International Journal of Computer Applications Technology and Research
Volume 5– Issue 2, 61 - 65, 2016, ISSN:- 2319–8656

Identifying Valid Email Spam Emails Using Decision Tree

Hamoon Takhmiri
Computer Science and Technology
Islamic Azad University
Kish International Branch
Kish Island, Iran

Ali Haroonabadi
Islamic Azad University
Kish International Branch
Kish Island, Iran

Abstract: The increasing use of e-mail and the growing trend of Internet users sending unsolicited bulk e-mail, the need for an anti-spam filtering or have created, Filter large poster have been produced in this area, each with its own method and some parameters are to recognize spam. The advantage of this method is the simultaneous use of two algorithms decision tree ID3 - Mamdani and Naive Bayesian is fuzzy. The first two algorithms are then used to detect spam Bagging approach is to identify spam. In the evaluation of this dataset contains a thousand letters have been analyzed by the software Weka charts provided in spam detection accuracy than previous methods of improvement.

Keywords: Spam; Fuzzy Decision Tree; ID3 Algorithm; Naive Bayesian; Anti-Spam

1. INTRODUCTION

Today, the problem of unintended emails called spam is turned to a serious problem that 80% of these unintended emails refer to spams. Spams make a lot of problems, in other words spams cause the creation of traffic and destroy storage space and authority. Spams cause that users spend a lot of time to divide and clean unintended emails and also cause users' feeling of lack of security. Spams cause some illegal problems such as pornography, pyramidal schemes and economic scams such as phishing sites. In recent years, the increasing popularity and low cost of emails have attracted the attention of direct marketing so that with a promise of winning in lottery and getting valuable prizes, they deceive users. Large lists of email addresses, usually are taken from web pages and archives of news groups, make it possible to send unintended emails to a thousand of receivers without any costs. Users receive large amount of spams that contain anything from holidays to projects of getting wealthy. The term unintended commercial email is used in books too[1]. Spam is used in a wider sense. Spams are annoying for most users because it wastes their time and unsettle their inbox. They also waste users' money by dialing connections, reduce bandwidth and maybe show unimportant subjects with inappropriate contents such as propaganda of vulgar sites. Ferris research institute estimated that economic losses resulting from unintended emails and spams have been over 50 million dollar [2].

2. RELATED WORK

Filters have usually relied on keyword patterns, to be more efficient and prevent the danger of accidental removal of ham messages which are called Ham or allowed messages. These patterns need to be checked with each user's received emails. However, detailed setting of such patterns needs time and proficiency which are unfortunately not always available [3].

Even characteristics of messages will change by the pass of time and need updating of keyword patterns. So, automatic processing of spam messages and allowed messages that have already been received is desirable. Note that text categorization methods can be effective in anti-spam filtering. Unlike most programs of text categorization, indiscriminate mass operation is an unintended message that makes it as spam. The phenomenon can be images, sounds or any other data. The point is that to be able to distinguish between different samples and react based on the type of each sample. Learning usually happens based on one of the following methods: statistically, combination, or neural.

Realizing statistical pattern by assuming that patterns are made based on a random system, is determined based on statistical characteristics of the patterns. Some of the most important reasons of sending spams are economic goals and also advertising for a product, a service or a special idea, deceiving users to use their private information, transmission of a malicious software to the users' computers, creating a temporary failure in email server, making traffic and broadcasting immoral contents [4].

Spams are always changing their contents and forms, so that the anti-spams can't realize them. Some methods to prevent propagation of spams are including:

- economic methods: pay to send emails: like email protocols

legislative methods: such as can-spam law, secure email transfer bed.

- change email transfer protocols and offer alternative protocols such as sending ID.
- control output and input emails
- filtering based on learning (statistics) by using mail features
- detecting a phishing mail (fraud page) by the help of fuzzy classification methods

3. SUGGESTED METHOD

To detect spams better, the first goal is finding behavioral characteristics of the spam, so we need the extraction of data and registration of events of spam's behavior like sender's IP, sending time, amplitude and etc. which are shown in table 1 These data are stored in database, so they are structural data [5].

