PROSTE ALGORYTMY KOMPRESJI BEZSTRATNEJ

Wprowadzenie • Algorytm ByteRun • ByteRun - przykład • Algorytm RLE • Przykład działania RLE • Algorytm LZW • Przykład kompresji LZW

Wprowadzenie do prostych algorytmów kompresji bezstratnej

Przez algorytmy proste, będziemy rozumieli takie, które nie dokonują analizy strumienia danych przed kompresją. Dzięki temu algorytmy te są po pierwsze proste, po drugie mają minimalne wymagania pamięciowe (możliwa jest np. bezproblemowa implementacja sprzętowa), ponieważ nie wymagają bufora danych. Oczywistą wadą tych algorytmów jest mniejszy stopień kompresji, w porównaniu do algorytmów analizujących. Przykładami takich prostych algorytmów są **ByteRun**, zastosowany między innymi w formacie zapisu obrazów IFF ILBM w 1985 roku, oraz **RunLengthEncoding (RLE)** zastosowany w formacie BMP. Oba te proste algorytmy potrafią jedynie skompresować powtarzające się ciągi jednakowych znaków, stąd ich zastosowanie ogranicza się w praktyce do kompresji obrazów o niewielkiej (<= 256) ilości kolorów.

Nieco bardzie zaawansowanym algorytmem jest **LZW**, nazwany tak od nazwisk twórców, **Lempela** i **Ziva**, którzy opracowali podstawy teoretyczne, oraz **Welsha**, który w 1984 roku opisał implementację, sprzętową zresztą. Algorytm LZW potrafi wykorzystać nie tylko powtarzalność pojedynczych znaków, ale również ich różnych zbitek, dzięki budowie tzw. słownika w czasie pracy algorytmu. Dzięki temu szczególnie dobrze LZW spisuje się przy kompresji tekstu, a zwłaszcza tekstów takich jak np. kody źródłowe programów. Co ciekawe słownik kodera nie musi być przesyłany wraz ze skompresowanymi danymi, bowiem dekoder jest w stanie go odtworzyć w czasie dekompresji. To dodatkowo podnosi efektywność metody. Algorytm LZW znalazł szerokie zastosowanie w praktyce:

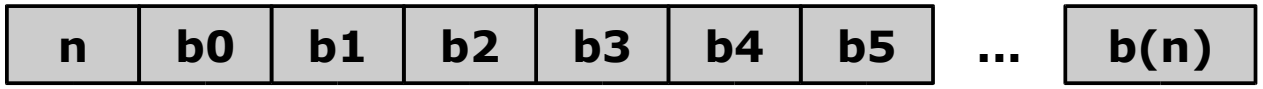
- Format zapisu obrazów GIF,
- Formaty zapisu dokumentów PostScript i PDF (kompresja bitmap),
- Programy do kompresji plików (ARC, compress).
- Transmisja modemowa (V.42bis, zmodyfikowany LZW).

Warto zaznaczyć, że przez dłuższy czas algorytm LZW był objęty patentem, co ograniczało jego stosowanie (patent zgłosiła firma Unisys, autor formatu GIF). Obecnie patent ten już nie obowiązuje.

Algorytm ByteRun

Algorytm ten daje kompresję tylko i wyłącznie w obecności powtarzających się symboli. Z założenia algorytm ten operuje na symbolach o rozmiarze bajtu. W algorytmie tym mamy do czynienia z dwoma sekwencjami:

sekwencja kopiowania bajtów



Pierwszy bajt, to liczba *n* z zakresu <0, 127>, określa ona ilość następujących *n*+1 bajtów, które mają być przy dekompresji skopiowane bez zmian.

sekwencja powtarzania bajtów

Tym razem *n* jest z przedziału <-1, -127>, co łatwo poznać po ustawionym najstarszym bicie. Bajt *b* jest przy dekompresji powtarzany (-*n*+1) razy, co pozwala na liczbę powtórzeń od 2 do 128. Pozostała, nieprzydzielona wartość *n* = -128, jest traktowana jako kod pusty (*no-op*).

Warto zauważyć że dla wielu strumieni danych kompresja ByteRun może dać zysk ujemny, na przykład w przypadku sekwencji ABABABABAB... „skompresowany” strumień będzie o 1 bajt dłuższy niż oryginał. Algorytm ByteRun stosowany był do kompresji obrazków w formacie IFF ILBM.

