Safe Internet Browsing Using Heuristic Based Technique

Vibhuti K. Patel[1], Prof. Hasmukh Patel[2]

[1]GTU, Gandhinagar, 382006, India
[2]HOD, LCIT Bhandu, Mehsana, India

***Abstract* - Today while people are depending upon Internet for their personal use as well as for business purposes, the fraud becomes a great problem for the users. Without the knowledge of the end users, attacker takes individuals confidential information by offering attractive offers through fake websites or spreading fake rumours. After fetching all these confidential information attacker misuses data or privileges of the end users. There are two ways to secure from these attackers. One way is to avoid use of Internet and second way is to identify the attacker and be secure from them. To avoid Internet is not the right way today because everything is depended on Internet. Hence it is necessary to detect the attacker and secure individuals private data. Till now many approaches are found to detect fake websites. Among them are Bacterial Foraging algorithm, Visual similarity based approach, Statistical Learning Theory etc. Every approach has some drawbacks or limitations such as less efficient or time consuming or no up-to-date blacklist or phishtank (Database having the fake websites' list). To detect the fake websites different techniques are proposed such as classifier technique, heuristics based technique, hybrid technique etc. In this paper, we propose a technique based on Heuristic based technique. In first Phase, website is identified based on different parameters like URL (Uniform Resource Locator), GTR (Google Top Rate), IP Address, Forms, age of domain. In second Phase, the visual similarity of the webpage is compared with the original website. The proposed approach also gives the suggestions to the users for their particular domain search that makes user more comfortable to use the system.**

***Key Words* – Phishing, Blacklist Generating and Updating, Visual Similarity, Classification Algorithm, TLD (Top Listing domain), Google's Toolbar Rank, Heuristic Value, Anti-Phishing Working Group, False Positive**

I. INTRODUCTION

Problem Statement

By way of people increasingly depend on internet for personal finance, business, investment; Internet fraud becomes a large threat. Internet fraud takes many forms to attack such as offered phony items for sale on ecommerce sites, to abusive rumours that manipulate stock prices etc. Through attractive websites and offers on Internet people want to take benefit and for this people share confidential data with the attacker. And attackers misuse the information and theft the privileges of the users. So it is necessary to identify fake websites for end user's security.

Phishing websites are the sites which have the identity of the genuine website. The term *Phishing is the variation of the term* '*Fishing',* replacing '*F'* with '*Ph'*. Phishing websites impersonate legal counterparts to attract users to visit their websites. Once the user visits the website, attacker can fetch the private information of the user. Websites which are created like a legal website and used for stealing the confidential data of individuals are Phishing websites So Phishing attacks are very serious problem to the users.

Till now many approaches are proposed for detecting phishing websites such as AZProtect, Black list Generator, Bacterial Foraging algorithm etc. All these approaches use different algorithms or methods for detecting fake websites. As per the websites' characteristics or behaviour we can identify whether it is fake or genuine. However people fail to identify the fake websites because nowadays attackers are also very clever. It is necessary to propose the best technique to identify the fake websites.

Background

The first consideration of the Phishing came in 1987, at the Interex conference. Jerry Felix and Chris Hauck presented a paper "*System Security: A Hacker's Perspective",* in which they discussed a method for a third party to replicate a trusted service [18]. In September 2001, in America, Attacker used an email asking for a post '9-11 ID check' to steal financial details from the E-Gold Digital currency service. This was the second Phishing attack. And the first was the same as the second in June 2001 in E-Gold at America. Both are initially seen as failure but they helped phishing to firmly on criminal organizations [18]. In 2004, Phishing was firmly established and between May 2004 and May 2005 it was estimated that $929 million (USD) was lost in phishing scam. After that the phishing attacks are taken place [18].

As per APWG report [1] phishing attacks are based on mostly banking sites. Through eBanking customers are attracted towards offers and giving account numbers and passwords without knowing of that website that whether it is fake or genuine. Below diagram gives the percentage of phishing attack in various phases on internet. Even social engineering websites are also made phishing to gain personal information about the users and to harass the users.

