

# Хеширане без колизии на ограничено множество от естествени числа (част 1)

30.10.2020 г.

(Perfect Hash)

Нека  $u \in \mathbb{N}$ ,  $S \subseteq \{0, 1, \dots, u - 1\}$ . Целта е да индексирате  $S$  с памет  $O(|S|)$ , така че да отговаряме на заявки *MEMBER* за  $O(1)$ , т.е.  $x \in S$ ? за константно време.

Интересен е случаят  $|S| \ll u$ .

Хеш функция е  $h : \mathbb{N} \rightarrow \{0, 1, \dots, m - 1\}$  с памет  $m$ , която можем да пресмятаме за време  $O(1)$ .

Казваме, че  $(x, y) \in \mathbb{N} \times \mathbb{N}$  е в колизия за  $h$ , ако  $x \neq y$  и  $h(x) = h(y)$ .

Когато е дадено множеството  $S$  ще търсим такава функция  $h$  и такова число  $n$ , така че  $\forall x, y \in S : (x, y)$  да не е в колизия за  $h$ .

Тогава " $x \in S$ ?":

1.  $h(x)$
2. проверяваме дали в масив с памет  $m$  на позиция  $h(x)$  е записано  $x$ .

„Лесни/прости“ хеш функции

$$h : \mathbb{N} \rightarrow \{0, 1, \dots, m - 1\}$$

$$1.1. h(x) = x \pmod{m}$$

$$1.2. h(x) = cx \pmod{m}$$

Ако  $m = 30$ , а  $c = 5$ , тогава  $h(x)$  ще използва само клетките 0,5,10,15,20,25, т.е. само 6 клетки общо, въпреки че в нашето множество ще има 30 елемента. Т.е. когато  $\gcd(c, m) \neq 1$  хеша от 1.2. ще е по-лош от този в 1.1. Избора на константата  $c$  в 1.2. трябва да е по-внимателен за да имаме добра ефективност ( $\gcd(c, m) = 1$ ).

Нека:  $p > u$  ( $S \subseteq \{0, 1, \dots, u - 1\}$ )

$$h_{c,m}(x) = (cx \pmod{p}) \pmod{m}, \text{ където } c \in \{1, 2, \dots, p - 1\}, \text{ а } m \in \mathbb{N}.$$

Ще изучаваме функциите  $h_{c,m}$  и ще покажем следните две неща:

1. съществува константа  $c$ :  $1 \leq c < p$ , за която  $h_{c,n^2}$  не съдържа колизия от  $S$
2. съществува константа  $c$ , за която  $h_{c,n}$  съдържа  $O(n)$  Колизии от  $S$

$$Col_S(c, n) = \{(x, y) \in S \times S \mid x \neq y \text{ и } h_{c,m}(x) = h_{c,m}(y)\}$$

$$Hit_S(c, n, i) = \{x \in S \mid h_{c,m}(x) = i\}$$

Наблюдения:

$$C = Col_S(c, m) : H_i = Hit_S(c, m, i)$$

$$1. |C| = \sum_{i=0}^{m-1} |H_i| (|H_i| - 1)$$

$$2. \sum_{i=0}^{m-1} |H_i| = |S| (= n)$$

Док-во:

$$1. (x, y) \in C \Leftrightarrow x \neq y \text{ и } h_{c,m}(x) = h_{c,m}(y) \Leftrightarrow \text{има } 0 \leq i \leq m : h_{c,m}(x) = i = h_{c,m}(y), x \neq y$$

$$C = \bigcup_{i=0}^{m-1} \{(x, y) | x, y \in H_i \text{ и } x \neq y\}$$

$$\Rightarrow |C| = \sum_{i=0}^{m-1} |\{(x, y) | x, y \in H_i, x \neq y\}| = \sum_{i=0}^{m-1} |H_i| (|H_i| - 1)$$

2. Тъй като  $h_{c,m}$  изобразява всяко число  $x$  в елемента от множеството  $\{0, 1, \dots, m-1\}$  то

$$S = \bigcup_{i=0}^{m-1} H_i, H_i \cap H_j = \emptyset \text{ за } i \neq j \Rightarrow |S| = \sum_{i=0}^{m-1} |H_i|.$$

Следствие:  $|C| + n = \sum_{i=0}^{m-1} |H_i|^2.$

# Хеширане без колизии на ограничено множество от естествени числа (част 2)

06.11.2020 г.

$$n \in \mathbb{N}$$

$$\text{Дадено: } S = \{0, 1, \dots, u-1\}$$

$$\text{Вход: } x \in \mathbb{N}$$

Изход: 1 ако  $x \in S$ , 0 в противен случай.