Przykład działania algorytmu ByteRun

Oto linia czterokolorowego obrazka, którą poddamy kompresji ByteRun:

Nadajmy numery kolorom z palety: **0 = niebieski**, **1 = czerwony**, **2 = zielony**, **3 = szary**. Pierwsze powtórzenie 3 niebieskich pikseli to oczywiście sekwencja [-2, 0], następnie 4 czerwone to [-3, 1]. Następnie mamy sytuację szczególną, mianowicie dwa piksele tego samego koloru (niebieskie) będą miały przed sobą (zielony piksel) i za sobą (szary, czerwony, szary) sekwencje kopiowania. W takim przypadku bardziej opłaca się „po całości” dać jedną sekwencję kopiowania. Oto dowód: klasycznie mielibyśmy: [0, 2] (kopiowanie zielonego piksela), [-1, 0] (dwa niebieskie) i [2, 3, 1, 3] (kopiowanie 3 pikseli), łącznie 8 bajtów. Zapisanie tego jako jednej sekwencji kopiowania daje nam [5, 2, 0, 0, 3, 1, 3] – o jeden bajt mniej. Dalej prosto, dwa zielone [1, 2], trzy niebieskie [2, 0], cztery szare [3, 3], później jest 7 pikseli do skopiowania [6, 1, 2, 1, 2, 3, 1, 2], dwa niebieskie [1, 0], trzy czerwone [2, 1] i dwa szare [1, 3].

Ostatecznie z 36 pikseli otrzymaliśmy ciąg [-2, 0, -3, 1, 5, 2, 0, 0, 3, 1, 3, 1, 2, 2, 0, 3, 3, 6, 1, 2, 1, 2, 3, 1, 2, 1, 0, 2, 1, 1, 3] składający się z 31 bajtów – kompresja zmniejszyła rozmiar danych do 86,1 % oryginału.

Dekodowanie jest bardzo proste, badamy znak liczby, jeżeli dodatni, to kopiujemy na wyjście następne n+1 bajtów, jeżeli ujemny, to powtarzamy następny bajt -n+1 razy (chyba, że n=-128, wtedy pomijamy ten bajt).

Algorytm RLE (Run Length Encoding)

Idea tego algorytmu jest identyczna jak ByteRun – eliminacja redundancji informacyjnej w postaci powtarzających się bajtów. Jest jednak kilka różnic, przede wszystkim założeniem jest przetwarzanie danych po 2 bajty, oprócz tego RLE posiada kilka sekwencji sterujących umożliwiających przeskakiwanie większych bloków danych. Sekwencja kopiowania danych wygląda następująco:

0x00	n	b0	b1	b2	b3	b4	b5	…	b(n-1)

Gdzie *n* przyjmuje wartości od 3 do 255. Warto zauważyć, że jeden bajt lub dwa bajty muszą być zapisane jako jedna lub dwie sekwencje powtarzania. Sekwencja powtarzania wygląda następująco:

n	b

Gdzie *n* jest dowolną liczbą od 1 do 255, a *b* powtarzanym bajtem. Pozostały nam jeszcze sekwencje specjalne. Sekwencje te są specyficzne dla formatu BMP i używane są przy kompresji maski przezroczystości: [0x00, 0x00] oznacza koniec linii obrazu, [0x00, 0x01] to koniec całego obrazu, [0x00, 0x02, x, y] to przesunięcie się w obrazie o wektor [x, y]. Przeskakiwane dane są traktowane jako przezroczyste.

RLE jest generalnie nieco mniej efektywny niż ByteRun, choć dzięki przetwarzaniu po 2 bajty, może być nieco szybszy. Na współczesnych procesorach jednak, różnica nie jest specjalnie widoczna.

Algorytm RLE – przykład działania

Kompresją RLE potraktujemy tę samą linię, co poprzednio.

