Vol. 6, No. 12, December 2015 ISSN 2079-8407
Journal of Emerging Trends in Computing and Information Sciences

http://www.cisjournal.org

Filtering Spam Using Fuzzy Expert System

[1] Siham A. M. Almasan, [2] Wadeea A. A. Qaid, [3] Ahmed Khalid, [4] Ibrahim A. A. Alqubati
[1, 2] Hodeidah University, Faculty of computer science and engineering, Yemen
[3, 4] Najran University, Community College, Computer Department, KSA

ABSTRACT

The rapid growth of users in the Internet and the abuse of e-mail by unsolicited users cause an exponential increase of spam in user mailboxes.. The techniques currently used by most anti-spam software are static; spammers simply examine the latest anti-spam techniques and find ways how to rip them off. This paper presents a fuzzy expert system to detect spam. Considering the pre-processing of the subject, content, the sender's email address and attachments of the email to be ranked by using common spam words list. These ranked items represent the input variables for the proposed system, which classify the email as spam or not. The fuzzy expert system performs to filter the spam and gives good results in terms of spam recall and precision.

Keywords: *Spam filtering, spam word ranking, spam classification, fuzzy logic.*

1. INTRODUCTION

Spam mail is generally regarded as the act of repeatedly sending unsolicited e-mails in large numbers to individuals with whom the sender has never had previous contact [15]. There are many varieties of spam. The most common can be called advertising, blank spam, image spam, and backscatter. Many small companies that have legitimate business use spam e-mail in order to advertise their products. The spam received without your consent. These messages can be quite harmless, or vice versa - to bring a potential threat [18, 19].

The problem of spam is becoming a pressing issue [13,10]. From lost productivity and the time responding to blocking spam from senders, to wasting valuable server storage space while the cost of sending spam is extremely low, spam can have a major impact on Internet users.

Recently automated anti-spam detectors have become a familiar method [2,7,4,3,14]. This is due to the imperfection of the legislation, and lack of a 100% effective method of fighting spam. Many different approaches for fighting spam have been proposed, ranging from various sender authentication protocols to charging senders indiscriminately, in money or computational resources, or traditional protection systems. Spammers have found a loophole and a workaround to continue to do their business [20,17,3].

Gyongyi and Garcia-Molina [9] stated that spammers have been successful in devising new sophisticated techniques to spread spam. Their goals are revenue generation, higher search engine ranking, promoting products and services, stealing information, and phishing.

Fuzzy logic is particularly attractive for spam detection, since they are capable of adapting to the evolving characteristics of spam, because there is no clear separation between spam and non-spam messages and fuzzy logic is a good way to deal with those fuzzy boundaries.

This paper presents a fuzzy expert system to detect spam using a data set of spam words to rank email subject, body, attachments and sender address.

The motivation of using fuzzy expert system for spam detection came from the fact that there is no clear separation between spam and non-spam messages and fuzzy logic is a good way to deal with those fuzzy boundaries. Fuzzy classification assumes the boundary between two neighboring classes as a continuous and overlapping area within which an object has partial membership in each class.

The rest of the paper is organized as follows: section two shows the details of the proposed fuzzy expert system to detect spam. Implementation and results discussion are expressed in section three. Finally section four present a paper conclusion.

2. FUZZY EXPERT SYSTEM FOR FILTERING SPAM

The concept of fuzzy logic was introduced in 1965 by LotfiZadeh. Fuzzy logic deals with fuzzy sets that allow a degree of membership. The membership in these vaguely defined sets is represented by the degree of relevance [12]. This provides flexibility in dealing with uncertainty in systems such as spam filtering. Fuzzy logic together with the appropriate rules of inference provides a good method for filtering spam.

The basic configuration of the fuzzy logic system consists of four main components: fuzzy rule base, fuzzy inference engine, fuzzifier and defuzzifier. The main advantage of fuzzy logic is designed on the basis of the human knowledge of the system behavior [16, 11].

