I.J. Information Technology and Computer Science, **2012, 8, 10-21**
Published Online July 2012 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2012.08.02

Trends, Issues and Challenges Concerning Spam Mails

Jitendra Nath Shrivastava
Deptt. of Computer Science & Engineering, Invertis University, Bareilly, India
& Research Scholar of Singhania University, Rajasthan, India
E-mail: jitendranathshrivastava@yahoo.com

Maringanti Hima Bindu
Deptt. of Information Technology, Jaypee Institute of Information Technology, India
E-mail:mhimabindu@yahoo.com, hima.bindu@jiit.ac.in

Abstract—Traditional correspondence system has now been replaced by internet, which has now become indispensable in everyone's life. With the advent of the internet, majority of people correspond through emails several times in a day. However, as internet has evolved, email is being exploited by spammers so as to disturb the recipients'. The entire internet community pays the price, every time there pops a spam mail. Online privacy of the users is compromised when spam disturbs a network by crashing mail servers and filling up hard disks. Servers classified as spam sites are forfeited from sending mails to the recipients'. This paper gives the broader view of spam, issues challenges and statistical losses occurred on account of spams.

Index Terms— Spam, Trojan horse, Botnet, ANN, HAM and Modem

I. Introduction

The internet, network of networks, makes communication very easy for two people on opposite sides of the world via e-mails. However, popularity of the email is affected by the spam mails in the e-mailbox. As most of the internet users are, in fact, inexperienced and they do not understand the challenges of spam, they are easily affected by Spammers.

Understated are the issues created by spam mails: [1]

- Spam reaches to client's inbox without his/her consent.
- Spam irritates internet users.
- Clients switch over ISP's continually looking for reliable email delivery.
- Users are less aware about spam.
- Spam badly affects internet performance and bandwidth.
- Millions of computers are compromised.
- Billions of dollars are lost globally.
- Identity theft.
- Increase in worms and Trojan Horses.
- Spam can crash mail servers and fill up hard drives.

In this paper, we have presented a comprehensive study regarding, spam, its classification, statistical analysis, spams filtering techniques, and future needs to deal with spams.

The paper is organized as follows, the spam and its classification along-with the spam statics in discussed in section 2. Section 3 describes the spam transferring methods. Role of botnet is discussed in section 4. The statistical figures regarding the spams are presented in section 5. Spam filtering techniques are discussed in section 6. The conclusions, of the paper are presented in section 7 of the paper.

II. SPAM and Its Classification

The term spam also refers to "Sending the same message to the large group of individuals in an attempt to compelling the message onto people who are un-willing to receive such messages." Receiving spam is a very common grudge of internet users as individuals attack users' email accounts through spam email. However, as spam is on the rise, various internet users still have partial knowledge as to what constitutes spam and what a spam email looks like. Every internet user is conversant with the word 'spam' as he gets it in his inbox approximately daily. The word 'Spam' is an acronym derived from the words 'spiced' and 'ham'. Year 1993 saw the term as 'Unsolicited or undesired bulk electronic messages. It follows, that Richard Dephew, the administrator of the world-wide distributed internet discussion system Usenet, wrote a program which mistakenly caused the release of dozens of recursive messages in the news, admin. policy newsgroup. The recipient's straightway found an appropriate name for these obtrusive messages – spam.

The fig.1 illustrates the spam communication system:

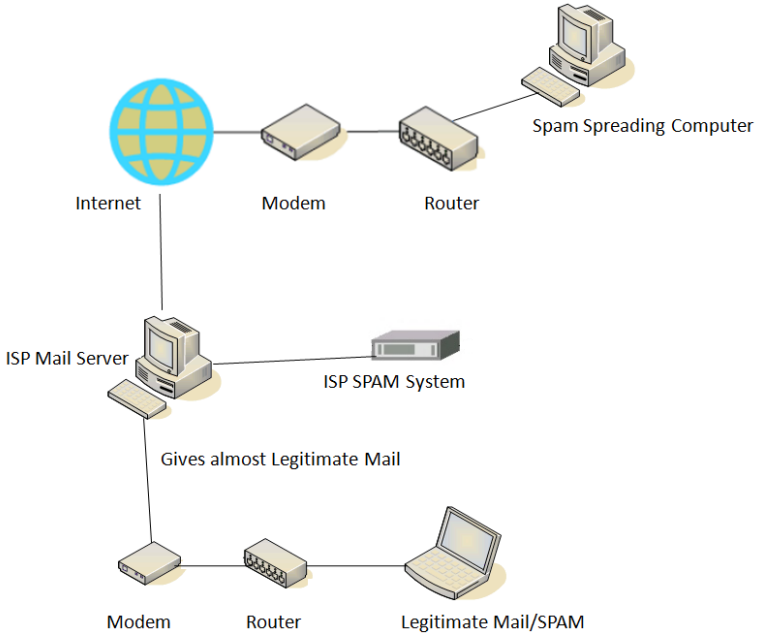

Figure 1:- Spam Communication System

Spams are of different styles and complexities; hence it is difficult to classify them. As some spam is a plain text with a URL; some are cluttered with images and attachments; some arrive with very little text and maybe only a URL. Also it arrives in various languages apart from English.

It is first required to look over the email headers to understand the language in which an email is composed. In January 2009, 96% of total spam was in English, but this has fallen very gradually over the year. Approximately, 10% of spam sent is currently in local languages [2].

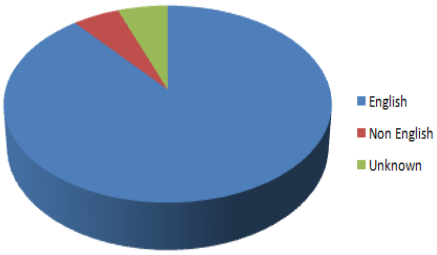

Figure 2: - Spam Compositions (%)

As the fig. 2 shows, the remaining of the non-English spam is in various languages. Since, January 2010, the second most accepted language in spam was Dutch. Spam in languages other than English is rising overall because the volumes are reasonably low when compared with English-language spam [2]. The fig. 3 shows the contribution of non-English language spam.

