

CLIENT SIDE SECURITY FOR URL AUTHENTICITY

SUBMITTED BY

S MUNIESHVAR (17MX112)

V AKHIL (17MX201)

15MX48 Mini Project I

DISSERTATION SUBMITTED IN PARTIAL FULFILMENT OF THE

REQUIREMENT FOR THE DEGREE OF

MASTER OF COMPUTER APPLICATIONS

ANNA UNIVERSITY



APRIL-2019

DEPARTMENT OF COMPUTER APPLICATIONS

PSG COLLEGE OF TECHNOLOGY

(Autonomous Institution)

COIMBATORE-641 004

PSG COLLEGE OF TECHNOLOGY

(Autonomous Institution)

COIMBATORE-641 004

15MX48 Mini Project I

CLIENT SIDE SECURITY FOR URL AUTHENTICITY

Bonafide record of work done by

S MUNIESHVAR (17MX112)

V AKHIL (17MX201)

Dissertation submitted in partial fulfilment of the requirements for the degree of

MASTER OF COMPUTER APPLICATIONS

ANNA UNIVERSITY

APRIL 2019

Faculty Guide

.....

Dr B Kalpana

ACKNOWLEDGEMENT

We indulged to thank **Dr.R.Rudramoorthy**, Principal, PSG College of Technology, Coimbatore for permitting us to take up this project.

We extend our earnest gratitude to **Dr.A.Chitra**, Professor and Head, Department of Computer Applications, PSG College of Technology, Coimbatore, for her support, encouragement and above all her ardent motivation.

We would like to express our gratitude to our guide **Dr.B.Kalpana**, Assistant Professor, Department of Computer Applications, PSG College of Technology, for her valuable suggestions, encouragement and guidance in the process of completion of the project.

We would like to express our gratitude to our tutor **Dr.N.Ilayaraja**, Assistant Professor, Department of Computer Applications, PSG College of Technology, for her support.

Finally, we express our gratefulness to all the faculty members of Department of Computer Applications, PSG College of Technology, and our friends for their encouragement and support.

SYNOPSIS

Phishing has become serious network security problem. Phishing has been mainly done by using spoof websites, which seems to be legitimate. Phishing causes loss of billions of dollars to both consumers and e-commerce companies. An approach to detect phishing URL using certain rules based on features of URL contents, which can be used as enterprise solution to anti-phishing.

A phishing URL and the corresponding page have several features which can be differentiated from a legitimate URL. The features such as URL-based, domain-based, page-based, content-based features are used differentiate spoofed URL from legitimate. Features of URL has been organised and a classifier algorithm has been used to identify phishing link from legitimate link.

The classifier methodology in this application evaluates and generates a set of rules for identifying spoofed links. The URL from the browser is segmented and passed through classifier algorithm. The browser will load the URL if it is legitimate or inform the user that the URL is spoofed. This approach is to effectively detect the spoofing URL with minimal false negatives at a speed adequate for online application.

CONTENTS

CHAPTER	PAGE NO
ACKNOWLEDGEMENT	i
SYNOPSIS	ii
1. INTRODUCTION	
1.1 Overview	1
1.2 Hardware Requirements	2
1.3 Software Requirements	2
1.4 Tools and Technologies	2
2. LITERATURE SURVEY	4
3. SYSTEM ANALYSIS	
3.1 Limitations of Existing System	6
3.2 Proposed System	6
3.3 Requirements Specification	
3.3.1 Functional Requirements	8
3.3.2 Non-Functional Requirements	9
4. SYSTEM DESIGN	
4.1 System Flow	10
4.2 Database Design	11
5. SYSTEM IMPLEMENTATION	
5.1 Working of application	12
5.2 Working with extension	15
6. SYSTEM TESTING	
6.1 Unit Testing	22
6.2 Integration testing	23
7. CONCLUSION	24
BIBILIOGRAPHY	25

CHAPTER 1

INTRODUCTION

This chapter gives a detailed study of the application. The development and deployment environments of the system are also specified in this chapter. It gives the brief description about various technologies and tools used in the development of the system.

1.1 OVERVIEW

Phishing is a cybercrime in which a target or targets are contacted by email, telephone or text message by someone posing as a legitimate institution to lure individuals into providing sensitive data such as personally identifiable information, banking and credit card details, and passwords.

A phishing is a form of fraud in which the attackers tries to get sensitive information from the users by creating a spoofed website by means of sending the URL of the spoofed website through mail which seems to be like a legitimate mail. The main reason for this type of attack is the lack of awareness of the users. But a system has to be developed to find the URL has been legitimate or spoofed.

Phishing is a serious problem that is achieved in a number of different ways. Email spoofing and website spoofing are two of the primary methods by which phishers acquire sensitive information from unsuspecting Internet users.

Website spoofing is used to make people believe that they are interacting with a trusted, legitimate company or person. Especially sophisticated methods of website spoofing can result in forged sites that appear nearly identical to their legitimate counterparts. If users are in a hurry, it is especially easy to fall prey to these sites. At a glance, they often appear to be real. Whenever user access a site through a link, it is important to be especially sceptical about it. Look closely at the URL. Keep in mind, however, that there are ways to cloak URLs.

A wide range of phishing techniques are used to create spoofed websites. As mentioned above, URL cloaking is a popular method. Through the use of specialized scripts, phishers can cover up the true URL with one that is associated with a trusted website.

Subdomains are also commonly used to confuse Internet users and to lend them false senses of security. Internationalized domains are increasingly being used in this way too. As with spoofed email addresses, URLs sometimes contained a few transposed letters. At a glance, they appear to be correct and are trusted by unsuspecting Internet users.

