

Algoritmy komprese dat

Kontextové metody



Matt Mahoney
Florida Institute of Technology

Kompresce dat: 2 fáze

① Vytvoření modelu

② Kódování

Model 0-tého řádu

- pravděpodobnosti výskytu izolovaných znaků abecedy
- žádná korelace mezi sousedními znaky

Model s konečným kontextem

Pravděpodobnost výskytu znaku závisí

- nejen na jeho četnosti výskytu
- nýbrž též na kontextu, v němž se vyskytuje

Původně navrženo pro kompresi textových souborů

Model řádu i - používá kontext délky i

Modely s konečným kontextem

Metody

- s pevnou délkou kontextu
- kombinované - používají kontexty různých délek
 - » úplné (všechny kontexty délek $i, i-1, \dots, 0$)
 - » částečně kombinované
- (statické), adaptivní

PPM - Prediction by Partial Matching

Cleary, Witten (1984) Moffat (1990)

- kombinace kontextového modelu & aritmetického kódování
- kombinovaný model řádu i

Pro znak z a kontext c

- určíme $f(z | c) = \text{četnost znaku } z \text{ v kontextu } c$

PPM

Kódování znaku z :

Bud' c kontext délky i

`read(z)`

```
if  $f(z|c) > 0$  then kóduj  $z$  s použitím  $f(z|c)$   
      else output(kód(ESC))  
           zkus kontext řádu  $i-1$ 
```

Nezbytný předpoklad

- pro jisté i musí být $f(z|c)$ pro všechny kontexty délky i definováno

Order $k = 2$ Predictions c p	Order $k = 1$ Predictions c p	Order $k = 0$ Predictions c p	Order $k = -1$ Predictions c p
ab \rightarrow r 2 $\frac{2}{3}$ \rightarrow Esc 1 $\frac{1}{3}$	a \rightarrow b 2 $\frac{2}{7}$ \rightarrow c 1 $\frac{1}{7}$ \rightarrow d 1 $\frac{1}{7}$ \rightarrow Esc 3 $\frac{3}{7}$	\rightarrow a 5 $\frac{5}{16}$ \rightarrow b 2 $\frac{2}{16}$ \rightarrow c 1 $\frac{1}{16}$ \rightarrow d 1 $\frac{1}{16}$ \rightarrow r 2 $\frac{2}{16}$ \rightarrow Esc 5 $\frac{5}{16}$	\rightarrow A 1 $\frac{1}{ A }$
ac \rightarrow a 1 $\frac{1}{2}$ \rightarrow Esc 1 $\frac{1}{2}$	b \rightarrow r 2 $\frac{2}{3}$ \rightarrow Esc 1 $\frac{1}{3}$		
ad \rightarrow a 1 $\frac{1}{2}$ \rightarrow Esc 1 $\frac{1}{2}$	c \rightarrow a 1 $\frac{1}{2}$ \rightarrow Esc 1 $\frac{1}{2}$		
br \rightarrow a 2 $\frac{2}{3}$ \rightarrow Esc 1 $\frac{1}{3}$	d \rightarrow a 1 $\frac{1}{2}$ \rightarrow Esc 1 $\frac{1}{2}$		
ca \rightarrow d 1 $\frac{1}{2}$ \rightarrow Esc 1 $\frac{1}{2}$	r \rightarrow a 2 $\frac{1}{3}$ \rightarrow Esc 1 $\frac{1}{3}$		
da \rightarrow b 1 $\frac{1}{2}$ \rightarrow Esc 1 $\frac{1}{2}$			
ra \rightarrow c 1 $\frac{1}{2}$ \rightarrow Esc 1 $\frac{1}{2}$			

abracadabra

PPM - pokračování

Jak definovat $f(z|c)$?

- # výskytů znaku z v kontextu c
- # případů, v nichž byl kontext c použit k predikci z

Princip **exkluze**

- x se vyskytne poprvé v kontextu abc
- $f(y|abc) > 0 \Rightarrow y$ lze vyloučit z modelu 2. řádu
- empirické údaje: 2x délka výpočtu
zlepšení komprese o 5%

Problém: pravděpodobnost znaku *esc*

PPMA (Cleary, Witten)

- pro kontext c platí $f(c) = n$
- $\Rightarrow P(\text{esc}|c) = 1/(n+1)$

PPMB

- $f(z|c)' = f(z|c) - 1$
- $abcx...abcx....abcy$
- $f(x|abc)' = 1$, $f(y|abc)' = 0$, $f(\text{esc}|abc)' = 2$
- $\Rightarrow P(\text{esc}|c) = 2/n$

Problém: pravděpodobnost znaku *esc*

PPMC (Moffat)