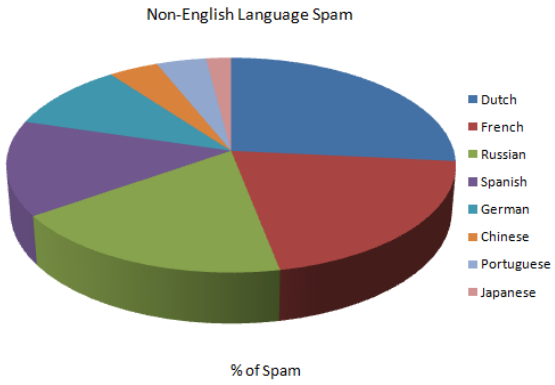

Figure 3: - Contribution of Spam by Non-English Language at Different Countries (%)

Brazil was the only nation observed where the most common language is neither “unknown” nor English. Roughly, 33% of spam sent to Brazilian recipients was in Portuguese. Brazil was one of the lowest percentages of English language spam at 25.6% [2]. The fig. 4 illustrates the percentage of spam in local languages.

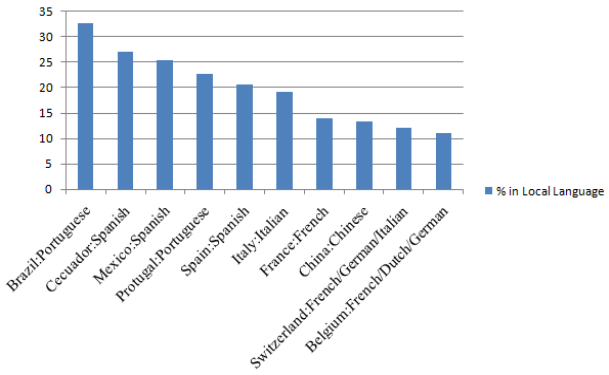

Figure 4: - Percentage of spam in local languages

Portuguese and Spanish are the most preferred languages than English for sending spam. Furthermore, it is observed that the proportion of all English spam is decreasing worldwide. However, spam in languages other than English is targeted to those countries that are also non-English in nature. For example, sending, Japanese spam to a primarily English, or German speaking country would be a waste of time.

III. Spam Transferring Methods

Practically, every email user receives few useless bulk mails regularly and is no way to protect an email from becoming a spam. Some of the important methods are discussed below:

3.1 Using URL Shortening Services

Sudden upsurge of social networking and micro-blogging services made URL shortening services more popular. The spammers' exploit of URLs from link shortening services became more popular during 2010. On July 28, 2009, 9.3% of spam comprised some sort of shortened hyperlink. On April 30, 2010, this highest figure almost doubled to 18.0% of spam, the present historical peak. Since mid to late August 2010, at least 1% of spam per day contained a shortened URL. For September 2010, the percentage of spam that contained a shortened URL reached 3% of spam for the month and to date; this figure has been tracking at about 2% of all spam.

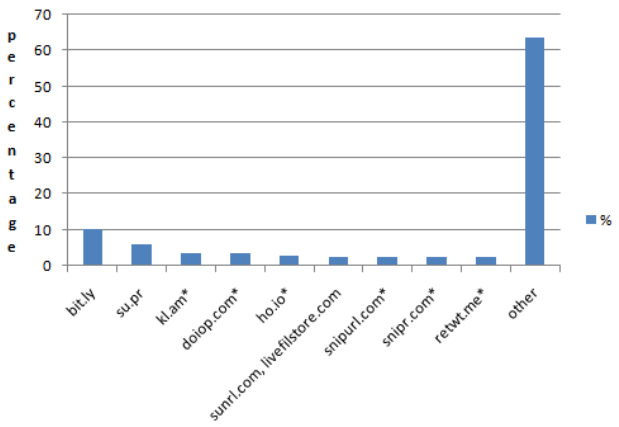

Figure 5: - Top short URL services used in spam (2010)

On average, 91% of spam included some kind of URL in 2009 and in 2010, which was roughly unchanged from 91.1%. An average 0.33% of all spam contained a short URL in 2009, and in 2010, this figure rose to 1.38% with an average of 1 in 66.1 of all URLs in spam being shortened in 2010 [2].

3.2 By considering Spam Message Size

Spam mails were made precise with a smaller content, enabling spammers to shoot several more mails, making more possible losses. Small file sizes are obtained by having small and simple emails that may contain a single line of text and a link to a web page. During the year 2010, roughly 72% of all spam was below 5kb in size. Regardless of this we can see some clear increases in the average size during April 2010 to August 2010. The increase in the average size was because of a long run of HTML format emails together with few attached images being sent by both the Rustock and Cutwail botnets.

IV. Role of Botnets in Spreading SPAM

A botnet Trojan is used to build new botnets. However, several, but not all botnets are considered to spread spam. For example, Zeus botnet is introduced to make financial fraud; this botnet is not at all used to send spam. Botnets are usually responsible for 80-90% of all spam sent worldwide. The overall average of sending spam from botnet in 2010 was 88.2%. The top three botnets have not changed in the latter half of 2010. Rustock remains the most dominant botnet; Grum being the second and Cutwail, the third largest. 2010 witnessed a large increase in spam emails from botnet infected machines. Cutwail was the most responsible botnet for sending massive volumes of Bredolab Trojan-infected emails throughout August 2010. Grum had also been sending a variety of malware-infected emails throughout the year [2].

