ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 2, February 2017

A Proposed System for E-Mail Spam Detection with Speech Tagging

Hiral Agravat[1], Rinku Kapdi[2], Ashutosh Abhangi[3]

M.E. Student, Dept. of Computer Engineering, Noble Group of Institutions, Junagadh, Gujarat, India[1]

M.E. Student, Dept. of Computer Engineering, Noble Group of Institutions, Junagadh, Gujarat, India[2]

Assistant Professor, Dept. of Computer Engineering, Noble Group of Institutions, Junagadh, Gujarat, India[3]

ABSTRACT: Email spam, also known as junk or unwanted bulk email is the unwanted emails that have created havoc in one's life. It has grown up from an avoidable issue to an unavoidable one. According to the latest estimates, spam consists of around 70% of the total emails being sent or received. It is utmost necessary to stop these unwanted and harmful emails as they contain different kinds of viruses and threats which may adversely affect the computer and its functioning. Another aspect of the problem is that due to very high number of incoming emails, it is very difficult to find one. To support ease of access, emails are needed to be categorized based on the type of information they contain which will help a person to know the content before even opening it.

KEYWORDS: Spam Mail, Non-Spam, Head Word, Filter

I. INTRODUCTION

In recent years, internet has become an important part of our life. With increased use of internet, numbers of email users are increasing day by day. It is probable that 294 billion emails are sent every day. This increasing use of email has created problems caused by unwanted mass email messages commonly referred to as Spam. It is assumed that around 90% of emails sent everyday are spam or viruses[2].

Email has now become one of the best ways for advertisements due to which spam emails are generated. Spam emails are the emails that the receiver does not wish to receive. A large number of identical message are sent to several recipients of email. Increasing volume of such spam emails is causing serious problems for internet users, Internet Service Providers, and the whole Internet backbone network. One of the examples of this may be denial of service where spammers send a huge traffic to an email server thus delaying legitimate message to reach intended recipients. Spam emails not only waste resources such as bandwidth, storage and computation power, but may contain fraudulent schemes, bogus offers and scheme. Apart from this, the time and energy of email receivers is wasted who must search for legitimate emails among the spam and take action to dispose the spam. Dealing with spam and classifying it is a very difficult task. Moreover a single model cannot tackle the problem since new spams are constantly evolving and these spams are often actively tailored so that they are not detected adding further impediment to accurate detection [1].

Spam E-mail is so complex that it cannot be detected by many of current techniques because the spammer can use new vulnerabilities which are never seen before10. There are a number of possible solutions to spam but not effective yet10.These ranges from communication-oriented approaches like authentication protocols over blacklisting to content-based filtering approaches which usually depend on some of Artificial Intelligence (AI) techniques11. Current AI algorithms are able to detect spamming E-mail based on fixed features and rules while a few number of machine learning algorithms design to work in online mode12. It is expected that errors will multiply with time particularly when handling zero-day spams in the classification process. Many studies these days heavily focus on spam E-mails detection [2].

We proposed to create a system which not only identifies email as a spam or non-spam (also known as Ham) but also which can find a concept present in the e-mail along with the category it falls the most into.

To improve the performance of system, we have removed all the unnecessary parts from the email like html tags and stop words. Again, to further increase the performance.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 2, February 2017

To categorize email into different categories we have selected Finance, Sports, Job/Occupation, Travel and Geography as the categories as they are the most common part of an e-mail. To categorize into these categories, we have selected few important keywords from the internet. We have created an algorithm which self-trains itself using the Wordnet database and the keywords provided in the file. The output of this algorithm is a database consisting of each word's definition along with its examples. These databases are used to find the matching keywords from the test e-mail which provides output for each category.

II. LITERATURE SURVEY

In the spam mail filtering, a new approach based on this strategy that how frequently words in the evidence are found by usage of their repetition number(frequency).The key sentences, those with the key words, of the incoming e-mails have to be tagged and thereafter the grammatical roles of the entire words in the sentence need to be determined, finally they will be put together in a vector in order to indicate the similarity between the received emails. So it takes advantage of an extraordinary algorithm called k-Mean algorithm to classify the received e-mails. It is Worthwhile to note that the so-called K-Mean algorithm follows some simple and understandable rules which are too easy to work with it. This method is executed on 189 e-mails.142 of these e-mails were non-spam e-mails and 49 spam e-mails were available among them. After repeating above steps for a repeated number of times the final precision for this method is obtained 83 percent [4].In other thing they have integrated the content based spam detection using Bayesian Classifier and phishing URLs detection using Decision Tree C4.5. Thus They found that performance evaluated for combination approach of Bayesian classifier and Decision Tree C4.5 are improved as compared to implementation using content based spam detection by Bayesian Classifier[14].

