IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-6, June 2018 E-ISSN: 2347-2693

Prevention of Phishing Attack using Hybrid Blacklist Recommendation Algorithm

Pritesh Saklecha[1*], Jagdish Raikwar[2]

[1] Computer Science, Institute of Engineering and Technology, DAVV, Indore, India
[2] Information Technology, Institute of Engineering and Technology, DAVV, Indore, India

[*]*Corresponding Author: saklecha.pritesh@gmail.com, Tel.: +917869899272*

Available online at: www.ijcseonline.org

Accepted: 07/Jun/2018, Published: 30/Jun/2018

***Abstract*—** This is the era of high end technologies which requires faster connectivity through internet and its variety of applications with prime concern of serving ease in transition, convey messages and data from one end of the world to another. These methods are consuming various sensitive information for their own record. Loose design formation and lower coupling with security approaches of web applications are waved of by attackers to get the system access from malicious activities which create the trouble in real life situations. The primary aim of this presented work is to study about various phishing techniques and their effects on our daily life additionally finding some acceptable and/ or adoptable detection and prevention techniques by which system automatically detects a phishing web URL uses data mining techniques. Along with the studying the work had also identified the problems associated with the current detection. As far as older systems are concerned detection are having larger ratio of false positive nature served with static patterns and rules. This work proposes a hybrid anti-phishing approach using some of the well-known phishing detection factors like MAC address of web pages. This works also elaborates the comparative study of most implemented recommendations algorithms to proposed Hybrid recommendations approach.

***Keywords*—** Blacklisting Recommendation System, Content Based, Collaborative Based, Knowledge Based System, Phishing, Pattern Analysis, MAC, Pattern Similarity Index (PSI).

I. INTRODUCTION

Phishing is the act of identifying the information related with the malicious sources such as names, password and credit card details for getting the access to the system by personating their trustworthy entities of electronic medium. Communications maintained by popular social web sites, auction websites, online payment processors or IT administrators are usually used to entice unpredictable public. The phishing source emails template contains the direct or indirect links to the infected sources and websites. Phishing is usually done by e-mail spoofing or instant messaging, and it often directs users to enter details on the fake website, whose format and experience is almost identical to legitimate. Phishing is a model of social building procedures used to mislead clients and adventures the poor ease of use of current web security innovations. Attempts to deal with the increasing number of reporting fishing incidents include law, user training, public awareness, and technical security measures. The Internet community has devoted tremendous efforts into creating defensive measures to fight phishing attacks. However, the problems constantly progressing and becomes more sophisticated since new online tricks are appearing on a regular basis. Therefore, an intelligent antiphishing solution based around computational and machine learning techniques is needed to differentiate among websites types. The number of available features that can be linked to a website or an email ismassive [3, 4].These features are associated with certain website's elements such as the URL, domain, and source code. One primary challenge in minimizing the phishing risk is to identify the smallest set of features before intelligently classifying the website as phishy or legitimate [5]. Networks can be private, such as within a company, and others that may be open for public access. Operational security, security and supervision on such networks being managed by the network administrator.

Large number of network attacker, hackers, and other unauthorized users effort technically to manipulate the normal behavior of network connections and attempt to expose, destroy, disable, alter, steal or gain unauthorized access to or make unauthorized use of an asset. Network administrator plays a vital role to achieve various security goals: data confidentiality, data integrity, and system availability. Data confidentiality means keep data and communication secret and privacy of personal financial/health records. Data integrity means protecting the reliability of data against tampering and make sure that people cannot change information they should not. Data and resources should be accessible and available when they needed and protection data against denial of service attacks are the meaning of system availability. For achieving these goal network administrator faces so many problems that how to predict the network attack, malicious activity, when the attacker attack, and how to pretend this type of activity. Researchers have done a huge amount of work in this area. Large numbers of approaches are proposed and continually proposed new approaches. Network administrators and researchers have a choice to collect

the previous attack logs, analyze them and create a list of such activities and block traffic from such sources. But this is a cumbersome task for a network administrator, because the data size is very large, continue changing and dynamic nature. Also administrator needs to rebuild such database routinely to prevent a network attack.

Rest of the paper is organized as follows, Section I contains the introduction of Blacklist Recommendation System , Section II contains the literature survey Section III presents the Problem definition behind the current study, Section IV explain Proposed solution through various steps involved, section V gives the result evaluation parameter Section VI concludes research work with future directions.

