Chapter 18

Using Data Mining Techniques and the Choice of Mode of Text Representation for Improving the Detection and Filtering of Spam

Reda Mohamed Hamou  
Dr. Moulay Taher University of Saïda, Algeria

Abdelmalek Amine  
Dr. Moulay Taher University of Saïda, Algeria

ABSTRACT

This chapter studies a boosting algorithm based, first, on Bayesian filters that work by establishing a correlation between the presence of certain elements in a message and the fact that they appear in general unsolicited messages (spam) or in legitimate email (ham) to calculate the probability that the message is spam and, second, on an unsupervised learning algorithm: in this case the K-means. A probabilistic technique is used to weight the terms of the matrix term-category, and K-means are used to filter the two classes (spam and ham). To determine the sensitive parameters that improve the classifications, the authors study the content of the messages by using a representation of messages by the n-gram words and characters independent of languages to later decide what representation ought to get a good classification. The work was validated by several validation measures based on recall and precision.

INTRODUCTION

Unsolicited email, or spam, is a wound of electronic communication currently can represent up to 95% of the volume of mail processed on some servers. This is an ethical and economic issue of role that effectively fight against this scourge. Although the decision “spam / non-spam” is most often easy to take for a human. A message in circulation prevents address manually sorting the acceptable mail and

DOI: 10.4018/978-1-4666-7272-7.ch018
other. Spam is a global phenomenon and massive. According to the CNIL (The Commission Nationale de L'Informatique and Freedoms), spam is defined as follows: “The” spamming “or” spam “is to send massive and sometimes repeated, unsolicited emails, to individuals with whom the sender has never had contact, he has captured the email address erratically. “It was not until the late 90s that the problem of detection and spam filtering by content, drew attention to three areas of research that were not directly affected by e-mail: the information retrieval, data mining and machine learning.

Detection and spam filtering is a binary classification problem in which the email is classified as either Ham or spam. This area has experienced a wide range of methods for the classification among its techniques; the use of Bayes’ theorem is the most famous. Bayesian filters work by establishing a correlation between the presence of certain elements (usually words) in a message and the fact that they usually appear in messages (spam) or in legitimate email (ham) to calculate the probability that the message is spam. Bayesian filtering spam is a powerful technique for the treatment of unwanted email. It adapts to the habits of mail each other and produces a false positive rate low enough to be acceptable.

This problematic leads us to do a study as to the representation of data (message corpus) to try to identify sensitive parameters that can improve the results of classification and categorization around detections and spam filtering. We know very well that supervised learning techniques yield the best results, it is for this reason that we have tried to inject a clustering algorithm (k-means) to try to minimize the intervention of expert.

STATE OF THE ART

Among the anti-spam techniques exist in the literature include those based on machine learning and those not based on machine learning.

The Techniques Not Based on Machine Learning

Heuristics, or rules-based, the analysis uses regular expression rules to detect phrases or characteristics that are common in spam, and the amount and severity of identified features will propose the appropriate classification of the message. The history and the popularity of this technology has largely been driven by its simplicity, speed and accuracy. In addition, it is better than many advanced technologies of filtering and detection in the sense that it does not require a learning period. Techniques based on signatures generate a unique hash value (signature) for each message recognized spam. Filters signature compare the hash value of all incoming mail against those stored (the hash values previously identified to classify spam e-mail). This kind of technology makes it statistically unlikely that a legitimate email to have the same hash of a spam message. This allows filters signatures to achieve a very low level of false positives. The blacklist is a technique that is simple common among almost all filtration products. Also known as block lists, blacklists filter e-mails from a specific sender. White lists, or lists of authorization, perform the opposite function, to correctly classify an email automatically from a specific sender. Currently, there is a spam filtering technology based on traffic analysis provides a characterization of spam traffic patterns where a number of attributes per email are able to identify the characteristics that separate spam traffic from non-spam traffic.