1.2 HARDWARE REQUIREMENTS

- 8MB of RAM
- Processor 1GHZ
- Ethernet connection or wireless adapter

1.3 SOFTWARE REQUIREMENTS

- Apache Tomcat or Glassfish
- Chrome Browser or Chromium Browser

1.4 TOOLS AND TECHNOLOGIES USED

The following are the tools and technologies used to develop the application:

Net beans

Net beans is an open source Integrated Development Environment developed by Sun Microsystems in 1999. It supports development of all java application types such as Java EE, Java SE, web applications, EJB, etc. This IDE has been used to develop a java application.

JSP

Java Servlet Page is a server-side programming technology that enables the creation of Dynamic, platform-independent method for building the web based application. JSP application has been used to identify that the given URL is spoofed or legitimate and it will decide whether the URL has to be loaded or not.

Java DB

Net beans provide support for creating and maintaining relational databases by embedding apache derby. Java DB has been used to store the domain of any URL and the type of URL such as legitimate or spoofed. So that once the URL type is identified, it is not necessary to check that URL every time for spoofing or not.

Jsoup API

Jsoup is a java library for working with real world HTML which provides a very convenient API for extracting and manipulating data using DOM. It implements WHATWG HTML specification and parses HTML to the same DOM as all browsers do. This API has been used to check if the webpage has input tag of type password.

Jazzy API

Jazzy is a command-line utility that generates documentation for swift or objective-C. It has been used to check all the words in the URL with the words in the dictionary. The URL has been converted into a statement and has been given as input for this API for which it will identify the spelling is typo-squatted or not.

Java Script

Java Script is a script language that runs on client machine and also acts as glue to bind JSP application with the browser action. All the chrome extensions are created by using the Java Script for which the browser action for every URL has been declared and defined.

JSON

JSON has been used to create a manifest file for chrome extension. It has been used to tell the behaviour of the extension application that is installed on system or mobile.

CHAPTER 2

LITERATURE SURVEY

Dr.V.Karamchand Gandhi[1] describes the most common features that are used to find the differentiation of legitimate and phishing Web Pages based on the URL features. By evaluating all the features, one can determine that the website which resembles the following features considered as phishing. The common features to develop a legitimate website are identified and these features are compared with phishing website features. This is done by the prediction algorithm. Differences are identified, an algorithm is developed by considering these features to differentiate and identify the legitimate website from the phishing website.

The algorithm in this system work as follows:

- If the age is less than 6 days and if there is no DNS record in the domain then it is labelled as spoofed website.
- If webpage rank is greater 100000 then it is labelled as suspicious.
- If webpage rank is less than 100000 then it is labelled as legitimate otherwise spoofed website.
- If any suspicious character such as '@' symbol is available in URL then it is labelled as spoofed website.

In this algorithm, there are lot of features of URL that has not been used. Also there are suspicious websites, where the algorithm can't able to find that it belongs to spoofed or legitimate.

Dr.Ebubekir Büber[2] by using the decision tree algorithm, has classified the URL as spoof website or legitimate website. There are 12 attributes, that are considered in this system. They are domain, count of subdomain, digits, special characters in URL, top level domain, age of domain, global page rank, meta tag, form features.

Decision Tree uses an information gain measure which indicates how well a given feature separates the training examples according to their target classification. The name of the method is Information Gain. High Gain score means that the feature has a high distinguishing ability. Because of this, the feature which has maximum gain score is selected as the root.

The Decision Tree Algorithm calculates this information for every feature and selects features with maximum Gain scores. To grow the tree, leaves are changed as a node which represents a feature. As the tree grows downwards, all leaves will have high purity. When the tree is big enough, the training process is completed.

The Tree created by selecting the most distinguishing features represents model structure for our detection mechanism. Creating mechanism which has high success rate depends on training dataset. For the generalization of system success, the training set must be consisted of a wide variety of samples taken from a wide variety of data sources. Otherwise, our system may working with high success rate on the dataset, but it cannot work successfully on real world data.

CHAPTER 3

SYSTEM ANALYSIS

This chapter comprises of system study which tells about the existing system and the proposed system. It explains the requirements specification which includes the functional and non-functional requirements.

3.1 LIMITATIONS OF EXISTING SYSTEM

In existing system, the spoof website detection is done only through domain based feature. Domain age, DNS record, Website traffic, '@' symbol presence, https in URL's domain are the only attributes that have been taken for the identification of spoofed website which detects only of about 45% of spoofed website. There are many more features of website that are to be noted. Phishers create spoofed site which inherits the design and outlook of the legitimate website. The existing system does not identify such type of spoof websites.

Machine learning algorithm takes more time to predict that the given URL is spoofed or legitimate. So when it comes to real time working to load a website user won't wait for so long time.

3.2 PROPOSED SYSTEM

In proposed system, many of the features of the website have been taken into account. The proposed system has the following features and an updated version of algorithm to identify the spoofed or legitimate websites.

The objective of this application is to identify the spoofed website and to block in client system and to shield users from malicious or unsolicited links.

Analysis of phishing URL:

<http://paypal.com-webusersuserid29348325limited.active-userid.com/webapps/89980>

The above URL has been analysed as follows:

- | | |
|---------------|-------------------|
| • Protocol | http:// |
| • Domain Name | active-userid.com |
| • Path | /webapps/89980/ |

- Subdomain item1 com-webappsuserid29348325limited
- Subdomain item2 paypal

A phishing URL and the corresponding page have several features which can be used to differentiate from a legitimate URL. The features are as follows:

- URL-Based Feature
- Domain-Based Feature
- Page-Based Feature
- Content-Based Feature

Some of the URL-Based features are as follows:

- Digit count in the URL
- Total length of the URL
- Checking for the URL is typosquatted or not.
- Checking whether it includes legitimate brand name or not.
- Number of subdomains.
- Top Level Domain (TLD) is commonly used or not.

