Introdução à Teoria da Informação

Período 2016.2

Professor: Leonardo

Trabalho 1 (16/03/2017)

Marcello Marques de Oliveira (11400986)

Uma comparação de Compressores-descompressores.

Os compressores-descompressores analisados foram:

1. Huffman não adpatativo, não contextual;
2. Huffman adaptativo, não contextual de decremento 1;
3. Huffman adaptativo, não contextual de decremento 1, com sqrt(N) etapas;
4. Huffman semi-adaptativo contextual de ordem 1;
5. WinRar.

Implementação:

Linguagem Utilizada: C++.

Para implementar o algoritmo de geração de árvores e de código de Huffman foram usados algumas propriedades de orientação a objeto, o uso das classes SNode (Simple Node) e LNode (Leaf Node) com herança direta de PNode (Parent Node), o polimorfismo garantido por essa herança para sobrecarregar o operador () e assim utilizar a implementação priority-queue da STL (min-heap based), foram facilitadores e minimizadores de linhas de código deixando a implementação portável, legível e de fácil debug.

O reuso de código também foi um expoente no que se refere a implementação realizada, após a criação do primeiro Huffman, os demais levaram 90% do código em média.

Os operadores new e delete foram utilizados em todos os momentos para garantir proveito máximo da memória principal. Em especial o operador new e a preferência por alocação dinâmica garantem a compressão teorica de dados de tamanho máximo possível reservado para o programa.

O tamanho do alfabeto utilizado nos algoritmos 1, 2, 3 e 4 possui 256 símbolos.

Mais detalhes sobre a implementação podem ser conferidos em: *github.com/marcello-oliveira/skynet/info-theory*.

Testes:

Os testes foram realizados com os arquivos do corpus Silesia disponível em: *sun.aei.polsl.pl/~sdeor/index.php?page=silesia*. No teste foram medidos (sempre que possível): razão de compressão, comprimento médio (bits/símbolo), entropia, tempo de compressão e de decompressão.

Para verificar a igualdade entre os arquivos (original e descomprimido) foi utilizado o programa diff do Ubuntu 15.04.

Informação retirada:

Arquivo de teste: dickens

(C/D) / Medida	R. Compress.	Comp. Médio	Entropia	T. Compress.	T. Descompre.
H NA NC	0.57	4.57 b/char	4.53 b/char	3550 ms	6725 ms
H A NC DE1	*	*	*	*	*
H A NC DE1 S	0.57	4.44 b/char	4.53 b/char	4302 ms	24820 ms
H SA CO O1	0.44	3.56 b/char	3.50 b/char	623 ms	5428 ms
WINRAR	0.30	*	*	794 ms	3000 ms

Arquivo de teste: mozilla

(C/D) / Medida	R. Compress.	Comp. Médio	Entropia	T. Compress.	T. Descompre.
H NA NC	0.78	6.24 b/char	6.22 b/char	19890 ms	151041 ms
H A NC DE1	*	*	*	*	*
H A NC DE1 S	0.78	6.24 b/char	6.22 b/char	26091 ms	179231 ms
H SA CO O1	0.61	4.74 b/char	4.68 b/char	5245 ms	51362 ms
WINRAR	0.30	*	*	2569 ms	644 ms

Arquivo de teste: mr

(C/D) / Medida	R. Compress.	Comp. Médio	Entropia	T. Compress.	T. Descompre.
H NA NC	0.46	3.70 b/char	3.68 b/char	4007 ms	17583 ms
H A NC DE1	*	*	*	*	*
H A NC DE1 S	0.46	3.69 b/char	3.68 b/char	5927 ms	24479 ms
H SA CO O1	0.42	3.40 b/char	3.21 b/char	477 ms	4905 ms
WINRAR	0.28	*	*	400 ms	277 ms

Arquivo de teste: nci

(C/D) / Medida	R. Compress.	Comp. Médio	Entropia	T. Compress.	T. Descompre.
H NA NC	0.30	2.43 b/char	2.42 b/char	11643 ms	17221 ms
H A NC DE1	*	*	*	*	*
H A NC DE1 S	0.30	2.43 b/char	2.42 b/char	11617 ms	42339 ms
H SA CO O1	0.25	2.03 b/char	1.94 b/char	1156 ms	7744 ms
WINRAR	0.06	*	*	825 ms	325 ms

Arquivo de teste: ooffice

(C/D) / Medida	R. Compress.	Comp. Médio	Entropia	T. Compress.	T. Descompre.
H NA NC	0.83	6.66 b/char	6.63 b/char	2503 ms	19607 ms
H A NC DE1	*	*	*	*	*
H A NC DE1 S	0.83	6.65 b/char	6.63 b/char	4146 ms	24630 ms
H SA CO O1	0.78	5.09 b/char	5.04 b/char	891 ms	6457 ms
WINRAR	0.37	*	*	363 ms	247 ms

