ALS1 – Přednáška 2

Dokonalé hašování

Hašování může mít skvělý výkon *v nejhorším případě* v případě, že je množina klíčů *statická*. Např.

- množina rezervovaných slov v programovacím jazyce,
- množina jmen souborů na CD-ROM.

Základní myšlenka: Použít dvojúrovňové schéma s univerzálním hašováním na obou úrovních.

- první úroveň – v podstatě stejné jako hašování s řetězením.
- druhá úroveň – místo seznamů použijeme *sekundární hašovací tabulky* S_j s asociovanou hašovací funkcí h_j. Vhodným výběrem můžeme zajistit, aby na sekundární úrovni nebyly žádné kolize.

Abychom zajistili, že na druhé úrovni nebudou žádné kolize, potřebujeme $m_j = n_j^2$, kde m_j je velikost sekundární tabulky ve slotu j, n_j je počet klíčů, které se tam nahašují.

To se může zdát hodně, ale uvidíme, že při vhodné volbě hašovací funkce na první úrovni bude očekávané množství použité paměti $\mathcal{O}(n)$.

Tu funkci vezmeme z $\mathcal{H}_{p,m}$. Klíče, které se hašují do slotu j jsou přehašovány do sekundární tabulky S_j velikosti m_j použitím hašovací funkce z $\mathcal{H}_{p,m_j}$

V následujím pujde o dvě věci:

- jak zajistit, že na druhé úrovni nebudou kolize.
- dokázat, že předp. množství použité paměti je $\mathcal{O}(n)$.

Theorem 1. *Pokud uložíme n klíčů do hašovací tabulky velikosti $m = n^2$ s použitím hašovací funkce náhodně vybrané z univerzální třídy hašovacích funkcí, pak pravděpodobnost, že nastane kolize je menší než* $1/2$.

Důkaz. Existuje $\binom{n}{2}$ párů klíčů, které mohou kolidovat; každý pár koliduje s pravděpodobností $1/m$, pokud je h vybraná z univerzální třídy $\mathcal{H}$ hašovacích funkcí. Nechť X je náhodná proměnná, která představuje počet kolízi. Pokud platí $m = n^2$, pak očekávané množství kolizí je

$$E[X] = \binom{n}{2} \cdot \frac{1}{n^2} = \frac{n^2 - n}{2} \cdot \frac{1}{n^2} < 1/2.$$

Použijeme Markovovu nerovnost $Pr\{X \geq t\} \leq E[X]/t$ pro $t = 1$ a je vymalováno.

Theorem 2. *Pokud uložíme n klíčů v hašovací tabulce velikosti $m = n$ použitím hašovací funkce h náhodně vybrané z univerzální třídy hašovacích funkcí, pak*

$$E\left[\sum_{j=0}^{m-1} n_j^2\right] < 2n,$$

kde n_j je počet klíčů hašovaných do slotu j.

Důkaz
Začneme následující rovností, která platí pro libovolné nezáporné celé číslo a:

$$a^2 = a + 2\binom{a}{2}.$$

Platí, že

$$E\left[\sum_{j=0}^{m-1} n_j^2\right] = E\left[\sum_{j=0}^{m-1} \left(n_j + 2\binom{n_j}{2}\right)\right] = E\left[\sum_{j=0}^{m-1} n_j\right] + 2E\left[\sum_{j=0}^{m-1} \binom{n_j}{2}\right] = E[n] + 2E\left[\sum_{j=0}^{m-1} \binom{n_j}{2}\right] = n + 2E\left[\sum_{j=0}^{m-1} \binom{n_j}{2}\right].$$

Suma $\sum_{j=0}^{m-1} \binom{n_j}{2}$ je vlastně celkový počet kolizí.
Podle vlastností univerzálního hašování, očekávaná hodnota této sumy je nejvýše

$$\binom{n}{2}\frac{1}{m} = \frac{n(n-1)}{2m} = \frac{n-1}{2},$$

protože $m = n$.
Takže

$$E\left[\sum_{j=0}^{m-1} n_j^2\right] \leq n + 2\frac{n-1}{2} = 2n - 1 < 2n.$$

Corollary 1. *Pokud uložíme n klíčů v hašovací tabulce velikosti $n = m$ použitím hašovací funkce h náhodně vybrané z univerzální třídy hašovacích funkcí a nastavíme velikost každé sekundární tabulky na $m_j = n_j^2$ pro $j = 0, 1, \ldots, m-1$, pak očekávané množství paměti potřebné k uložení všech sekundárních hašovacích tabulek v perfektním hašování je méně než $2n$.*

Důkaz. Protože $m_j = n_j^2$ pro $j = 0, 1, \ldots, m-1$, předchozí věta dává

$$E\left[\sum_{j=0}^{m-1} m_j\right] = E\left[\sum_{j=0}^{m-1} n_j^2\right] < 2n.$$

