International Journal of Electronic and Electrical Engineering.
ISSN 0974-2174 Volume 7, Number 1 (2014), pp. 71-82

http://www.irphouse.com

Conception and Evolution of Spamming

Preeti Choudhary[1] and Dr. Meenu Dave[2]

[1]M.Tech. Scholar, Department of Computer Science
Jagannath University, Jaipur, India
[2]Assistant Professor, Department of Computer Science
Jagannath University, Jaipur, India
E-mail: [1]preeteei.choudhary@gmail.com, [2]meenu.dave@jagannathuniversity.org

Abstract

Spamming can be understood as the activity of generation of 'unsolicited mails' around the world for personal benefits like money, challenge or revenge. Spam mails not only waste one's time and resources, but the exponentially enlarged figure of generated e-mails causes the problem of carbon footprint and thus contributes to unsustainable environment. This paper highlights the reasons behind the spamming activity and discusses the spam flow tracking techniques, along with a comparative view of supervised and unsupervised learning techniques for spam filters.

Keywords: Ham, Machine Learning Approach, Spam, Spambot, Spammer, Zombie.

1. Introduction

The right to communicate, express oneself freely, and share information through internet has changed the world altogether. The advent of email out rightly transformed the globe into a local village and the problem of snail mail or the slow exchange of information vanished forever. The benefits of emails are numerous, but the other side of the coin is equally disturbing and alarming.

The majority of emails generated per day all around the world are not legitimate or rather it can be said that they are mails based on consent and not content. Such mails are termed as spam mails. A little curiosity on the part of an email account holder which ends up in consenting to view the spam mail finally results in irritation and anger. Spammers and hackers benefit from spamming and a small group of spammers

can earn a lot. Spammers keep on researching for new ways to attract the account holder's attention.

The goal of this research is to elaborate the appropriate industry to segregate spam from ham (legitimate mails). Study of various scholars has been put forward to think about the exact prototype of desired spam filter. This paper in section II discusses the various motivations and benefits behind the spamming activity. Section III reveals the related facts and figures and tries to highlight the actual status of spamming at present. Section IV discusses spam flow tracking techniques and section V gives a comparative view between the works carried out by different scholars for spam filtering through supervised and unsupervised learning techniques.

2. SPAM: Motivation and Benefits

As given in [1], spam emails are mostly generated as an alternative to spending on the expensive commercials. It is definitely a profitable business and apparently quite lucrative. Spamming and television commercials, both are meant for advertisement but are distinct for countless reasons. Commercials are used to create good picture of the product in consumers' mind, to draw the attention of the viewer and not to persuade someone to order the product "immediately", whereas the goal of spam is to lure the account holder to consider the offer immediately. Thus the email account holder who gets attracted to these spam mails, either orders the promoted product/service or at least visits the promoted web site. Not only this, spam emails usually are from companies and/or products that are not selling well and/or are in trouble in its subject line.

According to [2], spammers are circulating spam mails all over the world to fulfill their common objectives.

- The main goal of spam circulation is connected to earning money through advertisements. Spammers receive payment when the account holder gets enchanted by these alluring mails and clicks or opens links as per the given directions. Sometimes, even per-click payments are made.
- Spammers are paid handsomely for advertising the products/services via web page advertisements or bulk emails.
- The spamming activity misguides the search engines, thus receiving higher ranks and in turn attracts more network traffic.
- Spammers even become hackers in certain cases. They gain backdoor access to the user's computer, execute Trojans to acquire required information and thus publish more user specific spam content afterwards.
- In addition to money, the inspiration for hacking has challenge, boredom or revenge as the motivational factors.

Spammers step into the hacker world in order to fulfill their objectives. Therefore, spammers are seen to have utilize hacker techniques to send spam emails, hide their footprint to bypass filters (e.g.: blacklisting), spread their malicious codes for

performing mass attacks in future, steal users' private information (e.g. email account username and password), and inject malicious code in legitimate websites [2].

The study by [3] elaborates that as their activities cause irritation and anger, spammers generally do not use valid return addresses on their mails. The phony return addresses are also directed to mail servers of companies who actually had nothing to do with the spam mails. Many a times, the junk mails also provide a message for removing or unsubscribing these mails, which is not true. Many spammers purposely send more junk mails to such delete or unsubscribe request addresses because they come to know of the clients who at least read or pay attention to their emails.