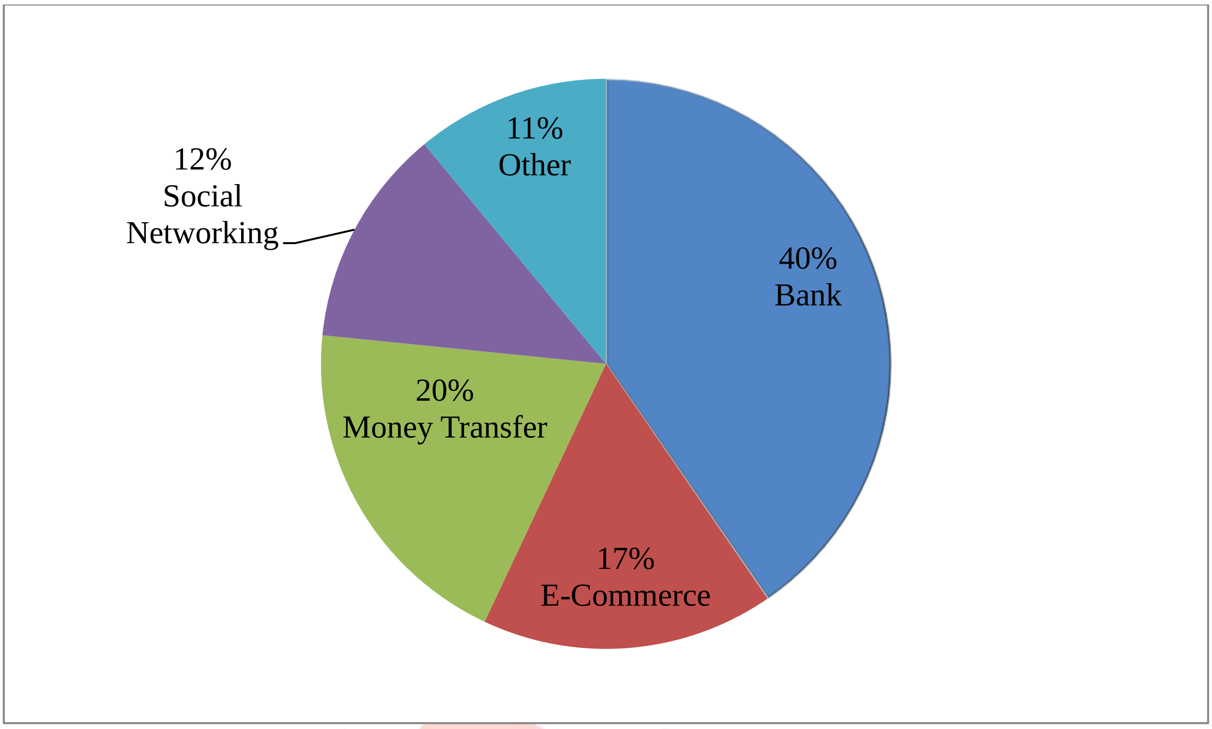

Fig. 1 Phishing sites ratio [1]

Among all attacks 20% of attacks only because of money transfer because while transferring money user does not know about the third party and giving all information to the website which is phishing site. Such a way to phishing attack also can be taken place while transferring money.

In 2013, Phishing attack numbers declined 20% from 2012 due to a precipitous drop in virtual server phishing attacks [19][1]. Most phishing occurs on hacked or compromised Web servers. The United States continued to be the top country hosting phishing sites during 2013 till. This is because of large fact that a large percentage of the world's websites and domains are hosted in the United States [1].

Various subdomains are used for phishing till 2013 as per APWG (Anti-Phishing Working Group) survey report. The list of those subdomains is as below:

Table 1 Phishing top-10 sub domain [1]

Rank	Attack	Domain	Provider
1.	833	net.tf	UNONIC.COM
2.	347	3owl.com	3owl.com
3.	290	usa.cc	Freeavailabledomains.com
4.	240	nazuka.net	nazuka.net
5.	226	altervista.org	altervista.org
6.	*193*	my3gb.com	my3gb.com
7.	155	kmdns.net	kmdns.net
8.	137	3eeweb.com	3eeweb.com
9.	92	p.ht	Hostinger
10.	89	cixx6.com	cixx6.com
Total	2601 Attacks		

Top-10 sub domains which are used for phishing are listed in above table: 1.1. Each subdomain service is effectively its own "domain registry" [1]. The subdomain services have many business models and unregulated services. Attackers use some TLD (Top Listing Domains) registries and registrars that can be implemented better anti-abuse policies and procedures.