Some of the Domain-Based features are as follows:

- Is the domain name or its IP address is blacklisted in well-known reputation services?
- Age of the domain.
- Registrant name hidden or not.

Some of the Page-Based features are as follows:

- Global page rank
- Country page rank

Some of the Content-Based features as follows:

- Page titles and Meta tags
- Form and Sensitive information
- Image clarity

All the features are useful for phishing domain detection. The features are used by the detection mechanism depends on the purpose of the application. The feature for detection mechanism has been selected carefully.

The features that are considered for the identification of URL are as follows:

- URL-Based Feature
 - ◆ Digit count – The number of digits available in the URL
 - ◆ Length – length of the URL
 - ◆ Typo-Squatted – URL has misspelled word or not
 - ◆ Number of subdomain in URL
 - ◆ Top Level Domain commonly used one
- Domain-Based Feature
 - ◆ IP address blacklisted or not
 - ◆ Age of domain
- Page-Based Feature
 - ◆ Page rank
- Content-Based Feature
 - ◆ Form and sensitive information – check if the website has form which asks for highly secured data such as password, OTP, CVV, etc.,

3.3 REQUIRMENTS SPECIFICATION

3.3.1 FUNCTIONAL REQUIREMENTS

- Detection of spoofed websites.
- Blocks the website in browser if it is spoofed.
- Allows loading of website only if it is legitimate.
- Should work as Google browser extension.
- Support all type of Operating System.
- Supports for chrome as well as chromium browser.
- Handles all exceptions.

3.3.2 NON-FUNCTIONAL REQUIREMENTS

- It will load the webpages with little higher time than required to load the webpage without this application.
- To be reliable extension.
- Easy to install.
- Requires less maintenance.

CHAPTER 4

SYSTEM DESIGN

This chapter includes the flow diagram of how URL is authenticated for client browsers.

4.1 WORK FLOW

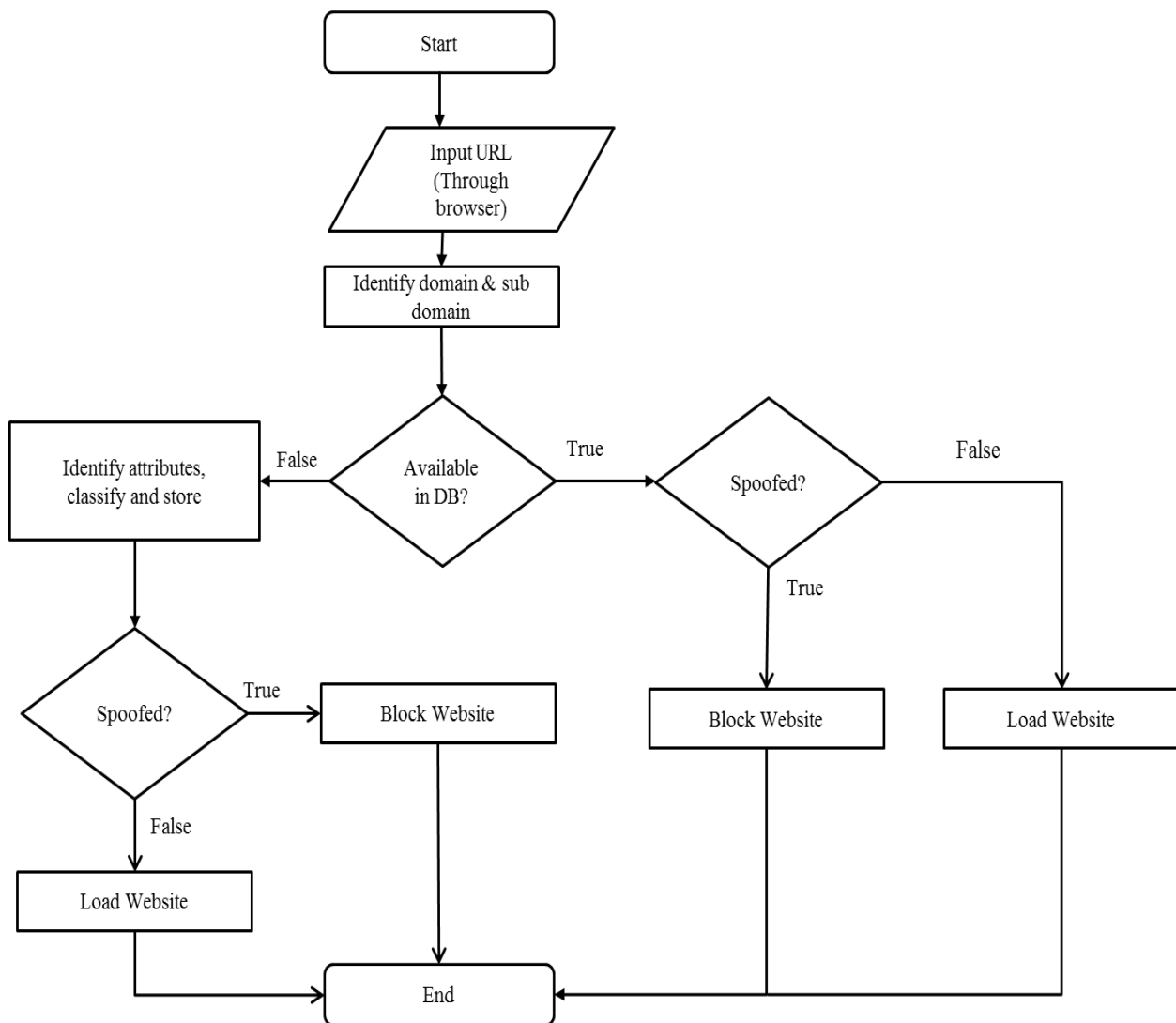


Fig 4.1 Work Flow of Anti-Phishing Application

4.2 TABLE DESIGN

Table name: URL_authenticity

Primary Key: Domain

Description:

This table has been used to store the already identified spoofed and legitimate websites where the domain along with the subdomain has been stored which act as primary key and the type of URL has also been stored.

