A Review on Different Spam Detection Approaches

Rekha [1], Sandeep Negi [2]

[1] *Mtech student, Department of CSE,*
Delhi Institute of Technology,
Gannaur, Sonepat
[2] *Assistant Professor, Department of CSE,*
Delhi Institute of Technology,
Gannaur, Sonepat

***Abstract*— Email is one of the crucial aspects of web data communication. The increasing use of email has led to a lucrative business opportunity called spamming. A spam is an unwanted data that a web user receives in the form of email or messages. This spamming is actually done by sending unsolicited bulk messages to indiscriminate set of recipients for advertising purpose. These spams messages not only increases the network communication and memory space but can also be used for some attack. This attack can be used to destroy user's information or reveal his identity or data. In this paper we discuss some approaches for spam detection.**

Keywords: - Spam Emails, Non-Spam Emails, Filters, Approaches.

I INTRODUCTION

In recent years, internet has become an integral part of our life. With increased use of internet, numbers of email users are increasing day by day. It is estimated that 294 billion emails are sent every day. This increasing use of email has created problems caused by unsolicited bulk email messages commonly referred to as Spam [1]. It is assumed that around 90% of emails sent everyday are spam or viruses.

Email has now become one of the best ways for advertisements due to which spam emails are generated. Spam emails are the emails that the receiver does not wish to receive. A large number of identical message are sent to several recipients of email. Increasing volume of such spam emails is causing serious problems for internet users, Internet Service Providers, and the whole Internet backbone network. One of the examples of this may be denial of service where spammers send a huge traffic to an email server thus delaying legitimate message to reach intended recipients. Spam emails not only waste resources such as bandwidth, storage and computation power, but may contain fraudulent schemes, bogus offers and scheme. Apart from this, the time and energy of email receivers is wasted who must search for legitimate emails among the spam and take action to dispose the spam. Dealing with spam and classifying it is a very difficult task. Moreover a single model cannot tackle the problem since new spams are constantly evolving and these spams are often actively tailored so that they are not detected adding further impediment to accurate detection.

A spam filter is a program that is used to detect unsolicited and unwanted email and prevent those messages from getting to a user's inbox. Like other types of filtering programs, a spam filter looks for certain criteria on which it bases judgments. For example, the simplest and earliest versions (such as the one available with Microsoft's Hotmail) can be set to watch for particular words in the subject line of messages and to exclude these from the user's inbox. This method is not especially effective; it may omit legitimate messages (called false positives) and passing actual spam message s.More sophisticated programssuch as Bayesian filters or other heuristic filters, attempt to identify spam through suspicious word patterns or word frequency.

Filter classification strategies can separated into two categories: those based on machine learning (ML) principles and those not based on ML (Fig 1). ML approaches are capable of extracting knowledge from a set of messages supplied, and using the obtained information in the classification of newly received messages.

Non-machine learning techniques, such as heuristics, blacklisting and signatures, have been complemented in recent years with new, ML-based technologies. In the last few years, substantial academic research has taken place to evaluate new ML-based approaches to filtering spam. ML filtering techniques can be further categorized into complete and complementary solutions. Complementary solutions are designed to work as a component of a larger filtering system, offering support to the primary filter (whether it be ML or non-ML based). Complete solutions aim to construct a comprehensive knowledge base that allows them to classify all incoming messages independently.

