Anti-phishing using Big Data

Vanita Khamkar
Computer Engineering
Saraswati COE,
Kharghar, Navi Mumbai

Payal Ingale
Computer Engineering
Saraswati COE,
Kharghar, Navi Mumbai

Dhanraj Walunj
Assistant Professor
Saraswati COE,
Kharghar, Navi Mumbai

ABSTRACT
Now a day's phishing attack has become one of the most serious issues faced by internet users, organizations and service providers. In phishing attack attacker tries to obtain the personal information of the users by using spoofed emails or by using fake websites or both. The internet community is still looking for the complete solution to secure the internet from such attacks.

The users will be victim for this kind of activities, because phishing web pages looks very similar to real ones, so finds difficult to distinguish between the fake website and ones, detecting this kind of webpage is very difficult because for identification it takes several attributes into consideration which user might not knowing those things. The existing phishing detection systems are highly dependent on database and they are very time consuming also. In this proposed system, Hadoop-Map Reduce is used for fast retrieval of URL attributes, which plays a key role in identifying phishing web pages and it is known for its time efficiency and throughput also can gained using this.

Keywords
Phishing, Anti-Phishing, Hadoop, Map Reduce, Information Retrieval, Data Mining.

1. INTRODUCTION
Phishing is the process used to acquire sensitive information such as username, password etc. through spamming or other deceptive means. Phishing often takes place in email spoofing or instant messaging. Phishing email contains messages like ask the users to enter the personal information so that it is easy for hackers to hack the information. Website based attack continued to generate billions of dollars in fraudulent revenue of expense of individual user and organization.

Phishing is like fishing in a lake, but instead of trying to capture fish, phishers attempt to steal the personal information of the user. Phishers create fake web pages that are meant to steal individual's information without brining notice to the user. Unknowingly he will give his information to the phishers.

Till today so many techniques are used for detecting the phishing websites mainly include authentication, filtering, attack tracing and analyzing, phishing report generating, and network law enforcement. There are many algorithms of data mining, machine learning, image processing is used for detecting the phishing algorithms effectively.

Because the phishing problem takes advantage of human ignorance or naivety with regards to their interaction with electronic communication channels (e.g. E-Mail,HTTP, etc...), it is not an easy problem to permanently solve. All the proposed solutions attempt to minimize the impact of phishing attacks. From a high-level perspective, there are generally two commonly suggested solutions to mitigate phishing attacks:

• User education; the human is educated in an attempt to enhance his/her classification accuracy to correctly identify phishing messages, and then apply proper actions on the correctly classified phishing messages, such as reporting attacks to system administrators.

• Software enhancement; the software is improved to better classify phishing messages on behalf of the human, or provide information in a more obvious way so that the human would have less chance to ignore it. The challenges with both approaches are:

• Non-technical people resist learning, and if they learn they do not retain their knowledge permanently, and thus training should be made continuous. Although some researchers agree that user education is helpful many other researchers disagree.

• Phishing is a semantic attack that uses electronic communication channels to deliver content with natural languages (e.g. Arabic, English, French, etc...) to persuade victims to perform certain actions

• Prevention — phishing prevention methods are defined differently in the literature depending on the context. In this survey, the context is attempting to prevent attackers from starting phishing campaigns in the future.

2. RELATED WORKS
The system uses a crawler method for detecting the URL's in the database where already existed and crawls webpage information then given to Map Reduce for predicting the authenticity. The Map Reduce technique improves the performance of the anti-phishing technique. But it requires blacklisted data. EMD (Earth Movers' Distance) [1] method takes the web pages and it compresses the image for reducing the intensity. The Multidimensional distributions are compressed into a fixed number of bins, which is of a predefined size, results in a histogram. The histogram is used for comparing the datasets against the existing datasets. It is an efficient method, requires huge amount of image compression technique. The textual and visual classification [1] is the technique where text and images is fused using Bayesian algorithm and entire document is parsed and then examined based on the dataset. This is a lengthy process because it needs entire search of a document.

DNS-based anti-phishing approach [2] is a technique where it is mainly based on blacklists, heuristic detection. But they do have some shortcomings. Blacklist is a DNS based anti-phishing approach is highly used technique by the browser. When user browsing the phishing sites, Internet Explorer7, Netscape Browser8.1, and Google Safe Browsing are the browsers gives the alert message to user. The blacklist must be updated each time for the proper result.

