LINK SPAM ANALYSIS USING HYBRID TECHNIQUES

Ms .Vinita Kule[1]
Ms.Darshana Ranbagle[2]
Ms.Reshma Dabade[3]

Abstract

Link spam is a form of spamming that recently became publicized most often when targeting weblogs, but also affects wikis, guest-books, and online discussion boards. The applications or the blogs that contain any hyperlinks which are entered or displayed by the user or visitor, then that application or blog may be the containing some malicious spam links which makes the other visitors as a target. Such hyperlinks can increase the page rankings of that respective page in the Google search engine. This may lead to the suppression of other relevant search web pages as these malicious webpages (with increased page ranks) will appear at the top of the search results. There are some approaches to detect such spam links. The techniques involved in Link Spam Analysis are either Content-based or Link-based. In this report, we will be implementing both these techniques to improve the efficiency of detecting spam links. Crawler is used to extract the attached links to the web page. These URLs will be studied based on Content-based algorithms. This content will be used to process the web page. Based on this results, we will get the result whether the entered URL contains spam links or not.

Keywords:*Spamdexing, Web Crawler.*

Introduction

Link Spam, also called as Spamdexing, is a method of manipulating the search engine indexes. In this method, the spammer inserts some fraudulent or fake links into the web page of a legitimate sites to allure the user to visit that fake links. By doing so, the page rank of those fraudulent links will increase, resulting in increase in the importance of that link. The motive of spammer behind this is to gather credentials of the user or steal some important information of the user. This data will be used by him to perform some illegal or fraudulent activities, against the user's understanding.

[1] vinitakule14@gmail.com Bharati Vidyapeeth College of Engineering Sector-7,C.B.D,Belpada,Navi Mumbai-400614,India.
[2] darshana.ranbagale@gmail.com Bharati Vidyapeeth College of Engineering Sector-7,C.B.D,Belpada,Navi Mumbai-400614,India.
[3] reshmadabade23@gmail.com Bharati Vidyapeeth College of Engineering Sector-7,C.B.D,Belpada,Navi Mumbai-400614,India.

Also, the increase in page rank will increase the number of victims, as it will be proved to be a legitimate site or link. This arises in an urge of finding a solution to deal with those issues. Detection of spam links is based on two major approaches: Content-based analysis and Link-based analysis. Content-based technique deals with the content present in the web page of that link and Link-based technique examines the link structure of that link.

Web Crawler is an Internet Bot which systematically browses the World Wide Web for the purpose of web indexing. It is a technique by which we can list out the hyperlinks attached to a particular web link. It first creates a list of the websites to be visited and that list is called as seed. Then these websites are searched in a systematic manner to detect all the hyperlinks associated with that web page. In this paper, we will be implementing a crawling technique on the entered link to gather all the hyperlinks related to that page. And on that links, we will be applying some classification based algorithms, which are discussed further, to detect whether the entered link is spam or not.

Proposed System

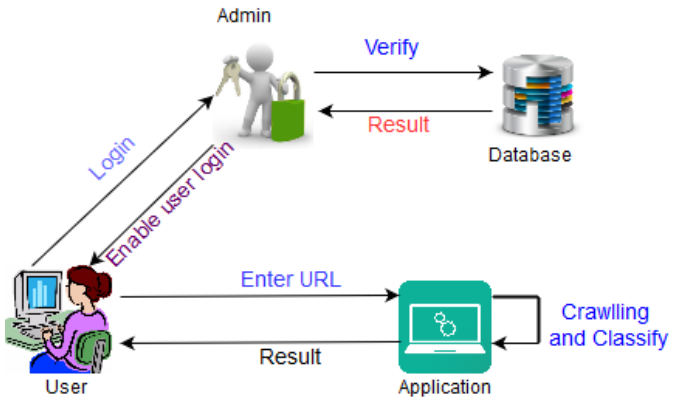

Fig. Architecture Diagram

In this paper, we will be detecting whether the entered URL is spam or not. For this, we will be using Crawling method to extract the hyperlinks associated with that URL. And on this output, we will be applying the content based analysis using the 6 algorithms which are discussed further. Based on this result, we will be deciding whether the entered URL is spam or not.