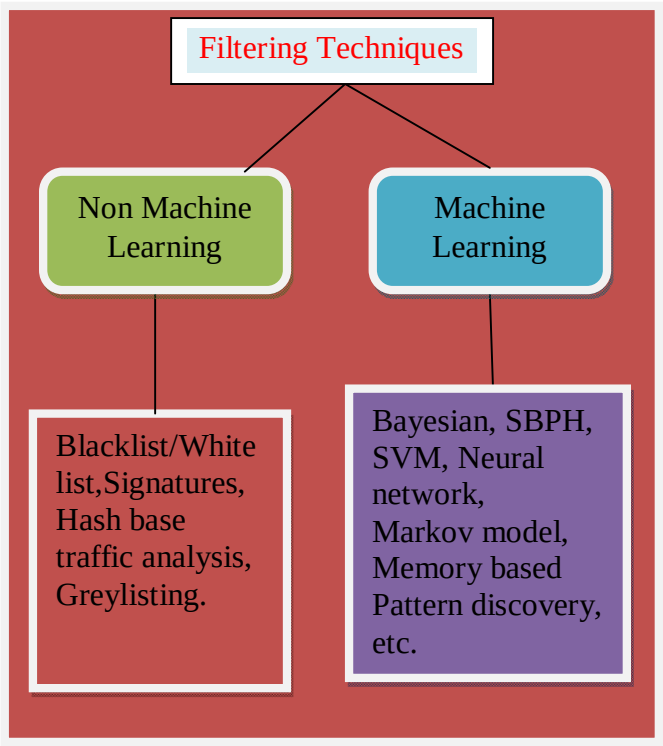

Fig 1. Classification of the various approaches to spam filtering

II LITERATURE SURVEY

Geerthik and Anish performed a work,**"** Filtering Spam: Current Trends and Techniques"**.** They give an overview about latest trend and techniques in spam filtering. They analyzed the problems which is introduced by spam ,what spam actually do and how to measure the spam .This article mainly focuses on automated, non-interactive filters, with a broad review ranging from commercial implementations to ideas confined to current research papers. The solutions using both machine and non –machine learning approaches are reviewed and taxonomy of different approaches is presented. While a range of different techniques have and continue to be evaluated in academic research, heuristic and Bayesian filtering, along with its variants provide the greatest potential for future spam prevention. [1].M.Basavaraju and Dr. R. Prabhakar performed a work," A Novel Method of Spam Mail Detection using Text Based Clustering Approach". A new spam detection technique using the text clustering based on vector space model is proposed in this research paper. By using this method, one can extract spam/non-spam email and detect the spam email efficiently. Representation of data is done using a vector space model. Clustering is the technique used for data reduction. It divides the data into groups based on pattern similarities such that each group is abstracted by one or more representatives.[2].Ann Nosseir , Khaled Nagati and Islam Taj-Eddin performed a work," Intelligent Word-Based Spam Filter Detection Using Multi-Neural Networks". They proposed an approach which is character-based technique. This approach uses a multi-neural networks classifier. Each neural network is trained based on a normalized weight obtained from the ASCII value of the word characters. Results of the experiment show high false positive and low true negative percentages. [3]. R. Kishore Kumar, G. Poonkuzhali, P. Sudhakar provides the analysis of email spam classifier through data mining techniques. In their work," Comparative Study on Email Spam Classifier using Data Mining Techniques " spam dataset is analyzed using TANAGRA data mining tool to explore the efficient classifier for email spam classification. Initially, feature construction and feature selection is done to extract the relevant features. Then various classification algorithms are applied over this dataset and cross validation is done for each of these classifiers. Finally, best classifier for email spam is identified based on the error rate, precision and recall. [4].

Rafiqul Islam and Yang Xiang performed classification of user emails form penetration of spam. In their paper," Email Classification Using Data Reduction Method" an effective and efficient email classification technique based on data filtering method is presented. They have introduced an innovative filtering technique using instance selection method (ISM) to reduce the pointless data instances from training model and then classify the test data. The objective of ISM is to identify which instances (examples, patterns) in email corpora should be selected as representatives of the entire dataset, without significant loss of information. They have used WEKA interface in our integrated classification model and tested diverse classification algorithms. Their empirical studies show significant performance in terms of classification accuracy with reduction of false positive instances. [5].
Asmeeta mali performed a work," Spam Detection using Bayesian with Pattern Discovery". In her paper she presents an effective technique to improve the effectiveness of using and updating discovered patterns for finding relevant and interesting information. Using Bayesian filtering algorithm and effective pattern Discovery technique we can detect the spam mails from the email dataset with good correctness of term. [6].
Vandana Jaswal proposes an image spam detection system that uses detect spam words. In her work," Spam Detection System Using Hidden Markov Model" filtering method are used to detect stemming words of spam images and then use Hidden Markov Model of spam filters to detect all the spam images. [7].

In year 2011, Saadat Nazirova performed a work," Survey on Spam Filtering Techniques". In this paper the overview of existing e-mail spam filtering methods is given. The classification, evaluation, and comparison of

traditional and learning-based methods are provided. Some personal anti-spam products are tested and compared. The statement for new approach in spam filtering technique is considered. [8].
As we are working on the approach that gives better result than other approaches to identify spam mail we need danger theory and dendritic cell algorithm. Here some other work on DCA is defined in literature.

Neha Singh performed a work,"Dendritic Cell algorithm and Dempster Belief Theory Using Improved Intrusion Detection System". To minimize false alarm rate she proposed novel dual detection of IDS based on Artificial Immune System that integrating the Dendrite Cell Algorithm and Dempster Belief theory in her work. [9].
In year 2007 Greensmith submitted his work,"The Dendritic Cell Algorithm". This is a novel immune-inspired algorithm based on the function of the dendritic cells of the human immune system. In nature, dendritic cells function as natural anomaly detection agents, instructing the immune system to respond if stress or damage is detected. Dendritic cells are a crucial cell in the detection and combination of 'signals' which provide the immune system with a sense of context. The dendritic Cell Algorithm is based on an abstract model of dendritic cell behaviour, with the abstraction process performed in close collaboration with immunologists.This algorithm consists of components based on the key properties of dendritic cell behaviour, which involves data fusion and correlation components. [10].

