CopyCat: Near-Duplicates Within and Between the ClueWeb and the Common Crawl

Martin-Luther-Universität Halle-Wittenberg¹  Bauhaus-Universität Weimar²  Leipzig University³
SIGIR, 11–15 July 2021
webis.de
CopyCat: Near-Duplicates in the ClueWeb and Common Crawl

Motivation (1)

- 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". Dogs are the most variable mammal on earth,
CopyCat: Near-Duplicates in the ClueWeb and Common Crawl

Motivation (1)

- Web crawls contain many near duplicates:
  [Fetterly et al.; LA-WEB’03]
CopyCat: Near-Duplicates in the ClueWeb and Common Crawl

Motivation (1)

- Web crawls contain many near-duplicates: [Fetterly et al.; LA-WEB’03]
Motivation (2)

- Impact of near-duplicates on the evaluation of search engines:
  [Bernstein et al.; CIKM’05; Fröbe et al.; ECIR’20]
  - Novelty principle:
    A document is irrelevant if it is content-equivalent to a document the user has already seen in the results.
CopyCat: Near-Duplicates in the ClueWeb and Common Crawl

Motivation (2)

- Impact of near-duplicates on the evaluation of search engines:
  [Bernstein et al.; CIKM’05; Fröbe et al.; ECIR’20]

  - Novelty principle:
    A document is irrelevant if it is content-equivalent to a document the user has already seen in the results.

  - Novelty principle on the TREC Web Tracks:
    - Average nDCG scores in 2012 decrease by 17%
    - “Leaderboard” changes ($\tau$ of 0.49 in 2010)
CopyCat: Near-Duplicates in the ClueWeb and Common Crawl

Motivation (2)

- Impact of near-duplicates on the evaluation of search engines:
  [Bernstein et al.; CIKM’05; Fröbe et al.; ECIR’20]
  - Novelty principle:
    A document is irrelevant if it is content-equivalent to a document the user has already seen in the results.
  - Novelty principle on the TREC Web Tracks:
    • Average nDCG scores in 2012 decrease by 17%  
    • “Leaderboard” changes ($\tau$ of 0.49 in 2010)

- Sampling bias in learning to rank:
  [Fröbe et al.; SIGIR’20]
  - Unintentional oversampling
  - Bias towards relevant near-duplicates
CopyCat: Near-Duplicates in the ClueWeb and Common Crawl

Resources

- The CopyCat resource simplifies deduplication in IR experiments
  1. Compilation of near-duplicate documents within widely used web crawls
     - Inclusion and exclusion lists
     - Covered crawls:
       - ClueWeb09, ClueWeb12
       - ClueWeb09B, ClueWeb12B13
       - Common Crawl 2015-11, Common Crawl 2017-04
  2. Software library for deduplication of arbitrary document sets
     - SimHash implemented in Apache Spark for large web crawls
     - CLI for smaller document sets
       - TREC run files
       - TREC qrel files
       - Assessment pools
Pilot study on canonical links:
- Canonical links allow authors of web pages to indicate duplicate content
- Between 0.3% and 49% of documents use canonical links

Ground-Truth:
- Semi-automatic assessment of 361 m document pairs
- Sampled from equivalence classes of canonical links
- Calculated the exact syntactic similarity for all document pairs
- Assessed 400 document pairs, choosing a precision-oriented threshold
- Document pairs with similarity above the threshold are near-duplicates
CopyCat: Near-Duplicates in the ClueWeb and Common Crawl

Deduplication of Large Web Crawls with CopyCat

- SimHash implemented in Apache Spark
- Fine-tuned on 361 m ground-truth document pairs
- CopyCat combines 2 strategies:
  - SimHash within equivalence classes of canonical links (Precision: 0.94)
  - SimHash within entire crawls (Precision: 0.97)
- Deduplication results with CopyCat:
  - Cluster: 135 nodes
  - Resulting inclusion/exclusion lists allow precision-oriented deduplication

