Kompresia sociálnych sietí
... od textu po grafy

Daniel Švoňava

Ontožúr 2009/2

Bojový plán

Čo je to entropia?

Miera neporiadku/náhodnosti v štruktúre systému.

Fyzici

Vysoká teplota $\longrightarrow$ vysoká entropia.

Informatici

Miera nepredpovedateľnosti hodnoty náhodnej premennej.

Biológovia

Živé organizmy majú nízku entropiu, skladba tela je predpovedateľná.

Naozajstná definícia

Očakávaná hodnota množstva informácie obsiahnutej v premennej X.

$$\begin{aligned} H(X) &= \mathrm{E}(I(X)) \\ &= \sum_{i=1}^{n} p(x_i)\, I(x_i) \\ &= -\sum_{i=1}^{n} p(x_i) \log_b p(x_i) \end{aligned}$$

kde $p(x_i)$ je pravdepodobnosť, že X nadobudne hodnotu práve x_i.

Kompresia postupností znakov

- dobre preskúmaná oblasť (vyše 50 r. výskumu)
- slovníková kompresia
- entropická kompresia

Entropická kompresia

Očakávaná dĺžka komprimovanej reprezentácie premennej X

$$\sum_{i=1}^{n} p(x_i)\, R(x_i)$$

kde $R(x_i)$ je dĺžka reprezentácie konkrétneho znaku x_i

- využijeme fakt, že $p(x_i)$ nemá rovnomernú distribúciu
- priradíme krátke reprezentácie znakom s veľkou pravdepodobnosťou
- prefix-free kódy

Príklad

Komprimovaný text

Kam si sa vybral jo_

x_i	$p(x_i)$	bitová rep.	$R(x_i)$
ž	0.6	0	1
g	0.1	10	2
r	0.01	1100	4
k	0.01	1101	4

...

Klasická reprezentácia grafu

Zoznamy susedov

- graf ako $G = (V, E)$
- vrcholy ako indexy z 0 až $|V| - 1$
- pre každý vrchol si pamätáme zoznam indexov jeho susedov

Transformácia na textovú kompresiu

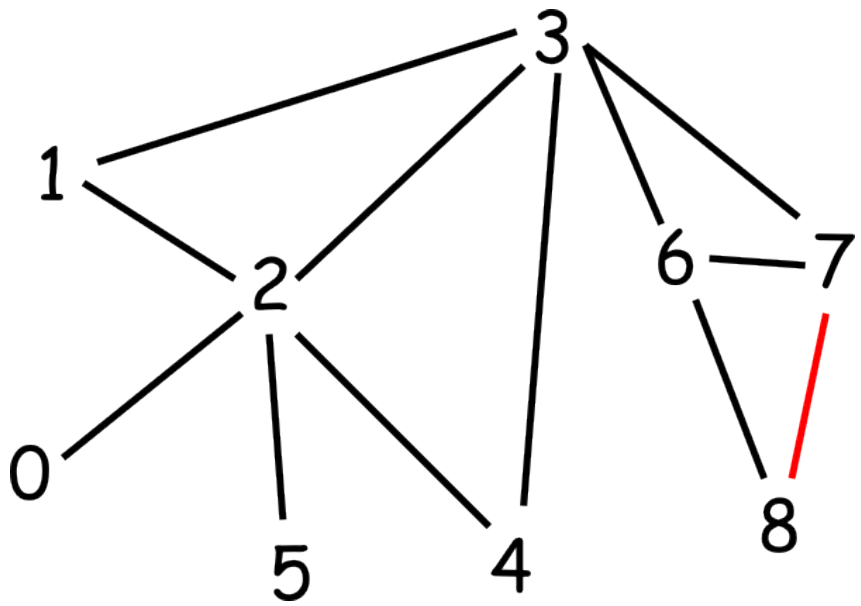
Zoznamy susedov

- graf ako $G = (V, E)$
- vrcholy ako indexy z 0 až $|V| - 1$
  $\longrightarrow$ **znaky** x_0 **až** x_{n-1}
- pre každý vrchol si pamätáme zoznam indexov jeho susedov
  $\longrightarrow$ **každý zoznam je jedno slovo textu, do abecedy sa pridá oddeľovač slov** x_n

Príklad

Komprimovaný graf

∅
∅
0 1
1 2
2 3
2
3
3 6
6 _



- riadok $i \longrightarrow$ susedia vrcholu i
- hrany nie sú orientované a sú zaznačené len raz

Príklad

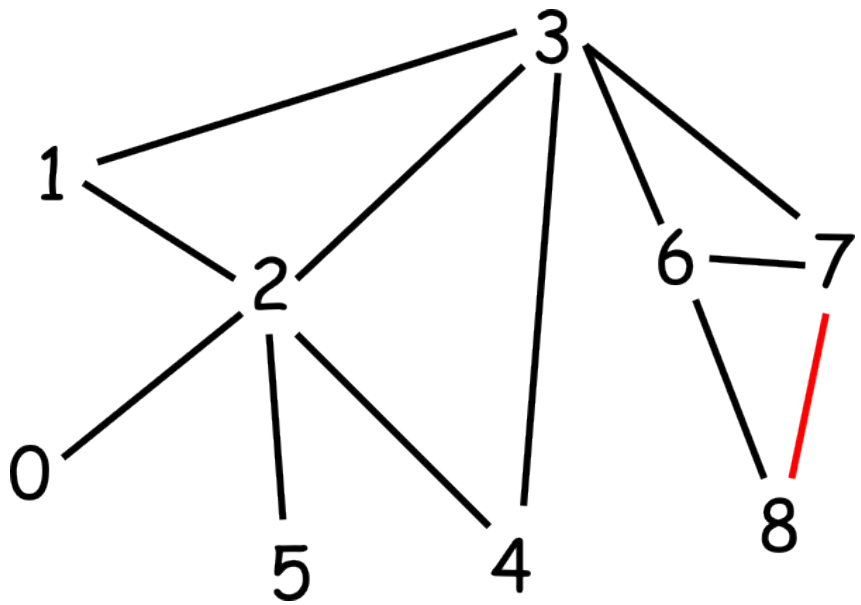
Komprimovaný graf

...

