Хеширование

Семейство хеш-функций —
$\mathcal{H} = \{h_i \mid h_i : A \to B\}$

Пример $\mathcal{H} = B^A$ — все ф-ии из $A \to B$

$x \neq y$ $Pr\{h(x) = h(y)\} = \frac{1}{|B|}$ $\quad Pr\{h(x_1) = z_1 \wedge h(x_2) = z_2 \wedge \dots \wedge h(x_n) = z_n\} = \frac{1}{|B|^n}$ $\quad x_i \neq x_j \ \forall i \neq j$

В начале программы $h \leftarrow U(\mathcal{H})$. Дальше пользуемся h.

Пример $\mathcal{H} = \{h_{m,x}\}_{x \in \mathbb{Z}_m} \ m \in \mathbb{P}$

$$h_{m,x}(s) = \left(\sum s_i \cdot x^{|s|-i-1}\right) \bmod m$$

Def. Универсальное сем. х-ф. — $\mathcal{H}$: $\forall \begin{smallmatrix} x_1 \in A \\ x_2 \in A \end{smallmatrix} \ x_1 \neq x_2$ $\quad \underset{h \leftarrow U(\mathcal{H})}{Pr}\{h(x_1) = h(x_2)\} \leq \frac{1}{|B|}$.

Th. Все операции в хеш-таблице с закр. адресацией и универс. хеш. имеют мат.ож. времени работы $O(\underbrace{\tfrac{n}{m}}_{\alpha} + 1)$,
$|B| = m$ — размер х-таблицы
n = # добавл. ключей

Д-во $k_1, k_2, \dots, k_n$ — различные добавл. ключи
Новый ключ: k

$$E\{\text{размера списка под номером } h(k)\} = E\left[\sum_{i=1}^{n} \mathbb{1}_{\{h(k_i) = h(k)\}}\right] =$$

$$= \sum_{i=1}^{n} E\{\mathbb{1}_{\{h(k_i) = h(k)\}}\} = \sum_{i=1}^{n} \left(Pr\{h(k_i) = h(k)\} \cdot 1\right)$$

1). k отл. от всех $k_i \Rightarrow \leq \sum_{i=1}^{n} \frac{1}{m} = \frac{n}{m} = O(\frac{n}{m} + 1)$

2). $k = k_j$, $k \neq k_i \ \forall i \neq j$ $\Rightarrow \leq \sum_{\substack{i \neq j \\ i = 1..n}} \frac{1}{m} + \underbrace{Pr\{h(k_j) = h(k)\}}_{=1} = \frac{n-1}{m} + 1 = O(\frac{n}{m} + 1)$

∎

Утв. $\mathcal{H}_{p \to m} = \{h_{a,b}\}_{\substack{a \in \{1..p-1\} \\ b \in \mathbb{Z}_p}}$ $\quad h_{a,b} : \mathbb{Z}_p \to \mathbb{Z}_m \quad p > m$
$h_{a,b}(x) = ((a \cdot x + b) \bmod p) \bmod m$
$\mathcal{H}_{p \to m}$ — универсальное сем. х-ф

Хеш. строк: $h_{m,x}(s) \to [0..m-1]$ $\quad m \sim 10^{18}$
$ht\{h_{m,x}(s)\} = \dots$
$h_{m,x}(s) \bmod ht\_sz$ $\quad ht\_sz \sim 10^6$

Д-во $x_1, x_2 \in \mathbb{Z}_p;\ x_1 \neq x_2$

1). $t(x) = (ax + b) \bmod p$
$|\mathcal{H}_{p \to m}| = p(p-1)$

* $ax_1 + b \not\equiv_p ax_2 + b$
$\underbrace{a}_{\neq 0}\underbrace{(x_1 - x_2)}_{\neq 0} \not\equiv_p 0$

