International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 6, Issue 1, No 3, 2018 ISSN: 2348-6600
Reference ID: IJCS-337 PAGE NO: 2239-2245

A Novel Framework Using Phishing Prevention and Detection

V.Geetha[#1], Dr.M.V.Srinath[*2], R.Suganthi[*3]

[#]*Head of the Department of Computer Science, Sengamala Thayaar Educational Trust Women's College, Mannargudi, Thiruvarur, Tamil Nadu*

[*]*Department of Computer Science, Sengamala Thayaar Educational Trust Women's College, Mannargudi, Thiruvarur, Tamil Nadu*

[3]aathilakshmi.ravi@gmail.com

Abstract— **A way to deal with identification of phishing pages in view of visual similitude is proposed, which can be used as a piece of a venture answer for hostile to phishing. A genuine page proprietor can utilize this way to deal with look the Web for suspicious website pages which are outwardly like the genuine site page. Our propose build up another neuro-fuzzy approach without utilizing IF-THEN guidelines to distinguish phishing. They are inspired by an earlier report that utilized the traditional neural system display. Consolidating the neural system with the fuzzy model, acquire a decent outcome as far as recognizable proof exactness. Besides, our framework can accomplish ongoing reaction and stable execution to recognize phishing URLs. This approach will break down the vindictive substance as opposed to include, enhancing the proficiency of phishing recognition. So this undertaking is gone for finding the control based identification technique for distinguishing phishing. Phishing assaults for the most part influence the unmindful and reckless individuals. Instructing individuals on phishing mindfulness turns out to be ineffectual, on the grounds that new and local clients add to the stream each day. So if the proposed framework is executed in the program, it will naturally identify the phishing sites and caution the client while perusing. Fundamental investigations demonstrate that the approach can effectively recognize those phishing website pages with a couple of false alerts at a speed satisfactory for online application.**

Index Terms— **Phishing detection, Phishing prevention, fuzzy model**

I. INTRODUCTION

Phishing is a criminal trap of taking casualties' close to home data by sending them mock messages asking them to visit a manufactured website page that resembles a genuine one of a honest to goodness organization and requests that the beneficiaries enter individual data, for example, Visa number, secret word and so forth. The casualties may at long last endure misfortunes of cash or different sorts. As per the reports of Anti-Phishing Working Group, the quantity of phishing assaults is expanding by half month to month and they can, as a rule, persuade 5% of the phishing email beneficiaries to react to them. By giving Internet exchange tasks, it is the commitment of the organizations to guard it. The organizations might be relied upon to bear the duty, take the activities to go out to effectively recognize those phishing messages and phishing sites, and after that forestall potential phishing assaults [3].

Phishing is a site falsification procedure with a goal to track and take the touchy data of online clients. The programmer tricks the client with social designing procedures, for example, SMS, voice, email, site and malware. Different methodologies have been proposed and actualized to recognize an assortment of phishing assaults, for example, utilization of boycotts and whitelists to give some examples. In this proposition, they propose a work area application called PhishSaver, which centers around the URL and site substance of the phishing website page. They go for identifying phishing sites with the assistance of a work area application named Phish Saver. Phish Saver utilizes a blend of boycott and various heuristic highlights to distinguish various phishing assaults. For boycott, they have utilized google API benefits that are Google safe perusing boycott as this rundown is continually refreshed and kept up by Google [4]. It is likewise conceivable to run Phish Saver as a daemon procedure that implies it can distinguish phishing assaults continuously as a client peruses the web. Phish Saver takes URL as information and yields the status of URL as phishing or honest to goodness site. The heuristics used to recognize phishing are footer joins with invalid esteem, zero connections in the body of the HTML, copyright content, title

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 6, Issue 1, No 3, 2018 ISSN: 2348-6600
Reference ID: IJCS-337 PAGE NO: 2239-2245

substance and site personality. PhishSaver can identify zero hour phishing assaults which may not have been boycotted and is quicker than visual based appraisal systems that are utilized as a part of recognizing phishing. They watch that PhishSaver has acquired a higher exactness rate and covers a more extensive scope of phishing assaults that outcomes in less false negative and false positive rate.