Motivation

Because of the fake websites and breaching the individuals' security I am inspired to propose a system which provides the security against the phishing attacks and make safe browsing on internet. I would like to form a better system for secure browsing. There are many techniques for detecting fake websites like look-up system, classifier system, heuristics technique and hybrid systems. But drawbacks of every approaches decrease the efficiency of the system. Hence the solution for this is to ease of attack from the attacker and making the system secure.

Objective

The objectives are Generation of blacklist of the fake websites, Comparison with the blacklist, Classification of the websites, check for the Domain Registration, and extract the domain name and Examination of the proposed system.

Scope

Scope is to identify fake websites and giving them suggestions for their particular domain search and to make users safe internet browsing and secure them from sharing the confidential data with attackers.

II. LITERATURE REVIEW

Phishing

Phishing is the attack to the victim for stealing the confidential data of individuals such as usernames, passwords, account numbers and other details. Phishing is an immorally fraudulent practice that includes unlawful attempt to acquire sensitive information such as usernames, passwords and credit card details by masquerading as a dependable entity in an electronic communication [2]. Through the phishing attack, attacker can gain all the information of the victim and making the website insecure. Successful phishing detection system would distinguish any phishing websites from legitimate websites [3].

A common phishing attack is to obtain a victim's authentication information corresponding to one website and then use this at another site. This is meaningful attack given to many end users to reuse passwords exact or with only some modifications.

Phishing attacks create serious problem to the e-banking and the e-commerce websites. Through these websites attacker takes the user name, passwords and the account number of the customer and fetch the privileges of the customer. Both consumer and the financial organizations are at threat for huge amount of fake transactions via stolen data. The treat is rapidly increasing, the victims being consumers or users of financial or banking organizations, trading corporations, and supplier of internet services.

Anti-Phishing

Anti-phishing is to identify the phishing attack. There are many anti-phishing techniques are proposed. PhishNet is the anti-phishing technique which is based on the predictive blacklisting to detect phishing attacks.[4] This system works on the blacklisted URLs which are depended on the components which are proposed. The first component of PhishNet is matching of the five enumerated simple combination of the URLs with the known phishing sites. The second component is based upon the approximate matched URLs to discover new phishing sites. The new URL entry is checked individually with every component and evaluated the URL for phishing, whether the URL is phishing or not.

For Detecting phishing site sometimes the characteristics of the URLs are considered. In given URLs, checking for the symbols that how many times they are repetitive. In URL, how many times the particular symbol is taking place and bound the limit. If the number of that limit is crossed, the site is considered as phishing. After that checking the particular URL's page that if it contains 'name', 'password', 'login' many times it means it is the login page. So the login page must has secure connection '*https://..*', if that login page has the '*http://,,*' connection, it shows that the URL is phishing URL [5]. The other characteristic to find the fake websites is checking for the links. URL is checked into the Phish tank (phish tank: includes all Phishing sites [3]).

Techniques of Fake website Detection

Table 2 Techniques of Detecting Phishing sites

No.	Technique	Description	Limitations
1.	Bacterial Foraging Algorithm[5]	Classifies websites as per their characteristics as well as their content	Less accurate, Time consuming
2.	Statistical Learning Theory[8]	Spoofed and concocted websites are classified as their performance through SLT based classifiers	Difficult to identify concocted sites, Less accurate for fake websites
3.	Phishnet[4]	Identify websites as per their content matching as well as their DNS and URL	No updated blacklist
4.	WhoIs Feature[3]	Classifiers classify as per websites' URL and content feature	Lack of regularity content, No updated blacklist
5.	Finite State Machine[10]	Demonstrating behavior or responses with respect to input submissions and classifies as per different heuristics	Efficient for only web applications, Not reported suspected websites directly
6.	Visual Similarity[13]	Detect only exact same fake web pages by comparing images with registered database.	Lack of prior knowledge about priori knowledge about web pages, Can only classify exactly same fake web pages
7.	Blacklist Generator[7]	Generating blacklist as per Google's top-10 search and creating blacklist	Not accurate for recent websites Cannot get exact search

