Survey on Spam Filtering Techniques and Mapreduce

Prajakta S. Patil[#1], Prof. Rashmi A. Rane[#2], Prof. Madhuri A. Bhalekar[#3]
Department of Computer Engineering
Maharashtra Institute of Technology Pune, India

Abstract*—Spam Email, also known as junk email , is a subset of electronic spam involving nearly identical messages sent to numerous recipients by email. The messages may contain disguised links that appear to be for familiar websites but in fact lead to phishing web sites or sites that are hosting malware. Spam email may also include malware as scripts or other executable file attachments. Spam is any unwanted and harmful mail. Separation of spam from normal mails is essential. This paper surveys different spam email filtering techniques. The different techniques are Machine learning based, list based, content based and hybrid or other. Machine learning based, is mostly used because of high accuracy and mathematical support.*

Keywords—*Spam filtering techniques, Machine learning based ,content based, word based.*

I. INTRODUCTION

The email system is one of the most used, modern day communication tools. Wide availability of an email system is working as a boon for business. Email is a quick as no need to wait for the response and it is straight forward way to stay in touch with the all. One threat to an email system is spam mail. The spam is nothing but the unwanted mail. The definition of spam is mail which is sent in bulk. Spam email, also known as junk email which has the abundant recipients. Normally, spam's contain links to phishing web sites or malware hosting web sites. Spam email may also include malware as scripts or other executable file attachments. Beside these, for checking legitimacy of mail consumes valuable time. According to the SMX email security provider, the live spam percentage is about 79.5% [1].

The average size of spam is 16 KB. For the separation of such spams from important mails, spam filtering is important. Amongst these, Naïve Bayesian classification, Support Vector Machine, K-Nearest Neighbor are most used and appreciated by researchers. Also, number of freeware and paid tools are available for spam filtering, which makes use of these techniques. Machine learning technique like Support Vector Machines (SVM) can be applied efficiently in spam filtering. The training process of SVM is the compute intensive process, so there is a lot of scope to introduce the Map Reduce platform for spam filter training. Map Reduce paradigm works with Map and Reduce tasks and these tasks are independent. During the SVM training, for each data point in hyperspace, maximize the margin in hyper plane.

Data mining is the process of discovering new patterns from large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics and database systems. This problem has been researched by many scholars in all kinds of application area for many years and many data mining methods have been developed and applied to practice. However, most classical data mining methods out of reach in practice in face of big data. Computation and data intensive scientific data analyses are increasingly prevalent in recent years. Support Vector Machines (SVMs) are powerful classification and regression tools, but their compute and storage requirements increase rapidly with the number of training vectors, putting many problems of practical interest out of their reach. Efficient parallel algorithms and implementation techniques are the key to meeting the scalability and performance requirements entailed in such large scale data mining analyses.

II. OVERVIEW OF EMAIL SYSTEM

In this section, a brief explanation of email protocol and the process of filtering will be elaborated. Simple Mail Transfer Protocol (SMTP) is the first protocol which transfers the emails by some commands. Figure illustrates SMTP commands. First, TCP/IP (Transmission Control protocol and Internet Protocol) connection starts between sender and the associated mail server. Following that, the SMTP commands begin with a Hello message and announcing the acceptance of the session between the client and the server. This process ends when the message is accepted by the mail server. TCP connection disconnects if there is no more message from the client to the mail server. When the email is delivered by the server, the filtering phase is started. Based on the server filtering policy, Blacklist and White list filtering is stared to examine if the email is a spam or a valid one. If the email is recognized as a valid one, it is sent to receiver's inbox otherwise the email is blocked or transferred to the spam folder. When a Grey list filtering is used in relevant mail server, the email is rejected for the first time. Afterward the body of the email is tested with content-based and rule-based filters according to the standards of the administrator

