International Journal of Scientific Research & Engineering Trends
Volume 3, Issue 4, July-2017, ISSN (Online): 2395-566X

An Approach to Detect Malicious URL through Selective Classification

Ghanshyam Sen*, Himanshu Yadav, Anurag Jain
Department of Computer Science & Engineering
Radharaman Institute of Technology and Science, Bhopal
*Email: gksen87@gmail.com

***Abstract* – A "malicious web page" refers to a web page that contains malicious content that can exploit a client–side computer system. Malicious website may be used as a weapon by cybercriminal to exploit various security threats such as phishing, drive-by-download and spamming. Malicious Web sites are hurdle on the way of Internet security. And used as a weapon to mount various security threat like phishing, drive-by-download and spamming. To handle there is need to develop an automatic system to recognized malicious website. To derive detection models for malicious web pages, distinguishing features of benign and malicious web pages are analyzed. This paper is a approach to find the malicious url using PSO approach.**

***Keywords* – Web Application, Prefetching, Web Mining.**

I. INTRODUCTION

Rapid growth of web application has increased the researcher's interests in this era. All over the world has surrounded by the computer network. There is a very useful application call web application used for the communication and data transfer. An application that is accessed via a web browser over a network is called the web application. Web caching is a well-known strategy for improving the performance of Web based system by keeping Web objects that are likely to be used in the near future in location closer to user. The Web caching mechanisms are implemented at three levels: client level, proxy level and original server level [1, 2]. Significantly, proxy servers play the key roles between users and web sites in lessening of the response time of user requests and saving of network bandwidth. Therefore, for achieving better response time, an efficient caching approach should be built in a proxy server.

Web caching and prefetching are the most popular techniques that play a key role in improving the Web performance by keeping web objects that are likely to be visited in the near future closer to the client. Web caching can work independently or integrated with the web prefetching. The Web caching and prefetching can complement each other since the web caching exploits the temporal locality for predicting revisiting requested objects, while the web prefetching utilizes the spatial locality for predicting next related web objects of the requested Web objects [1]. Prefetching is used as an attempt to place data close to the processor before it is required, eliminating as many cache misses as possible. Caching offers the following benefits: Latency reduction, Less Bandwidth consumption, Lessens Web Server load. Prefetching is the means to anticipate probable future requests and to fetch the most probable documents, before they are actually requested. It is the speculative retrieval of a resource into a cache in the anticipation that it can be served from the cache in the near future, thereby decreases the load time of the object.

II. WEB APPLICATION

Web Applications are software applications deployed by the World Wide Web. They use a single client-server model, and run in a Web browser on the client computer. Once a new release of a Web Application is installed on the server, this release is available to all users. This immediate deployment characteristic is probably one of the most powerful characteristics of a Web Application. There are different names in use for what here is called a Web Applications. Names in use are Web Sites, Web-based applications and Web Applications. Some authors are also using different names to indicate different types of Web Applications. In this article the term Web Application is used to represent all types.

III. WEB MINING

There is a lot of information on web pages and content links. These pages can be accessed by users, and the name of the network and, therefore, a new set of data records is created. These records contain the user access patterns. The techniques used to extract these records by the way the information discovery and identification of records. Therefore mining inputs come online in various fields, such as databases and data retrieval, machines and litigation instinctive acquisition talks

Drilling techniques can be classified into three types of network and

1. Web content mining,
2. Web structure mining, and
3. Web usage mining.

Web Content Mining (WCM)

The content of the web is of various types of data such as a combination of structured data, semi data structures

and unstructured data further this new data can be text, images, audio or video. And he asks the class of algorithms that detect useful information from this type of data, documents, Web content mining.

The main objectives of WCM include assistance in finding information, including user information on user profiles, and view the database in the WCM simulates the information on the network and the integration of a large number of issues complex. Many of the devices and provides a set of agents on the Internet by researchers to try and retrieve information and a higher level of abstraction of the semi-structured web by using data mining techniques data. Text and multimedia skills analysis data mining, drilling a useful topic in the numbering of the network. And some of these efforts are summarized as follows.

