Spammer Detection by Extracting Message Parameters from Spam Emails

Acquin Dmello
Student,
St. Francis Institute of Technology,
Mumbai, India

Gaurang Mhatre
Student,
Sardar Patel Institute of Technology,
Mumbai, India

Rohan Lopes
Student,
St. Francis Institute of Technology,
Mumbai, India

Haince Pen
Alumni,
St. Francis Institute of Technology,
Mumbai, India

ABSTRACT

Traditional and present methods to detect spam emails have been working quite well but they take no measures to detect and occlude the malicious actions of the spammers. In this paper a combination of certain parameters of an email is considered to cluster legit emails and spam emails. Initially, this approach tries to cluster spam emails. Based on their sources, the spam emails are clustered using their Message subjects, Attachments, Number of Hyperlinks, Message length, Stylistic and Semantic parameters. Since emails from same source have certain similarities, they are clustered together. These clusters are then mapped to their respective domains and their IP address is retrieved which is then reported to Anti-Spam Agencies.

General Terms

Algorithm, Clustering, Feature Extraction, Spammer Detection, Security, Algorithm, WEKA.

Keywords

Detection, Email Parameters, Information Extraction, Spam

1. INTRODUCTION

Spam related cyber-crimes are one of the most ever growing threats to the society. Spamming contribute to illegal earnings by selling various products and also spread malwares which serves as a medium to steal confidential data from a user's computer or makes their functioning ambiguous. Present methods to combat spamming have been quite lethargic since they serve as a temporary means to occlude the spamming effect. The best system to completely prevent spam emails is to stop the source of a spam, that is, to detect the spammer.

At present there are three important techniques which are used by spammers to send spam emails [1]:

- Open relays and Proxies
- Botnets
- Short-lived BGP announced routers

Open relays are a type of mail servers which allow any internet hosts to connect and send emails through them. Botnets are collections of machines that communicate with other machines to perform a similar task under one centralized controller. But these techniques are not so powerful, as current Anti-spam technologies are able to mitigate their effect. The most complex technique is the Short-lived BGP announced routers technique, in which a spammer announces an IP space, sends spam emails and then the IP space vanishes after some time [1], [2]. In this way, spammers manage to remain in dark.

This research proposes a combined approach towards detecting the source of spam and reports it to Anti-spam agencies. To block spammers' from sending spam emails, their supporting architecture should be eradicated. Hindering the functionality of spam hosts will highly abate spammers' revenue from illegal email campaigns and obstruct their ability to do spam email cyber-crimes. This research promotes a combination of algorithms for clustering spam email domains based on the hosting IP addresses and other emails parameters. This combination of algorithms detects potential spammer source over certain period of time. Evaluated experimental results show that when domain names are examined, it is found that many unrelated spam emails are actually related. By using wildcard DNS records and constantly replacing old IP domains with new IP domains, spammers can efficiently spoof URL or domain based blacklisting. Spammers also change their IP addresses occasionally, but not as frequently as domains. The domains and IP addresses that are identified using the method proposed in this paper, cyber crime investigators can be forwarded to trace the identities of spammers and the investigators can shut down the related spamming architecture. This paper illustrates how data mining and clustering techniques can help to detect spam domains and their hosts for anti-spam forensic purposes.

2. RELATED WORK

Today, researchers on spam are interested in identifying and obstructing the source of spam emails and not just identify the spam emails. Spam can be more effectively stopped by disrupting its source, such as the C&C and hosting servers by taking legal actions. This paper takes the same concept into consideration and proceeds with the research. The goal of this spam research is to cluster spam emails and identify spamming infrastructure that belongs to the same spamming group. In this paper, the related research is reviewed, including anti-spam and clustering algorithms on data streams. Spammer detection ability can be increased by considering more parameters for clustering.

According to Halder et al [3], spam emails have some identical styles and semantics within them. He proposed that spam emails can be identified using stylistic and semantic approach and hence identifies the spammer with the help of feature extraction using Data mining.

Li et al [4] claimed that spam emails are generally sent in groups having certain similarities in between them with respect to their domain, URLs present within them or prototype. Hence their research specified that different spam campaigns around the world can be grouped under a small group of spammers.

Also Chun et al [5] deliberated on similar tendencies of spammer behavior and inferred that clustering emails together based on their subjects and IP addresses can prove to be an effective strategy in determining spammer source. They did not just look for exact similarity, but they also explored fuzzy similarity.

This paper comes up with the idea of clustering spam emails by considering more features in order to make more specific clusters of spam emails. These specific clusters are then mapped to their respective IP addresses and then reported to take legal actions, thereby completely mitigating the source of spam.

