Získavanie metadát z webových sídiel

Bc. Milan Lučanský
Vedúci práce: Ing. Marián Šimko

Motivácia

- Neustále sa zvyšujúci počet informácií na webe
  - 152 miliónov blogov[1]
  - 255 miliónov webových stránok[1]
    - nárast za rok 2010 o 21,4 milióna

- Automatické zavedenie sémantiky do webu.

- RDF, OWL, Microformat majú takmer nolovú používanosť v rámci top 1 000 000 najnavštevovanejších stránok[2]

[1] http://royal.pingdom.com/2011/01/12/internet-2010-in-numbers/
[2] http://trends.builtwith.com/docinfo/RDF

Hypotézy

1. **Kaskádové štýly** a HTML značky dokážu pomôcť ATR algoritmom získať lepšie kľúčové slová.

2. Zdieľaná znalosť na webe dokáže rozšíriť množinu kľúčových slov o také, ktoré sa nenachádzajú priamo na skúmanom webovom dokumente.

3. Použitím synonymického slovníka (WordNet) dokážeme rozšíriť množinu kľúčových slov o také, kotré sa nenachádzajú priamo na skúmanom webovom dokumente.

Hypotéza 1

- HTML značky `title`, `h1-5`, a dokázali zvýšiť kvalitu kľúčových slov[3], budeme uvažovať ďalšie značky, ktoré formátujú dokument: `b, em, i, u, strong` a ďalšie.

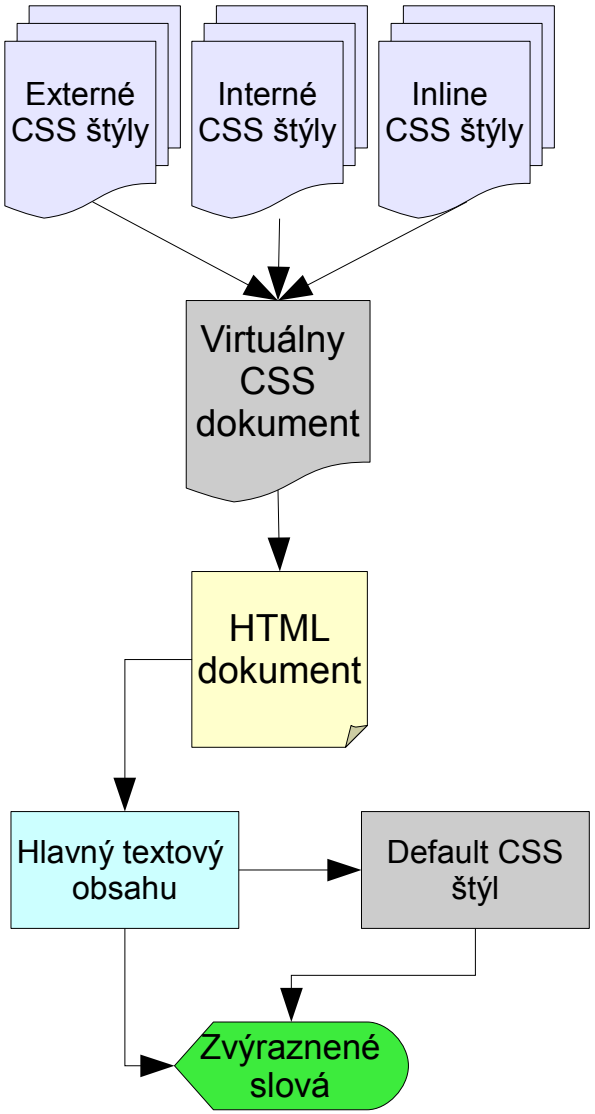
- Anchor text externých liniek.

- CSS štýl, použitie až desiatich atribútov na formátovanie textu, napr: `font-family, font-size, font-weight, color` a ďalšie.

dôraz na
preskúmanie vlastností CSS atribútov

[3] Lučanský M., Šimko M.: Získavanie obsahu webových sídiel. In: Bakalárska práca. 2010. 75 s. FIIT-5212-36261.

