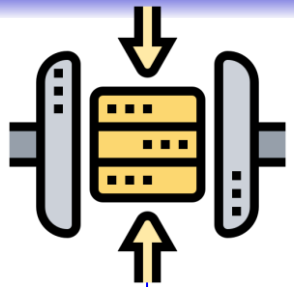
Algoritmos e Estruturas de Dados III

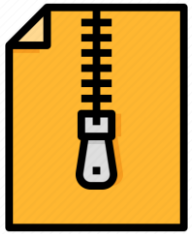
Aula 9 – Compressão de Dados

Prof. Felipe Lara

PUC Minas

Roteiro do Conteúdo





Compressão de Dados

- Introdução
- Classificação
- Simetria, perda e adaptabilidade
- Métricas

Tipos de Compressão de Dados

- Compressão com e sem perdas
- Redução de quantidade e tamanho de símbolos
- Codificação RLE
- Métodos Estatísticos
- Métodos de Dicionário

Compressão de Dados
Introdução

PUC Minas

Compressão de Dados - Introdução

Cristiano Ronaldo no deja de quemar registros en el campo y en los despachos. A sus 40 años, el portugués renovó su contrato con **Al-Nassr de Arabia Saudita,** lo que le coloca dentro del prestigioso equipo de los atletas más ricos del planeta.

Relacionados

Liga Saudí. Cristiano Ronaldo (40 años) no ve su final: "Me dicen que pare y que para qué quiero marcar 1.000 goles"

Tras más de dos décadas de prolífica carrera profesional militando en los principales clubes europeos **(Real Madrid, Manchester United o Juventus),** el capitán de la selección portuguesa ha apilado un palmarés deslumbrante además de lucrativos contratos de patrocinio con marcas como **Nike, Binance, Tag Heuer, Herbalife o Louis Vuitton.**

Compressão de Dados - Introdução

Objetivo

Codificar um conjunto de informações de modo que o código resultante seja **menor** que o original

Justificativa

Redução do espaço ocupado
Aumento da velocidade de transmissão dos dados

Compressão de Dados - Introdução

A	B	C	D	E	F	G	H	I	J	K	L	M
												

N	O	P	Q	R	S	T	U	V	W	X	Y	Z
												

Table 1.1: The 26 Braille Letters.

and	for	of	the	with	ch	gh	sh	th
								

Table 1.2: Some Words and Strings in Braille.

Compressão de Dados - Introdução

Conceito

Compressão é um processo de codificação de dados que busca **reduzir o número de bits** necessários para se representar uma informação.

A compressão de dados envolve dois processos:

- Em um deles os dados são comprimidos (codificados) para terem seu tamanho reduzido
- No segundo processo eles são descomprimidos (decodificados) para voltar ao formato original

