Recognizing Spam Domains by Extracting Features from Spam Emails using Data Mining

Kavita Patel
Thakur College of Engg and Tech
Mumbai, Maharashtra, India

ABSTRACT
This paper attempts to develop an algorithm to recognize spam domains using data mining techniques with the focus on law enforcement forensic analysis. Spam filtering has been the major weapon against spam, but failed to reduce the number of spam emails sent to an indiscriminate set of recipients. The proposed algorithm accepts as input, spam mails of personal account and extracts features such as stylistic, semantic, related email subjects and URLs present in the emails. The individual features are then clustered and evaluated. Further, these clusters are mapped with their respective domains. These spam domains are the URL of the webpage that spammer is trying to promote. The WHOIS information of the domain helps to get information about the source of that domain. Parameters like overall purity and the number of emails present in the cluster with highest purity is used to measure result of the individual features. An Experimental result shows that clustering of spam mails by stylistic and semantic parameter 20% less pure than other two features of spam mails.

General Terms
Spam, Semantics, Stylistics, Data Mining, Clustering

1. INTRODUCTION
Spam email is irrelevant or unsolicited messages sent over the Internet, typically to large numbers of users, for the purposes of spreading malware, advertising, phishing etc. which are a serious hazard to society. According to Kaspersky lab [3] (IT security firm) the share of spam in email traffic was approximately 72% in 2012. However, the increase in the worldwide use of email comes with an overwhelming increase in spam mails. The cost of spam mails consists of several components: the cost of bandwidth wasted by spam, the loss of productivity (as people have to spend time on spam), the cost of storage and network infrastructures, etc. [4] The FBI has released a new report showing that the cost of spam and its related scams is rising. In 2012 that figure rose to $485.3 million. The profit allows spammers to develop new trends and send out more spam messages.

Traditionally, the efficient way of controlling spam emails at the instant is filtering. However, filters can only distinguish spam emails or non-spam emails but cannot tell the origins of spam. One of the methods to reduce spam is domain blacklisting.

The motivation of this paper is to analysis of spam emails. During analysis [1], system extracts some features from spam emails and clusters them according to their parameter similarity. From these clusters spam domains are identified. The domains of the spam emails are analyzed and reported to the blacklist. Further incoming emails containing blacklisted domains will be blocked. Measures can also be taken to shut down the domains.

The rest of the paper is organized as follows: Section 2 provides background on spamming and an overview of previous related work. In Section 3, it describes proposed work for spam domain recognition. In Section 4 presents the analytical results. The conclusions and future directions are presented in Section 5.

2. BACKGROUND AND RELATED WORK
In this section, backgrounds of influential studies on spamming techniques are provided. Further in same section, discussion about spam methods & mitigation of it and the related research done.

2.1 Background
Now days, interest of researchers is developed in interrupting the source of spam emails and not just distinguish the spam emails. This paper supports the same concept and starts with the research [2]. The purpose of this paper is to use data mining's clustering techniques to cluster identical spam mails to identify spam domain that belongs to the unique spamming group. In this section, the similar concept is reviewed, including anti-spam methods and different clustering algorithms on data sets. Recognition of spam domain can be raised by considering more features for grouping.

According to Halder et al [1], analyzed spam emails have similarity in styles and semantics of them. They proposed that spammer can be recognized by clustering identical spam emails according to stylistic, semantic and combined features. Further, these clusters are relating to the internet protocol (IP) of these URLs and the whois information of the IP addresses help to get detail about the origin of spam.

Chun et al [2] studied on indistinguishable tendencies of spammer behavior and depicted that clustering same emails together depend upon the subject of the email. Later the IP address can be used as a better way to trace spammers. In their research, they look for exact similarity, at same time they also used fuzzy similarity.

Also Li et al [5] researched that spam emails are usually sent in bulk having specific equalities in between them with respect URLs presents in mail, to their respective domain or prototype which were used. Therefore, their research concluded that many different spam campaigns across the world can be merged under a tiny group of spammers.

This paper comes up with the idea of clustering spam emails by considering four features in order to make more specific clusters of spam emails. These specific clusters are then mapped to their respective domains and then reported to take

legal actions, thereby approximately reducing the source of spam.

