Scientific Journal of Impact Factor (SJIF): 4.72

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

"Emerging Technologies in the Computer World", January -2017

Security in Email Systems using Spam Filtering

Support Vector Machine and Map Reduce Concept

Pravin Khedkar, Rohit Reshi

Computer Engineering, AISSMS IOIT
Computer Engineering, AISSMS IOIT

Abstract — *The most dangerous online threat now-a-days is the threat in the mailing system. The spam are the threat in the mailing system which are any unwanted and harmful mail for the security purpose. The spam mails should separate from the rest of mails which are useful to the users. This paper surveys different spam filtering techniques. Support Vector Machine (SVM) training problems and need to introduce Map Reduce Hadoop to train SVM. Techniques to separate spam mails are word based. Content based, machine learning based and hybrid. The machine learning techniques are most popular because of high accuracy and mathematical support. SVM is used mostly for machine learning based technique because of its ability to handle data with large attributes, there are also some hurdles in the training process of SVM, that can't be given as input, both of these problem should be solved by implementing the training algorithm on map reduce (Hadoop) framework which gives up to 6 times speedup than sequential algorithm.*

Keywords- *Spam filtering techniques; Word based; Content based; Machine learning based; Support Vector Machine (SVM); Map Reduce; Web Services*

I. INTRODUCTION

In modern day communication tools, e-mail system is mostly used. E-mail became boon for business because of its wide availability-mail is fastest way of communication as there is no need to wait for response. The main danger for the e-mail is spam mail. Unwanted mail is also known as spam mail. Mails which are sent in bulk are spam mails. Phishing websites, malicious attachments are sent by spam e-mails. Spam e-mails also include malicious scripts and executable attachments [7]. The threat and the major problem like this of getting unwanted and malicious programs motivated us to build the system which separates spam but also blocks the accounts sending it.

II. GOALS AND OBJECTIVES

- With increasing security measures in network services, remote exploitation is getting harder.
- Therefore efficient filtering methods for spam messages are needed.
- In this project, we introduce a more proactive approach that allows us to detect valid and spam mail.
- The main aim is to design and develop a spam detecting system for emails using classification algorithm i.e. SVM.

III. EXISTING DIFFERENT SPAM FILTERING APPROACHES

3.1. Whitelist/Blacklist.

In this approach the list is used. In the whitelist includes the email address or entire domains. Blacklist is exactly opposite to the whitelist. It contains addresses which are harmful to users. Automatic list management tool is use in this method [5].

3.2. Bayesian Classifier.

In this type of approach probability and prediction is used. Particular words are occurring in the both spam emails and non-spam emails. These words have the probability of occurring frequently. The filters which are used doesn't know this words in advance .So we have to train words first so it can build them up. After training probabilities of words are used to calculate the probability that an email having particular set of words in it belong to either spam or valid emails. Each particular word or the interesting word have the probability of occurring then this words are taken. This contribution is called as posterior probability [6]. This posterior probability is calculated by Bayes' theorem. Then, the emails spam probability is computed all over the word in the emails. If this total value exceed over certain threshold then the filters will mark emails as spam.

3.3. Signatures.

In this approach signature is generated in which each spam message have unique hash value signature. The filters then compare the values with the previous spam mails values stored. It is impossible to have same values to the valid e-mails.

3.4. Mail Header Checking.

This is very known approach in the spam filtering. In this approach the mail headers are checked with the simply set of rules .This rules are decided at first level. This approach is easily implemented for the spam filtering.

IV. DISADVANTAGES OF EXISTING SPAM FILTERING APPROACHES

- In the whitelist and blacklist spam filtering approach the e-mail can be easily penetrated by spammer.
- In the Signature based approach the spam emails are unable to identified as first it has to report as spam and hash distributed.
- In mail header checking approach there is high positive false rate and rejecting connection requires additional information
- In Bayesian classifier approach it is rely on Naïve Bayes filtering which assumes event occurs independent of each other.

V. SPAM FILTERING TECHNIQUES USED IN THE STANDARD SYSTEMS AS YAHOO AND GMAIL

Gmail supports many authentication systems such as SPF (Sender Policy Framework), Domain Keys, and DKIM (Domain Keys Identified Mail).

5.1. SPF (Sender Policy Framework).

SPF is the email validation system designed to detect the spoof mails by providing mechanism to allow receiving mail exchangers to check that incoming mails are from authorized host domain networks or not.

