International Journal on Applications in Engineering and Technology
Volume 8 : Issue 4 : November 2022, pp 6 – 8 www.aetsjournal.com ISSN (Online) : 2455 - 0523

An Analysis of Spam Mail Using K-Means Clustering Algorithm

Mr K.Nagarajan, Mr. V. Vinod Kumar , Mr. M. Mohammedkasim , Mr. T Prabu

Abstract— Email is a method of exchanging digital messages between people using digital devices such as computers, tablets and mobile phones. Email spam also known as junk email is unsolicited bulk messages send through email. The use of spam has been growing in popularity since 1990s and is a problem faced by most email users. Recipients of spam often have had their email addresses obtained by spambots which are automated programs that crawl the internet looking for email addresses. Origin blacklisting is used to detect and filter these kinds of emails. The sources of emails are provided by origin detection. In this origin identity, the spam blocking system finds the source and if any source matches with the user identity, the spammer is blocked to send email. But the system cannot check the contents of spam. The Propose system verifies the mail contents. If the spam content matches with the spam database cluster, then it blocked while sending to receiver. In this system blocks the mail and prevents the flow of data in a network. The porter stemmer and k-means clustering algorithms are using in this propose system.

Keywords Origin blacklist, Porter stemmer , K-Means clustering

I. INTRODUCTION

E mail is the one of the most important application of network users. But Mail has facing problems like Hacking attacks, phishing attacks and malicious attacks to attempt fraud and deception motivation. They are using the emails to obtain personal credentials of user financial gain. The genuine contents of mail may include the Phishing URLs to steal the useful data which is also known as Spam. Spammers use the Spambots to create email distribution lists. Typically the Spammer sends an email to millions of email addresses, with the expectation that only a small number will respond or interact with the messages. Fraudulent spam also comes in the forms of phishing mails, which are emails disguised as official communication from banks, online payment processors or any other organizations a user may trust. The spam mail may also deliver other types of malware through file attachments or scripts, or contain links to website s hosting malware. There are many types of spam filtering techniques such as List based techniques, content based techniques, challenge –response, keyword-based filtering, etc. The content based techniques, which are very popular techniques to avoid spam mails. The spam emails are evaluated for terms or words to detect as spam or legitimate mail. These techniques consist of both Word based and heuristic filters. Word based, which block mail as a spam, if the mail has certain word having spamicity character. The spam mails contain the words that are rarely used in legitimate mails. So it can be easy to block the spam mails. Then the heuristics filters, which outperforms the normal word based filter. Mostly the score and point is criteria to classify the mail as spam or legitimate. Heuristic filters make use of different types of algorithms to examine the emails.

Mr K. Nagarajan, Assistant Professor, Department of Electronics And Communication Engineering, Nehru institute of Engineering and Technology , Coimbatore , India . (Email Id : naguambani@gmail.com)

Mr.Vinod Kumar V, Assistant Professor, Department of Electronics And Communication Engineering, Nehru institute of Engineering and Technology , Coimbatore , India . (Email Id: vinodvijayan0289@gmail.com)

Mr. Mohammedkasim M, Assistant Professor, Department of Electronics And Communication Engineering, Nehru institute of Engineering and Technology , Coimbatore , India . (Email Id : mohammedkasim1983@gmail.com)

Mr. T Prabu, Assistant Professor (SG) , Department of Electronics And Communication Engineering, Nehru institute of Engineering and Technology , Coimbatore , India . (Email Id : tprabu19@gmail.com)

II. PROPOSED SYSTEM

In this paper, we examine whether we can find malicious networks in a systematic manner using existing blacklists. The systematic detecting of disproportionately malicious networks may be used to build to determine if the networks are harboring a significant amount of malicious activity. These kind of metrics may offer several practical benefits. ISBs could use them to build identification of malicious networks into their peering agreements also when receiving traffic, the destination network mite prioritize traffic based on the cleanliness of Ass, which can help metrics to estimate.

