ELECTRONICS AND ELECTRICAL ENGINEERING

ISSN 1392 – 1215 2007. No. 6(78)

ELEKTRONIKA IR ELEKTROTECHNIKA

ELECTRONICS

T170

ELEKTRONIKA

Behavior Statistic based Neural Net Anti-spam Filters

D. Puniškis

Department of Electronics engineering, Kaunas University of Technology,
Studentu st. 50, 51368 Kaunas, Lithuania, phone: +370 686 19904, e-mail: danius.puniskis@stud.ktu.lt

R. Laurutis

JSC „Information Avenue“,
Elnio st. 10, 76344 Siauliai, Lithuania, phone: +370 685 28295, e-mail: remigijusl@aleja.lt

Introduction

Everybody will agree that e-mail has become one of the most common technologies of nowadays communication. However, due to its' SMTP protocol imperfectness, it gives a variety of problems. Unsolicited marketing messages or spam, virus and worms spreading and phishing attacks are nowadays reality - utilizing more than a half of the total daily message traffic [1].

While unsolicited traffic is a commercial trend, there is a big potential of intrusion in PC to convert it to spambot. However, this is performed with high carefulness using novel exploits of existing vulnerabilities of operating systems in limited dimension. The trend is that spammers are not trying to infect as more as possible machines. Having the small groups of infected PCs, it is easer to stay unnoticed, i.e. the antivirus developer wont be interested to react for complain of not massive security problem.

Recently, the methods for email systems detection limits to examination of characteristics of incoming messages. Where spam detectors calculate statistical features on received email for classification usually dealing with corpus composed of messages from several distinct users. Thus it is not possible to profile appropriate user's behavior. For characterization the users' normal email behavior the outgoing email traffic should be observed, after which comparing different behavior patterns the abnormal could be detected and suspended.

This individual-user based analysis, when combined with technologies that examine incoming mail, could form an extremely strong defense against the spread of spam and phishing messages and even novel stealth intrusions.

Feature descriptions

The appropriate user's email activity or behavior could be described by collecting some statistic of suitable feature. The part of them could be calculated from a single email (e.g. incoming email activity of a single user) and the rest that examine several email over fixed amount of time (e.g. the average character number of the single user's email subjects).

The purpose is to select the number of features which best describe abnormal sending (or receiving) behavior. Each feature can be expressed by either a continuous or polynomial value. For example, a frequency calculation returns a number, where feature responsible of email attachments is represented as an array of bits, where each bit represents the presence of specific type of attachment.

E-mail message features

Here are described numerical values calculated on a per e-mail message basis. The categories of features that represent their output as one or more bits, i.e. polynomial values, are presented in Table 1.

Table 1. Polynomial value single email features

Feature per-email	Description
Presence of HTML	Buggy HTML exploits usually used to overcome text classifiers.
Presence of script attributes	Useful in detecting potential security risks
Presence of images	Images are often used by spammers exploiting image processing vulnerability.
Presence of hyperlinks	Spam always goes with them
Number of attachments	Usually ordinary user do not attach many files to their email, however mail worms, due to its spreading mechanism, are sending attachments.
Number of word/characters in the subject and body	This features help build a basic profile o the user's writing characteristics. Spam messages have distinct form of automatically generated text.

The numerical values calculated over a fixed amount of time, typically consisting of user's last thirty massages, are continuous statistics. The category of feature that represents values, are presented in Table 2.

Table 2. Continuous value single email features

Number of emails sent/received	Spam-bots or worms tend to send emails faster than average user.
Number of unique email recipients:	It counts addresses in the To:, Cc:, and Bcc: headers.
Number of unique sender addresses	Many users have multiple active accounts on the same machine. However, a single machine sending from a large number of addresses at a high rate could indicate abnormal activity.
Average number of word/characters per subject body; average word length	It captures trends in email wording allowing to separate normal email from malicious activity.
Variance in number of word/character per subject, body; variance in word length.	It is suspicious when subject and body is written in capital letters
The set of distinct word frequency	The words which are frequent in spam messages.
Ratio of emails with attachments	Most users do not send large amount of emails with attachments in sequences.

Histogram distance metrics

Feature set is designed to capture specific elements of user email behavior. For the individual contributions of each feature to the overall effectiveness an analysis of feature ability to capture information specific to individual behavior is required.

