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

IJCSMC, Vol. 4, Issue. 4, April 2015, pg.451 – 457

RESEARCH ARTICLE

MASK SPAM DETECTION USING DIFFICULT KEYWORD IDENTIFICATION AND RELATION COMPLETION

Sankar K V[1], Dr. S.Uma[2], Subin P S[3], Thiyagarajan Abhimannan[4]

[1]PG scholar department of computer science and engineering HIT Coimbatore

[2]Head of the department PG department of computer science and engineering HIT Coimbatore

[3]PG scholar department of computer science and engineering HIT Coimbatore

[4]PG scholar department of computer science and engineering HIT Coimbatore

PG DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

HINDUSTHAN INSTITUTE OF TECHNOLOGY, COIMBATORE, TAMILNADU, INDIA

***ABSTRACT:** With the increasing popularity of electronic mail (or e-mail), several people and companies found it an easy way to distribute a massive amount of unsolicited messages to a tremendous number of users at a very low cost. These unwanted bulk messages or junk emails are called spam messages. The majority of spam messages that has been reported recently are unsolicited commercials promoting services and products including cheap drugs and herbal supplements, health insurance, travel tickets, hotel reservations, and software products. They can also include offensive content and can be used as well for spreading rumors and other fraudulent advertisements such as make money fast. E-mail spam has become an epidemic problem that can negatively affect the usability of electronic mail as a communication means. The similarity of spam filters with text categorization problems and the success of machine learning techniques in solving these problems have intrigued several researchers to investigate their applicability in filtering spam. One subtle difference is that a false positive would be a more serious error than a false negative as a false positive would mean that an important e-mail was identified as spam and rejected. In this paper, a masked spam detection using keyword concatenation and synonym relation completion spam filtering method is introduced. This method considers the content of the message to predict its category rather than relying on a fixed pre-specified set of keywords. Thus, it could adapt to spammer tactics and dynamically build its knowledge base for filtering spam.*

***KEYWORDS:** Spam, Anti-Spam, Naïve Bayesian, Keyword, Data mining*

I. INTRODUCTION

Email spam, also known as junk email or unsolicited bulk email (*UBE*), is a subset of electronic spam involving nearly identical messages sent to numerous recipients by email. Clicking on links in spam email may send users to phishing web sites or sites that are hosting malware[14]. Spam email may also include malware as scripts or other executable file attachments. Definitions of spam usually include the aspects that email is unsolicited and sent in bulk. E-mail spam slowly but exponentially grew for several decades to several billion messages a day[19]. Spam has frustrated, confused, and annoyed e-mail users. Laws against spam have been sporadically implemented, with some being opt-out and others requiring opt in e-mail. Spammers frequently seek out and make use of vulnerable third-party systems such as open mail relays and open proxy servers. SMTP forwards mail from one server to another—mail servers that ISPs run commonly require some form of authentication to ensure that the user is a customer of that ISP[18]. Open relays, however, do not properly check who is using the mail server and pass all mail to the destination address, making it harder to track down spammers. Spam can also be hidden inside a fake "Undelivered mail notification" which looks like the failure notices sent by a mail transfer agent (a "MAILER-DAEMON") when it encounters an error. Spam is an e-mail message that is unwanted and is basically the electronic version of junk mail that is delivered by the postal service. Solutions to the proliferation of spam are either technical or regulatory[13]. Technical solutions include filtering based on sender address or header content. Support vector machines (SVM's) in classifying e-mail as spam or no spam by comparing it to three other classification algorithms: Ripper, Rocchio, and boosting decision trees[10]. These four algorithms were tested on two different data sets: one data set where the number of features were constrained to the 1000 best features and another data set where the dimensionality was over 7000[8]. SVM's performed best when using binary features. This can improve the accuracy in at least two ways: the user can generate a list of acceptable senders that is always noted as no spam no matter what the subject and body contents. Furthermore, return e-mail that is a response to a user query will always be accepted as non-spam[9]. The increasing popularity and low cost of electronic mail have intrigued direct marketers to flood the mailboxes of thousands of users with unsolicited messages. These messages are usually referred to as spam or, more formally, Unsolicited Commercial E-mail (UCE), and may advertise anything, from vacations to get-rich schemes[7]. The performance improves as the size of the training corpus increases, which is an indication that a larger training corpus might lead to even better results[16]. Email spam, also known as junk email or unsolicited bulk email (UBE), is a subset of electronic spam involving nearly identical messages sent to numerous recipients by email. Clicking on links in spam email may send users to phishing web sites or sites that are hosting malware. Spam email may also include malware as scripts or other executable file attachments. The word "spam" is used with a broader meaning that does not exclude unsolicited bulk e-mail sent for non-commercial purposes. Using readily available bulk-mailing software and large lists of email addresses, typically harvested from web pages and newsgroup archives, it is now possible to send blindly unsolicited messages to thousands of recipients at essentially no cost. Anti-spam filtering is expected to become an important member of an emerging family of junk filtering tools for the Internet, which will include – among others – tools to remove unwanted advertisements, and block hostile or offensive content[12].Anti-spam legal measures are gradually being adopted, but they have had a very limited effect so far and of more direct value are anti-spam filters, software tools that attempt to block automatically spam messages[6].