As a simple and powerful statistical method, this is consisting of desirable features. This can minimizes the num of observations required to reach the decision among all the non sequential and sequential statistical tests with greater error rates. This means that the spam detection system can identify the spam content or not by using Data Mining techniques. The users of the system can be selecting the desired thresholds, weight and term frequency is computed to detect the spam content. The propose system efficiently identify the spam data and spam mail will be blocked to overcomes the network and zombies attack. By using this method, It can be improves the network performance and the network traffic is reduced with secure authentication. Once the administrator verifies the message as spam monitors the spam sender and stores their activity in the routing table which helps to easily find the spam

sender. Then this system is used autonomous system from unknown privileged user attacks.

III. PORTER STEMMER ALGORITHM

In this algorithm, Stems using a set of rules, or transformations, applied in a succession of steps. It is an excellent trade-off between speed, readability, and accuracy. Then there is no recursion occurs in this Porter stemmer algorithm. The following steps are in this stemmer algorithm,

Gets rid of plurals and –ed or –ing suffixes.

Turns terminal y to i when there is another vowel in the stem.

Maps double suffixes to single ones:

-iztion,-ational,etc.

Deals with suffixes, -ful, -ness, etc.

Takes off –ant, -ence, etc.

Removes a final –e.

Application of Stemmer algorithm:

Stem both document indexes and queries

Any situation where one is interested in grouping words into semantically similar sets.

Often can increase recall without decreasing precision.

IV. K-MEANS CLUSTERING

It is heuristic method. Here each cluster is represented by the center of the cluster"k" Stands for number of clusters, it is typically a user input to the algorithm and some criteria may be used to automatically estimate K.

The K-Means algorithm has following four steps:

Select the Initial centroids at random.

Assign each object to the cluster with the nearest centroids.

Compute the each centroid as the mean of the objects assigned to it.

Repeat previous 2 steps until no change.

The x1,...,xN are the data points or vectors of observations. Each observation (vector xi) will be assigned to one and only one cluster, The C (i) denotes cluster number for the i th observation. K-Means minimize the Within-cluster point scatter:

$$W(C) = 1/2 \sum_{K=1}^{K} \sum_{C(i)=k} \sum_{C(j)=K} |\!|xi-xj|\!|^2 = \sum_{K=1}^{K} N_k \sum_{C(i)=k} |\!|xi-mk|\!|^2$$

Where,

M_K is the mean vector of the K^thcluster.

N_K is the number of observation in K^thcluster.

Advantages

This algorithm is efficient in Computation.

It is easy to implement.

V.SYSTEM ARCHITECTURE

In this figure shows that the computational learning process is done then it go to recipient mail, if any spam mail passes after filtering the spam mail. The recipient can help the server to learn and identify the sender as a spammer according\ to the sender feedback. The Sender sends the mail to the receiver through mail server, which is used to identify the mail whether it spam or legitimate by using spam filter.

VI. MODULES

Mail Server Creation
Admin
User
Web Access Security
Accessibility and Verification
Remote Login Limitation
Cookies Verification
Preprocessing
Removing Stop words and Stem words.
Term Frequency

Spam Detection

Module Description

A. Mail server creation

The mail server is created to communicate with the difference persons through email server. The mail server also has the option for viewing sent mails, spam mail and detected mails.

B. Users

Users are end persons who are making or initialize the communication with the server.

It can be split as,

Legitimate users

Attackers

C. Web Access Security

Due to their ubiquitous use for personal data, web services have always been the target of attacks.

Accessibility and verification

Remote login Limitation

Cookies verification

D. Preprocessing

The 1st step towards controlling and analyzing textual data formats is to be the text based information available in free formatted text documents.

Removing Stop words and stem words

Initially, Its removing the unnecessary information available in the form of stop words.(e.g. is,am,the,of.,).

The stemming words e.g. 'Achieve', 'Achieving' and 'Achieved' are stemmed to 'Achieve'.

