Notas sobre o "hash" universal

• "Hash"

• Implementações

- Dicionários estáticos
- Dicionários dinâmicos

• Parâmetros

- Universo U de cardinalidade $|U| = u = 2^v$, por exemplo, o conjunto de todas os strings de 100 letras, $|U| = 256^{100} = 2^{800}$, $v = 800$ (bits).
- Conjunto $N \subseteq U$ de elementos a representar na tabela de "hash", por exemplo, os nomes dos alunos da Faculdade de Ciências.
- Tabela A de "hash" de cardinalidade $a = |A| = 2^b$, $a \ll u$, e, por exemplo, $a = 2n$. Exemplo: $a = 2^{15}$, $b = 15$ (bits), $v = 800$ (bits).
- Função de "hash" $h : U \to A$
- $h(y) = h(x)$, colisões. Resolução por métodos externos e internos.

• Propriedades desejáveis da função de "hash": "Espalhamento", tabela pequena, computação rápida.

• Variantes do método de "hash"

"Hash" clássico: a função de "hash" está fixa e os dados têm uma determinada distribuição probabilística. Consegue-se **tempo médio** $O(1)$. Mas, no pior caso, as operações básicas (pesquisa, inserção e eliminação) são de ordem $O(n)$.

"Hash" universal, aleatorização na função de "hash". Consegue-se **tempo médio** $O(1)$**, quaisquer que sejam os dados**.

"Hash" perfeito, os valores a memorizar são conhecidos previamente. Consegue-se determinar h por forma que **tempo** $O(1)$ **mesmo no pior caso**.

Aleatorizar o algoritmo do "hash", da mesma forma que aleatorizamos o algoritmo clássico do "quick sort"!

• "Hash" universal

H: conjunto de funções de "hash" de U em $\{1, 2, \ldots, a\}$ associada a uma distribuição probabilística. Dizemos que H é **universal** se para todo o par de valores $x \neq y$ de U é

$$\text{prob}_{h \in_U H}[h(x) = h(y)] \leq 1/a$$

A probabilidade de haver uma colisão entre 2 quaisquer elementos distintos não excede $1/a$, onde $a = 2^b$ é o tamanho da tabela de "hash".

Teorema Se H é universal então, para qualquer conjunto $N \subseteq U$ e para qualquer $x \in U$, o valor esperado, relativamente a uma escolha uniforme $h \in_U H$, do número de colisões entre x e os outros elementos de N não excede n/a (número de elementos de N a dividir pelo tamanho da tabela de "hash").

• **Construção do "hash" universal com matrizes aleatórias**

$$\text{matriz } H \text{ aleatória uniforme, } v \times b: \quad \forall x,\ h(x) = Hx$$

$v = 5$, $b = 3$, $x = [0,1,0,1,1]^t$

$$h(x) = Hx = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix} \times \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$$

Por exemplo, o segundo elemento de $h(x)$ é obtido a partir da segunda linha de H

$$(1 \times 0) + (0 \times 1) + (0 \times 0) + (1 \times 1) + (1 \times 1) = 0 + 0 + 0 + 1 + 1 = 0 \pmod 2$$

Somamos (módulo 2) alguns elementos da segunda linha de H. Quais? Os das linhas i para as quais $x_i = 1$. No exemplo acima, trata-se das linhas 2, 4 e 5 de x; isto é, temos a soma dos elementos de H de cor vermelha, $0 + 1 + 1 = 0 \pmod 2$.

Qual a probabilidade de ser $h(x) = h(y)$?

- $x, y \in U$, $x \neq y$, diferem no bit i, por exemplo, $x_i = 0$ e $y_i = 1$.
- Atribuir valores a todos os elementos de H, excepto aos da coluna i. $h(x)$ fica fixado
- Cada uma das 2^b colunas i possíveis vai originar valores de $h(y)$ diferentes... porquê?

$$\begin{matrix} & & & i & \\ \end{matrix}\begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 \end{bmatrix} \times \begin{matrix} x \\ \end{matrix}\begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 1 \end{bmatrix} = \begin{matrix} h(x) \\ \end{matrix}\begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} \quad \Bigg| \quad \begin{matrix} & & & i & \\ \end{matrix}\begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 \end{bmatrix} \times \begin{matrix} y \\ \end{matrix}\begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} = \begin{matrix} h(y) \\ \end{matrix}\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

Sendo i um bit em que x e y diferem com $x_i = 0$ e $y_i = 1$, as 2^b colunas i possíveis de H dão origem a 2^b valores de $h(y)$ distintos; assim,

- Para qualquer pré-escolha das colunas $\neq i$: $h(x)$ fica fixado e a probabilidade de $h(y)$ ter um determinado valor é $1/a$ (ou ... $h(y)$ fica fixado e...).
- $h(x)$ e $h(y)$ são variáveis aleatórias independentes.

A probabilidade de ser $h(x) = h(y)$ é $1/2^b = 1/a$.

Teorema Seja $|A| = a = 2^b$ o tamanho da tabela de "hash" e $|U| = u = 2^d$ o tamanho do universo. Para $x \in U$ seja $h(x)$ definido por $h(x) = Hx$ onde H é uma matriz aleatória e uniforme de 0's e 1's com b linhas e v colunas, e o produto matricial é efectuado no corpo $\{0, 1\}$. Então, este conjunto de funções h é um conjunto universal de "hash".