2.2 Spam Methods

In this section, different methods used by spammers to send mail. Spammers use different techniques to send large volumes of mail while remains as undetected as possible, including:[6], [7]

2.2.1 Direct spamming

Intruders usually buy domain or upstream connectivity from spam-friendly Internet service provider (ISPs), which faiths blindly on any activity. Probably, intruders/spammers purchase this connectivity and forward spam through ISPs that do not interrupt this activity and are told to change ISPs. To remain undetectable in these situations, spammers most of time include a pool of dialup IP addresses, uses proxy to reverse traffic from dialup connection, and forward outgoing traffic through the high bandwidth connection.

2.2.2 Botnets

A botnet includes a network of compromised computers controlled by an attacker. These bots can be controlled remotely to perform wide scale send spam, deliver Trojans, send phishing emails, distributed denial of service (DDoS) attacks, distribute copyrighted media or conduct other illegal activities[22]. Many bots have a centralized infrastructure. i.e., they are connected to a command and control (C&C) server. In such an infrastructure, the C&C server acts as a master point of failure for the botnet. That is, the whole botnet can be disrupted if the defender finds the C&C server. To avoid this weakness, bot masters are now shifting to different infrastructures. In a peer-to-peer (P2P) infrastructure a node can act as a client and a server so, there is no centralized point for C&C.

2.2.3 Open relays and proxies

Also referred to as an open relay server, an SMTP e-mail server that allows a third party to relay e-mail messages, i.e., sending and/or receiving e-mail that is not for or from a local user. Open relays make it possible for mobile users to connect to corporate networks by going first through a local ISP, which then forwards the message to their home ISP, which then forwards the mail to the reach point. However, a downside of open relay technology is the proliferation of its usage by spammers looking to obscure or even hide the source of the large-volume e-mails they send. Open relay other known identity is third-party relay, spam relay or non-secure relay.

2.3 Spam Mitigation

To hamper spammer's potential of forwarding spam; their standing architecture needs to be disturbed, such as peer-to-peer, C&C and hosting servers by applying legal actions. This proposed work focuses on the hosting domains because it is not compulsory for the email recipient to search the source of a spam email in order to process the mail, but it is useful that the intruder has an original website where the buyer can purchase his product. If the buyer cannot make contact with the sale website, no transaction can happen. The target-of-sale websites are where intruders cause most of their profits.

Methodologies for originating the trend of spam are as differed as the methods to send spam. The most popular used anti-spam method is filtering, which usually identifies email based on its content; content-based filtering uses parameters such as an email message's headers or body of mail to recognize whether an email is probably a spam or not. This type of filters is incorporated by famous spam filters such as SpamAssassin [9], have been less successful till date at minimizing the quantity of spam that generally reaches a recipients inbox. On other side, content-based filtering has drawbacks [6]. System administrators and users must regularly modify their filtering protocols and use huge collection of spam for training; in reverse, spammers continue to arise with different ways of changing the bodies of an email to bypass these filters. The cost of bypassing content-based filters for them is negligible, since spammers can smoothly modify email body to bypass these filters. In opposite, modifying the network parameters of where spam is being sent through, and how it is being sent, is expansive. All the research that has done on making filters depend on email body/contents, focus has been directed to the network-level properties combined with spamming behavior. Additionally, content-based checks carried out, most mail filters, including SpamAssassin, also scan to find whether the sending IP address is included in a blacklist or not. Blacklists of known open relays, open proxies, spammers alive today and continue to exist one of the top-tier spam filtering techniques. There are many widely used blacklists are still in use; individual of these lists is not together maintained, and inclusion into these lists is depending on many types of parameters (e.g., sending mail to a spam trap, operating an open relay, etc.). The result of this paper is to give the domain name that the spammer is trying to promote and naturally this strong-supporting technique for filtering spam is likely to become much less efficient after some time.

3. PROPOSED WORK

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

3.1 File Upload

Spam email data are zipped because, individual .eml file will take more time to upload and also increase the load on system. This zip folder is uploaded to the system and unzipped for further process. The zip logic is used so that numerous numbers of emails can be efficiently uploaded without taking much time.

