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

The Resource in a Nutshell

Web crawls contain many near-duplicates (i.e., documents with different URLs but very similar content). ChatNoir-CopyCat-21 is a precision-oriented near-duplicate resource consisting of (1) lists of near-duplicates within the ClueWeb09, the ClueWeb12, and two Common Crawl snapshots, as well as between selections of these crawls, and (2) a software library to deduplicate arbitrary document sets (e.g., TREC runs).

Compilation of Near-Duplicates

We compile the near-duplicates found in the ClueWeb09 (CW09), the ClueWeb12 (CW12), and the Common Crawls 2015-11 (CC15) and 2017-04 (CW17) into inclusion and exclusion lists for easy deduplication.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Web Crawl</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compressed size</td>
<td>CW09 4.0 TB</td>
<td>CW12 4.5 TB</td>
</tr>
<tr>
<td>Documents</td>
<td>CW09 1.0 b</td>
<td>CW12 731.7 m</td>
</tr>
<tr>
<td>SimHash duplicates</td>
<td>CW09 145.6 m</td>
<td>CW12 201.2 m</td>
</tr>
<tr>
<td>Canonical links</td>
<td>CW09 3.0 m</td>
<td>CW12 67.1 m</td>
</tr>
<tr>
<td>Crawled duplicates</td>
<td>CW09 15.0 m</td>
<td>CW12 11.2 m</td>
</tr>
<tr>
<td>CopyCat duplicates</td>
<td>CW09 145.8 m</td>
<td>CW12 204.3 m</td>
</tr>
</tbody>
</table>

Parameter-Tuning on Canonical Links

We conduct a pilot study to semi-automatically label 360 million document pairs sampled from equivalence classes of canonical links.

Deduplication in IR Experiments

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

Impact of the Novelty Principle:
- System effectiveness is overestimated
- Ranking of systems changed

Example: The 47 relevant ClueWeb09 documents for the query designer dog breeds contain 40 near-duplicates of the Wikipedia article on the left.

Transfer of Relevance Labels

Motivation (cw09 Web Track):
- 73,883 relevance judgments
- Estimated effort: 4 to 8 months

Idea:
Find near-duplicates of documents judged in an older crawl in a newer crawl to reuse existing topics while “saving” relevance judgments.

Example:
The pair shows a ClueWeb09 document judged as relevant for the query used car parts and its ClueWeb12 near-duplicate.