Email Classification using Classification Method

Mis.Elifenesh Yitagesu[1], Prof Manisha Tijare[2]
[1]computer Science and Engineering, Symbiosis institute of technology, Pune, Maharashtra, India
[2]computer Science/information technology, Symbiosis institute of technology, Pune, Maharashtra, India

Abstract

Electronic mail message became progressively vital and widespread technique communication due to its time speed. And its traffic has risen exponentially with the appearance of WWW. The more and more in email traffic comes additionally with an increasing the utilization emails for illegitimate purpose sizable amount of spam email are inflicting significant issue for web user and web service. Spam is unwanted or a bad email. And email users daily receive more spam email rather than ham email. For this reason we should have to use effective spam filtering technique are based on data classification. This technique is used to classify email as spam and legitimate. To test various classification rules we have use the WEKA interface.

Key Words: *classifier, email, spam and ham email, future selection*

I. INTRODUCTION

E-mail is an efficient, quick and low-cost communication approach. E-mail Spam is non-requested data sent to the E-mail boxes. Spam could be a huge drawback each for users and for ISPs. According to investigation nowadays user receives a lot of spam emails then non spam emails. To avoid spam/irrelevant mails we'd like effective spam filtering strategies. Spam mails area unit used for spreading virus or malicious code, for fraud in banking, for phishing, and for advertising. Therefore it will cause major problem for web users like loading traffic on the network, wasting looking out time of user and energy of the user, and wastage of network information measure etc. There are several approaches are used for spam email classification consistent with data outside of the content of email messages those completely different e-mail classification techniques are together with rule based mostly} and content based techniques. Rule based mostly techniques are: black list, white list, challenge response system. , WL (white list) is a kind of associate mail user agent (MUA) level rule-based filtering method; wherever a white list (WL) filtering method may be a register containing a group of contacts from that e-mail message will be accepted. If associate degree e-mail arrives however doesn't return from one amongst the contacts within the WL, then the email treated as spam and placed within the spam folder. BL (black list) is different from white list. The black list (BL) contains lists of proverbial spammers. Basically once a user gets spam email, the user adds the sender of the spam email to the black list (BL).The new arrived e-mails area unit first checked then if the sender of the email is on the BL, the e-mail is mechanically classified as spam [1]. The second technique is content based techniques .content based techniques are: Using classification algorithms this is the other methods or approaches have been applied in spam classification according to context such as: Naïve Bayes, Decision trees, Neural Network, K-nearest neighbor (KNN), Genetic algorithm and other methods [2]. To solve the problem there is a growing need for emails classification solution.

II. RELATEED WORK

There are some research work that apply machine learning methods in e-mail classification, filtering classification strategies can separated into two categories those based on Machine learning and those are not machine learning. Under machine learning there are different classification algorithms such as: Bayesian, SBPH, SVM, Neural Network, Markova model, memory based pattern discover etc. And under non Machine learning also have different classification approaches such as Black list\White list, signatures, hash base, traffic Analysis, grant listing [3].

They demonstrate that review some of the most popular machine learning methods in email classification. Such as: KNN (K-nearest neighbor) classifier method, Naïve Bayes classifier method, Artificial Neural Networks (ANN) classifier method, Support Vector Machines (SVM) classifier method, Artificial Immune System classifier method, Rough sets classifier method. This machine learning is performs: experiment implementation, and detail algorithm steps those steps are email preprocessing this is the first and most important step, description

of the feature extracted, spam classification, and the last one is performance evaluation [4].

Major three spam filtering technique namely Bayesian Theorem, SVM and K-NN respectively.

Naïve Bayes classifier use the Bayes Theorem of conditioned probability to recognize an email to be spam or not. Email classification process can be divided in to two different phases: the first phase is Training phase and the second is Classification phase. In the first Training Phase each email is first individually categorized to a category (spam/ham). So we must remove html tags, stop words, special characters, articles, proverbs. Extract keywords calculate the frequency of keywords and then save it to database for the selected category. In Classification Phase: In this phase the newly arrived mail is first converting to lower case and stop words, html tags, special characters, articles, proverbs are removed. Extracts the keywords from mails and then calculate the probability of these words from the learning dataset. If the spam probability is higher than the mail is spam otherwise it is no spam/legitimate mail.

SVM can be used to represent a document in vector space where each feature (word) represents one dimension. Identical feature denotes same dimension.

K-NN is a classification algorithm which classifies objects based on K objects having closest pattern in the training sample [5].

In different paper within the concept of email mining, the spam detection is used to identify unsolicited bulk or spam emails by using the data mining method. Generally supported the data chiefly used, spam detection strategies is divided into 2 classes, those classes are content based detection and sender based detection. In Content based detection as the name indicates this class is to spot spamming emails according to the content of the email. The second detection classes, sender based detection is to identify spamming emails not by the content of the email but using the email sender information [11].