PhishSaver is a work area application to viably distinguish phishing sites and practices. PhishSaver is competent to identify the greater part of the phishing procedures and goes for giving full-confirmation security from a wide range of phishing assaults. PhishSaver is anything but a conventional antivirus programming and does not ensure any insurance from any sort of infection assaults. PhishSaver depends on the possibility that clients ought to have the capacity to peruse the web securely and get to sites without getting worried about the authenticity of the sites. The framework works by taking a contribution from the client as URL of the site for which authenticity should be resolved. The framework at that point yields the condition of the site as phishing, real or obscure. So if the proposed framework is executed in the program, it will naturally identify the phishing sites and caution the client while perusing. Our build up another neuro-fluffy approach without utilizing IF-THEN standards to distinguish phishing. They are inspired by an earlier report that utilized the ordinary neural system demonstrate. Joining the neural system with the fluffy model, acquire a decent outcome as far as recognizable proof precision.

II. Related Work

A. Design and Evaluation of a Real-Time URL Spam Filtering Service

In this research, Grouping works freely of the setting where a URL shows up (e.g., blog remark, tweet, or email), offering to ascend to the likelihood of spam URL sifting as an administration. They mean the framework to go about as a first layer of protection against spam content focusing on web administrations, including interpersonal organizations, URL sharpeners, and email. They demonstrate the general planned task of Monarch. Ruler keeps running as an autonomous administration to which any web administration can give URLs to filter and order. Amid the period it takes for Monarch's order to finish, these administrations can either postpone the dispersion of a URL, disseminate the URL and retroactively square guests if the URL is hailed as spam (gambling a little window of presentation), or utilize a heavier-weight confirmation procedure to implement significantly stricter prerequisites on false positives that are ensured by grouping.

B.PhishDef: URL Names Say It All

In analyzing the execution of a few clumps construct learning calculations in light of grouping malevolent URLs, which incorporate phishing URLs and URLs exhibit in spam messages. The calculations are assessed when working with different capabilities, for example, have based highlights, for example, highlights from WHOIS inquiries, and lexical highlights. This work demonstrates that the mix of host-based and lexical highlights brings about the most astounding characterization precision. This work additionally implied that utilizing lexical highlights may prompt high exactness; notwithstanding, it didn't examine this course in adequate profundity. Our work expands on this underlying perception. They broadly assess how both cluster based and online calculations perform when utilizing just lexical highlights contrasted with full highlights. In a subsequent work, think about the execution of clump based calculations to online calculations when utilizing full highlights. The creators locate that online calculations, particularly Confidence-Weighted (CW), beat group based calculations. Our fundamental distinction from is that they center around lexical highlights rather than full highlights. they propose confusion safe lexical highlights, demonstrate that online calculations outflank cluster based calculations when working with lexical highlights, and give a point by point examination of the datasets to clarify why this is the situation.

C.Large-Scale Automatic Classification of Phishing Pages

Phish Tank characterizes phishing as "a false endeavor, typically made through email, to take individual data". With a specific end goal to secure end clients against the broadest arrangement of phishing tricks, they utilize a to some degree more broad meaning of phishing than this. They characterize a phishing page as any page that, without authorization, charges to follow up for an outsider with the goal of befuddling watchers into playing out an activity with which the watcher would just confide in a genuine specialist of the outsider. Note that these activities incorporate, yet are not constrained to, submitting individual data to the page. As it were, our meaning of phishing is nearer to "web falsification". This definition surely covers the run of the mill instance of phishing pages showing illustrations identifying with a money -related organization and asking for a watcher's login qualifications. This definition additionally covers phishing pages which show a confided in organization's logos and demand that the watcher download and execute an