With the rapid extension of the Internet, e-mail has been considered as one of the most efficient and convenient way of communication[5]. The rapid growth of users in the Internet and the abuse of e-mail by unsolicited users cause an exponential increase of e-mails in user mailboxes. The increasing popularity and low cost of sending an e-mail make it very attractive to the direct marketers. It is now become very easy to send unsolicited messages blindly to thousands of people at no cost at all by using easily available bulk-mailing software and large lists of email addresses typically harvested, even purchased or rented from web pages and newsgroup archives. The most popular and direct way to prevent spam is the anti-spam filters, software tools that block spam messages automatically[11]. These anti-spam filters vary in functionality from blacklist (frequent spammer list) and white list (trusted user list) to content-based filters. The latter is more powerful since spammers generally use false addresses.

A trainable Fuzzy classifier is used to build an automatic anti-spam filter[15]. Trainable fuzzy system is a fuzzy logic based system that derives the (fuzzy) classification from training data using learning techniques. The motivation of using fuzzy logic 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. The use of fuzzy model allows us to integrate domain specific expert knowledge to the learning task and make the system more adaptive.

Text categorization is the task of assigning predefined categories to free-text documents. A fuzzy similarity approach is used in text classification task[2]. The approach is originated from Rocchio algorithm adapted to solve this problem. In the fuzzy similarity approach, a fuzzy term-category relation is developed, where the set of membership degree of words to a particular category represents the cluster prototype of the learned model. Once the membership values of fuzzy term-category relation are known, A way is needed to measure the similarity between a test document to be categorized and the category's cluster centers, which are represented by the membership values of terms in the same category. The use of fuzzy similarity is motivated by the fact that the category of a document cannot be determined only from a single term, rather it is determined from a set of terms that co-occur in training documents classified as the same category. This scheme is different from fuzzy information retrieval that uses composition rules to compute the set of documents relevant to a query given fuzzy binary relation between terms and documents. The effectiveness of various fuzzy conjunction and disjunction operators used in fuzzy similarity formula and several document representations were evaluated using test sets from three text document collections.

The problem with filtering is that sometimes a valid message may be blocked. Thus, it is not our intent to automatically reject e-mail that is classified as spam. It is highly desirable that if the user decides that e-mail messages be rank-ordered by degree of confidence that the rank ordering be reliable. A spam filter is a program that is used to detect unsolicited and unwanted email and prevent those messages from getting to a user's inbox[4]. Spam filters help us in our fight against spam and spammers. But there isn't a perfect filter. An ideal spam filter should generate a false positive and a false negative[1]. Like other types of filtering programs, a spam filter looks for certain criteria on which it bases judgments. The existing system for spam filtering is signature based spam detection method that works based on signature or structures of the spam. Naive Bayesian Classifier is used to classify the spam[3]. The Naive Bayesian Classifier can be subjected to Bayesian poisoning where the Ketwords are manipulated by spammers to overcome the security provided by Naive Bayesian algorithm.

II. CONTRIBUTIONS MADE

Find out difficult keywords from input mails

Find out the meanings for that keywords.

Relation Completion Process i.e find the synonyms and create a rule to identify spam based on the synonym rule

The process mentioned analyze the data in-depth, which avoids the false positive results.

