I.J. Intelligent Systems and Applications, **2014, 02, 54-60**
Published Online January 2014 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijisa.2014.02.07

E-mail Spam Filtering Using Adaptive Genetic Algorithm

Jitendra Nath Shrivastava
Research Scholar, Singhania University, Jhunjhunu, Pacheri Bari, Rajasthan, India
E-mail : jitendranathshrivastava@yahoo.com

Maringanti Hima Bindu
Deptt.of Computer Science and Applications, North Orissa University, India
E-mail : mhimabindu@yahoo.com

***Abstract*—** Now a day's everybody email inbox is full with spam mails. The problem with spam mails is that they are not malicious in nature so generally don't get blocked with firewall or filters etc., however, they are unwanted mails received by any internet users. In 2012, more that 50% emails of the total emails were spam emails. In this paper, a genetic algorithm based method for spam email filtering is discussed with its advantages and dis-advantages. The results presented in the paper are promising and suggested that GA can be a good option in conjunction with other e-mail filtering techniques can provide more robust solution.

***Index Terms*—** Spam Filtering, Genetic Algorithm, SPAM and HAM

I. Introduction

Spam is any unsolicited email sent against the interest and knowledge of the recipient, usually with no intention of a response other than to visit a website or sell a product. In general these are broadcast messages send to a large number of peoples. However, it is important to differentiate between unsolicited email, which can be labeled as Spam and solicited email. Solicited email may have the same goals as unsolicited email, but you may receive a solicited email that the sender has deemed to be in your interest, or related to a previous interest. Spam email, however, is usually sent without any knowledge or consideration of the recipients interests, and is sent out only with the desired result in mind. Spams are not only wastage of money, bandwidth also very annoying for the users [1].

Recently, as per the Kaspersky Lab, the share of spam in email traffic decreased steadily throughout 2012 to hit a five-year low and the average for the year stood at 72.1% which is - 8.2 percentage points less than in 2011 (Fig.1) [2]. However, such a prolonged and substantial decrease in spam levels is unprecedented and spam will hold substantial part of email received [2]. However, the proportion of emails with malicious attachments remains as it is and fell only slightly to 3.4% [2]. This is a very large proportion, here; only emails with malicious attachments are considered and ignore other spam emails containing links to malicious websites. Previously malicious users relied on fake notifications from web hosting services, social networking sites, delivery services like courier etc., and messages from government and non-government organizations [2]. However, in 2012, they expanded their repertoire to include fake messages from a variety of airlines/train, hotel/resurgent reservation services, and various coupon services [2].

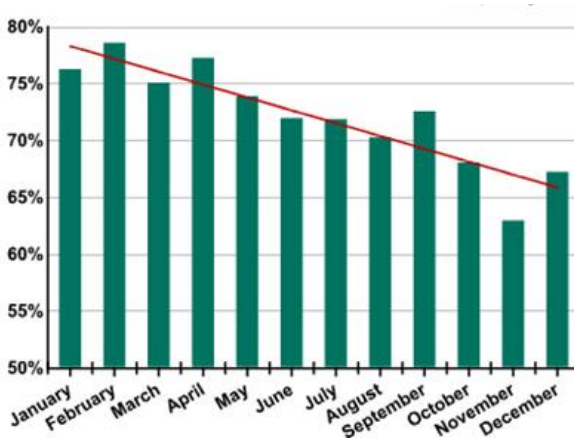

Fig. 1: E-mail spam trend in 2012 as per Kaspersky Lab [fig. source [2]]

1.1 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 [2]. 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 [2].

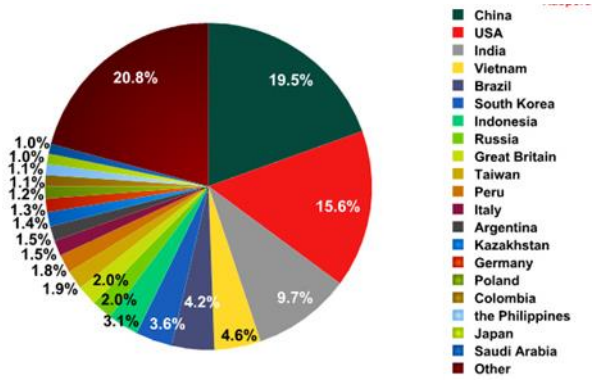

Fig. 2: countrywide distribution of the SAPM generation [fig. source [2]]

To fight spam, various spam filtering techniques are adopted. However, each scheme has its advantages and disadvantages, and in a nut shell none of them is very effective. The summary of the some of the techniques are detailed below:

1.2 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 [3].

