International Journal of Scientific Research & Engineering Trends
Volume 3, Issue 5, Sept.-2017, ISSN (Online): 2395-566X

Web URL Classification and Malicious Activities: A Review

Anshika Bansal
Scholar CSE, AITR, Bhopal
Email: anshika23bansal@gmail.com

***Abstract* – 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 Web Applications. 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. This paper throw some light on the different approaches can be apply in this manner.**

***Keywords* – Web Applications, www, Phishing.**

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.

The proposed work will focus on the web application in order to increase the server response. It is possible with the help of using the concept of web prefetching. The whole dissertation task is surrounding with the proposed concept. The implementation is done in Matlab R2010 which is a very popular language in this time.

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. CATEGORIZATION OF WEB APPLICATIONS

There are several ways to categorize Web Applications. The first categorization is found in[1]. Web Applications are divided into two groups: Web Application that have state, and that use some server-side logic and Web Sites that only have client-side logic. According to the terminology used in this paper, both are Web Applications. This two categories are rather broad and do not describe different feature of the large number of different types Web Applications with much precision. What are different types of Web Applications in other categorizations, will be the same type of Web Application in this categorization.

Here Web Applications are divided along two dimensions:

- The amount of control logic, and

- The amount of data processed.

The author of this categorization uses four different categories, but the number of categories could easily beex-tended.

- **Brochure** –no control logic and no data processed. An example for this category is a simple homepage.
- **Service oriented applications –** Some control logic, and a small amount of data processed. These sites are dedicated to provide services to its users, like email on the web (e.g. Hotmail).
- **Data intensive applications –** Web Applications that provide an interface to browse and query large quantities of data. An example is google.com.
- **Information system applications –** A mix of Service oriented applications and Data intensive applications (e.g. Amazon.com).

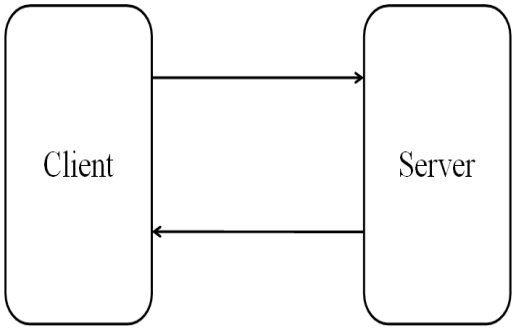

Fig.1. Client Server Architecture

A web application is any application that uses a web browser as a client. The application can be as simple as a message board or a guest sign-in book on a website, or as complex as a word processor or a spreadsheet.

A web application relieves the developer of the responsibility of building a client for a specific type of computer or a specific operating system. Since the client runs in a web browser, the user could be using an IBM-compatible or a Mac. They can be running Windows XP or Windows Vista. They can even be using Internet Exploreror Firefox, though some applications require a specific web browser.

Web applications commonly use a combination of server-side script (ASP, PHP, etc) and client-side script (HTML, Javascript, etc.) to develop the application. The client-side script deals with the presentation of the information while the server side script deals with all the hard stuff like storing and retrieving the information.

A Client Server architecture in which each computer or process on the network is either a client or a server. Servers are powerful computers or processes dedicated to managing disk drives(file servers),printers (print servers), or network traffic (network servers). Clients are PCs or work stations on which users run applications. Clients rely on servers for resources, such as files, devices, and even processing power.

Another type of network architecture is known as a peer-to-peer architecture because each node has equivalent responsibilities. Both client/server and peer are widely used, and each has unique advantages and disadvantages. Client-server architectures are sometimes called two-tier architectures.

IV. MACHINE LEARNING

Machine learning is a type of artificial intelligence that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of computer programs that can teach themselves to grow and change when exposed to new data.

The process of machine learning is similar to that of data mining. Both systems search through data to look for patterns. However, instead of extracting data for human comprehension -- as is the case in data mining applications machine learning uses that data to improve the program's own understanding. Machine learning programs detect patterns in data and adjust program actions accordingly. For example, Facebook's News Feed changes according to the user's personal interactions with other users. If a user frequently tags a friend in photos, writes on his wall or "likes" his links, the News Feed will show more of that friend's activity in the user's News Feed due to presumed closeness.

Machine learning and data mining are research areas of computer science whose quick development is due to the advances in data analysis research, growth in the database industry and the resulting market needs for methods that are capable of extracting valuable knowledge from large data stores. Machine learning is a set of tools that, broadly speaking, allow us to “teach” computers how to perform tasks by providing examples of how they should be done. For example, suppose we wish to write a program to distinguish between valid email messages and unwanted spam. We could try to write a set of simple rules, for example, flagging messages that contain certain features. However, writing rules to accurately distinguish which text is valid can actually be quite difficult to do well, resulting either in many missed spam messages, or, worse, many lost emails. Worse, the spammers will actively adjust the way they send spam in order to trick these strategies. Writing effective rules and keeping them up-to-date quickly becomes an insurmountable task. Fortunately, machine learning has provided a solution. Modern spam filters are “learned” from examples: we provide the learning algorithm with example emails which we have manually labeled as “ham” (valid email) or “spam” (unwanted email), and the algorithms learn to distinguish between them automatically.

