IJARCCE

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

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 8, August 2016

Phishing Detection using Map-reduce and PART Algorithm

Mr. Rakshith Raj K. R.[1], Prof. Prabhakara B. K[2]

PG Student, Computer Science and Engineering, Sahyadri College of Engineering, Mangaluru, India [1]

Professor, Computer Science and Engineering, Sahyadri College of Engineering, Mangaluru, India[2]

Abstract: Phishing is the fraudulent activity that is done by the phishers, in order to gain information of users such as their user IDs, passwords and credit card details through online. The users will be victim for this kind of activities, because phishing web pages looks very similar to real ones, so finds difficult to distinguish between the fake website and ones, detecting this kind of webpage is very difficult because for identification it takes several attributes into consideration which user might not knowing those things. The existing phishing detection systems are highly dependent on database and they are very time consuming also. In this proposed system, Hadoop-Map Reduce is used for fast retrieval of URL attributes, which plays a major role in identifying phishing web pages and it is known for its time efficiency and throughput also can gained using this. The PART algorithm is used for classifying and predicting the phished pages, which is more efficient and accurate than the algorithms used in existing systems. The main goal is to provide security to the user's data while browsing.

Key words: Phishing, Anti-Phishing, Hadoop, Map Reduce, Information Retrieval, Data Mining.

I. INTRODUCTION

Phishing is similar to fishing in a lake, but instead of trying to capture fish, phishers attempt to steal the personal information of the user. Phishers create fake web pages that are meant to steal individual's information without brining notice to the user. Unknowingly he will give his information to the phishers.

The Anti-Phishing Working Group [15] reported that there were almost 40,000 phishing attacks between January 2014 to and March 31, 2014. Almost 95% of users do not know about phishing and becoming victims for Phishing. No Anti-Phishing techniques are able control the Phishing activity completely. There is rise in the phishing activity in recent years. The recent statistics tells how the phishing is a major threat to the personnel data and loss of money due to unknowingly becoming victim for the phishing. To tackle phishing many of the software have been developed and deployed in the system. Industrial toolbar based anti-phishing, user interface based anti-phishing, and web content based anti-phishing is the anti-phishing techniques to avoid phishing. But none of the technique has been succeeded in finding complete solution to the phishing attacks. Till today so many techniques are used for detecting the phishing websites mainly include authentication, filtering, attack tracing and analyzing, phishing report generating, and network law enforcement. There are many algorithms of data mining, machine learning, image processing are used for detecting the phishing algorithms effectively.

Hadoop [13] is a free, Java-based programming framework that supports the processing of large data sets in a distributed computing environment. It is a framework for distributed processing of large datasets across clusters of computers, it can scale up to thousands of machines, where each machine offers computation and data storage. Hadoop distributes vast amount data among several hundreds of nodes for processing. It is distributed file system that allows data transfer among nodes very quickly and incase of any node failure it will not affect the operation. Even though many number of node becomes non-responsive, it does not make any change in the operation, this quality of Hadoop reduces the risk of catastrophic system failure significantly.

Hadoop was an idea developed from Google's MapReduce, a software framework which breaks large set of data into numerous small parts. Map takes a set of data and converts it into another set of data, where individual elements are broken down into key/value pairs. Secondly, reduce task, which takes the output from a map as an input and combines those data tuples into a smaller set of Key/Value pairs or tuples. These tuples can run on any node in the cluster. The Hadoop Distributed File System (HDFS) is designed to store very large data sets reliably, and to stream those data sets at high bandwidth to user applications. In a large cluster, thousands of host directly attached to storage and execute user application tasks.

In this proposed approach, the system is faster in detecting the phished webpages than the available existing anti-phishing techniques. Here in proposed system MapReduce is used for URL's information collection through various nodes in distributed environment which will be faster. And it is an open-source software program for storage and

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 8, August 2016

large-scale processing of datasets on clusters of community hardware. The PART algorithm will predict the authenticity of the URL based on the data given by the Hadoop MapReduce.