3. Spam Related Facts and Figures

As in [4], the simple name SPAM came from a contraction of "spiced ham," which was the winning entry in a 1937 contest to re-brand the Hormel-made lunchmeat. SPAM was a famous name at that time and has governed its trading portion ever since. In 1970, according to [5], the word 'spam' acquired the global popularity due to BBC comedy sketch Monty Python, where the text SPAM appeared for 108 times.

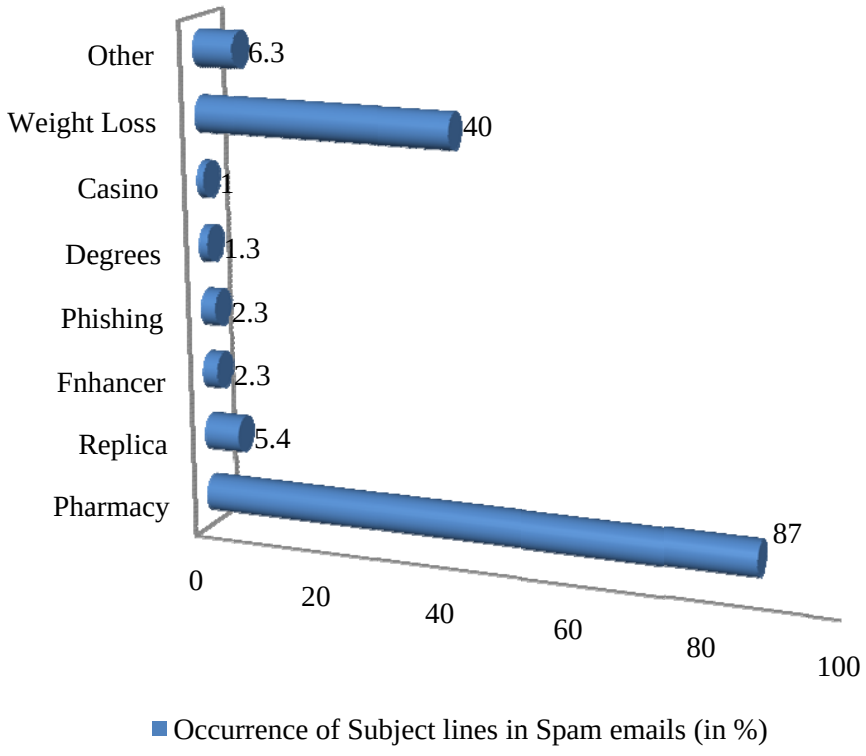

Fig. 1: Occurrence of subject line of spam mails in percentage.

After that, the irritating and fraudulent snippet of words was called spamming. Eventually the word 'spam' was treated as a synonym for 'Internet Junk Mail'.

As described in [6], initially Internet was opened world wide as a medium to communicate and exchange the information. In 1995, US Congress opened it for e-commerce purpose, this led to increase the users and business on the internet which resulted in tremendous growth in number of emails, and simultaneously in the volume of unsolicited bulk email (UBE) and unsolicited commercial email (UCE), also known as spam.

According to [5], in February 2005, Anthony Greco was the first person who was arrested for spamming. He was accused of sending 1.5 million of unsolicited ads for porn and mortgages and for blackmailing MySpacee.com in order to get marketing contract with them. Even a small business of spamming [7] which employs 50 people can make $50,000 per year easily. Figure 1 shows the frequency of subject lines appearing in spam emails among which pharmacy subject is the leading one.

In September 2010, it was stated in [7], that three big networks namely Verizon, AT&T and VSNL International host most of the spammers, hosting approximately 305,000 zombies per day. Most of the zombies come from Brazil.

It was estimated in [8] that in 2012 spam emails could increase by 58 billion which may cost around $198 billion, but the actual figures are different and outrageous.

