# METODY SŁOWNIKOWE

KODOWANIE DANYCH, A.Przelaskowski

- Koncepcja metod słownikowych
- LZ77
- Modyfikacje (LZSS)
- LZ78
- Modyfikacje (LZW)
- Wykorzystanie
- Testy

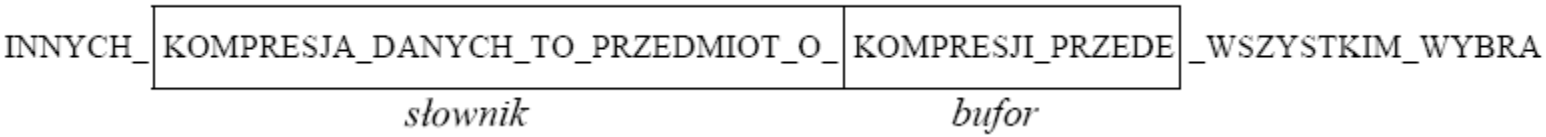
# Koncepcja

- Ciąg symboli o zmiennej długości → indeks o (prawie) stałej długości
- Dokładniej: ciąg symboli = identyczna fraza słownika → indeks frazy
- Efektywność: CR =bitowa długość frazy/bitowa długość indeksu (średnio)
  - dłuższe frazy → nieograniczona długość fraz
  - krótsze indeksy → mały słownik
- Klucz: koncepcja słownika
  - statyczna (*a priori*, np. słownik językowy zewnętrzny)
  - póładaptacyjna (słownik na podstawie analizy danych, konieczność kodowania słownika)
  - dynamiczna (przyczynowa adaptacyjność)

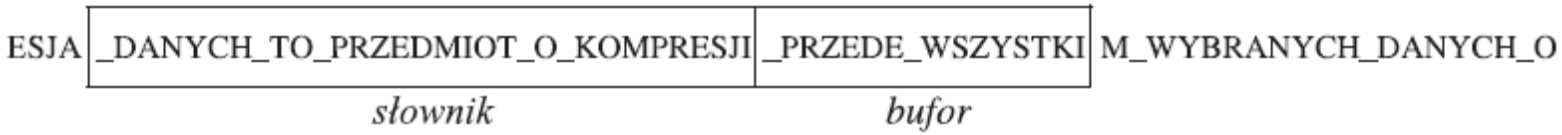
# LZ77

- słownik jako okno przesuwne:
  - dynamiczny
  - o ustalonym rozmiarze
  - struktura nasuwana na strumień ostatnio zakodowanych danych (model przyczynowy)
  - ograniczony rozmiar bufora frazy
  - indeks: (wskaźnik położenia frazy w słowniku, długość frazy, pierwszy symbol po kodowanym łańcuchu)
  - po zakodowaniu przesuwamy słownik (o długość frazy +1)
- Wady:
  - ograniczona długość kodowanego łańcucha
  - długi indeks
  - uwzględnienie jedynie ‘najbliższej historii’

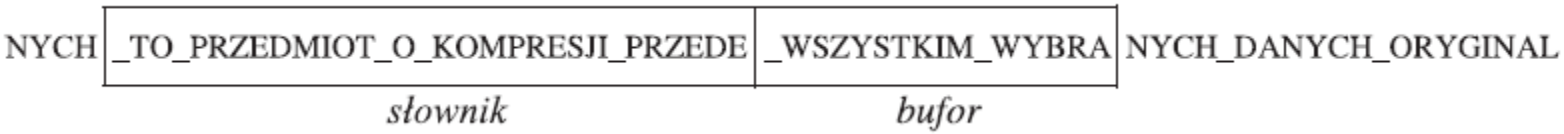
# Przykład (LZ77)




Indeks (1,8,”I”)




Indeks (11,6,”E”)




Indeks (1,1,”W”)

**Rozmiar indeksu**: np. 12bitów wskaźnika (4096 elementów słownika) +

+ 5bitów długości frazy (32 symbole) + 8bitów symbolu = 25 bitów

# Modyfikacje (LZSS)

- Modyfikacje:
  - dwa rodzaje indeksów
    - fraz krótszych: (bit, symbol)
    - fraz dłuższych: (bit, wskaźnik, długość)
  - efektywna w przeszukiwaniu struktura słownika
    - uporządkowane drzewo binarne z kolejnymi frazami słownika w węzłach: „KOMPRESJA DANYCH”, „OMPRESJA DANYCH ”, „MPRESJA DANYCH T”, „PRESJA DANYCH TO”, ....
  - inne, np. LZFG (wykorzystanie kodu unarnego do zapisu długości frazy)

