International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391

Spam Mail Detection Using Relevance Feature Discovery

Sayarabanu B. Nadaf, Anil D. Gujar

[1, 2]Computer Engineering Department, TSSM's Bhivarabai Sawant College of Engineering and Research, Pune

Abstract: ***Electronic mails are used very widely to share the information quickly. Various domains like business, organizations, academics, political and social uses Electronic mails as they are playing a crucial role in daily communications. The mails which cause the insecurity to the data are called as Spam mails. Spam mails are rapidly spreading all over the mail systems. They may lead to the financial loss and cause the inconvenience to the recipients. To handle these issues the spam mails should be detected efficiently and an appropriate action should be taken on them. Sometimes the spam filters are not able to capture the spam mails accurately as these filters capture data from only particular part of the mail. Relevance Feature discovery is an effective and latest approach for pattern mining. This carries out the mining of the positive, negative and general patterns. This approach includes the various processes like Text 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 is proposed. It is based on an innovative Relevance Feature Discovery model. It will scan through the contents of all mails and it will separate patterns in to positive, negative or general categories. Then it will analyze 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. It will detect the attached image to segregate it into spam or ham image.***

Keywords: spam mails, spam filters, Relevance feature discovery, pattern mining

1. Introduction

Data is rapidly growing now days which has given utmost importance to data mining in today's world. A pattern gives the user preference exactly which needs to be discovered for classifying the text documents. Term based approach is used by many text mining methods. But these methods are not so much useful as they have polysemy and synonymy drawbacks. To avoid these drawbacks Relevance Feature Discovery model was proposed as an innovative solution. It recognizes the patterns in the text documents. These patterns are categorized into three types of classes namely positive patterns, negative patterns and general patterns. Positive patterns are patterns of legitimate mails. The spam mails patterns are referred as Negative patterns. General patterns are the patterns which occurs both in positive and negative patterns. It also performs the term classification and term weights are updated with respect to the specificity in patterns and their distribution. Now days the spam mails are not only come with content in mails but also the spam information is hidden in the image attachment. So these images should also be filtered in to spam and ham mails. Generally the image attachment contains the text in the image format. Section 2 gives the related work done and section 3 explains the basic terminologies used in the paper. Section 4 gives the proposed model .Section 5 describes the conclusion and future scope of the paper.

2. Literature Survey

This section gives the different work performed on the text mining and spam mail detection. Feldman, Moshe Fresko,Yakkov Kinar proposed that text mining [2]is done at term level. Preprocessing the document collection and extracting the terms from the documents covers the text mining process. A set of terms and annotations specifying the document represents each document. This method identifies the number of occurrences of the terms but it can not deal with the large texts.Ning Zhong, Yuefeng Li, and Sheng-Tang Wu have come up with an innovative and effective pattern discovery technique [3] which consists of the processes of deploying of patterns to improve the usage electiveness and updating of discovered patterns which will find the interesting information with respect to the user and which is relevant information. It gives better performance than term based approach.Yuefeng Li, Abdul Mohsen Algarni, Ning Zhong discovered a technique for both positive and negative patterns in text documents .It considers the terms as well as patterns features. It works on long texts. It handles the terms and patterns both.M.Basavaraj and Dr.R. Prabhakar [9] suggested a new spam detection technique using vector space based text clustering is suggested by them. By using this method, we can process spam/non-spam email and detect the spam email effectively. In this method the data is represented using a vector space model. For data reduction clustering technique is used. It classifies the data into various categories based on similarities in patterns. [4]Ann Nosseir,Khaled Nagati and Islam Taj-Eddin proposed a concept which is based on Multi-Neural Networks. They come with an approach which is character-based technique. Multi-neural networks classifier is used by them. Based on a normalized weight each neural network is trained. [5] R. Kishore Kumar, G.Poonkuzhali, P. Sudhakar proposed that data mining methods are used for the analysis of email spam classifier. Spam dataset is analyzed with TANAGRA data mining tool for the efficient spam email classification. Rafiqul Islam and Yang Xiang worked [6] on classification of user emails form. Email classification which uses Data Reduction Method is a most effective email classification technique which is based on data filtering method. Instance selection method (ISM) is used to decrease the pointless data instances from training model and then the classification of test data is done. Asmeeta Mali [7] proposed that Pattern Discovery can be used for spam mail detection work. She

presented a unique technique which will increase the effectiveness of updating patterns which are discovered to specify the relevant and information in which users are interested.Vandana Jaswal [8] worked on spam detection on image which detects words which are spam. She used Hidden Markov Model as a filtering method which detects stemming words in images which are spams and then Hidden Markov Model is used. [1]As suggested by Mubarak Albathan ,Yuefeng Li, Yan Shen, Abdulmohsen Algarni and Moch Arif Bijaksana the text mining is done using the features which are relevant as well as irrelevant.As suggested by Zin Mar Win and Nyein Aye [11] the spam images can be detected by using Hough transform method efficiently.

3. Basic Terminologies used for Data Mining

3.1 Frequent and Closed patterns:

Let T1 =\{t1,t2,...,tm\} be the set of terms(words)which are drawn from mails and Coverset(X) is defined as the covering set of X in which the terms in X are defined. A pattern X (also a term set) is called closed if and only if
X =term set (cover set(X)).
Let X be a closed pattern. We have sup (X1) < sup (X) for all patterns X1 is subset of X.

3.2 Sequential Patterns:

A sequential pattern S =< t_1,……, tr > is an ordered list of items.

3.3 Specificity Function

A term's specificity is defined according to its appearance in a given training set. A term's relative specificity describes the extent to which the term focuses on the topic that users want. The term's specificity depends on users' perspectives hence it is difficult to derive it.