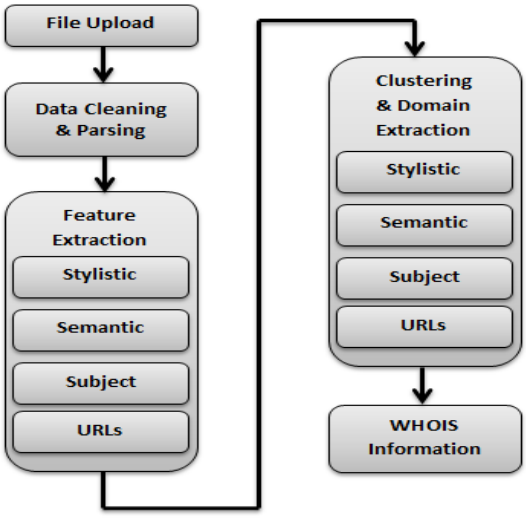

Fig.1 Flow of the approach to recognize spam domain. Note that "Fig.1" is abbreviated. There are sub sections of each step.

3.2 Data Cleaning and Parsing

In this step, the emails that had images, attachments and other than English language were filtered. At the same time, spam mails are parsed to yield following information viz. Unique ID, File_Name, Message_Id, Mail_Date, Delivery_Date, From_Mailer, In_Reply_to, Return_Path, References, Subject, and Content.

The initial dataset had approximately 3600 spam emails of all categories and languages. After data cleaning approximately 67% of the original data collected. Filtered data are divided into three different sizes and performed feature extraction and clustering. Data sets are 600 emails, which were considered as first dataset than 1200 and 2400 emails as the second and third dataset respectively.

3.3 Feature Extraction

The features that can help in clustering similar spam mails together are identified in this step. It considers four main features viz. stylistic, semantic, subject of mails and URLs present in the emails.

3.3.1 Stylistic Parameters

Stylistic parameters are based on the way in which the email is written (as emails produced from the identical botnet or composed by the identical spammer should have uniqueness in writing style). The main motive behind generating a spam campaign is to follow the users to buy a product or infect them with malwares. For this motive these emails constantly have a URL or email id inserted in the content and this can be used as a stylistic parameter of the email. Such distinct parameters of mail can be used to detect spam emails and group them in cluster. It considers some of stylistic parameters which 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 email ids used in the email, total count of URLs in the email body, types of different punctuations used in the email.

3.3.2 Semantic Parameters

Semantic parameters refer to the features that provide us understanding about the semantic and logical meaning of the emails. For this purpose, it uses the two classes of semantic parameters. First, the Trfr(Term frequency) and Indfr(Inverse Document frequency) for the top n most frequent words used in the dataset and second, the count of the top n bigrams used in the dataset, where n is the number that is decided based upon the cutoff of the minimum frequency count. Trfr & Indfr is a statistical measure that can be used to represent the importance of a term in a document [1]. It first remove all the stop words from the emails and then Trfr-Indfr can be calculated.

Equation 1, calculates the term frequency (trfr) of each term in a given document.

$$Trfr_{x,y} = n_{x,y} / \sum_i n_{i,y} \quad (1)$$

Where, trfrx,y is the term frequency of the term x in document y. ‘nx,y’ is the number of times the term x occurred in the document y and the ‘ni,,y’ is the sum of the number of occurrences of all the terms in the document y. ‘i’ in the above equation is any term in the document y.

Equation 2, calculates the Inverse documents frequency (Indfr) of each term in a given document.

$$Indfr_x = log|D| / |\{d : tr_x \in d\}| \quad (2)$$

Above equation provides the general significance of the term in the whole entity by performing mathematical division of the total number of documents by the number of documents containing the term. In Indfr equation, D is the total number of documents in the data set and the denominator is the number of documents d where a term trx appears. Finally, Trfr-Indfr is the product of the results obtained in Equation 1 and 2.

If it knows the most recurrent bigrams then it can do a better analysis of its content. Bigram is known as any two adjacent words that occur consecutively in a document. For the second class of semantic feature, it prepared an entity of top n most persistently used bigrams from the whole data set and then made a feature vector that keeps the count of those bigram occurrences for each of the documents. ‘n’ is the cut off value that is based on a pre-decided minimum frequency of the total bigrams.