Hypotéza 1



Výpočet CssIndexu pre:

- element div, font-size = 20 px, obsahuje 15 slov
- default font-size = 16 px

$$CssIndex(div) = \left(1 + \frac{1}{\log(wc_{div})}\right) \cdot vf$$

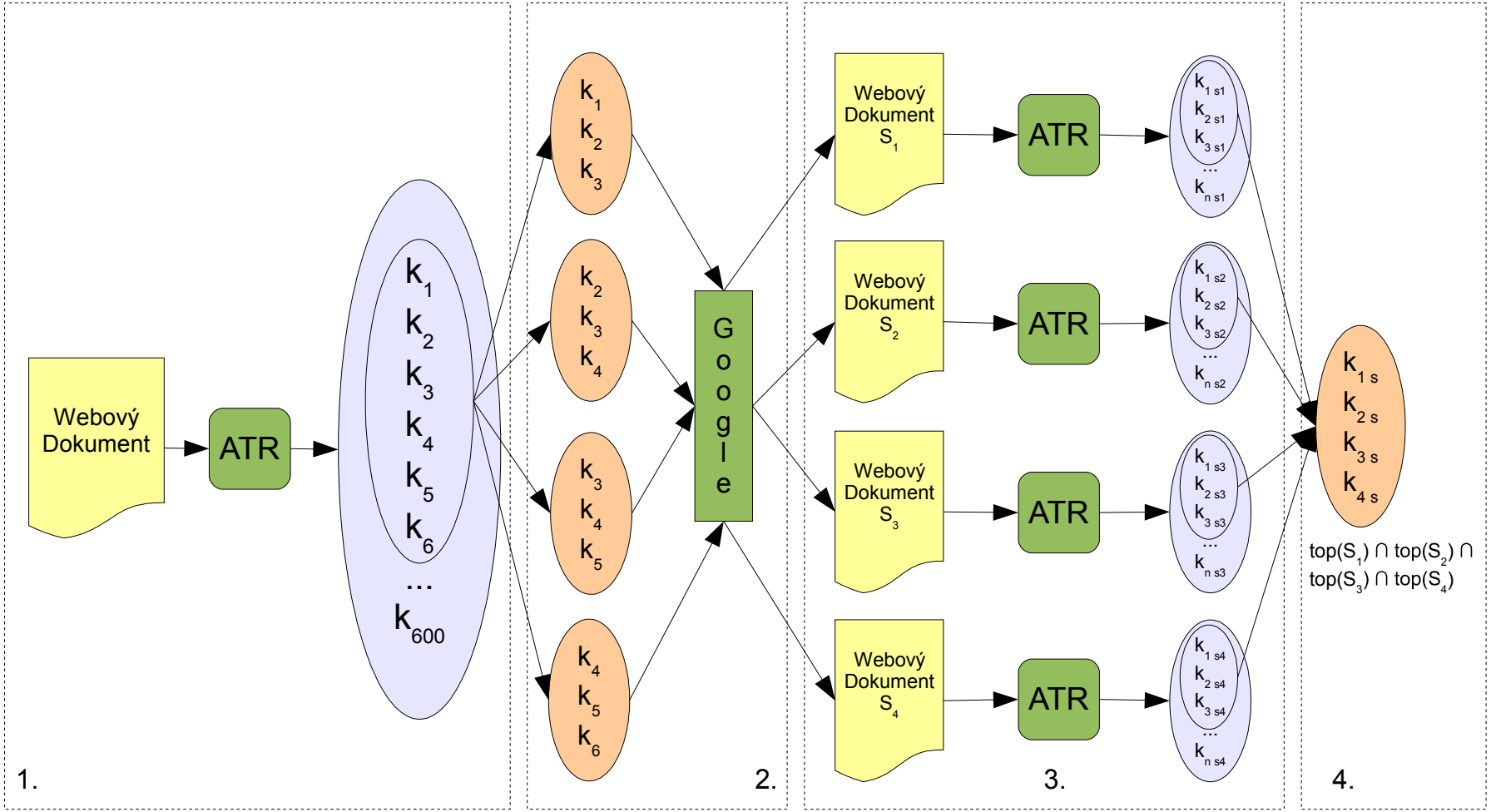
$$CssIndex(div) = \left(1 + \frac{1}{\log(15)}\right) \cdot \frac{20}{16} = 2{,}31$$

Visibility faktor:

