Volume 5, Issue 10, October-2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Using Fuzzy Logic for Email Spam Filtering

Harshita Kotian*
Computer Engineering, VJTI/ Mumbai University, India

Kaushalya Gupta
Computer Engineering, SFIT/ Mumbai University, India

Jeba Jino Stephy
Computer Engineering, SFIT/ Mumbai University, India

Abstract- Email is used today by millions of people to communicate around the world. Over the last few years, spam mail has become a major problem for users. A large amount of spam is flowing into users' mailboxes daily. Spam not only affects most email users but also strains the IT infrastructure of organizations and costs businesses billions of dollars in lost productivity. There are several methods developed for filtering spam. The approach described in this paper is to classify the spam mails according to degree of spam content using spam word ranking database and fuzzy rules. This work classifies the emails based on the degree of the threat that each word possess combined with other existing techniques.

Keywords— Spam, Ham, Fuzzy, FIS, Tokenization

I. INTRODUCTION

E-mail spam, also known as unsolicited bulk e-mail (UBE), junk mail, or unsolicited commercial e-mail (UCE), is the practice of sending unwanted e-mail messages, frequently with commercial content, in large quantities to an indiscriminate set of recipients [1]. Spam in e-mail started to become a problem when the Internet was opened up to the general public in the mid-1990s. It grew exponentially over the following years, and today composes some 80 to 85 percent of all the e-mail in the World.

By definition, email spam is any email that meets the following three criteria [8]:

- Anonymity: The address and identity of the sender are concealed
- Mass Mailing: The email is sent to large groups of people
- Unsolicited: The email is not requested by the recipients

Spam mails sent by the spammers adversely affect the email communication process. Unsolicited commercial mails are often sent by the spammer to illegally promote a service or product. Spam became an issue when the internet was opened to the general public. Internet users are forced to receive spam mails in their inbox. Spammers harvest the address of internet users from various sources and cause inconvenience to the users. The user's inbox are flooded with enormous spam mails.

Spamming has many negative consequences for the Internet like [9]:

- It consumes a lot of Internet resources.
- It threatens the very utility of email as a form of communication.
- It reduces the effectiveness of legitimate advertising.
- It exposes children to inappropriate material.
- It consume the bandwidth of network, wastes memory, time of user and causes financial loss to the users and the organizations

With the increasing popularity and low cost of sending an e-mail makes it very attractive to the direct marketers. It is now become very easy to send unsolicited messages blindly to thousands of people at no cost at all by using easily available bulk-mailing software and large lists of email addresses typically harvested, even purchased or rented from web pages and newsgroup archives [3]. Therefore, the volume of this unsolicited bulk e-mail or spam that shows up in the user's mailbox daily has been increasing exponentially. Thus the necessity for an effective spam filter increases.

The most common form of communication these days, in organizations especially and for consumers is email. In 2013, there were 929 million mailboxes for business email accounts and it is expected to grow and reach over 1.1 billion in 2017. Also in 2013, the majority of email traffic comes from the business emails, which accounts for 100 billion emails per day [6]. Majority of business communication happens over email. According to spam statistics for recent years, the percentage of email sent over the Internet has increased from 36% in 2002(Clifford et al., 2003) , 45% in 2003 to 64% in 2004(Jung and Emil, 2004) to 80% in 2006(Siponen and Stucke, 2006, Leavitt, 2007, Jaeyeon and Emil, 2004), 92% in 2009(J. and T., 2008) and 95% in 2010 (Gina, 2010) [6].Since emails are so rigorously used, they come with problems.

E-mail provides a perfect way to send millions of advertisements without any effort for a sender, and this fortunate fact is nowadays extensively exploited by several organizations. As a result, the e-mail boxes of millions of people get cluttered with all these "spam" or "junk mail".

It is now become very easy to send unsolicited messages blindly to thousands of people at no cost at all by using easily available bulk-mailing software and large lists of email addresses typically harvested, even purchased or rented from web pages and newsgroup archives [1].

According to Symantec, in 2014, spam comprised 64% of all emails sent globally. I.e. for every solicited mail received (ham), you got two others that were spam [11]. In a new survey conducted by Information Security/SearchSecurity.com, lost productivity (92 percent) and clogged e-mail servers (62 percent) were cited as the most egregious consequences of spam. More than half of the respondents were concerned about malware infections [10]. U.S.A is leading spam relaying country with 18.3% [4].