III SPAM DETECTION APPROACHES

There are several approaches to identify incoming messages as spams are, Whitelist/Blacklist, Bayesian analysis, Mail header analysis, Keyword checking etc. some of them are defined below:

- **Whitelist/Blacklist: - These** approaches simply create a list. A whitelist is a list which includes the email addresses or entire domains which the user knows. An automatic white list management tool is also used by user that helps in automatically adding known addresses to the whitelist. A blacklist is the opposite of whitelist. In this list we add addresses that are harmful for users.

- **Mail Header Checking: -** This approach is very known approach. In this we simply consist of set of rules that we match with mail headers. If a mail header matches, then it triggers the server and return mails that have empty "From" field, that have too many digits in address that have different addresses in "To" field from same source etc.

- **Signatures: -** This approach is based on generating a signature having unique hash value for each spam message. The filters compare the value of previous stored values with incoming emails values. It is probably impossible for legitimate message having same value with spam message value stored earlier.

- **Bayesian Classifier: - There** are particular words used in spam emails and non spam emails. These words have particular probability of occurring in both emails. The filters that we used don't know these probabilities in advance; we must train them first so it can build them up. After training the word probabilities are used to compute the probability that an email having particular set of words in it belong to either spam or legitimate emails. Each particular word or only the most interesting words contribute to email's spam probability. This contribution is known as the posterior probability and is computed using 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.

TABLE I
COMPARISON OF DIFFERENT SPAM DETECTION APPROACHES.

Approach	Advantage	Disadvantage
Whitelist/Blacklist	Simplistic in nature	Easily penetrated by spammer
Signatures	Low level of false positives	Unable to identify spam until email reported as spam & its hash distributed.
Mail Header Checking	Easily implemented	High false positive rate and rejecting connections require additional information/policies.
Bayesian Analysis	State-of-the-art approach (wide - spread implementation).	Rely on 'naive' Bayesian filtering (which assumes events occurred independent each other).

IV CONCLUSION

Spam emails are the biggest problem for the web data. This paper explored different approaches to deal with this problem. From all of these approaches no one can provide 100% result. Some of the approaches provide high false positive rates and false negative rates. There is very much scope for identifying mail as spam emails or legitimate mails for text as well as multimedia messages.

REFERENCES

[1] Geerthik. S and Anish .T.P, "Filtering Spam: Current Trends and Techniques", International Journal of Mechatronics, Electrical and Computer Technology Vol. 3(8), Jul, 2013, pp 208-223, ISSN: 2305-0543 © Austrian E-Journals of Universal Scientific Organization.

[2] M. Basavaraju, "A Novel Method of Spam Mail Detection using Text Based Clustering Approach", International Journal of Computer Applications (0975 – 8887) Volume 5– No.4, August 2010.

[3] Ann Nosseir , Khaled Nagati and Islam Taj-Eddin, "Intelligent Word-Based Spam Filter Detection Using Multi-Neural Networks", IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 2, No 1, March 2013 ISSN (Print): 1694-0814 | ISSN (Online): 1694-0784.

[4] R. Kishore Kumar, G. Poonkuzhali, P. Sudhakar," Comparative Study on Email Spam Classifier using Data Mining Techniques", Proceedings of the International MultiConference of Engineers and Computer Scientists 2012 Vol I, IMEC2012, March 14-16,2012, Hong Kong, ISBN: 977-988-19251-1-4.

[5] Rafiqul Islam and Yang Xiang, member IEEE, "Email Classification Using Data Reduction Method" created June 16, 2010.

[6] Asmeeta Mali, "Spam Detection Using Baysian with Pattren Discovery", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-2, Issue-3, July 2013.

[7] Vandana Jaswal, " Spam Detection System Using Hidden Markov Model", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 7, July 2013 ISSN: 2277 128X.

[8] 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).

[9] Neha Singh," Dendritic Cell Algorithm and Dempster Belief Theory Using Improved Intrusion Detection System "International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 7, July 2013 ISSN: 2277 128X.

[10] Julie Greensmith, "The Dendritic Cell Algorithm", Thesis submitted to the University of Nottingham for the degree of Doctor of Philosophy October 2007.

[11] Christina V, Karpagavalli S, Suganya G, "A Study on Email Spam Filtering Techniques", International Journal of Computer Applications (0975 – 8887) Volume 12– No.1, December 2010.