Търсена сложност  $O(1)$  за заявка и  $O(u)$  за построяване на индекс.

$$h_{c,m} : \mathbb{N} \rightarrow \{0, 1, \dots, m-1\}$$

$p > u$ ,  $p$ -просто число

$$h_{c,m}(x) = (cx \pmod{p}) \pmod{m}$$

разбърква елементите на  $S$

$$H[0 \dots m-1]$$

$$H[i] = s, \text{ ако } s \in S, h_{c,m}(s) = i$$

$$Col_S(c, m) = \{(x, y) \subseteq S^2 \mid x \neq y, h_{c,m}(x) = h_{c,m}(y)\}$$

$$Hit_S(c, m, i) = \{x \in S \mid h_{c,m}(x) = i\}$$

$$C = Col_S(c, m)$$

$$|C| + n = \sum_{i=0}^{n-1} |H_i|^2, \text{ където } H_i = Hit_S(c, m, i)$$

Искаме да минимизираме  $C$  при  $u$  и  $S$  постоянни. За да ограничим минимума от броя на

колизиите  $\min_{1 \leq c < p-1} |Col_S(c, m)|$  ще оценим  $\sum_{c=1}^{p-1} |Col_S(c, m)|$  и

$$\min_{1 \leq c < p-1} |Col_S(c, m)| \leq \frac{1}{p-1} \sum_{c=1}^{p-1} |Col_S(c, m)|$$

$$\sum_{c=1}^{p-1} |Col_S(c, m)| = \sum_{c=1}^{p-1} \sum_{\substack{(x, y) \in S^2 \\ x \neq y \\ h_{c,m}(x) = h_{c,m}(y)}} 1 = \sum_{\substack{(x, y) \in S^2 \\ x \neq y}} \sum_{\substack{c=1 \\ h_{c,m}(x) = h_{c,m}(y)}}^{p-1} 1.$$

Връщаме се на израза  $h_{c,m} = (cx \pmod{p}) \pmod{m}$ , но вече с фиксирани  $c$  и  $m$ .

Нека  $x, y \in S$ ,  $x \neq y$  са фиксирани,  $m$  и  $p$  също са фиксирани.

Питаме се кога  $h_{c,m}(x) = h_{c,m}(y)$ . Т.е. колко решения ще има последното уравнение в интервала между 1 и  $p$ .

Знаем, че  $0 \leq x, y < u < p$  : ?  $c \subseteq \{1, \dots, p-1\}$  за които  $h_{c,m}(x) = h_{c,m}(y)$

$$\left\lceil \frac{p-1}{m} \right\rceil$$

Т.е.

$$(cx \pmod p)(\pmod m) = (cy \pmod p)(\pmod m) \Leftrightarrow ((c(x-y)) \pmod p)(\pmod m) = 0$$

$c(x-y) \pmod p \in \{0, 1, \dots, p-1\}$ . От тези остатъци ни интересуват онези, които се делят на  $m$ :  $0, m, 2m, \dots, \left\lfloor \frac{p-1}{m} \right\rfloor m$ .

Т.е. се питаме: Колко са онези  $c$  между 1 и  $p$ , ( $0 \leq c < p$ ), за които

$$c(x-y) \pmod p \in \{0, m, \dots, \left\lfloor \frac{p-1}{m} \right\rfloor m\}$$

Тъй като  $x \neq y$ ,  $0 \leq x, y < p$ , то  $\gcd(x-y, p) = 1 \Rightarrow$  за всяко  $k \in \{0, p-1\}$  има единствено  $c_k$ :  $c_k(x-y) \equiv k \pmod p$

Тъй като  $c_0 = 0$ , то онези  $c$ :  $1 \leq c \leq p-1$  и за които

$$c(x-y) \pmod p \in \{0, m, 2m, \dots, \left\lfloor \frac{p-1}{m} \right\rfloor m\} \text{ са } c_m, c_{2m}, \dots, c_{\left\lfloor \frac{p-1}{m} \right\rfloor m}$$

$$\Rightarrow \text{търсеният брой решения за } c \text{ е } \left\lfloor \frac{p-1}{m} \right\rfloor$$