Attribute	Data Type
Domain	Varchar(100)
URL_Type	Varchar(10)

Table 4.2 URL Authenticity

CHAPTER 5

SYSTEM IMPLEMENTATION

This chapter includes the working of anti-phishing application and chrome extension to connect this application with browser.

5.1 WORKING OF APPLICATION

Anti-phishing application will work based on the attributes of any website. They are as follows:

Detection of hidden IP address

To know that if the IP address of a website is hidden, `getByName()` has been used. It returns the `InetAddress` of the given host. If the host is a literal IP address, then only its validity is checked. Fetches public IP Address of the host specified. It takes the host as an argument and returns the corresponding IP address. If it throws exception so that it is confirmed that IP address is not found.

Protocol type

Using `getProtocol()` the protocol of the given URL has been found. There are two protocols such as HTTP and HTTPS. The function will return HTTP or HTTPS for the given URL.

Top Level Domain is commonly used one

The top twenty TLD has been declared by Google has been taken. The top twenty TLD are as follows:

- .com – Commercial
- .org – Noncommercial
- .edu – US accredited postsecondary institutions
- .gov – United States Government
- .uk – United Kingdom
- .net – Network services
- .ca – Canada
- .de – Germany

- .jp – Japan
- .fr – France
- .au – Australia
- .us – United States
- .ru – Russian Federation
- .ch – Switzerland
- .it – Italy
- .nl – Netherlands
- .se – Sweden
- .no – Norway
- .es – Spain
- .mil – United States Military

If the URL doesn't have any of the above mentioned TLD. Then the URL doesn't have the commonly used TLD. To have some high restrictions and limitations only top twenty TLD declared by Google has been considered.

Age of the domain

By using the whoIs() function all the details about the domain such as owner name, registered date, expiry date has been retrieved. From that the registered date has been taken and age is calculated. The age is calculated by finding difference between the registered date and current date.

Typo-Squatted

URL has been first converted as a string first then each word of the string is compared with the words in the dictionary using JAZZY API. Then the misspelled words are replaced and string is returned. If the return string and the input string is same then there will be no typo-squat in the URL.

Number of subdomains

By counting number of dots (.) leaving the TLD in the URL, the number of subdomains has been taken count. As the level of subdomain increases there will be a more chance that it might be a spoofed website.

Digit count

In casual most business applications never use digits in their domain and subdomains. So by considering that, number of digits in the URL is counted. Also some of the webpages uses year in their domain name so by considering that the digit count is checked if it is more than 4 or not.

Form with password type

It has been checked that the given form has element of input of type password or not. By using JSOUP API, the website is treated as document by which it will check whether the form is available or not. Then by iterating through each element in the form in the document the type has been checked. If there exist password type, then the website has some sensitive information requested to the user.

By considering all the above features, an efficient algorithm for identifying the spoofed website has been designed. All the features are combined in a way that any legitimate site will not be identified as spoofed and any spoofed website will not be identified as legitimate.

The algorithm used in proposed system is as follows

IF (IP address is not found) THEN

 “URL is SPOOFED”

ELSE IF (http protocol is used) THEN

 IF (form has sensitive data AND TLD is not commonly used AND age is less than 10) THEN

 “URL is SPOOFED”

 ELSE IF (typo squatted AND TLD is not commonly used AND subdomains more than 1 AND digit count greater than 4 AND age is lesser than 10) THEN

 “URL is SPOOFED”

 ELSE

 “URL is LEGITIMATE”

 END IF

ELSE

 “URL is LEGITIMATE”

END IF

Storing in the Database

For all the time it is not possible to check if the given URL is spoofed or legitimate. So once if the URL has been tried, that will be stored in the database along with type of URL. So that next time when the same URL is entered it will check with the Database first after that it will go for the features.

Step by step working of anti-phishing application

Step 1:

Get the URL from user.

Step 2:

Retrieve domain name from the URL.

Step 3:

Check that the domain is available in the database or not.

Step 4:

If domain is available in the database check for the type of URL and based on that action has been taken otherwise goto step 5.

Step 5:

If domain is not available in the database then all the attributes for the URL has been found and using classifier algorithm the type of URL is found and based on that action has been taken.

5.2 WORKING WITH EXTENSION

Extensions are nothing more than a web page that has access to browser APIs. They enable users to tailor Chrome functionality and behaviour to individual needs or preferences.

User interfaces should be minimal. But in this application there is no need of user interface. Here extension is used, to access anti-phishing application. Through extension anti-phishing application has been identified.

Extensions are distributed through the Chrome Developer Dashboard and published to the chrome Web store. But before publishing, to work with the extension certain steps are to be followed.

Installing Extension carries the following steps:

Step 1

Select menu button, in more tools options select extensions.