$$vf = \frac{font - size}{default\ font - size}$$

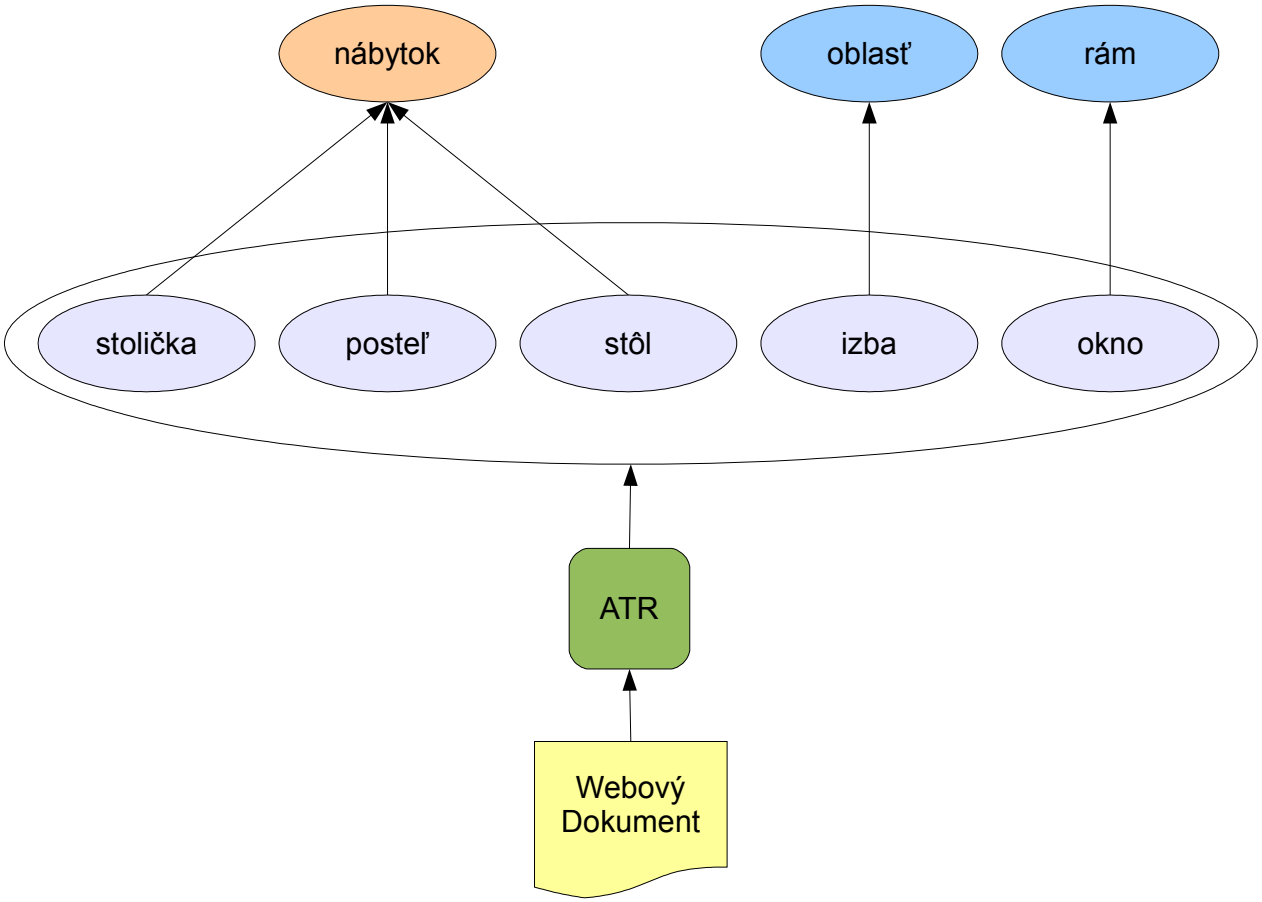
Hypotéza 2

Inšpirované PANKOW algoritmom[4]



[4] Cimiano, P.: Ontology Learning and Population from Text Algorithms, Evaluation and Applications. Springer. University of Karlsruhe, Germany. 2006. s. 347.

Hypotéza 3



Hľadanie spoločných hyperným pomocou synonymického slovníka WordNet.

Prototyp

- Implementovali sme program, ktorý
  - Stiahne webovú stránku
  - Získa kaskádové štýly a aplikuje ich na stránku
  - Získa hlavný textový obsah z webovej stránky
  - Získa obsah elementov formátovaných vymenovanými CSS atribútmi
  - Získa ováhované kľúčové slová

Overenie

Rozhodnutie o relevantnosti extrahovaných slov, napr. pomocou hlasovanie cez web formulár.

Overenie na rôznych typoch stránok

- Blogy
- Webové sídla
- News portály

Porovnanie

- Jednotlivých typov indexov voči sebe (TagIndex VS. CssIndex VS. LinkIndex).
- Jednotlivých metód voči sebe.

Záver

- Analyzovali sme problematiku získania kľúčových slov vo webových dokumentoch.
- Navrhli sme 3 prístupy extrakcie kľúčových slov z webu.
  - Dva z troch prístupov majú ambíciu nájsť aj kľúčové slová, ktoré sa na stránke nemusia nachádzať.
- Implementovali sme
  - Zlúčenie štýlov do virtuálneho štýlu
  - Získanie hlavného textového elementu
  - Nájdenie default CSS štýlov pre halvný text element