CASIA at WSC2008

Institute of Automation
Chinese Academy of Sciences

Guanggang Geng              Xiaobo Jin
  guanggang.geng@ia.ac.cn  xbjin@nlpr.ia.ac.cn
Chunheng Wang
  chunheng.wang@ia.ac.cn
Detection Framework

- WebGraph
- Feature Extraction

- Content
- Feature Extraction

- HostGraph
- Feature Extraction

- Feature fusion

- Spam Detection (ERUS)

- Detection Result
Host Level link analysis Features

\[ F_1(h) = Measure(h) \]

\[ F_2(h) = \sum_{v \in \text{Inlink}(h)} Measure(v) \times weight(v, h) \]

\[ F_3(h) = \sum_{v \in \text{Outlink}(h)} Measure(v) \times weight(h, v) \]

\[ F_4(h) = \frac{\sum_{v \in \text{Inlink}(h)} Measure(v) \times weight(v, h)}{\sum_{v \in \text{Inlink}(h)} weight(v, h)} \]

\[ F_5(h) = \frac{\sum_{v \in \text{Outlink}(h)} Measure(v) \times weight(h, v)}{\sum_{v \in \text{Outlink}(h)} weight(h, v)} \]

Measures???

HostRank,

TrustRank,

Truncated PageRank(TP) \( (T=1,2,..K) \)

\( \text{weight}(h,v)=f(n), \) \( n \) is the number of hyperlinks from host \( h \) to host \( v \)

we use boolean weight
Host Level link analysis Features

\[ F_6(h) = \frac{\sum_{v \in \text{Inlink(Outlink}(h)) \text{ Measure}(v)}{|\text{Inlink(Outlink}(h))|} \]

\[ F_7(h) = \frac{\sum_{v \in \text{Inlink(Outlink}(h)) \text{ Measure}(v)}{|\text{Inlink}(\text{Outlink}(h))|} \]

\[ F_8(h) = \frac{\sum_{v \in \text{Outlink(Outlink}(h)) \text{ Measure}(v)}{|\text{Outlink}(\text{Inlink}(h))|} \]

\[ F_9(h) = \frac{\sum_{v \in \text{Outlink(Outlink}(h)) \text{ Measure}(v)}{|\text{Outlink}(\text{outlink}(h))|} \]

\[ F_{10}(h) = \text{SiteSupporter}_d(h) \quad d \in \{1, 2, \cdots, k\} \]

We extract 9 *4 (HostRank, trustRank, TP (T=1,2)) + 4(d=1,2,3,4)=40 host level link features
Performance with different features on WEBSPAM-UK2006(Set1+Set2)(5-CV)

ROC with Bagging(C4.5)  ROC with Adaboost(stump)
Performance with different features on WEBSPAM-UK2006 (Set1+Set2) (5-CV)

<table>
<thead>
<tr>
<th>Features</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-measure</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content(C)</td>
<td>0.807</td>
<td>0.712</td>
<td>0.756</td>
<td>0.915</td>
</tr>
<tr>
<td>WebGraph(W)</td>
<td>0.771</td>
<td>0.781</td>
<td>0.776</td>
<td>0.931</td>
</tr>
<tr>
<td>HostGraph(H)</td>
<td>0.775</td>
<td>0.857</td>
<td>0.814</td>
<td>0.941</td>
</tr>
<tr>
<td>C+W</td>
<td>0.839</td>
<td>0.828</td>
<td>0.833</td>
<td>0.959</td>
</tr>
<tr>
<td>C+W+H</td>
<td>0.852</td>
<td>0.873</td>
<td>0.862</td>
<td>0.969</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Features</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-measure</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content(C)</td>
<td>0.839</td>
<td>0.740</td>
<td>0.786</td>
<td>0.931</td>
</tr>
<tr>
<td>WebGraph(W)</td>
<td>0.793</td>
<td>0.816</td>
<td>0.804</td>
<td>0.942</td>
</tr>
<tr>
<td>HostGraph(H)</td>
<td>0.805</td>
<td>0.860</td>
<td>0.831</td>
<td>0.949</td>
</tr>
<tr>
<td>C+W</td>
<td>0.845</td>
<td>0.832</td>
<td>0.838</td>
<td>0.960</td>
</tr>
<tr>
<td>C+W+H</td>
<td>0.855</td>
<td>0.887</td>
<td>0.871</td>
<td>0.971</td>
</tr>
</tbody>
</table>

Bagging(C4.5)  Adaboost(stump)
Detection Strategy---Ensemble Random Under-Sampling (ERUS)

How imbalance???

Non-spam:Spam
WEBSPAM-UK2006-Set1  7:1
WEBSPAM-UK2007-Set1  18:1
Detection Strategy---Ensemble Random Under-Sampling (ERUS)

Integration is based on the predicted spamicity of all the resampled training sets. In our experiment, the resample times $T=15$. 
ERUS Performance---F1-Measure (5-CV)

WEBSPAM-UK2006-Set1

WEBSPAM-UK2007-Set1

2008-5-9

WSC2008
ERUS Performance---ROC (5-CV)

\[ K = \frac{|\text{non-spam}|}{|\text{spam}|}, \text{ (after under-sampling)} \]

WEBSPAM-UK2006-Set1    WEBSPAM-UK2007-Set1

2008-5-9                WSC2008
## ERUS Performance (5-CV)

<table>
<thead>
<tr>
<th>Measures</th>
<th>AdaBoost</th>
<th>ERUS_Ada</th>
<th>Bagging</th>
<th>ERUS_Bag</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1-measure</td>
<td>0.762</td>
<td>0.781</td>
<td>0.763</td>
<td>0.778</td>
</tr>
<tr>
<td>AUC</td>
<td>0.967</td>
<td>0.972</td>
<td>0.968</td>
<td>0.972</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Measures</th>
<th>AdaBoost</th>
<th>ERUS_Ada</th>
<th>Bagging</th>
<th>ERUS_Bag</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1-measure</td>
<td>0.307</td>
<td>0.394</td>
<td>0.312</td>
<td>0.424</td>
</tr>
<tr>
<td>AUC</td>
<td>0.841</td>
<td>0.855</td>
<td>0.829</td>
<td>0.851</td>
</tr>
</tbody>
</table>
Thank you!