Arquivo de teste: osdb

(C/D) / Medida	R. Compress.	Comp. Médio	Entropia	T. Compress.	T. Descompre.
H NA NC	0.82	6.61 b/char	6.59 b/char	4179 ms	32006 ms
H A NC DE1	*	*	*	*	*
H A NC DE1 S	0.82	6.61 b/char	6.59 b/char	4167 ms	37833 ms
H SA CO O1	0.73	5.13 b/char	5.08 b/char	1604 ms	11348 ms
WINRAR	0.33	*	*	510 ms	336 ms

Arquivo de teste: reymont

(C/D) / Medida	R. Compress.	Comp. Médio	Entropia	T. Compress.	T. Descompre.
H NA NC	0.60	4.86 b/char	4.84 b/char	2393 ms	14307 ms
H A NC DE1	*	*	*	*	*
H A NC DE1 S	0.60	4.86 b/char	4.84 b/char	4167 ms	19360 ms
H SA CO O1	0.41	2.81 b/char	2.71 b/char	470 ms	2899 ms
WINRAR	0.24	*	*	351 ms	238 ms

Arquivo de teste: samba

(C/D) / Medida	R. Compress.	Comp. Médio	Entropia	T. Compress.	T. Descompre.
H NA NC	0.76	6.12 b/char	6.09 b/char	8146 ms	61194 ms
H A NC DE1	*	*	*	*	*
H A NC DE1 S	0.76	6.12 b/char	6.09 b/char	11725 ms	71980 ms
H SA CO O1	0.58	4.30 b/char	4.21 b/char	2027 ms	17662 ms
WINRAR	0.19	*	*	708 ms	450 ms

Arquivo de teste: sao

(C/D) / Medida	R. Compress.	Comp. Médio	Entropia	T. Compress.	T. Descompre.
H NA NC	0.57	7.54 b/char	7.52 b/char	3222 ms	27045 ms
H A NC DE1	*	*	*	*	*
H A NC DE1 S	0.57	7.54 b/char	7.52 b/char	4886 ms	32112 ms
H SA CO O1	0.85	5.87 b/char	5.83 b/char	1516 ms	10745 ms
WINRAR	0.75	*	*	470 ms	280 ms

Arquivo de teste: webster

(C/D) / Medida	R. Compress.	Comp. Médio	Entropia	T. Compress.	T. Descompre.
H NA NC	0.93	5.00 b/char	4.97 b/char	15417 ms	53787 ms
H A NC DE1	*	*	*	*	*
H A NC DE1 S	0.93	5.00 b/char	4.97 b/char	17049 ms	105377 ms
H SA CO O1	0.43	3.48 b/char	3.41 b/char	2398 ms	22414 ms
WINRAR	0.23	*	*	2493 ms	468 ms

Arquivo de teste: xml

(C/D) / Medida	R. Compress.	Comp. Médio	Entropia	T. Compress.	T. Descompre.
H NA NC	0.69	5.55 b/char	5.51 b/char	2073 ms	13195 ms
H A NC DE1	*	*	*	*	*
H A NC DE1 S	0.69	5.55 b/char	5.51 b/char	2593 ms	14758 ms
H SA CO O1	0.45	3.53 b/char	3.48 b/char	304 ms	2886 ms
WINRAR	0.09	*	*	200 ms	100 ms

Arquivo de teste: x-ray

(C/D) / Medida	R. Compress.	Comp. Médio	Entropia	T. Compress.	T. Descompre.
H NA NC	0.82	6.62 b/char	6.60 b/char	3369 ms	27533 ms
H A NC DE1	*	*	*	*	*
H A NC DE1 S	0.82	6.62 b/char	6.60 b/char	5518 ms	31681 ms
H SA CO O1	0.71	5.61 b/char	6.57 b/char	766 ms	8120 ms
WINRAR	0.48	*	*	454 ms	311 ms

O huffman adapativo não-contextual de decremento 1 revelou-se incapaz de comprimir/descomprimir arquivos com mais de 500kB em tempo hábil.

Outros Testes:

No github apresentado na introdução há disponível corner-cases utilizados para validar o código.

Conclusão:

No huffman não-adaptativo não-contextual percebemos uma forte correlação entre a entropia e o comprimento médio. Também há forte correlação entre tempo de compressão e tempo de descompressão.

No huffman adaptativo não-contextual percebemos as mesmas correlações que o huffman não-adaptativo não-contextual, e valores de entropia, comprimento médio e razão de compressão iguais. No geral tomou mais tempo que o huffman anterior para realizar a mesma compressão. É preferível o huffman anterior por: simplicidade e tempo.

No huffman semi-adaptativo contenxtual de ordem 1 foi observada a mesma correlação e um tempo muito menor além de uma razão de compressão alta para a maioria dos casos. Um detalhe da implementação: a opção para o 1º símbolo foi de joga-lo direto para saída e faze-lo de contexto inicial.

No Rar/Unrar há correlações entre as variáveis mencionadas porém não com a mesma força. A razão de compressão foi menor que nos outros Algoritmos e os tempos, principalmente o de descompressão, também foram menores.