ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 8, August 2015

Spam Web Page Detection based on Content and Link Structure of the Site

Kiran Hunagund[1], Santhosh Kumar K L[2]

M.Tech Scholar, Department of CS&E, Nitte Meenakshi Institute of Technology, Bangalore, India[1]

Assistant Professor, Department of CS&E (PG), Nitte Meenakshi Institute of Technology, Bangalore, India[2]

Abstract: Spam Web page is a website which does not contain any useful information. Spammer will create such spam pages for fun or to increase page rank in turn to generate their revenue. The Spam webpage Detection is one of the top challenges for the search engines. There are two different approaches for the detection of spam web page such as Link and Content based analysis. In this paper, we mainly focus on Content based analysis. We have used parameters such as average length of a word, keyword stuffing, and content of a body, number of stop words, unique count for body and title of page are used to identify spam.

Keywords: Search engine, Web mining, Spam web page, Content based analysis.

I. INTRODUCTION

Web mining is a part of data mining technique. In web mining it extracts the content from the web documents/services. Abstraction is the useful information in world- wide-web.The WWW is huge, widely distributed and global information service centre for Information services such as: news, advertisements, consumer information, financial management, education, government, e-commerce, Hyper-link information access and usage information etc. As the web users increased, the spammers also increased to create the spam web page where it contains no useful information. To detect spam web page, we have to abstract the content of a web page and analyze the content.

Search engine produces thousands of results for a given query. The search engine have to produce top ten results for a single query which is helpful for the end users which will be challenging task for the search engine. The spammer creates spam web page to increase the page value this may mislead to the end users. The search engine have to give good pages to users otherwise they will get fed up and change search engine. There are two type of spam web page detection

i. Content based method
ii. Link based method

Where, the content based method is used to abstract the content of a given web page and used to make analysis on it. The Link based method analysis is done based on link structure of the sites which are connected to the site and are measuring spam of web page.

II. RELATED WORK

We considered some of the papers of related to our work. The spam detection on the content of a web page [1], is independent of any language methods. "Number of words in the page" - the spammer simply adds the keywords which are unrelated to rest of the page. For detection we have to count the number of keywords in a page and also percentile of keywords in page content. "Number of words in the page title" - more number of keywords leads to the spam. The search engine searches the title of a page, where title represents content of a web page spammer, the popular keywords which are unrelated to the page. As the number of keywords increases more the spam."Average length of words"- in this, spammer combines the popular words like free videos, where it contents two words free and videos it increase the length of a word."Amount of anchor text"- anchor text redirect to another page, where spammer adds more number of anchor text which redirects to same page."Compressibility"- search engine add more value, which have more keywords. So spammer adds single keyword several times. If keyword is multiplied more than threshold level it may considered as a spam web page."Fraction of page drawn from globally popular words" - popular words are stop-words, which we use frequently. The non-spam contain at least 10% of stop-words, if not then it is spam web page."Fraction of globally popular words"- so spammer simply adds more no of stop-words to increase the webpage rank. If it is more than 75% then it must be a spam web page.

Content Spam [6] and "Title Spamming" [8] overstuffed with words in the title of a web page. "Body spamming"- overstuffed with more keywords that are query will be searched by the user."Meta-Tags Spamming" – where Meta tag describes the content of web page, so search engine looks for the Meta tags. "Anchor Text Spamming" - in this, spammer adds popular links in to the web page. The spammer adds thousands of links in to a web page that directs to another webpage.

III.PROPOSED METHODOLOGY

The content based spam web page detection is based on characters of web page .The spam web pages behave oddly compared to the standard web page. We have used six approaches to find the spam page. At the end, we have combined all those methodologies in our system.

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 8, August 2015

A. Standard length of a word

The standard length of a word is one of parameter for the content based spam web page detection method. The webpage can be considered as the spam, if the standard length is greater than maximum likelihood of spam page, because the spammer simply mixes the two or more keywords like freevideos, where there are two words free and videos because spammer stuffs the mixed words in to the body of a page. The spammer tags the malicious words in to the body of a page.

