Spam Detection using Adaptive Neural Networks: Adaptive Resonance Theory

*David Ndumiyana, Richard Gotora, and Tarisai Mupamombe

Bindura University of Science Education, Computer Science Department, P. Bag 1020, Bindura, Zimbabwe.

Downloaded 5 February, 2013

Today no single technology has achieved a hundred percent spam detection with zero false positives despite claims by vendors and suppliers of anti–spam software. However, no one can deny that machine learning heuristics in general and artificial neural networks to be specific have proven extremely effective and reliable at accurately detecting spam and reducing errors to an acceptable minimum level. This paper presents an enhanced experimental system for detection of unsolicited email messages using artificial intelligence and pattern recognition derived from an adaptive resonance theory algorithm. The system divides and groups email corpus into several categories to reduce the size of its input and increase output and recognition speed. Adaptive resonance theory is a type of neural network theory which is appropriate for pattern recognition especially when integrated with other modules can achieve high performance during detection and improvements on performance during training phase.

Key words: Adaptive resonance theory, automatic classifier, neural networks, spams filtering, adaptive learning.

INTRODUCTION

Spam is defined as all email the user does not want and has not asked to receive, meaning spam is in the eye of recipient, hence it’s all unwanted emails the user cannot stop from reaching email-boxes according to Hershkop et al. [1]. Email is one of the world’s popular and frequently used ways of communication because it is a very cost effective, generally accessible and fast message transfer method of marketing legitimate products or services to millions of users worldwide Hershkop [2]. Spam is also commonly known as unsolicited bulk mail has already caused numerous problems such as filling email-boxes thereby engulfing important personal mails, wasting network bandwidth, consuming user’s time and energy to sort through them, and not to forget other concerns associated with junk emails such as crashed mail servers, pornographic advertisements received by children among others in agreement with Duncan et al. [3]. Spam filtering technique is used to detect unsolicited emails automatically from the recipient’s mail stream. The exponential growth in the use of email for communication purpose has unfortunately triggered an increase in the number of threats meant to destroy its existence. The huge amount of spam sent daily demands speed on researchers to come up with new spam filtering systems that reduce the impact of these unwanted emails. Spam is now such a serious problem that has brought negative economic impact on companies. Research previously done by Rejeb [4] reported that US companies lost an estimated $20 billion annually due to spam as well as a decrease in employee productivity since they spent time dealing with junk mail. Coalition against Unsolicited Bulk Mail [5] also estimated a huge loss of revenue amounting to billions of dollars to spam because customers, suppliers and other business partners lost faith in the ability of companies to stop an influx of spam into email boxes.

This paper will begin by outlining the problem statement before moving on to the impact spam has on both individuals and companies including the general importance of email communication. Next is a discussion on common filtering techniques with each accompanied

*Corresponding Author’s E-mail: ndumiyanad@gmail.com. Tel.: +263 772 417 329.
by advantages and disadvantages before moving on to a brief analysis of literature on work previously done on the same subject by other researchers. This takes us to the description of a new method of spam detection and prevention designed to give at least relief to both individuals and companies. Results of administrating the proposed spam filter are fully illustrated and analysed in the last few pages before a conclusion was finally drawn at the end.

**Problem Statement**

Spam is a very serious problem for all email users and the concern is that, it comes from human beings who want to make money by economically disadvantage companies doing legitimate business. Spam violates other people's culture by exposing children to pornographic materials [3]. These unwanted email messages are exposed to children in countries where such materials are banned even for adult people. In countries where such advertisements may not violate culture, common sense tells us that young people should reach a certain age to be able to watch prohibited materials and there is time when pornographic films or documents could be viewed. Moreover, no nation under the sky is purely indigenous, diplomats, tourists and immigrants are common in every country. Hunt et al. [6] noted that, “millions of spam messages are sent every day advertising pornographic websites, drugs or software products or of fraud (phishing)”. Spam has economic effects on both users and service providers thereby prompting the need to reduce its impact by developing filtering system through research. A study conducted by SpamFighter [7] revealed that in 2007 industry lost more than $20 billion to spam including wasting network bandwidth when dealing with spam and a loss in employee productivity. A detailed summary of the problems spam has brought to both users and service providers is given below.