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 6, Issue 1, No 3, 2018 ISSN: 2348-6600
Reference ID: IJCS-337 PAGE NO: 2239-2245

obscure paired. Destinations guaranteeing to unblock an email or talk account when given the login certifications fall under this last classification. Note that on the off chance that one of these destinations is endorsed by the outsider, at that point it would be appropriately approved and thusly not a phishing page. The different proposition has likewise analyzed the issue of naturally ordering phishing website pages, however they have portrayed exploratory frameworks, not frameworks in dynamic utilize. Additionally, none of these endeavors utilized a loud preparing dataset. Our framework, then again, gives characterizations to a crowd of people of over a hundred million web clients continuously. They can distribute these characterizations without additional audit in light of the fact that our classifier makes less false positive arrangements than alternate frameworks in spite of our loud preparing information. Displayed a framework for utilizing Google web scan as a channel for phishing pages however utilized just 2519 dynamic URLs in their most reasonable dataset. Likewise, their traditionalist classifier showed a false positive rate of 3%, too high to possibly be reasonable without additionally audit. Portrayed a framework for grouping phishing messages that offers numerous likenesses with our framework, in spite of not arranging website pages.

D. Hybrid Phish Detection Approach by Identity Discovery and Keywords Retrieval

Phishing designs advance always, and it is normally hard for a discovery strategy to accomplish a high evident positive rate. While keeping up a low false positive rate (FP). In this proposition, they propose a novel half breed discovery strategy in light of IE and IR methods trying to accomplish a decent harmony amongst TP and FP. In another heuristics-based work, proposed a technique going for extricating the website page character from chosen DOM properties, (for example, the page title, Meta depiction upon the separated personality a rundown of highlights. With the extricated personality, a help vector machines (SVM) demonstrate was prepared on 50 phishing and 50 genuine pages, accomplishing a normal FP of around 12% and more than 90% TP on a testing set of 50 pages more than 7 runs. This is the main work managing personality extraction that they are aware of. Notwithstanding, its suspicion that the circulation of the personality words, for the most part , strays from that of the customary words is faulty, which is shown by their high false constructive rate. Indeed, even in DOM questions, the most successful term frequently does not correspond with the web character.

E.An Anti-Phishing Tool: PhishProof

Phishing is the craft of attracting clients to a deceitful page by taking on the appearance of a dependable association and getting their own data, for example, Visa number, government disability number, secret word or whatever other data which can be utilized to pick up advantage.

Notwithstanding having a few against phishing instruments the quantity of casualties has expanded drastically finished most recent couple of years as web clients disregard cautioning alarms and the vast majority of the arrangements accessible depend on client input. The proposed arrangement endeavors to enable clients to separate amongst phishing and true blue site pages with insignificant exertion. This venture incorporates a nitty gritty writing audit of existing hostile to phishing procedures. What's more, points of interest and restrictions of every strategy have been examined to comprehend where existing arrangements are deficient. A larger part of the phishing assaults have the aim of getting a monetary advantage and are going for getting clients' close to home subtle elements identified with money related establishments like banks. From the simple initially revealed phishing assault in the 90's the pattern has been set for aggressors to target budgetary foundations. A great many lives are influenced as individuals lose their cash to such tricks. An individual can likewise confront lawful issues if individual data assembled through phishing is utilized to disregard the law. The assaults are not simply constrained to monetary foundations brands were focused by assailants in February and March 2012, which is another unsurpassed high.

