© INNOVATIVE RESEARCH THOUGHTS | REFEREED | PEER REVIEWED
ISSN : 2458 – 308X | Volume : 03 , Issue : 01 | January - March 2017

Spam Detection Using FICANEURO Approach

Vikram Singh

***Abstract*—Spam detection is required to deal with the harmful effect of spam mail on user directly or indirectly. The directly effect can be in term of time, storage space and network bandwidth and indirectly effect can be defined in term of privacy and security. Several technical solutions like commercial and open source product have been used to alleviate the effect of this issue. We use the FICANEURO approach for the spam detection in this paper. It is combination of Advanced featured of Individual Component Analysis and Neural Network. The approach is mainly based on content based filtering. The result of this approach enhances the accuracy with increase in file size.**

Keywords-Spam, Neural Network, Individual Component Analysis, Content based Filtering, IP and Header, Principal Component Analysis

I. INTRODUCTION

Spam is defined as an unwanted of electronic message posted blindly to many numbers of recipients. The unwanted mail which is neither requested nor subscribed would consider as spam mail. It consumes more than half bandwidth of mailboxes, spam frustrates, confuse and annoy email users by wasting valuable resource and time. Spam even provides ways for phishing attacks and distributing harmful content i.e. viruses, Trojan Horses, worms and other malicious code.

At present majority of internet user received Emails that try to sell product of company that is no baneficial for you. Some of email even consists of the spyware,virus, malware and promo. These cause wastage of both time and money.Without a spam filter, one email might receive over hundreds of mails daily and find that most of them are of spam category. Spam Filtering can be of two types:

A. Non Machine Learning Based

The Non-Machine Learning Based technique consists of White list, blacklist and set of keywords like “you have won”. As these methods are dependent on list so these can be easily resolved by spammers. These methods also require manual update and sometime these methods misclassify legitimate mail as spam mail which is more dangerous than no filtering.

B. Machine Learning Based

The machine learning techniques first analysis the message content and then perform classification of mail as spam or not spam. Various machine learning anti-spam methods are Support Vector Machine, Memory Based Learning; Boosting Decision based learning, Fully etc.

Various Techniques have been developed to combat the problem of spam but still effective and efficient technique is required which will have very low false positive and false negative.

II. STRUCTURE OF EMAIL

An electronic mail structure made with four vital components: the envelope, the body of message and the header and IP

A. Body

The main part of the mail that includes details of conversation..

B. Header

This part contains routing information, including sender and recipient, date and subject. Email address has two parts separated by ‘@’.Username on the left side and domain name for the host server at right side.

C. Envelope

This contains the rourting information and hidden from the user..

D. IP Address

It contains the sender Internet Protocol address.

III LITERATURE REVIEW

Due to many adverse effects of spam there is a need of approach that can classify mail as a SPAM or Not Spam. This section provides the review of paper, which deals with such anti spam technique

A. *Access Filtering*

It verifies and authenticates header information of an email

B. *Economic Filtering*

Two main categories of economic solutions are computing-time-based systems and money-based systems. Computing-time-based systems stimulate spammer to spend considerable computing resources to send a single spam message. Money-based systems charge a small amount of money from every email sent.

C.*Content-based Filtering*

The systems that implement content-based filtering perform filtering when message is fully received. These systems can use rule-based filtering, Naive Bayesian

D.Characteristics based Filtering
It finds out distinct characteristics between good emails and spam.

Abhimanyu Lad [5] describe a spam detection user level program called spamnet which use the heuristic rules, artificial neural network as classifier and most important principal component analysis to detect spam by analysis mail content. The detection process is automated which retrains its system every 7 days so that it adapt the changes as new mails pattern changes. The process first pre-process the message through extractor module whose work is vital in whole process. Its main functionality is to remove of stop words and parses the message according to rules. This phase is also responsible in providing input to PCA which are the outputs of extractor. The input to PCA is feature vector which are created from words with the help of volatile vocabulary. The main role of PCA is to transform feature vector to optimal representation which has done by computing eigenvector and improve the performance and efficiency by reducing the feature set. And in final neural network with 6 inputs classify the mails through its output signal spam or non-spam.

Alex Brodsky et al. [3] proposed distributed, content independent, spam classification system called TRINITY that is specifically aimed at botnet generated spam and can be used in combination with existing spam classifiers [4]. They mainly focus on method which protect from the botnets attack. The computers which are connected to internet are first hacked and from their system IP address which are assigned dynamically at boot time, a numerous junk emails send to users. Trinity is a distributed spam detection system whose main objective along with detection process is to keep some of points in consideration while developing this approach are that it should be easily installable ,pluggable, within existing infrastructure. No need to modify the existing protocols. With the use of distributed system, central point of failure clause removed and most vital component it would be user controllable. The main goal of this approach is to identify the source an email which sends the spam and immediately update the distributed database about the sender information so that whole system disable the breach.

Dominic Langlois et al. [7] presents two Independent component analysis (ICA) algorithms which are Infomax and FastICA algorithms. With the implementation of ICA algorithm the results were also compared between them and also along with principal component analysis. By taking ‘cock tail party effect’ the result obtained by PCA and ICA are compared which state that for gaussianity ICA is use to represents a linear combination of the original variables or non-gaussian data so that the components would be statistically independent. The main goal of this paper is to find the independent component. ICA is a technique which used for source signal extraction but there implies some condition which has to follow for effective use. To accomplish this task mixture of three images has taken as example and the algorithm which extracts components from the mixture image would prove is more efficient than other.

A. Nosseir et al. [10] present approach which is based on character-word based technique and for classifying multiple neural networks used. The content based filtering technique check the words which consist 3, 4 and 5 character. After stop-word and noise word removal, the stemming process changes the plural form of nouns to root word. The list of words then classified according to the length of and categorizes them as good and bad words. Then for each word according to different length ,neural network has been trained which then classify which words in message are good and bad by giving output 0 and 1 respectively. Through this approach it has been proved that in comparison with the white list and black list this approach according to word length filtering perform better and along with it trains it sub network for improved performance.

IV. PROBLEM FORMULATION

Spams are the textual context of the system which can damage our system. Our basic problem is to protect our system from such unwanted files. To save our system form such kind of failures we need to design a system which can recognize the spams and can let you know on the basis of a training system.

A. Previous Work
Principal Component Analysis 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:

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

V. CONCLUSION AND FUTURE SCOPE

After completing our research work we conclude that the basic purpose of our research work is to detect spam by using Fast ICA algorithm with neural network as it provides the better solution regarding spam detection than the previous approach in term of accuracy and computational complexity. Our Research work shows that the results are better as the ICA is based on the signal processing basically. Now it is easy to compare the signal to another to match the pattern. The drawback of this system is that it has no rule set for processing.

In future, if somebody combines Fuzzy Logic with Neural Network this drawback can be removed and the result would be efficient. If somebody wants to experiment, Neural Network classification can be replaced with Bayesian Classification for better result.

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

[10]. Ann Nosseir , Khaled Nagati and Islam Taj-Eddin," Intelligent Word-Based Spam Filter Detection Using Multi-Neural Networks",IJCSI, Vol. 10, Issue 2, No 1, March 2013.