Web Spam Challenge Proposal for Filtering in Archives

András A. Benczúr¹, Miklós Erdélyi¹, Dávid Siklósi¹
Julien Masanès²

¹Hungarian Academy of Sciences (MTA SZTAKI)
Data Mining and Web Search Group

²European Archive Foundation, France
Emerging need for filtering in archives

Current situation:

“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.”

Filtering practice:

A nordic national library is spending 4 man months on filtering after each of its domain crawls

... and more seen in the morning session
Archive specialties 1: time series

• Series of crawl snapshots of the same domain
• Maintain filtering quality as time evolves
• Classify newly appeared hosts (crawl-time filtering)
• Classify based on content, ownership change
• Catch parked domain use
Archive specialties 2: model transfer

Archives operate on a moderate budget, thus they need to cooperate, share efforts

• Transfer model compiled by one archive to another
  • Train on one top level domain (TLD), classify on another (language, link usage, size)
• Adjust for different crawl coverage (e.g. initial small bootstrap crawl)
  e.g. WEBSPAM-UK2006: 10K hosts
  WEBSPAM-UK2007: 114K hosts
Proposed new tasks

1. Time series classification
   - Feature generation based on temporal change
   - Classify a new snapshot based on earlier labels

2. New site classification

3. Model transfer
   - Train on existing .uk labels
   - Test labels
     - for hosts of a new .uk crawl
     - for hosts of .eu
   - Different languages (English and non-English classification tasks)
Expected methods 1: language

• Tf.idf will not work for model transfer
  Tf.idf with SVM is too strong
• Handling a mixture of languages (.eu)
  • Forget tf.idf based classifiers?
  • Detect language, then translate?
  • Some “public” features may be crucial and may need more variants, e.g. word statistics normalized across languages
Expected methods 2: temporal change

• Time series should be useful - 3 AIRWeb 2009 papers on this topic
  • Graph evolution [Chung, Toyoda, Kitsuregawa]
    3 yearly .jp snapshots
  • Content change [Dai, Davison, Qi]
    content history from Wayback Machine
  • Change, variance of “public” features [ours]
    13 UbiCrawler.uk snapshots
Expected methods 3: normalization

- Precompiled “public” feature set hard to beat over UK2007 (AUC 0.80 by itself)
- But worse performance if coverage differs

UK2007 AUC 0.73 trained over 10K UK2006 labels

<table>
<thead>
<tr>
<th></th>
<th>UK2006</th>
<th>UK2007</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hosts</td>
<td>10,660</td>
<td>114,529</td>
</tr>
<tr>
<td>Spam %</td>
<td>19.8</td>
<td>5.3</td>
</tr>
</tbody>
</table>

Internet Archive has over 2M .uk domains!

- Need new tricks to normalize, e.g. compare with snapshot global average, etc.
New data sets?

• 13 .uk snapshots, maximum 400 pages per host extract, compressed 500GB, copied from Milan server in 2 weeks
  • Could be distributed by shipping disks
• More crawls with careful selection of hosts
  • .uk - EA estimates over 2M .uk hosts compared to 110K in UK2007
  • .eu another 3.2M hosts
• We have code for most “public” features
New labels?

- Training from existing UK2006-7 labels
- Creating test label set would need additional volunteer+EA assessment
- New UK2007 host labels
  - More newly appeared host labels (260 in UK2007, AUC only 0.699)
  - Label the same host in multiple snapshots with large content change (e.g. for ownership change, parked domain detection)
Summary

• Single-crawl spam filtering task may have lost attention because
  • Tf.idf based features work well enough (hard to beat)
  • Larger coverage appears to deteriorate graphical learning methods (performed better for UK2006)
  • “Public” feature set is too strong

• Multiple-crawl filtering provides new tasks to
  • Define temporal change and time series based features
  • Apply language independent or cross-lingual methods
  • Normalize, stabilize features across different crawl instances, coverage and domains
Questions?

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

... and also Julien Masanès
julien@europoparchive.org