Harnessing Web Archives to Tackle Selected Societal Challenges

The Oral Exam of
Johannes Kiesel
To Obtain the Academic Degree of
Dr. rer. nat.

Web Technology & Information Systems Group
Bauhaus-Universität Weimar

www.uni-weimar.de www.webis.de
Societal challenges

Issues that concern most if not all members of a society, either now or in a likely future.

Well-known challenges:

- Critical assessment of information
- Protection of the environment
- Preservation of culture
- Ensuring public health
- Security and privacy

* Taken from European Commission (Horizon 2020), World Economic Forum, Gesellschaft für Informatik
Harnessing Web Archives to Tackle Selected Societal Challenges

Societal challenges

*Issues that concern most if not all members of a society, either now or in a likely future.*

Well-known challenges:*

- Critical assessment of information
- Protection of the environment
- Preservation of culture
- Ensuring public health
- Security and privacy

Web archives

- Allow for large-scale analyses
- Allow to trace changes
- Allow to replicate analyses

* Taken from European Commission (Horizon 2020), World Economic Forum, Gesellschaft für Informatik

Source: DOMO, Reddit, GDELT, Wikipedia

© Kiesel 2022
Harnessing Web Archives to Tackle Selected Societal Challenges

Main contributions

1. Preservation of digital culture
   - 10K pages high-fidelity archive (FAIRest dataset award)
   - Reproduction assessment task
   - 9K pages segmentation dataset
   - Segmentation evaluation measures

2. Critical assessment of information
   - Revert-based vandalism detection
   - 30K edits Wiki vandalism dataset
   - 1M hyperpartisan news dataset
   - Style-based polarity detection
   - Hyperpartisan news challenge (SemEval, 42 teams)

3. Online security and privacy
   - 3B web sentences dataset
   - Position-dependent language model
   - Security estimate: mnemonic passwords
   - Personal archiving tool

Tailored web archiving technology (Webis Web Archiver)
Harnessing Web Archives to Tackle Selected Societal Challenges

Main contributions

1. Preservation of digital culture
   - 10K pages high-fidelity archive (FAIRest dataset award)
   - Reproduction assessment task
   - 9K pages segmentation dataset
   - Segmentation evaluation measures

2. Critical assessment of information
   - Revert-based vandalism detection
   - 30K edits Wiki vandalism dataset
   - 1M hyperpartisan news dataset
   - Style-based polarity detection
   - Hyperpartisan news challenge (SemEval, 42 teams)

3. Online security and privacy
   - 3B web sentences dataset
   - Position-dependent language model
   - Security estimate: mnemonic passwords
   - Personal archiving tool

Tailored web archiving technology (Webis Web Archiver)

→ New tasks
Harnessing Web Archives to Tackle Selected Societal Challenges

Main contributions

1. Preservation of digital culture
   - 10K pages high-fidelity archive (FAIRest dataset award)
   - Reproduction assessment task
   - 9K pages segmentation dataset
   - Segmentation evaluation measures

2. Critical assessment of information
   - Revert-based vandalism detection
   - 30K edits Wiki vandalism dataset
   - 1M hyperpartisan news dataset
   - Style-based polarity detection
   - Hyperpartisan news challenge (SemEval, 42 teams)

3. Online security and privacy
   - 3B web sentences dataset
   - Position-dependent language model
   - Security estimate: mnemonic passwords
   - Personal archiving tool

Tailored web archiving technology (Webis Web Archiver)

→ New tasks → New or improved algorithms
Harnessing Web Archives to Tackle Selected Societal Challenges

Main contributions

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>10K pages high-fidelity archive (FAIRest dataset award)</td>
<td>Revert-based vandalism detection</td>
<td>3B web sentences dataset</td>
</tr>
<tr>
<td>Reproduction assessment task</td>
<td>30K edits Wiki vandalism dataset</td>
<td>Position-dependent language model</td>
</tr>
<tr>
<td>9K pages segmentation dataset</td>
<td>1M hyperpartisan news dataset</td>
<td>Security estimate: mnemonic passwords</td>
</tr>
<tr>
<td>Segmentation evaluation measures</td>
<td>Style-based polarity detection</td>
<td>Personal archiving tool</td>
</tr>
<tr>
<td></td>
<td>Hyperpartisan news challenge (SemEval, 42 teams)</td>
<td></td>
</tr>
</tbody>
</table>

