

Odporúčanie kľúčových slov na základe kontextu sociálnej siete

Tomáš Kramár

Viacznačnosť slov

- Jedno slovo – viac významov („kohútik“, „jaguár“)
- Nie vždy vieme („c strings“) alebo si uvedomujeme že slovo je viacznačné
- Podľa viacerých výskumov je väčšina dopytov krátka (1-3 slová)

Príčina problému

- Vyhľadávače pracujú ako databázy
 - Indexujú
 - Ohodnocujú
 - Zoradujú
- Často krát nájdeme to čo chceme len vďaka tomu, že ten „náš“ význam je „dominantný“

Ako sme to už skúšali riešiť

- Alternatívne ohodnocovanie – Topic-sensitive PageRank
- Preusporadúvanie výsledkov – využitie sociálnych prepojení, UserRank
- Automatic query expansion – analýza lexikálnych afinít, synonymické slovníky, analýza dokumentov na desktope.

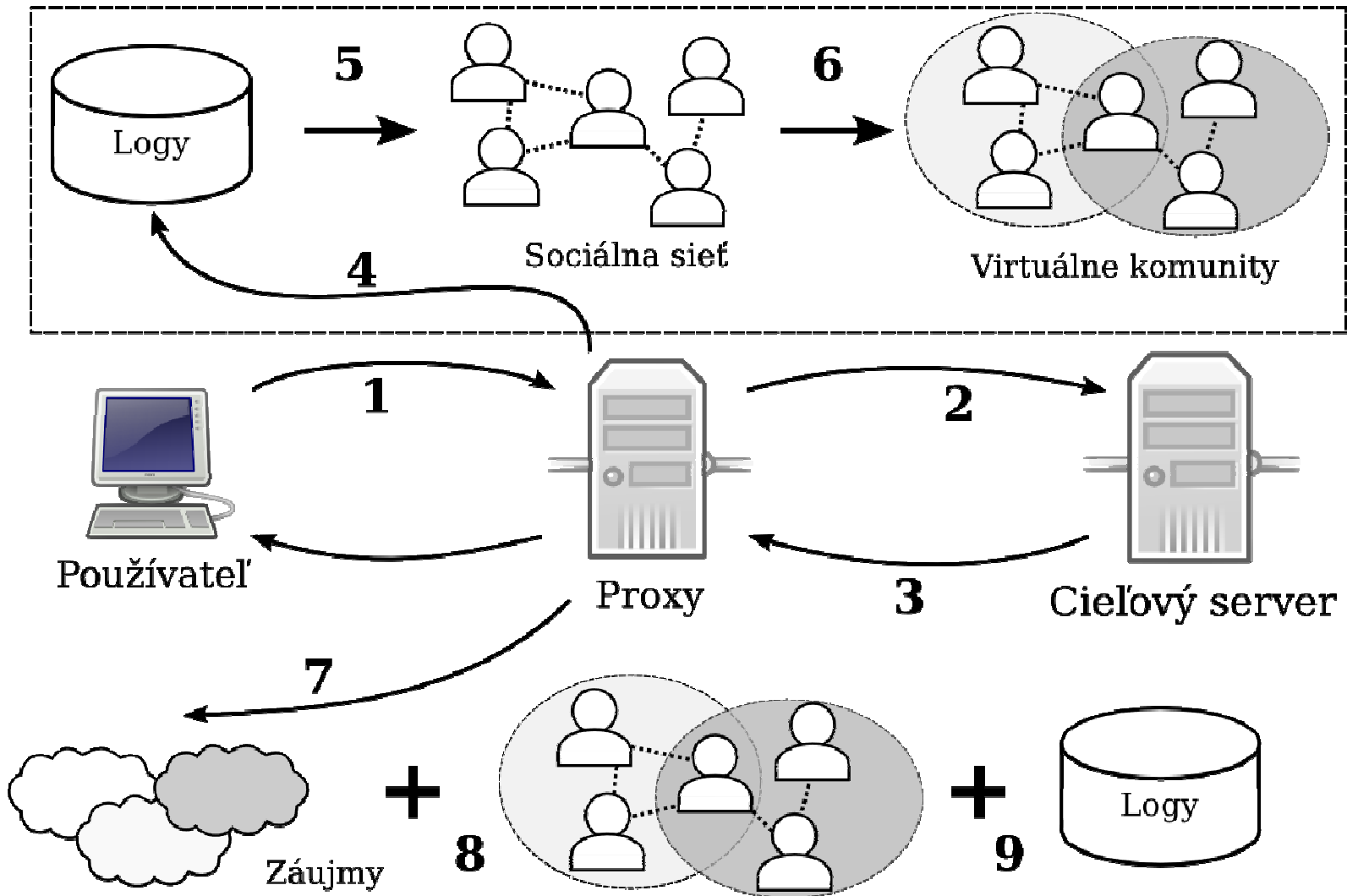
Návrh riešenia

- Ak som nič nenašiel, tak svoj dopyt doplním o ďalšie kľúčové slovo/slová tak, aby som ho zjednoznačil.
- **Prečo by tento krok za mňa nemohol spraviť niekto iný?**

Návrh riešenia

- Zistia sa všeobecné záujmy používateľa z jeho aktivity na webe
- Určia sa skupiny podobných ľudí (komunity)
- Zistí sa aktuálny záujem používateľa
- Podľa aktuálneho záujmu sa určia komunity s podobnými záujmami – tie budú tvoriť kontext
- Z kontextu sa doplnia kľúčové slová

Prehľad procesu expanzie dopytu



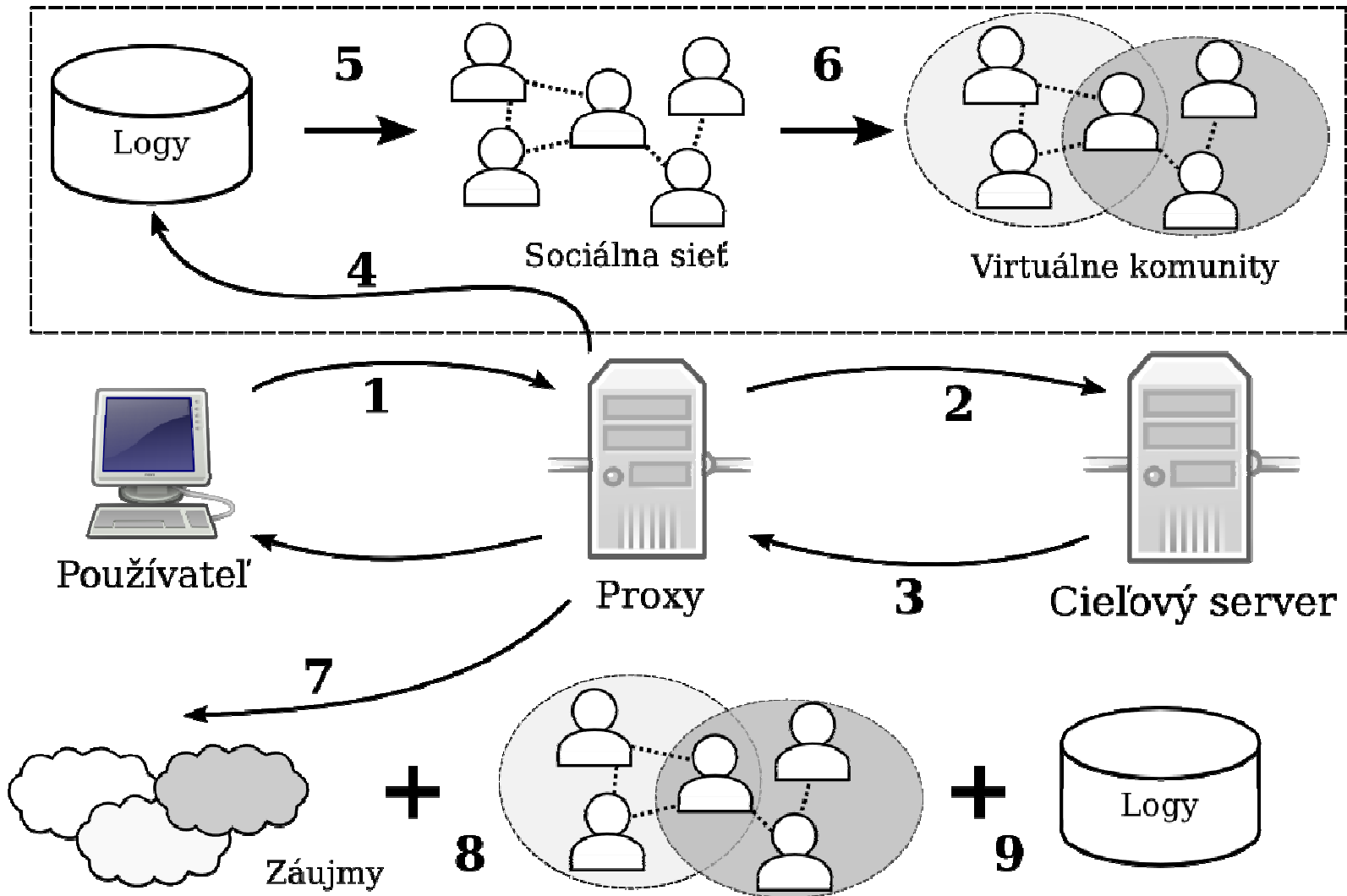
Všeobecné záujmy používatele I.

