ISSN 1392 – 1215 ELECTRONICS AND ELECTRICAL ENGINEERING
ELEKTRONIKA IR ELEKTROTECHNIKA 2008. No. 5(85)

T180 TELECOMMUNICATIONS ENGINEERING
TELEKOMUNIKACIJŲ INŽINERIJA

Artificial Intelligence for Greylisting Anti-spam

D. Puniškis
Department of Electronics engineering, Kaunas University of Technology,
Studentu str. 50, 51368 Kaunas, Lithuania, tel. +370 686 19904; e-mail: danius.puniskis@stud.ktu.lt

R. Laurutis
JSC „Information Avenue“
Elnio str. 10, 76344 Siauliai, Lithuania, tel. +370 685 28295; e-mail: remigijusl@aleja.lt

Introduction

According to the last MessageLabs spam statistics report [1], the volume of spam e-mail messages transmitted by the Internet has reached 75%. The problem of electronic junk still exists and became more sophisticated to combat, gaining the graphical, audio or even video shape as the force majoure of direct trading.

Varieties of email classification techniques are able to control the problem partly. The false positive (FP) is unacceptable, while important ham message treated as spam will be lost. However, due to none zero of false positives of single classifier there is a demand of methods to combine the different anti-spam techniques, to lower FP.

Having such huge load of spam, the question of recourses is also trivial. Considering this was developed simple, but promising greylisting (GL) technique [2], which has some disadvantage too. More serious is that it losses legitimate mail, and less one, needlessly delays it.

We propose a different email filtering technique where combination of ANN and greylisting are used, to exploit as more as possible the positive features of them.

Conventional Greylisting (CGN)

While spammers currently don't bother to use RFC-compliant software to send out messages, the greylisting technique of temporary rejection is still in power. At the moment, most spammers use a simple mechanism to send out spam, which does not react to temporary email rejections.

A CGN proceeding steps are shown in Fig. 1. The sending SMTP server initiates the process of e-mail transfer; the receiving SMTP server records a characteristic triplet of the message, usually consisting of:

1. The IP address of the host attempting the mail delivery,
2. The sender's address,
3. The recipient's address.

The receiving server then searches for this characteristic triplet in a local database. If no existing record matches, the message is refused with a "temporary failure" response (return code "451") and the triplet is stored locally. If, on the other hand, the triplet matches an existing record, the message is accepted and delivered to the final recipient.

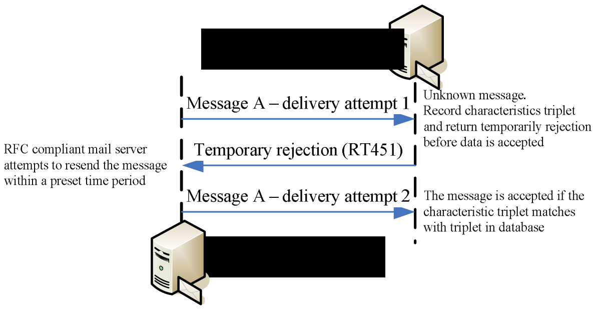

[redacted]
Message A – delivery attempt 1
Unknown message.
Record characteristics triplet and return temporarily rejection before data is accepted
RFC compliant mail server attempts to resend the message within a preset time period
Temporary rejection (RT451)
Message A – delivery attempt 2
The message is accepted if the characteristic triplet matches with triplet in database
[redacted]

Fig. 1 Main steps diagram of conventional greylisting

It is important to require that the repeated sending of a message occurs within a certain time period. It could be handled by introducing a time periods setting the minimal and maximum validity intervals of characteristic triplet after first sending attempt [3, 4].

This idea is very effective against spam; nevertheless there are several weaknesses in the conventional greylisting process:

- Greylisting introduces delays in the mail delivery process.
- How to define whether an SMTP session refers to a previous delivery attempt?
- Spammers can easily adapt and bypass CGN by resending messages or by sending potentially different messages with identical characteristic triplets successively.

Overview of email categorization using ANN algorithms

Each algorithm can be viewed as searching for the most appropriate classifier in a search space that contains all the classifiers it can learn. All machine learning algorithms require the same instance representation. The instances are messages and each message is transformed into a vector ($x_1, \ldots, x_m$), where $x_1, \ldots, x_m$ are the values of the attributes $X_1, \ldots, X_m$, [5, 6].

