The Impact of Main Content Extraction on Near-Duplicate Detection

Maik Fröbe, Matthias Hagen, Janek Bevendorff, Michael Völske, Benno Stein, Christopher Schröder, Robby Wagner, Lukas Gienapp, Martin Potthast

OSSYM, 11–13 October 2021
webis.de
Motivation

- Web crawls contain many near-duplicates:
  [Fetterly et al.; LA-WEB’03]

**Dog breed**

From Wikipedia, the free encyclopedia

A *dog breed* is a particular strain that was purposefully bred by humans to perform specific tasks, such as herding, hunting, and guarding. When distinguishing breed from type, the *rule of thumb* is that a breed always "breeds true".¹ Dogs are the most variable mammal on earth,
Impact of Main Content Extraction on Near-Duplicate Detection

Motivation

- Web crawls contain many near-duplicates:
  [Fetterly et al.; LA-WEB’03]
Impact of Main Content Extraction on Near-Duplicate Detection

Motivation

- Web crawls contain many near-duplicates:
  [Fetterly et al.; LA-WEB’03]
Impact of Main Content Extraction on Near-Duplicate Detection

Overview

- Near-Duplicates in Web Crawls: Risks and Potentials
  - Risk: Evaluation of Search Engines
    [Bernstein et al.; CIKM’05; Fröbe et al.; ECIR’20]
  - Risk: Training of Learning to Rank Models
    [Fröbe et al.; SIGIR’20]
  - Potential: Transfer of Relevance Labels
    [Fröbe et al.; SIGIR’21]

- Improve Near-Duplicate Detection with Main Content Extraction?
  - Precision vs. Recall
Impact of Main Content Extraction on Near-Duplicate Detection

Risk: Evaluation of Search Engines
[Bernstein et al.; CIKM’05; Fröbe et al.; ECIR’20]

- Cranfield paradigm: Query-document pairs labeled in isolation
- Web Track Topic 194: designer dog breeds
  - 40 of 47 relevant documents duplicate the same Wikipedia article
Impact of Main Content Extraction on Near-Duplicate Detection

Risk: Evaluation of Search Engines
[Bernstein et al.; CIKM’05; Fröbe et al.; ECIR’20]

- Cranfield paradigm: Query-document pairs labeled in isolation
- Web Track Topic 194: designer dog breeds
  - 40 of 47 relevant documents duplicate the same Wikipedia article

- Novelty Principle:
  A document is irrelevant if it is content-equivalent to a document the user has already seen in the results.
Impact of Main Content Extraction on Near-Duplicate Detection

Risk: Evaluation of Search Engines
[Bernstein et al.; CIKM’05; Fröbe et al.; ECIR’20]

- Comparison:
  - Evaluation without the novelty principle
  - Evaluation with the novelty principle (rankings deduplicated)

<table>
<thead>
<tr>
<th>Web Track</th>
<th># Runs</th>
<th>Duplicates removed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$\Delta_{nDCG}$ $\tau$</td>
</tr>
<tr>
<td>2009</td>
<td>71</td>
<td>-6.8% 0.91</td>
</tr>
<tr>
<td>2010</td>
<td>56</td>
<td>-9.9% 0.57</td>
</tr>
<tr>
<td>2011</td>
<td>37</td>
<td>-3.4% 0.92</td>
</tr>
<tr>
<td>2012</td>
<td>28</td>
<td>-12.4% 0.81</td>
</tr>
<tr>
<td>2013</td>
<td>34</td>
<td>-1.8% 0.90</td>
</tr>
<tr>
<td>2014</td>
<td>30</td>
<td>-4.5% 0.94</td>
</tr>
</tbody>
</table>

- Problems caused by near-duplicates:
  - Effectiveness is overestimated
  - “Leaderboard” changes

  ⇒ Risk: Training of Learning to Rank Models  [Fröbe et al.; SIGIR’20]
Impact of Main Content Extraction on Near-Duplicate Detection

Risk: Evaluation of Search Engines
[ Bernstein et al.; CIKM’05; Fröbe et al.; ECIR’20 ]

- Comparison:
  - Evaluation without the novelty principle
  - Evaluation with the novelty principle (rankings deduplicated)

<table>
<thead>
<tr>
<th>Web Track</th>
<th># Runs</th>
<th>Duplicates removed</th>
<th>$\Delta_{nDCG}$</th>
<th>$\tau$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009</td>
<td>71</td>
<td>-6.8%</td>
<td>0.91</td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td>56</td>
<td>-9.9%</td>
<td>0.57</td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td>37</td>
<td>-3.4%</td>
<td>0.92</td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td>28</td>
<td>-12.4%</td>
<td>0.81</td>
<td></td>
</tr>
<tr>
<td>2013</td>
<td>34</td>
<td>-1.8%</td>
<td>0.90</td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>30</td>
<td>-4.5%</td>
<td>0.94</td>
<td></td>
</tr>
</tbody>
</table>

- Problems caused by near-duplicates:
  - Effectiveness is overestimated
  - “Leaderboard” changes