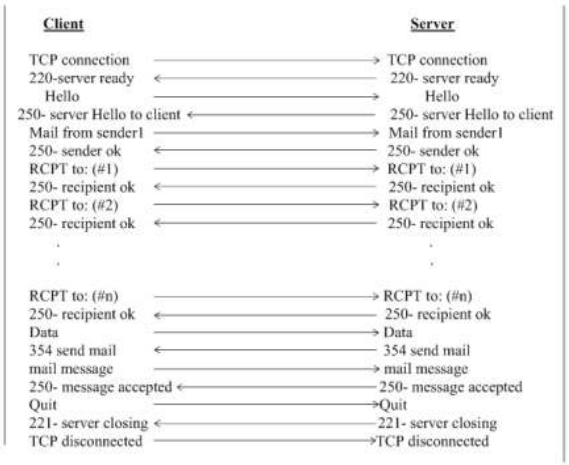

Fig 1: Overview of email system

III. SPAM FILTERING TECHNIQUES

In this section some mostly used spam filtering techniques are discussed. They are mainly classified in four classes as, a list or word based, content based, machine learning based and others or hybrid.

1. Machine Learning Based Techniques

This is mostly adopted and trusted technique in the scientific community. Strong mathematical background is reason behind success and popularity of machine learning based spam filtering[1].

A. K- Nearest Neighbor:

Non parametric nature and lazy learning algorithm are features of K Nearest Neighbor (KNN) . It is non parametric technique as it does not make any consideration on the distribution of original data. KNN algorithm is also famous for its lazy learning phase. In other words, the training phase is quite fast or there is no training phase. KNN keeps all the training data with it and this nature of KNN is termed as 'lack of generalization'. This training data is used further in the testing phase. While taking decision based on the entire training data set are taken under consideration. The technique called SVM works exactly opposite of it, where all non support vectors can be discarded. The training phase is costly in terms of memory as we need to store all training data.

The working of this technique resolves around concept called
characteristics vector. The characteristic vectors are measure of similarities among all messages. Any new mails are classified in any of spam or ham class on basis of distance of that mail from both classes. Normally, this technique does not use a separate training phase. The time complexity of KNN classification process is *O (vl).* Where, *v* is the size of the characteristic vector and *l* the sample size. False classification is headache in KNN which is wiped with *t/k* rule. If at least *t* mails in *k* neighbors of the mail *m* are spam, then mail *m* is classified as spam, otherwise, it is ham.

B. Neural Networks:

In the biological nervous system, the foundation element is 'neuron'. For information processing like human brain an Artificial Neural Network (ANN) model is often used that is inspired by biological nervous systems [5]. In training phase ANN builds a model for classification by learning similar examples. ANN can be used in data classification or pattern recognition. In the process of spam filtering, ANN builds a model which classifies new mails as spam or ham. NN can have two forms, namely the Perceptron and the Multilayer NN on basis of hidden layer. Basic Neural Network with just two layers, namely, the input layer and output layer is called as Perceptron. Multilayer NN have one or more hidden layers. Output of Perceptron paradigm is a function f(x) = *Xw*+b. Where, *w* is a weight vector and *b is* a bias or threshold vector. The training phase of NN is iterative in nature. During an each iteration of the training phase in spam filtering, weight vector and bias vectors are adjusted, to correctly classify new instance in training dataset or sample.

Let, *x* is any vector or instance of mail from an input sample, for which the NN fails to classify. Also, let wi and *bi* are the eight vector and bias respectively during the i th iteration. For every failed instance during the training phase, training continues until its correct classification and *wi* and *bi* are updated. In ANN training phase is said to be completed if all instances in input sample are correctly classified as a spam or ham. In this case, we say that the NN converges. Spam filter tools which work on neural network are NAGS spam filter, Spam Bouncer etc.

C. Support Vector Machine:

The Support Vector Machine [1,2] is one of the most modern techniques used in mail classification. In, abstract view, it is a kernel machine with the strong mathematical base. It is a technique of pattern recognition and data analysis. The training sample is a set of vectors of *n* attributes. At the end of training phase, we can say that, we are in hyperspace having dimensions equal to the number attributes. In process of spam filtering, SVM builds hyperspace with two classes, namely, spam and ham. These two classes are separated by a hyperplane. Every mail instance is treated as a single point with *n* dimensions in hyperspace. The distance between the hyperplanes and points of each class, is kept maximum, for good separation.

