Web search cloaking

Juha Härkönen, Oula Heikkilä

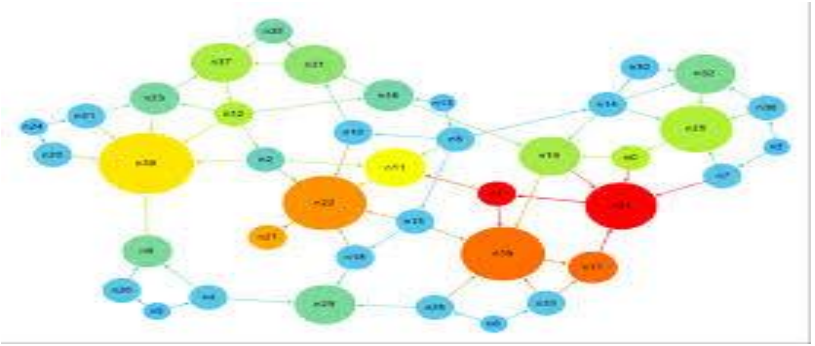
Search engines and optimizations

Search engines are based on indices list they collect publicly available web-pages.

For example, Google goes through the pages for search crawlers (Googlebot) all the time, and reads the content and links pages.

Each search engine has its own algorithms, with an index reading list based on your search.

Optimization techniques to influence the index page to list.



Referrence: http://computationalculture.net/article/what_is_in_pagerank

PageRank was developed by Google's search algorithm, and it assigns a numerical weighting to each element of a hyperlinked set of documents.

According to Google:PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites.

Cloak

Cloaking is a search engine optimization technique in which

- the content presented to the search engine spider is manipulated
- the content presented to the user's browser is different as the search engine spider

The purpose of cloaking is sometimes to deceive search engines so they display the page when it would not otherwise be displayed (black hat SEO).

It is also a functional technique for informing search engines of content they would not otherwise be able to locate (non-textual, video)

Cloaking is the simplest done by delivering content based on the IP addresses or the user-agent HTTP header of the user requesting the page.

Cloaking methods

Cloaking methods

- IP cloaking: IP cloaking is the process of a web server delivering a specific web page based on the visitors IP address. For example, search engines can be identified by the IP -addres that.
- User-agent cloaking: User Agent Cloaking is similar to IP cloaking in the sense that the cloaking script compares the User Agent text string which is sent when a page is requested.
- Repeat cloaking: Repeat cloaking returns the scamsite first time, then benign content afterwards (using cookies or tracking client IP)
- Referrer cloaking: Link can be controlled to bring the different sides of the header field indicates. Control can be done with php, for example Visit jyu.fi. When the user clicks on the link, the “goto-jyu.php” file uses a php redirect to send the user wherever php file redirects. Example **<?php** header(‘Location http://users.jyu.fi/~timoh/TIES327/security.html’) ;**?>**
- JavaScript Redirection Cloaking: As search engine crawlers could not emulate a real browser which could execute JavaScript code in web page, some web sites embed JavaScript redirection code in the page to redirect users to another web site.

