Jurnal Informatika Universitas Pamulang **ISSN: 2541-1004**
Penerbit: Program Studi Teknik Informatika Universitas Pamulang **e-ISSN: 2622-4615**
Vol. 5, No. 3, September 2020 (237-246) **10.32493/informatika.v5i3.6672**

A Survey on Phishing Website Detection Using Hadoop

Muhammad Rayhan Natadimadja[1], Maman Abdurohman[2], Hilal Hudan Nuha[3]

School of Computing, Telkom University, Jl. Telekomunikasi Terusan Buah Batu, Bandung, Indonesia, 40257
e-mail: [1]mrayhann@student.telkomuniversity.ac.id, [2]abdurohman@telkomuniversity.ac.id, [3]hilalnuha@telkomuniversity.ac.id

Submitted Date: August 31st, 2020 Reviewed Date: September 22nd, 2020
Revised Date: September 26th, 2020 Accepted Date: September 30th, 2020

Abstract

Phishing is an activity carried out by phishers with the aim of stealing personal data of internet users such as user IDs, password, and banking account, that data will be used for their personal interests. Average internet user will be easily trapped by phishers due to the similarity of the websites they visit to the original websites. Because there are several attributes that must be considered, most of internet user finds it difficult to distinguish between an authentic website or not. There are many ways to detecting a phishing website, but the existing phishing website detection system is too time-consuming and very dependent on the database it has. In this research, the focus of Hadoop MapReduce is to quickly retrieve some of the attributes of a phishing website that has an important role in identifying a phishing website, and then informing to users whether the website is a phishing website or not.

Keywords: Phishing; Hadoop; Website; Information Security; Phishing Detection

1. Introduction

Some people will do everything they can to get what they want and some of them will use their knowledge in a bad way like phishers. They make fake websites that are made to steal personal data from those accessing the site such as user IDs, passwords, and debit/credit cards. Average internet users may not be able to identify whether the websites are phishing or not because the websites are almost identical to the real one. Phishing activity is almost the same as fishing, but when fishing catches fish, whereas phishers capture personal information from a person or organization (Pham, Nguyen, Tran, Huh, & Hong, 2018). They made fake websites with the aim of stealing their personal data and without the user knowing they had given information to phishers.

From the report of Anti-Phishing Working Group, there are 266,387 website phishing in the third quarter of 2019 (Figure. 1**)**. This is 46 percent increased from the second quarter of 2019, which amounted to 182,465 (Anti-Phishing Working Group, 2019). Therefore, phishing is still a big crime because it can result in substantial losses.

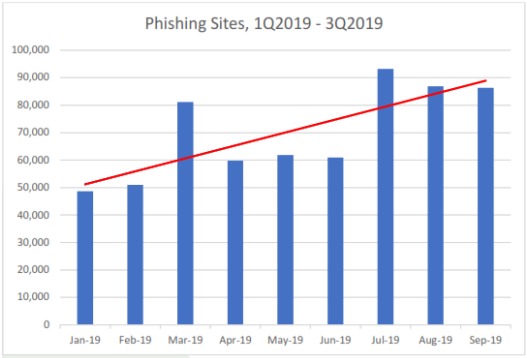

Figure 1. 2019 Phishing web report

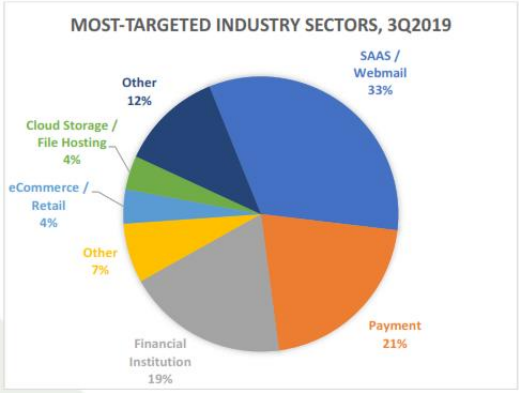

Figure 2. Most phisher target diagram

According to Figure. 2, MarkMonitor, member of APWG made an observation and got results that SAAS/Webmail is the target of largest phisers in the third quarter of 2019 (Anti-Phishing Working Group, 2019). Attacks on site File hosting and eCommerce are less popular in the third quarter of 2019, but attacks on payment sites are still among the second largest after SAAS/Webmail.

Until now, there have been many techniques used to detect phishing sites. As is usually paired into e-mail and browsers such as Google Safe Browser and SmartScreen Filter. Because phishing attacks take advantage of human ignorance of the internet, this is a difficult problem to be solved permanently. All these anti-phishing experiments were developed with the aim of minimizing the impact of phishing attacks.

2. Literature Review

In the technology industry that is developing today, which is very influential on this security problem has given anxiety to some users both at work and at home. Incident that exploit human vulnerability have increased in recent years (Dunlop, Groat, & Shelly, 2010). In this era , there are many developments in the field of security systems aimed at ensuring that security is the top priority and that preventive action must be taken as quickly as possible to avoid being hacked by people who wish to commit crimes in cyberspace. Some of cyber security workers are currently using a reliable and stable detection technique to be their phishing website detection technique (Mahajan & Siddavatam, 2018).

