© 2016 IJSRSET | Volume 2 | Issue 3 | Print ISSN : 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

A Research on Web Content Extraction and Noise Reduction through Text Density Using Malicious URL Pattern Detection

Charmi Patel, Prof. Hiteishi Diwanji

Information Technology, L. D. Engineering College, Ahmedabad, Gujarat, India

ABSTRACT

A Web Page has large amount of information including some additional contents like hyperlinks, header footer, navigational panel; advertisements which may cause the content extraction to be complicated. Page Segmentation is used to detect the noisy content block by detecting malicious URL from Web Pages. Main aim of this research is detecting malicious URL during content extraction by checking different patterns of URL. Performance is analysed based on precision, recall, execution time and noise detected using proposed algorithm.

Keywords: Page segmentation, Malicious URL, URL patterns, Text density

I. INTRODUCTION

Web Mining is Data Mining techniques which automatically discovers and extract information from World Wide Web. Web Mining is used to capture relevant data about consumer, individual user and several others [1].Web Mining decomposed into Resource Discovery, Information Selection & Pre-processing, Generalization and Analysis. It is very useful for pre-processing the data in web mining, expert system, knowledge discovery recommendation system, decision making. Tasks such as false advertisement Detection, demand forecasting, and comment extraction on product reviews use this [3].

Web Pages contain both information parts and noisy parts. The noisy parts may harm the web content mining. The noisy data means the links which are not actual when user click on that links users are redirected to the webpage which are fake. This will reduce efficiency of the webpage. There are various techniques used for content extraction and noise removal. Each method has different percentage of content extraction and noise removal. The proposed algorithm will be the enhancement in the existing algorithm for noise detection.

II. METHODS AND MATERIAL

A. Related Work

There are many techniques available for content extraction and noise detection from webpage. Each has its own advantages and disadvantages.

Warid Petprasit and Saichon Jaiyen[3] proposed web content extraction technique based on subject detection and node density. Algorithm to identify the subject node is based on maximum weight. Weight is calculated for each candidate node using the tag name (Ni), the keywords in meta tag(Ti) and title tag(Ki), and some properties in cascading style sheet (CSS) including font weight, font-size, and display properties. Node Density method is used to find the data rich region node using threshold value.

Shuang Lin, Jie Chen, Zhendong Niu[10] proposed Visual Clues Concept for Extracting Main Data from Deep Web Pages. To meet the reading habits of human beings deep Web pages designers always arrange the data records and the data items with visual regularity of them. That visual clues are used for content extraction.
Dandan Song, Fei Sun, Lejian Liao[6] proposed a hybrid approach for content extraction with text density and visual importance of DOM nodes.here, two kind of

information is consider : the textual information and the visual information. Accordingly, Text Density and Visual Importance are defined for the Document Object Model (DOM) nodes of a web page and content is extracted.

Nupur S. Gawale, Nitin N. Patil[7] proposed A System To Detect Malicious URLs for Twitter. They introduces a system to provides the security to multiple users of twitter by sending some alert mails. The goal is to download URLs in real time from multiple accounts. Then it get entry points of correlated URLs.This system finds such malicious URLs by using features like initial URL, similar text, friend follower ratio and relative URLs.

Tiliang Zhang, Hua Zhang, Fei Gao[12] proposed a Malicious Advertising Detection Scheme Based on the Depth of URL Strategy. The proposed Malicious advertising detection scheme contains three modules: suspicious URL extraction module, filter module and logging system.

B. Proposed Methodology

In proposed work noisy data means the links which are not actual or fake links. When user click on that links users are redirected to the webpage which are fake. The proposed algorithm will be the enhancement in the CECTD-DS algorithm for malicious URL detection. Webpage and patterns are given as input. The algorithm will match the pattern of malicious URL with given input pattern. This technique will increase malicious URL detection accuracy and so detection of noise is more accurate.

