Advances in Computer Science and Information Technology (ACSIT)
Print ISSN: 2393-9907; Online ISSN: 2393-9915; Volume 2, Number 6; April-June, 2015 pp. 580-583

http://www.krishisanskriti.org/acsit.html

Identifying E-mail Headers Suitable for Machine Learning based Spam Filtering

Nadia Nisar[1] and Zaffar Kanth[2]

[1]M.Tech Scholar, Department of Computer Science & Engineering Al-Falah School of Engineering and Technology Dhauj, Faridabad, Haryana
[2]Researcher, National Institute of Electronics and Information Technology
E-mail: [1]nadia.nisar786@gmail.com

Abstract—*Spamming brings annoyance which can not only degrade the e-mail communication but also lead to various problems including financial losses due to various kinds of attacks like phishing and spoofing. Spammers use sophisticated techniques to fool the common spam filters and by-pass the classification techniques applied by the filters. One of the ways spammer employs to by-pass spam filters is by modifying various e-mail header fields. The typical e-mail sent or received has two sections viz, e-mail Header section and the e-mail body section. Both the sections of the e-mail message could hold valuable information for a spam filter to classify a mail as either a spam or a non-spam. In this paper we identify potential header fields and possible feature set for formation of a dataset for machine learning based spam filtering. Machine learning approach has extensively been employed in filtering rules based on e-mail body. We propose a converged dataset based on attributes of e-mail body and some potential fields like received, subject etc in the e-mail headers. Incorporating header fields in the dataset could simplify, otherwise a time-consuming filtering process.*

Keywords: *Spam, Machine-Learning, E-mail Header, Classifier, Dataset.*

1. INTRODUCTION

Spam is a continuous nuisance an e-mail user encounters while interacting with an e-mail application. People spend a lot of time and effort in order to get away with the unsolicited emails received as a substantial part of the received e-mails in the user inbox. The frequency and the magnitude of the received spam messages can sometimes be of several orders of magnitude in comparison to the solicited e-mail messages. Spamming can just not only be a reason of worry to e-mail users but a large amount of junk mails could consume considerable amount of network bandwidth. Further, spams can cause various security issues like phishing and spoofing attacks which can lead to money laundering and other monetary losses. Spammers use sophisticated techniques to fool the common spam filters and by-pass the classification techniques applied by the filters. Spammers can use malware combined with the power of botnets to launch large scale spamming missions causing major traffic increase and leading to enormous economical loss [3]. One of the ways spammer employs to by-pass spam filters is by modifying various e-mail header fields as most of the fields in the e-mail headers can be forged. The typical e-mail sent or received has two sections including e-mail Header section and the e-mail body section. The header section comprises of various fields such as To, From, Subject, Content Type, MIME-Version, Message-ID etc. The body contains the actual message contents. The SMTP protocol is used to transfer e-mail's as set of 7-bit ASCII characters from one machine to another. MIME , an Internet standard is utilized to send e-mails with content written in different languages with additional character sets and various formats.

Some of the important header fields are listed below:

From
Lists whom the message is from and can be easily tampered with. This is the least reliable header field.

Subject
Although optional field, Spam as well as non spam mails usually have this field as the title of the e-mail.

To
Shows to whom the message was destined, but the recipient's address may not be present

Return-Path
The email address for return mail. This is the same as "Reply-To:"

Delivery Date
This shows the date and time at which the email was received by your (mt) service or email client.

Received
The received is the most reliable and important email header field. It lists all the servers through which the message traversed.

The received field is parsed from bottom to top.

Message-id

A unique string generated by the e-mail system when the message is created.

X-Spam-Level

Displays a spam level set by mail client/service.

Fig. 1 depicts a sample snapshot of the e-mail header as captured in gmail. The header looks like set of (name, value) pairs.