The proposed system relies on fuzzy rule based filtering approach which includes fuzzy inference system with fuzzy rules for classifying spam mail. First the email will be Pre-processing and ranking; where the information in the email is divided into the header (contains general information on the message, such as the subject, sender and recipient), body (the actual contents of the message)

and attachments. Then ranked the value of the body subject attached and sender address are used as the input variables in the fuzzy expert system to detect spam mail figure (1) shows the flowchart of the system.

The pre-processing is done by extracting the data from a message as follows:

1. Extract the sender address from the header;
2. Extract the attachments from the message;
3. Extract the words from the message header and body by the following:
   3.1 Tokenization;
   3.2 Change the words to their root forms;
   3.3 Remove stop words such as (to, in, etc.).

The ranking is done by:

1. Count the number of spam words by comparing them with a common list of spam words which is used in spam detection literature [8] as shown in equation (1).

$$W_{s \in \{body,\ subject\ ,attached,\ sender\ address\}} = \frac{Ms}{Ns} \quad (1)$$

Where:

Ms is the number of spam words in {body, subject, attached, sender address} Ns is total words in {body, subject, attached, sender address}.

W_s is the ranked value for the input variables body, subject attached and sender address.

The rank values of subject, body, attachments and sender address are used as input variables to construct the fuzzy rules that are used in the fuzzy expert system to classify the email to not spam, a little dangerous, medium dangerous, strongly dangerous.

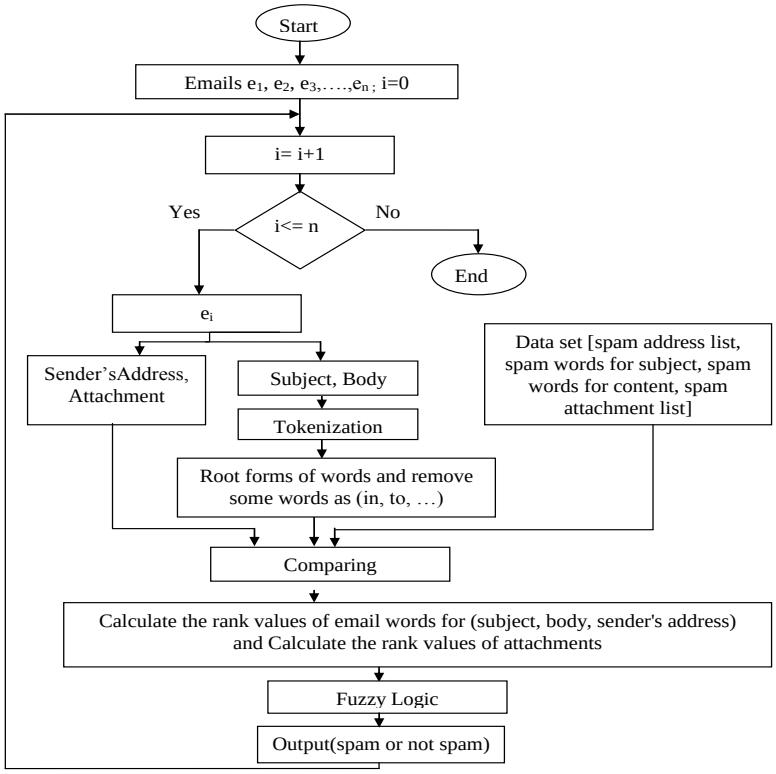

Figure 1: Proposed System

3. IMPLEMENTATION AND RESULTS DISCUSSION

Fuzzy inference system in LABVIEW is used to implement the proposed fuzzy expert system. Fuzzy inference system maps input space to an output space by using fuzzy logic. The ranked values for subject, body, attachments and sender' address words represent the inputs as shown in table 1. Figs.2.A. to Fig.2.D. show the membership of the input variables while fig.3 shows the membership of the output variable. Table 2. Shows the examples of the rules which are generated by Mamdani's fuzzy inference method as it is simple for min-max operations [5, 6].