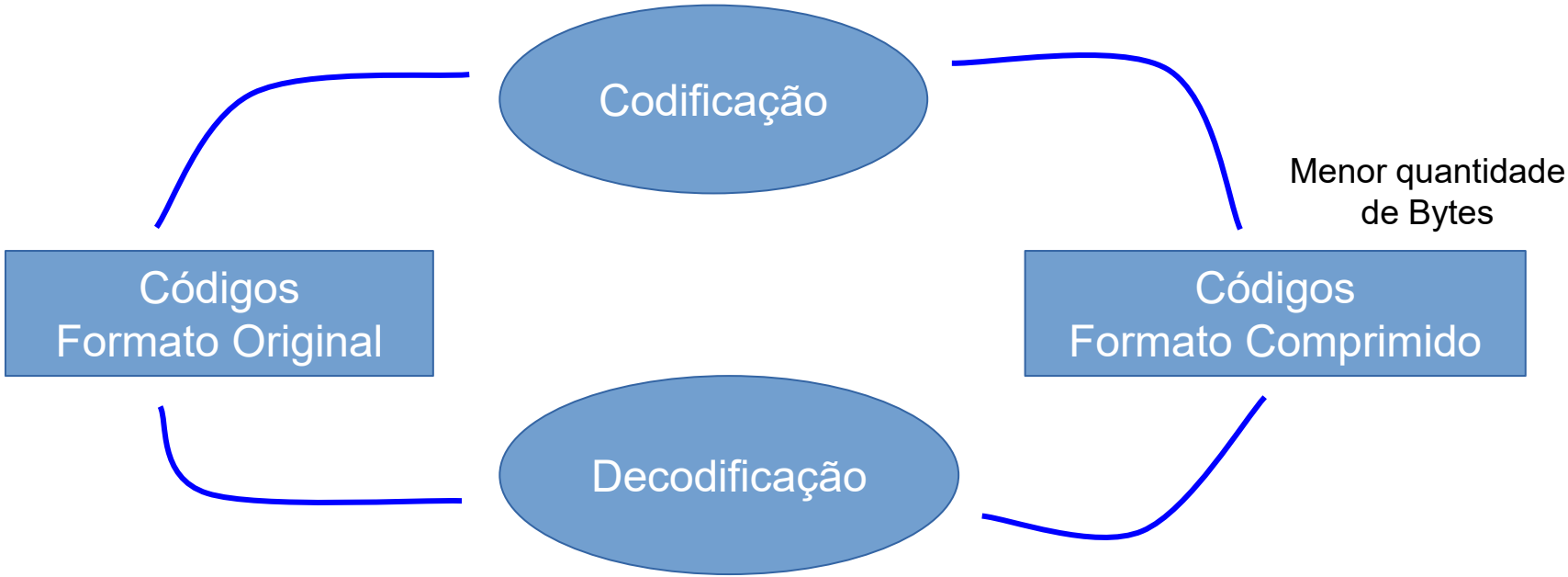
Compressão de Dados - Introdução

Problema 21.5: ***Compressão de Dados***

Entrada: $\langle X, \Sigma, f \rangle$, onde X é um arquivo com caracteres pertencentes a um alfabeto Σ e cada $i \in \Sigma$ possui uma frequência f_i de aparição.

Saída: Sequência de bits (código) para representar cada caractere de modo que o arquivo binário tenha tamanho mínimo.

Compressão de Dados - Introdução



Exemplo

"AAAAAAAAAABBBBBCCD"

(são 10 letras A, 5 letras B, 2 letras C e 1 letra D, totalizando 18 letras)

Se A = 00, B = 01, C = 10 e D = 11,

Qual a compactação resultante?

Quantos bytes serão usados?

Qual a economia?

Exemplo

"AAAAAAAAAABBBBBCCD"

(são 10 letras A, 5 letras B, 2 letras C e 1 letra D, totalizando 18 letras)

Se A = 0, B = 10, C = 110 e D = 111

Qual a compactação resultante?

Quantos bytes serão usados?

Qual a economia?

Exemplo

"AAAAAAAAAABBBBBCCD"

(são 10 letras A, 5 letras B, 2 letras C e 1 letra D, totalizando 18 letras)

Se A = 00, B = 01, C = 10 e D = 11, então a mensagem completa gastará 18*2 = 36 bits.

Se A = 0, B = 10, C = 110 e D = 111, então a mensagem completa gastará 10*1 + 5*2 + 2*3 + 1*3 = 29 bits.

Exemplo

"AAAAAAAAAABBBBBCCD"

(são 10 letras A, 5 letras B, 2 letras C e 1 letra D, totalizando 18 letras)

Se A = 00, B = 01, C = 10 e D = 11, então a mensagem completa gastará 18*2 = 36 bits.

Se A = 0, B = 10, C = 110 e D = 111, então a mensagem completa gastará 10*1 + 5*2 + 2*3 + 1*3 = 29 bits.

Exemplo

!	11
A	0
B	1
C	01
D	10
R	00

- Qual o texto da seguinte codificação: 01000010100100011?

Exemplo

!	101
A	0
B	1111
C	110
D	100
R	1110

- Qual o texto da seguinte codificação: 011111110011001000111111100101?

Exemplo

- Considere um arquivo com um alfabeto de tamanho 40.
- Faça o agrupamento de 3 caracteres.
- Quantos bits são necessário para representar cada tripleta?
- Houve compressão de dados?

