ISSN:0975-9646

Krutika Rani Sahu / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (4) , 2014, 5861-5864

An Enhanced Method for Spam Filtering Based on Basiyan Classifier

Krutika Rani Sahu
Shri Vaishnav Institute of Technology & Science
Computer Science Department

Abstract – **Phishing is an emerging type of social engineering crime on to the Web. Most phishes initiate attacks sending by emails to potential victims. The rapid development and evolution of phishing techniques pose a big challenge in Web identity security for researchers over all phishing attacks spoof users from the visual level and semantic level, and then make the appearances of web pages look similar to the real ones and make the web links and web page contents semantically related to the real ones. When emails lure users to access fake websites or unusualness website and induce them to expose sensitive and our private information In the Visual Assessment Approach, we detect phishing attacks using visual features of web pages. Study in demonstrated that visual assessment based detection is successful, but it fails in advanced phishing attacks This implies that unlike with other spam-filters, normal, personal, e-mail messages cannot be characterized as spam, since they do not confirm any of the two imperative rules. Also, newsletters confirm to only one of these two rules, thereby making it possible for newsletters to pass the filter as well, while all spam is flagged as being spam. This way Spam-Grid protects your system from unwanted e-mails while making sure the messages you do want arrive at their correct destination.**

Keywords – Content based spam filtering, phishing Detection and prevention, Bayesian classifiers, neural network.

I. INTRODUCTION

Spam-Grid functions like a proxy-server, meaning that your e-mail client will connect to Spam-Grid, which will in turn connect to your e-mail server. When e-mail messages are retrieved from the server, Spam-Grid will send a non-reversible encoding of those e-mails messages to the Spam-Grid server. The encoding is none reversible in order to protect the privacy of your messages. By analysing those encodings, the server will determine whether the e-mail distribution is a mass e-mail distribution. Once a distribution has been characterized as a mass distribution [10], its geographical distribution is analysed. Based on these two factors, a message (and subsequent messages of the same distribution) can be characterized as spam.

This implies that unlike with other spam-filters, normal, personal, e-mail messages cannot be characterized as spam, since they do not confirm any of the two imperative rules. Also, newsletters confirm to only one of these two rules, thereby making it possible for newsletters to pass the filter as well, while all spam is flagged as being spam [6]. This way Spam-Grid protects your system from unwanted e-mails while making sure the messages you do want arrive at their correct destination.

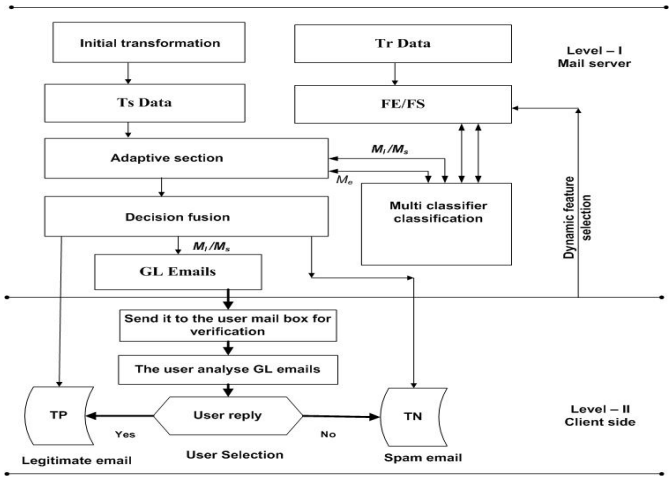


Fig 1 grid Base phishing

Grid based network executes its technology to discover the DNA of recent spam thereby gathering the know-how to come up with different techniques and applications to destroy them permanently. An accumulation of DNA of all spam are collected from the cyber world and being processed to develop the anti-spam and is sent to the server from where the client-machines are able to download to their systems and safeguard them from any external attacks.

II. BACKGROUND STUDY

1. Phishing Attacks Threat Modelling

Although phishing attacks have caused serious financial damage and have reduced users' confidence in the security of e-commerce, there is still a lack of methods to systematically analyze a given user authentication system for both system and user side vulnerabilities. As far as I am aware, the only published work that analyses the usability vulnerabilities of a system is by Josang et al. [10].Josang et al. present four "security action usability principles":

1. Users must understand which security actions are required of them.
2. Users must have sufficient knowledge and the ability to take the correct security action.
3. The mental and physical load of a security action must be tolerable
4. The mental and physical load of making repeated security actions for any practical number of instances must be tolerable.
5. The mental load of deriving the security conclusion must be toler- able.
6. The mental load of deriving security conclusions for any practical number of instances must be tolerable.

2. Timing Attack Techniques

Victims' names, with which they have bank account,[3] have described two timing–attack methods which can be used to obtain private information and have discussed methods for writing web application code that resists these attacks.