Algorithm : Pattern Matching Function

```
INPUT: Input pattern of malicious URL, webpage
OUTPUT: percentage of noise detection
  1. Let memory available.
  2. Apply page segmentation on webpage.
  3. Remove noisy content like copyrights, header,
     footer etc.
  4. Traverse each segment of webpage one by one
  5. If (URL pattern of current segment matches with
     input pattern of malicious URL)
  6. Count=count+1;
  7. Mark that URL as malicious URL
  8. End if
  9. Apply CECTD-DS technique for original content
     extraction and during that marked malicious URL
     are not extracted.
 10. End
```

Here, lexical and network based some URL features are taken into consideration to detect whether URL is malicious or not. There are some common patterns found if we analyse different malicious URL. Let us discuss some of them.

Domain Age : Domain age can be found by using WHOIS properties which use the date of creation of that domain. As the site is older there are less chances for that site to be malicious and harmful.

Traffic Received: Each website has its own user traffic. This web traffic of website is calculated by ALEXA rank.

Presence of Suspicious Symbol: A symbol like @ present in the URL. In normal programming one cannot use a symbol like @ in the URL. Whenever a @ symbol is used in the URL, all the text before it is ignored.

E.g.: www.paypal.com@pqr.com.

Even though this looks like the link to paypal.com, the user is taken to pqr.com.

Misplaced Top Level Domain:

E.g.: http://a9s7px4xfdgdfhfciy4x.Opu.ru/https/www.paypale.fr/Client175414541

In the above URL, we can see that the word PayPal has been used. At a quick glance, it might look fine. But a closer look shows that the actual domain name is very weird and the word PayPal is present in the sub section of URL, and even it is also spelled wrong. So we say that the familiar top-level domain (here PayPal) is out of position.

Number of dots present: In any URL if more number of dots are present than there is more possibility that site is phishing.

Presence of IP address in URL:

E.g., http://185.24.44.67/Ibchileperfigfdgfiento/Process?MID=&AID=LOGIN-0004&RQI=500
2345125BE97.

When we want to access any website we are not using IP address, original IP address is masked by domain name and we are using domain name to fetch any website in internet. If IP address is present in the URL than there are more chances of it to be phishing one as IP address in URL only used in intranet. In internet it might have been used to hide the phishing domain name.

Log Records: By analysing log records of any website we can identify xss attack and advertisement URLs.

In these proposed methodology we are combining CECTD-DS algorithm and lexical based URL features like presence of suspicious symbol, no. of dots present in URL for efficient content extraction and noise removal.

III. RESULTS AND DISCUSSION

Implementation and Result

The proposed technique will be implemented using the MATLAB Tool in which mathematical toolbox will be used for implementing. MATLAB is a multi-paradigm numerical computing environment with fourth- generation programming language.

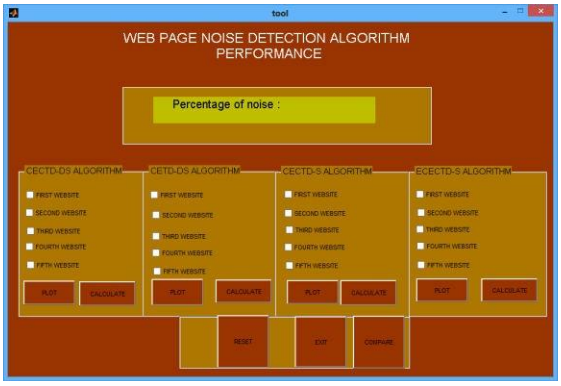


Figure 1. Tool

This is the tool generated for proposed system, in which three algorithms CECTD-DS, CETD- DS, CECTD-S and new proposed algorithm ECECTD-S is implemented. Five different websites are taken as input and noise is calculated using these algorithms.

