A Survey Paper on Spam Mail Detection Using RFD

Sayarabanu B. Nadaf[1]
Computer Engineering
TSSM's Bhivarabai Sawant College of Engineering and Research
Pune, India

Prof. Anil D. Gujar[2]
Computer Department
TSSM's Bhivarabai Sawant College of Engineering and Research
Pune, India

Abstract: Electronic mails are playing a vital role in daily communications across various domains like business, organizations, academics, political and social. Spam mails are the emails which contains the malwares. These will lead to insecurity of the data. Spam mails are significantly spreading all over the mail systems. They may lead to the financial loss and cause the inconvenience to the recipients. To avoid these issues the spam mails should be detected efficiently. Sometimes the spam filters are not able to capture the spam mail accurately as these filters capture data from only subject line or contents of the mail. Relevance Feature discovery is a novel pattern mining approach. This approach mines the positive, negative and general patterns. This approach includes the processes like NLP Pre-processing, Sequential pattern detection, assigning weights to patterns and then saving it to data sets. To filter the emails which are the spam emails efficiently a new approach which is based on an innovative Relevance Feature Discovery model. It will scan through all mails and it will categorize patterns in to positive, negative or general .Then it will analyse whether they are spam mails or not depending on the type of patterns and process accordingly. It will also synchronize with the Email server and manage emails for users on their system.

Key words: spam mails, spam filters, Relevance feature discovery, pattern mining, NLP.

I. INTRODUCTION

As the data is drastically increasing day by day data mining has gained the most importance in today's world. A pattern which will be more accurate to the user preference needs to be discovered for classifying the text documents. The most popular text mining and classification methods uses the term based approaches for describing the preferences of user .But these methods are not so much useful as they have issues of polysemy and synonymy .To overcome these issues Relevance Feature Discovery model was proposed as an innovative solution. It basically identifies the patterns in the text documents. These patterns are divided into three types of classes namely positive patterns, negative patterns and general patterns. It also classifies the terms and updates the term weights with respect to the specificity in patterns and their distribution in patterns. Specificity of terms can be defined as the exact words which will describe about the topic in which the user is interested. It may change as per user perspective. Spam emails are steadily growing from 1990's.These are the mails which may consists the untrusted links which initially gives impression to users that they are familiar but actually they lead to phishing web sites which have the malware. Spam mails also includes the malware scripts and harmful executable files. Spam mails are causing big financial losses all over the world. There are spam filters which are currently used by the various mail systems. According to a Cyberoam report published in 2014, there are averages of 54 billion spam messages sent every day across the world.

II. LITERATURE SURVEY

This section mainly gives the different work performed on the text mining and spam mail detection.[1] Feldman,Moshe Fresko,Yakkov Kinar proposed that text mining [1]is done at term level. The mining process starts by preprocessing the document collection and extractting the terms from the documents. Each document is represented as a set of terms and

annotations characterizing the document. This method gives the number of occurrences of the terms but it cannot handle the large texts.Ning Zhong, Yuefeng Li, and Sheng-Tang Wu have proposed an innovative and effective pattern discovery technique which includes the processes of pattern deploying and pattern evolving, to improve the electiveness of using and updating discovered patterns for finding relevant and interesting information with effective patterns as per the users requirements. It works better than term based approach.Yuefeng Li, Abdul Mohsen Algarni, Ning Zhong proposed a technique discovers both positive and negative patterns in text documents .It consider the terms as well as patterns feaures.It Works on long texts. It handles the terms and patterns both. M.Basavaraj and Dr. R. Prabhakar proposed," 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 by them. By using this method, we can extract spam/non-spam email and detect the spam email effectively. Data is represented using a vector space model. Clustering is the technique used for data reduction. It partitions the data into various groups based on pattern similarities.[4]Ann Nosseir , Khaled Nagati and Islam Taj-Eddin suggested a concept," Intelligent Word-Based Spam Filter Detection Using Multi-Neural Networks". They suggested an approach which is character-based technique. It uses a multi-neural networks classifier. Each neural network is trained based on a normalized weight. [5] R. Kishore Kumar, G.Poonkuzhali, P. Sudhakar provides the analysis of email spam classifier through data mining methods. Spam dataset is analyzed using TANAGRA data mining tool to the efficient classification of spam email.Rafiqul Islam and Yang Xiang worked on classification of user emails form penetration of spam. 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. Asmeeta Mali [7] performed a work," Spam Detection using Bayesian with Pattern Discovery". She presented an effective technique to improve the effectiveness of using and updating discovered patterns for finding relevant and interesting information. Vandana Jaswal [8] works on 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.

III. CONCLUSION

Efficient pattern detection in spam mail filtering plays crucial role .RFD gives the spam, non spam and general patterns which easily identifies mail is spam or non-spam. The current methods which uses the pattern detection method does not include the general patterns.RFD gives the general patterns of which user can decide to determine whether he wants to put the mail as spam or non-spam to avoid the loss

ACKNOWLEDGEMENT

References

1. Feldman, Moshe Fresko,Yakkov Kinar ,"Text Mining at the Term Level"

2.   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.

3. 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.

4. 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.

5. 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.

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

7. Asmeeta Mali, "Spam Detection Using Bayesian with Pattern Discovery", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-2, Issue-3, July 2013.

Sarayabanu et al.,

International Journal of Advance Research in Computer Science and Management Studies
Volume 4, Issue 1, January 2016 pg. 46-48

8. 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.

9. Saadat Nazirova, "Survey on Spam Filtering Techniques", Communications and Network, 2011, 3, 153 160,doi:10.42 36/cn.2011.33019 Published Online August 2011 (http://www.SciRP.org/journal/cn).

AUTHOR(S) PROFILE

Sayarabanu Nadaf, is pursuing the M.E degree in Computer Engineering from BSCOER .She is doing her post-graduation from Pune University.She received her B.E. degree from Walchand Institute Of Technology,Solapur from Shivaji University in 2005.Her area of interest are Data mining.

Anil Gujar, received the M.Tech (IT) degree from the Department of Computer Engineering He is currently working as Asst. Professor with Department of Computer Engineering, Bhivarabai Sawant College of Engineering and Research, Pune, MAH, India.