The reasons of existence of Spam are the low cost to send Spam mail and the ease of sending it through various software tools. The most popular and direct way to prevent spam is the anti-spam filters, software tools that block spam messages automatically

II. SPAM FILTERING TECHNIQUES

Anti-spam methods can be categorized into two approaches [7]:

- Pre-send methods act at the sender side to reduce or prevent transporting the e-mail over the network which means the problem should be prevented before it occurs.
- Post-send methods act at the receiver side after consuming the networks resources because the email has been transferred to the receiver side which means in this case, the problem should be addressed after it occurred. Also, Post-send methods can be categorized into two types:
  - Machine learning
  - Non-machine learning techniques

The non-machine learning techniques uses a set of created or predefined rules to classify the message as spam or ham such as heuristics (rule based), signature, and blacklisting techniques, whereas the machine learning techniques don't need to define rules explicitly, but they need training data or samples to learn the classifier in order to use them in classification process.

The different techniques are as follows:

- **Whitelists/Blacklists**

The functionality of these filters is simple: a whitelist is a list, which includes all addresses from which we always wish to receive mail. We can add email addresses or entire domains, or functional domains. An interesting option is an automatic whitelist management tool that eliminates the need for administrators to manually input approved addresses on the whitelist and ensures that mail from particular senders or domains are never flagged as spam. A blacklist works similarly: this is a list of addresses from which we never want to receive mail. The drawback is that spammers generally use false addresses [2]. Spammers are also evading IP-based blacklists with nimble use of the IP address space (e.g., stealing IP addresses on the same local network, stealing IP address blocks with BGP route hijacking [5].

- **Mail header checking**

Mail header checking consists of a set of rules that, if a mail header matches, triggers the mail server to return messages that have blank "From" field, that lists a lot of addresses in the "To" from the same source, that have too many digits in email addresses. It also enables to return messages by matching the language code declared in the header.

- **Bayesian analysis**

The word probabilities (also known as likelihood functions) are used to compute the probability that an email with a particular set of words in it belongs to either category. This contribution is called the posterior probability and is computed using Bayes' theorem. Then, the email's spam probability is computed over all words in the email, and if the total exceeds a certain threshold (say 96%), the filter will mark the email as a spam. The Bayesian Algorithm analysis needs to be more accurate and suffers from Bayesian poisoning [3.]

- **Content Based filtering**

Existing content-based filters can be categorized as rule-based, key-word based and learning based [1].

Keyword based

The keyword based filter utilizes a dictionary of common spam phrases and search for a particular pattern in the messages. While they perform well, they need to be maintained and tuned constantly since the characteristics of spam messages change over time.

Rule based

Rule-based filters generally use a wide range of tests to recognize spam features and assign a 'spam score' to every email. They also require periodic update and maintenance. As classification rules are often fixed and since the classification of 'good' and 'bad' spam often differs from person to person, fixed classification methods are unlikely to provide good performance for all users

Learning based

This category of anti-spam filters are still emerging and automatically learn how to block spam messages by processing previously received spam and legitimate messages.

To evade content-based filters, spammers have adopted techniques such as image spam and emails explicitly designed to mislead filters that "learn" certain keyword pattern [5].

III. PROPOSED SYSTEM

The proposed system is using the Fuzzy Logic Technique. The motivation of using fuzzy logic for spam detection came from the fact that there is no clear separation between spam and non-spam messages and fuzzy logic is a good way to deal with those fuzzy boundaries. A Trainable Fuzzy classifier is used to build an automatic anti-spam filter. Trainable fuzzy system is a fuzzy logic based system that derives the (fuzzy) classification from training data using learning techniques. Figure 3.1 below shows a model for fuzzy inference system for spam filtering.

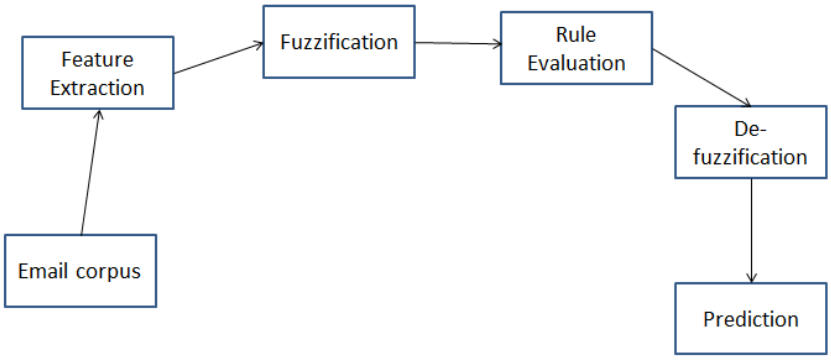


Fig 1: Spam filtering Model using fuzzy inference system

The email contents are first taken and features are extracted from it. Words are extracted from the content of the emails which, this work compares them against a list of spam words stored in the database ranked with its values and categorizes the words in accordance to the ranking. Fuzzy inference system finally takes the input value from the above ranking. They are then converted from crisp to fuzzy inputs. Fuzzy rules are then evaluated and classify the output as weakly, moderately and highly spam mail. Defuzzification is then applied to the output and prediction is given for presence of spam.

