E-mail Classification Using Genetic Algorithm with Heuristic Fitness Function

Jitendra Nath Shrivastava[#1], Maringanti Hima Bindu[*2]

[#1] *Research Scholar, Singhania University, Jhunjhunu, Pacheri Bari, Rajasthan, India*
[*2] *Prof. & Head, Deptt. of Computer Science and Applications, North Orissa University, Orissa, India*

***Abstract*— Internet users use e-mail to communicate over internet. They rely on the mail system to deliver their mails to the recipient. Spam has made the mail system more unreliable and unpredictable. It is the biggest threat to the e-mailing users. Spam has increased enormously in the last few years. Presently, the spam has an important role for users of email. It is really very hard to design an anti-spam solution that could be useful in stopping the spam completely. Most mail can get falsely caught by spam filters on the way to the recipient or it can drown among spam in the recipients' inbox. A general definition of spam does not occur because spam is different for every user. The Internet community needs to work to prevent spam. Possible ways to do this are through the law and the legal system, technical solutions and user awareness. In past, several statistical methods have been used, and have shown great performance, excelling in adapting to the ever changing content of SPAM e-mail. Still a perfect solution is not available till date. In this paper, a genetic algorithm based e-mail filtration method is presented, and method is tested on 1108 emails and efficiency of the method is found to be nearly 82%.**

***Keywords*— Hierarchical Matching, Calligraphic Retrieval, Skeleton Similarity**

1. INTRODUCTION

Despite of beneficial features of the Internet, it is also gradually becoming in support of malicious activities. Over the years, worms, viruses and security issues have been seen as the major problems that information technology is facing. However, spam is slowly taking a different dimension on a daily basis and, therefore, becoming as problematic as worms and viruses. The distribution of bulk messages by unknown senders cost little or no cost. Hence, e-mail users still spend lots of time and effort to recognize legitimate mails and deleting it from their mailbox. However, it is irritating when users check their mail box and find over thousands of mails from unknown sender with a reasonably good heading but irrelevant content that does not concern the users. Hence, this is just for spammers to promote their unethical markets and may be called spam.

The spammers activities has increased during the past few years. This has subsequently decreased the worth of e-mail. We are not only experiencing irritation but also are losing millions in the form of human resources and server time. Fighting spam is called a war. The entire world is giving its precious time and money to stop spam, which could be used for any number of constructive projects.

Presently several organizations are being deployed worldwide to fight the war against spam. Several software's are being developed to stop spam. These softwares are mainly some type of filters. Despite several solutions to spam, users are reluctant or unable to use them. Primarily, this is due to the lack of transparency and relatively difficult use of the solutions. The use of filters to stop spam is beyond the ability of many ordinary users.

Virtually, all email users are more likely to receive spam but how does the mail find its way into the users' respective inboxes remains unknown to the user. Spam comes from different spammers living in different countries of the world. Fig. 1 below shows the percentage of spam in top 20 world ranking [1].

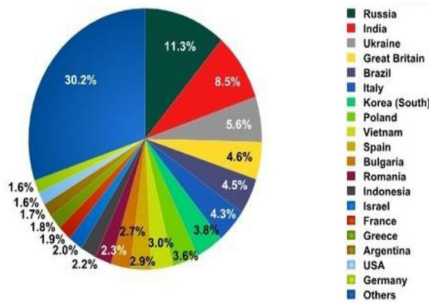

Fig. 1 countrywide distribution of the SAPM generation.

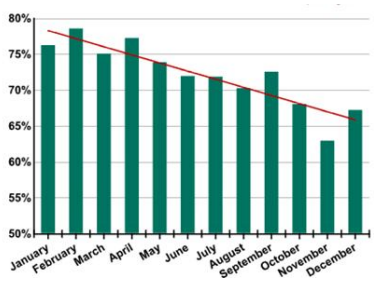

Fig. 2 E-mail spam trend in 2012 as per Kaspersky Lab.