Agent–Based Approach

There are three agent based Web mining systems named as:

a. Intelligent Search Agents.
b. Information Filtering/ Categorization.
c. Personalized Web Agents.

a. Intelligent Search Agents

Agentive various network functions are built well informed until looking up the relevant universe using the characteristics of Knowledge Base and visions client to prepare and submit the data to ensure.

b. Information Filtering/Categorization

The use of agentive functions different ways of data recovery functions Network explicit readable text automatically activates to recover mechanically and evaluation.

c. Personalized Web Agents

Many agents learn interests of Internet users, according to Internet usage and preferences based on interests and patterns of discovery.

To derive detection models for malicious web pages, distinguishing features of benign and malicious web pages are analyzed. Nevertheless, these artifacts are under constant evolution. This evolution is typically because of two reasons.

- First one, cyber-criminals constantly revamp their strategies to craft attack payloads in malicious web pages not only aimed at making attacks more complex but also to evade existing countermeasures.
- Second, benign web pages evolve because of new content, new functionalities, or changes to the underlying technologies used in building the web pages.

Both types of the evolution impact the precision of the detection techniques, rendering detection models out-of-date, in turn, resulting in malicious web pages that escape detection. In this dissertation we proposed an approach that leverages best features for web environment by utilizing ranker based searching and optimization to align learning-based detection models with evolving new web page artifacts.

To achieve his goal of more precise analysis and detection of malicious web pages standard machine learning algorithms are used. These techniques are based on discriminative features extracted from: URL string, HTML content, JavaScript code, and reputation metadata of web pages on social networking websites and score parameter.

IV. PARTICLE SWARM OPTIMIZATION

Swarm Intelligence (SI) is an innovative distributed intelligent paradigm for solving optimization problems that originally took its inspiration from the biological examples by swarming, flocking and herding phenomena in vertebrates.

Particle Swarm Optimization (PSO) incorporates swarming behaviors observed in flocks of birds, schools of fish, or swarms of bees, and even human social behavior, from which the idea is emerged. PSO is a population-based optimization tool, which could be implemented and applied easily to solve various function optimization problems, or the problems that can be transformed to function optimization problems. As an algorithm, the main strength of PSO is its fast convergence, which compares favorably with many global optimization algorithms like Genetic Algorithms (GA), Simulated Annealing (SA) and other global optimization algorithms. For applying PSO successfully, one of the key issues is finding how to map the problem solution into the PSO particle, which directly affects its feasibility and performance.

A swarm is a large number of homogenous, simple agents interacting locally among themselves, and their environment, with no central control to allow a global interesting behavior to emerge. Swarm-based algorithms have recently emerged as a family of nature-inspired, population-based algorithms that are capable of producing low cost, fast, and robust solutions to several complex problems [1][2]. Swarm Intelligence (SI) can therefore be defined as a relatively new branch of Artificial Intelligence that is used to model the collective behaviour of social swarms in nature, such as ant colonies, honey bees, and bird flocks. Although these agents (insects or swarm individuals) are relatively unsophisticated with limited capabilities on their own, they are interacting together with certain behavioural patterns to cooperatively achieve tasks necessary for their survival. The social interactions among swarm individuals can be either direct or indirect [3]. Examples of direct interaction are through visual or audio contact, such as the waggle dance of honey bees. Indirect interaction occurs when one individual changes the environment and the other individuals respond to the new environment, such as the pheromone trails of ants that they deposit on their way to search for food sources. This indirect type of interaction is referred to as stigmergy, which essentially means communication through the environment [4]. The area of research presented in this depth paper focuses on Swarm Intelligence. More specifically, this paper discusses two of the most popular models of swarm intelligence inspired by ant's stigmergic behavior and birds'flocking behavior.

International Journal of Scientific Research & Engineering Trends
Volume 3, Issue 4, July-2017, ISSN (Online): 2395-566X

V. PROPOSED METHODOLOGY

The proposed works provide an evolutionary algorithm based machine learning approach for malicious URL classification. This work use BAT algorithm for features extraction forms all URLs from the URL list.

