BAYESIAN FILTER TECHNIQUE FOR SPAM E-MAIL DETECTION: AN OVERVIEW

Sayali Wavhal

B.Tech Electrical, Veermata Jijabai Technological Institute, Matunga, Mumbai

Email: sayali.wavhal@gmail.com

Abstract— As web is expanding day by day and people generally rely on web for communication, e-mails are the fastest way to send information from one place to another. E-mail is an effective tool for communication as it saves a lot of time and cost. Spam, also known as Unsolicited Commercial E-mail, is an unfortunate problem on the Internet. Spam increases the load on the servers and the bandwidth of the Internet Service Providers and the added cost to handle this load must be compensated by the users. Mailbox management has become a big task because these unwanted emails clog the inbox. In this paper, we review one of the most popular Machine Learning methods using text categorization, Bayesian filter for Spam E-mail Detection

Keywords— Spam, unsolicited, text categorization, machine learning, spam e-mail detection, Bayesian filter.

INTRODUCTION

The e-mailboxes of millions of people are cluttered with unsolicited bulk email known as 'spam'. Many email spam messages may also contain disguised links that appear to be for familiar websites but in fact lead to phishing web sites or sites that are hosting malware. Being incredibly cheap to send, spam causes a lot of trouble to the Internet community: large amounts of spam-traffic between servers cause delays in delivery of legitimate email, people with dialup Internet access have to spend bandwidth downloading junk email. Sorting out the unwanted messages takes time and introduces a risk of deleting legitimate mail by mistake. According to a Cyberoam report in 2014, there are an average of 54 billion spam messages sent every day. Although, it is commonly believed that a change in Internet protocols can be the only effective solution to the spam problem, it is acknowledged that this cannot be achieved in a short time. In recent years, anti-spam filters have become necessary tools to face up the continuously growing spam phenomenon.

SPAM FILTERS

There are two general approaches to mail filtering: knowledge engineering and machine learning.

- Knowledge Engineering (KE): A set of rules is created according to which messages are categorized as spam or legitimate mail. Some of the methods following this approach are as follows:
  1. Domain filter: They allow mails from specific domain only. Keeping track of domains that are valid for the user is cumbersome.
  2. Blacklisting filters: They use the database of known abusers and filter unknown addresses as well. This requires constant updating of the database.
  3. Whitelist filters: Mailer programs learns all the contacts of a user and let mail from those contacts through. This means everyone should first communicate his/her e-mail ID to the user and only then send email.

  The major drawback of this method is that the set of rules must be constantly updated, and maintaining it is not convenient for most users.
- Machine Learning (ML): The machine learning approach does not require specifying any rules explicitly. Instead, a set of pre-classified documents is needed. A specific algorithm is then used to "learn" the classification rules from this data.

Use of text categorization techniques based on machine learning and pattern recognition approaches for email semantic content analysis is an effective way of spam detection compared to the knowledge engineering approach. In this work, we focus on Bayesian spam filters based on textual content analysis. The advantages of these techniques are the automatic construction of classification rules and their potentially higher generalisation capability with respect to manually encoded rules.

TEXT MINING

With regard to the analysis of the semantic content of e-mails, several researchers in recent years have investigated text categorisation techniques based on the machine learning and pattern recognition approaches due to their potentially higher generalization capability.

The first step, named tokenization, consists of extracting a plain text representation of document content. Prepare a corpus, which is a collection of all documents. Inspect the corpus and identify the real words. Different text may contain “Hello!”, “Hello,” “hello…” etc. for example. We would consider all of these the same. Clean the corpus by translating all letters to lower case, remove numbers, punctuation and non-content words like “I”, “me”, “you” etc. and finally remove the excess white space. In the end, tokenize the corpus. A token is a single element in a text string, in most cases a word. From here, Naïve Bayes classifier is used to build a spam filter based on the words in the message.

NAÏVE BAYES CLASSIFIER

Bayesian filters, considered the most advanced form of content-based filtering, employ the laws of mathematical probability to determine which messages are legitimate and which are spam. Naive Bayes classifiers work by correlating the use of tokens, with spam and non-spam e-mails and then using Bayes' theorem to calculate a probability that an email is or is not spam.