Countrywide Distribution of Sources of Spam

In the year, 2012 some major changes among the countries from which spam originates takes place. China, which was not

even in the top 20 sources of spam in 2011, took second place in 2012, accounting for 19.5% of all unsolicited mail [1]. Spam originating in the US increased 13.5 percentage points, to 15.6% - enough to take third place. Asia remains the leading region for spam generation and distribution. Over the past year, the region's share of the world's junk mail rose 11.2 percentage points to more than 50%. Due to the increased spam contribution from the United States, North America stood second place in the top 10 with rise to 15.8% — up from just 2% in 2011. The spam originating in Latin America fell by 8 percent and now down at 11.8%. Europe also dropped down the ranks. In 2012, the total amount of spam originating in Europe was just half of contribution came in 2011 [1].

Spam Classification

The spam can be classified as solicited e-mail and unsolicited emails.

Solicited e-mail

Solicited email is something that you've asked for, voluntarily or otherwise Solicited commercial email is any commercial message, newsletter, draft or posting sent only to recipients who have requested it and can choose to opt out of receiving the mailing. Solicited email can come from companies you've bought products from or related companies that have bought mailing lists from that company. Check the privacy policy of sites you are intending on submitting your email address to, it may provide insight to how often they will sell your address, if at all. Spammers use this distinction as a defense, claiming that all their mailings are solicited.

Unsolicited e-mail

Unsolicited bulk email ('SPAM') is one of the most prevalent threats to network integrity on the public Internet. It causes denial of service at the network level, by flooding network with unwanted emails thus wastage of bandwidth and overloading email hosts. It reduces productivity both of mail administrator and of end users. Unsolicited commercial email, commonly known as spam, is any message or posting, regardless of its content, that is sent to multiple recipients who have not specifically requested the mail. The amount of unsolicited email that is sent and received over the Internet increases every day. Spammers the people who send bulk unsolicited emails are beginning to put telemarketers to shame as it becoming annoying for unwanted users. Hence, an adaptive technique is necessary which can combat with ever changing email structure [2].

2. GENETIC ALGORITHM

The details of how Genetic Algorithms work are explained below, and the schematic of the algorithm is shown in fig. 3.

2.1 Initialization

In genetic algorithm initial population is *generated randomly*. However, some research has been done to produce a higher quality initial population more useful for a particular problem. Such an approach is used to give the GA a good start point and speed up the evolutionary process.

2.2 Reproduction

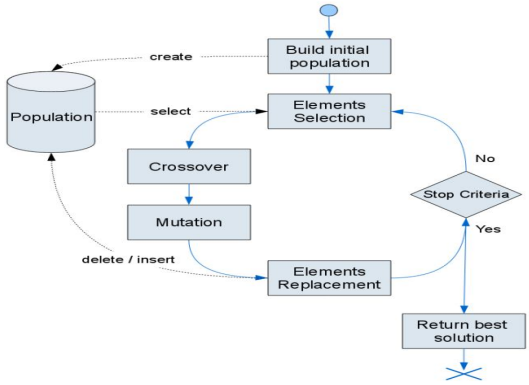

Fig. 3, Schematic of the genetic algorithm

In generational reproduction, the complete population is replaced in each generation. In this method, two mate of the old generation are coupled together to produce two children. This procedure is repeated N/2 times and thus producing N newly generated chromosomes.

2.3 Crossover Operator

The crossover is the most important operation in GA. Crossover as name suggests is a process of recombination of bit strings via an exchange of segments between pairs of chromosomes. There are various kinds of crossover like one point, multi-point crossover etc.. one point crossover is shown below:

.

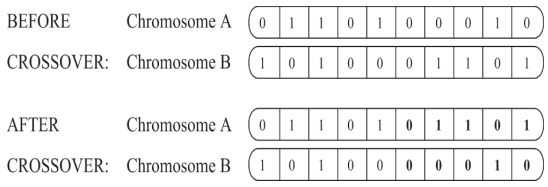

Fig. 4 Schematic of one point crossover

In uniform cross-over, each gene of the first parent has a definite probability (generally 0.5) of swapping with the corresponding gene of the second parent. However, in non-uniform crossover the probability value is different form 0.5.

