KNOWLEDGE BASE COMPOUND APPROACH AGAINST PHISHING ATTACKS USING SOME PARSING TECHNIQUES

Gaurav Kumar Tak[1], Gaurav Ojha[2], Udit Kumar[3]

[1]School of Computer Science & Information Technology, Lovely Professional University, Phagwara, Punjab – 144402, India
gauravtakswm@gmail.com
[2,3]Department of Information Technology, Indian Institute of Information Technology and Management, Gwalior – 474010, India

ABSTRACT

The increasing use of internet all over the world, be it in households or in corporate firms, has led to an unprecedented rise in cyber-crimes. Amongst these the major chunk consists of Internet attacks which are the most popular and common attacks are carried over the internet. Generally phishing attacks, SSL attacks and some other hacking attacks are kept into this category. Security against these attacks is the major issue of internet security in today's scenario where internet has very deep penetration. Internet has no doubt made our lives very convenient. It has provided many facilities to us at penny's cost. For instance it has made communication lightning fast and that too at a very cheap cost. But internet can pose added threats for those users who are not well versed in the ways of internet and unaware of the security risks attached with it. Phishing Attacks, Nigerian Scam, Spam attacks, SSL attacks and other hacking attacks are some of the most common and recent attacks to compromise the privacy of the internet users. This paper discusses a Knowledge Base Compound approach which is based on query operations and parsing techniques to counter these internet attacks using the web browser itself. In this approach we propose to analyze the web URLs before visiting the actual site, so as to provide security against web attacks mentioned above. This approach employs various parsing operations and query processing which use many techniques to detect the phishing attacks as well as other web attacks. The aforementioned approach is completely based on operation through the browser and hence only affects the speed of browsing. This approach also includes Crawling operation to detect the URL details to further enhance the precision of detection of a compromised site. Using the proposed methodology, a new browser can easily detects the phishing attacks, SSL attacks, and other hacking attacks. With the use of this browser approach, we can easily achieve 96.94% security against phishing as well as other web based attacks

KEYWORDS

Internet attacks, Phishing, SSL, Cyber-crime, Knowledge Base, and Parsing.

1. INTRODUCTION

On the World Wide Web, Cyber-crime is one of the major security issues, troubling internet security. These crimes can be defined as immoral actions performed with the use of internet. They include illegal access of data, illegal interception of data, eavesdropping of authorized data over

Jan Zizka (Eds) : CCSIT, SIPP, AISC, PDCTA - 2013
pp. 33–40, 2013. © CS & IT-CSCP 2013 DOI : 10.5121/csit.2013.3604

an information technology infrastructure , data interference (which includes unauthorized damaging, deletion, deterioration, alteration or suppression of computer data), Unethical access of web services, Disturbance of social-peace, systems interference (interfering with the functioning of a computer system by inputting, transferring, destroying, removing, deteriorating, altering or suppressing computer data), misuse of devices, forgery (ID theft), and electronic fraud.[1][4]

A Knowledge Base is the modelling of previously occurred events in order to predict future events by employing some artificial intelligence techniques [13]. It is a sort of database for knowledge management, providing the means for the computerized collection, organization, and retrieval of knowledge. They are basically artificial intelligent tools providing intelligent decisions. Knowledge is obtained and represented using various knowledge representation techniques rules, frames and scripts. The basic advantages offered by such system are documentation of knowledge, intelligent decision support, self-learning, reasoning and explanation. [14]

As we all know that phishing attack is a URL based attack which happens between the Internet user and the browser, so our proposed methodology gives the new security layer between browser and the User using the Knowledge Base and some parsing operations.

2. LITERATURE REVIEW

Commonly, anti-phishing tools use two major approaches for mitigating phishing sites. The first approach is based on heuristics to check the host name and the URL for common spoofing techniques. The second method lists out some blacklist phishing URLs. The heuristics approach is not 100% accurate since it produces low false negatives (FN), i.e. a phishing site is mistakenly judged as legitimate, which implies they do not correctly identify all phishing sites. The heuristics often produce high false positives (FP), i.e. incorrectly identifying a legitimate site as fraudulent. Blacklists have a high level of accuracy because they are constructed by paid experts who verify a reported URL and add it to the blacklists if it is considered as a phishing website. [1][4][8]

Delayed password disclosure [15] is another new method to avoid phishing attacks. This is based on the feedback generated by the interface as user enters the password; hence if the feedback generated is not according to the authentic website an alarm is triggered.