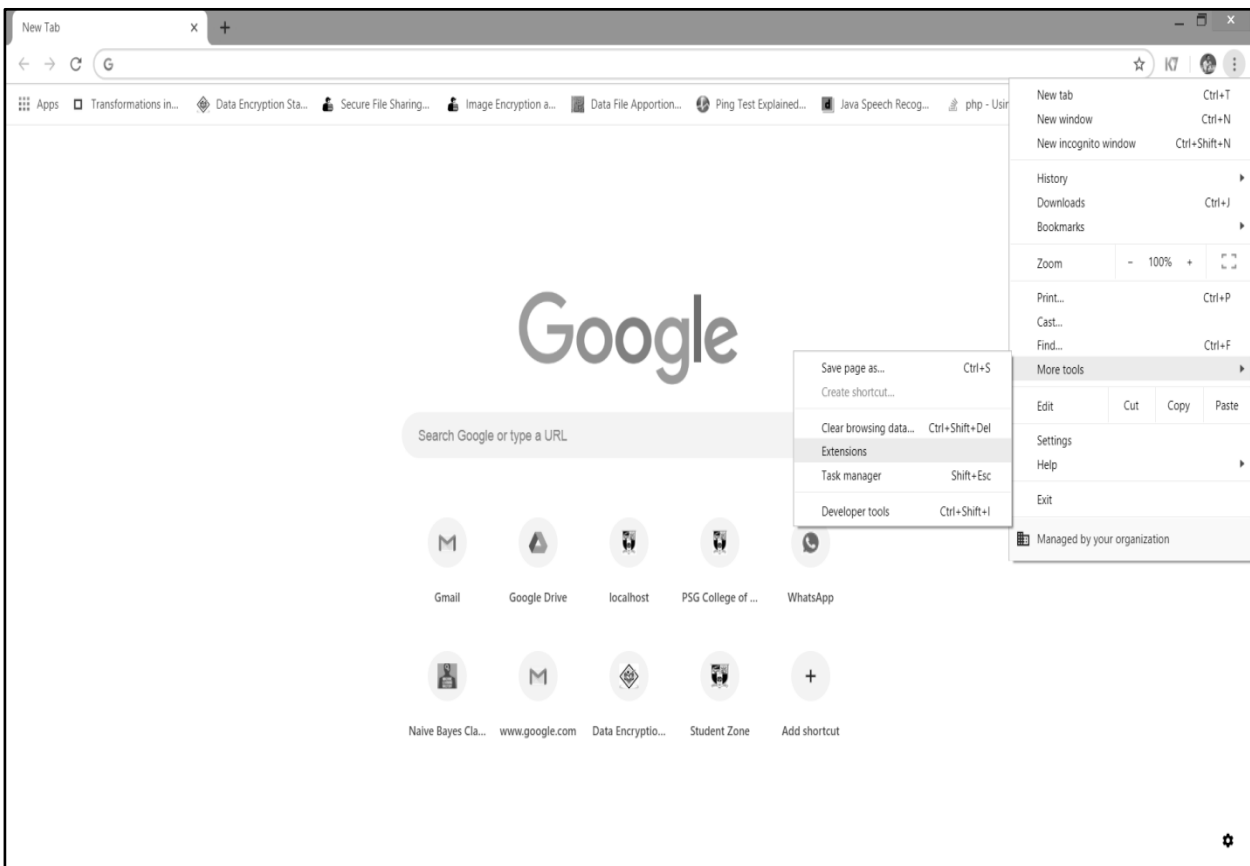


Fig 5.1 Extension Menu

Step 2

On the top right corner click on the developer mode.