This system uses a crawler (Rakshith & Prabhakara, 2016) to detect URLs in the database and web pages that will be checked, then given to MapReduce to be checked for authenticity. MapReduce is used to improve the performance of phishing site searches. This MapReduce technique improves the performance of phishing site searches. The method is done by taking a page from a phishing website and then compressing the image to reduce the intensity (Tangy, Uz, Caiy, Mamoulisy, & Chengy, 2013). The results of the compress are distributed into several containers whose size has been set, to produce a histogram. This histogram is used to compare datasets with existing datasets.

The current detection of phishing attacks is mostly in two categories namely, detecting and filtering phishing emails, and detecting and filtering phishing websites both approaches are very important to counter phishing attacks. Phishing e-mails and websites must be considered more because of their unpredictable nature, therefore sometimes phishing attacks can escape filters that have been installed. Apart from that, there are several tools used by phishers to bypass phishing emails and websites such as SMS, malware, social media and also online games. (Hong, 2012). In this study, we will discuss more about phishing attacks through websites, there are also several detection techniques that have been used or have been suggested.

The detection technique used in existing browsers such as Firefox and Chrome is to blacklist websites that have been registered in the database. The main weakness of the Blacklist is that it was created by volunteers who found it, therefore the blacklist must be frequently updated manually and the process takes a long time and therefore this technique is weak against new websites created that day (Jain & Gupta, 2016).

Another well-known technique in phishing detection is Visual Cryptography (Kumar & Kumar, 2015) which is a detection technique using images. Others use logos and textual content from a web page (Chiew, Chang, Sze, & Tiong, 2015). A frequent example is captcha that will block interruptions coming from other machines but is not very effective to prevent interruptions from humans.

Detection using Heuristic technique is also a technique that has been used to deal with phishing websites. Heuristics is a technique that estimates whether a web page has heuristics characters (Zhu, Chen, Ye, Li, & Liu, 2019). This technique can recognize phishing websites based on a series of features extracted from them (Tan, Chiew, Wong, & Sze, 2016). But just relying on heuristics will not be enough, because the phishers can outsmart their website so that could not be detected by heuristic techniques. Website visitors can be fooled easily because of its resemblance to the original website.

Cantina + is one example of a well-known heuristic-based approach. They propose the detection of phishing websites using Google PageRank, but only by relying on the value of PageRank (Sunil & Sardana, 2012). It is difficult to identify whether the site is really a phishing website or not, because the website could be an official website that was newly created or a low rank blog website.

Aaron Blum, Brad Wardman, Thamar Solorio proposed research (Blum, Wardman, Solorio, & Warner, 2010) focusing on the idea of limiting the source of features that can facilitate

information extraction through the host. The URL will be considered a binary feature vector. The vector is entered into the algorithm, then from the vector it will be found whether the URL is phishing or not.

Ramesh Gowtham and Ilango Krishnamurthi proposed. Anti-phishing system with filtering mechanism based on 15 heuristic features (Gowtham & Krishnamurthi, 2014). However, the accuracy of the login window must match the features provided. According Rakesh Verma and Keith Dyer, proposed a set of lexical URLs, and also how many letters are in them (Verma & Dyer, 2015). However, if the URL does not have spelling errors, then this feature may not work properly.

Machine learning based detection techniques also one of techniques to used to detect phishing websites. Machine learning techniques rely on a set of features being extracted onto every web pages and further require the genuine website for training data to be retrieved as well as a phishing website to be checked. (Qabajeh, Thabtah, & Chiclana, 2018). the accuracy of the result greatly affected by the quality of websites in the training set (Rao & Pais, 2019). Despite these challenges, the approach of using machine learning techniques has become an active subject of discussion for this phishing website detection research. Several studies have been carried out using varied data sets and using different classification algorithms (Abdeljaber, Mohammad, Thabtah, & McCluskey, 2013; Feng et al., 2018; Sahingoz, Buber, Demir, & Diri, 2019). The accuracy of algorithm is affected by features used in classification, but some study thought of the choice of intelligent method features properly. (Rajab, 2018). Choice of feature is an important task to build a good, generalized phishing detection. Currently, a feature that is widely used as an option is heuristics (Babagoli, Aghababa, & Solouk, 2019).

URL-based detection technique is also one of the techniques used to detect phishing websites. This technique analyzes the features from URL and inform if there any dangerous websites. Marchal et al. proposes a phishing detection system, in which the system uses lexical analysis of URLs as well as queries from search engines (Marchal, Francois, State, & Engel, 2014). But queries sent across the network can increase the space as well as the costs involved. While James et al., they do research on lexical-based phishing detectors as well as the information they get on the web page (James, Sandhya, & Thomas, 2013). this feature relies on special features made for certain websites, therefore this feature is not suitable for large-scale datasets.