# LZ78

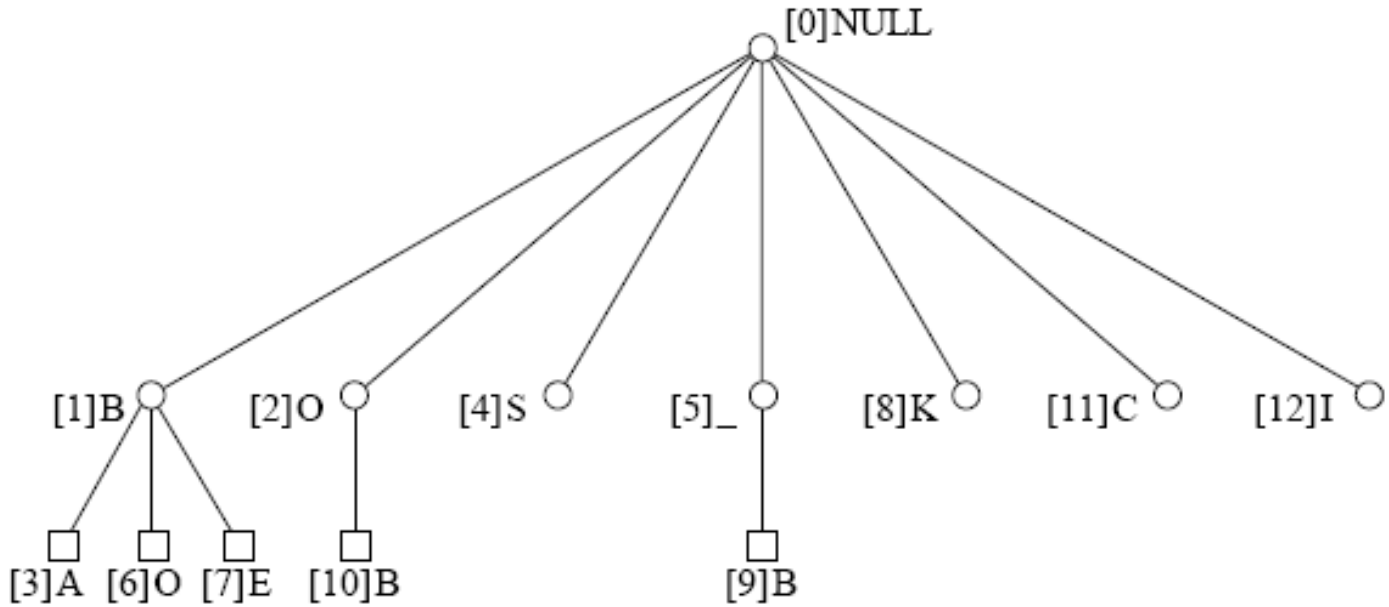
- Nieograniczony słownik zewnętrzny
  - dynamiczny
  - początkowo pusty z symbolem NULL
  - możliwy zmienny rozmiar (rosnący indeks)
  - wpisywane kolejno, coraz dłuższe frazy
  - nieograniczona długość frazy
  - ograniczanie rozmiaru słownika (poprawa efektywności)
  - indeks: (wskaźnik położenia frazy w słowniku, pierwszy symbol po kodowanym łańcuchu)
  - po zakodowaniu łańcucha wprowadzamy nową frazę do słownika (łańcuch plus symbol)

- Wady:
  - początkowo mało efektywny słownik (niewiele pozycji, krótkie frazy)
  - niewykorzystany rozmiar indeksu