Increased spam in India in 2010, made it the largest single source of spam from one country, at 8.5% of global botnet spam. Spam from the Russian federation rose to approximately 17% of global botnet spam at the end of 2010. Botnets have also been successful in recruiting new bots to their operation because of the emergence of high-speed broadband connection. This has opened the opportunity to infect new machines by cyber criminals. The following chart represents the percentage of spams, which are sent globally from various botnet at the end of 2010 [2].

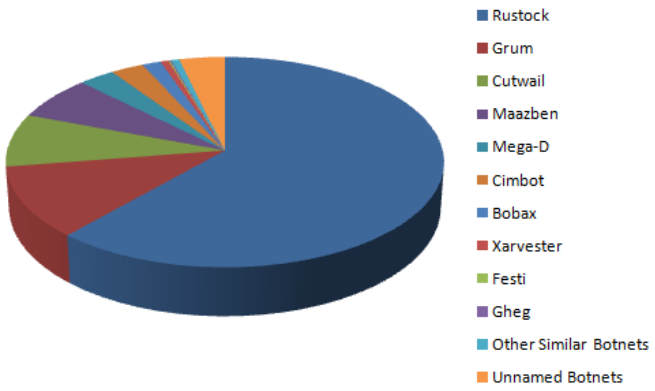

Figure 6: Spam sent from various botnets (2010)

The brief listing of important botnets during the year 2010 are mentioned below:

Rustock

As shown in the diagram, Rustock remained the most dominant botnet and had peaked up just over 80% of all spam, in mid-August 2010. The US remained the main source of infection for Rustock. It was the second and third largest source of infection in Brazil and India respectively.

Grum

The Grum botnet ranked second in the list of the most active spam-sending botnets and was accountable for approximately 9% of botnet spam.

Cutwail

It was the largest source of spam emails containing the Bredolab Trojan, ranked the third position and was responsible for approximately 6% of global spam.

Maazben

It had moved up to fourth position, responsible for over 5.2% of global spam.

Mega-D

By 2009, the Mega-D botnet disappeared from the spam-sending landscape for many days. However, it returned much strongly with a larger number of brand new IP addresses from which it was sending spam. It was responsible for almost 15.7% of global spam.

Storm

This botnet was responsible for 11.8% of all the spam containing shortened hyperlinks.

Lethic

This was in hibernation during September 2010 and the first half of October 2010. It was responsible for as much as 3% of all spam.

The figure 7 shows the infection percentage of important botnets in India:-

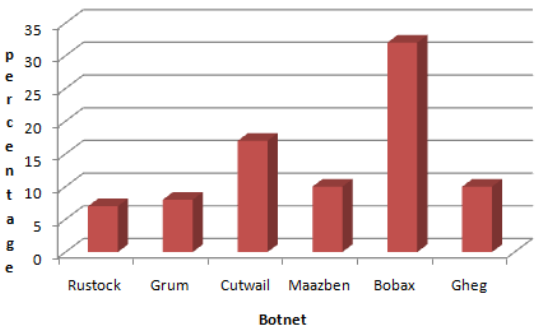

Figure 7: Infection percentage by botnets in India

V. SPAM: The Statistical Figures

Use of the internet has become prevalent among millions of users on account of rapid growth of broadband. However, still these users are facing different types of threats because of their little awareness about the computer security, and new users are quickly becoming infected with malware as their computers are subjected to botnets. Some important facts about emails are stated below [1]:

Table 1: Status of email accounts

Email accounts	By the year 2011	Expected in 2015
Number of email accounts	3.1 billion	4.1 Billion
Number of wireless email users	531 million	1.2 billion

Table 2: Percentage of email users globally

Worldwide email users	
Country/Area	Percentage
Asia/Pacific	47%
Europe	23%
North America	14%
Rest of the World	16%

Consumer email accounts have a substantial portion of worldwide emails. As a matter of fact, during 2011, consumer email accounts held 75% of worldwide mailboxes, while the percentage of corporate email accounts was 25% of worldwide mailboxes. These consumer email accounts are usually given by ISPs, Portals and a variety of hosting site providers free of charge. Over the next four years, it is presumed that the corporate email accounts will grow faster than the consumer email accounts owing to the increase of reasonable cloud-based email services and that the typical corporate email user will send and receive approximately 105 emails daily [1].

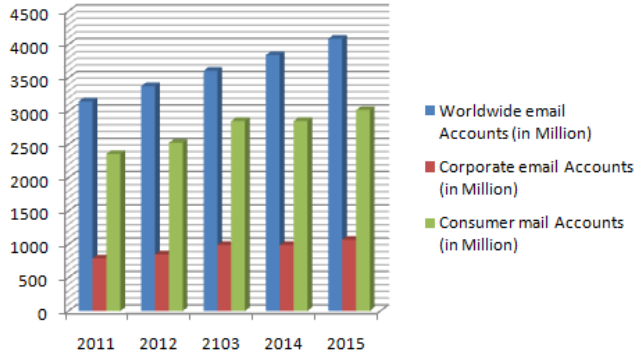

Figure 8:- Corporate Vs. Consumer email accounts, 2011-2015

Roughly, 19% of total emails are considered as spam despite having spam filters. While spam is an annoyance for users, it is a considerable expense for corporations. According to projections, a typical 1, 000-user organization can spend upwards of $3.0 million a year to fight and manage spam [1].