There are also other phishing website detection techniques such as user habits, according Srinvasa Rao and Alwyn R Pais, they exploit the phishing web pages to find out what happens when they enter data on the website, such as entering fake credentials and also observing the contents of the login page to get the desired results. (Rao & Pais, 2017). However, there are some limitations regarding the login system, for example, some websites can only enter an incorrect password three times. Also, in some websites the login column cannot be detected correctly, so false credentials cannot be sent automatically.

Finding phishing targets is useful for analyzing the behavior of an attacker and can help users to access legitimate web pages. In (Ramesh, Gupta, & Gamya, 2017), they propose to classify hyperlinks from suspicious web pages according to the related domain. However, this method requires analysis of many links and candidates for phishing targets which may not be included in the hyperlink group. In (Wenyin, Fang, Quan, Qiu, & Liu, 2010), they detect phishing targets from suspicious web pages using the consideration of the Sematic Link Network and their construction. with this method web page detection can be done. however, it requires a fairly high cost.

Due to the use of the open internet to carry out various online activities. Users must be prepared from the threat of cyber crime. There are many types of cyber crimes, and one of them is phishing. phishing is one of the most popular cyber crimes. (Pujara & Chaudhari, 2018).

Phishing will remain a dangerous attack despite extensive research on phishing website filters (Gutierrez et al., 2018). Therefore, a monthly report to record phishing attacks is produced by the Anti-Phishing Working Group (APWG), and another group that plays a role in fighting phishing is Phishtank. Phishtank is a web-based application that provides crowdsourcing services aimed at reporting and validating a website (Dobolyi & Abbasi, 2016). Phishtank users can add websites suspected of being phishing websites with the aim of indicating that website is a phishing website, and if true then that site's URL will be entered into the Phishtank database.

2.1 Phishing

Phishing is a method of committing fraud by tricking the target with the intention of stealing the target account (Mao, Tian, Li, Wei, & Liang,

ISSN: 2541-1004
e-ISSN: 2622-4615
10.32493/informatika.v5i3.6672

2017). Phishing is also often known as website violence (Satish & K, 2013). The term comes from the word fishing which means to lure the victim to be trapped into his trap. This phishing was created with the aim of stealing important information of a person or an organization such as their personal and financial information. Phishing is a serious crime and web threat because it can cause large financial losses (Mohammad, Thabtah, & McCluskey, 2015; Thabtah & Kamalov, 2017). The purpose of phishers is to deceive users into being able to provide their sensitive information (Abdelhamid, Ayesh, & Thabtah, 2014).

To trap internet users who frequently visit websites, attackers create phishing web pages (which are similar to social media) so that victims can enter their personal information on those web pages. Attackers usually publish links from their phishing website address on social media intended to trick users into visiting their phishing pages. Because of social media being an easy place to catch inexperience users and are diverted so that users access their websites.

Stolen information is usually in the form of a password or information about a user's credit card (Baykara & Gürel, 2018). With the help of a website display that resembles an official site, average users will enter their personal data into the phishing site. Information that is often stolen by these websites are, user's account number, user's password and username, credit card information, and user e-banking information. Phishing like this is also often found in users' e-mails.

In studies of user experience from phishing attacks, users are fooled by phishing websites (Volkamer, Renaud, Reinheimer, & Kunz, 2017) for these five reasons. Users lack knowledge of URLs, users do not know which website can be trusted, users do not see the full URL, because there is a redirection or hidden URL, users do not have time to ask the authenticity of a website, or users accidentally enter the website, users cannot distinguish phishing website from official website.

Although caution and user experience are important to avoid phishing, users may not be able to completely avoid phishing scams (Greene, Steves, & Theofanos, 2018). Because before they carry out an attack, the attacker also takes into account the habits and characteristics of the user (Curtis, Rajivan, Jones, & Gonzalez, 2018). Cyber-attacks can cost up to billions of dollars in losses as well as the loss of confidential user information (Shaikh, Shabut, & Hossain, 2017). In addition, attackers can also attack the user's mobile device, especially at this time, where the use of smartphones are increasing (Goel & Jain, 2018).

2.2 Hadoop

Hadoop Is a framework or Java-based open source platform under Apache to support applications that run on big data. Hadoop is used to handle large amounts of data, be it structured, semi-structured, or unstructured data. Hadoop replicates the data in several clusters so that if there is a problem in one cluster then the other clusters are still alive. The name hadoop itself comes from the elephant doll owned by Doug Cutting's son, then Hadoop was developed by Mike Cafarella and Doug Cutting in 2005.

2.3 MapReduce

Google introduced a programming model that aims to process large datasets called MapReduce (Zhang & Chen, 2014). The framework of MapReduce is used to process large dataset using many nodes, commonly called clusters or grids. The process can occur in a filesystem or database. MapReduce usually consists of three stages, Map, Shuffle, and Reduce.

2.4 Phishtank

Phishtank was launched in October 2006, Phishtank is a community-based service that provides a place to report and verify phishing websites. Users can report a website URL that is suspected to be a phishing site, then the Phishtank community will vote whether the URL is phishing or not. Phishtank is used by Opera web browser, online reputation, and internet security service browser plugin Web of Trust, Yahoo! Mail, the McAfee antivirus, and Kaspersky. The blacklist that has been approved by Phishtank can be downloaded as a JSON file.

