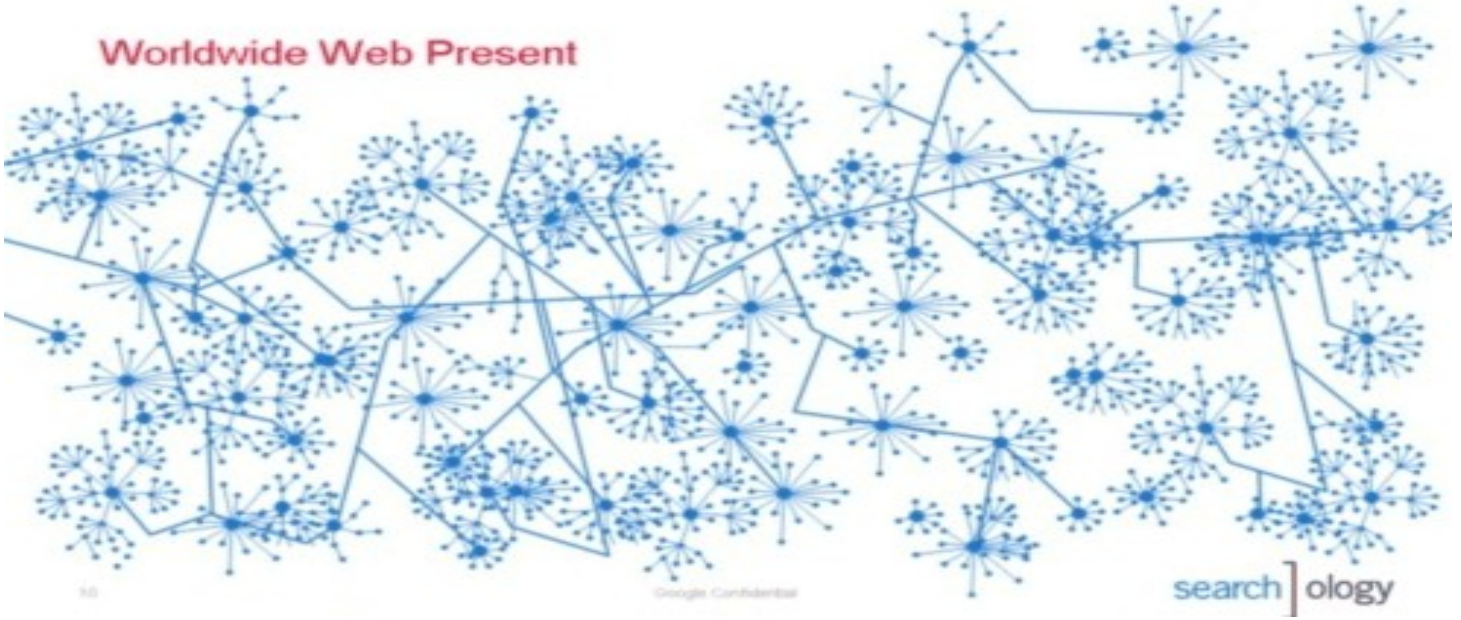
Web Mining

Tomáš Kramár

Web Mining

Tomáš Kramár

Web



Web Mining

- Web structure mining
- Web content mining
- Web usage mining

Web Structure Mining

- Ktoré webové stránky sú dôležité?
- Ako sú stránky na webe prepojené? Existujú nejaké zhluky?

Web Content Mining

- Ktoré stránky sú si navzájom podobné?
- Akej téme sa stránky venujú?
- Aké objekty sa vyskytujú na stránkach?
- Aké názory prezentujú?

Web Usage Mining

- Koľko ľudí videlo stránku?
- Existujú skupiny (komunity) podobných používateľov?
- Ako ľudia používajú túto stránku?
- Existujú nejaké vzory používania?

Obsah

Vzory používania

Tomáš	Šefčovič: Slováci ukázali, že vedia byť aj euroskeptickí
Dušan	Líbya prehodnocuje ropné kontrakty
Mišo	Umravniť zlých žiakov môže dohoda
Tomáš	Cibulková postúpila v Moskve suverénne do finále
Mišo	Tréner milánskeho AC: Chceme ísť vyššie
Jano	Tréner milánskeho AC: Chceme ísť vyššie
Dušan	Líbya prehodnocuje ropné kontrakty
Jano	Umravniť zlých žiakov môže dohoda
Tomáš	Tréner milánskeho AC: Chceme ísť vyššie
Dušan	Cibulková postúpila v Moskve suverénne do finále
Tomáš	Umravniť zlých žiakov môže dohoda
Tomáš	Líbya prehodnocuje ropné kontrakty

Vzory používania

Ak si niekto prečítal článok

“Tréner milánskeho AC: Chceme ísť vyššie”

tak si prečítal aj článok

“Cibulková postúpila v Moskve suverénne do finále”

Vzory používania

Ak si niekto prečítal článok

“Tréner milánskeho AC: Chceme ísť vyššie”

tak si v **90% prípadov** prečítal aj článok

“Cibulková postúpila v Moskve suverénne do finále”

Vzory používania

Tomáš	A, B, C, D
Mišo	C, D, E, D, A, B
Jano	F, G, E, C, F, B, F, A
Dušan	H, I, J, A, K, B, F
Jožko	I, J, K, H, E, D

Vzory používania

Tomáš	**A**, **B**, C, D
Mišo	C, D, E, D, **A**, **B**
Jano	F, G, E, C, F, **B**, F, **A**
Dušan	H, I, J, **A**, K, **B**, F
Jožko	I, J, K, H, E, D

Ak **A** tak **v 100%** prípadoch aj **B**.

Vzory používania

Tomáš	A, B, C, D
Mišo	C, D, E, D, A, B
Jano	F, G, E, C, F, B, F, A
Dušan	**H**, I, J, A, **K**, B, F
Jožko	I, J, K, C, E, D

Ak **H** tak **v 100%** prípadoch aj **K**.