3. METHODOLOGY

The approach to detect and report the spammer takes place in six important steps which are demonstrated in the figure below.

3.1 Data Collection

Data was collected from an Institution's local server which had approximately 20000 spam emails. The collected data originated from the mail servers containing the email accounts of the employees of the institution. When an email is received by the mail server from a non-existent email account but on the expected domain then the mail is moved to a different account on the mail server called the catch all accounts using an in-house-built heuristics rule that are specified to filter spam emails. It is assumed that the mails that are collected in this account are spam. The contents with uniform resource locators (URLs) were also checked to see if they were on the Institution server's known harmful sites blacklist database. Harmful sites can be "phishing" sites, sites that download malicious software onto user's computers, or spam sites that request personal information.

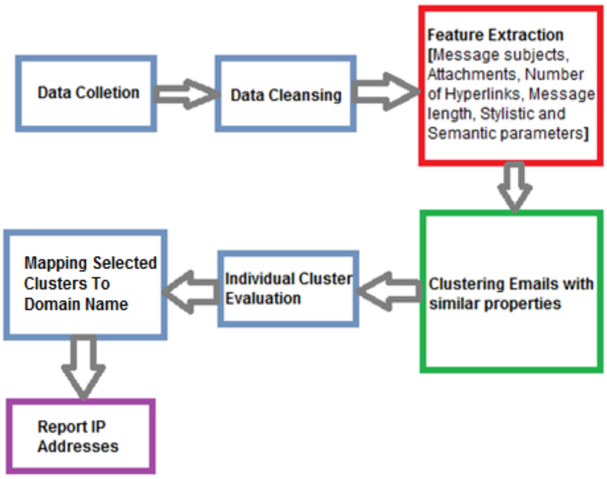

Data Colletion
Data Cleansing
Feature Extraction [Message subjects, Attachments, Number of Hyperlinks, Message length, Stylistic and Semantic parameters]
Clustering Emails with similar properties
Individual Cluster Evaluation
Mapping Selected Clusters To Domain Name
Report IP Addresses

Fig. 1: Flow of the approach to detect and report spammer. Note that "Fig.1" is abbreviated. There are sub sections of each step.

To identify spam mails, classification machine learning methods which consist of a training process and a testing process can be applied. During the training process, learning takes place to generalize information from a given data set which contains a large number of attributes. Often, using too many attributes may cause over-fitting of data during training which can hinder classifiers from classifying "new" data correctly. Using too few attributes may not be powerful enough to generalize characteristics of data. The collected data consisted of Seventeen content attributes and thirteen user behaviors attributes, all numeric values, from the raw data. It is critical to use important and relevant attributes and remove redundant and irrelevant ones for a chosen machine learning algorithm to obtain high classification rates.

Content attributes consisted of the following: emails replied, emails with spam words, emails with URLs, along with the mean, median, min, and max of the followings: the number of spam words per sentence on each email, black listed URLs in each email, characters per email, numeric characters per email, words per email and the number of times an email is sent, Removal of words with length < 3, Removal of stop words.

3.2 Data Cleansing

The next step that precedes data collection is Data Preprocessing. Initially, the emails that contained gibberish and lack of any one language of communication are removed. Next, the emails that had only web links or attachments were removed. Then all the emails that had at least one line of text and with more than four words were considered. Porter's stemming and stopping algorithm was used for this purpose. Stemming and stopping are done to reduce the vocabulary size which helps information retrieval and classification purposes.

Preprocessing left us with approximately 30% of email data collected. Filtered emails were branched into data sets of different sizes. In the work considered in this paper, an efficient clustering algorithm incorporating the features of K-Means and Expectation Maximization are used.

3.3 Feature Extraction

The features that can help in clustering similar documents together were identified. The calculation of clusters is done considering attributes of the email such as: Message Subject, Attachments, Number of Hyperlinks, Message length, Stylistic and Semantic features.

3.3.1 Message Subject

Emails sent by spammers around the world have maximum probability of containing trigger words (Examples: Meet Singles, Work from Home, Business Opportunity, Hello, weight Loss, Buy Pills).In this experiment 104 different trigger words are taken into consideration and mails are clustered according to the number of trigger word found in the subject of the mail.

3.3.2 Attachments

Attachments serve as the base of information for the spam mails to bypass spam filters. They may be distorted up to a certain extent, so that the spammer can specify the address or message on which he wants the traffic to be directed. This distorted image is scanned using image scanner module and that particular email is clustered in its respective cluster.