Zaczynamy od sekwencji powtarzania: [3, 0], [4, 1]. Następnie mamy drobny dylemat, czy dwa niebieskie zapisać jako powtórzenie, czy nie... Przy powtórzeniu będziemy mieli [1, 2], [2, 0], i później albo 6 bajtów na powtórzeniach, albo też 6 na sekwencji kopiowania ([0, 3, 3, 1, 3, 0], końcowe, zbędne zdawałoby się, zero jest skutkiem założenia przetwarzania danych 2-bajtowymi porcjami, każda sekwencja musi mieć parzystą liczbę bajtów). Tak czy tak – 10 bajtów. Jedna długa sekwencja kopiowania [0, 6, 2, 0, 0, 3, 1, 3] ma tylko 8 bajtów. Następnie mamy proste sekwencje powtarzania [2, 2], [3, 0] i [4, 3], po niej zaś sekwencję skopiowania 7 bajtów [0, 7, 1, 2, 1, 2, 3, 1, 2]. I znów powtórzenia [2, 0], [3, 1] i [2, 3]. Ostatecznie otrzymujemy 33-bajtową sekwencję [3, 0, 4, 1, 0, 6, 2, 0, 0, 3, 1, 3, 2, 2, 3, 0, 4, 3, 0, 7, 1, 2, 1, 2, 3, 1, 2, 2, 0, 3, 1, 2, 3]. To więcej niż dla RunLength (winne jest tu przede wszystkim 2-bajtowe wejście do sekwencji kopiowania), skompresowany ciąg ma długość 91,6% oryginału.

Dekodowanie RLE nie nastręcza większych trudności. Typowe obrazki BMP nie zawierają sekwencji specjalnych zaczynających się od 0, a wyłącznie sekwencje powtarzania i kopiowania. Należy jedynie pamiętać o tym, żeby pominąć nadmiarowe zera wyrównujące sekwencje kopiowania do parzystej liczby bajtów.

Kompresja RLE posiada również odmianę RLE4 przeznaczoną do kompresji obrazków 16-kolorowych (w bajcie mieszczą się 2 piksele), kompresji ulegają wtedy nie tylko ciągi jednakowych pikseli, ale również rastry typu ABABAB, często używane do symulacji większej ilości kolorów.

Alfabet i słownik kompresora LZW

Przed omówieniem algorytmu kompresji LZW, zdefiniujmy dwa pojęcia, szeroko używane w teorii kompresji bezstratnej.

Alfabet danego kompresora, to zestaw wszystkich znaków, jakie mogą się pojawić na jego wejściu (w danych nieskompresowanych). Przykładowo dla 4-kolorowego obrazka alfabet to [0, 1, 2, 3], dla tekstu zaś alfabet składał się będzie z użytego zestawu znaków. W danym zastosowaniu kompresora alfabet jest najczęściej stały i znany z góry.

Słownik kompresora to tablica (lub inna struktura) zawierająca fragmenty wiadomości wejściowej i przypisane im kody na wyjściu kompresora. Słownik z reguły zależy od treści kompresowanej wiadomości i powstaje dynamicznie w czasie kompresji (wyjątkiem są algorytmy RLE i ByteRun, gdzie słownik jest stały i znany z góry). Niektóre algorytmy kompresji wymagają przesłania słownika w skompresowanej wiadomości, inne (jak np. LZW) są w stanie dynamicznie odtworzyć słownik w czasie dekompresji.

Inicjalizacja słownika w algorytmie LZW polega na:
1. Określeniu ilości bitów niezbędnych do zakodowania wszsystkich symboli z alfabetu.
2. Wstawieniu całego alfabetu do słownika.

Ponieważ alfabet jest znany zarówno przez kompresor jak i dekompresor, nie zachodzi potrzeba przesyłania części słownika zawierającej alfabet. Pozostała część słownika (zawierająca symbole złożone) jest odtwarzana przez dekoder na bieżąco, więc również nie trzeba jej przesyłać.