3. Web Page Attack

Most phishers would create web pages that are visually similar (or identical) to the targeted ones to increase the disguising level. They can simply download the web pages from the real websites or manually recreate them. We found most of the early phishing web pages are very similar to the real web pages by investigating their appearances and HTML code. Hence it is possible to detect phishing web pages by evaluating the visual similarity of web pages. Liu et al proposed the HTML based visual similarity assessment in [8] to address this problem. They have a system running in an internal server of City University of Hong Kong [5]. However, this system cannot detect phishing web pages that are created with different source codes, because Flash, visual components can be embedded into the web pages instead of HTML. Therefore, a real web page can correspond to countless fake web pages with different coding Appendix A demonstrates that visually identical web pages can be composed by totally different coding. Homographic web page is one of our major motivations of investigating into the phishing detection method at the graphical level.

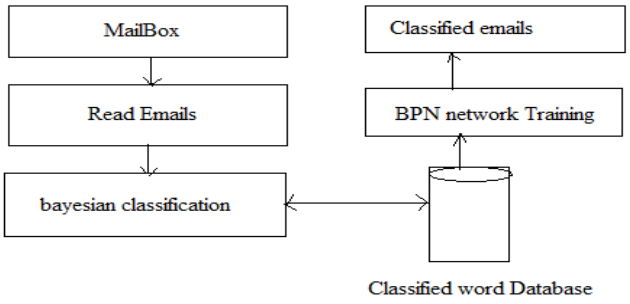
4. in-Session Phishing Attacks

The phishing attacks which have been described so far all need to actively engage users via a communication channel. In–session phishing [9], a more recently reported type of attack, uses a more passive mode, and yet is still very effective.

III PROPOSED WORK

In this proposed work we study around various phishing technique and we conclude as the main root of a phishing is going through spam emails.

- Filtering of spam (rate > 99 %)
- Limited number of false positive, efficiently managed anyway in any case (by notifying the sender and the recipient)
- Secure your company and maintains the integrity by filtering out the necessary unwanted e-mails thereby your mail server is not directly visible any more from the internet
- It integrates directly into your existing environment, without parameter setting
- It does not over-load any more your connection and your servers,
- Service Level Agreement (SLA) on filtering quality and service availability so that the business profitability is consistently maintained.



- Figure 2 proposed model

Accuracy

Accuracy of the system is estimated in different time during various different conditions and web users at the different time duration.

$$Accuracy(\%) - \frac{Total\ correctly\ classified}{Total\ provided\ samples\ to\ classify} X100$$

Error rate

Error rate reflects the performance of the system which is not correctly classified during algorithm execution of learning. That is achieved by using the below given formula.

$$Error\ rate(\%) = \frac{total\ incorrectly\ classified}{Total\ provided\ samples\ to\ classify} X100$$

Or

$$Error\ rate(\%) = 100 - accuracy$$

The accuracy of the system is evaluated using the results that are evaluated during the selection of the omitted results after result processing.

Memory

In this section of the performance evaluation we provide the two cost parameters for the system which is used to evaluate the time and memory constrain, this system covers the memory consumed by the system and time required to develop the model.

Memory is the total memory cost which is used to successfully execution of the system is given using the peak value consumed by the system.

IV. IMPLEMENTATION AND RESULTS

This section of the paper provides the key factors of implementation and results evaluation. The proposed scheme is implemented using the visual studio dot net technology which is a developer friendly development environment.The visual studio contains a rich class library and programmer can easily embed these techniques using code.

The machine learning technique is used for classify the mails in legitimate and spam category. The given model is therefore first learn from previous experience and after training provides the decision for classification of emails.

To implement the complete given techniques we first implement a simple Bayesian classifier to extract the email contents and evaluate the word frequency and probability to get how much frequent a word is occurred in spam mails and in a legitimate mail. if a word appears more frequently in a spam mail than legitimate mail then a combined probability of combination of different words are estimated which is find in both kinds of mail.

After evaluating the words and their probability to be in spam and in legitimate, a training session is conducted using the neural network to train with words and their corresponding probability of the words. After training of the neural network system is ready to classify the mails in two classes' first spam and second legitimate.

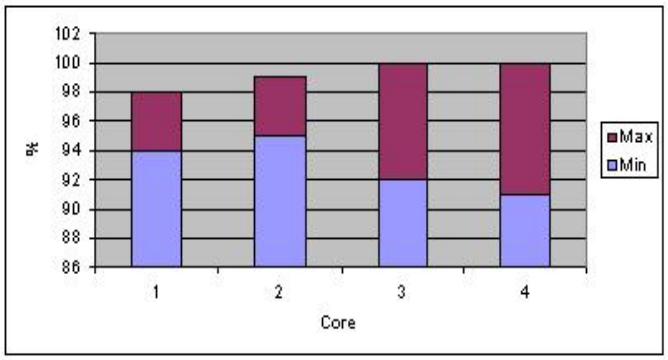


Figure 3 shows the memory of the system

And the time consumed by the system is given using the below given graph and figure.