5.2. DKIM (Domain Keys Identified Mail).

It is the method designed for email authentication used to detect email spoofing. It is allows to check on receiver side that the incoming email was from authorized domain or not. Verification is done using the signer's public key published in the DNS.

VI. PROPOSED TECHNIQUE

6.1. Support Vector Machine.

The most recent technique used in text classification is known as Support Vector Machine. In 1995, Vapnik proposed a novel method called support vectors machine (SVM) to perform pattern categorization. It is technique of pattern recognition and data analysis. It is the training sample which is a set of vectors of n attributes in the machine learning. We then assume that we are in a hyperspace of n dimensions, and that the training sample is a set of points in the hyper-space. Now let us take the example of case of two classes (as in the spam problem). Using Support Vector Machine we have to classify the problem using the hyper plane that would separate the class points of one class from the points of other classes. The distance of the points from the hyper plane has to be maximum for good separation [3].

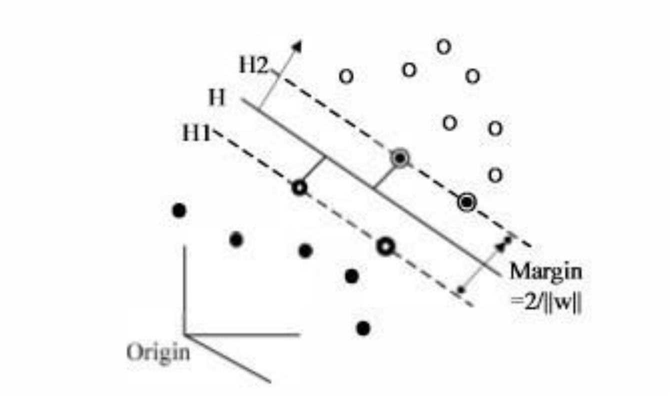

Figure 1. The optimal separating hyper plane under linear separability

In Figure. 1 Plane H1 is a good classifier and Plane H2 is not so good or even doesn't classify. The distance between the two planes is called the interval or the margin. In case we can't find the hyper plane, when the point are not linearly separable, the hyperspace could be extended to required distance according to the need. If all the samples in training corpus can be correctly plotted out by a hyper plane and the distance of proximate vectors away from plane H, we call it different-vectors, achieves the largest value, this plan is deemed to the most optimal categorization hyper plane. Its equation is $W_tX + b = 0$, where w is the normal of categorization plan in that vector w is. We call this different-vectors support vector, as the point including double circle. One of the most interesting features of SVM technique is that to find the appropriate plane, SVM method just explores the nearest of all the points.

6.2. Map Reduce.

Training the SVM set completely is a bit difficult, so the input data is divided into various sub-sets of data and is individually being worked on. Generally Sequential Minimal Optimization (SMO) is used to do that. The only problem with SMO is that it can't handle large data sets. So this problem could be eradicated by using the Map Reduce framework. A Map Reduce framework usually splits the training data-set to various different independent chunks of data sets which are processed by the map tasks in a completely parallel manner. The framework sorts the outputs of the maps, which are then input to the reduce tasks. The Map Reduce framework is most popularly used in the research fields like Mars, Phoenix, and Hadoop etc. Amongst above implementations Hadoop is mostly used in research field because of it 1 Terabyte sort achievement and support [2].

The Map Reduce programming paradigm has the massive scalability of carrying hundreds and thousands of nodes in a Map Reduce cluster. Using Map Reduce frameworks large training input data sets of SVM are splitted into small size data chunks. These chunks produced by SVM training data sets are then assigned to map tasks. The figure below shows the splitting of SVM input file and working of Map Reduce paradigm.