- Určené aktivitou na webe
- Charakteristický vektor obsahu
 - klíčové slová,
 - tagy,
 - ODP kategória

Tvorba siete a detekcia komunít

- Tvorba sociálnej siete – vrcholy používatelia, hrany váhované, vyjadrujú silu vzťahu medzi používateľmi
- Váha hrany sa určí počítaním navštívených domén, dokumentov a zo zhody vektorov obsahu
- Detekcia komunít šírením aktivácie

Prehľad procesu expanzie dopytu



Určenie aktuálneho záujmu

- *Vektor obsahu* aktuálne prezieraného dokumentu sa porovná s *vektorom obsahu* aktuálnych záujmov – podľa veľkosti prieniku sa nastaví aktuálny záujem
- Ak nenastane žiadny prienik, vytvorí sa nový záujem, neaktívne záujmy sa odstraňujú
- Aktuálny záujem sa namapuje na komunity, získavame kontext

Analýza query-streamov

- Vychádza z toho ako vyhladávame
- Query stream – postupnosť vyhladávaní, ktoré majú spoločné kľúčové slová
 - Jaguar, jaguar speed, jaguar car speed
- Nájdienie úspešných query streamov v kontexte, použitie posledného predefinovania

Analýza výskytov kľúčových slov

- Hľadanie ktoré kľúčové slová sa často vyskytujú spolu
- V rámci kontextu

Experimenty

- Cieľ: overiť či tieto dva prístupy dokážu doplniť dopyt tak, aby vracal relevantnejšie výsledky ako pôvodný dopyt

Query stream analýza

- dáta z AOL – 20 miliónov vyhľadávaní

Pôvodný dopyt	Predefinovaný dopyt
java history	history of java indonesia
jaguar	jaguar cat
jaguar	jaguar car
sphinx	sphinx cat
sphinx	sphinx egypt

Kľúčové slová

- Dáta z proxy servera, 310 dokumentov, 4 používatelia

Pôvodný dopyt	Doplnený dopyt
passenger	passenger apache
fiit	fiit stuba
xml	xml document
branch	branch git

Ďalšia práca

- Využitie implicitného hodnotenia pri generovaní soc. siete
- Obohatenie tagov o ich nadradené kategórie
- Overenie celého procesu na platforme proxy servera

Sumár

- Spojenie sociálnych vzťahov a dopĺňania dopytov – väčšia miera personalizácie
- Dynamické zisťovanie záujmu a mapovanie na komunity
- Kontext vyhľadávania – virtuálne komunity použí vateľov s podobnými záujmami
- Z kontextu
 - Query stream analýza
 - Výskyt kľúčových slov