Each attribute represents a single token (e.g., “$$$”), of Boolean variables:

$$X_i = \sum_{0-opposit}^{1-are\ tokens} . \qquad (1)$$

The key concept of email classification using machine learning algorithms can be categorized into two classes, $y_i \in \{-1,1\}$, and there are N labeled training examples: $\{(x_1, y_1),\ldots,(x_n, y_n)\}, x \in R^d$ where d is the dimensionality of the vector [5].

Preparing the corpus for neural net classifiers, where frequency of selected n-Gram attributes extracted. We had build the dataset consisting of 57 n-Grams with appropriate frequency value of appearance (e.g. in the first message labeled as spam the token of word “you” is 2,15). The data set consist 4600 entries, with 1810 data points labeled as spam [6,7].

For content classification we have trained and tested three different neural nets: MLP, GFF, and SVM.

Dual greylisting technique

The proposed technique consists of dual independent classification stages, based on different message analysis.

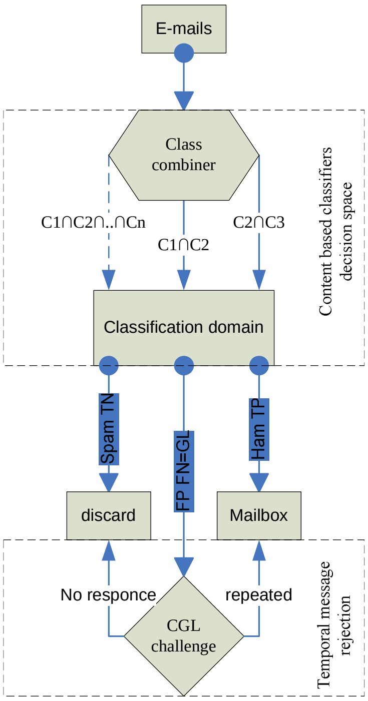


Fig. 2. Flow diagram of the email categorization process

As shown in Fig. 2 of email flow diagram: the first stage is represented by the combination of n classifiers based on content analysis of message, the second stage represents the CGN solution.

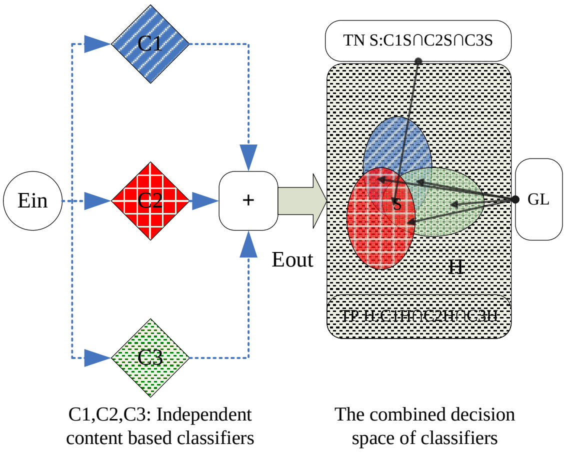
Firstly, the email corpus is transformed and indexed using learning algorithms. The transformed incoming emails are sent to the classifier domain for categorization. And depending on classification algorithm response additional conventional greylist challenge is used.

The classifier will categorize the email data and send to the output folder based on the identification of the email.

The figure 3 represents the GL generating system of different classifiers [8]. In this figure, where all classifiers give the same result ant its generated output sets overlaps each other, represents *TN* – pure spam mail *(S)*. All decision space which is outbound of any classifiers decision sets represents *TP* – pure ham mail *(H)*. The remaining regions of the output sets represent the *GL*, because not unique decisions come from all classifiers.

The output of the classifier will be categorized into three different parts:

- Common legitimate outputs from different classifiers, which is considered as *TP*
- Common spam outputs from different classifiers, which is considered as *TN*
- Different outputs comes from different classifiers, which is considered as *GL*



C1
C2
C3
Ein
+
Eout
TN S:C1S∩C2S∩C3S
S
H
GL
TP H:C1H∩C2H∩C3H
C1,C2,C3: Independent content based classifiers
The combined decision space of classifiers

Fig. 3. Output sets combinations of *n* classifiers

The all decisions of classification methods that do not directly get as *TP* or *TN* we consider as greylist too. The block diagram of three classifier (*n=3*) combination and their corresponding sequential output sets are given in figure 4. Every classifier has two sets of outputs $C_i S$ and $C_i H$ *(i=1…3)*. Considering this it is clear that inimitable result from classifiers goes only to the top and bottom section. The topmost section is *TP*, because all classifiers resolve positive and the bottommost section is *TN*, because all classifiers resolves negative. The remaining sections of this diagram are mixed outputs named as GL.

