Global Journals LaTeX JournalKaleidoscope™
Artificial Intelligence formulated this projection for compatibility purposes from the original article published at Global Journals. However, this technology is currently in beta. *Therefore, kindly ignore odd layouts, missed formulae, text, tables, or figures.*

Bayesiane Filter for Detecting a Spam

Dr. Elma Zanaj[1], Dr. Elma Zanaj[2] and Bledi Shkurti[3]

[1] Faculty of Information Technology, Polytechnic University of Tirana.

Received: 16 December 2011 Accepted: 2 January 2012 Published: 15 January 2012

Abstract

The detected of spam messages in terms that better having a spam email in the inbox than a ham message in the junk, has been investigated recently. The main contribution of the paper consists in comparing three antispam filters used more nowadays, and will find that which is filter is of the future. By using filters we will also create some patterns as the result of training with different number of emails. Simulations show that due to the trainging of the filters it will be easier to detect the spams.

Index terms— filter; Bayesian; spam; ham.

1 INTRODUCTION

pam emails or otherwise UCE (Unsolicited Commercial E-mail) have no exact definition. Most spam emails can be considered as undesirable but not all unwanted emails are spam [1]. A different name may be unwanted commercial emails, but unfortunately there are just spam advertising messages. Fig. 1 shows the set of all email messages and the place that spam occupy. In the email context, the term spam refers to the electronic equivalent of junk email: a set of unwanted messages. Many definitions present spam as messages sent to many people without the consent of recipients. UBE (Unsolicited Bulk Email) are messages sent to many addresses without their request and approval. Examples of UBE are political or religious emails, hate messages for different groups or fraud on the Internet. UBE includes UCE, and then commercial messages sent to many persons can be considered UBE and UCE. While we refer to messages that are not spam as ”not spam” or ”legitimate messages”, a fitting name could be ”ham”.

The history of spam can be divided into three different parts: ? The early spam letters addressed and sent manually

? The sender that use machines, which led to a dramatic increase in the number of spam ? The last part is when ”machine learning” came for filtering spam and made filtering more effective.

The use of Bayesian networks, does filtering not only more effective, but it also learns the characteristics of the message received. Even if spam senders do a lot of new tricks, Bayesian filters have good chances to filter them. Fig. 2 shows the results of the questionnaire issued by the spam to a user group **??**2].Statistics in December of 2010 show that 33% of Internet users do not like spam and 52% of them not only dislike but are frustrated of them, while only 15% of users do not have those problems. It is worth noting that 85% of users have problems with spam.

2 Irritated

If a company has no filter and the employees receive 6 spam messages per day, it must require at average 5 seconds reading and deleting each spam message, which means that the worker will spend 3 hours a year to read and delete spam. These estimates are prudent and do not take into account the time required for discussing the spam and their differences, or the contacting of the specialists for other problems caused by spam.

This paper will present a study, on the basis of which a filter can be selected, so less spams can be in our inbox.

The remainder of the paper is organized as follows. In Section 2, are presented some statistics and fact about the spam. Section 3 outlines some methods of fighting the spams, while Section 4 describes the three filters that we are going to use.Our simulation results are presented in Section 5, and the Conclusions concludes the paper.

3 STATISTICS AND FACTS ABOUT SPAM

Today are counted 14.5 billion spam per day distributed in whole world **??**3]. We can say that spam are 45% of all emails. In fact some companies have estimated that spam occupy a high percentage of all global email communications, such as 73%. United States of America are the first to send spam, with the largest number, followed by Korea comes as second distributor of unwanted messages. Most used type of spam is online advertising, which occupy 36% of all spam. The second largest are links to adult which constitute 31.7% of all spam. Unwanted emails relating to financial matters occupy third place with 26.5% of spam. In fact, observations indicate that spam has reduced public confidence towards online communications. More people have completely lost confidence in Internet communications because of spam.

