Spam Behavior Recognition Based on Session Layer Data Mining*

Xuan Zhang, Jianyi Liu, Yaolong Zhang, and Cong Wang

School of Information Engineering,
Beijing University of Posts and Telecommunications, Beijing, China, 100876
{zhangx, liujy, zyl, wangc}@nlu.caai.cn

Abstract. Various approaches are presented to solve the growing spam problem. However, most of these approaches are inflexible to adapt to spam dynamically. This paper proposes a novel approach to counter spam based on spam behavior recognition using Decision Tree learned from data maintained during transfer sessions. A classification is set up according to email transfer patterns enabling normal servers to detect malicious connections before mail body delivered, which contributes much to save network bandwidth wasted by spams. An integrated Anti-Spam framework is founded combining the Behavior Classification with a Bayesian classification. Experiments show that the Behavior Classification has high precision rate with acceptable recall rate considering its bandwidth saving feature. The integrated filter has a higher recall rate than either of the sub-modules, and the precision rate remains quite close to the Bayesian Classification.

1 Introduction

Knowledge Discovery has been broadly applied in Email Management and Spam Detecting [1], but until now applications here mainly concerned with mail body processing and user modelling. In this paper, we propose a framework using Decision Tree learning to find the patterns of spam transfer in session layer. Therefore, a Behavior Classification can be set up against spam flooding while saving the bandwidth.

Spam, officially called unsolicited bulk email (UBE) or unsolicited commercial email (UCE) [2], has long become a social problem and even brought serious threat to the Cyber Security. These annoying mails not only waste the end users' time to deal with them, but also consume large amount of bandwidth as well as enormous volume of server storage.

Several approaches have been proposed and typical server-side techniques exist in following three main categories:

Black/White-list filtering blocks bulk mail delivery during the transport process by checking blackhole list or white list when suspicious mail servers

* This work was supported by the National Natural Science Foundation of China under the Grant No. 60575034.

L. Wang et al. (Eds.): FSKD 2006, LNAI 4223, pp. 1289–1298, 2006.

connect [3]. This kind of techniques are quite simple and effective but somehow too "radical" to accommodate the exceptions.

Content-based Filtering targets the body of a mail message or the message header [3] using keywords matching or statistic methods. However, automatic software, often called "bulk mailers", usually add "innocent" content to outsmart content-based filters [1]. Furthermore, this kind of techniques can't prevent these unsolicited mails from taking up bandwidth.

Traffic monitoring technique, usually known as rate limiting, monitors and limits the mail traffic by observing the traffic flows of various hosts in a network [4], aims to defend mail server from large-scale spam attack. Impacts brought by junk mails mitigate but quantity of received spam not reduced.

Our approach is put in force during the email transfer session to recognize anomalous connections based on Behavior Classification which is set up through session log mining using Decision Tree algorithm. In this way, individual mail server can be more active in defending the spam flow according to its own "communication" history without getting the whole mail message and therefore not wasting bandwidth.

The rest of paper is organized as follows: In Section 2 we introduce related works and investigate whether the behavioral difference between spam and ham (normal mail) can be detected. Section 3 presents our approach in detail. Then behavior patterns discovered in a real mail server is shown in Section 4, effectiveness of the Behavior Classification (and also the integrated filter combined with Bayesian Classification) is evaluated here. Finally, we come to the conclusion in Section 5.

2 Related Works and Research Basis

To counter spam during the email delivery transaction is a new trend for the related research [5,6,7,8,9]. To detect abnormal email behavior in training phase, there are three typical approaches:

Man-Made Rule method: define classification model and rules manually through analyzing the possible characteristics in spam transfer.

Blacklisting method: query both a local black list file as well as Internet databases in real-time to defeat unsolicited traffic [9].

Filtering method: using statistical method to scan the data of mail to calculating spam probability of an email message [9].

These approaches are to some extent either unhandy or empirical. Our approach also focuses in the delivery phase, but applied at the mail server side in session layer and use Decision Tree learning in training phase. And it can be integrated with various server side techniques. Our opinion is based on the following reasons.

Email does not work so differently than it used to when it first appeared [1]. Sending email relies on SMTP(Simple Mail Transfer Protocol), including commands in Table 1 [10] during a typical interaction.

Table 1. Command Message in a typical SMTP Session

Command	Specification
HELO/EHLO	Sender's host name
MAIL	Sender of the messager
RCPT	Intended recipient of the message
DATA	Body of the mail
QUIT	Goodbye message

Unfortunately, SMTP is so simple that it can be easily faked and cheated. Spammers employ anonymous or compromise (spam trojan) server to mass mail junk mails without going through a specific mail server or a particular ISP, instead, they connect to the desination mail server directly or use a so-called open relay [3].

