Data Analytics and Artificial Intelligence
Vol: 2(5), 2022
REST Publisher; ISBN: 978-81-948459-4-2
Website: http://restpublisher.com/book-series/data-analytics-and-artificial-intelligence/

Spam Emails Filtering

* T. **Angalaeswari, M. Logeswari**
St. Joseph's College of Arts and Science for Women, Hosur, Tamil Nadu, India.
*Corresponding author Email: **angalaeswari5@gmail.com**

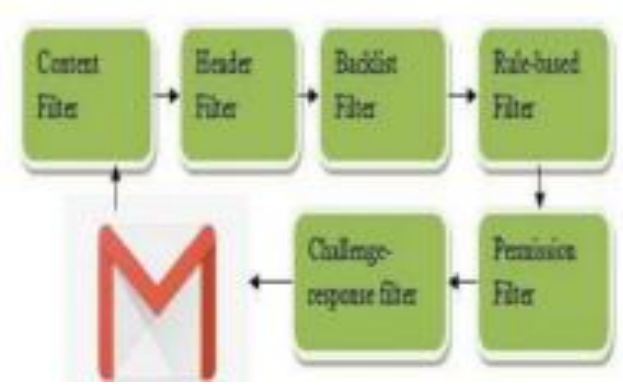
Abstract. The increasing volume of unsolicited bulk e-mail (also known as spam) has generated a need for reliable anti-spam filters. Machine learning techniques now days used to automatically filter the spam e-mail in a very successful rate. Descriptions of the algorithms are presented, and the comparison of their performance on the Spam Assassin spam corpus is presented. - E-mail is one of the most secure medium for online increase in popularity; the number of unsolicited data has also increased rapidly. To filtering data, different approaches exist which automatically detect and remove these untenable messages. There are several numbers of email spams filtering technique such as Knowledge-based technique, Clustering techniques, Learning-based technique, Heuristic processes and so on. This paper illustrates a survey of different existing email spam filtering system regarding Machine Learning Technique (MLT) such as Naive Bays, SVM, K-Nearest Neighbour, Bays Additive Regression, KNN Tree, and rules. However, here we present the classification, evaluation and comparison of different email spam filtering system and summarize the overall scenario regarding accuracy rate of different existing approaches.
Keywords: Spam, E-mail classification, unsolicited bulk email; spam filtering methods; machine learning; algorithm..

1. Introduction

Recently unsolicited commercial bulk e-mail also known as spam, become a big trouble over the internet. Spam is waste of time, storage space and communication bandwidth. The problem of spam e-mail has been increasing for years. In recent statistics, 40% of all emails are spam which about 15.4 billion email per day and that cost internet users about $355 million per year. A specific algorithm is then used to learn the classification rules from these e-mail messages. Machine learning approach has been widely studied and there are lots of algorithms can be used in e-mail filtering. They include Naïve Bayes, support vector machines, Neural Networks, K-nearest neighbour, rough sets and the artificial immune system. Generally, Spam email called as junk email or unsolicited message which sent by spammer through Email. The process is, collected the address on the web and sends the message through domain's username. The most effective and useful email filtering is Spam filtering which performs through antispam technique. As spammers are proactive natures and using dynamic spam structures which have been changing continuously for preventing the anti-spam procedures and thus making spam filtering is a challenging task. Spam filtering is a process to detect unsolicited massage and prevent from entering into user's inbox. Now days, various systems have been existed to generate anti-spam technique for preventing unsolicited bulk email. Most of the anti-spam methods have some inconsistency between false negatives (missed spam) and false positives (rejecting good emails) which act as a barrier for most of the system to make successful antis is system. Therefore, an intelligent and effective spam-filtering system is the prime demand for web users. Machine learning approach has been widely studied and there are lots of algorithms can be used in e-mail filtering. They include Naïve Bays, support vector machines, Neural Networks, K-nearest neighbor, Rough sets and the artificial immune system.

2. Several Email Spam Filtering Methods

At present, number of spam email has increased for several criteria such as an advertisement, multi-level marketing, chain letter, political email, stock market advice and so forth. For restricting spam email, several methods or spam filtering system has been constructed by using various concept and algorithms. This section concluded by describing few of spam filtering methods to understand the process of spam filtering and its effectiveness.