# Przykład (LZ78)

| | | Sekwencja wyjściowa | | Słownik | |
|---|---|---|---|---|---|
| $i$ | Sekwencja wyjściowa | Indeks | Symbol | Indeks | Fraza |
| 0 | – | – | – | [0] | NULL |
| 1 | „B” | 0 | „B” | [1] | „B” |
| 2 | „O” | 0 | „O” | [2] | „O” |
| 3 | „BA” | 1 | „A” | [3] | „BA” |
| 4 | „S” | 0 | „S” | [4] | „S” |
| 5 | „–” | 0 | ’–’ | [5] | „–” |
| 6 | „BO” | 1 | „O” | [6] | „BO” |
| 7 | „BE” | 1 | „E” | [7] | „BE” |
| 8 | „K” | 0 | „K” | [8] | „K” |
| 9 | „–B” | 5 | „B” | [9] | „–B” |
| 10 | „OB” | 2 | „B” | [10] | „OB” |
| 11 | „C” | 0 | „C” | [11] | „C” |
| 12 | „I” | 0 | „I” | [12] | „I” |
| 13 | „O” | 2 | *EOF* | | |

# Struktura słownika




Przyspieszenie: funkcje numerujące (hashing)

# Modyfikacja (LZW)

- Wstępne zapełnienie słownika alfabetem
- Krótsze słowo: (wskaźnik)
- Rozbudowa słownika (sytuacja krytyczna, dodawanie dłuższych fraz do słownika – LZWM (łańcuch,łańcuch), LZAP (łańcuch,przedrostki kolejnego łańcucha)

# LZW (kodowanie) - przykład

| | | | | Słownik | |
|---|---|---|---|---|---|
| $i$ | Sekwencja wejściowa | Sekwencja wyjściowa | Pamięć | Indeks | Fraza |
| 0 | – | – | – | [0]-[255] | Kolejne symbole alfabetu źródła |
| 1 | „BO” | Ind(B) | „O” | [256] | „BO” |
| 2 | „B” | Ind(O) | „B” | [257] | „OB” |
| 3 | „A” | Ind(B) | „A” | [258] | „BA” |
| 4 | „S” | Ind(A) | „S” | [259] | „AS” |
| 5 | „_” | Ind(S) | „_” | [260] | „S_” |
| 6 | „B” | Ind(_) | „B” | [261] | „_B” |
| 7 | „OB” | 256 | „B” | [262] | „BOB” |
| 8 | „E” | Ind(B) | „E” | [263] | „BE” |
| 9 | „K” | Ind(E) | „K” | [264] | „EK” |
| 10 | „_” | Ind(K) | „_” | [265] | „K_” |
| 11 | „BO” | 261 | „O” | [266] | „_BO” |
| 12 | „BC” | 257 | „C” | [267] | „OBC” |
| 13 | „I” | Ind(C) | „I” | [268] | „CI” |
| 14 | „O” | Ind(I) | „O” | [269] | „IO” |
| 15 | – | Ind(O) | – | – | – |

# LZW (dekodowanie) - przykład

| | | | | | Słownik | |
|---|---|---|---|---|---|---|
| $i$ | Indeksy wejściowe | `POPRZEDNI_INDEKS` | Wyjściowy łańcuch symboli | `PIERWSZY_SYMBOL` | Indeks | Fraza |
| 0 | – | – | – | – | `[0]-[255]` | Kolejne symbole alfabetu |
| 1 | Ind(B) | – | „B” | „B” | – | – |
| 2 | Ind(O) | Ind(B) | „O” | „O” | `[256]` | „BO” |
| 3 | Ind(B) | Ind(O) | „B” | „B” | `[257]` | „OB” |
| 4 | Ind(A) | Ind(B) | „A” | „A” | `[258]` | „BA” |
| 5 | Ind(S) | Ind(A) | „S” | „S” | `[259]` | „AS” |
| 6 | Ind(_) | Ind(S) | „_” | „_” | `[260]` | „S_” |
| 7 | 256 | Ind(_) | „BO” | „B” | `[261]` | „_B” |
| 8 | Ind(B) | `[256]` | „B” | „B” | `[262]` | „BOB” |
| 9 | Ind(E) | Ind(B) | „E” | „E” | `[263]` | „BE” |
| 10 | Ind(K) | Ind(E) | „K” | „K” | `[264]` | „EK” |
| 11 | 261 | Ind(K) | „_B” | „_” | `[265]` | „K_” |
| 12 | 257 | 261 | „OB” | „O” | `[266]` | „_BO” |
| 13 | Ind(C) | 257 | „C” | „C” | `[267]` | „OBC” |
| 14 | Ind(I) | Ind(C) | „I” | „I” | `[268]` | „CI” |
| 15 | Ind(O) | Ind(I) | „O” | „O” | `[269]` | „IO” |

