Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 109C (2017) 281–288

Procedia Computer Science

www.elsevier.com/locate/procedia

The 8th International Conference on Ambient Systems, Networks and Technologies (ANT 2017)

Using Case-Based Reasoning for Phishing Detection

Hassan Y. A. Abutair[a,*], Abdelfettah Belghith[a]

[a]*Computer Science Department, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia*

Abstract

Many classifications techniques have been used and devised to combat phishing threats, but none of them is able to efficiently identify web phishing attacks due to the continuous change and the short life cycle of phishing websites. In this paper, we introduce a Case-Based Reasoning (CBR) Phishing Detection System (CBR-PDS). It mainly depends on CBR methodology as a core part. The proposed system is highly adaptive and dynamic as it can easily adapt to detect new phishing attacks with a relatively small data set in contrast to other classifiers that need to be heavily trained in advance. We test our system using different scenarios on a balanced 572 phishing and legitimate URLs. Experiments show that the CBR-PDS system accuracy exceeds 95.62%, yet it significantly enhances the classification accuracy with a small set of features and limited data sets.

Peer-review under responsibility of the Conference Program Chairs.

Keywords: Phishing Detection, Machine Learning Classifiers, Case-Based Reasoning, Features Selection.

1. Introduction

Phishing is a method in which the attackers trick users into revealing sensitive information to be used in committing fraudulent activities. Many sophisticated techniques have been used by phishers to deceive unaware users such as leveraging social engineering tactics and technology to deliver well-crafted emails to delude users that emails are legitimate. In spite of phishing threats are on the rise, until now, there is no phishing detection system or technique that perfectly can detect or dynamically can adapt to differentiate between phishing and legitimate websites, this refers to the changeable nature and to the short life cycle of phishing websites. To thwart phishing attacks with a remarkable success, we should (1) educate people about different scenarios and forms of phishing attacks and (2) deploy detecting and filtering systems because phishing attacks are considered as semantic attacks [1], in which, attackers can easily trick innocent users by crafting deceptive semantic techniques. Phishing attacks mainly depend on deceptive techniques of social engineering to deliver successful phishing attacks; that is the more deceptive or trickier a phishing email is, the more successful the attacker can get user credentials back or to entice him into clicking on malicious links to download malware. The phishing attack vector starts by social engineering to craft a convincing email and then by utilising technology to deliver phishing emails. The typical phishing vector or attack through emails includes three phases which are Lure, Hook, and Catch [2].

* Corresponding author. Tel.: +966-11-4697353 ; fax: +966-11-4696452.
E-mail address: habualteer.c@ksu.edu.sa

Peer-review under responsibility of the Conference Program Chairs.
10.1016/j.procs.2017.05.352

- **The Lure** is a well-crafted email that looks legitimate and official, it contains an embedded link pointed to the hook to direct the user to a fake website that may belong to a Bank or to an organization.
- **The Hook** is a fake website that mimic the legitimate website in which the user can reveal his credentials.
- **The Catch** includes the use of sensitive information collected by the attackers in fraudulent activities.

In this study, we propose the CBR-PDS system that can learn and detect phishing attacks. It is also adaptive and dynamic with regards to new cases in contrast to other techniques. CBR-PDS can work effectively with a relatively small set of features and data sets as well. Our work is compared with that of[3]. The latter introduces inter-related domain similarities to detect phishing websites. However, it stands short to detect new cases of phishing websites for which it has not been trained. CBR-PDS can easily adapt and dynamically cover new cases and solve new problems based on past experiences.

1.1. Related Works

In phishing research literature, Machine Learning (ML) techniques are used as a binary classification to classify websites into either legitimate or fake websites[4]. The aim of these approaches is to bridge the user training gap due to user ignorance or inability to recognize phishing attacks[5]. Machine learning techniques are based on data sets that consist of previous experiences or a collection of examples. Each example is represented by a set of characteristic attributes or features[6]. There are several machine learning approaches used for combating phishing attacks. These approaches include Support Vector Machine (SVM), Bayesian Additive Regression Trees (BART), Random Forests (RF), Classification and Regression Trees (CART), Logistic Regression (LR), and Artificial Neural Networks (ANN). Liu et al.,[7] proposed an approach to identify phishing websites by mining the associated webpage set for a website. They explored the relationship to the given website with respect to some features like text similarity, ranking relationship, link relationship, and webpage layout similarities. Their experiments showed 91.44 percent accuracy rate with a false alarm rate of about 3.40 percent. Abu-Nimeh et al.[8] compared different machine learning techniques for predicting phishing emails; namely BART, RF, CART, LR, SVM, and ANN. They used 2889 malicious and legitimate emails samples with 43 extracted features. The study shows that LR dominated all the others used techniques. In similar recent study, Basnet and Doleck[9] conducted a comparison among seven machine learning techniques and showed that RF performs the best while SVM performs the worst. Also, SVM was investigated in[10,11] for phishing detection. Marchal et al.[3] introduced a new approach based on the relationships between the low-level domain (the registered domain) and the upper-level domain (path or query) parts of an URL. They defined a new concept called intra-URL relatedness, also they leveraged the Google[1] and Yahoo[2] search engines queries to establish the relatedness between the words. They evaluated this concept using features extracted from the words composing the URL. Their results show a correct classification rate of 94.91% with only 1.44% false positives. SVM also require high computations in order to train the data and it is also sensitive to noisy data and prone to overfitting. RF can deal with a large number of variables, and can also estimate the missing values. The major drawback of RF is the reproducibility in which it builds a random forest and it is difficult to interpret the final model and subsequent results[4].

Attacker nowadays depend on obfuscated URLs to trick users. Authors in[12,13,14,15,16,17,18] used URL based features for detecting phishing websites. Our work concentrates on URL features extraction and key words similarities in the main domain name and the remaining keywords in the URL such as the path, subdomains, and query part.

Phishing fighting is an ever continuous cyberwar. Phishers daily produce new phishing URLs and deploy various smart ways to trick users into revealing their sensitive information. The crux of phishing websites lies in their frequent changes with different forms (in particular URL obfuscations). As such, machine learning based phishing techniques stay short in detecting new produced URLs as their classifiers have not learned the structure and nature of these new developed phishing URLs. This is indeed the rationale behind proposing CBR-PDS system as it:

- Has a better performance over machine learning techniques as it has the capability to avoid past errors and can focus on the main important parts of the problem.
- Does not need a deep understanding of the domain as it has the ability to adapt to solve new problems.
- Can be used independently of the data set size.

[1] www.google.com
[2] www.yahoo.com