Search engine optimization: Black hat Cloaking Detection technique

Patel Trupti[1], Kachhadiya Kajal[2], Panchani Asha[3], Mistry Pooja [4]

Shrimad Rajchandra Institute of Management and Computer Application
Uka Tarsadia University, Bardoli, Surat.

Abstract

The Search Engine is not only a tool for searching the Web, but also an advertising platform for ones business and services of companies. Cloaking is Search Engine spamming technique that is used in website to deliver one page to Search Engine Optimization for indexing and serving an entirely different page to users browsing the site. In this paper we show different types of cloaking like IP cloaking , User Agent cloaking and show the comparison of different types of existing cloaking checker tools. We analyzed the output of available tools, but no any existing tools provide facility like how much cloaking done in website in terms of percentage. So we consider some parameter that are use to implement cloaking to identify the existence or non- existence of cloaking.

1. Introduction

Internet Marketing is an activity that provides facilities for promoting your products and services online [13]. It plays important role in getting huge amount of traffic through Search Engines.

Search Engine Optimization (SEO) is the process of improving the quality and volume of web traffic to a website which also help to achieve a higher ranking to a website with major Search Engines when certain keywords are put in the search field [14]. In Internet marketing strategy SEO include how Analysis of Search Engine works, what people search, and the particular terms of keywords typed into Search Engines.SEO has one practice that is Black Hat SEO is referred as SEO spamming that are used to get higher page ranking in an unethical manner. It violates Search Engine (Search Engine Optimization) rules and regulation and also unethically presents content in different visual or non-visual way to Search Engine spiders and Search Engine user. Spider is crawl the website or a program which automatically fetch the web pages [17]. It is also known as web crawler. Search engines gather data about a website by sending' the spider or bots to read' the site and copy its content. This content is stored in the search engine's database [16].

Black Hat SEO used a techniques like doorway pages, hidden text,cloaking,keyword stuffing,etc by using these techniques it increase SEO poisoning. Black Hat SEO is mostly used by those who are looking for a quick ROI(Return On Investment).

SEO poisoning is one type of attacks method in which cyber criminals create malicious website for taking advantages of users [15]. SEO poisoning attack is also known as Black Hat SEO attack when hackers manipulate Search Engine results to make there links appear higher than legitimate results. As user search for related terms, the related links appear near the top of the search results; make a greater number of clicks to malicious websites. In this paper we focus on one Black Hat SEO technique that is cloaking and we try to improve detection of Cloaking from website.

In some situation SEO practitioners give high rank to the website in unethical way by using Cloaking. Cloaking is a Black hat SEO technique in which the “benign” content presented to the Search Engine spider and scam content to user's browser who is referred via a particular search request [1]. This is done by delivering content using IP addresses or the User-Agent header of the user requesting the page. When a website gets a hit, it will call a robot.txt file that tells the site what to display depending on the User Agent identified or the IP address [12]. Robots.txt is a text file you put on your site to tell search robots which pages you would like them not to visit.

There is also ethical way of using cloaking. The use of Cloaking in ethical way is to hide the HTML code of high rank pages from people so that it can't be stolen. It directs the webpage to a place that are not visible on domain and the spiders are analyzing for it, and provides the use of fack pages that are loaded with the keyword and other content that is Search Engine optimized.

1.2 Types of Cloaking

IP Delivery Cloaking: Different content deliver based on IP address. This involves use of IP database which contains list of IP address of all known Search Engine spider. When visitor request a page if the IP address is not present then provide page to the human. If the IP address is match with database then provide to the Search Engine spider page [1].

User-Agent Cloaking: User Agent Cloaking is same as IP cloaking in the sense that the cloaking script compares the User Agent text string which is sent when a page is requested with its list of search engine User Agent and then provide the appropriate page. The difficulty with User Agent cloaking is that Agent names can be easily forge. User Agent cloaking is much more unsafe than IP based cloaking [11][1].

Repeat Cloaking: In Repeat Cloaking the Web site stores state on either the client side (using a cookie) or the server side (e.g., tracking client IPs). This mechanism allows the site to determine whether the visitor has previously visited the site, and to use this knowledge in selecting which version of the page to return [1].

1.3 Objective

Our main objective is to improve detection of the cloaking in website that is used in unethical or ethical way. There are many existing tools to detect the cloaking websites but these tools not provide proper output like cloaking used for ethical or unethical way. So provide the extra facilities rather than existing tools.