2.5 Phishing Website

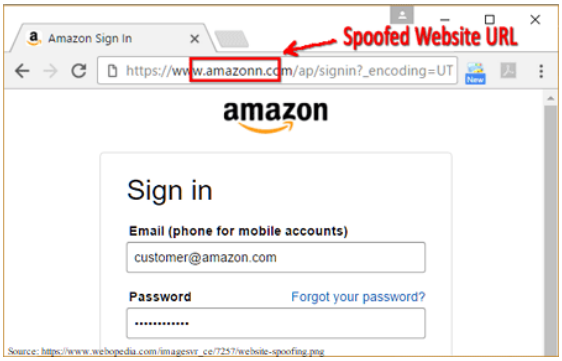

Figure 3. Example of phishing website.

ISSN: 2541-1004
e-ISSN: 2622-4615
10.32493/informatika.v5i3.6672

Phishing website pages have a similar interface to the original website, but they have different URLs. A cautious and experienced user can distinguish official and genuine websites only from their URLs. However, due to time constraints, some users do not see the entire URL, because they believe that the URL from social media is a genuine website. By using this kind of fraud, phishers try to obtain sensitive information and victim's personal data. (Gupta, Arachchilage, & Psannis, 2018). If user has entered this website, which they believe that this website is genuine like Figure. 3.

Users can easily provide their personal information without suspicion because of the similarity of the website with the original.

3. Writing Method

This research is a type of literature study obtained/studied from reliable sources relating to phishing, the techniques used to detect phishing and how to handle it using Hadoop. The writing of this paper begins with the lack of literature that summarizes the phishing detection techniques and method, and solution to speed up phishing detection.

4. Result and Disscussion

The general description of the system is to use the MapReduce technique to generate attributes of a phishing website. Users enter the URL that user want to visit, then the website will be analyzed, and the value of the attribute will be calculated. The dataset from Phishtank will be used to compare the attributes that have been obtained by MapReduce and after comparison it will produce results that the website is included as phishing web site or not.

The overview of how the phishing detection system work based on Figure. 4 are as follows, users enter the URL they want to check, Hadoop MapReduce will extract the attributes from the URL that has been given, the results of the extracted attributes will be made into a comparison material with data in the dataset, data in the dataset will be given to the classifier to make a rule to be used as a comparison, the classifier will forward the data and rules to predict, then predict will produce results, the website is a phishing or not.

4.1 Dataset

To create this phishing detection system, data sets that can represent URLs on the internet are needed. Therefore, we need a large dataset and the URL that can represent the internet. To build reliable dataset, the URLs used on this system are from the Phishtank website. This URLs are the core for the rule-making algorithm on this system whose attributes will be used in the construction of a phishing website detection system. Dataset will be provided as input to the Classifier that is applied in the WEKA machine learning data mining tool. Data sets are arranged hierarchically.

4.2 Attribute Generator

In this proposed system, the attribute generator is a module that has an important role in determining the genuineness of a URL. Attributes considered consists of three layers. This system uses the Layered attribute. where the first layer contains identity of URL and domain. While layer two consist of Security and encrytion, and source code and script. And layer three consist of web adress bar, page style, social human factor, and content.

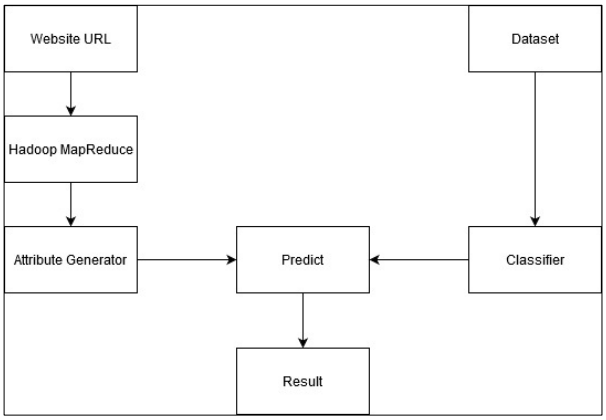

Figure 4. Overwiev of proposed system

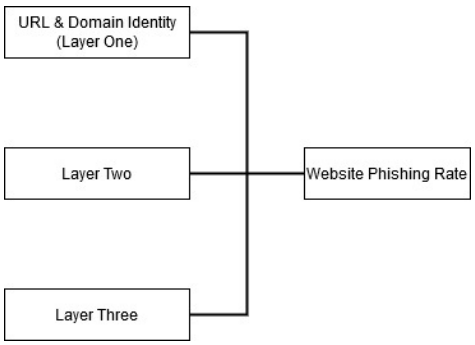

Figure 5. Architecture of Attribute that will be used.