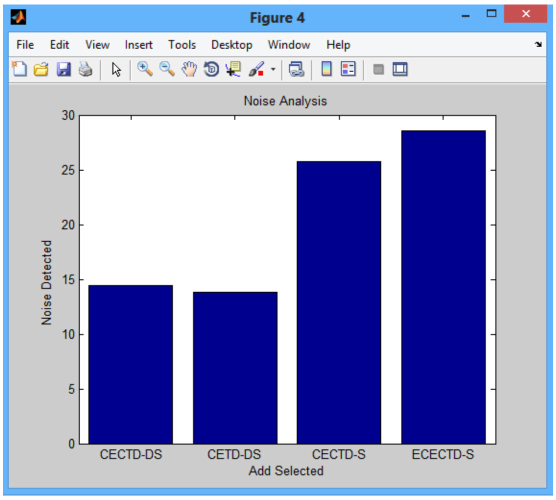


Figure 2. Comparison of Noise detected

As shown in graph percentage of noise detected using different algorithms are plotted.

Newly proposed algorithm detects more noise than existing algorithms. There are other graphs plotted for precision, recall and execution time required for analysis purpose.

Here, precision is the number of true positives (i.e. the number of URLs correctly identified as malicious) divided by the total number of URLs selected for comparison (i.e. the sum of true positives and false positives).

$$\text{Precision} = \frac{tp}{tp + fp}$$

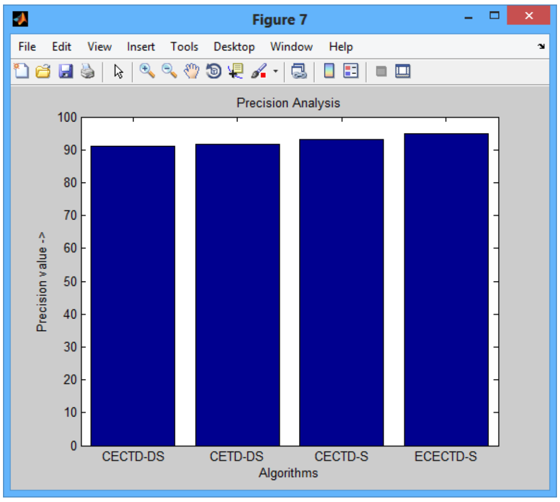


Figure 3. Comparison of Precision of different algorithms

Recall is defined as the number of true positives divided by the total number of malicious URLs that are actually in the webpage (i.e. the sum of true positives and false negatives, which are URLs that were not labeled as malicious but should have been).

$$\text{Recall} = \frac{tp}{tp + fn}$$

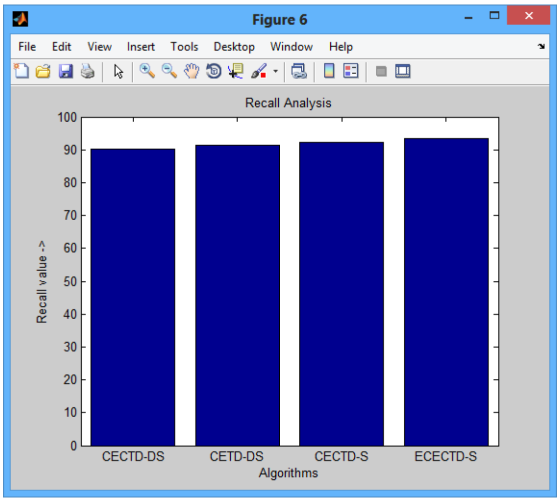
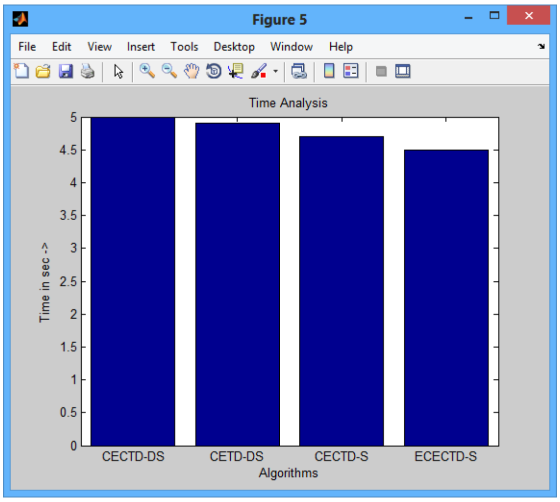


Figure 4 Comparison of Recall of different algorithms



Here comparison table is given for one website as input and its output values are shown in percentages.

