International Journal of Computer Applications Technology and Research
Volume 5– Issue 3, 132 - 136, 2016, ISSN:- 2319–8656

Identification of Spam Emails from Valid Emails by Using Voting

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: In recent years, the increasing use of e-mails has led to the emergence and increase of problems caused by mass unwanted messages which are commonly known as spam. In this study, by using decision trees, support vector machine, Naïve Bayes theorem and voting algorithm, a new version for identifying and classifying spams is provided. In order to verify the proposed method, a set of a mails are chosen to get tested. First three algorithms try to detect spams, and then by using voting method, spams are identified. The advantage of this method is utilizing a combination of three algorithms at the same time: decision tree, support vector machine and Naïve Bayes method. During the evaluation of this method, a data set is analyzed by Weka software. Charts prepared in spam detection indicate improved accuracy compared to the previous methods.

Keywords: Spam emails; tree decision; Naïve Bayes; support vector machine; voting algorithm

1. INTRODUCTION

In recent years, the increasing use of e-mails has led to the emergence and increase of the problems caused by unwanted bulk email messages commonly known as spam. By changing the common and aggressive content of some of these messages from a minor harassment to a major concern, spams began to reduce the reliability of emails. Personal users and companies that are affected by spams because of network bandwidth spent a lot of time for receiving these messages and distinguishing spam messages from standard messages (legally certified) for users. A business model based on spams for buying and selling is usually beneficial because the costs for the sender is little, so a lot of messages can be sent by maximizing replies and this aggressive performance is a feature of known spammers. Economic functions of spams have forced some countries to legislate for them. In addition, problems of pursuing the transmitters of these messages can limit the performance of such laws. In addition to legislation, some institutions have imposed changes in protocols and practical models. Another approach being implemented is using spam classifiers which by analyzing message content and additional information try to identify spam messages. For using this function once, messages are identified usually based on the settings used by classifier [1]. If classifiers are used by a single user, as a customer-focused classifier, messages are usually sent to a folder that only contain messages under the title of spam and this makes it easier to identify the messages. On the other hand, if the classifier works on the Email server, by checking multiple users' messages, they can be tagged as spam or get deleted. Another possibility is a multi-user setting in which classifiers running on different machines share information about the received messages to improve their performance. Generally, the use of classifiers has created an evolutionary scenario in which spammers use instruments with different ways, in particular, regulated methods for reducing the number of messages identified. Initially, spam classifiers were based on the user's known laws and were designed on the basis of arrangements that are easily seen in such messages. [2].

2. RELATED WORK

Filters have been relied on key word patterns. In order to be efficient and avoid the risk of accidental deletion, non-spam messages are called as legitimate messages or Ham. These patterns should be checked manually by each user's email. However, fine adjustment of patterns requires time and expertise which is unfortunately not always available. Even messages' features change over time which requires key word patterns to be updated regularly [3]. Thus, processing messages and detecting spams or non-spams automatically is desirable. It should be noted that text categorization methods can be effective in anti-spam filtering. Sending bulk unsolicited message makes it a spam message, not its real content. In fact, posting a bulk message carelessly makes the message a spam. Phenomena can be images, sounds, or any other data, but important thing is to distinguish between different samples and have a good reaction for every sample. Learning is usually used in one of the following ways: Statistical, synthetic or neural [4].

Recognition of Statistical pattern by assuming that these patterns are created by a possible system is determined based on the statistical properties of patterns. Some of important reasons for spamming include economic purposes, as well as promoting a product, service or a particular idea, tricking users to use their confidential information, transmission of malicious software to the user's computer, creating a temporary email server crash, generating traffic and broadcasting immoral contents. Spams are constantly changing their content and form to avoid detection by anti-spams. Some ways to prevent spamming include:

• Economic methods: getting cash to send e-mail: Zmail Protocols

• Legislative procedures: such as CAN-SPAM laws, securing email transmission platform

• Changing the e-mail transmission protocol and providing alternative protocols such as sending id and features

• Controlling your outgoing emails versus controlling incoming mails.

