International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 7, Issue 4, April 2017)

A Review on Spam Detection Techniques over Internet

Meenakshi[1], Dr. Neetu Sharma[2]

[1]M. Tech Scholar, [2]A.P., CSE Dept, Ganga Institute of Technology & Management, Kablana, Jhajjar, Haryana, Maharashi Dayanand University, Rohtak, Haryana, India

Abstract: **- Email access client 'hope' spam to be a figure or data within the body of the email and not an attachment. Excel is another extremely common file-type in use and users are very familiar with this format. Since many businesses use Microsoft Excel for spread sheets, Spam bases and so on, email users will have to check each document otherwise they risk losing important documentation. With most anti-spam software products on the market geared towards filtering the email itself and not attachments, 'Excel' spam has a longer shelf-life within a network. So we provide a system which detect the excel file and any text file. In this paper we study existing technique of spam detection and carry out problem statement.**

Keyword: **Spam, E-mail, Detection technique**

I. INTRODUCTION

Spam detection problem is becoming more serious now days. It consumes more than half bandwidth of mailboxes. Spam frustrates, confuse and annoy email users by wasting valuable resources and time. Spam even provides ways for phishing attacks and distributing harmful content such as viruses, Trojan horses, worms and other malicious code. Without a spam filter, one email user might receive over hundreds of mails daily and find that most of them are of spam category. The spam mails are with no use of email users. Due to this, serious attention has given to this issue in mailboxes. Several technical solutions like commercial and open-source products have been used to alleviate the effect of this issue.

II. QUERY BASED ANTI-SPAM TECHNIQUE

A query based cross layer approach to detect spam follow some steps:

- *Analyze the mail content:* This approach analyze the mail content and sender mail address of the mail, then cross analyze and compare the content and sender address of the previous spam mails. If content and sender address are already present then it declares that mail as a "spam".
- *Trusted Knowledge Base:* In trusted knowledge base, database of trusted sender is stored over the inbox based on the frequency of the communication of mails. If sender is not the trusted sender then following steps are needed to execute to identify the spam mails.
- *Keywords knowledge base:* This stores the spam keywords. When any mail is received by system, this approach analyzes the keywords of mails with keywords knowledge base of spam. Then spam is declared as spam or useful mail on the basis of result.
- *Sender mail address:* It verifies mail address of sender using mail header.
- *Sender Location:* This approach finds the location of mail server and compares the location with spam mails location.
- *Misbehavior of incoming mail:* Artificial Neural Network is used to predict any misbehavior of incoming mails.
- *Cross Validation:* In this step, system will verify the sender that sender is a genuine human user or machine generated user using some cross request.
- This approach needs a large amount of memory and much hardware for execution, so workload increases.

III. LITRETURE SURVEY

DBN uses a greedy layer-wise unsupervised algorithm to initialize the weight of deep neural network based on use of RBM [1]. Results of DBN are much better than SVM but technique for selecting number of hidden layers is still required.

Other techniques are:

A rule-based approach has developed for spam detection. It uses training and testing phases of data [2]. This approach improves efficiency of spam filtering as compared to previously proposed techniques but time complexity is higher due to rules generation and their execution. So digests were used in this approach to detect spam mails. A social network has to construct based on email exchanges between various users. Spammers are identified by observing abnormalities in the structural properties of the network.

Another novel approach has been proposed which creates a Bayes network out of email exchanges to detect spam. Bayesian classifiers scan the contents of the email to calculate probability distributions for every node in the network.

Other content-based filters provide some temporary relief from spam. But these filters are not robust enough against spammers. Spammers can easily fool these filters. So content-independent filter is needed [3]. A distributed, content independent, spam classification system called trinity has been proposed. Trinity is based on the following observation: Bots send a large number of e-mails in a short amount of time. If an e-mail is received from an unknown source that has sent many e-mails in a short period of time, then the likelihood of this being spam is high. But in this approach, right tradeoff between the security and weight of the protocols is needed. Trinity must have large number of peers to become effective.

PCA has also been used for spam detection. SpamNET has been introduced for effective spam detection which uses heuristic rules, PCA and neural networks [5]. This program is able to adapt itself according to environment in which various users send mails. This program retrains itself after every seven days. SpamNET has used Bayesian classifier as well as neural networks so processing power is high.

Various types of neural networks have been used to detect spam [9]. Neural Networks are able to detect features which can be detected by human. They state complex relationships between input and output. They make system adaptable so that system can adjust it according to changing environment.

A huge number of techniques and solutions have proposed to detect spam but every technique has some pitfalls.

IV. Problem Formulation

PCA has been used with neural networks for spam detection. PCA is dimensionality reduction technique which reduces inputs which are to be fed into neural networks. As a consequence, neural network is able to detect spam efficiently. But PCA takes eigenvectors that are highly correlated to each other. Some other problems may be faced during use of PCA:

- False positive rate is very sensitive to differences in features of mails.
- Effectiveness of PCA is sensitive to level of aggregation of traffic of incoming mails.
- If a strong virus may inadvertently pollute the processing of PCA
- PCA contains linear combinations of variable so there must be high-correlation between variables.

To remove inefficiencies created by PCA, FastICA can be used. It works in non-gaussian environment.

V. Proposed Method

As single technique is not sufficient to combat this issue so multiple methods should be used. In this research, FastICA will be used for spam detection with neural networks. FastICA is signal processing technique used for analysis of several types of data and feature extraction. It combats the problem faced during using PCA. False positive rate is very sensitive to small differences in the number of principal components in normal subspace in case of PCA. So FastICA can be used to detect spam even those which are independent of each other and have nothing in common.

VI. Conclusion

Spam emails are the biggest problem for the web data. This work explored optimal approach to deal with this problem. This approach performs well only on words of text and excel file. Spam filter system is implemented using Matlab platform. Based on results of neural classifier seem to perform better in classification. This system can be further enhanced by including the functionality of reading Web Pages so that web browsing can be made a spam-free experience by avoiding advertisements and unwanted malicious websites.

REFERENCES

[1] Grigorios Tzortzis and Aristidis Likas, "Deep Belief Networks for spam filtering", 19th IEEE International Conference on Tools with Artificial Intelligence, GR 45110, Ioannina Greece (2007)

[2] Gaurav Kumar Tak and Shashikala Tapaswi, "Query Based approach towards spam attacks using artificial neural network", International Journal of Artificial Intelligence & Applications, October 2010

[3] Alex Brodsky (Canada) and Dmitry Brodsky (USA), "A distributed content independent method for spam detection".

[4] A.Hyvarienen and E.Oja, Independent Component Analysis and Applications, Neural Networks 13(4-5):411-430, 2000

[5] Abhimanyu Lad, SpamNET Spam Detection using PCA and Neural Network

[6] http://www.cis.legacy.ics.tkk.fi/apo/papers/IJCNN99_tutorial web/node 32.html

[7] Dominic Langlois, Sylvain chartier and Dominique Gosselin, An introduction to Independent Component Analysis: Infomax and FastICA Algorithm (2010)

[8] Sasmita Kumari Behra (2009) "FastICA for blind source separation and its implementation", Rourkela

[9] Martin, Spam Filtering using Neural Networks, http://www.web.umr.edu/~bmartin/378Project/report.html.