ISSN 2395-1621

Content Based Classification Algorithms for Email Filtering

[#1]Tushar Bhagat, [#2]Ravikiran Tikale, [#3]Kiran Ugalmugale

[1]tusharbhagat313@gmail.com
[2]ravikiran.tikale@gmail.com
[3]onlyukboss@gmail.com

[#123]Department of Computer Engineering

JSPM's
Imperial College Of Engineering & Research,
Wagholi, Pune – 412207

ABSTRACT

Electronic mail is used daily by millions of people to communicate around the globe and is a mission-critical application for many businesses. Over the last decade, unsolicited bulk email has become a major problem for email users. An overwhelming amount of spam is flowing into users‟ mailboxes daily. Not only is spam frustrating for most email users, it strains the IT infrastructure of organizations and costs businesses billions of dollars in lost productivity. The necessity of effective spam filters increases. In this paper, we presented our study on various problems associated with spam and spam filtering methods, techniques. We also filter the upcoming mail from other user, if any unwanted mail coming then we automatically block that mail for security purpose.

ARTICLE INFO

Article History

Received: 1st April 2017

Received in revised form :

1st April 2017

Accepted: 4th April 2017

Published online :

4th April 2017

Keywords: Spam mail, Spammers, Security, Content Based Filtering.

I. INTRODUCTION

The spammed emails can be expressed as unbidden and unnecessary Emails sent with the motive of pecuniary profit or hazardous threats to the system. They may be used to distribute viruses or fake announcements that cause responders an average loss of 25 USD per reply. The recently done survey forecast that the 1 user among 50,000 users open or give replies to the spammed mails Moreover, the fact that 58 billion out of the 90 billion emails which are composed or sent daily are the fake mails or the threatened mails this underlines for both the instant and importance of creating and adopting effective processes for received emails. Identification and prevention of the spam is the most crucial work in the pattern recognition and data mining advancement as furious development has been done already for developing algorithms which are capable enough to identify the spam from the fake or the threatened emails. Emails are mainly filtered on the basis of its content in the body, it may include text and photos or videos or other stuff which may provide the detailed information about the sender of the email.In this paper we have compared some proposed content based filtering algorithms that are based on the text classification for determining whether the mail is threatened or not. In this paper, we compare among some proposed content based filtering algorithms that rely on text classification to decide whether an email is spam or not. We present and analyze results of these experiments.

Social networking sites which are available now a days are most famous and easy way for communication, sharing a huge amount of information about the people. Generally daily and continuous communications implies sharing of the numerous types of materials which includes texts, images, audio clips and video clips too. According to recent survey the statistics of Facebook and Twitter users shows there average users sharing approximately 90-95% of data every month .The dynamic character of these data creates the premise for the employment of web content mining strategies aimed to automatically discover useful information dormant within the data. They are many instrumental to provide an active support in complex and sophisticated tasks involved in OSN (Online Scouting Network) management, such instance access control and information filtering. Information filtering has explored for what concerns text documents and web content. However, aim of majority of all these proposals is to provide users a classification mechanism to avoid they are over useless data. In OSNs, information filtering can also be used for a different, more sensitive, purpose. This is due to the fact that in OSNs there is the possibility of posting or commenting

other posts on particular public/private areas, called in general walls. Information filtering can be used to give users the ability to work automatically control messages written on their own walls, by filtering out unwanted messages. We believe that this is a key that has not been provided so far. The aim of the present work is therefore to propose and experimentally evaluate an automated system, called Filtered Wall (FW), able to filter unwanted messages from OSN user walls. We exploit Machine Learning (ML) text categorization techniques automatically assign with each short text message a set of categories based on its content. The major efforts in building a robust Short Text Classifier (STC) are concentrated in the extraction and selection of a set of characterizing and discriminate features.

II. RELATED WORK

In Content-Based filtering, user is assumed to operate independently. As a result Content-Based Filtering system select information items base on correlation between content of items and user can also prefer to oppose collaborative filtering system that select items based on correlations between peoples with similar preference. While electronic mail was original domain of early work on information filtering process. Document processed in content-based filtering is mostly textual in nature and can makes content-based filtering close to text classification. This activity of filtering can modelled, as case of single labelled, binary classification and partition of incoming document in the relevant and non-relevant categories.

The various spam filtering techniques:

Rule based filtering:
Evaluate a large number of patterns-mostly regular expressions against a candidate message. Some matched patterns add to a message's score, while others subtract from it.

Bayesian classifier:
Particular words have particular probabilities of occurring in spam email and in legitimate email. The filter doesn't know these probabilities in advance, and must first be trained so it can build them up.

Content based Spam Filtering Techniques:
In Content-Based filtering, each user is assumed to operate independently. As a result, a Content-Based filtering system selects information items based on the correlation between the content of the items and the user preferences as opposed to a collaborative filtering system that chooses items based on correlation between people with similar preferences. While electronic mail was the original domain of early work on information filtering, subsequent papers have addressed diversified domains including newswire articles, Internet "news" articles, and broader network resources. Documents processed in content-based filtering are mostly textual in nature and this makes content-based filtering close to text classification. The activity of filtering can be modeled, in fact, as a case of single label, binary classification, partitioning incoming documents into relevant and no relevant categories.

III.EXISTING SYSTEM

Cache architecture consists of two lists namely black list and white list.