Prevalence of different cloaking methods

Reference: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6724370

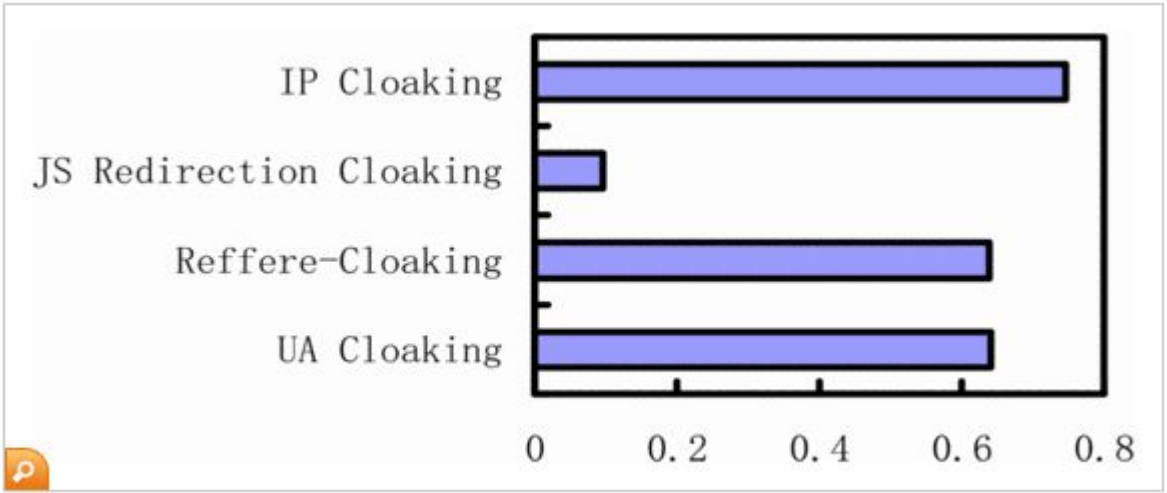


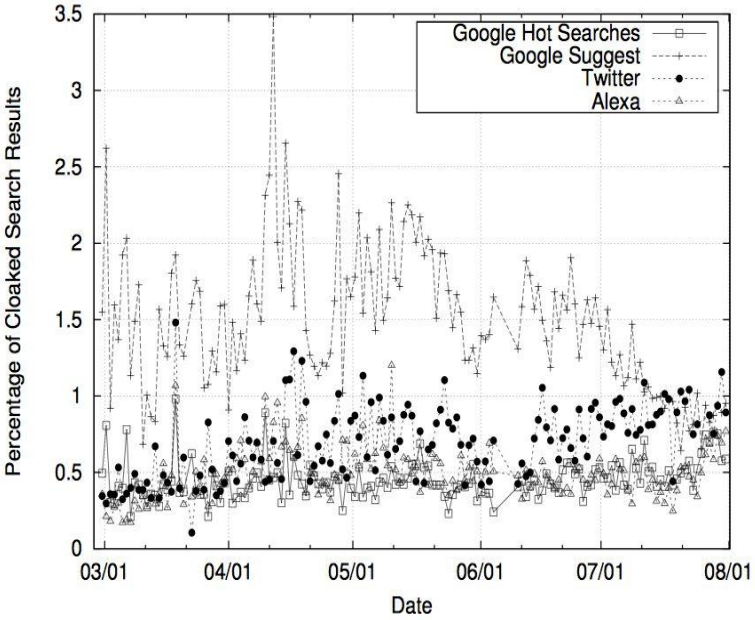
Figure 6. Distribution of combined cloaking techniques.

Dagger

University of California in San Diego have been studied cloaking frequency and developed Dagger -method which takes into account all the cloaking style.

Search engines tend to filter out the results of scam sites out.

The study keywords are generated by Google, Yahoo and Bing hot searches and the most popular search terms from Twitter and Alexa. For second set of search terms, have been use a set of terms catering to a specific domain pharmaceuticals.



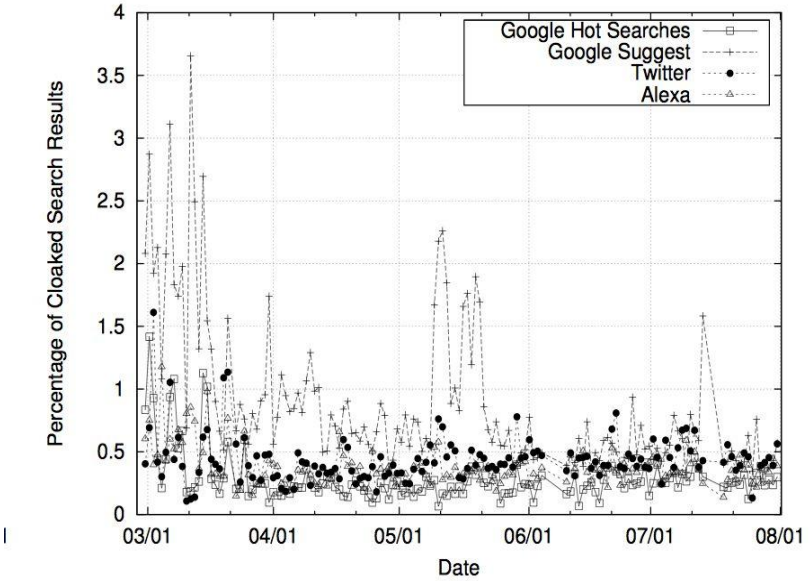
Google search and cloaked pages

Dagger

The study found that 0,5-3,5% of the search engines most popular searches lead to a page which cloaking was used.

Medical keywords resulted in more often cloaked page. In some aprodisiac robe following words up to 80% of the page received was cloaked.

No significant differences were observed between the results of search engines, how many of the proposed page was cloaked.



Yahoo search and cloaked pages

Detection methods

1. tag based
2. text-based
3. url-based

all compare how similar contents of different pages are

4. crawling the same page multiple times and comparing differences
5. loading the page multiple times and calculating the differences between copies of the same page

Uncovering Cloaking Web Pages with Hybrid Detection Approaches

College of Information Science and Engineering Hunan University, Changsha, China

Presents a new system for detecting cloaked web pages. It combines text, tag and URL based methods. The resulting system is then compared to existing systems.

