

Caio Magno Aguiar de Carvalho

**Estudo e Desenvolvimento de Algoritmos de
Compressão Sem Perda sobre Dados
Uniformemente Distribuídos**

São Luís, Maranhão, Brasil

08/2022

Caio Magno Aguiar de Carvalho

Estudo e Desenvolvimento de Algoritmos de Compressão Sem Perda sobre Dados Uniformemente Distribuídos

Tese de Doutorado submetida ao Programa de Pós-Graduação em Engenharia de Eletricidade da Universidade Federal do Maranhão como parte dos requisitos necessários à obtenção do grau de Doutor em Engenharia Elétrica.

Universidade Federal do Maranhão - UFMA

Centro de Ciências Exatas e Tecnologia

Programa de Pós-Graduação em Engenharia Elétrica

Orientador: Allan Kardec Barros Filho

São Luís, Maranhão, Brasil

08/2022

Ficha gerada por meio do SIGAA/Biblioteca com dados fornecidos pelo(a) autor(a).
Diretoria Integrada de Bibliotecas/UFMA

de Carvalho, Caio Magno Aguiar.

Estudo e Desenvolvimento de Algoritmos de Compressão
Sem Perda sobre Dados Uniformemente Distribuídos / Caio
Magno Aguiar de Carvalho. - 2022.

80 f.

Orientador(a): Allan Kardec Duailibe Barros Filho.

Tese (Doutorado) - Programa de Pós-graduação em
Engenharia Elétrica/ccet, Universidade Federal do
Maranhão, São Luís, Maranhão, 2022.

1. Compressão de dados. 2. Concatenação. 3. Dados
uniformes. 4. Teoria da informação. I. Barros Filho,
Allan Kardec Duailibe. II. Título.

Caio Magno Aguiar de Carvalho

Estudo e Desenvolvimento de Algoritmos de Compressão Sem Perda sobre Dados Uniformemente Distribuídos

Tese de Doutorado submetida ao Programa de Pós-Graduação em Engenharia de Eletricidade da Universidade Federal do Maranhão como parte dos requisitos necessários à obtenção do grau de Doutor em Engenharia Elétrica.

Trabalho aprovado. São Luís, Maranhão, Brasil, 16/09/2022:

Allan Kardec Barros Filho

Orientador

Ewaldo Éder Carvalho Santana

Convidado 1

Francisco da Chagas de Souza

Convidado 2

Hugo Valadares Siqueira

Convidado 3

São Luís, Maranhão, Brasil

08/2022

Dedico esse trabalho primeiramente ao Deus do Céu, que "dá sabedoria e ciência"(Dn 2.21), à minha esposa Aline Kessia e meu filho Raul que são o tempero e o consolo da minha vida, à minha mãe Ayla Aguiar e meu pai Hermes Alberto (in memorian), pela educação que não se pode aprender na escola.

Agradecimentos

Direciono meus agradecimentos primeiramente ao professor Allan Kardec, que me propôs tal desafio, por sua disposição e paixão pelo trabalho. Sua dedicação certamente será um ideal a ser alcançado em minha carreira.

Aos amigos do laboratório de Processamento da Informação Biológica (PIB): Jonathan Queiroz, Letícia Correia, Marta Barreiros, Felipe Gomes, George Vagner e Daniel Luna. Obrigado pelo apoio manifestado em críticas, contribuições, ideias e revisões.

*"Quem encerrou o mar com portas (...) e disse: Até aqui virás, e não mais
adiante(...)" Jó 38: 8, 11*

Resumo

A alta produção e consumo de informação digital em ritmo cada vez mais acelerado não acompanha as atuais ofertas de armazenamento e transmissão de dados, ou seja, produzimos mais conteúdo digital do que podemos armazenar e comunicar, e essa corrida aparentemente não será equilibrada facilmente. As técnicas de compressão de dados foram desenvolvidas afim de otimizar os mecanismos de armazenamento e comunicação de forma que a informação ocupe o mínimo de espaço em um sistema de gerenciamento de arquivos ou o mínimo de largura de banda em um canal de comunicação. Tais técnicas estão baseadas na Teoria da Informação proposta por Shannon, nas quais as estatísticas do sinal a ser comprimido desempenham papel fundamental na representação eficiente da informação. Repetição e estrutura são características fundamentalmente exploradas por algoritmos de compressão. Entretanto sequências de dados uniformemente distribuídos, independentes e identicamente distribuídos (i.i.d) rompem esses dois pilares que fundamentam a compressão estatística. É sabido também que idealmente a saída codificada de um algoritmo de compressão é uniformemente distribuída, portanto, estudar a possibilidade de compressão de distribuições uniformes é abrir a possibilidade de compressão recursiva. O presente trabalho tem como objetivo explorar essa possibilidade através da observação do problema da compressão fora do campo estatístico, mas a partir da redundância inerente da codificação binária padrão, proposta pelo algoritmo da Concatenação e da perspectiva geométrica através do método SVD-esfera-espiral. O algoritmo da Concatenação aproveita as frações de bits não utilizadas na representação binária padrão, tendo o seu desempenho máximo quando o tamanho do alfabeto dos dados comprimidos é $2^N + 1$. Os experimentos foram conduzidos sobre os dados da RAND Corporation, os quais são dados uniformes produzidos por processos físicos com alfabeto de tamanho 10. Os resultados mostraram que é possível obter até 12,5% de compressão sobre esse conjunto.

Palavras-chaves: compressão de dados. Dados uniformes. teoria da informação. concatenação.