IV. IMPLEMENTAION

The design of the Fuzzy based spam filter system is as shown in figure 2.

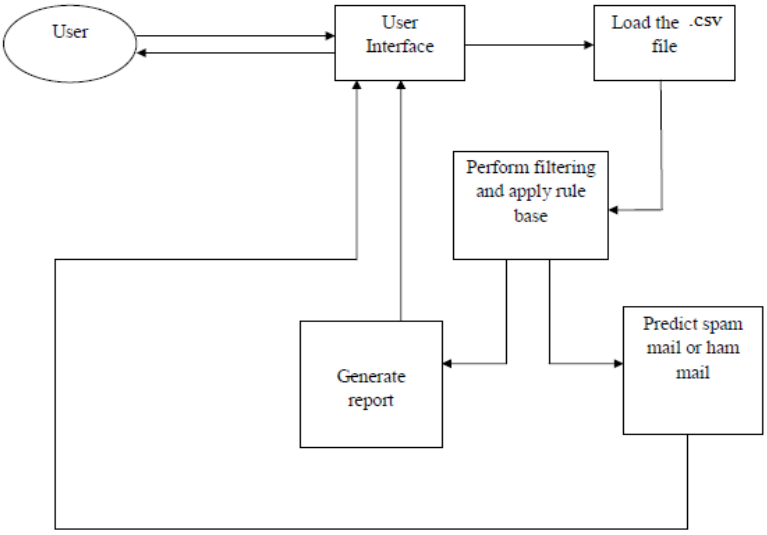


Fig. 2 Fuzzy based email filtering system design

The steps performed are:

Feature Extraction

In general, a message is represented as a collection of features derived from the message, or from extrinsic information related to the message. Feature selection represents the most important step of email spam filtering technique. In this step, we study information available in the email message and carefully select some of them to be among the features used for classification. It is also important to include most of the mandatory and optional email fields in order to fill any gab or missing information that is required for email classification.

The process of building a feature vector of an email starts by preprocessing of email messages to convert them into a standard format. After that, we select the required features and build the feature vector which summarizes all the needed information from an email.

The following are the features that is considered: Received Field, Sender Address, Number of Receivers, Email Subject and body

Tokenization

Tokenization is dividing of a string of written language into its component words. In English text one may reasonably identify words as strings of consecutive letters separated by punctuation or spaces. Some minor details, such as whether words may contain apostrophes or digits can also be found. For content based analysis tokenization step is a necessary step to be performed. Spammers have employed many obfuscation methods in an attempt to defeat spam filter tokenization which in turn have given rise to deobfuscation techniques that aim to recover the original tokens [3].

Fuzzification

In this system, for each input and output variable selected, we define two or more membership functions (MF), normally three but can be more i.e. to define a qualitative category for each one of them, for example: low, normal or high. Then fuzzification of all the variables is done. Any of the values will belong to at least one MF with a certain degree of membership.

Rule base (decision matrix) definition

Once the input and output variables and MF are defined, we design the rule-base (or decision matrix of the fuzzy knowledge-base) composed of *expert* IF <antecedents> THEN<conclusions> rules. These rules transformed the input variables to an output that indicate the risk of operational problems (this output variable, risk of a problem, also had to be defined with MF, usually low, normal and high risk). Depending on the expert knowledge, one could have several inputs and several outputs.