• filtering based on learning (statistical) and by using the mail features

• Mail detection phishing (fake pages) to help fuzzy classification methods

• Controlling methods: Controlling the features of a mail before sending it by e-mail server

3. SUGGESTED METHOD

For better identification of spams, the aim is initially finding behavioral features of spam, so the data mining and logging of spam behavior is required at first, such as detecting the IP of sender, sending time, frequency, number of attachments and so on. This information is stored in the database so that they are structure information. Behavioral attributes of spams can be extracted from these reports created in the mail server [5].

Before extracting data, analyzing the features of e-mails from reports is required. Information obtaining technology is selected to analyze these features and the main feature is obtained. Some features of fewer information and weaker relations are deleted. The behavioral feature of an individual e-mail is as follows:

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

There are no entirely clear features in real world, so that after fuzzification, normal and logical degrees below horizon can be explained for characterization of samples. After preprocessing, the value of information is as follows:

A) Customer IP: is used only to calculate the frequency of sender and extracting the common behavioral pattern of sender and is not involved in the decision tree computing.

B) Receive Time: the value of day and night time is a common value and requires fuzzification for horizontal degree (0, 1).

C) Context Length: the value of short and long for the size of email is also a common value feature and requires fuzzification too.

D) Protocol Validation: is Boolean and when it complies with the sender is (1) and in case of non-compliance is (0).

E) Context Type (CT): The value a text or Multipart is (1) for the text and (0) for the Multipart.

F) Receiver Number (RN): the value of many and less is a common value feature and also requires fuzzification.

G) Frequency (FRQ): the value of often and seldom frequency is a common value feature and also requires fuzzification.

H) Attachment Number (AN): the value of many and less is a common value feature and also requires fuzzification. Table 1 lists some examples of the results after preprocessing.

Table 1 Result Of Data Processing

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

It is assumed that (A, B) are fuzzy subsets defined in a confined space (F). “If A so B” is named as a fuzzy rule and simply recorded as (A → B), which is called fuzzy sets of conditions, and (B) is called fuzzy sets of conclusion. Based on the knowledge of decision tree, rules were classified as fuzzy and are in the form of "if - then". A rule is created for each path from the root to the leaves. Simultaneously with a special path, any value features are as a pair of (And) piece of a rule that is called prior rule. (Section If) predicts leaf node classification, so forms the compliance rule. (Section Then) of "if - then" rule are easier to understand, especially when the tree is large [6].

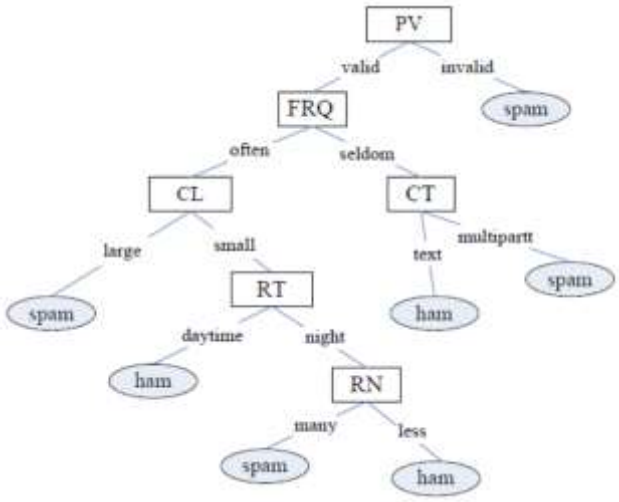

Figure 1 Decision tree

After checking decision trees and identifying important characteristics of a mail from the proposed decision tree in Figure 1 decision tree, Mamdani decision tree rules are generated as follows:

1. If the protocol (PV) of sending email is not authentic, the email is spam.

2. If e-mail protocol (PV) is valid, frequency (FRQ) is less and type (CT) is Multipart, the e-mail is spam.

3. If e-mail protocol (PV) is valid, frequency (FRQ) is many and context length (CL) is long, the e-mail is spam.

4. If e-mail protocol (PV) is valid, frequency (FRQ) is many, context length (CL) is short and receiver number (RN) is many, the e-mail is spam.

