Identifying Web Spam
With User Behavior Analysis

Yiqun Liu, Rongwei Cen, Min Zhang, Shaoping Ma, Liyun Ru
State Key Lab of Intelligent Tech. & Sys.
Tsinghua University
2008/04/23
Introduction – simple math

• How many spam pages are there on the Web?
  – Over 10% (Fetterly et al. 2004, Gyöngyi et al. 2004)
  – Web has 152 billion pages (How Much Info project 2003)

• How many can a search engine index?
  – Tens of billions (Google: 8 billion@2004, Yahoo: 20 billion@2005)

• #(spam) is equal to/more than search engines’ index sizes

• Search index will be filled with useless pages without spam detection.

• We have developed lots of spam detection methods

However …
Introduction

- Search “N95 battery time” with a certain Chinese search engine on 08/04/17
Introduction

- Problem: spam detection has been an ever-lasting process
  - Good news for anti-spam engineers!
  - Bad news for Web users / search engines

- Are detection methods not effective?
  - No! Lots of works report over 90% detection accuracy (Ntoulas et al. 2006, Saito et al. 2007, Lin et al. 2007, ...)

- Are detection methods not timely?
  - Yes! When one kind of spam appears, it takes a long time for anti-spam engineers to realize the appearance.
Introduction

- How does spam make a profit?

For a certain kind of Web spam technique

UV / Profit

Time

T1 T2
Introduction

• Important: find new kind of spam as soon as possible

Detect a new kind of Web spam technique timely

Reduce the spam profit

When profit < cost, spam stops
User-behavior Features

- Users will at first realize the existence of a new spam page
  - How to use the wisdom of crowds to detect spam?
    - Social annotation? (possible noises)
    - Web access log analysis.
  - Web access logs
    - Collected by a commercial search engine
    - July 1st, 2007 to August 26th, 2007
    - 2.74 billion user clicks in 800 million Web pages
User-behavior Features

• The behavior features we propose
  – How many user visits are oriented from search engine?
  – How many users will follow links on the page?
  – How many users will not visit the site in the future?
  – How many user visits are oriented by hot keyword searches?
  – How many pages does a certain user visit in the site?
  – How many users visit the site?
  – …
User-behavior Features

• Search engine oriented visiting rate (SEOV rate)
  – Web spam are designed to get “an unjustifiably favorable relevance or importance score” from search engines. (Gyongyi et. al. 2005)
  – Assumption:
    Most user visits to Web spam are from search engine result lists
  – Definition:
    
    $$SEOV(p) = \frac{\#(\text{Search engine oriented visits of } p)}{\#(\text{Visits of } p)}$$

User-behavior Features

- SEOV rate distribution

Some spam don’t receive many UV from search engines, either.

Most ordinary pages’ user visits are not from search engines.
User-behavior Features

• Source page rate ($SP$ rate)
  – Spam pages are usually designed to show users ads/low-quality information at their first look.
  – Users don’t trust hyperlinks on spam pages
  – Assumption:
    Most Web users will not follow hyperlinks on spam pages
  – Definition:
    $$SP(p) = \frac{\#(p \text{ appears as the source page})}{\#(p \text{ appears in the Web access logs})}$$
User-behavior Features

- **SP rate distribution**

  - Half of spam pages have very small *SP* values
  - User clicks hyperlink on some spam page, too. (users may be cheated by anchor texts)
User-behavior Features

• Short-time Navigation Rate (SN rate)
  – Users cannot be cheated again and again during a small time period
  – Assumption:
    Most Web users will not visit a spam site many times in a same user session
  – Definition:
    $$SN(s) = \frac{\#(Sessions\ in\ which\ users\ visit\ less\ than\ N\ pages\ in\ s)}{\#(Sessions\ in\ which\ users\ visit\ s)}$$
    $N$: parameter
User-behavior Features

- SN rate distribution ($N = 3$)

A number of ordinary pages also receive few UVs in a session. (redirection sites, low-quality sites, …)

Few spam pages are visited over 2 times in a session.
User-behavior Features

• Correlation values between these features
  – Different assumption
  – Different information sources
  – Relatively low correlation
  – Possible to use Bayes learning methods

<table>
<thead>
<tr>
<th></th>
<th>SEOV</th>
<th>SP</th>
<th>SN</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEOV</td>
<td>1.0000</td>
<td>0.1981</td>
<td>0.1780</td>
</tr>
<tr>
<td>SP</td>
<td>0.1981</td>
<td>1.0000</td>
<td>0.0460</td>
</tr>
<tr>
<td>SN</td>
<td>0.1780</td>
<td>0.0460</td>
<td>1.0000</td>
</tr>
</tbody>
</table>
Detection algorithm