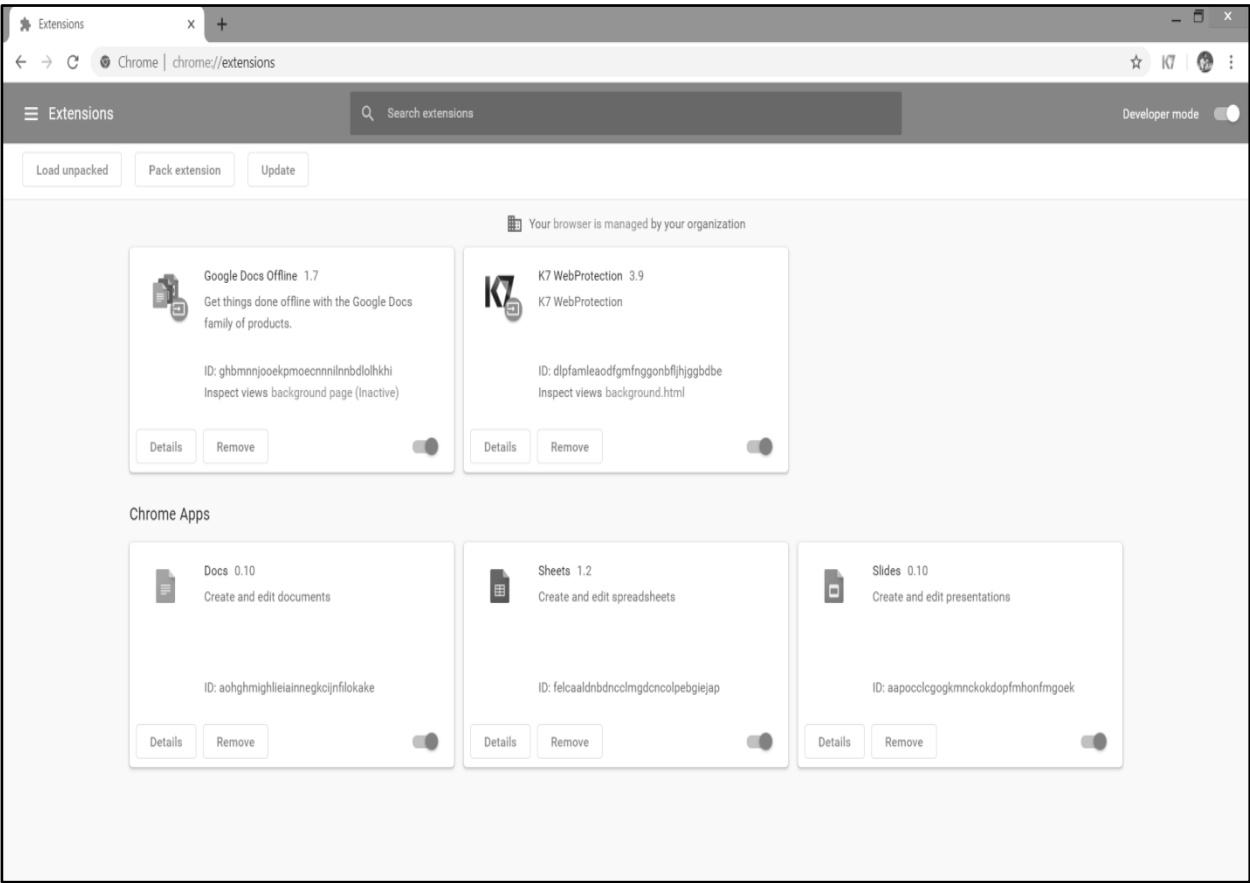


Fig 5.2 Extension Control Panel

Step 3

On the top left corner click on load unpacked and select the extension folder and click ok. Now the extension will be added.

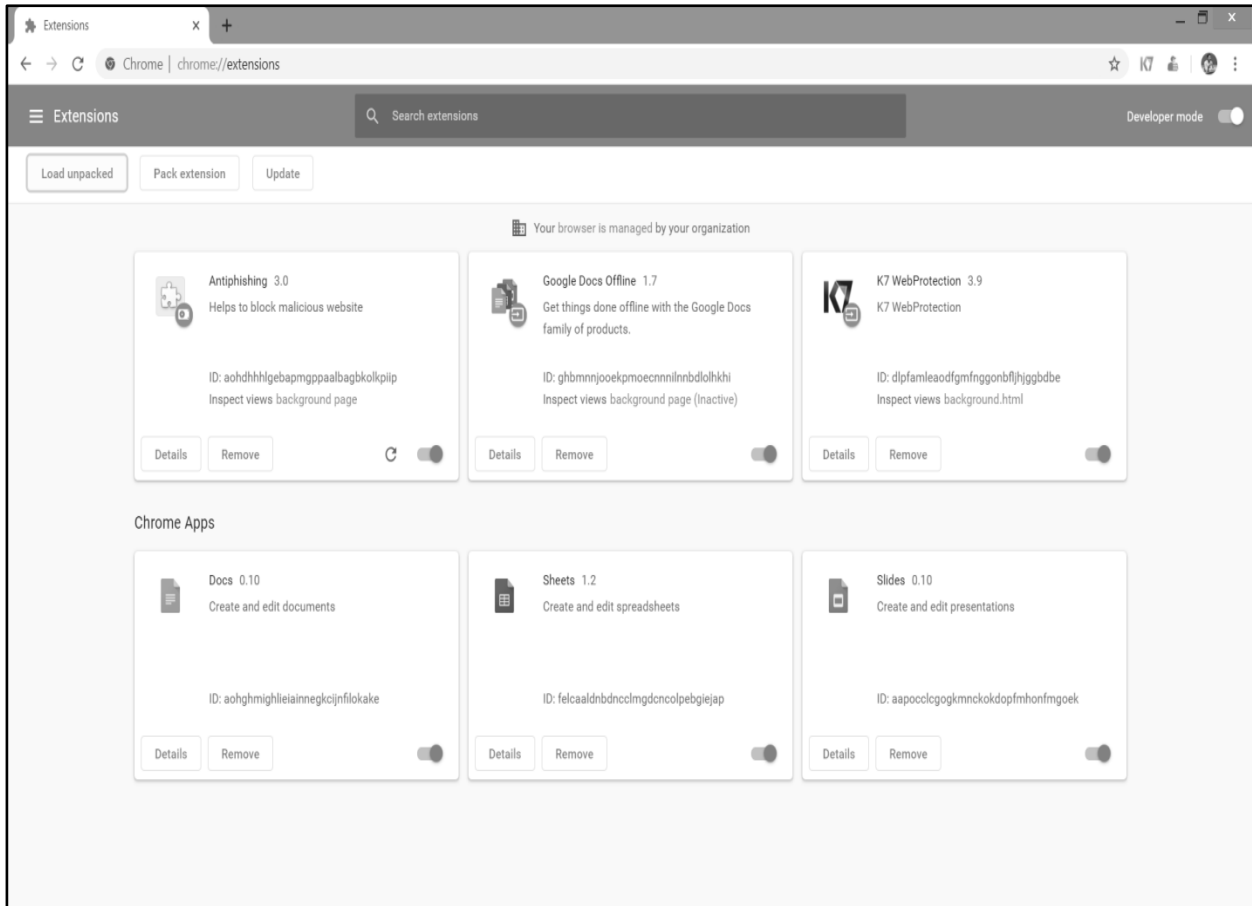


Fig 5.3 Adding Extension

On the top right side, the icon of the extension will be found.