A standard process of Email spam filtering system Standard Spam Filtering Method Email Spam filtering process works through a set of protocols to determine either the message is spam or not. At present, a large number of spams filtering process have existed. Among them, Standard spam filtering process follows some rules and acts as a classifier with sets of protocols. Figure.1 shows that, a standard spam filtering process performed the analysis by following some steps [14]. First one is content filters which determine the spam message by applying several Machines learning techniques [8, 10, 15-18]. Second, header filters act by extracting information from email header. Then, backlist filters determine the spam message and stop all emails which come from backlist file. Afterward, "Rules-based filters" recognize sender through subject line by using user defined criteria [19]. Next, "Permission filters" send the message by getting recipients pre-improvement. Finally, "Challenge response filter" performed by applying an algorithm for getting the permission from the sender to send the mail. Client Side and Enterprise Level Spam Filtering Methods A client can send or receive an email by just one clicking through an ISP. Client level spam filtering provides some frameworks for the individual client to secure mail transmission. A client can easily filter spam through these several existing frameworks by installing on PC. This framework can interact with MUA (Mail user agent) and filtering the client inbox by composing, accepting and managing the messages.

3. Below are some of the most popular machine learning methods

Naïve Bayes classifier: It is a supervised machine learning algorithm where words probabilities play the main rule here. If some words occur often in spam but not in ham, then this incoming e-mail is probably spam. Naïve bays classifier technique has become a very popular method in mail filtering software. Bayesian filter should be trained to work effectively. Every word has certain probability of occurring in spam or ham email in its database. If the total of words probabilities exceeds a certain limit, the filter will mark the e-mail to either category. Artificial Neural Networks classifier: An artificial neural network (ANN), also called simply a "Neural Network" (NN), is a computational model based on biological neural networks. It consists of an interconnected collection of artificial neurons. An artificial neural network is an adaptive system that changes its structure based on information that flows through the artificial network during a learning phase. The ANN is based on the principle of learning by example. Supervised: Here, the network is given a set of inputs and matching output patterns, known as training dataset, to train the network. Unsupervised: In this instance, the network trains itself by producing groups of patterns. There is no earlier set of training data given to the system. Support Vector Machines classifier: Support Vector Machine" (SVM) is a supervised machine learning algorithm that is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is a number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well. They can be easily trained and according to some researchers, they outperform many of the popular email spam classification methods Decision Tree: Decision Tree Classification generates the output as a binary tree like structure called a decision tree, in which each branch node represents a choice between a number of alternatives, and each leaf node represents a classification or decision. Naïve Bayes classifier method Bayesian classifier is working on the dependent events and the probability of an event occurring in the future that can be detected from the previous occurring of the same event [12]. This technique can be used to classify spam e-mails; words probabilities play the main rule here. If some words occur often in spam but not in ham, then this incoming e-mail is probably spam. Naïve bayes classifier technique has become a very popular method in mail filtering software. Bayesian filter should be trained to work effectively. Every word has certain probability of occurring in spam or ham email in its database. K-nearest neighbour classifier method The k-nearest neighbour (K-NN) classifier is considered an example-based classifier, that means that the training documents are used for comparison rather than an explicit category representation, such as the category profiles used by other classifiers. As such, there is no real training phase. When a new document needs to be categorized, the k most similar documents (neighbours) are found and if a large enough proportion of them have been assigned to a certain category, the new document is also assigned to this category, otherwise not. Additionally, finding the nearest neighbours can be quickened using traditional indexing methods.

4. Email spam filtering process

An email message is made up of two major components which are the header and the body. The header is the area that have broad information about the content of the email. It includes the subject, sender and receiver. The body is the heart of the email. It can include information that does not have a pre defined data. Examples include web page, audio, video, analog data, images, files, and HTML markup. The email header is comprised of fields such as sender's address, the recipient's address, or timestamp which indicate when the message was sent by intermediary servers to the Message Transport Agents (MTAs) that function as an office for organizing mails. The header line usually starts with a "From" and it goes through some modification whenever it moves from one server to another through an in-between server. Email Spam Filtering: A Systematic Review surveys current and proposed spam filtering techniques with particular emphasis on how well they work. The primary focus is on spam filtering in email, while similarities and differences with spam filtering in other communication and storage media - such as instant messaging and the Web - are addressed peripherally. Email Spam Filtering: A Systematic Review examines the definition of spam, the user's information requirements and the role of the spam filter as one component of a large and complex information universe. The work reported in this paper was motivated by our belief that to realize an effective personal E-mail filter in the framework of text classification, the following issues should be fully taken into account. An E-mail filter is personalized and the knowledge used by each personal filter is subjective. Therefore, classifying personal E-mail messages is more challenging than using a priori knowledge to filter commercial junk messages that are often characterized by symbols and words like '$', "free", "saving", etc. An in-depth

study on the distinct type of E-mail documents is needed to make full use of the information embedded in them. Feature selection is the key issue. Typical text classification techniques should be examined and compared to enable better understanding of the capabilities and characteristics of these techniques to perform the task of a personal E-mail filter. A relatively large amount of real E-mail data from individuals with different interests should be used in experiments. For the problem of classifying E-mail doc

