Query-log mining for detecting spam queries

Carlos Castillo\textsuperscript{1}, Claudio Corsi\textsuperscript{2}, Debora Donato\textsuperscript{1}, Paolo Feraggina\textsuperscript{2}, Aristides Gionis\textsuperscript{1}

\textsuperscript{1}Yahoo! Research Labs, Barcelona, Spain
\textsuperscript{2}University of Pisa, Italy
Query logs provide valuable information for queries and for documents
  - implicit tags
  - wisdom of crowds

Human-constructed directories provide high quality classification labels for (a subset) of documents

⇒ Identify spam by combining information contained in query logs and in web directories and usage mining
Query graphs: bipartite graphs between queries and documents

Extract features from query graphs

“Semantic” features obtained by propagating web-directory topic labels on the query graph

Use obtained features to improve accuracy of spam detection

Characterize also queries as spam-attracting
Example query log entry: 
$q_1$: “shoes” => $d_1$: shoes1.com, $d_2$: shoes2.com [clicked], $d_3$: shoes3.com

Click-Graph

View-Graph

Anticlick-Graph
syntactic features

- degree of a node (query or document)
- for document $d$: $\text{topQ}_x(d)$ the set of queries adjacent to $d$ and being among the fraction $x$ of the most frequent queries in the query log
- for document $d$: $\text{topT}_y(d)$ the set of query terms which compose the queries adjacent to $d$ in $G$ and being among the fraction $y$ of the most frequent terms in the query log
intuition: multi-topic attractor has potential of being spam

- topic labels can be obtain from a web directory
- ...but not for all documents
intuition: multi-topic attractor has potential of being spam
- topic labels can be obtain from a web directory
- ...but not for all documents
propagation

Read result at each node as a distribution, and compute its entropy
propagation by weighted average

\[ \text{score}_{v+1}^{i}(c) = \alpha^{i-1} \sum_{(v',v) \in E} \text{score}_{v'}^{i}(c) \times f(v', v) \]

and normalization

- propagation by random walk
  - inspired by topic-sensitive PageRank

- “Semantic features”: entropy of the distribution of topic scores (documents and queries)
query-log: sample of 1.6m queries from Yahoo! query log
web directory: DMOZ, 4.2m documents
labeled spam collection: the WEBSPAM-UK2006 dataset
**statistics on the query graphs**

<table>
<thead>
<tr>
<th></th>
<th>Document-level</th>
<th></th>
<th>Host-level</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$C_d$</td>
<td>$A_d$</td>
<td>$V_d$</td>
<td>$C_h$</td>
</tr>
<tr>
<td>Queries</td>
<td>1.59M</td>
<td>0.75M</td>
<td>2.78M</td>
<td>1.59M</td>
</tr>
<tr>
<td>Docs/hosts</td>
<td>2.75M</td>
<td>1.31M</td>
<td>23.47M</td>
<td>0.83M</td>
</tr>
<tr>
<td>Edges</td>
<td>3.69M</td>
<td>1.67M</td>
<td>40.71M</td>
<td>3.50M</td>
</tr>
<tr>
<td>$C_D(0)$</td>
<td>0.05</td>
<td>0.08</td>
<td>0.03</td>
<td>0.28</td>
</tr>
<tr>
<td>$C_Q(1)$</td>
<td>0.18</td>
<td>0.24</td>
<td>0.39</td>
<td>0.58</td>
</tr>
<tr>
<td>$C_D(2)$</td>
<td>0.22</td>
<td>0.22</td>
<td>0.45</td>
<td>0.70</td>
</tr>
<tr>
<td>$CC_{\text{max}}$</td>
<td>0.32</td>
<td>0.19</td>
<td>0.92</td>
<td>0.80</td>
</tr>
<tr>
<td>$</td>
<td>CC</td>
<td>$</td>
<td>0.21</td>
<td>0.23</td>
</tr>
<tr>
<td>Feature set</td>
<td>Features</td>
<td>TP</td>
<td>FP</td>
<td>F₁</td>
</tr>
<tr>
<td>----------------</td>
<td>----------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>Content (C)</td>
<td>98</td>
<td>75.8%</td>
<td>9.8%</td>
<td>0.692</td>
</tr>
<tr>
<td>Links (L)</td>
<td>139</td>
<td>84.2%</td>
<td>9.5%</td>
<td>0.739</td>
</tr>
<tr>
<td>Usage (U)</td>
<td>61</td>
<td>54.2%</td>
<td>7.4%</td>
<td>0.557</td>
</tr>
<tr>
<td>C ∪ L</td>
<td>237</td>
<td>83.9%</td>
<td>8.6%</td>
<td>0.756</td>
</tr>
<tr>
<td>C ∪ U</td>
<td>159</td>
<td>68.4%</td>
<td>6.6%</td>
<td>0.693</td>
</tr>
<tr>
<td>L ∪ U</td>
<td>200</td>
<td>78.5%</td>
<td>6.5%</td>
<td>0.757</td>
</tr>
<tr>
<td>C ∪ L ∪ U</td>
<td>298</td>
<td>78.9%</td>
<td>6.2%</td>
<td>0.765</td>
</tr>
</tbody>
</table>
define “spamicity of a query”: fraction of spam results shown to the user

Task 1: predict if query spamicity is “< 0.5” or “≥ 0.5”
AUC: 0.798, true positive rate: 73.7%, false positives: 29.0%

Task 1: predict if query spamicity is “= 0.5” or “≥ 0.5”
AUC: 0.838, true positive rate: 74.0%, false positives: 22.1%
Use query-log mining and DMOZ class labels for spam detection
Detect spam that has already “fooled” the search engine
Propagation method can be useful in other tasks, too
Future: extract better features and improve the results
Use query-log mining and DMOZ class labels for spam detection
Detect spam that has already “fooled” the search engine
Propagation method can be useful in other tasks, too
Future: extract better features and improve the results
Thank you!