

Získavanie metadát o vzťahoch a obsahu na webe

Tomáš Uherčík
Ing. Marián Šimko, PhD.

Motivácia

▶ Problém

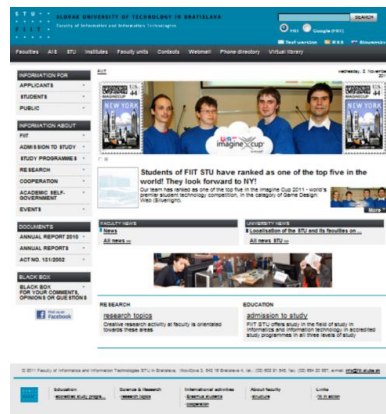
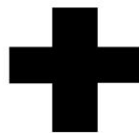
automatické získavanie metadát o internetových zdrojoch (prepojeniach) na webe

▶ Kde hľadať metadáta ?

- Obsah
- Prepojenia
- Záznamy o používaní webu
- Externé zdroje, ktoré obsahujú referenciu na *URL*
 - mikroblogy

Návrh metódy

- ▶ Získanie kľúčových slov z mikrobloggerov
- ▶ Získanie kľúčových slov z textovej analýzy zdroja
- ▶ Výber najlepších z nich



Metadáta

Kľúčové slovo 1

Kľúčové slovo 2

Kľúčové slovo 3

Kľúčové slová z mikrobloggerov

- ▶ Relevancia kľúčového slova:
 - Relevancia v texte pípnutí
 - Text Rank
 - Alchemy API
 - Ohodnotenie používateľa, ktorý pípnutie napísal
 - Tunk Rank

$$TunkRank(X) = \sum_{Y=nasledovníci(X)} \frac{1 + p \cdot TunkRank(Y)}{|nasledovníci(Y)|}$$

kde p je pravdepodobnosť znovu pípnutia



Úpravený Tunk Rank – ARank

- ▶ Tunk Rank obohatený o závislosť od periódy publikovania príspevkov:

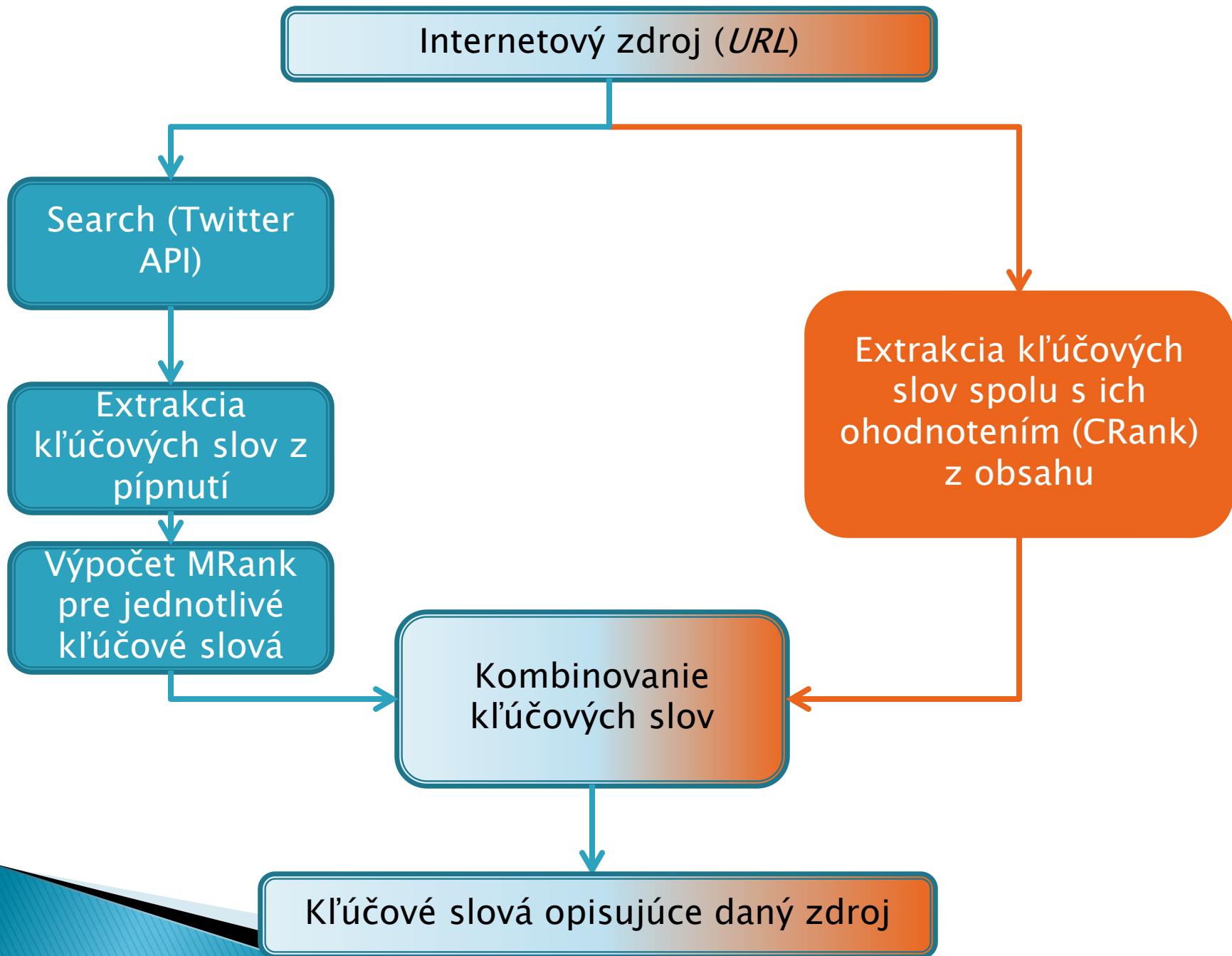
$$ARank(X) = \sum_{Y=\text{nasledovníci}(X)} \frac{1 + \frac{p}{\log(T)} \cdot ARank(Y)}{|\text{nasledovníci}(Y)|}$$

