# Hash

Definição, endereçamento aberto e encadeamento

Prof. Edson Alves - UnB/FGA 2019

### Sumário

- 1. Definição de hash
- 2. Exemplos de funções *hash*
- 3. Endereçamento aberto
- 4. Encadeamento

Definição de hash

# Motivação para a criação da hash

- ullet Estruturas lineares permitem N armazenar elementos sem que o valor de N seja conhecido *a priori*, em tempo de compilação
- ullet Estas estruturas, poém, não são eficientes na busca do elementos (complexidade O(N))
- As árvores de busca, como as estruturas lineares, também permitem o armazenamento de um número arbitrário de elementos (limitado somente pela memória disponível)
- $\bullet$  Em árvores busca perfeitamente balanceadas a ordem de complexidade da busca é  $O(\log N)$
- Porém a organização de memória das árvores não é contígua, levando à perda de eficiência em relação à cache
- A idéia da hash é deduzir o índice de um elemento em um vetor a partir apenas da informação armazenada pelo elemento, reduzindo a ordem de complexidade busca para O(1)

#### Hash

#### Definição

Uma função h é uma função de hash se ela transforma uma chave K no índice do elemento que contém K na tabela. Se h transforma chaves distintas em índices distintos, ela é uma função de hash perfeita.

- Para se criar uma função h de hash perfeita, a tabela deve conter, no mínimo, o número de elementos cujas chaves serão transformadas pela função h
- Uma colisão ocorre se duas chaves distintas  $K_1$  e  $K_2$  gerarem o mesmo índice, isto é, se  $h(K_1) = h(K_2)$  com  $K_1 \neq K_2$
- A idéia é encontrar uma função h que gere o mínimo de colisões mas que não seja sofisticada ao ponto de seu cálculo interferir na performance do programa

3

Exemplos de funções hash

## Soma dos elementos

Chave	Uma string $K$
Algoritmo	Atribui-se um código numérico para cada um dos caracteres que aparecem na string $K$ e somam-se estes valores
Nível de colisão	Alto

```
int h(const string& K)
2 {
3    int v = 0;
4
5    for (const auto& c : K)
6       v += c;
7
8    return v;
9 }
```

## Resto da divisão

Chave	${\sf Um\ inteiro}\ K$
Algoritmo	Obtêm-se o resto da divisão da chave $K$ pelo tamanho $T$ da tabela. De preferência, $T$ deve ser um número primo
Nível de colisão	Inversamente proporcional a ${\cal T}$

```
unsigned long h(unsigned long K, size_t T)
{
    return K % T;
}
```

# Enlaçamento deslocado

Chave	Uma string $K$
Algoritmo	Divide-se a string em $N$ partes de, no máximo, $m$ caracteres, e computa-se a chave aplicandose a operação XOR em todas as partes
Nível de colisão	Médio

```
1 unsigned long h(const string& S) {
2    unsigned long v = 0, p = 0, m = 4, i = 0;
3
4    for (const auto& c : S) {
5         p |= (c << 8*i++);
6
7         if (i == m) v ^= p, i = p = 0;
8    }
9
10    return v ^ p;
11 }</pre>
```

# Enlaçamento no limite

Chave	Uma string $K$
Algoritmo	Variante do enlaçamento deslocado. Divide-se a chave eme 3 partes, e se enlaça os extremos com a parte do meio invertida
Nível de colisão	Médio

# Meio quadrado

Chave	Um inteiro $K$
Algoritmo	Eleva-se ${\cal K}$ ao quadrado e toma-se a parte central do resultado
Nível de colisão	Médio

```
unsigned long h(int K)

{
    auto s = K*K;

    return (s & 0x00FFFF00) >> 8;
}
```

### **Polinomial**

Chave	Uma string $K$ com $N$ caracteres
Algoritmo	$g(x)$ é um polinômio de grau $N-1$ com coeficientes $a_i=K[i]$ e $h(K)=g(p)$ (mod $T$ ), onde $T$ é o tamanho da tabela e $p\neq T$ é primo
Nível de colisão	Baixo

```
unsigned long h(const string& K, size_t p, size_t T)

unsigned long h = 0;

for (int i = K.size() - 1; i >= 0; --i)

h = (h * p) % T;
h = (h + K[i]) % T;

return h;
```

