Bezstratne algorytmy kompresji

Złożoność obliczeniowa algorytmow

Mateusz Kojro

January 16, 2022

Kompresja bezstartna - Definicja

Algorytmy kompresji bezstratnej - klasa algorytmów przekształcających ciąg danych w inny - krótszy ciąg danych z ktorego możliwe jest nastepnie odczytanie początkowych danych bez żadnych zmian

Kompresja bezstartna - Ograniczenia

- **Uniwersalna bezstratna kompresja nie istnieje** - korzystając z tzw. Szufladkowej zasady Dirichleta w bardzo prosty sposób udowodnić można ze nie istnieje algorytm pozwalających na uniwersalne kompresowanie dowolnego ciągu danych. Bezstratne algorytmy danych kompresują zawsze tylko określony rodzaj danych (dane innego rodzaju po zastosowaniu danego algorytmu powiekszaja swój rozmiar).
- **Dane losowe są niekompresowalne** - Kompletnie losowe dane nie dają sie kompresować (W teorii złożoności Kołmogorowa właśnie nie kompresowalne dane definiują losowość)

Metody pomiarów - Badane wartości

- Czas kompresji ozn. T_k
- Czas dekompresji ozn. T_d
- Stosunek wielkości skompresowanego pliku do oryginalnego ozn. C_r

Metody pomiarów - Rodzaje danych

Wszystkie algorytmy przetestowano na poniższych danych:

- Tekst po angielsku
- Zdjęcia PNG
- Pliki audio w formacie MP3
- Dane wytworzone przez losowe maszyny Turinga - `http://mattmahoney.net/dc/uiq/`
- Kolejne cyfry π (Jako łatwo replikowalne źródło liczb losowych)

Dla każdego rodzaju testowano wycinki o rozmiarach 100b, 500b, 1Kb, 50Kb, 100Kb

Run-length encoding (RLE)

RLE - Zasada dzialania

Algorytm RLE opiera się na zastępowaniu ciągu powtarzających się symboli alfabetu przez symbol oraz ilość jego powturzeń.
Na przykład: Mając dany ciąg znaków:

$$AAAAABBBBBCCCDDE$$

korzystając z najbardziej naiwnej odmiany algorytmu RLE moglibyśmy zapisać go jako:

$$5A5B3C2D1E$$

Dekompresja ciągu jest trywialna.

RLE - Podsumowanie

- Asymptotyczna złożoność czasowa: $O(n)$
- Osiąga dobre wyniki przy niektórych obrazach kompresji obrazów bardzo słabe natomiast podczas kompresji tekstu
- Stosowana jako jedna z metod w plikach JPEG

Kompresja LZW (Lempel-Ziv-Welch)

LZW - Zasada dzialania

Kompresowanie słowa alfabetu z wykorzystaniem kompresji LZW opiera sie na rozszerzaniu początkowego alfabetu (standardowo 256 znaków kodu ASCII) o dodatkowe symbole składające sie z kombinacji tych znaków. Jego zaleta jest fakt ze w celu dekompresji nie potrzebne jest przekazywanie żadnych dodatkowych danych oprócz ustalonego początkowego słownika.

LZW - Podsumowanie

- Stosowana jako domyślna kompresja w plikach PDF, GIF
- Domyślny algorytm kompresji w Unixowej funkcji compress
- W profesjonalnych zastosowaniach stosowana w połączeniu z np kodowaniem Huffmana
- Rozwinięcie starszego algorytmu LZ78

Kodowanie Huffmana

Kodowanie Huffmana - Zasada dzialania

1. Elementy sortowane są ze względu na częstość występowania
2. Tworzone jest drzewo binarne wszystkich symboli wykorzystanych w kompresowanym słowie
3. Słowo kodowane jest następnie jako indeksy wierzchołków drzewa, które jest przekazywane łącznie z zakodowanym słowem

Kodowanie Huffmana - Podsumowanie

- Jedna z najprostszych metod bezstratnej kompresji danych.
- Udowodniona jako najbardziej optymalny algorytm kompresji bezstratnej (Podczas kompresji dzielącej strumień na pojedyncze symbole)
- Istnieje rozszerzenie do przechowywania w drzewie słów zamiast symboli
- W profesjonalnych zastosowaniach stosowana w połączeniu z np LZW
- Asymptotyczna złożoność: $O(n \log n)$

Biblioteka Zlib

Porównanie z Zlib

W celu porównania prostych implementowanych algorytmów z de facto standardową biblioteka używana do kompresji danych (stosowana między innymi w kernelu linux, plikach PNG czy w serwerach Apache) dane testowe zostaną przekazane do biblioteki Zlib - `https://zlib.net/`