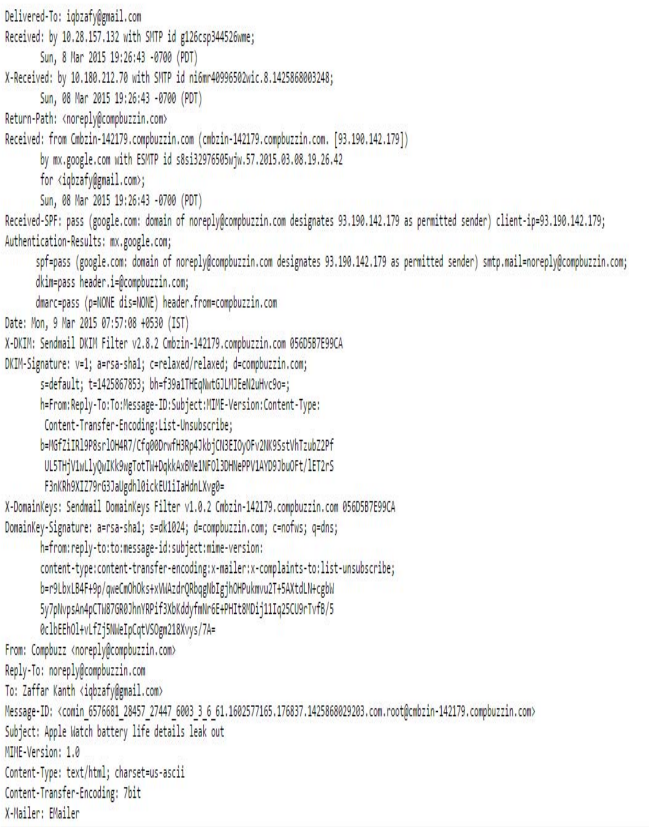

Fig. 1: Gmail Header

2. MOTIVATION

It has been analyzed that the spam filtering done on the principles of Message body semantics is time consuming and rigorous process. Modern day spam filtering techniques also include rules based on information contained in the email headers. There are various fields in e-mail headers that could be utilized to classify spams. For example, header information can reveal if the email is from a recognizable domain that is associated to the actual sender name. Finding out the IP reputation using Return Path E-mail Header is another example to utilize header information for spam classification. Although, extensive work has been done on machine learning approach applied to e-mail contents (e-mail body), little research has been done on the application of classifiers like naïve-bayes on e-mail headers. In contemporary spam filters Header based filtering is used in conjunction with e-mail body semantics based techniques. However, novel machine learning principles applied on e-mail headers could simply the filtering process and avoid rigorous and time consuming techniques applied on e-mail body.

3. RECENT ADVANCES

Both the sections of the e-mail message could hold valuable information for a spam filter to classify a mail as either a spam or a non-spam. An extensive research based on whitelisting/blacklisting, Bayesian probability analysis, mail body keyword tokenization and checking etc have already being conducted by many researchers. One of the common methods for spam detection is Bayesian spam filtering (Thomas Bayes) which is a statistical technique for e-mail filtering. In this technique a naïve Bayes classifier is used to identify spam e-mail. Bayesian classifier work by correlating the use of tokens with spam and non-spam e-mails and then using Bayesian inference to calculate a probability that an e mail is spam or not.

4. PROPOSED METHODOLOGY

Considering the existence of scope to improve the spam filtering activity we propose spam filtering based on e-mail headers complimented with machine learning approach applied on e-mail body. The proposed work is based on Naïve Bayes Classification technique which includes machine learning based filtering applied not only on the message body but also on some chosen fields in e-mail headers. The methodology uses custom developed dataset for header fields and an open source dataset spambase for e-mail body dataset testing. The work done has identified potentially relevant fields in the e-mail headers like:

- Received
- Number of receivers
- Message-ID
- Subject

Feature selection of Header-based email spam filtering technique was made on the basis of careful selection of some of them to be among the features used for classification. The selection was done after analyzing large number of publicly available datasets. The fields in the header along with the possible feature set considered are as follows:

Received

Each email can contain more than one Received field and is most reliable field. This field is typically used for email tracking by reading it from bottom to top. The bottom represents the first mail server that got involved in

transporting the message, and the top represents the most recent one, where each received line represents a handoff between machines. Hence, a new received field will be added on the top of the stack for each host received the email and transport it, and to which host the message will be delivered, in addition to the time and date of passing. The potential feature set that we extracted from this field includes features like:

Number of sender fields(S)

Total time

Number of receivers (N)

