Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 2170 - 2176
Received 15 April 2021; Accepted 05 May 2021.

Malicious URL Prediction Using Machine Learning Techniques

Harsha Vardhan Sai Aalla[1], Nikhil Reddy Dumpala[2], M. Eliazer[3]

[1]Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India.

E-mail:aa7673@srmist.edu.in

[2]Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India.

E-mail: dd7385@srmist.edu.in

[3]Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India.

E-mail:eliazerm@srmist.edu.in

ABSTRACT

The usage of malicious website is a serious security threat faced by users while surfing data in internet. It is offensive and it belonging to criminal activities. Requirement of safeguard activities to help end-user is much needed. The need of understanding about protocol, uniform resource locator (URL) and other features of webpage are non-negligible. The purpose of this work is to findmalicious webpage from lexical and to resolve uncertainties faced by users. The study upon identifying features of websites is vulnerable and how malicious attack will occur is reported. To maximise the accuracy in prediction, machine learning technique is intruded. In recent years Phishing, botnet and malicious threats are more common in internet world and by disguising URL to trust it as non-offensive one.In concentrating with future concern and providing solution to real time problems faced by end-user the proposed work initiate two different algorithms namely decision tree and logistic regression. A total of 420000 webpages are taken as input data which included both affected and legitimate website. The time taken for prediction and accuracy is calculated in the form of testing data set. Thus logistic regression achieved higher efficiency with accuracy of 97.5% in an effective manner.

KEYWORDS

Malicious Website, Uniform Resource Locator (URL), Phishing, Botnet.

Introduction

Malicious website threatening is usage of codes in the form of URL by attackers to collect personal information and unauthorised access of user database. It involves collecting data about passwords enrolled for email, bank account details, and pin number for either credit or debit card and little other necessary information. Attackers may trick user to get their information without sense. As like hacking, this method also takes control over user computer in the form of breaking defence system used in computer. The malicious links are spread through e-mail which having details about organisation, job vacancies, and online purchasing offers and also it looks like legitimate websites. So the user can easily attract much better than what are all the things presented in lexical.

For every year, the arising of phishing content websites count increasing and it is unstoppable. As per the details gathered from banking society, phishing attack of 0.47% raising early once. Attackers feels ease to attack unsuspected users and who do not aware of it. In a following way the attackers exploring their URL without doubt. The popular web page login portal is intimated by attackers to cover users and it totally appears as legitimate webpage. Once unsuspected user visits the link, the script running behind it extracts data and makes use of it by attacker. In figure 1 the steps carried out by attacker to theft information from user is clearly visualized.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 2170 - 2176
Received 15 April 2021; Accepted 05 May 2021.

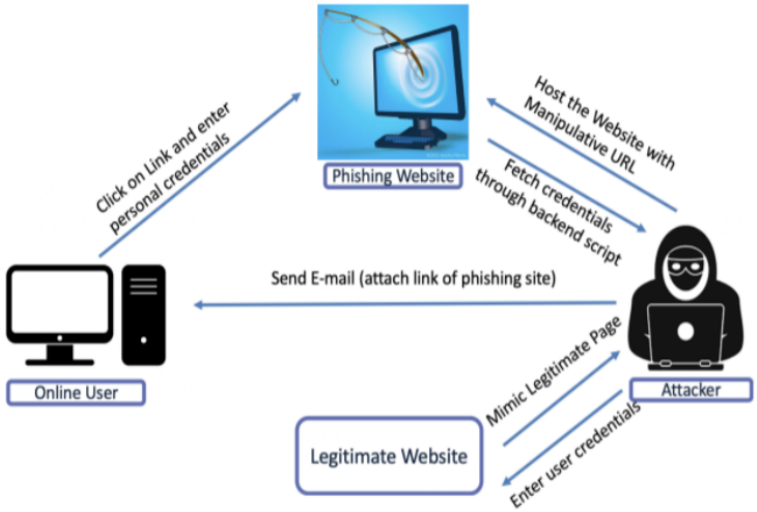

Fig.1. Mechanism behind data theft

It is represented by technical terms in the name of cyber-squatting, typo-squatting respectively. First one is relevant to hacking but it is done in the form of URL. In already existing company name attacker should have cloned domain. Second one having website as same as that of lexical but typographical error present in that.

Then needed for detecting malicious web page and reduce attack over user brings more attention. Machine learning technique taking input of lexical with malicious. From that it would separate good one.