F.Anomaly Based Web Phishing Page Detection

The objective of phishing assaults is to take client personalities and qualifications. Phishing assailants utilize different strategies to bait or seize a program to visit sham locales. They may pick a social building, for example, phishing messages, or/and specialized subterfuge, for example, ponies. A typical factor among all phishing destinations is that they perniciously misdirect clients to trust that they are other genuine locales. In this way, recognizing phishing pages is basically a validation issue amongst people and servers. In a perfect world, when a client visits a site, he or she can verify the connected with the web server, regardless of whether the server is gotten to out of the blue. Shockingly, no current specialized instrument completely takes care of this issue. It gives no assurance the HTML records sent by the web server are not misdirecting. A few plans are proposed to empower a client to validate a server with from the earlier security affiliation. Be that as it may, those plans are not appropriate to sites which a client visits

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 6, Issue 1, No 3, 2018 ISSN: 2348-6600
Reference ID: IJCS-337 PAGE NO: 2239-2245

out of the blue. Earlier learning of the objective site, which is sadly not generally accessible. All the more imperatively, since phishing aggressors can refresh the temptation systems to get around those plans, the adequacy of these plans isn't persuading. In a proactive way, an arrangement of methods is intended to catch phishing locales on the Internet. Propose a way to deal with identifying the phishing site pages in view of visual comparability (i.e. square level likeness, format comparability and general style closeness). In any case, they don't straightforwardly guard phishing assaults for clients. A personality of a site is characterized as an arrangement of words (i.e. character strings) which interestingly recognize the site's possession on the internet. In a website page, the character is demonstrated in various items or properties. No current work has tended to the issue of character recovery from web objects. In our plan, they build up a calculation by making utilization of watchword extraction strategies from data recovery writing. Note that a character isn't precisely the watchwords showing up in a page.

III. METHODOLOGY

In this research, propose a way to deal with the discovery of phishing site pages in light of visual likeness. An essential component of phishing site pages is that they resemble the genuine ones in the parts of the website pages and (alternatively) their URLs. Something else, the casualties would not trust them. Consequently, a honest to goodness website page proprietor can look for every single suspicious Url and contrast the relating site pages and the genuine one in visual perspectives. On the off chance that the visual closeness of a site page to the genuine website page is higher than an edge, the proprietor will be cautioned and would then be able to take whatever activities to quickly avert potential phishing attack. In our approach, the visual likeness between two site pages is estimated in three measurements: square level similitude, design comparability, and general style closeness, since a phishing page typically mirror the genuine one by utilizing comparative key locales/squares, comparable page format, and comparable styles (e.g., textual style family, size, adornment, and even dispersing). All these three visual closeness measurements are characterized in light of site page deterioration into an arrangement of remarkable squares.

In our examinations, gathered 8 phishing website pages focusing on 6 genuine authority site pages, together with other 320 ordinary site pages of business banks as the test dataset of suspicious site pages. A utilize each objective site page as an inquiry and attempt to recognize the phishing site pages focusing on it from the 320+8 pages in the dataset. For each suspicious website page, if its likeness to the question in any of the three viewpoints is higher than a particular edge, a phishing case is accounted for. The test comes about to demonstrate that our proposed approach can effectively recognize the phishing site pages with a couple of false alerts. They likewise test the unadulterated content highlights on this dataset yet acquire a less clear limit between the genuine phishing website pages, what's more, other non-phishing pages. Our design input features, which combine three useful URL features (PrimaryDomain, SubDomain, Path-Domain) and three web traffic features (PageRank, AlexaReputation and google Index). Especially, our use google Index instead of using Google search results like traditional methods to improve the accuracy.

