Novel Survey on Email Spam Filtering Methods

Diksha M. Bhalerao
Bharati Vidyapeeth College of Engineering
University of Mumbai,
bhaleraodiksha21@gmail.com

Prof. Dr Dayanand R. Ingle
Bharati Vidyapeeth College of Engineering
University of Mumbai,
dringleus@gmail.com

***Abstract*:**Spam emails are causing major resource wastage by unnecessarily flooding the network links.The cost of spam is borne mostly by the recipient, so it is a form of postage due advertising. This paper describes how different methods can be used for spam filtering.To protect against unsolicited e-mails there are number of techniques presented with goal of efficient, accurate spam filtering. Few previous spam filters can meet the requirements of being user-friendly, attack-resilient, and personalized. This paper presents a literature survey into the state of research on spam filtering methods and how it is useful for user’s lives.

Keywords:*Spam Filtering, Social networks, Support Vector Machine, Artificial neural network, content spam*

*****

1. INTRODUCTION

Now day’s electronic mail is becoming popular as 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 sexual enhancers, cheap drugs and herbal supplements, health insurance, travel tickets, hotel reservations, and software products. They can also include offensive content such pornographic images and can be used as well for spreading rumours and other fraudulent advertisements such as make money fast[11].

E-mail spam has continued to increase at a very fast rate over the last couple of years. It has become a major threat for business users, network administrators and even normal users. A study in July 1997 reported that spam messages constituted approximately 10% of the incoming messages to a corporate network. More recently, Message Labs stated in its 2006 Annual Security Report that spam activity has increased significantly in 2006 with levels that reach 86.2% of the e-mail traffic. The report has also indicated that largely due to the increased sophistication of robot networks, a.k.a. botnets, the spam volumes have increased by 70% over the last quarter of 2006 which in turn increased the overall email traffic by a third. Based on projections of current analysis and trends, it was expected that by the end of 2007, spam will continue to rise, reaching a plateau at around 92% of e-mail traffic. There is a prediction that by year 2015 spam will exceed 95% of all e-mail traffic. Although these figures might not be accurate enough, what can be concluded is that spam volume is dramatically increasing over years.

Spam can be very costly to e-mail recipients; it reduces their productivity by wasting their time and causing annoyance to deal with a large amount of spam. According to Ferris Research, if an employee got five e-mails per day and consumes 30 seconds on each, then he/she will waste 15 hours a year on them. Multiplying This by the hourly rate of each employ in a company will give the cost of spam to this company. In addition, spam consumes the network bandwidth and storage space and can slow down email servers. Spam software can also be used to distribute harmful content such as viruses, Trojan horses, worms and other malicious codes. It can be a means for phishing attacks as well. As a result, spam has become an area of growing concern attracting the attention of many security researchers and practitioners. In addition to regulations and legislations, various anti-spam technical solutions have been proposed and deployed to combat this problem. Front-end filtering was the most common and easier way to reject or quarantine spam messages as early as possible at the receiving server.

The rest of the paper is organized as follows: section 2 describes the Literature Survey. Section 3 describes the final conclusion based on Literature Survey.

2. LITERATURE SURVEY

This study is used to show how the different spam filtering methods can be also useful in the day to day communications .Ten papers have been studied which uses various methods, which are described as follows:

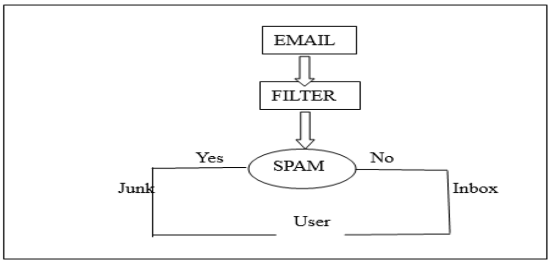

fig.1 Process of Spam Filtering

In paper “An Email Spam detection using SVM and RBF”,[1] Reena S. et al., focuses on to make a RBF NN technique and then compared it with SVM based on two parameters i.e. precision and accuracy. This is an efficient spam filtering technique which gives high precision and accuracy. Here the author has proposed the RBF technique. It is the neural network technique which uses hidden neurons to process the input and to give the output. In this technique the RBF has collected the spam words and formed the spam word dictionary. These words are used for training and testing. The Liebenberg algorithm is used in this technique. Results obtained from the RBF are compared with the SVM.

