METODY SŁOWNIKOWE

Materiały KODA, A.Przelaskowski

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

Koncepcja

- Wykorzystanie słownika fraz
- Ciąg symboli o zmiennej długości → indeks (frazy słownika) o (prawie) stałej długości
- Zasada (deterministyczna – zamiast przewidywania, jest wyszukiwanie):

  ciąg symboli == identyczna fraza słownika → indeks frazy
- Efektywność:

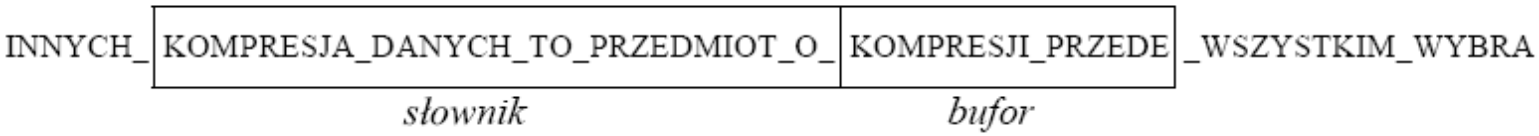
  CR = bitowa długość frazy/bitowa długość indeksu (średnio)
  - dłuższe frazy → nieograniczona długość fraz ?
  - krótsze indeksy → mały słownik?
- Klucz: koncepcja słownika
  - statyczna (*a priori*, np. słownik językowy zewnętrzny)
  - póładaptacyjna (słownik na podstawie analizy danych, konieczność kodowania słownika)
  - dynamiczna (budowanie słownika z adaptacją przyczynową)

LZ77

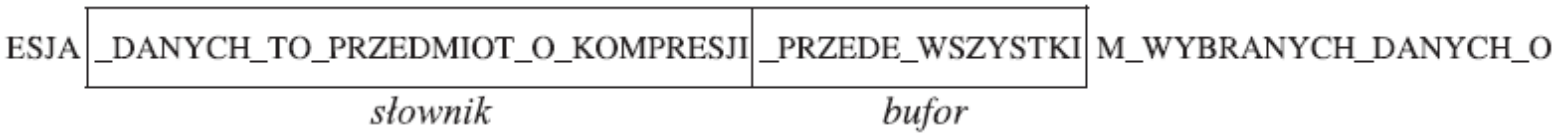
- słownik jako okno przesuwne:
  - dynamiczny
  - o ustalonym, ograniczonym rozmiarze (< ciąg zakod. dotąd symboli)
  - struktura nasuwana na strumień ostatnio zakodowanych danych (model przyczynowy)
  - ograniczony rozmiar (bufora) frazy
  - indeks: (wskaźnik położenia frazy w słowniku, długość frazy, pierwszy symbol po kodowanym łańcuchu)
  - po zakodowaniu przesuwamy słownik (o długość frazy +1)

- wady:
  - długi indeks
  - uwzględnienie jedynie ‘najbliższej historii’
  - ograniczona długość kodowanego łańcucha

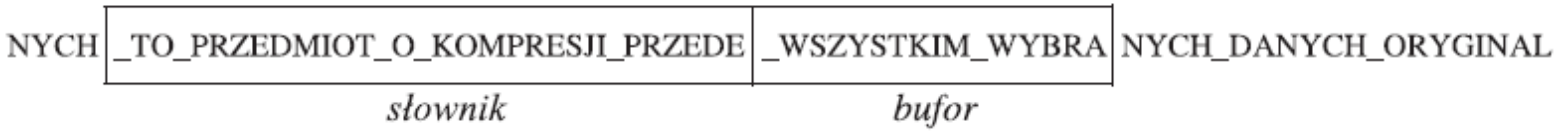
Przykład (LZ77)



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)

Modyfikacje: LZMA - indeksy

packed code (bits)	packet description
0 + byteCode	A single byte encoded using an adaptive binary range coder. The range coder uses context based on some number of the most significant bits of the previous byte. Depending on the state machine, this can also be a single byte encoded as a difference from the byte at the last used LZ77 distance.
1+0 + len + dist	**A typical LZ77 sequence describing sequence length and distance.**
1+1+0+0	A one-byte LZ77 sequence. Distance is equal to the last used LZ77 distance.
1+1+0+1 + len	An LZ77 sequence. Distance is equal to the last used LZ77 distance.
1+1+1+0 + len	An LZ77 sequence. Distance is equal to the second last used LZ77 distance.
1+1+1+1+0 + len	An LZ77 sequence. Distance is equal to the third last used LZ77 distance.
1+1+1+1+1 + len	An LZ77 sequence. Distance is equal to the fourth last used LZ77 distance.