The flow analysis of detecting live websites and process taken into account during prediction is presented it in figure 2.

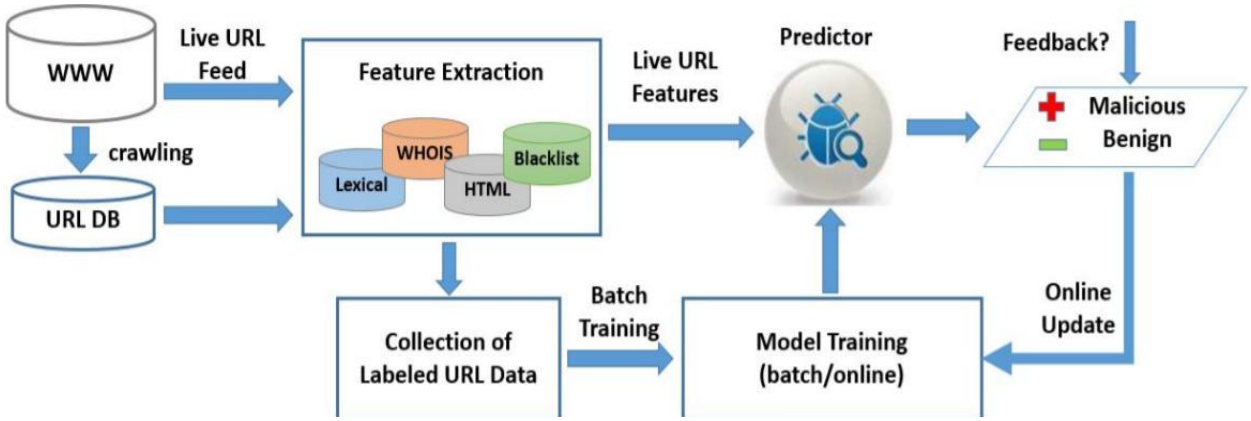

Fig.2. Generalised structure of process involved in prediction

In section II, survey upon different control schemes used in detecting and alert user is explained clearly in section II. The training, testing and process involved in proposed algorithm is presented in section III. The input data set validation and results obtained for randomly chosen websites are discussed in section IV. The impacts of malicious URL and actions taken to handle it with reviewing about best predictor among proposed method are concluded in section V.

Literature Survey

Rohit Verma et.al, reviewing about malicious URL prediction carried out by providing data set in [1].This work regarding physical structure of URL and it does not relevant to content based properties. A supervised learning

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 2170 - 2176
Received 15 April 2021; Accepted 05 May 2021.

algorithm is presented in this paper. Support vector machine (SVM) is trained and tested to detect malicious web page at speed of response is said to be high. Nearly 18 features are chosen for analysis. The classification is categorised by passing URL continuously in machine learning mechanism.

Lekshmi et.al, in [2] discussing about uncertainties in traditional method and reviewing advantages of exploring a technique with the help of machine learning. Blacklist is the oldest and simplest method used to predict malicious URL; but it could not have enough tendencies to detect malicious URL available nowadays. To make it as effective, one among black list method so called heuristic method is also intruded. But not suit for any type of attack. It is well developed and trained to detect few attacks only. To predict it and resolve it in a proper way with high rate of accuracy, machine learning is introduced.Batch learning and online learning are the two algorithms presented to detect malicious attack. Further batch learning adopts two algorithms, one is SVVM and next one is naïve bayes.

Teena Varma et.al, explaining random forest algorithm to resolve issues caused by attackers and prevent use of malicious web page in [3]. Initial stage followed after getting data is data slicing. After that it is diverted to training and testing.If it completes training, accuracy is calculated by doing cross validation. If it satisfies the necessity, then it is moved to treat testing dataset which one we need to test is stored in data set. Then it proceeds to classification. The presented algorithm makes multiple decision trees. The number of trees generated is used to treat different samples taken from a single dataset; it can handle different attributes to create a tree during each time. By completing creation of tree it would start classification in association with result obtained from each tree. Finally make it available to class. These are all the steps carried out in random forest prediction method.

Clayton Johnson et.al in [4] compared differential classifier algorithms namely KNN, random forest and CART. From analysis they representing random forest method are well suit to improve prediction as well as it consumes less time than other methods.

