Web Spam Filtering in Internet Archives

Miklós Erdélyi\textsuperscript{1}, András A. Benczúr\textsuperscript{1}, Dávid Siklósi\textsuperscript{1}
Julien Masanès\textsuperscript{2}

\textsuperscript{1}Hungarian Academy of Sciences (MTA SZTAKI)
Data Mining and Web Search Group

\textsuperscript{2}European Archive Foundation, France
Part I: Archival Institutions

- Web archives: 39 institutions under International Internet Preservation Consortium (IIPC) and more
- Loose collaboration, code and technology sharing
- Operated usually from moderate budget
- Effort sharing is crucial
Web spam in Archives

- Slightly different policies
  - May want to archive spam to preserve whole picture
  - Might be worried more about false positives
  - Will perhaps not serve general search queries to users
- But increasingly affected by spam becoming more and more costly if not fought against:
  - 10+% of sites, near 20% of HTML pages
- We have conducted a survey...
Survey results (1)

- Participants: 20 archival institutions from all around the Globe
  - National and other libraries
    - Library of Congress
    - National Library of Denmark
    - ...
    - Internet Archive
    - Documentation Centre for Dutch Political Parties
    - Virtual Knowledge Studio
    - ...
  - “Is spam or fake Web content a problem in your crawling and capturing process?”
    - Yes (39%)
    - No problem (4%)
    - Only 1 respondent expects no problem by spam even in the future
  - The type of spam met by archives, counter measures ...
Survey results (2)

• “If you do meet spam during capturing, of what type is that spam?”

<table>
<thead>
<tr>
<th>Type</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blog comment spam</td>
<td>20% (2)</td>
</tr>
<tr>
<td>Link farms</td>
<td>50% (5)</td>
</tr>
<tr>
<td>Copied content</td>
<td>60% (6)</td>
</tr>
<tr>
<td>Garbage content</td>
<td>70% (7)</td>
</tr>
</tbody>
</table>

• “If spam has impact on your Web archiving process, what actions do you undertake?”

<table>
<thead>
<tr>
<th>Action</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>We drop pages with spam or fake content</td>
<td>18,20% (2)</td>
</tr>
<tr>
<td>We drop sites with spam or fake content</td>
<td>45,50% (5)</td>
</tr>
<tr>
<td>We apply filters to avoid such noise</td>
<td>54,50% (6)</td>
</tr>
<tr>
<td>After capturing we manually correct the crawl</td>
<td>27,30% (3)</td>
</tr>
<tr>
<td>We see no options to avoid noise</td>
<td>27,30% (3)</td>
</tr>
</tbody>
</table>
Survey results (3)

• Low resources on spam filtering but...
• “If you undertake actions to diminish the spam problem in the Web archive of your institute, can you estimate how much you invest in this?”
  • “difficult to estimate”
  • “I would spend perhaps 3 or 4 days creating lists of seeds to filter out of the forthcoming crawl.”
  • “10 minutes - 1 hour per site”
  • “We use 2-5 minutes per website when going through the list of potential spam sites.”
  • “We do not do anything to edit captured content. We foresee that this would not scale, and that it would invite questions about the archive’s authenticity.”
Archive specific needs

- Filtering
  - Analyze and train by a “bootstrap” crawl
  - Filter newly appeared hosts crawl time
  - Aid manual assessment (active learning)
- Collaboration
  - Aid information and label sharing
  - Use a filter model trained possibly at another institution
  - Catch spam farms that span more top level domains
Part II: Spam hunting in Archives

- Dataset
  (WEBSPAM-UK snapshots)
- Temporal features
- Results
Dataset

- 13 UbiCrawler .uk snapshots
Temporal features (1)

- Transformations - without generating new features
  - Normalization: centralized feature values
  - Variance of feature values across snapshots
- Content change
  - Simple bag-of-words model
  - Similarity between two snapshots
  - Aggregated by average, maximum and variance
- Classification stability
  - On average, how easy it is to classify the given host?
Sample histograms

- Host-level standard deviation of top 200 corpus recall
- Average host content similarity in the bag of words model
- Variance of the probability of correctly predicted spamicity
Results (1)

- Using “public” content features, classification by C4.5

- Related feature sets:

<table>
<thead>
<tr>
<th>Setup</th>
<th>Challenge</th>
<th>New host</th>
<th>2006 → 2007</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training set size</td>
<td>1,201</td>
<td>4,000</td>
<td>10,662</td>
</tr>
<tr>
<td>Public content</td>
<td>0.753</td>
<td>0.699</td>
<td>0.730</td>
</tr>
<tr>
<td>BOW</td>
<td>0.619</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Stability</td>
<td>0.776</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Variance</td>
<td>0.618</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Results (2)

- Combination by log-odds averaging based random forest:

<table>
<thead>
<tr>
<th>Combinations</th>
<th>Challenge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content + BOW</td>
<td>0.729</td>
</tr>
<tr>
<td>Content + stability</td>
<td>0.766</td>
</tr>
<tr>
<td>Content + variance</td>
<td>0.726</td>
</tr>
<tr>
<td>Content + BOW + stability + variance</td>
<td>0.777</td>
</tr>
</tbody>
</table>

- **Conclusion**: temporal change based features seem to be useful by these preliminary experiments
Questions?

Miklós Erdélyi

datamining.sztaki.hu/
miklos@ilab.sztaki.hu