A Data Mining approach to Deal with Phishing URL Classification Problem

Sonam Saxena
Department of Computer Science Engineering
Swami Vivekanand College of Engineering
Indore, India

Amit Shrivastava
Assistant Professor
Department of Computer Science Engineering
Swami Vivekanand College of Engineering
Indore, India

Vijay Birchha
HOD Professor
Department of Computer Science Engineering
Swami Vivekanand College of Engineering
Indore, India

ABSTRACT
Data mining is readily growing and accepted technology in recent years. It is utilized for finding instant decisions by analyzing the historical records. A formal decision making technique can also be helpful for information security. In this presented work the demonstration of a data mining application is provided. The proposed data mining application contributes on the information security. Therefore URL classification problem is taken in consideration. In this context we can apply here the any supervised learning algorithm but in this work the association rule mining based technique is proposed for solving the URL classification. That technique is used for analyzing the URL patterns of two kinds of class labels i.e. phishing and legitimate. In this context a rule based classification technique is proposed. That technique is computing the association rules and we can use these patterns to classify the URL data. The Idea is taken from [1] where apriori algorithm is implemented for generation and classification of phishing URLs. Apriori algorithm is computationally complex and requires significant amount of time and memory for generating candidate sets. Therefore we usages the FP-Tree algorithm which efficient develops the association rules with less resource requirements. The system can be used for designing the phishing tool bars. This technique is used with the phish tank dataset with different set of data for experimentations. The obtained results shows the proposed technique requires less amount of time and memory. In near future it is tried to reduce time and improve the accuracy of the proposed phishing URL classification system.

Keywords
classification, rule based classification, association rule mining, phishing URLs, FP-Tree;

1. INTRODUCTION
A significant amount of users are utilizing the internet services, among them some of the users are new and some of them are experts. The new users are not much aware about the negative aspects of internet services. Therefore they are soft target for the different cyber criminals. The cyber criminals are technology experts are they are able to find loop holes of the technology and can deploy various kinds of attacks for new users. Among them the phishing attack is one of the serious and complex attacks. In this attack attacker tried to get baking and financial attributes from the end or target users. Therefore a malicious URL is pushed to the user and user visit the URLs and provides their confidential information to the attacker. Attacker is usages this information for cheating the target user. Therefore in most of the attacks the attacker utilizing the URLs so the URLs classification is a good strategy for identifying the phishing URLs. There are a number of techniques available for classifying or identifying the phishing URLs, among most of them either inefficient or not much accurate. Therefore the proposed work is motivated to design an efficient and accurate phishing URL classification technique. The proposed technique reduces the resource consumption and improves the accuracy of classification of the existing classification technique. In this context the proposed work is motivated to use the association rule mining technique for employing in this task. The FP-Tree is one of efficient association rule mining technique thus the proposed work incorporate the study of the FP-Tree algorithm and their applicability in rule based classification.

2. PROPOSED WORK
The proposed work is aimed to use the data mining techniques for classifying the Phishing URLs using the association rule mining technique. This chapter provides the understanding about the proposed functional data model. Therefore the functional aspects of the proposed system are discussed in this chapter.

2.1 System Overview
The phishing is a cyber crime which acts to harm the target web user in terms of their financials, by using the user's personal information. In this context, attacker is usage different techniques and tricks to still the banking and/or credit card information by sending the false links. These links contains the forged web pages and malicious scripts by which the user passes their confidential information to the attacker. In most of the phishing cases the attacker communicates with the target attacker by e-mail, SMS or any other communication medium to deliver the malicious links. The literature demonstrates the various efforts and techniques which are currently being used for identifying such of these malicious URLs. But most of the time the attacker change their tricks, therefore the awareness is prevention of the phishing attack.

But after a time it is required to develop the enhance technique which recover the gap and accurately classify the phishing URL patterns. In this context the data mining is helpful for analyzing the large amount of data and recovers the application targeted patterns. The data mining techniques are having the ability to compare and obtain the properties which are helpful for target pattern identification. In this presented work the association rule mining based machine learning technique is proposed for study and system design. The association rules mining find the relationship among two or more attributes (features) to identify the phishing URL patterns. Therefore it is a rule based classification approach for classifying the patterns. This section provides the basic understanding about the proposed data mining based phishing URL classification system. The next section involves the details of the proposed

system and their working process.