On the other hand, spam abuse the transferring protocol thus has distinct characters in behavior level, difference between bulk mailer and normal server is discriminated, for example, spammers send large amount of similar mails continuously and have to alter the message (e.g. domain info in command ”HELO”) at intervals to hide the track or spoof anti-spam filter while an ordinary mail server rarely. More details are given in Section 3.1.

3 Spam Behavior Recognition

3.1 Data Collecting and Pre-processing

The data collector works as an MTA (Mail Transfer Agent) so that command messages of each mail connection can be recorded. In fact, we add a Bayesian filter [11] to the collector for data pre-processing. Samples of the session data are listed in Table 2.

These items are almost all we can get during the SMTP session besides the user account and password, in which the item ”IP” shows the mailer's IP address, ”Domain” records the hostname of the mailer claimed well after the ”HELO” command, ”Sender” and ”Receiver” represent the sender and receiver available in the ”MAIL” and ”RCPT” command respectively, and ”Time” is the time stamp of each email.

The filter mentioned above checks the legitimacy of each mail first. And then we verify the result in manual additionally by looking through the subject of the mails (and only the subject items due to the privacy consideration). Finally, we make all these records clustered according to the ”IP” field, and the abnormal behavior set of mail connections is separated from the alternative one.

What should be pointed out is that samples in the spam collections does not mean these spams were sent from the exact domains we list, they were just sent as such by spam trojan.

Table 2. Sample of Session Data

Legitimate Sessions

IP	Domain	Sender	Receiver	Time
202.108.3.174	smtp.vip.sina.com	-@vip.sina.com	-	2005-9-26 18:38:57
202.108.3.174	smtp.vip.sina.com	-@vip.sina.com	-	2005-9-27 16:54:03
202.108.3.174	smtp.vip.sina.com	-@vip.sina.com	-	2005-11-09 9:07:29

Spam Sessions

IP	Domain	Sender	Receiver	Time
59.64.128.128	sina.com	david@sina.com	joe@-	2005-3-25 13:10:28
59.64.128.128	pro.ro	joe@pro.ro	michael@-	2005-3-25 13:23:28
59.64.128.128	tom.com	brenda@tom.com	fred@-	2005-3-25 13:31:26
...	...	...	...	...
202.105.72.231	zx433dan	i8f9d9g1@microsoft.com	-	2005-3-25 16:22:15
202.105.72.231	zx738tuj	n1s2e9w8@msa.hinet.net	-	2005-3-25 17:16:26
202.105.72.231	MISS	Mk2o0l2h7@msn.com	-	2005-3-25 17:55:55

Note: due to the privacy issue, legitimate sender, receiver and domain of the target server is presented as "-" here.

3.2 Feature Selection

The feature selection is very important for building efficient decision tree [12]. Bulk mailers tamper with command messages in SMTP session to pretend innocently, therefore we'd better not learn the patterns of spamming behavior from each entry of the record directly, but clues lie here.

As shown in Table 2, mails from IP "59.64.128.128" were obviously spams for the reason that the mailer altered it's "domain" from "sina.com" to "pro.ro" in such a while, the possibility of a normal server to behavior like this is very little (normal server may modify this but the change is usually tiny), therefore we check the consistency of claimed domain from the same IP and named the feature as "ip2domain".

The feature "ip2domain", "ip2sender", "domain2sender" check the consistency of "domain", postfix of "sender"(e.g. sina.com) and the correlation between each other compared with the previous record of the same IP, as we have pointed out, normal servers behave consistently while bulk mailers send various inconsistent message pretending that they are innocent in every individual connections. It is easy to be cheated once but only once if consistency is checked.

"Open Relay" is a term used to describe mail servers that allow unauthorized users to send email through them [13], now it become a backdoor for spammers relaying bulk mails. Though most mail servers have forbidden open relay for the secure consideration, spammers usually intrude to the normal servers to modify the option to allow open relay. Feature "ip2receiver" is to detect the trick.

"Senderlth" and "Receiverlth" may not be noticeable but the pattern mining proved they are useful. Explanation is given in Section 4.1 .

Table 3. Behavior Features and Specifications

Feature	Specification
ip2domain	y, if ”domain” consists with previous records’ of the same IP (”consists” for short in following items); n, otherwise
ip2sender	y, if postfix in ”sender” consists; n, otherwise
ip2receiver	y, if postfix in ”receiver” is correct; n, otherwise (check if ”open relay”)
domain2sender	y, if ”domain” associate with postfix in ”receiver” consists; n, otherwise
senderlth	length of characters before ”@” in ”sender”
receiverlth	length of characters before ”@” in ”receiver”

In fact, the malicious mailer also behave abnormally that it often send mails to email addresses not exist in the target server, like data in Table 2, all receivers of mails sent from ”59.64.128.128” are nonexistent in our mail server. This is one of the ways spammers collect their ”Mailing List”. We do not take the authenticity of target address as feature in our training phase because the validation of receiver can be accomplished by the mail server and this kind of connections will not take up much bandwidth (but may imply the mail attack). Besides, it is not worth to maintain a list of valid addresses in the Behavior Recognition phase and that will also reduce the flexibility of the Behavior Classification. Servers that attempt to make nonexistent-receiver mailings frequently can be simply added into the black list.

