International Journal of Pure and Applied Mathematics
Volume 117 No. 20 2017, 993-997
ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue

Various Techniques

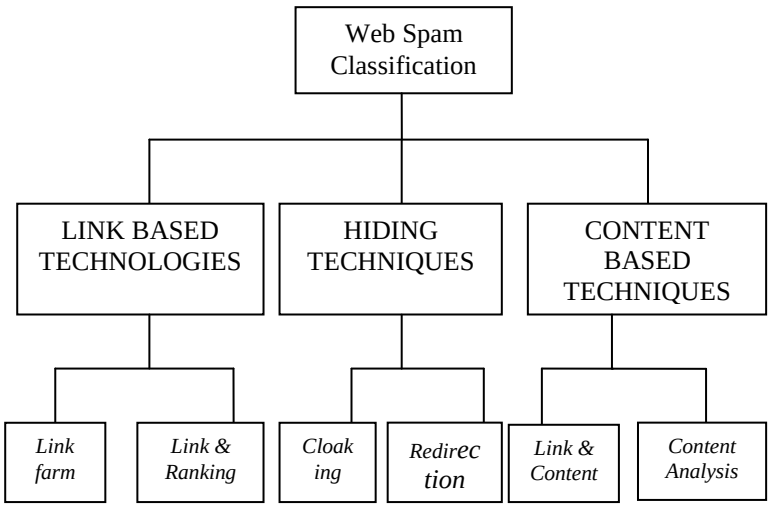
Moolchand Sharma[1], Dr. Ashish Khanna[2],Prerna Sharma[3]
[1]Assistant Professor, [2]Assistant Professor, [3]Assistant Professor
[1,2,3]Department of Computer Science,
MAIT, Delhi
[1]sharma.cs06@gmail.com, [2]ashishk746@yahoo.com, [3]prernasharma@mait.ac.in

Abstract— Current web search engines are built to serve all users independent of the special needs of any individual user. With the exponential growth of the available information on the World Wide Web, a traditional search engine, even if based on sophisticated document indexing algorithms, has difficulty meeting efficiency and effectiveness performance demanded by users searching for relevant information Web surfers trust search engines. They expect only the most relevant responses will be listed in the top ranking positions. But due to search engine spamming it won't happen. Without taking action, results from search engines will be greatly harmed. We have analyzed various solutions proposed by various authors and explore in detail the effective solutions to some search engine spam techniques, such as link stuffing and Cloaking. We have proposed completely different scheme for automatic elimination of spam through entry level spam check.

Keywords— Cloaking, Spamming,Search Engine Spam

I. INTRODUCTION

Search Engine Optimization (SEO) is a process of improving visibility and quality of a website or a web page to get in top of search results of a search engine. SEO is generally being seen in terms of "White Hat" (ethical) and "Black Hat" (unethical) approach towards search. Search Engine Spam or Web Spam is a behavior that attempts to deceive search engine ranking algorithms. Search spammers (or web spammers) refer to those who use questionable search engine optimization (SEO) techniques to promote their low-quality websites into top search rankings.
Examples of web spam users are flash technology based websites, image gallery websites, sites contain only categories header and product name.



Categories of web Spamming:

1. **Content based spamming** Content based spam changes the textual content of web pages to be ranked higher. Some most content based spam techniques include repeating keywords, unwanted unrelated keywords and adding an entire dictionary at the bottom of a web page.
2. **Link based spamming** Link based spam changes the link structure of web pages to attack search engines using link based ranking algorithms. The most popular link based spamming techniques is link farms, link exchanges, link bombs and adding comment spam in blogs and wiki systems.
3. **Page hiding based spamming** Page-hiding based spam hides the whole or part of a web page to search engines to achieve better ranking for the page. Cloaking and deliberate use of redirect are well known examples of page-hiding based spamming techniques

4. **Cloaking** : Showing different page to crawler, from users

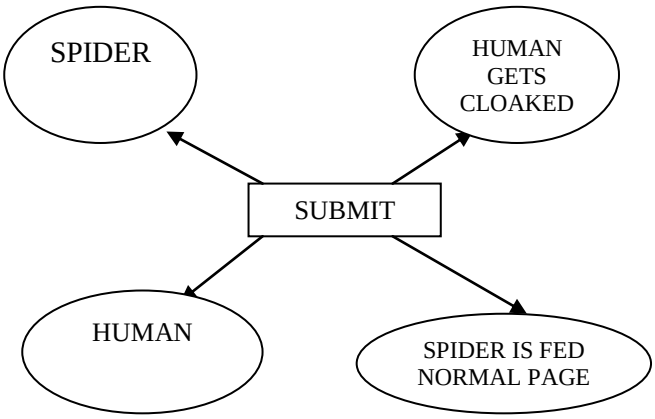


Figure 1.2. Mechanism of Cloaking

II. LITERATURE REVIEW

The term "cloaking", as applied to search engines, has an uncertain history, but dates to at least 1999 when it entered the vernacular of the emerging search engine optimization (SEO) market.1 The growing role of search engines in directing Web traffic created strong incentives to reverse engineer search ranking algorithms and use this knowledge to "optimize" the content of pages being promoted and thus increase
ver, since the
rch rankings

