Private and Secure Hyperlink Navigability Assessment in Web Mining Information System

Kavita Sharma

Department of Computer Science & Engineering,
Ambedkar Institute of Technology,
Geeta Colony, Delhi, India

Dr. Vishal Bhatnagar

Department of Computer Science & Engineering,
Ambedkar Institute of Technology,
Geeta Colony, Delhi, India

***Abstract*— Information explosion in World Wide Web has increased the interest in Web usage mining techniques in both commercial and academic areas. Study of interested web users; provide valuable information for web designers to quickly respond to their individual needs and for the efficient organization of the website. Among the several approaches like Association rule mining, classification, clustering, to extract knowledge from user's navigation data, this paper uses classification of log data to discover knowledge from web log files and extract knowledge mining from content and index page. In this paper, we present detection of Remote file inclusion Attack & create secure session; based on content page. At the end we proposed a model for Web Mining based secure Navigability Evaluation of Website.**

Key Words- Website Navigability, Web Usage Mining, Web Structure Mining, Remote file inclusion, Privacy.

I. INTRODUCTION

Now a day's internet is most emerging technology in the world. The use of internet needs to follow some specific protocol that is given by our service provider. A website is a collection of related web (World Wide Web) pages containing images, videos or other digital assets. Every website is hosted by at least one web server [13]. A web server is a program that, using the client/server model and the World Wide Web's Hypertext Transfer Protocol (HTTP), serves the files that form web pages to web users [12]. Whenever any web user surf that website user's some information is stored in web log which is in web server. Web log is storing information of the user activity which was performed on the website. Web log contain information about user name, IP address, Time Stamp, Access Request, Success Rate.

In this paper, we discuss secure website navigability and detecting the Remote File Inclusion (RFI) attack. Many programming languages helps us for writing a source code to detect that hacker/attacker was attempt the attack on the website or not. The code implement on the web log data for detecting RFI attack. Further, detecting RFI attack, filtering the data and creating the session, we get secure session. We represent secure website navigation. It is hard to detect hacker's attack by using signatures as far as security issues in website. The RFI is a good example of such as attack. We detect RFI attack in our website and ensure that our evaluation model will be very much secure and private.

The Remote file inclusion is a technique used to attack web applications from a remote computer. RFI attacks allow malicious users to run their own code on a vulnerable web server. The attacker succeeds in running malicious code on a web page by including code from a URL located on a remote server. When an application executes the malicious code it may lead to a backdoor exploit or technical information retrieval [3].

Every website is a collection of hyperlink. That's why we need to evaluate the hyperlink navigability and to evaluate the hyperlink navigability we need to understand a technique i.e. web mining. Web Mining is also useful to extract the information on the web [1]. Web Mining is based on knowledge discovery from web. Web Mining is categories into three parts:

1. Web Content Mining (WCM)
2. Web Structure Mining (WSM)
3. Web Usage Mining (WUM)

Now we identify what sort of problem occurred and how many problems were faced by the user while surfing the website. We propose a model for secure website navigability assessment with the help of web mining. The web log mining is representing as web usage mining then we secure web log file. We need to work on raw web logs, then to detect RFI attack and after filtration, we are able to get an enhanced session in a secure manner along with it an another parallel process also went on resulted secure website navigation through web structure mining resulted content page and structure of website representation, that was utilized by us to bring an effective result of secure surfing pattern which was further utilized with structure of website representation through applying matrix to evaluate result.

The rest of this paper is arranged as follows: Section 2 gives an overview about the background and related work in the area of web logs and website navigability. In section 3 the details of the model for the website navigation based on web mining Section 4 results of our model by applying any objective matrix. Finally, some conclusion and prospect are put forward in Section 5.

II. BACKGROUND & RELATED WORK

Web usage mining, the art of analyzing user interactions with a web page, has been dealt by several researchers using different approaches. Some researchers including [5], [6] have used classification algorithms for detecting web usage patterns. The authors [7] used similarity upper approximation clustering technique on web transactions from web log data to extract the behavior pattern of user's page visits and order of occurrence of visits.