LZMA – kodowanie wskaźnika położenia i długości frazy

Położenie (dist) kodowane jest na 11 bitach:
- 6 bitów określa klasę (segment)
- 5 bitów wskazuje liczbę bitów bezpośredniego odczytu położenia

Kodowanie długości frazy (len):

Length code (bits)	Description
0+ 3 bits	The length encoded using 3 bits, gives the lengths range from 2 to 9.
1+0+ 3 bits	The length encoded using 3 bits, gives the lengths range from 10 to 17.
1+1+ 8 bits	The length encoded using 8 bits, gives the lengths range from 18 to 273.

LZ78

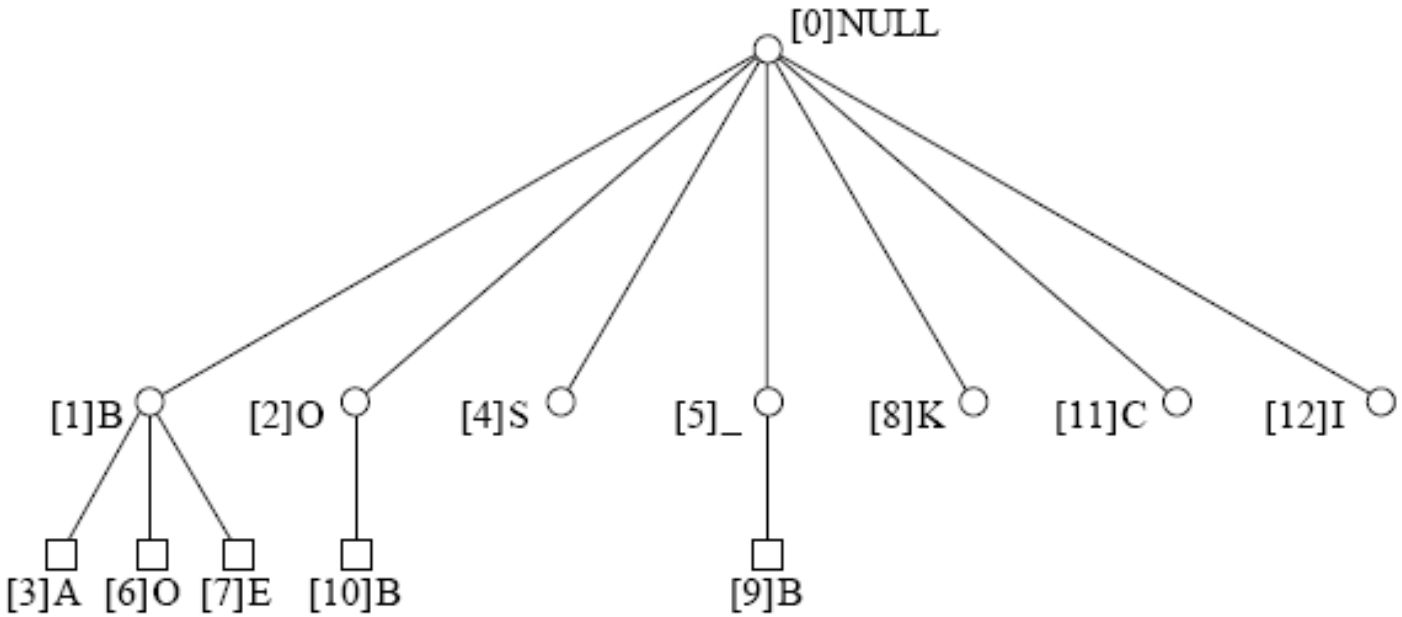
- Nieograniczony słownik zewnętrzny (koncepcja)
  - dynamiczny
  - początkowo pusty z symbolem NULL
  - wpisywane kolejno, coraz dłuższe frazy
  - ‘nieograniczona’! długość frazy
  - możliwy zmienny rozmiar (rosnący indeks)
  - ograniczanie rozmiaru słownika (do części aktywnej)
  - indeks: (wskaźnik położenia frazy w słowniku, pierwszy symbol po kodowanym łańcuchu)
  - po zakodowaniu łańcucha wprowadzamy nową frazę do słownika (łańcuch plus symbol); możliwe są inne frazy jako kombinacja wcześniejszych łańcuchów
- Wady podstawowej realizacji LZ78:
  - początkowo mało efektywny słownik (niewiele pozycji, krótkie frazy)
  - niewykorzystany rozmiar wskaźnika

