International Journal of
Emerging Research in Management &Technology
ISSN: 2278-9359 (Volume-4, Issue-5)

Research Article | **May 2015**

A Review Paper on an Analysis of Misclassification Error Detection in Mails Using Data Mining Techniqes

Ommerajan, Heena Khana
Panchkula engineering college
Hryana, India

Abstract—

This research is to classify and filter the large amount of data. The main purpose of this research is to reduce the error rate of the data and to improve the accuracy. In the previous techniques of classification there may be some miss classification. But in this research the problem of misclassification is reduced. The work is presented by this research is some modifications in the classification technique. Therefore, it's a good enterprise solution for filtering. This will optimize the system performance and make some improvements on the previous algorithm. This will give the better results from the previous one.

Keywords-- Email, SMTP, filtering, Bayesian filters, Spam.

I. INTRODUCTION

Email is electronic device. It is method of exchanging digital messages from source to destination. The exchange of messages from an author to one or more. Email messages can be text files, graphics images and sound files. Email messages are usually encoded in the ASCII text. But now-a-days, the problem in the email is spam and security also. Text editor is included in the email systems to compose the messages. When one send the message to the on specified address then one can also send the same message to the several users and this is called broad casting. As we know that emails are easy to use. Emails are fast and language used in emails is simple can be formal or informal. Message through email delivered at once. There is no paper work while using email. It contains friendly environment and can also have pictures, audio files, video files etc. There is also auto responders in email. Products can be advertised, so that companies can reach a lot of people and can advertise their product in a very short time. But having all these advantages emails have some disadvantages too like emails can carry viruses. Unknown and unwanted people can also send messages called spams. Through emails ones systems can get crashed. Mailbox may get flooded with emails after a certain time so one have to empty it from time to time.

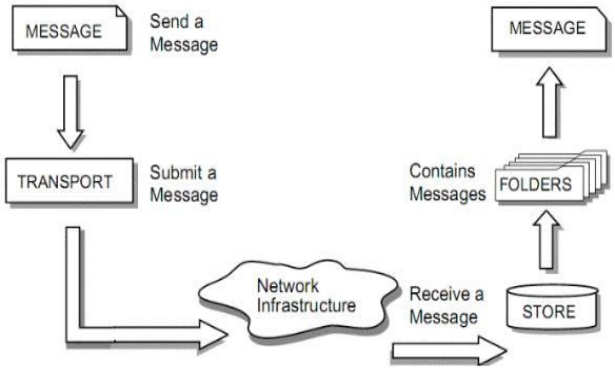

Fig. 1 Working of email server.

Their messages are modules which are added in the working of the email system. In the processing of the email, SMTP is also used. In shorts, the steps are:

- Message is sent by email client.
- Email server contacted to the recipients email server.
- Username's validity is checked by the email server.
- If valid username is typed, email is sent to the email server of the address.
- When the recipient sings in to his mailing account, he finds his email.

II. LITERATURE REVIEW

Androutsopoulos Ion, Paliouras Georgios, they had proposed and novel approach for An efficient method to construct automatically anti spam filters has been recently classified by Naive Bayesian algorithm. In this proposed research the memory based learning approached is used in which the comparison is done. In the proposed paper, the performance of the two machine learning algorithms in the context of ant spam filtering is investigated. So the method achieved very accurate spam filtering, outperforming undoubtedly the keyword based filter a of broadly used e-mail reader [1].