2.2 Methodology

The proposed system is described using the figure 1. This diagram contains the different functional steps that evaluate and transform data in each stage and provide the outcomes for next steps. The details of the system are given as:

2.2.1 PhishTank Dataset: The phish tank is an online database which is open to access by directly using the application using API (application programming interface) or by using the downloading the file in CSV format. The different security agencies and institutions when a new URL detects then they report it to the phish tank database. The database contains the following key attributes in their database.

Phish ID: That is a unique ID assigned by phish tank database as most of the database consists of the unique IDs for data instances for recognizing the phishing URLs by ID.

URL: The URL which is used for targeting the innocent web users is given in URL attributes. In other terms the attacker usages this URL for targeting the web user.

Phish detail URL: This attribute contains an additional URL which contains a web page for providing details about the identified phishing URL.

Submission date: This attribute provide a date when the URL is identified as phishing URL first time.

Verification time: This attribute demonstrate the time when the URL is verified as phishing URL

Online: This attribute provide the status of the URL and availability of target URL

Target: This attribute define the target company or organization for that the clone is prepared and hosted

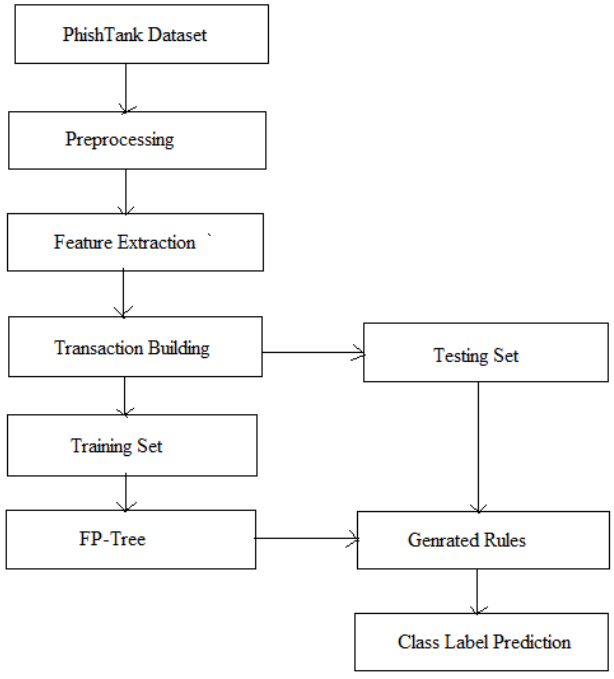

Figure 1 Proposed system architecture

2.2.2 Preprocessing: The preprocessing technique is used for transforming the datasets. The dataset transformation is performed in such a way by which, machine learning algorithm easily employed on the datasets therefore the preprocessing is aimed for optimization of the dataset. The obtained dataset from the phish tank contains a number of unused data and attribute. These attributes and features are not essential for preparing the proposed URL classification strategy therefore the URL is kept preserved and remaining attributes are removed from the initial dataset.

2.2.3 Feature selection: The URLs are the unstructured kind of data; therefore direct classification of these data is not feasible. Thus the essential attributes are needed to be recovering for accurate classification of URLs. The article [1] contains 14 functions for describing the properties of phishing URLs the given features and their phishing conditions are reported in table 1.

Table 1 Target features

S. No.	Features	Description
1	URL length	If URL length is higher than the 25 character then it may be a phishing URL
2	top level domain	The 66% of malicious URLs are not containing the Top level domain
3	number of dots in the path of the URL	If the number of dots in the URL path is found to be less than two then it may possible it is a phishing URL
4	certain keyword in the URL	According to the evaluation 91% of phishing URLs contains certain keywords
5	hyphen in the host name of the URL	The number of hyphen in the host part of the URL is found to be greater than 1 for phishing URLs.
6	Subdomain	The 64% of phishing URLs contains the sub-domains in there URLs
7	Unicode in URL	It is observed that the 64% of phishing URLs contains the uni-code characters
8	transport layer security	99% of phishing URLs are not using the HTTPS protocol
9	length of the host URL	The average length of the phishing URL is found to be greater than 75

10	dots in host name of the URL	If the number of dost are greater than or equal to 4 then it may be a phishing URL
11	IP address	94% of phishing URLs contains the IP addresses
12	special characters	If the URL host address contains the special characters then it may possible it is a phishing URL
13	number of slashes in URL	If a URL contains total number of slashes greater or equal to 5 then it may be phishing URL
14	number of terms in the host name of the URL	If the host name contains more than four terms then it may be a phishing URL

2.2.4 Transaction building: The feature computation of the system included 14 features (attributes). Additionally it is not feasible to have all the properties in a URL, therefore the features are which extracted in previous phase is used for generating the transactions for processing the URL samples using the association rule mining algorithms.