3.3.3 Subject of Spam Emails

A subject may contain a sequence of non-blank characters known as tokens, which are separated by spaces [11]. The number of tokens will be defined as the subject length. In this section, string match score used on set of spam email subjects to identify related messages. To find subject similarity, it uses two approaches. First, subject matching score based on partial token matching, in which the similarity of subject x and y is computed as Kulczynski(x, y) [21], x and y are matched as two strings, where each token in x and y is treated like a character in a string.

$$Kulczynski(x, y) = (ILD(x, y)/|x| + ILD(x, y)/|y|) / 2$$

Where, ILD(Inverse Levenshtein Distance) [21] is resulting number of matches between tokens in subjects x and y. For example,

Token x1: F R E E _

Token y1: F _ E E L

There are three matching letters, therefore ILD(x1,y1) = 3.

Some subjects are longer than others containing more tokens. The chances of two long subjects matching each other, while yielding approximately the same match score. Therefore, in second approach [11], a coefficient is introduced to adjust the matching score based on subject length. The reason for insertion of coefficient is to decrease the credit given to small subjects that match each other.

$$MatchScore(x, y) = Co * Kulczynski(x, y)$$

Where $Co = sqrt(min((|x| + |y|) / (2 * maxlen), 1))$

From the above equation, it finds subject match score based on which clustering is performed. For example, look at the following two subjects:

Subject 1: Your Profile is Shortlisted

Subject 2: Get Your Profile Shortlisted

Therefore, when matching a pair of tokens, it allows token to partially match each other if they have the same length. In particular, if two tokens x and y have the same number of characters, say n characters: length(x) = length(y) = n, it defines match(x, y) = m/n where m is the number of matching characters. If characters ($x_1, x_2, ..., x_n$) in x and ($y_1, y_2, ..., y_n$) in y are similar then match score(x, y) = 1. Thus the matching score for the above example is 0.78, because ‘get’ is not matched with ‘is’.

3.4 URLs in Emails

Spam email normally has hyperlinks to websites where the actions can be taken for spammers to generate revenue. Mails with less number of links have highest possibility of being a spam mail in comparison to mails with links as well as other content. In this section, it extracts URLs from spam emails contents and also shows the respective mailer domains.

3.5 Clustering and Domain Extraction

The clusters are manually evaluated with the ground truth data that was collected. It performs four different sets of clustering with respect to above extracted features [21]. Simultaneously, domain names are fetched from all spam mails.

To construct cluster using stylistic and semantic parameter, K-means algorithm is used. K-means [18] is a clustering partition method which first creates an initial set of k partitions, where parameter k is no. of partitions to construct. It then uses iterative relocation technique that attempts to improve the partitioning by moving objects from one group to another. For subject [11] and URLs in mails, it uses seed selection algorithm to create clusters. In this algorithm, it selects an arbitrary subject, usually the first one and match it with rest of the subjects and group of subjects are formed based on similarity threshold and remove that group from the set of total subjects. For similarity purpose, it uses the 'MatchScore' defined in 3.3.3 section. This process is repeated until the subject set is empty and so, they are partitioned into clusters of different sizes. It presents four different clustering results for two algorithms on respective dataset.

3.6 WHOIS Information

In this section, the extracted domain name from above step is given to WHOIS (Domain registration information). WHOIS is a query and response protocol that is used for querying databases that store the registered users or assignees of an internet resource, such as a domain name, an IP address block, or an autonomous system and wider range of other information [12], [13], [14] and [15]. The law enforcement and anti-spam communities can take appropriate action against spammers once they have WHOIS information.

4. ANALTYTICAL RESULTS

The spam emails were collected from personal account of known people. To collect spam mails there are many open source tools available, it uses "MailStore Home" tool to collect spam emails. MailStore Home supports a long list of mail servers, replicating and backing up messages in a straightforward interface--complete with folder tree and reading pane. To show, the results of individual features are considered based on their purity percentage.

4.1 Purity

Purity is a simple and transparent evaluation measure. It is required to measure cluster quality. The purity percentage is evaluated by following equation [1]:

Purity (%) = sum of correctly clustered instances / total number of instances.

