Fuzzy Systems and Soft Computing
ISSN : 1819-4362

PHISHING ATTACK DETECTION USING DECISION TREE ALGORITHM

Dr. A. Vinoth, Assistant Professor, Department of Information Technology, Sri Krishna Adithya College of Arts & Science, Coimbatore.
Akil S, Ashik Ahamed A, III B. Sc.IT, Department of Information Technology, Sri Krishna Adithya College of Arts & Science, Coimbatore.

Abstract

Attackers and other cybercriminals are making the internet hazardous as the majority of human activities shift online by posing a severe risk to customers and businesses, endangering global security, and undermining the economy. Nowadays, phishes are always coming up with fresh ways to trick users into disclosing their private data. It is crucial to build phishing detection algorithms in order to prevent falling prey to online crooks. For phishing detection, machine learning or data mining techniques are utilised, such as classification that divides online users into dangerous or safe users, or regression that forecasts the likelihood of being attacked by some online criminals in a specific time frame. In the past, a number of solutions for phishing detection have been put out, but the search for a better solution is still ongoing due to the dynamic nature of some of the numerous phishing schemes used by cybercriminals. In this project, we can put the framework for phishing website classification using machine learning techniques like decision trees into practise. Applied using benchmark datasets that are gathered from KAGGLE websites, experimental findings demonstrate that the suggested method offers better accuracy rate compared to the current techniques.

KEYWORDS: KAGGLE,PII, URL, IP, DNS

I. INTRODUCTION

After a hacker attack on an America Online account in 1996, the term "phishing" was first used to describe a type of online identity theft. The first phishing lawsuit was brought in 2004 against a California teenager who imitated the America Online website in order to obtain sensitive user data, including credit card information, which resulted in significant financial losses for the victims.

Phishing is a type of online fraud that includes impersonating a trusted source through email, instant messaging, and other communication channels in order to get private information such as usernames, passwords, and account information. Since the fraudsters are dynamic in their operations and continually alter their activities to evade any form of detection, the old methodologies employed by the majority of email filters for recognizing these emails are weak to deal with the newest growing patterns of phishing.

Phishing scams involve sending fake emails to victims under the guise of legitimate and well-known institutions like banks, universities, communication networks, etc., instructing them to update some personal information, such as their passwords and usernames, in order to maintain access to certain services offered by the institution. Phishing uses this method to collect users' sensitive information, which is then used to gain access to their vital accounts and cause identity theft and financial loss.

Phishing is primarily thought of as a criminal action that is carried out by an invader or hacker using technological deceptions and social engineering that are well-versed. The only goal is to steal the customers' or users' financial information and sensitive PII (Personal Identity Information). To deceive the beneficiaries into disclosing their financial information, the social engineering method uses baiting and counterfeit communications. The received mail gives the impression that it is from a reliable user or business, which leads the beneficiary to believe that it is authentic.

Another intriguing detail is the technological ruse that involves installing harmful software on the user's computer system in order to obtain their personal login information. As a result, this form of data leak has emerged as one of the biggest internet security threats. In order to defend and preserve computer systems from intruders or hackers, cyber security measures alter and develop daily. On the other hand, as security measures grow, hackers and invaders also build ever-more sophisticated security measures to get around the corresponding protected systems.

The cyber security systems evaluate numerous data patterns with the aid of machine learning techniques and afterwards learn from the training and analysis. The aforementioned analysis and training help the cyber security systems thwart similar online assaults.

PROBLEM IDENTIFICATION

When conducting a Web phishing assault, the attacker creates phishing websites that mimic real websites in an effort to trick people into disclosing their sensitive financial and personal information. Clicking a link in an email is how the phishing assault is originally launched. In order to update or verify their information, victims receive an email with a link. The Web browser will take the target users to a phishing website that looks just like the real website if they click on this link. Due to the fact that phishing websites urge users to submit critical information, attackers can subsequently steal the essential data from the users.