One of the methods to classify behavior is histogram analysis. It is possible between histograms of a specific feature over the data of two separate users to estimate how similar they are. Histograms are compared to one another to find similarity or abnormal behavior between different users' accounts, and within the same account (i.e. long-term profile.

For every pair of histograms h_1, h_2 there is a corresponding number $D(h_1,h_2)$ which indicates the distance between h_1 and h_2 .

To demonstrate the difference in e-mail message feature distributions among individual users we used simplified histogram intersection method according to formula:

$$D_1(h_1,h_2)=\sum_{i=0}^{n-1}\left|h_1[i]-h_2[i]\right| \cdot \tag{1}$$

The graph in Fig. 1 shows normalized histograms of two users in the dataset of the values for the features calculating the number of distinct addresses email is sent to. By taking the absolute value of the difference of each bin over these two users' histograms, we generate the graph in Fig. 2, which represents how different the two user's behavior is, according to selected feature.

According to the opposite metric of per-feature user similarity, we can plot how features separate individual behavior, considering different pairs of users. The data should be normalized, thus the maximum value of this difference is 1 (when compared histograms do not have any overlaps) and the minimum value is 0 (when histogram completely overlaps each other). The following figures illustrate the differences between all combinations of users for appropriate feature.

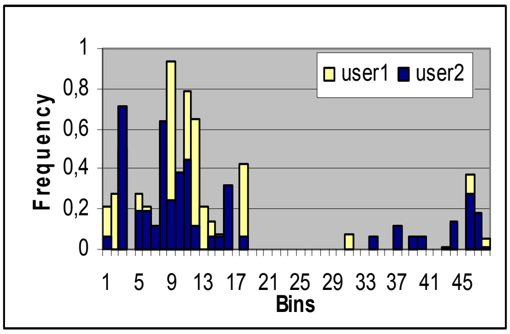

Fig. 1. Normalized histograms for two users of the values for the features

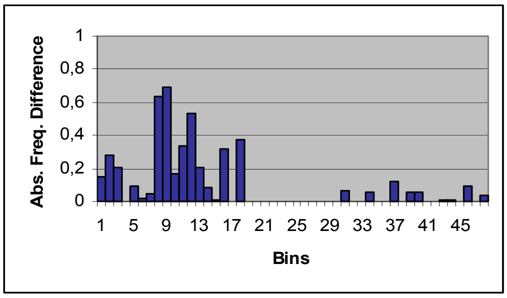

Fig. 2. The difference of feature histograms of two users

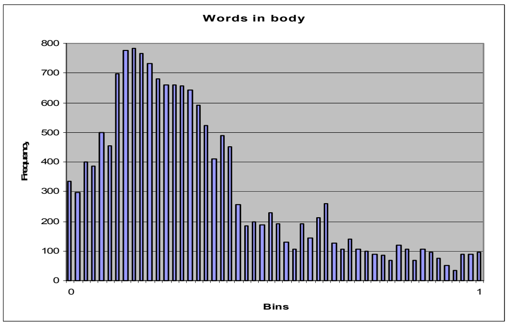

Fig. 3. The frequency of *Words* in email body distribution

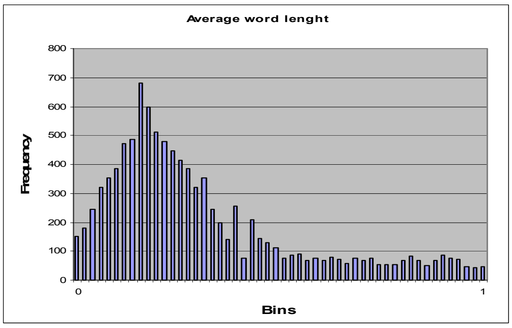

Fig. 4. Average word length distribution

In such distribution characteristics, could be expressed all selected features, and from them concludes that user behavior features are different per user, and each feature behaves slightly differently over all users, i.e. some statistics vary more widely than others.

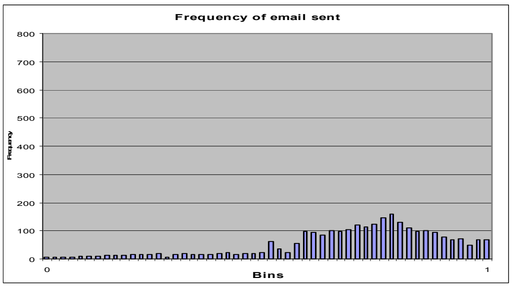

Fig. 5. Email sending frequency distribution

Spam detection

For classification problem we used several data sources: a corpus created by a custom real-time email interception framework that collected data from 27 co-workers within private company. We constructed training and test data sets where emails are first marked and labeled by the user indicating whether they are spam or normal. This information might also be retrieved from user behavior observation (i.e. when they delete a message prior to opening it, or move it to a "spam" folder).

