A new algorithm for detecting Web spam using page features and hyperlinks

Jacob Abernethy, UC Berkeley
(Thanks to Yahoo! Research for two internships and a fellowship!)

Joint work with Olivier Chapelle and Carlos Castillo (Chato) from Yahoo! Research
How to Be a Spammer
Learning to Find Spam

- Not a typical learning problem:
  - Web page contents are probably generated adversarially, with the intention of fooling the indexer
  - Given a hyperlink graph, BUT it’s not clear what purpose each link serves: may be natural, may be used for spam, or may simply be there to confuse the indexer
Which of the Blue Hosts are Bad?
One Key Fact

• An extremely useful observation for spam detection:

Good hosts almost NEVER link to spam hosts!!
Good does NOT link to Bad!
Graph-based methods try to compute the "spamicity" of a given page using only the hyperlink graph.

Perhaps most well-known is TrustRank, based on the PageRank algorithm.
Content-Based Methods

- Train a classifier based on page features:
  1. # words in page
  2. Fraction of visible words
  3. Fraction of anchor text
  4. Average word length
  5. Compression rate
WITCH

Web spam Identification Through Content and Hyperlinks
Key Ingredients

- Support Vector Machine (SVM) type framework
- Additional slack variable per node
- “Semi-directed” graph regularization
- Efficient Newton-like optimization
• Standard **SVM**: fit your data, but make sure your classifier isn’t too complicated (aka has a large margin)

\[
\Omega(w) = \frac{1}{l} \sum_{i=1}^{l} [1 - y_i w \cdot x_i]^2 + \lambda w \cdot w
\]
Graph Regularized SVM: fit your data, control complexity, AND make sure your classifier “predicts smoothly along the graph”

\[
\Omega(w) = \frac{1}{l} \sum_{i=1}^{l} \left[ 1 - y_i w \cdot x_i \right]_+ + \lambda w \cdot w + \gamma \sum_{(i,j) \in E} a_{ij} (w \cdot x_i - w \cdot x_j)^2
\]
Graph Regularized SVM with Slack: Same as before, but also learn a spam weight for each node.
Better Graph Regularization:

- When A links to B, penalizing the spam score as $(S_A - S_B)^2$ isn’t quite right. This hurts sites that receive links from spam sites.

Intuitively, this should be better.

Undirected Regularization

$(S_A - S_B)^2$

Directed Regularization

$max(0, S_A - S_B)^2$
NOT TRUE!!

• Interestingly, the issue is more complex

A *mixture* of the two types of regularization is better!
Optimal Regularizer

Semi-Directed Regularization
Seems Strange, BUT…

- Why didn’t simple directed regularization work?
- It will **fail** on certain cases:

  ![Diagram]

  - All in links come from bad guys
  - All out links go to good guys
Optimization

• Roughly a Newton-method type optimization.
• Hard part is computing the Newton Step
• Can be accomplished using linear conjugate gradient, ~50 passes over data to get one approximate Hessian.
• Requires roughly 10 Newton steps
WITCH Performance Results
Performance Comparison

![Graph showing Performance Comparison]
Web Spam Challenge

- Organized By Researchers at Yahoo! Research Barcelona and University Paris 6
- Used a web spam dataset consisting of 10,000 hosts including:
  - 1,000 labelled hosts, roughly 10% spam
  - A Hyperlink graph
  - Content-based features
Web Spam Challenge

• We won the 2\textsuperscript{nd} Track of the Web spam Challenge 2007 (measured by AUC, host-level only)
• Our algorithm outperforms the winner of the Track I competition (we were too late to compete).
## Performance Results

<table>
<thead>
<tr>
<th>Training Algorithm</th>
<th>AUC 10%</th>
<th>AUC 100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM + stacked g.l.</td>
<td>0.919</td>
<td>0.953</td>
</tr>
<tr>
<td>Link based (no features)</td>
<td>0.906</td>
<td>0.948</td>
</tr>
<tr>
<td>Challenge winner</td>
<td>–</td>
<td>0.956</td>
</tr>
<tr>
<td>Only Features</td>
<td>0.859</td>
<td>0.917</td>
</tr>
<tr>
<td>Features + GR</td>
<td>0.874</td>
<td>0.917</td>
</tr>
<tr>
<td>Slack + GR</td>
<td>0.919</td>
<td>0.954</td>
</tr>
<tr>
<td>WITCH (Feat. + Slack + GR)</td>
<td><strong>0.928</strong></td>
<td><strong>0.963</strong></td>
</tr>
</tbody>
</table>
Final Thoughts
“No Good $\rightarrow$ Bad Links” Assumption?

• Perhaps good sites will link to bad sites occasionally:
  ▪ Blog spam
  ▪ “link swapping”
  ▪ Harpers (thanks to reviewer for pointing this out!)

• How can we deal with this?
Thank You!!

Questions?

(and thanks to Alexandra Meliou for the PowerPoint Animations)