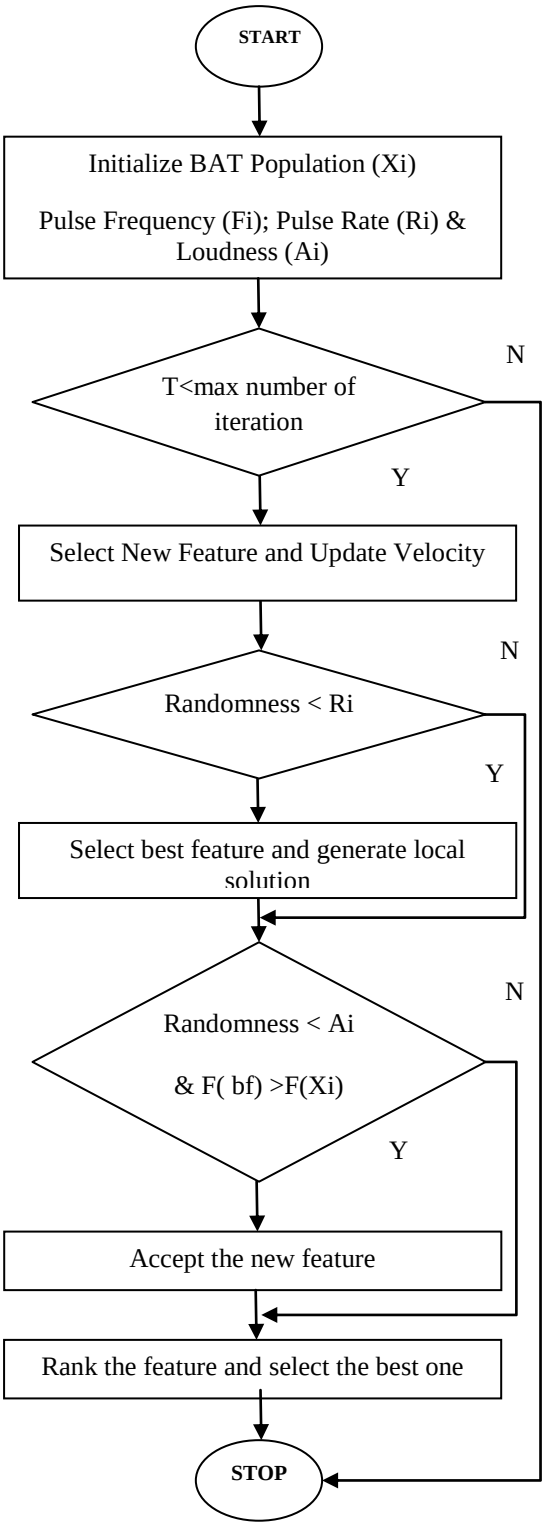

Fig.1. Flow graph

It includes identifying and determining the features based approach to classify different types of malicious URLs. The proposed approach employed the extracted data to Support Vector Machine classifier for classification of testing data to labels as benign, spam and phishing.

VI. PROPOSED ARCHITECTURE

Proposed framework initially use combine Web page data set contain Alexa [22] white listed URL, phishing URL provided by Phish Tank[21] and spam url [20]. In proposed framework initially initialized data set with random partition. Then BAT approach tends to generate random feature for malicious url. This BAT is recertifying through randomness. If randomness of relevant feature is low then it's acceptable otherwise feature is not to be consider. Acceptable Feature is denoted as relevant feature and apply for classify malicious. If malicious URL have high false negative rate then whole process is initialed by random partition and if malicious URL have low true positive rate then new feature is generated on same partition.

These chapters conclude the proposed technique for malicious URL classification that is based on one of the evolutionary algorithm known as BAT and apply SVM for classification. With the help of BAT we calculate the degree of uncertainty and again on the basis of randomness with the classify the Web URL, web url having higher entropy, is regarded as the "malicious", and generate the alarm.

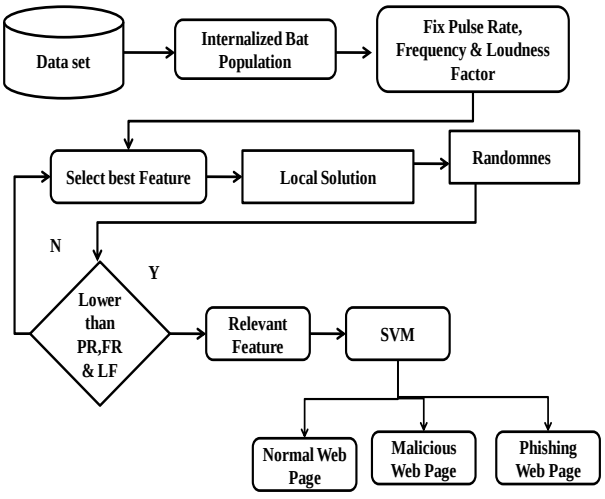

Fig.2. Proposed Architecture

VIII. RESULT ANALYSIS

Comparing to our algorithm (BAT) and PSO. The results of our algorithm's classification rate performance are shown in Table 6.2, and the results of PSO-SVM classification performance are shown in second column Table 6.2. We can find our algorithm outperforms than