<table>
<thead>
<tr>
<th>Web crawl</th>
<th>cw09</th>
<th>cw12</th>
<th>cc15</th>
<th>cc17</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compr. size</td>
<td>4.0 TB</td>
<td>4.5 TB</td>
<td>28.1 TB</td>
<td>54.0 TB</td>
</tr>
<tr>
<td>Documents</td>
<td>1.0 b</td>
<td>731.7 m</td>
<td>1.8 b</td>
<td>3.1 b</td>
</tr>
<tr>
<td>Duplicates</td>
<td>145.8 m</td>
<td>204.3 m</td>
<td>951.2 m</td>
<td>1.0 b</td>
</tr>
</tbody>
</table>
Showcase (1): Duplicates in Run Files

- We used the CopyCat CLI to deduplicate run files submitted to the TREC Web Tracks

<table>
<thead>
<tr>
<th>Web track</th>
<th>Near-dupl. in runs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year</td>
<td>Runs</td>
</tr>
<tr>
<td>2009</td>
<td>71</td>
</tr>
<tr>
<td>2010</td>
<td>56</td>
</tr>
<tr>
<td>2011</td>
<td>37</td>
</tr>
<tr>
<td>2012</td>
<td>28</td>
</tr>
<tr>
<td>2013</td>
<td>34</td>
</tr>
<tr>
<td>2014</td>
<td>30</td>
</tr>
</tbody>
</table>
CopyCat: Near-Duplicates in the ClueWeb and Common Crawl

Showcase (2): Relevance Label Transfer

- ClueWeb09: 73,883 relevance judgments (estimated effort: 4-8 months)
- Idea: Transfer relevance judgments to newer crawls (“save” judgment effort)
CopyCat: Near-Duplicates in the ClueWeb and Common Crawl Showcase (2): Relevance Label Transfer

- 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
CopyCat: Near-Duplicates in the ClueWeb and Common Crawl

Showcase (2): Relevance Label Transfer

- 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
CopyCat: Near-Duplicates in the ClueWeb and Common Crawl
Showcase (2): Relevance Label Transfer

- Experiment settings
  - Qrels deduplicated with CopyCat CLI
  - Precision-oriented near-duplicate threshold
CopyCat: Near-Duplicates in the ClueWeb and Common Crawl

Showcase (2): Relevance Label Transfer

- Experiment settings
  - Qrels deduplicated with CopyCat CLI
  - Precision-oriented near-duplicate threshold

![Graph showing transferred judgments over years]
CopyCat: Near-Duplicates in the ClueWeb and Common Crawl

Showcase (2): Relevance Label Transfer

- Experiment settings
  - Qrels deduplicated with CopyCat CLI
  - Precision-oriented near-duplicate threshold
CopyCat: Near-Duplicates in the ClueWeb and Common Crawl

Takeaways

- CopyCat simplifies deduplication in IR experiments
  - Ready-to-use inclusion and exclusion lists of near-duplicates
  - Software library

- Showcase on relevance label transfer
  - Few near-duplicates of judged documents occur in newer crawls
  - New relevance judgments needed to evaluate “old” topics on new crawls

- Future work:
  - Increase recall with main content extraction
  - Consider near-duplicates of relevant passages between documents

- Code, Paper, Slides: webis.de/publications
CopyCat: Near-Duplicates in the ClueWeb and Common Crawl

Takeaways

- CopyCat simplifies deduplication in IR experiments
  - Ready-to-use inclusion and exclusion lists of near-duplicates
  - Software library

- Showcase on relevance label transfer
  - Few near-duplicates of judged documents occur in newer crawls
  - New relevance judgments needed to evaluate “old” topics on new crawls

- Future work:
  - Increase recall with main content extraction
  - Consider near-duplicates of relevant passages between documents

- Code, Paper, Slides: webis.de/publications

Thank You!