According to [9], the facts related to spamming are quite astounding and have a far more giant face than expected. As surveyed, the new facts of spamming measured in August 2012 are stated below:

- More than 1 billion users' email accounts are on Hotmail, Yahoo mail or Gmail.
- On an average, more than 100 trillion emails are sent per year.
- million email messages are sent per second.
- 90% of the trillions of emails are spam or viruses.
- Spam costs businesses over $20 billion in form of decreased productivity and technology expenses.
- Around 20% increase in average email volumes was observed in 2012 as compared to 2011.

Based on the details provided by [7], Table 1 shown below gives year wise comparative view with reference to the massive growth in spam emails.

Table 1: Year Wise Comparative View of Increase in Spamming

Email related worldwide details	**Year**				
	2008	**2009**	**2010**	**2011**	**2012**
Number of Email accounts (in billions)	1.3	1.4	2.9	3.146	2.2
Total email traffic per day (in billions)	210	247	294	349	419
Spam percentage (out of total emails sent)	70%	90.4%	89.1%	71%	68.8%
Estimated global Spam cost	about $750 per user per year				

As per the details provided by [10], there are some important facts about spamming which become the year landmarks in the Internet world.

1978: First junk e-mail was sent to the open network
1982: First e-mail chain letter
1991: Craig Shergold e-mail chain letter
1993: MAKE.MONEY.FAST spam chain letter
1994: Siegel & Canter “Green Card Lottery” spam kick-starts the commercial spam era
1994: Mistrustful undesired e-mail coined as “SPAM”
1994: “Good Times Virus” e-mail hoax
1995 : "Spamware" (spamming software) starts developing
1995: Around 2 million of email addresses were bid tosale
2000: Taiwan (.tw) becomes the spam capital of the world
2000: Nigerian scam of spam becomes popular
2001: Around 209 million email addresses were bid for sale
2007: 25% of the 600 million computers on the Internetcould be spambots.
2008: Spam makes up 70% of all e-mails sent
2008: Up to 25% of all computers could be being used asSpambots
2009: Spam predicted to cost $130 billion globally

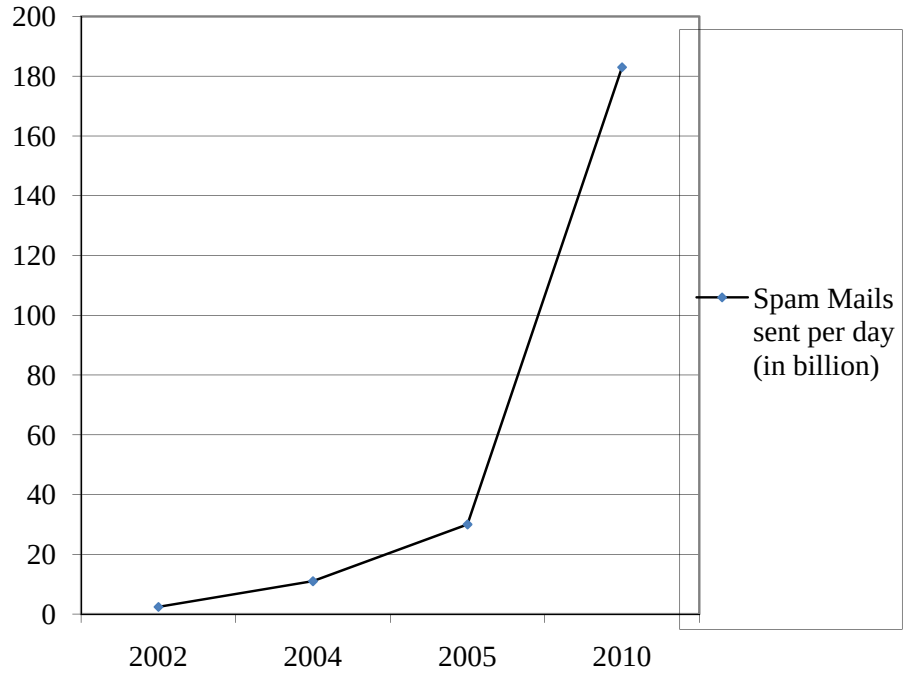

Fig. 2: Yearly rise in Spam Mails.