Another method to create awareness amongst users against phishing is Trust bar construction [16]. This method associates logos with the public key of the website being visited hence easing the way of authentication of website. Passmark™ is a similar method currently being used by Bank of America.

The detection and identification of phishing websites in real-time, particularly for e-banking/payment gateway website, is a very complex and dynamic problem which involves many factors and criteria. Many methods like improving site authenticity, one time passwords, having separate login and transaction passwords, personalized e-mail communication, user education about phishing are being implemented to prevent phishing attacks, but they don't provide high security.

3. PROPOSED METHODOLOGY

Here we propose a knowledge base approach against phishing attacks which also uses some parsing techniques to detect the attack.

3.1. Knowledge Bases

Our methodology uses some knowledge bases which are I, T, A, B & C. Knowledge Base Initial or KBI stores the pattern and other detection methods of previously detected phishing attacks and other web attacks. It validates the URL and also relates the URL with the previously detected phishing attacks. If pattern of new URL matches with the previously stored Phishing attacks, then it generates a phishing alert before visiting the URL. Knowledge Base Trusted or KBT maintains all the trusted and secure URLs which are previously visited on the same browser. The user can further manually add the frequently visited legitimate websites to this knowledge base for whom he wishes not to carry out security checks every. Knowledge Base A defines all the URL-pattern based phishing and SSL attacks which have detected previously by the browser till date. This Knowledge Base is used before the operation of 'Parser-1'.

Knowledge Base B stores the all information (like license year, rating of the domain, popularity of the domain etc.) of the URLs which is previously visited and detected as the Phishing attacks. Knowledge Base C stores the result of Fraud Check analysis of URLs and generates queries when the URL is analyzed using fraud check with the previous history. Finally, Knowledge Base D is responsible for maintaining the history of the all the URLs which are previously visited.

3.2. Parsers

Some Parsers are also used in the detection of URL based attack in the proposed methodology. Parser 1 is used to detect the pattern based URL attacks. This parser provides the security against phishing attacks as well as SSL attacks. It also analyzes the usage of some special character (like '-','.' etc.) in the URL to detect the attacks. This parser's operation is based on the fact, that phishing attackers use the some fraction of the actual legitimate URL so as to generate a close to real phishing URL. Then we have Parser 2, which pulls out all the details of the website such as license year, rating of the domain, popularity of the domain etc. when a URL is passed through it. Using these details parser-B can declares if the URL is phishing website URL or a legitimate website URL. This parser takes account of the fact that phishing URLs are newly registered one with low rating and popularity. Hence if the URL is newly registered, then it can be a phishing attack on any existing URLs.

Parser 3 performs an important step for security against the phishing attacks. It performs the fraud check analysis of an URL and generates a warning message if URL is not secure. Parser 4 searches for other URLs whose pattern matches with the requested URL. It finds all details of the other similar URLs and compares all the details (like year of domain registration, rating of the domain, popularity of the domain etc.) with the requested URL details. It then displays all the results in the preference on the browser screen before visiting the requested URL.

In implementation of parser 4 and 5, the Open Source Crawler "crawler4j" has been used. [17]

3.3. Re-visit Policy

In the proposed methodology, the parsers also use the re-visit policy when needed because web has its dynamic nature. The re-visit policy can be easily understood using the freshness function described in two ways viz. Freshness and Age. Freshness is used as binary measure which indicates whether the local copy is accurate or not. The freshness of any page 'p' in the repository at time t is defined as:

$$F_p(t) = \begin{cases} 1 & \text{if } p \text{ is equal to the local copy at time } t \\ 0 & \text{otherwise} \end{cases}$$

Age is a measure which indicates how outdated the local copy of page is. The age of a page 'p' in the repository, at time t is defined as:

$$A_p(t) = \begin{cases} 0 & \text{if } p \text{ is not modified at time } t \\ t - \text{modification time of } p & \text{otherwise} \end{cases}$$

4. EXECUTION OF THE PROPOSED METHODOLOGY

Execution of proposed methodology depends on the sequence of knowledge bases and corresponding parsers. Final result of the proposed methodology is not affected by the sequence of the operations. Sequence affects only the space complexity and time complexity of the methodology.