We can extract the behavioral characteristics of spams from their mail servers. Before the extraction of data, we need the

analysis of characteristics of emails from their reports. Obtaining data technology is chosen to analyze these characteristics, then the main characteristic is obtained and characteristics with less data and weaker connection are deleted. Behavioral features and characteristics of a single email is as follows:

- Customer IP (CIP)
- Receive time (RT)
- Context Length (CL)
- Frequency (FRQ)
- Context Type (CT)
- Protocol Validation (PV)
- Receiver Number (RN)
- Attach number (AN)
- Server IP (SIP)

Table 1 Mail Log Format

Time	IP	Sender	Receiver	Size	Subject	Status
...						
15:18:30	IP1	lzleon79@21cn.com	jsjxy	4987	中京国际物流	spam
15:19:33	IP2	gtfhg65@163.com	gjbjb	890	信息	spam
15:19:35	IP3	cugeuvoxler@euvox.com	0geq00 mthsmejc		Mailbox not exist	spam
15:19:36	IP4	Q0paolino@quadrugby.nl	chk	2442	The wor-ld's large-st selec-tion of online meds available	spam
15:19:39	IP5	30zhangke8@online.ln.cn	chengrw	1303	To chengrw	normal
...						

Features do not exist entirely clear in real world to explain making character for the samples logically and naturally. Data value after preprocessing is as follows:

A) Customer IP (CIP): is used only to calculate the frequency of the transmitter and to extract common pattern of transmitter's behavior, and is not used in calculations of decision tree.

B) Reaching Time (RT): the value of time of day and night is a common value and needs fuzzy making for the degree of transverse (1,0).

C) Context Length (CL): short value, long value and the size of the email are common values and need fuzzy making.

D) Protocol Validation (PV): is Boolean type and when matches with the sender (1) and in case of mismatch (0).

E) Context Type (CT): value in text/Html, multipart. (1) and when tye is text (0).

F) Receiver Number (RN): more value and less value, is a feature of common value and needs fuzzy making.

G) Frequency (FRQ): often or seldom frequency is a feature of common value and needs fuzzy making.

H) Attachment number (AN): more and less value, is a feature of common value and needs fuzzy making. Table 2 lists some examples of after preprocessing results.

Table 2 Attributes From Mail Logs

CIP	RT	CL	FRQ	CT	RN	PV	AN	SIP
...	...	...	...	...	...	...	...	...
IP1	15	4987	2	text	3	valid	0	SIP1
IP2	15	890	4	html	1	valid	1	SIP2
IP3	15	1298	1	html	1	invalid	0	SIP1
IP4	16	2442	3	multipart	2	valid	2	SIP3
...	...	...	...	...	...	...	...	...

Assuming that (A,B) are defined fuzzy subsets in a limited space (F). If A and B are named as a fuzzy rule and recorded as (A→ B) and named as fuzzy condition sets, so B is called fuzzy conclusion sets. The presented knowledge of each fuzzy decision tree shows that the rules are classified as (if - then).

For each path from root to leaves, a rule and a specific path are made. Each value of features is a pair of a part of the piece (and) of a law which is called prior law. The IF part predicts the node of the classification leaf, and so makes the following law (then part). Laws of if-then are for easier understanding, especially when the tree is big[6].

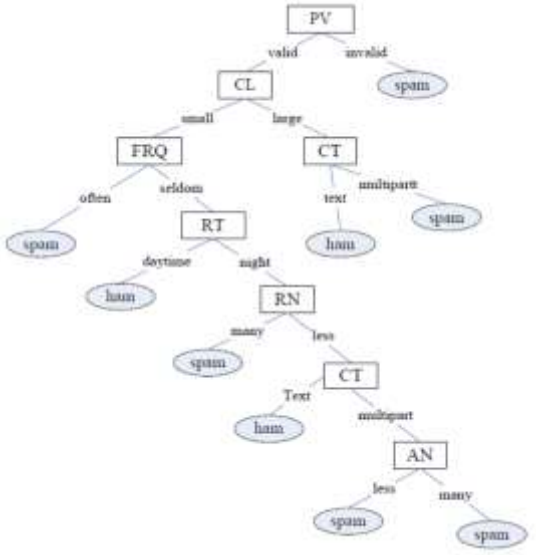

Figure 1 Decision Tree

After examining the decision tree and identifying important features of a mail by the proposed decision tree in figure 1, mamdani's generated decision tree rules are as follows:

1) If the protocol (PV) of email is not reliable, then the email is a spam.

2) If the protocol of email (PV) is valid, context length (CL) is large and context type (CT) is multipart, then the email is a spam.