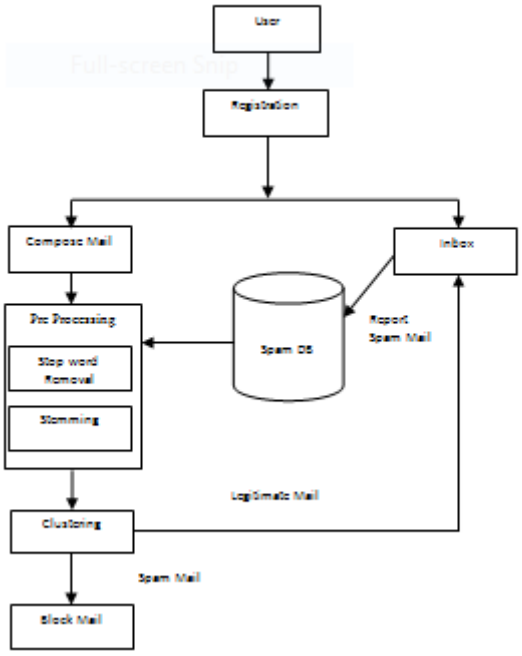
Term Frequency

The data is given a suitable representation based on terms or words defined in the text.

The term frequency (TF) and Inverse Document Frequency (IDF) methods may be used in this level of information processing.

E. Spam Detection

The Spam data are clustered with the data mining techniques. In this K-means clustering is applied to group all the spam mails which is related with the new emails. If the new mail matches with threshold value, then the mail is blocked to flow in the network. This system prevents the storage of spam mail from memory. If the mail doesn't match with spam data, then it will send the data to receiver.



Data flow diagram

VII. CONCLUSION

In this proposed system, Porter stemmer algorithm is used to stemming the words and detect the spam .Then K-means clustering algorithm is used to calculate the tf, idf, weight of terms in the data and finally it will be blocked. In our proposed system, the Spam mails can be easily detected and it can improve the performance of networks. The system is used autonomous system from privileged user attacks.

VIII. REFERENCES

1. R Lu et al., "PReFilter: An efficient privacy-preserving relay filtering scheme for delay tolerant networks, "In Proc. IEEE Conf. Comput.Commun.

2. K.Zhang, X.Liang, R. Lu, & X.Shen, "SAFE: A social based updatable filtering protocol with privacy-preserving in mobile social networks, "In Proc.IEEE Conf.Comput. Commun.

3. K.Wei, M.Dong, K. Ota, and K.Xu,"CAMF: Context-aaware message forwarding in mobile social networks,"IEEE Trans.Parallel Distrib.Syst., vol. 26.

4. X.Liang, X.Li, K.Zhang, R.Lu, X.Lin, and X.Shen,"Fully anonymous profile matching in mobile social networks,"IEEE J.Sel.Areas Commun.vol.31.

5. R Lu et al." Social Aware Multicast In Disruption-Tolerant Networks"."In Proc. IEEE Conf.Comput. Commun.

6. Dhanaraj S., Dr.V.Karthikeyani."A Study on email-image spam filtering techniques."In Pattern Recognition, Informatics and mobile Engineering (PRIME), 2013 International Conference.

7. Y Gao, A Choudhary and Gang Hua," A Comprehensive Approach to Image Spam Detection: From Server to Client Solution ", Information.

8. Chirita, P.A. Jo, D. and Nejdl, W. Mailrank,"Using ranking for spam detection,"Proceeding of the 14th ACM International Conference on Information and Knowledge Management, CIKM 2005

9. Medlock, B., "An adaptive approach to spam filtering on a new corpus, "Proceedings of the Third Conference on Email and Anti-spam, CEAS'2006.

10. Cook, Duncan, Jacky Hartnett, Kevin Manderson, and Joel Scanlan."Catching spam before it arrives: domain specific dynamic blacklists."In Proceedings of the 2006 Australasian workshops n Grid Computing and e-research-Volume 54, 2006.