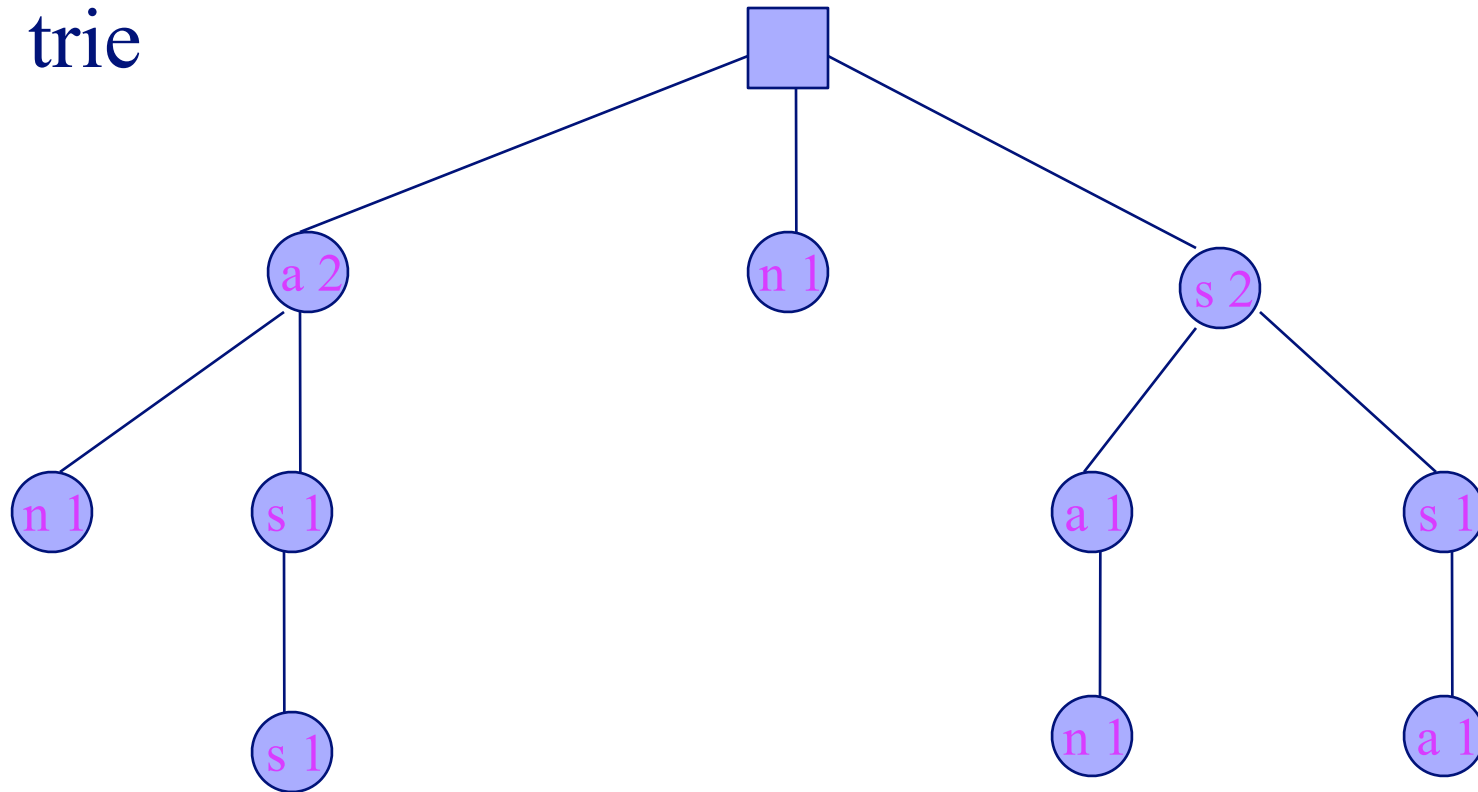
- pro každý kontext c skupina znaků, pro něž $f(x|c) > 0$
- $f(esc|c) := \#$ znaků ve skupině
- $\Rightarrow P(esc|c) = f(esc|c) / (n + f(esc|c))$

PPMD (P.G.Howard, J.Vitter)

- $f(esc|c) := (\# \text{ znaků ve skupině}) / 2$
- $f(x|c) := 1$. výskyt má váhu $1/2$, další 1