II. RELATED WORKS

The system uses a crawler [1] method for detecting the URL's in the database where already existed and crawls webpage information then given to Map Reduce for predicting the authenticity. The Map Reduce technique improves the performance of the anti-phishing technique. But it requires blacklisted data. EMD (Earth Movers' Distance) [2] method takes the web pages and it compresses the image for reducing the intensity. The Multidimensional distributions are compressed into a fixed number of bins, which is of a predefined size, results in a histogram. The histogram is used for comparing the datasets against the existing datasets. It is an efficient method, requires huge amount of image compression technique. The textual and visual classification [3] is the technique where text and images is fused using Bayesian algorithm and entire document is parsed and then examined based on the dataset. This is a lengthy process because it needs entire search of a document.

Visual Cryptography [5] is the image based authentication, where image captcha is decomposed into two shares. They are stored in separate database such that original captcha available only when both the shares are available together. Once the user gets original image captcha he can use that as password. This avoids machine based intrusion, but not suitable for human intruder much effectively.

DNS-based anti-phishing approach [2] is a technique where it is mainly based on blacklists, heuristic detection. But they do have some shortcomings. Blacklist is a DNS based anti-phishing approach is highly used technique by the browser. When user browsing the phishing sites, Internet Explorer7, Netscape Browser8.1, and Google Safe Browsing are the browsers gives the alert message to user. The blacklist must be updated each and every time for the proper result.

Heuristic-based anti-phishing technique is to estimate whether a page has some phishing heuristics characteristics. For example, some heuristics characteristics used by the SpoofGuard [3] toolbar include checking the host name, checking the URL for common spoofing techniques, and checking against previously seen images. If you only use the Heuristic-based technique, the accuracy is not enough. Besides, phishers can use some strategies to avoid such detection rules. The user may be deceived by the phishing website because the phishing website imitates a legitimate website. Its pages are often similar with the legitimate sites. Therefore, some researchers proposed a similarity assessment method to detect phishing sites.

The image is decomposed into shares which expose the original image after decryption [6]. Anti-phishing working group (APWG) generates and reports monthly about the current phishing attacks. It also has a large number of partners which has made huge contributions in this work. Phishtank.com [7] corpus is updated by various users who report phishing websites. This data is available for free for developers developing applications. E-mail server based technique extract all the suspected URLs from inbox as well as from spams and examine them as most attacked users are from phishing E-mails.

III.PROPOSED SYSTEM

The proposed system makes use MapReduce technique for attribute generation of the URL. The MapReduce will extract from attribute of the URL from different nodes through the network. This attributes are given as input to the PART algorithm. Training Data from Phish tank is given to the Classifier model. The classifier model will generate some set of rules with the help of PART algorithm. The output of the classifier model and attribute generated by the Map Reduce will be given to the Prediction model. The prediction model will generate the result whether web page is normal or phished page.

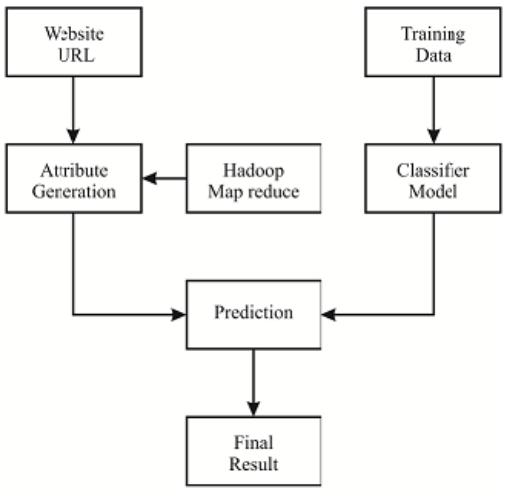

Fig 1 System Architecture

The Fig.1 shows the system architecture consists of:

Website URL: This is the URL of the web page the user visits through his device.
Attribute Generator: There are some attributes which has major role in defining authenticity of the URL.
Hadoop Map Reduce: Apache Hadoop is free software framework for storing and processing of large datasets on clusters.
Training Data: It is the dataset from the phishtank.com parameters of URL's to distinguish the authentic webpage from phished webpage.
Classifier Model: This module is used for classifying the dataset and generating some set rules for finding the phished webpage.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 8, August 2016

Prediction Module: Prediction module makes the decision based on the input it gets from the attribute generator & classifier model and result will displayed to the user about authenticity of the web page.