Black List: In our research we have put a few domains and email ids in the black list which were presumed of causing danger or threat. For example those websites can be put in blacklist which have a past record of fraudulent or which exploits browser's vulnerabilities. In creating a filter; if the sender of mail has its entry in the black list then that mail is undesirable and will be considered as spam.

White List: It is opposite to the black list concept. It consists of the list of entries which can penetrate through and are authorized. These mails are considered as ham mails and can be accepted by the user. It has a set of URLs and domain names that are legitimate. After creating both the lists when any email arrives the 'To' and 'From' field is extracted from its subject to check if it is in the black list or the white list. The main rule applied here is that if the sender is from the black list then it will be considered as a spam mail.

IV.PROPOSED SYSTEM

A fundamental problem we often study is about leveraging the secrecy of a small piece of knowledge into the ability to perform cryptographic functions (e.g. encryption, authentication) multiple times

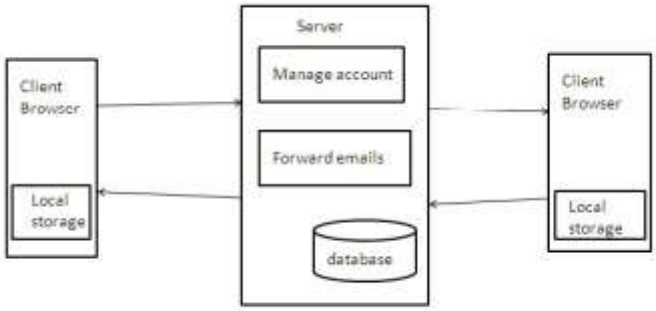

Fig 1: System Architecture

Module:

Client:
Client menace number of user cans browse the emails. He can perform different activity like manage account forward emails etc.

Server:
Server can perform interface between client browser and database. Server is strong parameter in our project, if server is slow down then the forwarding or receiving emails also slow down.

Database:
Database can stored the all reporting data; user can perform in sending or receiving all types of emails.

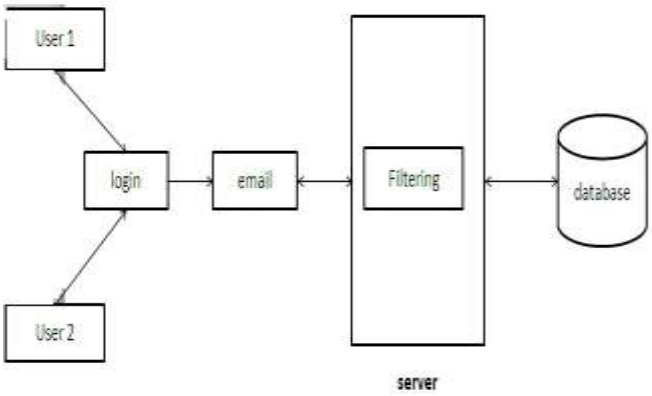

Fig 2: Functionality working

V. RESULT

Fig 3. Home Page

Fig 4. Compose mail

Fig 5. Block receive mail

VI. CONCLUSION

In this paper we have presented a basic study of the SVM, Decision Trees, Neural Networks and Artificial Neural Network classifies on the basis of the perfection, accuracy performance of the algorithms. The better approach of this paper will be the additional features added for classifying the ham or spam mails using advanced email Filtering algorithms.

REFERENCE

[1] Miszalska, I., Zabierowski, W., & Napieralski, A. (2007, February). "Selected Methods of Spam Filtering in Email." In CAD Systems in Microelectronics, 2007. CADSM'07. 9th International Conference-The Experience of Designing and Applications of (pp. 507-513). IEEE.

[2] Scholar, M. (2010). "Supervised learning approach for spam classification analysis using data mining tools." organization, 2(08), 2760-2766.

[3] Youn, S., & McLeod, D. (2007). "A comparative study for email classification." In Advances and Innovations in Systems, Computing Sciences and Software Engineering (pp. 387-391). Springer Netherlands.

[4] Xiao-li, C., Pei-yu, L., Zhen-fang, Z., & Ye, Q. (2009, August). "A method of spam filtering based on weighted support vector machines." In IT in Medicine & Education, 2009. ITIME'09. IEEE International Symposium on (Vol. 1, pp. 947-950). IEEE.

[5] Sculley, D., & Wachman, G. M. (2007, July). "Relaxed online SVMs for spam filtering." In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 415-422).

[6] Chan, T. Y., Ji, J., & Zhao, Q. "Learning to Detect Spam: Naive-Euclidean Approach." International Journal of Signal Processing, 1.

[7] Puniškis, D., Laurutis, R., & Dirmeikis, R. (2006). "An artificial neural nets for spam e–mail recognition." Elektronika ir Elektrotechnika (Electronics and Electrical Engineering), 5(69), 73-76.

[8] Drucker, H., Wu, D., & Vapnik, V. N. (1999). "Support vector machines for spam categorization." Neural Networks, IEEE Transactions on, 10(5), 1048-1054.

[9] Provost, J. (1999). "Naıve-Bayes vs. Rule-Learning in Classification of Email." University of Texas at Austin.

[10] Medlock, B. (2006, July). "An Adaptve, Semi-Structured Language Model Approach to Spam Filtering on a New Corpus."