Lista de ilustrações

Figura 1 – Quantidade de informação produzida e armazenada no período 1986-2007. Fonte: Hilbert e López (2011)	13
Figura 2 – Fluxo do processo de compressão e decompressão de um arquivo	15
Figura 3 – Dois níveis de compressão com perdas JPG: A imagem mais a esquerda é a imagem original. A imagem ao meio é a representação comprimida em nível mediano. A última imagem é a representação original comprimida ao máximo. Fonte: www.lifewire.com/the-effect-of-compression-on-photographs-493726	22
Figura 4 – Modelo de sistema de comunicação proposto por Shannon	23
Figura 5 – Exemplos de códigos unicamente decodificáveis e códigos instantâneos	26
Figura 6 – Exemplo de representação de um <i>prefix code</i> como uma árvore binária	27
Figura 7 – Arvore binária que ilustra o teorema de Kraft-McMillan	28
Figura 8 – Ilustração do procedimento da construção da árvore de Huffman supondo que o relacionamento entre os pesos seja $p_2 < p_3 < p_1 \lll p_4$	30
Figura 9 – Codificação de cada símbolo pela árvore de Huffman	30
Figura 10 – Exemplo de codificação	33
Figura 11 – Codificação completa de uma mensagem usando LZ77	34
Figura 12 – Partição representando a distribuição dos símbolos de \mathcal{A}	35
Figura 13 – Codificação da sequência $[s_1, s_2, \dots]$	36
Figura 14 – Codificação da sequência $[s_1, s_2, \dots]$ para $P(s_1) = 0.7$, $P(s_2) = 0.1$ e $P(s_3) = 0.2$	37
Figura 15 – Símbolo s_0 gerado por uma variável latente y	40
Figura 16 – Exemplo de rotação em três etapas a partir da rotação do eixo Z, seguida da rotação do eixo Y e por fim eixo X	46
Figura 17 – Vetor unitário localizado na esfera unitária através de suas coordenadas esféricas	47
Figura 18 – Projeção de uma espiral sobre uma esfera	48
Figura 19 – Cobertura da esfera em função do número de revoluções da espiral esférica.	49
Figura 20 – Vetor \mathbf{u} (azul) sendo aproximado por um vetor $\hat{\mathbf{u}}(t)$ (vermelho)	50
Figura 21 – Economia percentual (acima) e tempo de codificação em segundos (abaixo) para k variando entre os divisores de 10^6 de 1 a 100.	55
Figura 22 – Diferenças nas taxas de codificação em função do tamanho do alfabeto	57
Figura 23 – Comparação do tempo de codificação entre os algoritmos	58
Figura 24 – Superfície determinada pelos vetores \mathbf{u}	59
Figura 25 – Distribuição de frequência de s	60
Figura 26 – Relação entre s e t	60

Figura 27 – Exemplo de relação S-T	61
Figura 28 – Distribuição de frequência dos tamanhos dos conjuntos de parâmetros t associados a cada parâmetro s em vetores 2×1	61
Figura 29 – Distribuição de frequência de S para vetores 3×1 sem permutação entre si	63
Figura 30 – Distribuição de frequência dos tamanhos dos conjuntos de parâmetros t associados a cada parâmetro s em vetores 3×1 sem permutação	64
Figura 31 – Taxa de codificação das matrizes de permutação em função do número de concatenações	65
Figura 32 – Ilustração de como o algoritmo da Concatenação funciona num alfabeto de 10 símbolos	68
Figura 33 – Aumento da redundância em função do tamanho do alfabeto	69
Figura 34 – Número de códigos não utilizados em função do tamanho do alfabeto .	69
Figura 35 – Bits economizados pela operação de concatenação em função do fator k para um alfabeto de tamanho 10	70
Figura 36 – Evolução do ganho de bits em relação ao tamanho do alfabeto e o fator de concatenação. A barra de cor abaixo indica a escala utilizada para medir o ganho de bits.	70

Lista de tabelas

Tabela 1 – Comparação entre os algoritmos tomando Economia Percentual e Tempo de Codificação como critérios na compressão do arquivo RAND	56
---	----

Sumário

1	INTRODUÇÃO	13
	Introdução	13
1.1	Motivação	17
1.2	Objetivos	18
1.2.1	Objetivo Geral	18
1.2.2	Objetivos Específicos	18
2	REFERENCIAL TEÓRICO	20
	Referencial Teórico	20
2.1	Definições de compressão de dados	20
2.2	Métricas de avaliação de compressão	20
2.3	Compressão com perdas e Compressão sem perdas	22
2.4	Informação	23
2.5	Codificação	26
2.5.1	A Desigualdade de Kraft-McMillan	27
2.6	Algoritmos de Codificação	28
2.6.1	Código de Huffman	29
2.6.2	Lempel-Ziv	32
2.6.3	Codificação aritmética	34
2.6.4	<i>Assymetrical Numeral System - ANS</i>	38
2.7	<i>Bits Back Coding</i>	39
2.8	Considerações Finais	40
3	METODOLOGIA PROPOSTA	42
	Metodologia Proposta	42
3.1	Método da Concatenação	42
3.2	Método Esfera-Espiral	45
3.3	Método SVD-Esfera-Espiral	45
3.3.1	Transformação Esfera-Espiral	47
3.4	Outras abordagens para a Esfera-Espiral	50
3.4.1	Vetores de ordem 3	51
3.4.2	Vetores de ordem 2	51
3.5	Conclusão	52