Keywords are analyzed accurately, to cover all difficult masked spams in emails.

It gives high accuracy in classification of masked spams.

III. EXISTING WORK

Unwanted messages are always a challenge to the users of any kind of mail or message delivery system, whether it is normal post or electronic mails. Email spam is a growing threat to the users of any electronic messaging or mail system. Many different kinds of measures have been taken to rectify this issue but there always seems to have a unaddressed problem. Spam filters which use algorithms are used to filter the messages and to remove spam. However no matter how evolved a spam detection system spammers always finds a way to bypass it.

Accuracy in spam detection poses a big challenge to those who develop spam filters and related algorithms. Mails mistaken as spam get moved in to the spam folder there by increasing the risk missing important mails. The primary driving force or the objective of this project is to allow the filtering of spam in a more accurate level .This is done by narrowing the search and identification of spam messages by using the basic weapon of spam filtering called keywords in a more advanced way i.e. to concatenate the various keywords to form new sentences or string collections that would identify spam and also use the synonyms of these key words to unmask the masked spam.

IV. ALGORITHM USED

NAIVE BAYESIAN ALGORITHM

The Naive Bayesian algorithm is a probabilistic classifier algorithm. It is based on Bayes Theorem. It has independent assumptions and hence the name Naive.

This algorithm uses a formula to identify spam messages. The formula used is given below.

$Pr(S|W)= Pr(W|S).Pr(S)/Pr(W|S).Pr(S)+ Pr(W|H).Pr(H)$ Where

$Pr(S|W)$ is the probability that a message is a spam, knowing that the word "replica" is in it

$Pr(S)$ is the overall probability that any given message is spam

$Pr(W|S)$ is the probability that the word "replica" appears in spam messages

$Pr(H)$ is the overall probability that any given message is not spam (is "ham")

$Pr(W|H)$ is the probability that the word "replica" appears in ham messages

If the above formula calculates a spam probability greater than or equal to 90% then the mail would be classified as spam.

V. DISADVANTAGES OF EXISTING SYSTEM

The Naïve Bayesian algorithm even though is an efficient one it can be subjected to Bayesian Poisoning by Spammers wherein the normal keywords that spam detection system’s use by comparing with data sets would be replaced by different word with same meaning thereby evading detection i.e usage of legitimate text in the place of otherwise normally spam keyword.

The spam filtering also has to deal with another problem called false positive[20]. A false positive occurs when a legitimate message is considered as spam by the spam detection system thereby causing such mails to be moved to spam folder and causing problems of blocking important messages.

Changing the pattern of a spam keyword also evades detection by spam detection systems.

VI. PROPOSED SYSTEM

Classifying an email as spam or not cannot be done at the mail server. It needs to be done at the email client. For example let's say there are 2 users - A and B. And A works for a Bank and B works as a Pharmacist. A mail with content "Reduce your mortgage loan" is spam for B but ham for A. And a mail "Solution for baldness" is spam for A but ham for B. So when the recipient receives the email, if he received a mail and he considers it as spam, he can "Mark it as Spam". This is not a big issue. On the other hand, if he noticed a mail that was ham went into his spam folder, he can "Mark it as NOT Spam". This is an issue, as the mail might be an important one and you might miss out on it (as its not showing in your inbox). So the spam detectors should be careful not to mark a ham as spam. Also, spam can be detected based on email content, email subject, sender email, recipient emails, etc. Let's see how they work.

In the industry we have a collection of thousands of ham/spam emails which can be used to build our Spam filter application. Download these emails into your data store. Run a job on it (Map-Reduce or batch) to go through the email message and split them as multiple words[17][21]. You might have to do additional tasks like removing special characters, quotes, converting to lower case, ignoring words of length less than 4, ignore common words, ignore

words with only letters, etc. Now the valid words you add it into a HashMap as Key. The value for the Map is a Node. The Node class has 3 fields - spamCount, hamCount and probability. So if I am reading a word "XYZ" from spam email and it is the first time I encountered this word, then the Node class would have spamCount=1, hamCount=0. We will calculate probability after the map is constructed. Note that the same word can appear in the ham list. Every time a word is put in the map, increment a class level variable totalSpam (or totalHam) by 1. After all the emails are read and the map is constructed, iterate the map and get each key. For the key get the spamCount and hamCount. Calculate probability using -

probability=(spamCount/totalSpam)/((spamCount/totalSpam) + (hamCount/totalHam))

Do this for all the keys. The probability is a floating point value between 0.0 and 1.0.
That completes the training step. Next is the filtering step.

