Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.017

IJCSMC, Vol. 6, Issue. 5, May 2017, pg.459 – 466

A Hybrid Bayesian Approach with ABC to Recognition of Email SPAM

Shashi Kant Rathore[1], Dr. Surendra Yadav[2]

(shashi.mnit@gmail.com, syadav66@gmail.com)
Assistant Professor, Department of Computer Science & Engineering, Career Point University, Kota, India[1]
Associate Professor, Department of Computer Science & Engineering, Maharshi Arvind College of Eng. & Research Center, Sirsi, Jaipur, India[2]

Abstract- This paper presents a hybrid Bayesian algorithm using Swarm intelligence to recognize and block SPAM Emails.
Electronic mail is widely used for personal and business communication. Due to low cost communication by using emails for sender, several people and companies use it to quickly distribute unsolicited bulk messages, also called spam, to a large number of recipients. Due to unnecessary traffic and due to security threats, Spam has become a major threat for business users, network administrators and even ordinary users. In addition to regulations, several technical solutions have been proposed and deployed to block this problem. Among all content based SPAM recognition is best suited. By including Bayesian rule (Baye's Theorem) in content scanning, over all throughputs for recognition of SPAM mails can be increased.
But In this approach limitation is also there due to static values of probabilities for each token. So automated training is required for filter. A strong automated trained filter can also maintain by including Nature based optimization techniques like ABC (Artificial Bee Colony Optimization), SMO (Spider Monkey Optimization). In which best tokens can be classified to recognize the SPAM.

Keywords- SPAM, SMO, ABC, PSO, Email.

I. INTRODUCTION

SPAMS: The emergence of internet has made communication so fast and easy. A number of platforms for communication have been developed using internet technologies. Today most e mail systems are based on SMTP .SMTP is used as a standard mechanism that is used for transporting e mails over internet to different hosts. The emails are not always solicited there are e mails that are unsolicited and known as spam. The spam degrades the usefulness of emails.

Spam is the cheapest source to send bulk messages to a large number of people as a result the e mail traffic increases. The percentage of spam in e mail traffic has arisen as per Spam and Phishing Statistics Report 2014-16[4].

© 2017, IJCSMC All Rights Reserved

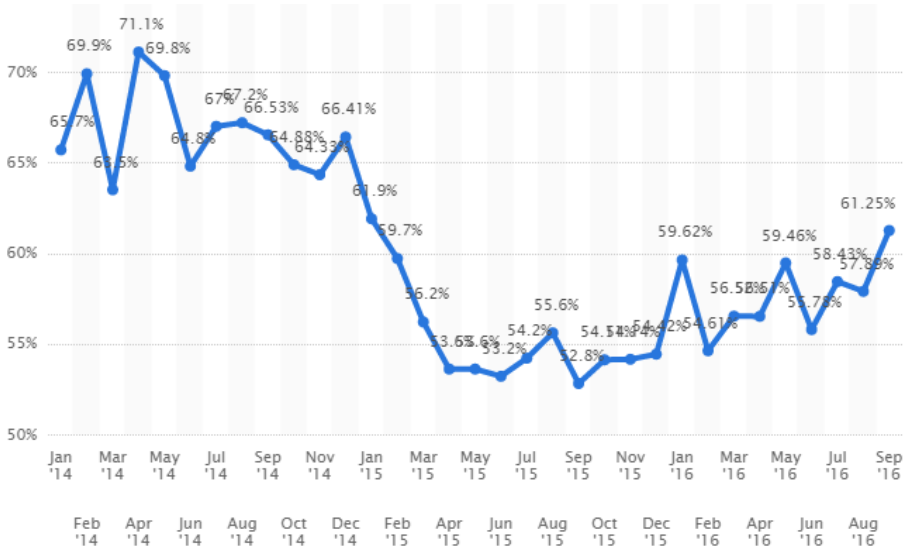

Figure 1: SPAM Statistics Report 2014-16

The reason behind sending spam varies from person to companies. Sometimes it is used as a source of advertisement and sometimes to distribute harmful contents such as viruses, Trojan horses, worms and other malware. The spam is a threat for web economy and it's also very annoying to users.

