ISSN (ONLINE) : 2395-695X
ISSN (PRINT) : 2395-695X
Available online at www.ijarbest.com

International Journal of Advanced Research in Biology, Ecology, Science and Technology (IJARBEST)
Vol. 1, Issue 1, April 2015

PREVENTION OF CYBERCRIME BY SUSPICIOUS URL DETECTION IN SOCIAL NETWORKS USING ENHANCED DBSCAN ALGORITHM

R. Ravi[1], **Dr. Beulah Shekhar**[2]
Department of Computer Science & Engineering, Francis Xavier Engineering College, Tamil Nadu, India[1]
Department of Criminalogy , Manonmanium Sundaranar University, Tamil Nadu, India[2]

Abstract – ***Cyber law is a term that refers to all legal and with its regulatory portion of World Wide Web and the internet. Since Cybercrime is a newly specialized field, growing in Cyber laws, a lot of development has to take place in terms of putting into place the relevant legal mechanism for controlling and preventing Cybercrime. Social Network is susceptible to malicious message containing URLs for spam, phishing and malware distribution. Mainly we are dealing with Twitter, where we find suspicious address by using application tool manually and from that manual coded tool, we are calling API. In this research, we propose Warning Tool, a suspicious address detection scheme for twitter. Our structure investigates correlation of URL which redirect the chains that are most regularly share the similar URLs. Here the developmental methods to determine correlated URL by using application tool. Phishing website detection can be said as new to the arena. Phishing websites are considered as one of the lethal weapon to embezzle one's personal information and use it for the cracker's benefits. The Detection course has been divided into two steps 1) Feature Extraction, where the ideal features are extracted to capture the nature of the file samples and the phishing websites. 2) Categorization, where exceptional techniques are used to automatically group the file samples and the websites into different classes.***

Key Words: Suspicious URL, DBScan, Clustering, Cyber Laws, SQL Injection.

I. INTRODUCTION

In the beginning of certain development era, people are well motivated to attain a good progress in existing technological activities. Since the start of civilization period, the man has been always motivated by need to make better progress in the existing technologies. These networks are steadily grew and then called as online Internet activity for sharing business regime where communication done in both modes is called as a Cyberspace[9]. Cyber law is a common term which refers to legal jurisdiction and other ways of preceding regulatory aspects in internet. It is a constantly generic process. Whenever an internet development strategy follows, we enforce numerous amendments while it grows and there are numerous legal issues will be raised.

The traffic hazards, allotment in distributions, dissemination of obscene material, and posting includes pornography with all its filthy exposure constitutes the most important known criminal cyber offence today. This is one who threatens to challenge the development of the younger creation in cybercrime and also leaving permanent scar and damage on the younger generation, if can't restricted. Clustering can be measured as the most important unconfirmed learning problem[14].

Every other problem of this kind deals with sentence a structure in a compilation of unlabeled information. A loose definition of cluster could be "the process of organize objects into groups whose member are similar in some way". The goal of clustering is to decide the inherent grouping in a set of unlabeled information. But to decide how and what will constitute a high-quality clustering, which can be exposed that present is no absolute "best" criterion which constitute would be self-governing of the final aim of the clustering. As a result, it is the user which must provide these measures, in such a way that the results of the clustering will ensemble their requirements. The linear combination strategy has been also used in other bioinformatics problems, such as gene clustering with multiple data (or constraints), including Gene Ontology, metabolic networks, and gene expression. Clustering algorithms can be functional in many fields, for occurrence such as Marketing, Libraries, Biology, Insurance, WWW for clustering weblog data to determine collection of related right to use pattern and other classification[10]. Software such as anti-malware, anti-virus and firewalls which are relied upon by users at small, home and large organizations around the world to defend against malware attacks which helps in identify and prevent the further extend of malware in the network.

Many early infectious programs with the original Internet Worm, which were written as experiment or trouble. Today, malware is used primarily to steal financial, sensitive information of personal, or business importance by black hat hackers with damaging intention.