In addition, frequency of connections from the same IP would be taken as a feature in situation that the throughput of mail server maintains in high level.

4 Experimentation

The total of the terms is 27500, 15000 for training, 2500 for pruning and 10000 for testing. For comparison, we also evaluate the integrated effectiveness of our Behavior Classification and Bayesian Filtering [11].

4.1 Pattern Mining and Analysis

Decision tree is chosen because they provide a comprehensible representation of their classification decisions [14]. C4.5 Decision Tree Algorithm, which has been widely used and tested [15], is applied to the training set to generate the decision tree. The Decision Tree retrieved is often a very large, complex that overfits the training data and incomprehensive to experts. It’s impossible to avoid noise in the training set, then overfitting the tree to the data in this manner can lead

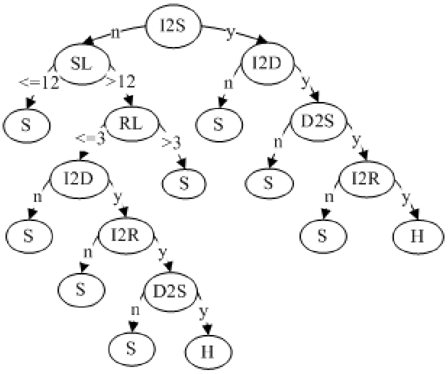


Fig. 1. Decision Tree after pruning. Nodes Specification(detail can be find in Table 3): I2S:ip2sender, SL:senderlth, RL:receivelth, I2D:ip2domain, I2R:ip2receiver, D2S:domain2sender, S:Spam, H:Ham.

to poor performance on unseen data [16]. Therefore, we pruned the tree in C4.5 algorithm firstly and then in manual way. Figure 1 shows the decision tree which was generated and pruned for characterizing the attribute post test.

From the Decision Tree, we can find:

1. Left branch of the Tree is the main branch, which contains approximately 58.13% of the whole entries.

2. The significance of each feature can be sorted as below: ip2sender, senderlth, receiverlth, ip2domain, ip2receiver, domain2sender.

3. The tree generated above may include some phenomena that are not noticeable. Length of email address are rarely taken into consideration in the anti-spam topic, however, the tree shows that length of mail sender and receiver can be valuable for classifying mail connections. For example, "senderlth $\leqslant$ 12", which can be explained that generally legitimate sender don't apply for long-name accounts from their Email Service Provider while spammers like to adopt such long names ("wdjhzfdomkvyotv" e.g.) to "produce" more different email addressed randomly. And "receiverlth $\leqslant$ 3", that's because the mail server under protected provides email accounts according to the initial letter of employees' name (Chinese name usually contains two to three characters). It is implied that the Data Mining approach do help discover unrevealed knowledge hide in behavior record.

4. The tree may differ from one mail server and another, from time to time after long intervals, so that new patterns of spam behavior can be detected dynamically. This approach of spam pattern learning can be "personalized" to every single mail server individually.

4.2 Evaluation

The integrated framework is illustrated in Figure 2, dashed frame presents the Behavior Recognition Process given in Section 3. Mail stream is filtered through Anti-Dos module, IP block module, Behavior Classification and Bayesian