Endereçamento aberto

#### Problemas com a colisão

- Como visto, as funções de hash podem gerar colisões, isto é, um mesmo índice para duas chaves distintas
- Naturalmente surge o seguinte questionamento: como inserir duas chaves que colidem em uma mesma tabela, e como resgatá-las em uma busca?
- Uma alternativa para o tratamento de colisões é o endereçamento aberto

### Endereçamento aberto

#### Definição

Se a chave K for mapeada para uma posição já ocupada da tabela, o endereçamento aberto utiliza a sequência de sondagem

$$N(h(K) + p(1)), N(h(K) + p(2)), \dots, N(h(K) + p(i)), \dots$$

onde p é a função de sondagem, i é o índice de sondagem e N a função de normalização, até que

- 1. se encontre uma posição desocupada
- 2. N(h(K) + p(j)) = N(h(K))
- 3. se verifique que a tabela está cheia

## Sondagem linear

- Na sondagem linear, temos a função de sondagem é a identidade, isto é,  $p(i)=i, \forall i$
- A função de normalização faz com que o índice resultante esteja dentro dos limites da tabela, usando o resto da divisão:

$$N(K)=K(\mathrm{mod}\ T),$$

onde T é o tamanho da tabela

- Se uma posição N(h(K)+p(i)) já estiver ocupada, tenta-se o próximo índice de sondagem (i+1) até que se encontre um espaço vago ou ocorra uma das outras condições
- Esta estratégia tende a formação de agrupamentos de chaves, com pontos de acumulação na tabela e intervalos contíguos não ocupados

