Analýza log záznamov pomocou MBA

Predspracovanie záznamov 1

- Log záznamy- server: www.einnews.com
- Pôvodný počet logov 1 500 000 !!!

- Tvar log záznamu :

ny8.shopwiki.com - - [01/Apr/2005:00:00:00 -0600]
"GET /tell_a_friend.php?url=%2Falgeria%2Fnewsfeed-**Business**& HTTP/1.1" 200 8123 "-"
"ShopWiki/1.0 (+http://www.shopwiki.com/)"
"ny8.shopwiki.com.18468111233519992 9" out 1

- Pomocou bash skriptu záznamy pretvorené na
  - Main_input so záznamami tvaru :
  - **ny8.shopwiki.com** **Business**
  - zostalo **2197** záznamov !!!

Predspracovanie záznamov 2

1. Odstránenie logov generovaných robotmi
2. Vygenerovanie main_input : kto kateg
3. Získanie všetkých kategorií (29)
4. Získanie všekých kategorií, na ktoré klikal daný užívateľ
5. Premenovanie nákupov na čísla
6. Vytvorenie „nákupu užívateľa“

„Nákup “ alebo kto kam klikal

- Jeden riadok obsahuje všetky kategórie, na kt. klikal jeden človek, jeho identita už nie je zaujímavá
- 17 19 21 22 24 29 3 4 7 8 9

  1 10 11 13 14 2 27

  12 25 28

  1 26

  .....

Ukážka prečíslovaných kategórii

- 10 Elections
  11 Energy
  12 Environment
  13 EuropeanUnion
  .....

Vytvorenie matice klikania

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	...
1	1	0	0	0	0	1	0	1	1	0	1	1	0	0	0	0	0	0	1	0	1	1	1	
2	1	0	1	0	1	0	1	1	1	1	0	0	0	0	1	1	0	1	0	1	0	1	0	
3	0	1	0	1	1	1	0	1	1	0	1	1	0	0	0	0	0	0	1	1	0	1	0	

- **MBA tovar <=> kategória**
- **MBA nákup** (zložený z tovarov) <=>
  záznam užívateľa(záznamy o kategóriách na ktoré klikal)

•Riadky odpovedajú kategóriám
•Stĺpce odpovedajú záznamom užívateľov

•Bitový vektor odpovedajúci i-tej kategórii (riadku) má na j-tom mieste **1**, ak užívateľ túto kategóriu prehliadal *inak* **0**

MBA

- Kým na vstupe ešte máme ntice (kategórií)
  - Generuj ntice
  - Spočítaj výskyt všetkých kategórii z ntice u všetkých užívateľov pomocou bitového andu
  - Spočítaj podporu => porovnaj s prahom_podpory
  - Spočítaj spoľahlivosť => porovnaj s prahom_spol.
  - Spočítaj zlepšenie >1 (OK), <1 (neguj),
    =0(nezávisle kategórie)
  - Generuj kombinácie => pridaj na vstup, ktoré vyhoveli
  - Generuj pravidlá

- float spolahlivost(ntica pole){

kopia=pole

return(cetnost(kopia)/cetnost(pole)*100);}

- float zlepsenie(ntica pole){

kopia= pole; posledny=pop(pole);

return(cetnost(kopia)*pn/cetnost(pole)/cetnost(pom))*100}

- float podpora(ntica pole){

return(cetnost(ntica)/pn);

}

- Četnosť(ntica) vracia koľko bolo užívateľov , ktorí klikali na vstupnú kombináciu kategórii

Výsledek

HumanRights => Crime	**R: 21 S: 0.724137931034483 C: 0.7** I: 0.52051282051282
Crime => HumanRights	**R: 21 S: 0.724137931034483** C: 0.538461538461538 I: 0.52051282051282
Economy => Crime	R: 17 S: 0.586206896551724 C: 0.548387096774194 I: 0.407775020678247
Crime => Economy	R: 17 S: 0.586206896551724 C: 0.435897435897436 I: 0.407775020678247
Crime => Banking	R: 16 S: 0.551724137931034 C: 0.41025641025641 I: 0.475897435897436
Crime => Accidents	R: 16 S: 0.551724137931034 C: 0.41025641025641 I: 0.440645773979107
Business => Automotive	R: 16 S: 0.551724137931034 C: 0.615384615384615 **I: 1.04977375565611**
Banking => Crime	R: 16 S: 0.551724137931034 C: 0.64 I: 0.475897435897436
Automotive => Business	R: 16 S: 0.551724137931034 **C: 0.941176470588235 I: 1.04977375565611**
Accidents => Crime	R: 16 S: 0.551724137931034 C: 0.592592592592593 I: 0.440645773979107

Záver

- Overili sme tvrdenie, že na predspracovanie dát sa venuje skoro 70% času
- Spracované údaje tvorili 1/681 pôvodných dát
- Lepšie voliť dáta, ktoré sa v jednej tranzakcii viackrát opakujú
- Zjednodušenie výpočtu použitím bitových vektorov
- Matica četností nebola použitá

- Ales Zoulek
  - ales.zoulek@gmail.com
  - ICQ: 82647256

- Jana Šefčíková
  - neollie@gmail.cz
  - ICQ 172649127