Codificação de Fonte (Código de Huffman)

Teoria da Informação - AULA 12

Profª. Verusca Severo

Universidade de Pernambuco
Escola Politécnica de Pernambuco

11 de agosto de 2021

Introdução

- O código de Huffman foi proposto por David Huffman em 1952. Na época, Huffman era estudante de doutorado no MIT.
- O método de codificação foi publicado por Huffman no artigo *"A Method for the Construction of Minimum-Redundacy Codes"*.
- **Curiosidade:** o método surgiu quando Huffman cursava a disciplina de teoria da informação com o prof. Fano (o responsável pelo método Shannon-Fano em parceria com Shannon).

Introdução

- Como no método proposto por Shannon-Fano, é considerado a existência de um alfabeto fonte, onde cada símbolo tem sua respectiva probabilidade de ocorrência.

- O objetivo do método também é associar códigos menores a símbolos mais prováveis e códigos maiores aos menos prováveis.

Introdução

- **Aplicação:** A codificação de Huffman é simples e rápida. É amplamente utilizada em aplicações de compressão, que vão de GZIP, PKZIP, BZIP2 a formatos de imagens como JPEG e PNG.

- Embora o algoritmo de Huffman seja ótimo para codificação símbolo a símbolo com uma distribuição de probabilidade conhecida, este não é ótimo quando a probabilidade por símbolo é desconhecida.

Introdução

- O código é obtido a partir de uma árvore construída recursivamente a partir da junção dos dois símbolos de menor probabilidade, que por sua vez são somados em símbolos auxiliares e recolocados no conjunto de símbolos.
- O processo termina quando todos os símbolos forem unidos em símbolos auxiliares, formando uma árvore binária.
- A árvore é então percorrida, atribuindo-se valores binários de 1 ou 0 para cada aresta, e os códigos são gerados a partir desse percurso.

Código de Huffman

- O código é feito da seguinte forma:

  1. Ordenação dos símbolos s_i por ordem decrescente de probabilidade $P(s_i)$;
  2. Junção dos dois símbolos de menor probabilidade num símbolo hipotético cuja probabilidade será a soma dos dois símbolos unidos. O novo símbolo deverá ser posicionado na listagem decrescente;
  3. Repetir o passo 2 até que todos os símbolos estejam agregados num único símbolo hipotético com probabilidade 1 (raiz da árvore).

- **Lembre-se:** Percorrendo a árvore da "raíz" até uma "folha", obteremos o código correspondente ao símbolo presente nessa "folha".

Código de Huffman

- **Exemplo 1:** Seja uma fonte discreta sem memória que emite símbolos do alfabeto $S = \{s_1, s_2, s_3, s_4, s_5\}$ com distribuição de probabilidades $P(S = s_1) = 0,4$, $P(S = s_2) = 0,2$, $P(S = s_3) = 0,2$, $P(S = s_4) = 0,1$ e $P(S = s_5) = 0,1$.
  a. Construa um código de Huffman para esta fonte e calcule a eficiência do código obtido.

Código de Huffman

Solução (Exemplo 1):

VER MATERIAL EM ANEXO!

Código de Huffman

- **Exemplo 2:** Uma fonte S emite os símbolos $S = \{s_1, s_2, s_3, s_4\}$ com probabilidades $P(s_1) = \frac{1}{2}$, $P(s_2) = \frac{1}{4}$, $P(s_3) = \frac{1}{8}$ e $P(s_4) = \frac{1}{8}$. Determine o código de Huffman binário para essa fonte e calcule a eficiência do código.

Código de Huffman

Solução (Exemplo 2):

VER MATERIAL EM ANEXO!

Código de Huffman

- **Exemplo 3:** Uma fonte S, discreta e sem menória, emite os símbolos $S = \{a, b\}$ com probabilidades $P(a) = \frac{9}{10}$ e $P(b) = \frac{1}{10}$
  a. Obtenha o código de Huffman binário para a extensão de ordem 3 da fonte. Calcule a eficiência
  b. Suponha que a fonte emite a mensagem "*aaaaaaabaaaa*", como ela deve ser codificada se usarmos o código anterior?

Código de Huffman

Solução (Exemplo 3):

VER MATERIAL EM ANEXO!

Código de Huffman

- **Exemplo 4:** Suponha que uma fonte discreta produz as cinco letras E, R, T, C e O com as probabilidades de ocorrência $P(E) = 0,5$, $P(R) = 0,09$, $P(T) = 0,15$, $P(C) = 0,01$ e $P(O) = 0,25$, respectivamente.

  a. Determine a entropia da fonte.

  b. Determine a sequência original de letras que deu origem à sequência codificada 00110100100010100001. O código usado foi o de Huffman, com (o símbolo E codificado pela palavra código 1 e nos ramos foi atribuído o *bit* 0 ao ramo superior de cada ramificação da árvore).

  c. Determine o número médio de bits por cada letra da fonte.

Código de Huffman

Solução (Exemplo 4):

VER MATERIAL EM ANEXO!

Código de Huffman

- **Desafio:** Uma palavra foi codificada usando o código de Huffman, tendo-se obtido a sequência binária

$$101110110101110011101 00$$

O alfabeto original era constituído pelas letras A, B, C, D, E, I, L, R e T e a letra I foi codificada como "00". Supondo que estas letras ocorriam com as probabilidades

$$P(A) = 0,26; \quad P(D) = 0,01; \quad P(L) = 0,01;$$

$$P(B) = 0,09; \quad P(E) = 0,07; \quad P(R) = 0,23;$$

$$P(C) = 0,08; \quad P(I) = 0,22; \quad P(T) = 0,03$$

Qual terá sido a palavra codificada?

Código de Huffman

Solução (Desafio):

VER MATERIAL EM ANEXO!