A. Content based detection

Classification and semi supervised clustering are the foremost typically used techniques in content based detection [11]

Classification method

Under this method there are different algorithms such as Naïve Bayes classification is widely used in email spam detection.

Support vector machine

Emails area unit classified into two different groups in email spam detection concept. Those are spam and non-spam by a hyper plan. The aim is to search out a hyper plan, which may increase the margin between spam and non-spam category. The key plan of the rule based classifiers is to classify the emails by a group of “IF THEN" rules [11].

Semi-super vised clustering method

In this method use the k mean algorithm this method is one of the widely used clustering algorithm [11].

B. Sender based detection

Classification method

Hear we use K-Nearest Neighbor (K-NN) Classifier and Naïve Bayes classifier.

Semi-super vised bunch technique and email sender name analysis technique area unit the foremost used technique in sender based mostly spam detection [11].

The classification drawback in sender based spam detection is similar to the content based detection drawback. The distinction between sender based spam detection and content based spam detection is that the options used for classification. In sender based detection, the email sender information such as the writing style and the email sender user name is used as the major features. In content based detection, terms extracted from the emails are the major features [6].

In other paper they are proposed to do a method of classifying emails using the concept of Association rule mining for finding interesting hidden pattern in large transactional database. And they are using the algorithms Apriori and FP tree growth there are many approaches have been used for spam classification. According to information outside of the content and in content of email message.

In outside the content of email message we have to use the approaches such as: black list, white list, challenge response system. Spam classification according to content such as Naïve Bayes, Decision trees, Neural Network, KNN, Genetic algorithms.

In the training phase there are different phases such as email classification, email preprocessing, tokenization process, Noise system elimination process, stop words elimination process, suffix stripping process [10].

III. CLASSIFICATION ALGORITHMS

1. Naive Bayes classifier method

Spam filtering is that the best better-known use of naïve theorem text classification it makes use of Naïve mathematician classification to spot spam email.Bayesian spam filtering could be a terribly powerful technique for managing spam.it is supervised learning technique still as a method for classification. Naive theorem text classification is quick, accurate, easy and straightforward to implement [7].

Bayes theorem

The Bayes theorem says that the likelihood of prevalence might rely on the availability or non-availability of another event. This dependency is written in terms of contingent probability.

$P(X|Y) = P(X \cap Y) / P(Y)$

$P(Y/X) = P(Y \cap X) / P(X)$ $P(X \cap Y) = P(X/Y) P(Y) = P(Y/X) P(X)$

The Bayesian classifier uses the theorem of Bayes theorem that says:

$P(Zj \mid d) = P(d|Zj) P(Zj) / P(d)$

Consider every attribute and sophistication label as a chance variable and given a record with attribute (X1, X2…Xn). The aim of this theorem is to predict category Z. we wish to seek out the worth of Z that maximizes P (Z|X1, X2…Xn).

The approach taken is to reckon the posterior likelihood P (Z|X1, X2… Xn) For all worth of Z victimization the Bayes theorem.

P(Z|X1,X2,…,Xn)= P(X1,X2,…,Xn|Z) P(Z) | P(X1,X2,…,Xn) therefore you selected the worth of Z that maximize P(Z|X1,X2…,An).this is similar to selecting the worth of Z that maximizes P(X1,X2,…..,Xn|Z)P(Z) [8].

2. K-nearest neighbor classifier method

In KNN classification the new email that the one to be classified is matched with each of the exemplar emails the ones that have been classified and stored in the system and the K nearest exemplars among are found. Then the class that has the biggest number of exemplars among these K nearest ones will be taken as the class of the new email. Hear we use a version of KNN classifier in which each vote is weighted by the similarity between the new email and the corresponding exemplar in the KNN [9].

KNN is a classification algorithm which classifier object based on K objects having closest pattern in the training sample. To identify the closest pattern objects a number of similarity measures are used among which the most popular is equivalent distance calculated as:

D (p1, p2) = ((X2-X1)2 + (Y2-Y1)2)1/2

Where p1 and p2 represents the points or objects in space having coordinates (x1, y1) and (x2, y2) respectively. Using such a model a training set can be generated for different classes associated with each point. Consider a New Object N is to be classified as one of the predefined classes A or B. The Euclidean distance of that objects will be calculated with the other objects in space and the class of K nearest neighbor is assigned to N [5].