2.2.5 Training set: The cross validation technique is machine learning requires a set of data for taking the training and an additional set of samples are required to test the performance of learning algorithm. Therefore 70% of random data instances are utilized for training or rule generation and remaining samples are consumed for performance evaluation of the developed system.

2.2.6 Testing set: The testing set is prepared by using 30% of entire samples which are processed before. The selection of these feature instances is performed on the basis of random order. These set of instances are used for predicting the class labels using the prepared data model. Additionally it is used also for computing the performance of classification system.

2.2.7 FP-Tree algorithm: The FP-Tree algorithm is an association rule mining algorithm which is used for frequent tree generation. That technique is efficient and scalable as compared to the apriori algorithm. Thus when we need efficiency in association rule computation then the FP-Tree algorithm is used. The FP-Tree algorithm is described in chapter 2 in detail.

2.2.8 Rule generation: The process of FP-Tree generates the association rules in form of tree. These rules can also be represented using the "IF-THEN-ELSE" rules where the else part contains the class labels of the rules. Thus this phase contains the set of rules which are generated by FP-Tree algorithm for classification task.

2.2.9 Class label prediction: That is the final outcome of the proposed phishing URL classification system. In this phase by traversing the generated rules the class labels of the URLs are identified / predicted.

2.3 Proposed Algorithm

The proposed algorithm is used for recognizing the class labels of the given URLs in terms of malicious or legitimate. Therefore the above given process is summarized in the steps of process involved. The table 2 contains the required steps of the proposed algorithm.

Table 2 Proposed algorithm

Input: phish tank dataset D

Output: URL class labels

Process:

1. $D_n = ReadDatasetInstances(D)$
2. $P_n = preProcessData(D_n)$
3. $for(i = 1; i \leq n; i++)$
    a. $for(j = 1; j \leq 14; i++)$
        i. $F_{i,j} = computeFeature(P_i, Fun_j)$
    b. $end\ for$
4. End for
5. $[Train, Test] = SplitDataset(F_{i,j}, 70{,}30, random)$
6. $Rules_m = FPTree.GenrateRules(Train)$
7. $C = Rules_m.Predict(Test)$
8. Return C

3. RESULTS ANALYSIS

After successfully implementation of the proposed technique the performance evaluation of the proposed URL classification system is conducted. Based on the obtained experimental outcomes the key parameters are computed and reported in this chapter.

3.1 Accuracy

The accuracy of a data mining algorithm is measurement of correctness of data classification. In this context the algorithm how accurately identify the given patterns are termed as the accuracy of algorithm. To compute the accuracy of algorithm the following formula can be used:

$$accuracy(\%) = \frac{total\ correctly\ classified\ pattern}{total\ pattern\ to\ classify} X100$$

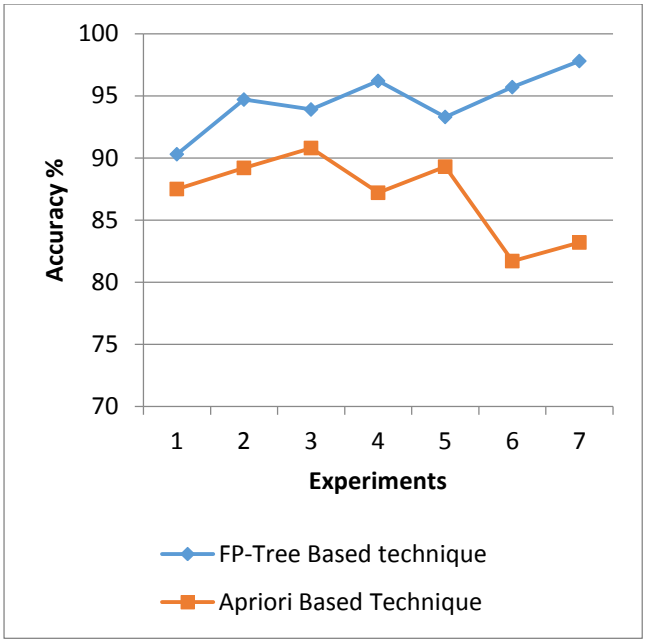

Figure 2 Accuracy (%)

