Link Spam Detection Based on Mass Estimation

Zoltán Gyöngyi, Pavel Berkhin, Hector Garcia-Molina, Jan Pedersen
Roadmap

- Search engine spamming
- Link spamming
- PageRank contribution
- Spam mass
  - Definition
  - Estimation
  - Algorithm
- Experiments
Spamming: Example

#1 search result for the query “austria ski”

<table>
<thead>
<tr>
<th>Austria ski/resorts</th>
<th>Swiss/ski/resorts</th>
<th>Italy/ski/resorts</th>
<th>France/ski/holidays</th>
<th>Last Minute</th>
</tr>
</thead>
<tbody>
<tr>
<td>ski-austria.com</td>
<td>skiswitzerland.com</td>
<td>skitaly.com</td>
<td>ski-france.com</td>
<td>dive-lastminute.com</td>
</tr>
<tr>
<td>stanton-austria.com</td>
<td>zermatt.com</td>
<td>aosta-italy.com</td>
<td>holiday-francais.com</td>
<td>golf-lastminute.com</td>
</tr>
<tr>
<td>austria-anthelberg.com</td>
<td>jungfrau-region.com</td>
<td>courmayeur.com</td>
<td>holidays-italy.com</td>
<td>holidays-lastminute.com</td>
</tr>
<tr>
<td>lech-austria.com</td>
<td>vbiers-switzerland.com</td>
<td>dolomites-italy.com</td>
<td>bigl italia.com</td>
<td>ski-lastminute.com</td>
</tr>
<tr>
<td>stubai-austria.com</td>
<td>zermattswitzerland.com</td>
<td>livigno-italy.com</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tirol-austria.com</td>
<td>holidays-switzerland.com</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>holidays-austria.com</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asia/activities/dest</td>
<td>Asia/activities/dest</td>
<td>Holidays Europe</td>
<td></td>
<td></td>
</tr>
<tr>
<td>travel-thailand.com</td>
<td>asiaandiveholidays.com</td>
<td>holidays-europe.com</td>
<td></td>
<td></td>
</tr>
<tr>
<td>bangkok-thailand.com</td>
<td>asianmp3.com</td>
<td>holidays-si-neurope</td>
<td></td>
<td></td>
</tr>
<tr>
<td>pattaya-thailand.com</td>
<td>mp3-thailand.com</td>
<td>europeareservations.com</td>
<td></td>
<td></td>
</tr>
<tr>
<td>phuket-thailand.com</td>
<td>thailandhealthtimes.com</td>
<td>croatia-coast-holidays.com</td>
<td></td>
<td></td>
</tr>
<tr>
<td>thailandgolfmaps.com</td>
<td>thailandpropertytimes.com</td>
<td>slovenia-icoast.com</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Best Price</td>
<td>Best Price</td>
<td>Best Price</td>
<td>Best Price</td>
<td>Alpine Sun</td>
</tr>
<tr>
<td>bestprice-austria.com</td>
<td>bestprice-thailand.com</td>
<td>bestprice-touring.com</td>
<td>alpineholidays.com</td>
<td>activelifestylewoman.com</td>
</tr>
<tr>
<td>bestprice-austria.com</td>
<td>bestprice-merin.com</td>
<td>bestpriceverbier.com</td>
<td>alpinescents.com</td>
<td>eurorski-on-line.com</td>
</tr>
<tr>
<td>bestprice-italy.com</td>
<td>bestprice-courmayeur.com</td>
<td>bestprice-airline-tickets.com</td>
<td>business-traveltoday.com</td>
<td></td>
</tr>
<tr>
<td>bestprice-switzerland.com</td>
<td>bestprice-skiing.com</td>
<td>bestprice-airline-tickets.com</td>
<td>bookhotelsdirect.com</td>
<td></td>
</tr>
<tr>
<td>bestprice-france.com</td>
<td>bestprice-golfing.com</td>
<td>bestprice-travel-network.com</td>
<td>activelifestyle.com</td>
<td></td>
</tr>
</tbody>
</table>

| Available Accommodation     | Available Accommodation   |        | global apartments                  | global apartments                  |
| availablerooms-thailand.com | availablerooms-switzerland.com |        | alpine-apartmentregister.com        | lakesmountainsapartments.com |
| availablerooms-zermatt.com  | availablerooms-italy.com   |        | apartment-italy.com                  | availablerooms-italy.com          |
| availablerooms-ski.com      | availablerooms-austria.com |        | apartment-austria.com                | availablerooms-switzerland.com    |
| availablerooms-france.com   | availablerooms-france.com  |        |        | availablerooms-france.com          |

Very Large Data Bases  •  Seoul, September 13, 2006
Spamming: Example

#1 search result for the query “austria ski”
Spamming: Example
Spamming: Introduction

Spamming = misleading search engines to obtain higher-than-deserved ranking
Spamming: Introduction

Spamming = misleading search engines to obtain higher-than-deserved ranking

Techniques

Spamming
  Term
  Link

Hiding
  Content Hiding
  Cloaking
  Redirection
Spamming: Introduction

Spamming = misleading search engines to obtain higher-than-deserved ranking

Link spamming = building link structures that boost PageRank score
Spamming: Our Target