5. Email spam filtering architecture

Spam filtering is aimed at reducing to the barest minimum the volume of unsolicited emails. Email filtering is the processing of emails to rearrange it in accordance to some definite standards. Mail filters are generally used to manage incoming mails, filter spam emails, detect and eliminate mails that contain any malicious codes such as virus, trojan or malware. The workings of email is influence by some basic protocols which include the SMTP. Some of the widely used Mail User Agents (MUAs) are Mutt, Elm, Eudora, Microsoft Outlook, Pine, Mozilla Thunderbird, IBM notes, K mail, and Balsa. They are email clients that assists the user to read and compose emails. Spam filters can be deployed at strategic places in both clients and servers. Spam filters are deployed by many Internet Service Providers (ISPs) at every layer of the network, in front of email server or at mail relay where there is the presence of firewall [25]. The firewall is a network security system that monitors and manages the incoming and outgoing network traffic based on predetermined security rules. The email server serves as an incorporated anti spam and anti-virus solution providing a comprehensive safety measure for email at the network perimeter [26]. Filters can be implemented in clients, where they can be mounted as add-ons in computers to serve as intermediary between some endpoint devices [27]. Filters block unsolicited or suspicious emails that are a threat to the security of network from getting to the computer system. Also, at the email level, the user can have a customized spam filter that will block spam emails in accordance with some set conditions. **Gmail filter spam** Google's data centers makes use of hundreds of rules to determine whether an email is valid or spam. Every one of these rules depicts specific features of a spam and certain statistical value is connected with it, depending on the likelihood that the feature is a spam. The weighted importance of each feature is then used to construct an equation. A test is conducted using the score against a sensitivity threshold decided by each user's spam filter. And consequently, it is classified as a lawful or spam email. Google is said to be using state of the art spam detection machine learning algorithms such as logistic regression and neural networks in its classification of emails. Gmail also use optical character recognition (OCR) to shield Gmail users from image spam. Also, machine-learning algorithms developed to combine and rank large sets of Google search results allow Gmail to link hundreds of factors to improve their spam classification. The evolving nature of spam over time revolves around factors such as domain reputation, links in message headers and others. These can make messages to unexpectedly end up in the spam folder. **Yahoo mail filter spam** Yahoo mail is the first free webmail providers in the world with over 320 million users. The email provider has its own spam algorithms that it uses to detect spam messages. The basic methods used by Yahoo to detect spam messages include: URL filtering, email content and spam complaints from users. Unlike Gmail, Yahoo filter emails messages by domains and not IP address. Yahoo mail uses combination of techniques to filter out spam messages. It also provide mechanisms that prevent a valid user from being mistaken for a spammer. Examples are ability of the users to troubleshoot SMTP Errors by referring to their SMTP logs.

Outlook email spam filter After Gmail and Yahoo mail, we discussed Outlook from Microsoft in this section and how it handles spam filtering. In 2013, Microsoft changed the name of Hotmail and Windows Live Mail to Outlook.com. Outlook.com was patterned after Microsoft's Metro design language and directly imitates the interface of Microsoft Outlook. Outlook.com is a collection of applications from Microsoft, one of which is Outlook webmail service. Outlook webmail service allows the users to send and receive emails in their web browser. It allows the users to connect cloud storage services to their account so that when they want to send an email with file attachments, they can select files from not only their computer and One Drive account but also from Google Drive, Box, and Drop box account. Detailed algorithm steps Step 1: Email pre-processing: The content of email is received through our software, the information is extracted then as mentioned above, then the information (Feature) extracted is saved into a corresponding database. Every message was converted to a feature vector with 21700 attributes (this is approximately the number of different words in all the messages of the corpus). An attribute n was set to 1 if the corresponding word was present in a message and to 0 otherwise. This feature extraction scheme was used for all the algorithms. Step 2: Description of the feature extracted Feature extraction module extract the spam text and the ham text, then produce feature dictionary and feature vectors as input of the selected algorithm, the function of feature extraction is to train and test the classifier [9]. For the train part, this module account frequency of words in the email text, we take words which the time of appearance is more than three times as the feature word of this class. And denote every email in training as a feature vector. Step 3: Spam classification Through the steps above, we take standard classification email documents as training document, pre-treatment of email, extract useful information, save into text documents according to fix format, split the whole document to words, extract the feature vector of spam document and translate into the form of vector of fix format. We look for the optimal classification using the selected algorithm which is constructed using the feature vector of spam documents. Step 4: Performance evaluation In order to test the performance of above mentioned six methods, we used the most popular evaluation methods used by the spam filtering researchers. Spam Precision (SP), Spam Recall (SR), Accuracy (A). Spam Precision (SP) is the number of relevant documents identified as a SP = # of Spam Correctly Classified_ Total # of messages classifies as