Table 3 Accuracy (%)

S. No.	FP-Tree Based technique	Apriori Based Technique
1	90.3	87.5
2	94.7	89.2
3	93.9	90.8
4	96.2	87.2
5	93.3	89.3
6	95.7	81.7
7	97.8	83.2

The comparative performance of both the algorithms i.e. traditional Apriori based technique and the proposed FP-Tree based approach is reported using table 3 and figure 2. The table 3 contains the observed experimental outcomes; additionally their line graph representation is given in figure 2. The blue line shows the accuracy of proposed FP-Tree based technique and red line shows the performance of the Apriori algorithm based technique. According to the results the proposed technique demonstrates the higher accuracy of classification as compared to the proposed technique of phishing URL classification.

3.2 Error Rate

The error rate is the measurement of incorrectness of a trained machine learning algorithm. Therefore that is defined by ratio of incorrectly recognized data patterns and total data samples provided for classification. In this context the percentage error measurement the following formula is used:

$$Error\ rate\ (\%) = \frac{Incorrectly\ classified\ data\ patterns}{total\ patterns\ to\ classify} X100$$

Or

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

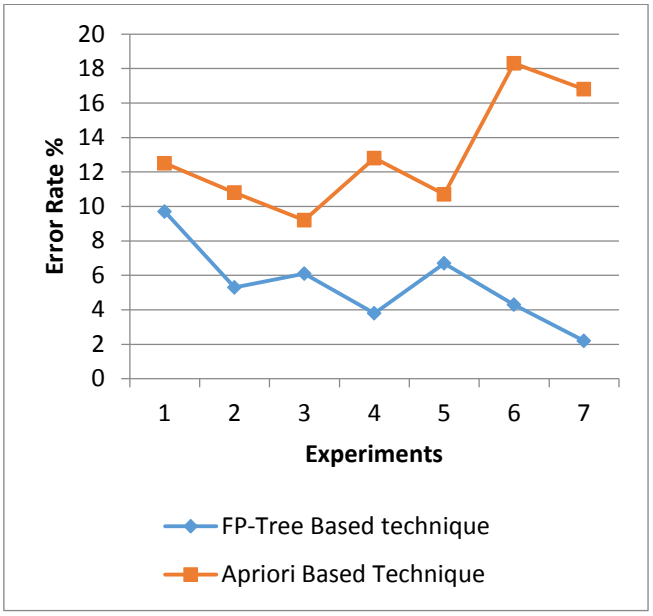

Figure 3 Error rate (%)

Table 4 Error rate (%)

S. No.	FP-Tree Based technique	Apriori Based Technique
1	9.7	12.5
2	5.3	10.8
3	6.1	9.2
4	3.8	12.8
5	6.7	10.7
6	4.3	18.3
7	2.2	16.8

The comparative performance in terms of percentage error rate for both the algorithms i.e. proposed FP-Tree based algorithm and the traditional Apriori algorithm is given in table 4 and figure 3. The table consists of the experimental values and their line graph representation is given in figure 3. The line graph shows the performance of proposed FP-Tree based technique using blue line and the red line is used for demonstrating the Apriori algorithm. According to the results the proposed technique produces less amount of error rate as compared to traditional Apriori based technique.

3.3 Memory Usages

The memory usages are also known as the space complexity of an algorithm. The memory usages of an algorithm in java based technology are computed by the total memory assigned to the process and the total free memory size. Therefore it is denoted using the following formula:

$$Memory\ usages = total\ assign\ memory - free\ memory\ space$$

The memory usages of the FP-Tree based and Apriori algorithm based technique is given in figure 4 and table 5. The memory usages for both the algorithms are computed in terms of KB (kilobytes). The table 5 shows the obtained experimental

values and that is represented using line graph as given in figure 5. The blue line of line graph shows the performance of the proposed FP-Tree based phishing URL classification technique and the red line shows the performance of the traditional apriori based algorithm. According to the visual results the proposed technique requires less amount of memory as compared to the traditional Apriori algorithm.