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.

1.3 Bayesian Classifier

Particular words have particular probabilities of occurring in spam email and in legitimate emails [4]. 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.

1.4 Support Vector Machine (SVM)

'Support Vector Machines' [5][6][7] 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 (ham) or RED (spam). 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).

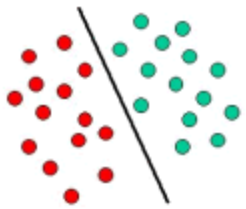

Fig. 3: Support Vector Machine

1.5 Content Based Spam Filtering Techniques

Neural networks are the best candidates for problems of classification [8][9][10][11]. The basic idea used in ANN is to create a word list w of the n most frequently used words in certain parts (initially just the message body) of the training corpus of spam and real messages. The elements i_n of the ANN input vector i for a given message are then derived as follows. If the email in question contains m_n instances of the word in the nth position of the global wordlist, and if the email in question has l words, then i_n will be set to m_n/l. The input vector is thus simply a representation of the presence or absence of the words in the global word list, weighted by the length of the message itself.

A "word" in the above context is defined as any series of upper or lower case alpha characters greater than three characters and less than x (generally an integer around 15) characters in length. When the words are placed into the wordlist, they are converted to all lower case, and all comparisons done against the wordlist are subsequently done in lower case.

In this scheme, no consideration of the proximity of certain word combinations is made. For example, the word "make" occurring close to "money" occurring close to "fast" would be a strong indication that the

email is a spam, but the technique would not pick up on the fact that they occurred close together.

These inputs were fed into a simple three-layer (input, hidden, and output) fully connected conventional artificial neural network. The network was trained for a certain number of training epochs for every input in the training set, with the weights adjusted using back-propagation.

Finally, it is not known how much the performance of the static wordlist approach described here would degrade classification over time. The results presented here might be artificially high because the wordlist was compiled from spam and real messages that sample the entire *timespam* of all messages in all data sets (training, validation, and test) [12]. As trends in spamming change over time, and as topics of conversation in a user's real emails vary over time, we might expect fewer words from future emails to match the wordlist, and consequently discrimination capability for the spam ANN might decline.

As stated above ANN technique is a good classifier technique. However, in SPAM filtering it success is very limited [13].

However, none of the above techniques stated above provide robust solution. This happens, because the structure of spam emails is changing continuously. To counteract such a problem, an adaptive technique is required. In this paper, we have used a Genetic algorithm for spam classification. The power of genetic algorithm lies in the fitness function, and incorrect fitness function will lead to wrong solution. In spam classification, the identification of fitness function is not easy, hence we did experiment, on 500 emails then we fixed our fitness function. In the next section, the genetic algorithm and its overview are presented.

The rest of the paper is organized as follows. In section II, idea of Genetic Algorithm based e-mail Spam classifications is presented. The genetic algorithm steps are detailed in section III of the paper. In section IV, genetic algorithm based e-mail classification is presented. The obtained results are presented in section V of the paper. Finally, major conclusions of the paper are discussed in the section VI of the paper.

II. Genetic Algorithm Based E-mail Spam Classifications

Genetic Algorithms can identify and exploit regularities in the environment, and converges on solutions (can also be regarded as locating the local maxima) that were globally optimal [14]. This method is very effective and widely used to find-out optimal or near optimal solutions to a wide variety of problems. Genetic algorithm does not impose any limitations required by traditional methods such as gradient descent search, random search etc. The Genetic Algorithm technique has many advantages over traditional non-linear solution techniques. However, both of these techniques do not always achieve an optimal solution. However, GA provides near optimal solution easily in comparison to other methods. The GA is very different from "classical" optimization algorithms-

a) It does the encoding of the parameters, not the parameters itself.

b) The search is more elaborative in a given amount of time.

c) As GA is probabilistic in nature, it may yields "different solutions on different set of simulations". To get an optimal solution Monte Carlo methods can be adopted

d) The solution space in multiple directions instead of in single direction.