8.	Hierarchical clustering[9]	Automatic phishing categorization by extracting different features of websites	Less efficient, Not accurate as well others
9.	CANTINA+[24]	Filters phish sites using hash value as well as login form filtering	More computation needed, More time consuming, Attackers can compromise legitimate domains
10.	PageRank Based[20]	Classifies fake websites as per heuristics and Google's top-10 searches	More heuristics, Calculation is complex

III. PROPOSED WORK

Problem Formulation

Phishing attacks are based on identity theft of genuine website or webpage or web application. Phishing attack is going through below strategy:

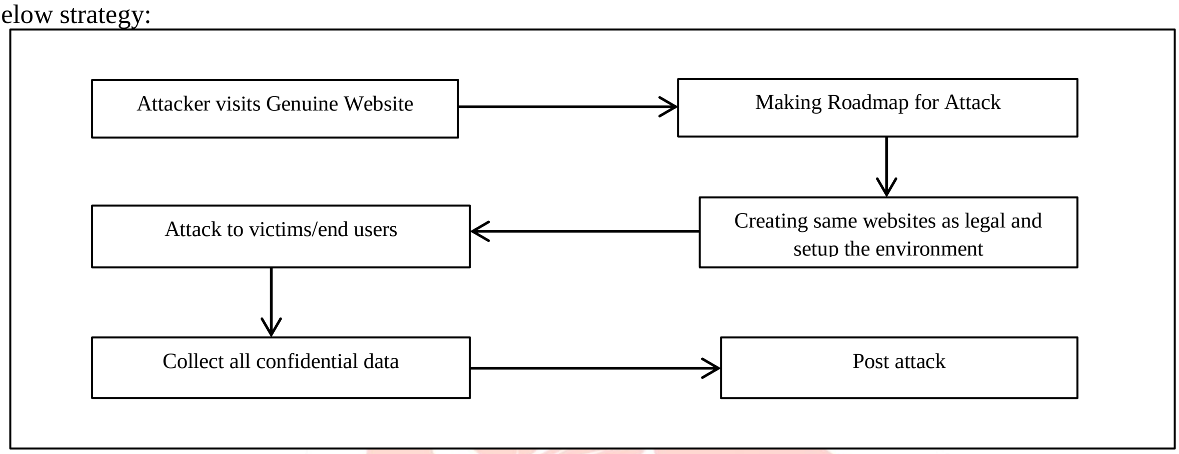
Attacker visits Genuine Website
Making Roadmap for Attack
Creating same websites as legal and setup the environment
Attack to victims/end users
Collect all confidential data
Post attack

Fig. 2 Flow of Attack

As per diagram, for phishing attack attacker visits genuine website and extract all the information about website. As per all information about webpages attacker creates same website for attracting the user. After creating phishing website attacker attacks to the end users and asks for confidential information about the user. Attackers gain all these personal information for misuse or for gaining privileges from the users.

There are three techniques to classify fake websites [6].

i. *Classifier Technique* in which blacklist is created and the heuristics are measured for classifying the fake webpages.
ii. *Lookup System* in which blacklist is created using IE Phishing filter and creating whitelist, too.
iii. *Hybrid System* in which combination of both techniques which are creating blacklist as well as whitelist and calculating heuristics.

Phishing solutions can be classified in below categories:

- *Blacklisting:* In this solution, comparing URL with the blacklist. If the URL matches with the blacklist then alert the user for threat.
- *Machine Learning:* It is for about creating white list as well as blacklist and giving the result about the fake website. It gives 100% true positive but cannot control false positive [5].
- *Heuristics:* In this approach classifies the URL's based heuristics and observing phishing sites but it is not giving guaranteed result for phishing sites like blacklisting.
- *Trusted Communication:* This technique is for authenticate the site for secure browsing.
- *Hybrid:* In this technique multiple features are combined for phishing site detection.