In absolute numbers, the average of spams sent per day increased from 2.4 billion in 2002 to 300 billion approx in 2009. More than 90% of incoming email traffic is spam. According to the National Technology Readiness Survey, the cost of spam in terms of lost productivity in the United States has reached US$ 21.58 billion annually, while the worldwide productivity cost of spam is estimated to be US$ 50 billion. On a worldwide basis, the information technology cost of dealing with spam was estimated to rise from US$ 20.5 billion in 2003, to US$ 198 billion in 2007[11].

Due to the huge amount of generated emails, misuse of resources resulting into increased carbon footprints adds to the downside of spamming. This in turn affects the environment and increases the level of pollution which may rise to hazardous levels. On one hand, measures are being adopted at the planning, designing, implementation and maintenance levels to increase the use of green computing in the IT world, and on the other hand activity like spamming is also increasing day by day which mitigates the efforts made in the direction of sustainable computing.

As discussed in [7], rise in spamming is tremendous and unbelievable. These incremented results have been the result of positive responses (due to the lack of awareness) towards spam mails and inefficient measures to filter hams from spams. Figure 2 denotes the exponential rise in spamming.

There is a vast market of spam. According to the survey in September 2010, there were total 13 countries which generate maximum volumes of spam emails. It has been observed that Brazil generates spam emails on the largest scale, while India acquires second position in spamming. Figure 3 shows the big pillars [7] of spam market.

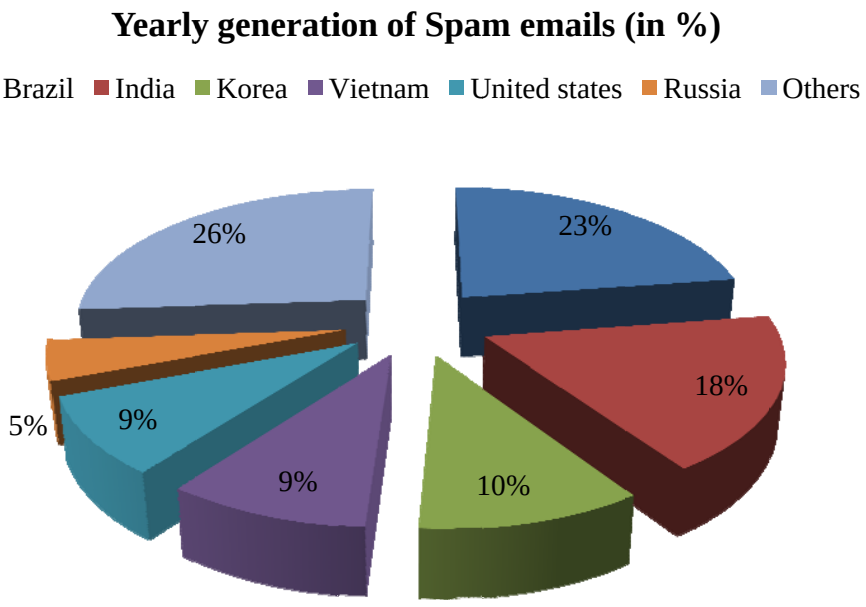

Fig. 3: % Spam E-mails generated yearly worldwide.

According to [7], domain gmail.com sends highest number of spam emails all over the world. This is followed by yahoo.com and hotmail.com. The order of 10 highest spam producing domains (from highest to lowest) is as follows:

- gmail.com
- yahoo.com
- hotmail.com
- expensetelle.com
- leafskip.info
- aim.com
- nbstrimming.com
- yahoogropus.com
- adaptedalcoh.com
- jmcastle.com

Most of the time, spam is sent from an email address designed to look as if it is coming from a legitimate domain, when it really is not the case. For example, only 1% of spams sent from gmail addresses are actually coming from gmail.

4. Spam Flow Tracking Techniques

As [12] and [3] state, it is impossible for a non-technical user to filter spam emails from ham emails whereas technical user can classify hams and spams, eliminating most of them. Still, it is difficult for them also to eliminate all spam emails without losing legitimate messages. Though some ISPs provide site-wise filtering but this method is not fully sufficient.