An email comes from a sender "X". So again, get the words (as described above) and for each word get the probability of the word in the map. If the word doesn't exist it the map, it means the spam filter is not trained for this word. So it could be a valid word, give it a value 0.5. Calculate the interest values I for each word as follows-

I = |0.5 - probability|

Once it is calculated for all the words, sort the I values in descending order (highest interest). Out of this take N values (N=15). For these I values, get the corresponding probabilities p1, p2, p3.. p15. Now calculate the total probability using the following formula

P = (p1*p2*p3..p15)/((p1*p2*p3..p15) + ((1-p1)*(1-p2)*(1-p3)....(1-p15)))

This value would be between 0.0 and 1.0. The nearer the value is to 0, the lesser the chances of it being spam. So we mark anything equal to or greater than 0.9 as spam.

Next comes machine learning. It could happen that, an email which is not marked spam needs is found to be spam. You mark it as spam. To do that, add the word back to map and calculate the probabilities again.

VII. CONCLUSION

Email spam is a growing threat to the internet community. The spam filters that are available now are efficient but the threats posed by spammers are evolving day by day and so the spam filters also need to evolve. The spam filters also face the challenge of false positives where in legitimate and important mails go to spam due the constraints imposed by the spam filters on the mails. So a context driven spam filtering is needed and this need is being addressed in this project "mask spam detection using difficult keyword query identification and relation completion " which deals with difficult keyword identification, concatenation and synonym identification that allows sets that allow partial membership in a set. Here in this project a spam filter that makes use of a dataset that contains keywords is developed. The keywords are identified and then concatenated to identify the probabilities of their occurrence and the synonyms of the keywords will be used to identify the context in which the keywords occur thereby narrowing the filtering of spam and thus avoiding false positives. This provides flexibility in dealing with uncertainty in systems such as spam filtering. Using this approach, a classification model is built from a set of pre-classified e-mail instances.

VIII. FUTURE WORK

Spams are not just limited to keywords in the form of text. Spam in the form of images in mails also pose threat to mail users. The spam filters performs filtering based on the keywords but a keyword in the form of image is not a text and therefore is difficult to detect. Future works on spam filtering could provide methods to identify the spam keywords hidden in images thereby making the spam filters more efficient against spam

The future enhancement could include technology that is not limited to just mails but that can be spread to every kind of message delivery system and technologies including cellphone messages.

IX. BRIEF AUTHOR BIOGRAPHY

Sankar K V received the B Tech Degree in Information Technology from Hindusthan College Of Engineering And Technology Coimbatore affiliated to Anna University in 2008 and is currently pursuing M E degree at Hindusthan Institute of Technology Coimbatore affiliated to Anna University.

Dr. S.Uma is Professor and Head of PG Department of Computer Science and Engineering at Hindusthan Institute of Technology, Coimbatore, Tamilnadu, India. She received her B.E., degree in Computer Science and Engineering in First Class with Distinction from PSG College of technology in 1991 and the M.S., degree from Anna University, Chennai, Tamilnadu, India. She received her Ph.D., in Computer Science and Engineering Anna University, Chennai, Tamilnadu, India with High Commendation. She has nearly 24 years of academic experience. She has organized many National Level events like seminars, workshops and conferences. She has published many research papers in National and International Conferences and Journals. She is a potential reviewer of International Journals and life member of ISTE professional body. Her research interests are pattern recognition and analysis of non linear time series data.

REFERENCES

[1]. Battista Biggio, Member, IEEE , Giorgio Fumera, Member, IEEE , and Fabio Roli, Fellow, *IEEE, "* Security Evaluation of Pattern Classifiers under Attack", , *IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 4, April 2014*

[2]. Dwi H. Widyantoro and John Yen ,"A Fuzzy Similarity Approach in Text Classification Task", *Department of Computer Science Texas A&M University*

[3]. Ester.M, Kriegel H, Sander J., and Xu.X, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," Proc. Second Int'l Conf. Knowledge Discovery and Data Mining (KDD), vol. 96, pp. 226-231, 1996.