Algorithm 1**:** STANDARD LENGTH OF A WORD

Step 1. Enter the URL
Step 2. Extract the content of a body tag of page
Step 3. Remove the stop-words in the page
Step 4. Count the number of words
Step 5. Sum of Count the each word length
Step 6. Find the standard of a word by Divide the sum of count of each word length to the number of word.
Step 7. If standard length>8 then
Step 8. it is spam web page count++
Step 9. Else non-spam

Standard length of a word is a keyword that is in the body tag, where we have to remove the stop-words because it does not content any keywords and finds the standard length. If standard length is greater than 8 then the site spam web page.

B. Meta Tag Keyword Padding

Meta tag gives the information about the site, which is given by the web master. The information like about the clients and search engine etc., are added in the <head> section of HTML.
There are three parts of Meta tag

- Description
- Keywords
- Robots

Code example :< Meta name=”description” content=”keywords of a page”/>

Where search engine searches the keywords of Meta tag. If the Meta data keywords are not matching to the site content then the site is called spam web page. So, spammer adds more Meta keywords in a body of a page.

Algorithm2: META TAG KEYWORD PADDING STEPS

Step 1. Enter the URL of a site
Step 2. Extract the content of meta tag
Step 3. Remove the stop words from the meta tag
Step 4. Extract the content of a body tag
Step 5. Remove the stop-words from the body tag
Step 6. Count the no matching words of meta tag with the body tag
Step 7. If matched cont is greater than 5 then
Step 8. Spam count++
Step 9. Else non-spam

We have to extract the content of Meta and body, then remove the stop-words count to match the element. If the count is greater than 5 then there is likelihood of spam otherwise it is determined as non-spam.

C. Number of Words in a Body

Number of words in the web page is another method to detect spam. In this the amount of applicable content of a page is measured. The web page is designed to give the information to the users. Here, we extracted the content of the page to find the number of words in the page tag. If it is less than 100 then we can consider measured as spam web page.

Algorithm 3: NUMBER OF WORDS IN A BODY

Step 1. Enter the url
Step 2. Extract the content of page
Step 3. Remove the stop words from the page
Step 4. Count the number of words in a page
Step 5. If count is greater than 100
Step 6. Non-spam else
Step 7. Spam web page count++

Spammer basically creates a web page adds only few well-liked words and adds a images in a page and urls that directs to another page and stop words in the page is removed .then counts the no words in a page. If count is greater than 100 then the page is good otherwise it is spam.

D. Number of Stop-words in a Title

Another method is finding the fraction of stop words in a title. if the stop words in a title is more, then it is valid web page because web page describes the information so it contains more stop-words. The spammer adds keywords but not the stop-words, so if fraction of stop-words is less than 10 fractions then it is a spam web page.

Algorithm 4**:** NUMBER OF STOP-WORDS IN A TITLE

Step 1. Enter the url
Step 2. extract the title of url
Step 3. Count the no stop-words from the web page
Step 4. Count the no keywords in a title of a page
Step 5. percentage stoop-words =100*number of stop-word/no. of keyword
Step 6. If percentage is greater than 10 then
Step 7. Non-spam else
Step 8. Spam web page count++

Title describes the type of web page. The web designer uses the stop-words in a title but, the spammer will simply add the keywords not the stop-words. So, if the stop word is greater than 10 then it is not a spam web page.

E. Number of Distinctive Count of a Word in a Body Tag

In this method, we have to calculate the fraction of unique count in a body because; spammer adds same well-liked keywords several times in body tag for example cricket, free, videos etc. The search engine usually searches the well-liked words so if fraction is greater than 20% then there are maximum chances of spam web page.

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 8, August 2015

Algorithm 5: NUMBER OF DISTINCTIVE COUNTS OF A WORD IN A BODY TAG

Step 1. Enter the url of a webpage
Step 2. Extract the body tag
Step 3. Remove the stop-words
Step 4. Count the unique word count
Step 5. Find fraction
Step 6. If fraction is greater then 20 then
Step 7. Spam web page
Step 8. Count++ else
Step 9. Non-spam

In this method, there is maximum possibility of finding spam web page. It finds the number of distinctive keywords in a web page. The maximum distinctive count leads to the spam web page.

F. Number of Distinctive Count of a Word in a Title

In this method, we have to calculate the percentage of unique count in a title because, spammer adds same well-liked keywords several times in title of a web page example cricket, free, videos etc. The search engine usually searches the well-liked words so if fraction is greater than 20% then there are maximum chances of spam web page.