Antes da Compactação

PUC Minas

Compressão de Dados - Introdução

Antes da Compressão – Racionalização da representação

- Eliminação de itens redundantes
- Uso de Notação Codificada
- Codificação de textos
- Supressão de espaços inúteis

Compressão de Dados - Introdução

Antes da Compressão – Racionalização da representação

- **Eliminação de itens redundantes**
- Uso de Notação Codificada
- Codificação de textos
- Supressão de espaços inúteis

Aluno {Nome, Matrícula, nota1, nota2, nota3, média}

=

Aluno {Nome, Matrícula, nota1, nota2, nota3}

Compressão de Dados - Introdução

Antes da Compressão – Racionalização da representação

- Eliminação de itens redundantes
- **Uso de Notação Codificada**
- Codificação de textos
- Supressão de espaços inúteis

17 de Abril de 2001 ⇒ 19 B
17/Abril/2001 ⇒ 13 B
17/04/01 ⇒ 3 B (Notação Binária)
00010001 00000100 00000001
Bits realmente utilizados : 10001 0100 0000001 ⇒ 2 B

Compressão de Dados - Introdução

Antes da Compressão – Racionalização da representação

- Eliminação de itens redundantes
- Uso de Notação Codificada
- **Codificação de textos**
- Supressão de espaços inúteis

Diminuição da ocorrência de textos em tabelas

Funcionário {CódF, Nome, **NomeDept**, DataNasc}
Equipamento {CódE, Nome, Desc, **NomeDept**}

Funcionário {CódF, Nome, **CodDept**, DataNasc}
Equipamento {CódE, Nome, Desc, **CodDept**}
Departamento {**CodDept**, **NomeDept**}

Compressão de Dados - Introdução

Antes da Compressão – Racionalização da representação

- Eliminação de itens redundantes
- Uso de Notação Codificada
- Codificação de textos
- **Supressão de espaços inúteis**

Campos com tam. fixo devem ser avaliados com muito cuidado.

Ex. Nome: 70 caracteres (70 bytes)
(José da Silva⇒ 13 bytes)

Classificação
da Compactação de Dados

PUC Minas

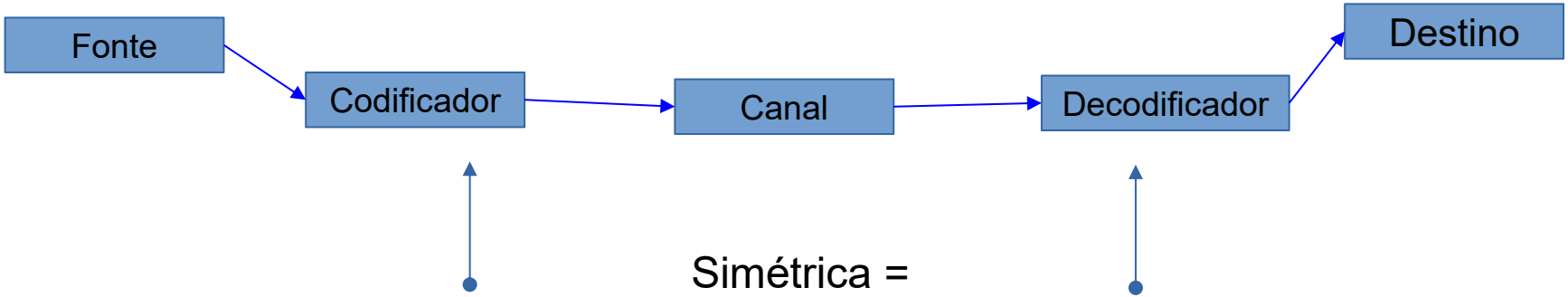
Compressão de Dados - Classificação

- Quanto a Simetria
- Quanto a Perda
- Quanto a Adaptabilidade

Compressão de Dados - Classificação

- **Quanto a Simetria**
- Quanto a Perda
- Quanto a Adaptabilidade

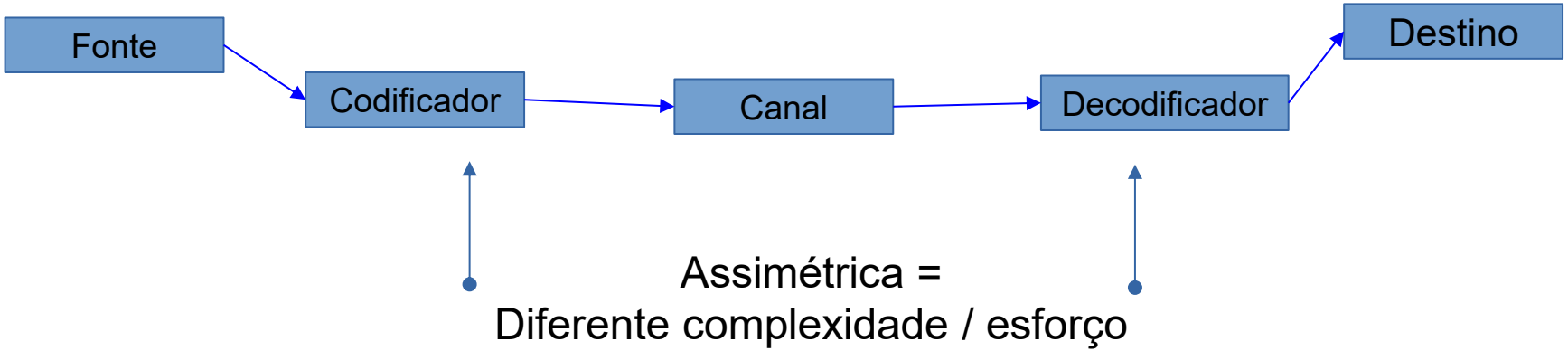
Procedimentos simétricos são indicados em aplicações que envolvem a transmissão e apresentação das imagens simultaneamente