II. LITERATURE SURVEY

Phishing is a type of social nature in which an attacker, otherwise called a malicious user, endeavors to falsely recover honest to goodness clients' classified or delicate qualifications by imitating electronic communications from a trustworthy or public governance in an automated fashion. A complete phishing attack involves three roles of phishers. Firstly, mailers sends out a large number of fraudulent emails (usually through botnets), which direct users to fraudulent websites. Secondly, collectors set up fraudulent websites (commonly hosted on compromised machines), which actively prompt users to supply confidential information. Finally, cashers use the confidential information to achieve a payout. Monetary exchanges often occur between those fishers. Phishing has spread beyond email to include VOIP, SMS, instant messaging, social networking sites, and even multiplayer games. Below are some major categories of phishing.

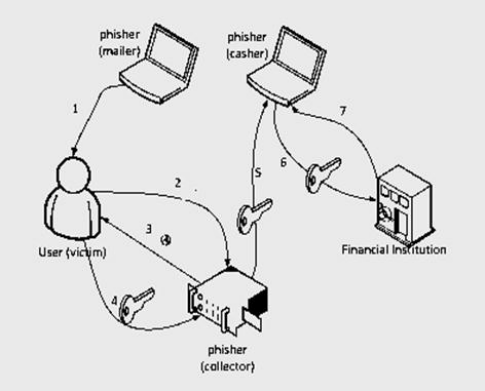

Figure 1 shows the phishing

The method which is commonly used is blacklist-based feature that lists all reported phishing URLs. This feature can be categorized into two types: i) URL blacklisting features, and ii) behavioural blacklisting features. Behavior based features keeps track of the sensitive information the attacker poses and what the user enters into the web forms. In this paper [2], authors study the problem of forecasting attack sources based on past attack logs from several contributors. Formulate this problem as an implicit recommendation system, and propose a multi-level prediction model to solve it. This model evaluates and combines various factors, that is: (i) Attacker-afflicted history using time-series, (ii) Victims interaction using neighbors and / or neighborhood models (iii) global patterns using singular value decomposition. Current research work evaluate combined method, referred to as Blacklisting Recommendation System (or BRS), on one month of logs from Dshield, and demonstrate that it improves significantly the prediction rate over state-of-the-art approaches as well as the robustness beside poisoning attacks. Along the way, evaluate the Dshield dataset, and reveal dominant designs of malicious traffic [2]. In this paper [4], customer preferences for products are drifting over time. Product perception and popularity are constantly changing as new selection emerges. This chapter has introduced the main data mining methods and techniques that can be applied in the design of a RS. This works have also surveyed their use in the literature and provided some rough guidelines on how and where they can be applied. This work mainly started by reviewing techniques that can be applied in the pre-processing step. First, there is the choice of an appropriate distance measure, which is reviewed in Section 2.2.1. This is required by most of the methods in the following steps. The cosine similarity and Pearson correlation are commonly accepted as the best choice. Although there have been many efforts devoted to improving these distance measures, recent works seem to report that the choice of a distance function does not play such an important role [4]. In this paper [5], Neighborhood-based algorithms are frequently used modules of recommender systems. Generally, the likelihood of the similarity measurement used to evaluate neighboring relationships for the success of such an approach is important .In this article authors propose a way to calculate similarities by formulating a regression problem which enables to extract the similarities from the data in a problem-specific way. Another popular approach to the recommendation system is the regular matrix factorization (RMF). Authors present an algorithm neighborhood-aware matrix strategy that efficiently includes neighborhood information in the RMF model.This leads to increased prediction accuracy. The proposed methods are tested on the Netix dataset[5]. In this paper [7], the authors were given brief reviews of various recommendation systems algorithms, which have been proposed in recent literature. First, authors were present the basic recommender systems challenges and problems. Then, give an overview of association rules, memory-based, model-based and hybrid recommendation algorithms. Finally, evaluation metrics to measure the performance of those systems were being discussed. In this paper, author discussed how recommendation algorithms could be evaluated in order to select the best algorithm from a set of candidates. This is an important step in the research attempt to find better algorithms, as well as in application design where a designer chooses an existing algorithm for their application. As such, many evaluation metrics have been used for algorithm

selection in the past. [7]. In this paper [8], a widely used defense practice against malicious traffic on the Internet is through blacklists: lists of prolific attack sources are compiled and shared. Blacklist aims to predict and block future attack sources. Existing blacklisting techniques have focused on the most effective attack sources, and recently, on collaborative blacklisting. In this paper, formulate the problem of forecasting attack sources (also referred to as "predictive blacklisting") based on shared attack logs as an implicit recommendation system. Compare the performance of existing approaches against the upper bound for prediction, and demonstrate that there is much room for improvement[8].

