

Algoritmy komprese dat

Slovníkové metody Metody třídy LZ78



Abraham Lempel
Technion - Israel Institute of Technology



Jacob Ziv
Technion - Israel Institute of Technology

Slovníkové metody komprese dat

☀ Idea

- opakující se fráze uloženy do slovníku
- výskyty fráze v textu → ukazatel do slovníku

Problémy a v praxi používaná řešení:

- NP-těžký problém konstrukce optimálního slovníku ⇒ hladový algoritmus
- slovník musí znát i dekodér ⇒ dynamické metody

Jacob Ziv, Abraham Lempel

[LZ77] A universal algorithm for data compression. *IEEE Trans. Inform. Theory*, IT-23(3):337-343, 1977.

[LZ78] A compression of individual sequences via variable-rate coding. *IEEE Trans. Inform. Theory*, IT-24(5):530-536, 1978.

LZ78

Achillova pata LZ77:

a

b	c	d	e	f	g	h	i	a	b	c	d	e	f
---	---	---	---	---	---	---	---	---	---	---	---	---	---

 g h

Jacob Ziv, Abraham Lempel (1978)

Posuvné okno → explicitní slovník

- opakující se fráze jsou uloženy do slovníku
- výskyty fráze v textu jsou nahrazeny ukazatelem do slovníku
- slovník se nepřenáší

LZ78

- ① počáteční slovník prázdný
- ② na vstupu načti nejdelší frázi f , která je ve slovníku
- ③ na výstup $\langle i, k(z) \rangle$
 - i je ukazatel (číslo řádku) do slovníku na frázi f
 - $k(z)$ je kód znaku z , který ve vstupním řetězci následuje f
- ④ přidej do slovníku fz

☀ Příklad

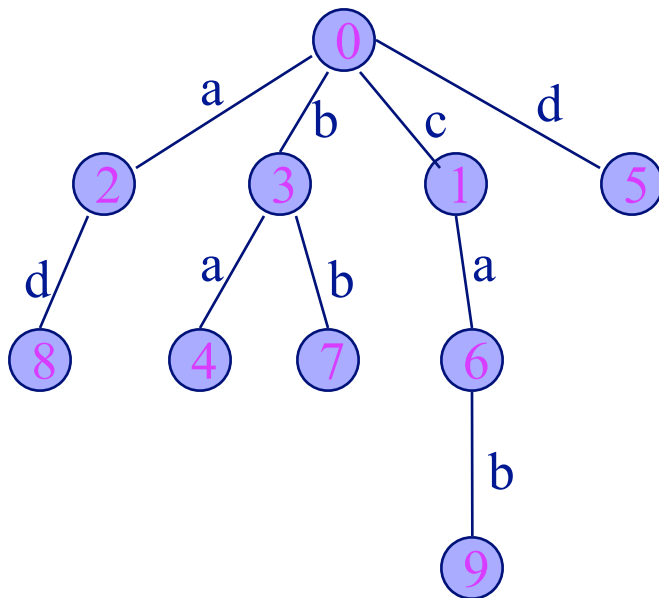
cabbadcabbadcab

$\langle 0,c \rangle \langle 0,a \rangle \langle 0,b \rangle \langle 3,a \rangle \langle 0,d \rangle \langle 1,a \rangle \langle 3,b \rangle \langle 2,d \rangle \langle 6,b \rangle$

index	fráze
1	c
2	a
3	b
4	ba
5	d
6	ca
7	bb
8	ad
9	cab

LZ78 - datové struktury

Slovník je reprezentován stromem (trie)



index	fráze
1	c
2	a
3	b
4	ba
5	d
6	ca
7	bb
8	ad
9	cab

Reorganizace slovníku

Nutná při vyčerpání prostoru pro slovník:

- vymazání celého slovníku
- vypuštění frází, které se
 - » vyskytovaly nejméně často (LFU)
 - » již dlouho nevyskytly (LRU)

Je třeba zachovat **prefixovou vlastnost**:

- je-li ve slovníku fráze f
- pak jsou v něm i všechny její předpony

Jiná strategie:

- když se začne zhoršovat kompresní poměr \Rightarrow
- vymaž slovník a začni znovu od počátečního nastavení

LZ78 - dekódování

$\langle 0,a \rangle \langle 0,b \rangle \langle 1,b \rangle \langle 3,a \rangle$

LZW

Terry Welch (1984)

LZ78 - citace & inovace

Inovace

- kód znaku následujícího frázi ve slovníku vyžaduje $\log n$ bitů, kde n je velikost abecedy
- nezávisí na vstupním řetězci

LZ78 s odloženou inovací

Algoritmus

- inicializace slovníku
 - » na počátku jsou do slovníku vloženy všechny znaky vstupní abecedy
- načti na vstupu nejdelší frázi f , která je ve slovníku
- $\text{output}(\text{index}(f))$
- do slovníku ulož fz , kde z je znak následující f ve vstupním řetězci

☀ Příklad

Vstup: abcabcabcba

index	fráze
1	a
2	b
3	c

LZW - příklad

Vstup: abcabcabcba

Výstup: 1 2 3 4 6 5 9 1

index	fráze
1	a
2	b
3	c
4	ab
5	bc
6	ca
7	abc
8	cab
9	acb

LZW - dekodování

Kódové slovo: 1 2 3 4 6 5 9 1

Problém:

- na vstupu je podřetězec *awawa*
- *a* - znak, *w* - slovo
- *aw* je ve slovníku
- *awa* není ve slovníku

Řešení:

- chybějící fráze má tvar *awa*
- *aw* je naposledy dekodované slovo

 **Problém:**

- najděte nejkratší vstup, na němž nastane popsaná situace

index	fráze
1	a
2	b
3	c
4	ab
5	bc
6	ca
7	abc
8	cab
9	bcb

LZC

compress 4.0 (Thomas et al, 1985)

- modifikace LZW

Kódování ukazatelů

- binární kód se vzrůstající délkou
- 9 - 16 b

Zaplnění paměti

- → komprese se statickým slovníkem

Monitoruje kompresní poměr

- zhoršení \Rightarrow vymaž slovník a začni znovu od počátečního nastavení

Protokol V.42bis (ITU-T) pro modemy

LZT

Tischer (1987)

Modifikace LZC

Kódování ukazatelů

- minimální binární kód s rostoucí délkou

Přeplnění slovníku

- \Rightarrow vyloučím frázi, která se nejdéle nevyskytla (LRU)

Slovník = LRU seznam

- indexovaný hašovací tabulkou
- nová fráze se vkládá na začátek seznamu
- poslední fráze se ze seznamu vylučuje

Miller, Wegman (1984)

☀ Idea:

- nová fráze = zřetězení *dvou* posledních frází
- délka frází se rychleji zvětšuje

Přeplnění slovníku \Rightarrow strategie LRU

Slovník nemá prefixovou vlastnost

- to komplikuje vyhledávání
- \Rightarrow backtracking

LZAP

Storer (1988)

☀ Idea:

- místo ST přidat všechny podřetězce St
- kde t je předpona T

☞ Vlastnosti:

- větší slovník \Rightarrow
 - » delší kódová slova
 - » ale širší výběr fráze
- rychlejší vyhledávání
 - » nevyžaduje backtracking