International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 2 Issue: 8 2127 – 2132

Framework for Phishing Detection in Email under Heave Using Conceptual Similarity

Sowndarya Karri, SSSN.Usha Devi N
Computer science Engineering, UCEK, Kakinada
Andhra Pradesh, India
e-mail: sowndaryakarri@gmail.com, usha.jntuk@gmail.com

***Abstract*--**Today everything is available in online. Every day so many users start their online transactions. The main reason behind this is number of alternatives and best deals are available there. They can choose according to their taste with cost effective manner. This is one side of a coin. The other side fully dealt with security problems and frauds in the online transactions. Among most of the online transactions email is the shortcut and flexible for both communication as well as for attack. So this paper mainly focuses on detection of phishing attacks and categorizes the emails based on specified and critical properties which give more information about the source of the phishing. In general most of the existing systems focus on email classification based on header part or body part. Most of the filters available today focus mainly on mail headers only. Sometimes this is not enough to detect the fraud. Some more studies focus on body part also. But they follow document clustering with term intensive similarity. First, to identify advanced phishing attacks blind term intensive similarity is not sufficient. Second, emails system is like online stream. So the nature of the phishing behavior may change time to time. In that case online learning is also required to handle concept drifts. This paper focuses on conceptual similarity along with term intensive similarity. We introduced a novel procedure named as "Framework for Phishing detection in email under heave using conceptual similarity" to adaptively classify the emails. Simulation results shows that our proposed approach effectively detect and isolate the emails with phishing attack by comparing underlying concept.

Keywords: *E-mail, Phishing, Term-similarity, Conceptual similarity, Clustering, Online learning*

*****

I. INTRODUCTION

Phishing is the social engineering attack of defrauding an online account holder of financial information by impersonating as a legitimate company in await of attempting to acquire information such as usernames, passwords, credit card numbers and sometimes money. Nowadays, Phishers use many techniques to lure millions of victims each year. One of the advanced techniques of phishing is spear phishing, a targeted phishing attack in which attackers obtain details about victims through phishing techniques to make themselves seem more trustworthy so that they would have a much higher chance of spreading the malware through the company's system. To conflict the threat of phishing attacks, researchers have investigated reason behind why people are falling for phishing by resolving which troops are more vulnerable to phishing. By using these analysis they determined how best to focus on anti phishing discipline.

Phishing is best understood as one of a number of distinct methods that identity thieves use to "steal" information through deception – that is, by enticing unwitting consumers to give out their identifying or financial information either unknowingly or under false pretences, or by deceiving them into allowing criminals unauthorized access to their computers and personal data. The United States and some other countries use the term "identity theft," and the United Kingdom often uses the term "identity fraud," to refer broadly to the practice of obtaining and misusing other's identifying information for criminal purposes. Identity fraud also can be used to refer to the subsequent criminal use of others' identifying information to obtain goods or services, or to the use of fictitious identifying information (not necessarily associated with a real living person) to commit a crime. Phishing is committed so that the criminal may obtain sensitive and valuable information about a consumer, usually with the goal of fraudulently obtaining access to the consumer's bank or other financial accounts. Often "phishers" will sell credit card or account numbers to other criminals, turning a very high profit for a relatively small technological investment.

II. RELATED WORK

Email phishing, in which someone tries to trick you into revealing personal, financial and sensitive information like social security numbers, credit card numbers and account passwords by sending fake emails that look legitimate [1], remains one of the biggest online threats that snare millions of victims each year based on details of attacks and statistics on the enormous growth in the number of attacks since the phenomenon first emerged [2]. Another new variation is called **Vishing**, which involves voice communication. Email may or may not be involved. As computer users have become more educated about the dangers of Phishing emails and have learned to avoid them, perpetrators have begun incorporating the telephone into their schemes. The latest statistics reveal that banks and financial institutions along with the social media and gaming sites continue to be the main focus of phishers. Some loyalty programs are also becoming popular among phishers because with them phishers can not only breech the financial information of victim but also use existing reward points as currency. U.S. remains the largest host of phishing, accounting for 43% of phishing sites reported in January 2012. Next was Germany at 6%, followed by Australia, Spain, Brazil, Canada, the U.K., France, Netherlands, and Russia [3]. A study of demographic factors suggests that women are more susceptible to phishing than men and users between the ages of 18 and 25 are more susceptible to phishing than other age groups [4].