Tailored web archiving technology (Webis Web Archiver)

→ New tasks  → New or improved algorithms  → More adequate evaluation measures
Harnessing Web Archives to Tackle Selected Societal Challenges

Main contributions

1. Preservation of digital culture
   - 10K pages high-fidelity archive (FAIRest dataset award)
   - Reproduction assessment task
   - 9K pages segmentation dataset
   - Segmentation evaluation measures

2. Critical assessment of information
   - Revert-based vandalism detection
   - 30K edits Wiki vandalism dataset
   - 1M hyperpartisan news dataset
   - Style-based polarity detection
   - Hyperpartisan news challenge (SemEval, 42 teams)

3. Online security and privacy
   - 3B web sentences dataset
   - Position-dependent language model
   - Security estimate: mnemonic passwords
   - Personal archiving tool

Tailored web archiving technology (Webis Web Archiver)

→ New tasks  → New or improved algorithms  → More adequate evaluation measures  → Larger and more accurate datasets
Challenge 1
Preservation of Digital Culture

Web Page Segmentation
(highlighting reproducibility)
Web Page Segmentation

Flashback: Supercut of Elton John singing 'Your Song' through the years

© Kiesel 2022
Web Page Segmentation

Visually distinct segments

Self-contained segments
Web Page Segmentation

Existing definitions (9): biased towards downstream tasks

- Segments are visual blocks (4), edge-delineated (2), visually distinct (1), self-contained (1), have a heading (1)

  → Problem: inconsistent evaluation methodology

  → No reliable benchmark of algorithms

Existing datasets (20): not re-usable

- The 12 with human annotations are small (max 1000 pages)
- Only 3 of these allow for algorithms based on computer vision
- None allow to reproduce page for browser-based algorithms
Web Page Segmentation

Existing definitions (9): biased towards downstream tasks
- Segments are visual blocks (4), edge-delineated (2), visually distinct (1), self-contained (1), have a heading (1)
→ Problem: inconsistent evaluation methodology
→ No reliable benchmark of algorithms

Existing datasets (20): not re-usable
- The 12 with human annotations are small (max 1000 pages)
- Only 3 of these allow for algorithms based on computer vision
- None allow to reproduce page for browser-based algorithms

Solution
- Segment concept based on human viewer (Gestalt principles)
- Dataset of 8490 archived web pages (5 segmentations each; reproducible in browser)
- Segmentation fusion method
- Evaluation measure, tweakable towards downstream tasks

Gestalt principles (selection)
- Proximity
- Similarity
- Closure
A web page segment is a part of a web page containing those elements that belong together as per agreement among a majority of viewers.

Elements $E = \{e_1, \ldots, e_n\}$

Segmentation $S = \{s_1, \ldots, s_m\}$ with segments $s_i \subseteq E$
A web page segment is a part of a web page containing those elements that belong together as per agreement among a majority of viewers.

Elements $E = \{e_1, \ldots, e_n\}$
Segmentation $S = \{s_1, \ldots, s_m\}$ with segments $s_i \subseteq E$

Large-scale human annotation (8490 pages \times 5)

→ Annotation of 600,000 segments in 4 months of full-time work
A web page segment is a part of a web page containing those elements that belong together as per agreement among a majority of viewers.

Elements $E = \{e_1, \ldots, e_n\}$

Segmentation $S = \{s_1, \ldots, s_m\}$ with segments $s_i \subseteq E$

Ground-truth fusion: hierarchical clustering (UPGMA)

Large-scale human annotation (8490 pages × 5)

→ Annotation of 600,000 segments in 4 months of full-time work
A web page segment is a part of a web page containing those elements that belong together as per agreement among a majority of viewers.

Elements $E = \{e_1, \ldots, e_n\}$

Segmentation $S = \{s_1, \ldots, s_m\}$ with segments $s_i \subseteq E$

Ground-truth fusion: hierarchical clustering (UPGMA)

Evaluation: $F_{B^3} \in [0, 1]$ (from clustering evaluation)

$\rightarrow$ Decomposition into $P_{B^3}, R_{B^3}$

$\approx$ errors of oversegmentation, undersegmentation

Large-scale human annotation (8490 pages × 5)

→ Annotation of 600,000 segments in 4 months of full-time work
A web page segment is a part of a web page containing those elements that belong together as per agreement among a majority of viewers.