3.3.3 Number of Hyperlinks

Spam email generally contains graphical or text links to websites where the vital actions take place for spammers to make a profit. Mails with only one or more hyperlinks have high probability of being a spam mail as compare to an email with hyperlinks along with other context. By identifying the numbers of hyperlinks mails can be distributed in respective clusters.

There are three types of URLs: static URLs, URLs with sub-URLs, and random-obfuscated URLs [6].Fixed URLs are the ones in which the spammer inserts the same URL in all the messages. Usually, these URLs correspond to small links with meaningful and readable names, such as buypills.com. Those are the short branches which end at depths 3 or 4, usually. A different strategy frequently observed is the selling of different products in the same URLs, what generates a set of URLs in which each URL correspond to a different product from the same website. For example, pills1.htm, pills2.htm and pills3.htm are different products associated with the same URLs. As long as the spammer keeps other parts of the URL and its layout fixed, these distinct messages will be grouped into the same URL cluster because the portion of the URL that specifies the product is infrequent compared to the other messages' characteristics. Finally, the third class of URLs is the one in which spammers constantly obfuscates their URLs inserting random fragments which are different for each message.

3.3.4 Message Length

During the study it was observed that spam emails generally are limited in terms of message length. As the primary goal of a spam mail is to direct traffic to the intended website, the spammers generally include URLs, to the intended website and no text. So this paper considers a mail with no text as spam and moves it to a different cluster for further evaluation.

3.3.5 Stylistic Parameters

Spammer most often use obfuscation or misspelling to bypass spam filters (For Example: HOW ARE YOU? Written as H0W 4R3 U?). Such distinct parameters of mail can be used to detect spam emails and group them in cluster. These parameters include: total number of word count of the text in the email, number of lines present in the email body, total number of the punctuations used in the email body, total count of contractions used in the email, total number of obscured words used in the email, total count of email ids used in the email, types of different punctuations used in the email etc.

3.3.6 Semantic Parameters

Semantic parameters are the parameters that give us semantic or logical meaning of the spam emails. There are two classes of Semantic features used in this approach [3]. These two classes are Tf-Idf (Term frequency-Inverse Document frequency). Tf-Idf is a statistical measure that can be used to represent the importance of a term or word in a document. Tf-Idf is used for the first x words, that are most frequent words used in the dataset and the count of the first x bigrams used in the dataset, where x is the number that is decided based upon the cut-off of the minimum frequency count. Using Tf, the term frequency of each term in the available document can be calculated. The Idf provides with the general significance of the term in the whole dataset by performing mathematical division of the total number of total number of documents by the number of documents containing the term.

4. Clustering

The clusters are manually evaluated with the ground truth data that was manually collected. Purity was calculated using equation given below.

Purity (%) = ∑#of correctly clustered instances /# of instances *OR* **∑#of clusters providing features similar to previously reported spam features**

The clusters` purity is evaluated based on correctly clustered instances and clusters having similarity with previously reported Spam emails. From the first part of equation (∑#of correctly clustered instances) it can be checked that clustered emails are correctly predicted to have either one of the above mentioned features or all. If first part of equation fails (i.e. if none of the features are identified for cluster) then a check is done whether the cluster's feature has similarity with previously reported spam features. Next, logical OR operation is performed on the two results. By doing this maximum purity from clusters can be expected. Also it is possible to recover the email which is not a spam. But if it accidently identifies a spam email as non-spam, then it can be considered as spam based on second part of Equation (∑#of clusters providing features similar to previously reported spam features). Later the clusters can be analyzed individually and the results can be presented for the cluster that gives the highest accuracy.

The threshold for the minimum number of emails in a cluster had to be at least 8% of the total emails in the dataset else the cluster was discarded. This was done to avoid false alarms given by singleton or small clusters claiming themselves to be 100% pure. Weka implementation of these two algorithms was used to do so [8].

4.1 Mapping Selected Clusters to Domain Name

The Hyperlinks in the spam emails are fetched and store in the Hyperlinks Table with the respective message ID. The WHOIS information of the Hyperlink can be found by mapping the Hyperlink to the IP address. This WHOIS information which is the information of the domain registrar is sent to Cybercrime investigators or legal authorities so that appropriate legal action against the spammer can be taken.