2. Related Work

David Y. Wang and Stefan Savage describe current state of search engine cloaking as used to support Web spam, determine new techniques for identifying cloaking (via the search engine snippets that identify keyword-related content found at the time of crawling) and most importantly and also explored the dynamics of cloaked search results and sites over time [1].

Gyongyi and Garcia-Molina [9] describe cloaking and redirection as spam hiding techniques. They noted that web sites can identify search engine crawlers by their network IP address or user-agent names. They additionally point out that some cloaking (such as sending the Search Engine a version free of navigational links and advertisements but no change to the content) is accepted by some engines.

Perkins argues that any agent-based cloaking is spam.No matter what kind of content is sent to Search Engine, the goal is to modify search engines rankings, which is an obvious characteristic of Search Engine spam [4].

Najork was awarded a patent for a method of detecting cloaked pages. He proposed the idea of detecting cloaked pages from users' browsers by installing a toolbar and letting the toolbar send the signature of user perceived pages to search engines. His method may still have difficulty in distinguishing rapidly changing or dynamically generated Web pages from real cloaking pages, and does not directly address semantic cloaking [5].

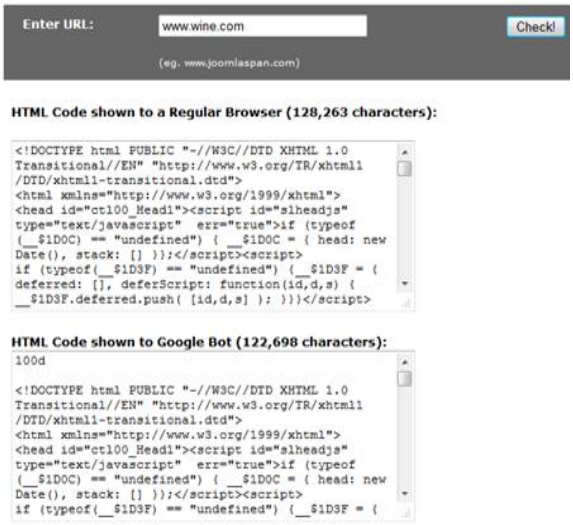
It introduced the idea of automatic detection of cloaking pages using more than two copies of the page. We differentiate semantic cloaking from strictly syntactic cloaking, and explored methods for syntactic cloaking recognition. None of the above papers discuss how many percentage of cloaking occur in website [2].

3.Comparative Study of Cloaking Detection Tool

In direction of the above problem we use the comparative study of different cloaking detection tools which is as below.

1)Joomla Span – URL Cloaking Checker

This tool checks a web page and shows how Search Engines see your site as different what a browser/ human would see [8].



Enter URL: www.wine.com Check!

(eg. www.joomlaspan.com)

HTML Code shown to a Regular Browser (128,263 characters):

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0
Transitional//EN" "http://www.w3.org/TR/xhtml1
/DTD/xhtml1-transitional.dtd">
<html xmlns="http://www.w3.org/1999/xhtml">
<head id="ctl00_Head1"><script id="slheadjs"
type="text/javascript"  err="true">if (typeof
(__$1D0C) == "undefined") { __$1D0C = { head: new
Date(), stack: [] }};</script><script>
if (typeof(__$1D3F) == "undefined") {__$1D3F = {
deferred: [], deferScript: function(id,d,s) {
__$1D3F.deferred.push( [id,d,s] ); }}}</script>
```

HTML Code shown to Google Bot (122,698 characters):