Sondagem linear, T = 11

0	1	2	3	4	5	6	7	8	9	10

Sondagem linear, T = 11

$$h(51) = 51 \; (\bmod \; 11) = 7$$

V	ı	2	3	4	3	0	/	0	9	10
							51			

Sondagem linear, T = 11

0	1	2	3	4	5	6	7	8	9	10
							51			

Sondagem linear, T=11

$$h(16) = 16 \pmod{11} = 5$$

0	1	2	3	4	5	6	/	8	9	10
					16		51			

Sondagem linear, T = 11

0	1	2	3	4	5	6	7	8	9	10
					16		51			

Sondagem linear, T=11

$$h(76) = 76 \; (\bmod \; 11) = 10$$

V	ı	2	3	4	5	0	/	0	9	10
					16		51			76

Sondagem linear, T=11

0	1	2	3	4	5	6	7	8	9	10
					16		51			76

Sondagem linear, T=11

$$h(35) = 35 \; (\bmod \; 11) = 2$$

0	1	2	3	4	5	6	7	8	9	10
		35			16		51			76

Sondagem linear, T=11

0	1	2	3	4	5	6	7	8	9	10
		35			16		51			76

Sondagem linear, T=11

$$h(-6) = -6 \pmod{11} = 5$$

V	ı	2	3	4	5	0	/	٥	9	10
		35			16		51			76

Sondagem linear, T = 11

$$N(h(-6)+1)=(5+1) \; (\bmod \; 11)=6$$

0	1	2	3	4	5	6	7	8	9	10
		35			16	-6	51			76

Sondagem linear, T = 11

0	1	2	3	4	5	6	7	8	9	10
		35			16	-6	51			76

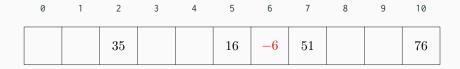
Sondagem linear, T = 11

$$h(49) = 49 \; (\bmod \; 11) = 5$$

Ø	ı	2	3	4	Э	0	/	٥	9	10
		35			16	-6	51			76

Sondagem linear, T=11

$$N(h(49)+1)=(5+1) \; (\bmod \; 11)=6$$



Sondagem linear, T = 11

$$N(h(49)+2)=(5+2) \; (\bmod \; 11)=7$$

0	1	2	3	4	5	6	7	8	9	10
		35			16	-6	51			76

Sondagem linear, T = 11

$$N(h(49)+3)=(5+3) \; (\bmod \; 11)=8$$

0	1	2	3	4	5	6	7	8	9	10
		35			16	-6	51	49		76

# Implementação da sondagem linear

```
1 #include <bits/stdc++ h>
3 using namespace std;
5 template<typename I, size_t T>
6 class HashSet {
7 private:
      size_t mod(const I& a, int b)
q
          return ((a % b) + b) % b;
10
      size_t h(const I& K) { return mod(K, T); }
14
      size_t N(const I& K, size_t i) { return mod(h(K) + i, T); }
16
     vector<I> xs;
      bitset<T> used;
18
20 public:
      HashSet() : xs(T) \{ \}
```

# Implementação da sondagem linear

```
22
      bool insert(const I& K)
24
         if (used.count() == T)
25
               return false;
26
          for (size_t i = 0; i < T; ++i)
28
          {
               auto pos = N(K, i);
30
               if (not used[pos])
32
                   xs[pos] = K;
34
                   used[pos] = true;
                   break;
36
38
39
          return true:
40
41
```

# Sondagem quadrática

Na sondagem quadrática, a função de sondagem é dada por

$$p(i) = (-1)^{i-1} \left[ \frac{i+1}{2} \right]^2,$$

para i = 1, 2, ..., T - 1

- A função de normalização é dada por  $N(K) = K \pmod{T}$ , onde T é o tamanho da tabela
- A sondagem quadrática pode ser interpretada como a sequência

$$h(K)+i^2, h(K)-i^2, h(K)+(i+1)^2, h(K)-(i+1)^2, \ldots$$
 para  $i=1,2,\ldots,(T-1)/2$ 

• Se T for um número primo da forma 4k+3, a sequência acima passa por todas as posições da tabela (Radke, 1970)

