Content Based Spam E-mail Filtering

Sandhya Dhakane, Apurva Badgujar, Sohel Bagwan, Pooja Gawali, Prof. Bharti Kudale
Genba Sopanrao Moze College of Engineering Balewadi, Pune

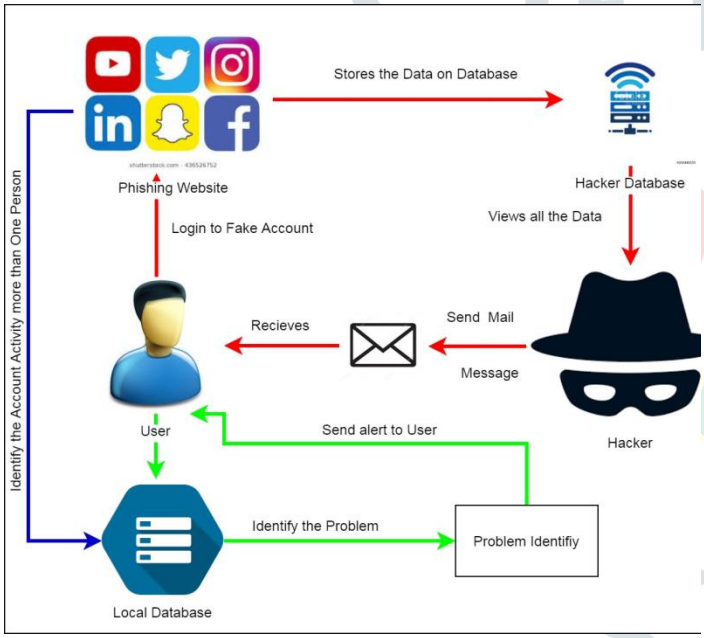
Abstract: Currently, E-mail is one in all the foremost necessary methods of communication. However, the increasing of spam emails causes holdup, decreasing productivity, phishing, that has become a heavy drawback for our society. And the variety of spam e-mail is increasing per annum. Therefore, spam e-mail filtering is a vital, meaningful and difficult topic. The aim of this analysis is to seek out associate effective resolution to filter doable spam e-mails. And as we know, in recent days, there are several techniques that spammers use to avoid spam-detection like obfuscation techniques. In this case, the subsequent projected approach uses email content only to create keyword corpus, Alongside some text process to handle obfuscation technique. The rule was evaluated using the CSDMC2010 SPAM corpus dataset that contained4327 emails within the coaching dataset and 4292 emails within the testing dataset. The experimental results show that the projected algorithm has ninety-two.8% accuracy.

Keywords: Spam E-mail, Ham E-mail, Keyword Corpus, Spam E-mail Filtering

Introduction: Currently, E-mail is one of the most important methods of communication. However, the increasing number of spam e-mails causes traffic congestion, decreases productivity, and phishing; which has become a serious problem for our society. Based on a recent Internet Security Threat Report, published by Symantec Corporation, in 2012 and 2013, the estimated Global Email Spam Volume per day is about 30 and 29 Billion, and the global average spam rate was 69% and 66%, respectfully. The real cost of spam emails is more than one can imagine. According to a paper, it cites the real cost of spam emails could at least add up to $20 billion annually to American firms and consumers, and the cost could be much higher. Therefore, spam e-mail filtering is an important and meaningful topic. Typically, E-mails are mainly consisted of two parts: header and body. In the header section, it contains many fields that can be categorized into two main types: mandatory and optional. Mandatory fields include “From”, “To and Cc”, “Sender” etc. Optional fields contain “Subject”, “Message-ID,” etc. Both mandatory and optional fields provide valuable information, like sender’s email address, number of recipients, and subject to help us to classify spam email. In practice, spam email filtering methods can be categorized into several categories. For example, blacklists and whitelists, IP blocking, header-based filtering and content-based filtering approach. Blacklists, whitelists and IP blocking are relatively the fast way, as compared to other detection approaches, to identify spammers. However, blacklists and whitelists or IP blocking have potential issues that the spammer could change current email account(s) or one IP to another one, in order to escape detection. In this case, normal methods could not easily filter these spam emails. Poor performance and low accuracy are the result of using these approaches. This paper proposes a content-based spam email filtering approach. The proposed algorithm contains two main phases: training phase and classification phase. In the training phase, individual users’emails are extracted from training datasets. After the email content is collected, the next step was to build a spam and ham keywords corpus that was used to compare with those keywords that were extracted from individual users’ email. Before comparing those extracted words with the spam and ham keywords corpus, in order to improve the accuracy and handle more possible spam techniques, some content processing methods are applied to handle obfuscation techniques that spammer(s) intentionally apply to elude keyword detection; for example, HTML tags removing, insignificant words, and infrequent words

filtering. Beside those approaches mentioned above, a weighed scheme for keyword detection is applied in order to improve the accuracy of classification. According to the experimental results, the proposed approach has 92.8% accuracy rate. The rest of the paper is organized as follows. In section 2, a brief review of present related works of spam email filtering approaches. Section 3 describes a proposed algorithm for content-based spam email filtering. Section 4 presents the experimental results of this work. Section 5 discusses the conclusion of this paper, the future work in spam filtering. Finally, Section 6 lists the references of this paper.