Fuzzy rule is of the form:

IF A1 is Low AND A2 is Mid AND A3 is High AND A4 is High THEN B is Spam

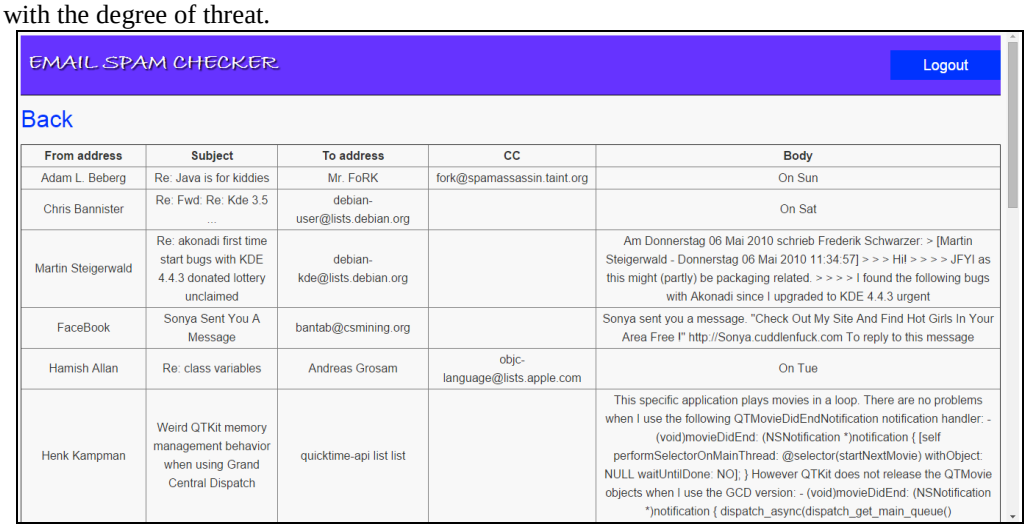
Where A1 to A4 are crisp inputs representing feature values, AND represents the fuzzy AND operation and B represents the fuzzy output

Defuzzification

In order to obtain a percentage of risk of a spam, the output is defuzzified

V. EXPERIMENTAL RESULTS

The results are displayed by using a dummy Email dataset. It displays the results of classification of mails into ham and Spam along with the degree of threat.



From address	Subject	To address	CC	Body
Adam L. Beberg	Re: Java is for kiddies	Mr. FoRK	fork@spamassassin.taint.org	On Sun
Chris Bannister	Re: Fwd: Re: Kde 3.5 ...	debian-user@lists.debian.org		On Sat
Martin Steigerwald	Re: akonadi first time start bugs with KDE 4.4.3 donated lottery unclaimed	debian-kde@lists.debian.org		Am Donnerstag 06 Mai 2010 schrieb Frederik Schwarzer: > [Martin Steigerwald - Donnerstag 06 Mai 2010 11:34:57] > > > Hi! > > > > JFYI as this might (partly) be packaging related. > > > > I found the following bugs with Akonadi since I upgraded to KDE 4.4.3 urgent
FaceBook	Sonya Sent You A Message	bantab@csmining.org		Sonya sent you a message. "Check Out My Site And Find Hot Girls In Your Area Free !" http://Sonya.cuddlenfuck.com To reply to this message
Hamish Allan	Re: class variables	Andreas Grosam	objc-language@lists.apple.com	On Tue
Henk Kampman	Weird QTKit memory management behavior when using Grand Central Dispatch	quicktime-api list list		This specific application plays movies in a loop. There are no problems when I use the following QTMovieDidEndNotification notification handler: - (void)movieDidEnd: (NSNotification *)notification { [self performSelectorOnMainThread: @selector(startNextMovie) withObject: NULL waitUntilDone: NO]; } However QTKit does not release the QTMovie objects when I use the GCD version: - (void)movieDidEnd: (NSNotification *)notification { dispatch_async(dispatch_get_main_queue()

Fig 3 View all mails page

The user can view all mails before filtering them. These mails are the email corpus that is provided as an input to the spam filter for processing.