frequently required content vastly different from the page being promoted, this encouraged SEO firms to serve different sets of page content to search engine crawlers than to normal users; hence, cloaking [5].

Types of Cloaking

For cloaking to work, the scammer must be able to distinguish between user segments based on some identifier visible to a Web server. The choice of identifier used is what distinguishes between cloaking techniques, which include Repeat Cloaking, User Agent Cloaking, Referrer Cloaking (sometimes also called "Click-through Cloaking"), and IP Cloaking.

In the case of **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. Thus first-time visitors are given a glimpse of a scam, in the hopes of making a sale, but subsequent visits are presented with a benign page stymieing reporting and crawlers (who routinely revisit pages). In contrast, **User Agent Cloaking** uses the User-Agent field from the HTTP request header to classify HTTP clients as user browsers or search engine crawlers. User agent cloaking can be used for benign content presentation purposes (e.g., to provide unique content to Safari on an iPad vs. Firefox on Windows), but is routinely exploited by scammers to identify crawlers via the wellknown User-Agent strings they advertise (e.g., Googlebot). Referrer Cloaking takes the idea of examining HTTP headers even further by using the Referer field to determine which URL visitors clicked through to reach their site. Thus, scammers commonly only deliver a scam page to users that visit their site by first clicking through the search engine that has been targeted (e.g., by verifying that the Referer field is http://www.google.com). This technique has also been used, in combination with repeat cloaking and chains of Web site redirections, to create one-time-use URLs advertised in e-mail spam (to stymie security researchers). However, we restrict our focus to search engine cloaking in this page [1].

We may think of email spam as a scourge—jamming our collective inboxes with tens of billions of unwanted messages each day—but to its perpetrators it is a potent marketing channel that taps latent demand for a variety of products and services. While most attention focuses on the problem of spam delivery, the email vector itself comprises only the visible portion of a large, multi-faceted business enterprise. Each visit on a spam link points to the fact that the start of a long and complex links, spanning a range of both technical and business components that together provide the necessary infrastructure needed to monetize a customer's visit. Botnet services must be secured, domains registered, name servers provisioned, and hosting or proxy services acquired. All of these, in addition to payment processing, merchant bank accounts, customer service, and fulfillment, reflect necessary elements in the spam value chain [4].

III. Problem Description

- In existing cloak detection technique separate classifier and filter is required, they are two step process.
- There is no independent approach for automatic elimination of cloaked pages in one time process.
- Difficulty in analyzing the justifiability measurement, i.e. which was done manually (features extraction were done manually).
- No scheme is available to measure lifetime of cloaked pages which are independent of any online available tool.

IV. Problem Statement

For web surfing you are completely depend on search engines and imagine if top results shown are not of your use which is result of artificial boosting of web pages. Spammers have created many different spamming techniques that make it difficult for search engines to find out. Also there are some legitimate spam sites which are detected by spam filters which is not genuine. Entry level check on web sites is required in order to get better efficiency.

V. Proposed System

An easy treatment of web spam is to simply remove it from the result list. We have proposed simple and best anti spamming approaches through which spam could be detected and blocked automatically so that relevancy can be improved.

Advantages

- Provides an easy way to differentiate spam and normal page and removes spam easily and automatically.
- Automatic detection as well as elimination spam in single process, so that users will get only the spam free results on searching.
- Our approach can combinable with online available tools to add more spam detection.

We have proposed a simple approach for measuring lifetime of cloaked pages for dynamic cloak reduction.