Jyoti Pandey, et al [14] proposed data mining based service would run in background mode. The service computes the web pages likely to be requested by the user, considering their past web access log history, using association rules and thus optimizing the access time.

Nakayama, et al. [8] discovered the gap between the website designer's expectations and visitor behavior. Their approach uses the inter-page conceptual relevance to estimate the former, and the inter-page access co-occurrence to estimate the latter. They focus on website design improvement by using multiple regressions to predict hyperlink traversal frequency from page layout features.

Xiao Fang, et al. [4] found web mining based objective metrics for measuring web site navigability. This approach is described as the extent to which a visitor can use a website's hyperlink structure to locate target contents successfully in an easy and efficient manner.

When user visit on the web sites then their all activity is store on the server side in web log file. It stores all activity which is done by the user on the web sites.

If you want to discover whether your website is being attacked, hack attempt identifier can help with that. Remember that just because an attack occurred doesn't mean it was successful, but it's still useful to know what you're up against [9].

III. WEBSITE NAVIGATION MODEL BASED ON WEB MINING

The website navigation is the process of monitoring and controlling the usage of website. To evaluate the Efficiency, Power and Load of the website, we need to follow website navigation. Here, we propose a model for evaluating the given objectives through web mining. By examining web logs in our model, we can easily detect when and how many times our website was attacked by hacker.

Here we classified our model into three parts and each parts of model perform secure and private evaluation.

A. Secure Website Navigation through WSM

Data Mining is applied to usage information such as web server, logs proxy server etc. Its techniques are used to discover the usage patterns from website application. Proposed techniques predict the user behavior when it is interact with the web. WSM is representing as a graph. Web pages are representing as a node and hyperlink is representing as an edges.

In website navigation we required the web parsing technique to extract the hyperlink on the website. In the website classification we performed two specific tasks i.e. web content page and index page. The content page is show the text content of web pages and index page is pointing the all hyperlink of the web page. By applying

web parsing technique also evaluate the index page by extracting hyperlink that will represent the structure of website through WSM.

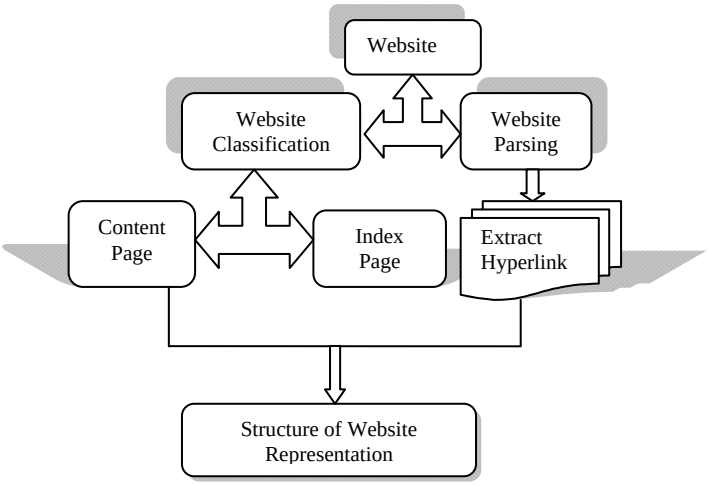

Fig. 1. Secure Website Navigation through WSM

B. Secure Website Navigation through WUM

We propose a model for evaluating website navigability that builds on the analysis and detect hacking detail of web logs and take into consideration both website structure and visitors' access/surfing patterns in secure manner.

The web log is store the activity which is done by the user on the website. The web log contains information about user name, IP address, time stamp, access request, number of bytes transferred, result status, URL that referred and user agent etc. as Common Log Format. The data present in the log file cannot be used as it is for the mining process. Therefore the contents of the log file should be performed these preprocessing step.

Web log file is store reside on the server. If user visits many times on the site then it creates entry many times on the server.

The contents of web log file are [11]:

1. Visiting Path – Paths which follow by the user to visit on the website.
2. User Name – Identify the user through IP address which provide by ISP. It is temporary address. Some website provides the facilities to create user own login user name and password whenever user access these site.
3. Success Rate- It is user activity which is done on the website that is Number of downloads and number of copies.
4. Path Traversed- The path identifies who is visit on the website through user.
5. Last visited Page- It store the last record that is visited by the user.
6. URL – It is may be HTML page and CGI program. This is accessed through the user.
7. Request Type- This is the Method which is performing on the website like GET and POST.

