Microsoft Silicon Valley Web Spam Challenge Entry

Steve Chien, Dennis Fetterly, Mark Manasse, Marc Najork, and Alexandros Ntoulas
Our Approach

• Use machine learning to combine
  – Features provided by the organizers
  – Features from our previous work
  – New features where we focus on those that have some creation cost
Additional Features

• **URL**
  – Features derived from the URLs in the UK 2006-05 dataset, such as the number of dots, dashes, and digits in the hostname

• **Page Content**
  – Features derived from word frequency analysis in the 77 million pages
  – Features based on grouping documents into sets of near-duplicate documents

• **Graph Structure**
  – Features indicative of link exchange based on the UK 2006-05 page-level and host-level web graph

• **Evolution**
  – Results from re-crawling the 77 million URLs in the UK2006-05 dataset
  – URL overlap with web crawl that occurred in 2002

• **Economic**
  – Features derived from the registrar records for the 7,707 domains in the UK 2006-05 dataset
  – Features based on the publisher ID of any Google AdSense advertisements embedded in the 77 million pages
Feature Selection

• Ended up with 322 features
• Avoid over-fitting by selecting only most discriminating features
• Used feature selection algorithms in WEKA
  – Evaluated features using several attribute evaluators, search methods, and cutoff values
• The 75 features used were identified using the Ranker search method and the ChiSquared attribute evaluator.
Top 10 Ranked Features

Observations:
• Validates that TrustRank (6/10) and neighborhood are important (2/10) for spam detection
• Economic features are also important

<table>
<thead>
<tr>
<th>Rank</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>1761.73</td>
<td>average_spamicity_neighbors_PASS2</td>
</tr>
<tr>
<td>1756.16</td>
<td>log_OP_trustrank_hp_div_pagerank_hp_CP_</td>
</tr>
<tr>
<td>1741.12</td>
<td>log_OP_trustrank_hp_div_indegree_hp_CP_</td>
</tr>
<tr>
<td>1518.52</td>
<td>average_spamicity_neighbors_PASS1</td>
</tr>
<tr>
<td>1283.88</td>
<td>L_trustrank_hp</td>
</tr>
<tr>
<td>1177.37</td>
<td>log_OP_trustrank_mp_div_pagerank_mp_CP_</td>
</tr>
<tr>
<td>1131.78</td>
<td>MeanHostsAdSenseld</td>
</tr>
<tr>
<td>1128.39</td>
<td>log_OP_trustrank_mp_div_indegree_mp_CP_</td>
</tr>
<tr>
<td>1032.10</td>
<td>L_trustrank_mp</td>
</tr>
<tr>
<td>838.15</td>
<td>STD_83</td>
</tr>
</tbody>
</table>
Evaluation

- Evaluation on the 5,622 labeled hosts
- Used ten-fold cross validation
- Best classifier used bagging in combination with a C4.5 decision tree

<table>
<thead>
<tr>
<th>Class</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-spam</td>
<td>96.8%</td>
<td>97.6%</td>
</tr>
<tr>
<td>Spam</td>
<td>70.5%</td>
<td>80.2%</td>
</tr>
</tbody>
</table>
Acknowledgements

• Challenge Organizers
• UK2006-05 collection coordinators