After the successful user login process, the user will enter the URL which he want to detect whether it is spam or not. Crawling technique will be used to extract the hyperlinks connected or associated with that web page. The below given algorithms will be used to classify that web URL whether it is Spam or not.

1. **Standard Length of Word**

The standard length of 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 page.

Algorithm 1: STANDARD LENGTH OF 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 the Count of each word length

Step 6. Find the standard length of word by dividing the sum of count of each word length to the number of word

Step 7. If standard length>8 then it is spam

Step 8. Else non-spam

Step 9. Web page count++

2. **Meta Tag Keyword Padding Steps**

Meta tag gives 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. 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.

Algorithm 2: 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 Meta tag

Step 6. Count the no matching words of Meta tag with the body tag

Step 7. If matched content is greater than 5 then Spam

Step 8. Else non-spam

Step 9. Web page count++

3. 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 user. 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 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, THEN Non-spam

Step 6. Else Spam

Step 7. Web page count++

4. 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.

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 number of stop words from the web page.

Step 4. Count the number of keywords in the web page.

Step 5. Percentage of stop words= (100*Number of stop words)/Number of Keywords.

Step 6. If Percentage is greater than 10, then non-spam

Step 7. Else Spam

Step 8. Web page count++.

5. Number Of Distinctive Counts 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 some well-liked words several times in a body tag, for example, cricket, free, videos, etc. The search engine usually searches the well-liked words.

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

Step 1. Enter the URL.

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>20, then Spam

Step 7. Else Non-spam

Step 8. Web page count++

6. **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 spammers add same well-liked keywords several times in a title of a web page, example, cricket, free, video, etc. The search engine usually searches the well-liked words.

Algorithm 6: NUMBER OF DISTINCTIVE COUNT OF A WORD IN A TITLE

Step 1. Enter the URL.

Step 2. Extract the contents of the title tag.

Step 3. Remove the stop words.

Step 4. Count the unique word count

Step 5. Find percentage

Step 6. If percentage>20, then Spam

Step 7. Else Non-spam

Step 8. Web page count++

Combining All Content Based Methods:

This method combines all the parameters mentioned above.

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 Spam web page

Step 4. Else Non-Spam web page.

Conclusion

Spam Link is becoming a serious network security problem, causing financial loss of billions of dollars to both consumers and e-commerce companies. And perhaps more fundamentally, Spam Link has made e-commerce distrusted and less attractive to normal consumers. In this paper, we have studied the characteristics of hyperlinks attached to the entered link using

crawling technique. We then used an anti-spam algorithm, which is a combination of six algorithms. This anti-spam algorithm is not only useful for detecting spam link attacks, but also can shield users form malicious or unsolicited links in Web pages. Our future work includes further extending the algorithm to improve the efficiency and accuracy of approach to detect spam links

References

[1] Rajendra Kumar Roul, Shubham Rohan Asthana, Mit Shah, and Dhruvesh Parikh, "Detection of spam web page using content and link-based techniques: A combined Approach," BITS, Pilani-K.K.Birla Goa Campus, Goa, India, *Sadhana*Vol. 41, No. 2 pp. 193-202, 2 February 2016.

[2] Kiran Hunagund, Santosh Kumar K L, "Spam Web Page Detection based on Content and Link Structure of the Site", Department of CS&E, Nitte Meenakshi Institute of Technology, Bangalore, India, Vol. 4, Issue 8, August 2015.

[3] Jing Wan, Mufan Liu, Xuechao Zhang, "Detecting Spam WebPages through Topic and Semantic Analysis", Beijing, China, 2015.