Detecting Duplicate Web Documents using Clickthrough Data

Filip Radlinski,
Paul N. Bennett, Emine Yilmaz

Microsoft
Motivation

• Near-Duplicates and duplicates are common.
• In many cases, showing them isn’t helpful.
• Usually identified based on page content.
  ➢ not sensitive to query.
• We show that clicks give a more context-sensitive duplication signal.
Previous approaches

• Detecting Duplication using Content
  – Fingerprints for exact duplicates:
    Compute a fingerprint for each document. If fingerprints match, check if the documents are the same.
  – Shingling to find near duplicates:
    For each ngram, compute a fingerprint. Measure the similarity between a summary of fingerprints.

• Reducing redundancy in search results
  – Given relevance and similarity, rank results by relevance minus redundancy

• These methods don’t depend on the query.
Detecting Duplicate Web Documents using Clickthrough Data, Radlinski, Bennett & Yilmaz, WSDM 2011

Redundancy Score

\[
bias_{uv} = \frac{c_{\hat{u}v}}{c_{\hat{u}v} + c_{u\hat{v}} + c_{\hat{u}\hat{v}}}
\]
Detecting Duplicate Web Documents using Clickthrough Data, Radlinski, Bennett & Yilmaz, WSDM 2011

Redundancy Score

\[ \text{bias}_{vu} = \frac{c_{\hat{v}u}}{c_{\hat{v}u} + c_{v\hat{u}} + c_{\hat{v}\hat{u}}} \]
Redundancy Score

- The redundancy score is the minimum of these two ratios:

\[ r(u, v) = \min \left( \frac{c^{uv}}{c^{uv} + c^{u\tilde{v}} + c^{\tilde{u}v}}, \frac{c^{\tilde{v}u}}{c^{\tilde{v}u} + c^{v\tilde{u}} + c^{\tilde{v}\tilde{u}}} \right) \]

- In words:
  
  Across both presentation orders, what is the minimum rate at which just the top result is clicked?

- High presentation bias \( \leftrightarrow \) High redundancy score
Redundancy Score

- If in both orders, the top result always gets clicked:

  ➢ The results are probably duplicate
  ➢ Or users always just click on the top result: we’ll check

Detecting Duplicate Web Documents using Clickthrough Data, Radlinski, Bennett & Yilmaz, WSDM 2011
Redundancy Score

• If one of the results is always clicked on, even when its lower:

 ➢ That result is preferred, these are not redundant

Most real document pairs are in between
Relationship to Click-Skip and FairPairs

• The Click-Skip approach:
  A clicked document is more relevant than skipped ones above.

• FairPairs:
  If you present documents in both orders, the one with higher bottom click rate is more relevant.

  ➔ A **bottom click** is a **relevance** signal

• This work:
  Present documents in both orders. If the top one is always clicked, the documents are duplicate.

  ➔ A **top click** is a **duplication** signal

Detecting Duplicate Web Documents using Clickthrough Data, Radlinski, Bennett & Yilmaz, WSDM 2011
Score Distribution

- Does real document pairs exist with a variety of redundancy scores?
Classes of Duplication

Inspecting pairs of documents with high redundancy score, three types of duplicates jump out:

- **Exact duplicates**: Both pages appear identical, perhaps with the exception of ads.
- **Content duplicates**: Both pages provide the same / very related information (**for this query**), but from different sources.
- **Navigational duplicates**: Getting from one page to the other is very easy.
Example: Navigational Duplicates

- Other common examples:
  - Bank homepage vs. online banking page
  - Related Amazon products or eBay auctions

Detecting Duplicate Web Documents using Clickthrough Data, Radlinski, Bennett & Yilmaz, WSDM 2011
Example: Content Duplicates

• Other common examples:
  – Competing song lyrics websites
  – Different recipe websites
  – Competing sofa manufacturers
Evaluation Approach

• Test if redundancy score tells us about duplication
  1. Sample tuples with variety of redundancy scores
  2. Judge the (query, url, url) triplets for duplication
  3. Measure agreement between score and judgments
  4. Train duplicate classifiers