Particular words have particular probabilities of occurring in spam email and in legitimate email. The filter does not know these probabilities in advance, and must first be trained so it can build them up. To train the filter, the user must manually indicate whether a new email is spam or not. For all words in each training email, the filter will adjust the probabilities that each word will appear in spam or legitimate email in its database. After training, the word probabilities (also known as likelihood functions) are used to compute the probability that an email with a particular set of words in it belongs to either category. Each word in the email contributes to the email's spam probability, or only the most interesting words. This contribution is called the posterior probability and is computed using Bayes' theorem. Then, the email's spam probability is computed over all words in the email, and if the total exceeds a certain threshold (say 95%), the filter will mark the email as a spam. Bayesian email filters utilize Bayes' theorem. Bayes' theorem is used several times in the context of spam.

- **Computing the probability that a message containing a given word is spam:**
  Let’s suppose the suspected message contains the word “X” is likely to be a spam.
  The formula used by the software to determine that is derived from Bayes’ theorem,

$$\Pr(S|W) = \frac{\Pr(W|S)\cdot\Pr(S)}{\Pr(W|S)\cdot\Pr(S)+\Pr(W|H)\cdot\Pr(H)}$$

Where :
Pr(S|W) is the probability that a message is a spam, knowing that the word “X” is in it;
Pr(S) is the overall probability that any given message is spam;
Pr(W|S) is the probability that the word “X” appears in spam messages;
Pr(H) is the overall probability that any given message is not spam(is ham);
Pr(W|H) is the probability that the word “X” appears in ham messages.

Of course, determining whether a message is spam or ham based only on the presence of the word "X" is error-prone, which is why Bayesian spam software tries to consider several words and combine their spamicities to determine a message's overall probability of being spam. Most reports have shown that Bayesian filters works correctly over 99 percent for one user.

SHORTCOMINGS OF BAYESIAN SPAM FILTERING

Depending on the implementation, Bayesian spam filtering may be susceptible to Bayesian poisoning, a technique used by spammers in an attempt to degrade the effectiveness of spam filters that rely on Bayesian filtering. A spammer practicing Bayesian poisoning will send out emails with large amounts of legitimate text (gathered from legitimate news or literary sources). Spammer tactics include insertion of random innocuous words that are not normally associated with spam, thereby decreasing the email's spam score, making it more likely to slip past a Bayesian spam filter. Another technique used to try to defeat Bayesian spam filters is to replace text with pictures, either directly included or linked. The whole text of the message, or some part of it, is replaced with a picture where the same text is "drawn". The spam filter is usually unable to analyze this picture. It is worth pointing out that this trick is often used in phishing e-mails, which are one of the most harmful kinds of spam.

SPAM FILTERING FOR TEXT EMBEDDED WITH IMAGES

Although the effectiveness of text categorisation techniques could be affected by tricks used by spammers for content obscuring, spam e-mails containing such kinds of tricks can be identified by other modules of spam filter, for instance by performing lexicographic analysis or analysis of syntactic anomalies. Carrying out semantic analysis of text embedded into images attached to e-mails first requires text extraction by Optical Character Recognition (OCR) techniques.