**Impact of Junk Email**

**Spam Cost Money**

Today more than 50% of emails received on the Internet are spam and this trend is said to have much more negative consequences in China and America [8]. A previous report by Rejeb [4] estimated that US companies lose around $20 billion annually due to spam including loss of productivity of employees spending time dealing with junk email, installing and upgrading anti-spam filtering system, bandwidth and network resources.

Since email has become an essential communication facility for business, a study conducted by SpamFighter [7] noted that more than 50% of the 50 billion emails sent daily in the year 2007 were spam and the rates had been raised year after year. This became a major concern to industry that the issue cost almost $20 billion per week.

**Spam Waste Resources**

A lot of network resources which are important for network users are being wasted on a daily basis resulting in unnecessary delays in accessing emails. Spam may also fill up the user email box thereby creating insufficient space for legitimate email to sail through, smoothly. In addition, spam influences the daily work on many employees by requiring them to spend a lot time dealing with it on a daily basis. More so, many current spam emails bring users unexpected email attachments that would seriously crack the user's system.

**Spam Violates Culture**

Spam has become a notorious communication tool used to violate other people’s culture; a large number of children from across the globe are now exposed to a huge amount of pornographic spam materials especially through commercial advertisements. This is made worse by the fact that spam authors do originate from countries which may not see anything wrong with cultural provocation caused by junk emails and distribution of unwanted materials enjoys no restrictions across different geographical locations. According to Carpinter et al. [6], “millions of spam email messages are sent every day, advertising pornographic websites, drugs or software products or of fraud”

**Spam Invades Privacy**

Privacy is defined as the right an individual has of minding their own business without interferences from external forces. The public Internet exposes users to much of the unwanted emails; no matter how small the event is, as long as your email address is available on everybody's eye [9].

“The saddest thing is that Internet users are now aware of what is going on, resulting in most users losing confidence in public participation. Many people do not want to subscribe to newsgroup or mailing list because they know they will get spam” [10].

**Spam and Email Deletion**

Recipients of unsolicited commercial email messages must spend a lot of their productive time deleting annoying and possibly offensive communications. One problem that can arise is the possibility of users deleting important legitimate electronic messages in their emailboxes according to Kufandirimbwa et al. [9]. This means the user could lose out completely on the contents of email if they cannot request a retransmission of previous mail possibly because the recipient could not remember the original email sender.
Spam Filtering Techniques

A spam filter is a set of instruction for determining if the received mail is spam email or not and they are used to prevent spam email from passing through to the recipient. The unfortunate development is that no spam filtering technique has achieved a 100% guarantee that every junk mail is clinically dealt with because every approach has its weaknesses and strengths as reported by Sabri et al. [11]. A successful anti-spam technique provides high accuracy of spam detection; meaning the ability of anti-spam method to differentiate spam from legitimate email messages. Spammers are improving on their attacking techniques of spam emails including the representation of plain text with diagrams [12] in order to prevent it from detection by rule-based anti-spam filters. More so, spammers are conducting word obfuscation on keyword to modify it so that it approximates another keyword. A common example is found when spammer changes Viagra to “vi@gra” or “v-i-a-g-r-a”. It is therefore important to establish effectiveness, accuracy and ease of administration of anti-spam solution. Effectiveness is a factor used to evaluate anti-spam filtering techniques and refers to the percentage of spam emails correctly filtered as spam. Software vendors of anti-spam are encouraged to continually monitor and fine tune their products from time to keep them strong.

Accuracy is another factor used for integral evaluation of anti-spam solution and refers to the ability of anti-spam solution to distinguish legitimate messages from spam emails. This is very important because more than one or two false positives–legitimate emails that poorly designed anti-spam solutions mistakenly classify as junk email–can represent the failure of a product for a computer user as reported by Kufandirimbwa et al. [9]. An increase in false positives is reciprocated by increase in the number of users who also lose confidence in spam blocking therefore clients are forced to manually wade through quarantines, and delete spam, a risky and frustrating exercise according to Rejeb [4].