Vzory používania

Tomáš	**A**, **B**, C, D
Mišo	C, D, E, D, **A**, **B**
Jano	F, G, E, C, F, **B**, F, **A**
Dušan	**H**, I, J, **A**, K, **B**, F
Jožko	I, J, K, H, E, D

Ak **A** tak **v 100%** prípadoch aj **B** a **stalo sa to v 80%** prípadov.

Ak **H** tak **v 100%** prípadoch aj **K** a **stalo sa to v 20%** prípadov.

Vzory používania

Nájdi všetky vzory, ktoré sa stali v **X%** prípadov,
kde pravdepodobnosť implikácie je väčšia ako **Y%**.

Asociačné pravidlá

antecedent → consequent

[support = 10%, confidence = 80%]

Nájdi všetky vzory, ktoré sa stali v **minsup** prípadov, kde pravdepodobnosť implikácie je väčšia ako **minconf**.

Vstupný formát

- Transakcie

Tomáš	Šefčovič: Slováci ukázali, že vedia byť aj euroskeptickí
Mišo	Tréner milánskeho AC: Chceme ísť vyššie
Mišo	Tréner milánskeho AC: Chceme ísť vyššie
Tomáš	Cibulková postúpila v Moskve suverénne do finále

Tomáš	Šefčovič: Slováci ukázali, že vedia byť aj euroskeptickí, Cibulková postúpila v Moskve suverénne do finále
Mišo	Tréner milánskeho AC: Chceme ísť vyššie, Tréner milánskeho AC: Chceme ísť vyššie

Zopakovanie

- Kvalita/cennosť pravidla
  - support (podpora) – ako často sa vzor nachádza v dátach
  - confidence (spoľahlivosť) – ako často je aplikovateľná dôsledková časť pravidla

Nájdite 4 najcennejšie vzory

Tomáš	Šefčovič: Slováci ukázali, že vedia byť aj euroskeptickí
Dušan	Tréner milánskeho AC: Chceme ísť vyššie
Mišo	Tréner milánskeho AC: Chceme ísť vyššie
Tomáš	Cibulková postúpila v Moskve suverénne do finále
Mišo	Cibulková postúpila v Moskve suverénne do finále
Jano	Cibulková postúpila v Moskve suverénne do finále
Dušan	Umravniť zlých žiakov môže dohoda
Jano	Tréner milánskeho AC: Chceme ísť vyššie
Tomáš	Tréner milánskeho AC: Chceme ísť vyššie
Dušan	Šefčovič: Slováci ukázali, že vedia byť aj euroskeptickí
Mišo	Šefčovič: Slováci ukázali, že vedia byť aj euroskeptickí

Vzory používania

Tomáš	Šefčovič: Slováci ukázali, že vedia byť aj euroskeptickí Cibulková postúpila v Moskve suverénne do finále Tréner milánskeho AC: Chceme ísť vyššie
Dušan	Tréner milánskeho AC: Chceme ísť vyššie Umravniť zlých žiakov môže dohoda Šefčovič: Slováci ukázali, že vedia byť aj euroskeptickí
Mišo	Tréner milánskeho AC: Chceme ísť vyššie Cibulková postúpila v Moskve suverénne do finále Šefčovič: Slováci ukázali, že vedia byť aj euroskeptickí
Jano	Cibulková postúpila v Moskve suverénne do finále Tréner milánskeho AC: Chceme ísť vyššie

Vzory používania

Tomáš	A, B, C
Dušan	C, D, A
Mišo	C, B, A
Jano	B, C

Vzory používania

Tomáš	A, B, C
Dušan	C, D, A
Mišo	C, B, A
Jano	B, C

$A \rightarrow C$ [support = 3/4, confidence = 3/3]

$B \rightarrow C$ [support = 3/4, confidence = 3/3]

$A \rightarrow B$ [support = 2/4, confidence = 2/3]

$A \rightarrow B,C$ [support = 2/4, confidence = 2/3]

Hľadanie asociačných pravidiel

1. Vygeneruj všetky frekventované množiny (frequent itemsets), pre ktoré platí *minsup*
   { A, B, C, D }, { A, E, F }, { G, H, I, A }
2. Z frekventovaných množín odvoď pravidlá, pre ktoré platí *minconf*
   A, B → C, D
   A → E
   G, H, I → A

downward (apriori) closure: Ak množina prvkov spĺňa podmienku pre *minsup*, tak ju musí spĺňať aj každá jej podmnožina

Apriori

```
C_1 ← pocetnosti(T)
F_1 ← filtruj-na-minsup(C_1)
for (k = 2; F_k-1 je neprázdna; k++)
    C_k ← generuj-kandidatov(F_k-1)
    foreach t z T
        foreach c z C_k
            if c je obsiahnuté v t
                c.count++
        endfor
    endfor
    F_k ← filtruj-na-minsup(C_k)
endfor

return všetky F_k
```

Generovanie kandidátov

- F3 = {1,2,3}, {1,2,4}, {1,3,4}, {1,3,5}, {2,3,4}

- C4' = {1,2,3,4}, {1,3,4,5}
- C4 = {1,2,3,4}

Vlastnosti Apriori

- Teoreticky exponenciálna zložitosť, vďaka riedkosti dát je ale efektívny
- Level-wise
- Deterministický výstup

- Alternatíva: FP-Growth

Generovanie pravidiel

Ak v rámci frekventovanej množiny platí nejaký dôsledok, tak musia platiť aj všetky jeho podmnožiny.

$A, B \rightarrow C, D$

$A, B, C \rightarrow D$

$A, B, D \rightarrow C$

V RapidMineri nájdite všetky vzory

Tomáš	Šefčovič: Slováci ukázali, že vedia byť aj euroskeptickí
Dušan	Tréner milánskeho AC: Chceme ísť vyššie
Mišo	Tréner milánskeho AC: Chceme ísť vyššie
Tomáš	Cibulková postúpila v Moskve suverénne do finále
Mišo	Cibulková postúpila v Moskve suverénne do finále
Jano	Cibulková postúpila v Moskve suverénne do finále
Dušan	Umravniť zlých žiakov môže dohoda
Jano	Tréner milánskeho AC: Chceme ísť vyššie
Tomáš	Tréner milánskeho AC: Chceme ísť vyššie
Dušan	Šefčovič: Slováci ukázali, že vedia byť aj euroskeptickí
Mišo	Šefčovič: Slováci ukázali, že vedia byť aj euroskeptickí

Kde je problém?

