Dobývání znalostí

Analýza serveru fotoaparat.cz

Andrej Kruták, 2008

Cieľ, motivácia I

- Na fa.cz je rozsiahla galéria fotografií
  - nie je však určená na fotky typu rajce.net
  - cieľom zdokonalovanie sa autorov, hľadanie chýb
  - obmedzenia na vkladanie/hodnotenie fotiek
    - 2 fotky za týždeň, prvý mesiac po registrácii zakázané hodnotenie starších fotiek
    - => obmedzenie ”zamorenia”, multiúčtov atď.
  - napriek obmedzeniam diskusie o objektívnosti hodnotenia...

Cieľ, motivácia II

- Systém hodnotenia:
  - 1-7 bodov (radšej nevystaviť – výnimočná fotka)
  - -1 nevhodné fotky (premazávané, nebudeme s nimi počítať)
  - Je možné zakázať hodnotenie (v takom prípade sa sústredíme iba na počet zobrazení)
- Motivácia:
  - Zistiť, či/ako závisí hodnotenie od okolností vloženia fotky
    - Čas vloženia, priemerná farba, hodnotenia udelené užívateľom

Získanie dát

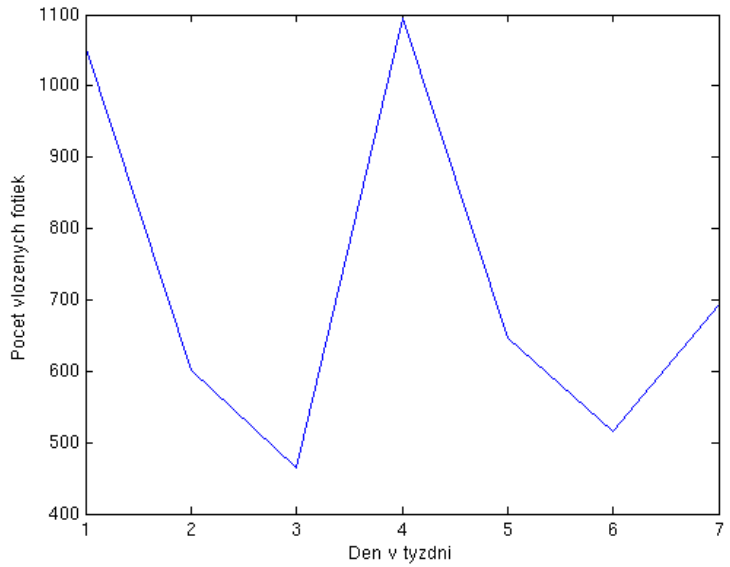
- Brute force – stiahnutie webov popisujúcich cca. 1/20 fotografií za cca. posledný rok
  - časť z toho odpad (zmazané fotografie)
  - rozparsovanie (nechutného) html stromu
    - bash+python [11kB kódu :-)]
  - získanie dát o autoroch, podobne ako o fotkách
  - = ~5000 fotografií, ~2700 autorov
    - traffic ~400 MB :-)

Príprava dát I

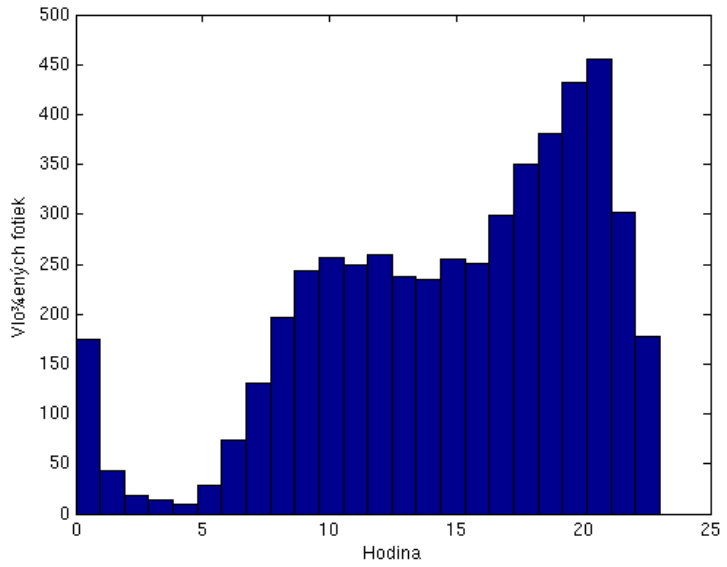
- Použité fotoaparáty – názvy nepoužiteľné
  - kompakty-profesionálne aparáty => 0-3
- Z dátumu použijeme len:
  - deň v roku
  - deň v týždni
  - hodina