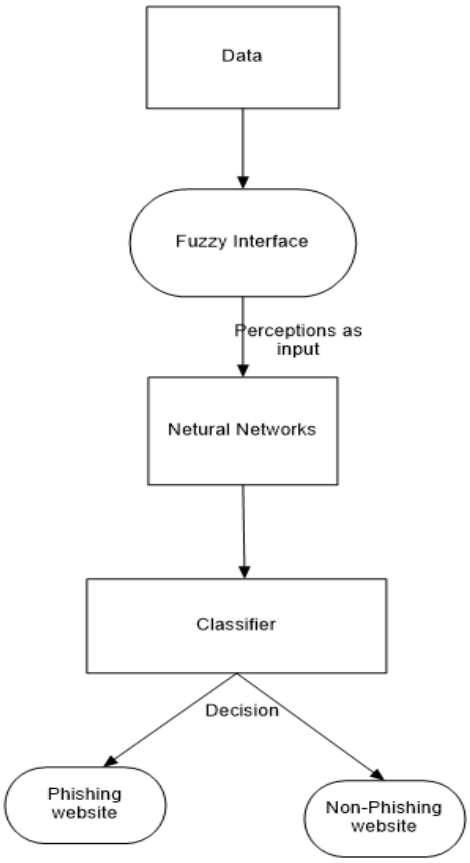

Fig.1 Phishing website detection model

These features are more readily available and faster than web-content features in terms of gathering features and detecting phishing websites. Develop a new neuro-fuzzy approach without using IF-THEN rules to identify phishing.

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 6, Issue 1, No 3, 2018 ISSN: 2348-6600
Reference ID: IJCS-337 PAGE NO: 2239-2245

They are motivated by a prior study that used the conventional neural network model. Combining the neural network with the fuzzy model, obtain a good result in terms of identification accuracy. Our Fi-NFN classification model enhances the classification accuracy to 98:36% and improves the convergence of the training phase. Furthermore, our system can achieve real-time response and stable performance to detect phishing URLs.

Input: d is a PrimaryDomain
Output: The heuristic value of PrimaryDomain
If d is IP then
    Return d belongs to a phishing site;
else
    Result = SuggestionGoogle(d);
if Result is NULL then
    Return d belongs to a legitimate site;
else
    value = Levenshtein(Result; d);
    Return value;
end
end

Calculating the heuristic value of Sub-Domain/PathDomain

Input: m is a SubDomain/PathDomain
Output: The heuristic value of SubDomain/PathDomain
If m is NULL then
    Return m belongs to a legitimate site;
else
    Result = SuggestionGoogle(m);
If Result is NULL then
    Return m belongs to a legitimate site;
else
    value = Levenshtein(Result;m);
Return value;
end
end

The Suspicious URL Detection/Generation Module can be executed by numerous techniques. Notwithstanding, it isn't our concentration in this research. Some heuristic standards can be connected to create every single suspicious Url to a given URL by supplanting a few characters with other comparable ones, e.g., 'o' with '0', as well as including some prefix/addition, e.g., "on the web" and "card". An elective route is to screen the email servers and break down each message got. Since most phishing wrongdoings are started by methods for email, email is the most imperative source to recognize potential phishing URLs. On the off chance that a message contains a catchphrase (e.g., the name of the client organization which asks for this phishing-discovery benefit), every one of the URLs implanted in the message is extricated as the yield consequence of the Suspicious URL Generation Module.

The square level likeness measures the visual comparability of two pages at the level of individual squares. It is characterized as the weighted normal of the visual similitudes of all coordinated square matches between two pages. Essentially, the substance of a square can be arranged as either content or picture. It utilizes distinctive highlights to speak to content squares and picture squares. The highlights for content squares incorporate content substance, shading, textual style, limit, route, and so on., and the highlights for picture squares incorporate alt tag, square shading, measure, source, route, and so forth. The highlights for each square are separated in the component extraction venture in the relating site page handling module and frame the list of capabilities of the square. In the event that the two contrasting squares are of various sorts, their comparability is essentially set to 0 in our present usage. Nonetheless, thinking about later on that I can extricate some textural data from the picture square utilizing an OCR module. For this situation, the picture square is changed over to a content square and their similitude can be figured utilizing the content square closeness estimation technique. Then again, the content square can be changed over to a picture square and the picture square comparability estimation strategy can be connected in a like manner. On the off chance that the two squares are of a similar kind, initially compute their closeness regarding each element in the list of capabilities and afterward utilize a weighted total of the individual component similitudes as the aggregate comparability of the two squares. The heaviness of each component implies its significance to the aggregate comparability and can be doled out observationally. In our usage, concentrate more on shading related highlights.

