Extraction of Segments from Web 2.0 Pages

URL → Genre Detection → Page Segmentation → Segment Classification → Output Format

Dipl. Inform. Renato Dominguez Garcia
Renato.Dominguez.Garcia@KOM.tu-darmstadt.de
Tel.+49 6151 165842

KOM - Multimedia Communications Lab
Prof. Dr.-Ing. Ralf Steinmetz (Director)
Dept. of Electrical Engineering and Information Technology
Dept. of Computer Science (adjunct Professor)
TUD – Technische Universität Darmstadt
Merckstr. 25, D-64283 Darmstadt, Germany
Tel.+49 6151 166150, Fax. +49 6151 166152
www.KOM.tu-darmstadt.de

© author(s) of these slides 2008 including research results of the research network KOM and TU Darmstadt otherwise as specified at the respective slide
Motivation

- The influence of the internet for the US president election has grown since 2004 (+11%) [1]
- Primary information source about the US president election for young Americans (18 – 29 years old) is the internet (42%) [1]
- Each third American read blogs [2]
- PR professionals recognize importance of blogs [3]

→ Information extraction from blogs can help to understand the public opinion
→ Automatically detection of blogs, wikis and forums may be useful
→ Information extraction from small segments is easier than from large web pages
→ Genre Detection and Information extraction can be used in other fields: Community Mining, Improvement of search results

Our Approach: Genre Detection

Genre Detection
- 6 Genres
  - Blogs (Start pages, post pages)
  - Wikis
  - Forums (Start pages, thread pages)
  - Others
- Based on the structure of web pages (Patterns)
- Machine Learning Techniques
  - Support Vector Machines
- 336 Features
- Corpus: ~ 33000 Web pages
- Evaluation
  - 1345 Instances (1000 Blogs/Wikis/Forums)
  - 87.5 % Correctly classified instances

URL → Genre Detection → Page Segmentation → Segment Classification → Output Format

- Based on the structure of web pages (Patterns)
- Machine Learning Techniques
  - Support Vector Machines
- 336 Features
- Corpus: ~ 33000 Web pages
- Evaluation
  - 1345 Instances (1000 Blogs/Wikis/Forums)
  - 87.5 % Correctly classified instances

Output Format:
- Number of detected patterns
- Number of outer patterns
- Ratio of patterned vs unpatterned Code
- Length of patterns
- Offset before first pattern starts
- Depth of patterns
...
Our Features: Examples

Blog’s start page

- Many long patterns
- Blog’s start page
- Wiki-page
- Forum’s start page

- Many short patterns

All posts share a similar structure

<table>
<thead>
<tr>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fat Thunder</td>
</tr>
</tbody>
</table>

All threads share a similar structure

Set of links

Many long patterns

Many short patterns
Our Approach: Segmentation

- Page segmentation (Four steps)
  - Pre-processing (cleaning HTML)
  - Segmentation based on the hierarchical structure of web Pages
  - Visual-based segmentation
  - Filtering based on heuristics

- Segment classification
  - Machine Learning Techniques
    - Random Forest
    - 139 Features
    - Corpus: ~ 500 instances

- Evaluation
  - Genres (blog posts, comments, others)
  - 97.2% correct classified instances