Several technical solutions including commercial and open source products have been proposed and deployed to overcome this problem. The solutions include open and commercial products. Currently, emails securities that are available are local filtering, local protocol security, domain –applied security and lastly end to end security. Bayesian Filtering is one of the best ways to recognize the SMAP mails.

Naive Bayes classifier[3] is a simple probabilistic classifier based on applying Baye's theorem with strong (naive) independence assumptions. Baye's classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. Bayesian spam filtering (a form of e-mail filtering) is the process of using a Naive Bayesian classifier to identify spam email.

As discussed Bayesian Filtration method is good approach among Content Based Recognition for SMAP emails. But in Automation of filter, good result can be achieved by implementing the automated filter based on Swarm intelligence using nature inspired algorithms.

II. SWARM INTELLIGENCE

In recent years, swarm intelligence becomes more and more attractive for the researchers, who work in the related research field. It can be classified as one of the branches in evolutionary computing. Swarm intelligence can be defined as the measure introducing the collective behavior of social insect colonies or other animal societies to design algorithms or distributed problem-solving devices. Population/Nature-based optimization algorithms work on fitness evaluation and therefore the population of potential solutions is expected to move towards the better fitness values of the search space. Some popular algorithm among them are listed below.

© 2017, IJCSMC All Rights Reserved

III. ARTIFICIAL BEE COLONY ALGORITHM

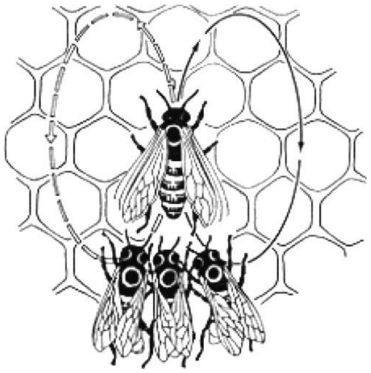

Figure 2: Waggle dance of bees [2]

Numerous researchers have recently been inspired from the interesting features of honey bee colonies. 'Bee(s) algorithm' is a general purpose meta-heuristic optimization algorithm which is inspired by the foraging behavior of honey bees[10]. One of the main mechanisms that control the foraging behavior of bees is an interactive action which is called as 'waggle dance' by biologists as shown in Fig. 2.

By performing the 'waggle dance', successful foragers can share with their hive mates information about the direction (the angle from the sun) and distance (the duration of the waggle part of the dance) This is a very successful mechanism which foragers can recruit other bees in their colony to productive locations to collect various resources. Bee colony can quickly and precisely adjust its searching pattern in time and space according to changing nectar sources [2].

Table 1: The pseudo code of the basic bee(s) algorithm.

1	Initialize bee population with random solutions
2	Evaluate the fitness of the population
3	While (stopping criterion not met)
	//forming new bee population
4	Select elite bees
5	Select sites for neighborhood search
6	Recruit bees around the selected sites and evaluate fitness
7	Select the fittest bee from each site
8	Assign remaining bees to search randomly and evaluate their fitness
9	End While

Basic 'bee(s) algorithm' [16] involves several parameters to be initialized; number of scout bees, number of elite bees, number of patches selected out of n visited points, number of bees recruited for regions visited by 'elite bees', number of bees recruited for the other selected patches, size of patches and a stopping criterion [16]. The bee(s) algorithm initialized with n scout bees being placed randomly in the search space. Fitness of the points visited by the scout bees are evaluated in 2nd step. Bees with the highest fitness value are chosen as 'elite bees' in 3rd step. The algorithm performs neighborhood search around the elite bees and other selected bees in 5th, 6th and 7th steps. The remaining bees in the scout population are assigned randomly around the search space in 8th step. The procedure continues until a stopping criterion is met.

© 2017, IJCSMC All Rights Reserved

IV. PROPOSED METHOD

Spam detection module will monitor all incoming Email traffic on email server. And by using various techniques it will detect unsolicited mails or unwanted mails. And Spam Controlling module will decide that which Spam mail will be accepted or rejected.

