Made cloaking algorithm adapt to various languages

Avoid: offline machine learning classifiers

Avoid running client side scripts (unbounded latency to data collection process)

Crawl three times: disguised as normal users, disguises as google bots.

1. Compute the similarity between the content

Return the empty site to google

1. Why decided this threshold 0.33
2. Use DOMS tree to detect the layout
3. Compute the content similarity and layout separately, need to explore the relationship between content and layout.

# Survey on Web Spam Detection: Principles and Algorithms

We categorize all existing algorithms into

three categories based on the type of information they use:

content-based methods, link-based methods, and methods

based on non-traditional data such as user behaviour, clicks,

HTTP sessions.

In turn, we perform a subcategorization of

link-based category into five groups based on ideas and principles

used: labels propagation, link pruning and reweighting,

labels refinement, graph regularization, and featurebased.

To distinguish users from crawlers spammers

analyze a user-agent field of HTTP request and keep track

of IP addresses used by search engine crawlers. The other

strategy is to redirect users to malicious pages by executing

JavaScript activated by page onLoad() event or timer. It

is worth mentioning that JavaScript redirection spam is the

most widespread and difficult to detect by crawlers, since

mostly crawlers are script-agnostic [29].