The advancement in this phishing attack is the "**Spear Phishing**" attack which uses phishing emails targeted at a specific company [5]. Recent studies of anti-phishing work group concluded in part that the presence of personal information does not significantly affect success rate of phishing attacks, which suggests the most people do not pay attention to such details [6].

As phishers have increasingly utilized various formats such as email title and sender's addresses to develop defraud plots, email recipients should be alert about these formats in evaluating incoming messages [7]. The improved corporate strategies against phishing attacks particularly spear phishing. Baker et al. suggested that most modern organizations lack effective responses against phishing attacks [8].

There is a great need for research that investigates the various ways that are required for anti-phishing. Most of the recent anti-phishing study focuses on the question: How do individuals process a phishing email and form their tendency to respond to the email? It explore how users attention to visual triggers and phishing deception indicators influence their decision making process and consequently their decision outcomes.

Visceral triggers (for example, stressing urgency of response) are motivational manipulations that scammers use to reduce the depth at which people process information, allowing decision errors to occur [9]. Phishing deception indicators (for example, poor grammar, spelling mistakes, sender address spoofing) are cues that reveal the inconsistency between the deceptive event and personal past experiences, and they help reveal the deception nature of email. So the recent research model [10] attempts to capture the interplays between phishing design features (such as visceral triggers and phishing deception indicators) and individual characteristics (such as knowledge of email-based scams), and investigates their impacts on email recipients.

III. PROPOSED APPROACH

Background:

Machine Learning:

Machine learning is a process of giving knowledge to the proposed learner. For that purpose initially proposed system is trained with some existing email corpus with heterogeneous content. A novel unsupervised machine learning algorithm named as "*Framework* for *Phishing detection in email under heave using conceptual similarity"* is used by the proposed system for effective handling of concept drift email streams. Before enter into details of algorithm we need to finish some pre-processing steps described as follows.

A. Email pre-processing: It is having following stages:

a) Email Header processing

Generally email header contains following information

1. From: It specifies source of the email sender. But it doesn't specify the actually source because it can be easily counterfeit and unreliable.
2. Subject: This is like title or abstract contains objective of the mail body and for which it is intended to.
3. Date: Specifies the composition date and time of email.
4. To: Destination mail id to which the message was addressed.
5. Return-Path: Similar to Reply-To and specifies return mail address.
6. Envelope-To: Describes address of mailbox to specified email id in "To"
7. Delivery Date: Specifies date and time of delivery of email to intended client or service.
8. Received: It specifies the stack trace address mechanism. i.e. it is the mail received path in hop by hop fashion. The oldest or first address is placed in the last part of the path and final address received is place in the first part of the path. It is most useful part of email header for email forensics.
9. Message-id: This is a non-repeatable string or unique id assigned to a new message at the time of creation by the mail system. But this is unreliable due possibility of tampering.
10. Content-Type: Specifies the MIME type such as html, plain text or image etc.
11. Content-Length: Size of the mail message
12. X-Spam-Status: Specifies probability of the current mail as spam.
13. Message Body: Actual content prepared by the sender in the email.

Among these fields proposed system utilizes Subject, Date, To, Delivery Date, Message Id, Received, Content type, Content length and Message Body fields for email statistics and analysis.

Initially Historical or training email corpus is loaded and extract the above mentioned fields from each email. Then these fields are store as a transaction table in the database for offline or online querying. For this transaction table a view is created with specified group by statements. This view is used to get frequency of particular path and content. So it is more helpful to filter the mails in basic level. Based on this statistics initial decision is made and mark some of the mail paths or ids are susceptible. Later body processing stage is initiated.

B. Body processing

In this stage body is identified based on the content type. It may be html, plain or other MIME type. Based on the content type process is also divided. It is having two stages of work again.

1. HTML Processing: In this body text is scanned for <A> and <SCRIPT> tags. Because, in most of the cases hyperlinks and script elements causes to phishing and virus injection. Every hyper link is basically scanned to identify the source of that link. This can be achieved through processing of host part of link with "trace route" like tools. The results of the "trace route" are compared with "phish tank" database. If any one of the address in the trace route belongs to phish tank database, proposed system removes that hyper link. This phish tank database is publicly available.

 ISSN: 2321-8169
2127 – 2132

2. Text processing of email body includes number of stages. This is heart of the conceptual document mining. It includes some natural language processing steps. They are described in following modules. Body is identified with <P> tag as well as <TD>tag. After that all the individual <P> [<TD>] plain text parts are merged into single document for each mail and placed in a mail body store locally. Here we assume that this local store is maintained in the firewall system.

