

Kodowanie i Kompresja Danych

Rafał Włodarczyk

INA 5, 2025

Contents

1	Literatura	2
2	Wykład I	2
2.1	Kodowanie	2
2.2	Kompresja	2
2.3	Miara Informacji	3
2.4	Entropia	3
2.5	Nierówność Krafta-McMillana	3
2.6	Kodowanie Shannon-Fano	4
3	Wykład II	4
3.1	Kodowanie Huffmana	4
3.2	Kodowanie dynamiczne	4

1 Literatura

1. K. Sayood, Kompresja danych - wprowadzenie - READ ME 2002 (ISBN 83-7243-094-2)
2. A. Przelaskowski, Kompresja danych - BTC 2005 (ISBN: 83-60233-05-5)
3. J. Adamek, Foundations of Coding - Wiley 1991 (ISBN 0-47-162187-0)
4. D. Salomon, G. Motta, Handbook of Data Compression, Springer-Verlag London 2010 (ISBN: 978-1-84882-903-9)

2 Wykład I

2.1 Kodowanie

Definition. Kodowanie. Przyporządkowanie elementom jakiegoś alfabetu ciągów binarnych (lub ciągów nad innym alfabetem).

Information. Początki ASCII. ASCII ma 127 znaków, ponieważ w dawnych czasach przesłanie ośmiu bitów. wiązało się z co najmniej jednym błędem - jeden bit został pozostawiony na parzystość.

2.2 Kompresja

Definition. Kompresja Stratna. Algorytm dopuszcza pewien poziom utraty informacji w zamian za lepszy współczynnik kompresji. Przykład - odcięcie małych składowych z kompresji SVD.

Definition. Kompresja Bezstratna. Z postaci skompresowanej zawsze można odtworzyć postać identyczną z oryginałem.

Definition. Kody stałej długości. Kody stałej długości np. kody ASCII o długości 8 bitów. Znamy podział na znaki

Definition. Kody zmiennej długości. Kody o różnej długości - Kody Prefixowe, Morse. Kod Prefixowy daje się odczytać tylko od końca.

Fact. Miary jakości kompresji. 1. Stosunek liczby bitów danych do bitów po kompresji.

2. Procentowy stosunek bitów po kompresji do bitów danych.

3. Czas kompresji i dekompresji - w zależności od sposobu użycia może być ważny.

Pierwsze aparaty cyfrowe robiły dużo drgań. Każda klatka była osobnym JPG-iem, pojawiały się zniekształcenia. Telefony nie kompresują dobrze, ponieważ nie są tak obliczeniowo wydajne, jak komputery.

Fact. Metody kompresji. • Metody różnicowe.

- Statystyczne właściwości danych.

W języku polskim nie możemy zejść poniżej lekko powyżej 5 bitów, w języku angielskim lekko powyżej 4 bitów.

Fact. Granice Kompresji. Kompresja bezstratna jest funkcją różnowartościową, różnych danych długości n jest 2^n (ciągi bitów). Różnych ciągów bitowych długości mniejszej niż n jest $2^n - 1$, $(\sum_{i=1}^{n-1} 2^i)$

- Istnieje ciąg, który nie może być skompresowany.
- Ponad połowa ciągów skróci się o co najwyżej 1-bit.

2.3 Miara Informacji

Definition. Teoria Informacji. Jeśli $P(A)$ jest prawdopodobieństwem wystąpienia informacji A to miara informacji:

$$i(A) = \log \frac{1}{P(A)} = -\log P(A)$$

Jeśli A i B są niezależne, to:

$$i(AB) = \log \frac{1}{P(AB)} = i(A) + i(B)$$

2.4 Entropia

Definition. Entropia. Średnia informacja w zbiorze wiadomości $A_1, \dots, A_n$, w którym prawdopodobieństwo wiadomości A_i wynosi $P(A_i)$ jest definiowana poprzez:

$$H = \sum_{i=1}^n P(A_i) i(A_i)$$

Minimalna entropia wynosi 0, a maksymalna dla 256 znaków wynosi 8. Wniosek. Kody jednoznacznie dekodowalne w modelu z niezależnymi wystąpieniami symboli muszą mieć średnią długość co najmniej równą entropii.

Fact. Test na jednoznaczność dekodowalności. Tworzymy listę słów kodowych. Dla każdej pary sprawdzamy czy jedno słowo jest prefiksem drugiego. Jeśli tak to do listy dodajemy sufiks drugiego słowa (chyba, że już dodaliśmy taki sufiks). Powtarzamy powyższą procedurę, aż do momentu kiedy znajdziemy na liście sufiks równy słowu kodowemu (kod nie jest jednoznaczny), albo nie można znaleźć nowych sufiksów (kod jest jednoznaczny)

2.5 Nierówność Krafta-McMillana

Theorem. Nierówność Krafta-McMillana. Niech C będzie kodem składającym się z N słów o długościach $l_1, l_2, \dots, l_n$. Jeżeli C jest jednoznacznie dekodowalny, to

$$K(C) = \sum_{i=1}^n \frac{1}{2^{l_i}} \leq 1$$

2.6 Kodowanie Shannon-Fano

Definition. Kodowanie Shannon-Fano. Algorytm kodowania, który przypisuje krótsze kody częściej występującym symbolom, a dłuższe rzadziej występującym. Działa poprzez:

1. Posortowanie symboli według malejącego prawdopodobieństwa.
2. Podział listy na dwie części o możliwie równych sumach prawdopodobieństw.
3. Przypisanie każdej części odpowiednio bitu 0 lub 1.
4. Powtarzanie procesu rekurencyjnie dla każdej części, aż do uzyskania jednoznacznych kodów dla wszystkich symboli.

Kody Shannon-Fano są zawsze jednoznacznie dekodowalne, ale nie zawsze optymalne.

3 Wykład II

3.1 Kodowanie Huffmana

Kody Huffmana to metody historycznie najstarsze, oraz często wykorzystywane na końcu procesu kompresji.

Information. Obserwacje Huffmana. Huffman wpadł na następujące pomysły:

- Dwa najrzadziej występujące symbole powinny mieć tę samą długość, ponieważ w przeciwnym przypadku dłuższy miałby redundantny bit.
- Częstsze symbole powinny mieć krótsze słowa kodowe.

Definition. Kody Huffmana. Kody możemy dostać różne, ale średnia długość kodu dla zadanych danych jest stała. Długości kodów mogą być różne. Algorytm jest optymalny, jeśli średnia długość kodu jest najmniejsza.

Aby szybko wyznaczyć średnią długość kodu można dodać wszystkie prawdopodobieństwa węzłów środkowych.

3.2 Kodowanie dynamiczne