After searching from many documents to find which attributes are needed for this system. The architectural model for the attributes in Figure. 5 – Figure. 8 is based on (Aburrous, Hossain, Dahal, & Thabatah, 2009). This model is used because the consideration of using visual aspects and this model is not used only for specific

purposes, it can also be used to determine the attributes of a general website. Attributes will be generated by several rules. The authenticity of a website will be inversely proportional to the value of the suspicious attribute that has been obtained. Hadoop MapReduce will separate the attributes. Using MapReduce will reduce the computation time for each dataset. then the separated attributes will be compared to determine the genuineness of the website.

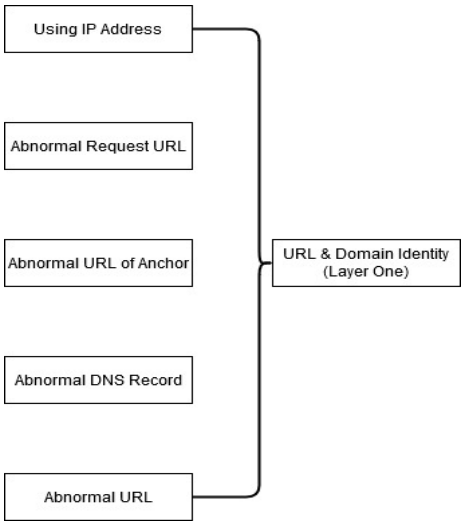

Figure 6. Detail of Attributes in layer one.

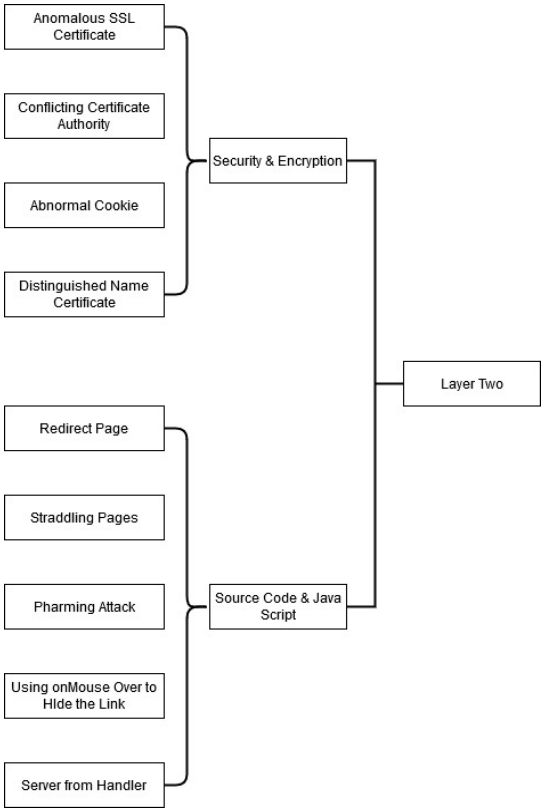

Figure 7. Detail of Attributes in layer two.

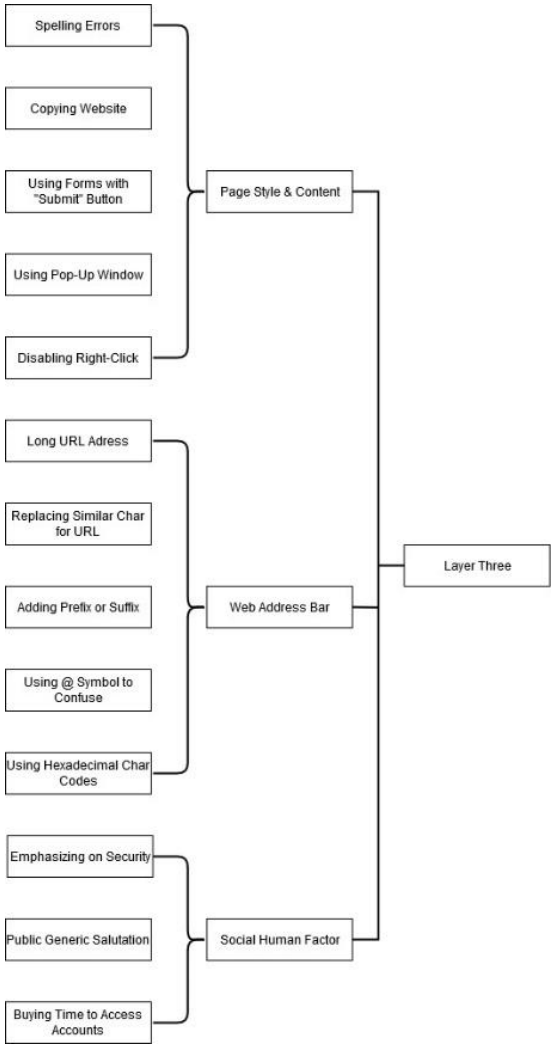

Figure 8. Detail of Attributes in layer three.

