International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Spam Detection Techniques: A Review

Reena Sharma[1], Gurjot Kaur[2]

[1, 2] Department of Computer and Science Engineering, Chandigarh University, Gharuan, Punjab, India

Abstract: ***Internet has become an important part of our lives which is used for nearly all purposes. Internet is like an umbrella providing various services at a single time. One of the important uses of this is, sharing of information. E-mails are the most commonly used utility of internet for communication. The emails contain some unsolicited mails called spam which create the problems for users .Detection of spam become a time consuming and lavish activity. Numerous ways have been inspected lately. Spam is in the form of text and images which can harm the system. To broadcast unsolicited facts Spam sender greatly misuses the E-mail. Thus, Spam can be termed as one of the most recurring problems to be tackled by an internet user. Many Techniques have been developed to overwhelm the spam. In this paper different spam detection techniques are discussed, following which a solution has been proposed to prevent this problem.***

Keywords: AdaBoost, Content Spam, Black and White Listing, Link Learning, Spam Filter

1. Introduction

E-mails are fast and inexpensive way of information sharing and communication in today's world. Reading E-mails becomes the regular habit of the peoples. E mail containing unwanted mails irritates the user and occupies the half of the bandwidth of the inbox. These mails are identified as spam. The problems of spam mails are a grim issue. E-mail spam refers to sending disparate, incorrect and spontaneous email messages to numerous users. [7]. the purpose of these mails is publicity, advancement and scattering backdoors or malicious programs. The time spends by people in reading and deleting the spam mail is waste. A spam mail cannot only annoying but also dangerous to recipients. Clicking on links contained in spam emails may send user to phishing and malware [1]. According to Symantec, globally 75.9% of email messages are spam [5].

Spammers are generally technically skilled persons that are hired by companies for sending spam. A third party is hired to prevent any legal action on the company itself. Spamming activity can cost attractively to a company, if done right. For example a company is selling wonky dolls for 50 dollars a doll. If company lets the spammer send out 10 million mails where the response rate turns out near to 0.1%, it may return around half of one millions dollars. Spammers get E-mail address by foraging them from internet, news etc.

A major problem in detecting spam stems from active adversarial efforts to frustrate classification. Spam sender

Uses a multitude of techniques based on knowledge of current algorithm, to evade detection. Dealing with spam incurs high cost for enterprises, investing efforts to try to reduce spam –related costs by installing spam filters [9].

The basic concept of spam filter can be illustrated in the given diagram [14]:

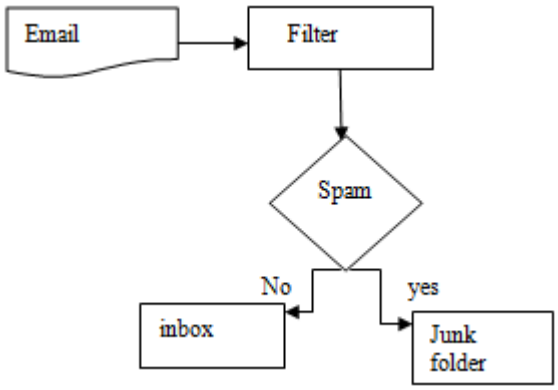

Figure 1: Spam filter

Spam E-mail characterized by three main features:

Anonymity [1]**:** The address and identity of the sender are concealed.

Mass Mailing [1]**:** The email is sent to large groups of peoples.

Unsolicited**:** The email is not requested by recipients.

According to studies undertaken by M86 security , the ascending trend that European IT has lately followed(as compared to North America) involves not only positive aspects; with it came less pleasant phenomena like spam[7]. Spam is not only limited to emails. It prevails in text messages services, newsgroups**,** and social networking sites and also in web search engines. Spams are detected by machine learning and non –machine learning algorithms

2. Related Work

We have read the various papers. The techniques are described below which are discussed in these papers.

Guang Gang Geng et al [6] in their work they proposed a link based semi supervised learning algorithms to boost the performance of a classifier. This classifier integrates the traditional self-training with topological dependency based link learning. The Experiments with a dataset WEBSPAM –

Uk20006 (http://chato.cl/webspam/datasets/uk2006/) benchmark showed that the algorithms are effective.

Link Based Learning Algorithm

The algorithm is based on the self- training and link learning. These two involve the complete use of classifier's self-learning capability and the topological dependency on the web graph.