Here in fig. 2, Plane1 is good classifier and Plane2 doesn't
classify all instances [1]. It may also happen that, we can't find good separator hyperplane (Plane 1) as in fig. 2. In such case, hyperspace is called as non-linearly separable. To obtain linear separation in the non-linearly separable hyperspace, it is extended to

more dimensions. SVM method considers only the nearest points in hyperspace to find hyperplane.

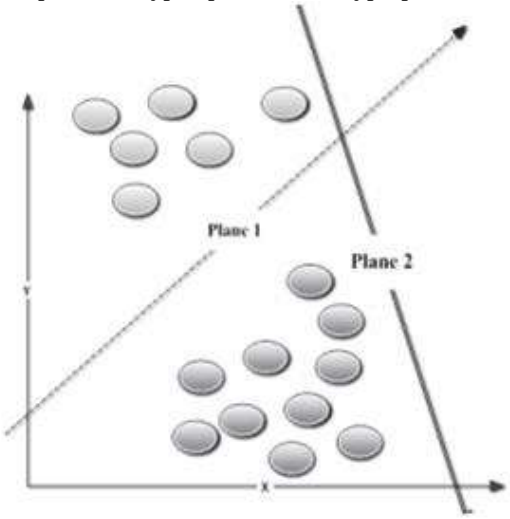

Fig. 2. Hyperplane that separate the two classes

D. *Naïve Bayes Classifier:*

In machine learning, naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independenceassumptions between the features[6].

Basically it operates with 'Bag of Words', which it represented as an unordered collection of words and discusse in the section below. The Naive Bayes classifier works by calculating spamicity of tokens or words in spam and non-spam e-mail. Bayesia inference is used later to calculate a probability that an email is or is not spam. Spamicity of a word is the probability of a word being spam.

For calculating spamicity of total message, this technique lets us combine the probability of multiple independent events into one number. Each word has particular probabilities of occurring in spam email and in legitimate email. The overall process has the following steps:

1. Train the filter.
2. Calculate the probability of words.
3. Combine the word probability to classify mail.

At the start, emails are manually classified as spam or ham. After training, the probability of each word in the spam and legitimate mail is calculated by the following formula. Then this data is stored by the spam filter in its database. Filter also maintains a spam token database and the non spam token database which contains count of each word in email. Disadvantages of this technique are

1. Words which occur in spam are misspelled.
2. Spammers insert sensitive words in the form of images in a spam mail and Bayesian Classifier can't analyze images.

2. CONTENT BASED TECHNIQUES

It is a very popular technique to avoid spam, in which mails are evaluated for words or phrases to determine mail as a spam or legitimate.

A. *Word Based:*

It blocks mail as a spam, if mail has certain word having spamicity character. Mostly, spams contains the terms which are rarely used in legitimate mails. So, it is easy to block the spams. One serious problem is that if a filter is configured to block mails containing more common words then it increases the false positivity. The list of such words is available online

B. *Heuristic filters:*

It outperforms the normal word based filter. The word heuristic refers to some intuitive criteria, rather than simple technical metrics. Mostly point and score is criteria to classify the mail as legitimate or spam. More points are assigned to words or a term which occurs frequently in spam. The terms frequently used in legitimate mails are assigned with low score. At the end, a score of mail is calculated. If the score is beyond some predefined threshold, it is marked as spam. Experienced spammers can easily by pass this type of filter by avoiding use of terms with the high score. Also Heuristic filters make use of various algorithms to examine the email.