The aim of this paper is to compare effectiveness of different classifiers and to show general degradation in performance as the feature set grows in size. Applying collected corpus to several classifiers.

Support Vector Machine

It is based on the idea that every solvable classification problem can be transformed into a linearly separable one by mapping the original vector space into a new one, using non-linear mapping functions. SVM's learn generalized linear discriminant functions of the following forms:

$$f(\vec{x}) = \sum_{i=1}^{m'} \omega_i \cdot h_i(\vec{x}) + \omega_0 \cdot \quad (2)$$

where m' – dimensionality of the vector space; $h_i(x)$ – the non-linear function that map the original attributes to the new ones. SVM applies a linear algorithm that attempts to maximally separate the "normal" data from the origin via a hyper plane boundary.

The SVM is trained with normal and abnormal user activity. Our process data consist of 4600 data points: 2990 for training, 1150 for cross validation and 460 for testing. We used a training set of 2990 data points with 27 incoming and 13 outgoing email features. The corpus consists 1810 data points containing actual spam activity. Data points are used for training using Gaussian Radial Bias Function (RBF) kernel option. The kernel function defines the feature space in which the training set examples will be classified.

Naïve Bayes

It is the simplest and most widely used algorithm that derives Bayesian Decision Theory. Considering condition that attributes $X_1,...,X_n$ are independent given the category C, and each the probability that a message with vector $\vec{x} = \langle x_1,...x_m \rangle$ belongs in category c is:

$$P(C = c \mid \vec{X} = \vec{x}) = \frac{P(C = c) \cdot \prod_{i=1}^{m} g(x_i; \mu_{i,c}, \sigma_{i,c})}{\sum_{c' \in \{c_L, c_S\}} P(C = c') \cdot \prod_{i=1}^{m} g(x_i; \mu_{i,c'}, \sigma_{i,c'})}, \quad (3)$$

where

$$g(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (4)$$

is a normal probability density function of continuous attributes.

Evaluation measurement

Detection rate – is the total number of spam detected divided by the total number of spam messages

$$Detection = \frac{tp}{tp + fn} \quad (5)$$

False positive rate – represents the number of non-spam email flagged as spam divided by the total non-spam messages.

False negative rate – represents the number of non-spam email flagged as spam.

Error rate – the amount of emails misclassified regardless of the mistakes made by the model.

Table 3. Individual classifier performance over spam experiments

Learning algorithm	Detection rate	False positive rate	False negative rate
Naïve Bayes	98,3 %	3,8 %	1,7 %
SVM	96,5 %	4,1 %	3,5 %

The results of different classifier accuracy to detect spam from user behavior and its dependency on feature set selection are presented in Figure 6.

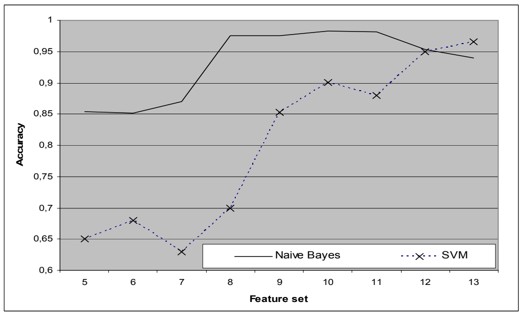

Fig. 6. The accuracy of classifiers on feature selection during training

Conclusions

This paper presents an approach to spam detection using feature generation on outgoing email traffic to build models of user behavior.

Initial analysis indicates that user behavior can be clustered into sets of common models that describe the general behavior patterns of most users making a large scale detection system feasible.

There are several features based on word distributions and social network analysis that can be included in our feature set for better prediction of user behavior. In addition, any deployable system will have to account for the temporal changes in user behavior via periodic retraining

The effectiveness of feature selection can be seen in the performance of abnormal mail sending detection via different structure classifiers, and the best results from our data set was reached applying Naive Bayes statistical method. There also obvious that increasing feature set the accuracy of classifiers doesn't changes or even reduces.

