Bc. Milan Lučanský
Ing. Marián Šimko, PhD.

Acquiring Web Site Metadata by Heterogeneous Information Sources Processing

What is the problem?

- too many web docuents emerge
- in december 2011 there were about 556 000 000 websites
- web documents structure differ from site to site
- contemporary approaches do not achieve satisfactory results

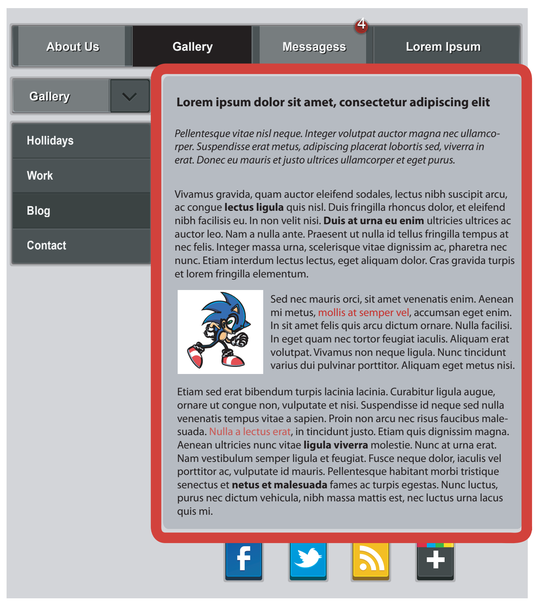
What to do about it?

- we need to utilize **features** of todays web pages
- web documents provide **implicit semantics** in style
- **keyword** paradigm seems apropriate to describe web documents

Why keyword paradigm?

- widespread in many disciplines (summarization and cetegorization)
- „de facto" standard in searching (google search)
- used in social media (e.g. deli.cio.us)
- used in advertising (e.g. AdWords)

Our approach to keywords extraction



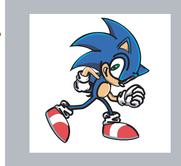
1. Find all CSS styles

2. Extract content

Lorem ipsum dolor sit amet, consectetur adipiscing elit

Pellentesque vitae nisl neque. Integer volutpat auctor magna nec ullamcorper. Suspendisse erat metus, adipiscing placerat lobortis sed, viverra in erat. Donec eu mauris et justo ultrices ullamcorper et eget purus.

Vivamus gravida, quam auctor eleifend sodales, lectus nibh suscipit arcu, ac congue **lectus ligula** quis nisl. Duis fringilla rhoncus dolor, et eleifend nibh facilisis eu. In non velit nisi. **Duis at urna eu enim** ultricies ultrices ac auctor leo. Nam a nulla ante. Praesent ut nulla id tellus fringilla tempus at nec felis. Integer massa urna, scelerisque vitae dignissim ac, pharetra nec nunc. Etiam interdum lectus lectus, eget aliquam dolor. Cras gravida turpis et lorem fringilla elementum.



Sed nec mauris orci, sit amet venenatis enim. Aenean mi metus, mollis at semper vel, accumsan eget enim. In sit amet felis quis arcu dictum ornare. Nulla facilisi. In eget quam nec tortor feugiat iaculis. Aliquam erat volutpat. Vivamus non neque ligula. Nunc tincidunt varius dui pulvinar porttitor. Aliquam eget metus nisi.

Etiam sed erat bibendum turpis lacinia lacinia. Curabitur ligula augue, ornare ut congue non, vulputate et nisi. Suspendisse id neque sed nulla venenatis tempus vitae a sapien. Proin non arcu nec risus faucibus malesuada. Nulla a lectus erat, in tincidunt justo. Etiam quis dignissim magna. Aenean ultricies nunc vitae **ligula viverra** molestie. Nunc at urna erat. Nam vestibulum semper ligula et feugiat. Fusce neque dolor, iaculis vel porttitor ac, vulputate id mauris. Pellentesque habitant morbi tristique senectus et **netus et malesuada** fames ac turpis egestas. Nunc luctus, purus nec dictum vehicula, nibh massa mattis est, nec luctus urna lacus quis mi.

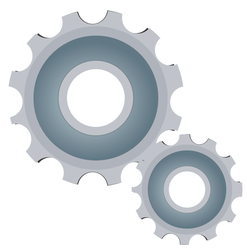
3. Find empasized phrases

1. empasized phrase
   + CSS features
2. empasized phrase
   + CSS features
   ...

4. Extract plain text

Lorem ipsum dolor sit amet, consectetur adipiscing elit Pellentesque vitae nisl neque. Integer volutpat auctor magna nec ullamcorper. Suspendisse erat metus, adipiscing placerat lobortis sed, viverra in erat. Donec eu mauris et justo ultrices ullamcorper et eget purus. Vivamus gravida, quam auctor eleifend sodales, lectus nibh suscipit arcu, ac congue lectus ligula quis nisl. Duis fringilla rhoncus dolor, et eleifend nibh facilisis eu. In non velit nisi. Duis at urna eu enim ultricies ultrices ac auctor leo. Nam a nulla ante. Praesent ut nulla id tellus fringilla tempus at nec felis. Integer massa urna, scelerisque vitae dignissim ac, pharetra nec nunc. Etiam interdum lectus lectus, eget aliquam dolor. Cras gravida turpis et lorem fringilla elementum...

5. ATR + some magic



6. Get relevant keywords

- ✓ Keyword 1
- ✓ Keyword 2
- ✓ Keyword 3
- ✓ Keyword 4
- ✓ Keyword 5

...

Magic explanation

- extract emphasized words and phrases from web document
- compute the *visibility factor*
- extract phrases and words from plain text
- improve weight of phrases, which are emphasized by CSS styles

Visibility factor

- add points for every style feature, which is different from main textual content.

Exprimental evaluation & Further work

So far we have evaluated
- 60 web documents
- from 11 different web sites
- together 2048 phrases

- determine important style features
- train SVM to automatically detect keywords

Base algorithms	Average relevance %	After utilizig CSS features	AVG relevance %	Delta %
CValue	72,98	*Cvalue*	74,71	**1,73**
TermEx	71,44	*TermEx*	73,64	**2,2**
Weirdness	63,61	*Weirdness*	70,26	**6,65**
GlossEx	55,37	*GlossEx*	63,1	**7,73**
TF-IDF	48,27	*TF-IDF*	69,98	**21,71**

PeWe@FIIT
personalized web group

STU FIIT