Cho Do Xuan et.al, describing about issues and in-secureness of network information; also evolving several steps to eliminate this circumstance is presented in [5].URL is the major one which is used to surfing data in internet. Some unauthorised persons (i.e, hacker) collecting information about an individual by using several techniques namely phishing, social engineering and pharming. These actions are carried out in the form of URL. Malicious URL is the inoffensive way of affecting user without any contact. In other end, unauthorised person can take action upon user's device especially like malware download, redirecting towards unnecessary websites.Nowadays prediction of malicious URL is much needed one. In recent surveys there are three techniques mostly used by hackers to spread these things in internet namely botnet, malicious and phishing. These are shared as file and message. In past few years, signature set is used to detect whether it is safe or not. Once URL is noted as malicious it should provide alarm signal. Due to some disadvantages and difficulties in that, it is left out.By proposing a URL with collection of its attributes and behaviour, the proposed work is done using machine learning as a key factor. Random forest and SVM are the two supervised algorithms chosen to rectify about discussed issues.Among various familiar algorithms, this method is chosen because of its advantages and requirement in accordance with attributes.

Gold Wejinya et.al in [6] describing about the process and functioning carried out in malicious URL detection by SVM, random forest, naïve bayes. The threats in webpages are pointed out and primitive measures taken to prevent data theft are presented.Jeena et.al, in [7] handled a statistic approach to maximise the accuracy. In a special concern algorithms are chosen and prediction is progressed. SVM and CNN are the two algorithms used in this case study. From above concern an idea about presenting various topologies to predict malicious.

Proposed Methodology

The proposed method follows steps to perform training and testing of algorithms is presented in the form of block diagram in figure 3.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 2170 - 2176
Received 15 April 2021; Accepted 05 May 2021.

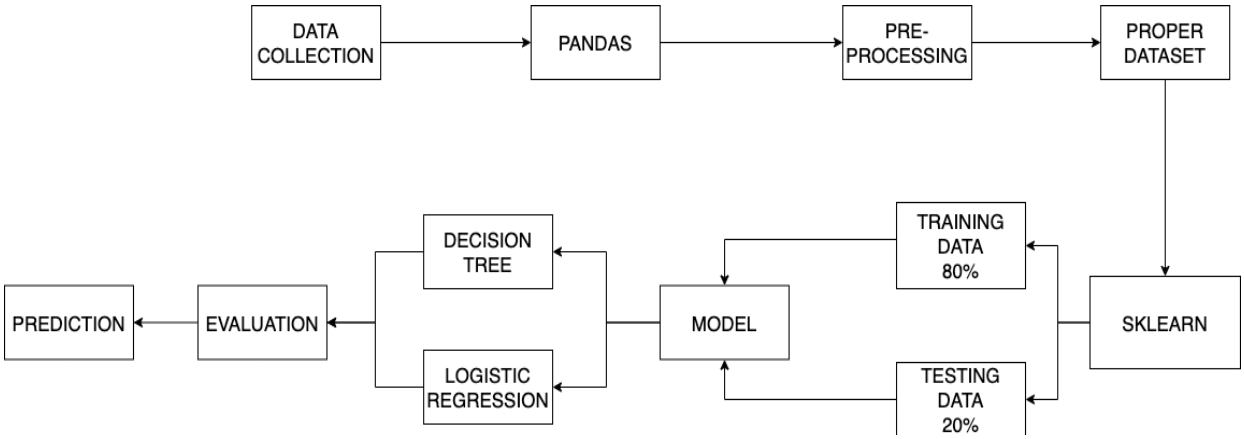

Fig.3. Flow analysis of proposed method

The detailed structure of algorithms used in our proposed work is represented below.

Decision tree: It is a well-known classifier algorithm.It has tree like structure and its each node represents features; branch denotes rule used for taking a decision; leaf node denotes outlet. Node present at top is called as root node. Depending upon attributes, the partition process is carried out and it is named it as recursive partitioning. It provides higher resolution to take action upon different data sets either data should be numerical / categorical. It is the ideal state to process relation between nonlinear attributes and classes. It can have capability to modelling both non-linear and unconventional datasets. It can affect interaction among predictors which enrolled in prediction due to binary structure. It having some drawbacks and also functioning upon new dataset is difficult. At that condition, the time needed to perform prediction is said to be high.