Compressão de Dados - Classificação

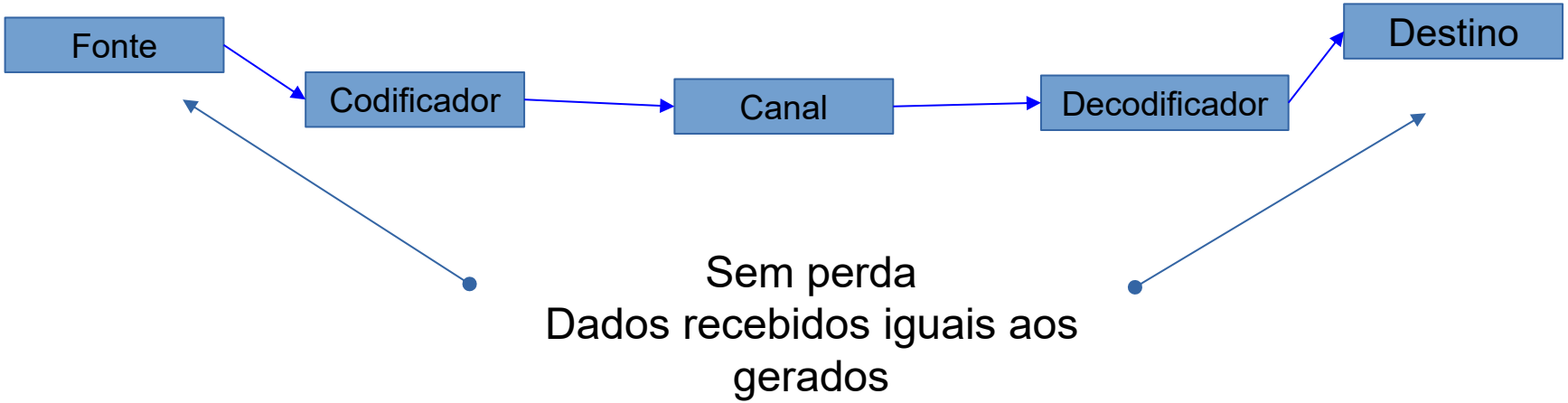
- **Quanto a Simetria**
- Quanto a Perda
- Quanto a Adaptabilidade

Procedimentos de backup tendem a privilegiar uma compressão mais rápida em detrimento do tempo necessário à descompressão