Purity is calculated in two different meaning mainly. First, the highest purity means highly pure cluster from total number of cluster for individual features and other is overall purity examines correctly clustered instances from total number of resulted cluster.

It presents the result for the cluster that gives the highest accuracy. The overall purity and highest purity of a cluster with respect to individual features are shown in figure 2, figure 3, figure 4 and figure 5.

Table 1: Stylistic Clustering Statics

	First Data set	Second Data Set	Third Data Set
Overall Purity (%)	58	65	62
Highest Purity (%)	66	80	72

Good results were obtained by stylistic clustering in the data set of 1200 because; the total no. of words in the emails of this dataset was low in content (i.e. where the total count of words was short).

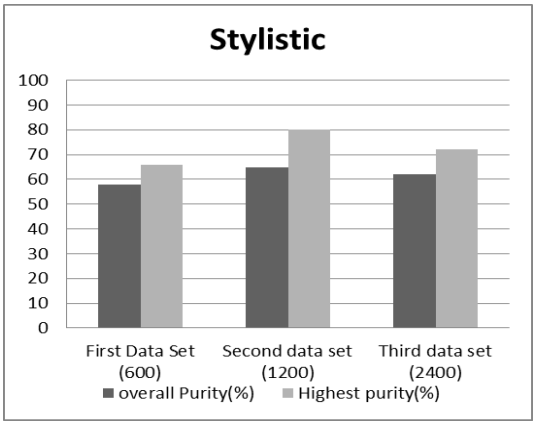

Fig.2 Graphs showing the 'Overall Purity' and 'Highest Purity of cluster' for Stylistic clustering

Most of the emails in 1200 dataset originally belonged to one cluster because of similar writing style. These reasons bring vast improvement in the overall accuracy of K-means algorithm in clustering that data set. The length of the emails also affects the type of the differentiating features.

Table 2: Semantic Clustering Statics

	First Data set	Second Data Set	Third Data Set
Overall Purity (%)	70	69	73
Highest Purity (%)	91	81	96

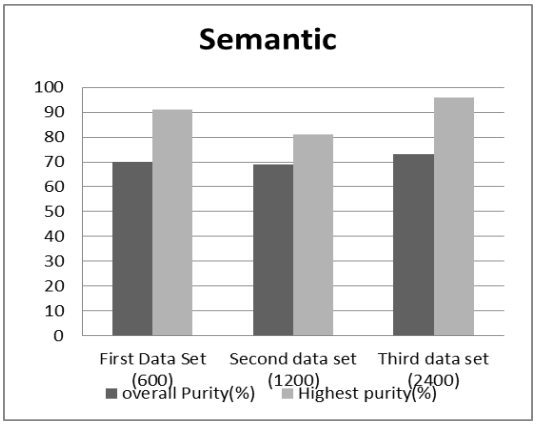

Fig.3 Graphs showing the 'Overall Purity' and 'Highest Purity of cluster' for Semantic clustering

Semantic clusters with high purity where extracted when the email body is rich in content. The data set of 2400 emails produces better results than the remaining ones because the emails in this data set were mostly rich in the text content hence; it was easier to extract bigrams from them.

For example, in 1200 emails dataset the stylistic clustering gave better results than semantic because many of the emails in that dataset were smaller in textual content and not enough to be distinguished by the semantic clustering approach.

Table 3: Subject-wise Clustering Statics

	First Data set	Second Data Set	Third Data Set
Overall Purity (%)	92	91	94
Highest Purity (%)	100	100	100

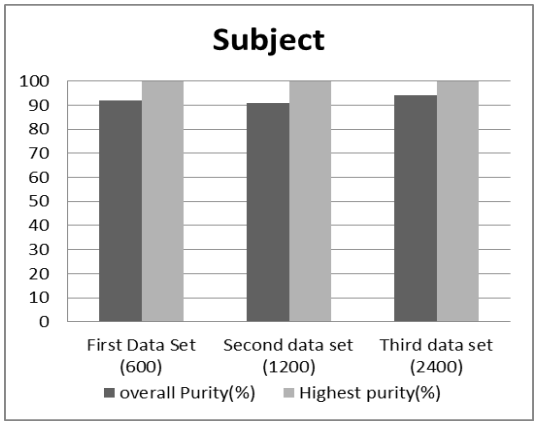

Fig.4 Graphs showing the 'Overall Purity' and 'Highest Purity of cluster' for Subject wise clustering

Subject and URL wise clustering gave us better results in all datasets when compared with rest of other two features. As shown in table 3, first datasets (i.e. 600) overall purity is increased by 34% compared to stylistic parameters of corresponding datasets whereas, when compared with semantic features it raised by only approximately to 22%.