Fig 5.4 Active Extension

To turn off the extension click on the on/off button in the extension

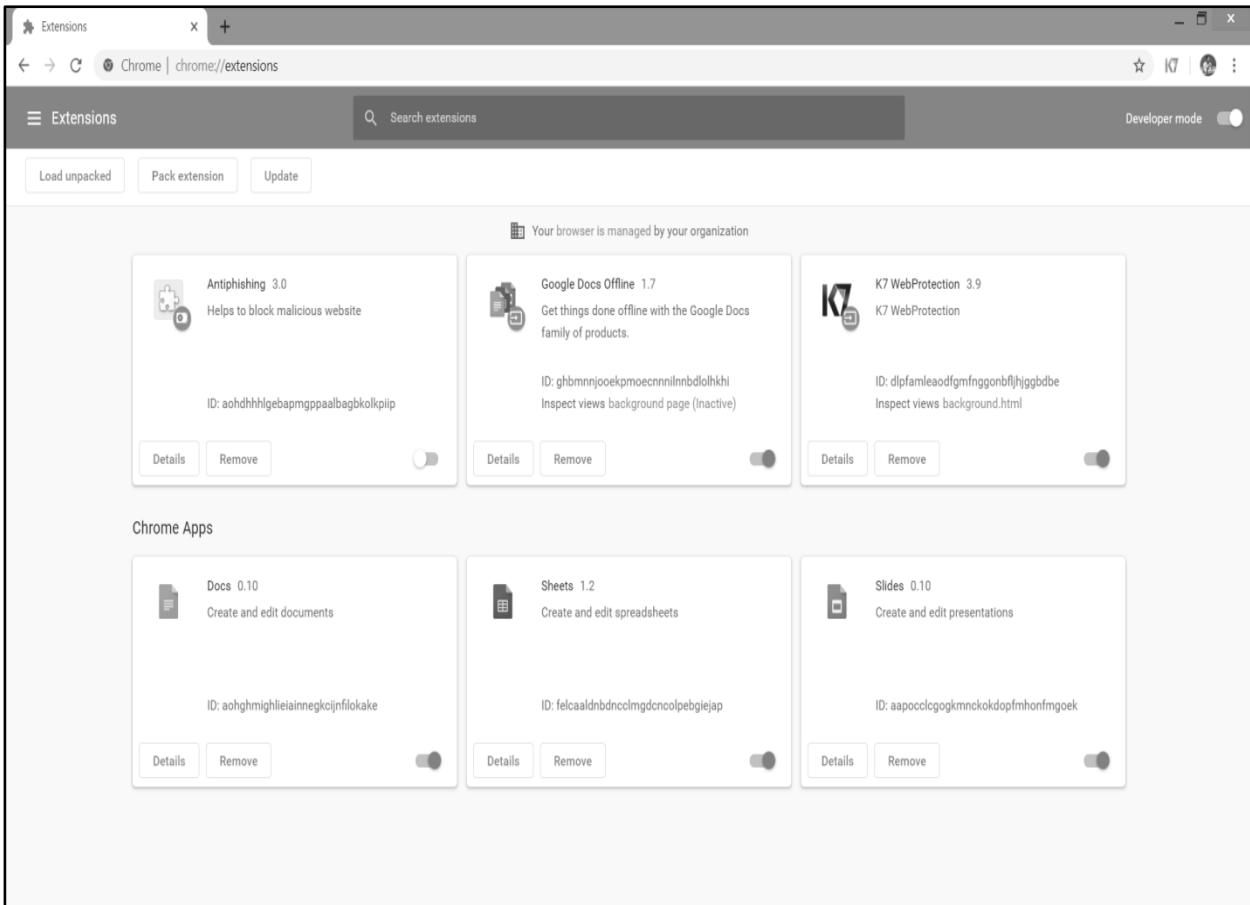


Fig 5.5 Disabling Extension

Now the extension icon is also disabled in the browser.

The above are the steps to install extension in the browser in developer mode. To install extension from chrome web store, visit the web store and search for the extension and click on add extension and the extension will be added directly to the browser.

Whenever a URL entered, extension will take that URL as input to the anti-phishing application and will be redirected to anti-phishing application. Then the anti-phishing application will check if the domain of the URL is spoofed or legitimate. If the URL is identified as legitimate then it will redirect to the URL given by the user otherwise the URL will not be loaded and will display a webpage showing that the website is spoofed and so it cannot be loaded.

This extension will not redirect every URL, as it will check only if the Domain changes because if the URL with a domain is already been checked when it is loaded for first time. So it is not necessary to check for every time when the menus of the webpage are clicked.

After connecting the anti-phishing application using the chrome extension. If any legitimate website is given then it will be loaded.

As in the below example psgtech.edu is the website given as sample and identified as legitimate website and loaded in the browser.