Elements $E = \{e_1, \ldots, e_n\}$

Segmentation $S = \{s_1, \ldots, s_m\}$ with segments $s_i \subseteq E$

Ground-truth fusion: hierarchical clustering (UPGMA)

Evaluation: $F_{B^3} \in [0, 1]$ (from clustering evaluation)

$\rightarrow$ Decomposition into $P_{B^3}, R_{B^3}$

$\approx$ errors of oversegmentation, undersegmentation
A web page segment is a part of a web page containing those elements that belong together as per agreement among a majority of viewers.

Elements $E = \{e_1, \ldots, e_n\}$

Segmentation $S = \{s_1, \ldots, s_m\}$ with segments $s_i \subseteq E$

Ground-truth fusion: hierarchical clustering (UPGMA)

Evaluation: $F_{B_3} \in [0, 1]$ (from clustering evaluation)

$\rightarrow$ Decomposition into $P_{B_3}, R_{B_3}$

$\approx$ errors of oversegmentation, undersegmentation

Elements of downstream tasks

Characters

DOM nodes

Pixels

Edge pixels

High agreement for all tasks

<table>
<thead>
<tr>
<th>Agreement measure</th>
<th>Characters</th>
<th>Nodes</th>
<th>Pixels</th>
<th>Edge pixels</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_{B_3}$</td>
<td>0.78</td>
<td>0.74</td>
<td>0.65</td>
<td>0.73</td>
</tr>
<tr>
<td>$\max(P_{B_3}, R_{B_3})$</td>
<td>0.97</td>
<td>0.95</td>
<td>0.94</td>
<td>0.96</td>
</tr>
</tbody>
</table>
A web page segment is a part of a web page containing those elements that belong together as per agreement among a majority of viewers.

Elements $E = \{e_1, \ldots, e_n\}$
Segmentation $S = \{s_1, \ldots, s_m\}$ with segments $s_i \subseteq E$

Ground-truth fusion: hierarchical clustering (UPGMA)

Evaluation: $F_{B^3} \in [0, 1]$ (from clustering evaluation)

→ Decomposition into $P_{B^3}, R_{B^3}$
≈ errors of oversegmentation, undersegmentation

Elements of downstream tasks

<table>
<thead>
<tr>
<th>Characters</th>
<th>DOM nodes</th>
<th>Pixels</th>
<th>Edge pixels</th>
</tr>
</thead>
</table>

High agreement for all tasks

<table>
<thead>
<tr>
<th>Agreement measure</th>
<th>Characters</th>
<th>Nodes</th>
<th>Pixels</th>
<th>Edge pixels</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_{B^3}$</td>
<td>0.78</td>
<td>0.74</td>
<td>0.65</td>
<td>0.73</td>
</tr>
<tr>
<td>$\max(P_{B^3}, R_{B^3})$</td>
<td>0.97</td>
<td>0.95</td>
<td>0.94</td>
<td>0.96</td>
</tr>
</tbody>
</table>

Insights into segmentation technology ($F_{B^3}$)

<table>
<thead>
<tr>
<th>Elements/task</th>
<th>1Seg</th>
<th>VIPS</th>
<th>HEPS</th>
<th>Cor.</th>
<th>MMD.</th>
<th>Meier</th>
<th>MV@2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Characters</td>
<td>0.52</td>
<td><strong>0.67</strong></td>
<td>0.50</td>
<td>0.61</td>
<td>0.61</td>
<td>0.50</td>
<td>0.62</td>
</tr>
<tr>
<td>Pixels</td>
<td>0.24</td>
<td>0.38</td>
<td>0.33</td>
<td>0.36</td>
<td><strong>0.42</strong></td>
<td>0.32</td>
<td>0.39</td>
</tr>
</tbody>
</table>
Challenge 2
Critical Assessment of Information

Spatio-Temporal Analysis of Vandalism in Wikipedia

(highlighting temporal dynamics)
Wikipedia Vandalism
Wikipedia Vandalism

Vandalism is a problem for Wikipedia

- 470 million edits to the English Wikipedia (14 years)
- 40 million (9.5%) are vandalism
  → Rate of today: a vandalism case every 5 seconds

How to fight vandalism?