Where we filter that particular web log by using some specific language and remove the error which is store in the web log. This technique is removing all noisy error in the web log. Through web log file we detecting the hacking attempt of that website before filtering. We analysis and examine the RFI attacks.

The RFI attack is a type of vulnerability most often found on websites, it allows an attacker to include a remote file usually through a script on the web server. It performs following action on the website [9]:

1. **Client side Attack**- In this attack hacker/attacker change the client side data request with the help of java script like as Cross Site Scripting Attack.
2. **Server side Attack**- In this type attacker/hacker changes the data and adds something which is sent to visitor according to the user request with help of PHP.

3. **Denial of Service Attack**- In this attack without hacking password files or stealing sensitive data, the DoS attacker creates network congestion by generating a large volume of traffic in the area of the targeting system by hacker [2].
4. **Manipulation of data**- In this hacker/ attacker adds new data bit in the original data.

After filtering we are ready to create the session but we must be aware that our sessions are secured. Session is the process of keeping track of a user's activity across session of interaction with the website and that can be done by following Andrew W. Trieger [10]. Through this method, we will get the secure session tracking method.

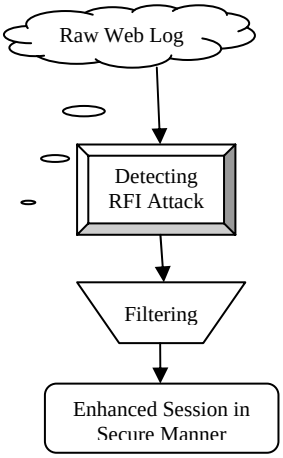

Fig. 2 Secure Website Navigation through WUM

C. Secure Website Surfing Pattern

In this model we represent the secure web surfing pattern. We gathered how to link one web page to other web pages and access the content of the web pages.

Here we apply the access log mining pattern and calculate the distance between one page to another web pages. In surfing pattern gathered information which pattern user visit on the website and which time access the content on the web pages to occur session. According to user access number of the web pages content then create number of session according to content page. Through the pattern mining in content page and session we get the surfing pattern.

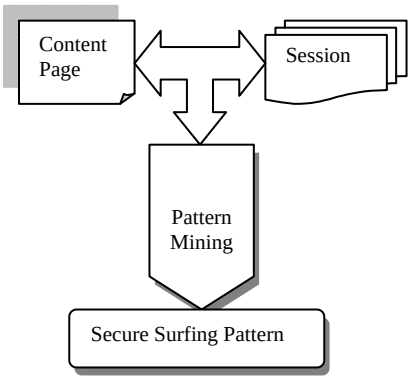

Fig. 3 Secure Website Surfing Pattern

These patterns enhanced by the structure of website and secure web session then get the result secure web surfing pattern.

IV. SECURE EVALUATION RESULT

Now we evaluate the secure result by applying matrix in structure of website and secure web surfing pattern as an input. Through this matrix we calculate the load, power and efficiency. In this characterized the probability of number of click and access number of hyperlinks on the website [4].

Efficiency is represent the user found the target content whereas load measure the problem when access the user targeted result on the website and decide the navigability direction in the course of access the target content. Power is represent the probability that user successfully access the target content.

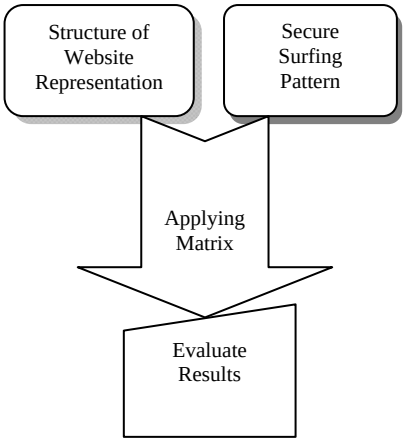

Fig. 4 Secure Evaluation Result

These are calculating according to user interaction on the website and using the identified prominent page access pattern and secure evaluation result.