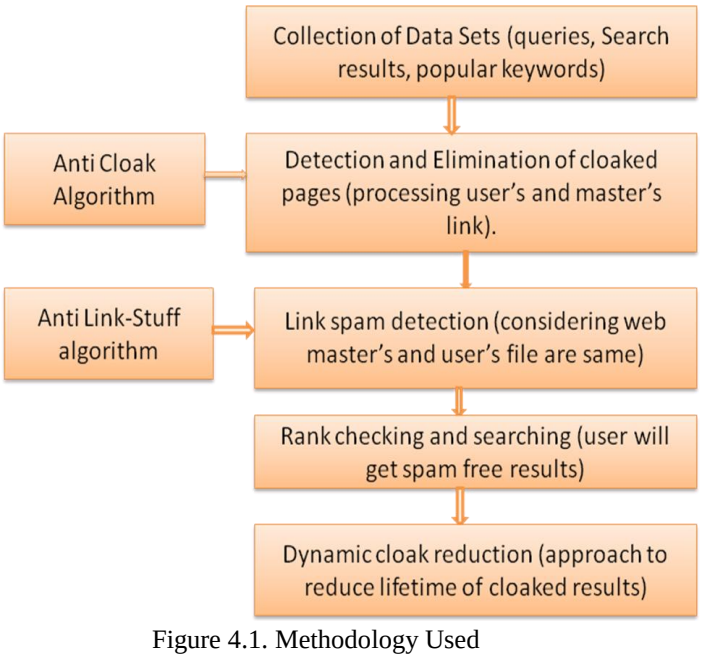


Figure 4.1. Methodology Used

- **Proposed Algorithm:**

Anti cloaking: We have divide our algorithm into three parts. Although they run in single process.

Part 1: Data Collection

1. Data: Search terms, Web Master's [m] and User's file [u] for each term.

2. Term bag [tb] ← frequency of occurrence of each word. {Replace all non words characters with blank and hence extract words out of it}.

3. Key dictionary [sk] ← dictionary of popular keywords generally used by spammers for keyword stuffing {taken according to Google's standards}.

Part 2: Entry Level Checking

1. Threshold ← 20 % sk

2. If m have more repeated words and m [tb] > t then page will discard else add to data base.

3. Also R= m [tb] - u [tb] U u[tb] – m[tb]

4. If R is more then normal term frequency difference then sends it for part 3 else add it to database {means pages have difference but not for deceiving search engine ranking algorithms}

Part 3: Cloak Status

1. If m [tb] > t then cloak status not clear else status is cleared {usually web master send more popular keywords to web master's copy}.

2. If cloak status is clear then values will add to data base else web page will delete from the data base.

Note- Added values will only available for searching algorithm.

Anti Link-Stuff Algorithm:

1. Data set {Collect pages for a given query}.

2. Enter the URL of the web page from the data set.

3. Link Extraction {Parse complete page}.

3.1 Extract links and store to an array.

3.2 Filter out hyperlinks to pages within the same page.

3.4 Do step 3 until complete web page will parsed.

4. Find Duplicate Penalty

4.1 Compare domain names of incoming and outgoing links.

4.2 If common nodes exist between incoming and outgoing links page will marked as "penalty", else marked as "free".

4.3 If penalty was marked then link stuffed else not stuffed

Note- **Not stuffed pages send for ranking and repeat the process.**

VI. CONCLUSION & FUTURE ENHANCEMENT

CONCLUSION

We give best approach for automatic cloak detection and elimination. We have given a combine approach for entry level keyword stuffed page detection, elimination of cloaked pages have surveyed many link combating approaches and implement our best approach by taking only best part of them. Our approach can also be implement with online available tools. Also dynamics of cloaked pages were discussed cloaked pages on a single platform. We have presented a better approach for the same.

FUTURE ENHANCEMENT

More types of spamming detection can be add in our present work.A scheme can be design through which more than one spamming detection works simultaneously. For dynamic cloak detection frequency of checking can be increased for efficient detection of lifetime of cloaked pages. Link stuffing detection can further be integrated to cloak reduction.

VII. LIMITATION

In our work we have done Keyword and link spam type of cloaking detection although there are several more types of cloaking which degrade the quality of the search engines.
We have assumed present available quality standards of the Google search engine for our work. We don't consider the new upcoming standards (anti web-spam algorithm updates).
For cloaking detection we have considered only the content based factors. Further work can be done by considering more factors such as redirection. Link Spam detection approach and cloaking detection approach works separately. Our approach works offline only.

VIII.REFRENCES

[1] David Y. Wang, Stefan Savage, and Geoffrey M. Volker "Cloak and Dagger: Dynamics of Web Search Cloaking" ACM 978-1-4503-0948-6/11/10-CCS11 -2011.

[2] Yuan Niu, Yi Mia Wang, How Chen Ming Ma & Francis Hsu. "Quantitative Study of Forum Spamming using context based analysis" University of California-2006.

[3] Baoning Wu & Brain D. Davison "Detecting Semantic Cloaking on the web"IW3C2 ACM 1595933239/06/0005 - May 2006.