4.3 Classifier

The function of this module is to fetch a data from the database and makes some rules for comparing website whether phishing or not. To make rules that can be trusted, a tool for mining data, WEKA (Hall et al., 2009) can be used to help the process. With all the data mining algorithms in it, it can help in determining the most suitable rules. Then the PART algorithm is used in this system. PART is short for Projective Adaptive Resonance Theory. This algorithm is very useful if faced with a large database. This system works in a way, gives the attributes that are received and given to the predict module. Then the layer will act as a coordinator between the rules that are made and the attributes that are accepted. By classifying attribute values correctly, layers can estimate the nature of a website. To make this system simple, websites are classified into three categories. Trustworthy, Suspicious, and Phishing.

4.4 Predict

The task of this module is to make decisions based on input obtained from the attribute generator and classifier. The rules from the classifier will be used as a decision maker. The next input is obtained from the attribute generator using Hadoop MapReduce. The attributes of the web page will be searched for by Hadoop MapReduce and then forwarded to the predict module (Baitule & Deshpande, 2014). Using Hadoop MapReduce, attributes from attribute generator can be compared with datasets that have been arranged according to rules.

Table 1. Result Summary

No	Paper	Method	Dataset	Accuracy	Remarks
1	Utilisation website logo (Chiew et al., 2015)	Heuristic/SVM	Phishtank and Alexa	93.40%	Can detect image-based phishing
2	Effective Phishing Websites Detection Model (Zhu et al., 2019)	Neural network and Optimal feature selection	UCI Dataset, Phishtank, and Alexa	99.93%	Continously change of features and can deal with phishing with sensitive feature
3	PhishWho (Tan et al., 2016)	Heuristic	Phishtank, OpenPhish, and Alexa	96.10%	Cannot address visual cloning, use three phases to detect phishing website, and loaded to client's browser.
4	PageRank (Sunil & Sardana, 2012)	Heuristic/ Google PageRank	Phishtank	98%	Only relying on value of Pagerank and cannot detect zero-day phishing.
5	Efficient feature-based machine learning framework (Rao & Pais, 2019)	J48, AdaboostM1, Random Forest, SVM, Bayers	Phishtank and Alexa	99.31%	training set depend on the quality, and using various algorithms.
6	Novel neural network (Feng et al., 2018)	Neural Network /Monte Carlo Algorithm	UCI repository	97.71%	All pages must be downloaded, using 30 features.
7	Machine learning based (Sahingoz et al., 2019)	K-star, kNN, SMO	Phishtank, Yandex	95.7%	Have a relatively huge dataset and using various algorithms.
8	Heuristic nonlinear regression (Babagoli et al., 2019)	Mete-heuristic/ Decision tree and Wrapper	UCI Datasets	92.8%	Use third-party service and use 20 features.
9	PhishScore (Marchal et al., 2014)	SVM, LMT, Jrip, PART	Phishtank	94.91%	Real-time phishing detecting system and using various algorithms.
10	Analyzing the feign relationship (Ramesh et al., 2017)	TVD algorithm	Google, Alexa, Netcrafts, Millersmiles, Phishtank, Reasonable-Phishing Webpage list.	99.54%	Minimal use of third-party service and low false positive rate.

5. Reported Output

Using Hadoop MapReduce will speed up the process of detecting phishing websites. because Hadoop MapReduce runs in a distributed environment, the attribute distributing process will run faster. so, the results will be obtained faster. Experimental results reported in the literature is summarized by Table 1.

6. Conclusion

The main objective of this proposed system is to improve the search performance of phishing websites, especially their speed. This can be achieved with the help of Hadoop MapReduce by spreading tasks through several different nodes, this way the user can find out if a URL is phishing or not more quickly, and also Hadoop MapReduce will speed up the overall system response.

References

Abdelhamid, N., Ayesh, A., & Thabtah, F. (2014). Phishing detection based Associative Classification data mining. *Expert Systems with Applications*, *41*(13), 5948–5959. https://doi.org/10.1016/j.eswa.2014.03.019

Abdeljaber, F., Mohammad, R. M., Thabtah, F., & McCluskey, L. (2013). *Predicting Phishing Websites using Neural Network trained with Back-Propagation*. 682–686.

Aburrous, M., Hossain, M. A., Dahal, K., & Thabatah, F. (2009). Modelling intelligent phishing detection system for e-banking using Fuzzy Data Mining. *2009 International Conference on CyberWorlds, CW '09*, 265–272. https://doi.org/10.1109/CW.2009.43

Anti-Phishing Working Group. (2019). *Phishing activity trends report, 3rd quarter 2019 phishing*.

Babagoli, M., Aghababa, M. P., & Solouk, V. (2019). Heuristic nonlinear regression strategy for detecting phishing websites. *Soft Computing*, *23*(12), 4315–4327. https://doi.org/10.1007/s00500-018-3084-2

Baitule, P. D., & Deshpande, S. P. (2014). *A Survey On Efficient Anti Phishing Method Based on Visual Cryptography Using Cloud Technique By Smart Phones*. *2014*, 11–15.