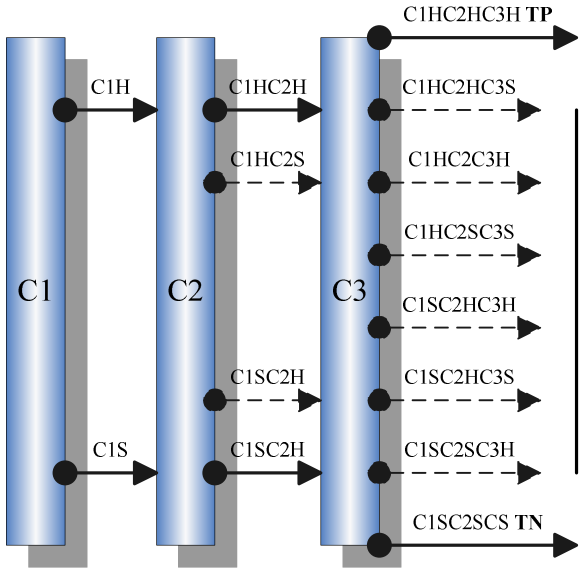


Fig. 4. Block diagram of combination of 3 classifiers

The combined output of overall *n* classifiers we can express by:

$$(C_1, C_2, ..., C_n) = \\ = \underbrace{\prod_{i=1}^{n} C_i(H)}_{TP} + \underbrace{\prod_{i=1}^{n} C_i(S)}_{TN} + \underbrace{\sum_{j=1}^{p} C_j(H)C_j(S)}_{GL} \;, \qquad (2)$$

where $p = 2^n - 2$ $\quad C_1, C_2, ..., C_n$ – classifiers.

From (2) we can derive the True Positive (*TP*), True Negative (*TN*) and the Greylist (*GL*) expressions.

For combination of three classifiers this follows as:
The number of ham outputs having *TP* is:

$$TP \Rightarrow \prod_{i=1}^{n} C_i(H) \Rightarrow C_1(H)C_2(H)C_3(H) \;, \qquad (3)$$

where $C_1 \cup C_2 : H \Rightarrow n(C_1H \cap C_2H)$,
$C_1 \cup C_3 : H \Rightarrow n(C_1H \cap C_3H)$,
$C_2 \cup C_3 : H \Rightarrow n(C_2H \cap C_3H)$,
$C_1 \cup C_2 \cup C_3 : H \Rightarrow n(C_1H \cap C_2H \cap C_3H)$.

The number of spam outputs having *TN* is:

$$TN \Rightarrow \prod_{i=1}^{n} C_i(S) \Rightarrow C_1(S)C_2(S)C_3(S) \;, \qquad (4)$$

where $C_1 \cup C_2 : S \Rightarrow n(C_1S \cap C_2S)$,
$C_1 \cup C_3 : S \Rightarrow n(C_1S \cap C_3S)$,
$C_2 \cup C_3 : S \Rightarrow n(C_2S \cap C_3S)$,
$C_1 \cup C_2 \cup C_3 : S \Rightarrow n(C_1S \cap C_2S \cap C_3S)$.

The outputs mixed from different classifiers, which mean some of the classifiers, are truly classified but some are misclassified. These sorts of output are considered neither *TP* nor *TN* but in the middle of them, which is called greylist. The total number of their combinations is:

$$GL \Rightarrow \sum_{j=1}^{p} C_j(H)C_j(S) \Rightarrow \\ \Rightarrow C_1(H)C_2(H)...C_{p-1}(H)C_p(S) + C_1(H)C_2(H)... \\ ...C_{p-1}(S)C_p(H) + ... + C_1(S)C_2(S)... \\ ...C_{p-2}(S)C_{p-1}(H)C_p(S) + C_1(S)C_2(S)... \\ ...C_{p-2}(S)C_{p-1}(S)C_p(H), \qquad (5)$$

where
$C_1 \cup C_2 : GL \Rightarrow n(C_1S \cup C_2H) + n(C_2S \cup C_1H)$,
$C_1 \cup C_3 : GL \Rightarrow n(C_1S \cup C_3H) + n(C_3S \cup C_1H)$,
$C_3 \cup C_2 : GL \Rightarrow n(C_3S \cup C_2H) + n(C_2S \cup C_3H)$,
$C_1 \cup C_2 \cup C_3 : GL \Rightarrow$
$\Rightarrow n[C_3H \cup (C_1S \cap C_2S) + C_1H \cup (C_3S \cap C_2S) +$
$+ C_2H \cup (C_3S \cap C_1S)] + n[C_3S \cup (C_1H \cap C_2H) +$
$+ C_1S \cup (C_3H \cap C_2H) + C_2S \cup (C_3H \cap C_2H)]$.