IP and domain name Validity

Ratio N/S

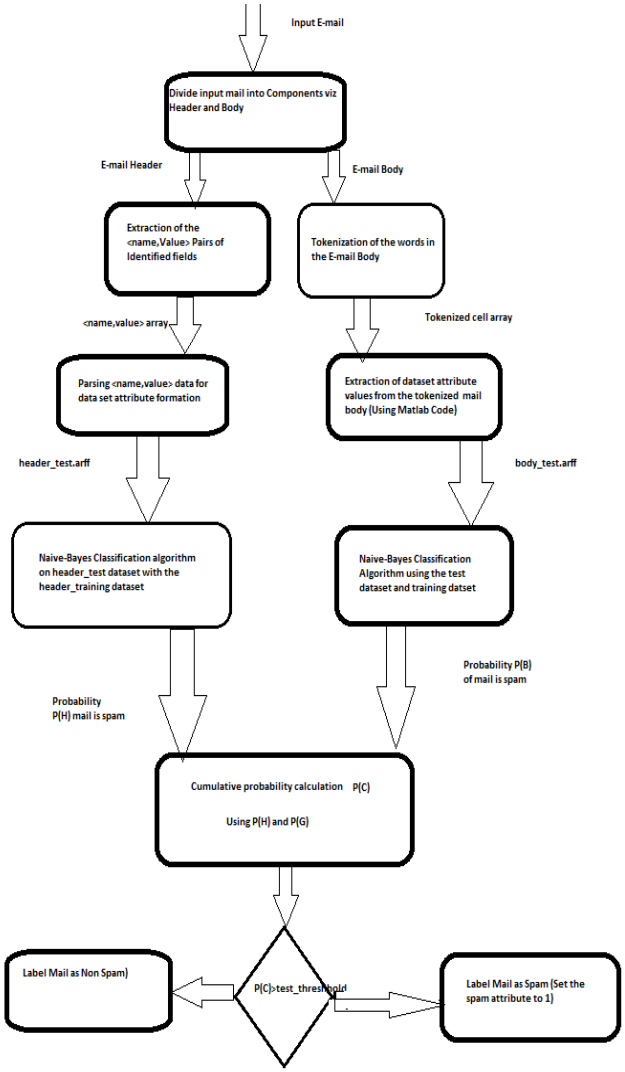

Fig. 2

Message-ID

Each message is assigned with a globally unique key by the Mail user agent. The Message-ID field is generated using the name of the machine and the date/time of the email when it is sent. This field is made up of two component parts with @ sign as the separator. The first part specifies the date and time when the mail was generated. The other part is the machine name. This field can be easily. Therefore, it is required to make sure that the domain name in the Message-ID field is consistent with the domain name in the from field. Inconsistency of this information would indicate a spamming behavior.

Subject

Spammers tend to use special characters, words and phrases in the limited size header fields. Words like Congratulations, Reward, and Money etc could be good indication of the mail being spam. In our approach presence of certain list of words can be used for feature selection

Using the above mentioned feature set a training dataset was populated using the attribute values. The dataset along with the spambase dataset could be used with test dataset separately for header attributes and body attributes respectively.

5. IMPLEMENTATION

Platform for performance benchmarking in MATLAB was setup. Open source machine learning dataset spambase was studied and converted to a format suitable for integration in a Matlab platform to take advantage of MATLAB's Strong visualization and analysis features. Spambase is created by Mark Hopkins and others at Hewlett-Packard Labs [4]. Now, It is offered by UCI- machine learning and is a repository of e-mails (blend of spam and non spam) from the postmaster and the people who have reported spams. The spambase dataset hasThe dataset usually, is used as ARFF file and contains various attributes based on E-mail body. Most of the attributes indicate whether a particular word or character was frequently occurring in the e-mail body. The run-length attributes measure the length of sequences of consecutive capital letters. The excerpts of the detailed attribute set are as shown below:

48 continuous real [0,100] attributes of type word_freq_

WORD = percentage of words in the e-mail that match WORD, i.e. 100 * (number of times the WORD appears in the e-mail) / total number of words in e-mail. A "word" in this case is any string of alphanumeric characters bounded by non-alphanumeric characters or end-of-string. 6 continuous real [0,100] attributes of type char_freq_CHAR]