As soon as phishing commences, the attackers can eventually commit money theft. Web phishing assaults must be stopped in their early phases since phishing websites will inevitably target online companies, banks, users of the Internet, and the government. Due to the numerous cutting-edge techniques phishing attackers employ to trick online consumers, identifying a phishing website is a difficult undertaking. The ability to reliably and quickly identify phishing websites is crucial for the effectiveness of phishing website detection systems. There are many traditional methods for identifying phishing websites that are based on established black and white listing databases. These methods are ineffective, though, as a new website may be established in a matter of seconds. As a result, the majority of these solutions cannot accurately determine if a new website is phishing or not on the fly. As a result, many recent phishing websites might be considered trustworthy websites.

II. LITERATURE STUDY

System analysis will be performed to determine if it is flexible to design information based on policies and plans of organization and on user requirements and to eliminate the weakness of present system. This chapter discusses the existing system, proposed system and highlights of the system requirements.

Identifying fake/phishing URLs presents a significant challenge in cyber security. These malicious links can employ various evasion tactics, including URL obfuscation, URL shortening, and mimicking legitimate domains, making them difficult to detect using traditional methods. Furthermore, the dynamic nature of malicious content, localization, and the constant evolution of attack techniques create a moving target for URL detection systems.

False positives can erode trust in these systems, and multi-step attacks further complicate identification. Additionally, URL detectors may not always account for user behavior or context, which is crucial in distinguishing between genuine and fake/phishing URLs. To address these issues, cyber security professionals rely on a combination of techniques, including machine learning, threat intelligence, and user education, to enhance the accuracy and effectiveness of fake/phishing URL detection.

Disadvantages of Existing System

- Time complexity is high
- Computational problems can be identified

False positive rate can be occurred

III. DEVELOPMENT OF PHISHING ATTACK DETECTION USING DECISION TREE ALGORITHM

Due to its constant evolution and the billions of cash it has syphoned off of governments, businesses, and everyday people, phishing has long been a challenging menace in every culture. It is identity theft that makes use of a specific type of social engineering assault to get crucial information from a person or group of people. We examine numerous aspects used in various phishing attempts in this essay. An improved form of trees first offered is the decision tree. It is typically utilized in classification tasks, where it serves as a classifier to translate an input pattern into a certain class.

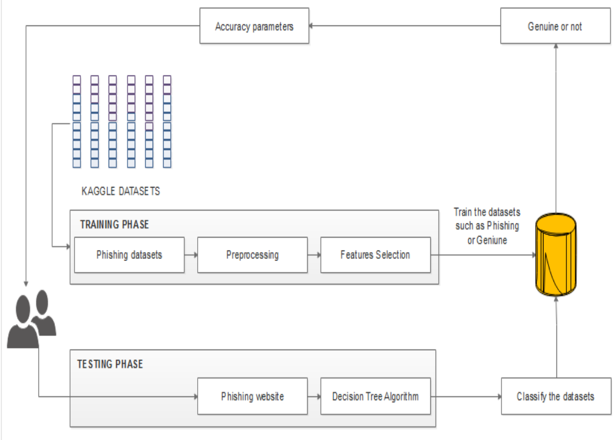
In order to enhance the performance of decision trees, a new supervised learning method uses a technique known as. When compared to the conventional implementations, it has a lot of strength. Its advantages include improved regularization capabilities that decrease over fitting, fast speed and performance since trees are produced in parallel, flexibility because of the customization of its optimization aims and assessment criteria, and built-in procedures for managing missing information.

Decision Tree is a great tool for many researchers in data science and machine learning due to these and many more benefits. A few of the researchers used this method. This algorithm uses a way to aggregate trees and is based on a tree model. The target variable will be predicted using Decision Tree repeatedly using the training data until the model's parameters is optimized.

