Malicious URL Detection based on Machine Learning

ABSTRACT

Currently, the risk of network information insecurity is increasing rapidly in number and level of danger. The methods mostly used by hackers today is to attack end-toend technology and exploit human vulnerabilities. These techniques include social engineering, phishing, pharming, etc. One of the steps in conducting these attacks is to deceive users with malicious Uniform Resource Locators (URLs). As a results, malicious URL detection is of great interest nowadays. There have been several scientific studies showing a number of methods to detect malicious URLs based on machine learning and deep learning techniques. In this paper, we propose a malicious URL detection method using machine learning techniques based on our proposed URL behaviors and attributes. Moreover, bigdata technology is also exploited to improve the capability of detection malicious URLs based on abnormal behaviors. In short, the proposed detection system consists of a new set of URLs features and behaviors, a machine learning algorithm, and a bigdata technology. The experimental results show that the proposed URL attributes and behavior can help improve the ability to detect malicious URL significantly. This is suggested that the proposed system may be considered as an optimized and friendly used solution for malicious URL detection.

EXISTING SYSTEM

A. Signature based Malicious URL Detection

Studies on malicious URL detection using the signature sets had been investigated and applied long time ago [6, 7, 8]. Most of these studies often use lists of known malicious URLs. Whenever a new URL is accessed, a database query is executed. If the URL is blacklisted, it is considered as malicious, and then, a warning will be generated; otherwise URLs will be considered as safe. The main disadvantage of this approach is that it will be very difficult to detect new malicious URLs that are not in the given list.

B. Machine Learning based Malicious URL Detection

There are three types of machine learning algorithms that can be applied on malicious URL detection methods, including supervised learning, unsupervised learning, and semisupervised learning. And the detection methods are based on URL behaviors. In [1], a number of malicious URL systems based on machine learning algorithms have been investigated. Those machine learing algorithms include SVM, Logistic Regression, Nave Bayes, Decision Trees, Ensembles, Online Learning, ect. In this paper, the two algorithms, RF and SVM, are used. The accuracy of these two algorithms with different parameters setups will be presented in the experimental results.

The behaviors and characteristics of URLs can be divided into two main groups, static and dynamic. In their studies [9, 10, 11] authors presented methods of analyzing and extracting static behavior of URLs, including Lexical, Content, Host, and Popularity-based. The machine learning algorithms used in these studies are Online Learning algorithms and SVM. Malicious URL detection using dynamic actions of URLs is presented in [12, 13]. In this paper, URL attributes are extracted based on both static and dynamic behaviors. Some attribute groups are investigated, including Character and semantic groups; Abnormal group in websites and Host-based group; Correlated group.

Disadvantages

- The system is not implemented Machine Learning Algorithm Selection.
- The system is not implemented URL Attribute Extraction and Selection.

Proposed System

- In the proposed system, machine learning algorithms are used to classify URLs based on the features and behaviors of URLs. The features are extracted from static and dynamic behaviors of URLs and are new to the literature.
- Those newly proposed features are the main contribution of the research. Machine learning algorithms are a part of the whole malicious URL detection system. Two supervised machine learning algorithms are used, Support vector machine (SVM) and Random forest (RF).

Advantages

- The proposed algorithms are suitable to utilized the usefulness of our new features selected for malicious URL detection.

- In the proposed work, SVM and RF are selected as an example to illustrate the good performance of the whole detection system, and are not our main focus. Readers are encouraged to implement some other algorithms such as Naïve Bayes, Decision trees, k-nearest neighbors, neural networks, etc.

SYSTEM REQUIREMENTS

- **H/W System Configuration:-**

Processor	Pentium –IV
RAM	4 GB (min)
Hard Disk	20 GB
Key Board	Standard Windows Keyboard
Mouse	Two or Three Button Mouse
Monitor	SVGA

<u>SOFTWARE REQUIREMENTS:</u>

Operating system	Windows 7 Ultimate.
Coding Language	Python.
Front-End	Python.
Back-End	Django-ORM
Designing	Html, css, javascript.
Data Base	MySQL (WAMP Server).