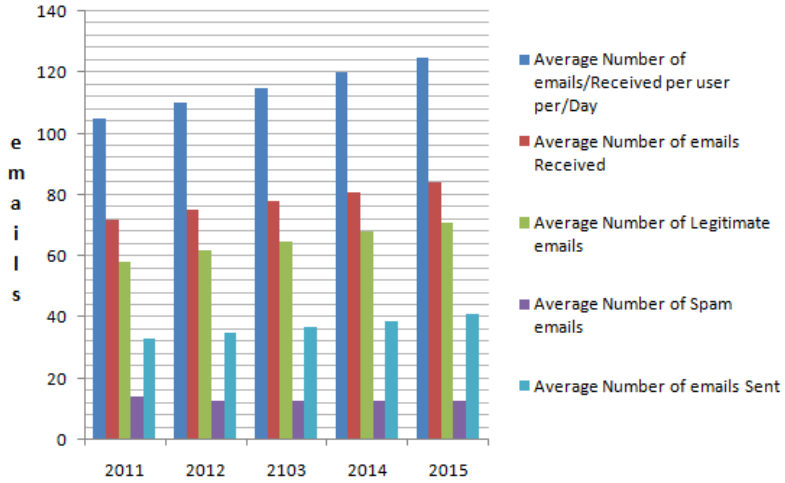

Figure 9: Expected corporate email sent and received (per user per day from 2011-2015)

In the past, there were many attempts to disrupt botnet activities; still, there are approximately five million spam-sending botnets worldwide. However, the average number of spam emails sent from each bot fell down from approximately 85 emails per bot per minute in 2009 to approximately 77 spam emails per bot per minute at the end of 2010 [2].

The table 3 states the important facts about the spam and other threats.

Table 3:- Threats in 2010 and 2011

Threats	Year 2010	Year 2009
Global Spam Rate	89.10%	87.70%
Average rate of malware in email traffic	1 in 284.2 emails	1 in 286 emails
Malware blocking with containing a malicious link (within the body of the message)	23.70%	15.1
Average ratio of email traffic blocked as phishing attacks	1 in 444.5	1 in 325.2
Average number of web sites blocked as malicious	3,188	2,465

During the year 2010, spammers produced several spam campaigns pertaining to major exciting events like the FIFA World Cup 2010. The spammers' exploitation of URLs became more and more spread during the year 2010, especially on April 30th, when roughly 18.0% of spam that day contained a shortened URL. Almost about 188.6 million phishing emails were blocked by Skeptic™ in 2010. Roughly, 95.1 billion phishing emails were predicted to be in flow during 2010. The most frequently spoofed phishing organization was an international bank, responsible for 14.9% of phishing attacks, blocked in 2010 [2].

Even legitimate applications may be vulnerable to being exploited by cyber criminals where vulnerabilities may exist in the web site. Some of the important facts related to few countries are mentioned below:

Table 4:- Important Facts

Year 2010					
Country	China	Nigeria	India	US	Brazil
Population of world population	19%	2%	17%	4.50%	2.9
Internet Pentration	32%	29%	7%	77%	
Sending worldwide botnets of total botnet spam	0.33%	0.03%	8.50%	8.70%	5.60%
Active bots	17000		487,000	368,000	318,000
Broadband users are the part of botnet	1 in 23,700	1 in 24,000	1 in 250	1 in 653	1 in 239
No. of botnet spam emails were sent per broadband user each day	2	NA	320	110	223

Europe has always been a major source of spam accounting for roughly 30% of worldwide spam. In 2009, it sent equal volumes of spam as Asia and South America did. However, in 2010, MessageLabs Intelligence observed that spam from South America had decreased while Asia continued much as before[2].

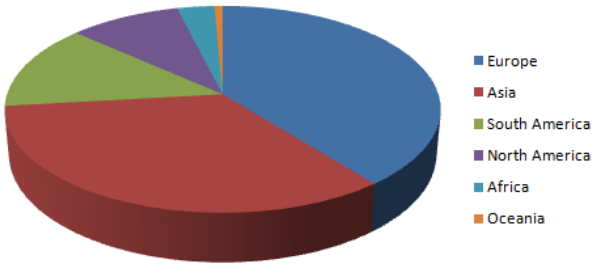

Figure 10:- Percentage of global spam continent wise (2010)

Spam from webmail services is not as common as it once was. Only 0.7% of the total spams during the year 2010 were sent from a webmail account thereby making webmail services an uncommon medium of spam. As almost all spam is now sent from botnet-infected computers. The characteristic feature of botnets is that they do not normally send much spam from genuine webmail accounts. It is observed that 89% of spam, which is sent from webmail accounts, does originate from botnets.

VI. SPAM Filtering Techniques

The various spam filtering techniques are adopted to get rid of the problem of spam. However, each scheme has its advantages and disadvantages, and in a nut shell, none of them is very effective. In the next section, various basic techniques are discussed as an overview.

6.1 Distributed adaptive blacklists

This approach blocks the emails coming from blacklisted servers as they are considered to be spam. These servers owing to their vulnerability are blacklisted in advance. Furthermore, the emails coming from blacklisted servers are deleted at the server level. The blacklist can also be maintained at the personal level.

Advantages:

The blacklist approach is beneficial when servers are compromised and used for sending spam to hundreds of thousands of users. This makes the method better and comparatively cost-effective to use at the ISP level along with some other filtering technique. Tools like Razor and Pyzor can be used for this purpose.

Disadvantages:

The criterion of any spam filter is not only efficiency in filtering spam but also doing the job with the minimum amount of 'false positives'. Marking a legitimate message as spam is a greater mistake than marking a spam as legitimate. The blacklist approach generates a large amount of false positives and hence the idea of barring a culprit server forever is not a good idea. A legitimate message arriving from a blacklisted server would always be considered a spam. MAPS RBL, probably the best-known blacklist, catches only 24 percent of all spam with a 34 percent of false positives. Moreover, there are many ethical issues involved in blacklisting a server, the worst scenario being that of blacklisting a server without knowing whether that server is a source of spam or not. Furthermore, a spammer is a moving target. While a spammer might use a compromised computer to send spam, as soon as he learns his computer is being detected, he can use a different computer until that one also gets detected. This can go on and on. The result is that while servers are shunned, the spammer still keeps spamming.

The solution to this approach has been the use of Distributed Adaptive Blacklists. Its basic working is to detect a spam message and inform all the recipients (which may run into millions) of that message about its status. Digests of spam are maintained at the server level. So, whenever a new message is received at the MTA, adaptive blacklists are called to detect whether the message is spam.