PSO. And we can see our algorithm overcomes some problems existing in PSO.

Table 1: Comparing To BAT Algorithm And PSO

	BAT	PSO
TPR	95.23	80.95
FNR	4.76	4.76
Accuracy	90.47	76.19

8.1 Comparative Graph

The above figure is a comparative graph for the both methods with three parameters. The conclusion comes from this graph is the BAT gives the batter results from the PSO. Here the FNR is equal but other parameter shows the result is batter then PSO.

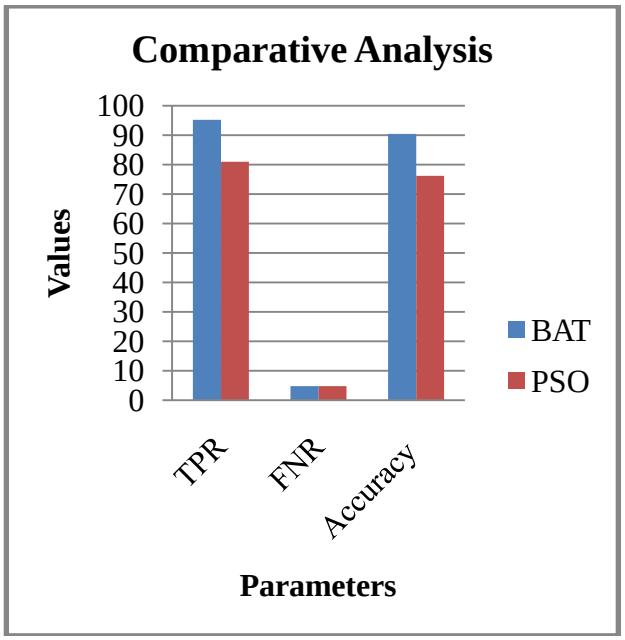

Fig.3. Comparative Graph

Accuracy is defined as the percentage of correctly classified result. In this dissertation accuracy is defined by the correctly classification of the URLs in their respective class i.e. Benign, spam and phishing. In terms of true positive and true negative accuracy is defined as:

$$\text{Accuracy} = \frac{\text{Total Positive}}{\text{Total Assessments}} = \frac{\text{TP} + \text{TN}}{(\text{P} + \text{N})}$$

Where

TP = Number of instances that are correctly predicted to their actual class.

TN = Number of instances that correctly rejected.

P+N = Total number of instances to be classified.

IX. Conclusion

Detection of malicious web has become a necessary and hot topic of research as numbers of internet users are increasing at a high pace. There are lots of challenges regarding this detection process. First the number of online URL is very large. Second web environment uses diverse platform and difficult to find security solution for them. Third now threats are become more and more complex and used various obfuscation techniques to bypass detection techniques. The existing detection techniques are focused only on single type of attacks only. New generated malicious web pages exploit multiple types of attacks for targeting the client. Cloaking type of attacks is difficult to detect because these web respond differently to browser and crawler. Size of web is a big challenge in the process. The proposedworks provide anevolutionary algorithm based machine learning approach for malicious URL classification. This work use BAT algorithm for features extraction forms all URLs from the URL list. It includes identifying and determining the features based approach to classify different types of malicious URLs. The proposed approach employed the extracted data to Support Vector Machine classifier for classification of testingdata to labels as benign, spam and phishing.

References

[1] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth “From Data Mining to KDD in Databases” pp. 0738-4602 1996.

[2] Thair Nu Phyu “Survey of Classification Techniques in Data Mining” Vol I Imecs 2009, March 18 - 20, 2009, Hong Kong.