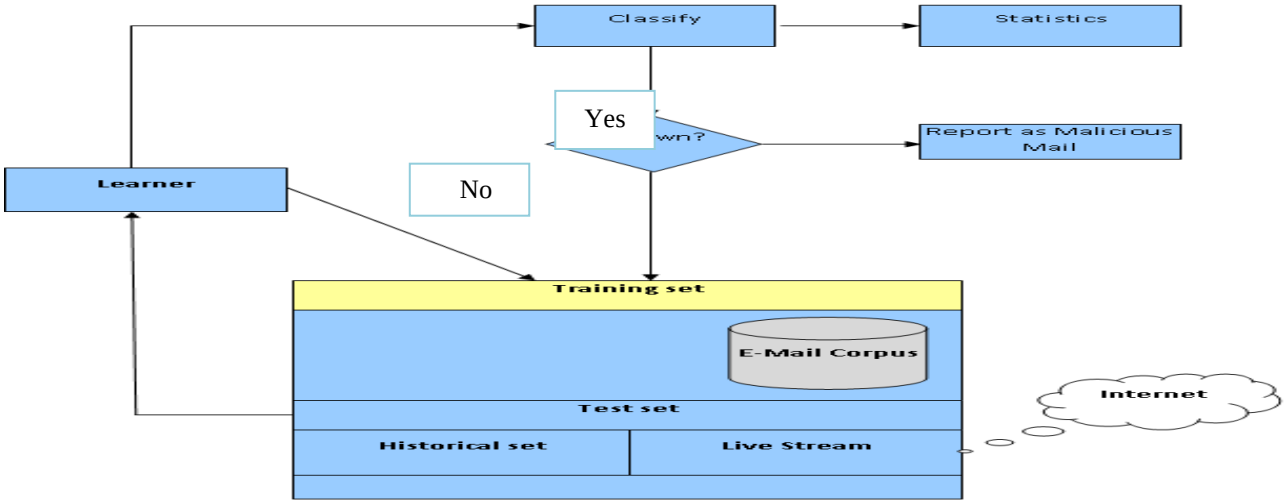


Figure 1: Proposed Architecture

After pre-processing is done documents are clustered as follows. The base of this algorithm is acquired from [11] and modified according to the demand of context presented in this paper.

Documents in training e-mail corpus are equal sized document vectors. So they can be directly compared with each other with familiar cosine similarity measure.
Along with cosine similarity, conceptual similarity also simultaneously compared as mentioned above.

1. First document itself is treated as Leader and form a new and first cluster.
2. From second document onwards, each document is compared with existing clusters with given similarity thresholds. Here similarity means both conceptual and cosine similarity. Each similarity value is compared with respective similarity thresholds.
3. If both of the similarities are above the given threshold then that document is placed in the current cluster.
4. Otherwise current document will form a new cluster and announce itself as leader to that cluster.
5. Same process is repeated for all the documents placed in the Training email corpus.

This is actual but slightly modified version of actual Leader follower algorithm. But it has some limitation. Let us consider some n^{th} document. That document doesn't have knowledge about the leaders created after its creation. So it may assign to a leader which is created before. But if any leader which is also having similarity greater than given threshold but nearest to the n^{th} document when compared to the current leader, then existing algorithm doesn't give solution to this. So in this paper we extend the algorithm to meet this requirement i.e. except leader documents remaining all documents are compared with those leader which are formed after their creation. If any leader or cluster is more nearer when compared to current cluster then this document will be removed from the current cluster and assigned to new cluster and update their leader. Otherwise no change has been made.

Here also there is limitation to handle concept drift that will be occurred in the mail stream processing. Existing mail corpus is static and it is fully offline process and all the documents are converted into equal sized document vectors. This will not raise any problem due to offline and more over static dataset. But this cannot meet the online as well as dynamic document vector size demand. So this can be achieved through our novel "*Framework for Phishing detection in email under heave using conceptual similarity*".

It includes following sub modules.

C. Mail Stream Processor Implementation

1. In this module proposed system asked the user to submit his email credentials to process the mails in inbox.
2. After successful login each unread mail is extracted by Mail stream processor by default. If user change the setting to "all' then both read and unread mails are extracted.
3. Email header part is processed first and compared with the existing training database. If it is matched with any existing mail header information then current mail is marked according to the existing header. Otherwise new entry will be submitted to the database.
4. Next Mail body is processed. Here all the text processing steps are executed as mentioned in the previous sections. But documents which are to be clustered may have varied length those cannot

directly compared with the existing clustered documents.