There are tools that ensure that the emails, which are different versions of the same spam, do not get identified as legitimate. In addition, maintainers of distributed blacklists create "honey-pot" addresses that are never used for legitimate purposes. The basic disadvantage of this approach is that it generates a considerable amount of false negatives. Thus, it is recommended that this approach be used in conjunction with another effective filtering technique.

This technique is implemented at the mail server. When a message is received by a message transfer agent (MTA), a distributed blacklist filter is called to determine whether the message is a known spam or not. These tools use clever statistical techniques for creating digests. Tools like Razor and Pyzor operate around servers that store digests of known spams [3].

6.2 Rule Based Filtering

As evident from the name, in a rule-based approach, each email is compared with a set of rules to determine whether it is a spam or not. A rule set contains rules with various weights assigned to each rule. Initially, each incoming email message has a score of zero. The email is, then, parsed to detect the presence of any rule, if it exists. If the rule is found in the message, its weight is added to the final score of the email. In the end, if the final score is found to be above some threshold value, the email is declared as spam [4].

Advantages:

This approach can be very effective with a given set of rules. It can achieve 90 to 95 percent efficiency. The filter is easy to install, as it merely requires copying the rule set without any training nor any sort of personal tuning. Furthermore, the rule set can be updated by copying an additional set of rules to challenge the current trend of spam.

Disadvantages:

The rigidity of the rule-based approach favors its biggest disadvantage. The spam filter is not intelligent as there is no self-learning facility available in the filter. Spammer if versed in the knowledge of the rule set can design a spam to deceive the method. For example, if there is a rule for classifying a message as a spam and the message contains the word "Viagra" more than five times, the spammer can easily circumvent the rule by using the term "V*i*a*g*r*a" instead of "Viagra." As a matter of fact, rules cannot be kept secret. The best option is to go through every spam and update the rule set by manually adding newly discovered rules. Unfortunately, this updating process is never ending, as the spammers continually devise new procedures to deceive the spam filters. This process requires personal effort, time, and some level of expertise, qualities that are absent in every email user.

The rule-based approach could be used in an integrated spam filter in combination with some other approach. In a rule-based approach, decisions as to whether to classify an email as spam or not are binary in nature. This classification process on its own does not give continuous confidence. Such confidence is critical because the cost of a false-positive classification (classifying the legitimate message as spam) is very high. Owing to the above reason, there is a need for a classification scheme based on probability, wherein all messages near a threshold value can be categorized as legitimate to avoid the danger of being 'false positives'. As far as the computational speed is concerned, the rule-based approach is faster than the use of blacklists, but it is slower than statistically-based approaches. 'Spam Assassin' is the most successful spam filtering tool available on the market that uses this approach.

The ReadMe file of 'Spam Assassin' states that it does between spam and non-spam correctly in 99.94 percent of the cases.

Patterns, mostly regular expressions are matched against a candidate message. Some matched patterns add to a message's score, while others subtract from it. If a message's score exceeds a certain threshold, it is filtered as spam; otherwise it is considered as legitimate. In certain cases, certain ranking rules are fairly constant over time. On the other hand, other rules need to be updated as the spam and other products evolve over a period of time.

6.3 Bayesian Classifier

Particular words have particular probabilities of occurring in spam email and in legitimate emails [5]. The filter must be trained in advance for these probabilities. After training the 'word probabilities' (also known as 'likelihood functions'), they in turn are used to compute the probability that an email with a particular set of words in it belongs to either of the category. Each word in the email contributes to the e-mail's 'spam probability', or only the most interesting words, may do so. This contribution is called the posterior probability and is computed using Bayes' theorem. Then, the e-mail's 'spam probability 'is computed over-all words in the email, and if the total percentage exceeds a certain threshold (say 95%), the filter marks the email as a spam. Some spam filters combine the results of both Bayesian spam filtering and other heuristics (pre-defined rules about the contents, looking at the message's envelope, etc.), resulting in even higher filtering accuracy, sometimes at the cost of adaptive-ness. Server-side email filters, such as DSPAM, Spam Assassin[6], Spam ayes[7], Bogofilter and ASSP, make use of Bayesian spam filtering techniques[8].

Let $C = c_1 ... c_m$ be m document classes. Given a new unlabelled document D and its corresponding word-list $W = w_1 ... w_d$ (defined in the same way as the wordlist for the training set), the naive Bayes approach assigns D to a class c^*_{Nb} as follows:

$$c^*_{NB} = \arg\max_{c_j \in C} P(c_j) \prod_{i=1}^{d^r} P(w_i \mid c_j)$$

Where $P(c_j)$ is the a priori probability of class c_j and $P(w_i \mid c_j)$ is the conditional probability of word w_i given class c_j. The underlying assumption of the naive Bayes approach is that for a given class c_j, the probabilities of words occurring in a document are independent of each other.

When the size of the training set is small, the relative frequency estimates of probabilities $P(w_i \mid c_j)$ will not be reasonable; if a word never appears in the given training data, its relative frequency estimate will be zero. Hence, the accuracy of the techniques depends on the type and size of the datasets.

6.4 K-means

The k-means formulation assumes that the clusters are defined by the distance of the points to their class centers only [9]. In other words, the goal of clustering is to find those k mean vectors $(c_1, ... c_k)$ and provide the

cluster assignment $y_i \in (1,...,k)$ of each point x_i in the set. The K-means algorithm is based on an interleaving approach where the cluster assignments y_i are established given the centers and the centers are computed given the assignments. The optimization criterion is as follows:

$$\min y_1,...y_m, c_1,..c_k \sum_{j=1}^{k} \sum_{y_i=j} \left\| x_i - c_j \right\|^2 \quad (1)$$

Assume that $(c_1,...c_k)$ are given form the previous iteration, then,

$$y_i = \arg\min_j \left\| x_i - c_j \right\|^2 \quad (2)$$

and next assuming that $y_1 \cdots y_m$ Cluster assignment are given, then for any set $S \subseteq \{1,...,m\}$.