Baykara, M., & Gürel, Z. Z. (2018). Detection of phishing attacks. *6th International Symposium on Digital Forensic and Security, ISDFS 2018 - Proceeding*, *2018-Janua*, 1–5. https://doi.org/10.1109/ISDFS.2018.8355389

Blum, A., Wardman, B., Solorio, T., & Warner, G. (2010). Lexical feature based phishing URL detection using online learning. *Proceedings of the ACM Conference on Computer and Communications Security*, 54–60. https://doi.org/10.1145/1866423.1866434

Chiew, K. L., Chang, E. H., Sze, S. N., & Tiong, W. K. (2015). Utilisation of website logo for phishing detection. *Computers and Security*, *54*, 16–26. https://doi.org/10.1016/j.cose.2015.07.006

Curtis, S. R., Rajivan, P., Jones, D. N., & Gonzalez, C. (2018). Phishing attempts among the dark triad: Patterns of attack and vulnerability. In *Computers in Human Behavior* (Vol. 87). https://doi.org/10.1016/j.chb.2018.05.037

Dobolyi, D. G., & Abbasi, A. (2016). PhishMonger: A free and open source public archive of real-world phishing websites. *IEEE International Conference on Intelligence and Security Informatics: Cybersecurity and Big Data, ISI 2016*, 31–36. https://doi.org/10.1109/ISI.2016.7745439

Dunlop, M., Groat, S., & Shelly, D. (2010). GoldPhish: Using images for content-based phishing analysis. *5th International Conference on Internet Monitoring and Protection, ICIMP 2010*, 123–128. https://doi.org/10.1109/ICIMP.2010.24

Feng, F., Zhou, Q., Shen, Z., Yang, X., Han, L., & Wang, J. Q. (2018). The application of a novel neural network in the detection of phishing websites. *Journal of Ambient Intelligence and Humanized Computing*, *0*(0), 1–15. https://doi.org/10.1007/s12652-018-0786-3

Goel, D., & Jain, A. K. (2018). Mobile phishing attacks and defence mechanisms: State of art and open research challenges. *Computers and Security*, *73*, 519–544. https://doi.org/10.1016/j.cose.2017.12.006

Gowtham, R., & Krishnamurthi, I. (2014). A comprehensive and efficacious architecture for detecting phishing webpages. *Computers and Security*, *40*, 23–37. https://doi.org/10.1016/j.cose.2013.10.004

Greene, K., Steves, M., & Theofanos, M. (2018). No phishing beyond this point. *Computer*, *51*(6), 86–89. https://doi.org/10.1109/MC.2018.2701632

Gupta, B. B., Arachchilage, N. A. G., & Psannis, K. E. (2018). Defending against phishing attacks: taxonomy of methods, current issues and future directions. *Telecommunication Systems*, *67*(2), 247–267. https://doi.org/10.1007/s11235-017-0334-z

Gutierrez, C. N., Kim, T., Corte, R. Della, Avery, J., Goldwasser, D., Cinque, M., & Bagchi, S. (2018). Learning from the ones that got away: Detecting new forms of phishing attacks. *IEEE Transactions on Dependable and Secure Computing*, *15*(6), 988–1001. https://doi.org/10.1109/TDSC.2018.2864993

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter*, *11*(1), 10–18.

Hong, J. (2012). The state of phishing attacks. *Communications of the ACM*, *55*(1), 74–81. https://doi.org/10.1145/2063176.2063197

Jain, A. K., & Gupta, B. B. (2016). A novel approach to protect against phishing attacks at client side using auto-updated white-list. *Eurasip Journal on Information Security*, *2016*(1). https://doi.org/10.1186/s13635-016-0034-3

James, J., Sandhya, L., & Thomas, C. (2013). Detection of phishing URLs using machine learning techniques. *2013 International Conference on Control Communication and Computing, ICCC 2013*, (Iccc), 304–309. https://doi.org/10.1109/ICCC.2013.6731669

Kumar, V., & Kumar, R. (2015). Detection of phishing attack using visual cryptography in ad hoc network. *2015 International Conference on Communication and Signal Processing, ICCSP 2015*, 1021–1025. https://doi.org/10.1109/ICCSP.2015.7322654

Mahajan, R., & Siddavatam, I. (2018). Phishing Website Detection using Machine Learning Algorithms. *International Journal of Computer Applications*, *181*(23), 45–47. https://doi.org/10.5120/ijca2018918026

Mao, J., Tian, W., Li, P., Wei, T., & Liang, Z. (2017). Phishing-Alarm: Robust and Efficient Phishing Detection via Page Component Similarity. *IEEE Access*, *5*, 17020–17030. https://doi.org/10.1109/ACCESS.2017.2743528

Marchal, S., Francois, J., State, R., & Engel, T. (2014). PhishScore: Hacking phishers' minds. *Proceedings of the 10th International Conference on Network and Service Management, CNSM 2014*, 46–54. https://doi.org/10.1109/CNSM.2014.7014140

Mohammad, R. M., Thabtah, F., & McCluskey, L. (2015). Tutorial and critical analysis of phishing websites methods. *Computer Science Review*, *17*, 1–24. https://doi.org/10.1016/j.cosrev.2015.04.001

Pham, C., Nguyen, L. A. T., Tran, N. H., Huh, E. N., & Hong, C. S. (2018). Phishing-Aware: A Neuro-Fuzzy Approach for Anti-Phishing on Fog Networks. *IEEE Transactions on Network and Service Management*, *15*(3), 1076–1089. https://doi.org/10.1109/TNSM.2018.2831197

Pujara, E. P., & Chaudhari, M. B. (2018). Phishing Website Detection using Machine Learning : A Review. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, *3*(7), 395–399.