= percentage of characters in the e-mail that match CHAR, i.e. 100 * (number of CHAR occurences) / total characters in e-mail

1 Continuous real [1,...] attribute of type capital_run_length_average
= average length of uninterrupted sequences of capital letters
1 continuous integer [1,...] attribute of type capital_run_length_longest
= length of longest uninterrupted sequence of capital letters
1 continuous integer [1,...] attribute of type capital_run_length_total
= sum of length of uninterrupted sequences of capital letters
= total number of capital letters in the e-mail
1 nominal {0,1} class attribute of type spam
= denotes whether the e-mail was considered spam (1) or not (0), i.e. unsolicited commercial e-mail.

Preprocessing of the spambase dataset was carried out using WEKA an open source, java based suite of tools meant for data mining tasks in machine learning. Binarization of the dataset was required and removal of certain irrelevant attributes was done in order to simplify e-mail body analysis.

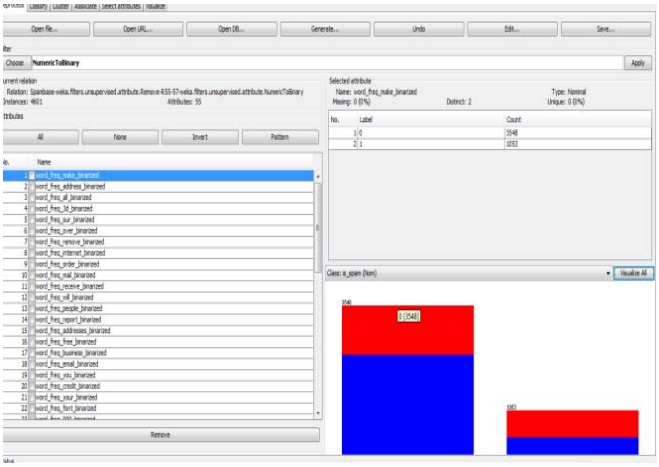

Fig. 3: Various attributes of the spambase.arff as seen in the weka tool. The plot shows the percentage of spam and non spam mails in the data set.

A custom made training dataset on the basis of feature set identified from the e-mail header fields was created using 500 e-mails. The training dataset was merged with simplified variant of spambase dataset.

The overall design of the filtering approach is shown in Fig. 2.

MATLAB platform was used to test the spam filtering methodology in a passive manner. Further, Matlab was used to calculate and compare the confusion matrix in both the scenarios. The matrix is based on set of test dataset (ARFF File) made up of 50 mails (1:1 ratio). The Matrix suggests no evidence of performance degradation with simplified spambase dataset. However, significant amount of time was saved to suggest that header based machine learning can reduce the time required to classify the mails. The subsequent integration with a Mail User Agent could be done to test the performance in the active manner.

Table 1: When only spambase dataset was used

	Actual	
Predication	**Spam**	**Non-Spam**
Spam	19	4
Non-Spam	6	21

Table 2: When simplified spambase was used with custom header based dataset

	Actual	
Predicition	**Spam**	**Non-Spam**
Spam	18	3
Non-Spam	7	22

6. CONCLUSION

When the naïve-bayes classification technique is applied using machine learning based dataset encompassing certain header field based attributes, the rigorous and time consuming filtering done on e-mail body can be simplified without affecting the spam filter performance. The conclusion was arrived after removing certain attributes and simplifying others from spambase dataset and incorporating the custom developed dataset.

REFERENCES

[1] Shukor Bin Abd Razak, Ahmad Fahrulrazie Bin Mohamad Identification of Spam Email Based on Information from Email Header, 2013 13th International Conference on Intelligent Systems Design and Applications (ISDA)
[2] Understanding an email header http://kb.mediatemple.net/
[3] Spamming Botnets: Signatures and Characteristics Yinglian Xie, Fang Yu, Kannan Achan, Rina Panigrahy, Geoff Hulten+, Ivan Osipkov+ Microsoft Research, Silicon Valley
[4] Mark Hopkins, Erik Reeber, George Forman, Jaap Suermondt Hewlett-Packard Labs, 1501 Page Mill Rd., Palo Alto, CA 94304