For phishing attack, attacker collects all the information about the genuine website and trying to create same as genuine website.

Proposed Solution

There are many approaches to detect phishing websites but every approach has some limitations like no updated blacklist for comparing phishing sites, less efficient, complex computations, more time consuming, not controlled false positives etc. To overcome all these drawbacks I propose an approach to detect phishing site and make safe internet browsing.

This approach is about finding URL's heuristics values and making the system effective and user friendly. Here, URLs heuristics are based on their characteristics like its registration, expiry date, validity etc. Then the URL is checked in Google's top-10 search and calculating its weight for the heuristic value and classifying the URL.

For classifying the websites there are three phases I have proposed:

i. After entering the URL checking into Blacklist

ii. Calculating Heuristics value
iii. Identifying website as per its Heuristics value

Phase-I, is to match URL for phishing directly to blacklist which is already generate before and updated as per requirement. Blacklist is generated as per Google's Top-15 search for particular popular domains and making blacklist though this search. If URL is matched with the blacklist, warn the user and giving suggestions about user's search by extracting its domain.

Phase-II is to calculate threshold based on weights of heuristics values. In this approach, considered heuristics are Google's PageRank, IP Address, Age of Domain, Dots and Suspicious URL. All this heuristics have different weights as per classification algorithm.

Phase-III is to identifying the website as per its weight of heuristic values. If the heuristic value is greater than or equals to threshold, URL is legitimate else URL is fake. For the fake URL, warn the user and giving suggestions as per its extracted domain through Google's Top-10 search.

Proposed Flow Chart

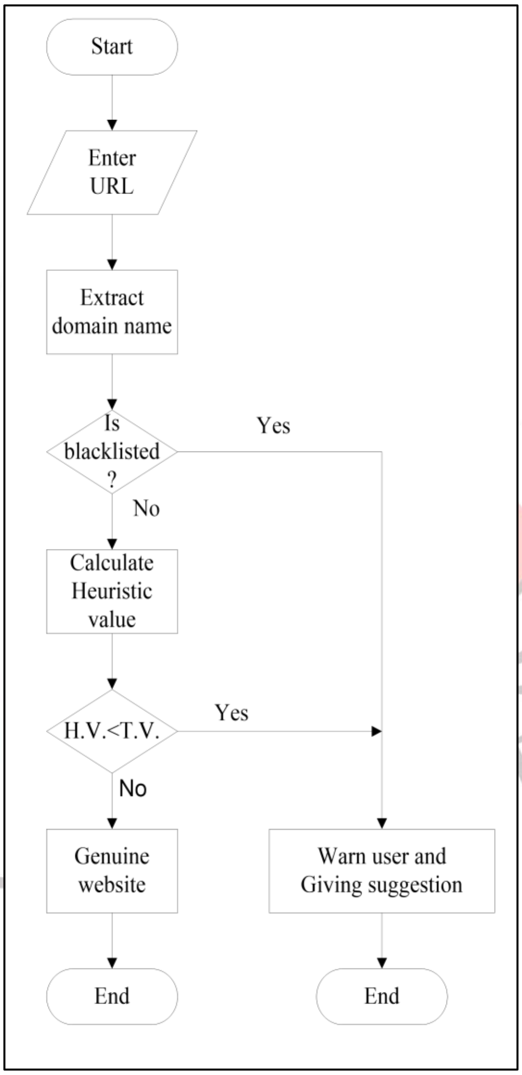

Fig 3 Flow Diagram

Functional Description

The proposed system is Heuristics based system in which identifying the characteristics of the URL and classify the website. Entered URL will be checked in blacklist, if the URL is within the blacklist, warn the user for the fake website. Otherwise calculate the heuristics values and match it with threshold value if the value is same as threshold value or greater than it, site will be genuine website else the site will be fake. For the fake website, extracting the domain name from URL and giving suggestions from Google Top ten searches.