V. 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

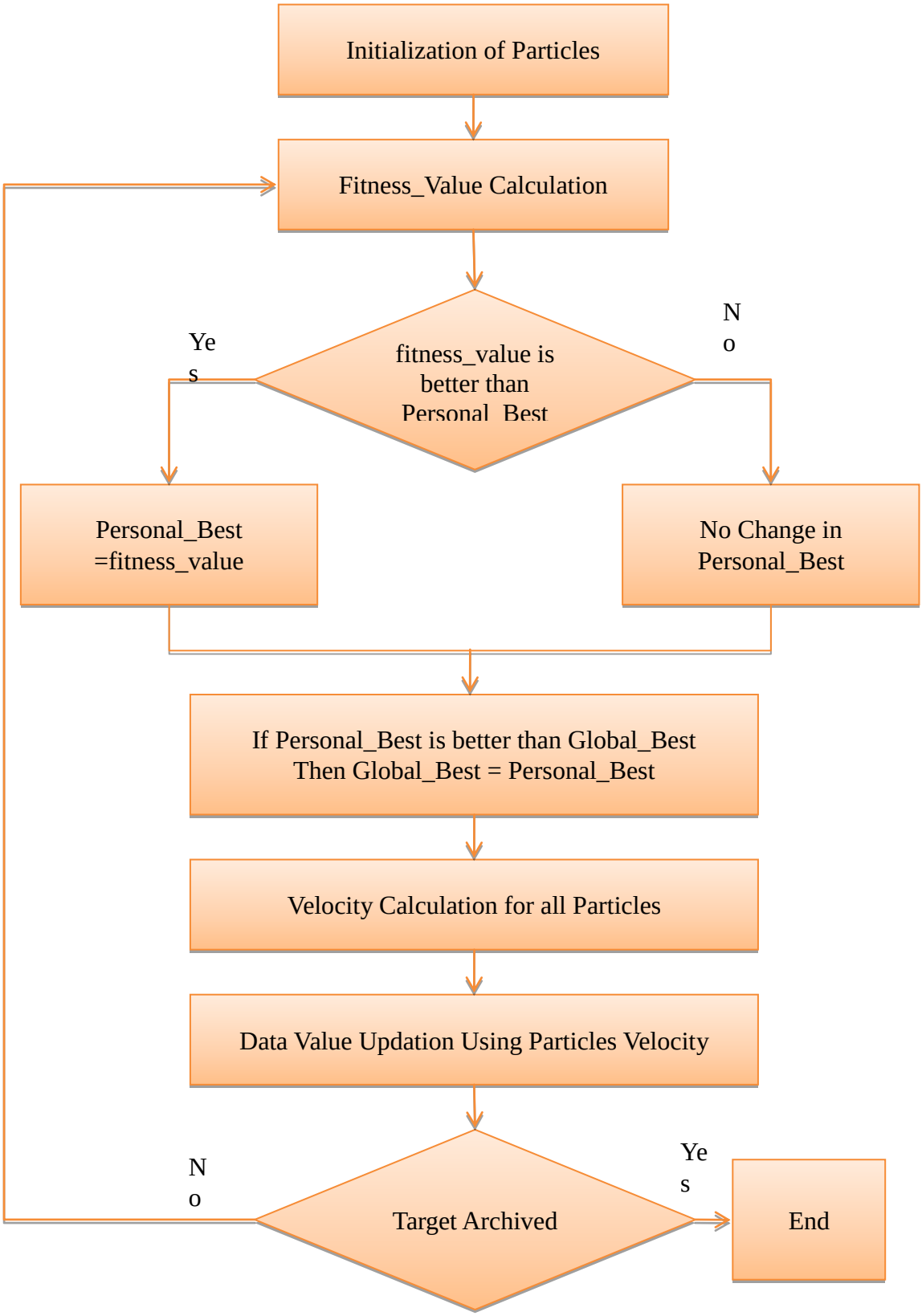

Fig.2. Procedures for particle swarm optimization

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.

Swarm intelligence models are referred to as computational models inspired by natural swarm systems. To date, several swarm intelligence models based on different natural swarm systems have been proposed in the literature, and successfully applied in many real-life applications. Examples of swarm intelligence models are: Ant Colony Optimization, Particle Swarm Optimization, Artificial Bee Colony, Bacterial Foraging, Cat Swarm Optimization, Artificial Immune System, and Glowworm Swarm Optimization. Here the primarily focus on two of the most popular swarm intelligences models, namely, Ant Colony Optimization and Particle Swarm Optimization.

1 Move in the same direction as your neighbor
2 Remain close to your neighbors
3 Avoid collisions with your neighbors
4 Particle Swarm Optimization (PSO) was originally inspired by the flockin behavior of birds. In terms of this bird flockin analogy, a particle swarm optimizer consists of a number of particles, or birds, that fl around and search space, or the sky, for the best location. The individuals communicate either directly or indirectly with one another search directions (gradients). Each of the particles in a swarm corresponds to a simple agent that moves through a multi-dimensional search space sampling an objective function at various positions. The best solution can be represented as a point or surface in the search space. Potential solutions are plotted in this space and seeded with an initial velocity. The motion of a given particle is dictated by its velocity which is continuously updated in order to pull it towards its own best position and the best positions experienced by the neighbors in the swarm. The performance of each particle is evaluated using a predefine fitnes function which encapsulates the characteristics of the optimization problem. Over time, particles accelerate towards those with better fitness values. PSO is a simple, but powerful search technique. It has few parameters to adjust and is easy to implement.

