Detecting Spam Web Pages through Content Analysis

Alex Ntoulas\(^1\), Marc Najork\(^2\), Mark Manasse\(^2\), Dennis Fetterly\(^2\)

\(^1\) UCLA Computer Science Department
\(^2\) Microsoft Research, Silicon Valley
Web Spam: raison d’etre

- E-commerce is rapidly growing
  - Projected to $329 billion by 2010
- More traffic → more money
- Large fraction of traffic from Search Engines
- Increase Search Engine referrals:
  - Place ads 😊
  - Provide genuinely better content 😊
  - Create Web spam … 😞
Web Spam
(you know it when you see it)

For excellent FREE PICTURE WHIPPING DRAWING of JAPANESE WORDS click above.

These links below are just here for the search engines. Anyway, enjoy all the material we have for you at x-stasy you're sure to be excited.

More links -
FREE PICTURE JAPANESE WORDS WHIPPING DRAWING FREE PICTURE JAPANESE WORDS WHIPPING DRAWING FREE PICTURE JAPANESE WORDS WHIPPING DRAWING FREE PICTURE
Defining Web Spam

- **Spam Web page:**
  A page created for the sole purpose of attracting search engine referrals (to this page or some other “target” page)

- Ultimately a judgment call
  - Some web pages are borderline cases
Why Web Spam is Bad

- Bad for users
  - Makes it harder to satisfy information need
  - Leads to frustrating search experience

- Bad for search engines
  - Wastes bandwidth, CPU cycles, storage space
  - Pollutes corpus (infinite number of spam pages!)
  - Distorts ranking of results
How pervasive is Web Spam?

Dataset

- Real-Web data from the MSNBot crawler
  - Collected during August 2004
- Processed only MIME types
  - text/html
  - text/plain
- 105,484,446 Web pages in total
Spam per Top-level Domain

Percentage of spam

Top-level domain

95% confidence
Spam per Language

Percentage of spam

French  German  English  Japanese  Chinese

95% confidence
Detecting Web Spam (1)

- We report results for English pages only (analysis is similar for other languages)
- ~55 million English Web pages in our collection
- Manual inspection of a random sample of the English pages
- Sample size: 17,168 Web pages
  - 2,364 (13.8%) spam
  - 14,804 (86.2%) non-spam
Detecting Web Spam (2)

- Spam detection: A classification problem
  - Given salient features of a Web page, decide whether the page is spam

- Which “salient features”?  
  - Need to understand spamming techniques to decide on features
  - Finding right features is “alchemy”, not science

- We focus on features that  
  - Are **fast to compute**  
  - Are **local** (i.e. examine a page in isolation)
Distribution of Word-counts in <title>

- Spam more likely in pages with more words in title
Visible Content of a Page

Visible Content = \frac{\text{size (in bytes) of visible words}}{\text{size (in bytes) of the page}}
Distribution of Visible-content fractions

- Spam not likely in pages with little visible content
This master of human resource management college online accredited doctoral degrees! About it or online or courses? Set online content medical content degrees sale mafm certificate on-line your courses? None mtm study major certified professional course master of human resource management college, therefore..

This master of human resource management college online accredited doctoral degrees! About it or online or courses? Set online content medical content degrees sale mafm certificate on-line your courses? None mtm study major certified professional course master of human resource management college, therefore.

This master of human resource management college online accredited doctoral degrees! About it or online or courses? Set online content medical content degrees sale mafm certificate on-line your courses? None mtm study major certified professional course master of human resource management college, therefore.

This master of human resource management college online accredited doctoral degrees! About it or online or courses? Set online content medical content degrees sale mafm certificate on-line your courses? None mtm study major certified professional course master of human resource management college, therefore.
zipRatio of a page

\[
\text{zipRatio} = \frac{\text{size (in bytes) of uncompressed non-HTML text}}{\text{size (in bytes) of compressed non-HTML text}}
\]
Distribution of zipRatios

- Spam more likely in pages with high zipRatio
Obscure Content

very rare words
Independence Likelihood Model

- Consider all $n$-grams $w_{i+1} \ldots w_{i+n}$ of a Web page, with $k$ $n$-grams in total.
- Probability that $n$-gram occurs in collection:

$$P(w_{i+1} \ldots w_{i+n}) = \frac{\text{number of occurrences of n-gram}}{\text{total number of n-grams}}$$

- Independence likelihood model

$$IndepLH = -\frac{1}{k} \sum_{i=0}^{k-1} \log P(w_{i+1} \ldots w_{i+n})$$

- Low $IndepLH \rightarrow$ high probability of page’s existence
Distribution of 3-gram Likelihoods (Independence)

- Pages with high & low likelihoods are spam
- Longer n-grams seem to help
Conditional Likelihood Model