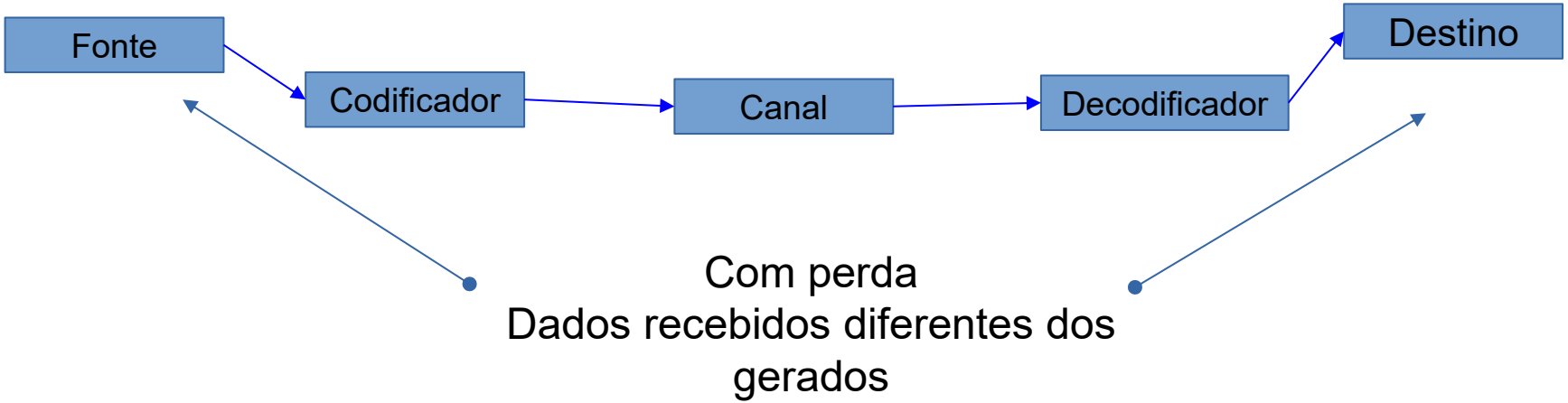
Compressão de Dados - Classificação

- Quanto a Simetria
- **Quanto a Perda**
- Quanto a Adaptabilidade



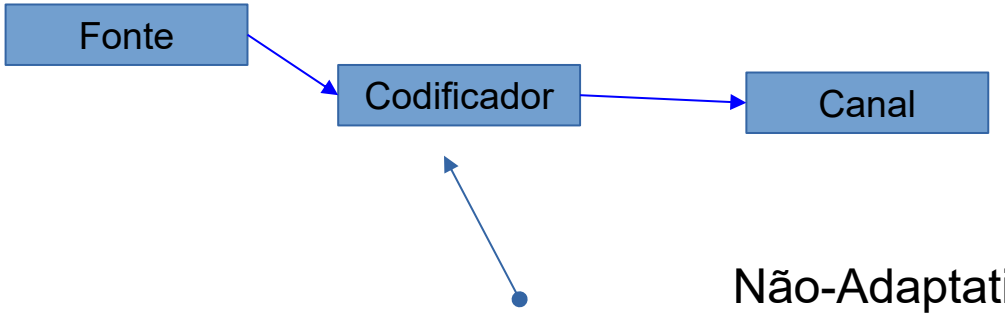
Compressão de Dados - Classificação

- Quanto a Simetria
- **Quanto a Perda**
- Quanto a Adaptabilidade



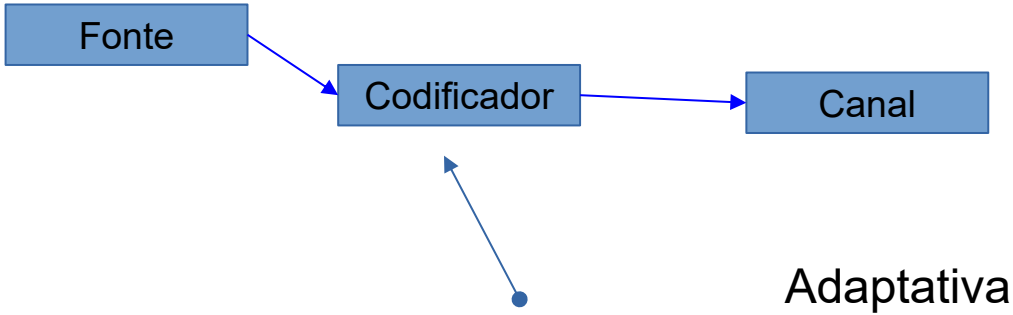
Compressão de Dados - Classificação

- Quanto a Simetria
- Quanto a Perda
- **Quanto a Adaptabilidade**



Compressão de Dados - Classificação

- Quanto a Simetria
- Quanto a Perda
- **Quanto a Adaptabilidade**



Compressão de Dados - Métricas

- Taxa de Compressão
- Fator de Compressão
- Percentual de Redução
- Taxa de Bits

Compressão de Dados - Métricas

- **Taxa de Compressão**
- Fator de Compressão
- Percentual de Redução
- Taxa de Bits

Tc = Tamanho Final ÷ Tamanho Inicial

Compressão de Dados - Métricas

- **Taxa de Compressão**
- Fator de Compressão
- Percentual de Redução
- Taxa de Bits

Tamanho original: 10.000 bytes
Tamanho comprimido: 2.500 bytes
Total de caracteres: 10.000
Tc = Tamanho Final ÷ Tamanho Inicial
Tc = 2.500/10.000 = 0.25