II Related Works

These goal of clustering helps in deciding the set of unlabeled information and also to decide to constitute a good clustering schemes. These can be stated as a "best" criterion which constitute would be self-governing of the final identity of the clustering. As a result, these users must supply this measure, in such a way that those results of the clustering will ensemble their requirements. For example, in finding the council for homogeneous cluster (in data reduction), which is used for deciding "natural clusters" and illustrate their unknown properties ("natural" data types), which is to dealt in finding practical and appropriate groupings or in decision in unusual data objects (outlier detection) could be interesting.

Traditionally, only local-content information of documents from the data set has been utilized for clustering, where each document is represented by "bag of words," resulting in a weighted vector according to the so called vector space model, and then, clustering is carried out on weighted vectors[7].

Clustering algorithms can be functional in many fields, for occurrence:

Marketing: finding groups of customers with parallel behavior in a given large database of customer data contain their past buying records and properties

Libraries: book ordering

Biology: classification of animals and plants using their features

City-planning: identification in groups of houses according to the detailed residence type, value and geographical location

Insurance: identifying clusters of motor insurance strategy holders with a high average claim cost, for recognize frauds

Earthquake studies: Identification of dangerous zones using clustering observed earthquake epicenters

WWW: clustering weblog data to determine collection of related right to use pattern, document classification

The requirements of a clustering algorithm should satisfy the following requirements are:

- discover clusters with subjective shape
- scalability
- dealing with different types of attributes

- minimal requirements for realm data to determine input parameter
- inattentiveness to order of input records
- ability to deal with clatter and outliers
- Interpretability and usability.
- high dimensionality

Clustering algorithms may be classified as enlisted below:
Overlapping Clustering

- Exclusive Clustering
- Probabilistic Clustering
- Hierarchical Clustering

BIRCH Clustering

BIRCH is a Balanced Iterative Reducing and Clustering using Hierarchies which is a data mining algorithm which is unsupervised and it is used to perform hierarchical clustering over large data-sets. BIRCH advantages is its ability to incrementally and cluster incoming links generated dynamically with its multi-dimensional measures data points in an challenge to make the best quality clustering from a given set of resources (identified by its memory and time constraints)[15]. In some situation BIRCH requires a single scan of the database application. In addition, BIRCH also claims to be the first to propose a clustering algorithm in the database storage area to handle including the noise (data points may not be a part of the primary model) effectively" beating DBSCAN by two months of its effort.

Step 1: In the first step of the algorithm, it scans all data and builds an initial CF tree memory using the given emory amount.
Step 2: In the second step, it scans all the leaf level entries which is in the initial CF tree to restructure a smaller CF tree, while grouping crowded sub clusters and removing outliers into larger ones.
Step 3: In third step, an existing clustering algorithm which is used to cluster all leaf entries. Here an agglomerative hierarchical clustering type of algorithm is functional directly to the sub clusters represented in their CF vectors.
Step 4: Step four, also afford us with an option of removal outliers. This is the reason which is too distant from its neighboring seed can be care for as an outlier.

III PROPOSED METHODOLOGY

Enhanced DBSCAN Algorithm

Enhanced Density-based spatial clustering of application with noise (EDBSCAN) is a data clustering algorithm which was proposed in 1996 by Martin Ester, Hans-Peter Kriegel, Xiaowei Xu and Jörg Sander as DBSCAN techniques. It is also a Enhanced density-based clustering algorithm because it locates a number of clusters opening from the predictable density allotment of corresponding nodes includes both pixel and points[15]. Enhanced DBSCAN is one of the large numbers of common clustering algorithms and also which was the most cited in technical writing. OPTICS can be noted as a generalization of EDBSCAN to multiple ranges, efficiently replacing the bound values with maximum search radius information.

This entire methodology works under the working principle of the clustering techniques in fig 3.1 used are the clustering methodologies of Enhanced DB-scan used in huge datasets as it is a better advantage in the world of internet.