2.4 Mutation

Mutation has the effect of ensuring that all possible chromosomes can maintain good gene in the newly generated chromosomes. With crossover and even inversion, the search is constrained to alleles which exist in the initial population so initial characters can be maintained. The mutation operator can overcome this by simply randomly selecting any bit position in a string and changing it if required. This is useful since crossover and inversion may not be able to produce new alleles if they do not appear in the initial generation and a new

type of chromosomes can be generated with old and new character.

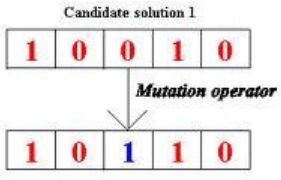

Fig. 5 Schematic of mutation

3. E-MAIL FILTERING PROCESS

Currently, in email filtering there are two methods which are used, one is filtering e-mail address, and the other is filtering e-mail content. But both of these technologies are lack of intelligence and adaptability for new and emerging spam, they must be manually re-amended to adapt to the new changes. With spammers and means of diversification springing up, the traditional filter based on the old technique is difficult to adapt to the new spam, the studying of email structure according to network information, as well as transmission information and so on to identify the characteristics of the spam. And then automatically set up and update new features and rules of the spam is the need of the hour. This can be achieved by using the Genetic Algorithm in the designing of e-mail filters. In this paper, Genetic algorithm is used as spam classifier. The collection of the e-mails is called corpus. Spam mails for the corpus are encoded into a class of chromosomes and these chromosomes undergo with genetic operations, i.e., crossover, mutation and fitness function etc.. The rules set for spam mails are developed using the genetic algorithm.

Rules for classifying the emails:

The weight of the words of gene in testing mail and the weight of words of gene in spam mail prototypes are compared and matched gene is find. If the matched gene is greater than some number let say 'x' then mail is considered as spam.

Fitness Function:

$$F = \begin{cases} 1 & \text{SPAM mail} \\ 0 & \text{Ham mail} \end{cases}$$

The basic idea is to find SPAM and HAM mails form the mails arriving in the mail box. As the fitness function is itself problem dependent and cannot be fixed initially in SPAM email filtering. For the evolution of the fitness function we carried out experiments on 500 mails which consist of pool of 300 SPAM and 200 HAM mails, and we found that the minimum score point was 3 for the correct identification of emails. Hence, we defined our fitness function as

$$F = \begin{cases} 1 & \text{Score point } \geq 3 \\ 0 & \text{Score point } < 3 \end{cases}$$

5.3.1 Procedure:

An email consists of header and message or body. In the header part Form, To, CC (carbon copy), BCC (black carbon copy) and Subjects are the fields. In genetic algorithm, header is irrelevant and only body part is taken into consideration. From the body of the mail, words are extracted. In the extraction of the word article like "a, an, the, for" and numerical numbers are discarded.

In genetic algorithm, first database is created which will classify spam and ham emails, and as per our choice database can be divided into several categories. It must be remember that as the size of the database increases, the number of word in the data dictionary also increases. The selection of categories depends on the classifications of the emails. However, if lesser number of categories is defined, still email can be identified as spam mail. However, the chances of false positive/negative increases. In our experiment we considered database of 2448 emails, out of which 1346 are SPAM mails and rest 1102 mails are HAM mails. In the data-dictionary 421 are considered which are divided into seven categories. The data dictionary is presented in appendix A. The procedure of calculating weights for a word of a particular group is detailed below:

Table 1 : Calculation of weights

Group	Word	Frequency	Normalized frequency of getting a word	Weight of word	Weight of group
C_1	Sex	113	0.268	0.102	0.062
C_1	Nude	23	0.055	0.021	
C_3	Free	694	1.648	0.63	0.391
C_3	Game	167	0.397	0.151	