Companies look at spam as a problem that reduces productivity and safety. About 52% of the interviewed companies rated the fight against spam as most needed and as the most important. Regarding a study done by "Radicati Research Group", a research company in California, spam cost to the business 20.5 billion $ annually in the reduction of productivity and technical expenses. Another study says that the average annual loss for an employee is approximately 1934$. Future forecast for the costs of spam are not very good, it is provided that 58 billion spam emails will be sent every day and that within the next 8 years, that overdrafts will cost to the businesses 198 billion $ every day. The number of spam emails depends on frequency of use of email. For example, if someone receives 10 emails every week, and has only a spam, then to him is simple and easy to delete unwanted email than to implement a filter for his emails. Spam usually causes problems for users who receive hundreds of emails a week, for who is really annoying to delete hundreds of spam emails every week. One of the characteristics and trends can be the language [4]. Most of the spams are written in English. Another characteristic is their hour of delivery. The time when a spam is sent, is during working hours in the United States of America. In the area of Washington DC, New York obtained an average of 50% spam over time 8 am to 2 pm, that in other times of the day. All these statistics can serve as a basis for solutions antispam filters. Spam are: advertising, finance, Phinishing.

4 III.

5 SIMPLE METHODS TO FIGHT SPAM

Some very simple methods are discovered during the evolution of spam filters. These are ideal to combine with more complex filters. Greater efficiency can be achieved by using small pieces together of these filtration methods:

1) The key words. The first chooice is to look for key words in the message subject.

2) Black list. It is necessary to make the difference between the two levels of the black list: the black list of network-level and black list at the address.

3) White List. These are the opposite of black lists.

Content filtering identifies spam, and white lists require the identification of users. A white list is a safe community contacts (not dangerous to send spam).

If the method of the black list and white list are used together, would necessitate a different filtering to find addresses that are not in any of the lists referred above. 4) Throttling. Throttling only slows down the speed at which a network or a machine can send traffic. In fact this is the most effective ways to fight spam. 5) Filtering in cooperation. This will allow individuals to communicate in reliable groups for infected messages with other group members.

6 6) Network filtering. Protocol Simple Mail Transfer

Protocol (SMTP) is the way that email servers communicate. This protocol was designed to function independently and to ensure the privacy of Internet users. Senders of spam have benefited from this protocol of servers to send spam emails anonymously. In fact, authenticated SMTP thought to be an answer to spam, but was seen to be necessary only to identify legitimate email senders. Authenticated SMTP requires to the users to give their password before sending email. Many spamsending today build their email servers and keep in non specious networks to send their emails and passing the required authentication during delivery [1]. SMTP provides several other opportunities for later use.

7) Fake Worker. The main idea of this solution is to create a fake email address will be used to set "traps" . This can be used especially with names of companies (eg.: xyz@company.com) [5]. 8) Project Honey Pot. Project Honey Pot is the first system capable to identify the senders of spam and robot programs, which are used to gather email addresses from web sites [6]. To participate in the project Honey Pot, web page builders must install the program on the server where the page is held. The rest is managed automatically. 9) SPF (Sender Policy Framework) [7]. The issue is the falsification of address of the sender. Falsification of the address of the sender is a problem for simple users or companies. It even reduces the trust in email communications because it reduces user's confidentiality and their trust.

To filter a spam, it starts with getting the email that is addressed to a particular receiver. The message is sent to antispam filter that classifies mail as spam or non-spam by using several methods. If the filter is classified ham (legitimate email) it is sent to the directory E ham message (inbox) in which the user receives his emails. Filters should be able to continuously update. Filters updating strategies are as diverse as filtration methods. The updates are two types: manual and selflearning (training). A manual update involves a person changing the filter parameters. In updating self-learning (training), the filter will find in the contents of the mail, the pieces that show for a spam or non-spam emails, from emails that are previously classified. The time that the filter can be updated, is every month or may be more often (after every new email). Online updating, the messages are processed and classified one after another. Before processing the next message, the filter can be adopted based on information received, and this is called training. When a filter is trained by more than one message at the same time, this process is called multiple training.