The Paper “Spam Mail Detection Using Artificial Neural Network”,[2]proposes a spam detection system to detect text as well as image based spam using ANN algorithm. In this system, pre-processing of email text before executing the algorithms is used to make them predict better. Using this system High level, low level, and combination of both the features of image in a spam mail can be predicted.

Deepak Agarwal [3] focuses on a popular machine learning algorithm SVM with different parameters using different kernel-functions. It is evaluated to get best accuracy. Different kernel functions are implemented for spam filtering. The author has used four types of kernels: linear kernel, polynomial kernel, RBF and sigmoid kernel. After this the accuracy is estimated for all the kernels at all the combinations of train files and test files. Various Machine learning methods are being used to classify spammer’s emails from legitimate emails.

In “A Novel Method for Detecting Spam Email using KNN Classification with Spearman Correlation as distance Measure”,[4]focuses on to make a RBF NN technique and then compared it with SVM based on two parameters i.e. precision and accuracy. This is an efficient spam filtering technique which gives high precision and accuracy. Spearman’s connection coefficient is a factual measure of the quality of monotonic relationship between matched information. Then KNN algorithm is used with Spearman Correlation. Spearman correlation coefficient is used as distance measure in KNN classification technique.

Rahul Bansod[5] proposes a technique to classify text and image spam mails using Artificial Neural Network with negative and positive weight measure based on their probability of being promotional or non-promotional word. The classification is based using ANN. OCR tool is used to extract images and texts from image. Accuracy, flexibility and speed are the main features that characterizes a good OCR system. Here the email contents are analyzed and list of words are weighted accordingly to their probability of being a spam oriented word. Based on the value of cognitive load author differentiates between spam and ham mails.

In [6] Savita T.et al., focuses on an algorithm for email classification based on naïve Bayesian theorem. The purpose of this is to automatically classify mails into spam and legitimate message. The mails are classified on the bases of email body. This algorithm found to be effective and reasonable method for email classification. The first number is the number of times that the number of times that the word has Occurred in legitimate emails. The second number is the number of times that the word has occurred in spam emails. The GA-SVM algorithm show an improvement from SVM algorithm for spam detection. In [7]T. Hemalatha et al., proposes an enhanced filtering measure by using a machine learning technique based on content filtering. In this a spam classification method based on machine learning and content feature.

Sufian Hameed et al., [8] proposes a novel spam system “LENS” which is used to select legitimate and authentic users from outside the recipient’s social circle and within pre-defined social distances. It is proved to be fast in processing emails and scales efficiently with increasing community size. LENS is quite fast in processing emails and also compared its performance with the most popular content based filter. In LENS, the MS is responsible for executing the protocol on the behalf of the email users. Each email user can explicitly control its community and can give feedback by reporting spam emails. All the LENS enabled MSs are assumed to be legitimate with a valid certificate issued from Trusted Authority.

In [9] Michael S. et al., proposes SocialFilter which is a trust-aware collaborative spam mitigation system. SocialFilter is a first collaborative unwanted traffic mitigation system that assesses the trustworthiness of spam reporters by both auditing their reports and by leveraging the social network of reporters administrators. These reports concern spamming hosts identified by their IP address. SocialFilter is a trust layer that exports a Measure of the system belief that the host is spamming. SocialFilter nodes are managed by human administrators. The nodes maintained by competent and trusted admins are likely to generate unreliable reports. The social relationship between admin known to be less competent are likely to generate unreliable reports.

In [10] Ze Li et al., proposes the filter which is accurate and user-friendly called “SOAP” (Social network Aided Personalized and effective spam filter). This filter integrates three components into the basic Bayesian spam filter: social closeness-based spam filtering, social interest based spam filtering and adaptive trust management .This filter is user-friendly, attack resilient and personalized. The first number is the number of times that the number of times that the word has Occurred in legitimate emails. The second number is the number of times that the word has occurred in spam emails. The GA-SVM algorithm show an improvement from SVM algorithm for spam detection. The Naïve Bayes classification is based on Bayes' rule of condition probability. Entire email classification process is divided into three phases or steps. Learning phase, training phase and execution phase.