- Supermarket predáva
  - potraviny
  - elektroniku
  - knihy
  - záhradnú techniku

- Chcem nájsť cenné vzory v nákupoch

Rare item problem

{Mixér, Panvica} *[sup = 0.006%]*

{Chlieb, Syr, Vajcia, Žemľa, Mlieko, Cukor, Maslo} *[sup = 0.007%]*

{Chlieb, Mlieko, Panvica} *[sup = 0.006%]*

2 riešenia

- ľahké
- komplikované

Multiple Minimum Support (MIS)

- MIS(Mixér) = 0.006%
- MIS(Pečivo) = 1%

Multiple Minimum Support (MIS)

- minsup pravidla = min(MIS) každého prvku
  - MIS(1) = 10%
  - MIS(2) = 20%
  - MIS(3) = 5%
  - MIS(4) = 6%
- neplatí downward closure
  - {1, 2} [minsup = 10%, sup = 9%]
  - {1,2,3} [minsup = 5%, sup = 6%]

MS-Apriori

- Výber MIS
  - na základe počiatočných početností:
    - MIS(i) = X.sup(i); 0 <= X <= 1
  - rozdeliť dáta do blokov
    - MIS(i) = MBS(block(i))
- Výber nutných položiek
  - nastavenie MIS na > 100%

Sekvenčné vzory

Používateľ	Čas transakcie	Transakcia
1	6.1.2011	A
1	8.2.2011	B
2	1.6.2011	C, D
2	2.6.2011	A
2	3.6.2011	C, E, F, G
3	14.5.2011	A, H, G, I
4	6.7.2011	A
4	9.7.2011	A, E, G, I
4	11.7.2011	J
5	21.6.2011	J

Sekvenčné vzory

Používateľ	Sekvencia
1	< {A}, {B} >
2	< {C, D}, {A}, {C, E, F, G} >
3	< {A, H, G, I} >
4	< {A}, {A, E, G, I}, {J} >
5	< {J} >

1-sekvencie	<{A}>, <{J}>
2-sekvencie	<{A} {E}>, <{A}, {G}>
3-sekvencie	<{A} {E, G}>, <{A, G, I}>

Dolovanie sekvenčných vzorov

- Generalized Sequential Pattern (GPS)
- PrefixSpan
- MS-GPS
- Traversal paths
  - A → B
  - A → C → D → B

Príprava dát

wiki.fiit.stuba.sk:80 67.195.115.36 - - [16/Oct/2011:06:27:39 +0200] "GET /research/seminars/pewe;revision?revision=23 HTTP/1.0" 200 8898 "-" "Mozilla/5.0 (compatible; Yahoo! Slurp; http://help.yahoo.com/help/us/ysearch/slurp)"

Príprava dát

wiki.fiit.stuba.sk:80 **67.195.115.36** - - [16/Oct/2011:06:27:39 +0200] "GET /research/seminars/pewe;revision?revision=23 HTTP/1.0" 200 8898 "-" **"Mozilla/5.0 (compatible; Yahoo! Slurp; http://help.yahoo.com/help/us/ysearch/slurp)"**

Príprava dát

wiki.fiit.stuba.sk:80 67.195.115.36 - - [**16/Oct/2011:06:27:39 +0200**] "GET /research/seminars/pewe;revision?revision=23 HTTP/1.0" 200 8898 "-" "Mozilla/5.0 (compatible; Yahoo! Slurp; http://help.yahoo.com/help/us/ysearch/slurp)"

Príprava dát

wiki.fiit.stuba.sk:80 67.195.115.36 - - [16/Oct/2011:06:27:39 +0200] "GET **/research/seminars/pewe;revision?revision=23** HTTP/1.0" 200 8898 "-" "Mozilla/5.0 (compatible; Yahoo! Slurp; http://help.yahoo.com/help/us/ysearch/slurp)"

Problémy

- Kto?
  - Identifikácia používateľov
    - IP adresa nestačí (proxy)
    - User Agent nemusí stačiť tiež
    - Heuristiky: IP adresa + UA + referrer + topológia
  - Šum – assets a boti
    - Statické zoznamy
- Čo?
  - Chýbajúce záznamy – cachovanie (klient, proxy)

Identifikácia sedení

Tomáš	Šefčovič: Slováci ukázali, že vedia byť aj euroskeptická
Dušan	Tréner milánskeho AC: Chceme ísť vyššie
Mišo	Tréner milánskeho AC: Chceme ísť vyššie
Tomáš	Cibulková postúpila v Moskve suverénne do finále
Mišo	Cibulková postúpila v Moskve suverénne do finále
Jano	Cibulková postúpila v Moskve suverénne do finále
Dušan	Umravniť zlých žiakov môže dohoda
Jano	Tréner milánskeho AC: Chceme ísť vyššie
Tomáš	Tréner milánskeho AC: Chceme ísť vyššie
Dušan	Šefčovič: Slováci ukázali, že vedia byť aj euroskeptická
Mišo	Šefčovič: Slováci ukázali, že vedia byť aj euroskeptická

Identifikácia sedení

Tomáš	Šefčovič: Slováci ukázali, že vedia byť aj euroskeptická
Dušan	Tréner milánskeho AC: Chceme ísť vyššie Umravniť zlých žiakov môže dohoda Šefčovič: Slováci ukázali, že vedia byť aj euroskeptická
Mišo	Tréner milánskeho AC: Chceme ísť vyššie Cibulková postúpila v Moskve suverénne do finále
Jano	Cibulková postúpila v Moskve suverénne do finále Tréner milánskeho AC: Chceme ísť vyššie
Tomáš	Cibulková postúpila v Moskve suverénne do finále Tréner milánskeho AC: Chceme ísť vyššie
Mišo	Šefčovič: Slováci ukázali, že vedia byť aj euroskeptická