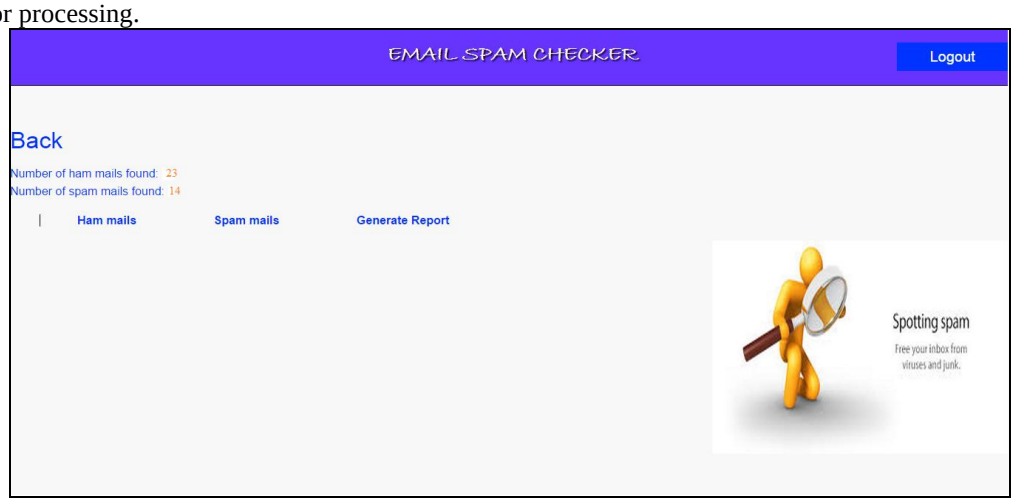
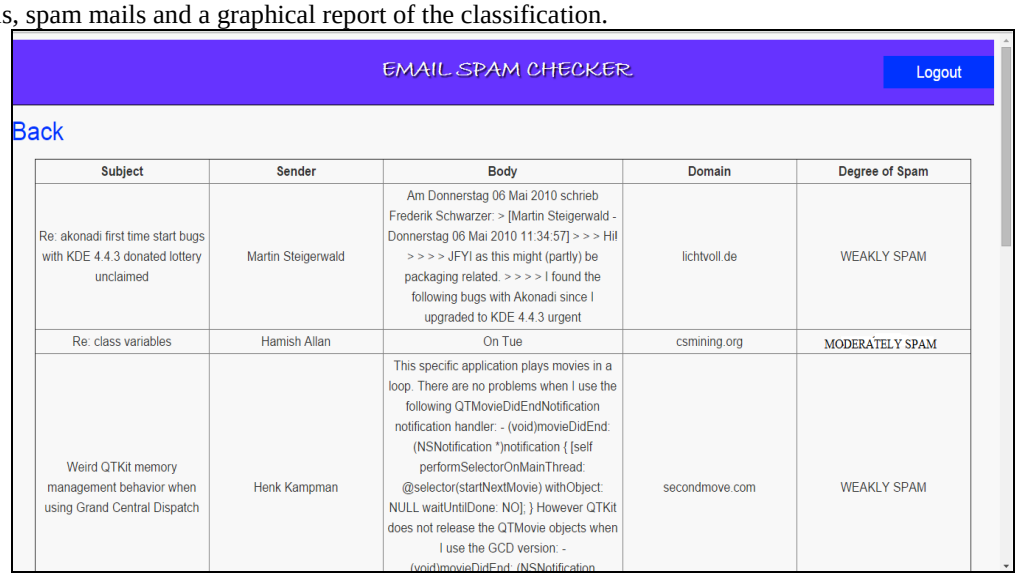


Fig 4 Result displayed after filtering process

This page displays the count of ham and spam mails in the input email dataset. The page provides options for viewing the ham mails, spam mails and a graphical report of the classification.



Subject	Sender	Body	Domain	Degree of Spam
Re: akonadi first time start bugs with KDE 4.4.3 donated lottery unclaimed	Martin Steigerwald	Am Donnerstag 06 Mai 2010 schrieb Frederik Schwarzer: > [Martin Steigerwald - Donnerstag 06 Mai 2010 11:34:57] > > > Hil > > > > JFYI as this might (partly) be packaging related. > > > > I found the following bugs with Akonadi since I upgraded to KDE 4.4.3 urgent	lichtvoll.de	WEAKLY SPAM
Re: class variables	Hamish Allan	On Tue	csmining.org	MODERATELY SPAM
Weird QTKit memory management behavior when using Grand Central Dispatch	Henk Kampman	This specific application plays movies in a loop. There are no problems when I use the following QTMovieDidEndNotification notification handler: - (void)movieDidEnd: (NSNotification *)notification { [self performSelectorOnMainThread: @selector(startNextMovie) withObject: NULL waitUntilDone: NO]; } However QTKit does not release the QTMovie objects when I use the GCD version: - (void)movieDidEnd: (NSNotification	secondmove.com	WEAKLY SPAM

Fig 5 Spam mails

The page lists all the spam mails with the degree of threat it possess (i.e. classification into Weakly Spam mail, Moderately Spam mail and Highly Spam mail) after the filtering process.