Classification of spam is based on Bayes theorem that establishes the link between conditional propabilities of two events [9]. Bayes theorem provides a way to calculate the probability of the hypothesis, when the event Y takes the training data, which appear with X:(/) (|) * ( ) / ( ) P Y X P X Y P X P Y = (**1**)

This theorem is the basis of statistics Bayesiane, which compute the probability of a new event based on other propabilities calculated before. When we test a new email message the starting point is a null hypothesis: "the message is spam", the alternative hypothesis is: "the message is not spam" [9].To classify a message as spam, is created the frequency distribution of their components and is compared with previous records of training (the corpus of spam) with a statistical test. Statistical tests will provide a probability value p , and if it is lower than a significance level, the null hypothesis is discarded and is verified that the message is not spam. Otherwise the null hypothesis is accepted. The probability of the Bayesiane statistics of a model based on data is calculated differently from classical statistics that estimate the probability of data by providing a hypothesis [8]. Bayesian classifier calculates the probability that a message is spam.

IV.

7 THE CHOICE OF ANTISPAM FILTER

In our experiment we will compare antispam filters used more nowadays, and will find that which is filter is of the future. By using filters we will also create some patterns as the result of training with different number of emails. For all antispam solutions that use Bayesian network, we must first train the filters. To compare the filters is important not only to find a way for training, but that all filters must be trained in the same manner and with the same email messages. So the emails used can be classified before.

During the training process, the values of the particles differ in particle vocabulary of the filter, to obtain a high accuracy. This makes spam messages known from filters.

We should note that the training process does not end after the conclusion of training, it continues during the filter testing. Moreover, the filter becomes more personalized and closer to our needs during its use. For testing and comparison purposes we will use some programs that are convenient. Having met the requirements of users, we will compare the programs that are free. Although there are many different filters, because of limits for testing, we will compare three of them that are used more nowadays. The filters are:

1) SpamAssassin, an open source program and the most famous and widely used [9]. It is among the most effective filters, especially when used with databases of spams. SpamAssassin is a spam filter based on a set of rules to identify spams. Each rule checks the emails to assess whether it is spam or not. When all rules are applied, the amount is compared with the threshold set by the user. A number greater or equal to the limit means that the message is spam, and a smaller number than the threshold indicates that email is ham. First to make the classification of spam or ham, we have to train SpamAssassin filter. So the filter known the characteristices of the messages and creates a decision database, which will be used in the testing phase. So we have to train regularly with new messages to keep updated with messaging characteristices. We must train with the spam and ham messages. Training only with one category will enable the filter to recognize messages. To make a good training, it is recommended to use 1000 ham and 1000 spam messages, if it is possible. From the tests performed till now, the training with more than 5000 messages does not make a difference to accuracy.

2) Mozilla Thunderbird. The second program that we will use is Mozilla Thundebord. Mozilla's products have filter very well implemented and designed **??**10]. The filter has an automatic opportunity to be trained, and learns quickly from training, giving positive results in many everyday situations. Thunderbird included a Bayesian filter, a white list and may make classifications effectively as a filter SpamAssassin on a server.

3) SpamProbe. The third program is SpamProbe, another open source filter, which is a statistical spam filter [11]. It have rules created by users, and is based on Bayesian analysis of the frequency of the words in spam or ham emails taken by a user. the particles will be registrated to the database and their values will be updated after each recived email. This solution provides the ability to adapt to frequent changes in the characteristics of email, but it requires considerably resources (processing power) for each email that gets to manage all the particles and change their values. b) training when there are errors: this refers to the idea that the frequent change of values can lead to more errors (incorrect values). In this case the values of the particle in the database are modified only if there is a reaction by the user. So the user controls whether the classification was correct or not. If the classification is incorrect, the user makes changes manually by placing email into the appropriate directory (inbox or spam). This requires less memory compared with the first mode of training. Negative aspects of this method

are the difficulties in adopting with the new features of emails. For example, a new kind of spam may require more time until it is classified correctly. c) training till to maturity: this solution is obtained by merging the two methods above. Initially pursued the idea of ”training to all” and till are obtained a sufficient knowledge, passed on ”training where no errors”, so the changes happen only when errors occur. This combination of methods provides the advantages of both solutions; the only problem is the determining when the filter has received enough knowledge to move to another method.

V.

8 SIMULATION RESULTS

In our simulations is used the way ”training to all”. Note here the training is done after each email, but in everyday use, this is done once a day, not to consume the processing power of machines. For a simple user, daily training do not brings any greater advantage, but at the company level it can be very important.