kde:

p – pravdepodobnosť znovu pípnutia

T – perióda ako často používateľ publikuje

X – používateľ

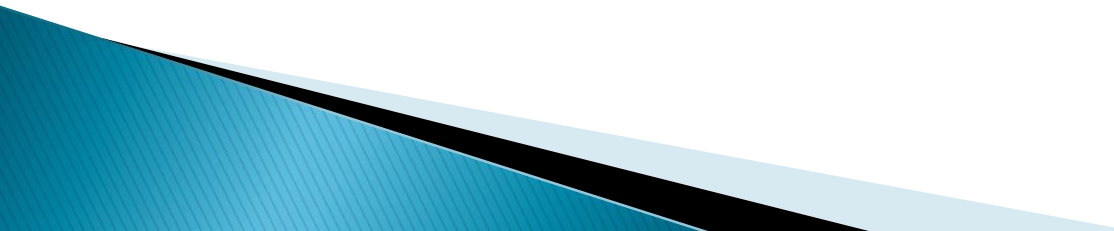


Implementácia

- ▶ RESTFUL webová služba na získanie kľúčových slov pre zadanú URL

Overenie

Presnosť získavania kľúčových slov z mikrobloggerov

- ▶ získanie kľúčových slov pre vybranú množinu kľúčových slov
 - ▶ označenie nesprávnych kľúčových slov
 - ▶ vyhodnotenie
- 

Miera obohatenia *sig*

$$sig(URL) = \frac{|C_T(URL) \setminus C_C(URL)|}{|C_T(URL) \cup C_C(URL)|}$$

- ▶ C_T – množina správnych kľúčových slov získaných z mikroblogu *Twitter*
- ▶ C_C – množina správnych kľúčových slov získaných z obsahu

URL	Twitter Keywords	Twit. Prec.	Content Keywords	Cont. Prec.	Aggr. Prec.	Sig
http://helloblueivycarter.tumblr.com/	cute, Jay-Z, Beyonce, Blue Ivy Carter, the most beautiful girl	100%	Blue Ivy Carter	100%	100%	80%
http://www.grammy.com/live	grammys, pre-telecast ceremony, awards, pre-telecast ceremony part, <i>non piu mesta</i>	80%	grammy awards, Tedeschi Trucks Band. electronic act skrillex, Kanye West, Foo Fighters	100%	90%	44%
http://www.liverpoolfc.tv/news/latest-news/luis-suarez-i-m-sorry	Luis Suarez, LFC Luis Suarez, didn't shake, official apology, <i>ordinary people</i>	80%	Old Trafford, <i>Patroce Eva</i> , game, <i>manager</i> , club	60%	70%	57%
http://www.tnz.com/2012/02/11/whitney-houston-dead/	Whitney Houston, Whitney Houston R.I.P.	100%	Houston, Beverly Hilton Hotel, <i>Bobby Brown</i> , Video, Music Awards	80%	90%	33%
http://edition.cnn.com/2012/02/21/opinion/wisniewski-google-privacy/index.html?hpt=hp_c3	Google, latest tracking news, Privacy	100%	Google, privacy, advertising, privacy protection, Chester Wisniecky	100%	100%	17%

Dosiahnuté výsledky pre 10 URL

- ▶ Priemerná presnosť: 86%
- ▶ Priemerná miera obohatenia: 46%

Ďalšia práca

- ▶ Získanie kľúčových slov pre vybranú reprezentatívnu množinu (cca 100 *URL*) z priemerne často sa vyskytujúcich *URL* v datase (30+ tweets)
- ▶ Vytvorenie webovej aplikácie na vyhodnotenie správnosti kľúčových slov
- ▶ Vyhodnotenie expertami
- ▶ Sumárne vyhodnotenie charakteristík

Ďalšia práca

- ▶ Overenie obohateného Tunk Ranku
 - Problém – chýbajúci aktuálny dataset – kompletný twitter graf
 - Overenie na staršom, menšom datasete

Zhodnotenie

- ▶ Dosiahnuté výsledky pre malú množinu *URL*
 - Priemerná presnosť: 86%
 - Priemerná miera obohatenia: 46%
- ▶ Nutnosť realizácie rozsiahlejšieho experimentu
 - Viac expertov
 - Väčšia množina *URL*
 - Webová aplikácia na vyhodnotenie správnosti kľúčových slov