Limitations:

Although because of its simplicity and classiness, Genetic Algorithm has proven themselves as efficient problem solving strategy. However, GA cannot be considered as universal remedy. Some limitations of GA are:

1) The method chosen for representing any problem must be strong and firm, it must withstand random changes or otherwise we may not obtain the desired solution.

2) In Genetic Algorithm, the Fitness function must be chosen very carefully. It should be able to evaluate correct fitness level for each set of values. If the fitness function is chosen poorly, then Genetic Algorithm may not be able to find an optimal solution to the problem, or may end up solving the wrong results.

3) Genetic Algorithms uses random parameter selection, hence it will not work well when the population size is small and the rate of change is too high.

4) In Genetic Algorithm solution is comparably better with, presently known solutions; it cannot make out "the optimum solution" of its own.

5) Sometimes over-fit of the fitness function abruptly decreases the size of population, and leads the algorithm to converge on the local optimum without examining the rest of the search space. This problem is also known as "Premature Convergence".

III. Genetic algorithms Steps

The details of how Genetic Algorithms work are explained below [14][15] [16].

3.1 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.

3.2 Reproduction

There are two kinds of reproduction: generational reproduction and steady-state reproduction.

3.2.1 Generational Reproduction

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.

3.2.2 Steady-state Reproduction

In this method, two chromosomes are selected and cross-over are performed and one or two children are produced. In some cases mutation is also applied and after crossover and mutation the newly generated offspring are then again added to the original population, and thus after some iterations old generation dies out.

3.3 Parent Selection Mechanism

In general probabilistic method is used for the parent selection. This process is stochastic in nature however it does not imply GA employs a directionless search. In general, the chance of each parent being selected is related to its fitness [17].

3.3.1 Fitness-based selection

The standard, original method for parent selection is Roulette Wheel selection or fitness-based selection. In this kind of parent selection, each chromosome has a chance of selection that is directly proportional to its fitness. The effect of this depends on the range of fitness values in the current population.

3.3.2 Rank-based selection

In the rank-based selection method, selection of a chromosome's is probabilistic and is based on relative rank or position in the population, rather than absolute fitness.

3.3.3 Tournament-based selection

The tournament based selection is to choose N parents in random manner and finally returns the fittest one of these parents.

3.4 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.

3.4.1 One-point Crossover

In one point cross-over, a bit position is randomly selected that need to change. In this process, a random number is generated which is a number (less than or equal to the chromosome length) as the crossover position. Here, the bits before the number keep unchanged and swap the bits after the crossover position between the two parents.

3.4.2 Two-point Cross Over

The two point cross-over, is similar to that of one-point crossover except that here two positions are selected and only the bits between the two positions are swapped. This crossover method preserves the first and the last parts of a chromosome and just swaps the middle part.

3.4.3 Uniform 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.

3.5 Inversion

Inversion operates as a type of reordering technique. As its name suggest, it operates on a single chromosome and inverts the order of the genes between two randomly chosen points on the chromosome. This operator is inspired by a natural biological process; hence some additional overhead is required.

3.6 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.

IV. E-mail Filtering Process

The current means of filtering technology is mainly divided into two types, 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, automatically set up and update new features and rules of the spam, using the improved Genetic Algorithm to the design of e-mail filters are the innovations. Genetic algorithm can be 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:**

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

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. Hence, we defined our fitness function as

$$F = \begin{cases} 1 & \text{Score point } \geq 3 \\ 0 & \text{Score point } < 3 \end{cases} \quad (2)$$

4.1 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 [18][19]. 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, but the chances of false positive/negative increases. 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.

V. Results

In this paper introductory results are produced by considering four 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 recipient,

Avangar Technologies announces the beginning of a new unprecendented global employment campaign. Due to company's exploding growth Avangar is expanding business to the European region. During last employment campaign over 1500 people worldwide took part in Avangar's business and more than half of

them are currently employed by the company. And now we are offering you one more opportunity to earn extra money working with Avangar Technologies.

We are looking for honest, responsible, hard-working people that can dedicate 2-4 hours of their time per day and earn extra £300-500 weekly. All offered positions are currently part-time and give you a chance to work mainly from home.

Please visit Avangar's corporate web site (http://www.avangar.com/sta/home/0077.htm) for more details regarding these vacancies.