Heuristic-based anti-phishing technique is to estimate whether a page has some phishing heuristics characteristics. For example, some heuristics characteristics used by the Spoof Guard [2] toolbar include checking the host name, checking

the URL for common spoofing techniques, and checking against previously seen images. If you only use the Heuristic-based technique, the accuracy is not enough. Besides, phishers can use some strategies to avoid such detection rules. The user may be deceived by the phishing website because the phishing website imitates a legitimate website. Its pages are often similar with the legitimate sites. Therefore, some researchers proposed a similarity assessment method to detect phishing sites.

The image is decomposed into shares which expose the original image after decryption [4]. Anti-phishing working group (APWG) generates and reports monthly about the current phishing attacks. It also has many partners which has made huge contributions in this work. Phishtank.com corpus is updated by various users who report phishing websites. This data is available for free for developers developing applications. E-mail server based technique extract all the suspected URLs from inbox as well as from spams and examine them as most attacked users are from phishing E-mails.

2.1Existing Methodologies

APD: ARM Deceptive Phishing Detector System. Phishing Detection in Instant Messengers Using Data Mining Approach Mohd Mahmood Ali and Lakshmi Rajamani introduced Association Rule mining technique fordeceptive phishing. The proposed approach is named as APD (Anti-phishingDetector), detects Phishing in Instant Messengers. Anti-phishing system (APD)dynamically traces out any potential phishing attacks when messages exchanged between clients of an Instant Messaging System. Author also usesApriori algorithm to detect deceptive phishing and Information retrieval system to extract frequently reoccurring words and the messages will be forwarded toARS Anti-Phisher component for further processing. ARMP-IM implemented using Apache TomCat 6.0 for Web Server for creating separate sessions for each user with Browser support. Authors conclude by saying that this approach can be enhanced for mobile Instant Messengers for 3G and 4G.Using Feature Selection and Classification Scheme for Automating Phishing Email Detection Isredza Rahmi A HAMID, Jemal ABAWAJY and Tai-hoon KIM used hybrid feature selection method to detect phishing email.

The main objective is to identify the behavior features in phishing email. This approach is based on the message provided in the message-id field. The message-id tags provided in the email header is used to identify the sender behavior. Using hybrid feature selection algorithm, features are extracted from the email. The author uses these features to

mine the sender behavior to identify whether the email came from legitimate sender or not. Prevention Schemes Against Phishing Attacks on Internet Banking Systems confuse and collect information‘s about phishers. The steps involved in are phishing mail detection, server authentication, early phishing site detection, two factor user authentications, and transaction authentications. Some limitations of anti-phishing techniques are identified and overcome by this honey pots framework. This framework is designed to attack phishers.

3. PROPOSED SYSTEM

Framework below in figure shows proposed framework. The user will enter the URL of the webpage, she wishes to visit. Using that URL, we will download the source code of the webpage & then decide the values of the attributes. For finding these values we will make use of Hadoop-MapReduce [3]. This will speed up the process of attribute value assignment. Basic word count example of Hadoop-MapReduce is used to search sensitive words in webpages. In same way wherever required help of Hadoop is taken. These calculated attributes are the input to the Prediction module. Based on the records stored from phishtank.com database, training data is prepared. All the characteristics of reported phishing website at phishtank.com corpus is studied and based on that attributes are decided and training data for machine learning algorithm is prepared [5]. Using training data machine learning algorithm generates set of rules based on which decision is to be made. Prediction module gets two inputs rules generated by machine learning algorithm and attribute found from requested URL. Prediction module finally predict URL falls under which category (Phishing, Legitimate, and Doubtful).