User will enter the URL for identification of the website whether it is legitimate or not. Entered URL is matched with the blacklist. If the URL is within the blacklist, user will be warned for phishing site and giving suggestions for that particular search. If the URL is not matched with the blacklist, calculating heuristic values. After this, obtain the values of the heuristics. The values of heuristics values of GTR and age of domain are obtained by parsing the pages which will give these values and the values of suspicious URL and IP address are obtained by checking the URL.

- *Bad Forms:* In this heuristic checking for the form actions and how many links are within the particular form. Genuine websites have links which are similar to their home page or domain name.
- *Pop-ups:* Generally no more pop-ups into legitimate websites. Here, we have considered pop-ups no more than five.
- *Suspicious URL:* For heuristics, checking whether the URL is containing '@' or '-'. However, legitimate site rarely uses '-'. In the URL after '@' string will be considered before '@' string part is discarded. Heuristics will check whether these conditions are satisfied or not for phishing site. If the conditions are satisfied then the site is suspicious else declared as legitimate site.
- *IP Address:* URL contains IP address as its domain is checked by this heuristics.
- *Dots:* How much dots are within the URL will be checked by this heuristics. Normally, legitimate URL has less number of dots. Here, checking for minimum five dots within the URL. If there are more than five dots, the URL is considered as not legitimate site and calculating values as per this strategy.

Classification algorithm will be applied after obtaining the heuristic values on the training dataset to obtain the weights using a simple forward linear model described below,

$$S=\sum f(w_i*h_i) \qquad (1)$$

Where h_i is the result of each heuristic, w_i is the weight of each heuristics.
After this calculate the weight for each heuristics. For this, the higher the weight will be given to it.

$$W_i=(e_i/\sum e_i) \qquad (2)$$

Where e_i is the effect of each heuristic and will be calculated as per above (2) equation.

From the score of (1), S of the URL will be calculated. If the value of S is greater than the threshold then it is legitimate site else warn the user as a phishing site. If the value is not equals to threshold or less than the threshold site will be declared as a phishing site and warn the user for that. After this, for user convince giving suggestions to the users. If all these heuristics will be satisfied by any website then the page source of the web page will be compared with original webpage of Google's Top-10 searches. If it matches with the original website, then declared that URL as legal web page, else legal webpage. For suggestions extract the domain name and through good search engine giving suggestions to the users which are safe for browsing

Scope for Future Work

In this approach, we can detect the fake websites and providing Google's top five links about the search. Here, detecting fake website is based upon the different heuristic values and trying to get the best result. In future the technique can be proposed by removing or adding more heuristics to gain high accuracy rate to classify the fake websites as well as legal websites. These techniques can be implied by combined other techniques and develop hybrid technique to safe Internet browsing.

IV. RESULT

Here, proposed approach is giving proper result for detecting fake website. We have considered parameters for accuracy like, time taken, miss-leading, accuracy for genuine website, and defined threshold. Among these parameters we have monitored the other systems and calculating the accuracy of those systems. As shown in below diagram the accuracy of the proposed system is achieved.