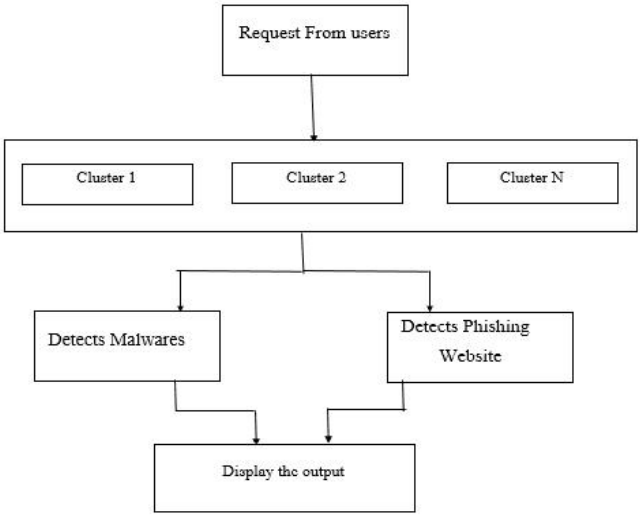

Fig 3.1 Architecture of Clustering in detection

IV ALGORITHM USED

The algorithm for the redirect state is given as follows.

```
Begin
    Accept URL
    Open Connection form URL
    Accept response code
    If (response code!=200)
        Begin
            Redirect=true;
        End
    End if
    While (redirect)
        Begin
            Get new URLlocation
            NumberHops++;
            If (NumberHops>5)
                Begin
                Block the webpage
                End
            End if
End
Go to webpage
End loop
End
```

Fig 4.1 Algorithm to block the WebPages with more than 5 redirects

```
DBSCAN(D, eps, MinPts(minimum points))
C = 0;
for each searched link P in dataelement D
Mark P as Linked
fellowPts = sectionQuery(P, eps)
if sizeof(fellowPts) < MinPts
mark P as NOISE
else
C = next cluster
expandCluster(P, NeighborPts, C, eps, MinPts)
add P to cluster C
 for each point P' in NeighborPts
 if P' is not visited mark P' as searched
fellowPts' = sectionQuery(P', eps)
if sizeof(fellowPts') >= MinPts
fellowPts = fellowPts joined with fellowPts' if P' is not a associate of some cluster then add P' link to cluster C
```

sectionQuery(P, eps) return all points inside P's eps-neighborhood node that are visited.

Fig 4.2 Enhanced DBSCAN Algorithm

The basic clustering techniques used in this context are changed in the projected proposed system. The clustering techniques used are DB scan and BIRCH, where BIRCH stands for balanced iterative reducing and clustering using hierarchies. The transform in the basic clustering expertise provides that this scheme can be used in better space. As the malware samples and phishing websites are proliferated, the ACS resulted in to capture them and also store in the database which is segregated for the further use. So it needs very bulky data sets which can be given that by the BIRCH and DB scan. The results are also excellent when compared to the existing system this detects the greatest quantity of the malware samples and phishing websites. The malware samples and phishing websites details were also taken from the internet for its verification.

Advantages of Proposed System

- By using these base clustering algorithms phishing and malwares can be detected in large dataset.
- By using large dataset techniques the whole data can be scanned only once and it provides the best result for the phishing websites and malware samples.
- These clustering techniques can be suspended, resumed and stopped.

V EXPERIMENTAL SETUP

Here the obtained comparison Figure shows the comparison of different clustering techniques used in the existing system and proposed system with the Cost, Implementation and Upgrade difficulties. In the graph the X-axis stands for Clustering techniques and the Y-axis stands for Maintenance. From the graph it is clear that the proposed clustering techniques is easy to implement, cost efficiency and easy to upgrade.

In India, Phishing and Malware Spreading are cognizable, bailable and compoundable with permission of the court before which the prosecution of such offence is pending and triable by any magistrate under Information Technology (Amendment) Act, 2008. In the future, the basic clustering techniques can be changed and it can be tested and also the feature extraction techniques can be changed and tested.

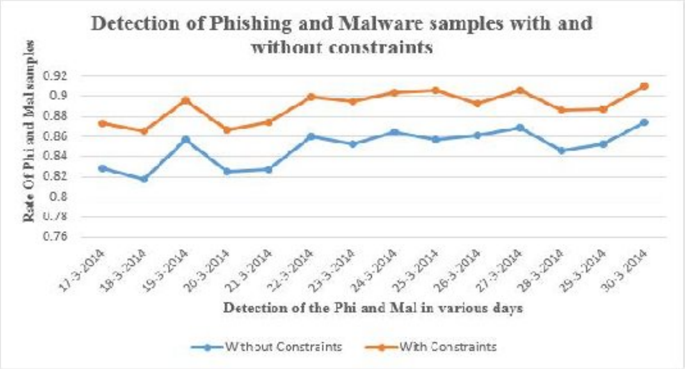

Fig 5.1 Comparison for detecting Phishing and Malware

With these techniques placed, the users also should have some intelligence about what webpage they are using and what they are clinking on the internet. They should at least know not to provide their SSN number, credit card details, Bank account number and passwords. The users also need to understand the risk if they were providing the passwords and other private matters to others or in unknown sites.