Vol. 6, No. 12, December 2015 ISSN 2079-8407
Journal of Emerging Trends in Computing and Information Sciences

http://www.cisjournal.org

Table 1: Shows the input and outputs variables and their linguistic values

Parameter	Linguistic values	Range	Shape
Sender' address	Very weak spam	(0 ; 0,05 ; 0,1)	Triangle
	Weak spam	(0,07 ; 0,2 ; 0,3)	Triangle
	medium spam	(0,25 ; 0,45 ; 0,55)	Triangle
	strong spam	(0,5 ; 0,6 ; 0,7)	Triangle
	very strong spam	(0,65 ; 0,8 ; 1 ; 1)	Trapezoid
Words Subject	Very weak spam	(0 ; 0,05 ; 0,1)	Triangle
	Weak spam	(0,07 ; 0,2 ; 0,35)	Triangle
	medium spam	(0,3 ; 0,45 ; 0,55)	Triangle
	strong spam	(0,5 ; 0,6 ; 0,7)	Triangle
	very strong spam	(0,65 ; 0,825 ; 1 ; 1)	Trapezoid
Words body	Very weak spam	(0 ; 0,05 ; 0,1)	Triangle
	Weak spam	(0,07 ; 0,2 ; 0,3)	Triangle
	medium spam	(0,25 ; 0,4 ; 0,55)	Triangle
	strong spam	(0,5 ; 0,6 ; 0,7)	Triangle
	very strong spam	(0,65 ; 0,85 ; 1 ; 1)	Trapezoid
Attchments	Very weak spam	(0 ; 0,05 ; 0,1)	Triangle
	Weak spam	(0,07 ; 0,2 ; 0,35)	Triangle
	medium spam	(0,3 ; 0,4 ; 0,55)	Triangle
	strong spam	(0,5 ; 0,6 ; 0,7)	Triangle
	very strong spam	(0,65 ; 0,85 ; 1 ; 1)	Trapezoid
Email	not spam	(0 ; 0,05 ; 0,08)	Triangle
	a little dangerous	(0,07 ; 0,15 ; 0,3)	Triangle
	medium dangerous	(0,2 ; 0,5 ; 0,7)	Triangle
	strong dangerous	(0,5 ; 0,85 ; 1 ; 1)	Trapezoid