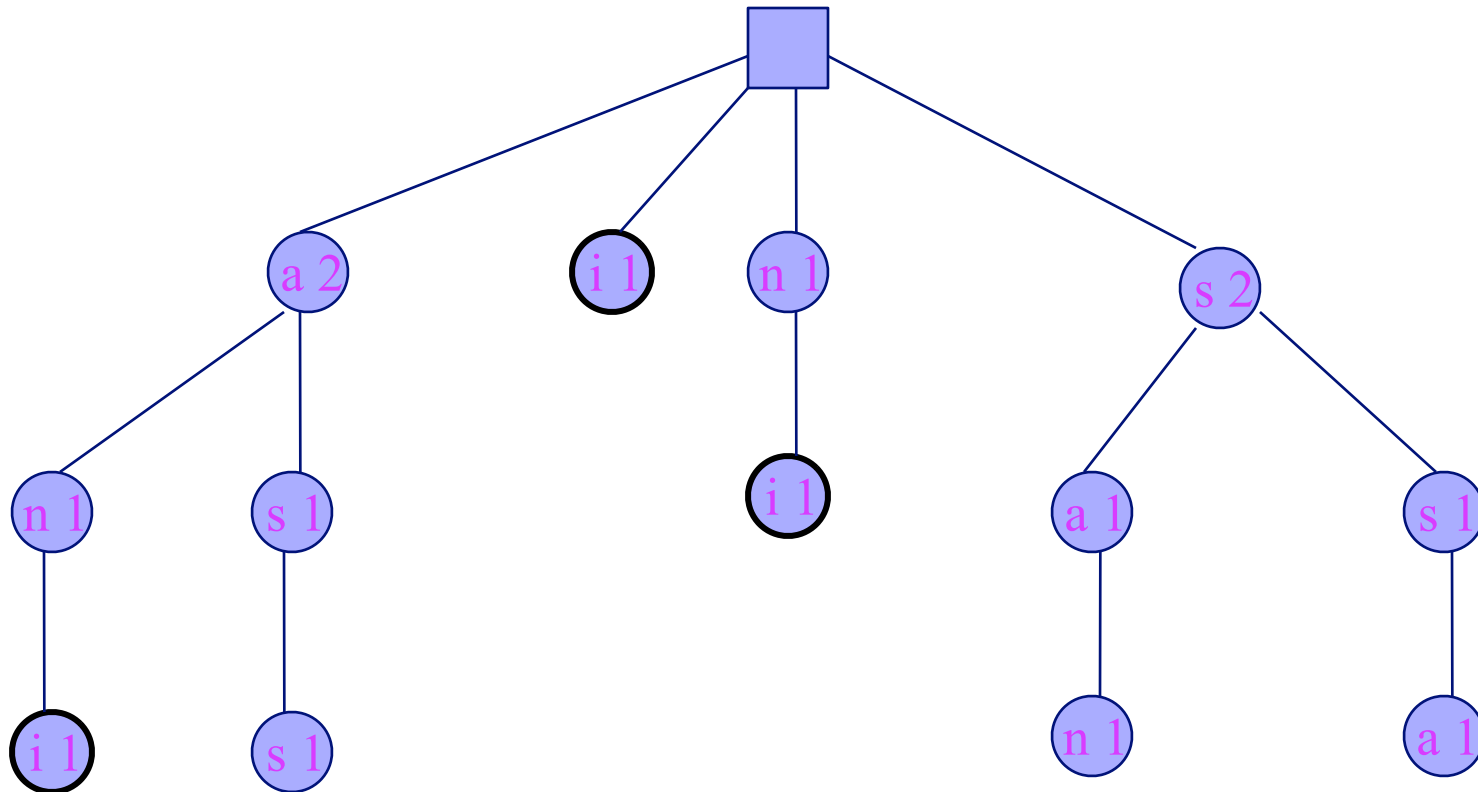
Datové struktury: kontextový strom

trie



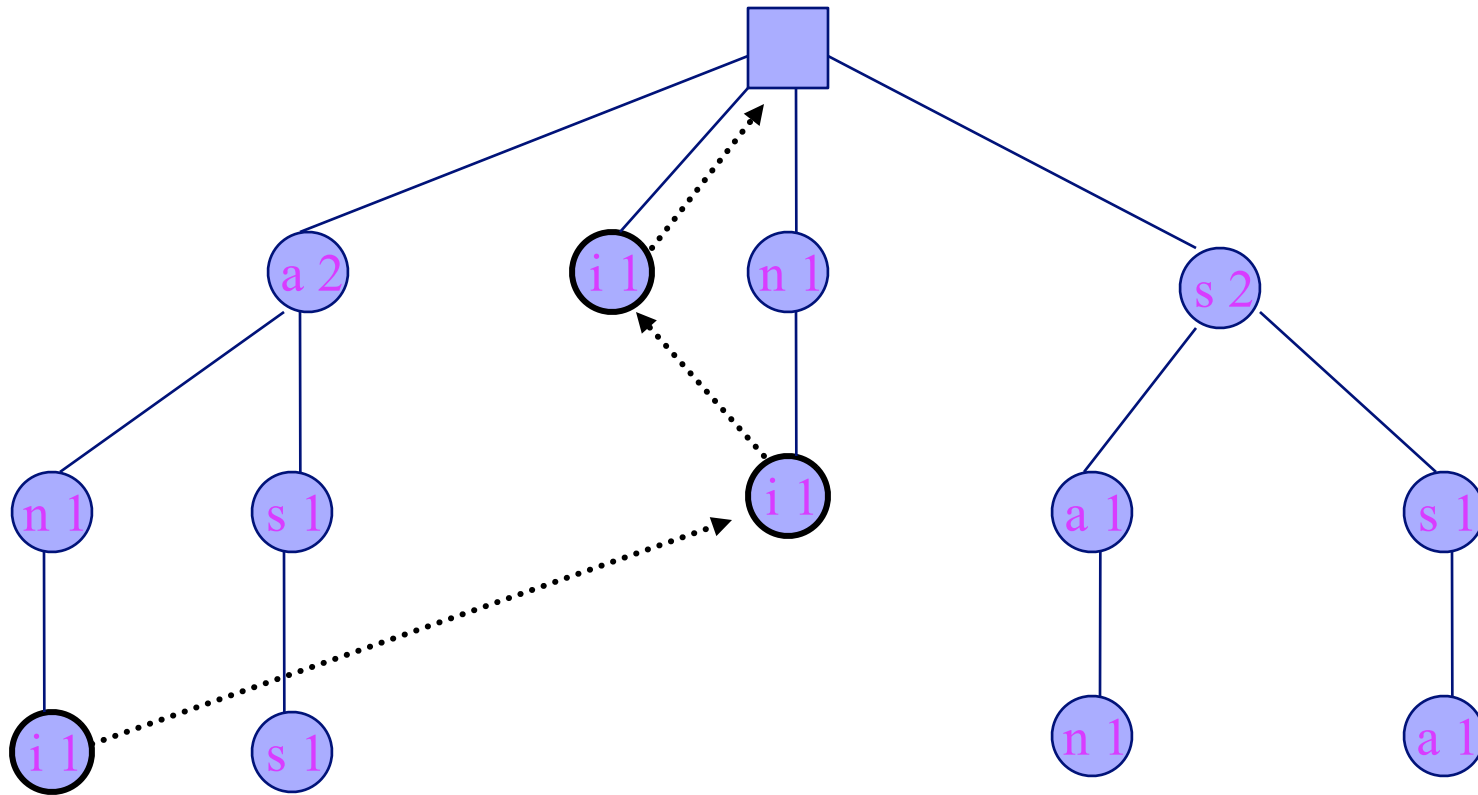
assan

Datové struktury: kontextový strom



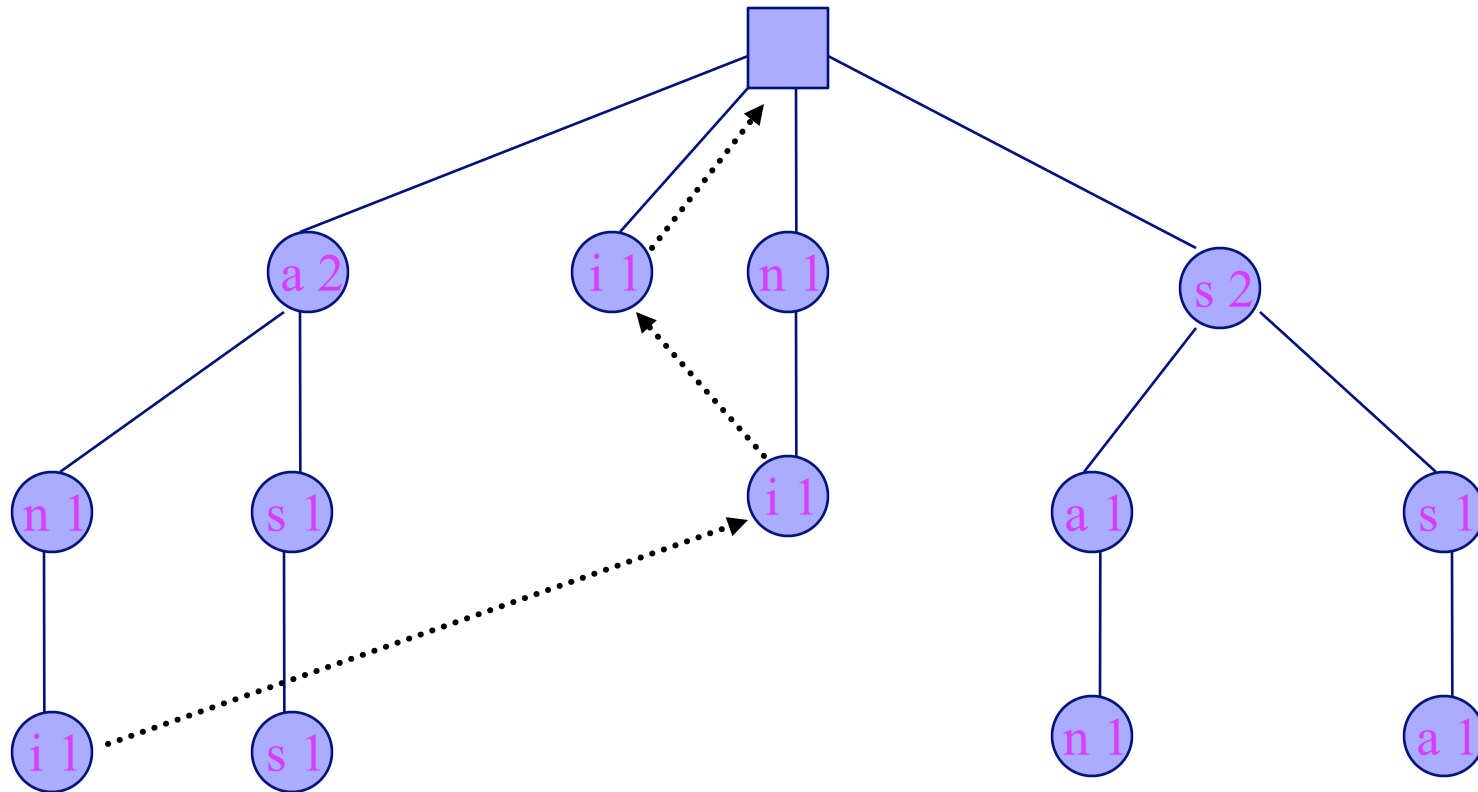
assani

Datové struktury: kontextový strom



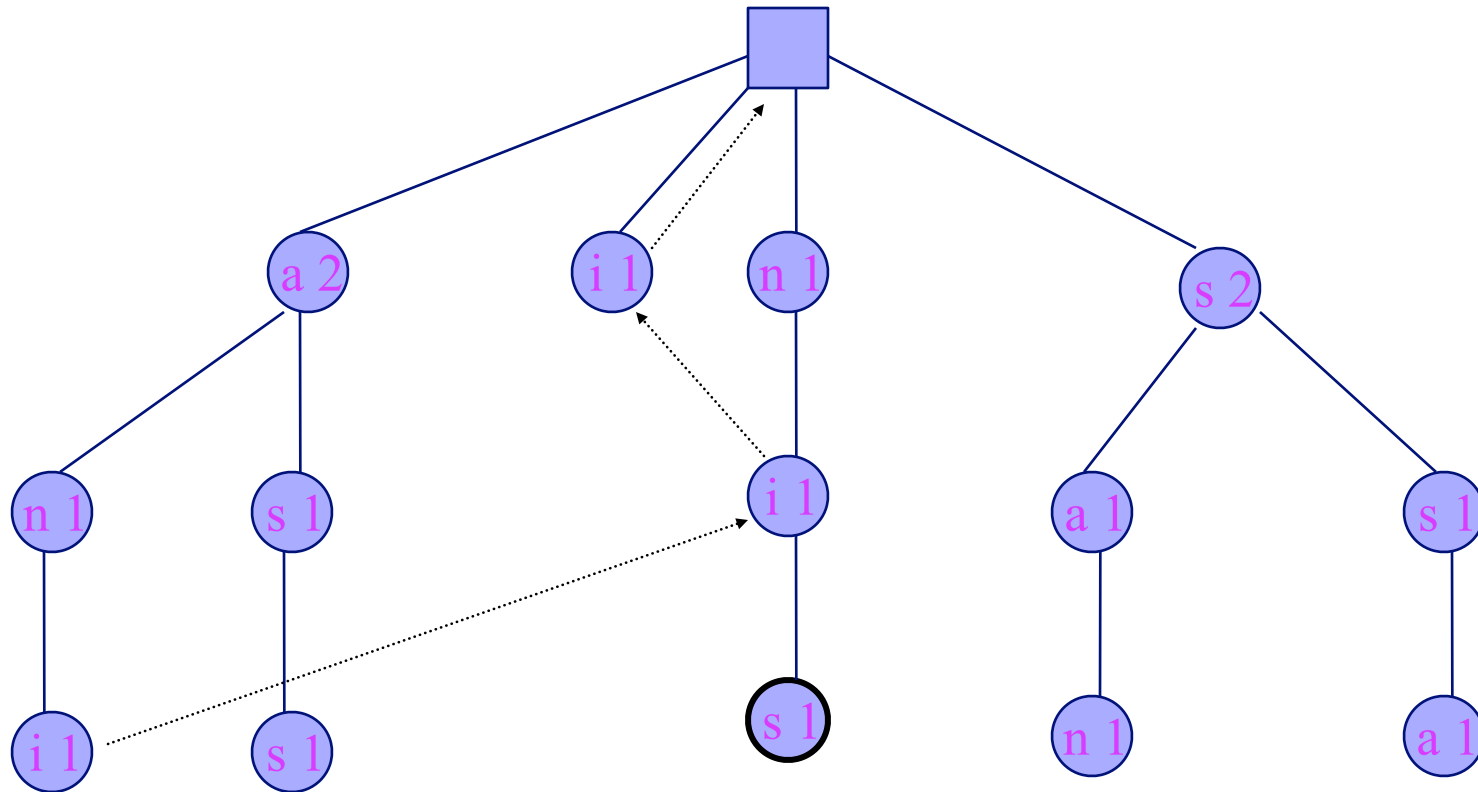
assani