3) If the protocol of email is valid (PV), context length (CL) is short and frequency (FRQ) is more, then the email is a spam.

4) If the protocol of email is valid (PV), context length (CL) is short, frequency (FRQ) is less or seldom, receive time (RT) is night, and receiver number (RN) is more, then the email is a spam.

5) If the protocol of email is valid (PV), context length (CL) is short, frequency (FRQ) is less or seldom, receive time (RT) is night, receiver number (RN) is less, context type (CT) is multipart, then the email is a spam.

6) If the protocol of email is valid (PV), context length (CL) is short, frequency (FRQ) is less or seldom, receive time (RT) is night, receiver number (RN) is less, context type (CT) is multipart, attachment number (AN) is less or more, then the email is a spam.

7) If the sender's mail server is not valid and reliable, then the email is a spam.

First, spam measures are determined which contain two implicit and tacit parts. Implicit measures are analyzed by Mamdani's fuzzy decision tree, such as protocol type, context length, context type, time, frequency, receiver number, attachment number and etc. Tacit measures are analyzed by Naïve Bayesian method such as frequency of free word repetition, money, three zeros in a row and etc. In fact, the considered data set is a combination of implicit characteristics that are in fuzzy_ Mamdani decision tree and tacit characteristics that are used in Naive Bayesian method. Implicit characteristics of the considered data set are analyzed by decision tree algorithms (ID3) and the results are completed by Fuzzy Mamdani rules [7].

Then tacit characteristics are examined in Naive Bayesian principles and finally, the obtained results from both algorithms are entered in Baking algorithm, that is each mail in a dataset enters the Naive Bayesian and decision tree and in the absence of correct diagnosis (FP and NP) a negative score is registered for the procedure. Finally, the optimal weight may be achieved through trial and error. The bonus rate should also be achieved. This means that the desired class level of the case (or a spam) is divided by the number of spam detection methods. And the result should be divided by the number of mails of the dataset to obtain bonus rate. Mails that are classified correctly are multiplied by bonus rate, and mails that are classified incorrectly are multiplied by bonus rate too. The obtained difference by multiplying the bonus rate in wrong and correct classifying is collected with initial weight (0.5%) This operation is done for Naive Bayesian and decision tree methods and because Naive Bayesian method's threshold is more favorable, it's considered as final threshold. To obtain the ultimate accuracy, each mail is entered in to two Naive Bayesian and decision tree [8].

The output of methods, if both methods have the same results, or in the case of difference, the priority of identification is given to Naive Bayesian method. And to obtain the ultimate accuracy, results are compared with the main class of the mail (spam or ham). When a new mail enters, after the recognition of both methods (Ham=0, spam=1) the output of each method is multiplied by the coefficient obtained from that method, and obtained values are gathered together, for example if just the tree realizes the spam and the other one doesn't realize it, the accuracy is in average and if the response of both methods are the same, for example both detect spam or both do not detect spam, the accuracy is desirable. In the final test by K-Fold method, the data set is divided in to four parts. The first part is for testing and the rest are for learning, in the next step the second part is for testing and the first, third, and forth parts are for learning, then the third part is for testing and the other parts are for learning and after that the forth part is for testing and other parts are for learning [9].

4. RESULT AND DISCUSSION

The dataset that the proposed method is implemented on it contains 1000 emails that 350 (35%) of them are spam and 650 (65%) of them are ham. The last column of this data set is class column and number 1 means spam and 0 means ham. Some examples of keywords for no implicit part of implementation on Naive Bayesian are as follow:

Money, Credit, 000, Internet, Edu, Talent, Free, Make ,# ,$, …

And the other part of this dataset contains implicit characteristics to use for the implementation on fuzzy decision tree, such as:

Sending time, Context type, Context length, Frequency, Receiver number, Sender's number, …

The goal of testing the mentioned dataset is to examine the accuracy of detection of the proposed method and showing a better detection of spams rather than efficiency of Naive Bayesian or decision tree methods. The method is that after analyzing dataset in Naive Bayesian method and extracting levels of efficiency, accuracy and dark bright points and areas, the same data set is analyzed by decision tree and levels of efficiency, accuracy and dark, bright points and areas are extracted, then the obtained results are voted based on Baking method, then the method that has got better comprehension is a priority and its further recognition is collected with the interface of the two methods. To implement in Naive Bayesian method, first the considered data set is implemented in Weka software, then the considered inputs are chosen among fields of dataset, The data set that the proposed method is implemented on it contains 1000 emails that 350 (35%) of them are spam and 650 (65%) of them are ham. The last

column of this dataset is class column and number 1 means spam and 0 means ham. Some examples of keywords for no implicit part of implementation on Naive Bayesian are as follow:

Money, Credit, 000, Internet, Edu, Talent, Free, Make ,# ,$, ...