We have that

$$\frac{1}{|S|} \sum_{j \in S} x_j = \arg\min_c \sum_{j \in S} \left\| x_i - c_j \right\|^2 \quad (3)$$

In other words, given the estimated centers in the current round, the new assignments are computed by the closest center to each point x_i, and then given the updated assignments the new centers are estimated by taking the mean of each cluster. Since each step is guaranteed to reduce the optimization energy, the process must converge to some local optimum.

6.5 K Nearest Neighbors

If at least t messages in k neighbors of the message m are unsolicited, then m is unsolicited email, otherwise, it is legitimate.

The nearest neighbour decision rule assigns the new unlabelled document D to the document class c_j if the training pattern closest to D is from class c_j. We use the TF–IDF (TF is the term frequency in a document and IDF is the inverse document frequency) weighting scheme and use the cosine similarity [10] instead of Euclidean distance to measure the similarity of the two documents. Given two documents D1 and D2, their corresponding weighted feature vectors are

$T_1 = (t_{1i}\delta_i 1)_{i=1}^{d}$ and $T_2 = (t_{2i}\delta_i 2)_{i=1}^{d}$, where δ_{ki} is the weight of word w_i in document k (TF-IDF). The similarity between D1 and D2 is then defined as:

$$S(D_1, D_2) = \frac{T_1^T T_2}{\|T_1\| . \|T_2\|} \quad (4),$$

where $\| \ \|$ denotes the norm of the vector.

6.6 Support Vector Machine (SVM)

'Support Vector Machines' [11][12] is based on the concept of decision planes that define decision boundaries. A decision plane is one that distinguishes among a set of objects having different class memberships. A schematic example is shown in the illustration below. In this example, the objects belong either to class GREEN or RED. The separating line defines a boundary on the right side of which all objects are GREEN and to the left of which all objects are RED. Any new object (white circle) falling to the right is labeled, i.e., classified, as GREEN (or classified as RED should it fall to the left of the separating line).

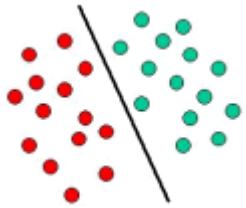

Figure 11:- Support Vector Machine

6.7 Content Based Spam Filtering Techniques

Neural networks are the best candidates for problems of classification [13][14][15][16][17]. Without being spread out over the model, we will retain in what follows the characteristics which contribute to the design of an anti-spam filter. To follow, if one makes a point of applying the technique of the perceptron, it is enough to choose a characteristic vector larger than that of the training sample to ensure the convergence. However, doing so will have a toll on the computation.

6.8 The Multi-Layer Networks

As implied in the name, the multi-layer neural net is a network of connected perceptrons which form a network with successive layers. The outputs of each perceptron are inputs of perceptrons of the following layer. The inputs of the neurons of the first layer are the components of the characteristic vector, while the outputs of the last layer are the results of the classification. The layers between the first and the last are called hidden layers. The function of each neuron is somewhat different from the simple perceptron, although the training is also made in an iterative way as the simple perceptron. The output function is:

$y = \phi\left(\sum_{i=1}^{k} w_i x_i + b \right)$, ϕ is a non-linear function such as $\frac{1}{1+e^{\alpha x}}$ or $\tanh x$

Figure 12 is the graphical representation of a multi-layer neural network. To train a neural network is to readjust the weights and biases in such a manner so as to minimize the sum of the errors of the output.

$$E(f) = \sum_{i=1}^{n} |f(x_i) - c_i|^2 \quad (5)$$

The tuning of these parameters is described in detail in [16][18].

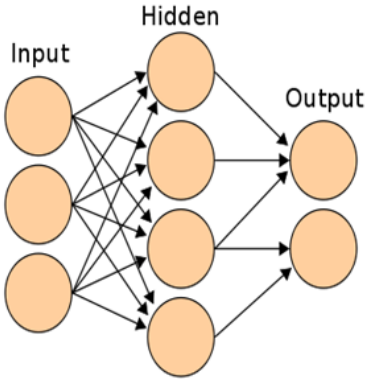

Figure 12:- ANN

6.9 Technique of Search Engines

When it acts on text e-mails, classification techniques of text seem to be efficient. However, spammers leave no table unturned to invent tricks to circumvent filters. One of these tricks is to include in the body of the message only the hyperlink to a Web page which contains the advertising text. This poses problem for web content classification. A proposed technique to overcome this kind of spam is to use the public search engines which offer a means to classify the websites [19][20][21]. The principle of this technique is to analyze automatically the contents of the pages referred by the links sent in the messages, likely to be spam.

6.10 Technique of Genetic Engineering

In the design of a Bayesian filter, the characteristic vector may include the frequencies of some words generally selected by human experts. As a matter of fact, this construction is sometimes decisive in the performances of the filter. Reference [22] underneath mentions Hooman proposing a method to build automatically the Bayesian filter. This method lays its foundation on the genetic programming. Thus, the frequencies of a word occurring in E-mail can debate the classification of the message as undisputed. As genetic programming, by Koza [23][24][25][26][27] the filter is represented by a syntactic tree where nodes are numbers that represent the frequencies, operations on numbers, words and operations on words. A syntactic tree of a filter should be built according to a precise syntax. Syntactic rules then can be used to check the correctness of the tree by checking whether we are able to reduce the tree to some number.

Prototype:

As depicted in Fig. 13, the system comprises of 2 major processes. The input e-mail then undergoes the process of keyword extraction. Analogous to the domain of genetics, the process of genetic algorithm then produces a chromosome representing the e-mail from the words as extracted above. The evaluation mechanism then creates spam mail prototypes as the 'end product 'of the system.