- Consider all \( n \)-grams \( w_{i+1} \ldots w_{i+n} \) of a Web page, with \( k \) \( n \)-grams in total
- Probability that \( n \)-gram \( w_{i+1} \ldots w_{i+n} \) occurs, given that \( (n-1) \)-gram \( w_{i+1} \ldots w_{i+n-1} \) occurs

\[
P(w_{i+n} \mid w_{i+1} \ldots w_{i+n-1}) = \frac{P(w_{i+1} \ldots w_{i+n-1}w_{i+n})}{P(w_{i+1} \ldots w_{i+n-1})}
\]

- Conditional likelihood model

\[
CondLH = -\frac{1}{k} \sum_{i=0}^{k-1} \log P(w_{i+n} \mid w_{i+1} \ldots w_{i+n-1})
\]

- Low \( CondLH \) \( \rightarrow \) high probability of page’s existence
Distribution of 3-gram Likelihoods (Conditional)

- Pages with high & low likelihoods are spam
- Longer n-grams seem to help
Spam Detection as a Classification Problem

- Use the previously presented metrics as features for a classifier
- Use the 17,168 manually tagged pages to build a classifier
- We show results for a decision-tree – C4.5 (other classifiers performed similarly)
Putting it all together

- size of the page
- static rank
- link depth
- number of dots/dashes/digits in hostname
- hostname length
- hostname domain
- number of words in the page
- number of words in the title
- fraction of anchor text
- average length of the words
- fraction of visible content
- fraction of top-100,200,500,1000 words in the text
- fraction of text in top-100,200,500,1000 words
- compression ratio (zipRatio)
- occurrence of the phrase “Privacy Policy”
- occurrence of the phrase “Privacy Statement”
- occurrence of the phrase “Customer Service”
- occurrence of the word “Disclaimer”
- occurrence of the word “Copyright”
- occurrence of the word “Fax”
- occurrence of the word “Phone”
- likelihood (independence) 1,2,3,4,5-grams
- likelihood (conditional probability) 2,3,4,5-grams
A Portion of the Induced Decision Tree

- \( \text{IhoodIndep5gram} \leq 13.73377 \)
- \( \text{IhoodIndep5gram} > 13.73377 \)
- \( \text{fracTop1KInText} \leq 0.062 \)
- \( \text{fracTop1KInText} > 0.062 \)
- \( \text{fracTextTop500} \leq 0.646 \)
- \( \text{fracTextTop500} > 0.646 \)
- \( \text{fracTop500InText} \leq 0.154 \)
- \( \text{fracTop500InText} > 0.154 \)
- spam
- non-spam
Evaluation of the Classifier

- We evaluated the classifier using 10-fold cross validation (and random splits)
- Training samples: 17,168
  - 2,364 spam
  - 14,804 non-spam
- Correctly classified: 16,642 (96.93 %)
- Incorrectly classified: 526 (3.07 %)
### Confusion Matrix & Precision/Recall

<table>
<thead>
<tr>
<th>classified as</th>
<th>spam</th>
<th>non-spam</th>
</tr>
</thead>
<tbody>
<tr>
<td>spam</td>
<td>1,940</td>
<td>424</td>
</tr>
<tr>
<td>non-spam</td>
<td>366</td>
<td>14,440</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>class</th>
<th>recall</th>
<th>precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>spam</td>
<td>82.1%</td>
<td>84.2%</td>
</tr>
<tr>
<td>non-spam</td>
<td>97.5%</td>
<td>97.1%</td>
</tr>
</tbody>
</table>
Bagging & Boosting

<table>
<thead>
<tr>
<th>class</th>
<th>recall</th>
<th>precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>spam</td>
<td>84.4%</td>
<td>91.2%</td>
</tr>
<tr>
<td>non-spam</td>
<td>98.7%</td>
<td>97.5%</td>
</tr>
</tbody>
</table>

After bagging

<table>
<thead>
<tr>
<th>class</th>
<th>recall</th>
<th>precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>spam</td>
<td>86.2%</td>
<td>91.1%</td>
</tr>
<tr>
<td>non-spam</td>
<td>98.7%</td>
<td>97.8%</td>
</tr>
</tbody>
</table>

After boosting
Related Work

- **Link spam**
  - B. Wu and B. Davison. *Identifying Link Farm Spam Pages*. [WWW 2005]

- **Content spam**

- **Cloaking**

- **e-mail spam**
Conclusion

- Studied properties of the spam pages (per top-level domain and language)
- Implemented and evaluated a variety of spam detection techniques
- Combined the techniques into a decision-tree classifier
- We can identify ~86% of all spam with ~91% accuracy
Thank you