

Acquiring Metadata about Web Content Based on Microblog Analysis

Tomáš Uherčík
supervised by
Marián Šimko
STU
FIIT

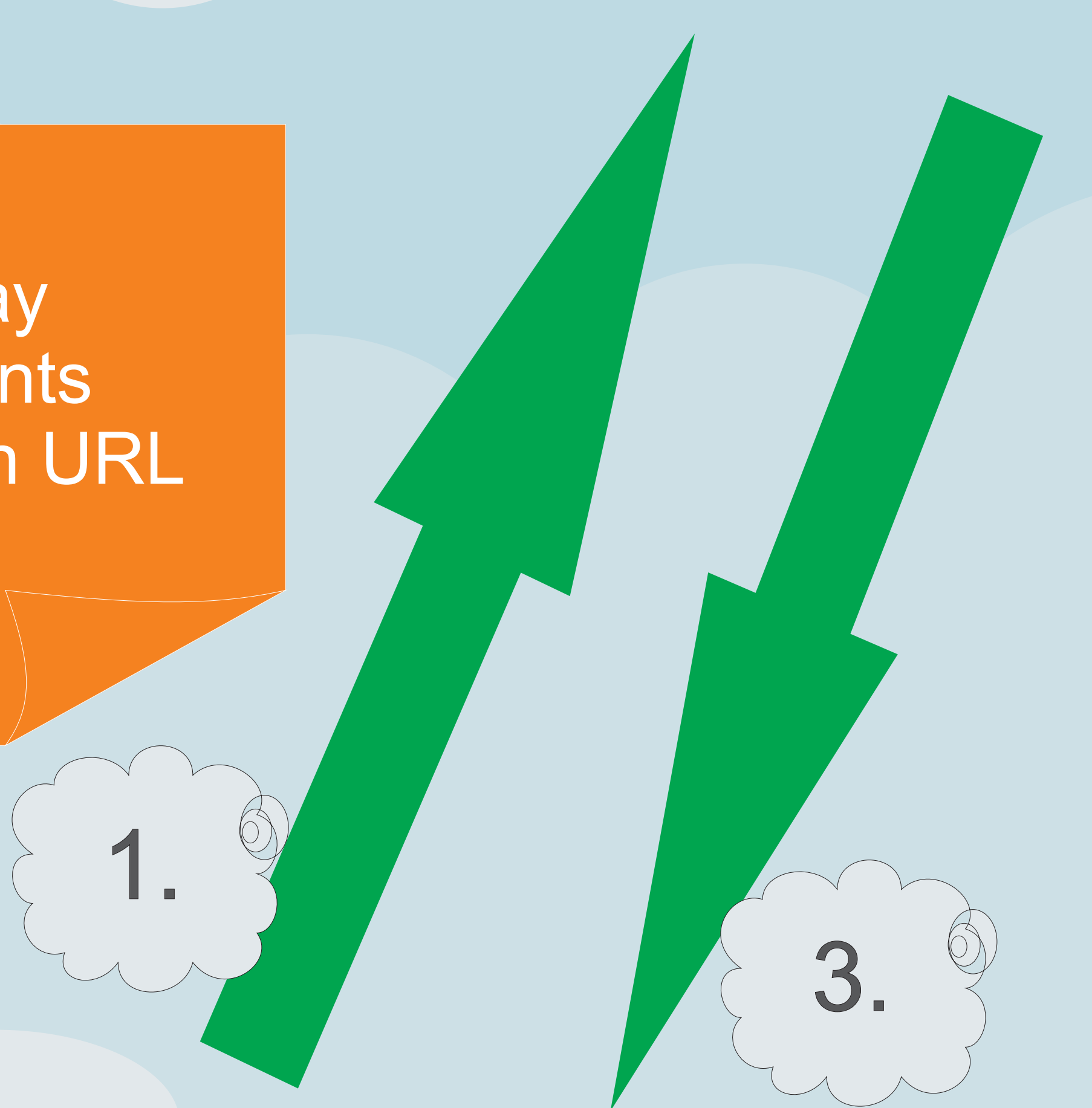
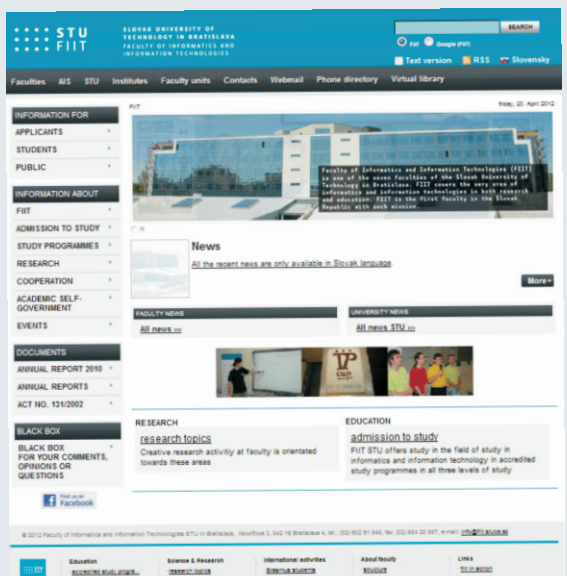
1. Web page entry
2. Processing
3. Keywords

Twitter

- 175 million tweets a day
- over 465 million accounts
- 25 % of tweets contain URL
- user graph based on follower's relationship

URL to analyze

?



Obtained Keywords

- FIIT
- Slovak University of Technology
- Informatics
- Imagine Cup
- Information Technologies

Our Approach

Principles

- search for URL in tweets to get external metadata
- rank authors by method based on TunkRank
- compute the ranking of keywords from their text relevance in tweets and rankings of their authors

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

$T = \text{Med}(T_0, T_1, \dots, T_n)$

T_i - the time gap between published tweets

Usage Example

URL	Twitter Keywords	Twit. Prec.	Content Keywords	Cont. Prec.	Aggr. Prec.	Impr.
http://helloworldivycarter.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>Patrice Eva</i> , game, <i>manager</i> , club	60%	70%	57%
http://www.tmz.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%

Evaluation

Precision

- we evaluated precision of our approach on small set of URLs
- we obtained average precision 86 %

Further Evaluation

- survey experiment in the progress to evaluate more URLs and get feedback from more people

Improvement Rate

- measures to what extent our method enriches keywords extracted from content
- we obtained 46 % for the test set of URLs

$$\text{impr}(\text{URL}) = \frac{|C_T(\text{URL}) \setminus C_c(\text{URL})|}{|C_T(\text{URL}) \cup C_c(\text{URL})|}$$

C_T - set of correct keywords acquired from Twitter

C_c - set of correct keywords acquired from content