In this spam there prime aim is to detect text as well as image based spam emails. To achieve the objective we applied three algorithms namely: KNN algorithm, Nai·ve Bayes algorithm and reverse DB SCAN algorithm. Pre-processing of email text before executing the algorithms is used to make them predict better. This paper uses Enron corpus's dataset of spam and ham emails. In this research paper, we provide comparison performance of all three algorithms based on four measuring factors namely: precision, sensitivity, specificity and accuracy. We are able to attain good accuracy by all the three algorithms. The results have shown comparison of all three algorithms applied on same data set. Although methods used by us have many advantages, it certainly does come with some disadvantages. The disadvantage of text filtering is that they are time consuming. The OCR based detection also has disadvantages like, the recognition is not always perfect, and works for certain fonts only, cannot predict for CAPTCHA images and obviously are expensive. In future, as it is an adaptable and scalable project thus we would like to detect threats found in emails that are viruses[5].

III. RELATED WORK

Spam filtering is a catch-all term that describes the steps that happen to an email between a sender and a receiver to distinguish between wanted and unwanted messages. Filtering companies and ISPs consider the underlying mechanisms behind spam filters confidential, so senders don't always get specific information about why an email was blocked or delivered to the bulk folder. Because filters are often a black box to email senders, they develop various theories about how or why spam filters work the way they do. Some of the theories are true, but many of them are closer to myth than reality. Senders that understand spam filters and the goals behind filtering can separate filtering myths from filtering truths, enabling them to create the most deliverable and effective emails for their target market. In thinking about filters, it's important to understand the goals of the filter. Filters aren't designed solely to block unsolicited bulk email, also known as spam. They're also useful for blocking malicious email, including virus infections and phishing emails. There are also user-specific filters, which can overrule any network filter. Mail that might otherwise go to the bulk folder will be delivered to the inbox [3].

K-means clustering algorithm:

k-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori. The main idea is to define k centers, one for each cluster. These centers should be placed in a

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 2, February 2017

cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest center. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as barycenter of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more. Finally, this algorithm aims at minimizing an objective function knows as squared error function given by[13]:

$$J(V) = \sum_{i=1}^{C} \sum_{j=1}^{Ci} (\|x_i - v_j\|)^2$$

where,

'||xi - vj||' is the Euclidean distance between xi and vj.
'ci' is the number of data points in ith cluster.
'c' is the number of cluster centers.

Algorithmic steps for k-means clustering:

Let X = {x1,x2,x3,……..,xn} be the set of data points and V = {v1,v2,……..,vc} be the set of centers.

1) Randomly select 'c' cluster centers.

2) Calculate the distance between each data point and cluster centers.

3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers.

4) Recalculate the new cluster center using:

$$V_i = (1/C_i) \sum_{j=1}^{C_i} (x_i)$$

Where, 'c_i' represents the number of data points in i^{th} cluster.

5) Recalculate the distance between each data point and new obtained cluster centers.

6) If no data point was reassigned then stop, otherwise repeat from step 3)[13].

Following Steps are:-

step 1 : All of the Existed e-mails are labelled by means of the Gate Soft Ware.

step 2 : Doing some pre-analysis functions in order to eliminate the extra character and find unnecessarywords

step 3: Making Si vector as follow:

Si = {Wi1,Wi2,….Wim}
Where Wij=weighted of the jth word in the e-mail
Calculate the value of Wij from the equation:
TF*IDF
Where TF = repetition word
IDF = e-mails (find from logarithm form)

step 4: Largest weight word selected. The sentences with these words will be chosen and the grammatical roles of these sentences will be stored vector.

Pi = {POSTi1, POSTi2, ……, POSTiq}
Where POSTiq = the grammatical role of qth word from Si e-mail.

step 5: Similarity criterion will be measured by COS method.

SIM_1 = COS (Si,Sj)
SIM_2 = COS (Pi,Pj)

Find the total similarity criterion of both the Si and Sj e-mails will be computed by the following formula:

SIM (Si,Sj) = SIM_1 * SIM_2

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 2, February 2017

step 6: E-mails will be divided into two categories by means of K-Mean algorithm[4].

IV. PROPOSED ALGORITHM

We have tried to create a system which not only identifies email as a spam or non-spam (also known as Ham) but also which can find a concept present in the e-mail along with the category it falls the most into. We have used different lexical, semantic and syntactic features in order to create the system. We have created a database of emails. Using this, we have trained our system to find the words and its count in each database. To improve the performance of system, we have removed all the unnecessary parts from the email like html tags and stop words. Again, to further increase the performance, we have used Wordnet database to just keep the most important words in the database. We have used the concept of unigram probability in order to identify email as a spam or ham. Unigram probability is calculated on a weighted basis which will be explained in further section. To categorize email into different categories we have selected Finance, Sports, Job/Occupation, Travel and Geography as the categories as they are the most common part of an e-mail. To categorize into these categories, we have selected few important keywords from the internet. We have created an algorithm which self-trains itself using the Wordnet database and the keywords provided in the file. The output of this algorithm is a database consisting of each word's definition along with its examples. These databases are used to find the matching keywords from the test e-mail which provides output for each category. This algorithm is slightly modified version of the Simplified Lesk Algorithm.To find the concept of an e-mail; we have used a common concept which says that always a concept is Noun. By using this concept, we have found a word that occurs the most in the test e-mail and is also a noun[3].