Corollary 2. *Pokud vybereme n klíčů v hašovací tabulce velikosti $m = n$ použitím hašovací funkce h náhodně vybrané z univerzální třídy hašovacích funkcí, a nastavíme velikost každé sekundární tabulky na $m_j = n_j^2$ pro $j = 0, 1, \ldots, m-1$, pak pravděpodobnost že celková paměť použitá pro sekundární hašovací tabulku překročí $4n$ je méně než 0.5.*

Důkaz. Použijeme Markovovu nerovnost, $Pr\{X \geq t\} \leq E[X]/t$, na nerovnost z předchozího důkazu s $X = \sum_0^{m-1} m_j$ a $t = 4n$:

$$Pr\left\{\sum_{j=0}^{m-1} m_j \geq 4n\right\} \leq \frac{E\left[\sum_{j=0}^{m-1} m_j\right]}{4n} < \frac{2n}{4n} = \frac{1}{2}.$$

Takže po otestování několika náhodně vybraných hašovacích funkcí najdeme takovou, která využívá rozumné množství paměti.

Binární vyhledávací stromy (binary search tree, BST)

Máme lineárně uspořádanou množinu klíčů. BST je strom, kde

- uzly jsou označeny klíčem,
- každý uzel má nejvýše dva potomky (nejvýše jednoho levého a nejvýše jednoho pravého).
- pokud má uzel s klíčem k levého potomka s klíčem l, platí $l < k$.
- pokud má uzel s klíčem k pravého potomka s klíčem p, platí $k < p$.

Theorem 3. *Očekáváný počet porovnání při vyhledávání v BST o N klíčích je asi*

$$2 \ln N \approx 1.386 \log_2 N.$$

To platí za předpokladu, že se jedná o náhodně postavený strom (randomly built BST; RBBST):

- pravděpodobnost, že N klíčů bylo vloženo v každém z $N!$ pořadí je stejná.
- ze stromu se nemazalo.

Označme:

- C_N – průměrný počet porovnání při úspěšném hledání v BST s N klíči.
- C'_N – průměrný počet porovnání při neúspěšném hledání v BST s N klíči.

$$C_N = 1 + \frac{C'_0 + C'_1 + \cdots + C'_{N-1}}{N} \tag{1}$$

Rozšířený binární strom – přidáme zvláštní uzly tam, kde měl původní strom prázdný podstrom.
délka vnější cesty (E – external path length): Součet vzdáleností kořene od všech vnějších uzlů.
délka vnitřní cesty (I – internal path length): Součet vzdáleností kořene od všech vnitřních uzlů.

Theorem 4. *Pro BST o N vnitřních uzlech platí, že*

$$E = I + 2N. \tag{2}$$

Pokud předpokládáme, že každý klíč je vyhledáván se stejnou pravděpodobností a že každý z $N + 1$ intervalů mezi klíči a vně extrémních hodnot klíčů je stejně pravděpodobný, dostáváme:

$$C_N = 1 + \frac{I}{N} \quad \text{and} \quad C'_N = \frac{E}{N+1}.$$

S použitím (2) dostáváme

$$C_N = \left(1 + \frac{1}{N}\right) C'_N - 1. \tag{3}$$

Z (1) a (3) dostáváme

$$(N+1)C'_N = 2N + C'_0 + C'_1 + \cdots + C'_{N-1}. \tag{4}$$

Zbavíme se rekurence – odečteme od (5) rovnici

$$NC'_{N-1} = 2(N-1) + C'_0 + C'_1 + \cdots + C'_{N-2}. \tag{5}$$

Dostaneme

$$(N+1)C'_N - NC'_{N-1} = 2 + C'_{N-1}.$$

Po úpravě

$$C'_N = C'_{N-1} + \frac{2}{N+1}.$$

Intermezzo: Harmonická čísla

$$H_n = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} = \sum_{k=1}^{n} \frac{1}{k} = \ln n + \mathcal{O}(1)$$

poslední krok se dá ukázat aproximací integrály.

Když se dá sumace vyjádřit jako $\sum_{k=m}^{n} f(k)$, kde $f(k)$ je monotónně klesající funkce, můžeme ji aproximovat:

$$\int_m^{n+1} f(x)dx \leq \sum_{k=m}^{n} f(k) \leq \int_{m-1}^{n} f(x)dx$$

Dolní hranice:

$$\sum_{k=1}^{n} \frac{1}{k} \geq \int_1^{n+1} \frac{1}{x}dx = \ln(n+1)$$

Horní hranice:

$$\sum_{k=2}^{n} \frac{1}{k} \leq \int_1^{n} \frac{1}{x}dx = \ln n \quad \text{a tedy} \quad \sum_{k=1}^{n} \frac{1}{k} \leq \ln n + 1$$

Harmonická čísla rekurzivně:

$$H_n = \begin{cases} 0 & \text{pro } n = 0, \\ H_{n-1} + \frac{1}{n} & \text{jinak.} \end{cases}$$

Takže $C'_N = 2H_{N+1} - 2$.

konec intermezza

Po dosazení do (3) a zjednodušení dostáváme

$$C_N = 2\left(1 + \frac{1}{N}\right) H_N - 3 \approx 2 \ln N.$$

Theorem 5. *Očekávaná výška RBBST o N uzlech je $\mathcal{O}(log(N))$.*

Nadefinujeme si tyto tři náhodné proměnné:

- výška RBBST o N prvcích X_N,
- exponenciální výška $Y_N = 2^{X_N}$
- klíč v kořeni RBBST R_N.