Logistic regression: For discrete set of classes it assigns observations. Using logistic sigmoid function, this transform output is evaluated. Thus a probability function is adapted to N number of discrete classes. It works well when data is linear or otherwise the system possess complex non-linear with variables, the system efficiency will affect. Other than this, this method needs statistical assumption.

Result and Discussion

This section describing about the step by step procedure taken into check URL either it is associated with malicious or not. For this analytical study nearly 420000 URL's are checked using two differential algorithms.

```
In [10]: urls_data
```

Out[10]:

	url	label
0	diaryofagameaddict.com	bad
1	espdesign.com.au	bad
2	iamagameaddict.com	bad
3	kalantzis.net	bad
4	slightlyoffcenter.net	bad
...	...	...
420459	23.227.196.215/	bad
420460	apple-checker.org/	bad
420461	apple-iclods.org/	bad
420462	apple-uptoday.org/	bad
420463	apple-search.info	bad

420464 rows × 2 columns

Fig.4. Describing number of URL and result

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 2170 - 2176
Received 15 April 2021; Accepted 05 May 2021.

The URL numbers are pointed out with status. Only few number of URL visualized having status of bad. In that more than 50000 URL possess malicious attack. Thus remaining websites can have are safe to use and surf data in internet. The following below flow chart describing about number of URL affected with malicious and good one.

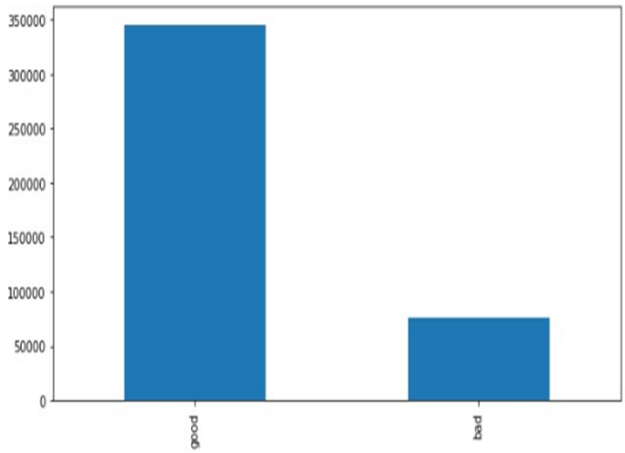

Fig. 5. Flowchart explaining about status of URL

This is the way to prediction process is carried out. The websites are chosen randomly and put it in below tool; it should check the URL with the help of trained algorithm. In below figure 6, for case study purpose randomly chosen URL is pasted it in tool. Then it take several actions upon it and it describes chosen URL is bad. Similarly the same process is carried out in figure 7 also. The following procedure is the easiest way to perform prediction.

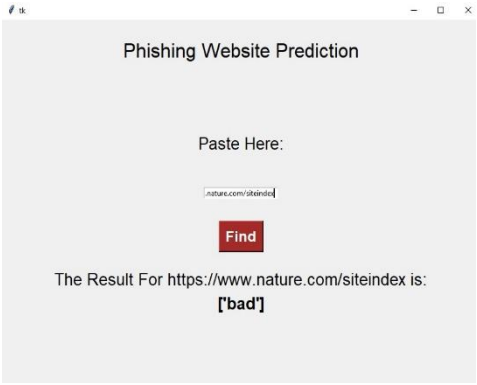

Fig.6. Random analysis 1

Fig.7. Random analysis 2

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 2170 - 2176
Received 15 April 2021; Accepted 05 May 2021.

Other than analysis, accuracy and time consumption is more important. From above facts, the accuracy prediction is carry over for both algorithms. Decision tree possess accuracy of 85% and it is somewhat lesser than conventional algorithms. The time needed for improving prediction process is high. Let move towards logistic regression; it having nearly 97.5% with quick response. Thus logistic regression is the only thing can improve prediction in severe condition also.

```
logit = DecisionTreeClassifier(criterion = "gini", random_state = 100,max_depth=3, min_samples_leaf=5)   #using Deci
logit.fit(X_train, y_train)
```

```
DecisionTreeClassifier(max_depth=3, min_samples_leaf=5, random_state=100)
```

```
print("Accuracy ",logit.score(X_test, y_test))
```

```
Accuracy  0.8582640647854162
```

Fig.8. Presenting accuracy of decision tree

```
LogisticRegression()
```

```
print("Accuracy ",logit.score(X_test, y_test))
```

```
Accuracy  0.9749087319990962
```

