Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 4, April 2015, pg.234 – 237

RESEARCH ARTICLE

Spam Filtering using the Social Anthropology and Data Mining Technique

Sanjay Kumar. N

Computer Science and Engineering, New Horizon College of Engineering, Bengaluru, India

sanjay.nadupanna@gmail.com

Rohini. T

Asst. Professor, Computer Science and Engineering, New Horizon College of Engineering, Bengaluru, India

rohini.antharmuki@gmail.com

Abstract— The electronic mails are playing important role in the todays professional and personal life. This has also increased the number of spam mails. Many techniques are evolved to filter the spam mails. Bayesian spam filter is one of the traditional technique which is very stable for the static keywords. Bayesian spam filter mainly focus on parsing the spam keywords. Bayesian filter can be easily deceived with synonyms of the spam keywords. In SOAP (SOcial Aided Personalized) effective spam filter, each node connects to its social friends forming a distributed overlay by directly using the social network as overlay links. It collects the information and check the spam autonomously in a distributed manner. SOAP incorporates the four components into the Bayesian spam filter: social closeness, social interest, adaptive trust management and friend notification. The time complexity is one the main drawback, when the mail is sent from the anonymous mailer. Here we propose Social Anthropology and Data Mining (SADM approach) by combining data mining technique like apriori algorithm and association rule mining with SOAP components to deal with the incoming e-mails in an efficient way. It also evades the dependency on the overlay links and makes easier and feasible for the implementation.

Keywords— Bayesian filter, overlay, SOAP, apriori algorithm, strong association rule

I. INTRODUCTION

The electronic mails is one of the most ubiquitous communication method in our daily life, both personally and professionally. Spam mails are increasing day-to-day in an explosive manner. The primary way to prevent a spam mail is to make it worthless sending them. The survey provides a statistic of 120 billion spam mails sent every day. Spam emails thwart with both the email service providers and the end users. Spam mails also occupy the space with legitimate emails thereby wasting the storage space.

Ultimately a spam filter should be efficient to stop the spam mails reaching the users inbox. Such a spam filter should be robust, personalized and user friendly.

Accurate filters should result in the less false positive and false negatives. False positive are the mails which are appropriate emails but treated as a spam mails. False negative are the spam emails that are not detected in the filtering process. The traditional spam filters can be deceived by using the innocuous (poison attack) words by the clever spammers. During the poison attack, spam filters may fail to detect the spam mails.

Along with the personalized and based on the user (dis)interest the mails can further be classified efficiently by incorporating the social context. First, the filter should consider the closeness between the receiver and the sender as well as trust level on Friend of Friend (FoF) [2]-[4]. Second, the (dis)interest and the domain of the individuals should be considered. The spam filter should be user-friendly, since an interest varies from person to person. Hence care should be taken while filtering the emails into different categories.

In this paper along with Bayesian and SOAP [1] approach we propose the implementation of the Data Mining technique to filter the emails efficiently. Apriori algorithm is used to find the, frequently repeated words in the email content. Using the Strong Association rule mining we further identify the combination of the words in the content, which enhances the identification of the email class.

II. RELATED WORK

The large quantity of spam has accelerated many spam filtering approaches. These approaches can be categorized into content-based and identity-based.

A. Identity-Based Approach

The identity-based approach is the simplest approach of spam filtering by maintaining the blacklist and whitelist [5]-[8]. Both blacklist and the whitelist maintain the list of addresses, to specify which address mails should be filtered before reaching the user inbox. The mails from the whitelist address would be representing the legitimate sender and the mails from the blacklist would be treated as spam mails. Another way of identity-based approach is blocking the mail address which would be sending the same mail to the large number of people frequently. This is one of the server-side solutions to identify such address. Email scoring mechanism [9] based on an email network augmented with reputation ratings. An email is considered spam if the reputation is very low of the sender. The identity-based approach deals only with the sender address rather than the content of the email.

B. Content-Based Approach

The primary approach of the content-based spam filtering is parsing the static keyword. In this approach the spam mails can be easily passed through the filtering process by using the synonyms of the spam keywords. The machine learning-based approach uses the training sample to detect the spam mails. This approach is not dynamic and takes more time to update for the new spam keywords.