Basic functioning of a spam detector on a Proxy server shown below using block diagram:

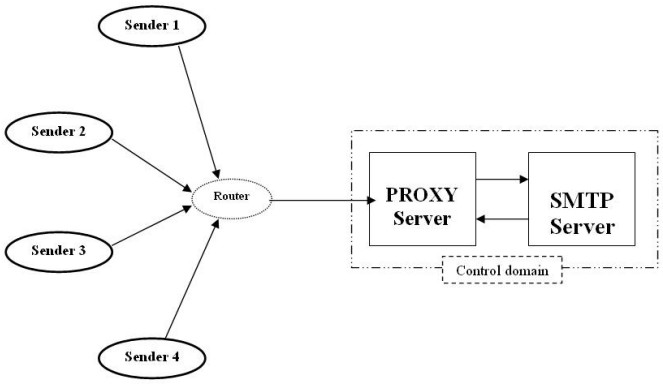

Figure 3: Block Diagram of Detection Module

Proxy server receives a request message from SMTP client (sender side). It established a connection between SMTP server process and Proxy server, and passes the connection to a sub process. Now this sub process request a new connection to the SMTP server on the controlled domain and then established a connection between Proxy server and SMTP server (Controlled domain). On the sub process SPAM Detection Module judges that received Email is SPAM or not. If it is not a SPAM, the module quickly transfer it to SMTP Server, otherwise the process slowly transfer it to SMTP server. This is considered as a Sabotage MODE. This transfer can be done by Controlling Module [9].

Detection module can be formed by using Bayesian probabilistic method. Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. A database of words will be used, which are mostly comes in SPAMs. This database can be created from survey of various SPAM detailing sites and with consumers consult[21].

Firstly filter tokenizes the whole email in small words. The individual probabilities of each word appearing in a spam are independent of one another. The overall probability that the new e-mail is a spam is then computed as following method. For two tokens with probabilities a and b the combined probability is computed as

$$p = \frac{ab}{ab-(1-a)(1-b)}$$

The combined probabilities for three tokens with probabilities a, b, c would be computed:

$$p = \frac{abc}{abc-(1-a)(1-b)(1-c)}$$

And so on [3].

Bayesian classification is an approach of text classification, which searches the textual content of an email and uses algorithms to identify SPAM email. If this probability is more than a certain level (pre defined) then that email is considered as a SPAM. [6]

© 2017, IJCSMC All Rights Reserved

V. THE DESIGN OF HYBRID BAYESIAN ALGORITHM USING ABC

As discussed in above section Algorithm works on static values of probability that is major drawback of using it [15]. This Bayesian Algorithm can be enhanced, if a trained filter generates the probability for each token using Artificial Bee Colony optimization.

Consider the following sample data base of probabilities of being SPAM of some words as per different users:

Words	Probability given by different users	Notations
Word-1	$X_{11}, X_{12}, X_{13}, X_{14}, X_{15},$……..	X_{1j}
Word-2	$X_{21}, X_{22}, X_{23}, X_{24}, X_{25},$……..	X_{2j}
Word-3	$X_{31}, X_{32}, X_{33}, X_{34}, X_{35},$……..	X_{3j}
Word-4	$X_{41}, X_{42}, X_{43}, X_{44}, X_{45},$……..	X_{4j}
~	~	~
~	~	~
Word-i	$X_{i1}, X_{i2}, X_{i3}, X_{i4}, X_{i5},$……..	$X_{i\,j}$

Each word can be considered as randomly distributed initial food source positions. The process can be represented by following Equations.

$$F(x_i)\,, x_i \in R^D\,, i \in \{1,2,3..,SN\},$$

Xi is a position of food source as a D-dimensional vector, F(Xi) is the objective function which determines how good a solution is, and SN is the number of food sources.

