ISSN 2395-1621

Email Spam Detection Using Association Rules and Extraction Techniques

[#1]Akash Wagh, [#2]Shipra Arya, [#3]Sonal Patil, [#4]Dhanashri Joshi

[1]akwagh06@gmail.com
[2]shipraarya.bittu@gmail.com
[3]patilsonal0202@gmail.com
[4]jdhanashrim@gmail.com

[#1234]Department of Computer Engineering, Flora Institute of Technology, Savitribai Phule Pune University, Maharashtra, India

ABSTRACT

Spam is one of the major issue especially email spam, of the todays Internet resulting in financial damage to companies and annoying the users and managing such kind of the mailbox has become a crucial one. These unwanted emails clogs the inbox as well as can be used for various attacks which may destroy user's information or to reveal his/her identity or data. In this paper we have presented a new technique for detecting a spam email using data mining techniques like clustering and generating association rules. Where vector space notations are used for representing the emails and the results obtained from the proposed technique's efficiency of detecting spam emails.

Keywords— Association rules, spam detection, Email spam, Text clustering

ARTICLE INFO

Article History

Received :16th April 2016

Received in revised form :

19th April 2016

Accepted : 21st April 2016

Published online :

26th April 2016

I. INTRODUCTION

In todays world of internet, emails has become a faster and effective way of communication in personal as well as in professional communication. It provides a more desirable way for internet user to transfer their data worldwide. But this emails can be wanted as well as unwanted basically called as spam emails. Spam emails are known as unsolicited commercial emails typically unwelcome messages often commercial or political in nature, transmitted via internet as mass mailing to a large number of recipients. Spam mail is subset of an electronic spam which includes nearly identical messages sent to several recipients through email. It may include a malware as script or executable file attachment. Usually there are two types of spam where each one has different consequences. Usenet spam is a similar message sent to 20 or more group of users, it specially aims at users who read newsgroups but rarely do post anything, these spam weaken the ability of system administrator to manage their topic of interest on their systems. The second one is type of spam email which targets the users directly. Spam increases the load on server also weakens the bandwidth of the ISP which in turn adds the cost to handle this load that should be compensated by the users. Moreover the time spent by the users on reading and deleting spam emails is worthless. Therefore, detecting email becomes an important aspect of research for automatically separating emails from the spam ones.Automatic email spam detection contains more difficulties because of the unstructured information containing large number of documents. As the usage increases all of these features will directly or indirectly affect performance in form of quality and speed. Even though various detecting techniques are been developed so far but still the results are not that satisfying. So identifying best spam algorithm itself becomes a tedious job because of features and drawbacks of every algorithm against each other. This paper basic idea is designing model of the email spam detector and also to evaluate performance on it.

II. LITERATURE SURVEY

The email detection offers an important contribution in analysing and detecting spam emails mainly includes

several solutions for detecting and filtering spam on client side. Many machine learning approaches used till now for same purpose. Some of are Bayesian classifiers such as Naïve Bayes, Ripper and also Support vector machine(SVM) and many more among data mining techniques. In many of the techniques, the Bayesian classifier is observed to be used widely and to be good one. Number of techniques makes use of clustering for detecting spam followed by KNN classifier or SVM classifier. So we have tried of using a different approach using clustering followed by generating association rules.

TABLE I. LITERATURE SURVEY

AUTHOR	TECHNIQUES USED	DATA SOURCE
N.S.Kumar et al .2012	Spam email detection technique	Ling-spam corpus
Mugdha More and Bharat Tidke	Provide Home service	Social media online opinion and spam detection and pervasive computing and Ling-spam corpus
Walin Ma et al. 2009	Email spam detection method, Antigen feedback mechanism	TREC07 Corpus
Leman Akoglu et al. 2013	FRAUDEAGLE framwork	App reviews, SoftWare marketplace(SWM) Dataset
Nitin Jindal et al. 2007	Spam detection technique	Manufactured product reviews

III.PROPOSED SYSTEM

It uses vector space model which is an algebraic model to represent text documents as vectors of identifier. The tf-idf weighting scheme is used here, it sets value to the number of times a particular word occurs in the document, the value increases proportionally to the number of time it appears in a document but is offset by the frequency of the word in the corpus. The value is computed as

$$tf*idf(t,d,D)=tf(t,d) \times idf(t,D)$$

Where tf is a term frequency and idf is a inverse document frequency

$$idf=Log(N/dft)$$

Where, N is the number of documents, dft is the number of documents that contain that term. Arbitrarily choose K email documents as initial centroids and assign each email document to the cluster to which it is the most similar, based on the decided similarity measure, Obtain a new centroid for each cluster. Repeat it until no change is seen. For distance-based clustering, the similarity criterion is distance: two or more objects belong to the same cluster if they are close according to a given distance. In our case, the distance measure that we use is the cosine distance between documents. It is the most popular similarity measure applied to text documents. Similarity is determined by using

$$Cos()= (A.B)/A\ B$$

The cosine distance of two documents is defined by the angle between their feature vectors.