[4]. Etzioni O., Banko M., Soderland S., and Weld D., "Open Information Extraction from the Web," Comm. ACM, vol. 51, no. 12, pp. 68-74, 2008.

[5]. Finkel.J, Grenager.T. , and C. Manning, "Incorporating Non- Local Information into Information Extraction Systems by Gibbs Sampling," Proc. 43rd Ann. Meeting on Assoc. for Computational Linguistics (ACL), pp. 363-370, 2005.

[6]. Fogla P, Sharif M., Perdisci R, Kolesnikov O., and Lee W., "Polymorphic Blending Attacks," Proc. 15th Conf. USENIX Security Symp., 2006.

[7]. Georgios Sakkis , Ion Androutsopoulos, Georgios Paliouras ,Vangelis Karkaletsis Constantine D. Spyropoulos, Panagiotis Stamatopoulos T, "A Memory-Based Approach to Anti-Spam Filtering for Mailing Lists" *Institute of Informatics and Telecommunications, National Centre for Scientific Research (NCSR) "Demokritos", Department of Informatics, Athens University of Economics and Business Institute of Informatics and Telecommunications, National Centre for Scientific Research (NCSR) "Demokritos", Department of Informatics, University of Athens*

[8]. Gorunescu, F*, Data Mining: Concepts, Models, and Techniques, Springer, 2011*.

[9]. Han, J., and Kamber, M., *Data mining: Concepts and techniques, Morgan-Kaufman, Series of Data Management Systems San Diego:Academic Press, 2001*.

[10]. Harris Drucker, *Senior Member, IEEE,* Donghui Wu, *Student Member, IEEE*, and Vladimir N. Vapnik , "Support Vector Machines for Spam Categorization"

[11]. Heikki, Mannila,*Data mining: machine learning, statistics and databases*, IEEE, 1996.

[12]. Ion Androutsopoulos, John Koutsias, Konstantinos V. Chandrinos and Constantine D. Spyropoulos , "An Experimental Comparison of Naive Bayesian and Keyword-Based Anti-Spam Filtering with Personal E-mail Messages", *Software and Knowledge Engineering Laboratory Institute of Informatics and Telecommunications National Centre for Scientific Research "Demokritos"*

[13]. Johnson P., Tan B., and Schuckers S., "Multimodal Fusion Vulnerability to Non-Zero Effort (Spoof) Imposters," Proc. IEEE Int'l Workshop Information Forensics and Security, pp. 1-5, 2010.

[14]. Lowd.D and Meek C., "Good Word Attacks on Statistical Spam Filters," Proc. Second Conf. Email and Anti-Spam, 2005.

[15]. Muztaba Fuad.M , Debzani Deb," A Trainable Fuzzy Spam Detection System" *Dept. of Computer Science Montana State University Bozeman, Montana, USA,* M. Shahriar Hossain *Dept. of CSE, Shahjalal University of Science & Technology Sylhet-3114, Bangladesh*

[16]. NeelamadhabPadhy, Dr.Pragnyaban Mishra and RasmitaPanigrahi, *"The Survey of Data Mining Applications and Feature Scope, International Journal of Computer Science, Engineering and Information Technology (IJCSEIT)", vol.2, no.3, June*

[17]. Shiwen Cheng, Arash Termehchy, and Vagelis Hristidis ,"Efficient Prediction of Difficult Keyword Queries over Databases", *IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 6, June 2014*

[18]. Townsend S.C., Zhou Y., and Croft B., "Predicting query performance," in *Proc. SIGIR '02*, Tampere, Finland, pp. 299–306.

[19]. Tran T., Mika P., Wang H., and Grobelnik M., "Semsearch ´S10," in *Proc. 3rd Int. WWW Conf.*, Raleigh, NC, USA, 2010.

[20]. Wittel G.L. and Wu .S.F., "On Attacking Statistical Spam Filters," Proc. First Conf. Email and Anti-Spam, 2004.

[21]. Zhixu Li, Mohamed A. Sharaf, Laurianne Sitbon, Xiaoyong Du, and Xiaofang Zhou, Senior Member, IEEE , "CoRE: A Context-Aware Relation Extraction Method for Relation Completion", *IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 4, April 2014*