1$$

$$\Pr[P_i = k] = \binom{n}{k}\left(\frac{1}{n}\right)^k\left(1-\frac{1}{n}\right)^{n-k}$$

$$\left(\frac{n}{k}\right)^k \le \binom{n}{k} \le \left(\frac{ne}{k}\right)^k$$

$$1/e \le \left(1-\frac{1}{n}\right)^{n-k} \le 1$$

$$\Pr[P_i = k] = \binom{n}{k} \left(\frac{1}{n}\right)^k \left(1 - \frac{1}{n}\right)^{n-k}$$

$$\left(\frac{n}{k}\right)^k \le \binom{n}{k} \le \left(\frac{ne}{k}\right)^k$$

$$1/e \le \left(1 - \frac{1}{n}\right)^{n-k} \le 1$$

$$\Pr[P_i = k] = \underbrace{\binom{n}{k}}_{\leq (ne/k)^k} \left(\frac{1}{n}\right)^k \underbrace{\left(1 - \frac{1}{n}\right)^{n-k}}_{\leq 1} \leq \left(\frac{\cancel{n}e}{k}\right)^k \cdot \frac{1}{\cancel{n}^k} = \left(\frac{e}{k}\right)^k$$

$$\Pr[P_i = k] = \underbrace{\binom{n}{k}}_{\geq (n/k)^k} \left(\frac{1}{n}\right)^k \underbrace{\left(1 - \frac{1}{n}\right)^{n-k}}_{\geq 1/e} \geq \left(\frac{\cancel{n}}{k}\right)^k \cdot \frac{1}{\cancel{n}^k} \cdot (1/e) = \frac{1}{ek^k}$$

$$1/ek^k \leq \Pr[P_i = k] \leq (e/k)^k$$

$$\Pr[P_i = k] = \frac{1}{k^{\Theta(k)}} = \frac{1}{e^{\Theta(k \log k)}}$$

- pre $k = \ln n / 3 \ln\ln n$ je aspoň $1/e\sqrt[3]{n}$
- $\Longrightarrow$ očakávaný počet pozícií, ktoré prekročia k je $\Omega(n^{2/3})$

$$1/ek^k \leq \Pr[P_i = k] \leq (e/k)^k$$

$$\Pr[P_i = k] = \frac{1}{k^{\Theta(k)}} = \frac{1}{e^{\Theta(k \log k)}}$$

- pre $k = \ln n / 3 \ln \ln n$ je aspoň $1/e\sqrt[3]{n}$
- $\Longrightarrow$ očakávaný počet pozícií, ktoré prekročia k je $\Omega(n^{2/3})$

$$1/ek^k \leq \Pr[P_i = k] \leq (e/k)^k$$

$$\Pr[P_i = k] = \frac{1}{k^{\Theta(k)}} = \frac{1}{e^{\Theta(k \log k)}}$$

- pre $k = \ln n / 3 \ln \ln n$ je aspoň $1/e\sqrt[3]{n}$
- $\Longrightarrow$ očakávaný počet pozícií, ktoré prekročia k je $\Omega(n^{2/3})$

$$1/ek^k \leq \Pr[P_i = k] \leq (e/k)^k$$

$$\Pr[P_i = k] = \frac{1}{k^{\Theta(k)}} = \frac{1}{e^{\Theta(k \log k)}}$$

- pre $k = \ln n / 3 \ln \ln n$ je aspoň $1/e\sqrt[3]{n}$
- $\Longrightarrow$ očakávaný počet pozícií, ktoré prekročia k je $\Omega(n^{2/3})$

$$\Pr[P_i \geq \ell] = \sum_{k=\ell}^{n} \underbrace{(e/k)^k}_{\leq (e/\ell)^k} \leq (e/\ell)^\ell (1 + e/\ell + (e/\ell)^2 + \cdots) = \Theta((e/\ell)^\ell)$$

$$\Pr[P_i \geq \ell] \leq (e/\ell)^{\ell} \cdot [1/(1-e/\ell)]$$

- pre $\ell = \lceil (3 \ln n)/\ln \ln n \rceil$ dostaneme $\Pr \leq 1/n^2$

$$\sum_{k=1}^{n} p_k O(k^2)$$

$$p_k \sim 1/\exp(k)$$

$\Longrightarrow \sum p_k O(k^2) = O(1)$

$$\Pr[P_i \geq \ell] \leq (e/\ell)^\ell \cdot [1/(1-e/\ell)]$$

- pre $\ell = \lceil (3 \ln n)/\ln \ln n \rceil$ dostaneme $\Pr \leq 1/n^2$

$$\sum_{k=1}^{n} p_k O(k^2)$$

$$p_k \sim 1/\exp(k)$$

$\Longrightarrow \sum p_k O(k^2) = O(1)$