To explain the Naïve Bayes method, this example is discussed as follows:

Figure 2 Separating non-spam mails from spams

P(Ci) : P(Class=Spam) = S / N

P(Ci) : P(Class=Ham) = H / N

The total occurrence of the word Free in spam

$$X1 = \frac{A}{Spam}$$

$$\sum_{i=0}^{n} Free(Ham) = B$$

$$X2 = \frac{B}{Ham}$$

After obtaining the threshold (X1, X2), if the number of occurrences of the word Free in 21st spam email is closed to the X1, that mail is spam and if it is closed to X2, it is HAM. The same operations on other components can be checked [7].

Using SVM (Linear support vector machine) in classification issues is a new approach that in recent years has become an attractive subject for many and is used in a wide range of applications, including OCR, handwriting recognition, signs recognition and so on. SVM approach is that in the training phase, they try to maintain the decision boundary in such a way that its minimum distance to each of the considered categories becomes maximum distance. This choice makes the decision practically to tolerate the noisy environment and have a good response. This boundary selection method is based on the points called support vectors. Thus in the training phase, general characteristics of spam are extracted according to the data analysis obtained from data collection and training is done based on it. Then testing phase was performed based on the mentioned cases and each time compared with original data until the results became optimized [8]. The proposed method is as follows: first measurement criteria for spam are determined which contains implicit (non-material) and explicit (content). Implicit cases that are analyzed by decision tree include protocol type, content length, content type, time, frequency, receiver number, etc. explicit cases are determined by Naïve Bayes and support vector, such as the repetition of words Victory, Win, three zeros in a row, and so on. In fact, the desired dataset is a combination of implicit properties inside the decision trees and explicit characteristics used in Naïve Bayes method and vector machine and the results of all three methods are surveyed by voting. In other words, the implicit characteristics of the data set were analyzed by decision tree algorithm and the obtained results are completed through fuzzy - Mamdani rules [9]. Then explicit characteristics in Naïve Bayes rule and support vector machine are evaluated and the results of all three methods are surveyed by voting. Each of the mails inside the data collection are entered into Naïve Bayes, support vector machine and decision tree. If at least two of the three proposed algorithms were determined correctly, or in other words, at least two of the three proposed algorithms are like minded, the result is acceptable. In this method, results are divided in two groups: high reliability for all three algorithms being likeminded and average reliability for two of the three proposed algorithms being likeminded. Ultimately in the final test, data set was divided into four parts by K-fold method and the first quarter was analyzed for testing and the rest for learning; next, the second quarter of data set was analyzed for testing and first, third and fourth quarters were analyzed for learning; then the third quarter was analyzed for testing and first, second and fourth quarters were analyzed for learning; also the fourth quarter was considered for testing and first, second and third quarters were considered for learning [10].

4. RESULT AND DISCUSSION

Dataset for implementing the proposed method contain 1000 emails in which 300 (30%) are spam and 700 (70 %) are non-spam. In the last column of this data set, there is a (Class) column and inserting number 1 in this column means spam and inserting zero means non-spam for every mail. Examples of keywords for the explicit section regarding the implementation of the Naïve Bayes method and support vector machine include:

Victory, Money, Win, Lottery, 000, ###,…

Another section of this data set containing the implicit characteristics may be used to implement the decision trees, including:

Sending time, type of text, text length, frequency, receiver number, sender number and...

The purpose of testing the proposed data set is to assess the accuracy of detection of the proposed method and showing better spam detections compared with Naïve Bayes method, support vector machine, or decision tree [11]. After analyzing the data set inside the Naïve Bayes method and extracting

efficiency and accuracy percentages and bright-dark spots, that same data set inside the decision tree and support vector machine is analyzed and its accuracy and efficiency percentages is calculated and the results will be analyzed by voting method. Then results with the like-mindedness of at least two of the three proposed algorithms algorithm are determined as the final results. To demonstrate the efficiency, f the proposed method is performed by one of the methods discussed [12]. F-Measure and accuracy criteria were compared so that the testing data set is divided into ten parts and hundred, two hundreds, three hundreds to thousand mails were tested. The results will be compared with the results of spam swarm optimization method that include negative selection algorithm and particle swarm optimization on measurement and accuracy metrics.

