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

30.10.2020 г.

(Perfect Hash)

Нека $u \in \mathbb{N}, \quad S \subseteq \{0,1,...,u-1\}$. Целта е да индексираме 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 $h : \mathbb{N} \to \{0,1,...,m-1\}$ 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

$h : \mathbb{N} = \{0,1,...,m-1\}$

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-лощ от този в 1.1. Избора на константата c в 1.2. трябва да е по-внимателен за да имаме добра ефективност $\big(\gcd(c, m) = 1\big)$.

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

$h_{c,m}(x) = \big(cx \pmod p\big) \pmod m$, където $c \in \{1,2,...,p-1\}$, 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

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

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

Наблюдения:

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

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

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

Док-во:

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

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

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

2. Тъй като $h_{c,m}$ изобразява всяко число x в елемента от множеството $\{0,1,...,m-1\}$ то $S = \bigcup_{i=0}^{m-1} H_i,\ H_i \cap H_j = \emptyset$ 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.

$n \in \mathbb{N}$
<u>Дадено</u>: $S = \{0,1,...,u-1\}$
<u>Вход</u>: $x \in \mathbb{N}$
<u>Изход</u>: 1 ако $x \in S$, 0 в противен случай.

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

$h_{c,m} : \mathbb{N} \to \{0,1,...,m-1\}$
$p > u$, p-просто число
$h_{c,m}(x) = \big(cx(\mod p)\big)(\mod m)$
разбърква елементите на S

$H[0...m-1]$
$H[i] = s$, ако $s \in S$, $h_{c,m}(s) = i$

$Col_S(c,m) = \{(x,y) \subseteq S^2 \,|\, x \neq y, h_{c,m}(x) = h_{c,m}(y)\}$
$Hit_S(c,m,i) = \{x \in S \,|\, h_{c,m} = i\}$
$C = Col_S(c,m)$
$|C| + n = \sum_{i=0}^{n-1} |H_i|^2$, 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

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

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

Връщаме се на израза $h_{c,m} = \big(cx(\mod p)\big)(\mod m)$, 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.

$\left[\frac{p-1}{m}\right]$

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

Т.е.

$$\big(cx(\mod p)\big)(\mod m) = \big(cy(\mod p)\big)(\mod m) \Leftrightarrow \Bigg(\big(c(x-y)\big)(\mod p)\Bigg)(\mod m) = 0$$

$c(x-y)(\mod p) \in \{0,1,...,p-1\}$. От тези остатъци ни интересуват онези, които се делят на m: $0,m,2m,\ldots,\left[\dfrac{p-1}{m}\right]$.

Т.е. се питаме: Колко са онези c между 1 и p, $(0 \le c < p)$, за които $c(x-y)(\mod p) \in \{0,m,\ldots,\left[\dfrac{p-1}{m}\right]m\}$

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

Тъй като $c_0 = 0$, то онези $c: \ 1 \le c \le p-1$ и за които $c(x-y)(\mod p) \in \{0,m,2m,\ldots m\left[\dfrac{p-1}{m}\right]m\}$ са $c_m, c_{2m}, \ldots, c_{\left[\frac{p-1}{m}\right]m}$

$\Rightarrow$ търсеният брой решения за c е $\left[\dfrac{p-1}{m}\right]$

Следователно:

$$\sum_{\substack{c=1 \\ h_{c,m}(x)=h_{c,m}(y)}}^{p-1} 1 = \left[\frac{p-1}{m}\right].$$

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

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

<u>Твърдение</u>: За всяко $m \ge 1$, $p > u$ има $1 \le c \le p-1$, за което броя на колизиите $|Col_S(c,m)| \le \dfrac{n(n-1)}{m}$.

Доказателвство: $\left[\dfrac{p-1}{m}\right] \le \dfrac{p-1}{m}$ и заместваме в израза

$$(p-1) \min_{1\le c\le p-1} |Col_S(c,m)| \le n(n-1)\left[\frac{p-1}{m}\right]$$

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

Доказателство: $|Col_S(c,m)| \overset{m=n^2}{=} |Col_S(c,n^2)| \le \dfrac{n(n-1)}{n^2} < 1$

<u>Следствие 2</u>. За $m = n$ или $m = 2n$ има такова c, за което $1 \le c \le p-1: \ |Col_S(c,m)| \le n$

Доказателство: $\min\limits_{1\le c\le p-1} |Col_S(c,n)| \le \dfrac{n(n-1)}{n} = n-1$

Метод:

1. Намерете константа c: $|Col_S(c,2n)| \leq n$
2. Нека $L[0...2n-1]$ е масив от списъци, като $L[i] = \{x \in S \,|\, h_{c,2n}(x) = i\}$
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 $2n-1$

   $H_i[0...2|L[i]|^2 - 1]$: $H_i[j] = x \Leftrightarrow x \in L[i]$ & $h_{c_i,2|L[i]|^2}(x) = j$

   $H[j] = -1$ ако няма x с горното свойство.