Konstantinos V. Chandrinos , Constantine D.Spyropoulos et al, in this paper, author compare the various different , to detect the spam Naive Bayesian is trained automatically. This approach is tested on a large collection of personal email messages which are made publically available in "encrypted" from contributing towards standard benchmarks. Appropriate Cost sensitive measures are introduced. In this approach Naive Bayesian filter is compared to see the performance, to filter which is part of widely used email reader. In this approach filtering/routing, text categorization, test collection keywords are used. In conclusion, it concluded after experiment results that cost sensitive evaluation suggests that neither the Naive Bayesian nor the keyword-based filter perform well enough to be used [2].
Jonas Hörnstein et al discussed the rise of the World Wide Web and the ever-increasing amounts of machine-readable text has caused text classification to become an important aspect of machine learning. One application that has the potential to have an effect on almost every user of the Internet is e-mail filtering. The World talk Corporation estimates that over 60 million business people use e-mail. Many more use e-mail purely on a personal basis and the pool of e-mail users is growing everyday. And, automated techniques for learning to filter e-mail have yet to significantly affect the e-mail market. Here, problems are attacked that plague practical e-mail filtering and suggest solutions that will bring us closer to the acceptance of using automated classification techniques to filter personal e-mail. Results are presented from a number of experiments and show that a system such as file could become a useful and valuable part of any e-mail client. This paper discusses practical concerns surrounding the application of text classification to the problem of mail filtering [3].
D. Štorekhad devolped Rajput Arjun et al In this approach, proper filtering is for the junk (spam) mails and irrelevant mails. So proper filter for these applied on email servers. If dynamic nature is used with available spam filters then the results of spam filtering can be unexpectedly better than earlier. So in this research some techniques for improvement of Bayesian filter is discussed. Keywords, hash map and spam probability is are used in this approach. In this approach, most of the spam detection techniques are unable to find these spams and the database of spam should be updated all the time. In this approach dynamic trains and classification is discussed [4].
Li-wei Zhang, Zhen-fang et al In this proposed paper, as they knew that naive Bayesian algorithms had been widely used. They basically used for the filtering of the spam. Texts can be classified more quickly and correctly, because the naive baysian algorithm is simpler. As we know there are two types of mails one is spam mail and another is legitimated mail, so the traditional method does not consider the different features between the Spam mail and the legitimate mail and also the loss of misclassifying legitimate mail as spam is not taken into account. So the improved algorithm in this research based on naive Bayesian. In this research boosting method is also proposed. But applying these methods better performance had been shown. As we studied this research combined Bayesian algorithm with boosting and this presented a new spam filtering algorithm. As earlier Bayesian algorithm ignores the important information but in the experimental results they had seen that this algorithm can overcome this problem also. And of course this algorithm is more feasible. Now this research paper is going to be the main base paper of my research. I will do some modifications in the same algorithm, as I discussed earlier, I will use expert system and then add cases and also add actions correspondence to that particular case. So this base paper and the algorithm is going to help me in my research [5].
Yang Song, Aleksander Kołc et al In this paper, Several improvements to the NB classifier have been proposed which is well suited to applications requiring high precision, such as spam filtering. A term-weighting function based on the correlation measure is introduced, which was demonstrated to perform very well on its own and as well as alternative to the typical multiplicative aggregation of several term weighting functions. The problem of feature sparsity is addressed for short documents a class dependent. To expand their attribute vectors CF technique was proposed. Improvement of the classifier performance in the low-FP region is shown. Finally, a novel two-stage NB cascade was introduced. This combines the ability to tackle the potential non-linearity of the decision boundary with an algorithm that jointly optimizes the decision thresholds of the terminal components of the cascade so as to achieve the best performance at a specified FP rate. For NB the proposed techniques have been shown to be quite effective, they are also applicable to other learners [6].
M. Basavaraju, Dr. R. Prabhakar et al The novel method of efficient spam mail classification using clustering techniques is presented in this research. E-mail spam is one of the main problems of the today's internet, bringing financial harm to companies and annoying individual users. In between the approaches developed to discontinue spam, filtering is an important and popular one. A new spam finding technique using the text clustering based on vector space model is proposed in this research paper. By using this method, one can take out spam/non-spam email and detect the spam email efficiently. Vector space model shows the representation of data. Clustering is the technique used for data reduction. It splits the data into the groups based on pattern similarities such that each group is abstracted by one or more representatives. In recent times, there is a growing emphasis on exploratory analysis of very large datasets to discover useful patterns. Each cluster is abstracted using one or more representative [7].