Many current models use variations on these rules, often implementing them by means of concentric "zones" around each animal. In the zone of repulsion, very close to the animal, the focal animal will seek to distance itself from its neighbors to avoid collision. Slightly further away, in the zone of alignment, the focal animal will seek to align its direction of motion with its neighbors. In the outermost zone of attraction Particle Swarm Optimization Bird flocking and fish schooling are the inspirations from nature behind particle swarm optimization algorithms. It was first proposed by Eberhart and Kennedy. Mimicking physical quantities such as velocity and position in bird flocking, artificial particles are constructed to “fly” inside the search space of optimization problems.

VI. LITERATURE SURVEY

Some criminals and malcontents try to take advantage of others by using malicious websites. As a result, many systems developed to prevent the end user to visit the malicious websites. It was a lot of methods applied in these systems, for example, it was built blacklisted by a number of techniques, including the guide reports, jars of honey, creeping web. Inevitably, not all malicious sites blocked. This points to the problem and has developed some client systems for the analysis of the content or behavior of a web site and visited it. But over the runtime can not be avoided. Compared with this approach, which is an effective approach to detect malicious sites. While this approach can get 95-99% accuracy, there is a need for private information. In this paper, it suggests a new strategy for the detection of malicious websites on the basis of privacy. We use structural and technical division of the unique value decomposition (SVD) to protect private information. (SVM) assessment is then used support vector machine. Our empirical results suggest that, compared with the new style of the original, the strategy has similar accuracy in the detection of a large number of malicious Web sites from your URL[8].

The objective of the creation of malicious software, and computers to obtain and perform malicious activities moves to show the attacker computer skills to earn money. Therefore, modern forms of malicious attackers infections take a more sophisticated and effective, including malware infection via malicious websites, as well as conventional farms, such as the spread of worms. Malicious Web sites trying to compromise machines attacked by loading a car that redirects users to sites that exploit and install malware necessarily on their computers by exploiting vulnerabilities in your web browser or additions solution. As a countermeasure to these malicious sites, blacklists for URLs, domains, which are great. However, attackers tend to change the URL addresses or domains within a short period to avoid the blacklist. Therefore, the blacklists system so that it can filter malicious Web sites is unknown is critical. In this paper, we first analyze the characteristics of malicious sites on the Internet your domain information, such as the (regime) AS, a block of IP addresses, IP address, domain registrar. Secondly, it is assessing the reputation of IP address blocks and recorders used by the attackers. Then blacklist system built from a combination of mass and registered a low reputation, which proposes an IP address, which is extensively used by the attackers. From the experimental results we have, and Web sites with the same combination with a low reputation appeared for a long time, pointing out that our proposed blacklist has a certain ability to filter malicious Web sites are not known.[9]

And it offers a new way to combine and add sporadic security incidents with Web browsing activity for the

production of new and expanded to storm with low false positive information. This method integrates the Internet, and events related to the storming of security, in the form of a series of time spaced irregularly, and then add to these common elements integrated through a population of equipment navigation events monitored. This compilation allows not only greater validation and knowledge about known security events, but also reveals the issue of security and the new activity is not known yet with very few false positives. This source of information-oriented allowed to take more effective defensive measures and increased security throughout the enterprise. Using data covering more than 24,000 computers and extends 6 months, and reflected the value of our approach. More importantly, the data show a decrease of 6.4 million Web for just 19 of the 10 areas of the Internet that require examination by a security analyst Due to our set of data in real-world application.[10]

Present Firstfilter, which new locations in terms of cost-sensitive that detects malicious Uniform Resource Locator (URL) in large-scale networks of seed companies. Firstfilter classified enter benign URL as it is known or malignant, and uses a cost matrix to identify the most important features and control model error in the classification. The matrix provides cost-effective tool to improve the key performance criteria for sensitive seeded cost. Rating Firstfilter widely with data sets that have been collected and an extensive network of enterprise network for a period of three months, covering 2015, from June to August 2015. The results of our evaluation to outperform always works Firstfilter and reduce the costs of other bilateral and classified.[11]

E-mail communication is growing rapidly. E- mail containing text and addresses of content. It may be the suspect text, from unwanted content that contains the addresses and locations required by the United Nations can be harmful, which redirects users to a Web phishing sites (malignant) sender. So to stop this spam activity detection and addresses of malicious sites that are beneficial to users by removing spam content and addresses of malicious sites required mail system. We have used the method to extract data such as supervised classification systems improves the accuracy and detects more than the amount of spam and addresses of malicious sites.[12]

VII. Conclusion

This paper is review of some of techniques which are used for detecting the malicious web pages. 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.

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 Conf.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.