$(a, b) \rightleftarrows (t(x_1), t(x_2))$ — биекция в мн-во $\{(y_1, y_2) \in \mathbb{Z}_p^2 \mid y_1 \neq y_2\}$
$\# = p(p-1)$ $\quad (t(x_1) \neq t(x_2))$* $\quad \# = p(p-1)$

$y_1 \equiv_p ax_1 + b \quad | \cdot x_2$
$y_2 \equiv_p ax_2 + b \quad | \cdot x_1$
известны

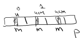
$x_2 y_1 - x_1 y_2 \equiv_p ax_1x_2 + bx_2 - ax_1x_2 - bx_1$
$b(x_2 - x_1) \equiv_p x_2 y_1 - x_1 y_2$
$b \equiv_p (x_2 y_1 - x_1 y_2) \cdot \underbrace{(x_2 - x_1)}_{\neq 0}^{-1}$

$\underbrace{y_1 - y_2}_{\neq 0} \equiv_p a \underbrace{(x_1 - x_2)}_{\neq 0}$

2). $h(x_1) = y_1 \bmod m$ $\quad y_1 = t(x_1)$ $\quad (y_1, y_2)$ — р/м распределена
$h(x_2) = y_2 \bmod m$ $\quad y_2 = t(x_2)$ $\quad$ на $\{(a_1, a_2) \in \mathbb{Z}_p^2 \mid a_1 \neq a_2\}$
$\forall a_1 \neq a_2 \ Pr\{y_1 = a_1 \wedge y_2 = a_2\} = \frac{1}{p(p-1)}$

Зафикс. $u = h(x_1) = y_1 \bmod m$

$$Pr\{y_2 \bmod m = u\} = \frac{k}{p-1} \leq$$



$$\leq \frac{\cancel{p-1}}{m\cancel{(p-1)}} = \frac{1}{m}$$

$\boxed{\begin{matrix} u + km < p \\ u + (k+1)m > p \end{matrix}}$ $\quad u, u+m, u+2m, \dots, \underset{y_1}{\overset{}{u+km}}$

$k < \frac{p-u}{m}$ $\quad k \leq \frac{p-1}{m}$
$\Downarrow \qquad \Uparrow$
$m \cdot k < p \Rightarrow mk \leq p-1$

∎

Хеш-таблица для хеширования строк полин. хешом

$h_{m,x}(s) \bmod ht\_sz$ $\quad (m \in \mathbb{P})$

Утв. $h(s) = \underbrace{h_{m \to ht\_sz}}_{\text{универс.}}(\underbrace{h_{m,x}(s)}) \quad m \in \mathbb{P}$ $\quad a, b \in \mathbb{Z}_{ht\_sz},\ a \neq 0$
$((a \cdot h_{m,x}(s) + b) \bmod m) \bmod ht\_sz$

$\mathcal{H} = \{h_{m,x,ht\_sz,a,b}\}_{x, a, b}$ — „почти универсально"
$x \in \mathbb{Z}_m$, $a \in \mathbb{Z}^{*}_{ht\_sz}$, $b \in \mathbb{Z}_{ht\_sz}$

Д-во $s \neq t$

$$Pr\{h(s) = h(t)\} = Pr\{h_{m,x}(s) = h_{m,x}(t)\} + Pr\{h_{m \to ht\_sz}(y_1) = h_{m \to ht\_sz}(y_2) \mid \underbrace{h_{m,x}(s)}_{y_1} \neq \underbrace{h_{m,x}(t)}_{y_2}\} \leq$$

$$\leq \underbrace{\frac{\max(|s|, |t|)}{m}}_{\text{маленькая}} + \frac{1}{ht\_sz} \overset{?}{\leq} \frac{1}{ht\_sz}$$

$m \gg$ длины строк

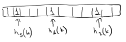
$ht\_sz = 10^6$
$m \sim 10^{18}$
$|s_i| \leq 10^6$
$\frac{10^6}{10^{18}} + \frac{1}{10^6} = 10^{-12} + 10^{-6} \sim 10^{-6}$