To compare the filters, the most useful training is 30 to 70 which means that the filter is trained with 70% of all received email (70% of all email messages contain different messages spam or ham). The filter uses ”knowledge” obtained from the training phase to determine whether an email is spam or legitimate.

After every processed mail the filter generates a binary result. In our case 1 means that the message is categorized as spam and 0 means it is classified as legitimate. This result is a binary vector which will be compared with results of other filters and also with the original classification.

The results were above our forecast, and the filtrate reacted very well during testing, and this was as the result of a large number of emails that were used for training. The efficiency of filters was above our forecasts for the entire group of emails, so it was reasonable to create models with different training. This brought the idea of training with 40% of all messages. In this case the testing is done with 30% of the messages, which are from the training group and the same for all filters.

This experiment aims to compare the filter with different training. For example, the message No. 1 was tested with trained Spam Assassin with 70% of messages; message No. 1 was also tested Spam Assassin filter that was trained with 40% of all messages. One should note that by switching to the filter with 40% training, the database of 70% model is deleted, and then the filter has other knowledge.

So in this experiment not only filters are compared with each other, but the same filter with different training. However the results were similar with acceptable accuracy.

In the end we trained the filter with 10% of messages. Although we had little training messages, the difference between models is small. Filter Thunderbird will be test without doing the training. This program has knowledge of the daily use. The main reason of this test is to evaluate the ability of the program without doing the training and to have information on the accuracy of filters that are trained by daily use. As we shall see later, a trained filter with messages to a user for 6 months (although most of the messages are written in Albanian) will have the accuracy of a trained filter manually. Remember that personal messages of 6 months are sent at the thunderbird filter and are made their manual classification. So this serves as training for the filter, but are not used messages that have trained other filters.

Filters are shown in Table 1 and the models are compared with each other. Mark ”X” indicates that the testing is done for the respective model. Training and testing time should not remain outside our attention because it is very important when filters will be used for large companies with more email traffic. Also is important the training time versus the time of testing. In everyday uses we are not able to measure the training time, so we can see only the testing times. The time it takes for each program is shown in Table 2. Our tests are done on a computer with processor frequency (CPU) 1.2GHz and 1GB of RAM memory. To analyze the statistical is used the margin of error. The number of errors is divided by the number of all messages tested. The error rate can be called failure rate. On the degree error is taken into account only the number of errors occurring. As expected more spam messages are allowed than legitimate emails filtered. This refers to the principle that better to have spam in the inbox, rather than to go legitimate messages to the directory spam, which in many cases we do not control at all, and so may lose forever these messages.

Thunderbird filter is used for 6 months no training is done with training emails, it results are poor. This was somewhat expected because the program was used with Albanian language messages, while messages for the testing are in English. However, after training the results are comparable with other filters. Spam that can pass without being filtered are called ”false negative”, while legitimate messages that are filtered are called ”false positive”. Remember that there are 603 messages used for testing, of these 413 are legitimate messages and 190 are spam messages. So we know the original classification of messages.

All three filters have a tendency to increase the accuracy with increased training. So the graphics being to decrease from left to right. So it is expected that for the Bayesian filters as much training to do the greater is the accuracy. Spam Assassin filter has a balanced result, so that not necessarily with more training brings greater precision. It is important to emphasize that Spam Probe gives us the best result and is the only filter with precision greater than 90%. Thunderbird has an intermediate result between the filters, but it is the simplest to use, because it serves even as a program to manage emails. Thunderbird which was not training, gave us a very poor result, but it will not take much into account, because to compare the filters should be put on equal terms.