Základná štatistika I

- Zaujímavé informácie už v rozparsovaných dátach
  - Kedy sa vkladajú fotky



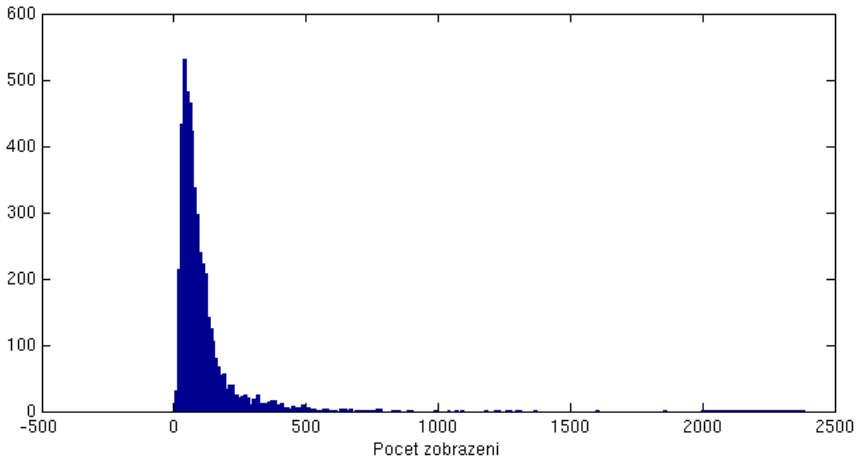
deň v týždni

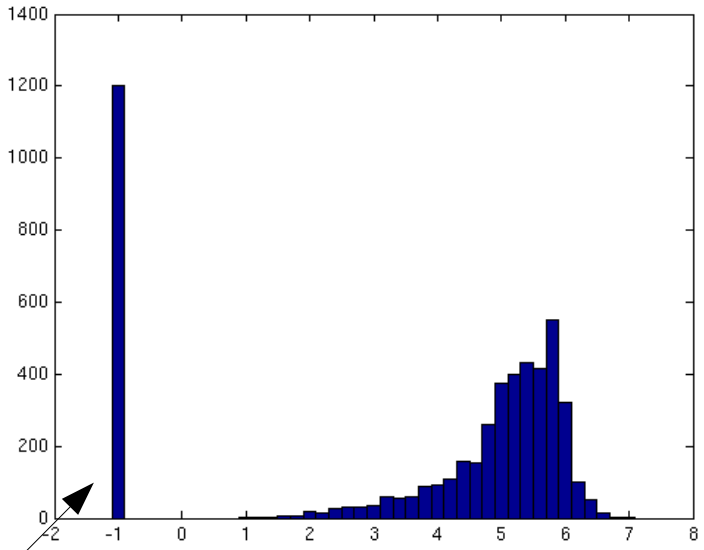


hodina

Základná štatistika II

- Zaujímavé informácie už v rozparsovaných dátach
  - Ako vyzerá priemerná návštevnosť / hodnotenie