Algorithm 6**:** NUMBER OF DISTINCTIVE COUNTS OF A WORD IN A TITLE

Step 1. Enter the URL of a webpage
Step 2. Extract the content of a title tag
Step 3. Remove the stop-words
Step 4. Count the unique word count
Step 5. Find percentage
Step 6. If percentage is greater then 20 then
Step 7. Spam web page
Step 8. Count++ else
Step 9. Non-spam

In this method, it finds the number of unique keywords in a web page, maximum the distinctive count leads to spam web page.

G. Combining All Content Based Methods

This method combines all the parameters mentioned above. It increases the robustness of the system.

Algorithm: COMBINED METHOD

Step 1. Input count (output of all the content based methods)
Step 2. Sum the all count values
Step 3. If count>3 then
Step 4. Spam web page else
Step 5. Non-spam page

Each method detect the spam web page or not on their own conditions. If a page is spam then the count increases to one if not then remains null if count is greater than 3, then the page will be declared as a spam.

IV. RESULTS AND ANALYSIS

The final output of combining the all six content method is shown in figure 1.The table shows the Serial no, url of the site to be detected, percentage of a unique count title, percentage of a unique count body, length of a body, average length of a word in a body, percentage of stop-words in a body, keyword plugging and final is results to that spam or non-spam.

	URL	PER_TITLE	PER_BODY	lenth_of_body	AVG_LENTH	PER_STOPWORD	keyword	SPAM_NONSPAM
1	http://digitalpictureservices.co.uk	100.0	14.970060348510742	1517.0	7	0.0	0.0	notspam
2	http://www.direct-car-finance.pedez.com	80.0	30.559610366882129	18254.0	12	18.0	0.0	spam
3	http://direct.companieshouse.gov.uk	100.0	12.038140296336035	3034.0	8	0.0	0.0	notspam
4	http://dolgam-snowdonia.co.uk	100.0	20.858434677124023	6354.0	6	9.0	0.0	notspam
5	http://domi.co.uk	100.0	17.732948303222656	33146.0	7	0.0	0.0	notspam
6	http://dreamking-productions.co.uk	50.0	21.834933362435758	58043.0	5	12.0	0.0	notspam
7	http://www.e-pubs.royalsoc.com	52.941177	13.718411445617676	5068.0	10	0.0	0.0	notspam
8	http://eastayrshire.boys-brigade.org.uk	100.0	100.0	8.0	8	0.0	0.0	spam
9	http://extra.exeter-airport.co.uk	100.0	14.102563858032227	1393.0	6	0.0	0.0	notspam
10	http://f4beauty.co.uk	100.0	6.140845298576709	11972.0	9	28.0	0.0	spam
11	http://ibiza-tickets.co.uk	33.333332	18.819776535034418	11404.0	9	0.0	0.0	notspam
12	http://ibiza-tickets.co.uk	33.333332	18.819776535034418	11404.0	9	0.0	0.0	notspam
13	http://moneyworld-fa.co.uk	75.0	3.9252927030331421	17971.0	7	0.0	0.0	notspam
14	http://www.personal-finance-uk.pedez.com	80.0	30.559610366882129	18254.0	12	18.0	0.0	spam
15	http://www.192directory.co.uk	100.0	6.928234577176955	22300.0	14	16.0	10.0	spam
16	http://www.2dolist.co.uk	100.0	100.0	8.0	8	0.0	0.0	spam
17	http://www.aaxon.com	60.0	4.567622184753418	38857.0	10	0.0	6.0	spam
18	http://www.ablepest.co.uk	100.0	17.76439666748047	26507.0	14	18.0	6.0	spam
19	http://www.acgl.co.uk	100.0	19.870410919189453	2207.0	9	25.0	0.0	spam
20	http://www.adboards.co.uk	77.27273	12.211981773376465	26549.0	12	0.0	0.0	notspam
21	http://www.adboards.co.uk	77.27273	12.161383628845215	26510.0	12	0.0	0.0	spam

Figure1. Result of Combined method for Spam detection

The accuracy of the methodology can be estimated using detection rate and false positive.

The detection rate is the percentage ratio of detected spam web page to total number of spam website tested. More the detection rate leads to efficiency of the spam detection.