• Problem:
  – Uniform sampling of negative examples (pages which are not spam) is difficult

• Solution:
  – Learning from positive examples (Web spam) and unlabelled data (Web corpus)
  – Calculate the possibility of a page \( p \) being Web spam using user behavior features

\[
P(p \in \text{Spam} \mid SEOV(p), SP(p), SN(p))
\]
Detection algorithm

• For a single feature $A$:

$$P(p \in \text{Spam} \mid p \text{ has feature } A) \propto \frac{\#(p \text{ has feature } A \cap p \in \text{Spam sample set})}{\#(\text{Spam sample set})} \Bigg/ \frac{\#(p \text{ has feature } A)}{\#(\text{CORPUS})}$$

• For three features $SEOV$, $SP$ and $SN$:
  
  – Features are approximately independent as well as conditionally independent given the target value

$$P(p \in \text{Spam} \mid p \text{ has feature } A_1, A_2, \ldots, A_n) \propto \prod_{i=1}^{n} \left( \frac{\#(p \text{ has feature } A_i \cap p \in \text{Spam sample set})}{\#(\text{Spam sample set})} \right) \Bigg/ \frac{\#(p \text{ has feature } A_i)}{\#(\text{CORPUS})}$$
Detection algorithm

• Algorithm Description

1. Collect Web access log (with information shown in Table1) and construct access log corpus $S$;
2. Calculate $SEOV$ and $SP$ scores according to Equation (1) and (2) for each Web page in $S$;
3. Calculate $SEOV$ and $SP$ scores for each Web site in $S$ by averaging scores of all pages in the site;
4. Calculate $SN$ score for each Web site in $S$ according to Equation (3);
5. Calculate $P(Spam \mid SEOV, SP, SN)$ according to Equation (9) for each Web page in $S$. 
Experimental Results

• Experiment setup
  – Training set:
    • 802 spam sites
    • Collected from the hottest search queries’ result lists
  – Test set:
    • 1564 Web sites annotated with whether it is spam or not
    • 345 spam, 1060 non-spam, 159 cannot tell
    • Percentage of spam is higher than the estimation given by Fetterly et. al. and Gyöngyi et. al. (we only retain the sites which are visited at least 10 times)
Experimental Results

• How to evaluate the performance
  – Focus: find the recently-appeared spam types (not to detect all possible spam types)
  1: Whether the spam candidates identified by this algorithm are really Web spam. (effectiveness)
  2: Whether this algorithm detect spam types more timely than current search engines. (timeliness)
  3: Which feature is more effective?
Experimental Results

- Detection performance (**effectiveness**)
  - Whether the top-ranked candidates are Web spam
  - 300 Pages with the highest $P(Spam)$ values
    - Only 6% are not Web spam (low-quality page, SEO page)
    - Many spam types can be identified. (**wisdom of crowds**)

<table>
<thead>
<tr>
<th>Page Type</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-spam pages</td>
<td>6.00%</td>
</tr>
<tr>
<td>Web spam pages (Content spamming)</td>
<td>21.67%</td>
</tr>
<tr>
<td>Web spam pages (Link spamming)</td>
<td>23.33%</td>
</tr>
<tr>
<td>Web spam pages (Other spamming)</td>
<td>10.67%</td>
</tr>
<tr>
<td>Pages that cannot be accessed</td>
<td>38.33%</td>
</tr>
</tbody>
</table>
Experimental Results

• Detection performance (timeliness)
  – Experiments with one of the most frequently-used Chinese search engines (use X to represent it)
  – Recent data: Access logs from 08/02/04 to 08/03/02
  – Top-ranked spam candidate sites
    • 723/1000 are spam sites (some failed to be connected)
    • X indexed 34 million pages from these 723 sites in early Mar.
    • 59 million pages were indexed by X at the end of Mar.

These spam are not detected by X, X spent lots of resources on these useless pages
Experimental Results

- Detection performance (algorithm & features)

AUC value of the detection algorithm is about 80%
Experimental Results

- Detection performance (algorithm & features)

Learning algorithm gains better performance than any single feature.
Experimental Results

- Detection performance (algorithm & features)

SN performs the worst:
Examples: Q&A portal, Audio/Video sharing sites.
Conclusions

• The amount of Web spam is perhaps over search engine index size

• Timeliness is as important as effectiveness in spam detection

• User behavior features can be used to find recently-appeared spam types timely and effectively