CONCLUSION

Manual Inspections represents the human-driven review that is used to test the security results of people, policies, and processes. It also includes the inspections of technology expectations such as architectural designs. These are conducted by analyzing documentation or performing dialogue with the system owners or designers. While the manual inspection is still simple, it is a most powerful and effective techniques. It allows the tester to determine its security concerns that are likely to be evident by asking the system designers how it works and even though it is implemented in a specific way.

REFERENCES

1. Abu-Nimeh S., Nappa D., Wang X., and Nair S. (2007), ‘A comparison of machine learning techniques for phishing detection’ in Proc. APWG eCrime Res. Summit, p. 60–69.

2. Aburrous M., Hossain M.A., Dahal K. and Thabtah K. (2010), ‘Predicting phishing websites using classification mining techniques with experimental case studies’ in Proc. 7th Int. Conf. Inf. Technol., pp. 176–181

3. Benevenuto F., Magno G., Rodrigues T. and Almeida V. (2010), ‘Detecting Spammers on Twitter’, Proc. Seventh Collaboration, Electronic Messaging, Anti-Abuse and Spam Conf. (CEAS).

4. Dolatabadi, H. , Shirazi, M.N. ; Hejazi, M.(2011) New mechanism to confront injection attacks.

5. Chou N., Ledesma R., Teraguchi Y., Boneh D., and Mitchell J.C (2004), ‘Client side defence against web-based identity theft’ in Proc. 11th Annual Network Distrib. Systems Security Symposium

6. Garzotto F., Baresi L., Paolini P. (2011) From Web Sites to Web Applications:

7. Dazeley R., Yearwood R.L., Kang B.H., and Kelarev A. V (2010), ‘Consensus clustering and supervised classification for profiling phishing emails in internet commerce security’ in Knowledge Management and Acquisition for Smart Systems and Service, Vol. 6232, pp. 235–246.

8. El-Bahlul Fgee., Ezzadean H.Elturki and A.Elhounie (2012) ‘My Security for Dynamic Websites in Educational Institution’ Sixth International Conference on Next Generation Mobile Applications, Services and Technologies

9. Elovici Y., Shabtai A., Moskovitch R., Tahan G., and Glezer C. (2007), ‘Applying machine learning techniques for detection of malicious code in network traffic’ in KI 2007: Advances in Artificial Intelligence, Vol. 4667, pp.44–50.

10. Ester M., Kriegel H.P., Sander J., and Xu X. (1996), ‘A density-based algorithm for discovering clusters in large spatial database with noise’ in Proc. ACM International Conference Knowledge Discovery Data Mining, pp. 226–231.

11. Fredrikson M., Jha S., Christodorescu M., Sailer R., and Yan X. (2010), ‘Synthesizing near-optimal malware specifications from suspicious behaviors’ in Proc. IEEE Symposium Security Private, Washington, DC IEEE Computer Society, pp. 45–60.

12. Grier c., Thomas k., Paxson v. and Zhang M. (2010), ‘@spam: The Underground on 140 Characters or Less’, Proc. 17th ACM Conf. Computer and Comm. Security (CCS).

13. Herley C. and Florencio D. (2008), ‘A profitless endeavor: Phishing as tragedy of the commons’ in Proc. New Security. Paradigms Workshop.

14. Klien F. and Strohmaier M. (2012), ‘Short Links under Attack: Geographical Analysis of Spam in a URL Shortener Network’, Proc. 23rd ACM Conf. Hypertext and Social Media (HT).

15. Lee S. and Kim J. (2012), 'Warning Bird: Detecting Suspicious URLs in Twitter Stream', Proc. 19th Network and Distributed System Security Symp. (NDSS).

16. WASP Testing Project (2013) https://www.owasp.org/index.php/Testing_Guide_Frontispiece