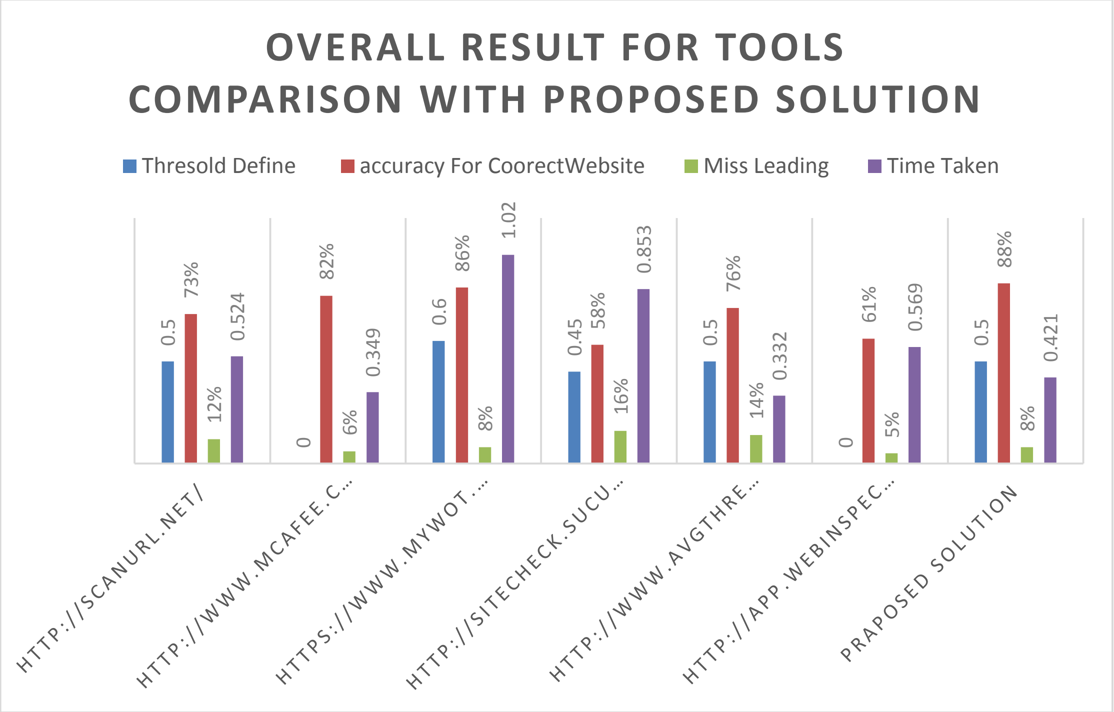

Fig 4 Comparison with other tools

As compared to other tools the proposed system has accuracy of 88% for genuine websites and taking time less as compared to others. Proposed solution has no more miss-leading websites.

V. CONCLUSION

Today phishing is very serious attack so it is necessary to detect it and making internet browsing secure. To detect fake websites there are many techniques are found like visual similarity, blacklist generator, PageRank, Bacterial Foraging algorithm etc. All this techniques have drawbacks like more time consuming, more computation etc. All these drawbacks affect the performance of the approach and giving less efficiency. So it is necessary to propose a new approach with high efficiency. This approach is based on classification of URL's heuristics based on their weights. First of all, matching the URL within the blacklist and identifies whether it is legitimate or not. After this calculate the weights of heuristics of the URL and identifying the website by classification algorithm. Here, used classification algorithm is simple linear model approach. Calculated heuristics will be compared with the threshold value. If the value is greater than the threshold, the URL will be declared as legitimate else warn the user. After warn the user giving some suggestions through extracting domain name and searching domain name with the help of good search engine. In this approach, I tried to overcome all these limitations and control false positives and making the users safe internet browsing.

Acknowledgement

Apart from the efforts of me, the success of any task depends largely on the encouragement and guidelines of many others. I take this opportunity to express my gratitude to the people who have been instrumental in the successful completion of this work. I would like to express my deepest gratitude to Shri AdityaSinha Sir, Team coordinator, C-DAC Pune, without his encouragement and guidance this work would not have materialized. I would like to show my greatest appreciation to my external guide Prof. Hasmukh Patel, HOD CE, LCIT, Bhandu. I can't say thank you enough for his tremendous support and help. I feel motivated and encouraged every time. I take immense pleasure in thanking Mr.Naresh Kumar Gardas, Course Coordinator, C-DAC, Mr.Jiggyasu Sharma, Project Engineer C-DAC and Mr.Ashwin Kevat, Coordinator, GTU for having permitted me to carry out this work and for all valuable assistance in this work.

Finally, yet importantly, I would like to express my heartfelt thanks to my beloved parents for their blessings, my friends/classmates for their help and wishes for the successful completion of this work.

REFERENCES

[1] Anti-Phishing Working Group, "Phishing Activity Trends Report, 1st Quarter, July 2013.

[2] Muhmmad Sheikh Sadi, Md. Mizanur Rahman Khan, Md. Merazul Islam, "Towards Detecting Phishing Web Contents for Secure Internet Surfing", published in IEEE 2012, International Conference on Informatics, Electronics and Vision.

