Algorytmy tekstowe

laboratorium 3 - kompresja tekstu

Algorytmy statycznego i dynamicznego kodowania Huffmana

Jakub Pinowski

1 Format plików wyjściowych

1.1 Statyczne kodowanie Huffmana

Plik rozpoczyna się danymi potrzebnymi do odkodowania i poprawnego przeczytania tekstu:

- 8 bitów na ilość kodów (N) jakie znajdziemy w tekście
- 8 bitów na ilość bitów na ilu zapisany jest najdłuższy kod (L_{MAX})
- 3 bity na ilość bitów, które musimy "uciąć" z końca tekstu, a zostały dopiane ze względu na bajtową reprezentację pliku na dysku
- N razy fragment w postaci *8 bitów na kod ASCII* + $\lceil log_2(L_{MAX}) \rceil$ *bitów na długość kodu L* + *L bitów na kod Huffmana*

Kolejne bity zawierają zakodowaną treść

1.2 Dynamiczne kodowanie Huffmana

Plik składa się wyłącznie z kodów ASCII i kodów z drzewa Huffmana. Przeglądanie pliku polega na sprawdzaniu kolejnych znalezionych kodów i aktualizowaniu drzewa oraz dodawaniu liter do tekstu

- 3 bity na ilość bitów, które musimy "uciąć" z końca tekstu, a zostały dopiane ze względu na bajtową reprezentację pliku na dysku
- Na początku znajduje się 8 bitów na kod ASCII pierwszej litery tekstu
- Jeżeli znaleziony kod jest kodem znaku specjalnego, kolejne 8 bitów jest kodem ASCII litery, która wcześniej nie wystąpiła

2 Porównanie czasów działania

Nazwa	Rozmiar	Statyczny		Dynamiczny	
		Kompresja	Dekompresja	Kompresja	Dekompresja
test1.txt	1.03 kB	0.003s	0.001s	0.017s	0.007s
test2.txt	10.27 kB	0.002s	0.007s	0.061s	0.051s
test3.txt	100.48 kB	0.018s	0.069s	0.584s	0.498s
testHp.txt	863.36 kB	0.149s	0.561s	5.268s	4.390s

Algorytm dynamicznego kodowania mocno traci w przypadku większych plików, jest to spowodowane potrzebą aktualizacji drzewa po każdej literze, co wymaga znalezienia węzła w drzewie z którym mógłby zamienić się węzeł aktualizowany. Dla dużego drzewa może to wymagać sporej ilości czasu.

Przy okazji mierzenia czasów, program sprawdza zgodność odkodowanego tekstu z zawartością oryginalnego pliku.

3 Stopień kompresji tekstu

Plik test1.txt 1.03 kB

- Static 39.2%
- Adaptive 43.1%

Plik test2.txt 10.27 kB

- Static 45.7%
- Adaptive 46.1%

Plik test3.txt 100.48 kB

- Static 46.4%
- Adaptive 46.4%

Plik testHp.txt 863.36 kB

- Static 42.7%
- Adaptive 42.7%

Widzimy że niezależnie od wariantu kodowania Huffmana i wielkości pliku, stopień kompresji jest dość podobny i oscyluje około wartości 40-45%. Czym większy plik, tym bliższe sobie są stopnie kompresji.