Architecture Diagram



Mathematical model:

- Consider S is a System.
- S={I,P,O}
- Where
  - I= input,
  - P= Procedure
  - O=Output
- Input
  - User=Using the Social Websites.
  - Hacker= Send the Fraud mail.
  - Server= Finding the Bugs.
- Procedure
  - Process1: User uses the Social Media.
  - Process2: Send Mail to user by Hacker.
  - Process3: User opens the fraud website and insert a personal information.
  - Process4: Hacker view all the information of user.
- Output:
  - O=System shows the fraud detection.

Literature survey:

Paper 1. Content Based Spam E-mail Filtering

Author Name: Pingchuan Liu and Teng-Sheng Moh

Description: Currently, E-mail is one in all the foremost necessary methods of communication. However, the increasing of spam emails causes holdup, decreasing productivity,phishing, that has become a heavy drawback for our society. And the variety of spam e-mail is increasing per annum. Therefore, spam e-mail filtering is a vital, meaningful and difficult topic. The aim of this analysis is to seek out associate effective resolution to filter doable spam e-mails. And as we know, in recent days, there are several techniques that spammers use to avoid spam-detection like obfuscation techniques. In this case, the subsequent projected approach uses email content only to create keyword corpus, Alongside some text process to handle obfuscation technique. The rule was evaluated using the CSDMC2010 SPAM corpus dataset that contained 4327 emails within the coaching dataset and 4292 emails within the testing dataset. The

experimental results show that the projected algorithm has ninety-two.8% accuracy.

Screen Shots:

Conclusion:

In this paper, we have a tendency to planned a content-based spam email filtering approach. The system uses keyword-based corpust hat were designed from coaching datasets to classify new incoming email message. so as to enhance the accuracy of our algorithmic rule, we have a tendency to came up with some completely different processes to handle obfuscated, insignificant, or infrequent words. We performed some experiments to guage our proposed work. As are often seen from our results, our proposed algorithm with associate accuracy, recall, false positive rate, false negative rate and f-measure of ninety two.8%, 93.9%, 84.6%, 7.8%,6.1% and 89.1% severally. However, there square measure still some fields we will improve in the future. for instance, presently we have a tendency to square measure solely target the email content, however, there square measure some helpful information we will use within the email header half like sender email address and informatics address, email subject, range of recipients or even time. Beside this, users' selection is additionally a good feature facilitate to discover spam emails. In some cases, even for the best algorithmic rule, the filter will still somehow misclassify some emails. Therefore, the e-mail receivers will get a chance to solve this drawback by distinguishing the e-mail through them. Later on, anti-spam system can keep updating the keywordcorpus or filter strategies based on the feedbacks that collectfrom users.Furthermore, currently, most anti-spam emailapproaches are client-side based filtering approaches.Therefore, all the emails are classified after the email hasalready been sent to the recipient. The sending anddelivering process already wastes the networks and

server'sefficiency. Therefore, if the email can be classified before itis sent to a receiver, it can help to reduce the workload ofboth networks and servers. For instance, a rating system canbe applied to determine if user is spammer or not based onuser historical behavior. The rating system keeps track ofuser behavior and set a threshold that how many emails areclassified as spam in given amount of time. If the number ofspam emails reach or exceed the threshold, system canautomatically either send a warning to customer or freezethis account. Then the workload of networks or servers canbe reduced from spams.

References

[1] 2014 Internet Security Threat Report, Volume19; Availablefrom: http://www.symantec.com/content/en/us/enterprise/other_resourcesstr_main_report_v19_21291018.en-us.pdf.

[2] The economics of Spam, 2012. Available from:http://pubs.aeaweb.org/doi/pdfplus/10.1257/jep.26.3.87.

[3] A.Qaroush, I.M. khater, M. Washaaha, "Identifying spam e-mailbased-on statistical header features and sender behavior,"Proceedings of the CUBE International Information TechnologyConference, pp. 771–778, 2012.

[4] P. Klangpraphant, P. Bhattarakosol, "E-mail authentication system: aspam filtering for smart senders," Proceedings of the 2nd International Conference on Interaction Sciences: InformationTechnology, Culture and Human, pp. 534–538, 2009.

[5] K. N. Junejo, A. Karim, "Automatic Personalized Spam Filteringthrough Significant Word Modeling," Proceedings of the 19th IEEEInternational Conference on Tools with Artificial Intelligence, Vol.2,pp.291–298, 2007.

[6] L. Fu, G. Gali, "Classification Algorithm For Filter Email Spams,"Recent Progress in Data Engineering and Internet TechnologyLecture Notes in Electrical Engineering, Vol 157, pp.149, Springer,2012.

[7] L. Firte, C. Lemnaru, R. Potolea, "Spam Detection Filter using KNNAlgorithm and Resampling," Intelligent Computer Communicationand Processing (ICCP), 2010 IEEE International Conference on, pp.27–33, 26-28 Aug 2010.

[8] http://csmining.org/index.php/spam-email-datasets-.html. Nov 2014.

[9] A. S. Ali, Y. Xiang, "Spam Classification Using Adaptive BoostingAlgorithm," Computer and Information Science, 2007. ICIS 2007.6th IEEE/ACIS International Conference on, pp. 972–976, 11-13July 2007.

[10] M. Sasaki, H. Shinnou, "Spam Detection Using Text Clustering,"Cyberworlds, 2005. International Conference on, pp. 319, 23-25 Nov2005.