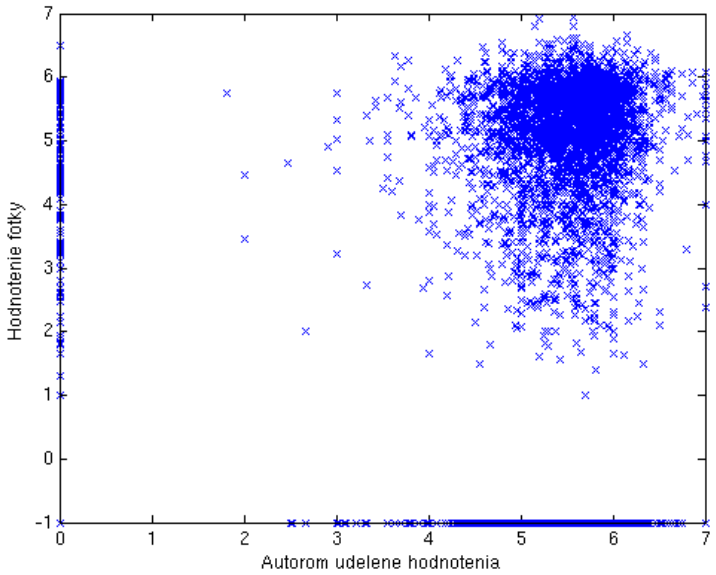




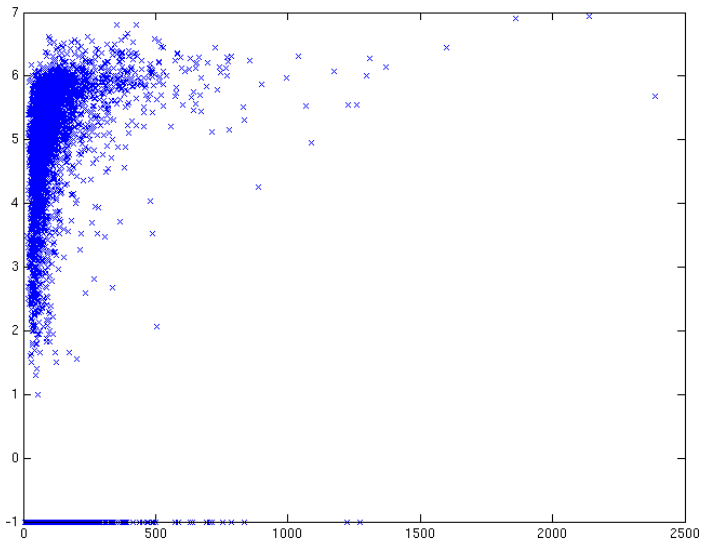
nehodnotené fotografie

Nepohodlná pravda I

- Grafy závislostí už také pekné nie sú...