# Exemplo de inserção utilizando sondagem quadrática

Sondagem quadrática, T = 11Elemento a ser inserido: 51

0	1	2	3	4	5	6	7	8	9	10

Sondagem quadrática, T=11

$$h(51) = 51 \; (\bmod \; 11) = 7$$

V	ı	2	3	4	5	0	1	0	9	10
							51			

Sondagem quadrática, T = 11Elemento a ser inserido: 16

0	1	2	3	4	5	6	7	8	9	10	
							51				

Sondagem quadrática, T=11

$$h(16) = 16 \; (\bmod \; 11) = 5$$

0	1	2	3	4	5	6	7	8	9	10
					16		51			

Sondagem quadrática, T=11

0	1	2	3	4	5	6	7	8	9	10
					16		51			

Sondagem quadrática, T=11

$$h(76) = 76 \; (\bmod \; 11) = 10$$

0	1	2	3	4	5	6	/	8	9	10
					16		51			76

Sondagem quadrática, T=11

0	1	2	3	4	5	6	7	8	9	10
					16		51			76

Sondagem quadrática, T=11

$$h(35) = 35 \; (\bmod \; 11) = 2$$

0	1	2	3	4	5	6	7	8	9	10
		35			16		51			76

Sondagem quadrática, T=11Elemento a ser inserido: -6

0	1	2	3	4	5	6	7	8	9	10	
		35			16		51			76	

Sondagem quadrática, T = 11Elemento a ser inserido: -6

$$h(-6) = -6 \pmod{11} = 5$$

V	1	2	3	4	5	0	/	0	9	10
		35			16		51			76

Sondagem quadrática, T=11

$$N(h(-6) + 1^2) = (5+1) \pmod{11} = 6$$

0	1	2	3	4	5	6	7	8	9	10
		35			16	-6	51			76

Sondagem quadrática,  $T=11\,$ 

0	1	2	3	4	5	6	7	8	9	10
		35			16	-6	51			76

Sondagem quadrática, T=11

$$h(49) = 49 \; (\bmod \; 11) = 5$$

0	1	2	3	4	5	6	7	8	9	10
		35			16	-6	51			76

Sondagem quadrática, T=11

$$N(h(49)+1^2)=(5+1) \; (\bmod \; 11)=6$$

0	1	2	3	4	5	6	7	8	9	10
		35			16	-6	51			76

Sondagem quadrática, T=11

$$N(h(49) - 1^2) = (5 - 1) \pmod{11} = 4$$

0	1	2	3	4	5	6	7	8	9	10
		35		49	16	-6	51			76

## Implementação da sondagem quadrática

```
1 #include <bits/stdc++ h>
3 using namespace std;
5 template<typename I, size_t T>
6 class HashSet {
7 private:
     size_t mod(const I& a, int b) { return ((a % b) + b) % b; }
9
     size_t h(const I& K) { return mod(K, T); }
10
     size_t N(const I& K, size_t i)
     {
          auto sign = i % 2 ? 1 : -1;
14
          auto i = (i + 1)/2:
16
          return mod(h(K) + sign * j * j, T);
18
     vector<I> xs;
20
     bitset<T> used;
```

# Implementação da sondagem quadrática

```
23 public:
      HashSet() : xs(T) {}
25
      bool insert(const I& K)
26
      {
27
         if (used.count() == T) return false;
28
          for (size_t i = 0; i < T; ++i)
30
               auto pos = N(K, i);
               if (not used[pos])
34
                   xs[pos] = K;
36
                   used[pos] = true;
                   break;
38
40
41
          return true;
42
43
```

#### hash duplo

- O hash duplo é uma das melhores estratégias de endereçamento aberto
- Isto porque a sequência de sondagem gerada tem muitas das características das sequências aleatórias
- No hash duplo a sequência de sondagem tem a forma

$$h(K) = (h_1(K) + ih_2(K)) \pmod{T}, \quad i = 0, 1, 2, \dots, T - 1$$

onde  $h_1(K), h_2(K)$  são duas funções de *hash* auxiliares, i é o índice de sondagem e T é o tamanho da tabela

- Diferentemente das sondagens lineares e quadráticas, a função  $h_2(K)$  depende do valor do chave K
- Deste modo, as sequências de sondagem para chaves  $K_1 \neq K_2$ , com  $h_1(K_1) = h_1(K_2)$ , tendem a serem diferentes

#### hash duplo

- A função  $h_2(K)$  deve gerar valores co-primos com o tamanho T da tabela
- Se T é uma potência de dois (isto é,  $T=2^k$  para algum k positivo), a função  $h_2(K)$  deve gerar apenas números ímpares
- Se T é um número primo,  $h_2(K)$  tem que gerar números positivos menores do que T
- Uma maneira de se obter isso é fazer  $h_1(K) = K \pmod T$  e  $h_2(K) = 1 + (K \pmod {T-1})$
- $\bullet$  O hash duplo tem melhor desempenho do que a sondagem linear e a sondagem quadrática, pois gera um número  $O(T^2)$  de sondagens possíveis, enquanto que as outras duas geram O(T) sequências de sondagem