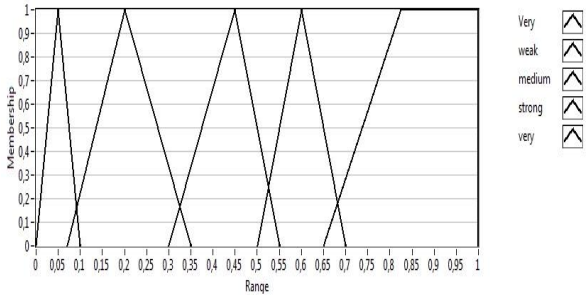

Figure 2A: Membership for the sender address

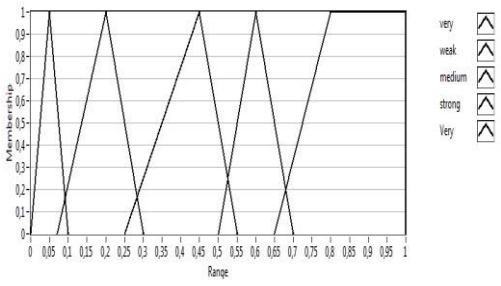

Figure 2B: Membership for the subject

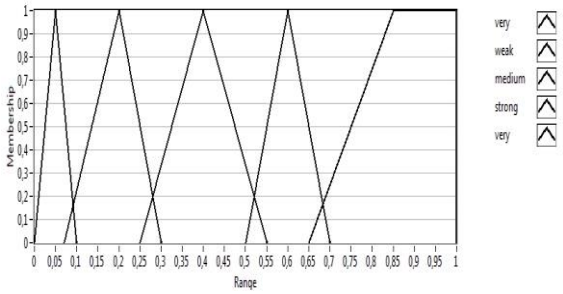

Figure 2C: Membership for the body

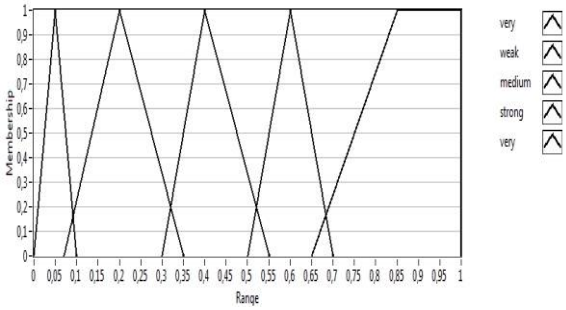

Figure 2D: Membership for the attachments

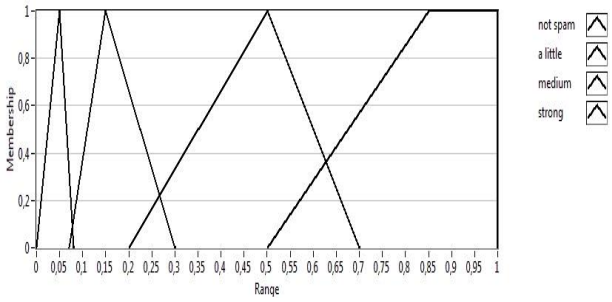

Figure 3: Membership functions for the output

Table 2: Example of fuzzy rules

RRules	Sender'address	Words subject	Words body	Attchment	Email
iif	……	……	'strong spam'	'medium spam'	'very dangerous'
Iif	……	……	'very strong spam'	'medium spam'	'very dangerous'
Iif	'medium spam'	'medium spam'	'medium spam'	'medium spam'	'medium dangerous'
Iif	'strong spam'	'very strong spam'	'medium spam'	'medium spam'	'very dangerous'
Iif	'Very strong spam'	'strong spam'	'medium spam'	'medium spam'	'very dangerous'
Iif	'Very strong spam'	'very strong spam'	'medium spam'	'medium spam'	'very dangerous'

Fig.4A. and fig. 4B. shows the interface of the proposed fuzzy expert system first, the system loads the email message to extract the number of the spam words in email subject, attachment, body and sender address. Then the system calculates the rank value for these variables, which are used to generate the fuzzy rules that are used to classify the email message as legitimate or spam as shown in fig.5A and fig.5B.