Следователно:

$$\sum_{c=1}^{p-1} 1 = \left\lfloor \frac{p-1}{m} \right\rfloor.$$

$$h_{c,m}(x) = h_{c,m}(y)$$

$$\sum_{\substack{(x,y) \in S^2 \\ x \neq y}} 1 = \sum_{\substack{c=1 \\ h_{c,m}(x) = h_{c,m}(y)}} 1 = \sum_{\substack{(x,y) \in S^2 \\ x \neq y}} \left\lfloor \frac{p-1}{m} \right\rfloor = n(n-1) \left\lfloor \frac{p-1}{m} \right\rfloor \left( \leq n(n-1) \left( \frac{p-1}{m} \right) \right)$$

$$\sum_{c=1}^{n-1} |Col_S(c, m)| \geq (p-1) \min_{1 \leq c \leq p-1} |Col_S(c, m)|$$

**Твърдение:** За всяко  $m \geq 1$ ,  $p > m$  има  $1 \leq c \leq p-1$ , за което броя на колизиите  $|Col_S(c, m)| \leq \frac{n(n-1)}{m}$ .

Доказателство:  $\left\lfloor \frac{p-1}{m} \right\rfloor \leq \frac{p-1}{m}$  и заместваме в изрази

$$(p-1) \min_{1 \leq c \leq p-1} |Col_S(c, m)| \leq n(n-1) \left\lfloor \frac{p-1}{m} \right\rfloor$$

**Следствие 1.** За  $m = n^2$  или  $m = 2n^2$  има такова  $c$ , за което  $|Col_S(c, m)| = 0$

$$\text{Доказателство: } |Col_S(c, m)| \stackrel{m=n^2}{=} |Col_S(c, n^2)| \leq \frac{n(n-1)}{n^2} < 1$$

**Следствие 2.** За  $m = n$  или  $m = 2n$  има такова  $c$ , за което

$$1 \leq c \leq p-1 : |Col_S(c, m)| \leq n$$

$$\text{Доказателство: } \min_{1 \leq c \leq p-1} |Col_S(c, n)| \leq \frac{n(n-1)}{n} = n-1$$

Метод:

1. Намерете константа  $c$ :  $|Col_S(c, 2n)| \leq n$
2. Нека  $L[0 \dots 2n - 1]$  е масив от списъци, като  $L[i] = \{x \in S \mid h_{c, 2n}(x) = i\}$
3. За всяко  $i = 0$  до  $2n - 1$  намираме константа  $c_i$ :  $Col_{L[i]}(c_i, 2|L[i]|) = 0$
4. За всяко  $i = 0$  до  $2n - 1$   
 $H_i[0 \dots 2|L[i]|^2 - 1] : H_i[j] = x \Leftrightarrow x \in L[i] \ \& \ h_{c_i, 2|L[i]|^2}(x) = j$   
 $H[j] = -1$  ако няма  $x$  с горното свойство.

$$|C| + n = \sum_{i=0}^{m-1} |H_i|^2$$

$$2n \stackrel{(1)}{\geq} |Col_S(c, 2n)| + n = \sum_{i=0}^{2n-1} |L[i]|^2$$

За пълнота ще разгледаме как се отговаря на заявки.  
Заявки:

```

procedure Query( $x$ ) {
     $i \leftarrow h_{c, 2n}(x)$ 
     $j \leftarrow h_{c_i, 2|L[i]|^2}(x)$ 
    if  $H_i[j] = x$  then return yes
    else return no
}

```

### Вероятностен анализ

Ще разгледаме само дискретни вероятности в този анализ, което означава, че имаме изброими множества. Дискретна вероятност:  $(X, p)$ , където  $X = \{x_i \mid i \in \mathbb{N}\}$  е произволно множество, а  $p : X \rightarrow [0, 1]$  е функция със свойството  $\sum_{i=0}^{\infty} p(x_i) = 1$ .

$x_i$  се наричат елементарни събития.

Събитие  $A$  в всяко подмножество във вероятностното пространство  $(X, p)$ .  $A \subseteq X$ .