• For each triplet, asked three questions:
  1. Which page is most relevant to the query?
  2. How similar is the utility of these pages for the query?
  3. Is it “easy” to navigate from either page to the other?
# Judging Duplication

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Both equally</td>
<td>Identical</td>
<td>-</td>
<td>6%</td>
</tr>
<tr>
<td>Both equally</td>
<td>any</td>
<td>Yes within</td>
<td>5%</td>
</tr>
<tr>
<td>Left or Right</td>
<td>any</td>
<td>Yes within</td>
<td>13%</td>
</tr>
<tr>
<td>Both equally</td>
<td>Very similar</td>
<td>No</td>
<td>16%</td>
</tr>
<tr>
<td>Left or Right</td>
<td>Very similar</td>
<td>No</td>
<td>8%</td>
</tr>
<tr>
<td>Both equally</td>
<td>Related</td>
<td>No</td>
<td>4%</td>
</tr>
<tr>
<td>Left or Right</td>
<td>Related</td>
<td>No</td>
<td>15%</td>
</tr>
<tr>
<td>Different intents</td>
<td>-</td>
<td>No</td>
<td>12%</td>
</tr>
<tr>
<td>Left or Right</td>
<td>Different</td>
<td>No</td>
<td>12%</td>
</tr>
<tr>
<td>Left/Right/Both</td>
<td>any</td>
<td>Yes across</td>
<td>2%</td>
</tr>
<tr>
<td>Neither Relevant</td>
<td>any</td>
<td>any</td>
<td>4%</td>
</tr>
<tr>
<td>other</td>
<td>other</td>
<td>other</td>
<td>3%</td>
</tr>
</tbody>
</table>

- Exact duplicates
- Navigational duplicates
- Content duplicates
- “Weak” content duplicates
- Not Duplicate
Judging Duplication

• Inter-judge agreement on a small set tells us it's tricky to make these judgments

<table>
<thead>
<tr>
<th>Judgment 2</th>
<th>Judgment 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exact (E)</td>
<td>0</td>
</tr>
<tr>
<td>Navigational (N)</td>
<td>19</td>
</tr>
<tr>
<td>Content (C)</td>
<td>31</td>
</tr>
<tr>
<td>Weak Cont. ($C_w$)</td>
<td></td>
</tr>
<tr>
<td>Not Duplicate</td>
<td>78</td>
</tr>
</tbody>
</table>

• The hard one is content vs. weak content vs. not duplicate.
  – The judgments often differ by one level
Detecting Duplicate Web Documents using Clickthrough Data, Radlinski, Bennett & Yilmaz, WSDM 2011

Judgments vs Redundancy Score

![Bar chart showing the fraction of tuples for different redundancy scores.]

- Not Duplicate
- Content Duplicate
- Navigational Duplicate
- Exact Duplicate

Redundancy Score, \( r(u,v) \)

Fraction of Tuples
Judgments vs Redundancy Score

- Exact Duplicate
- Navigational Duplicate
- Content Duplicate
- Not Duplicate

Redundancy Score, $r(u,v)$

Fraction of URL Pairs
Learning Duplicate Detection

- Training on all our data, different click scores distinguish between the different classes:
Learning Duplicate Detection

![Precision vs Recall Graph](image)

- **CART**
- **MetaCost-CART**
- **LogReg**
- **MetaCost-LogReg**

Source: Bennett & Yilmaz, WSDM 2011
Acting on Duplication

• Assuming that we can detect the classes of duplicates, what should we do?
  – Exact duplicates: Remove them.
  – Navigational duplicates:
    • Probably pick just one, but the right one!
  – Content duplicates:
    • Maybe tweak the UI to show them as alternatives?

• Beyond modifying search result rankings:
  – Better relevance from clicks
  – Clean up training/evaluation data
Conclusions & Open Questions

- We proposed a taxonomy of duplication.
- Clicks can be used to distinguish among the classes.
- Presentation bias has a limited effect on non-duplicates.
- Sometimes (near-)duplicates are useful. How can we measure how useful they are?
- How to obtain more reliable evaluation data?
- Investigate how other duplication signals (e.g. content) help classification.