We will introduce the feature extraction. WebGraph related features and HostGraph related features.

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.

![Flow chart of the spam detection strategy](image)

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.

$$F_1(H) = Measure(H)$$

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
