Universelles Hashing

Def.: Eine Klasse H von Hash-Fkt. heißt universell bezüglich Größe m der Hash-Tabelle falls:

$\forall$ Schlüsselpaare $x, y \in U$ mit $x \neq y$ gilt:

$$\left|\{h \in H \mid h(x) = h(y)\}\right| \leq \frac{|H|}{m}$$

Satz: Wählt man $h \in H$ zufällig und gleichverteilt und ist $n \leq m$, so ist

$E(\#\text{Kollisionen mit } x) < 1$ für alle x.

Frage: Existiert so eine universelle Klasse H?

Antwort: Ja

Wähle zunächst Primzahl $p \geq x \; \forall x$.
Für $a \in \mathbb{Z}_p^* = \{1, \dots, p-1\}$ und $b \in \mathbb{Z}_p = \{0, 1, \dots, p-1\}$ definiere

$$h_{a,b}(x) := ((a \cdot x + b) \bmod p) \bmod m$$

Satz: Dann ist $H = \{h_{a,b} \mid a \in \mathbb{Z}_p^*, b \in \mathbb{Z}_p\}$ universell bezüglich m.

Beweis: Betrachte zwei Schlüssel $x \neq y$ aus $\{0, \dots, p-1\}$ und $h_{a,b} \in H$. Definiere

$$\boxed{\begin{aligned} r &:= a \cdot x + b \bmod p \\ s &:= a \cdot y + b \bmod p \end{aligned}} \quad (*)$$

Beh: $x \neq y \Rightarrow r \neq s$

Bew: Annahme: $r = s \Rightarrow a \cdot x + b \equiv a \cdot y + b \bmod p$

$\overset{a \in \mathbb{Z}_p^*}{\Rightarrow} x \equiv y \bmod p$

$\Rightarrow x = y$. ↯ □

Beh: a und b sind eindeutig durch x, y, r, s bestimmt.

Bew.: Aus (*) folgt

$(r-s) \equiv a \cdot \underbrace{(x-y)}_{\in \mathbb{Z}_p^*} \bmod p$

$\Rightarrow a = (x-y)^{-1} \cdot (r-s) \in \mathbb{Z}_p$

Zu b: $r = a \cdot x + b \Rightarrow b = r - (x-y)^{-1} \cdot (r-s) \cdot x$ □

Es gibt $(p-1) \cdot p$ Möglichkeiten für die Wahl von a und b.

Es gibt $(p-1) \cdot p$ Paare $r, s \in \{0, \dots, p-1\}$ mit $r \neq s$.

$\Rightarrow$ Es gibt Bijektion zwischen möglichen Paaren (a, b) und (r, s).

$\Rightarrow$ Wählt man eine zufällige Hashfkt. aus H, also ein zufälliges Paar (a, b), so erhält man ein zufälliges Paar (r, s).

Es gilt:

$\Pr\left(h_{a,b}(x) = h_{a,b}(y)\right) = \Pr(r \equiv s \bmod m)$

Für festes r gibt es $\leq \lceil \frac{p}{m} \rceil - 1$ mögliche $s \neq r$ mit $r \equiv s \bmod m$.
Es gilt

$$\lceil \frac{p}{m} \rceil - 1 \leq \frac{p+m-1}{m} - 1 = \frac{p-1}{m}$$

Daraus folgt: Die W'keit dafür, dass ein zufälliges $s \in \{0, \dots, p-1\} \setminus \{r\}$ zu einer Kollision führt ist $\leq \frac{1}{m}$.
$\Rightarrow$ H ist universell bezüglich m. $\square$

Perfektes Hashing:

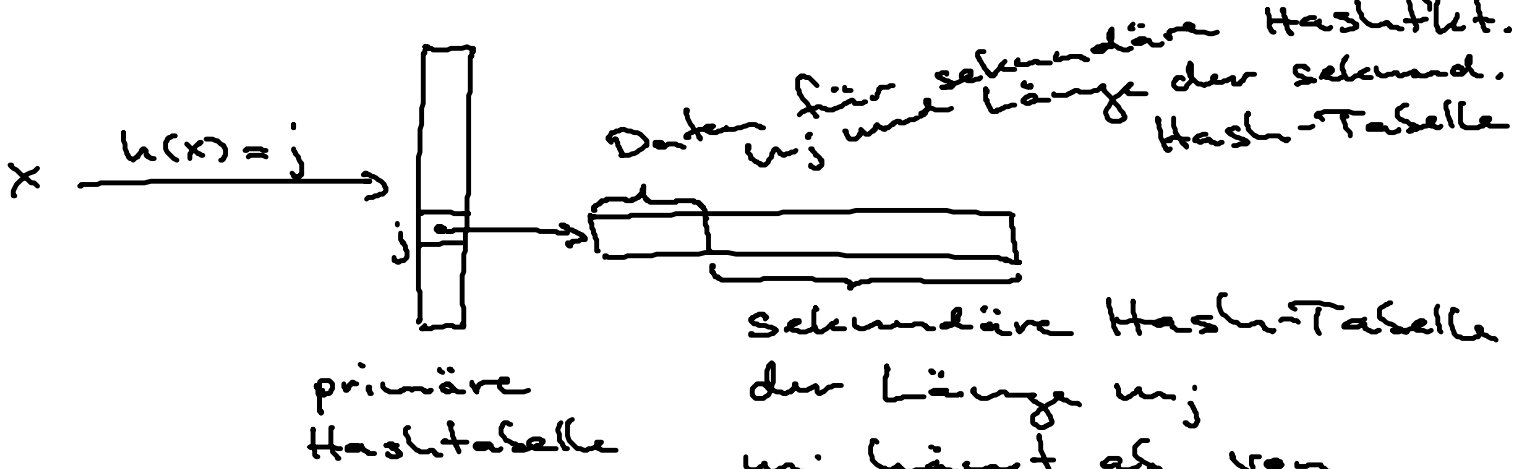
Nehme an, dass Schlüsselmenge statisch und a priori bekannt ist:

z.B.:
- reservierte Worte in Java (Compiler verwendet Hashing)
- Namen aller Dateien auf DVD.

Ziel: Verbessere Worst-Case Aufwand für die Suche (im Mittel Aufwand $O(1)$).

Def.: Ein Hashing-Verfahren heißt **perfekt** :$\Leftrightarrow$ Worst Case Aufwand für das Suchen ist $O(1)$.

Wir realisieren ein solches Hashing-Verf. mittels eines zweistufigen Verfahrens.



m_j hängt ab von

$m_j := \#$ Schlüssel mit $h(x) = j$

Konkret: $m_j = n_j^2$

Bem.: m_j kann man vorab (nach Wahl von h) bestimmen, da alle Schlüssel bekannt.

Primäre Hashfkt. h und sekundäre Hashfkt. h_j, $j = 1, \dots, m$, sind aus universeller Klasse gewählt.

Konkreter:

$$h(x) = ((a \cdot x + b) \bmod p) \bmod m$$

(m: Länge primärer Hashtabelle; $h \in H_{p,m}$; p: Primzahl $> x$ $\forall$ Schlüssel x)

$$h_j(x) = ((a_j \cdot x + b_j) \bmod p) \bmod m_j$$

$\uparrow \in H_{p,m_j}$ $\qquad m_j = m_j^2$

Spezialfall: Falls $m_j = 1$, wähle $a_j = b_j = 0$,
$m_j = m_j^2 = 1$

Zwei mögliche Probleme:

A: Wie stellt man sicher, dass beim sekundären Hashing keine Kollision auftritt?

B: Kann der benötigte Speicherplatz irgendwie beschränkt werden?

zu A:

Satz: (Eigenschaft der sek. Hashfkt mit $m = m_j$, $m = m_j$)
n Schlüssel werden mit zufällig gewählter Hashfkt. $h \in H_{p,m}$ in Hashtabelle der Länge $m = m^2$ gespeichert. Dann ist die Wahrscheinlichkeit dafür, dass eine Kollision auftritt $< \frac{1}{2}$.

Dieser Satz besagt, dass bei wiederholter zufälliger Wahl von $h \in H_{p,m}$ nach wenigen Wiederholungen eine kollisionsfreie Hashfkt. gefunden wird.

mit großer Wahrscheinlichkeit.

Beweis des Satzes:

Es gibt $\binom{n}{2}$ mögliche Kollisionen.

$H_{p,m}$ universell $\Rightarrow x \neq y$ kollidieren mit W'keit $\leq \frac{1}{m}$.

Sei X Anzahl der Kollisionen, dann gilt

$$E(X) = E\left(\sum_{x \neq y} C_{x,y}\right) = \sum_{x \neq y} E(C_{x,y}) \leq \binom{n}{2} \cdot \frac{1}{m}$$

$$= \frac{n \cdot (n-1)}{2} \cdot \frac{1}{n^2} < \frac{1}{2}$$

Markov'sche Ungleichung:

Für eine Zufallsgröße $X \geq 0$ und Zahl $t > 0$ gilt:

$$Pr(X \geq t) \leq \frac{E(X)}{t}.$$

In unserem Fall:

$$Pr(X \geq 1) \leq \frac{E(X)}{1} < \frac{1}{2}$$

□

→ Problem A gelöst!