V. CONCLUSION AND FUTURE WORK

This paper describe model to prevent on data loss in website navigability. Data loss can be occurred due to noise and RFI attack. By web log file we identify the user activities on the website. We propose a model for secure website navigability through web mining. To this end, the data prevention and detecting attack must be linked with security related technology in website navigation. Here, we get the secure evaluation result by applying objective matrix.

In future, we plan to implement this model with algorithm and mathematical solution. In addition to it, we plan to evaluate this model against real and synthetic datasets.

ACKNOWLEDGMENT

The authors would like to grateful for constructive suggestions and thoughtful comments from reviewers who improved the content of the paper.

REFERENCES

[1] Kavita Sharma, Gulshan Shrivastava, Vikas Kumar, “Web Mining: Today and Tomorrow” In Proceedings of the IEEE 3rd International Conference on Electronics Computer Technology, 2011.

[2] Gulshan Shrivastava, Kavita Sharma, Swarnlata Rai, “Technical Overview Dos and DDos Attack” in Proceeding of International Conference in Computing 2010, ACRS, Pp 274-282, 2010.

[3] Or Katz, “Detecting remote file inclusion attack”, Breach Security, Inc., May 2009.

[4] Xiao Fang, Michael Chau, Paul J. Hu, Zhuo Yang, “web mining–based objective metrics for measuring web site navigability”, Twenty-Seventh International Conference on Information Systems, Milwaukee 2006.

[5] M. Perkowitz, and O. Etzioni, “Adaptive Websites”, Communications of the ACM, Vol. 43, Pp. 152-158, 2000.

[6] A. Picariello, and C. Sansone, “A web usage mining algorithm for web personalization”, Intelligent Decision Technologies, Vol. 2, Issue 4, Pp. 219-230, 2008.

[7] K. Santhisree, and A. Damodaran, “Clustering on Web usage data using Approximations and Set Similarities”, International Journal of Computer Applications (ICJA), Published By Foundation of Computer Science, No. 4, Article 5, Pp. 27-31, 2010.

[8] T. Nakayama, H. Kato, and Y. Yamane, “Discovering the gap between website designers’ expectations and users’ behavior” In Proceding of the Ninth Int’l World Wide Web Conference, Amsterdam, May 2000.

[9] http://25yearsofprogramming.com/blog seen on March 2011.

[10] Andew W. Trieger, “Secure Session Tracking Method and System for Client-Server Environment” In United States Patent, Patent id: US006226750B1, May 1, 2001.

[11] L.K. Joshila Grace1, V.Maheswari2, Dhinaharan Nagamalai “Analysis of Web Logs and Web User in Web Mining“International Journal of Network Security & Its Applications (IJNSA), Vol.3, No.1, January 2011.

[12] James B. Lingan, http://whatis.techtarget.com seen on March 2011.

[13] http://en.wikipedia.org seen on March 2011.

[14] Jyoti Pandey, Amit Goel, Dr. A K Sharma, " A Framework for Predictive Web Prefetching at the Proxy Level using Data Mining", IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.6, June 2008.

AUTHORS PROFILE

Kavita Sharma, has obtained a degree of M.Tech. in Information Security from Ambedkar Institute of Technology, New Delhi after completing her B.Tech. & Polytechnic in the field of Information Technology. She is an acknowledged academic researcher and profilic author. She has contributed to numerous book, journal and conference publications in the area of Information Technology. She has participated in different National & International Workshop. Her area of interest includes Website Designing, Data Structure and Algorithm, Web & Cloud Mining and Data Security.

Dr. Vishal Bhatnagar, Associate-Professor (CSE), has obtained his Ph.d. degree in 2010 and has done M.Tech. (IT) from Punjab University in the year 2005 and completed his B.E. (CSE) from Nagpur University in the year 1999. He has more then 12 years of experience. His area of Interest is Database and Data Mining, Data Warehouse, and application of DWDM in business domain. He joined as an Assistant Professor (CSE) in the department of Computer Science and Engineering in Ambedkar Institute of Technology, Geeta Colony, Delhi. He is currently working as a Associate Professor and HOD (CSE Deptt.) in A.I.T., New Delhi.