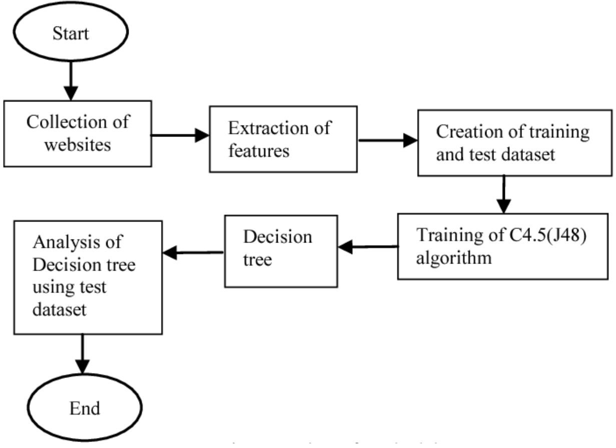
Advantages:

- Improved accuracy in detection
- Time complexity can be reduced
- Analyze unsupervised datasets
- Computational complexity can be reduced

System Architecture



System Flow Diagram


Start
Collection of websites
Extraction of features
Creation of training and test dataset
Training of C4.5(J48) algorithm
Decision tree
Analysis of Decision tree using test dataset
End

MODULE DESIGN

Datasets acquisition

The internet today has a huge selection of electronic resources, many of which include information of a high calibre. The internet typically offers more information than is required, though. The user needs to choose the best data set in the shortest amount of time for a certain information need. One use of information retrieval is machine learning, which involves condensing the input text into a shorter version while maintaining its information content and overall meaning. The summary of documents based on a query using a similarity metric has received a great deal of attention. This module consists of the upload of any common csv file. You may compile a sizable number of phishing datasets in this module. We may upload user-submitted datasets and kaggle datasets in this module. A data set is a collection of data, albeit this form is less common in modern dictionaries nowadays. Each variable's values are listed in the data set, including the IP address, URL, domain, DNS records, mouse click events, and label values.

Preprocessing

The [data mining] method includes a crucial phase called data pre-processing. The techniques used to collect data are frequently not tightly regulated, which leads to out-of-range numbers, impossible data combinations, missing values, etc. Data analysis that has not been thoroughly checked for these issues may yield false findings. Thus, before doing an analysis, the representation and quality of the data come first. Knowledge discovery during the training phase is more challenging if there is a lot of redundant, irrelevant information available or noisy data. The tasks involved in data preparation and filtering might take a long time to process. In this module, we may estimate the missing values of the data as well as remove the unnecessary values. Offer organized datasets in the end.

Features selection

The process of limiting the inputs for processing and analysis, or of locating the most significant inputs, is referred to as feature selection. The practise of removing relevant information or features from existing data is referred to as feature engineering (or feature extraction) in a related phrase. Utilizing statistical analysis, filter feature selection techniques score each feature. The characteristics are sorted according to the score and either chosen to be preserved in the dataset or rejected. It may be applied to create data from numerous time series. Choose the various characteristics from the uploaded datasets for this module. Then develop feature values for the datasets based on relevance ratings.

Classification

as a supervised learning task, detecting phishing involves predicting a target variable using training data (xi). Generally speaking, compared to certain other classification and regression approaches, tree-based models don't do as well. Nonetheless, the prediction ability of trees may be greatly enhanced by merging several trees using a technique called boosting. Using the boosting approach, decision tree is a tree-based model that aggregates trees. Using decision tree, we iteratively predicted the target variable yi using the training data xi until the model's parameters were optimized.

Phishing detection

This module helps to detecting phishing website with the pre-trained dataset and machine learning model with decision tree algorithm and provides result approximately.

IV. RESULTS AND DISCUSSION

The implementation phase focuses how the engineer attempts to develop the system. It also deals with how data are to be structured, how procedural details are to be implemented, how interfaces are characterized, how the design will be translated into programming and hoe the testing will be performed. The methods applied during the development phase system e will vary but three specific technical tasks should always occur.

- ➢ The software design
- ➢ Code generation
- ➢ Software testing

The system group has changed with responsibility to develop a new system to meet requirements and design and development of new information system. The source of these study facts is variety of users at all level throughout the organization.