Execution of proposed methodology in divided into several steps which are described as follows in the following sections

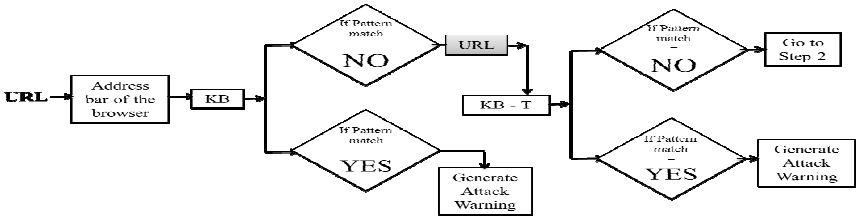

Figure 1. Flowchart of Step 1 of the proposed methodology

4.1 Historical Attack Detection

This step is composed with 2 operations which are occurred using 'Knowledge Base I' and 'Knowledge Base T'.

Knowledge Base Initial (KBI) is used to detect the attacks which has the same pattern with the previous detected attacks stored in it. Knowledge Base Trusted (KBT) is used to find the trusted status of requested URL which was previously declared by the user. In Historical attack detection the browser first tallies the URL with the KBI to check if its pattern matches that of any frequent phishing attack stored in the knowledge base. If it is safe then it proceeds to match up with the KBT. In this knowledge base it matches the URL against the trusted URLs stored by the user.

4.2 URL Pattern based Attack Detection

It is composed 2 operations which are related to 'Knowledge Base A' and 'parser 1'. This Step 2 provides the security against those attacks which are purely URL-pattern based phishing as well as SSL attacks. Knowledge Base A detects only those attacks which were detected previously by the browser and were stored in its database. During the step 2, 'parser 1' scans the requested URL and finds the occurrence of special characters ('-','.' etc) and their repetition in the URL. It is used to detect the pattern based phishing attacks. Working of step2 is represented in Figure 2.

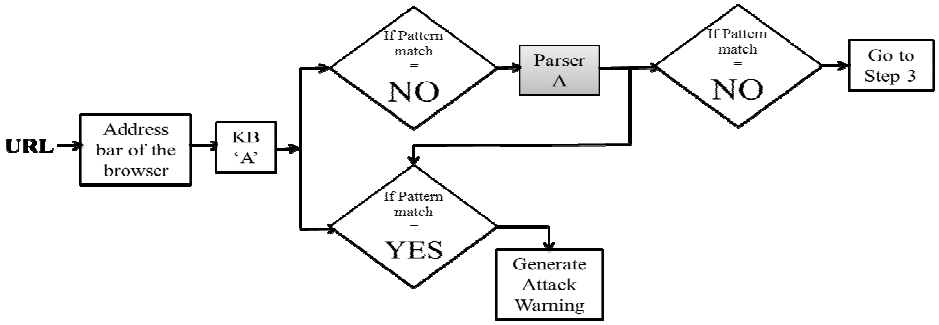

Figure 2. Flowchart of Step 2 of the proposed methodology

4.3 URL Information Analysis

URL information can be very helpful in detection of the phishing attacks. This step is based on the fact that the phishing URLs are newly registered and have lower rating and popularity over the internet. Figure 3 represents the working of URL information analysis step. In this step requested URL is analyzed with the Knowledge Base B and information of URL is analyzed using the historical data of URL (if the URL was visited previously) and displays the results and generates warning if URL is phishing attack based URL.

If the URL is not present in the history of Knowledge Base B then it goes to the parser 2 for the information analysis. 'Parser 2' works to finds the information of URL as a web crawler (which is described above) and performs the proper analysis after crawling for the details of the URL over the internet.