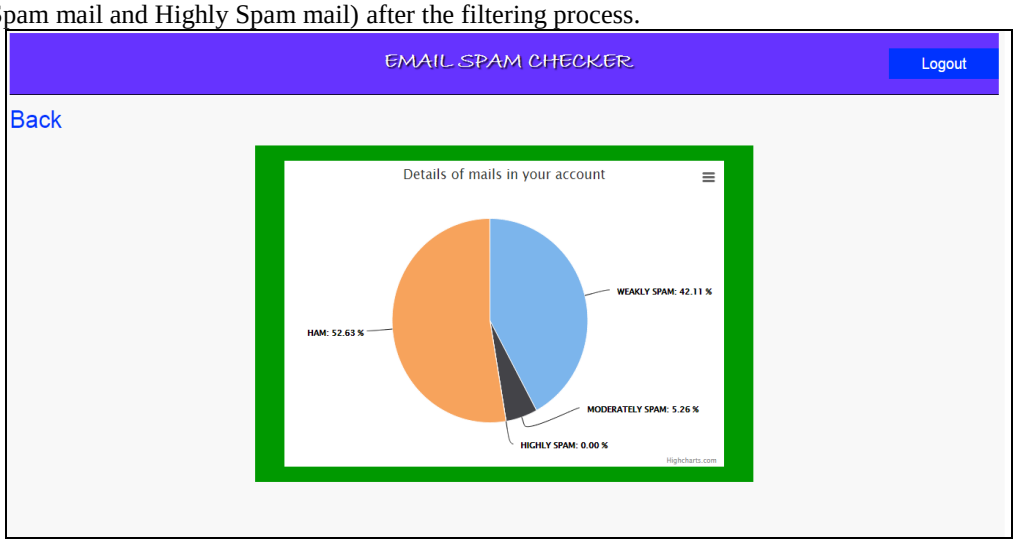


Fig 6 Classification result in form of Pie-Chart

The analysis of mails in the dataset after the filtering process is shown using a Pie chart showing the percentage of mails in each of the category i.e. (Ham mails, Weakly Spam mail, Moderately Spam mail and Highly Spam mail).

VI. CONCLUSIONS

There are many classifiers and filters available for classifying and filtering of email. A content based classification of spam mails with fuzzy word ranking is proposed which is an evolutionary approach for detecting the spam mails. The proposed work used sets of linguistic terms for ranking and classifying spam mails. This method has extracted only four features from an email instead of extracting all the features from the mail.

The system can further be enhanced in the future by training it and selecting more additional features like audio spam detection, mails with only HTML in its body, multiple language detection. Also text, video and image in the attachments can be scanned.

REFERENCES

[1] Christina V, Karpagavalli S, and Suganya G, "A Study on Email Spam Filtering Techniques", International Journal of Computer Applications (0975 – 8887), Volume 12– No.1, December 2010.

[2] Vijayan R, Viknesh S T G M, and Subhashini S, "An Anti-Spam Engine using Fuzzy Logic with Enhanced Performance Tuning", International Journal of Computer Applications (0975 – 8887), Volume 16– No.3, February 2011.

[3] M. Muztaba Fuad , Debzani Deb , M. Shahriar Hossain " A Trainable Fuzzy Spam Detection System" .

[4] G.Santhi, S. MariaWenisch and Dr. P. Sengutuvan," A Content Based Classification of Spam Mails with Fuzzy Word Ranking" IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 3, No 2, May 2013.

[5] Anirudh Ramachandran, Nick Feamster, and Santosh Vempala,"Filtering Spam with Behavioral Blacklisting", Proceeding CCS '07 Proceedings of the 14th ACM conference on Computer and communications security,2007.

[6] Kamini Bajaj and Josef Pieprzyk," A Case Study of User-Level Spam Filtering", Proceeding AISC '14 Proceedings of the Twelfth Australasian Information Security Conference - Volume 149, 2014.

[7] Aziz Qaroush, Ismail M. Khater, Mahdi Washaha,"Identifying Spam E-mail Based-on Statistical Header Features and Sender Behavior ", Proceeding CUBE '12 Proceedings of the CUBE International Information Technology Conference, 2012.

[8] What is Email Spam? [Online]. Available: http://emailmarketing.comm100.com/email-marketing-ebook/email-spam.aspx.

[9] Gammadyne [Online]. Available: http://www.gammadyne.com/spam.htm

[10] Using email spam filtering techniques to get rid of spam [Online]. Available: http://searchsecurity.techtarget.com/Using-email-spam-filtering-techniques-to-get-rid-of-spam

[11] 2014 Estimated Global Email Spam Rate is 64%. That's Almost 2 out of 3 Emails [Online]. Available: http://www.business2community.com/email-marketing/2014-estimated-global-email-spam-rate-64-thats-almost-2-3-emails-0875585.