Datové struktury: kontextový strom



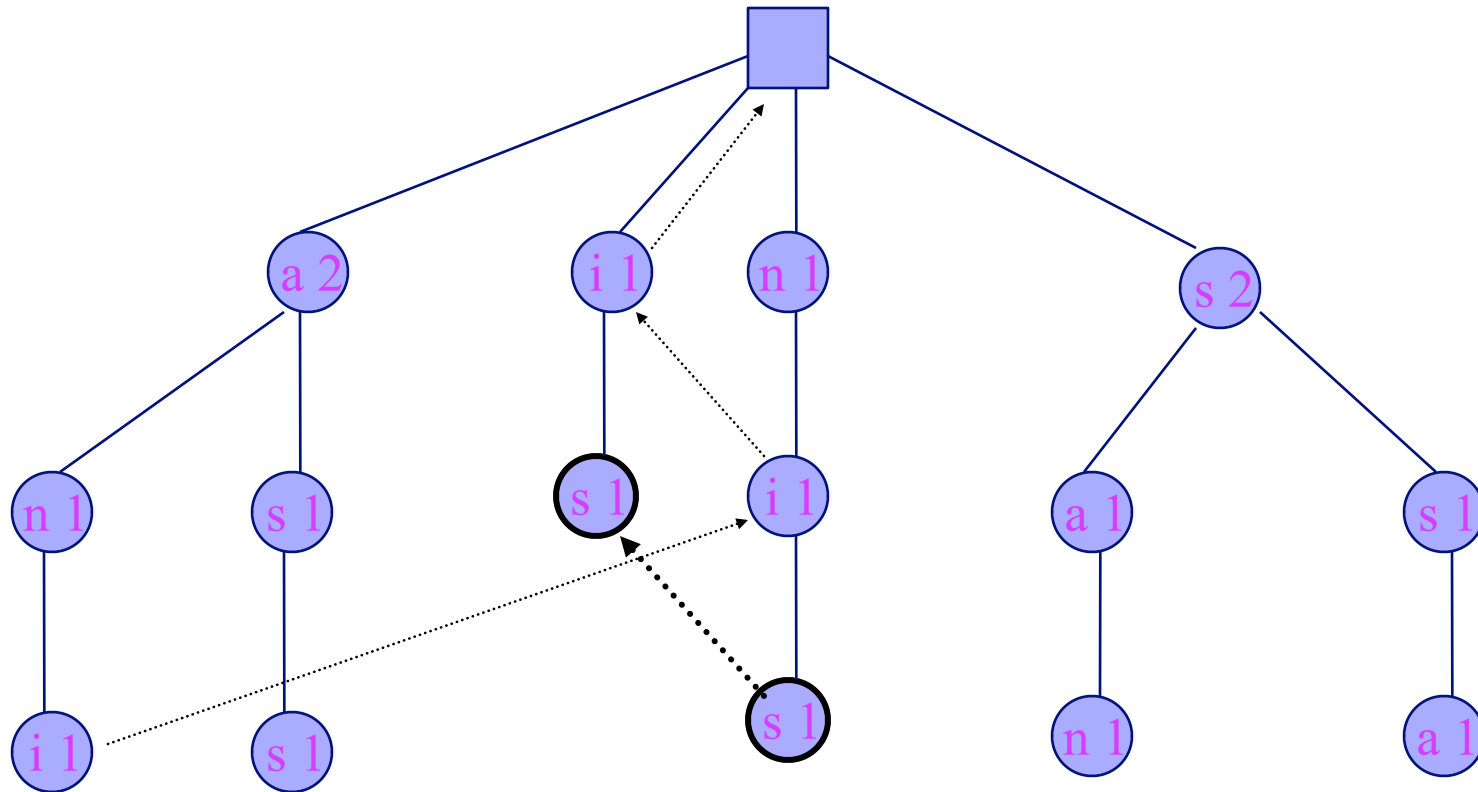
assanis

Datové struktury: kontextový strom



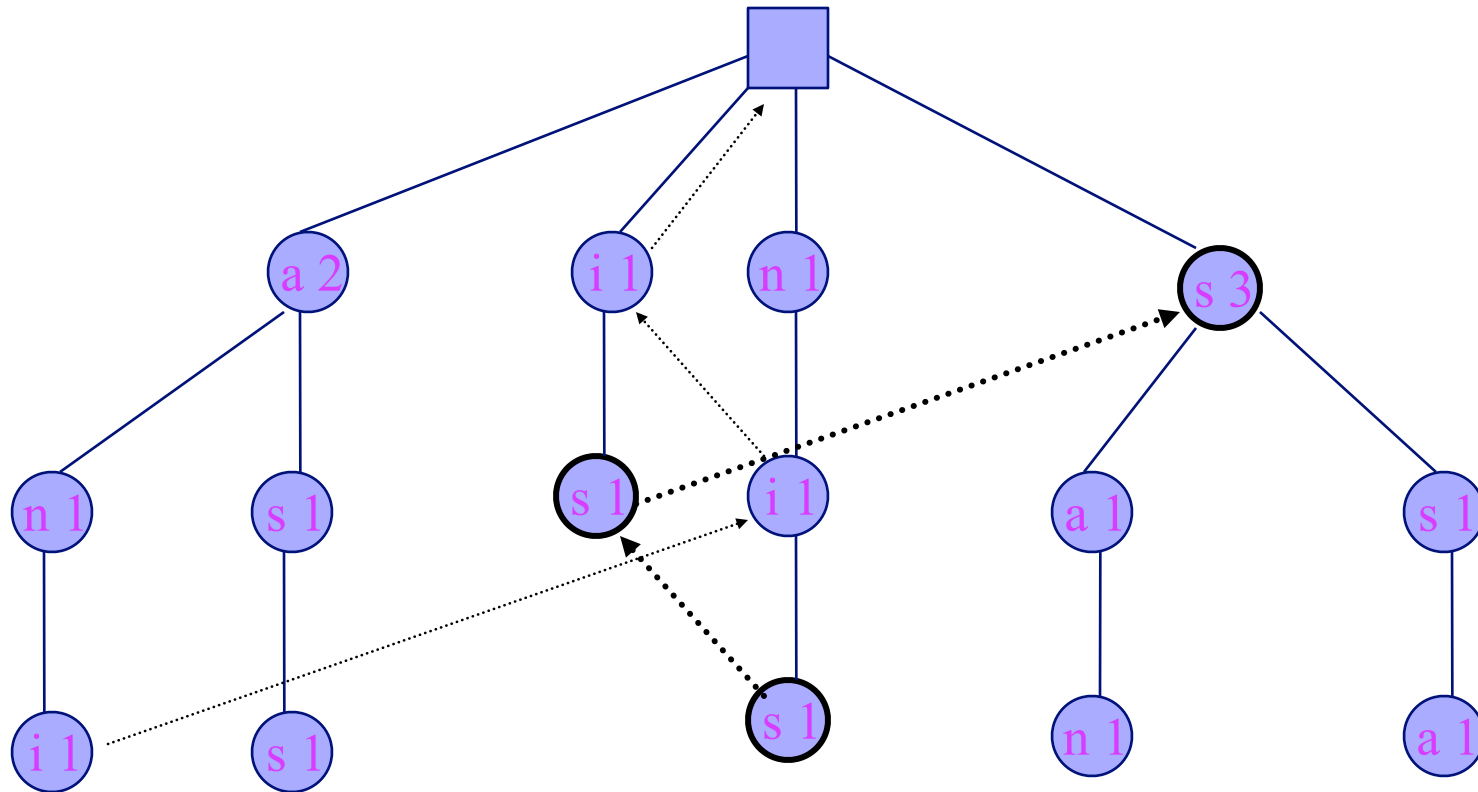
assanis

Datové struktury: kontextový strom



assanis

Datové struktury: kontextový strom



assanis

Paměťová omezení