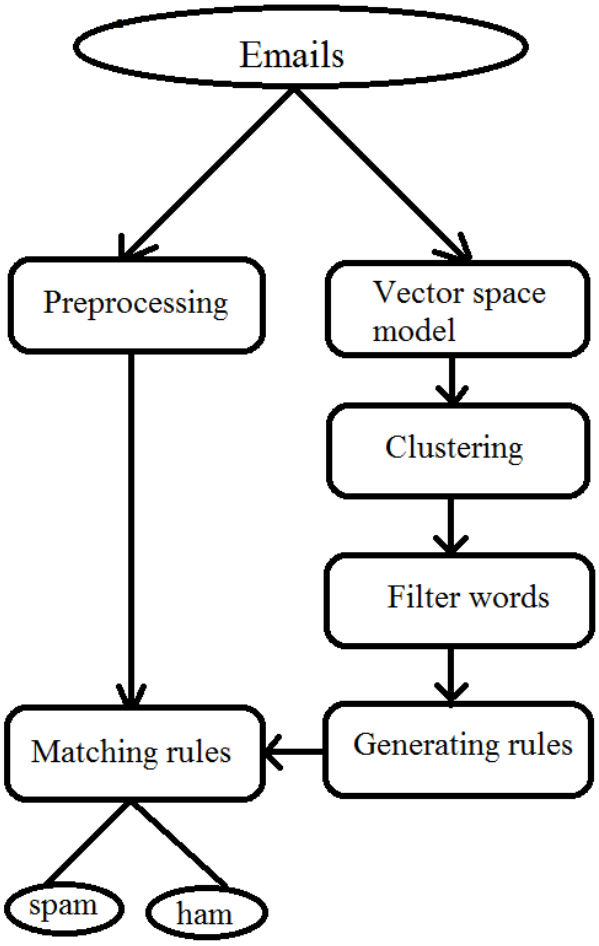

FIGURE II. SYSTEM ARCHITECTURE

Where, "." denotes the dot-product of the two frequency vectors A and B. A denotes the length of a vector. Document similarity is based on the amount of overlapping content between documents. The resulting similarity ranges from -1 meaning exactly opposite to 1 meaning exactly the same, with 0 usually indicating independence, and in-between values indicating intermediate similarity or dissimilarity. Apart from being simple, k-means algorithm is efficient for operating on large data sets. But the initial choice of the centroids can give varied outcomes. After applying the k-means algorithm the cluster with spammy emails is to be selected and spam words are to be extracted from them which give us the list of spam words. This forms a filtering step. Only the words that occur in the list are retained and the rest of the text is deleted. At the end of this operation, we get the various combinations in which the

spam words occur in the set of emails. This list can be updated periodically to include new words that can be considered as spam words. At this point we can go ahead and generate the association rules using the Apriori algorithm. Once we have the rules, they can be matched against new emails to decide whether it is spam or not. Association rules are if/then statements that help uncover relationships between seemingly unrelated data in a relational database or other information repository. For example, the rule egg, bread → butter found in the sales data of a supermarket would indicate that if a customer buys egg and bread together, he or she is also likely to buy butter. The two main important terms comes under association rules are support and confidence. Support is an indication of how frequently the item appears in the database. Confidence indicates the number of times the if/then statements have been found to be true. Support and confidence values can be changed to control the number of rules that are generated. It attempt to find subsets which are common to at least a minimum number C of the item sets. Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found. The purpose of the Apriori Algorithm is to find associations between different sets of data. In our technique, the associations that we are interested are between the spam words that occur in emails. Support and Confidence are used to control the number of rules generated, they are not the evaluation criteria. Instead, four measures precision, recall, specificity and accuracy were used in the evaluation process. Accuracy value might be high but it might be labeling only one of spam or non-spam emails correctly. If the precision is high, it means that the false positives are less. If he Recall is more, it means that the system recognizes most of the spam messages. A high value of specificity means that very few non-spam messages are associated as spam. A perfect predictor would be described as 100 percent sensitivity and 100 percent specificity. Christian Borglets implementation of the Apriori algorithm is used for generating frequent item sets. By varying the values used for support and confidence, the number of rules generated can be controlled. The proposed system is setup on a machine with the following hardware: Processor Intel(R) Core(TM) i3-2328M CPU @ 2.20GHz, 500 GB HDD. Set of email documents is converted into the vector space notation. Entries in the vectors are the tf-idf values. It may happen that some emails in the set have more length than others. In such emails, same terms are likely to appear more times i.e. they may have more term frequencies. Plus, such emails may also have more terms that can be considered as spam..