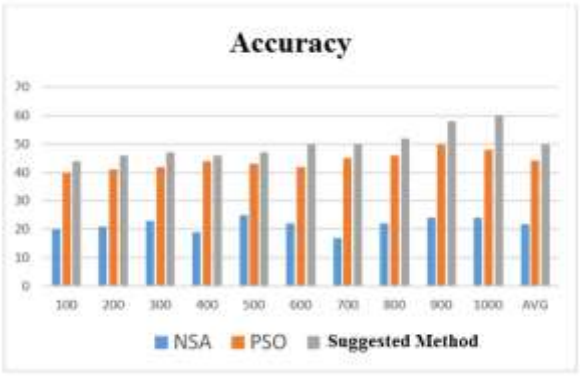

Figure 3 Accuracy Compare Between Methods

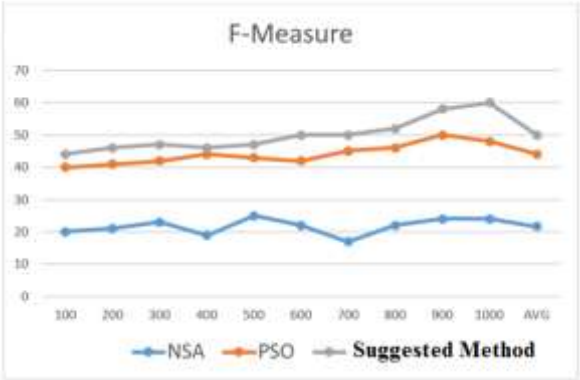

Figure 4 F-Measure

5. CONCLUSION

This method provides a new approach to detect spam using a combination of decision tree, Naïve Bayes method, support vector machine method and surveying by voting to extract the behavioral patterns of spam. Since there are no entirely clear features in the real world, degrees below horizon is normal and reasonable to explain the behavioral characteristics [13]. Decision tree begins identifying the mail and spams in dataset using fuzzy - Mamdani rules. Then Naive Bayes method is performed on the selected data set through Bayes formula of this operation. Then support vector machine algorithm analyzes the explicit data and voting method by the obtained percentages from all three methods and fining like-mindedness of at least two of the three proposed algorithms, shows the accuracy of detection. Proposed method not only shows a better performance compared with the independent use of each of the three methods, but divide's the detection into two groups of almost reliable and very reliable by using the majority of votes in detecting spam and normal mails (common TP and TN of all three methods).

One of the most important things in determining the optimal method to detect spam is to minimize the number of non-spams mails that are known as spam, since finding and deleting spams between safe e-mails known is simple for the user while finding a safe mail between spams is usually difficult and time consuming. To improve the accuracy of spam detection results, three methods were used and better statistics were provided for spam detection through voting method. Comparing the proposed approach with some of previous methods that have already been done show a better performance regarding the accuracy of results. Adding a fuzzy preprocessing level for processing email contents for user was done by using mail classifications into categories based on content, subject, sender, sending time, receiver number, sender number, etc. and the integration of Naïve Bayes method, decision tree and support vector machine based on implicit and explicit components of a mail for all the three methods and classification was done by voting. Using false positive and negative rates increased the accuracy of statistical filters for spam detection accuracy and lowered detection error rate..

6. RECOMMENDATIONS AND FUTURE WORK

To improve the performance of proposed method in the future, more details can be evaluated by the development of decision tree. For example, more detailed non-content cases including: sending time, sending protocol, content length, content type, time zone, receiver number, frequency and number of attachments can increase the accuracy of decision tree in spam detection. For Naïve Bayes and support vector machine methods, adding content details, some parts of a mail such as: subject, content, sender, keywords and user interests, results in improved performance of these two methods regarding the classification of spam and non-spam mails. Finally, using all the three methods in voting algorithm will show a better efficiency percentage. More K-fold divider and learning percentage increases proposed method's accuracy of detection. In other words, K-fold amount and learning percentage considered for the proposed method have a direct relationship with the accuracy of detection. Of course, if details and K-fold amount exceeds a certain extent, implementing the method will be more complex.

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.