And the other part of this dataset contains implicit characteristics to use for the implementation on fuzzy decision tree, such as:

Sending time, Context type, Context length, Frequency, Receiver number, Sender's number... ,

The goal of testing the mentioned dataset is to examine the accuracy of detection of the proposed method and showing a better detection of spams rather than efficiency of Naive Bayesian or decision tree methods. The method is that after analyzing dataset in Naive Bayesian method and extracting levels of efficiency, accuracy and dark, bright points and areas, the same data set is analyzed by decision tree and levels of efficiency, accuracy and dark, bright points and areas are extracted, then the obtained results are voted based on Baking method, then the method that has got better comprehension is a priority and its further recognition is collected with the interface of the two methods. To implement in Naive Bayesian method, first the considered data set is implemented in Weka software, then the considered inputs are chosen among fields of dataset [10].

To show the efficiency, the proposed method is discussed with one of these methods. A comparison is done based on accuracy and measurement criteria so that the examined dataset is divided in to ten sections and is examined in groups of 100,200,300,....,1000 mails. The obtained results are compared with the results of spam particle swarm optimization method which contains negative selection method and particle swarm optimization method [11].

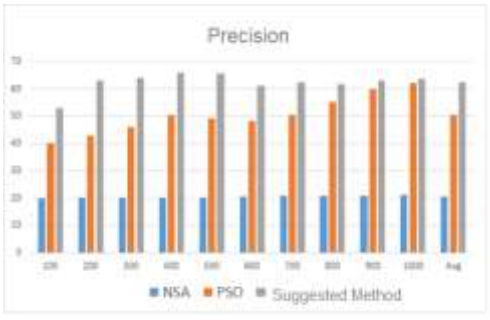

Figure 2 Precision Compare Between Methods

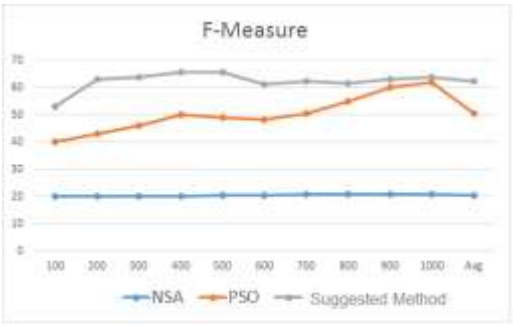

Figure 3 F-Measure

5. CONCLUSION

This method presents a new solution to detect spams by the use of fuzzy decision tree, Naive Bayesian, and Baking voting algorithm to extract spam's behavioral patterns. Because completely clear characteristics don't exist in real world, the degree of crosslinking to explain characters are neutral and rational. Fuzzy decision tree detects spam and ham mails by the use of fuzzy Mamdani rules, then Naive Bayesian method by the use of Bayesian formula does the same operation on chosen dataset, then Baking method by dividing votes in to smaller sections, gaining optimized weight and implementing it on obtained percentages will achieve the level of accuracy and health[12]. The proposed method not only shows a better efficiency in comparison with using each method separately, but also by the use of common interface of spam and ham emails detection (common TPs and TNs of both methods) divides detection in to two categories of reliable and highly reliable. One of the most important items in determining the optimal method of spam detection is minimizing the number of ham mails that are detected as spam mails because finding and deleting a spam among ham mails is easy for the users while finding a ham mail among spam ones is typically difficult and time consuming. To improve accuracy of spam detection results, two methods are used and by the use of Baking voting method and dividing votes, a better spam detection is provided. As mentioned in previous chapter, the comparison of suggested method with some methods that have been done before, shows better performance in terms of obtained accuracy results. Adding a preprocessing fuzzy level to process contents of emails for users by the use of categorizing mails based on content, subject, sender, time, receiver's number, sender's number, and etc. and combining three Naive Bayesian, decision tree, and Baking algorithm methods based on tacit and implicit components of a mail, categorizing has been done based on two methods and voting has been done by Baking algorithm, and false positive and negative rates cause an improvement in the accuracy of statistical filters to detect spams and a decrease in error detection [13].