[4] Kirill Levchenko, Andreas Pitsillidis, Neha Chachra "ClickTrajectories: End to End Analysis of Spam value Chain. 1081- 6011/11 DOI 10.1109/SP 2011.24- IEEE 2011.

[5] Gabriele D'Angelo, Fabio Vitali, Stefano Zacchiroli-Marc "Content Cloaking: Preserving Privacy with Google Docs and other Web Applications" ACM987-1-60558-638-0/10/03 - 2010.

[6] Yi Li a, Jonathan J. H. Zhu b& Xiaoming Li c "Survey of major techniques for combating link spamming, JICS-2010.

[7] What Google knows: Privacy and Internet Search Engines? by Omer Tene- 2007.

[08] Wael H. Gomaa, Aly A. Fahmy, "A Survey of Text Similarity Approaches," International Journal of Computer Applications, Vol. 68, No. 13, pp. 13-18, 2013.

[10] Md. Abu Kausar, Md. Nasar, Sanjeev Kumar Singh, A sing Genetic Information,

Vol.1, No.3, pages 122-127, 2013.

[11] N. Hashimah Sulaiman and D. Mohamad, A Jaccard Based Similarity Measure for Soft Sets, Proc. of IEEE Symposium on Humanities, Science and Engineering Research, pages 659-663, 2012.

[11] Sung-Hyuk Cha, "Comprehensive Survey on the Distance/Similarity Measures between Probability Density Functions," International Journal of Mathematical Models and Methods in Applied Sciences, Vol. 1, Issue 4, pp. 300-307, 2007.

[13] J. Singh, P. Singh, Y. Chaba, Performance Modeling of Information Retrieval Techniques Using Similarity Functions in Wide Area Networks, International Journal of Advanced Research in Computer Science and Software Engineering, Vol.4, Issue12, pp.786-793, 2014.

[14] David E. Goldberg, Genetic algorithm in search, optimization, machine learning (Adison Wesley, 1989).

[15] J. H. Holland, Adaptation in natural and artificial systems (2nd ed., MA: MIT Press, Cambridge, 1992).

[16] Z. Michalewicz, Genetic algorithm + data structure = evolution programs (Springer, 1996).

[17] L. Egghe and C. Michael, Strong Similarity Measures for the Ordered Sets of Documents in Information Retrieval, Information Processing and Management, 38(6), 823-848, 2002.

[18] William P. Jones, George W. Furnas, Pictures of Relevance: A Geometric Analysis of Similarity Measures, Journal of American Society for Information Science, Vol. 38, No.6, pages 420-442, 1987.

[19] Michael Gorden, Applying Probabilistic and Genetic Algorithms for Document Retrieval, Communications of ACM, Vol.31, No. 10, pages. 1208-1218, 1988.

[20] V. N. Gudivada, V. Raghavan, W. I. Grosky and R. Kasanagottu, Information Retrieval on the World Wide Web, IEEE Internet Computing, pages 58-68, 1997.

[21] R. Baeza-Yates and B. Ribiero-Neto, Modern information retrieval (Addison Wesley, New York, 1999).

[22] G. Salton and C. Buckley, Improving Retrieval Performance by Relevance Feedback, Journal of the American Society for Information Science, 41(4), pages 288-297, 1990.

[23] C. Lopez-Pujalte, V.P. Guerrero-Bote and F. Moya-Anegon, A Test of Genetic Algorithms in Relevance Feedback, Information Processing and Management, 38(6), pages 795-807, 2002.

[24] C. Lopez-Pujalte, V.P. Guerrero-Bote and F. Moya-Anegon, Order-Based Fitness Function for Genetic Algorithms Applied to Relevance feedback, Journal of the American Society for information Science and Technology, 54(2),pages 152-160,2003.

[25] [illegible]
and I

Search: A Study of Different Mutation Rates, ACM Trans. Inter. Tech., 4(4), pages 378–419, 2005.

[26] P. Sihombing, A. Embong and P. Sumari, Comparison of Document Similarity in Information Retrieval System by different formulation, Proc. 2nd IMT-GT Regional Conference on Mathematics, Statistics and Applications, Malaysia, Penang 2006.

[27] M. Zolghadri Jahromi, and M.R. Valizadeh, A proposed query-sensitive similarity measure for information retrieval", Iranian Journal of Science & Technology, Shiraz University, Vol. 30, no. B2, pages.171-180, 2006.

[28] A. Ahmad, A. Radwan, A. Bhagat, Abdel Latef, Abdel Mgeid A. Ali and Osman A. Sadek, Using Genetic Algorithm to Improve Information Retrieval Systems, World Academy of Science, Engineering and Technology, pages 1021-1027, 2008