One of the possible concerns for an anti-spam filtering solution is false positive problem where a spam filter may wrongly identify a legitimate email as a spam and block from reaching the destination. On the other hand, anti-spam system may also face false negativity where a spam filter may identify a spam as a legitimate email and allow that email to sail through to the intended recipient [13]. A mean is used to quantify the effectiveness of a spam filter, it can be based on percentage of spam email being blocked while allowing legitimate emails to pass straight to the user’s mailbox, hence every spam filtering system should be able to balance threshold and minimise the number of false positives and false negatives according to Hassan et al. [14].

The following is short description of common filtering techniques with the advantages and disadvantages of each:

Whitelist Filtering Techniques

The user can only receive email messages from senders whose email address or domain or IP address listed on a previously designed whitelist [12].

**Advantage:** Minimizes false negatives since all emails received are from legitimate and trusted sources.

**Disadvantages:**
- This method is too aggressive and can lead to high number of false positives because it prevents the recipient from accessing a legitimate email messages from a new contact not listed on the whitelist [15].

Blacklist Filtering Techniques

In this technique, users do prepare in advance a list of email addresses; domain or IP addresses which the system automatically considers as spam [12,13].

**Advantage:** This technique has a very low CPU overhead and is considered easier to implement since it only requires DNS lookup [3].

**Disadvantage:** there is great need to update blacklist on a regular basis since it expands rapidly thereby keeping pace with spammers who are always changing their identities.

Response System Filtering

The technique maintains a whitelist for known sender’s address and a blacklist for known spammers’ address. Every incoming email message whose source address is listed on whitelist will be allowed to pass through to the user email box. On the contrary, any incoming email message from a site listed on the blacklist is immediately rejected and deleted. If email message is incoming from a source whose address does not appear in either whitelist or blacklist, then a request is sent back to the sender to authenticate themselves as legitimate. The source that proves its identity as legitimate is then listed on the whitelist. From that moment the previous email and future communication will send straight to inbox.

**Advantage:** This method reduces negative impact associated with traditional whitelist technique by minimizing the number of false positive. It also reduces the action of spammers by putting a blockade on the incoming spam email.

**Disadvantages:** The technique is known to generate a deadlock problem whenever two new parties are communicating for the first time and both are running response system at the same time, the request to prove identity sent by the recipient’s system is caught by the sender’s system, then neither part will have time to give
appropriate reply in agreement with Delanny [15].

Checksum Filtering Technique

A spam filtering solution can use a checksums (fingerprints or signatures) of known spam to identify in order to detect spam with accuracy. If an email matches these fingerprints then it is a likely spam [16].

Advantages: Ordinary users may help identify spam and not only network administrators thereby increasing the number of tools available to fight against spam emails.

Disadvantages: Spammers have developed a technique to evade the method by adding random text into the message subject or the body.

Heuristic method/Rule-based Technique

The anti-spam software searches email messages for characteristics that indicate spam. The features may include specific words or even phrases, badly formed message headers and a huge number of exclamation marks and capital letters [13]. An email message is classified as spam if its point value surpasses a set threshold.

Advantage: The rules generated by rule – based software can be shared to others, meaning that the knowledge generated on spam can be globally disseminated and distributed quickly.

Disadvantages: Authors of spam emails can easily overcome this technique by performing word obfuscation such as changing "virus" to "v-i-r-u-s".

Case – Base Reasoning (CBR) Method

It is a lazy technique to machine learning where induction is delayed to run-time. Therefore CBR can be updated continuously and new training dataset is immediately available to the induction process [17].

Advantage: Individual users can maintain their own case bases to represent their personal, subjective interests as outlined by Duncan et al. [3].

Disadvantages: It is slower in doing computation therefore more time is required to perform filtering process.

Bayesian Filtering Technique

This is defined as the process of using a Naïve Bayes classifier to detect spam email messages. The Delanny [15] asserts that algorithms are capable of classifying the occurrences of certain words and phrases in terms of how and where they appear in the email message, and not only by their existence.

Advantage: It can be trained on per-user basis. The Bayesian spam filter eventually assigns a higher probability based on the individual user's specific attributes or characteristics according to Liang [18].