IV. SVM AND MAPREDUCE

Training phase of SVM needs much more time than actual classification phase. SVM training time minimization is the global minimization problem. During training, entire dataset is used to get the final output, so there is a lot of room for parallelization. Generally algorithm Sequential Minimal Optimization (SMO) is used for implementing SVM. General approach is to split the training data and use a number of SVMs to process individual data chunks. Another problem with SMO is its scalability [4]. It can't handle dataset of large size. We can wipe out these two problems by introducing MapReduce framework . There are popular implementations of MapReduce like Mars , Phoenix , Hadoop etc. Amongst above implementations, Hadoop is mostly used in research field because of it 1 Terabyte sort achievement and support.

In industry, giants like Microsoft, Yahoo, and Google etc use MapReduce in the process of spam filtering. MapReduce programming paradigm allows for massive scalability across hundreds or thousands of nodes in a MapReduce cluster. This approach minimizes memory requirements drastically to store the matrix of input mail instances. With MapReduec large input file for training SVM is spited into small size data chunks. Input data chunks are treated as Key (K) and Value (V). Each map task works with one single data chunk, so numbers of data chunks are normally equal to the number of map tasks.

Each Map task can almost independently run serial SMO on their respective training set. Output produced by map tasks is passed through shufflers, combiners, sorters etc to group same Keys (K) at near to each other. MapReduce gives output in the form of {key, value} pair. The Reduce task has {key, value} pairs as an input, generated by each Map task. Then it combines the result of all Map tasks to get final output.

All map and reduce tasks run independently.Input data set is divided in n data chunks as Data*1*…Data*n*. Later, Map tasks as Map*1*…Map*n* are generated for each data chunks respectively. Each map task generates support vectors as SV*1*…SV*n*. The number of reducer tasks can be 1 to n depends on dataset and hardware availability. The Reducer task combines map tasks given as input to it to generate final.

V. CONCLUSION

Neural Networks follow how the human brain works by having a network of neurons. Each neuron is equivalent to the logistic regression unit. NN is very good at learning non linear functions. One disadvantage with NN is that training time is very long as compared to others. The major feature of NN is that multiple outputs can be learned at the same time. SVM is supervised learning technique and provides binary classification mechanism. Training SVM is easy compared to NN. It works best when you have a small set of input features. The big issue is that, it can't deal with the large number of input data, so Map Reduce (Hadoop) can be used effectively for training. The Naïve Bayes classifier views problem as conditional probability estimation. Strengths of Naïve Bayes are high scalable nature and incremental learning. In case of KNN, the nearest data points are calculated as Euclidean distance. All training data is rem embered. Plus points are simplicity as no models need to be trained.

In the real world, mostly data have the large number of
attributes and SVM is good at handling that data as compared to other techniques. Also in comparison to NN and Naïve Bayes classifier, SVM is good and clearer for solving complex nonlinear functions. To minimize training time Map Reduce can be effectively used.

ACKNOWLEDGMENT

We express our sincere thank to all the authors, whose papers in the area of spam filtering are published in various conference proceedings and journals, and to all authors and
organizations of referred websites.

REFERENCES

[1] Amol G. Kakade1, Prashant K. Kharat2, Anil Kumar Gupta, Survey of Spam Filtering Techniques and Tools, and Map Reduce with SVM, International Journal of Computer Science and Mobile Computing Vol.2 Issue. 11,November-2013, pg. 91-98.

[2] Puch-Tran Ho , HEE Su Kin, Application of Sim Hash Algorithm and Big Data Analysis in Spam Email Detection System, International Journal of Computer Applications (0975 8887) Volume 39 No.6, February 2014.

[3] Sahil Puri1, Dishant Gosain2, "COMPARISON AND ANALYSIS OF SPAM DETECTION ALGORITHMS,International Journal of Application or Innovation in Engineering and Management,Volume 2, Issue 4, April 2013.

[4] Godwin Caruana, Maozhen Li, Yang Liu, An ontology enhanced parallel SVM for scalable spamlter training, Neuro computing Elsevier, vol. 108, pp.45-57, 2013.

[5] L. Zhang, J. Zhu, T. Yao, An evaluation of statistical spamltering techniques , ACM Transaction on Asian Language Information Process, vol. 3, pp.243269,2004.

[6] Available:http://en.wikipedia.org/wiki/Naive