```
100d

<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0
Transitional//EN" "http://www.w3.org/TR/xhtml1
/DTD/xhtml1-transitional.dtd">
<html xmlns="http://www.w3.org/1999/xhtml">
<head id="ctl00_Head1"><script id="slheadjs"
type="text/javascript"  err="true">if (typeof
(__$1D0C) == "undefined") { __$1D0C = { head: new
Date(), stack: [] }};</script><script>
if (typeof(__$1D3F) == "undefined") {__$1D3F = {
```

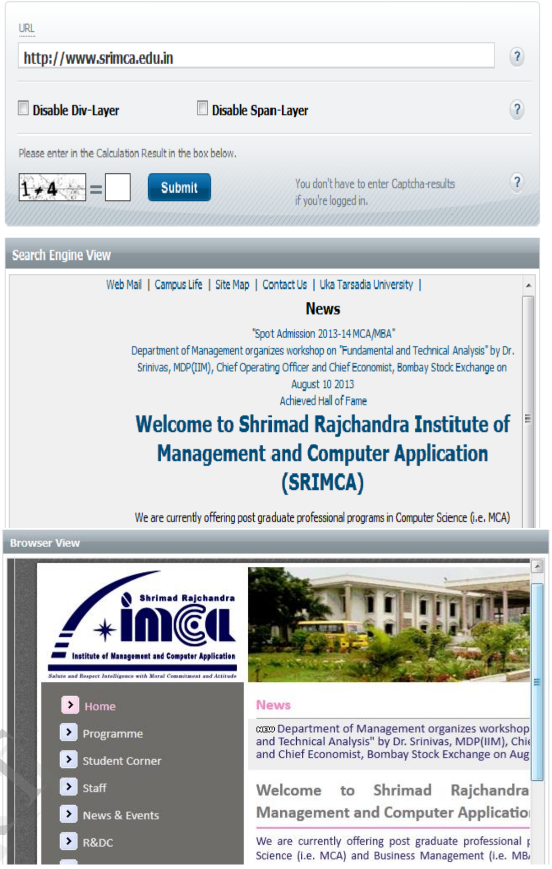
Figure1: Joomla Span: URL Cloaking Checker.

By using above tool we experiment with www.wine.com and checked the output .It displays two different views like Googlebot and Regular browser. And it will display HTML code with total character.

If the total character difference is minor then it indicates that there is no clocking in the website. If the total character difference is major then it indicates that there is a chance of clocking in the website.

2) Link Vendor – Cloaking Detector

This cloaking detector copy the Googlebot based on User-Agent to detect cloaked content. So you can indentify websites on which the content display to the search engine is different from that presented to the Googlebot or user [9].



URL

http://www.srimca.edu.in

□ Disable Div-Layer □ Disable Span-Layer

Please enter in the Calculation Result in the box below.

1+4 = □ Submit

You don't have to enter Captcha-results if you're logged in.

Search Engine View

Web Mail | Campus Life | Site Map | Contact Us | Uka Tarsadia University |

News

"Spot Admission 2013-14 MCA/MBA"

Department of Management organizes workshop on "Fundamental and Technical Analysis" by Dr. Srinivas, MDP(IIM), Chief Operating Officer and Chief Economist, Bombay Stock Exchange on August 10 2013

Achieved Hall of Fame

Welcome to Shrimad Rajchandra Institute of Management and Computer Application (SRIMCA)

We are currently offering post graduate professional programs in Computer Science (i.e. MCA)

Browser View

Shrimad Rajchandra
imca
Institute of Management and Computer Application
Salute and Respect Intelligence with Moral Commitment and Attitude
Home
Programme
Student Corner
Staff
News & Events
R&DC

News

NEW Department of Management organizes workshop and Technical Analysis" by Dr. Srinivas, MDP(IIM), Chie and Chief Economist, Bombay Stock Exchange on Aug

Welcome to Shrimad Rajchandra Management and Computer Applicatio

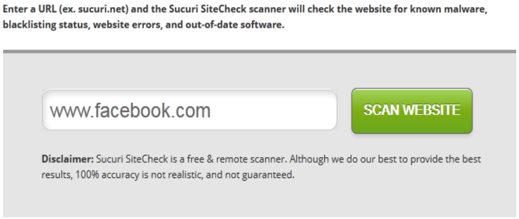
We are currently offering post graduate professional Science (i.e. MCA) and Business Management (i.e. MB

Figure2: Link Vendor – Cloaking Detector.

By using above tool we experiment with www.srimca.edu.in and check the output. It display two different views like Browser view and search engine view. In search engine view it will not display images and in browser view it will display images. There is no difference of any content in both the view.

3)Sucuri Security - Site Check Malware Scanner

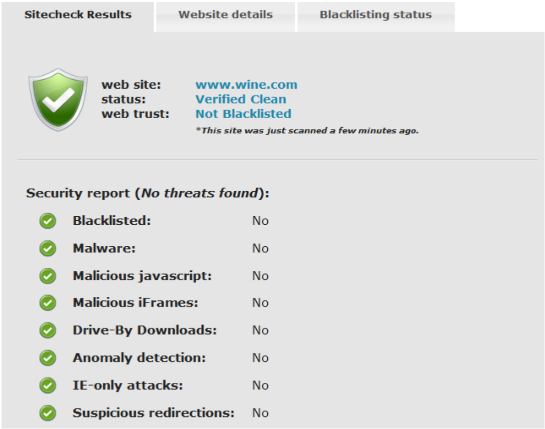
This is one type of word press plugin. It will detects various types of malware,website errors, disabled sites, Obfuscated JavaScript injections, Cross Site Scripting (XSS),Hidden,Malicious, iFrames, MaliciousRedirects, Backdoors,Anomolies, IP Cloaking, IE-only attacks[10].



Enter a URL (ex. sucuri.net) and the Sucuri SiteCheck scanner will check the website for known malware, blacklisting status, website errors, and out-of-date software.

www.facebook.com SCAN WEBSITE

Disclaimer: Sucuri SiteCheck is a free & remote scanner. Although we do our best to provide the best results, 100% accuracy is not realistic, and not guaranteed.



web site: www.wine.com