5. Results

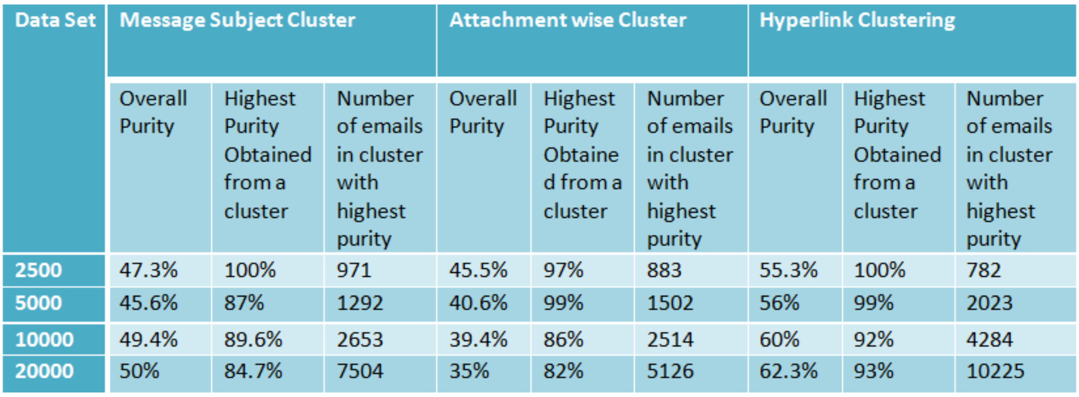
Table 1, Table 2 and Table 3 shows the clustering accuracy of DBSCAN [9] and CURE [10] algorithm combined on the dataset. The purity of the combined feature set is better than Message Subject, Attachment, Hyperlinks, Message Length, stylistic and the semantic features individually. The accuracy of the results is increasing as the size of the data set grows.

Stylistic clustering yielded good output when the email length was short (i.e. where the total count of words was less) Emails that include differentiating punctuations like a sentence that always end with an exclamation mark (!) or question mark (?) or style are easy to identify using this stylistic clustering. Semantic clusters yield good results when the semantic body is rich in content. The length of the emails also affects the type of the differentiating features.

The data set of 20000 emails yields better results than 5000 emails data set for two reasons. Initially, the number of sample set will increase because of which the emails can be more specifically classified. Secondly, DBSCAN algorithm will filter impurity from a larger amount of emails that are available to compare from.

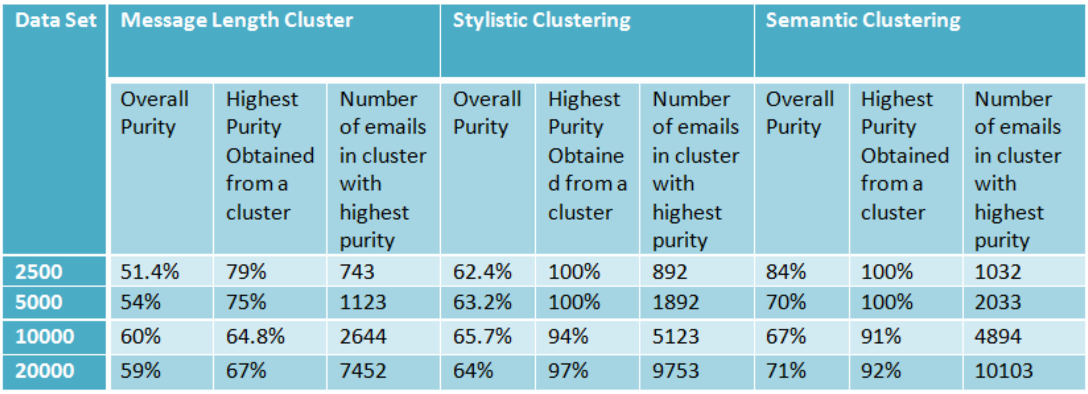
IP address mapping of the clusters give successful results for last 10000 set of emails because most of the domains of the previous 10000 data set were inactive by then. These last 10000 emails mapped to 54 unique IP addresses and it was possible to collect the WHOIS information for these addresses.

Table 1: Table showing the purity of clusters using DB Scan and Cure algorithm on the data set



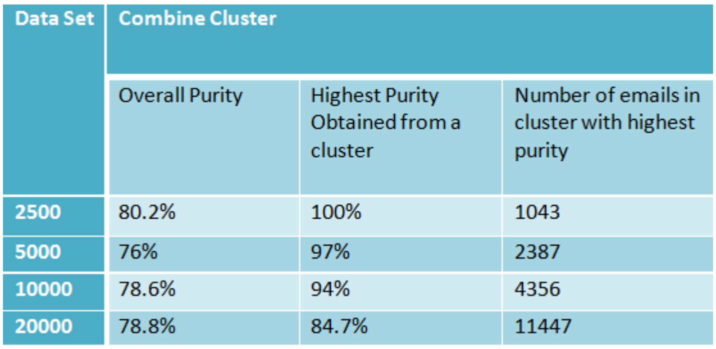
Data Set	Message Subject Cluster			Attachment wise Cluster			Hyperlink Clustering		
	Overall Purity	Highest Purity Obtained from a cluster	Number of emails in cluster with highest purity	Overall Purity	Highest Purity Obtained from a cluster	Number of emails in cluster with highest purity	Overall Purity	Highest Purity Obtained from a cluster	Number of emails in cluster with highest purity
2500	47.3%	100%	971	45.5%	97%	883	55.3%	100%	782
5000	45.6%	87%	1292	40.6%	99%	1502	56%	99%	2023
10000	49.4%	89.6%	2653	39.4%	86%	2514	60%	92%	4284
20000	50%	84.7%	7504	35%	82%	5126	62.3%	93%	10225