Stage of Development of a System

- ➢ Feasibility assessment
- ➢ Requirement analysis
- ➢ External assessment
- ➢ Architectural design
- ➢ Detailed design
- ➢ Coding
- ➢ Debugging
- ➢ Maintenance

Feasibility Assessment

In Feasibility this stage problem was defined. Criteria for choosing solution were developed, proposed possible solution, estimated costs and benefits of the system and recommended the course of action to be taken.

Requirement Analysis

During requirement analysis high-level requirement like the capabilities of the system must provide in order to solve a problem. Function requirements, performance requirements for the hardware specified during the initial planning were elaborated and made more specific in order to characterize features and the proposed system will incorporate.

External Design

External design of any software development involves conceiving, planning out and specifying the externally observable characteristic of the software product. These characteristics include user displays, report formats, external data source and data links and the functional characteristics.

Internal Design Architectural and Detailed Design

Internal design involved conceiving, planning out and specifying the internal structure and processing details in order to record the design decisions and to be able to indicate why certain alternations were chosen in preference to others. These phases also include elaboration of the test plans and provide blue prints of implementation, testing and maintenance activities. The product of internal design is architectural structure specification.

The work products of internal design are architectural structure specification, the details of the algorithm, data structure and test plan. In architectural design the conceptual view is refined.

Detailed Design

Detailed design involved specifying the algorithmic details concerned with data representation, interconnections among data structures and packaging of the software product. This phase emphasizes more on semantic issues and less synthetic details.

Coding

This phase involves actual programming, i.e, transacting detailed design into source code using appropriate programming language.

Debugging

This stage was related with removing errors from programs and making them completely error free.

Maintenance

During this stage the systems are loaded and put into use. They also get modified accordingly to the requirements of the user. These modifications included making enhancements to system and removing problems.

V.CONCLUSION AND FUTURE ENHANCEMENT

The rapid increase in usage of the Internet and web services has led to a drastic increase in number of web attacks. Phishing is a web attack where the phisers try to acquire user's sensitive information for fraudulent purposes. Phishers target user's sensitive information through a fake website that appears similar to a legitimate site in terms of interface and URL address. Hence, there is an increase in victims falling prey to the phishing sites. In this project, we have clearly shown and demonstrated the boost in accuracy of phishing detection by a careful selection of only some features using Decision Tree algorithm.

This research contributes to knowledge by boosting the accuracy that outperforms other widely known algorithms like: Decision tree algorithm trained on the same dataset with the same evaluation criteria for fairness; the most widely used algorithms use 17 features which is standardized for detecting phishing for detection but in this work we have demonstrated that some features are absolutely useless and hampers on the accuracy of the detection and also slows down the detection process. The method helps us weed out those useless features and only uses the important features to boost the accuracy of detection

Scope for Future Enhancement

The recent achievements of deep learning techniques in complex natural language processing tasks make them a promising solution for phishing website detection too. Future work proposes a novel hybrid deep learning model that combines convolutional and recurrent neural networks for fake news classification. And also extend the framework to implement various deep learning algorithms to improve the accuracy and reduce the complexity in classification to analyse phishing images, audio or video

VI. REFERENCES

BOOK REFERENCES

1. Van Rossum, Guido, and Fred L. Drake. The python language reference manual. Network Theory Ltd., 2011.
2. Van Rossum, Guido, and Fred L. Drake. The python language reference manual. Network Theory Ltd., 2011.
3. Dierbach, Charles. Introduction to Computer Science using Python: A Computational Problem-Solving Focus. Wiley Publishing, 2012.
4. James, Mike. Programmer's Python: Everything is an Object Something Completely Different. I/O Press, 2018.
5. Reges, Stuart, Marty Stepp, and Allison Obourn. Building Python Programs. Pearson, 2018.

WEBSITE REFERENCES

1. https://docs.python.org/3/tutorial/
2. https://www.w3schools.com/python/
3. https://www.tutorialspoint.com/python/index.htm
4. https://www.programiz.com/python-programming
5. https://www.learnpython.org/