status: Verified Clean
web trust: Not Blacklisted
*This site was just scanned a few minutes ago.

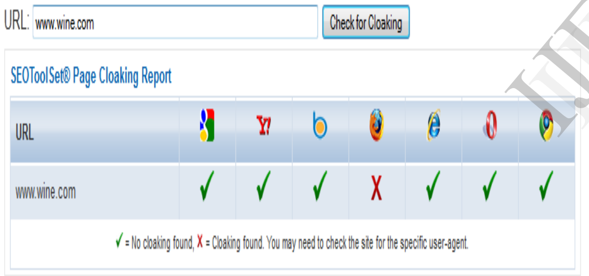
Security report (*No threats found*):

Blacklisted:	No
Malware:	No
Malicious javascript:	No
Malicious iFrames:	No
Drive-By Downloads:	No
Anomaly detection:	No
IE-only attacks:	No
Suspicious redirections:	No

Figure3: Sucuri Security - Site Check Malware Scanner

4) SEO Cloaking Checker

This tool showing one version of a Web page to users or Googlebot and a different version of Web page usually stuffed with other keywords, to the Search Engine .This is a secret method that Google and the other Search Engines consider misleading, since it attempts to prejudice the spiders into ranking the Web page not deserve higher or for a different keyword term [7].



SEOToolSet® Page Cloaking Report

URL							
www.wine.com	✓	✓	✓	X	✓	✓	✓

✓ = No cloaking found, X = Cloaking found. You may need to check the site for the specific user-agent.

Figure4: SEO Cloaking Checker

By using above tool we experiment with www.wine.com in and check the output. it will check cloaking for all browser like Mozilla fire fox,bing,internet explorer,etc.it also check a site for specific user agent. We get cloaking in **www.wine.com** for Mozilla.

4. Conclusion

Cloaking has become a standard tool and it Adds significant complexity for differentiating legitimate Web content from fraudulent pages.Our work has examined the output of cloaking detection tool and also done the comparative study of that tools but there are some lacking in output like no tools provide the information about how much cloaking is implement on site in terms of percentage and also which content are cloaked in website. So we try to implement this lacking think in our future work by using some parameter like <no script>, <display: none> and <iframe>.By which we can indentify there is a chances of presence of cloaking in the website.

5. Future Work

We also indentify that no one of existing tools define that cloaking is use for ethical purpose or unethical purpose. So we try to implement this lacking think in future work. We try to create our own tool for it.

6. References

[1] David Y. Wang, Stefan Savage, and Geoffrey M. Voelker. "Cloak and Dagger: Dynamics of Web Search Cloaking". Deptartment of Computer Science and Engineering, University of California, San Diego, 2011.

[2] Wu and B. D. Davison. "Cloaking and redirection: A preliminary study". In Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), May 2005

[3] Z. Gyongyi and H. Garcia-Molina."Web spam taxonomy". In First International Workshop on Adversarial Information Retrieval on the Web(AIRWeb), Chiba, Japan, 2005

[4] A.Perkins.White paper:"The classification of search engine spam", Sept. 2001.

[5] M. Najork. System and method for identifying cloaked web servers, June 21 2005.

[6] http://en.wikipedia.org/wiki/Cloaking

[7] http://www.akamarketing.com/search-engine-cloaking-and-stealth-technology.html.

[8] http://www.joomlaspan.com/webtools/url-cloaking -checker.php.

[9] http://www.linkvendor.com/seo-tools/cloaking detector.html.

[10] http://sitecheck.sucuri.net/scanner/

[11] http://www.seotools.com/seo-cloaking-checker/

[12] http://www.projectblackhat.com

[13]http://www.webopedia.com/TERM/I/
internet marketing.html.

[14] http://en.wikipedia.org/wiki/Search_engine_optimization

[15] http://whatis.techtarget.com/definition/search-poisoning.

[16] http://www.webdesign.org/sitemaintenance/
se-optimization/role-of-spider-in-seo.14400.html

[17] http://forums.hostsearch.com/showthread.php?271
15-Define-Spider-in-seo