

Compressão de dados (Implementação) – 11/06/2023

Este trabalho tem como objetivo sedimentar conceitos vistos em sala sobre compressão de dados. Para tanto, visa-se à implementação, tanto do compressor quanto do descompressor, de um dos métodos estudados (Huffman estático, Huffman dinâmico, LZ77, LZ78 ou LZW).

O trabalho pode ser feito em equipe, composta de no máximo três membros. Os membros de cada equipe devem ser informados até no máximo 18/06/2023, em formulário que será aberto no Ava Moodle. Caso contrário, assume-se que a equipe conterà apenas um membro.

O trabalho deve ser entregue, via Ava Moodle, até 11/07/2023. A entrega deverá conter os códigos fontes usados na implementação e um relatório. O compressor e o descompressor devem ser entregues como dois programas independentes. O relatório deve abordar o seguinte conteúdo:

- Justificativa para a escolha do método pela equipe, contendo um breve comparativo entre os conceitos usados pelos métodos
- Descrição do método escolhido e de sua implementação
- Análise experimental, contendo:
 - Taxa de compressão do método em relação à entropia. Para estimar a entropia, considerar a estatística dos símbolos considerando-os como sequência de 8 bits.
 - Taxa de compressão comparando implementações de métodos encontradas no mercado.
 - Variação do tipo de arquivo objeto de compressão (texto, código fonte, imagens etc).

As restrições para a entrega do trabalho com relação aos membros da equipe, possíveis detecção de plágio e linguagem de programação são similares ao primeiro trabalho.

Taxa de compressão

Exemplo:

41 41 41 41 41 41 62 5A 3B 47 72 56 56 56 56 56 56

Resultado: 85 41 04 62 5A 3B 47 72 86

Obs.: Taxa de compressão: 18:9 (redução de 50% aprox)

Entropia

1. Calcule a frequência de ocorrência de cada símbolo no resultado comprimido. Isso pode ser feito contando o número de vezes que cada símbolo aparece.

2. Normalize as frequências dividindo cada frequência pelo tamanho total do resultado comprimido. Isso fornecerá a probabilidade de ocorrência de cada símbolo.

3. Calcule a entropia utilizando a fórmula da entropia de Shannon:

$$H = - (p * \log_2(p))$$

onde H é a entropia e p é a probabilidade de ocorrência de cada símbolo.

4. Para cada símbolo no resultado comprimido, calcule $p * \log_2(p)$, some todos os valores e, em seguida, multiplique o resultado por -1 para obter a entropia.

A entropia resultante será medida em bits por símbolo ou bits por caractere e indicará a quantidade média de informação necessária para representar cada símbolo no resultado comprimido. Quanto menor a entropia, maior é a compressão e maior é a eficiência do algoritmo em reduzir o tamanho dos dados.