- Explain why people vandalism
- Analyze when people vandalize
- Analyze where these people are
Wikipedia Vandalism

Vandalism is a problem for Wikipedia

- 470 million edits to the English Wikipedia (14 years)
- 40 million (9.5%) are vandalism
  → Rate of today: a vandalism case every 5 seconds

How to fight vandalism?

- Explain why people vandalize
- Analyze when people vandalize
- Analyze where these people are

Language-independent detection approach

- Take all 1.2 billion edits to the 7 most-edited Wikipedias (english, german, french, spanish, russian, italian, japanese)
- Historical geolocation of anonymous editors (77% of edits by cross-checking RIR, IPligence, and IP2Location)
- Vandalism detector based on revert patterns (community behavior)
  → Spatio-temporal analysis per local time of anonymous editors
Not all reverts indicate vandalism

- Prior work: use only reverts whose comment indicates vandalism
  → Underestimates vandalism; language-dependent

- Our approach: identify revert patterns indicating non-vandalism

Revert to blank page
Empty revert
Self-revert
Revert correction (enlargement)
Revert reverting more than one editor
Reverted revert
Interleaved reverts (edit war)
Not all reverts indicate vandalism

- Prior work: use only reverts whose comment indicates vandalism
  → Underestimates vandalism; language-dependent

- Our approach: identify revert patterns indicating non-vandalism

- Only 46% of reverted edits are vandalism
- Human evaluation: precision 82.8%, recall 84.7%
  (4 times the recall of prior work)
Not all reverts indicate vandalism

- Prior work: use only reverts whose comment indicates vandalism
  → Underestimates vandalism; language-dependent

- Our approach: identify revert patterns indicating non-vandalism

  - Revert to blank page
  - Empty revert
  - Self-revert
  - Reverted revert
  - Interleaved reverts (edit war)
  - Revert correction (enlargement)
  - Revert reverting more than one editor

- Only 46% of reverted edits are vandalism

- Human evaluation: precision 82.8%, recall 84.7%
  (4 times the recall of prior work)
Spatio-temporal vandalism analysis

- English Wikipedia from United States
- French Wikipedia from France
- Japanese Wikipedia from Japan

- Monday - Thursday
- Friday
- Saturday
- Sunday

© Kiesel 2022
Challenge 3
Online Security and Privacy

Security Estimate for Mnemonic Passwords

(highlighting volume)
The mnemonic password advice
(as per German BSI, Google, etc.)

1. Create a sentence
2. Memorize it
3. Concatenate the first characters of each word
4. Use the string as password

When I walked to the grocery store,
there were camels flying overhead!

Password: wiwttgstwcfo
The mnemonic password advice
(as per German BSI, Google, etc.)

1. Create a sentence
2. Memorize it
3. Concatenate the first characters of each word
4. Use the string as password

Passwords that require a botnet \((H_1 \approx 65 \text{ Bit})\):
- 14 random lowercase letters (out of 26)
- 10 random ASCII characters (out of 96)
- 5 random words (out of 7776)

And for mnemonic passwords?

When I walked to the grocery store, there were camels flying overhead!

Password: wiwtgswcfo

![Graph showing probability distribution of word initials](image)
The mnemonic password advice
(as per German BSI, Google, etc.)

1. Create a sentence
2. Memorize it
3. Concatenate the first characters of each word
4. Use the string as password

When I walked to the grocery store, there were camels flying overhead!

Password: wiwtgstwcfo

Passwords that require a botnet ($H_1 \approx 65$ Bit):

- 14 random lowercase letters (out of 26)
- 10 random ASCII characters (out of 96)
- 5 random words (out of 7776)

And for mnemonic passwords?

Depends on password distribution (Kerckhoffs’ principle) → model distribution from a billion passwords
The mnemonic password advice
(as per German BSI, Google, etc.)

1. Create a sentence
2. Memorize it
3. Concatenate the first characters of each word
4. Use the string as password

Passwords that require a botnet \((H_1 \approx 65\text{ Bit})\):
- 14 random lowercase letters (out of 26)
- 10 random ASCII characters (out of 96)
- 5 random words (out of 7776)

And for mnemonic passwords?