Identifikácia sedení

- Časové okno
  - 25,5 minúty
- Rozdelenie na navigačné/obsahové stránky
- Maximálnym dopredným odkazom
- Štatistické jazykové modely

Objavovanie štruktúry webu

- klasické IR – ee
- demokratická štruktúra webu
- hľadanie autorít a hubov
- hľadanie vplyvných dokumentov

HITS

- query-dependent

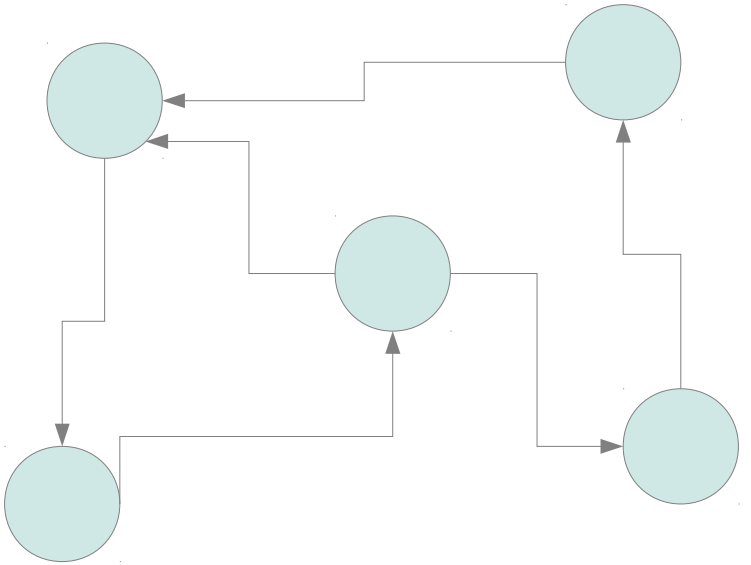
- nájdi top N dokumentov vyhovujúcich dopytom
- pre každý dokument stiahni jeho okolie W
- získaj *base set* S
- Nájdi autority a huby

HITS – nevýhody

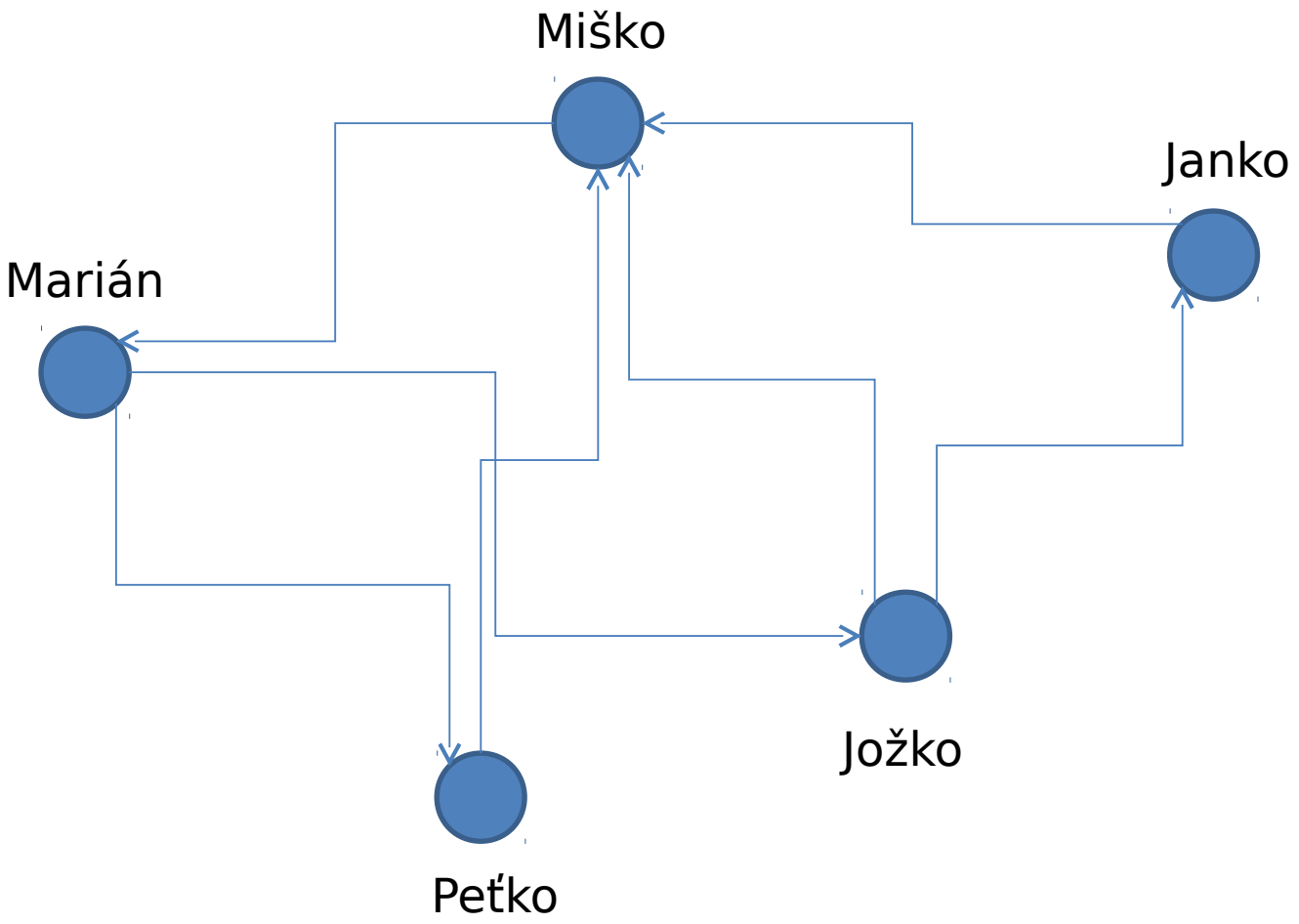
- nevie sa vysporiadať so spamom
- topic-drift
- výkon