Spam filters can be implemented at all layers [13]; firewalls at the network level provide an integrated solution of Anti-Spam and Anti-Virus offering complete email protection from the unwanted spam emails. At MDA (Mail Delivery Agent) level also spam filters can be installed as a service to all of their customers. At email client, users can have personalized spam filters that then automatically filter email according to the chosen criteria.

i) Block Listing: As described in [6], the first spam filter was based on blocking of emails from nasty addresses, so it came to be known as block listing or blacklisting. But this type of filtering provided limited use when spammers started using zombies or botnets. The inverse of blacklisting is whitelisting, in which the emails from certain legitimate addresses are allowed in. This filter provides better solution, but it does not allow new legitimate addresses.

If a spam message is forwarded to multiple recipients, it is known as bulk mailing block. Such message must be identified which is sent in bulk. The problem is that these simple-minded methods can be fairly easily defeated. To circumvent bulk mailing blocks, some commercial emailers routinely route emails through an opaque maze of servers, so the message ends up not looking like a bulk message; in fact, they often rely

on sophisticated scanners to locate insecure networks of unsuspecting third parties through which messages can be relayed. To render tracking ineffective, they change their originating address often, and liberally, jumping from one of a large block of sender addresses to the next when complaints flow. They also routinely use specialized software to replace the originator's identification in e-mail headings with a false one. This has spawned a new cottage industry peddling "spamware," or high-tech tools needed by spammers, such as cloaking technology and stealth e-mailing products. One such company even throws in 25 million free "fresh e-mail addresses" with any bulk email software purchase [14].

ii) **RuleBased Predicates**: Rulebased techniques are more unsusceptible to black and white listing, since they do not depend on sender addresses. They generally scan the subject and/or the body of the e-mail message for some predetermined keywords, or strings of keywords (keyphrases), or specific layout features, the presence of which tags the message as suspicious[14].

iii) Tokenization: In [15]**, t**okenizer or in other words parser pulls out the interesting and noteworthy chain of letters from both subject and body of emails. This method is known as tokenizer. This method is equipped to understand various HTML tags and URLs' encoding methods like Base64 and also removes whitespaces among the words.

iv) SMTP Approaches: In [16], there is the exploration of everything to detect junk emails like email traffic observation, email exchange route verification, and authenticated SMTP sessions in SMTP based methods. The security is extended with a new proposed protocol based on SMTP, known as Differentiated Mail Transfer Protocol (DMTP). In DMTP, the recipient firstly acknowledged with an intent message and subsequent message will be retrieved if the user opens the previous one.

v) Machine Learning Approaches: As spammers adapted to simple filtering, new filtering techniques have been proposed that use more complicated rules, sometimes computer generated**.** These techniques give better results than others as they are based on automatic learning, modifying, and evaluating the relevant parameters [6]. Examples of such techniques are Bayesian classifiers, boosting trees, support vector machines, and combinations thereof, are machine learning techniques that have been applied to text categorization in general and email filtering in particular [14]. Any content filtering technique may be used in this approach, an example of which would be the Bayesian approach. It is assumed that this is the best available text categorization rule. It is used to decide whether the incoming Email is spam or legitimate. Most spam filters in use are now based on Bayesian filtering. The disadvantages of using this filtering are the text categorization rules require initial setup, the colloquial and popularly used words need to be updated in the Spam Word's dictionary and most recent technique of using similar looking words, for e.g. @ instead of a, or \ / instead of V, will also need to be identified [8].

According to [17], there is lot of work accomplished by research communities for detecting and classifying spam emails from the legitimate emails. Most of them use machine learning techniques but along with the technological advancements, the

structure and dispersion of email content is continuously changing. These changes make the mail classifying models built on old examples, inadequate for classifying new emails and the models become obsolete gradually. So detecting these changes and regular updating of the model is an important issue in the context of spam filtering which comes under the supervised machine learning. The learned model of a classifier should adapt itself to classify new emails correctly. This change in content distribution is known as concept drift. Concept drift detection methods help to find critical situations in which the email classify model should be updated.

vi) Legislative Action: After years of trying voluntary measures only to see the amount of junk increase, some people have decided that legislation is the answer. The US Congress is considering several bills to restrict junk emails. One weakness of any legislative solution is that spam is an international problem, and it's easy to send email from outside the United States. Still, U.S. laws could be applied as long as the spammer companies are American, and at the moment most of them are. Yahoo has a page with updates on many of these suits (http://headlines.yahoo.com/Current-Events/Spam-Wars/) [3]. The Government of Canada approved the world's toughest anti-spam legislation. The corporations can get fines up to 10 million dollars and individuals up to 1 million dollars for anti-spam law violation [5].