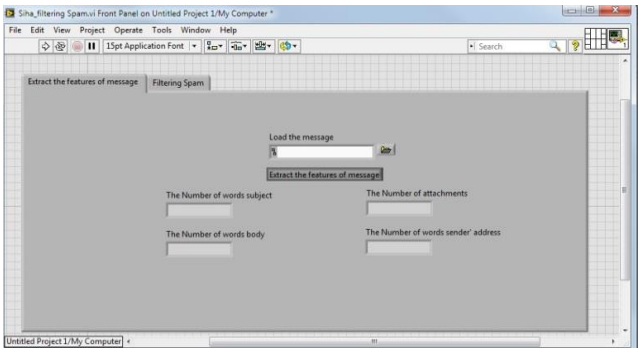

Figure 4A: Interface for reading email message and calculate the spam words

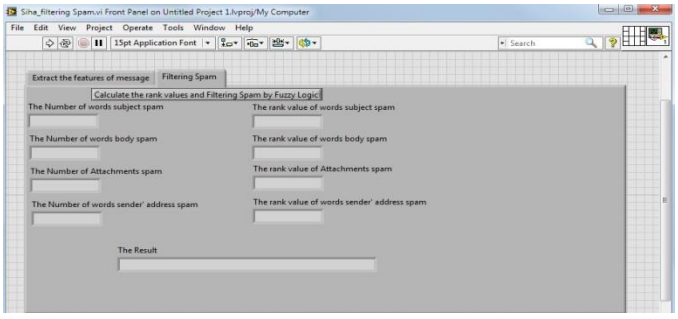

Figure 4B: Interface for ranking input variables and classify the message

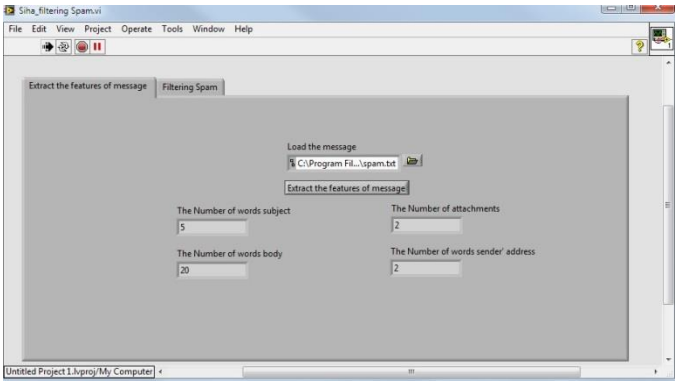

Figure 5A: An example of reading message and calculating spam words

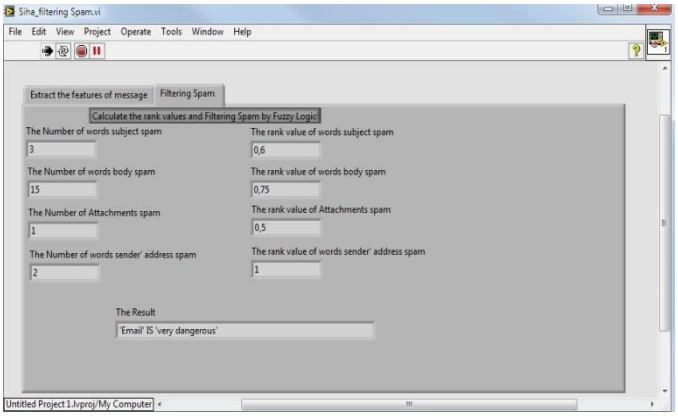

Figure 5B: Example of classifying message

Fig. 5A shows the results of processing an email message where the number of spam words in the subject are 5, attachment are 2, email body are 20 and sender address are 2. These numbers are ranked as shown in

fig.5B to get the rank value of the subject is 0.6, the body is 0.75, the attachments are 0.5 and the sender's address is

The system used these values to classify the email message as very dangerous.

To evaluate the proposed system performance, spam precision, which denotes the percentage of messages in the used test data classified as spam which truly are spam, and spam recall, which denotes the proportion of actual spam messages in the used test data that are categorized as spam by the classifier. Spam precision (Sp) and spam recall (Sr) can be defined as follows:

$$S_p = \frac{n_{s\to s}}{n_{s\to s}+n_{l\to s}} \quad (2)$$

$$S_r = \frac{n_{s\to s}}{n_{s\to s}+n_{s\to l}} \quad (3)$$

Where

$n_{s\to s}$ is the spam email that classified as spam.
$n_{l\to s}$ is the legitimate email that classified as spam
$n_{s\to l}$ is the spam email that classify as legitimate

A data set of 188 email messages are used to evaluate the fuzzy expert system, 112 of these messages are spam and 76 are legitimate messages. Table3. shows that the fuzzy expert system achieved impressive spam recall and precision results.

Table 3: Spam recall and precision

	Spam
Recall	98%
Precision	100%

Classifying a legitimate message as spam is generally more severe an error than classifying a spam message as legitimate [1]. The precision results correspond to this statement.

4. CONCLUSION

This paper presented a fuzzy expert system for classification spam emails. The proposed system ranked the spam words that exist in the subject, body, attachment and sender address. The ranked values are passed to the fuzzy inference system. FIS classifies the spam and produce, the output. The results show that the approach of using the fuzzy logic with the expert system is the best way to get rid of the spam messages. The fuzzy logic performs to filter spam and gives good result in terms of recall and precision. The results reveal that the proposed Fuzzy logic algorithm is more efficient than the other algorithms and simpler in application.

REFERENCES

[1] Ahmed Khalid , Izzeldin M. Osman " A Multi-Phase Feature Selection Approach for the Detection of SPAM", World of Computer Science and Information Technology Journal (WCSIT) ISSN: 2221-0741 Vol. 1, No. 3, 96-99, 2011

[2] Androutsopoulos, I., Paliouras, G., Michelakis, E. "Learning to filter unsolicited commercial e-mail.", Tech. report, 2004/2, NCSR.