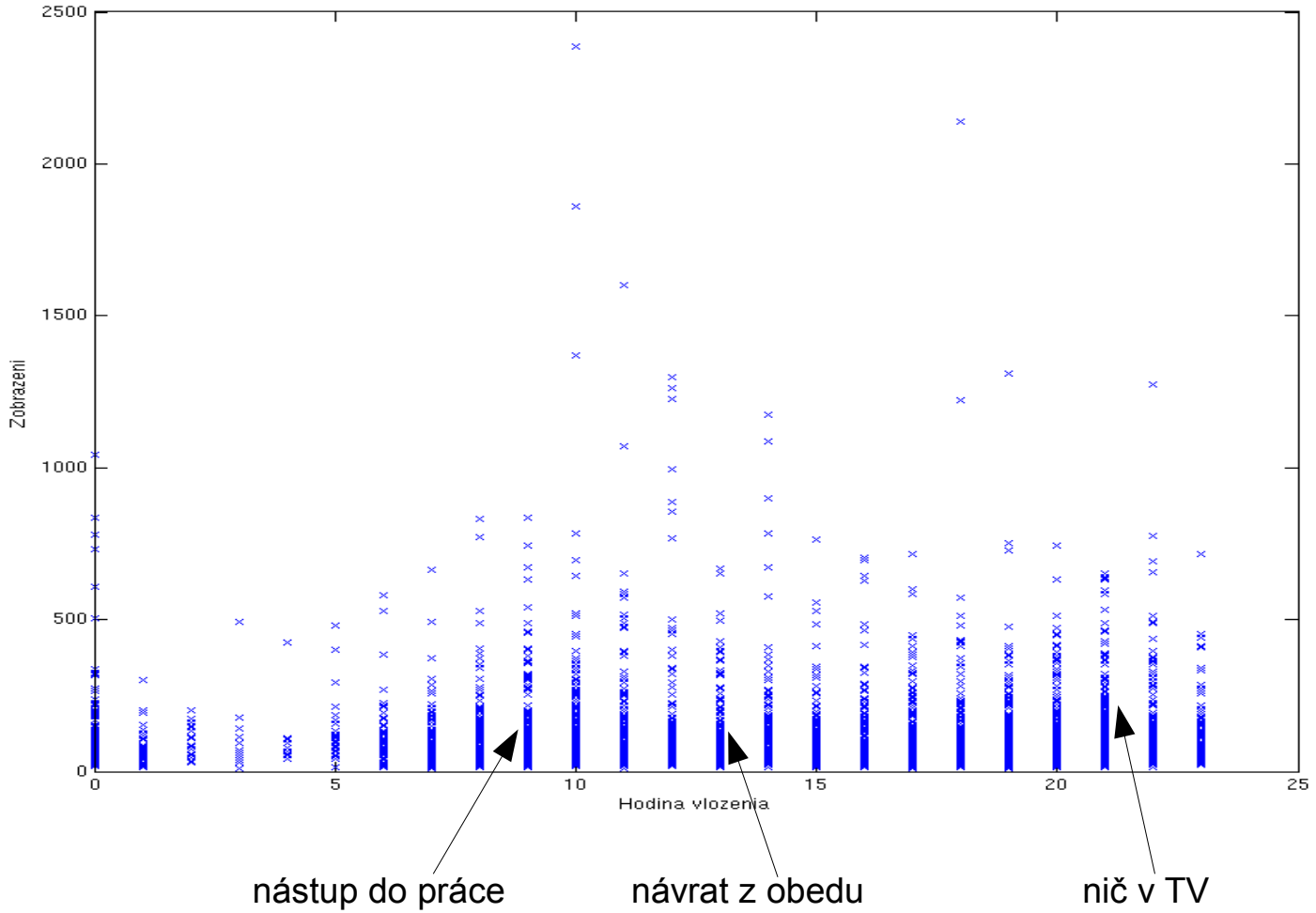
udelené vs. prijaté hodnotenia



hodnotenie vs. zobrazenia

Nepohodlná pravda II

- Jeden trošku rozumnejší graf



NowWhat...

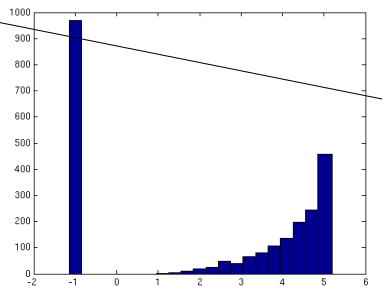
- Je zrejmé, že dáta sú výrazne korelované
  - nik normálny nebude vkladať fotky o 3 v noci :-)
  - získané body vs. priemerné hodnotenie
  - vysoké hodnotenie na serveri svedčí o
    - vysokej úrovni autorov – vkladajú len dobré fotky
    - nízkej úrovni hodnotiacich
    - o niečom inom
- Snažiť sa spočítať výslednú známku bez analýzy fotky je samozrejme utópia...