⇒ Risk: Training of Learning to Rank Models [Fröbe et al.; SIGIR’20]
Impact of Main Content Extraction on Near-Duplicate Detection

Risk: Evaluation of Search Engines
[Bernstein et al.; CIKM’05; Fröbe et al.; ECIR’20]

- Comparison:
  - Evaluation without the novelty principle
  - Evaluation with the novelty principle (rankings deduplicated)

<table>
<thead>
<tr>
<th>Web Track</th>
<th># Runs</th>
<th>Duplicates removed</th>
<th>$\Delta_{nDCG}$</th>
<th>$\tau$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009</td>
<td>71</td>
<td>-6.8%</td>
<td>0.91</td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td>56</td>
<td>-9.9%</td>
<td>0.57</td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td>37</td>
<td>-3.4%</td>
<td>0.92</td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td>28</td>
<td>-12.4%</td>
<td>0.81</td>
<td></td>
</tr>
<tr>
<td>2013</td>
<td>34</td>
<td>-1.8%</td>
<td>0.90</td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>30</td>
<td>-4.5%</td>
<td>0.94</td>
<td></td>
</tr>
</tbody>
</table>

- Problems caused by near-duplicates:
  - Effectiveness is overestimated
  - “Leaderboard” changes

⇒ Risk: Training of Learning to Rank Models [Fröbe et al.; SIGIR’20]
Impact of Main Content Extraction on Near-Duplicate Detection

Potential: Transfer of Relevance Labels
[Fröbe et al.; SIGIR’21]

- ClueWeb09: 73,883 relevance judgments (estimated effort: 4-8 months)
- Idea: Transfer relevance judgments to newer crawls (“save” judgment effort)
Impact of Main Content Extraction on Near-Duplicate Detection

Potential: Transfer of Relevance Labels

[Fröbe et al.; SIGIR’21]

- ClueWeb09: 73,883 relevance judgments (estimated effort: 4-8 months)
- Idea: Transfer relevance judgments to newer crawls (“save” judgment effort)

– Relevant for query “used car parts” in ClueWeb09
Impact of Main Content Extraction on Near-Duplicate Detection

Potential: Transfer of Relevance Labels

[Fröbe et al.; SIGIR’21]

- ClueWeb09: 73,883 relevance judgments (estimated effort: 4-8 months)
- Idea: Transfer relevance judgments to newer crawls ("save" judgment effort)

- Relevant for query “used car parts” in ClueWeb09
- Near-Duplicate in ClueWeb12 is also relevant
Impact of Main Content Extraction on Near-Duplicate Detection

Potential: Transfer of Relevance Labels

[Fröbe et al.; SIGIR’21]

- Experiment: Transfer relevance judgments from ClueWeb09 to ClueWeb12
Impact of Main Content Extraction on Near-Duplicate Detection

Potential: Transfer of Relevance Labels
[Fröbe et al.; SIGIR ‘21]

- Experiment: Transfer relevance judgments from ClueWeb09 to ClueWeb12
Impact of Main Content Extraction on Near-Duplicate Detection

Motivation: Main Content Extraction for Near-Duplicate Detection

- Perfect main content extraction would improve precision and recall
  - Recall: Identical content in different boilerplate
Impact of Main Content Extraction on Near-Duplicate Detection

Experimental Setup

- Select 186,819 document pairs from the ClueWebs
  - Group documents by (canonical) URL
  - Select 5,000 groups at random
  - Consider all possible pairs within a group

- We label 500 document pairs as near-duplicate (identical content) or not
  - Uniform sampled over $S_3$ similarity

<table>
<thead>
<tr>
<th></th>
<th>$S_3 \geq 0.9$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>Full Content</td>
<td>0.98</td>
</tr>
<tr>
<td>JusText</td>
<td><strong>1.00</strong></td>
</tr>
<tr>
<td>Jericho</td>
<td>0.97</td>
</tr>
<tr>
<td>Trafilatura</td>
<td>0.77</td>
</tr>
<tr>
<td>Boilerpipe</td>
<td>0.61</td>
</tr>
</tbody>
</table>
Impact of Main Content Extraction on Near-Duplicate Detection

Takeaways

- Not accounting for near-duplicates introduces risks
  - Evaluation of retrieval models
  - Learning to rank

- Transfer of relevance judgments as use-case of near-duplicate-detection

- CopyCat
  - Resource to simplify deduplication in TREC-style experimentals
  - Open source: github.com/chatnoir-eu/chatnoir-copycat

- Future work:
  - Expand experiments
  - Study transfer of relevance labels on MS Marco
Impact of Main Content Extraction on Near-Duplicate Detection

Takeaways

- Not accounting for near-duplicates introduces risks
  - Evaluation of retrieval models
  - Learning to rank

- Transfer of relevance judgments as use-case of near-duplicate-detection

- CopyCat
  - Resource to simplify deduplication in TREC-style experimentals
  - Open source: github.com/chatnoir-eu/chatnoir-copycat

- Future work:
  - Expand experiments
  - Study transfer of relevance labels on MS Marco

Thank You!