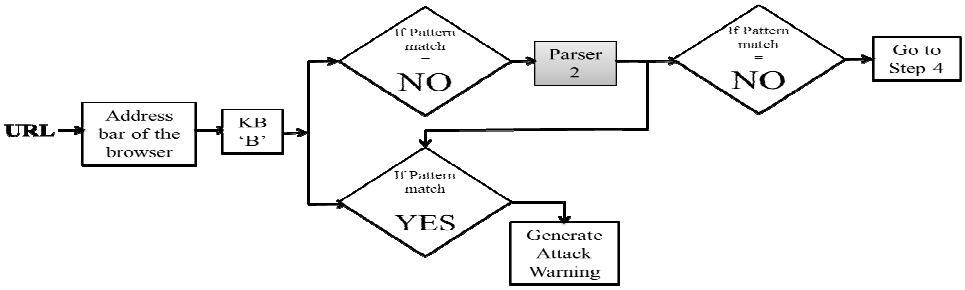

Figure 3. Flowchart of Step 3 of the proposed methodology

4.4 Fraud URL Detection

This step is performed by the Knowledge Base C and parser 3.Knowledge Base C performs the fraud check analysis of the requested URL (if it is available in the history of Knowledge base). It displays the result and appropriate messages. If the URL is not visited previously then parser 3 performs the Fraud check analysis to provide security against phishing attacks (or other web attacks) using some security algorithms. Figure 4 describes the fraud check analysis.

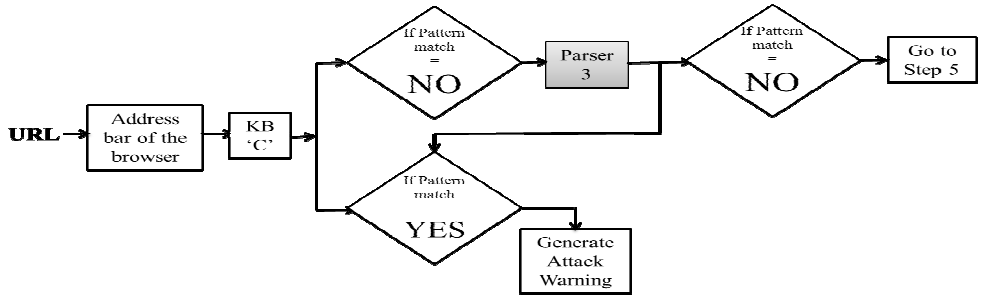

Figure 4. Flowchart of Fraud Check Analysis

5. IMPLEMENTATION AND RESULTS

We have implemented the proposed methodology in a virtual scenario, where we explored all the visited URLs of browsers on different machines using the history feature. All the URLs have been stored in a database for detect the phishing attacks and perform the analysis. We are planning to implement this methodology with some new add-ons to install in present web browsers (like other Firefox add-ons).

We have analyzed the URLs visited over the 5-months. In the initial stage of implementation, security risks are more because of absence of data in the different knowledge base. The implemented scenario provides 97.98 % security against phishing attacks and some hacking attacks. We have not executed our proposed methodology for the duration of Dec, 2012 and Jan, 2013 but during Feb, 2013 to April, 2013, we have executed the above methodology.

The following table data represents the recorded activities of the Web URLs in the implemented scenario and detected the phishing attacks and some hacking attacks.

Table 1: URL and some Web Attacks Analysis (2012-13)

Month	**Dec**	**Jan**	**Feb**	**Mar**	**Apr**
No. of URLs visited	978	967	897	1023	1218
Phishing Attacks	21	18	17	22	24
Detected phishing attacks with the browser	15	12	15	20	24
SSL Attacks	19	13	12	9	14
Detected SSL attacks with the browser	13	11	12	8	14
Execution Time (in minutes)	0	0	165	204	284

The above table shows the number of phishing attacks encountered and the execution time taken by our methodology from December 2011 to April 2012. The execution time for the first two months is actually zero as we have not implemented our methodology then. We have implemented our methodology from February 2012 onwards.

Kindly note that the approximate time of execution per URL visit, for the first month comes out to be 11 seconds. This increases to 12 seconds in the second month and to 14 seconds in the third month. This gradual increase can be attributed to the fact that the knowledge base is increasing in size hence the browser searches for more security attack then before

6. CONCLUSIONS

We have recorded the web URLs activities of with the usage of proposed methodology and without usage of proposed methodology over 5 months. From the data, we have analyzed the attacks and detected attacks over the time. Our system indicated that the 97.98% security against phishing attacks as well as SSL-attacks over the browsing. Table 1 represents the recorded data over the 5 months' time period. Limitations of the proposed method are that due to various parsing operations, its time complexity and space complexity is higher. So many times, it increases the browsing time of web browser. Due to slower speed of browsing, generally web users avoid this type of higher web security.