III. EMAIL FILTERING SCHEME

For a better understanding of the task of Email filtering is the processing of email to systematize it according to the exact criteria. Most often this refers to the automatic processing of incoming messages, but the term is also used to the involvement of human intelligence in addition to anti-spam techniques. Bayesian spam filtering is a statistical method of e-mail filtering. Bayesian spam filtering makes use for Naive Bayes classifier to make out spam e-mail. Work is classified by Bayesian to compare the use of tokens i.e typically words, or we can say irregularly other things, with spam and non-spam e-mails. Bayesian spam filtering is a extremely powerful technique for constricting with spam, that can adapt itself to the email needs of individual users, and gives low false positive spam finding rates that are generally acceptable to users.

Bayesian filtering is one of the most effective and bright solutions to fight with spam email nowadays. Spam is a trouble faced by all email users and it reflects no sign of slowing down anytime shortly; in fact, the number of spam emails is growing daily. Added to this, spammers are becoming more complicated and are continuously managing to outsmart 'static' methods of fighting spam

There are basically two types of filtering, inbound filtering and outbound filtering. In inbound filtering, email messages are sheltered by the filtering system. In this type of filtering, message scanning process is involved. In case of outbound filtering, scanning email messages from local users before any potentially unsafe messages can be delivered to others on the Internet. Outbound email filtering is commonly used by Internet service providers is transparent SMTP, in which email traffic is intercepted and filtered by means of a transparent proxy within the network.

IV. EMAIL FILTERING TECHNIQUE

The techniques currently used by most anti-spam software are static, meaning that spammers simply examine the latest anti-spam filtering techniques and hit upon ways how to cut them, usually done by simply change the message a little. This gave anti spam developers a new challenge – come up with a new anti spam technique; one that was familiar with spammers' tactics as they vary over time, and that is capable to adapt to the particular organization that it is protecting from spam. There are different emails filtering methods.

1) Blacklist: Blacklist comes under the list based filters. This is spam filtering method attempts to stop unwanted email by blocking messages from the list of sender. Blacklist contains the records of email addresses. In this when in coming message arrives, the spam filter checks to see if its IP or email address is on the blacklist. Then it considers the message as a spam and then reject it.

2) Whitelist: Whitelist blocks spam using a system almost exactly opposite to that of blacklist. In this if an unknown sender's email address is checked against the database, if they have no history of spamming, their message is sent to inbox and then they added to the whitelist.

3) Word based filtering : Word based filtering comes under the content based filtering it is the simplest form of filtering .word based filtering is the capable technique for fighting junk email. For example, if the filter has been set to stop all messages containing the word "acbd". But spammers often purposefully misspell keywords in order to evade word based filtering and this is the main problem in this type of filtering

4) Heuristic filters: This type of filtering contain multiple terms instead of containing one term based on the word based filtering. In this filter adds up all the points and then calculates the total score, the heuristic filters work fast, minimize email delay, and quite effective. But the problem is this technique is some spammers might learn which words to avoid including, thus fooling the heuristic filter into believing they as begin senders.

5) Bayesian filters: Bayesian filters technique is the most advance content based technique. It employs the laws of mathematical probability to settle on which message are real and which message is spam. In this, filter takes words and phrases finding legitimate mails ad adds them to the list. This method acquires a training time period before it starts running well. There are other filtering methods like challenge/response system, collaborative filters**.** Bayesian spam filtering is the process of using a naive Bayes classifier to identify spam e-mail. It is depended on the principle that most events are dependent and that the probability of an event occurring in the future can be inferred from the previous occurrences of that event. This similar method can be used to classify spam. If some content of text helds often in spam but not in legitimate mail, then it would be reasonable to predict that this email is almost certainly spam. Bayesian spam filtering has become a popular mechanism to distinguish illegitimate spam email from legitimate email. Nowadays most of mail clients apply Bayesian spam filtering. Bayesian filters must be 'trained' to work effectively. Particular words have probabilities (also known as likelihood functions) of occurring in spam email but not in real email. For instance, most email users will commonly encounter the word Viagra in spam email, but will rarely see it in another email. Before mail will be filtered using this technique, the user needs to create a database with words and tokens (such as the $ sign, IP addresses and domains,), collected from a sample of spam mail and valid mail (referred to as 'ham'). For all words in each training email, the filter will regulate the probabilities that each word will appear in spam or legitimate email in its database. After training, the word probabilities are used to calculate the probability that an email with a particular set of words in it belongs to either category. If the sum of word probabilities exceeds a threshold, the filter will take the email as spam. Users can then make a decision whether to move email marked as spam to their spam folder or whether to just delete them.In recent years, the increasing popularity and low cost of email have attracted the attention of direct marketers. Using readily available bulk-mailing software and large lists of email addresses, typically harvested from web pages and newsgroup archives, it is now possible to send blindly unsolicited messages to thousands of s at essentially no cost. As a result, it is becoming increasingly common for users to receive daily large quantities of unsolicited bulk email, known as spam, advertising anything from vacations to get-rich schemes. The term Unsolicited Commercial E-mail (UCE) is also used in the literature. We use "spam" with a broader meaning that does not keep out unwanted bulk e-mail sent for non-commercial purposes.