Let's for an example an email consists of four words namely 'sex', 'nude', 'free' and 'game'. Out of these four words sex and nude belongs to categories C_1 and Free and Game belongs to categories C_3 (see Appendix -A). Let us consider an email with 1103 words, out of which 997 words are sex, nude, free and game. These words are taken so large in number to make sure that the considered mail is a spam mail as the spam database is very small as it contains only 421 words. The extracted words form the emails are first classify as whether they belongs to any spam database category. Once if words in email match word in spam data dictionary then the probability of getting a word from the spam database is obtained by dividing the frequency of a spam word by total number of words in data dictionary. In our case "nude' occurs 23 times, hence probability of getting 'nude' word is 23/421=0.268. The weight of the word (W_w) is calculated by

$$W_w = \frac{F_w / T_{WD}}{\sum p_W} \times \frac{S_{WM}}{T_{WM}}$$, where

F_w : Frequency of spam word

T_{WD} : Total word in data dictionary

S_{WM} : Total spam word in e-mail

T_{WM} : Total word in e-mail

$\sum p_W$: Probability of getting a word

The p_w f or the word 'sex' is

$$W_w = \frac{F_w / T_{WD}}{\sum p_W} \times \frac{S_{WM}}{T_{WM}}$$

$$W_w = \frac{113/421}{0.268+0.055+1.648+0.397} \times \frac{997}{1103}$$

$$W_w = 0.102$$

The weight of the category is calculated by taking the average of the category for example the weight of category C_1 is (0.102+ 0.021)/2=0.062. Then after normalization the weights are converted in the range of 0.000 to 1.000. And using the hex representation we have

The weight of the gene can be encoded as

Binary 0000000000 represents weight 0.000

Binary 0000000001 represents weight 0.001

Binary 0000000010 represents weight 0.002

……

…….

Binary 1111100111 represents weight 0.999

Binary 1111111000 represents weight 1.000

C_1	C_2	C_3	C_4	C_5	C_6	C_7
W_1	W_2	W_3	W_4	W_5	W_6	W_7

C_1	C_2	C_3	C_4	C_5	C_6	C_7
Sex Nude	✖	Free Game	✖	✖	✖	✖

C_1	C_2	C_3	C_4	C_5	C_6	C_7
0.062	0	0.391	0	0	0	0

Fig. 6 SPAM chromosomes prototype

Once, chromosomes are constructed for the incoming mails. The process of genetic algorithm starts and crossover takes place. As discussed above there are various ways by which cross-over can be performed. In crossover is only allowed for bit of gene in particular category only. In our algorithm, both multi-point and single point is done and positions of bits are selected randomly. In each generation of chromosomes only 12% are crossed. The next process is mutation, here to recover some of the lost genes or in our case it is done to recover some of the lost data, here only 3 % of genes are mutated.

The weight of the words of gene in testing mail and the weight of words of gene in spam mail prototype are compared to find the matched gene. If number of matched gene, is greater than or equal to three, than spam mail prototype will receive one score point. If the score point are greater than some threshold score points than the mail is considered as spam mail. However, the threshold point can be manually adjusted to get the appropriate results as we fixed it by doing experiments on 500 emails.

4. RESULTS

In this paper introductory results are produced by considering three mail prototypes. As in this method the body text is very important in the classifications of mail. We selected three different classes of e-mails.

Mail Prototype 1:The below mail is an example of SPAM mail.

Dear,Sir/Madam

It?s with every sense of humility, sincerity and fairness do I implore this medium to reach you at this time.

In the first place, my names are Vijay Patel. 28 Years of age. My ground parents migrated from India to the UK in 1932 and my parents and his siblings were all born here in the UK.

My father Mr. Dinesh Patel Died as a result of heart attack he heard after losing his Gold shop here in Birghmirgham during the UK Riot by the Angry street guys and claimed a lot of our belonging including lives and property.