3. Support vector machine (SVM) classifier

Combine the SVM with the genetic algorithm to enhance the performance of SVM. In its simplest form SVM can be used to represent a document in vector space where each feature (word) represents one dimension. Identical feature denotes same dimension. Two of the parameter particularly term frequency (TF) and TF-inverse document frequency (TF-IDF) add price to those vectors. Wherever TF the quantity of times a word occur in a very document TF-IDF uses the on top of TF multiplied by the IDF (inverse document frequency). DF (document frequency) is that the variety of times that word happens altogether the documents. The IDF is outlined as: [5].

IDF (WI) = log (|D| / DF (WI))

SVM could be a powerful state of the art rule with robust theoretical foundation [2]. SVM classifier includes the property of robust information regularization and this SVM might simply handle high dimensional feature areas.

WEKA interface

Weka is one of the data mining tools it may use for classification and clustering. It's a collection machine learning algorithm like classification, regression, clustering, and association rules to accomplish the data mining tasks [2]. The interface can link with email information to gather the information for pre-processing then generate the coaching and take a look at data sets then we have a tendency to convert each set into rail format. We have a tendency to pass coaching set to the rail library to coach the classifier then take a look at the effectiveness with take a look at set [12].

CONCLUSION AND FUTURE WORK

Email is considered as powerful evidence and spam is an every growing menace that can be very powerful. Use effective filtering techniques to avoid spam or irrelevant mail. Those techniques area unit rule based and content based technique. In rule based mostly techniques wherever a white list (WL) may be a register containing a set of contacts from that e-mail message may be accepted. If an e-mail arrives however doesn't return from one in every of the contacts within the white list (WL), then it's treated as unsolicited email and placed within the spam folder.[1] BL (black list) contains lists of better-known spammers. Basically once a user gets spam, the user adds the sender of the spam to the black list (BL). The whole domain of the sender of the spam may be another to the BL. The new coming e-mails area unit checked, and if the sender list is previously stored on the black list (BL), the e-mail is mechanically classified as spam. In content based techniques use the classification algorithm such the basic algorithms are Naïve Bayes, K nearest neighbor (KNN) and Support Vector Machine (SVM).Bayesian classifier is one of the most important and widely used classifier and also its the simplest classification method due to its manipulating capabilities of tokens and associated probabilities according to the users' classification decision and empirical performance. In the future work we have a plan to implement other algorithm to our classification method to achieve better performance.

Acknowledgment

First of all I would like to thank my God whatever he did for me and I would like to thank my supervisor Prof. ManishaTijare. Her support, comments on, and contribution to my work together with encouragement during my studies has been invaluable for me. I also want to thank the people in the Computer Science Department. Thanks for support and interesting discussions. Further, I would like to thank my colleagues at the Department of Computer Science and engineering, for a pleasant and friendly working environment. I am also thankful of Computer Science non-teaching but technical staff, always helped for hardware and software issue.

REFERENCE

[1] MinyiGuo,Yang Xiang Mid Rafiqul Islam wanlei Zhou, "An innovative analyser for multi classifier e-mail classification based on grey list analysis," Journal of Network and computer Application, feabruary 2008.

[2] Rafiqul Islam and Yang Xiang, "Email Classification using Data Reduction method".

[3] Rekhan, SandeepNegi, "Areview on different spam detection approaches ," international journal of engineering trend and technology, vol. 11, may 2014.

[4] W.A. Awad1 and S.M. ELseuofi, "Machine Learning Methods for Spam E-mail Classification," international jornal of computer science and information technology, vol. 3, February 2011.

[5] savitapundalkiteki, santoshkumarbiradar, "effective email classification for spam and non-spam," international journal of advanced research in computer science and software engineering, vol. 4, no. 6, june 2014.

[6] Mrs. Pranjal S. Bogawar, Dr. Kishor. K. Bhoyar, "Email mining: A Review," international journal of computer science issues, vol. 9, no. 1, pp. 429-434, january 2012.

[7] AbhaSuryavanshi, ShishirShandilya, "spam filtering and removing spam content from message by using naive bayesian," international journal of computational engineering and managment , vol. 15, pp. 104-109, july 2012.

[8] S. Roy, A. Patra, S.Sau, K.Mandal, S. Kunar, "an efficient spam filtering techniques for email account," american journal of engineering research, vol. 02, no. 10, pp. 63-73, 2013.

[9] Matthew Chang, Chung Keung Poon *, "using phrases as features in email classification," the journal of system and software, january 2009.

[10] Mohammed A.Naser, AtharH.Mohammed, "Emails classification by data mining techniques ," Journal of Babylon University/Pure and Applied Sciences, vol. 22, 2014.

[11] Jian Pei, and Wo-Shun LukGuanting Tang, "Email Mining: Tasks, Common Techniques, and Tools," april 2013.

[12] Ravi KalkindriSujeet More, "evaluation of deceptive mails using filtering and weka," IEEE sponsored 2nd international conference on innovation in information embedded and communication system ICIIECS'15, 2015.