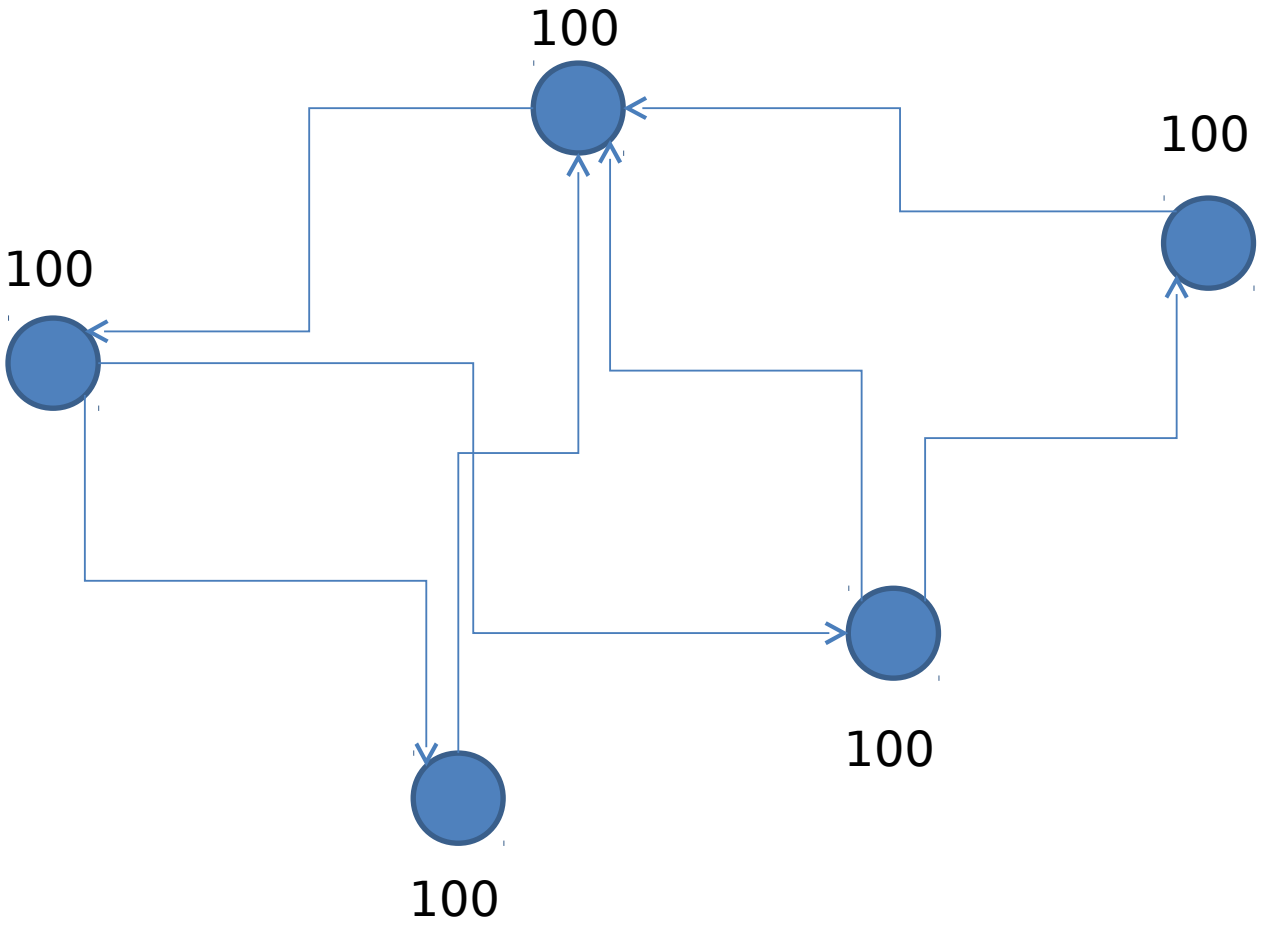
PageRank

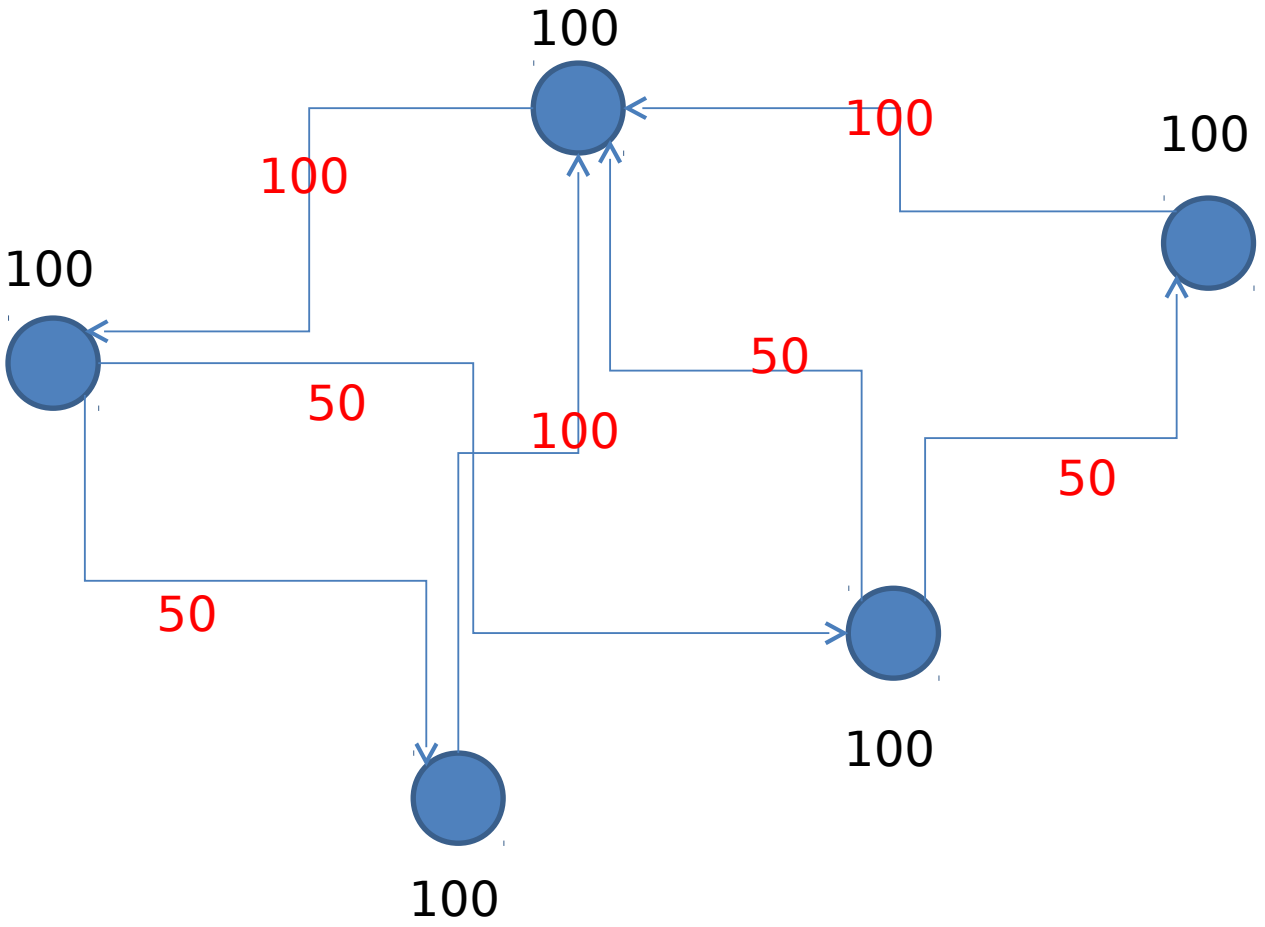
- query-independent
- závislosť hub-autorita