Experimental results

We have used three classification algorithms as MLP, SVM and GFF based on artificial neural nets (ANN). Every one of them was tested individually and in combined approach together with conventional greylist technique.

In the following table are presented comparative results of each ANN structures if they were applied separately and in the combined manner together with CGL.

Table 1. ANN's structure and proposed technique Confusion Matrix data

Classifier structure	True positive, %	True negative, %	False positive, %	False negative, %
MLP 57-25-1	96,2	98,3	3,8	1,7
SVM 57-1	99,7	99,72	0,3	0,28
GFF 57-20-15-1	95,9	96,5	4,1	3,5
3ANNs+CGN Combination	99,63	91,8	0,37	8,2

Conclusions

The main objective of proposed email classification technique is to reduce *FP* and achieve better accuracy. According to the email flow diagram showed in Fig. 2 and Fig. 3, and also keeping in mind the CGL algorithm, we can analyze four scenarios:

1. The incoming message is pure legitimate. The all three neural text *n-Gram* analyzers give decision as ham (*TP*). Email message is delivered to the inbox avoiding CGL analysis, thus not being delayed by unnecessary rejection.

2. The incoming message being pure legitimate is not correctly recognized by one classifier. Than the message labeled as *GL* is delivered to CGL stage where after valid response to resend request is delivered to the inbox.

3. The incoming message is pure spam. The all three neural classifiers detects *TN*. Message is immediately discarded. The additional resend request to the sending server from CGL stage is avoided. The processing load of the servers is reduced.

4. The incoming message is suspiciously spam. The classifiers give different decisions, and for the message labeled as *GL* at the CGN stage temporary rejection technique is applied, and sender's server RFC compliance is checked.

The combination of few different classifiers allows reducing the false positive.

Additional analysis of generated *FP* in conventional greylist, resolves the delay problem of pure legitimate message, and improves the *TP* of classifier.

The FP result of 3ANNs+CGL classifier as stated in Table 1, is non zero. It happens when there are messages with non textual content. The *TP* result is also controversial; however the technique of combined technology of greylisting and classification techniques gives advantage over any individual technique comparing the exploit of recourses and the delivery delay, having as minimal as possible *FP* result.

References

1. **MessageLabs Intelligence**. Spam review // Annual security report 2007. – Accessed at: http://www.messagelabs.com, last accessed on February, 2008.
2. **Levine R.** Experiences with Greylisting // Taughannock Networks. – Trumansburg, New York, 2006.
3. **Janecek A., Gansterer W., Kumar K.** Multi-level reputation-based greylisting // ARES Barcelona. – 2008.
4. **Harris E.** The next step in the spam control war: Greylisting. – Accessed at: http://projects.puremagic.com/greylisting/whitepaper.html. – 2003.
5. **Drucker H., Shahrary B., Gibbon D.** Support vector machines: relevance feedback and information retrieval // Inform. Process. Manag. – 2002. – No.38. – P. 305–323.
6. **Puniškis D, Laurutis R, Dirmeikis R.** An Artificial Neural Nets for Spam E-mail Recognition // Electronics and Electrical Engineering. – Kaunas: Technologija, 2006. – No. 5(69). – P. 73–76.
7. **Laurutis R., Puniškis D.** Neural networks for computer virus epidemics recognition. // Electronics and Electrical Engineering. – Kaunas: Technologija, 2005. – No. 4(60). – P. 28–32.
8. **Rafiqul I, Wanley Z**. An Innovative Analyser for Email Classification Based on Grey List Analysis // IFIP International conference on Network and Parallel computing, Dalian, 2007.

Submitted for publication 2008 02 11

D. Puniškis, R. Laurutis. Artificial Intelligence for Greylisting Anti-spam // Electronics and Electrical Engineering. – Kaunas: Technologija, 2008. – No. 5(85). – P. 49–52.