Fig.9. Presenting accuracy of logistic regression

Table I. Reviewing about testing and training

Dataset	Predicted safe URL	Accuracy (%)	Precision (%)	Recall (%)	Training time (s)	Testing time (s)
420000 URLs	Logistic regression	97.5	98	95	1.78	0.01
	Decision tree	85	87	81	2.62	0.01

Conclusion

People prefer internet to share their information and it provides more convenient services every day; it prefer more things much needed in day-to-day life.Increase in number of webpages and applications accessed by users in internet are the initial stage for attackers to collect data. Instead of spreading malware it is easy to fetch malicious code in websites for attackers. The rise in malicious URL and its usage attains peak. The URL prediction is considered as critical one in online shopping, online banking, cyber security applications, trading, etc. Thus a promising technique so called machine learning took major part.

The experimental analysis of malicious URL prediction and step by step procedure involved in it is discussed. In proposed method, comparative analysis is carried out in helping the user before losing their privacy and sharing their information without their permission. The systematic rule followed in our analysis is presented with detailed study upon conventional URL prediction.Other than that what are all the algorithms used by others to improve prediction accuracy and reducing time are also reviewed. The future prediction depending upon machine learning can adopt improved prediction taken with logistic regression will be a better choice. Absolutely the way of proposed prediction method can predict well and there is no doubt in it.

With the developing computer and system technologies, people exchange information over the Internet that attracts people due to the convenience of services they offer day by day and beyond that, they do many other things related to daily life. During these processes, users have intelligence and critical information such as descriptive usernames and passwords. Most network applications detect their users with them.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 2170 - 2176
Received 15 April 2021; Accepted 05 May 2021.

The rapid increase of the web pages and applications caused them to become the primary target for the attackers. Today, the number of malicious websites has increased considerably. Malicious behavior of trusted or malicious users threatens network applications. Users who are unaware of anything become a victim only by visiting these harmful pages. Attackers can exploit the web environment more easily by uploading or embedding malicious code on the web page instead of spreading the malware. According to the Google Research Center, over 10% of web pages contain malicious code. Therefore, the detection of harmful web pages has become very important to protect the users of the web environment from these threats. In this respect, determining whether web pages directed to users are used for malicious behavior is of great importance for the institution and individual users to overcome the situation with minimum damage. Recent years have witnessed detecting Malicious URL has a significant role in cybersecurityapplications. Malicious URL has been a severe threat to cybersecurity. Without any questions, CPS can be considered as a crucial step in the development of data-accessing and data-processing services available on the Internet.

References

[1] "Detection of Malicious URLs using Machine Learning Techniques" *Immadisetti Naga Venkata Durga Naveen, Manamohana K, Rohit Verma; (IJITEE)* March, 2019.

[2] "Detecting malicious urls using machine learning techniques: a comparative literature review" Lekshmi A R, Seena Thomas, (IRJET) June 2019.

[3] "Malicious URL Detection using ML" Mrs. Teena Varma, Pratik Zinjad, Shreeniket Vast, Idris Vohra, A.Hannan Sunsara, *(IRJET)* May 2020.

[4] "Towards Detecting and Classifying Malicious URLs Using Deep Learning" Clayton Johnson, Bishal Khadka, Ram B. Basnet, and Tenzin Doleck, *Journal of Wireless Mobile Networks*, Dec. 2020.

[5] "Malicious URL Detection based on Machine Learning" Cho Do Xuan, HoaDinh Nguyen, Tisenko Victor Nikolaevich, (IJACSA), Vol. 11, No. 1, 2020.

[6] "Machine Learning for Malicious URL Detection" Gold Wejinya, Sajal Bhatia, December 2020, (AISC)

[7] "Malicious URL Detection Using Machine Learning Techniques" R.Jeena, G.Preethi, A.Praveena, A.Preethi, *IJIRSET*, 2019.

[8] "Detecting malicious URLs using machine learning techniques" Frank Vanhoenshoven, G. Nápoles, M. Köppen; 2016 (SSCI).

[9] "Classification of URL into Malicious or Benign using Machine Learning Approach" Deebanchakkarawarthi G, Parthan AS, Sachin Lal, Surya A, *IJARCCE*, February 2019.

[10] "Exploring Malware Behavior of Webpages Using Machine Learning Technique: An Empirical Study" AlhanoofFaizAlwaghid and Nurul I. Sarkar, Electronics 2020.