The classifier is first trained with the small labeled data set in link learning. The trained classifier is used to classify the unlabeled data and then give a predicted spamicity value. In link learning step, the link spamicity of unlabeled data will be calculated according to their neighbor's. Largest samples and smallest samples are converted into labeled ones with their predicted tables which are based upon the value of link spamicity

$Ps(x) = p_{spam}(x)/p_{spam}(x) + p_{normal}(x)$……(1)

$Ls(h) = \Sigma_{v\in Nh}(ps(v))\times weight(h,v))/\Sigma_{v\in N(h)}\times weight(h,v)$..(2)

Where v, h are the hosts, weight(h ,v) is the weight of host h ,v ,weight(h,v)€ {1,n, log(n)}, where n is number of hyperlinks between h and v between h and v. N(h)€ in link (h) or outlink(h). Inlink (h) represent the link set of h, and outlink (h) is the outline set of h.

Link Training Steps

1. Weight initialization
2. While i< no of iterations
3. Train labeled training set
4. Detect unlabeled set and compute their spam values
5. Annotate host level graph with ps value
6. Compute link learning process
7. Select largest spam samples in each iteration according to Ls value.
8. End while
9. Train the classifier on labeled train set
10. Test the samples with trained classifier.

Faraz Ahmed et al [5]. Markov clustering based approach for the detection of spam profiles on OSN's is presented in this paper. Work is based on a real dataset of Facebook profiles, which include both benign and spam profiles. A set of three features is identified and used for to model social interaction of OSN user using a weighted graph. MCL is applied to exploit the behaviour similarity of profiles and unearth the cluster present in profile data set.

Markov Clustering

This Technique does not require no of cluster to be supplied by the user like K-means algorithm. Data is collect from the wall post, pages, and tags. Then the feature identification has done on the basis of the weights of the wall post, pages and Tags. The weights are assigned with the help of a weighted graph. Markov clustering is used for clustering and F-measure is calculated from the cluster of the data set.

R.Malarvizhi et al.in [14] an overview for spam filtering, and the ways of evaluation and comparison of different filtering methods is present in the paper**.** Fisher Robinson Inverse chi square, Ad boosts classifier, Bayesian classifiers are discussed. Bayessian method is used to create the spam filter in this paper.

AdaBoost Classifier

A machine learning algorithm proposed by Freund and Robert Schapiro**.** The Meta algorithm which can be used in aggregation with other learning algorithms to improve the performance of this algorithm**.** Confidence based label sampling is used by the AdaBoost classifier which works with the concept of active learning. Classifier is trained by the variance and obtains a scoring function which can be used to classify the mail as spam or non- spam [14**]**. Classifier is trained with help of labelled data. This indicates the data has originally been classified as spam or ham [14]. Required functions which classify the message as spam are generated by the trained classifier. By using this algorithm training process is improved. This uses a classifier recursively in a series of rounds n=1…,N, for each call a distribution of weights D(n) is updated that indicates the importance of each record in the data corpus for the classification[14].

1. Weight distribution initialization.
2. For i = 1: no of rounds
   Train a learner using distribution.
3. Calculate error.
4. if error>1.2 then break
5. Determine the current weight
6. $D_{t+1}(i)= Dt/Zt$
7. Weight distribution updating.
   End

Loredana Firte et al [10] present a new approach for spam detection filter. The solution developed is an offline application that uses the K-Nearest Neighbour algorithm and pre-classified email data set for the learning process. Messages are classified with the KNN algorithm based on a set of features from the email properties sand content.

Mails are selected from the inbox. These selected mails are analysed and the feature extraction is don**e**. The features that are extracted are Number of recipients, No of replies, message size, and number of attachments. Then resampling of data is done. With the help of KNN algorithm classification of data has done. F-measure is computed and the result is obtained.

M. Basavaraju et.al [11] proposed a new spam detection technique using the text clustering based on vector space model is proposed in the research paper. With help of this method we can extract and detect spam/non spam email. The suggested method contains the space between all of the attributes of an email

Words are considered as the features. Words represented with discrete values based on statistics of the occurrence or

the non-appearance of words**.** Classification of data is done with the help of **BIRCH** (Balanced Iterative Reducing and Clustering using Hierarchies) and K-NNC is used. Porter's algorithm is used for the pre-processing of the data. Then vocabulary is built with the help of vector space model and then clustering of data has done. Precision, Accuracy is calculated at the end.

Siddu.Pacingill. Algur et.al, in [17] proposed a system in which link and content spam detection are used to detect the web pages as spam. System also classifies the web page as spam based on threshold set by statistical method**.**

Unsupervised Web spam Detection is Studies. The two spam detection modules, the link spam detection module and content spam detection are considered module. The modules gather related facts of the target URL. Web page's content statistics are extracting by the content spam dector. It has four sub modules, web page, and loader, and parser, stop word remover and content spamicity finder. The link structure of web pages is extracted by link spam detector. It also has four sub modules web page loader, link extractor, PageRank and link spamicity finder. The third sub module of both modules return link and content spamicity total of web pages and then it is compared with threshold value which is computed by statistical measure.