Hodnota R_N je se stejnou pravděpodobností kterýkoli z prvků $\{1, \ldots, N\}$. Pokud $R_N = i$, pak

- levý podstrom je RBBST o $i - 1$ prvcích,
- pravý podstrom je RBBST o $N - i$ prvcích,
- $Y_N = 2 \cdot \max(Y_{i-1}, Y_{N-i})$

Jako krajní případy Y_N máme $Y_1 = 1, Y_0 = 0$.
Dále definujeme náhodné proměnné $Z_{N,1}, Z_{N,2}, \ldots, Z_{N,N}$, kde $Z_{N,i} = I\{R_N = i\}$.

Máme $\Pr\{Rn = i\} = \frac{1}{n}$ pro $i = 1, 2, \ldots, n$, a tedy

$$E[Z_{N,i}] = \frac{1}{n} \text{ pro } i = 1, 2, \ldots, n.$$

Protože právě jedna hodnota $Z_{N,i} = 1$ a všechny ostatní jsou 0, máme taky

$$Y_N = \sum_{i=1}^{N} Z_{N,i}(2 \cdot \max(Y_{i-1}, Y_{N-i})).$$

Ukážeme, že $E[Y_N]$ je polynomická v N, z toho pak vyplyne, že $E[X_N] = \mathcal{O}(\log N)$.
Náhodná proměnná $Z_{N,i}$ je nezávislá na Y_{i-1} a Y_{N-i}.
Když vybereme $R_N = i$, levý podstrom (s exp. výškou Y_{i-1}) je náhodně postaven z $i-1$ klíčů, které jsou menší než i. Tento podstrom je jako jakýkoli jiný podstrom postavený z $i-1$ prvků:

- pouze počet prvků v něm je závislý na volbě R_N,
- jeho struktura není nijak závislá na volbě R_N. stejně tak pro pravý podstrom.

$$\begin{aligned}
\mathrm{E}[Y_N] &= \mathrm{E}\left[\sum_{i=1}^{N} Z_{N,i}(2 \cdot \max(Y_{i-1}, Y_{N-i}))\right] = \sum_{i=1}^{N} \mathrm{E}\left[Z_{N,i}(2 \cdot \max(Y_{i-1}, Y_{N-i}))\right] \\
&= \sum_{i=1}^{N} \mathrm{E}\left[Z_{N,i}\right] \cdot E\left[(2 \cdot \max(Y_{i-1}, Y_{N-i}))\right] = \sum_{i=1}^{N} \frac{1}{N} \cdot \mathrm{E}\left[(2 \cdot \max(Y_{i-1}, Y_{N-i}))\right] \\
&= \frac{2}{N} \sum_{i=1}^{N} \mathrm{E}\left[\max(Y_{i-1}, Y_{N-i})\right] \leq \frac{2}{N} \sum_{i=1}^{N} \mathrm{E}\left[Y_{i-1}\right] + \mathrm{E}\left[Y_{N-i}\right]
\end{aligned}$$

V posledním výrazu se každý term $\mathrm{E}[Y_0], \mathrm{E}[Y_1], \ldots, \mathrm{E}[Y_{N-1}]$ vyskytuje dvakrát – můžeme ho zjednodušit na

$$E[Y_N] \leq \frac{4}{N} \sum_{i=0}^{N-1} \mathrm{E}[Y_i].$$

Ukážeme, že pro všechna kladná N je to ekvivalentní s

$$E[Y_N] \leq \frac{1}{4}\binom{N+3}{3}.$$

Pro $N = 1$ toto platí

$$1 = Y_1 = \mathrm{E}[Y_1] \leq \frac{1}{4}\binom{1+3}{3} = 1$$

Dále

$$\begin{aligned}
\mathrm{E}[Y_N] &\leq \frac{4}{N} \sum_{i=0}^{N-1} \mathrm{E}[Y_i] = \frac{4}{N} \sum_{i=0}^{N-1} \frac{1}{4}\binom{i+3}{3} = \frac{1}{N} \sum_{i=0}^{N-1} \binom{i+3}{3} \\
&= \frac{1}{N}\binom{N+3}{4} = \frac{1}{N} \cdot \frac{(N+3)!}{4!(N-1)!} = \frac{1}{4} \cdot \frac{(N+3)!}{3! \cdot N!} = \frac{1}{4} \cdot \binom{N+3}{3}
\end{aligned}$$

Platí, že $2^{\mathrm{E}[X_N]} \leq \mathrm{E}[2^{X_N}] = \mathrm{E}[Y_N]$.

Z toho dostáváme, že

$$2^{\mathrm{E}[X_N]} \leq \frac{1}{4}\binom{N+3}{3} = \frac{(N+3)(N+2)(N+1)}{24}.$$