 $\it Hash \ duplo, \ T=11$ 

0	1	2	3	4	5	6	7	8	9	10

 $\it Hash \ duplo, \ T=11$ 

$$h(51) = h_1(51) = 51 \ (\mathsf{mod}\ 11) = 7$$

0	ı	2	3	4	5	б	/	8	9	10
							51			

 $\it Hash \ duplo, \ T=11$ 

0	1	2	3	4	5	6	7	8	9	10
							51			

Hash duplo, T=11

$$h(16) = h_1(16) = 16 \; (\bmod \; 11) = 5$$

0	ı	2	3	4	5	6	/	8	9	10
					16		51			

 $\it Hash \ duplo, \ T=11$ 

0	1	2	3	4	5	6	7	8	9	10
					16		51			

Hash duplo, T = 11

$$h(76) = h_1(76) = 76 \text{ (mod } 11) = 10$$

0	1	2	3	4	5	6	/	8	9	10
					16		51			76

 $\textit{Hash duplo}, \ T=11$ 

0	1	2	3	4	5	6	7	8	9	10
					16		51			76

Hash duplo, T=11

$$h(35) = h_1(35) = 35 \ (\mathsf{mod}\ 11) = 2$$

0	1	2	3	4	5	6	7	8	9	10
		35			16		51			76

 $\it Hash \ duplo, \ T=11$ 

0	1	2	3	4	5	6	7	8	9	10
		35			16		51			76

Hash duplo, T=11

$$h(-6) = h_1(-6) = -6 \pmod{11} = 5$$

V	1	2	3	4	5	0	/	0	9	10
		35			16		51			76

Hash duplo, T = 11

$$h(-6) = (h_1(-6) + h_2(-6)) \pmod{11} = 5 + 5 = 10$$

0	1	2	3	4	5	6	7	8	9	10
		35			16		51			76

Hash duplo, T = 11

$$h(-6) = (h_1(-6) + 2h_2(-6)) \pmod{11} = 5 + 10 \pmod{11} = 4$$

0	1	2	3	4	5	6	/	8	9	10
		35		-6	16		51			76

 $\textit{Hash duplo}, \ T=11$ 

0	1	2	3	4	5	6	7	8	9	10
		35		-6	16		51			76

Hash duplo, T = 11

$$h(49) = h_1(49) \pmod{11} = 5$$

0	1	2	3	4	5	6	/	8	9	10
		35		-6	16		51			76

Hash duplo, T = 11

$$h(49) = (h_1(49) + h_2(49)) \; (\mathsf{mod} \; 11) = 5 + 10 \; (\mathsf{mod} \; 11) = 4$$

0	1	2	3	4	5	6	7	8	9	10
		35		-6	16		51			76

Hash duplo, T = 11

$$h(49) = (h_1(49) + 2h_2(49)) \text{ (mod } 11) = 5 + 20 \text{ (mod } 11) = 3$$

v	'	2	3	7	J	O	,	0	5	10
		35	49	-6	16		51			76

## Implementação do hash duplo

```
1 #include <bits/stdc++ h>
3 using namespace std;
5 template<typename I, size_t T>
6 class HashSet {
7 private:
     size_t mod(const I& a, int b) { return ((a % b) + b) % b; }
9
     size_t h1(const I& K) { return mod(K, T); }
10
     size_t h2(const I& K) { return 1 + mod(K, T - 1); }
     size_t N(const I& K, size_t i)
14
     {
          return mod(h1(K) + i*h2(K), T):
16
18
     vector<I> xs;
19
     bitset<T> used;
20
```

## Implementação do hash duplo

```
HashSet() : xs(T) {}
24
      bool insert(const I& K)
25
26
         if (used.count() == T) return false;
27
28
          for (size_t i = 0; i < T; ++i)
30
               auto pos = N(K, i);
32
               if (not used[pos])
34
                   xs[pos] = K;
                   used[pos] = true;
36
                   break;
38
39
40
          return true;
41
42
```

**Encadeamento** 

#### Resolução de colisão por encadeamento

- Uma outra forma de se tratar as colisões é o uso do encademento
- A ideia é que cada célula da tabela corresponda a uma lista duplamente encadeada
- Cada chave K tal que h(K) = j, onde  $j = 0, 1, \dots, T-1$ , é adicionada à lista que ocupa a posição j
- O uso de listas duplamente encadeadas permite uma remoção mais eficiente
- Se a opção de remoção não for implementada, uma lista simplesmente encadeada ou um vector também podem ser utilizados
- ullet Deve-se tomar cuidado, porém, porque no pior caso todas as chaves colidem em uma mesma posição, de modo que a busca e a inserção passam a ter complexidade O(N), porém ocupando mais espaço em memória que uma única lista