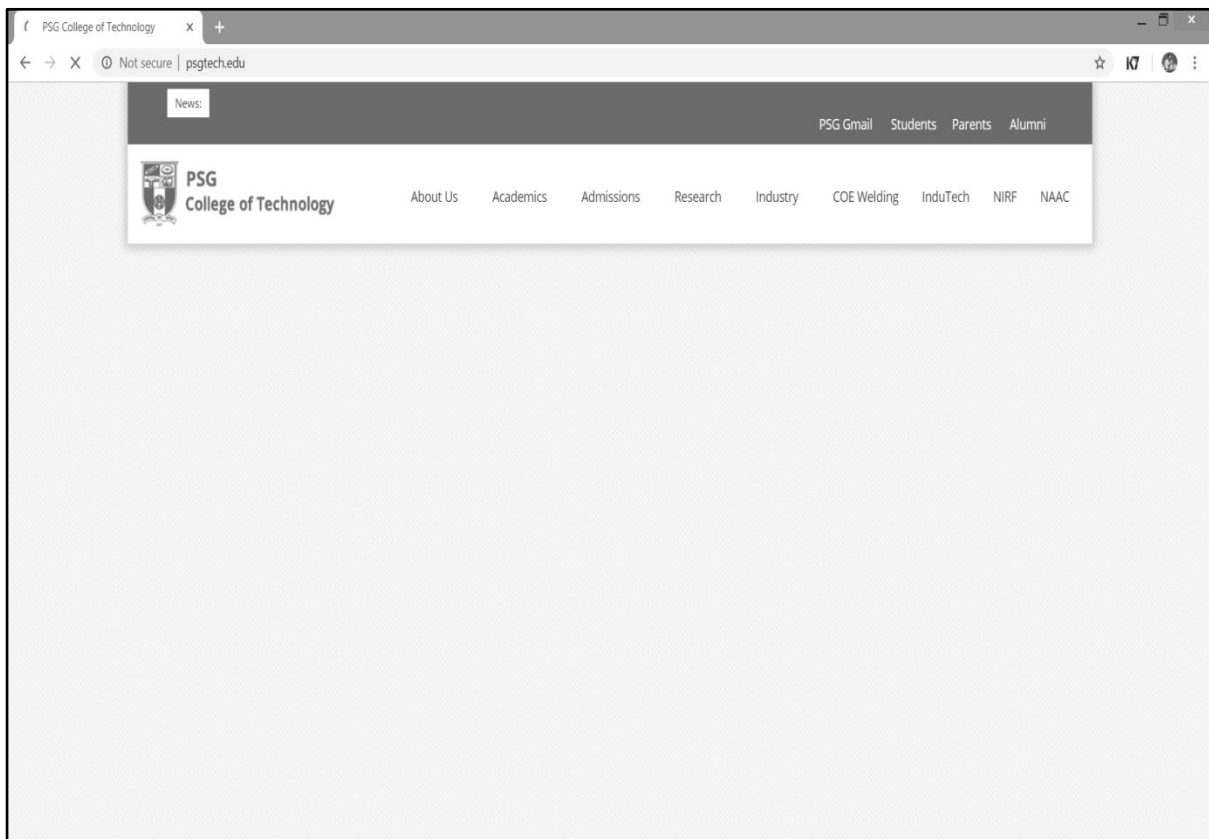


Fig 5.6 Loading legitimate URL

If the given URL is identified as spoofed website then the website will not be loaded and it will show the message in the browser as below



Fig 5.7 Blocking of Spoofed URL

CHAPTER 6

SYSTEM TESTING

This chapter includes various test cases and test reports that help in identification and removal of errors which were unnoticed during development.

6.1 UNIT TESTING

In unit testing, the anti-phishing application has been tested before it is connected to browser. Using the URL data set found in the “phishtank” website the application has been tested. The test case consists of all possible types of URL that has been tested.

Test case consists of a set of test inputs, execution condition and expected results developed for a particular objective, for instance to exercise a particular path to verify compatibility with a specific requirements. Test case data are taken from a “phish-tank” to determine the system is processing correctly.

Test ID	Input URL	Type of URL & Reason	Expected Result	Actual Result
1	paypal.com.webscr.123457.logi n.submit.dispatch.se	Spoofed, as it has malicious content.	Should not be loaded	Not loaded
2	paypal-manager-login.net	Spoofed, as it has malicious content.	Should not be loaded	Not loaded
3	google.com	Legitimate	Should be loaded	Loaded
4	1000projects.org	Legitimate	Should be loaded	Loaded
5	paypal-manager-loesung.net	Spoofed, as it has malicious content.	Should not be loaded	Not loaded
6	paypal-manager-account.net	Spoofed, as it has malicious content.	Should not be loaded	Not loaded
7	facebook.com	Legitimate	Should be loaded	Loaded
8	live.com	Legitimate	Should be loaded	Loaded

9	W3schools.com	Legitimate	Should be loaded	Loaded
10	paypal.com.laveaki.com.br	Spoofed, as it has malicious content.	Should not be loaded	Not loaded
11	paypal.com.client.identifiant.compte.clefs.informations.upgarde.mon.com	Spoofed, as it has malicious content.	Should not be loaded	Not loaded
12	Javatpoint.com	Legitimate	Should be loaded	Loaded
13	Wikipedia.org	Legitimate	Should be loaded	Loaded
14	Google.co.in	Legitimate	Should be loaded	Loaded
15	163.com	Legitimate	Should be loaded	Loaded
16	Psgthiran.in	Legitimate	Should be loaded	Loaded
17	Awardspace.com	Legitimate	Should be loaded	Loaded
18	000webhost.com	Legitimate	Should be loaded	Loaded
19	Yaxz.in.ed	Spoofed, as it has malicious content.	Should not be loaded	Not loaded
20	Apple.com.adminconsole.in.com	Spoofed, as it has malicious content.	Should not be loaded	Not loaded

Table 6.1 Testing Report

6.2 INTEGRATION TESTING

Integration testing is the process of testing the application by integrating all modules as single applications. As the chrome, has been integrated with anti-phishing application through extension. Whenever a URL is entered in browser then the extension will take the URL to the anti-phishing application and identified the type and has been loaded based on the result.

For all the test cases mentioned above, after integration of this application it gives the same result as in the unit testing. So after integration there is no impact on the anti-phishing application.

CHAPTER 7

CONCLUSION

This application for client side security, using URL authenticity has been successfully implemented. This application can be used in any type of device and for that the user has to download the extension for their browser. It supports multiple platform devices. Anti-phishing application has been deployed as a web extension where the browser will redirect the entered URL to the anti-phishing application where it will find the type of URL and blocked if it spoofed otherwise it will be loaded.

This application can be enhanced in future with some additional capabilities. In some legitimate website, which contains embedded URLs like advertisements or other page load URLs we can collect all URLs in the page and verify if each URL is spoofed or not. Whenever the user tries to click on the URL a popup menu shows the message if the URL is spoofed.

BIBLIOGRAPHY

References:

1. Dr V Karamchand Gandhi, "An Efficient Algorithm To Identify Phishing Sites Using URL Domain Features", International Journal of Advanced Research in Computer Science, 8 (7), July-August 2017,508-510
2. Dr Ebubekir Büber, "Phishing URL Detection with ML", International Journal of Advances In Computer Science and Cloud Computing, Volume-3, Nov-2015.

Web Links:

1. <https://developer.chrome.com/extensions/devguide>
2. <https://www.javatpoint.com/java-bean>
3. <https://www.w3schools.com/js/>
4. https://www.w3schools.com/xml/dom_intro.asp