V. CONCLUSIONS

The presented work shows the email filtering and clustering of the filtered email saved in the database. The email filtering is based upon the filtering techniques like white list and black list filtering techniques or mainly Bayes theorem and support vector machine (SVM). The filtering of the email is based upon the extensions of the files, like doc files, jpg files, zip files. After the filtering, emails which are filtered and saved in the database are clustered. Clustering of the emails is based upon the K-means algorithm and SVM algorithm. Email filtering and clustering is the main work in my

research. Previous algorithm was used to filter the email only for the spam. The work is enhanced by the previous one and is implemented to give better results by filtering the email and clustering of the filtered email.

ACKNOWLEDGMENT

Working on this thesis **of Analysis of Misclassification Error Detection in Mail Using Data Mining Techniques** provided a unique experience and analysis, I feel great pleasure and privilege in working over this research. I am deeply indebted to **"Panchkula Engg. College"** for the invaluable guidance, support and motivation for the many other aids without which it would have been impossible to complete this project.

I have no words to express my deep sense of gratitude for Heena Khanna (Mentor) mam for her enlightening guidance, directive encouragement, suggestions and constructive criticism for always listening to our problems and helping us out with their full cooperation.

Last but not the least,Father Haji Bashir Ahmad Dar Mother Jahan Ara Brothers Nisar Ahmad Mir, Ommer Bashir who have given me that much strength to keep moving on forward every time, we are greatly thankful to them and have no words to express my gratitude to them.

REFERENCES

[1] Androutsopoulos, I., Paliouras, G., Karkaletsis, V., Sakkis, G., Spyropoulos, C. D., & Stamatopoulos, P. (2000). Learning to filter spam e-mail: A comparison of a naive bayesian and a memory-based approach. arXiv preprint cs/0009009.

[2] Basavaraju, M., & Prabhakar, R. (2010). A novel method of spam mail detection using text based clustering approach. International Journal of Computer Applications, 5(4).

[3] Hovold, J. (2005, July). Naive bayes spam filtering using word-position-based attributes. In Proceedings of the 2nd Conference on Email and Anti-Spam (CEAS 2005).

[4] Jin, X., Wang, L., Lu, Y., & Shi, C. (2003). MC-tree: Dynamic index structure for partially clustered multi-dimensional database. Tsinghua Science and Technology, 8(2), 174-180.

[5] Liu, P. Y., Zhang, L. W., & Zhu, Z. F. (2009). Research on e-mail filtering based on improved Bayesian. Journal of Computers, 4(3), 271-275.

[6] Rajput, A., & Toshniwal, D. Adaptive Spam Filtering based on Bayesian Algorithm.

[7] Rennie, J. (2000, August). ifile: An application of machine learning to e-mail filtering. In Proc. KDD 2000 Workshop on Text Mining, Boston, MA.

[8] Sahami, M., Dumais, S., Heckerman, D., & Horvitz, E. (1998, July). A Bayesian approach to filtering junk e-mail. In Learning for Text Categorization: Papers from the 1998 workshop (Vol. 62, pp. 98-105).

[9] Song, Y., Kołcz, A., & Giles, C. L. (2009). Better Naive Bayes classification for high-precision spam detection. Software: Practice and Experience, 39(11), 1003-1024.