Table 2: Table showing the purity of clusters using DB Scan and Cure algorithm on the data set



Data Set	Message Length Cluster			Stylistic Clustering			Semantic Clustering		
	Overall Purity	Highest Purity Obtained from a cluster	Number of emails in cluster with highest purity	Overall Purity	Highest Purity Obtained from a cluster	Number of emails in cluster with highest purity	Overall Purity	Highest Purity Obtained from a cluster	Number of emails in cluster with highest purity
2500	51.4%	79%	743	62.4%	100%	892	84%	100%	1032
5000	54%	75%	1123	63.2%	100%	1892	70%	100%	2033
10000	60%	64.8%	2644	65.7%	94%	5123	67%	91%	4894
20000	59%	67%	7452	64%	97%	9753	71%	92%	10103

Table 3: Table showing the purity of clusters using DB Scan and Cure algorithm on the data set using combined approach



Data Set	Combine Cluster		
	Overall Purity	Highest Purity Obtained from a cluster	Number of emails in cluster with highest purity
2500	80.2%	100%	1043
5000	76%	97%	2387
10000	78.6%	94%	4356
20000	78.8%	84.7%	11447

6. CONCLUSION

Email technology has massively transformed the communication medium between humans. People find it easy to communicate with their peers electronically. It's fast, efficient and reliable. Due to the overuse of this technology spammers are attracted towards exploiting all the means of making illegal profits from commercial/advertisings. This is a threat to email technology.

The methodology of spammer detection mentioned in this paper not only identifies the spam source or spammer, but also reports them to legal authorities. Clusters with very high purities point to the leading spam tendencies for the period. There is an immediate requirement of a high speed system as spam emails need to be identified and reported immediately. Moreover, there is a need to deal with the real time data meaning that an immediate need of a technology that is fast, efficient, reliable and generalized. The spam emails keep on changing their prototype, their structure and also their method to spoof spam filters. So the technique to cluster these spam emails needs to adopt those changes. If these systems are implemented, spammer can be completely eradicated. In future, research in the field of email spam can consider more properties of a spam email to clustering method to detect the spam source.

7. REFERENCES

[1] Marios Kokkodis and Ting-Kai Huang. 2006. An empirical study of spam and spammers behaviour, University of California, Riverside.

[2] Anirudh Ramachandran and Nick Feamster. 2006. Understanding the Network Level Behaviour of Spammers.

[3] Soma Halder, Richa Tiwari, Alan Sprague. 2011. Information Extraction from Spam Emails using Stylistic and Semantic Features to Identify Spammers. IEEE.

[4] F. Li, M. Hseieh. 2006. An Empirical Study of Clustering Behavior of Spammers and Group Based Anti-Spam Strategies.

[5] C. Wei, A.P. Sprague, G. Warner and A. Skjellum. 2010. Clustering spam domains and targeting spam origin for forensic analysis, J. Digital Forensics, Security, and Law (Vol: 5), ADFSL.

[6] Pedro H. Calais, Douglas E. V. Pires Dorgival Olavo Guedes, Wagner Meira Jr., Cristine Hoepers, Klaus Steding-Jessen. 2008. A Campaign-based Characterization of Spamming Strategies.

[7] C. Liu, S. Stamm, 2007. Fighting Unicode Obfuscated Spam, InProc. Of the anti-phishing working groups 2nd annual eCrime Researchers Summit, USA.

[8] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann,I. H. Witten, 2009. “The WEKA data mining software: An update”,SIGKDD Explorations, Volume 11, USA.

[9] Henrik Bäcklund, Anders Hedblom, Niklas Neijman, 2011. A Density-Based Spatial Clustering of Application with Noise**.**

[10] Sudipto Guha, Rajeev Rastogi, Kyuseok Shim, 2001. Cure: An Efficient Clustering Algorithm For Large Databases