

# Classificação e Pesquisa de Dados

Aula 26

Compressão de Arquivos: Codificação de Huffman

UFRGS

INF01124

## Código de Huffman

Baseia-se no fato de que, normalmente, os caracteres ocorrem com frequências diferentes nos dados que temos a representar, ou seja, o código gerado é função da frequência de ocorrência do caracter no texto.

Este método de compressão trabalha com códigos de tamanho variável de modo que, os códigos de menor comprimento são atribuídos aos caracteres **mais frequentes**, e os códigos de maior comprimento são atribuídos aos caracteres de **menor frequência**.

## Código de Huffman

A frequência pode ser obtida:  
1- analisando todo o texto  
2- analisando uma amostra

- ◆ Supõe-se que o texto seja constituído de um conjunto de caracteres (ou símbolos)  $S = \{s_1, \dots, s_n\}$ ,  $n > 1$
- ◆ É conhecida a frequência  $f_i$  de cada símbolo  $s_i$  no texto
- ◆ Nenhum código pode ser prefixo de algum outro
- ◆ Os códigos em questão são representados como em uma árvore binária de prefixo
  - Cada símbolo  $s_i$  está associado a uma folha da árvore
  - Os códigos dos símbolos são seqüências binárias

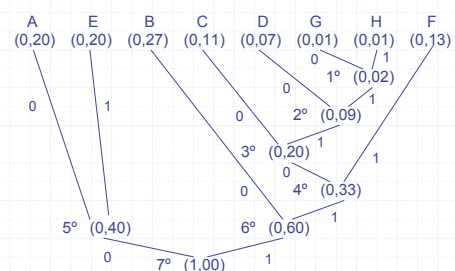
## Código de Huffman

Ch	Frequência (f)
A	0,20
B	0,27
C	0,11
D	0,07
E	0,20
F	0,13
G	0,01
H	0,01

## Construção da árvore de Huffman "Algoritmo guloso"

- ◆ O processo inicia com a definição de  $n$  sub-árvores, com um único nó cada, contendo o símbolo  $s_i$ . A frequência de cada nó é igual a frequência do símbolo a ela correspondente
- ◆ O passo geral, iterativamente, realiza:
  - Seleciona os dois nós  $T'$  e  $T''$  de menor frequência
  - Funde os dois nós em uma única árvore, criando um novo nó  $T$ , cujos filhos esquerdo e direito (indiferentemente) são as raízes das subárvores  $T'$  e  $T''$ .
  - $F(T) = f(T') + f(T'')$
- ◆ O algoritmo termina quando restar apenas uma única subárvore

## Construção da árvore de Huffman



# Código de Huffman

Ch	Frequência (f)	Código	Comprimento (l)	$l * f$
A	0,20	00	2	0,40
B	0,27	10	2	0,54
C	0,11	1100	4	0,44
D	0,07	11010	5	0,35
E	0,20	01	2	0,40
F	0,13	111	3	0,39
G	0,01	110110	6	0,06
H	0,01	110111	6	0,06

Nro. médio de bits por caracter = 2,64

# Exercício

Determine os Códigos de Huffman para os caracteres abaixo com as seguintes frequências

E	O	B	A	C	X	w	F
(0,18)	(0,20)	(0,12)	(0,22)	(0,11)	(0,03)	(0,01)	(0,13)