III. PROBLEM DEFINITION

None of these algorithms are perfect to recommended accurate result for the problem. Each and every algorithm has some porn and cons and need to evaluate problem within the existing algorithms and correct them. Some evolution on existing algorithms is enlisted below.

Problem with Content Based Recommendation Technique:

a. Can only be effective in limited circumstances. It is not straightforward to recognize the subtleties in content.
b. Depend entirely on previous selected items and therefore cannot make predictions about future interests of users.
c. These shortcomings can be addressed by collaborative filtering (CF) techniques.

Problem with Collaborative Based Recommendation Technique:

a. Cold Start: There needs to be enough other users already in the system to find a match.
b. Sparsity: If there are many items to be recommended, even if there are many users, the user/ratings matrix is sparse, and it is hard to find users that have rated the same items.
c. First Rater: Cannot recommend an item that has not been previously rated.
    - New items
    - Esoteric items
d. Popularity Bias: Cannot recommend items to someone with unique tastes.

The two main problems of user-based CF are that the whole user database has to be kept in memory and that expensive similarity computation between the active user and all other users in the database has to be performed.

IV. PROPOSED SOLUTION

To overcome the existing blacklist recommendation problem Hybrid Recommendation Algorithm has been proposed. It is a fusion of Content and Collaborative recommendation algorithms. Content Based Blacklist Recommendation algorithm use Naive Bayes Classifier and Collaborative Based Blacklist Recommendation algorithm Use Herlocker and Reidl Neighbor-hood based filtering method. Blacklists are widely used to deal with several types of malicious activity. For example, IP and DNS blacklists help to block unwanted web content, pam producers, and phishing sites.

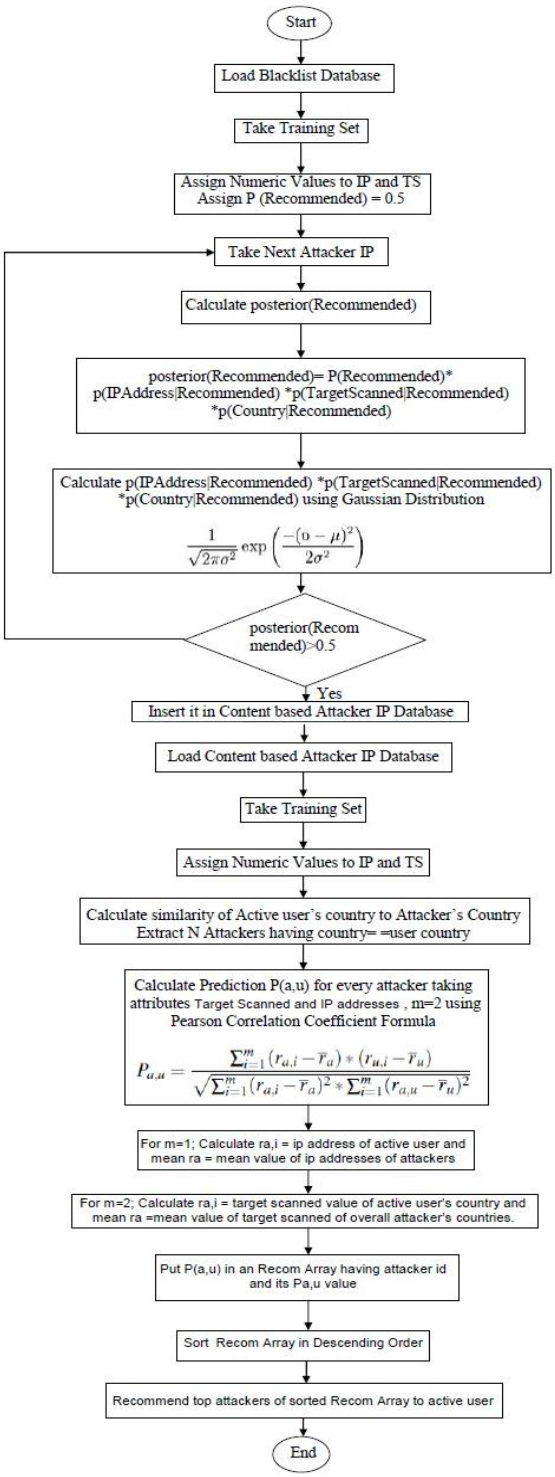

Figure 4.1: Flowchart of Hybrid Blacklist Recommendation Algorithm

V. RESULT EVALUATION

Table 6.1 Comparative Parameters Of Content, collaborative and Hybrid Algorithm:

Algorithms	No. of Inputs	Accuracy	Precision	Recall	F-Measure
ContentBased	60	0.61	.19	.07	.11
CollaborativeBased	60	.81	1	.39	.56
Hybrid	60	.89	1	.64	.78
ContentBased	340	.62	.23	.14	.21
CollaborativeBased	340	.9	1	.62	.77
Hybrid	340	.9	1	.62	.77

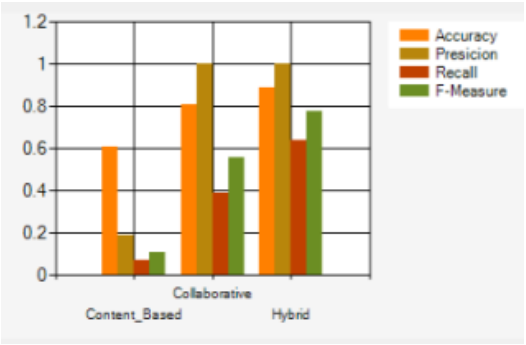

Fig 5.1 Result Chart

VI. CONCLUTION

This work covered the original of our recommendation experiments, and examined how to include information in the Recommendation algorithm from Metadata. In addition, examined if it was possible to combine the recommendations generated by our algorithms and improve recommendation performance by exploiting the complementarities of different algorithms. Here quantified how common these problems are, and proposed algorithms for the automatic detection of IP address and domain name. Also investigated the influence they can have on recommending items for victim logs. At last come to a conclusion of this work in this final part.

This work has examined how recommendation systems can be applied to the domain of IP address blacklisting. In this work, study about the malicious activity has been done on the basis of given past observations. As part of this study, analysed a real dataset of 1-month logs from Dshield, available through a shared repository of logs from different victims/contributors. Propose Hybrid Algorithm, which is apply on the repository of attackers and victim logs and gives the blacklist to future attacks. It brings improvement of degree of prediction rate. Alternatively, the network administrators of these hosts could be contacted and warned so that they can take action and fix their systems.

REFERENCES

[1] Artus Krohn-Grimberghe, Alexandros Nanopoulos, Lars Schmidt-Thieme, "A Novel Multidimensional Framework for Evaluating Recommender Systems" Barcelona, Spain, Sep 30, 2010. pp.34-51.

[2] Fabio Soldo, Anh Le, Athina Markopoulou "Blacklisting Recommendation System: Using Spatio-Temporal Patterns to Predict Future Attacks," IEEE Journal on selected areas in communication vol.29,no.7 Aug 2011.pp.1423-1437.

[3] J. Zhang, P. Porras, and J. Ullrich, "Highly predictive blacklisting," in USENIX Security, San Jose, CA, USA, Jul. 2008.pp.16-32.

[4] Yehuda Koren, "Collaborative Filtering with Temporal Dynamics," in KDD'09, June 28–July 1, 2009, Paris, France.

[5] Andreas Töscher, Michael Jahrer, and Robert Legenstein "Improved Neighborhood-Based Algorithms for Large-Scale Recommender Systems," in 2nd Netflix-KDD Workshop, August 24, 2008, Las Vegas, NV, USA.

[6] A. Ramachandran, N. Feamster, and S. Vempala, "Filtering spam with behavioral blacklisting," in ACM CCS, Alexandria, VA, Oct 2007.

[7] Zan Huang , Daniel Zeng and Hsinchun Chen, "A Comparative Study of Recommendation Algorithms in Ecommerce Applications," in Proc. of ACM KDD '04, Seattle, WA, USA, Aug. 2004, pp. 79–88.

[8] F. Soldo, A. Le, and A. Markopoulou, "Predictive Blacklisting as an Implicit Recommendation System," in IEEE INFOCOM, San Diego, CA, USA, Mar. 2010.pp.1-11.

AUTHOR PROFILE

Mr Pritesh Saklecha received the B.E. degree in 2014 from Medicaps Institute of Science and Technology, Indore, M.P. and M.E. (Software Engineering) in 2018 from IET DAVV Indore, India.

Dr. Jagdish Raikwar received B.E. degree in 2005 from RGPV Bhopal, M.P. and M.Tech. (information Technology) in 2008 from RGPV Bhopal, India and currently working as Assistant Professor in IET DAVV Indore, India.