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

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

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 $X = \{x_i \,|\, i \in \mathbb{N}\}$ е произволно множество, а $p : X \to [0,1]$ 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 $V : X \to \mathbb{R}$

Пример $\left(\{1,2,...,p-1\}, \Pr\}\right)$ $\Pr(c) = \dfrac{1}{p-1}$ за $1 \leq c \leq p-1$

$$\underbrace{Col_S(m)}_{V} : \{0,1,...,p-1\} \to \mathbb{R}$$

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

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

Обратно към примера:

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

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

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

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

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

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

Неравенство на Марков

(X, p) - дискретно вероятностно пространство.

$V : X \to \mathbb{R}_0^+ = \{r \in \mathbb{R} \,|\, r \geq 0\}$. Тогава за всяко $a > 0 \in \mathbb{R}$ е изпълнено, че $p([V > a]) \leq \dfrac{\mathbb{E}V}{a}$.

Доказателство:

$$\mathbb{E}V = \sum_{x \in X} p(x)V(x) = \sum_{\substack{x \in X \\ V(x) \leq a}} \underset{\geq 0}{p(x)} \underset{\geq 0}{V(x)} + \sum_{\substack{x \in X \\ V(x) > a}} p(x) \underset{\geq a}{V(x)} \geq 0 + \underbrace{\sum_{\substack{x \in X \\ V(x) > a}}}_{(*)} p(x) a \geq$$

$$\geq 0 + \sum_{\substack{V(x) > a \\ x \in X}} p(x) \,.\, a = a \sum_{x \in \{y \in X | V(y) > a\}} p(x) = ap\big(\{y \,|\, V(y) > a\}\big) = ap\big([V > a]\big).$$

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

Пример: $\big(\{1,2,...,p-1\}, \Pr\big)$ $\Pr(c) = \dfrac{1}{p-1}$

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

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

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

$$\left.\begin{array}{l} c \leftarrow random\ uniform\ in\ \{1,2,...,p-1\} \\ while\ Col_S(c,2n) > n\ do \\ c \leftarrow random\ uniform \\ \qquad (independent)\ in\ \{1,2,...,p-1\} \end{array}\right\}$$ събитие в някякво вероятностно пространство

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

Произведение на вероятностни пространства:

(X_1, p_1) и (X_2, p_2). Тогава дефинираме $X = X_1 \times X_2$
$p(x_1, x_2) = p_1(x_1) \,.\, p_2(x_2)$
(X, p) - декартово производение 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:
$A_1 \times X_2 \cap X_1 \times A_2 = A_1 \times A_2$

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

В частност ако $A_2 = X_2$:

$$p(A_1 \times X_2) = \sum_{a_1\in A_1} p_1(a_1) \underbrace{\sum_{a_2\in X_2} p_2(a_1)}_{1} = p_1(A_1)$$ и аналогично $p(X_1 \times A_2) = p_2(A_2)$.

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

Обратно към нашия анализ за перфектния хеш.

Нека $(X_i, \Pr_i)$ са вероятностните пространства:

$$(X_i, \Pr_i) = \underbrace{\big(\{1,2,...,p-1\}, \Pr\big) \times \big(\{1,...,p-1\}, \Pr\big) \times \ldots \times \big(\{1,2,...,p-1\}, \Pr\big)}_{i}$$

Нека $A_i = \{(c_1, c_2, \ldots, c_i) \in X_i \,\big|\, |Col_S(c_j,2n)| > n$ за $j \le i\}$

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

Сега ако разгледаме случайната величина

$$Time : \bigcup X_i \to \mathbb{N}, \quad Time\big((c_1, \ldots, c_i)\big) = \begin{cases} i+1, \; if \;\; j : |Col_S(c_j, 2n)| > n \\ else \min_{j \le i}\{j \,|\, Col_S(c_j, 2n)| \le n\} \end{cases}$$

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

Следователно очакването

$$\mathbb{E}(X_i, \;\Pr_i)Time \overset{def.}{=} \sum_{j=1}^{i+1} \Pr_i(Time = j)\,.\,j = \sum_{j=1}^{i+1}\sum_{s=1}^{j} \Pr_i(Time = j) =$$

$$= \sum_{s=1}^{i+1} \underbrace{\sum_{j=s}^{i+1} \Pr_i(Time = j)}_{\underbrace{\Pr_i(Time \ge S)}_{\Pr_i(Time > S-1)}} = \sum_{s=1}^{i+1} \Pr_i(Time \ge s - 1) \le \sum_{s=0}^{i} \frac{1}{2^s} = 2 - \frac{1}{2^i} \le \mathbf{2}.$$

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 константана 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.