C. SOAP:SOCIAL NETWORK BASED BAYSEIAN SPAM FILTER

The SOAP leverages social information including personal (dis)interest and social information. It mainly uses the nodes forming the overlay by connecting to their friends. SOAP integrates the four components into the Bayesian filter: social closeness-ness spam filtering, social interest-interest based spam filtering, adaptive trust management, and friend notification. This approach includes legal issues for accessing the details of the different overlay links. Moreover many social network proprietaries do not provide their user details, due to information confidentiality. Even though this approach provides fine filtering of the spam mails, the time complexity is the major issue in worst case of the new sender address.

III. PROPOSED METHOD

In this paper we discuss Social Anthropology and Data Mining (SADM) approach along with the selected methods of the Bayesian and SOAP approach. The email service provider can implement this technique without depending on any overlay nodes. The proposed technique includes the following components: (1) members list (2) members friend list (3) global list (4) content-based data mining filter. Figure 1 shows the architecture of the overall spam filtering components and Figure 2 shows the flow of spam filtering process.

A. Members List

When a person receives the email, first the sender address will be verified in the receiver contacts. If the sender is in the receiver contact list, then the mail will be treated as

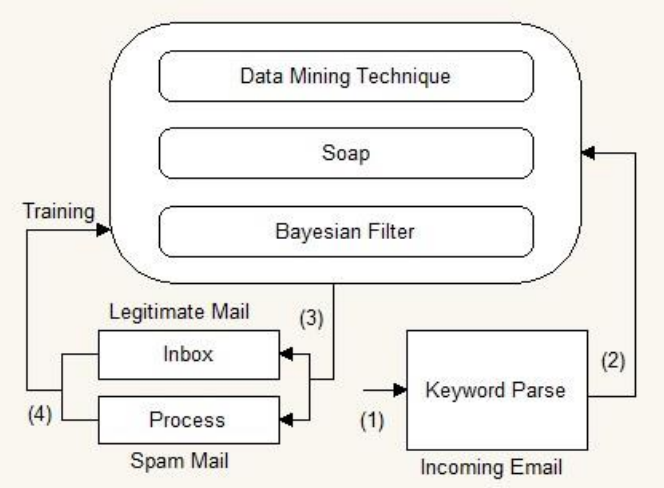

Fig. 1. Architecture of Filter.

legitimate email, there by excluding the further filtering process. If the sender is in the blacklist of the receiver then the mail will be treated as spam mail. This is one of the straight forward method based on identity. In the case, if the sender is not in the member contact list or blacklist, then the next level of filtering technique will be instantiated.

B. Member Friends List

Email received from the sender who is not in the member contact list will be verified in the member friends contact list. The sender will be validated with the friends contact list to assure that the sender is not a spammer. If the sender address is in the friends contact list then the mail will be treated as a legitimate email. Since a friend of friend (FoF) is less likely to send a spam mail. If the sender is in the blacklist of the friends list then, the mail will be treated as the spam. Here we specify the threshold to decide whether the sender should be really treated as spammer, since the interest differs from person to person. This section implements the concept of Adaptive trust management of the SOAP but without depending on any overlay nodes. Still, if the sender address is not present in the friends contact list or blacklist then the mail will be further processed to the next level of filtering.

C. Global List

When the sender is not present in the both member list and the member friends list, the sender address will be validated across the global member list. Here the sender address will be verified in the global member contact list. If any of the members has received the email from the particular sender then we check the status or trust level of the global member. However if the more number of members have treated the sender as spammer then automatically the mail is blocked reaching to the receiver inbox. In case, more number of people has treated the sender as genuine then, the mail from the sender is treated as legitimate. To avoid the false negative rate we can specify the threshold for members trust level. Say, among the members who received a mail from the particular sender in which more than 50 percent people treat it as a spam then we can decide the sender as spam. Based on the number of user we can set the threshold limit to decide the class of the email. If the average of members is less than the specified threshold, then the email will be sent to the next level of filtering process, where the content of the email will be parsed.