Uncovering Cloaking Web Pages with Hybrid Detection Approaches - Components

1) Data crawling Component
   a) Collect Hot Search Terms
   b) Crawl URLs
   c) Crawl HTML
2) Detection algorithms
   a) Text Based Algorithm
   b) Tag Based Algorithm
   c) URL based Algorithm

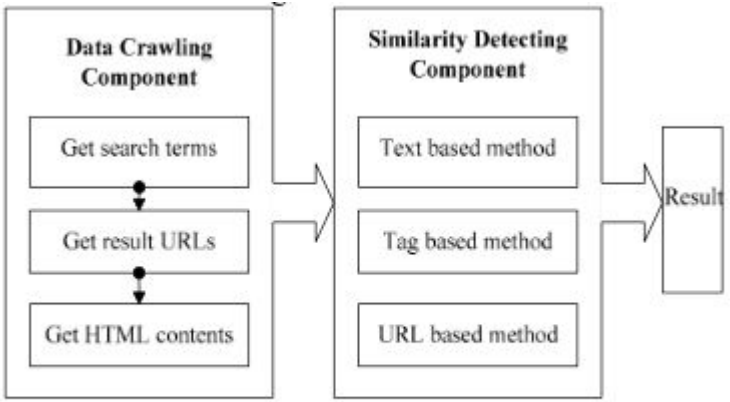


Figure 1. System Architeture.

Uncovering Cloaking Web Pages with Hybrid Detection Approaches - Results

Precision: percentage of URLs marked as cloaked among all the URLs that be detected as cloaked

Recall: the percentage of URLs that were detected as cloaked among all the URLs marked as cloaked

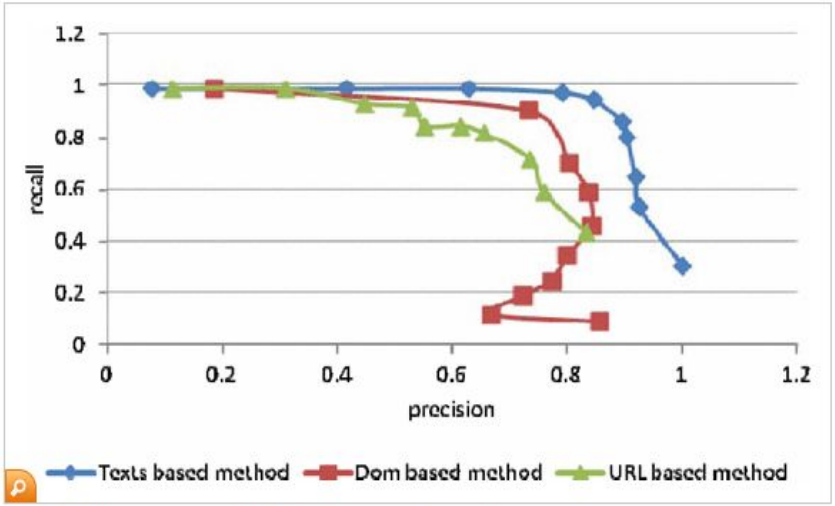


Figure 3. Precision and recall of different detecting methods

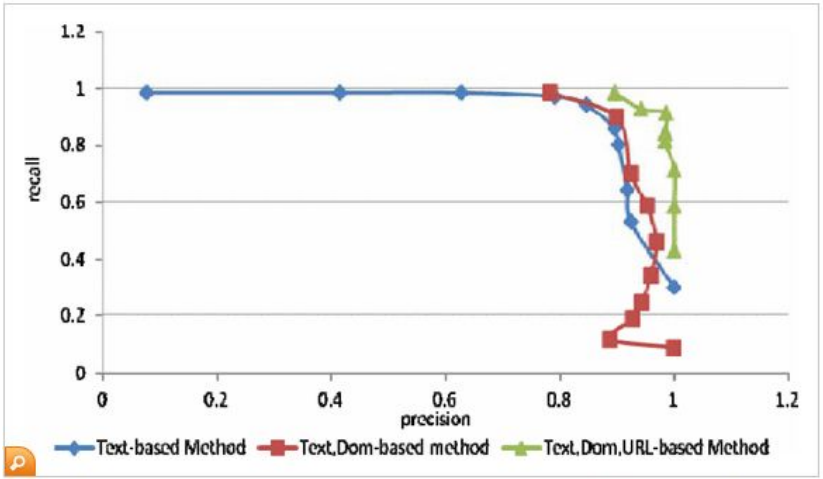


Figure 4. Precision and recall in Text-based method, Text-Dom-based method, and Text-Dom-URL-based mthod.

References

1. Cloak and Dagger: Dynamics of Web Search Cloaking, David Y. Wang, Stefan Savage, and Geoffrey M. Voelker; Department of Computer Science and Engineering University of California, San Diego; http://cseweb.ucsd.edu/~dywang/pubs/ccs298-wang.pdf
2. Uncovering Cloaking Web Pages with Hybrid Detection Approaches, Issue Date: 24-26 Aug. 2013, Written by: Jun Deng; Hao Chen; Jianhua Sun; College of Information Science and Engineering Hunan University, Changsha, China; http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6724370

Thanks!