The proposed algorithm is consists of six main steps.

Step – 1: Collect the test emails.

Step – 2: Remove the following from emails:

1) Html tags
2) Symbols
3) Unimportant word and remaining are converted in dictionary file

Step – 3: Now here using the lesk algorithm to find the repeated word and list out the total number of word count from the dictionary file.

Lesk Algorithm steps:

Let your sentence be A B C

Let each word have synsets i.e. {A:(a1, a2, a3), B:(b1), C:(c1, c2)}

Now form possible synset sets: (a1, b1, c1), (a1, b1, c2), (a2, b1, c1) ... (a3, b1, c2)

Define function F(a, b, c) which returns the distance (score) between (a, b, c).

Call F on each synset set.

Pick the set with the maximum score.

Step – 4: Total number of word count is checked which word are spam and which are ham/non-spam.

Step – 5: Collect spam and ham mail are put in the unigram weighted formula to find out whether mails are spam or ham.

$$Unigram \quad \Pr obability \quad (Weighted \quad)_{spam} = \frac{\frac{C_{spam}\ TestFile}{C_{spam}\ Total}}{\frac{C_{spam}\ TestFile}{C_{spam}\ Total} + \frac{C_{ham}\ TestFile}{C_{ham}\ Total}}$$

Fig. 1. Formula for finding spam mail

$$Unigram \quad \Pr obability \quad (Weighted \quad)_{ham} = \frac{\frac{C_{ham}\ TestFile}{C_{ham}\ Total}}{\frac{C_{ham}\ TestFile}{C_{ham}\ Total} + \frac{C_{spam}\ TestFile}{C_{spam}\ Total}}$$

Fig. 2. Formula for finding ham mail

Step – 6: Total number of word count we process it in the word net dictionary which gives the definition of that word and defined the category.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 2, February 2017

V. CONCLUSION AND FUTURE WORK

Spam emails are the biggest problem for the web data. This paper explored different features to deal with this problem. There is very much capacity for identifying mail as spam emails or non-spam mails for text as well as their different categories.

REFERENCES

1. Rekha, Sandeep Negi, "A Review on Different Spam Detection Approaches" , International Journal of Engineering Trends and Technology (IJETT) – Volume 11 Number 6 - May 2014
2. Ammar Almomani, Atef Obeidat, Karim Alsaedi, M. Al-Hazaimeh Obaida and Mohammed Al-Betar, "Spam E-mail Filtering using ECOS Algorithms" , Indian Journal of Science and Technology, Vol 8(S9), 260–272, May 2015
3. https://wordtothewise.com/wp-content/uploads/2014/02/WttW_SpamFiltering.pdf
4. Mohammand Reza Parsaei & Mohammad Salehi. "E-Mail Spam Detection Based on Part of Speech Tagging", 2015 2nd Internation Conference on Knowledge-Based Engineering and Innvovation(KBEI),November 5-6,2015.
5. Anirudh Harisinghaney, Arnan Dixit, Saurabh Gupta, Anuja Arora, "Text and Image Based Spam Email Classification using KNN, NaIve Bayes and Reverse DBSCAN Algorithm" , 2014 International Conference on Reliability, Optimization and Information Technology ICROIT 2014, India, Feb 6-8 2014
6. Siti-Hajar-Aminah Ali & Seiichi Ozawa, Junji Nakazato & Tao Ban, Jumpei Shimamur, "An Autonomous Online Malicious Spam Email Detection System Using Extended RBF Network", 978-1-4799-1959-8/15/$31.00 @2015 IEEE
7. Yang Gao, Guyue Mi and Ying Tan," Variable Length Concentration based Feature Construction Method for Spam Detection", 978-1-4799-1959-8/15/$31.00 @2015 IEEE
8. Arie Wahyu Wijayanto ,Takdir, " Fighting Cyber Crime in Email Spamming: An Evaluation of Fuzzy Clustering Approach to Classify Spam Messages", International Conference on Information Technology Systems and Innovation (ICITSI) 2014 Bandung-Bali, 24-27 November 2014 ISBN: 978-1-4799-6526-7
9. https://www.tutorialspoint.com/data_mining/dm_overview.htm
10. https://en.wikipedia.org/wiki/Spam
11. http://www.webopedia.com/TERM/S/spam.html
12. http://www.computational-logic.org/iccl/master/lectures/summer06/nlp/part-of-speech-tagging.pdf
13. https://sites.google.com/site/dataclusteringalgorithms/k-means-clustering-algorithm
14. Sunil B. Rathod, Tareek M. Pattewar,"A Comparative Performance Evaluation of Content Based Spam and Malicious URL Detection in E-mail", 2015 IEEE International Conference on Computer Graphics, Vision and Information Security (CGVIS), 978-1-4673-7437-8/15/$31.00 ©2015 IEEE

BIOGRAPHY

Hiral Jagdishbhai Agravat is a Master of Computer Engineering from Noble College of Engineering, Junagadh and she completed her Bachelor of Engineering from Marwadi Education Foundation Group of Institutation College of Engineering, Rajkot.