6 _

Pravdepodobnostná tabuľka ako pri texte

i	$p(i)$	bitová rep.	$R(i)$
7	0.3	00	2
3	0.3	01	2
0	0.05	1000	4
5	0.05	1001	4



- šírenie aktivácie
- najkratšia cesta

Symbolický príklad komprimovanej reprezentácie

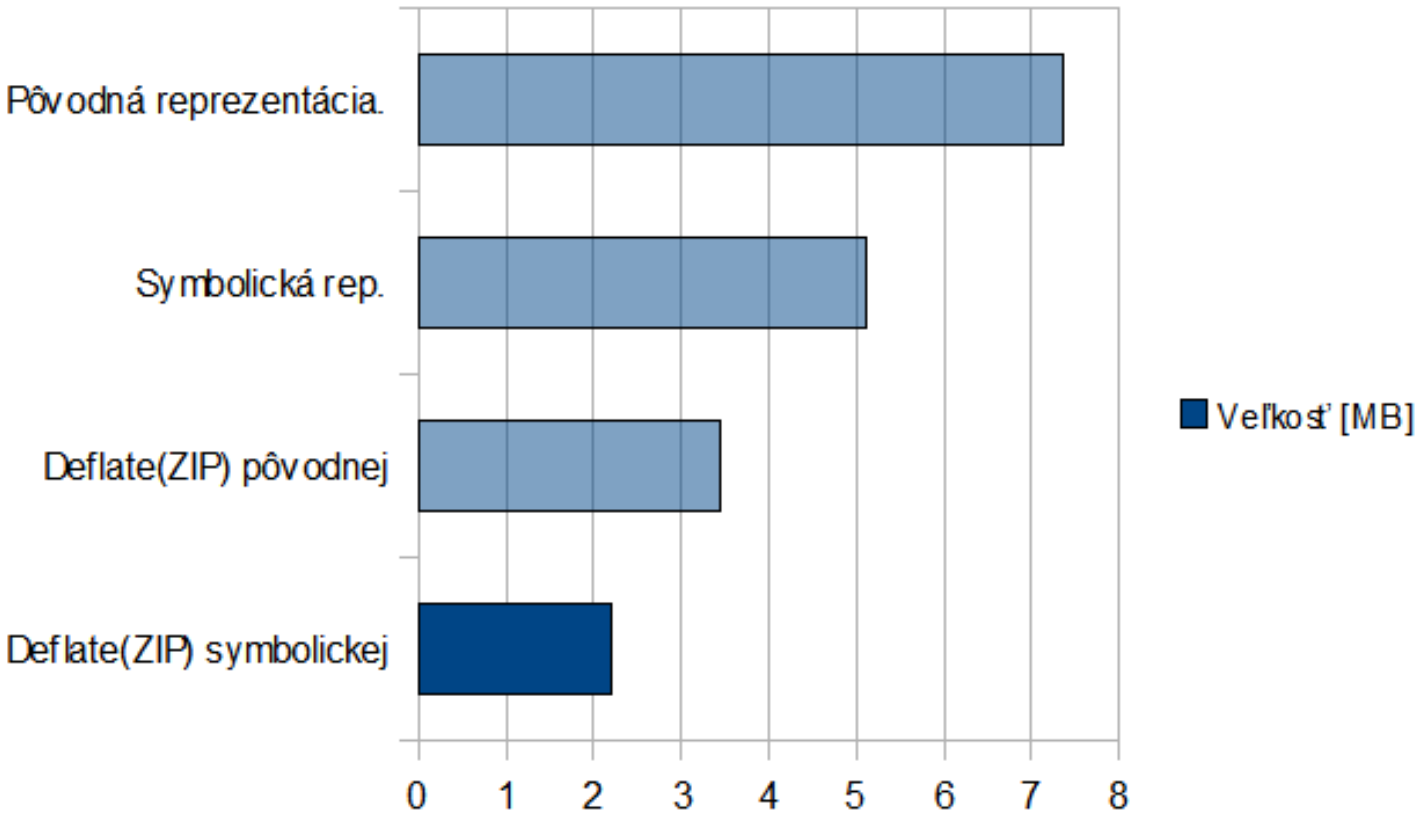
Pôvodná reprezentácia

```
...
20248 9426 22452 24673 5935 20423 13148 11079 22460
14363 33817 14387 14347 5937
29472 50562 14387 36689 5937
16410 27676 5938 17121 19182 47423 13518
...
```

Komprimovaná reprezentácia

```
...
20248 860 39 25771 9 11 7 8 15
14363 1052 47 3 16
29472 50573 16100 101 103
16410 50 8 4 4 5 24
...
```

Symbolický príklad komprimovanej reprezentácie



Záverom

Ďalšia práca

- operácie nad komprimovanou reprezentáciou bez úplnej dekomprimácie
- objavovanie štruktúr/vlastností sociálnych sietí
- analýza vhodných pravdepodobnostných modelov (šírenie aktivácie, random walks with restart, dĺžka cesty, ...)
- analýza kódovaní
- priestorová efektívnosť reprezentácie, časová efektívnosť algoritmov
- integrácia s existujúcimi nástrojmi a ľuďmi ;-)

- Otázky? Diskusia?