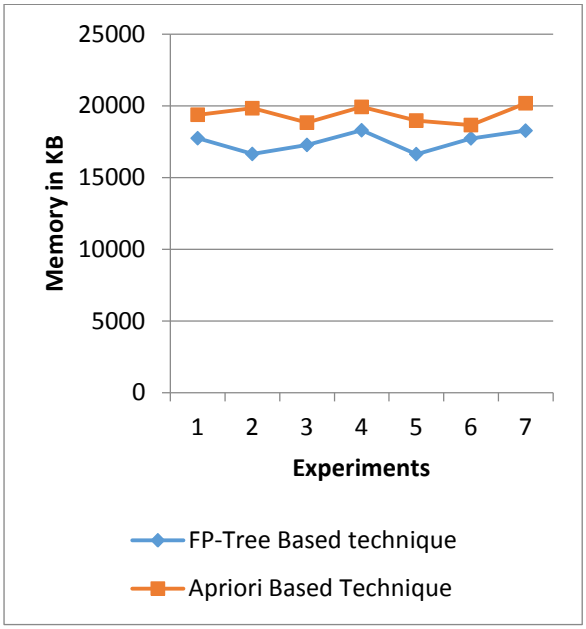

Figure 4 Memory usages in KB

Table 5 Memory usages in KB

S. No.	FP-Tree Based technique	Apriori Based Technique
1	17736	19372
2	16639	19826
3	17269	18821
4	18291	19917
5	16628	18964
6	17725	18652
7	18264	20173

3.4 Time Complexity

The time complexity is the amount of time which is required to process all the input data according to algorithm behavior. This parameter is computed on the basis of the time difference between algorithm initialization and finalization. Therefore the following formula is used:

$$time\ complexity = end\ time - start\ time$$

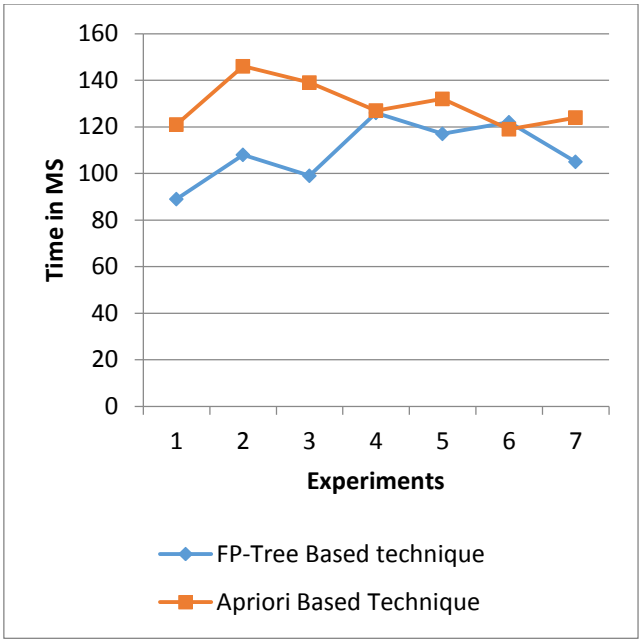

Figure 5 Time complexity

The time utilization of both the algorithms is measured in terms of milliseconds (MS). The obtained time consumption of both algorithms is reported in table and the graphical representation is given in figure 5. The X axis of this diagram shows the number of experimental observations and the Y axis shows the consumed time in terms of milliseconds (MS). According to the obtained results the proposed technique requires less amount of time as compared to the traditional apriori based URL classification.

4. CONCLUSION AND FUTURE WORK

The proposed efficient and accurate phishing URL classification system is implemented and evaluated successfully. According to the obtained experimental outcomes the summary of the entire efforts conducted prepared. In addition of the future extension of the work is also reported in this chapter.

4.1 Conclusion

The data mining techniques are currently utilized in various real world problems for predicting, classifying and recognizing the patterns. Therefore the computational algorithms are applied on data for evaluation. In this context the techniques supports the classification, clustering and association pattern mining techniques. In this presented work the data mining techniques are used for discovering the malicious URL patterns. These malicious patterns of URLs are used for phishing attack deployments. During such kind of attacks the attacker prepares the false web URLs to get the essential and user sensitive information. In this data the user's account information or credit card details are primary which is captured by the end innocent web users.

The proposed work is focused on classification of malicious URLs or phishing URLs using the data mining techniques. In this context the association rule mining technique is employed for identifying the phishing URL patterns. In order to prepare the data model the PhishTank Database is used for discovering the features from URLs. After concluding the feature set the URLs are transformed into the 2D vector. Using these vectors the data is converted into the item-sets and transactions. The transactions are used with the FP-Tree algorithm for generating

the association rules. The generated association rules are finally consumed for evaluating the patterns for discovering the phishing URLs in real world. That technique is suitable for classifying the URLs efficiently and accurately, therefore using this technique the anti-phishing tools can be prepared.

