CASIA at Web Spam Challenge 2008 Track III

Guang-Gang Geng
Institute of Automation
Chinese Academy of Sciences
Beijing 100080, P. R. China
guanggang.geng@ia.ac.cn

Xiao-Bo Jin
Institute of Automation
Chinese Academy of Sciences
Beijing 100080, P. R. China
xbjin@nlpr.ia.ac.cn

Chun-Heng Wang
Institute of Automation
Chinese Academy of Sciences
Beijing 100080, P. R. China
chunheng.wang@ia.ac.cn

1. DETECTION STRATEGY

Figure. 1 is the flow chart of our web spam detection strategy. The detection is based on the content analysis features, WebGraph related features and HostGraph related features. Next,we will introduce the feature extraction.

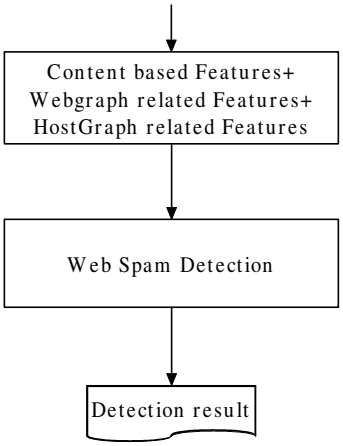

Figure 1: Flow chart of the spam detection strategy.

1.1 Content Analysis Features and WebGraph Related Features

The content-based features and WebGraph related features we used is provided by the Challenge 2008. We don't use some of the features, such as all the combined features, since combined features are redundant in feature selection sense.

1.2 HostGraph Related Features Extraction

The HostGraph G is defined as $G = (V, E, weight)$, where V is the set of hosts, $weight = f(n)$ is a weighting function, n is the number of links between any page in host u and any page in host v, and E is the set of edges with non-zero weight. Based on the facts of topological dependencies of spam and normal nodes, the following HostGraph related features are extracted.

$$F_1(H) = Measure(H) \tag{1}$$

$$F_2(H) = \frac{\sum_{h \in Inlink(H)} Measure(h) * weight(h, H)}{\sum_{g:g \in Outlink(h)} weight(h, g)} \tag{2}$$

$$F_3(H) = \frac{\sum_{h \in Outlink(H)} Measure(h) * weight(H, h)}{\sum_{g:g \in Inlink(h)} weight(g, h)} \tag{3}$$

$$F_4(H) = \frac{\sum_{h \in Inlink(Inlink(H)} Measure(h)}{|Inlink(Inlink(H))|} \tag{4}$$

$$F_5(H) = \frac{\sum_{h \in Inlink(Outlink(H))} Measure(h)}{|Inlink(Outlink(H))|} \tag{5}$$

$$F_6(H) = \frac{\sum_{h \in Outlink(Intlink(H))} Measure(h)}{|Outlink(Inlink(H))|} \tag{6}$$

$$F_7(H) = \frac{\sum_{h \in Outlink(Outlink(H))} Measure(h)}{|Outlink(outlink(H))|} \tag{7}$$

$$F_8(H) = SiteSupporter_d(H) \quad d \in \{1, 2, \cdots, k\} \tag{8}$$

where $Measure \in$ {PageRank, Trustrank, Degree-related measures,Truncated PageRank}, $\{h, H, g\} \subseteq V$, d represent the distance of hyperlinks (For instance, with respect to the inlink, $d = 1$). $Inlink(H)$ represents the inlink set of H, and $Outlink(H)$ is the outlink set of H correspondingly. $weight(h, g)$ is the weight of host h and g, $weight(h, H) \in \{1, n, log(n)\}$, where n is the number of hyperlinks between h and H. The results we submitted are based on the boolean weight.

1.3 Detection Algorithm

The detection algorithms we used in the experiment are bagging with ERUS strategy, which have been proven to be effective for spam detection[1][2]. The weak classifier for bagging is C4.5. ERUS is a detection strategy for class-imbalance learning.

2. REFERENCES

[1] G.G. Geng, C.H. Wang, X.B. Jin, Q.D. Li, and L. Xu. IACAS at Web Spam Challenge 2007 Track I, *Web Spam Challenge* 2007.

[2] L.Breiman.Bagging Predictors. *Machine Learning*. 24(2),1996.

AIRWeb'08, April 22, 2008, Beijing, China.
.