5. So mail body document vector is sub divided or extended based on the actual cluster document vector size. This sub document vectors are individually processed and compared with existing leaders.
6. In this case sub document vectors may be assigned to different clusters. To clearly specify the cluster to which main document is belongs to is evaluated based on the weight calculated at each cluster for sub document vectors. If overall probability is less than given threshold then this document itself form new cluster with those sub document vectors which are not fit to any other cluster.

D. Data Post Processing

Finally with aggregating all of the results retrieved from mail header statistics, body processing and attachment processing susceptible mails will be finalized. Now to label the clusters with proper type, phishing and spam word database has to be used. Every cluster is scanned for these words and calculates the probability of these words in each mail body document. Which of the clusters are having probability of malicious documents > given malicious threshold they are labeled with Spam or Phish or Malware accordingly and rest of the clusters are marked as normal. All these clusters are again useful of training for next generation or upcoming mails. This is called as online learning.

E. Mail Body Document Clustering

Let a document **d** with m features w1, w2, . . . ,wm be represented as an m-dimensional vector, i.e., **d** = < d1, d2, . . . , dm>. If wi, $1 \leq i \leq m$, is absent in the document, then di = 0.Otherwise, di > 0. The following properties, among other ones, are preferable for a similarity measure between two documents:

1) The presence or absence of a feature is more essential than the difference between the two values associated with a present feature. Consider two features wi and wj and two documents **d**1 and **d**2. Suppose wi does not appear in **d**1but it appears in **d**2. Then wi is considered to have no relationship with **d**1 while it has some relationship with **d**2.In this case, **d**1 and **d**2 are dissimilar in terms of wi. If wj appears in both **d**1 and **d**2. Then wj has some relationship with **d**1 and **d**2 simultaneously. In this case, **d**1 and **d**2 are similar to some degree in terms of wj. For the above two cases, it is reasonable to say that wi carries more weight than wj in determining the similarity degree between **d**1and **d**2. For example, assume that wi is absent in **d**1, i.e., d1i = 0, but appears in **d**2, e.g., d2i = 2, and wj appears both in **d**1 and **d**2, e.g., d_{1j}= 3 and d_{2j} = 5. Then wi is considered to be more essential than wj in determining the similarity between **d**1 and **d**2, although the differences of the feature values in both cases are the same.

2) The similarity degree should increase when the difference between two non-zero values of a specific feature decreases. For example, the similarity involved with d13 = 2 andd23 = 20 should be smaller than that involved with d13 = 2and d23 = 3.

3) The similarity degree should decrease when the number of presence-absence features increases. For a presence/absence feature of **d**1 and **d**2, **d**1 and **d**2 are dissimilar in terms of this feature as commented earlier. Therefore, as the number of presence-absence features increases, the dissimilarity between **d**1 and **d**2 increases and thus the similarity decreases. For example, the similarity between the documents < 1, 0, 1> and < 1, 1, 0 > should be smaller than that between the documents < 1, 0, 1 > and< 1, 0, 0 >.

4) Two documents are least similar to each other if none of the features have non-zero values in both documents. Let **d**1 =< d11, d12, . . . , d1m > and **d**2 = < d21, d22, . . . , d2m >. If $d_{1i}d_{2i} = 0$, $d_{1i} + d_{2i} > 0$ for $1 <= i <= m$, then d1 and d2 are least similar to each other. As mentioned earlier, d1 and d2 are dissimilar in terms of a presence-absence feature. Since all the features are presence-absence features, the dissimilarity reaches the extremity in this case. For example, the two documents <x, 0, y > and < 0, z, 0 >, with x, y, and z being non-zero numbers, are least similar to each other.

5) The similarity measure should be symmetric. That is, the similarity degree between d1 and d2 should be the same as that between d2 and d1.

6) The value distribution of a feature is considered, i.e., the standard deviation of the feature is taken into account, for its contribution to the similarity between two documents. A feature with a larger spread offers more contribution to the similarity between d1 and d2.