Current methods for detecting email system mostly works by examining content characteristic of incoming messages. Due to huge impact to recourses, such as bandwidth wasting, increased processing load, the greylist technology was developed to exploit the incompatibility of spammers mail servers, when they doesn't respond on the request to repeat message resend. Both of these methods, applied to the spam problem separately, give some disadvantages. The first one, analyzing the n-grams of text symbols in emails, gives some portion of misclassification, being unable due to some reasons make correct decision. The second one needlessly delays legitimate mail forcing to wait appropriate time gap for response or even losses it, if the sending server is not properly configured. Combining both techniques into one system, we can improve spam filtration effectiveness. The first stage of classifiers eliminates the unnecessary rejection and delay, when email is legitimate. If the message appears being spam and will be correctly labeled by all classifiers, it will be immediately discarded before reaching conventional greylist stage. If the spam or legitimated message is detected partly, i.e. with some misclassification it will be treated as greylist message and temporary rejection technique applied at the greylisting stage, where it will be discarded or delivered to the inbox. Ill. 4, bibl. 8 (in English; summaries in English, Russian and Lithuanian).

Д. Пунишкис, Р. Лаурутис. Анти-спам фильтр на основе нейронных классификаторов и серых списков // Электроника и электротехника. – Каунас: Технология, 2008. – № 5(85). – С. 49–52.

Современные методы обнаружения спама обычно работают, анализируя параметры содержания входящих сообщений. Для автоматической блокировки спама тоже используются серые списки, основанные на том, что «поведение» программного обеспечения, предназначенного для рассылки спама, отличается от поведения обычных серверов электронной почты. Если почтовый сервер получателя отказывается принять письмо и сообщает о «временной ошибке», сервер отправителя обязан позже повторить попытку. Спаммерское программное обеспечение в таких случаях, обычно, не пытается это делать. Оба метода, используемые отдельно, дают некоторые недостатки. Первый – ошибочно принимает решения из-за недостаточной точности текстового анализатора. Второй – придает задержку или совсем теряет сообщение, если сервер отправителя не отлажен корректно. Используя эти технологии в сочетании, можно увеличить точность определения спама и устранить положительную ошибку. Нейронные классификаторы, определяя чисто–нормалное письмо, устранит ненужную задержку. Если сообщение определяется как подозрительная, требование об ответе на временной ошибке, проверит сервер отправителя к согласности RFC стандарту. Ил. 4, библ. 8 (на английском языке; рефераты на английском, русском и литовском яз.).

D. Puniškis, R. Laurutis. Neuroninių klasifikatorių ir pilkųjų sąrašų technologijos el. pašto filtras // Elektronika ir elektrotechnika. – Kaunas: Technologija, 2008.– Nr. 5(85). – P. 49–52.

Šiuolaikinės elektroninio pašto filtravimo sistemos veikia turinio filtravimo pagrindu. Dalis pašto žinučių gali būti klasifikuojamos ir naudojant vadinamuosius pilkuosius sąrašus. Metodas paremtas RFC protokole numatytu laikinu žinutės atmetimo ir kartojimo užklausos principu ir šiandien efektyviai išnaudoja el. šiukšlių generavimo strategijos netobulumus. Taikant abu šiuos metodus atskirai, susiduriama su tam tikrais trūkumais. Teksto klasifikatorius ne visada iš susidarytų n-Gramų dažnių mokymo aibės geba priimti teisingą sprendimą (teigiama klaida). Pilkųjų sąrašų metodas susijęs su žinutės vėlinimu, kai per nustatytą periodą yra laukiama pakartotinio atsakymo iš siuntėjo tarnybinės stoties. Siekiant sumažinti teigiamą klaidą ir pasiekti didesnį filtravimo efektyvumą, siūloma abu metodus sujungti. Klasifikavimo lygmenyje nustačius, kad žinutė yra normalaus turinio, ji į pilkųjų sąrašų analizatorių nebepatenka ir nėra vėlinama. Jei nustatoma, kad žinutė įtartinai panaši į nepageidaujamo turinio žinutes, pilkųjų sąrašų analizatorius klausia siuntėjo tarnybinės stoties pakartotinio patvirtinimo ir priklausomai nuo gauto rezultato priima atitinkamą sprendimą. Il. 4, bibl. 8 (anglų kalba; santraukos anglų, rusų ir lietuvių k.).