Automatic Web Content Annotation

Jakub Ševcech
Supervisor: Professor Mária Bieliková

What?

Method for automatic creation of annotations into web pages in Slovak.

Annotations attached to important words in the text of web page.

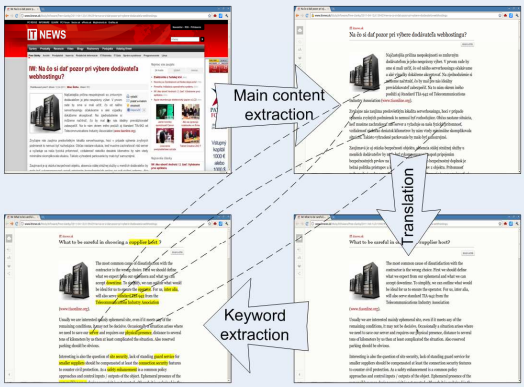
Annotations providing additional information.

Publicly available services used to search for information to fill the annotation.

How?

1. Search for candidate words to assing the annotation
2. Mapping candidate words
3. Search for information to fill the annotations
4. Annotation visualization and adaptation

Search for candidate words to assign the annotation



- Removal of redundant parts of pages as various menus and advertisements.
- Translation of text into English.
- Extraction of keywords - various ATR algorithms or methods from the field of Natural Language Processing.
- Search for mapping between extracted English keywords and their equivalents in Slovak text

Mapping candidate words

Bilingual dictionary and Levenshtein distance used to find mappings between equivalent words in Slovak text and its translation.

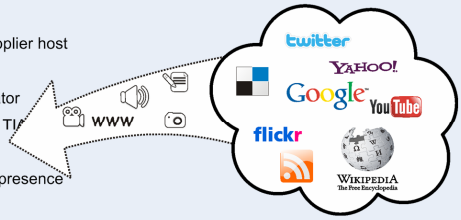
Different shapes of words processed.

Levenshtein distance - the minimal number of edit operations needed to transform one string to another.

1. kitten → sitten (substitution of 's' for 'k')
2. sitten → sittin (substitution of 'i' for 'e')
3. sittin → sitting (insertion of 'g' at the end).

Function	*Correct*	*Incorrect*	*More*	*Mapped words / all words*
Basic	92.75%	5.82%	1.60%	45.38%
Position of words taken into account	55.14%	32.92%	11.93%	84.85%
Stemmed dictionary	92.45%	5.58%	1.95%	63.59%
Both improvements	64.07%	24.95%	10.96%	96.08%

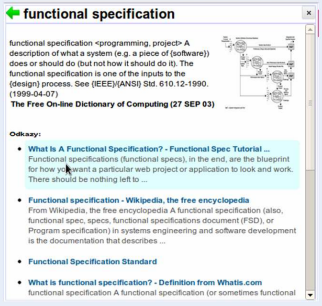
Dodávateľa webhostingu - Supplier host
Výpadky - Downtime
Prevádzkovateľ - Operator
Štandard TIA-942 - Standard TIA
Server - Server
Fyzická prítomnosť - Phisical presence
...



Search for information to fill the annotation

Various publicly available services used for information search.

- Annotation method is independent on domain of processed text.
- No pretreated database of information needed.
- Various types of information in content of annotations.
- Quality of annotations heavely depends on the quality of used services.



Annotation Visualization and Adaptation

Visualization in form of word definition and links to related resources.

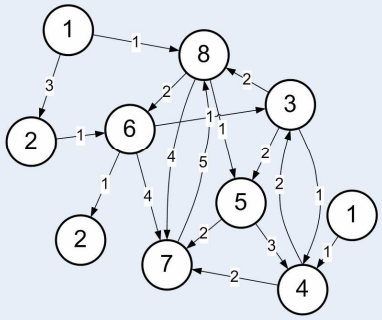
More relevant resources should be on higher positions.

Reordering based on clickthrough data.

- Clicks - statements that one link is better than another.
- Weights of links - PageRank algorithm on statements graph.

Evaluation in educational system ALEF.

Order of links created using implicit feedback compared to order gathered through explicit feedback.



Institute of Informatics and Software Engineering, Fakulty of Informatics and Information Technologies,
Slovak University of Technology, Ilkovičova 3, 842 16 Bratislava 4
sevo_jakub@yahoo.fr

PeWe@FIIT
personalized web group