Вероятност на събитието  $A$  е  $p(A) = \sum_{x \in A} p(x)$

Случайна величина  $V : X \rightarrow \mathbb{R}$

Пример  $(\{1, 2, \dots, p-1\}, \text{Pr})$   $\text{Pr}(c) = \frac{1}{p-1}$  за  $1 \leq c \leq p-1$

$Col_S(m) : \{0, 1, \dots, p-1\} \rightarrow \mathbb{R}$

$$\underbrace{[Col_S(m)](c)}_{V(c)} = |Col_S(c, m)|$$

Очакване на случайна величина във вероятностно пространство  $(X, p)$  ( $V : X \rightarrow \mathbb{R}$ ) е  $\mathbb{E}V = \sum_{x \in X} p(x)V(x)$ .

Обратно към примера:

$$(\{1, 2, \dots, p-1\}, \text{Pr}), \text{Pr}(c) = \frac{1}{p-1}, V(c) = |Col_S(c, m)|.$$

$$\mathbb{E}V = \sum_{c=1}^{p-1} \text{Pr}(c)V(c) = \sum_{c=1}^{p-1} \frac{1}{p-1} |Col_S(c, m)| = \frac{1}{p-1} \sum_{c=1}^{p-1} |Col_S(c, m)| = \frac{1}{p-1} n(n-1) \left[ \frac{p-1}{m} \right]$$

В случая когато  $m = 2n^2$  получаваме, че  $\mathbb{E}V \leq \frac{n(n-1)}{2n^2} < \frac{1}{2}$ , а когато  $m = 2n$ ,

получаваме че  $\mathbb{E}V \leq \frac{n-1}{2}$ .

Означение:  $V : X \rightarrow \mathbb{R}, a \in \mathbb{R}$

$[V = a]$  е събитието  $\{x \in X \mid V(x) = a\}$

$[V \leq a] : \{x \in X \mid V(x) \leq a\}$

### **Неравенство на Марков**

$(X, p)$  - дискретно вероятностно пространство.

$V : X \rightarrow \mathbb{R}_0^+ = \{r \in \mathbb{R} \mid r \geq 0\}$ . Тогава за всяко  $a > 0 \in \mathbb{R}$  е изпълнено, че

$$p([V > a]) \leq \frac{\mathbb{E}V}{a}.$$

Доказателство:

$$\begin{aligned} \mathbb{E}V &= \sum_{x \in X} p(x)V(x) = \sum_{\substack{x \in X \\ V(x) \leq a}} p(x)V(x) + \sum_{\substack{x \in X \\ V(x) > a}} p(x)V(x) \geq 0 + \sum_{\substack{x \in X \\ V(x) > a}} p(x)a \geq \\ &\geq 0 + \sum_{\substack{V(x) > a \\ x \in X}} p(x) \cdot a = a \sum_{x \in \{y \in X \mid V(y) > a\}} p(x) = ap(\{y \mid V(y) > a\}) = ap([V > a]). \end{aligned}$$

$$\text{Тъй като } a > 0 \Rightarrow \underline{\underline{p([V > a]) \leq \frac{\mathbb{E}V}{a}}}.$$

$$\text{Пример: } (\{1, 2, \dots, p-1\}, \text{Pr}) \quad \text{Pr}(c) = \frac{1}{p-1}$$

$$V : \{1, 2, \dots, p-1\} \rightarrow \mathbb{R}_0^+$$

$$V(c) = Col_S(c, 2n), \mathbb{E}V \leq \frac{n-1}{2}. \text{ Тогава от неравенството на Марков:}$$

$$\text{Pr}([V] > n) \leq \frac{n-1}{2n} < \frac{1}{2}.$$

$$\left. \begin{array}{l} c \leftarrow \text{random uniform in } \{1, 2, \dots, p-1\} \\ \text{while } \text{Col}_S(c, 2n) > n \text{ do} \\ c \leftarrow \text{random uniform} \\ \quad (\text{independent}) \text{ in } \{1, 2, \dots, p-1\} \end{array} \right\} \begin{array}{l} \text{събитие в някакво} \\ \text{вероятностно пространство} \end{array}$$

[Да не се бърка с" несъвместими събития  $A \cap B = \emptyset$  или  $p(A \cap B) = 0$ ]

Произведение на вероятностни пространства:

$(X_1, p_1)$  и  $(X_2, p_2)$ . Тогава дефинираме  $X = X_1 \times X_2$

$p(x_1, x_2) = p_1(x_1) \cdot p_2(x_2)$

$(X, p)$  - декартово произведение на вероятностните пространства  $(X_1, p_1)$  и  $(X_2, p_2)$ .