Table 1. Summary

Sr No	Paper title	Author	Method	Advantages	Limitations
1.	Email Spam Detection Using SVM and RBF	Reena S.	RBF Neural Network and SVM	1. Gives high precision and accuracy.	1. SVM does not perform better as compared to RBF.
2.	Spam Mail Detection Using Artificial Neural Network.	Harshal D. et al.	ANN Algorithm.	1. High level, low level, and both the features of image in a spam mail can be predicted.	1. It is not attack-resilient.
3.	Spam Filtering using SVM with Different Kernel Functions.	Deepak A. et al	SVM Algorithm	1. Accuracy obtained is Higher.	1. Speed and size for training and testing is more.
4.	A Novel Method for Detecting Spam Email using KNN Classification with Spearman Correlation as Distance Measure	Ajay S.	K nearest neighbour classification	1. Achieves high accuracy.	1. More time is required for execution.
5.	Email Spam Classification Using Artificial Neural Network with Weight Measures.	Rahul Bansode	ANN Method	1)Easy to implement. 2)Self-learning capability.	1) Learning can be slow.
6.	Effective Email Classification for Spam and Non-Spam	Savita Teli et al.	Naïve Bayes Classification Algorithm	1. Accuracy increases for large datasets.	1. It is not user friendly. 2. Vulnerable to poison attack.
7.	Spam Filtering in Online Social Networks Using Machine Learning Technique.	T. Hemlata et .al	Machine learning(RBFNN) and content features	1. Performance is better.	1. Word combinations are not used which can give better classification.
8.	LENS: Leveraging Social Networking and Trust to Prevent Spam Transmission.	Sufian Hameed et al.	LENS	1. Resilient to poison attacks. 2. User Friendly.	1. Vulnerable to impersonation attack. 2. Not personalized.
9.	Social Filter: Introducing Social Trust to Collaborative Spam Mitigation	Michael S. et al.	Social Filter	1. Resilient to Sybil attack.	1. Spammer may use dynamic IPs.
10.	SOAP: A Social Network Aided Personalized and Effective Spam Filter to Clean Your E-mail Box	Ze Li et al.	SOAP	1. It is personalized, attack-resilient and user-friendly.	1. Vulnerable to impersonation attack.

In [1] Liebenberg algorithm is used and RBF performs much better than SVM. Two lists are used: Black list and White List [2] and spam mail is predicted. [3] evaluates SVM to get better accuracy. In [4] author focuses on to make a RBF NN technique and then compared it with SVM based on two parameters i.e. precision and accuracy. [5] proposes a technique to classify text and images. In [6] the author proposed a system which automatically classify spam messages and non-spam messages. [7] uses a technique based on machine learning and content feature. [8] uses a method which drastically reduces the consumption of internet bandwidth by spam. In [9] SocialFilter system that enables nodes with no email classification functionality to query the network on whether a host is a spammer. SOAP exploits the social relationship among email correspondents to detect the spam adaptively and automatically [10].

3. CONCLUSION

Based on these papers, it is studied that spam filtering plays a vital role in everyday life. It helps to understand how spam filter should be attack-resilient, and user friendly. In the future it will improve time and accuracy of process of spam detection in Bayesian filter.

REFERENCES

[1] Reena Sharma, Gurjot Kaur" Email Spam Detection using SVM and RBF", 7 April 2016.

[2] Harshal Deshmukh, Chetan Nandeshwar, Sagar Wanjari,Pankaj Bhardwaj,Devendra Ramtekkar, Rajesh Nasare "Spam Mail Detection Using Artificial Neural Network", 2016.

[3] Deepak agarwal, Rahul Kumar "Spam filtering using SVM with Different Kernel Function", 2016.

[4] Ajay Sharma, Anil Suryawanshi "A Novel Method for Detecting Spam Email using KNN Classification with Spearman Correlation as Distance Measure" ,2016

[5] Rahul Bansod,R.S.Mangrulkar,V.G.Bhujade" Spam Classification using Artificial Neural Network with Weight Measures", 2016.

[6] Savita Teli,Santoshkumar Biradar "Effective Email Classification for Spam and Non-Spam", VOL.4, Issue 6, JUNE 2014.

[7] T.Suganya,T.Hemlatha"Spam Filtering in Online Social Networks using Machine Learning Technique",Vol 2-Issue 1,Jan 2014.

[8] Sufian HAmeed,Xiaoming Fu,Pan Hui,Nishanth Shastry "LENS: Leveraging Social Networking and trust to Prevent Spam Transmission", 2014.

[9] Michael sirinivasan,Kyungbaek Kim,Xiaowei Yang"SocialFilter: Introducing Social Trust to Collaborative Spam Mitigation",2011.

[10] Ze Li,Haiying Shen "SOAP: A Social Network Aided Personalized and Effective Spam Filter to Clean Your E-mail Box", 2011.

[11] Haiying Sheh, Ze Li "Leveraging Social Networks for Effective Spam Filtering", November 2014.