This occurrence led me to talk with my father?s Lawyer over the Will of my father and he gave me a blue print which stated that I am the apparent heir of his Account with the HSBC BANK UK and at present, I don?t feel safe or secure anymore here in the UK. I deem it necessary for me to come to India which is the Country of my Fathers and settle down and also get married and settle instead of staying in the UK and peradventure lose the remaining inheritance willed to me by my late father.

I need an honest and truthful citizen of India who shall help me in area of Investment of my fortune which is the sum of Three Million, Seven hundred and Ninety Pounds My proposal is a profit oriented venture. Therefore, I do need your corporationand do update me with the norms that has to do with an investment like in Real Estate or founding and Academic Institution or any other venture that will be profit incline. Our sharing formula is negotiable though I have drafted it to be 70/30% in the profit sharing! Conversely, your utmost corporation is required since I am

ready to dispatch this fund as peace and order has been restored here in the UK. It?s necessary for us to work as a formidable team and build a business that will yield and better our future. Kindly reply me and do not fail to write your bio data shortly. Here is my contact number for easy and fast communication .I can also speak in hindi

Faithfully
Vijay Patel
EMAIL:vijaypatel945@yahoo.co.uk
MOBILE:+447014239568

As all of us know that such a mail are SPAM that are easily available in anybody mail box. The above email is tested with our generated system and the score point was 114.

Mail Prototype 2:The below mail is an example of HAM mail.

Dear Dr. Srivastava,

Associate Editor Gilberto Brambilla invites you to review this new submission to IEEE Photonics Technology Letters. If you are unable to review this manuscript, it would be greatly appreciated if you could please suggest alternative reviewers.

This is the abstract of the manuscript we would like you to review:

Agreed: http://mc.manuscriptcentral.com/ptl-ieee?URL_MASK=MRXPcGTstT9c5jDmYrmG

Declined: http://mc.manuscriptcentral.com/ptl-ieee?URL_MASK=bTGYddXDn8kfnb93J7Ph

Unavailable: http://mc.manuscriptcentral.com/ptl-ieee?URL_MASK=F9sqcXDdBD68Q9Tm2fT3

The site is located at:
http://mc.manuscriptcentral.com/ptl-ieee

Please reply to Sylvia Flores at s.j.flores@ieee.org with your answer as to whether or not you agree to review this paper (please do reply; we would rather have a "no" response than no response at all). If you agree to reviw it, you will receive an e-mail notice within a day instructing you to access the Manuscript Central website and download the paper. Thank you very much for your valuable service to the community.

Sincerely,

Gilberto Brambilla
Associate Editor
IEEE Photonics Technology Letters

The above email is tested with our generated system and the score point was zero. Our proposed algorithm treats this mail as a HAM mail. Indeed it is a HAM mail.

Mail Prototype 3: The below mail is an example of false positive mail.

Congratulation!! dear winner, we are using this medium to officially notify you: open the attachment in your mail box fill the form and send it back to US.nokiaclaimdept2013@live.co.uk

Regards
Dr. Darwin Payton
Event Manager
TEL: (+44) 7017048564

The above email is tested with our generated system and the score point was zero. Our proposed algorithm treats this mail as a HAM e-mail. However, it is a SPAM mail. Hence, this is an example of false positive.
The above email is tested with our generated system and the score point was zero. Our proposed algorithm treats this mail as a HAM e-mail. However, it is a SPAM mail. Hence, this is an example of false negative. This is happening because in our data-dictionary the work like 'congratulation', 'winner', 'claim' are not present.
As stated above, Genetic Algorithms do not work well when the population size is small and the rate of change is too high. As we have taken only 421 words dictionary, hence population size is very small, and the rate of change will be very high as e-mail types are countless.
We did this experiment again by adding these words 'congratulation', 'winner', 'claim' in data dictionary and we found that our system works well now with score point 4, and treated this mail as SPAM mail.
In our early results we found that, if number of words in the mail is larger, then more correct classification is possible. We have checked our algorithm on large corpus of 2248 mails out of which 1346 were SPAM mails and rest of them were HAM mails.. The results on such a large email corpus are taken into account to see more accurate classifications of mail and effectiveness of GA algorithm. However, we did this experiments on the high end machine to get more clear and accurate picture of the GA. In our experiments we found that the nearly 82% mails are correctly classified by our method. The score point varies from 4 to 137; however, it can go further beyond 137 depending on the number of words in the e-mail. In the future work, the in-depth analysis of the GA parameters and size of spam database on SPAM filtering is presented.