According to [18], many solutions have been offered from distinct backgrounds like economical, legislative (the CAN-SPAM act in the U.S.) as well as technical. On technological side, special kind of software is required to be installed either on client or server side to discriminate the spam emails and handle them appropriately.

5. Spam Filters and Supervised/ Unsupervised

5.1 Learning Techniques: A Comparative View

By seeing the studies carried out by various scholars, it is clear that supervised techniques to filter spam mails works well than unsupervised techniques. It is possible to use an unsupervised approach along with the supervised approach to cut some cost but the earlier one could not prove itself the efficient one, hence cannot be used as a standalone approach.

In 2003, as presented by [14], the conventional mail filtering techniques based on unsupervised learning were used where the classification is done on the basis of keyword matching. But, if the spammers change the composition of spam emails, the old classifiers will not be able to give the accurate results. That is the worst part of the unsupervised learning. Along with this, machine learning techniques based on supervised learning works well where the classifiers are regularly fed with the changing patterns of spam mails with different data sets.

In 2006, [18] discusses that instead of rulebased mining, text categorization and pattern identification techniques for email text analysis generates good results. The concept of extracting the text embedded into images sent as attachments has been used as base.

In 2009, as explained in [8], a mix and match approach of various spam filtration techniques has been used. In this both machine learning techniques are used concurrently. It classifies emails through various parallel arrangement of filters. Filters used in this study are black and whitelisting, content based filtering and forging filtering. In forging, sender's IP address is checked and domain name of email sending server is validated at server level with its IP address or Reverse DNS Lookup.

In 2010, the supervised learning was used and promoted. In this research two variants of Naïve Bayes classifier have been applied for spam filtering: multinomial and multi-bernoulli. In this, classifiers are updated dynamically through various learning techniques, to classify new emails and help in concept drift detection [17].

In 2011, in [19] a study was carried out to accomplish the goal and to speed up spam filters while keeping high classification accuracy. In the study, the overall acceleration came from three improvements: 1) Approximate pruning, which reduces the latency of duplicate token search by approximating membership checking with Bloom filter. 2) Approximate lookup, which allows us to replace memory intensive dictionary lookup with extended Bloom-filterbased value retrieval. 3) Approximate scoring, which replaces intensive floating point, logarithm operations with lookups on small cache-resident table. In particular, the major gain of the speedup comes from “Approximate lookup”, which is enabled by two novel techniques. The first technique approximates the dictionary lookup with hash-based Bloom filter lookup, which trades off memory accesses for increase in computation. Our second approximation method uses lossy encoding, which applies lossy compression to the statistical data by limiting the number of bits used to represent them. The goal is to increase the storage capacity of the Bloom filter and control its misclassification rate [19].

In 2012, [20] gave an image spam filtering technique, called Image Texture Analysis-Based Image Spam Filtering (ITA-ISF), that makes use of low-level image features for image characterization. The paper evaluates the performance of several machine learning-based classifiers and compares their performance in filtering image spam based on low level image texture features. These classifiers are: C4.5 Decision Tree (DT), Support Vector Machine (SVM), Multilayer Perception (MP), Naïve Bays (NB), Bayesian Network (BN), and Random Forest (RF). The experimental studies based on two publicly available datasets show that the RF classifier outperforms all other classifiers with an average precision, recall, accuracy, and F-measure of 98.6% [20].

6. Conclusion

Spam war is still on and needs an awareness to be spread all around because precautions are better than cure. Spam classifiers can only drain the existing waste but cannot stop the spammers. It can be possible, only if the email account holders themselves become unresponsive to spam mails, which could decrease their number.

Almost every scholar whose work discussed above is agreed on a common point that the supervised learning is more adopted over unsupervised learning; moreover combination of the two could do well in spite of using unsupervised learning alone. Sometimes, the solutions lack expertise due to ever increasing tricks and sharpness of attackers. It bounds the world to rethink about this global problem and the best remedy to save the world from the ghost, called spam.