Detection rate= (Detected spam web site/total number of spam web site tested) * 100

Our Detection rate result = (27/35)*100=77.14

TABLE I ESTIMATION OF DETECTION RATE

Total no of spam site tested	Spam site detected	Detection rate
35	27	77.14

The false positive is defined as percentage ratio of total no of normal sites classified as spam and number of normal website. Lesser the false positive leads to the more efficiency of the spam detection.

False positive= (# of normal sites classified as spam/# of normal website)

Our False positive result= (13/65)*100=20.0

In our approach, we tested on 100 web sites; we got 77.14 % correct results in terms of detection rate and 20% in terms of false positive rate.

TABLE II ESTIMATION OF FALSE POSITIVES

Total of normal sites classified as spam	Total number of normal website	False positives
13	65	20.0

V. CONCLUSION

In this paper, the content based methods for spam detection is used. Content based method is implemented based on six different approaches. At the end we combined all the six approaches and got 77.14% accurate detection and 20% of false positive rate. As the future

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 8, August 2015

work, we can also consider the Link-based approach to make the system more robust.

VI. ACKNOWLEDGMENT

The authors express their sincere gratitude to **Prof. N R Shetty,** Director, Nitte Meenakshi Institute of Technology and **Dr. H C Nagaraj**, Principal, Nitte Meenakshi Institute of Technology for providing encouragement, support and the infrastructure to carry out the research.

REFERENCES

[1] Luca Becchetti, Carlos Castillo, "Linki-based and content-based Techniques", Universita di Roma "La Sapienza" 'via Ariosto 25, 00185 Roma, Italia
[2] Sergey Brin and Lawrence Page "The Anatomy of a Large-Scale Hypertextual Web Search Engine" Computer Science Department, Stanford University, Stanford, CA 94305, USA.
[3] Vijay Krishnan, Rashmi Raj, "Web Spam Detection with Anti Trust Rank",AIRWEB'06, August 10, 2006, Seattle, Washington, USA.
[4] Zolt´an Gy¨ongyi, Hector Garcia-Molina, Jan Pedersen, " Combating Web Spam with Trust Rank ", Proceedings of the 30th VLDB Conference, Toronto, Canada, 2004.
[5] Nikita Spirin, Jiawei Han, "Survey on Web Spam Detection: Principles and Algorithms"
[6] Alexandros Ntoulas, Marc Najork, "Detecting Spam Web Pages through Content Analysis", May 23–26, 2006, Edinburgh, Scotland.
[7] Brin, Sergey, and Lawrence Page. "Reprint of: The anatomy of a large-scale hyper textual web search engine", Computer Networks, 2012.
[8] Maria Soledad Pera. "Identifying Spam Web Pages Based on Content Similarity", Lecture Notes in Computer Science, 2008.
[9] Dennis Fetterly. "Detecting spam web pages through content analysis", Proceedings of the 15th international conference on World WideWeb - WWW 06 WWW 06, 2006.
[10] Alka Jindal. "Contrast of link based web ranking techniques", 2008 International Symposium on Biometrics and Security Technologies, 04/2008
[11] Carlos Castillo. "A reference collection for web spam", ACM SIGIR Forum, 12/1/2006
[12] Guo, G.M... "A computer-aided bibliometric system to generate core article ranked lists in interdisciplinary subjects", Information Sciences, 20070901
[13] http://www.minterest.org/free-seo-keyword-density-tool-checker/
[14] Pruthi, Jyoti. "Anti-Trust Rank: Fighting Web Spam", International Journal of Computer Science Issues (IJCSI)/16940784.

BIOGRAPHIES

Kiran hunagund received B.E in Computer Science and Engineering from Visvesvaraya Technological University. He is currently studying M.Tech Final Year in Computer Science & Engg. At Nitte Meenakshi Institute of Technology, Bangalore. His areas of interest are Web Mining, Data Mining.

Santhosh Kumar K L received B.E and M.Tech in Computer Science and Engineering from Visvesvaraya Technological University. He is currently working as Assistant Professor in the Department of Computer Science & Engg (PG)., Nitte Meenakshi Institute of Technology, Bangalore. His areas of interest are Image Processing, Video Shot detection & Video Summarization, Activity Tracking and Facial Expression Recognition system.