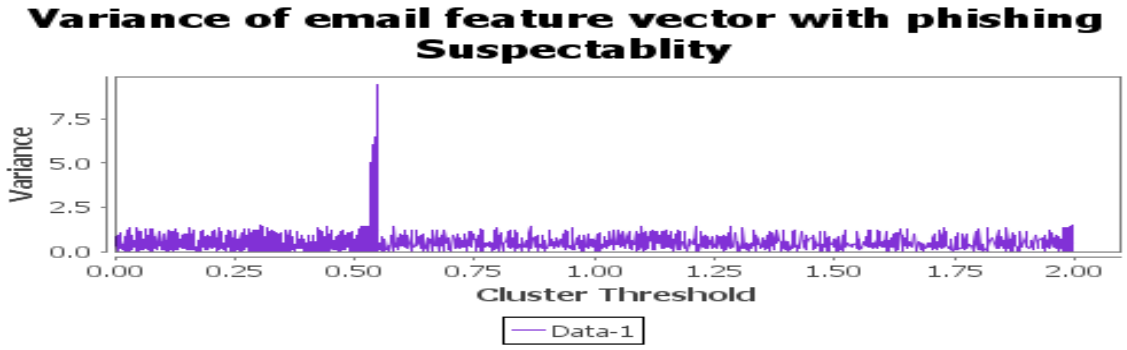
F. Validating cluster quality using Davies-Bouldin Index

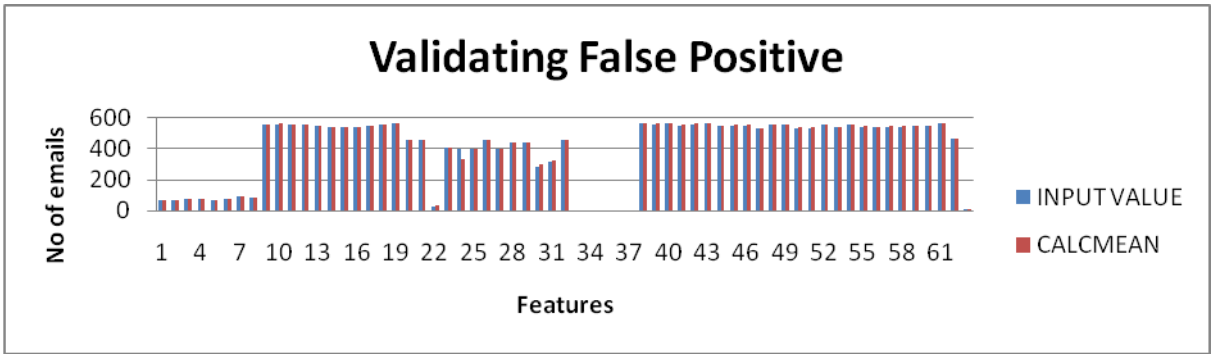
Davies-Bouldin index [12] is a function which gives the ratio of sum of within cluster scatter to between cluster separations. The application of Davies-Bouldin index can be explained in two phases. Initially we have to find the separation within the objects of the cluster, usually this is done by measuring the distance between the centroid of the cluster and the objects within the cluster, but to obtain better results this distance measure should be similar to one that is used in the algorithm and hence we use the average of similarity within the cluster. In the next step we need to find the similarity between the clusters that in turn determines the separation between the clusters. The two steps are averaged over the number of clusters. By this explanation it is obvious that the lower value returned by the Davies Bouldin index indicates the better clustering because in the value of the ratio between the cluster scatter to the cluster separation, the initial step which returns dispersion of objects among the cluster should be low when compared to the separation between the clusters. But if the value of index is 0 it indicates that no clusters are formed and hence we cannot consider 0 as a better cluster quality index.