- Depends on password distribution (Kerckhoffs' principle) \(\rightarrow\) model distribution from a billion passwords

Approach: substitute mnemonics by web sentences
- 3 billion web sentences corpus from a standard web archive
- Statistically align the sentence corpus to mnemonics
- Estimate password distribution using position-dependent language models
  \(\rightarrow\) Security estimates against offline \((H_1)\) and online attacks \((H_0, \lambda_n)\)
Sentence acquisition for password distribution estimate

<table>
<thead>
<tr>
<th>5,000</th>
<th>Mnemonics Study by Yang et al., 2016</th>
</tr>
</thead>
<tbody>
<tr>
<td>80,000</td>
<td>Sentences The Bible</td>
</tr>
<tr>
<td>5,000,000</td>
<td>Sentences Encyclopedia Britannica</td>
</tr>
<tr>
<td>70,000,000</td>
<td>Passwords Largest password corpus</td>
</tr>
<tr>
<td>730,000,000</td>
<td>Web pages ClueWeb12, 27.3 TB</td>
</tr>
<tr>
<td>3,400,000,000</td>
<td>Sentences Extracted and filtered</td>
</tr>
<tr>
<td>500,000,000</td>
<td>Sentences And aligned to mnemonics</td>
</tr>
</tbody>
</table>

Alignment in sentence complexity (≈ readability)

- Model from the mnemonics study
- Distribution in the Web

Syllables

Sentences of length 12
Security estimates (per character)

<table>
<thead>
<tr>
<th>Language model</th>
<th>Lowercase letters</th>
<th>ASCII</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$H_1$</td>
<td>Ppl.</td>
</tr>
<tr>
<td>Uniform</td>
<td>4.70</td>
<td>26.0</td>
</tr>
<tr>
<td>Order 0</td>
<td>4.15</td>
<td>17.8</td>
</tr>
<tr>
<td>Order 8</td>
<td>3.71</td>
<td>13.1</td>
</tr>
<tr>
<td>Order 8, position-dependent</td>
<td>3.65</td>
<td>12.6</td>
</tr>
</tbody>
</table>
Security estimates (per character)

<table>
<thead>
<tr>
<th>Language model</th>
<th>Lowercase letters</th>
<th>ASCII</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$H_1$</td>
<td>Ppl.</td>
</tr>
<tr>
<td>Uniform</td>
<td>4.70</td>
<td>26.0</td>
</tr>
<tr>
<td>Order 0</td>
<td>4.15</td>
<td>17.8</td>
</tr>
<tr>
<td>Order 8</td>
<td>3.71</td>
<td>13.1</td>
</tr>
<tr>
<td>Order 8, position-dependent</td>
<td>3.65</td>
<td>12.6</td>
</tr>
</tbody>
</table>

Reaching $H_1 = 65$ Bit with mnemonic passwords

- Lowercase letters from 13+ words sentence 54 Bit
- 7-bit visible ASCII (incl. %, !, @, #, etc.) 8 Bit
  (adds on average 2 characters $\approx 6.4$ Bit)
- Word replacements (and $\rightarrow$ &), to $\rightarrow$ 2, etc.) 2 Bit
- Different characters (last of each word) 0 Bit
- Complex sentences (rich vocabulary) + 2 Bit

Total 66 Bit
Harnessing Web Archives to Tackle Selected Societal Challenges

Summary

1. Preservation of digital culture
   - 10K pages high-fidelity archive (FAIRest dataset award)
   - Reproduction assessment task
   - 9K pages segmentation dataset
   - Segmentation evaluation measures

2. Critical assessment of information
   - Revert-based vandalism detection
   - 30K edits Wiki vandalism dataset
   - 1M hyperpartisan news dataset
   - Style-based polarity detection
   - Hyperpartisan news challenge (SemEval, 42 teams)

3. Online security and privacy
   - 3B web sentences dataset
   - Position-dependent language model
   - Security estimate: mnemonic passwords
   - Personal archiving tool

Tailored web archiving technology (Webis Web Archiver)

Highlighted aspects:
- Reproducibility
- Temporal dynamics
- Volume