9 VI.

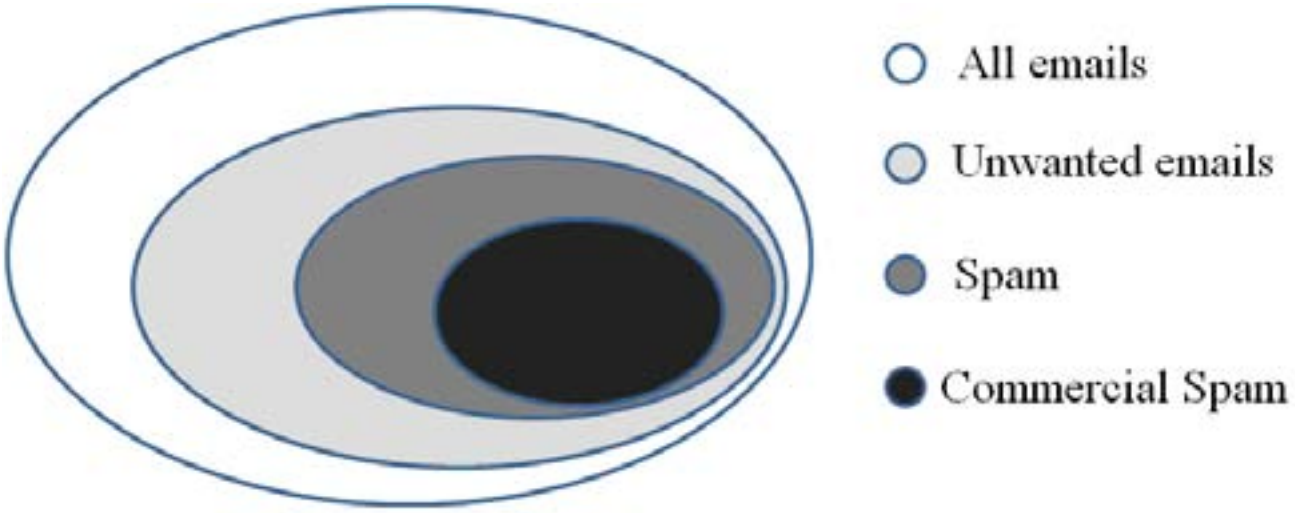
10 CONCLUSIONS

The conclusions are that all three filters have advantages and disadvantages. Cannot conclude what is the best filter. A brief summary is presented in Table 3. As can be seen, it is not clear which filter to choose, or which one is the best. There are several points of view to be taken into account. However, it may be a prediction that filtering with only one program does not yield results that will receive the filter with more programs. In fact for future work can be considered the union of two filters with one another to achieve greater precision. [1]

1

Figure 1: Figure 1 :

[1] © 2012 Global Journals Inc. (US)



2

Figure 2: Figure 2 :

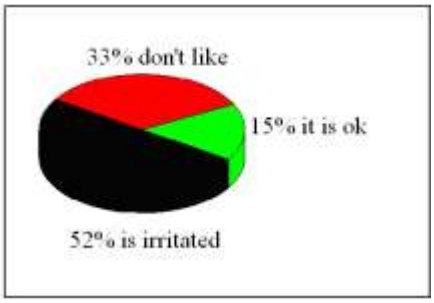


Figure 3: All

1

	70% training	40% training	10% training	0% training
SpamAssassin	X	X	X	
SpamProbe	X	X		
Thunderbird	X			X

Figure 4: Table 1 :

2

	70% training	30% training
SpamAssassin	5 min	12 min
SpamProbe	3 min	7 min
Thunderbird	0 min	15 min

Figure 5: Table 2 :

2

Figure 6: Table 2

3

Filters	advantages	Disadvantages
SpamAssassin	Good ability to training	Poor results with little training emails
SpamProbe	Good results with little training emails	More training does not lead to increased accuracy
Thunderbird	Easy to use	Requires more time

Figure 7: Table 3 :

.1 Global Journal of Computer Science and Technology

Volume XII Issue X Version I

[Zdziarski ()] 'Ending Spam -Bayesian Content Filtering and the Art of Statistical Language Classification'. Jonathan A Zdziarski . http://www.sg.hu/cikkek/41288/sophos_messze_meg_a_spam_halala *SpamCop statistics*, Jul 5, 2005. 2. No Starch Press.

[Kågström ()] *Impoving naïve Bayesian spam filtering*, Jon Kågström . 2005. (Master Thesis)

[Kornblum] 'SMTP Path Analysis -Exposing Zombie Spammers'. Aaron E Kornblum . *CEAS 2005*,

[SpamAssassin] http://www.mozilla.com/thunderbird/ *SpamAssassin*,

[SpamProbe] http://spamprobe.sourceforge.net/ *SpamProbe*,