Przykład (LZ78)

s_{WE}='BOBAS_BOBEK_BOBCIO'

		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

  Efekt: krótsze słowo: (wskaźnik)
- Rozbudowa słownika
  - ‘sytuacja krytyczna’
  - rosnący indeks
  - szybsza rozbudowa słownika: dodawanie dłuższych fraz:
    - LZMW (łańcuch,łańcuch)
    - LZAP (łańcuch,przedrostki kolejnego łańcucha)
- …

LZW (kodowanie) - przykład

s_{WE}='BOBAS_BOBEK_BOBCIO'

				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

i	Sekwencja wejściowa	Sekwencja wyjściowa	Pamięć	Słownik: Indeks	Słownik: 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)	–	–	–

					Słownik	
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 dowolne 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 kodu Huffmana

Testy efektywności

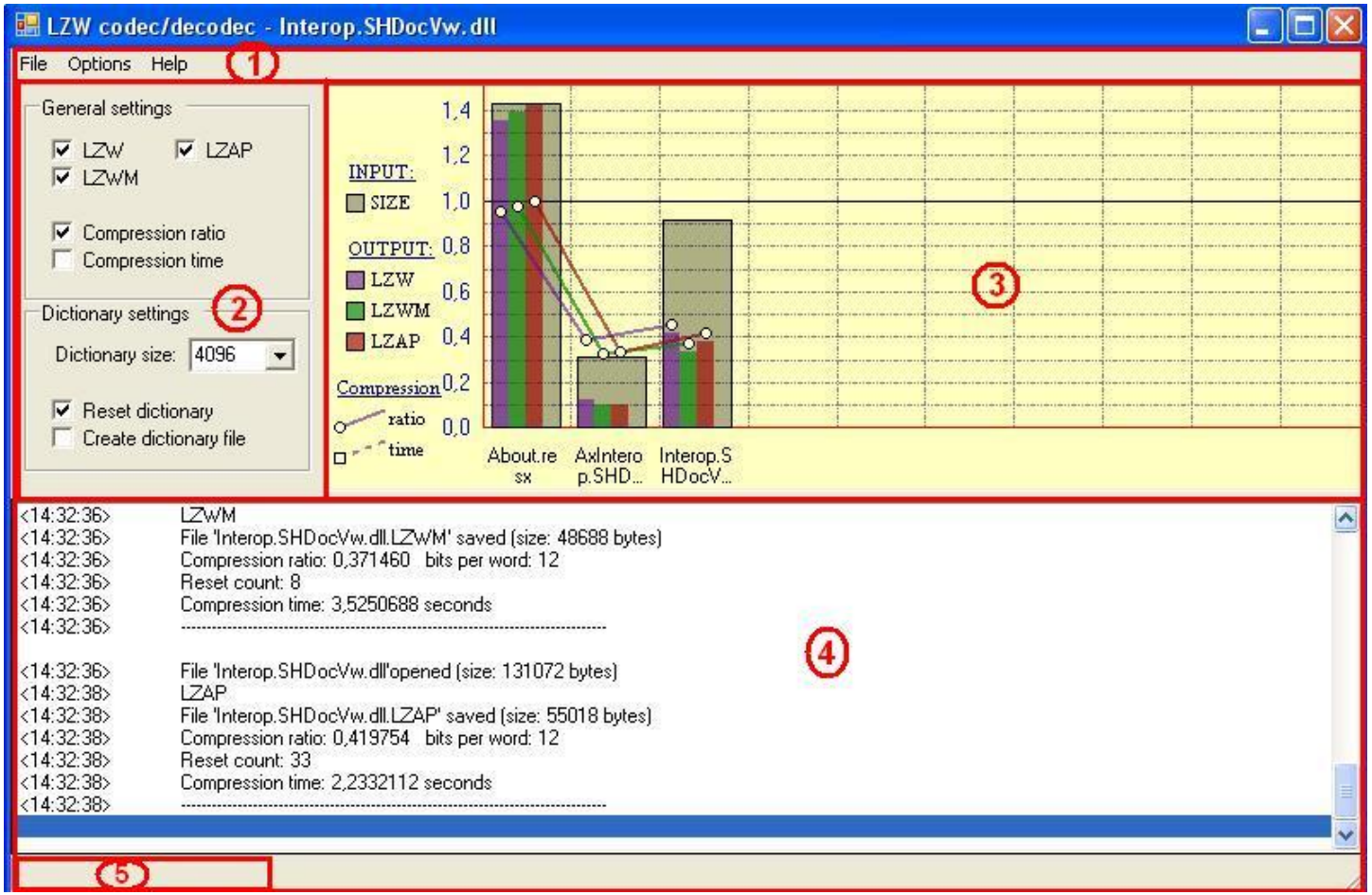
	Zbiór						
Koder	Z1 (10 kB)	Z2 (100 kB)	Z3 (1 MB)	Z4 (4 MB)	Z5 (256 kB)	Z6 (256 kB)	Średnio
LZSS	4,66	7,70	5,49	2,44	8,20	7,88	6,06
LZW	4,69	10,29	8,97	2,64	9,18	9,00	7,46
LZW_ZS	4,44	9,55	4,40	1,52	8,22	7,23	5,89
LZSS+HUF	3,43	6,76	4,18	1,18	7,29	6,96	4,97
LZSS+ARI	3,41	6,74	4,16	1,18	7,25	6,93	4,94