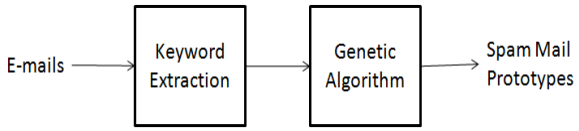

Figure 13 System architecture

'Corpus is the term quoted for the collection of emails considered as 'Spam'. Spam mails forming corpus are encoded to chromosomes and undergo the genetic operations, that of 'crossover' and 'mutation' and are then evaluated by a 'fitness function'. To follow, as a result of genetic algorithm, mail prototypes are obtained. Figure 14 shows a flowchart of spam mail prototypes construction.

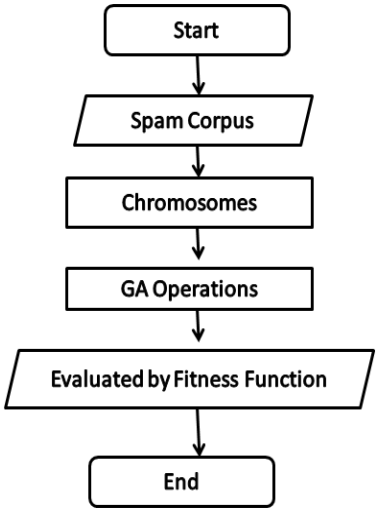

Figure 14 Flow chart

Genetic Operations

Crossover

In general, the crossover is allowed for bits of gene within the same group only. Multi-point crossover enables us to randomly select the position to cross. In each generation, approximately 15 percent of chromosomes are crossed.

Mutation

Mutation guarantees to preserve some data. Mutation is done by changing random bit position. To approximate, each generation at least mutates 2 percent of the chromosomes.

Evaluation

As mentioned in prior, after emails from corpus had been encoded to chromosomes and underwent the operations of genetic algorithm, they become 'argument' of the fitness function. The fitness value obtained and used for ranking spam prototypes can be computed from Eqn. as under:

$$FF = \sum_{i=1}^{n} \frac{\text{No.of Keyword } i \times W_i}{\text{Total keyword in an email(n)}}$$

Where the training weight (w_i) is the summation of weight of any word (w_i) found in each spam mail divided by total e-mails in corpus which we use for training and w_i is calculated by count number of any word i in each e-mail and divided by total words in that e-mail.

Selection

After all chromosomes are 'through' the fitness function, the system selects appropriate chromosomes for filtering incoming e-mails. Roulette wheel technique is used as a selection method.

Rules set for classifying e-mails

The weight of words of gene in test mail and the weight of words of gene in spam mail prototype are compared to find the matching gene. The proposed system assigns one spam score point to the spam mail prototype on the condition that the number of matched gene is greater than or equal to 3. The aforesaid classification of spam mails is shown in Figure15. The comparison finally yields the sum of spam score points f all prototypes. To follow, if the percentage of spam score point is greater than the percentage of threshold, then this test mails falls under the category of spam mails. Normally in the experiments, we set the threshold value at 30% so that this threshold value can also be manually adjusted to the appropriate value for optimal result.

6.11 Adaptive Techniques

In Static data methods the data is first passed to train the model. The encoded message then contains the trained model, followed by the data which is encoded using this model. The decoder in the first place reads the model and then uses it to decode the remaining data. The advantage of adaptive methods is that they require only a single pass over the data. However, it is very important that system should be intelligent enough to catch the spam. As the type and classifications of the spam change very frequently, the techniques implied should have a capability to adhere with the change. The ANN based technique has the potential to filter the spam, but due to the Euclidean distance criterion this method is not very effective. However, this method can be improved significantly, by taking into account, other distance criterion.

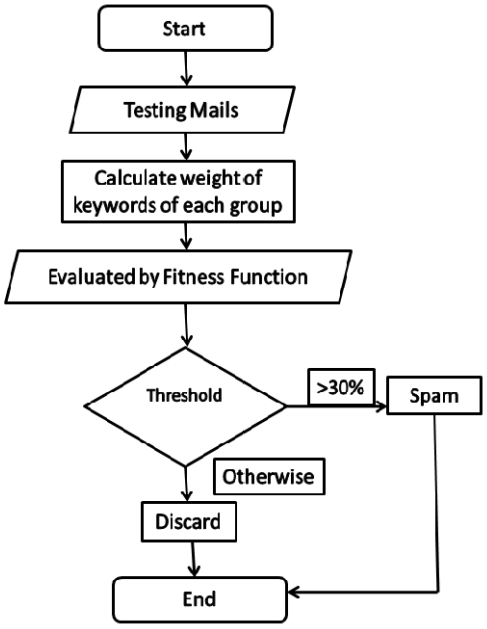

Figure 15 Flow charts (selection)

Limitations:

A number of anti-spam procedures are presently working to differentiate spam from legitimate e-mails. However spammers and phishers continuously strive to use dynamic spam structures to 'adulterate' email content to circumvent these procedures. Apart from other technological procedures, different adaptive filters have been built that are capable to permit an algorithm to constantly monitor the kind of e-mail's or their content a recipient would usually process and what to observe in normal course of its business. These filters are comprised of the complex statistical techniques that categorize future e-mails based on the word content of accepted e-mails.

Owing to the fact that several emails can be classified as spam and sent to the junk mail folder, it is necessary that a manual search be made of both the inbox and spam folders to check for 'false negatives'. This however is again a waste of time and money.

VII. Conclusion

To conclude on the above, the crux of the discussion is that the spams not only put the internet user in trouble but due to its financial loss is also incurred. Hence, every user/company must review the impact of spam on the vast field of IT and also on the employees' output so as to determine the suitable action to fight against it. It is unfortunate that spam and junk email is cluttering up

everyone's email inbox. This paper focuses on the spam, its classification and statistical losses due to the spam. As this is a review paper, throughout the content, there is a continuous effort to make learn to the reader about spams, its classification, its detrimental effect, how to stop these spams.