References

1. **MessageLabs Ltd**, Spam review // Research report. – MessageLabs Ltd.; http://www.messagelabs.com last accessed February, 2007.
2. **Shawe-Taulor, J., Christiani, N.** Kernel methods for pattern analysis, 2004.
3. **Guyon, I., Elisseeff, A.** An introduction to variable and feature selection // Journal of Machine Learning Research. – 2003 . – No. 3. – P. 245–271.
4. **Hauser. S**. 2003. Statistical Spam Filter Review. http://www.sofbot.com/article/Spam_review.html; last accessed February, 2007.
5. **Weiss, A.** Ending Spam's Free Ride // netWorker. – 2003. – No. 7(2). – P. 18–24.
6. **D. Puniskis, R. Laurutis, R. Dirmeikis.** 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.

Submitted for publication 2007 02 27

D. Puniškis, R. Laurutis. Behavior Statistic based Neural Net Anti-spam Filters // Electronics and Electrical Engineering. – Kaunas: Technologija, 2007. – No. 6(78). – P. 35–38.

Current methods for detecting email system mostly work by examining characteristic of incoming messages. Spam detectors calculate statistical features on received email for classification usually dealing with corpus composed of messages from several distinct users. Thus it is not possible to profile that user's behavior. To characterize the user's normal email behavior the outgoing email traffic can be observed, after which abnormal behavior caused by a compromised machine can be detected and contained at the source. The effectiveness of feature selection can be seen in the performance of abnormal mail sending detection via different structure classifiers, and the best results from our data set was reached applying Naive Bayes statistical method. There are also discovered that increasing feature set, the accuracy of classifiers doesn't changes or even reduces. For false positive reduction and gaining classifier accuracy it is essential to combine several distinct methods of user based behavior and content analysis over bidirectional mail traffic. It could form an extremely strong defense against the spread of spam. Ill. 6, bibl. 7 (in English, summaries in English, Russian and Lithuanian).

Д. Пунишкис, Р. Лаурутис. Анти-спам фильтры на основе нейронных сетей и статистических классификаторов // Электроника и электротехника. – Каунас: Технология, 2007. – № 6(78). – С. 35–38.

Современные методы обнаружения спама обычно оснавываются на анализе параметры, статистическьй проверки их содержания, обычно входящих сообщений. Это ограничивает оценку поведения пользователя в телекоммуникационных сетях. Для характеристики поведения пользователей должен проводиться анализ исходящего трафика, после чего, обнаружив аномалию, его можно блокировать. Для обнаружения аномалий в сети использованы классификаторы различной структуры, где наилучшие результаты получены статистическим методом Бэсена. Установлено что с увеличением числа параметров точность классификатора не всегда возрастала. Для увеличения точности и снижения положительной ошибки осмыслено использовать методы поведения в сочетании с анализаторами содержания. Ил. 6, библ. 7 (на английском языке; рефераты на английском, русском и литовском яз.).

D. Puniškis, R. Laurutis. Nepageidaujamo pašto filtrai neuroninių tinklų ir statistinių klasifikatorių pagrindu // Elektronika ir elektrotechnika. – Kaunas: Technologija, 2007. – Nr. 6(78). – P. 35–38.

Šiuolaikinės elektroninio pašto filtravimo sistemos veikia turinio filtravimo pagrindu, kai žinutės priskiriamos vienai ar kitai klasei pagal sukurtus statistinius parametrus, dažnai apdorojant tik keleto vartotojų duomenis. Taikant tokį metodą, sumodeliuoti vartotojo elgsenos tinkle negalima. Vartotojų elgsenos tinkle modeliams sudaryti turime sukaupti duomenis apie siunčiamą duomenų srautą, registruoti atitinkamus įvykius ir, pastebėję neįprastus veiksmus vartotoją blokuoti. Įvykių tinkle analizei ir detekcijai naudojami skirtingų struktūrų klasifikatoriai. Geriausių rezultatų pasiekta statistiniu Beseno metodu. Be to, nustatyta, kad didinant parametrų rinkinį nebūtinai didėja klasifikatoriaus tikslumas. Siekiant sumažinti teigiamą klaidą ir pasiekti didesnį filtravimo efektyvumą, būtina sujungti skirtingom ir turinio ir elgesio stebėjimo metodikom ir pagrįstus klasifikatorius į bendrą sistemą, kas leistų gerokai sumažinti nepageidaujamo pašto kiekį. Il. 6, bibl. 7 (anglų kalba; santraukos anglų, rusų ir lietuvių k.).