∎

Фильтр Блума

add(k)
exists(k) → True/False

$k \in S \Rightarrow$ ответим True
$k \notin S \Rightarrow Pr\{\text{ответим False}\} \geq 1 - 0.7^s$
$Pr\{\text{ошибки}\} \leq 0.7^s$

z 

Пример: поиск слов в документах

$k \quad h_1(k) \quad h_2(k) \ \dots \ h_s(k)$

Mem = m битов
Time = s подсч. х-ф
Предполагаем, что $\forall k_1, \dots, k_n : k_i \neq k_j \ \forall i \neq j$ $\quad h_a(k_i) \sim U[0..m-1]$
$\forall a \in [1..s]$ $\quad$ независимы в совокупн.
(например, ключи даны случайн. незав.)

$Pr\{\text{ошибки}\} \leq ?$

$k \to s$ значений оказались $= 1$
$k \notin S$

$t_1, t_2, \dots, t_s$
$Pr\{z\{t_1\} == 1\} = 1 - Pr\{z\{t_1\} == 0\} =$
$= 1 - (Pr[h_a(k) \neq t_1])^{sn} = 1 - Pr\{h_a(k) \neq t_1\}^{sn} =$

$= 1 - \left(1 - \frac{1}{m}\right)^{sn}$ $\quad \left(Pr\{h_a(k) \neq t_1\} = 1 - \frac{1}{m}\right)$

$$Pr\{\text{ошибки}\} = \left(1 - \underbrace{\left(1 - \frac{1}{m}\right)^{sn}}\right)^s \leq \left(1 - e^{-\frac{sn}{m-1}}\right)^s \underset{s = \dots}{=} \left(1 - \frac{1}{2}\right)^{\frac{m-1}{n}\ln 2} = \left(\frac{1}{2^{\ln 2}}\right)^{\frac{m-1}{n}} \leq 0.63^{m/n}$$

$\boxed{e^x \geq 1 + x, \ x \in \mathbb{R}}$

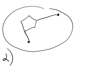
$$\left(1 - \frac{1}{m}\right)^{sn} = \left(\frac{m-1}{m}\right)^{sn} = \left(\frac{m}{m-1}\right)^{-sn} = \left(1 + \frac{1}{m-1}\right)^{-sn} \leq e^{-sn/(m-1)}$$

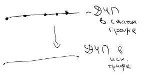
$s = \frac{m-1}{n} \cdot \ln 2$ $\quad e^{-\frac{sn}{m-1}} = \frac{1}{2}$

Паросоч. в произв. графах
(невзвеш.)

(1) $\nexists$ ДЧП $\Leftrightarrow$ п/сог max (в любом графе)

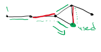
Алгоритм сжатия соцветий:





$O(V^3)$, $O(VE \cdot \alpha)$

(2) dfs для двуд. графов
нашелся ⇒ ура



Перемешаем у ∀ вершины все ребра

```
while 1
    for i=1...V
        if i - св.: dfs(i)
    used = ∅
```

```
dfs(i):
    shuffle(edges[i])
```

(3) Матрица Татта

$$M_{ij} = \begin{cases} 0, & \text{ребра } (i\,j) \text{ нет} \\ x_{ij}, & i > j \ (i\,j) \in E \\ -x_{ij}, & i < j \ (i\,j) \in E \end{cases}$$

n^2 переменных
$\det M$ — мн-чл. от n^2 перем.

Th. $\det M \equiv 0 \Leftrightarrow \nexists$ соверш. п/сог.

Lm Шварца – Зиппеля: $x_{ij} \to$ случ. $\mathbb{Z}_p$

(если $\not\equiv 0$) $\quad Pr\{\det M(x_{ij}) = 0 \bmod p\} \leq \frac{\deg \det M}{p} = \frac{n}{p}$

$p = 10^{18} \quad n = 10^5$
$Pr \leq 10^{-13}$

Rem G – двудольный ⇒ теорема для матрицы Татта очевидна.