The component comparability of each extraordinary element compose is characterized in an unexpected way. In the event that the conceivable estimations of a component are enumerative or discrete (e.g., the textual style family), the closeness esteem for that element is twofold. On the off chance that the conceivable estimations of a component are generally ceaseless (e.g., text dimension or shading), the comparability esteem for that element is just characterized as 1 short the standardized distinction of the element esteems. Two squares are considered as coordinated if their similitude is higher than a limit. After I get the likeness estimations of all sets of conceivable coordinating squares, attempt to locate a coordinating plan between the two pages' squares. This is really a bipartite chart coordinating issue, and an avaricious

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 6, Issue 1, No 3, 2018 ISSN: 2348-6600
Reference ID: IJCS-337 PAGE NO: 2239-2245

like coordinating calculation is utilized to locate the coordinating squares for the genuine page hinders in the plummeting request of their weights. The heaviness of each square in the genuine site page is created in the Key Block Labeling and Weighting venture in the True Webpage Processing module which is talked about. The key squares are coordinated first. When attempting to locate a coordinating for a staying genuine page obstruct, its most comparative suspicious site page hinder in the rest of the hopeful rundown is chosen as its coordinating square and afterward expelled from the rest of the applicant list.

IV. Result and Discussion

Besides, to test our approaches capacity to maintain a strategic distance from false cautions, likewise gathered an arrangement of 200 typical site pages, which are record pages from the official sites of 100 business banks. These 420 ordinary website pages together with the 5 phishing pages are utilized to frame the test dataset. All these test website pages can be downloaded at. The 6 genuine pages are viewed the inquiry to look for outwardly comparative pages in the test dataset. Once the similitude in term of any of the three measurements between a page and the inquiry is bigger than a limit t, the framework will report that this page is most likely a phishing page. As a matter of fact, this is the way how our way to deal with phishing site page identification works. I trust that the approach can be utilized as a piece of an against phishing system in an undertaking arrangement.

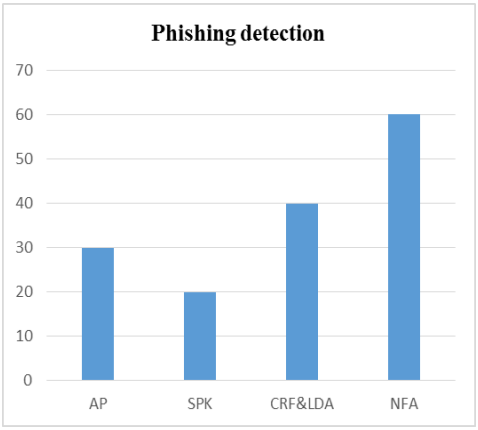

Fig 2. Phishing detection analysis

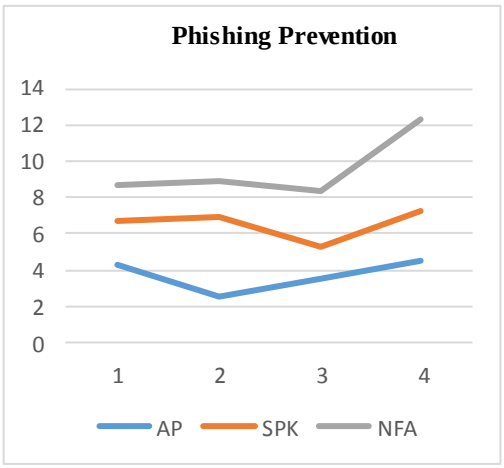

Fig 2. Phishing prevention analysis

A genuine website page proprietor can utilize this way to deal with identify phishing site pages. Besides, the use of the proposed approach isn't restricted to a discovery of this sort of phishing assaults. It can be utilized for discovery of all perniciously manufactured site pages impersonating the website pages of any association and individual. Our proposed NFA give efficient phishing prevention and detection compared with other existing approaches.