[3] Insoon Jo, Eunjin (EJ) Jung, Heon Y. Yeom, " You're not who you claim to be: website identity check for phishing detection", IEEE 2010.

[4] Pawan Prakash, Manish Kumar, Ramana Rao Kompella, Minaxi Gupta, " PhishNet: Predictive Blacklisting to Detect Phishing Attacks", presented as conference at IEEE INFOCOM, 2010.

[5] Radha Damodaram, M.L. Valarmathi, " Bacterial Foraging Optimization for Fake Website Detection", TIJCSA, January 2013.

[6] Ahmed Abbasi, Hsinchun Chen, "A Comparison of Tools for Detecting Fake Websites", published by the IEEE computer society, 2009.

[7] Mohsen Sharifi and Seyed Hossein Siadati, "A Phishing Sites Blacklist Generator", IEEE 2008.

[8] Ahmed Abbasi, Zhu Zhang, David Zimbra, "Detecting Fake Websites: The Contribution of Statistical Learning Theory", MIS Quarterly 2010.

[9] Weiwei Zhuang, Qingshan Jiang, "An Intelligent Anti-Phishing Strategy Model for Phishing Website Detection", IEEE 2012, published in 32nd International Conference on Distributed Computing Systems Workshops.

[10] Hossain Shahriar and Mohammad Zu;kernine, "Phish Tester: Automatic Testing of Phishing Attacks",2010 IEEE, Fourth International Confenrence on Secure Software Integration and Reliability improvement.

[11] C.L. Lai, K.Q. Xu, Raymond Y.K. Lau, "Toward a Language Modeling Approach for Consumer Review Spam Detection", IEEE 2010.

[12] Ali Darwish, Ahmed El Zarka and Fadi Aloul, "Towards Understanding Phishing Victims' Profile", IEEE 2013.

[13] Masanori Hara, Akira Yamada, Yutaka Miyake, "Visual Similarity-based Phishing Detection without Victim Site Information", IEEE 2009.

[14] Tareq Allan, Justin Zhan, "Towards Fraud Detection Methodologies", IEEE 2010.

[15] Abdullah Alnajim and Malcolm Munro, "An Evaluation of Users' Tips Effectiveness for Phishing Websites Detection", IEEE 2008.

[16] Maher Aburrous, M.A. Hossain, Keshav Dahal, Fadi Thabtah, "Predicting Phishing Websites using Classification Mining Techniques with Experimental Case Studies", published in seventh international conference on Information Technology, IEEE 2010.

[17] Ahmed Abbasi, "A Comparison of Fraud Cues and Classification Methods for Fake escrow Website detection", Springer, Sept. 2009.

[18] http://www.brighthub.com/internet/security-privacy/articles/82116.aspx

[19] http://www.apwg.org/ Anti-Phishing Working Group

[20] A.Naga Venkata Sunil, Electronics and Computer Engineering Department, IIT Roorkee,"A PageRank Based Detection Technique for Phishing Websites", published on 2012 IEEE Symposium on Computer & Informatics.

[21] Peter Likarish, Don Dunbar, and Thomas E. Hansen, "B-apt: Bayesian anti-phishing toolbar", ICC 2008. IEEE International Conference, 2008, pp 1745-1749, May 2008.

[22] Zang Y., Hong J. and Cranor, L. "CANTINA: A Cloud-Based Approach to Detect Phishing Web sites". In Proceedings of the 16th World Wide Web Conference (WWW'07), May 2007, pp 639-648.

[23] Nuttapong Sanglersinlapachai and Arnon Rungsawang, "Using Domain Top-page Similarity Feature in Machine Learning-Based Web Phishing Detection", 3rd International Conference on Knowledge Discovery and Data Mining, 2010, pp 187-190.

[24] Gaung Xiang, Jason Hong, Carlyn P. Rose, "CANTINA+: A Feature-rich Machine Learning Framework for Detecting Phishing Web sites", ACM Journal Name, Vol. 3,May 2011, pp 567-599