Disadvantages: This method is susceptible to Bayesian poisoning, a technique used by authors of spam emails to degrade accuracy and effectiveness of spam filters that rely on Bayesian filtering. More so, it is generally difficult to interpret the message content which contains graphical images.

Related Literature

Ian et al. [19] in his research used a neural network method on a data set of email messages from a single participant user. Descriptive characteristics of words and messages similar to those required to identify spam messages were used as feature for defining spam messages. Ian used a total corpus of 1654 emails received by one of the authors over undisclosed number of months. Results of comparisons between his neural network filter and Naïve Bayesian technique indicated that the developed neural network needed fewer features to achieve results produced by Naïve Bayesian approach.

Levent et al. [20] developed anti-spam filtering methods for agglutinative languages in general and for Turkish to be specific. He used dynamic methods based on Artificial Neural Networks and Bayesian Networks and his algorithms is user – specific. Further to that the algorithms have the advantage of adjusting themselves with characteristics of incoming email messages. According to his findings, a total of 750 emails including 410 spams and 340 hams were used in the experiments where a success rate of approximately 90% was achieved.

In a related development James Clark [21] presented a paper on neural network based system for automated email classification. In addition to that he also presented linger which can be defined as a neural network based system used for automatic email categorization problems. His research showed that neural networks can successfully be used for automated email filing into mailboxes and spam mail filtering.

One of the authors who did research based on neural networks is Puniskis et al. [22] who used neural network approach to the classification of spam. The technique employs attributes composed of descriptive features of evasive patterns that spamming community employs rather than using the context or frequency of keywords in the email message. With availability of 2788 ham and 1812 email spam in his possession received during a certain period of months, he noted that artificial neural
networks are satisfactory but they are adequate for using alone as a spam filtering tool.

The cited literature is not the only available ones in the market with a history of tremendous achievement but serves to buttress the fact that artificial neural networks have become one of the new approaches that can be used to reduce concerns caused by authors of email spam. Therefore the use of neural networks for the spam filtering system used is supported by clear evidence from quoted literature, so the project would be a success as well.

**Neural Networks for Spam Detection**

A neural network is defined as a group interlinked physical devices called nodes or neuron and a good example is the human brain which is able to make very rapid and reliable decisions within a fraction of a second. Junk email presents a unique challenge for traditional spam filtering software solutions in terms of huge numbers of messages which reach millions on a daily basis and the nature of content in these unwanted email messages such as pornography, finance, products and services [10]. The general principle utilised in any spam filtering technique either heuristic or keyword-based is the same, spam email messages generally appear distinct from good messages. The ability to detect these differences is the first step toward detecting and stopping spam from spreading.

The use of neural networks technique for spam filtering provides a more refined, more mathematical and potentially a far more accurate and reliable approach for distinguishing between spam and good email messages [10,23]. The techniques attempts to accurately copies the way humans visually separate spam from legitimate messages without entirely exposed to every spam messages generated; the ordinary user very quickly learns to identify spam messages from legitimate emails. By exposing neural networks software techniques to a variety of email messages content both bad and legitimate email messages on a regular basis equips the approach with skills and ability to make fast and accurate decisions on what spam is and what is not junk email.

**Neural Networks Operation**

The basic operating principle on neural networks is based on pattern recognition, which means that each message is categorised according to its characteristics or attributes. For effective and accurate pattern recognition the neural network system must be trained first by subjecting it to computational analysis of message content with large representative samples of both spam and legitimate messages. This training enables neural network spam filter to recognize spam email messages from non-spam emails [10,11]. Training datasets for spam and non-spam emails is done by examining each class of email message according to definition of spam. By using statistical methodologies, all keywords that are unique to each class of email are identified computationally. The neural network system goes on to pre-process each email in the respective datasets to determine words that are common in each class of email.

When training session is complete the neural network is now used to scan live email message streams and each message is scanned to identify relevant words for processing by the filter. In order to maximize detections and minimize false positives, an accurately designed heuristic engine will cater for a variety of sensitivity thresholds in detecting spam email messages [10,24]. This means the dividing point between spam and non-spam can be adjusted so that the likelihood of false positives can be greatly reduced.