Because, the Levenshtein distance, Kulczynski similarity of subject and seed selection strategy in the algorithm was able to pick up variations with the help of MatchScore. Experimental results showed the recursive seed clustering algorithm out-performed the simple algorithm when there were variations in the pattern. The simple algorithm regarded each variation as a separate cluster.

Table 4: URL-wise Clustering Statics

	First Data set	Second Data Set	Third Data Set
Overall Purity (%)	100	100	100
Highest Purity (%)	100	100	100

URL wise clustering gave excellent result when compared with all remaining three features. From table 4, it shown that extraction of URL from content is beneficial. As the dataset size increased, the purity also increased with it. Although less number of clusters based on URLs were obtained. Since, it has not considered the emails with images which contain large number of referencing URLs in it.

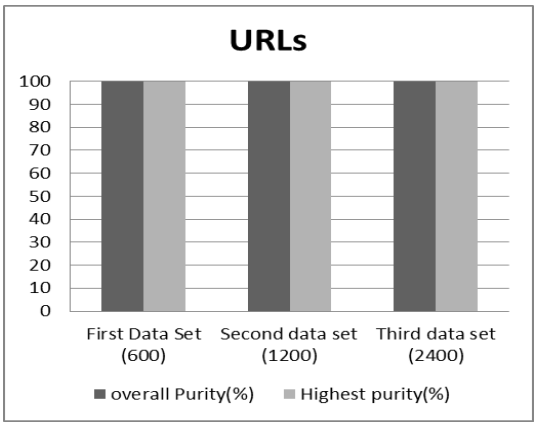

Fig.5 Graphs showing the 'Overall Purity' and 'Highest Purity of cluster' for URLs clustering

By studying domains instead of URLs in emails, it can effectively compress the data while not losing valuable information. URL wise clusters overall purity result is increased by more than nearly 35% when compared with stylistic and semantic feature results.

5. CONCLUSION AND FUTURE SCOPE

This research aims to recognize spam domain using data mining to assist the termination of spam emails. The research proposes an extraction of features and detecting significant spam domains from a large number of spam emails.

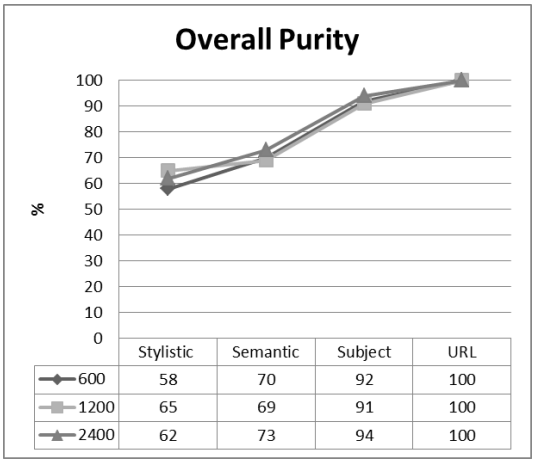

Fig.6 Overall Purity for all four features

From Figure 6 and 7, it concludes that clustering spam mails of overall purity with subject wise and URL wise are approximately 20-30% whereas, Highest purity are around 10-15% more when compared with stylistic and semantic features. The extraction of subject and URL from spam mails is increasing the efficiency of clustering by 10% and it also helpful to raises the productivity of cyber-crime investigator. The recognition of domain can improve the effectiveness of domain blacklist by detecting new spam domains. The identified domain is used by investigators as leads to trace the identities of spammers.