Sledování velikosti volné paměti, pokud klesne pod určitou mez

- \Rightarrow **zmrazení** modelu
 - » aktualizuj četnosti již existujících kontextů
 - » ignoruj nové kontexty
- \Rightarrow **rekonstrukce** modelu
 - » k inicializaci použij bezprostřední historii, uloženou ve vyrovnávací paměti

Paměťová omezení

Vylepšení

- kromě volné paměti sleduj též relativní úspěšnost komprese
- pokud začne klesat \Rightarrow *rekonstrukce* modelu

Efektivnější datová struktura: DAWG

- Directed Acyclic Word Graph
- sloučení ekvivalentních kontextů

Experimentální výsledky (Moffat, Turpin, 2002)

	prostor (MB)						
řád	1	2	4	8	16	32	64
1	3,38						
2	2,44						
3	1,91	1,90					
4	1,85	1,71	1,66				
5	2,02	1,81	1,67	1,60	1,58		
6	2,18	1,96	1,78	1,66	1,59	1,56	
7	2,31	2,09	1,90	1,76	1,66	1,58	1,56

Jednotky: b/znak

Soubor: bible.txt

Metoda: PPMD

Obsluha paměti: při vyčerpání paměti rekonstrukce

PPM* (Cleary, Teahan, Witten, 1995)