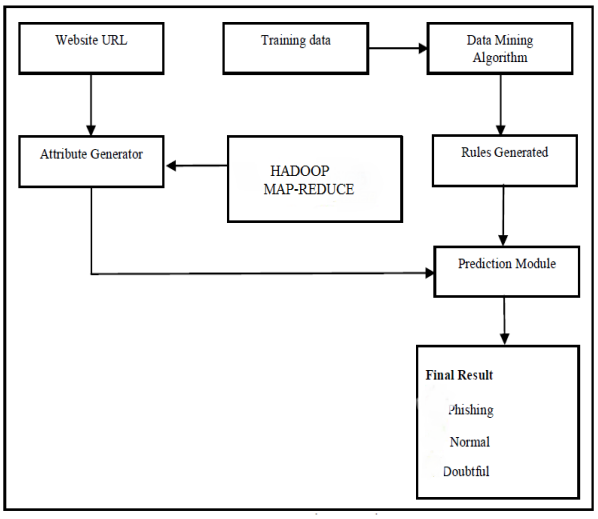

Fig 1: System architecture

3.1 System algorithms

Input: - The input given is a link which is present
e.g. http://www.example.com

Output: - Determination and warning by the given system whether a given page is phished, normal or a doubtful page.

Link to be detected <-- link
Search the link which is given by the user in the dataset

if link in dataset then
print “The link is a phished page”
else
generate attributes of that page through keywords present in that page
Map and Reduce those keywords of the given page and generate a pattern through the attributes generated.

Compare the pattern with the existing dataset.

if pattern generated has a higher threshold value then print “Phished”
else if
pattern generated has less threshold value then print “Doubtful”

else

print “normal page”.

Macintosh, use the font named Times. Right margins should be justified, not ragged.

3.2 Hadoop

An open source software framework that supports data intensive distributed applications. Hadoop makes it possible to run applications on systems with thousands of commodity hardware nodes, and to handle thousands of terabytes of data. Its distributed file system facilitates rapid data transfer rates among nodes and allows the system to continue operating in case of a node failure. This approach lowers the risk of catastrophic system failure and unexpected data loss, even if a considerable number of nodes become inoperative. Consequently, Hadoop quickly emerged as a foundation for big data processing tasks, such as scientific analytics, business and sales planning, and processing enormous volumes of sensor

4. CONCLUSION

Main goal of the system is to achieve speed up in existing anti-phishing system by some means. Using Hadoop-MapReduce in integration with anti-phishing technique we have achieved considerable time speedup. Even if the phishing webpage is not showing phishing characteristics very clearly at first layer it might show characteristics in the next layer so that no phishing webpage will pass through our system. This is the advantage of having layered architecture of attributes. Hadoop-MapReduce will increase the response time of the system considerably. This system is very effective in securing network from phishing attach even at its best. There is a lot of scope for improvement of this system. One can improve the performance of system by converting this as a cloud service. As per type of organization we are protecting from phishing attach change the attributes to be considered for making effective decision about the phishiness of the system.

5. FUTURE SCOPE

Hadoop-MapReduce will help the system respond much quickly. Another valuable addition to this system could be a cloud based service.

This cloud based service would deploy an image captcha to a page which will determine the authenticity of the webpage. This is even more helpful a more number of users are switching into handheld devices. Hence authenticity of users must be ensured even to a higher level.

6. REFERENCES

[1] Fu, A.Y.; Liu Wenyin; Xiaotie Deng, "Detecting Phishing Web Pages with Visual Similarity Assessment Based on Earth Mover's Distance (EMD)," Dependable and Secure Computing, IEEE Transactions on , vol.3, no.4, pp.301,311, Oct.-Dec. 2006.

[2] Hong Bo; Wang Wei; Wang Liming; Geng Guanggang; Xiao Yali; Li Xiaodong; Mao Wei, "A Hybrid System to Find & Fight Phishing Attacks Actively," Web Intelligence and Intelligent Agent Technology (WI-IAT), 2011 IEEE/WIC/ACM International Conference on vol.1, no., pp.506,509, 22-27 Aug. 2011.

[3] J. Dean and S. Ghemawat, "MapReduce: simplified data processing on large clusters," Communications of the ACM - 50th anniversary issue: 1958 - 2008, vol. 51, no. 1, pp. 107-113, 2008.

[4] Aburrous, M.; Hossain, M.A.; Dahal, K.; Thabatah, F., "Modelling Intelligent Phishing Detection System for E-banking Using Fuzzy Data Mining," CyberWorlds, 2009. CW '09. International Conference on, vol., no., pp.265,272, 7-11 Sept. 2009.

[5] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

[6] Tanenbaum, A. S. (2010). Computer Networks (5th Edition). Prentice Hall; 5 edition (October 7, 2010).