[3] Androutsopoulos, I. Koutsias, J. Chandrinos, K.V. Spyropoulos, C. D. "An Experimental Comparison of Naïve Bayesian and Key-Based Ant-Spam Filtering with Personal Email Messages" in Proceedings of 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval July 24,2000, pp 160-167.

[4] Aradhye, H., Myers, G., Herson, J. Image analysis for efficient categorization of image-based spam e-mail. In Procintconf doc analysis and recog (Vol. 2), 2005.

[5] Bolata, F., Nowr, A. "From fuzzy linguistic specifications to fuzzy controllers using evolution strategies", IEEE International Conf. on Fuzzy Systems, pp.1089-1094. 1995.

[6] Bova, S., Codara, P., Maccari, D., Marra, V. A., "logical analysis of Mamdani-type fuzzy inference theoretical bases", IEEE International Conference on Fuzzy Systems, Barcelona, pp. 1-8, 2010.

[7] Bratko, A., Filipic, B., Cormack, G. V., Lynam, T. R., Zupan, "Spam filtering using statistical data compression models", Journal of Machine Learning Research, 7, 2006. p.p. 2673–2698.

[8] G.Santhi, S. Maria Wenisch, P. Sengutuvan3, "A Content Based Classification of Spam Mails with Fuzzy Word Ranking", IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 3, No 2, May 2013

[9] Gyongyi, Z., Garcia-Molina, H,"Web spam taxonomy", First international workshop on adversarial information retrieval on the web AIRWeb, 2005.

[10] J. Graham-Cumming. "The spammer's compendium", In Proc. of the Spam Conference, Cambridge, January 2003.

[11] Kaoru Hirota, Hajime Yoshino, Ming QiangXu, Yan Zhu, Xiao Yi Li, Daigo Horie. "A Fuzzy Case Based Reasoning System for legal inference", Fuzzy Systems Proceedings, IEEE World Congress

on Computational Intelligence, 1998, vol.2: 1350 – 1354.

[12] Kobersi, I.S., Finaev, V.I., Almasani, S.A., Kaid, W.A.A. "Control of the Heating System with Fuzzy Logic", World Applied Sciences Journal 23 (11): 1441-1447, 2013ISSN 1818-4952, 2013.

[13] L.F. Cranor, B.A. LaMacchia. "Spam! Communications of the ACM", Vol. 41, No. 8 (Aug. 1998), Pages 74-83.

[14] Sahami, M., Dumais, S., Heckerman, D. and Horvitz, E. "A Bayesian Approach to Filtering Junk E-Mail" in Learning for Text Categorization workshop. AAAI Technical Report WS-98-05, 1998, pp 55-62.

[15] Seung-HoonYoo, Chul-Oh Shin , Seung-Jun Kwak, "Inconvenience cost of spam mail: a contingent valuation study", ISSN: 1350-4851, 1466-4291 Online Journal homepage: http://www.tandfonline.com/loi/rael20

[16] Siham A. M. Almasani, Wadeea A. A. Qaid, Ahmed Khalid, Ibrahim A. A. Alqubati. "Fuzzy Expert Systems to Control the Heating, Ventilating and Air Conditioning (HVAC) Systems", International Journal of Engineering Research Technology (IJERT), Vol. 4, 2015.Issue 08:pp.808-815.

[17] Thiago S. Guzella, Walmir M. Caminhas. A review of machine learning approaches to Spam filtering. Expert Systems with Applications", Elsevier, 2009, vol.36. p.p. 10206–10222.

[18] Yue, X., Abraham, A., Chi, Z.-X., Hao, Y.-Y., & Mo, H Artificial immune system inspired behavior-based anti-spam filter", Soft Computing, 2007. 11, p.p. 729–740.

[19] Zhang, L., Zhu, J., Yao, T.,"An evaluation of statistical spam filtering techniques", ACM Transactions on Asian Language Information Processing, 2004, 3(4), p.p.243–269.

[20] Zorkadis, V., Karras, D. A., "Efficient information theoretic extraction of higher order features for improving neural network-based spam e-mail categorization", Journal of Experimental and Theoretical Artificial Intelligence, 2006, 18(4), 523–534.