D. Content-Based DataMining Approach

Once the classification of the email cannot be identified in the above three process then, the content of the email will be parsed to identify its class. Here we can also introduce the domain concept of filtering mails based on the occupation or area of interest. For example, if a person is working as software developer then, the mails regarding the new technology, seminars, webinar sessions, tools so on will be a relevant to the end user. Similarly, in this section we can also consider different domain based filtering. Hence while parsing the keyword it can also be compared with the domain keywords to decide the class of the email. First, in this process the useless words are eliminated. The keywords are parsed and calculate the weight of the each word which is frequently repeated using the apriori algorithm. The calculated words which satisfy the minimum support are compared with training sample of the Bayesian filter, which specifies the class it belongs to. To make it more precise we further use the Strong Association mining rule [10]. Here we find the number of times the association of words occurred in the email content. If the

combination of the words present is equal to or more than the specified minimum confidence; then, that email can be easily classified into particular class. This technique helps to filter the email in finer granularity and result in reducing the false positive rate and false negative rate.

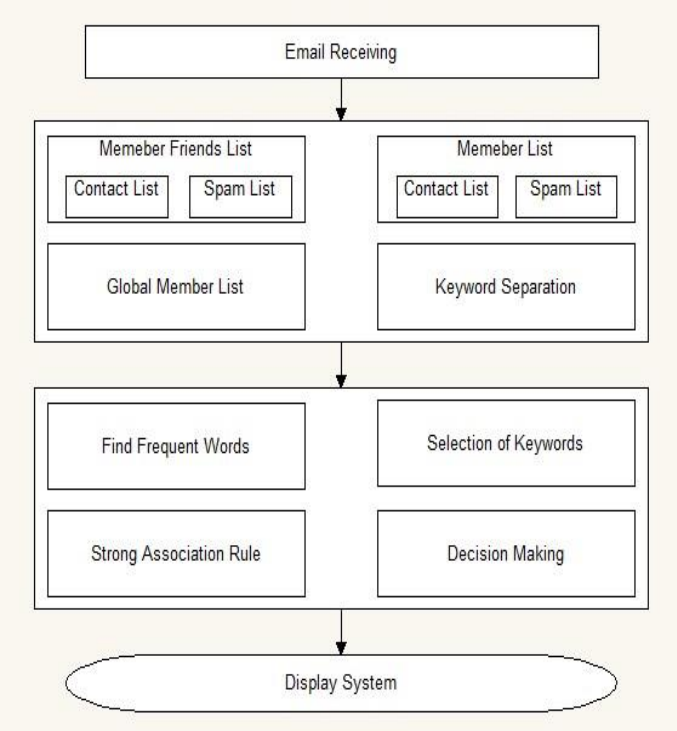

Fig. 2. Flowchart of filtering process.

IV. CONCLUSION

This paper introduces the SADM approach along with traditional approaches to efficiently filter the spam mails. Compared to other approaches this paper specifies the independent way of filtering spam without depending on any other overlay efficiently. Also overcome the drawback of SOAP approach w.r.t time complexity by making it feasible in the implementation. The SADM approach also decreases the false positive and false negative rate than the traditional filtering approaches. Further the efficiency can be compared using the different data mining algorithm likes FP Growth, Decision tree and so on.

References

[1] Haiying Shen, "Leveraging Social Networks for Effective Spam Filtering," in IEEE Transaction on Computers, vol.63, No. 11, November 2014.

[2] S. Hameed, X. Fu, P. Hui and Satry, "LENS:Leveraging Social Networking and Trust to Prevent Spam Transmission, "Proc. Int'l Conf. Network Protocols (ICNP), pp. 13-18, 2011.

[3] S. Garriss, M. Kaminsky, M.J. Freedman, B. Karp, D. Mazieres, and H. Yu, "Re: Reliable Email," Proc. Nat'l Securities Depository Limited (NSDL), pp. 1-10, 2006.

[4] P. Oscar Boykin and V.P.Roychowdhury, "Leveraging Social Networking to Fight Spam," Computer, vol. 38, no. 4, pp. 61-68, 2005.

[5] "SpamCop Blocking List," http://spamcop.net/bl.shtml, 2013.

[6] "DNS Real-time Black List," http://dnsrbl.net/, in 2015.

[7] Spamhaus, http://www.spamhaus.org/sbl/index.lasso, in 2015.

[8] Blars.org, http://www.blars.org/, in 2015.

[9] J. James and J. Hendler, "Reputation Network Analysis for Email Filtering," Proc. Conf. Email and Anti-Spam (CEAS), pp. 1-10, 2004.

[10] Jiawei Michelin Kamber, "Data Mining Concepts and Techniques", Morgan Kauf Mann Publishers.