6. RECOMMENDATIONS AND FUTURE WORK

To improve the proposed method, we can expand branches and leaves of decision tree to enter more details. In fact detailed fuzzy making of a mail includes: sending time,

sending protocol, context length, context type, time zone, number of receivers, frequency, and number of attachments, which increase accuracy performance of decision tree in detecting spams.

Operations such as adding more characteristics to fuzzy Mamdani decision tree and increasing Mamdani's laws improve the efficiency. Adding no implicit details to different parts of a letter such as subject, content, sender, effective keywords in Naive Bayesian method cause the performance improvement of Naive Bayesian method in the field of classifying letters. Finally, the use of both methods in baking algorithm show a better performance percentage. The more the K-Fold divider, the higher the detection accuracy of proposed method is. In other words, the amount of considered K-Fold in proposed algorithm correlates with the accuracy of diagnosis. More attention to details of spam detection and correct classification of mails, results in the increase of accuracy. On the other hand, detection and division of implicit and no implicit characteristics of a mail that each one is detected in its own related method, help a better classification of emails. Note that more attention to details of a mail in detection of a spam will increase accuracy and decrease simplicity and understanding of the method.

7. REFERENCES

[1]. Wu, C.T., Cheng, K.T., Zhu, Q., and Wu, Y.L., 2008, "Using Visual Features For Anti-Spam Filtering", In Proceedings of the IEEE International Conference on Image Processing, Vol. 29, Iss. 1, pp. 63-92.

[2]. Goodman, J., and Rounthwaite, R., 2004, "Stopping Outgoing Spam", In Proceedings of the 5th ACM Conference on Electronic Commerce, pp. 30-39.

[3]. Siponen, M., and Stucke, C., 2006, "Effective Antispam Strategies In Companies: An International Study", In Proceedings of the 39th IEEE Annual Hawaii International Conference on Transaction on Spam Detection, Vol. 6, pp. 245-252.

[4]. Cody, S., Cukier, W., and Nesselroth, E., 2006, "Genres Of Spam: Expectations And Deceptions", In Proceedings of the 39th Annual Hawaii International Conference on System Sciences, Vol. 3, pp. 48-51.

[5]. Golbeck, J., and Hendler, J., 2006, "Reputation Network Analysis For Email Filtering", In Proceedings of the First International Conference on Email and Anti-Spam, pp. 21-23.

[6]. Liang, Z., Jianmin, G., and Jian, H., 2012, "The Research and Design of an Anti-open Junk Mail Relay System", In Proceedings of the First IEEE International Conference on Computer Science and Service System, pp. 1258-1262.

[7]. Feamster, N., and Ramachandran, A., 2006, "Understanding The Network-Level Behavior Of Spammers", In Proceeding of the 3th ACM Conference on Email and Anti-Spam, Vol. 36, Iss. 4, pp. 291-302.

[8]. Lili, D., and Yun, W., 2011, "Research And Design Of ID3 Algorithm Rules-Based Anti-Spam Email Filtering", In Proceedings of the Second IEEE International Conference on Software Engineering and Service Science, pp. 572-575.

[9]. Zhitang, L., and Sheng, Z., 2009, "A Method for Spam Behavior Recognition Based on Fuzzy Decision Tree", In Proceedings of the Ninth IEEE International Conference on Computer and Information Technology , Vol. 2, pp. 236-241.

[10]. Duquenoy, P., Moustakas, E., and Ranganathan, E., 2005, "Combating Spam Through Legislation: A Comparative Analysis Of Us And European Approaches", In Proceedings of the Second International Conference on Email and Anti-Spam,pp. 15-22.

[11]. Jones, L., 2007, "Good Times Virus Hoax FAQ", Available: http://cityscope.net/hoax1.html, [Accesed: Jul. 10, 2015].

[12]. Singhal, A., 2007, "An Overview Of Data Warehouse, Olap And Data Mining Technology", Springer Science Business Media, LLC, Vol. 31, pp. 19-23.

[13]. Ismaila, I., and Selamat, A., 2014, "Improved Email Spam Detection Model With Negative Selection Algorithm And Particle Swarm Optimization", Elsevier Journal of Alliance and Faculty of Computing, Vol. 22, pp. 15-27.