Detect pages that achieve high PageRank through link spamming

\[ k >> m \]

\[ s_0 \]

\[ s_1 \]

\[ s_2 \]

\[ s_{k-1} \]

\[ s_k \]

\[ g_1 \]

\[ g_m \]
PageRank Contribution
PageRank Contribution

\[ p_0 \]
PageRank Contribution

\[ p_0 \]
PageRank Contribution
PageRank Contribution
PageRank Contribution

\[ p_0^+ = 2 \, c^2 \, (1 - c) / n + 2 \, c \, (1 - c) / n \]

\[ p_0^- = 6 \, c^2 \, (1 - c) / n + c \, (1 - c) / n \]
Spam Mass: Definition

- **Absolute mass**
  - Amount (part) of PageRank coming from spam
  - \[ a.m. = p_0^- = 5 \]

- **Relative mass**
  - Fraction of PageRank coming from spam
  - \[ r.m. = \frac{p_0^-}{p_0} = \frac{5}{7} \]
  - More useful in practice
Spam Mass: Estimation

Ideally...

\[ p_0 \]
Spam Mass: Estimation

In practice...

- Approximate the set of good nodes by a subset called **good core**

\[ p_{0}^{+} \]
Spam Mass: Estimation

In practice...

- Approximate the set of good nodes by a subset called **good core**

\[ p_0^- = p_0 - p_0^+ \]
Spam Mass: Algorithm

1. Create good core
2. Compute PageRank scores $p_i$ and $p_i^+$
3. Compute estimated relative mass $m_i$ as $(p_i - p_i^+) / p_i$
4. For all pages $i$ with large PageRank
   Mark page as spam if $m_i > \text{threshold}$
Experiments: Data

- Yahoo! web index → host graph
  - 73.3M nodes
  - 979M links
- Good core
  - High-quality web directory: 16,780
  - Governmental hosts: 55,320
  - Educational hosts: 434,000
Experiments: Data

- **Sample**
  - 0.1% of nodes with PageRank > 10x minimum
  - 892 nodes
  - Manually labeled good, spam

- **Relative mass groups (approx. same size)**
  - Group 1: 44 samples with smallest rel. mass
  - ...
  - Group 20: 40 samples with largest rel. mass
Experiments: Relative Mass

- Anomalies
  - *.alibaba.com
  - *.blogger.com.br
  - Polish hosts → only 12.pl in good core
Experiments: Relative Mass

Total number of hosts above threshold

Estimated precision

Relative mass threshold

Anomalous hosts excluded

Anomalous hosts included
Experiments: Core Size

![Graph showing the relationship between estimated precision and relative mass threshold for different core sizes. The graph includes lines for 100% core, 10% core, 1% core, 0.1% core, and core. The y-axis represents estimated precision, ranging from 0 to 0.8, and the x-axis represents the relative mass threshold, ranging from 0.98 to 0.1.]
Related Work

- PageRank analyses
  - [Bianchini+2005], [Langville+2004]

- Link spam analyses
  - [Baeza+2005], [Gyöngyi+2005]

- Link spam detection
  - Statistics: [Fetterly+2004], [Benczúr+2005]
  - Collusion detection: [Zhang+2004], [Wu+2005]

- TrustRank
  - [Gyöngyi+2004], [Wu+2006]
Conclusions

- **Search engine spamming**
  - Manipulation of search engine ranking
  - Focus on link spamming

- **Spam mass**
  - ~ PageRank contribution of spam
  - Useful in link spam detection

- **Strong experimental results**
  - Virtually 100% of top 47K nodes spam
  - 94% of top 105K nodes spam
Link Spamming: Model

- Spam farm
Link Spamming: Model

- Spam farm
  1. Target node

$S_0$
Link Spamming: Model

- Spam farm
  1. Target node
  2. Boosting nodes

Ski Austria travel...

Great cheap ski
Switzerland Italy travel
best rates winter sports
hotels
Link Spamming: Model

- Spam farm
  1. Target node
  2. Boosting nodes
  3. Hijacked links from good nodes

Joe's Blog

Comments
Great pictures! See my Austria ski vacation. (by as7869)
Link Spamming: Model

- Spam farm alliances
**PageRank**

- **Probabilistic model:** $p = c U^\top p + (1 - c) v$
  - $U = U(T, v)$ stochastic transition matrix
  - $|v| = 1$

- **Linear model:** $(I - c T^\top) p = (1 - c) v$
  - No adjustment for nodes without outlinks (transition matrix $T$ has all-zero rows)
  - Advantages
    - For $p = PR(v)$ and $v = v_1 + v_2$, $p = p_1 + p_2$ where $p_1 = PR(v_1)$ and $p_2 = PR(v_2)$
    - Faster to compute
PageRank Contribution

- Walk $W$ from $x$ to $y$: $x = x_0, x_1, \ldots, x_k = y$
  - Weight $\pi(W) = \text{out}(x_0)^{-1} \cdots \text{out}(x_{k-1})^{-1}$

- Contribution of $x$ to $y$ over $W$:
  $c^k \pi(W) (1 - c) / n$

- PageRank contribution $p_y^x$ of $x$ to $y$—over all walks
  - Possibly infinite # of walks if there are cycles
  - $p_y^x = \text{PR}($random jump to $x$ only$)$

- See also [Jeh+2003]