Сега ако  $A_1 \subseteq X_1, A_2 \subseteq X_2$

Въпросът  $p(A_1 \cap A_2)$  е безсмислен в същия случай.

Но

$A_1$  в  $X = X_1 \times X_2$  ще изглежда като  $A_1 \times X_2$

$A_2$  в  $X = X_1 \times X_2$  ще изглежда като  $X_1 \times A_2$

Сега в  $X = X_1 \times X_2$ ,  $A_1$  и  $A_2$  да настъпват едновременно:

$A_1 \times X_2 \cap X_1 \times A_2 = A_1 \times A_2$

$$\begin{aligned} p(A_1 \times A_2) &= \sum_{(a_1, a_2) \in A_1 \times A_2} p(a_1, a_2) = \sum_{\substack{a_1 \in A_1 \\ a_2 \in A_2}} [p_1(a_1)p_2(a_2)] = \left( \sum_{a_1 \in A_1} p_1(a_1) \right) \left( \sum_{a_2 \in A_2} p_2(a_2) \right) \\ &= (p_1(A_1)p_2(A_2)) \end{aligned}$$

В частност ако  $A_2 = X_2$ :

$$p(A_1 \times X_2) = \sum_{a_1 \in A_1} p_1(a_1) \underbrace{\sum_{a_2 \in X_2} p_2(a_1)}_1 = p_1(A_1) \text{ и аналогично } p(X_1 \times A_2) = p_2(A_2).$$

Следователно събитията  $A_1$  и  $A_2$  могат да бъдат разглеждани като независими в декартовото произведение  $X_1 \times X_2$ .

Обратно към нашия анализ за перфектния хеш.

Нека  $(X_i, \text{Pr}_i)$  са вероятностните пространства:

$$(X_i, \text{Pr}_i) = \underbrace{(\{1, 2, \dots, p-1\}, \text{Pr}) \times (\{1, \dots, p-1\}, \text{Pr}) \times \dots \times (\{1, 2, \dots, p-1\}, \text{Pr})}_i$$

Нека  $A_i = \{(c_1, c_2, \dots, c_i) \in X_i \mid |\text{Col}_S(c_j, 2n)| > n \text{ за } j \leq i\}$

От горните разсъждения имаме, че  $\Pr_i(A_i) = \prod_{j=1}^i \Pr \left( [|Col_S(c_j, 2n)|] \right) < \left( \frac{1}{2} \right)^i$

Сега ако разгледаме случайната величина

$$Time : \bigcup X_i \rightarrow \mathbb{N}, \quad Time((c_1, \dots, c_i)) = \begin{cases} i+1, & \text{if } j : |Col_S(c_j, 2n)| > n \\ \text{else } \min_{j \leq i} \{j : |Col_S(c_j, 2n)| \leq n\} \end{cases}$$

Тогава вероятността  $\Pr_i(\underbrace{Time > j}_{A_j \times \{1, 2, \dots, p-1\}}) \leq \frac{1}{2^j}$ .

Следователно очакването

$$\begin{aligned} \mathbb{E}(X_i, \Pr_i) Time &\stackrel{\text{def.}}{=} \sum_{j=1}^{i+1} \Pr_i(Time = j) \cdot j = \sum_{j=1}^{i+1} \sum_{s=1}^j \Pr_i(Time = j) = \\ &= \sum_{s=1}^{i+1} \sum_{j=s}^{i+1} \Pr_i(Time = j) = \sum_{s=1}^{i+1} \Pr_i(Time \geq s-1) \leq \sum_{s=0}^i \frac{1}{2^s} = 2 - \frac{1}{2^i} \leq 2. \end{aligned}$$

$$\underbrace{\Pr_i(Time \geq S)}_{\Pr_i(Time > S-1)}$$

Т.е. в очакване ще направим най-много две итерации в цикъла от описания по-горе алгоритъм, за да намерим константа  $c$ , която да има свойството, че броят на колизиите ще бъде не по-голям от  $n$ . По абсолютно същия начин в останалите няколко фази, където ще трябва да разхвърляме елементите, които евентуално са попаднали в колизия при този избор на константата  $c$ , може да докажем, че в очакване броя на итерации, които са необходими за да намерим подобна константа  $c$  е ограничен отгоре от 2. Съществено е независимия случаен избор на променлива  $c$  измежду числата от 1 до  $n$ .