5. CONCLUSION

In this paper, genetic algorithm is presented in detailed and it has been discussed how GA can be beneficial in SPAM email

classifications. In genetic algorithm, first database is created which will classify spam and ham emails, and as per our choice database can be divided into several categories. It must be remember that as the size of the database increases, the number of word in the data dictionary also increases. The selection of categories depends on the classifications of the emails. However, if lesser number of categories is defined, still email can be identified as spam mail. However, the chances of false positive /negative increases. Many experiments have been performed to fix some of the important parameters of GA. The fitness function is selected very carefully using doing a set of experiments. The proposed idea has been tested on 2248 mails and the overall efficiency is nearly 82%.

APPENDIX-A

Database [5][6]

Group	Content	Example of keywords in each group
C1	Adult	adult, aphrodisiac, big, cam, climax, company, cum, desire, erotic, fantasy, fuck, gay, girl, greate, guy, hard, hardcore, heaven, hot, huge, long, man, max, maxlength, nude, etc.
C2	Financial	Account, accountant, alert, analyst, attorney, bank, bankruptcy, benefit, bill, billing, broker, budget, building, cash, cheque, commission, consolidate, court, credit, creditor, currency, customer, deposit, etc.
C3	Commercial	college, commerce, computer, cost, deliver, discount, especial, expensive, express, fantastic, free, furnishing, furniture, game, gif, gift, great, guarantee, inexpensive, etc.
C4	Beauty and diet	after, age, amaze, anti-aging, appetite, beauty, become, before, believe, blood, body, botanic, breast, build, burn, Diet calorie, capsule, card, cell, change, chemical, cholesterol, confirm, course, diet, difference, dose, drug, effect, effective, eliminate, energy, enhance, exercise, eye, face, fast, etc.
C5	Traveling	book, deluxe, excite, guide, holiday, honest, hotel, luxury, meal, package, plan, problem, relax, relief, reserve, resort, summer, temple, ticket, tour, train, travel, traveler, trip, vacation,
C 6	Home-Based	address, astonishment, base, broadcast, bulk, business, comfort, connect, demo, domain, downline, download, Business earn, email, emailing, ethernet, facemail, fresh, home, homebased, homeworker, host, income, interest, international, etc.
C7	Gambling	action, award, bet, bonus, casino, challenge, extra, gambling, gold, hunt, las, lucky, millionaire, player, poker, prize, reward, rich, vegas, win, lottery, etc.

REFERENCES

[1] http://www.kaspersky.com/about/news/spam/2013/Spam_in_201 Continued_Decline_Sees_Spam_Levels_Hit_5_year_Low

[2] Blanzieri E. and Bryl A. 2008. A Survey of Learning -Based Techniques of Email Spam Filtering, Conference on Email and Anti-Spam.

[3] Koprowski G. J. 2006. Spam accounts for most e-mail traffic, Tech News World. Available: http://www.technewsworld.com/story/51055.html

[4] Tang K.S. et.al. 1996. Genetic Algorithm and Their Applications, IEEE Signal Processing magazine, pp.22-37.

[5] Sanpakdee U. et.al. 2006. Adaptive Spam Mail Filtering Using Genetic Algorithm.

[6] Spam Assassin, http://spamassassin.org.

AUTHORS

Jitendra N. Shrivastava received his Master of Technology (M.Tech) degree in Information Technology from Indian Institute of Information Technology (IIITA), Allahabad, India in 2007. 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 a Professor in North Orissa university, 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.