**RESEARCH METHODS**

An adaptive neural network spam detector was designed (Figure 1) using Adaptive Resonance Theory algorithms adopted from Grossberg [25]. The ART architecture was adopted from Cosoi [26] with improvements on effectiveness and accuracy of spam detection. ART architecture models are able to self-organize in real time and can generate stable recognition while receiving inputs attributes beyond those originally stored. Self–organize refers to the ability of artificial neural networks of creating their own representation of information they received during learning period. Adaptive Resonance Theory is a paradigm that has the ability to learn without supervised training and is very consistent with cognitive and behavioural models. ART can be described as an unsupervised paradigm based on competitive learning which is capable of automatically discovering classes and creating new ones whenever they are needed.

**The Training Phase**

An automated process was created (Figure 1) which gathered spam and ham corpus over a period of time during which patterns had been studied and learned without human input. The neural network received input data with heuristics that supplied information about the content of email. The system had greater affinity on keywords such as Viagra, free and obfuscated words extracted from the email header and other parts of the filter. A train corpus which contained both spam and legitimate email was generated and tests were done. The corpus of unwanted mail was divided into several categories such as Adult, Education, Fraud, Finance and Internet.

The size of both input and output of each neural network was limited by building a tree hierarchy of neural networks with each working on a different type of spam. This way the neural network increased detection and
enhanced rate of recognition. During training period if neural network failed to identify a class for a particular pattern, a new one will be created. If the neural networks specialised on one type of spam even when the number of heuristics was increased, the size of output could not grow rapidly resulting in a precise and refined analysis.

**Testing Phase**

The analysis would start with arrival of a new email. The general tree of heuristics was run first on the email in order to determine the category that could accept it. If none accepted, then email is considered legitimate, otherwise it was passed to the next neural networks which dealt with such spam. The algorithm was repeated before information was extracted, but on input associated with a chosen neural network. In the event that the next level is another neural network, the information is passed further repeating the algorithm also. If the email cannot be classified, then it is legitimate. Should the next level be the final one, then the email has been classified and analysis process was terminated at that point.

**RESULTS AND DISCUSSION**

Three classes of spam were used in this finding, namely; Health, Fraud, and Financial. The hierarchy contained four neural networks which were; a general category for processing general spam heuristics for a pre-classification of data and specialised networks, each corresponding to a class. The role of general category was to classify spam into general categories, Health, Fraud or Finance. Thereafter the input was passed on the specialised networks where dedicated heuristics were carried out. Table 1 shows detection of all three neural networks and division of heuristics for each class.

Information presented on Table 1 indicates detection for each category after each class received 2300 samples for training purpose and tested with10 000 test samples. Results for computation on false positives rate was close to 0.07%. There is clear evidence from data presented that the rate of detection improved consistently with an increase in input data sets. There was no limit to the number of heuristics that could be received by the neural network filter. Much attention on detection rate
The training phase continued for some time until it was realised that even if more training items are added, the results were now in the constant range of 97.56% – 99.99% detection rate according to test data items with a false positive rate of 0.01% according to results obtained and recorded (Table 3).

Although the number of trained items used by the neural networks spam filter of 128 000 is very small as compared to the millions of spam emails that are released into circulation on a daily basis, at least the adaptive neural networks spam filter has demonstrated that it can be trained and is ready to reduce negative effects caused by unwanted emails received by users in the whole world.

Conclusion

The neural networks spam filter used can achieve incredible 100% test set accuracy on machine learning standards because it could make rapid accurate predictions in a short space of computer time, it is efficient since it uses a few training attempts, and not forgetting its flexibility because learning allows new learning on one or more repositories without compromising prior knowledge until network memory is all used up. A new automatic process that can learn the patterns of new spam emails which does not compromise detection rate on traditional spam is now available in the market.

Machine–learned heuristics in general and neural networks to be precise have proved significantly effective and reliable at accurately detecting spam and reducing errors to a minimum acceptable level.

REFERENCES


and other Kernel Based Learning Methods Cambridge University Press, 2003; 24(2): 105–106