V. CONCLUSION AND FUTURE WORK

- Adapted easily with entire software usable for anti-spam.
- Appliance at server side done with less complication.
- Restrict all undesirable network traffic caused by unsolicited or junk e-mails from penetrating user's mail in-box.
- Barring the unwanted e-mails by make allowance for predetermined filters.
- "Friends" list ensures your friend's e-mails are never blocked.
- Initialized when computer load.
- Suitable with other e-mail software like SMTP, POP3, MIME etc.
- Appropriate with other distributed computing environment like Lotus, Microsoft Outlook likewise.
-

There were different research directions for advancement of this work:

(i) Distributed computing – As computing paradigm also changes from centralized to distributed computing for parallel working. This technique of quick searching and discarding of such useless mails save the network jamming and stop the floating of junk e-mails. It utilizes ideal computational power to counter spam thereby reducing the actual power needed.

(ii) Database upgrades automatically – probability count of spam words at database should maintained regularly and count must be increment at regular interval of time.

Up gradation and the advancement in this field always gain the consideration of both the researchers and the spammers. And the level of distributed computing like grid computing, peer to peer (p2p), social networking and the web semantics shows furtherance and facilities for utilization of resources.

	Email	Email	Email	Email
Exe Time	1.4ms	1.5ms	1.4ms	1.4ms
Comp Time	0.4 ms	0.10ms	0.4ms	0.09ms
Comm Time	1ms	0.5ms	1ms	0.5ms
Speed Ratio	50%	19%	50%	20%
C/C	0.4	2	0.4	1.8
Time Complex	3	3	3	3
Cost	1.3	1.5	1.4	1.4

REFERENCES

[1] Theodore S. Rappaport, "Wireless Communications: Principals and Practice," 2nd ed., Pearson Education (Singapore) Pte. Ltd., India, 2002.

[2] C. Berrou, A. Glavieux and P. Thitimajshima, "*Near Shannon limit error-correcting coding and decoding: turbo code*," Inter. Conf Commun., pp.1064-1070.1993.

[3] D.C. MacKay, "*Near Shannon limit performance of low density parity check codes*," Electronics Letters, Vol. 32, pp. 1645–1646, Aug. 1966.

[4] M. K. Simon and M.-S. Alouini, "*Digital Communication over Fading Channels: A Unified Approach to Perform Analysis*," John Wiley & Sons, 2000.

[5] A. F. Naguib and R. Calderbank, "*Space-time coding and signal processing for high data rate wireless communications*," IEEE Signal Processing Magazine, Vol. 17, No. 3, pp. 76-92, Mar. 2000.

[6] C. Abad. The economy of phishing: A survey of the operations of the phishing market. First Monday, 10(9), 2005.

[7] K. Albrecht, N. Burri, and R. Wattenhofer. Spamato - An Extendable Spam Filter System. In 2nd Conference on Email and Anti-Spam (CEAS), Stanford University, Palo Alto, California, USA, July 2005.

[8] P. Behera and N. Agarwal. A confidence model for web browsing. In Toward a More Secure Web - W3C Workshop on Transparency and Usability of Web Authentication, 2006.

[9] T. Betsch and S. Haberstroh, editors. The routines of decision making.Mahwah, NJ ; London : Lawrence Erlbaum Associates., 2005.

[10] A. Bortz and D. Boneh. Exposing private information by timing web applications. WWW '07: Proceedings of the 16th international conference on World Wide Web, pages 621–628, 2007.

[11] J. Brainard, A. Juels, R. L. Rivest, M. Szydlo, and M. Yung. Fourth-factor authentication: somebody you know. In CCS '06: Proceedings of the 13th ACM conference on Computer and communications security, pages 168–178, New York, NY, USA, 2006. ACM.

[12] K. Cameron and M. B. Jones. Design rationale behind the identity metasystem architecture. Technical report, Microsoft, 2006.

[13] W. Cathleen, J. Rieman, L. Clayton, and P. Peter. The cognitive walk- through method: A practitioner's guide. Technical report, Institute of Cognitive Science, University of Colorado, 1993.

[14] M. Chandrasekaran, R. Chinchain, and S. Upadhyaya. Mimicking user response to prevent phishing attacks. In IEEE International

Symposium on a World of Wireless, Mobile, and Multimedia networks, 2006.

[15] Pierre Viland, Gheorghe Zaharia and Jean-Francois Helard, "*Coset Partitioning for the 4- PSK Space-Time Trellis Codes*," IEEE Conference on "Signals, Circuits and Systems, 2009.

[16] N.Kumaratharan, S.Jayapriya and P.Dananjayan, "STTC *based STBC Site Diversity Technique for MC-CDMA system*," IEEE Second International Conference on Computing, Communication and Networking Technologies, pp. 1-5, 2010.

[17] Pierre Viland, Gheorghe Zaharia and Jean-Francois Helard, "*Improved Balanced 2n-PSK STTCs for Any Number of Transmit Antennas from a New and General Design Method*," IEEE Conference on Vehicular Technology, pp. 1-5, 2009.