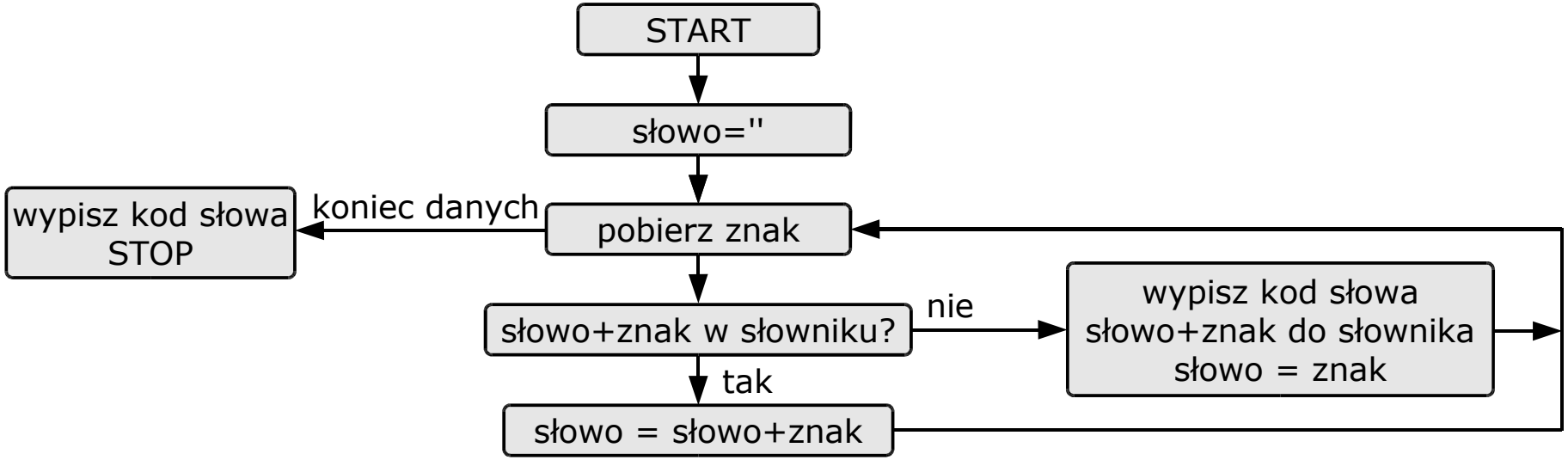
Kompresja LZW w akcji (1)

Algorytmem LZW skompresujemy tę samą linię obrazu, która posłużyła nam przy poprzednich algorytmach.

Alfabetem będą tu poszczególne kolory, zakodowane tak jak poprzednio [0, 1, 2, 3]. Do zakodowania alfabetu wystarczą nam 2 bity. Tak więc pierwsze wolne miejsce w słowniku będzie miało kod 4. Oto więc słownik na starcie pracy algorytmu:

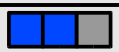
	0
	1
	2
	3

W czasie pracy algorytm wykonuje sekwencję czynności. Dane są przetwarzane znak po znaku, a więc nie jest wymagany swobodny dostęp do dowolnego elementu kompresowanej wiadomości.



Kompresja LZW w akcji (2)

A oto kolejne etapy kompresji linii przedstawione w formie tabeli:

znak	słowo	słownik ?	do słownika	na wyjście
		tak		
		nie	4	0
		tak		
		nie	5	4
		nie	6	1
		tak		
		nie	7	6
		nie	8	1
		nie	9	2
		tak		
		nie	10	4
		nie	11	3
		nie	12	1
		nie	13	3
		nie	14	2
		tak		
		nie	15	9
		tak		
		tak		
		nie	16	10
		nie	17	3
		tak		
		nie	18	17
		tak		

Kompresja LZW w akcji (3)

znak	słowo	słownik ?	do słownika	na wyjście
		 nie	 19	8
		 tak		
		 nie	 20	8
		 tak		
		 nie	 21	11
		 tak		
		 nie	 22	9
		 nie	 23	0
		 tak		
		 tak		
		 nie	 24	7
		tak		
koniec				17

Na wyjście został wyemitowany ciąg [0, 4, 1, 6, 1, 2, 4, 3, 1, 3, 2, 9, 10, 3, 17, 8, 8, 11, 9, 0, 7, 17], liczący sobie 22 bajty. Ciąg został skompresowany do 61,1% oryginalnej wielkości. Skuteczność kompresji rosłaby wraz ze wzrostem długości ciągu (i wielkości słownika). W praktyce ogranicza się wielkość słownika tak, że najdłuższy kod zajmuje 12 bitów, co daje 4096 elementów słownika. Dalsza kompresja przebiega tak samo, ale do słownika nie są już dopisywane kolejne pozycje. Najczęściej też kody skompresowane są zapisywane na polach bitowych o minimalnej długości, co dodatkowo w istotny sposób polepsza kompresję, szczególnie dla małych alfabetów.