REFERENCES

[1] Ollmann G., The Phishing Guide Understanding & Preventing Phishing Attacks, NGS Software Insight Security Research

[2] Aburrous ,Maher Ragheb, Alamgir, Hossain,Keshav Dahal, Thabatah, Fadi, "Modelling Intelligent Phishing Detection System for E-banking Using Fuzzy Data Mining," cw, pp.265-272, 2009 International Conference on CyberWorlds, (2009)

[3] Abu-Nimeh, S.; Nair, S., "Bypassing Security Toolbars and Phishing Filters via DNS Poisoning," Global Telecommunications Conference, 2008. IEEE GLOBECOM 2008. IEEE , vol., no., pp.1-6, (Nov. 30 2008-Dec. 4 2008)

[4] Yu, W.D.; Nargundkar, S.; Tiruthani, N., "A phishing vulnerability analysis of web based systems," Computers and Communications, 2008. ISCC 2008. IEEE Symposium on, vol., no., pp.326-331, 6-9 (July 2008)

[5] Alnajim, A. and Munro, M. 2009. An Anti-Phishing Approach that Uses Training Intervention for Phishing Websites Detection. In Proceedings of the 2009 Sixth international Conference on information Technology: New Generations (2009). ITNG. IEEE Computer Society, Washington, DC, 405-410. DOI= http://dx.doi.org/10.1109/ITNG.2009.109

[6] Chen ,Juan and Guo ,Chuanxiong, Online Detection and Prevention of Phishing Attacks, in Proc. Chinacom 06

[7] Beginning PHP5, Apache, and MySQL Web Development by Elizabeth Naramore, Jason Gerner, Yann Le Scouarnec, Jeremy Stolz, Michael K. Glass; ISBN: 9780764579660

[8] Sophos White Paper, Phishing and the threat to corporate networks,(2005)

[9] PHP, AJAX, MySql and JavaScript Tutorials, http://www.w3schools.com/

[10] Prentice Hall - Deitel - Java How to Program, 4th Edition, Java_2_Complete_Reference_5E , Java - How To Program, 6th Edition

[11] Ahn ,Luis von, Blum, Manuel, Hopper, Nicholas, and Langford ,John. CAPTCHA: Using Hard AI Problems for Security. In Eurocrypt

[12] Gedam,Dhiraj Nilkanthrao,RSA Based Confidentiality And Integrity Enhancements in SCOSTA-CL, A thesis report,Department of Computer Science and engineering,Indian Institute of Technology ,Kanpur, India, (July,2009)

[13] J. Ullman, Database and knowledge base systems, In Database and knowledge-base systems, volume 2, Computer Science Press, 1989

[14] Akerkar RA and Sajja Priti Srinivas: "Knowledge-based systems", Jones & Bartlett Publishers, Sudbury, MA, USA (2009)

[15] Delayed password disclosure Markus Jakobsson, Steven Myers, Palo Alto Research Center, 3333 Coyote Hill Road, Palo Alto, CA 94303, USA. School of Informatics, Indiana University, Bloomington, IN, USA Journal: Int. J. of Applied Cryptography 2008 - Vol. 1, No.1 pp. 47 - 59

[16] Herzberg, A. and Gbara, A. (2004) Technical Report 2004-23, Protecting (even) Naïve Web Users, or: Preventing Spoofing and Establishing Credentials of Web Sites, DIMACS, October 30, Available at: http://dimacs.rutgers.edu/TechnicalReports/2004.html.

[17] The java project of the "crawler4j" can be downloaded from here: http://www.code.google.com/p/crawler4j/

Authors

Gaurav Kumar Tak is an assistant professor in the School of Computer Engineering, Lovely Professional University. His primary research areas of interest are Cyber-crime and Security, Wireless Ad-hoc Network and Web Technologies. He has written several research papers in these areas and continues to work for improving the security of web applications and making the web safe to surf.

Gaurav Ojha is a student in the Department of Information Technology, Indian Institute of Information Technology and Management Gwalior. His areas of interest are Web technologies, Open source software and Internet Security.