Chapter 15
Spam Mail Filtering Using Data Mining Approach:
A Comparative Performance Analysis

Ajay Kumar Gupta
https://orcid.org/0000-0001-9666-5047
Madan Mohan Malaviya University of Technology, India

ABSTRACT

This chapter presents an overview of spam email as a serious problem in our internet world and creates a spam filter that reduces the previous weaknesses and provides better identification accuracy with less complexity. Since J48 decision tree is a widely used classification technique due to its simple structure, higher classification accuracy, and lower time complexity, it is used as a spam mail classifier here. Now, with lower complexity, it becomes difficult to get higher accuracy in the case of large number of records. In order to overcome this problem, particle swarm optimization is used here to optimize the spam base dataset, thus optimizing the decision tree model as well as reducing the time complexity. Once the records have been standardized, the decision tree is again used to check the accuracy of the classification. The chapter presents a study on various spam-related issues, various filters used, related work, and potential spam-filtering scope.

INTRODUCTION

SPAM (Attri, 2012) is one of the electronic messaging systems which includes most broadcast media through which it sends or receives the unsolicited messages on the computer, mobile or PDA etc. indiscriminately. Junk e-mail (E-mail spam), is a subset of spams that involves approximately same e-mail messages transmitted to no. of recipients. Spam (Attri, 2012) is use of electronic messaging system to send unsolicited bulk messages indiscriminately. When the number of messages in your inbox started to increase, it became annoying for us to remove the unwanted e-mail. IE- mail spam is also known as unsolicited bulk e-mail (or junk e-mail). The current survey shows an increasing trend for amount of incoming spam and scammer attacks are becoming targeted, and consequently more of a threat. When

DOI: 10.4018/978-1-7998-2491-6.ch015

targeted attacks first emerged five years ago, Symantec message labs intelligence tracked between one or two attacks per week. Subsequently, attacks have increased to 10 per day to 60 per day in 2010. The number of spam sent by the countries of Europe will increase to 40 percent to 45 percent of all spam. These facts state that the spam is a big problem for today and also for tomorrow and it actually makes sense to investigate new effective methods against spam. The purpose of this work is to discover the techniques to filter the spam from incoming emails. Filtering spam is a technique to categorize all the incoming emails in network into spam and ham messages. Here, important issues related to spam filtering, the applicable steps for classification, methods and the evaluation measures in the spam filtering are discussed in detail. A lot of works have been done before in this spam filtering domain. These include Bayesian Networks, Decision Tree, K-Nearest Neighbor etc. (Ma, 2009), (Razmara, 2012) with some extra features or with some additional methods in it. With advancement, Spammers frequently change their email's external sign to misguide spam filtering systems, so, there arises a need for adaptive filtering systems, which have the power of quick reaction to the changes and provides fast and qualitative self-tuning with a new set of features. The study so far concludes that there are many of the filtering techniques which are based on text categorization methods but none of them can claim to provide an ideal solution i.e. zero percent false positive and zero percent false negative. Still, there are lots of scopes for research in classifying text messages as well as multimedia messages. This is not possible to maintain 100% accuracy and efficiency of filtering spam. But, one should try to make sure that the model is more efficient, reliable and accurate as possible. Classifier should avoid the following two cases to be more accurate.

- **Ham Misclassification:** The genuine mail should not be classified as a spam mail. Due to this misclassification, the receiver may get unaware of important mails which may be very damaging sometimes by causing serious risks.
- **Spam Misclassification:** The spam should not be classified as important mails as it causes many more financial and behavioral damage.

Process of Spam Filtering

A spam may be of different forms as image spam, blank spam, sms spam, email spam etc. The spam mail usually contains advertisement contents. As per common aspect, the filter focuses on the modules of emails to primarily classify the spam and hams. On that basis the spam, filters are of 3 types on the strategy of focusing on emails to classify spam.

1. Subject of message
2. Body content of message (content based filter)
3. Senders status (sender's reputation based on past history as spammer or not)

A general machine learning based spam filter (Zhong, 2010) consists of at least the following sequences.

1. **Collection of Emails:** First of all, all the network emails are collected from individual users which are considered as both spam and legitimate email.
2. **Pre-Processing:** The next is the transformation process. In this phase, the task of pre-processing is usually defined by the author what strategies she/he is using. Generally, it consists of removal of conjunctions, stop words etc. It also has tokenization process in it.