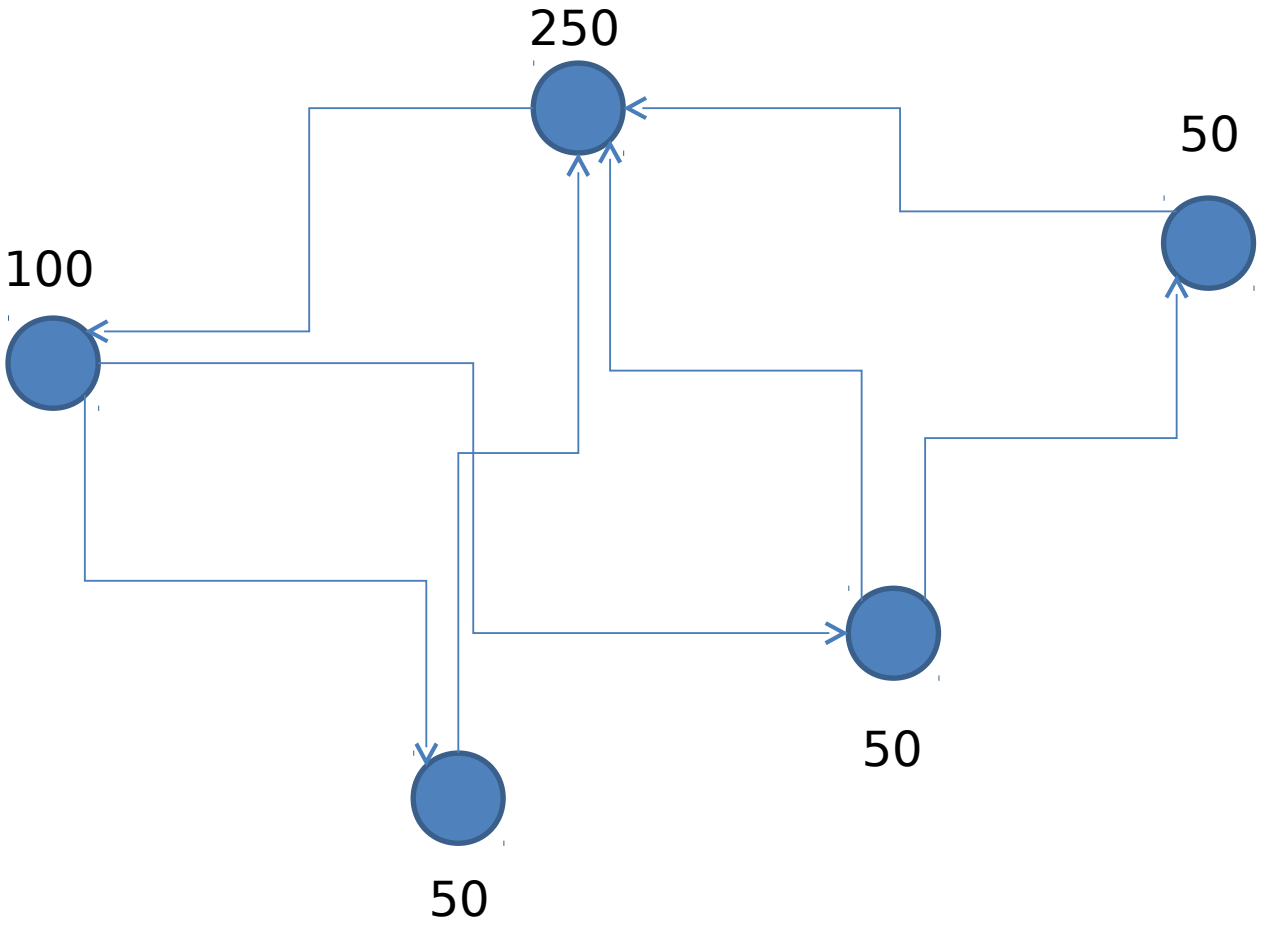


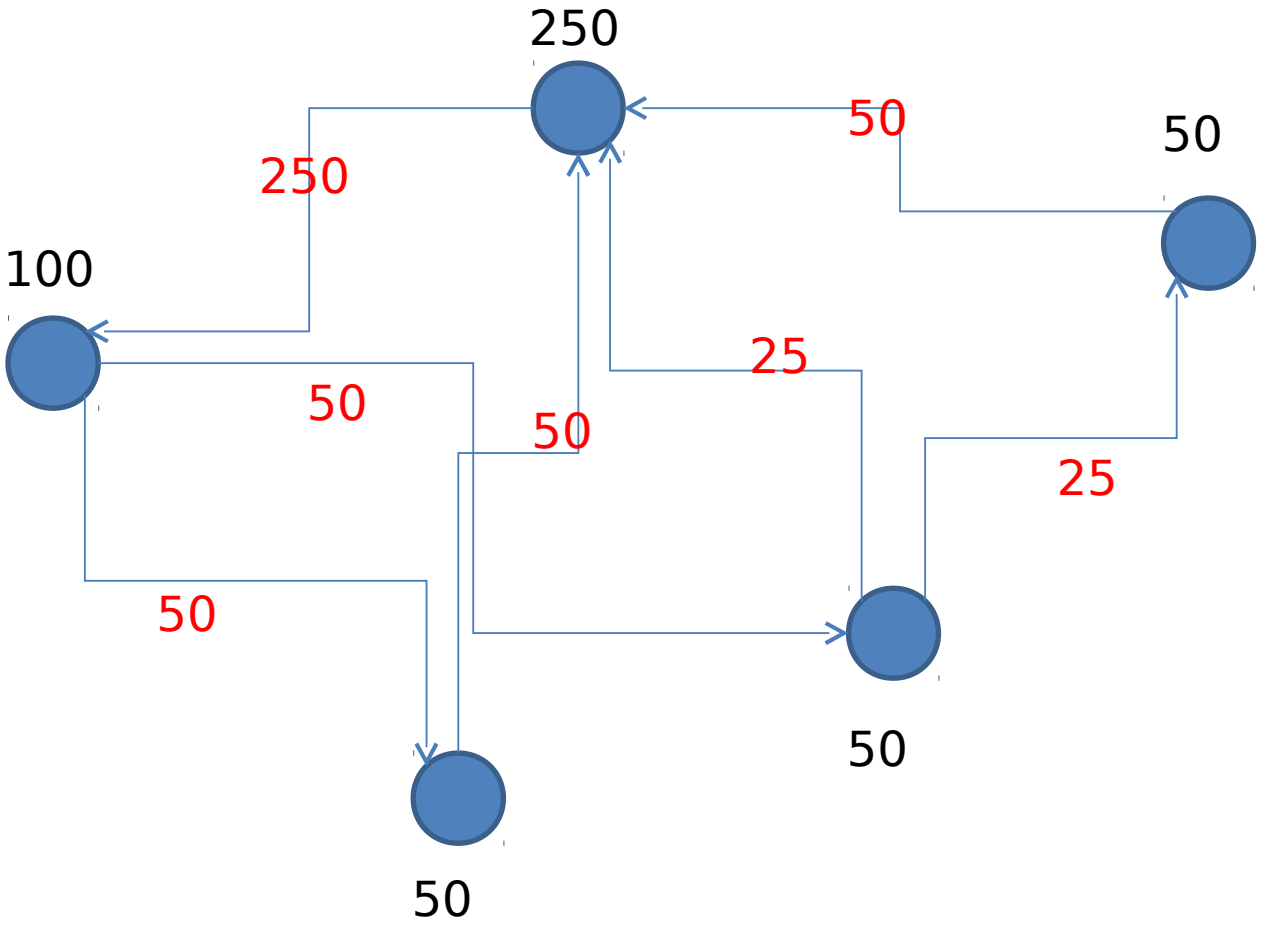
(c) Ján Suchal

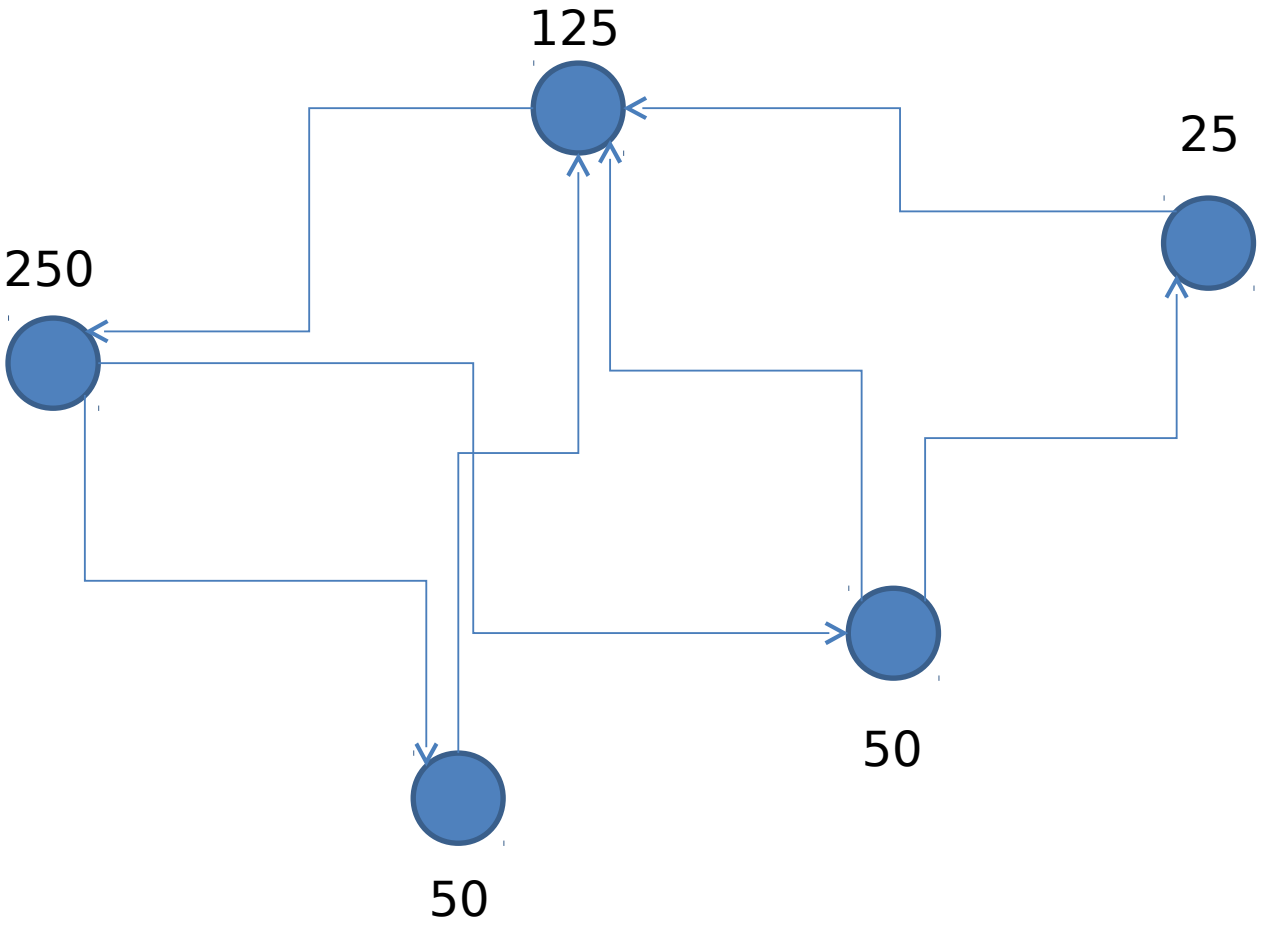


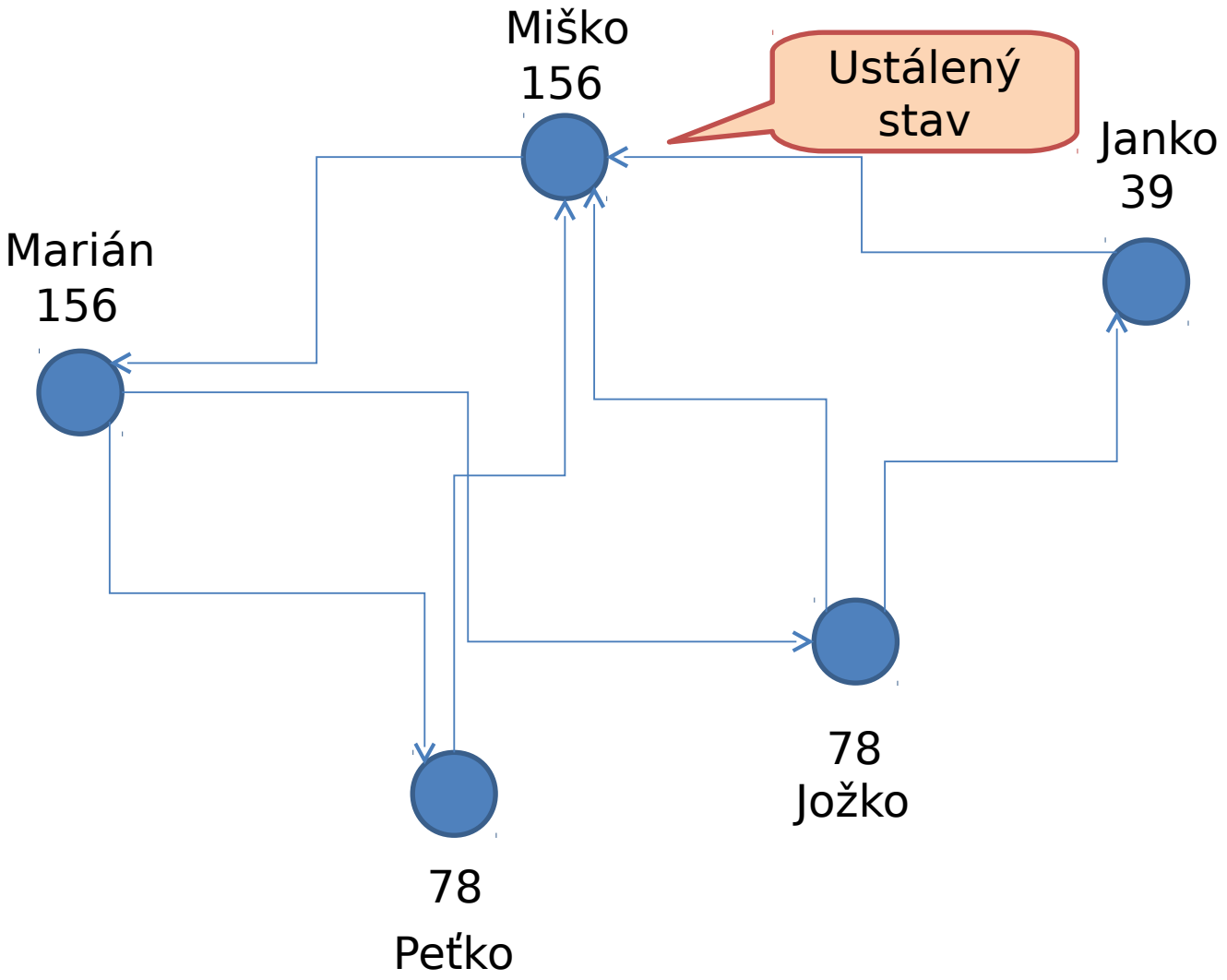

100
100
100
100
100


100
100
100
100
100
100
50
50
50
100
50
100
100


250
50
100
50
50


250
50
50
250
100
25
50
50
25
50
50
50


125
25
250
50
50


Miško
156
Ustálený
stav
Janko
39
Marián
156
78
Jožko
78
Peťko

PageRank

- random surfer model
  - s P(d) skočím po odkaze
  - s P(1-d) skočím mimo
- Vlastnosti
  - anti-spam
  - query-independent (výhoda aj nevýhoda)
  - nezohľadňuje čas
    - Timed PageRank