The implementation of the proposed enhanced association rule based phishing classification technique is implemented using JAVA technology. Additionally for preserving the obtained performance the MySQL server is used. Based on experiments the proposed technique and existing technique is compared using the different performance parameters. The summary of comparative parameters is given in table 6.

Table 6 Performance summary

S. No.	Parameters	FP-Tree algorithm	Apriori algorithm
1	Accuracy	90.3 – 97.8 %	81.7 – 90.8 %
2	Error rate	2.2 – 9.7 %	9.2 – 18.3 %
3	Memory usages	16628 – 18264 KB	18652 – 20173 KB
4	Time consumption	89 – 126 MS	119-146

According to the obtained performance summary as given in table 6.1 the proposed technique is efficient and accurate as compared to the Apriori algorithm based technique. Thus the proposed work is acceptable for utilizing the proposed algorithm for real world problem of phishing detection or classification.

4.2 Future Work

The main of the proposed work is to enhance the existing approach of phishing URL classification technique is successfully implemented. The experimental evaluation of the implemented approach demonstrates the superiority of the proposed system. In order to improve the given approach the following future extension of the work is proposed:

1. The existing approach utilizes the similar technique of association rule mining for classifying the phishing URLs, in near future the technique is replaced with the supervised learning approach for more accurate outcomes
2. The proposed technique usages the classical algorithms for classification in near future the existing technique is improved using the deep learning concepts.

5. REFERENCES

[1] S. Carolin Jeeva and Elijah Blessing Rajsingh, "Intelligent phishing url detection using association rule mining", Hum. Cent. Comput. Inf. Sci. (2016) 6:10, DOI 10.1186/s13673-016-0064-3

[2] Chapter 3: Data Mining: an Overview, available online at: http://shodhganga.inflibnet.ac.in/bitstream/10603/11075/7/07_chapter3.pdf

[3] Mohammed J. Zaki and Wagner MeiraJr, "Data Mining and Analysis Fundamental Concepts and Algorithms", Cambridge University Press Hardback, 2014 [Book]

[4] Michael Goebel and Le Gruenwald ―A Survey of Data Mining and Knowledge Discovery Software Tools‖, ACM, 1999

[5] Neelam adhabPadhy, Dr. Pragnyaban Mishra, "The Survey of Data Mining Applications and Feature Scope", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), PP. 43-58 Vol.2, No.3, June 2012.

[6] Sundaravaradan, Naren, Manish Marwah, Amip Shah, and Naren Ramakrishnan. "Data mining approaches for life cycle assessment." In Sustainable Systems and Technology (ISSST), 2011 IEEE International Symposium on, pp. 1-6. IEEE, 2011.

[7] Manoj and Jatinder Singh, "Applications of Data Mining for Intrusion Detection", International Journal of Educational Planning & Administration. Volume 1, Number 1 (2011), pp. 37-42

[8] M. Rajalakshmi, M. Sakthi, "Max-Miner Algorithm Using Knowledge Discovery Process in Data Mining", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 3, Issue 11, November 2015

[9] Smriti Srivastava & Anchal Garg, "Data Mining For Credit Card Risk Analysis: A Review", International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR), Vol. 3, Issue 2, Jun 2013, 193-200

[10] Dipti Verma and Rakesh Nashine, "Data Mining: Next Generation Challenges and Future Directions", International Journal of Modeling and Optimization, Vol. 2, No. 5, October 2012

[11] Gaurav Varshney, Manoj Misra and Pradeep K. Atrey, "A survey and classification of web phishing detection schemes", SECURITY AND COMMUNICATION NETWORKS, Security Comm. Networks 2016; 9:6266–6284, Copyright © 2016 John Wiley & Sons, Ltd

[12] Hassan Y. A. Abutaira, Abdelfettah Belghitha, "Using Case-Based Reasoning for Phishing Detection", Procedia Computer Science 109C (2017) 281–288, 2017 The Authors Published by Elsevier B.V.

[13] Rakesh Verma, Avisha Das, "What's in a URL: Fast Feature Extraction and Malicious URL Detection", IWSPA '17, March 24-24 2017, Scottsdale, AZ, USA, c 2017 Copyright held by the owner/author(s). ACM ISBN 978-1-4503-4909-3/17/03.