[3] B.N. Lakshmi. #1, G.H. Raghunandhan. #2 “A conceptual Overview of Data Mining” Proceedings of the National Conference on Innovations in Emerging Technology-2011 Kongu Engineering College, Perundurai, Erode, Tamilnadu, pp.27-32. India.17 & 18 February, 2011.

[4] Han J. and M. Kamber (2000), Data Mining: Concepts and Techniques, Academic Press, San Diego, CA.

[5] R. Kosala and H. Blockheel, “Web Mining Research: A Survey”, In SIGKDD Explorations, Volume 2, Number 1, pages 1-15, 2000.

[6] P. Adriaans, D. Zantinge, “Data Mining” Addison Wesley Longman Limited, Edinbourgh Gate, Harlow, CM20 2JE, England. 1996.

[7] S. Chakrabarti, “Data mining for hypertext: A tutorial survey”. ACM SIGKDD Explorations, 1(2):1-11, 2000.

[8] M. Wu and M. Yang, "Privacy Preservation for Detecting Malicious Web Sites from Suspicious URLs," *2011 International Conference on Business Computing and Global Informatization*, Shanghai, 2011, pp. 400-403.

[9] Y. Fukushima, Y. Hori and K. Sakurai, "Proactive Blacklisting for Malicious Web Sites by Reputation Evaluation Based on Domain and IP Address Registration,"*2011IEEE 10th International Conference on Trust, Security and Privacy in Computing and Communications*, Changsha, 2011, pp. 352-361.

[10] D. Kent and L. M. Liebrock, "Statistical detection of malicious web sites through time proximity to existing detection events," Resilient Control Systems (ISRCS), 2013 6th International Symposium on, San Francisco, CA, 2013, pp. 192-197.

[11] L. Vu, P. Nguyen and D. Turaga, "Firstfilter: A cost-sensitive approach to malicious URL detection in large-scale enterprise networks," in*IBM Journal of Research and Development*, vol. 60, no. 4, pp. 4:1-4:10, July-Aug. 2016.

[12] S. B. Rathod and T. M. Pattewar, "A comparative performance evaluation of content based spam and malicious URL detection in E-mail,"*2015 IEEE International Conference on Computer Graphics, Vision and Information Security (CGVIS)*, Bhubaneswar, 2015, pp. 49-54.

[13] Z. Li-xiong*et al.*, "Malicious URL prediction based on community detection,"*Cyber Security of Smart Cities, Industrial Control System and Communications (SSIC), 2015 International Conference on*, Shanghai, 2015, pp. 1-7.

[14] M. S. Lin, C. Y. Chiu, Y. J. Lee and H. K. Pao, "Malicious URL filtering A big data application," *Big Data, 2013 IEEE International Conference on*, Silicon Valley, CA, 2013, pp. 589-596.

[15] H. K. Pao, Y. L. Chou and Y. J. Lee, "Malicious URL Detection Based on Kolmogorov Complexity Estimation," *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2012 IEEE/WIC/ACM International Conferences on*, Macau, 2012, pp. 380-387.

[16] M. K. K. Leung, A. Delong, B. Alipanahi and B. J. Frey, "Machine Learning in Genomic Medicine: A Review of Computational Problems and Data Sets," in*Proceedings of the IEEE*, vol. 104, no. 1, pp. 176-197, Jan. 2016.

[17] M. Nickel, K. Murphy, V. Tresp and E. Gabrilovich, "A Review of Relational Machine Learning for Knowledge Graphs," in Proceedings of the IEEE, vol. 104, no. 1, pp. 11-33, Jan. 2016.

[18] S. Gunduz, B. Arslan and M. Demirci, "A Review of Machine Learning Solutions to Denial-of-Services Attacks in Wireless Sensor Networks," 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), Miami, FL, 2015, pp. 150-155.

[19] N. M. De Mel, H. H. Hettiarachchi, W. P. D. Madusanka, G. L. Malaka, A. S. Perera and U. Kohomban, "Machine learning approach to recognize subject based sentiment values of reviews,"2016 Moratuwa Engineering Research Conference (MERCon), Moratuwa, 2016, pp. 6-11.

[20] "Spam URLs," [Online]. Available: http://www.joewein. de/sw/bl-text.htm. [Accessed 10 September 2016].