# LZW (sytuacja krytyczna)

Powód: przesunięcie o jedna pozycję zawartości słowników kodera i dekodera

kodowanie

| | | | Słownik | |
|---|---|---|---|---|
| Sekwencja wejściowa | Sekwencja wyjściowa | Pamięć | Indeks | Fraza |
| ... | ... | „_" | ... | ... |
| „SPORT" | Ind(_SPOR) | „T" | [1000] | „_SPORT" |
| „_TO" | Ind(T_T) | „O" | [1001] | „T_TO" |
| ... | ... | ... | ... | ... |
| ... | ... | „_" | ... | ... |
| „SPORT_" | 1000 | „_" | [2000] | „_SPORT_" |
| „SPORT_T" | 2000 | „T" | [2001] | „_SPORT_T" |

(znak,łańcuch)

(znak,łańcuch,znak, łańcuch,znak)

dekodowanie

| | | | | Słownik | |
|---|---|---|---|---|---|
| Indeksy wejściowe | POPRZEDNI_INDEKS | Wyjściowy łańcuch symboli | PIERWSZY_SYMBOL | Indeks | Fraza |
| ... | ... | ... | ... | ... | ... |
| Ind(_SPOR) | Ind(...) | „_SPOR" | „_" | [999] | „..._" |
| Ind(T_T) | Ind(_SPOR) | „T_T" | „T" | [1000] | „_SPORT" |
| ... | ... | ... | ... | ... | ... |
| ... | ... | ... | ... | ... | ... |
| 1000 | ... | „_SPORT" | '_' | [1999] | „..._" |
| 2000 | 1000 | ? | ? | [2000] | ? |

# Wykorzystanie

- PNG: predykcja+deflate
- Deflate: LZ77 (32kB i 258)+kod Huffmana
- LZ77: *LHA* (*LHarc*), *zip*, *gzip*, *ARJ*, *RAR*, *7-Zip* i inne
- LZW+kod Huffmana: Compress, PKArc, WinZip
- GIF (LZW)
- inne

# Deflate

- Ogólnodostępna biblioteka ZLIB
- Wersja LZ77 z podziałem na frazy krótkie i długie
- Słownik 32 kB, bufor 258 bajtów
- Cykliczna kolejka jako okno przesuwne
- Przeszukiwanie: tablica z numerowaniem na 3 symbolach przedrostka
- Podział danych wejściowych na bloki (max 64kB)
- Drzewa Huffmana statyczne, dynamiczne, częściowo predefiniowane, jedne do symboli i długości fraz, drugie do wskaźnika
- Drzewa dynamiczne zapisywane są na początku bloku za pomocą kody Huffmana

# Testy efektywności

| Koder | Zbiór | | | | | | |
|---|---|---|---|---|---|---|---|
| | Z1 (10 kB) | Z2 (100 kB) | Z3 (1 MB) | Z4 (4 MB) | Z5 (256 kB) | Z6 (256 kB) | Średnio |
| *LZSS* | 4,66 | 7,70 | 5,49 | 2,44 | 8,20 | 7,88 | 6,06 |
| *LZW* | 4,69 | 10,29 | 8,97 | 2,64 | 9,18 | 9,00 | 7,46 |
| *LZW_ZS* | 4,44 | 9,55 | 4,40 | 1,52 | 8,22 | 7,23 | 5,89 |
| *LZSS+HUF* | 3,43 | 6,76 | 4,18 | 1,18 | 7,29 | 6,96 | 4,97 |
| *LZSS+ARI* | 3,41 | 6,74 | 4,16 | 1,18 | 7,25 | 6,93 | 4,94 |

| Koder | Zbiór | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Z1 (4530) | Z2 (3953) | Z3 (2415) | Z4 (2048) | Z5 (2048) | Z6 (2048) | Z7 (3096) | Z8 (2877) | Średnio |
| *gzip* | 2,37 | 2,41 | 2,40 | 1,25 | 0,78 | 0,36 | 2,52 | 2,35 | 1,81 |
| *Compress* | 2,22 | 2,84 | 3,27 | 1,03 | 0,57 | 0,25 | 2,42 | 2,26 | 1,86 |

W przybliżeniu równy potencjał obu rodzin koderów słownikowych