References

[1] Grunch, M., "Article About Spam", http://www.articles-about-spam.com/benefits-of-spam.shtml. Last accessed on March 13, 2014.

[2] Hayati, P. and Potdar, V, 2009, "Spammer and hacker, two old friends", 3rd IEEE International Conference on Digital Ecosystems and Technologies, pp. 290-294.

[3] Ivey, K. C., April 1998, "Information superhighway-spam: the plague of junk e-mail", pp. 15-16, http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=00659621. Last accessed on March 13, 2014.

[4] The Douglass Report, November 23, 2004 , "Healthy benefits of spam", http://douglassreport.com/2004/11/23/healthy-benefits-of-spam, Last accessed on March 13, 2014.

[5] Smith, V. , November 16, 2012 , Sysaid IT Professional Community Forum, "5 Amazing facts about spam", , http://www.sysaid.com/Sysforums/posts/list/9038.page, Last accessed on March 13, 2014.

[6] Hoanca, B., 2006, "How good are our weapons in the spam wars?", IEEE Technology and society magazine-spring, pp. 22-30.

[7] Vallente, D., September 7, 2010, Wikibon Blog, "Facts about spam: A visual journey", http://wikibon.org/blog/facts-about-spam, Last accessed on March 13, 2014.

[8] Bhuleskar R., Sherlekar , A. and Pandit , A., 2009, "Hybrid spam e-mail filtering", First International Conference on Computational Intelligence, Communication Systems and Networks, IEEE Computer Society, pp. 302-307.

[9] Schneider, A., August 23, 2012, Mass Transmit Broadcast email marketing blog, "13 Amazing facts about email", http://masstransmit.com/broadcast_blog/12-amazing-facts-about-email. Last accessed on March 13, 2014.

[10] Nash Network Inc IT consulting, page, March 2009, "Strangling the internet: why spam matters and what it costs", http://www.nashnetworks.ca/why-spam-matters-and-what-it-costs.htm. Last accessed on March 13, 2014.

[11] Almeida, T. A. and Yamakami, A., 2010, “Content-based spam filtering”, IEEE.
[12] Yunos, F., “Whitepaper|spam and ham- a simple guide”, Open kod, no. 12, pp. 1-8, http://www.openkod.com/ver3/images/pdf/antispam.pdf. Last accessed on March 13, 2014.
[13] Christina, V., Karpagavalli , S. and Suganya, G., September 2010 “Email spam filtering using supervised machine learning techniques”, International Journal on Computer Science and Engineering (IJCSE), vol. 02, no. 9, pp. 3126-3129.
[14] Bellegarda, J. R. , D. Naik, K. E. and Silverman A., 2003, “Automatic junk e-mail filtering based on latent content”, IEEE, ASRO, pp. 465-470.
[15] Luo, Y., January 2010 , “Workload characterization of spam email filtering system”, International Journal of Network Security & Its Application (IJNSA), vol.2, no.1, pp. 22-41.
[16] Caruana, G. and Li, M., February 2012, “A survey of emerging approaches to spam filtering”, ACM Computing Surveys, vol. 44, no. 2, article 9, pp. 1-27.
[17] Hayat, M. Z., Basiri, J. , Seyedhossein, L. and Shakery, A. , 2010, “Content-Based concept drift detection for email spam filtering”, IEEE, 5th International Symposium on Telecommunications (IST'2010), pp. 531-536.
[18] Fumera G., Pillai I. and Roli F. , December 2006, “Spam filtering based on the analysis of text information embedded into images”, Journal of Machine Learning Research, vol. 7, pp. 2699-2720.
[19] Zhong, Z. and Li, K., January 2011, “Speed up statistical spam filter by approximation”, IEEE transactions on computers, vol. 60, no. 1, pp. 120-133.
[20] Al- Duwairi, B., Khater, I. and Al- Jarrah, O., September/December 2012, “Detecting image spam using image texture features”, International Journal for Information Security Research (IJISR), vol. 2, issues 3/4, pp. 344-353.