	Zbiór								
Koder	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

Narzędzie: LZW codec

(Piotr Frączek, Krzysztof Jankowski, Paweł Leszkiewicz, Mariusz Zyśk)



Wyniki

	dokument.doc	RAND.txt	REGULAR_small.txt	REGULAR.txt	TPDE.doc	motylek.avi	Cynthia.bmp	
Długość (kB)	1320	1420	1420	1900	707	712	1417	Średnia (bit rate)
dyn LZW (4096)	5,92	9,59	1,38	1,38	2,77	10,47	0,59	4,59
dyn LZWM (4096)	5,62	9,86	1,01	1,01	2,22	10,5	0,57	4,4
dyn LZAP (4096)	5,82	9,82	1,52	1,53	2,62	10,48	1,03	4,68
dyn LZW (8192)	6,02	9,71	1,27	1,28	2,7	10,74	0,56	4,61
dyn LZWM (8192)	5,74	10,09	0,99	0,98	2,16	10,89	0,54	4,48
dyn LZAP (8192)	5,89	10,05	1,38	1,39	2,46	10,85	0,98	4,71
dyn LZW (16384)	6,06	9,65	1,18	1,18	2,62	10,79	0,54	4,57
dyn LZWM(16384)	5,86	10,24	0,95	0,96	2,16	11,09	0,52	4,68
dyn LZAP (16384)	5,96	10,18	1,29	1,29	2,39	11,05	0,89	4,72
dyn LZW (32768)	5,92	9,38	1,1	1,2	2,66	10,49	0,55	4,47
dyn LZWM(32768)	6,19	10,28	0,94	0,94	2,24	11,02	0,52	4,59
dyn LZAP (32768)	6,23	10,21	1,21	1,22	2,34	10,97	0,78	4,71
stat LZW	6,31	10,22	1,48	1,48	2,95	11,16	0,63	4,83
stat LZWM	5,96	10,36	1,11	1,1	2,36	11,19	0,61	4,67
stat LZAP	6,22	10,33	1,54	1,43	2,84	11,19	1,1	4,95
RAR	1,91	7,21	1,08	1,07	1,36	6,54	0,58	2,82
ZIP	4,04	7,08	1,1	1,11	1,63	7,7	0,76	3,35
Huffman	6,29	6,96	3,6	3,6	4,23	7,81	1,74	4,89
Arytmetyczny	6,28	6,95	3,54	3,54	4,42	7,75	1,76	4,89