Hashing Universal y Perfecto

Note Title 11/3/2009

$h : [0, M-1] \to [0, N-1]$

$M >> N$

h "bien equilibrada" para $S \subset [0, M-1]$

↳ si $|S| \leq N$, no hay conflictos

si $|S| > N$, conflictos están repartidos

Problema $\forall h \, \exists S$ con muchas colisiones

Soluciones ① Elegir h al azar → Universal hash

② Elegir h en función de S → Perfect hashing

I Hashing Universal

① **Definicion**

Una familia H de funciones : $[0, M-1] \to [0, N-1]$
es 2-universal si
$\forall\ x \neq y \in [0, M-1]$ $\Pr_{h \in H}\left[h(x) = h(y)\right] \leq \frac{1}{N}$

$\leadsto$ idea es de elegir h al azar en H

Notaciones

$$\cdot\ C_{x,y} = \begin{cases} 1 & \text{si } h(x) = h(y) \\ 0 & \text{sino} \end{cases} \Rightarrow E(C_{x,y}) = Pr(C_{xy} = 1) \quad (*)$$

$$\leq \begin{cases} \frac{1}{n} & \text{si } x \neq y \\ 1 & \text{sino} \end{cases}$$

$$\forall x\quad C_{x,S} = |\{y \in S,\ C_{x,y} = 1\}| = \sum_{y \in S} C_{xy}$$

Resultado $\forall S \subset [0, M-1]$ $\forall h$ al azar en H

Si H es 2-universal y $x \in S$

$$E(C_{x,S}) = \sum_{y \in S} E(C_{xy}) \overset{*}{\leq} 1 + \frac{|S|-1}{N} < 1 + \frac{|S|}{N}$$

linearidad de E

Corolario Si elijo $|S| = \Theta(N)$ y un h al azar en H universal el costo esperado de todas las operaciones es $O(1)$

Nota No hay garantía sobre el costo en peor caso..

③ Ejemplo de Familia Universal

$$\mathcal{H}_{p,N} = \left\{ h_{a,b} \;/\; \begin{array}{l} 1 \leq a < p \\ 0 \leq b < p \end{array} \right\}$$

donde - $h_{a,b} = \big((a x + b) \bmod p \big) \bmod N$

- p primo $> N$

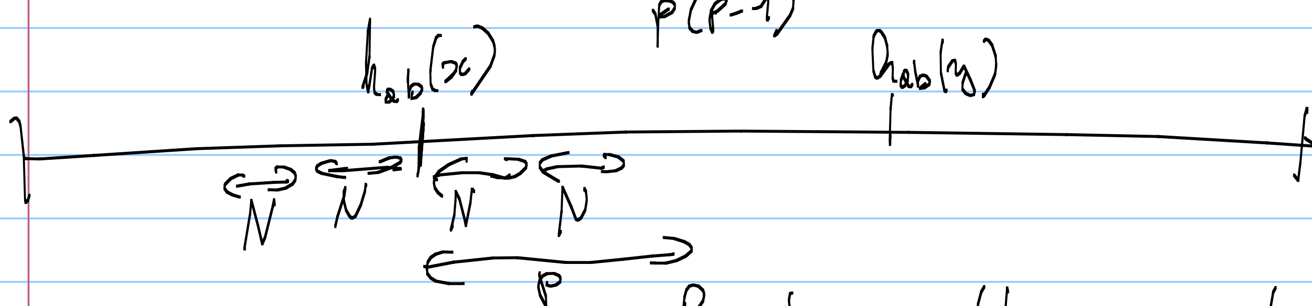
(Nota : en practica p primo "con alta probabilidad")

Lemma H_{pN} es Universal

Intuition de la prueba

Queremos mostrar $\forall x \neq y \ \Pr_{a,b}\left[h_{ab}(x) = h_{ab}(y)\right] \leq \frac{1}{N}$

$\Pr[a,b \text{ elegidos}] = \frac{1}{p(p-1)}$



Prueba completa en apuntes

2) Hashing Perfecto

Dado S, ¿como elegir h?

Una funcion de hashing $h : [0, U-1] \to [0, N-1]$ es **perfecta** para $S \subset [0, U-1]$ si no genera ninguna colisiones para $x \neq y \in S$

Objectivo Construir tal h (para tener peor caso $\in O(1)$) $(|S| \leq N)$

- en tiempo polinomial deterministica (e.g. N^2)
- al azar (algoritmo de Las Vegas)

(a) <u>Caso fácil</u> Espacio $|S|^2$

* estamos dispuestos a pagar $|S|^2$ de espacio
* eligamos $h \in H$ universal al azar
* Pr[h perfecta para S] ?

$$\forall x, y \quad \Pr[h(x) = h(y)] \leq 1/N^2 \quad \text{porque } h \in H \text{ universal}$$

$$\Pr[\exists x \neq y \in S \;\; h(x) = h(y)] \leq \frac{N(N-1)}{2} * \frac{1}{N^2} < \frac{1}{2}$$

$\Rightarrow$ Algoritmo de tipo Las Vegas, $O(1)$ iteraciones en promedio, costo $O(|S|)$ (para verificar)

(b) Caso Real $\qquad O(N) = O(|S|)$

(1) Elijo $h \in H$ universal $[0, n-1] \longrightarrow [0, N-1]$

(2) $B_i \triangleq$ # elementos tq $h(x) = i \qquad \forall i \in [0, N-1]$

(3) Si $\sum B_i^2 > 4n$, vuelvo al paso (1)

* Para cada Celda $i \in [0, N-1]$:

- tenemos un conjunto de tamaño B_i
- podemos "gastar" B_i^2 espacio
- aplicamos el "caso fácil" para cada celda

$\Rightarrow h_0, \ldots, h_{N-1}$ funciones de hash "perfectas" en tiempo lineal en promedio, espacio $5n$

$\gamma \quad Pr\left[\sum B_i^2 < 4n\right]$?

$$\sum_i B_i^2 = \sum_{x,y} C_{xy} \qquad B_i^2 = \left|\{(i,j),\ i,j \in B_i\}\right|$$

$$= \{(i,j),\ C_{ij} = 1\}$$

$$E\left(\sum B_i^2\right) = E\left(\sum C_{xy}\right)$$

$$= \underbrace{E\Big(\sum_{x=y} \underbrace{C_{xy}}_{1}\Big)}_{N} + E\Big(\sum_{x \neq y} C_{xy}\Big)$$

$$\sum_{x\neq y} E(C_{xy}) \qquad Pr[C_{xy} = 1] \leq \frac{1}{N}$$

$$= N \qquad N(N-1) \times \frac{1}{N}$$

$$< 2N$$

Si el promedio es $< 2N$, al menos la mitad tiene valor $< 4N$

$\Rightarrow$ con probabilidad $\frac{1}{2}$ $E\left(T = \sum B_i^2\right) \in O(1)$

$\Rightarrow$ Algo de Las Vegas para encontrar hash perfecto para S dado.