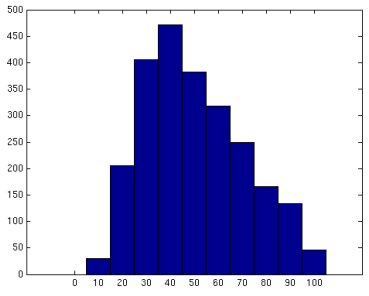
Príprava dát II

- Použijeme len časť z dobytých atribútov
- Je potrebné aj skonvertovať hodnoty atribútov
  - parametre obrázku → veľký/malý/neštandardná veľkosť | farba/čb
  - hodina+deň vloženia →
    0-6 | 7-12 | 13-19 | 20-23 + po-ut | st-pi | víkend
  - autorom v priemere prijaté / udelené hodnotenia →
    prijaté>udelené (0-0.8) | prijaté ~= udelené | prijaté < udelené (1.2-...)
  - počet prehliadnutých fotiek → 0-100 | 101-1000 | 1001-10k | viac
  - počet zobrazení → 0-20 | 21-50 | 51 – 100 | 101 - ...
  - výsledná známka → <1-3) | <3-4) | <4-5) | <5-6) | <6-7>

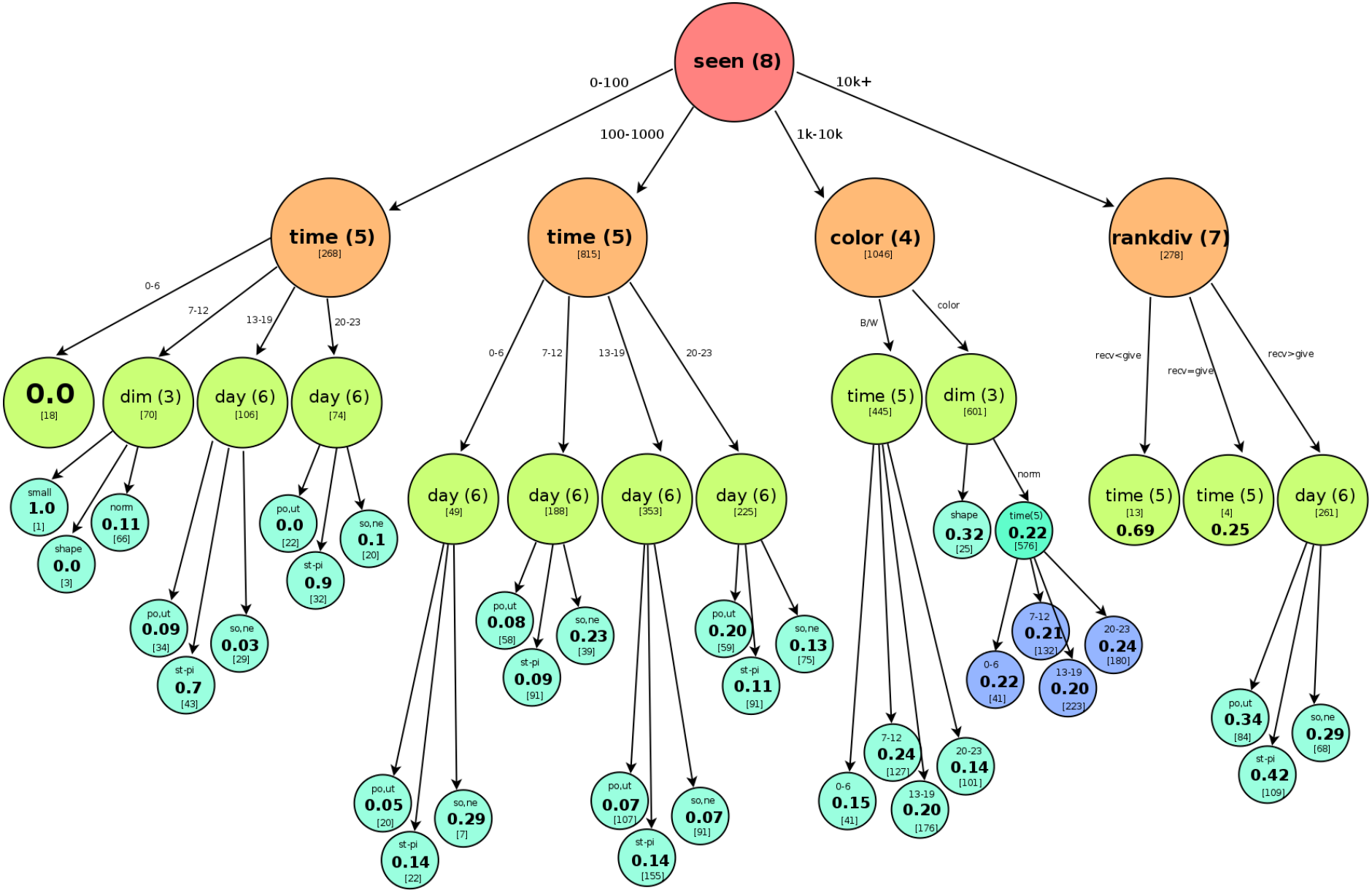
Spracovanie

- Na samotné dobytie použijeme rozhodovacie stromy (mierne modifikovaný ID3 algoritmus)
  - Medzi jednotlivými atribútmi sú určite závislosti, bolo by teda možno zaujímavé použiť sofistikovanejšiu metódu – Bayesove / neurónové sieťe...
- Cieľový atribút: f(výsledná známka+# zobrazení)
  - vyradíme iff (známka>5.2 || #>100) => 2400 záznamov (zrejme ”neprefesionálna vrstva”)
- => výpočet :-)





Výsledok :-)


seen (8)
0-100
100-1000
1k-10k
10k+
time (5) [268]
time (5) [815]
color (4) [1046]
rankdiv (7) [278]
0-6
7-12
13-19
20-23
0.0 [18]
dim (3) [70]
day (6) [106]
day (6) [74]
small 1.0 [1]
norm 0.11 [66]
shape 0.0 [3]
po,ut 0.09 [34]
so,ne 0.03 [29]
st-pi 0.7 [43]
po,ut 0.0 [22]
so,ne 0.1 [20]
st-pi 0.9 [32]