V. Conclusions

In this research, propose a novel approach to identify phishing site pages in light of visual comparability. The approach initially breaks down the website pages into notable squares as indicated by visual prompts. The visual likeness between two website pages is then estimated in three perspectives: square level similitude, design closeness, and general style comparability. A website page is accounted for as a phishing suspect if any of these likenesses to the genuine site page is higher than a limit. In view of the benefits of the mist design and the enhanced neuro-fuzzy approach, propose a phishing ID demonstrate, called Fi-NFN, to secure neighborhood gadgets effortlessly and rapidly. Without expending numerous assets from nearby gadgets, our Fi-NFN demonstrate straightforwardly ensures clients progressively, as well as enhances the nature of administrations at the edge of the system.

References

[1] Y. Zhang, J. I. Hong, and L. F. Cranor, “Cantina: a content-based approach to detecting phishing web sites,” in Proceedings of the 16th international

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 6, Issue 1, No 3, 2018 ISSN: 2348-6600
Reference ID: IJCS-337 PAGE NO: 2239-2245

conference on World Wide Web. ACM, 2007, pp. 639–648.
[2] PhishTank. Accessed Nov 2015. [Online]. Available: http://www.phishtank.com/stats/2014/01/.
[3] S. Sheng, B. Wardman, G. Warner, L. F. Cranor, J. Hong, and C. Zhang, "An empirical analysis of phishing blacklists," in Proceedings of Sixth Conference on Email and Anti-Spam (CEAS), 2009.
[4] V. M. Gandhimathi K1, "Identifying similar web pages using scoring methods for web community mining," in Proc. of Int. Conf. on Advances in Computer Science, AETACS, 2013.
[5] B. B. Gupta, A. Tewari, A. K. Jain, and D. P. Agrawal, "Fighting against phishing attacks: state of the art and future challenges," in Neural Computing and Applications, 2016.
[6] Fog computing and the internet of things: Extend the cloud to where the things are. [Online]. Available: http://www.cisco.com/c/dam/en us/solutions/trends/iot/docs/computing-overview.pdf.
[7] Y. Li and M. Chen, "Software-Defined Network Function Virtualization: A Survey," IEEE Access, vol. 3, pp. 2542–2553, 2015.
[8] G. Xiang, J. Hong, C. P. Rose, and L. Cranor, "Cantina+: a feature-rich machine learning framework for detecting phishing web sites," AC Transactions on Information and System Security (TISSEC), vol. 14, no. 2, pp. 1–28, 2011.
[9] W. Hadi, F. Aburub, and S. Alhawari, "A new fast associative classification algorithm for detecting phishing websites," Applied Soft Computing, vol. 48, pp. 729 – 734, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1568494616303970.
[10] L. T. C. James, J. ; Sandhya, "Detection of phishing urls using machine learning techniques," in Control Communication and Computing (ICCC), 2013 International Conference on, 2013, pp. 304–309.
[11] M. Aburrous, M. A. Hossain, K. Dahal, and F. Thabtah, "Intelligent phishing detection system for e-banking using fuzzy data mining," Expert systems with applications, vol. 37, no. 12, pp. 7913–7921, 2010.
[12] A.N. V. Sunil and A. Sardana, "A pagerank based detection technique for phishing web sites," in Computers & Informatics (ISCI), 2012 IEEE Symposium on. IEEE, 2012, pp. 58–63.
[13] Chowdhury A. Duplicate data detection (an overview). AOL. http://ir.iit.edu/~abdur/Research/Duplicate.html.
[14] Chowdhury A., Frieder O., Grossman D., and McCabe. M. Collection statistics for fast duplicate document detection. ACM Trans. on Information Systems (TOIS) 20(2): 171–191, 2002.
[15] Hoad T.C. and Zobel J. Methods for identifying versioned and plagiarised documents. Journal of the American Society for Information Science 54(3): 203–215, 2003.
[16] Salton G., Wong A., and Yang C.S. A vector space model for information retrieval. Journal of the American Society for Information Science 18(11), pp. 613–620, 1975.