Qabajeh, I., Thabtah, F., & Chiclana, F. (2018). A recent review of conventional vs. automated cybersecurity anti-phishing techniques. *Computer Science Review*, *29*, 44–55. https://doi.org/10.1016/j.cosrev.2018.05.003

Rajab, M. (2018). An anti-phishing method based on feature analysis. *ACM International Conference Proceeding Series*, 133–139. https://doi.org/10.1145/3184066.3184082

Rakshith, K. R., & Prabhakara, B. K. (2016). Phishing Detection using Map-reduce and PART Algorithm. *International Journal of Advanced Research in Computer and Communication Engineering*, *5*(8), 492–494. https://doi.org/10.17148/IJARCCE.2016.58101

Ramesh, G., Gupta, J., & Gamya, P. G. (2017). Identification of phishing webpages and its target domains by analyzing the feign relationship. *Journal of Information Security and Applications*, *35*, 75–84. https://doi.org/10.1016/j.jisa.2017.06.001

Rao, R. S., & Pais, A. R. (2017). Detecting phishing websites using automation of human behavior. *CPSS 2017 - Proceedings of the 3rd ACM Workshop on Cyber-Physical System Security, Co-Located with ASIA CCS 2017*, 33–42. https://doi.org/10.1145/3055186.3055188

Rao, R. S., & Pais, A. R. (2019). Detection of phishing websites using an efficient feature-based machine learning framework. *Neural Computing and Applications*, *31*(8), 3851–3873. https://doi.org/10.1007/s00521-017-3305-0

Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. (2019). Machine learning based phishing detection from URLs. *Expert Systems with Applications*, *117*, 345–357. https://doi.org/10.1016/j.eswa.2018.09.029

Satish, S., & K, S. B. (2013). Phishing Websites Detection Based on Web Source Code and URL in the Webpage. *International Journal of Computer Science and Engineering Communications*, *1*(1), 1–5. https://doi.org/10.5281/zenodo.821732

Shaikh, A. N., Shabut, A. M., & Hossain, M. A. (2017). A literature review on phishing crime, prevention review and investigation of gaps. *SKIMA 2016 - 2016 10th International Conference on Software, Knowledge, Information Management and Applications*, 9–15. https://doi.org/10.1109/SKIMA.2016.7916190

Sunil, A. N. V., & Sardana, A. (2012). A PageRank based detection technique for phishing web sites. *2012 IEEE Symposium on Computers & Informatics (ISCI)*, 58–63. https://doi.org/10.1109/ISCI.2012.6222667

Tan, C. L., Chiew, K. L., Wong, K. S., & Sze, S. N. (2016). PhishWHO: Phishing webpage detection via identity keywords extraction and target domain name finder. *Decision Support Systems*, *88*, 18–27. https://doi.org/10.1016/j.dss.2016.05.005

Tangy, Y., Uz, L. H., Caiy, Y., Mamoulisy, N., & Chengy, R. (2013). Earth mover's distance based similarity search at scale. *Proceedings of the VLDB Endowment*, *7*(4), 313–324. https://doi.org/10.14778/2732240.2732249

Thabtah, F., & Kamalov, F. (2017). Phishing Detection: A Case Analysis on Classifiers with Rules Using Machine Learning. *Journal of Information and Knowledge Management*, *16*(4), 1–16.

https://doi.org/10.1142/S0219649217500344

Verma, R., & Dyer, K. (2015). *On the Character of Phishing URLs*. 111–122. https://doi.org/10.1145/2699026.2699115

Volkamer, M., Renaud, K., Reinheimer, B., & Kunz, A. (2017). User experiences of TORPEDO: TOoltip-poweRed Phishing Email DetectiOn. *Computers and Security*, *71*, 100–113. https://doi.org/10.1016/j.cose.2017.02.004

Wenyin, L., Fang, N., Quan, X., Qiu, B., & Liu, G. (2010). Discovering phishing target based on semantic link network. *Future Generation Computer Systems*, *26*(3), 381–388. https://doi.org/10.1016/j.future.2009.07.012

Zhang, K., & Chen, X. W. (2014). Large-scale deep belief nets with mapreduce. *IEEE Access*, *2*, 395–403. https://doi.org/10.1109/ACCESS.2014.2319813

Zhu, E., Chen, Y., Ye, C., Li, X., & Liu, F. (2019). OFS-NN: An Effective Phishing Websites Detection Model Based on Optimal Feature Selection and Neural Network. *IEEE Access*, *7*, 73271–73284. https://doi.org/10.1109/ACCESS.2019.2920655