To understand how the new emails will be processed, lets take an example. Assume that the words lottery and gambling occur in the list of spam words. So there may be a rule of the form lottery > gambling. Any new email that has both these words in its content will be treated as spam. Likewise, several other rules may match for some test email. Emails which are not spam will not contain any words from the list of spam words or will not contain all the words that form a rule. Such emails will be identified as non-spam by the system.

IV. OBSERVATION RESULTS

The purpose of this paper is to present a technique to identify email messages as spam or non-spam. Once that is done, the accuracy of the method is tested to see how many mails were correctly categorized. Emails messages are represented as vectors. Clustering is then applied to group together spam and non-spam emails.These is then applied to the new mails to see whether they are spam or not. True positives, true negatives, false positives, and false negatives are the four possible outcomes when an email is associated by the system. Its false positive when the email is incorrectly identifed as spam, but in fact it is non-spam. Its false negative when the email is incorrectly identified as non-spam when it is spam in fact. True positives and true negatives are correct identifications. The corpus used for training and testing is the ling-spam corpus. In ling-spam, there are four subdirectories, corresponding to 4 versions of the corpus, which are: bare, lemm, lemm_stop and stop. Where lemming is similar to stemming and stop-list tells whether stopping is done on the text or not. The percentage of spam in this corpus is 18%. Final results are obtained by taking the average of the scores obtained in each experiments done.

Results of applying proposed model the on the ling-spam data set are like : Presicion was of 60% and Accuracy of 90%. As seen above 90% of the total emails were correctly identified and the Precision value were lesser compared to accuracy thus more false positives are generated.

V. CONCLUSION

A new technique to effectively detect spam emails using clustering and association rules was suggested. Clustering is used as a data reduction step to find the spammy clusters out of all the emails. Using these rules, we can then associate an incoming email as spam or non-spam.

ACKNOWLEDGEMENT

We are very grateful to all authors in reference section. Their methods, algorithms, conceptual techniques are very helpful for our research. All papers in the reference section are very useful for our proposed system.

REFERENCES

[1] N.S.Kumar, D.P.Rana, R.G.Mehta, Detecting Email Spam using spam word associations, IJETAE,ISSN 2250-2459, Volume 2,Issue 4,April 2012.

[2] More, Mugdha and Bharat tidke. “Social media opinion summarization using ensemble technique.” Pervasive Computing (ICPC), 2015 International Conference on IEEE 2015.

[3] Walin Ma, Dat Tran, Dharmendra Sharma, A Novel Spam Email Detection system based on Negative Selection, 2009 Fourth International Conference on Computer Science and Convergence Information Technology.

[4] Nitin Jindal and Bing Liu, Review Spam Detection, Proceedings of 16th International World Wide Web

conference, WWW '07. Ban_, Alberta, Canada 2007.

[5] Prabhakar, R. and Basavaraju, M. 2010. A Novel Method of Spam Mail Detection Using Text Based Clustering Approach. Phil. Trans. Roy. Soc.London, vol. A247, pp. 529551.

[6] Alguliev, R.M., Aliguliyev, R.M. and SNazirova, S.A.Classification of Textual E-mail Spam Using Data Mining techniques. Applied Computational Intelligence and Soft Computing, vol. 2011.

[7] Snehal Dixit and A. J. Agrawal, Survey On Review Spam Detection, International Journal of Computer Communication Technology ISSN, 2013.

[8] "Apache Mahout: Scalable Machine Learning and Data Mining." Internet:http://mahout.apache.org/ [Mar. 23, 2012].

[9] Firte, L, Lemnaru, C. and Potolea, R. 2010. Spam Detection Filter Using KNN Algorithm and Resampling,‖ in Intelligent Computer Communication and Processing (ICCP), 2010 IEEE International Conference

[10] Kyriakopoulou, A. and Kalamboukis, T. 2006. ―Text Classification Using Clustering,‖ in ECML-PKDD Discovery Challenge Workshop Proceedings.