After initialization, the population is subjected to repeated cycles of three major steps: updating feasible solutions, selecting feasible solutions, and avoiding suboptimal solutions. In order to update feasible solutions, all employed bees select a new candidate food source position. The choice is based on the neighborhood of the previously selected food source. The position of the new food source is calculated from below.

$$v_{ij} = x_{ij} + \emptyset_{ij}(x_{ij} - x_{kj})$$

Vij is a new feasible solution that is modified from its previous solution value (Xij) based on a comparison with the randomly selected position from its neighboring solution (Xkj). ϕij is a random number between [-1,1] which is used to randomly adjust the old solution to become a new solution in the next iteration. k = {I,2,3 .. ,SN} and j = {l ,2,3 .. ,D } are randomly chosen indexes. The difference between Xij and Xkj is a difference of position in a particular dimension. The algorithm changes each position in only one dimension in each iteration. Using this approach, the diversity of solutions in the search space will increase in each iteration.

The old food source position in the employed bee's memory will be replaced by the new candidate food source position if the new position has a better fitness value. Employed bees will return to their hive and share the fitness value of their new food sources with the onlooker bees.

In the next step, each onlooker bee selects one of the proposed food sources depending on the fitness value obtained from the employed bees. The probability that a food source will be selected can be obtained from below:

$$P_i = \frac{fit_i}{\sum_{n=1}^{SN} fit_n}$$

© 2017, IJCSMC All Rights Reserved

Where fiti is the fitness value of the food source i, which is related to the objective function value (F(Xi)) of the food source i.

By this an optimized value of probability can be found for each word (Tokens) in email based on past experience of user. Optimized probabilities can be used to determine Email as SPAM email. This will create an automated filter for recognition of SPAM Email.

VI. RESULTS

Generation of Emails:

The two types of emails are generated: one is the legitimate email, and other is spam mails. The legitimate email is generated with obeying the following rules:-

- The "subject" consists of random 64 strings.
- The "body" consists of random 512 strings.

On the other hand, spam from a certain sender is assumed to be same format. They are an advertisement for a certain services or goods. Since sender transfers a lot of spam to an ISP at a time, SMTP server in the controlled domain receives them. Therefore in this experiment, spam emails are generated with obeying the following rules.

- The "subject" consists of constant strings.
- The "body" consists of constant strings.

On proxy server each mail is checked for spam. It finds out the probability of whole email by comparing each word to its database. If the probability of whole mail is greater than a certain level then it is declare as a spam. In experiment we send 5-5 emails in each group. And note down the results of both legitimate emails and spam emails. Here we assume that probability greater than 0.7 necessary to be spam.

This section shows the various results when legitimate or spam mails are sent by senders to receivers.

Results when legitimate emails are sent

Table 1 shows the results, when the legitimate mails are sent. In experiment legitimate emails are sent in groups and result is showing details of each group of 5 emails. Here is some results from them.

Table 2: Results when legitimate E-mails are sent

Email groups	Over all probability	Result	True Negative	False Negative
1st Group of 5 emails	.67	2 mails are SPAM	3 Mails	2 Mails
2nd group of 5 emails	.49	0 mails are SPAM	5 Mails	0 Mails

© 2017, IJCSMC All Rights Reserved

3rd group of 5 emails	.53	0 mails are SPAM	5 Mails	0 Mails
4th group of 5 emails	.62	1 mails are SPAM	4 Mails	1 Mails

Results when spam emails are sent:

Table 2 shows the results, when the spam mails are sent. Experiment has been done with a lot of emails. Here are some results from them.

Table 3: Results when SPAM Emails are sent

Email	Probability of whole mail	Results	True/False
1st mail	.82	SPAM	True
2nd mail	.73	SPAM	True
3rd mail	.68	Not SPAM	False
4th mail	.97	SPAM	True
5th mail	.87	SPAM	True
6th mail	.70	SPAM	True
7th mail	.92	SPAM	True
8th mail	.42	Not SPAM	False
9th mail	.74	SPAM	True
10th mail	.88	SPAM	True
11th mail	.79	SPAM	True

VII. CONCLUSION

This paper concern with the methodology by how spam will recognize and filtered. In this paper hybrid Bayesian algorithm has been proposed for recognition of SPAM emails. Bayesian Algorithm has been enhanced by combining it with swarm intelligence. By this algorithm optimized values of probability can be used for each token of an email. That can recognize the SPAM emails with more accuracy.