Encadeamento, T=11

Elemento a ser inserido: 51

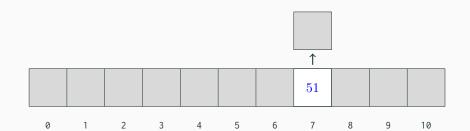


1 2 3 4 5 6 7 8

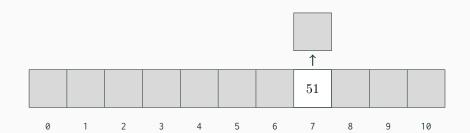
9

10

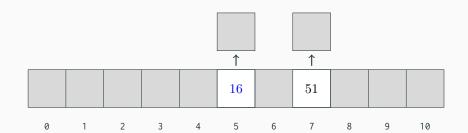
Encadeamento, T = 11



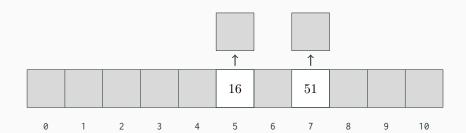
Encadeamento, T=11



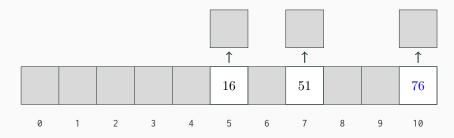
Encadeamento, T = 11



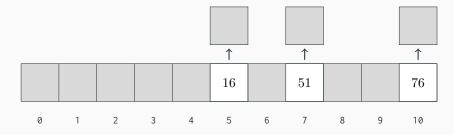
Encadeamento, T = 11



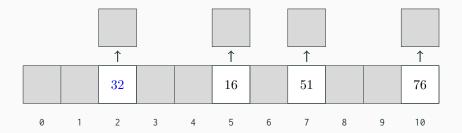
Encadeamento, T = 11



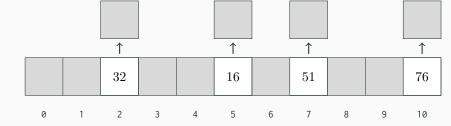
Encadeamento, T = 11



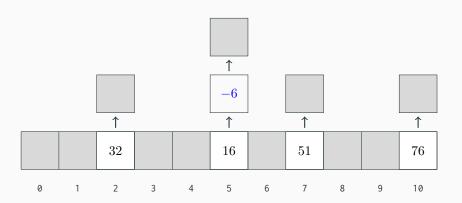
Encadeamento, T = 11



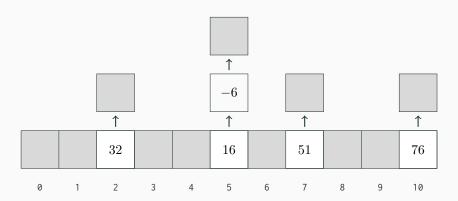
Encadeamento, T = 11

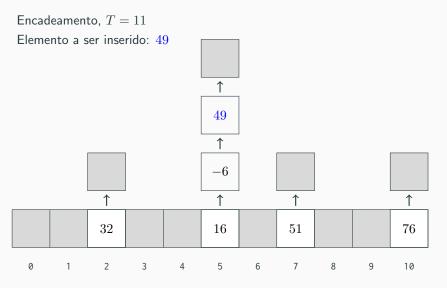


Encadeamento, T = 11



Encadeamento, T = 11





# Exemplo de implementação do encadeamento

```
1 #include <bits/stdc++ h>
3 using namespace std;
5 template<typename I, size_t T>
6 class HashSet {
7 private:
     size_t mod(const I& a, int b) { return ((a % b) + b) % b; }
9
     size_t h(const I& K) { return mod(K, T); }
10
     vector<list<I>>> xs;
14 public:
     HashSet() : xs(T) {}
16
     void insert(const I& K)
18
          xs[h(K)].push_back(K);
20
```

#### Referências

- CORMEN, Thomas H.; LEISERSON, Charles E.; RIVEST, Ronald L.; STEIN, Clifford. *Introduction to Algorithms*, The MIT Press, 3rd edition, 2009.
- 2. **DROZDEK**, Adam. *Algoritmos e Estruturas de Dados em C++*, 2002.
- RADKE, Charles E. The Use of Quadratic Residue Research, Communications of the ACM, volume 13, issue 2, pg 103–105, 1970<sup>1</sup>.
- 4. **STROUSTROUP**, Bjarne. *The C++ Programming Language*, 2013.
- 5. C++ Reference<sup>2</sup>.

<sup>&</sup>lt;sup>1</sup>https://dl.acm.org/citation.cfm?id=362036

<sup>&</sup>lt;sup>2</sup>https://en.cppreference.com/w/