Table 1 Comparison Table

Parameter	CECTD-DS	CET D-	CECTD-S	ECECTD-S
Noise(%)	25.6747	27.519	12.0847	25.9522
Execution Time(sec)	5.33	4.42	4.2	4.23
Recall(%)	90.4	92.8	92.2	94
Precision(%)	91.9	93	95	95.2

IV. CONCLUSION AND FUTURE WORK

In this research new technique for content extraction and noise detection from webpage is proposed which uses page segmentation technique, text density and URL pattern matching. In content extraction using text density algorithm some time malicious URL are taken as legitimate URL and vice versa. Therefore in proposed work we have added URL pattern matching function. Lexical and network based features of URL are used to detect suspicious URL in content extraction. Lexical based features are used in malicious URL matching function. By this proposed work we will improve the accuracy of noise detection and decrease the false positive and false negative value for malicious URL detection.

In future both lexical and network based features are used parallel for malicious URL detection which will increase the accuracy for noise detection in webpage.

V. REFERENCES

[1] Shuang Lin, Jie Chen, Zhendong Niu, “Combining a Segmentation-Like Approach and a Density-Based Approach in Content Extraction” ,TSINGHUA SCIENCE AND TECHNOLOGY, ISSNll1007- 0214ll05/18llpp256-264 Volume 17, Number 3, June 2012

[2] A.F.R.Rahman, H.Alam and R.Hartono, “Content extraction from HTML documents”, International workshop on Web Document Analysis, pp. 7-10, 2001.

[3] Warid Petprasit and Saichon Jaiyen, “Web Content Extraction Based on Subject Detection and Node Density”, 978-1-4799- 6049-1/15/$31.00 ©2015 IEEE

[4] W3C Document Object Model (2009) Website. http://www.w3.org/DOM

[5] F. Sun, D. Song, and L. Liao, "DOM Based Content Extraction via Text Density," Special Interest Group on Information Retrieval, ACM, 2011

[6] Dandan Song, Fei Sun, Lejian Liao, "A hybrid approach for content extraction with text density and visual importance of DOM nodes" , Springer-Verlag London 2013

[7] Nupur S. Gawale, Nitin N. Patil, "Implementation of A System To Detect Malicious URLs for Twitter Users" ,IEEE- ICPC 2015

[8] Aanshi Bhardwaj, Veenu Mangat, "A Novel Approach for Content Extraction from Web Pages", 978-1-4799-2291-8/14/$31.00 ©2014 IEEE

[9] Yogesh W. Wanjari, Vivek D. Mohod, Dipali B. Gaikwad, Sachin N. Deshmukh, "Automatic News Extraction System for Indian Online News Papers", 978-1-4799- 6896-1/14/$31.00 ©2014 IEEE

[10] Shuang Lin, Jie Chen, Zhendong Niu, "Combining a Segmentation-Like Approach and a Density-Based Approach in Content Extraction", TSINGHUA SCIENCE AND TECHNOLOGY ISSNll1007- 0214ll05/18llpp256-264 Volume 17, Number 3, June 2012

[11] Mr. Satish J. Pusdekar and Prof. Shaikh. Phiroj," Using Visual Clues Concept for Extracting Main Data from Deep Web Pages",IEEE,2014

[12] Tiliang Zhang, Hua Zhang, Fei Gao, "A Malicious Advertising Detection Scheme Based on the Depth of URL Strategy", IEEE,2013