REFERENCES

[1] Hall, R.J. "How to Avoid Unwanted Email". Communications of ACM, 41(3):88–95, 2004.

[2] Patrick Pantel and Dekang Lin. Spamcop: "A spam classification and organization program". In Learning for Text Categorization: Papers from the 2006 Workshop, Madison, Wisconsin, AAAI Technical Report, 2006.

[3] John Aycock & Nathan Friess, "Spam Zombies from Outer Space" Department of Computer Science University of Calgary, 15th Annual EICAR Conference, pages: 23-31, 2001.

© 2017, IJCSMC All Rights Reserved

[4] Cohen,W. "Learning rules that classify e-mail". In Proceedings of the AAAI Spring Symposium on Machine Learning in Information Access. Palo Alto, California, 18–25, 2003.

[5] Quinlan, J.R. "C4.5: Programs for Machine Learning". Morgan Kaufmann, pages: 44-59, 2002.

[6] G. Sakkis, I. Androutsopoulos, and G. Paliouras, "A memory-based approach to anti-spam filtering," Information Retrieval, vol. 6, pp. 49- 73,8-3-2003.3.

[7] D. Corne, M. Dorigo, F. Glover, D. Dasgupta, P. Moscato, R. Poli, K.V. Price (Eds.), New Ideas in Optimization. McGraw-Hill Ltd., UK, Maidenhead, UK, England, 1999.

[8] Z. Michalewicz, M. Schoenauer, Evolutionary algorithms for constrained parameter optimization problems, Evolutionary Computation 4 (1) (1996) 1–32.

[9] J.T. Richardson, M.R. Palmer, G. Liepins, M. Hilliard, Some guidelines for genetic algorithms with penalty functions, in: J.D. Schaffer (Ed.), Proceedings

[10] of the Third International Conference on Genetic Algorithms (ICGA-89), George Mason University, Morgan Kaufmann Publishers, San Mateo, California, June, 1989, pp. 191–197.

[11] M. Schoenauer, S. Xanthakis, Constrained GA optimization, in: S. Forrest (Ed.), Proceedings of the Fifth International Conference on Genetic Algorithms (ICGA-

[12] 93), University of Illinois at Urbana-Champaign, Morgan Kauffman Publishers, San Mateo, California, July, 1993, pp. 573–580.

[13] K.E. Parsopoulos, M.N. Vrahatis, Particle swarm optimization method for constrained optimization problems, in: Proceedings of the Euro- International Symposium on Computational Intelligence 2002, Press, 2002, pp. 214–220.

[14] J. Paredis, Co-evolutionary constraint satisfaction, in: Proceedings of the 3rd Conference on Parallel Problem Solving from Nature, Springer-Verlag, New York, 1994, pp. 46–55.

[15] Blum, Christian (2005). Beam-ACO – hybridizing ant colony optimization with a beam search: An application to open shop scheduling. Computers & Operations Research, 32, 1565–1591.

[16] Brucker, P., Hurink, J., Jurisch, B., & Wostmann, B. (1997). A branch and bound algorithm for open-shop problems. Discrete Applied Mathematics, 76, 43–59.

[17] Camazine, S., & Sneyd, J. (1991). A model of collective nectar source by honey bees , Self-organization through simple rules. Journal of Theoretical Biology, 149, 547–571.

[18] D. Karaboga, An idea based on honey bee swarm for numerical optimization, Technical Report-TR06, Erciyes University, Engineering Faculty, Computer Engineering Department, Turkey, 2005.

[19] D. Karaboga, B. Basturk, A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm, Journal of Global Optimization 39 (2007) 459–471.

[20] F. Kang, J. Li, Q. Xu, Structural inverse analysis by hybrid simplex artificial bee colony algorithms, Computers and Structures 87 (2009) (2009) 861– 870.

[21] X.S. Yang, Engineering optimizations via nature-inspired virtual bee algorithms, in: Lecture Notes in Computer Science, Springer (GmbH), 2005, pp. 317–323.

© 2017, IJCSMC All Rights Reserved