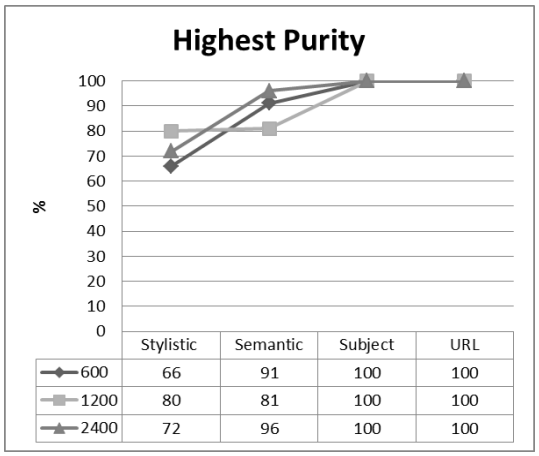

Fig.7 Highest Purity for all four features

In the future, it would like to parallelize the task of initial email parsing to receive real time spam feeds from major recipients and to use different features such as, images present in the emails, attachments, etc to identify clusters conforming different spam use cases. Also, it would like to experiment on feature analysis. It is imperative to analyze which features can give efficient result in this area because, the amount of data or spam emails is huge and the feature extraction and clustering can take a long time. It will be helpful to identify which of the features can be more essential and can use to save computational time. If it can find which set of features gives good results and can work on different kind of emails, it could help the computer forensics experts to get hold of the primary spammer. This proposed system can be used as complementary tool for existing anti-spam system to efficiently identify spammers.

6. ACKNOWLEDGMENTS

Hearty thanks to my guide Dr. R R Sedamkar for his support and guidance.

7. REFERENCES

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

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

[3] Kaspersky, Internet security Center, threats report statistics. http://usa.kaspersky.com/internet-security-center/threats/spamstatistics-report-q2-2013#.Uq6poM5P_rQ

[4] All Spammed up, Anti-spam in a business environment. http://www.allspammedup.com/2012/05/the-cost-of-spam-is-rising/

[5] F. Li, M. Hseieh, "An Empirical Study of Clustering Behavior of Spammers and Group Based Anti-Spam Strategies", In Proc. of the 3rd Conf. on Email and Anti-Spam, USA, 2006.

[6] Anirudh Ramachandran and Nick Feamster "Understanding the Network Level Behavior of Spammers", 2006, Georgia Tech.

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

[8] G. Warner A.P. Sprague and C. Wei "Clustering malware-generated spam emails with a novel fuzzy string matching algorithm", In Proc. of SAC '09.Honolulu, Hawaii, U.S.A.

[9] SpamAssassin, 2005. http://www.spamassassin.org/.

[10] C. Wei, A.P. Sprague, G. Warner and Anthony Skjellum "Mining Spam Email to Identify Common Origins for Forensic Application", SAC'08, March 16-20, 2008, Fortaleza, Ceará, Brazil. Copyright 2008 ACM 978-1-59593-753-7/08/0003

[11] C. Wei, A.P. Sprague, G. Warner and Anthony Skjellum "Identifying New Spam Domains by Hosting IPs: Improving Domain Blacklisting", Copyright 2006 ACM 238-7-59463-783-7/08/0007

[12] Spamhaus DBL. http://www.spamhaus.org/dbl/

[13] Dietrich, C. and Rossow, C. "Empirical research on IP blacklisting", ISSE 2008 Securing Electronic Business Processes, 163, 2009.

[14] SURBL. http://www.surbl.org

[15] URIBL. http://www.uribl.com

[16] Dietrich, C. and Rossow, C. "Spam, Domain Names and Registrars", MAAWG 12th General Meeting, San Francisco February 18th-20th, 2008.

[17] Project Honey Pot. http://www.projecthoneypot.org/.

[18] Wikipedia http://en.wikipedia.org/wiki/Cluster_analysis

[19] Calton Pu and Steve Webb. "Observed Trends in Spam Construction Techniques: A Case Study of Spam Evolution". CEAS 2006 Third Conference on Email and AntiSpam, July 2728, 2006, Mountain View, California USA.

[20] P. Tan, M. Steinbach, V. Kumar, Introduction to DataMining, (First Edition), Addison-Wesley Longman Publishing Co., USA, 2005, pp 496-515.

[21] Chun Wei, Clustering Spam Domains and Hosts: Anti-Spam Forensics with Data Mining, Dissertation, 2010.

[22] Jeet Morparia, "Peer-to-Peer Botnets: Analysis and Detection" 2008.