References

[1] Radicati S. Email Statistics Report, 2011-2015[R]. The Radicati Group, Inc. A Technology Market Research Firm, USA.

[2] Symantec's MessageLabs Intelligence: 2010[R]. Annual Security Report.

[3] Zhang L, Zhu J and Yao T. An evaluation of statistical spam filtering techniques[J]. ACM Transactions on Asian Language Information Processing (TALIP), 2004, 3(4):243–269,

[4] Cristiatnini N and Shawe-Taylor J. An introduction to Support Vector Machines and Other Kernel-Based Learning Methods[B]. Cambridge University Press, 2003.

[5] Sahami M. Learning limited dependence Bayesian classifiers [C]. Second International Conference on Knowledge Discovery in Databases, 1996.

[6] Meyer T A. and Whateley B. Spambayes: Effective open-source, bayesian based, email classification system[C]. Proceedings of the First Conference on Email and Anti-Spam (CEAS), 2004. Available: http://www.ceas.cc/papers-2004/136.pdf

[7] Spamassassin available: http://spamassassin.org.

[8] Zhang L., Zhu J. and Yao T. An evaluation of statistical spam filtering techniques [J] ACM Transactions on Asian Language Information Processing (TALIP), 2004, 3(4):243–269.

[9] Musat C N. Layout Based Spam Filtering [C]. World Academy of Science, Engineering and Technology, 2005, 12, 371-374,

[10] Glick, M. and Rumelhart D. Neuroscience and Connectionist Theory. The Development in Connectionist Theory[B]. Erlbaum Associates, Hilldsdale, NJ, 1989.

[11] Cortes C. and Vapnik V. Support vector networks [A]. Machine Learning, 1995, 20, 273-297.

[12] Burges C J C. A tutorial on support vector machines for pattern recognition[C].Data Mining and Knowledge Discovery, 1998 2(2):121–167.

[13] McCulloch W S. and Pitts W H. A logical calculus of the ideas immanent in nervous activity [J]. Bulletin of Mathematical Biophysics 5:115-133.

[14] Rosenblatt F. Principles of Neuro dynamics [B]. Spartan Books, Washington, 1958.

[15] David Reby, Sovan Lek, Ioannis Dimopoulos Jean Joachim, Jacques Lauga,Ste ́phane Aulagnier Artificial neural networks as a classification method in the behavioural sciences[J]. Behavioral process (Elsevier), 1997, 40, 35–43.

[16] Holland J. Adaptation in Natural and Artificial Systems [B]. Ann Arbor: The University of Michigan Press, 1975.

[17] Shavlik J. Mooney R. and Towell G. Symbolic and neural learning algorithms: An experimental comparison [A]. Machine Learning 1991, 6:111–143.

[18] Sahami M. A bayesian approach to filtering junk email [W]. Proceedings of AAAI-98 workshop on Learning for Text Categorization, Madison, Wisconsin, USA, 1998.

[19] Haykin S. Neural Networks: A comprehensive foundation [B]. Printice Hall, 1998.

[20] Steele R. Techniques for Specialized Search Engines[C]. In: Proc. Internet Computing 2001, Las Vegas.

[21] Kolesnikov O., Lee W., and Lipton R. Filtering spam using search engines[R]. Technical Report GITCC-04-15, Georgia Tech, College of Computing, Georgia Institute of Technology, Atlanta, GA30332, 2004-2005.

[22] Katirai H. Filtering junk e-mail: A performance comparison between genetic programming and naive bayes [B]. Carnegie Mellon University, 1999.

[23] Koza J R. Hierarchical genetic algorithms operating on populations of computer programs[C]. In the Proceedings of the Eleventh International Joint Conference on Artificial Intelligence IJCAI-89, 1989, 1, 768–774.

[24] Koza J R. Genetic programming: A paradigm for genetically breeding populations of computer programs to solve problems[R]. Technical Report STAN-CS-90-1314, Dept. of Computer Science, Stanford University, June 1990.

[25] Koza J. Genetic Programming: On the Programming of Computers by Means of Natural-Selection [B]. MIT Press, MA, 1992

[26] Koza J R. Genetic Programming II: Automatic Discovery of Reusable Programs [B]. MIT Press, Cambridge, MA, 1994.

[27] Ahluwalia M. and Bull L. A Genetic Programming-based Classifier System[C]. Proceedings of the Genetic and Evolutionary Computation Conference, 1999, 1, 11-18, Orlando, Florida, USA.

Authors

Jitendra Nath Shrivastava received his Master of Technology (M.Tech) degree in Information Technology from Indian Institute of Information Technology (IIITA), Allahabad, India in 2007. He is working as an Associate Professor in the Department of Computer Science & Engineering at Invertis University. Presently he is doing his research work in Singhania University in the area of spam prevention techniques. His research interests are Data Mining and Artificial Intelligence. He has published two books and research papers. He is board of studies member for various autonomous institutions and universities. He can be contacted by email jitendranathshrivastava@yahoo.com

Maringanti Hima Bindu received doctorate (Ph.D.) Artificial Intelligence from Indian Institute of Information Technology, Allahabad, India in 2009. She has worked with BHABHA Atomic Research Institute, ISM, Dhanbad, IIIT, Allahabad. Presently she is working as an Assistant Professor in Information Technology Department of Jaypee Institute of Information Technology, Noida, India. Her research areas of interests are Artificial Intelligence, Image Processing and Pattern Recognition, Natural Language Processing and Cognitive Science. She has published many papers in national and international conferences and journals. She is the review board member of various reputed journals. She is board of studies member for various autonomous institutions and universities. She can be contacted by email mhimabindu@yahoo.com, hima.bindu@jiit.ac.in