Compressão de Dados - Métricas

- Taxa de Compressão
- **Fator de Compressão**
- Percentual de Redução
- Taxa de Bits

Fc = Tamanho Inicial ÷ Tamanho Final

Compressão de Dados - Métricas

- Taxa de Compressão
- **Fator de Compressão**
- Percentual de Redução
- Taxa de Bits

Tamanho original: 10.000 bytes
Tamanho comprimido: 2.500 bytes
Total de caracteres: 10.000
Fc = Tamanho Inicial ÷ Tamanho Final
Fc = 10.000/2.500 = 4

Compressão de Dados - Métricas

- Taxa de Compressão
- Fator de Compressão
- **Percentual de Redução**
- Taxa de Bits

$$Pr = 100 \times (1 - T c)$$

Compressão de Dados - Métricas

- Taxa de Compressão
- Fator de Compressão
- **Percentual de Redução**
- Taxa de Bits

Tamanho original: 10.000 bytes
Tamanho comprimido: 2.500 bytes
Total de caracteres: 10.000
$Pr = 100 \times (1 - T c)$
$Pr = 100 \times (1 – 0.25) = 75\%$

Compressão de Dados - Métricas

- Taxa de Compressão
- Fator de Compressão
- Percentual de Redução
- **Taxa de Bits**

bpc ≡ bits por caracter
bpp ≡ bits por pixel

Compressão de Dados - Métricas

- Taxa de Compressão
- Fator de Compressão
- Percentual de Redução
- **Taxa de Bits**

Tamanho original: 10.000 bytes
Tamanho comprimido: 2.500 bytes × 8 bits = 20.000 bits
Total de caractéres: 10.000
bpc = Total de bits/Total de caractéres = 20.000/10.000 = 2bpc

Tipos de Compressão de Dados

PUC Minas

Compressão sem Perdas x Com Perdas

Sem Perdas

- Permite a recuperação exata dos dados originais após o processo de descompressão
- Remoção (recuperável) das redundâncias
- Aplicada na compressão de dados, textos, programas, imagens médicas ...
- Exemplos: ZIP

Com Perdas

- Informação obtida após a descompressão é diferente da original (antes da compressão)
- Informação suficientemente "parecida" para que seja de alguma forma útil.
- Eliminação de detalhes
- Aplicada na compressão de imagens, áudio, vídeo, …
- Exemplos: JPEG, MP3, MP4

Compressão - Quantidade e Tamanho dos Símbolos

Redução da Quantidade de Símbolos

- Um símbolo passa a representar um conjunto de outros símbolos
- Ex.:
  - Ao invés de indexarmos cada letra, indexamos palavras
  - Um pixel pode representar um conjunto de pixels

Redução do Tamanho dos Símbolos

- Um símbolo pode ser representado com menos bits do que o usual
- Ex.:
  - Podemos usar menos de 1 byte para representar uma letra
  - Um pixel pode usar menos de 3 bytes

Compressão com Perdas

- Transformada Discreta de Cosseno
- Desenvolvida em 1974 por N. Ahmed, T. Natarajan and K. R. Rao
- Uma das ferramentas mais usadas em processamento de imagens
- https://ieeexplore.ieee.org/document/1672377/

Compressão com Perdas

- Em plena Pandemia
- Com vídeo-chamadas em grande uso
- Uma homenagem aos que permitiram este feito

https://www.youtube.com/watch?v=W29r-zJtqcY

Exemplos e Métodos de Compressão de Dados

PUC Minas

Codificação RLE

- RLE = Run-length Encoding
- Compressão sem perda de dados
- Sequências longas de valores repetidos são armazenadas como um único valor e sua contagem no lugar de sua sequência original.
- Útil em dados com muitas repetições de valores
    - aaaaabbbbbbbbbbccccdddeeeeee
    - 5a10b4c3d6e

Métodos Estatísticos

- Utilizam códigos de comprimentos variáveis.
- Dados na informação original que aparecem com maior freqüência são representados por palavras-código menores
- Dados de menor incidência são representados por palavras-código maiores
- Ex: Shannon-Fano / Huffman

Métodos de Dicionário

- Os símbolos (ou conjunto de símbolos) são substituídos por códigos a partir de um “dicionário”
- Os códigos possuem tamanho fixo
- Os dicionários podem ser estáticos ou dinâmicos
- Ex: LZ77 / LZ78 / LZW