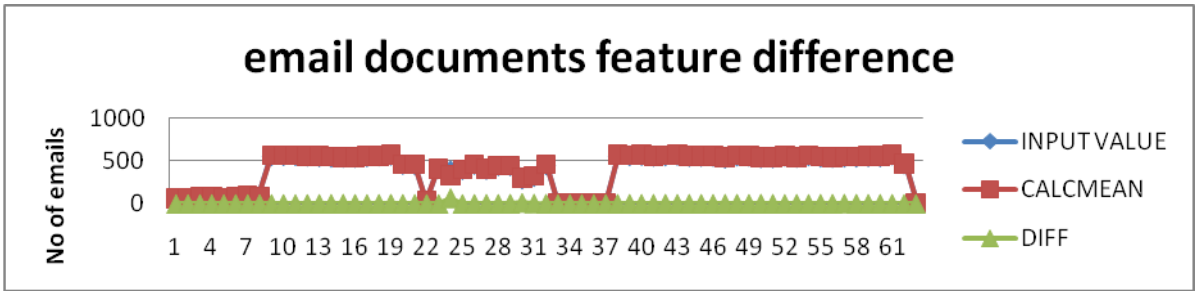
 ISSN: 2321-8169
2127 – 2132

IV. SIMULATION RESULTS

Basic Clustering			Term wise Clustering			Conceptual Similarity wise Clustering		
Threshold Value	Number of Clusters	Davies-Bouldin Index	Threshold Value	Number of Clusters	Davies-Bouldin Index	Sub-Threshold Value	Number of Clusters	Davies-Bouldin Index
0.1	42	1.832	0.1	43	0.588	0.05	43	0.044
0.2	5	0.59	0.2	42	0.574	0.1	42	0.059
0.3	1	0	0.3	41	0.561	0.2	41	0.072
0.4	1	0	0.4	39	0.534	0.3	39	0.1
0.5	1	0	0.5	38	0.521	0.4	38	0.115
0.6	1	0	0.6	38	0.521	0.5	38	0.115







V. CONCLUSION

This project focuses on similarity of email messages along with mail header to detect and separate normal mails from malicious mails. For that purpose a novel "*Framework* for *Phishing detection in email under heave using conceptual similarity"* is introduced to handle concept drifts, mail similarity, dynamic building of cluster.

VI. FUTURE WORK

Here two more things have to be considered. Firstly, new document may have new set of terms those are not placed in the Hash map. So update the dictionary i.e. Hash map with new terms with their respective synset ids. This is used in conceptual similarity. Secondly existing similarity thresholds may not fit in case of concept drift occurs. To overcome these limitations Evolutionary computing is introduced. It is used to find the best fit threshold for given context. According to this method initial population is given and this process will be terminated based on the fitness function value. Here fitness function is cluster validity measure which is used to find the quality of the clusters as well as clustering scheme for given similarity thresholds.

References

[1] Markus Jakobsson and Steven Myers. Phishing and countermeasures: understanding the increasing problem of electronic identity theft. John Wiley & Sons, Inc., 2007.

[2] Jagatic, Tom; Nathaniel Johnson, Markus Jakobsson, Filippo Menczer (October 2007). "Social Phishing".

[3] PhishTank. Phishtank stats-Jan 2012. http://www.phishtank.com/stats/2012/01/?y=2012&m=01

[4] Steve Sheng, Mandy Holbrook, Ponnurangam Kumaraguru, and Lorrie Cranorand Julie Downs1. Who falls for phish? a demographic analysis of phishing susceptibility and effectiveness of interventions. In 28th international conference on Human factors in computing systems, Apr 2010.

[5] Bank, David (August 17, 2005). "Spear Phishing Tests Educate People About Online Scams".

[6] Markus Jakobsson and Jacob Ratkiewicz. "Designing Ethical Phishing Experiments", 2006.

[7] R. B. Horowitz and M. G. Barchilon, "stylistic guidelines for e-mail." IEEE trans. Prof. Commun, vol.37, no. 4, pp. 207-212, Dec. 1994.

[8] E. M. Baker, W. H. Baker, and J. C. Tedesco, "Organizations respond to phishing: Exploring the public relations tackle box," commun. Res. Rep., vol. 24, pp. 327-339, 2007.

[9] J. Langenderfer and T. A. Shimp, "Consumer vulnerability to scams, swindles, and fraud: A new theory of visceral influences on persuasion," psychol. Market., vol.18, pp. 763-783, 2001.

[10] Jingguo wang, Tejaswini Herath, Rui Chen, Arun viswanath and H. Raghav Rao, "Phishing susceptibility: An investigation into the processing of a targeted spear phishing email", IEEE trans. Vol 55, No.4, December 2012.

[11] P. A. Vijaya Department of Computer Science and Automation, Intelligent Systems Lab, Indian Institute of Science, Bangalore 560 012, India,M. NarasimhaMurty Department of Computer Science and Automation, Intelligent Systems Lab, Indian Institute of Science, Bangalore 560 012, India,D. K. Subramanian Department of Computer Science and Automation, Intelligent Systems Lab, Indian Institute of Science, Bangalore 560 012, India

[12] Davies, David L.; Bouldin, Donald W. (1979). "A Cluster Separation Measure". IEEE Transactions on Pattern Analysis and Machine Intelligence. PAMI-1 (2): 224–227. doi:10.1109/TPAMI.1979.4766909