6. Summary of Existing E-mail Spam Classification Approaches

Since last few decades, researchers are trying to make email as a secure medium. Spam filtering is one of the core features to secure email platform. Regarding this several types of research have been progressed reportedly but still there are some untapped potentials. Over Stime, still now e-mail spam classification is one of the major areas of research to bridge the gaps. Therefore, a large number of researches already have been performed on email spam classification us using several techniques to make email more efficient to the users. The two common approaches used for filtering spam mails are knowledge engineering and machine learning. Emails are classified as either spam or ham using a set of rules in knowledge engineering. The person using the filter, or the software company that stipulates a specific rule-based spam filtering tool must create a set of rules. Using this method does not guarantee efficient result since there is need to continually update the rules. This can lead to time wastage and it is not suitable especially for naïve users. . Machine learning field is a subfield from the broad learning one tries to uncover hidden regularities (clusters) or to detect anomalies in the data like spam messages or network intrusion. In e-mail filtering task some features could be the bag of words or the subject line analysis. Thus, the input to e-mail classification task can be viewed as a two dimensional matrix, whose axes are the messages and the features. E-mail field of artificial intelligence, this aims to make machines able to learn like human. Learning here means understood, observe and represent information about some statistical phenomenon. In unsupervised classification tasks are often divided into several sub-tasks. First, Data collection and representation are mostly problem specific (i.e. e-mail messages), second, e-mail feature selection and feature reduction attempt to reduce the dimensionality (i.e. the number of features) for the remaining steps of the task. Finally, the e-mail classification phase of the process finds the actual mapping between training

7. Performance evaluation measures

Spam filters are usually evaluated on large databases containing ham and spam messages that are publicly available to users. An example of the performance measures that are used is classification accuracy (Acc). It is the comparative number of messages rightly classified; the percentage of messages rightly classified is used as an added measure for evaluating performance of the filter.

8. Conclusion

In this paper we review some of the most popular machine learning methods and of their applicability to the problem of spam e-mail classification. Descriptions of the algorithms are presented, and the comparison of their performance on the Spam Assassin spam corpus is presented, the experiment showing a very promising results specially in the algorithms that is not popular in the commercial e-mail filtering packages, spam recall percentage in the six methods has the less value among the precision and the accuracy values, while in term of accuracy we can find that the Naïve bayes and rough sets methods has a very satisfying performance among the other methods, more research has to be done to escalate the performance of the Naïve bayes and Artificial immune system either by hybrid system or by resolve the feature dependence issue in the naïve bayes classifier, or hybrid the Immune by rough sets. Finally hybrid systems look to be the most efficient way to generate a successful antispam filter nowadays.

References

[1]. M. Awad, M. Foqaha Email spam classification using hybrid approach of RBF neural network and particle swarm optimization Int. J. Netw. Secur. Appl., 8 (4) (2016)

[2]. D.M. Fonseca, O.H. Fazzion, E. Cunha, I. Las-Casas, P.D. Guedes, W. Meira, M. Chaves Measuring haracterizing, and avoiding spam traffic costs IEEE Int. Comp., 99 (2016).

[3]. Visited on May 15, 2017 Kaspersky Lab Spam Report (2017) 2012 https://www.securelist.com/en/analysis/204792230/Spam_Report_April_2012

[4]. E.M. Bahgat, S. Rady, W. Gad An e-mail filtering approach using classification techniques The 1st International Conference on Advanced Intelligent System and Informatics (AISI2015), November 28-30, 2015, Springer International Publishing, BeniSuef, Egypt (2016), pp. 321-331 View PDFCrossRefView Record in Scopus

[5]. N. Bouguila, O. Amayri A discrete mixture-based kernel for SVMs: application to spam and image categorization

[6]. Awad, W. A., & ELseuofi, S. M. (2011). Machine Learning methods for E mail Classification. International Journal of Computer Applications, 16(1).

[7]. Saad, O., Darwish, A., & Faraj, R. (2012). A survey of machine learning techniques for Spam filtering. International Journal of Computer Science and Network Security (IJCSNS), 12(2), 66.

[8].Chen, Y., Jain, S., Adhikari, V. K., Zhang, Z. L., & Xu, K. (2011, April). A first look at inter-data center traffic characteristics via yahoo!datasets. In INFOCOM, 2011 Proceedings IEEE (pp. 1620-1628). IEEE