3.4 Spam Detection

We assume that our system consists of input set I= {data set, mails} where the data set is collection of training data D= {D1, D2,…Dm}and the mails are the mails which are received by the recipients M={M1,M2,….Mn}

4. Proposed Model

4.1 System Overview

A new email spamming filter is proposed which is built on pattern mining technique. It effectively determines relevant as well as irrelevant patterns from training data. It uses both term based and pattern based technique's advantages together. An evaluation performance is increased by using classifier based technique to determine relevant and irrelevant patterns from training data and a classifier to detect as mail as spam or not. We will also detect the spam mails which consist of images as attachments. Spammers are constantly challenging the anti-spam technology by creating new methods; image-based spam is the When a picture contains the text messages which are spam and the file type is specified as an image file it is called as an image spam. Due to this the text based spam filters are not capable of detecting and blocking such spam messages. There are various techniques available for detecting image spam (DNSBL, Gray Listing, Spamtraps etc).We propose an innovative image spam detection mechanism which uses file properties, histogram and Hough transform to detect image spam or ham more effectively.

4.2 System Architecture

The proposed model focuses on below phases:
A. Initialization phase
B. Email Application phase

A. Initialization phase
Inputs to the proposed system are the spam and non-spam mail datasets. In this phase below processes will be carried out.

Text Processing:
In this process the all the functions related to text processing like removal of stop words, parts of speech, parsing are carried out.

Sequential Pattern Detection:
In this, the sequential patterns are detected. When the terms are ordered in list it is called as sequential pattern. Term Specificity in patterns using Relevance feature model is defined. Weights are assigned to patterns depending on their occurrences using RFD model. These patterns are saved in database. Performing all above processes we get the training dataset of spam, non-spam and general patterns.

B. Email Application phase

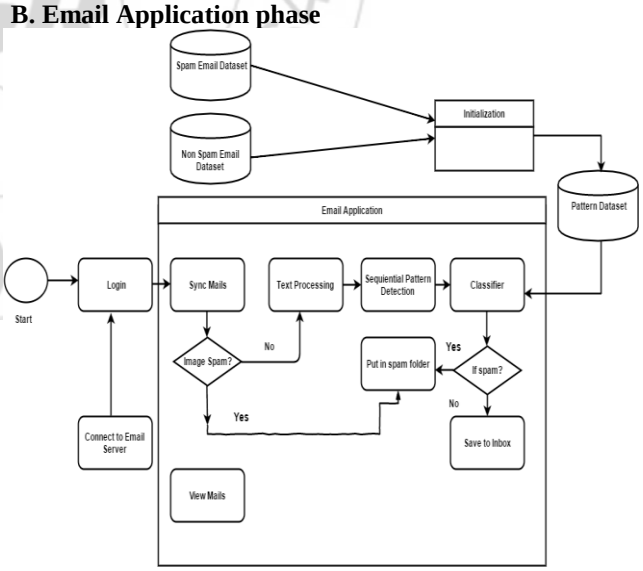

Figure 1: System Architecture

When a new mail arrives in the mail system it goes through text processing, sequential pattern detection to find the patterns in the mails. Classifier classifies these patterns to give spam, non-spam and general mails. Spam mails are placed in spam folder. Non spam mails are sent to Inbox folder. This system detects the spams which are in image forms. Currently to break the spam mail detection methods the images are used as spams. So these spams are also

detected very effectively.

4.3 Algorithm Used

Nave Bayesian classifier is the best classifier where we assume that the mails will come one by one . It calculates the Prior probability of the cases. Then it considers the vicinity which means that new cases which will belong to particular new cases. It calculates the Likelihood of given cases. Then it considers the Posterior probability.

Prior probability of spam mails=Number of spam mails/Total number of mails

Prior probability of non-spam mails=Number of non-spam mails/Total number of mails

Likelihood of both spam and non-spam mails are calculated. Then posterior probability of spam mails and non-spam mails is calculated.

Posterior Probability= Prior probability x Likelihood

4.4 Results and Discussion:

Figure 2: Comparison of spam detection

As shown in the above figure we can say that Bayesian classifier is the best classifier which is used to classify the patterns. It works better than KNN classifier.

5. Conclusion and Future Scope

Efficient pattern detection in spam mail filtering plays crucial role. Using RFD model spam detection gives the spam patterns, non –spam patterns and general patterns which easily identify the whether the mail is spam or ham. The current method 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 of important mails. The images which are in forms of spams are also detected using File Properties, Histogram and Hough Transform. The current proposed system is for English language mails but as future scope we can design the system for multiple languages.

References

[1] Yuefeng Li, Abdulmohsen Algarni, Mubarak Albathan, Yan Shen, and Moch Arif Bijaksana," Relevance Feature Discovery for Text Mining"IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 27, NO. 6, JUNE 2015
[2] Feldman, Moshe Fresko,Yakkov Kinar ,"Text Mining at the Term Level", Ronen Feldman , Moshe Fresko , Yakkov Kinar , Yehuda Lindell , Orly Liphstat , Martin Rajman , Yonatan Schler , Oren Zamir
[3] Y. Li, A. Algarni, and N. Zhong, "Mining positive and negative patterns for relevance feature discovery," in Proc. ACM SIGKDD Knowl. Discovery Data Mining, 2010, pp. 753–762
[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,
[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
[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] 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
[10] Saadat Nazirova, " Survey on Spam Filtering Techniques", Communications and Network, 2011, 3, 153 160,doi:10.42
[11] Zin Mar Win and Nyein Aye,"Detecting Image Spam Based on File Properties, Histogram and Hough Transform", Journal of Advances in Computer Networks, Vol. 2, No. 4, December 2014

Author 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.