"bespeakplur"

The above email is tested with our generated system and the score point was 5. Our proposed algorithm treats this mail as a SPAM e-mail as it is giving too much money for part time job.

Mail Prototype 2

The below mail is an example of HAM mail.

Dear Dr. Srivastava

You may have received several emails regarding the NCC2013 paper review request. We will highly appreciate if you please accept and complete the review as early as possible. If you have any difficulties, please delegate the review to some colleague who would be able to do the review in next 7 days. Regards,

Swades De, Aaditeshwar Seth

NCC 2013 Networks Symposium Co-Chairs

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 [20][21], 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. We run our code 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.

VI. Conclusion

In this paper, a Genetic Algorithm based e-mail spam classification algorithm is presented. In this paper some basic results are presented. This algorithm successfully distinguishes spam and ham e-mails. The efficiency of the process depends on the dataset and GA parameters. The efficiency of the algorithm is more than 82%. In the future work, some advanced results that relate with the characterization of the GA parameters will be presented, and how the false positive/negative results can be minimized. Hence, in the general GA in conjunction with other e-mail filtering techniques can provide more accurate SAPM filtering technique.

References

[1] Enrico Blanzieri and Anton Bryl, "A Survey of Learning-Based Techniques of Email Spam Filtering" Conference on Email and Anti-Spam, 2008.

[2] http://www.kaspersky.com/about/news/spam/2013/Spam_in_2012_Continued_Decline_Sees_Spam_Levels_Hit_5_year_Low

[3] Youn, Seongwook, and Dennis McLeod, 2007, A Comparative Study for Email Classification, Advances and Innovations in Systems, Computing Sciences and Software Engineering, pp. 387-391

[4] Liu Pei-yu, Zhang Li-wei and Zhu Zhen-fang, "Research on Email Filtering Based on Improved Bayesian", Journal of Computers, v4, 2009, pp. 271-275.

[5] C. Cortes and V. Vapnik, "Support-vector networks. Machine learning", v20, 1995, pp. 273–297.

[6] N.Cristiatnini and J. Shawe-Taylor, "An introduction to Support Vector Machines and Other Kernel-Based Learning Methods," Cambridge University Press, 2003. http://www.support-vector.net.

[7] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery," v2, 1998, pp.121–167.

[8] H. T. Siegelmann and E. D. Sontag, "On the computational power of neural nets. Journal of Computer and System Sciences," v50, 1995, pp.132–150.

[9] W. S. McCulloch and W. H. Pitts, "A logical calculus of the ideas immanent in nervous activity"

[10] F. Rosenblatt, "Principles of Neurodynamics," Spartan Books, Washington, 1958.

[11] M. Basavaraju and Dr. R. Prabhakar, 2010, A Novel Method of Spam Mail Detection using Text Based Clustering Approach, IJCA, pp. 15-25.

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

[13] Rich Drewes "An artificial neural network spam classifier", Project home page: www.interstice.com/drewes/cs676/spam-nn.

[14] L. Zhang, J. Zhu, and T. Yao, "An evaluation of statistical spam filtering techniques" ACM Transactions on Asian Language Information Processing (TALIP), v3, 2004, pp.243–269.

[15] P. Ferragina and R. Grossi, "Improved dynamic text indexing," J. Algorithms, v31, 1999, pp. 291–319.

[16] J. K¨arkk¨ainen and E. Ukkonen. Lempel-zivparsing and sublinear, "Size index structures for string matching" in Proc WSP'96, Carleton University Press, 1996, pp. 141-155.

[17] Goldberg, D. E., and Deb, K. 1991. A comparative analysis of selection schemes used in genetic algorithms. Foundations of Genetic Algorithms. Morgan Kaufmann.

[18] Koza, J. R. 1992. Genetic Programming: On the Programming of Computers by Means of Natural Selection.MIT Press.

[19] Koza, J. R. 1994. Genetic Programming II: Automatic Discovery of Reusable Programs. MIT Press.

Authors' Profiles

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, hima.bindu@jiit.ac.in

How to cite this paper: Jitendra Nath Shrivastava, Maringanti Hima Bindu,"E-mail Spam Filtering Using Adaptive Genetic Algorithm", International Journal of Intelligent Systems and Applications(IJISA), vol.6, no.2, pp.54-60, 2014. DOI: 10.5815/ijisa.2014.02.07