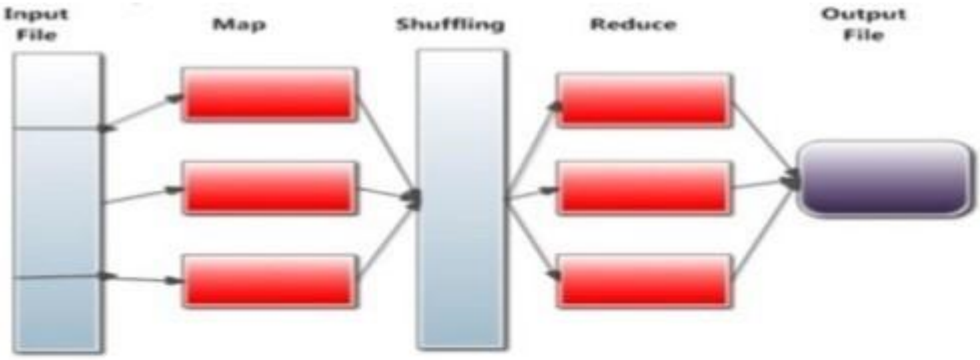

Figure 2. Working of Map Reduce paradigm

Each task of the Map Reduce paradigm can perform the serial SMO independently on their respective training sets. Map Reduce gives output in form of {key, value} pair. Reduce task has {key, value} pairs generated by each Map task as input and combines result of all Map tasks to get final output. All map and reduce tasks run independently.

VII. THE ADVANTAGES OF THE SVM TECHNIQUE CAN BE SUMMARIZED AS FOLLOWS

- By introducing the kernel, SVMs gain flexibility in the choice of the form of the threshold separating different companies.
- Since the kernel introduced by the SVM, a non-linear transformation assumes nothing about the form of the data transformation, which makes data necessarily separable.
- SVM provide a good out-of-sample generalization.
- SVM deliver a unique solution, since the optimality problem is convex. This is an advantage compared to Neural Networks, which have a number of solutions associated with local minima.

VIII. DISADVANTAGE OF THIS TECHNIQUES

A common disadvantage of techniques such as SVM is that is doesn't has the transparency to show the results. Being the dimensions too high, the score of different or all companies can't be represented by simple parametric functions of the financial ratios. The financial ratios don't have the constant weights. Thus the variable score is generated from the marginal contribution of each financial ratio.

IX. CONCLUSION

- To develop a mailing system capable of isolating the spam mails and containing them separately in spam folder. Also it includes the blocking of particular account for sending spams emails in bulk.
- The disadvantages of the existing email system spam filtering systems would be overcome in this project.
- The email system which will take input as the email received by its users , the received email is then validated as 'spam' or 'ham' by comparing it to the keywords used as a spam mail using the algorithms.

- As the spam emails can be send or are sent in a bulk in normal email system, in our system the account sending spam mail in bulk will be blocked.

X. FUTURE SCOPE

The email system which will take input as the email received by its users , the received email is then validated as 'spam' or 'ham' by comparing it to the keywords used as a spam mail using the algorithms. As the spam emails can be send or are sent in a bulk in normal email system, in our system the account sending spam mail in bulk will be blocked

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] Saadat Nazirova," Survey on Spam Filtering Techniques", Communications and Network, 2011, 3, 153-160 doi:10.4236/cn.2011.33019 Published Online August 2011 (http://www.SciRP.org/journal/cn)

[3] Amol G. Kakade, Prashant K. Kharat," Spam filtering techniques and Map Reduce with SVM: A study", 2014 Asia-Pacific Conference on Computer Aided System Engineering (APCASE).

[4] J. Vijaya Chandra, Dr. Narasimham Challa,, Dr. Sai Kiran Pasupuleti, "A Practical Approach to E-mail Spam Filters to Protect Data from Advanced Persistent Threat", 2016 International Conference on Circuit, Power and Computing Technologies [ICCPCT]

[5] Tarjani Vyas, 2Payal Prajapati, & 3Somil Gadhwal," A Survey and Evaluation of Supervised Machine Learning Techniques for Spam E-Mail Filtering",

[6] Wanqing You, Kai Qian, Dan Lo, Prabir Bhattacharya, Minzhe Guo, Ying Qian," Web Service-enabled Spam Filtering with Naïve Bayes Classification", 2015 IEEE First International Conference on Big Data Computing Service and Applications.

[7] 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.

[8] Godwin Caruana1, Maozhen Li1,3 and Man Qi2," A MapReduce based Parallel SVM for Large Scale Spam Filtering", 2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)

[9] Amol G. Kakade1, Prashant K. Kharat2, Anil Kumar Gupta," Survey of Spam Filtering Techniques and Tools, and MapReduce with SVM ",IJCSMC, Vol. 2, Issue. 11, November 2013, pg.91 – 98 .

[10] Salwa Adriana Saab, Nicholas Mitri, Mariette Awad," Ham or Spam? A comparative study for some Content-based Classification Algorithms for Email Filtering", 17th IEEE Mediterranean Electrotechnical Conference, Beirut, Lebanon, 13-16 April 2014