day (6) [49]
day (6) [188]
day (6) [353]
day (6) [225]
po,ut 0.05 [20]
so,ne 0.29 [7]
st-pi 0.14 [22]
po,ut 0.08 [58]
so,ne 0.23 [39]
st-pi 0.09 [91]
po,ut 0.07 [107]
so,ne 0.07 [91]
st-pi 0.14 [155]
po,ut 0.20 [59]
so,ne 0.13 [75]
st-pi 0.11 [91]
B/W
color
time (5) [445]
dim (3) [601]
0-6 0.15 [41]
7-12 0.24 [127]
13-19 0.20 [176]
20-23 0.14 [101]
shape 0.32 [25]
norm
time(5) 0.22 [576]
0-6 0.22 [41]
7-12 0.21 [132]
13-19 0.20 [223]
20-23 0.24 [180]
recv<give
recv=give
recv>give
time (5) [13] 0.69
time (5) [4] 0.25
day (6) [261]
po,ut 0.34 [84]
so,ne 0.29 [68]
st-pi 0.42 [109]

Interpretácia

- základ je **učiť sa od druhých**
  - užívatelia s najmenej prezretými fotkami majú obecne najmenej úspešných fotografií
  - ”**začiatočníci**” by mali vkladať fotky **skôr večer**
- pokročilí vkladajú podobne veľa ČB i farebných
  - najúspešnejšie - atypické farebné fotky (ale málo)
- najaktívnejší – najväčšia úspešnosť
  - výrazne väčšia úspešnosť ”rozdávačov bodov” :-(
  - inak najlepšie vkladať fotku v **stredu-piatok**

ToDo

- Čo by sa dalo zlepšiť:
  - množstvo dát
  - brať do úvahy ďalšie údaje
    - kategórie fotografií
    - detekovať rámiky
    - hľadať multiúčty :-)
  - dni: po+st | ut+št+pi | so+ne (po uzávierke :-))
  - požiadať o dáta priamo od prevádzkovateľa
  - použitý algoritmus (ktorý by bral do úvahy spojitosť)
    - korelácie...

Fin

Ďakujem za pozornosť :-)