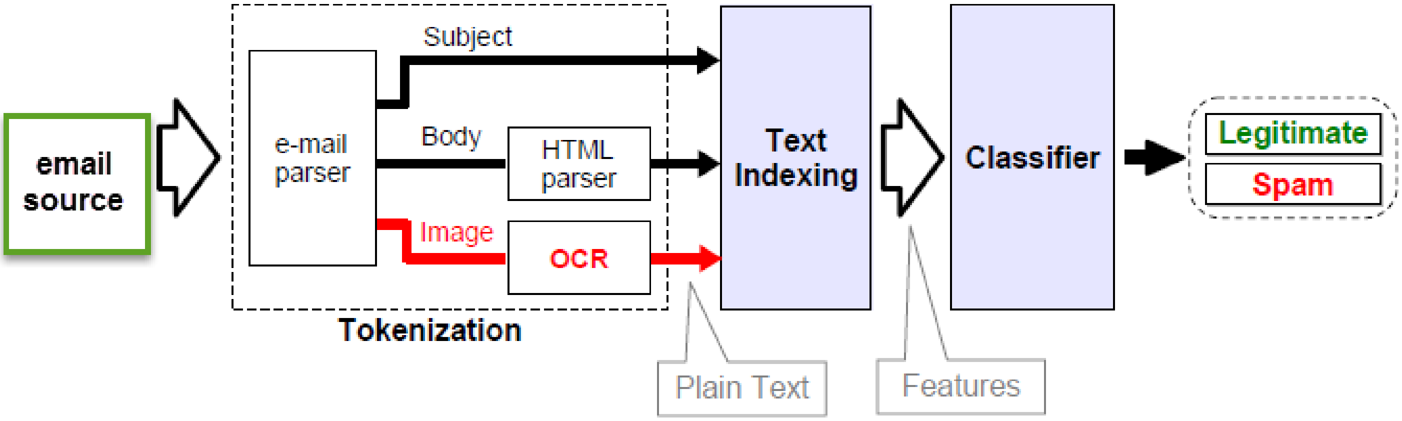

Figure: Block diagram for spam filter with text information embedded into images

The text contained in the body of the e-mail and that embedded into images can be viewed simply as a different "coding" of the message carried by an e-mail. Accordingly, semantic analysis of text embedded into images using text categorization techniques is similar to the ones applied to the body of the e-mail. The basic idea is to extend the phase of tokenization in the document processing steps by including plain text extraction from attached images, as well as from the subject and body fields.

However, it should be taken into account that a vocabulary in which clean digital text is mixed with noisy text (due to OCR) could affect the generalization capability of a text classifier. To avoid including spurious terms generated by OCR noise in the vocabulary, only the terms coming from the subject and body fields could be used to create it. Terms extracted from images can instead be used only at the indexing phase, when the feature vector representation of e-mails is constructed. Consider now the indexing phase. For e-mails containing text embedded into images, a possible choice is to use both the terms belonging to such text and the ones belonging to the subject and body fields to compute the feature vector. However, if terms belonging to images attached to training emails are not included in the vocabulary, it could be better not to use them even for indexing training e-mails. The rationale is again to avoid that OCR noise affects the generalization capability of the text classifier. In this case, the whole training phase of the text classifier would be carried out without taking into account text extracted from images. Such text would be used only for indexing testing e-mails at the classification phase.

Indexing of testing e-mails can also be performed in different ways, to take into account that in spam e-mails with attached images the whole spam message is often embedded into images, while the body field contains only bogus text or random words. One possibility is the following: if an e-mail does not contain attached images, its feature vector is computed as usual from the text in the subject and body fields; otherwise it is computed using only text extracted from attached images. In other words, text in the subject and body is disregarded at classification phase, if the e-mail has text embedded into an attached image. A more complex strategy can also be used: both the above feature vectors can be computed, namely one taking into account only terms in the subject and body fields, and the other one taking into account only terms in the text extracted from images. These two feature vectors are then independently classified, and the two classification outcomes (either at the score or at the decision level) are then combined either within the considered module of the spam filter to yield a single decision for that module, or at a higher level outside that module. This strategy could be effective if the spam message is often (but not always) carried only by text embedded into images. For instance, the maximum of the two scores could be taken, assuming that the text classifier is trained to give higher scores to spam e-mails.

CONCLUSION

Whatever new filtering capabilities arise, it is just a matter of time before spammers find ways to evade them. Because of this text distortion and image spam, spam filtering are not just simple text classification and information retrieval problems anymore. This paper represents a critical analysis of spam e-mail filtering and how Bayesian filter can adapt to new distortion patterns to develop new techniques for spam filtering is addressed.

REFERENCES:

[1] Konstantin Tretyakov, "Machine Learning techniques in Spam Filtering", Data Mining Problem-oriented Seminar, MTAT.03.177, May 2004, pp. 60-79.

[2] Giorgio Fumera, Ignazio Pillai, Fabio Roli, "Spam Filtering Based On The Analysis Of Text Information Embedded Into Images", Journal of Machine Learning Research 7 (2006) 2699-2720.

[3] Eric Chen, “Data Mining Applied to Email SPAM Detection and Filtering”, CS445 Hw1

[4] https://en.wikipedia.org/wiki/Naive_Bayes_spam_filtering