IV. MODULES

A. Attribute Generation

In this system, Attribute generation is the important aspect in finding the authenticity of the URL. The attributes which are taken into consideration are IP address in the domain of the URL, @ symbol in the domain, shortened URL, Long URL, Presence of SSL Certificate, Redirect Pages etc. The attributes are obtained by the Hadoop Map Reduce. The Map Reduce will collect all the information from the distributed environment and will give a value for those obtained attributes. Computation time required for the Map Reduce is less. This attributes are then given to the Classifier model.

B. classifier model

The training data are generated based on the phishtank.com database. The phishtank.com will have the all reported phishing websites are being analysed and based on that, dataset is prepared. This trained data will be given to Classifier model, which will generate set rules for that attributes with the help of PART (Projective Adaptive Resonance Theory) algorithm.

C. Prediction Module

The Prediction model takes the attribute generation and classifier model as input. Based on the input it gets from the attribute generation and classifier model it will make the decision. The training data is the dataset from phishtank.com and classifier model generated a set of rules using Data Mining algorithm called PART Algorithm. The attribute generation values are analyzed with the dataset of classifier model. Using the attribute generation and classifier model as input prediction model will make use of PART algorithm for generating the result.

V. EXPECTED OUTCOME

Using the Hadoop Map Reduce is going to speed up the processing speed of the phishing detection system. Since the Hadoop Map Reduce is working on the distributed environment attribute generation will be very speed, so outcome will be faster than other existing system. The PART algorithm is the efficient machine learning and Data Mining Algorithm in giving solution, with the help of dataset from phishtank.com the PART algorithm predicts the result efficiently. This system is able in finding newborn phishing sites.

VI. CONCLUSIONS AND FUTURE WORK

Main aim is to make the system to speed up the system from existing anti-phishing system. Using the Hadoop MapReduce will increase the throughput and speed up the processing. The usage of PART algorithm will help in detecting the phishing websites efficiently.

The Hadoop MapReduce can be replaced by other framework Spark, which is more advanced than the MapReduce. This spark would deploy in the system for collecting the attribute values. This is more advanced and faster than the MapReduce, so it will helps in increasing the execution speed. Hence authentic websites and safe browsing can be provided for user.

REFERENCES

[1] Mayura G, Derick S, Abhishek P, Anush S" Antiphishing Using Hadoop Framework", Feb. 04 – 06, 2015.
[2] Yu Tang, Leung Hou U, Yilun Cai, Nikos Mamoulis, Reynold Cheng, "Earth Mover's Distance Based Similarity Search At Scale", 2013 .
[3] Zhang H, Liu G, Chow TW, Liu W "Textual And Visual Content-Based Antiphishing: A Bayesian Approach", Oct 2011
[4] Pradeepa A., Dr. Antony Thanamani "Significant Trends of Big Data Analytics in Social Network", August 2013.
[5] Divya James and Mintu Philip, "A Novel Anti Phishing framework Based On Visual Cryptography", International Journal of Distributed and Parallel Systems, January 2012.
[6] K. A. Aravind, R. MuthuVenkataKrishnan "Anti-Phishing Framework for Banking Based on Visual Cryptography", JCSMA, Jan 2014.
[7] http://www.phishtank.com - "Join the fighting against phishing"
[8] http://www.netcraft.com/anti-phishing/.
[9] Jyoti Chakra, Ritu Dahiya, Neha Garg, Monica Rani "Phishing And Anti-Phishing Techniques: Case Study", IJARCSSE, May 2013
[10] http://www.michael-noll.com/tutorials/running-hadoop-On-ubuntu-linux-single-node-cluster/ -"Running Hadoop on Ubuntu Linux (Single-Node Cluster)".
[11] docs.apwg.org/reports/apwg_trends_report_q1_2014.pdf
[12] Prasad Baitule, Swapnil Deshpande "A survey On Efficient Anti Phishing Method Based on Visual Cryptography Using Cloud Technique by Smart Phones", ICAET, 2014.
[13] http://www.aosabook.org/en/hdfs.html
[14] Yue Zhang, Serge Egelman, Lorrie Cranor, Jason Hong "Phinding Phish: Evaluating Anti-Phishing Tools", 2006..
[15] http://www.antiphishing.org/apwg-news-center/data-logistics/