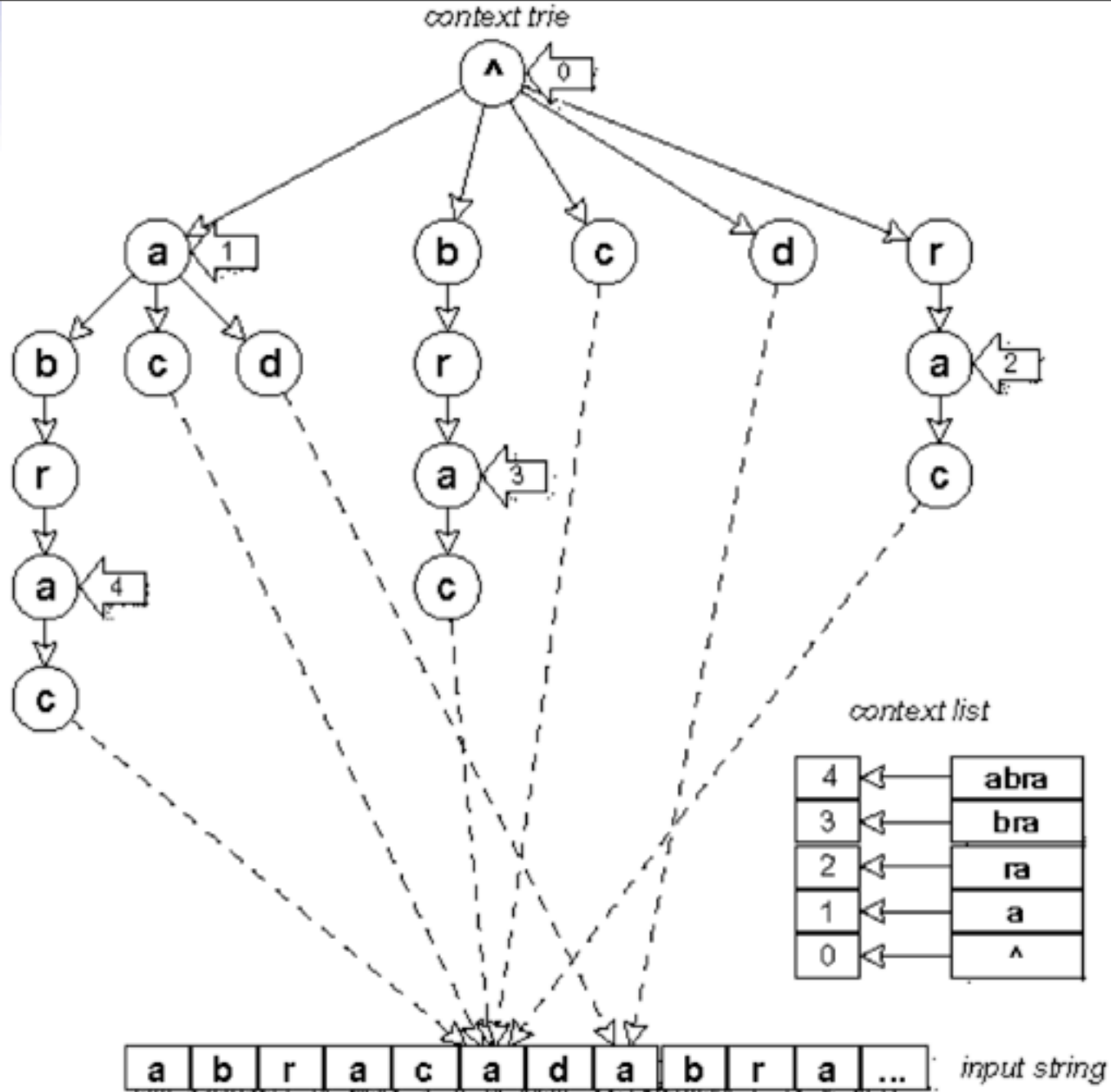
Kontexty libovolné délky

- výběr „nejvhodnějšího“ kontextu
- dekodér musí být schopen učinit stejnou volbu jako kodér, i když nezná následující znak

Deterministický kontext - v minulosti se v něm vyskytoval vždy stejný znak

Strategie navržená CTW

- použij nejprve vždy nejkratší deterministický kontext
- pokud takový neexistuje, použij nejdelší v seznamu



Další modifikace

PPMII (D.Škarin, 2002)

- *PPM: One Step to Practicality*, DCC '02
- kontexty dědí od svých předků
- 3 třídy kontextů s různými modely
- škálování četností
- RAR

Další modifikace

PAQ

- Matt Mahoney (PAQ1, 2002)
 - » PAQ8PX, 2009, Jan Ondruš
- binární abeceda
- kontextový model + aritmetický kodér
- context mixing
 - » vážený průměr odhadů psti z různých modelů