Mohammed Mikki et al [1] .In their research they proposed an improved filtering technique**.** Technique is based on the improved digest algorithm and DBSCAN clustering algorithm.

Digest of e-mails has taken and then clustering is performed. DBSCAN (Density Based spatial Clustering of Application with Noise) algorithm is used for clustering purpose. Two parameters ε and minimum number of points required to form a cluster are used by the DBSCAN.

Vandana et al [9].proposed an image spam detection system is introduced. Hidden markov model was used to detect all the spam images

Spam detection system has designed which detect the spam on the basis of content. The files which are used in spam detection are text files and excel files. Feature extraction of content is done by removing the stemming and stopping words.

Saadat Nazirova [15] survey on spam detection techniques is performed. The techniques discussed in this paper are Image Based Spam filtering, Method based on acceptance of sender as spammer, image based spam filtering etc.

3. Proposed Work

A major requirement is to protect the user from the spam mails. Various clustering algorithms are used to detect the spam. The previous methods have few limitations of having less accuracy or precision. This problem would be solved by using the RBFN algorithm. The result obtained by using this algorithm may be compared with the SVM.

4. Conclusion

During the survey of various spams detection techniques it has been widely observed that there are numerous spam detection techniques available around us. These technique either lack in performance or level of accuracy. The above proposed methodology may be adopted to enhance the precision quotient of the existing spam detection.

References

[1] Alaa H. Ahmed and Mohammed Mikki, "Improved Spam Detection using DBSCAN and Advanced Digest Algorithm", International Journal of Computer Applications, Vol. 6 May 2013.
[2] Ann Nossier, khaled kagati, and Islam Taj-Eddin, "Intelligent Word- Based Spam Filter Detection Using Multi Neural Networks", International Journal of Computer Science Issues, Vol 10, Issues 2, No 1, March 2013 ISSN(Print): 1694 -0814|ISSN(Online): 1664-0784
[3] Asmeeta Mali, "Spam Detection Using Bayesian with Pattern Discovery", International Journal of Recent Technology and Engineering, ISSN: 2277-3878, Vol.2, Issue-3, July 2013.
[4] Erik D. Demaine, Fried helm Meyer auf der Heide, U. Paderborn, Rasmus Pagh, Mihai P˘atra¸scu, "On Dynamic Dictionaries Using Little Space", ARVIX, 2005.
[5] Faraz Ahmed and Muhammad Abulaish, "An MCL Based Approach for Spam Profile Detection in Social Network", IEEE 11th International Conference on Trust, Security and Privacy.
[6] Guang-Gang Geng, Qiu-Dan Li and Xin-Chang Zhang, "Link Based Small Sample Learning for Web Spam Detection", ACM, April 2009
[7] Junod .J, "Serves to Spam: Drop Dead", Computer and Security Elsevier, Vol.16, 1997
[8] J. Klensin, "RFC2821: Simple Mail Transfer Protocol", April 2001
[9] J. Vandana and Nidhi Sood, "Spam Detection System Using Hidden Markov Model", International Journal of Advanced Research in Computer Science and Software Engineering, Volume-3,Issue-7, July-2013
[10] Loredana Firte, Camelia Lemnaru and Rodica Potolea, "Spam Detection Filter Using KNN Algorithm and Resampling", IEEE Conference, 2010
[11] M.Basavaraju and Dr.R.Prahakar, "A Novel Method of Spam Mail Detection using Text Based Clustering Approach", International journal of Computer Applications, Volume-5, August 2010
[12] M.Soranamageswari and C.Meena, "Statistical Feature Extraction for Classification of Image Spam Using Artificial Neural Network", Second International Conference on Machine Learning and Computing, February 2010
[13] P.A.Chitira, J.Diederich and W.Nejdl, "MailRank: using Ranking for Spam Detection", ACM 14th International Conference on Information and Knowledge Management, October2005

[14] R.Malarvizhi and K.Saraswathi, "Content – Based Spam Filtering and Detection Algorithms-An Efficient Analysis and Comparison", International Journal of Engineering Trends and Technology, Vol.4,Issue 9,Septmber 2013
[15] Saadat Nazirova, "Survey on Spam Filtering Techniques", Communication and Network, August 2011
[16] Sender Policy Framework
[17] Siddu p.Algur and Neha tarannumPendari, "Hybrid Spamicity Approach to web Spam Detection", IEEE Conference on Pattern Recognition, Informatics and Medical Engineering, March 2012
[18] T.-J. Liu, W.-L. Tsao and C.-L. Lee, "A High Performance Image-Spam Filtering System", 19th International Symposium on distributed Computing and Application to Business, Engineering and Science, August 2010