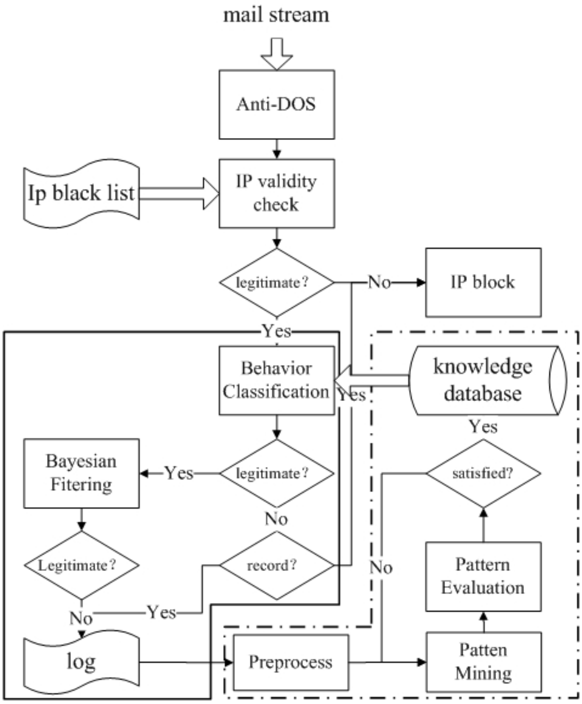
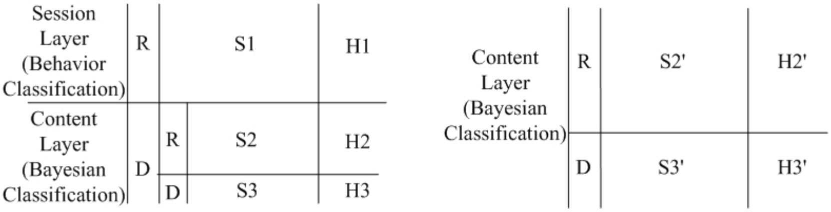


Fig. 2. Integration of Two Classfication



Session Layer (Behavior Classification)	R		S1	H1
Content Layer (Bayesian Classification)	D	R	S2	H2
		D	S3	H3

(a)

Content Layer (Bayesian Classification)	R	S2'	H2'
	D	S3'	H3'

(b)

Fig. 3. Recall and Discharge. (a) Integrated Classification (b) Bayesian Classification R denotes Recall, D denotes Discharge while S denotes Spam, H denotes Ham.

Filtering. For evaluation, Anti-Dos and IP block module are turned off, and the Behavior Classification and Bayesian Filtering Module are evaluated individually and integratively.

From Figure 3, recall rate R and precision rate P can be calculated. Results are listed in Table 4. Binary tree traversal algorithm is used to present the Decision Tree generated above in Figure 1, judging whether a mail session is legitimate or not in classification implementation [17].

Table 4. Evaluation comparing with Bayesian Filtering

Items	Behavior Classification(%)	Bayesian Classification(%)	Integration(%)
Precision	98.00	92.03	91.71
Recall	53.27	73.47	78.32

4.3 Behavior Classification Evaluation

According to the experiment, the precision rate (98.00%) of Behavior Classification is quite high while the recall rate (53.27%) is relatively low. As mentioned above, the behavior patterns of spam will alter with time going, and bulk mailer would also adjust its behavior against the anti-spam's strategy. So, to always recall most spam patterns is ideal but hard to realize, and we should always emphasize on the precision rate taking the quality of mail service in consideration.

On the other hand, the advantage of Behavior Recognition in the session layer is that it's time saving and can protect bandwidth resource against junks. In this aspect, more than 50% of the malicious connection be detected meet the requirement in reality. By adopting strategies such as reject connection or rate limiting, the mail server being protected can benefit a lot from Behavior Classification.

4.4 Integrated Evaluation

The evaluation of the integrated filtering focus in the real line frame in Figure 2. We select Bayesian Classification in email content to be integrated with Behavior Recognition, not only because they work in different phases of email transfer, but also for the reason that the Bayesian filtering usually have high recall rate.

As shown in Table 4, the integrated recall rate is higher compared to either of the classification individually, but precision rate is 0.32% lower than the Bayesian Classification (the lower one). In fact, for the two layer filtering, there are (Consist with the Figure 3, subscript 1 for Behavior Classification, 2 for Bayesian Classification individually, 2'for Bayesian Classification works after the Behavior Classification and *intg* for the integrated filter):

$$\max(S_1, S'_2) \leqslant S_1 + S_2 \leqslant \min(S_1 + S_2 + S_3, S_1 + S'_2) \qquad (1)$$

$$\max(H_1, H'_2) \leqslant H_1 + H_2 \leqslant H_1 + H'_2 \qquad (2)$$

So, we can get:

$$\frac{1}{1 + \frac{H1 + H'_2}{\max(S_1, S'_2)}} \leqslant P_{intg} \leqslant \frac{1}{1 + \frac{\max(H_1, H'_2)}{\min(S_1 + S_2 + S3, S_1 + S'_2)}} \qquad (3)$$

$$R_{intg} \geqslant \max(R_1, R_2) \qquad (4)$$

Taken high precision (98.00% here) of the front Module (Behavior Classification), the integrated precision is bounded from 90.82% to 94.03% according to Equation 3 in our experiment. And the recall rate would be higher than $\max(R_1, R_2)$, helping to save much of time and bandwidth. The performance tally with our expectation that the Behavior Classification can be integrated with current various anti-spam techniques.

5 Conclusion

In this paper, an approach is given to discover different patterns of Spam Behavior in transfer phase using Data Mining method, upon which a Behavior Classification was set up. Evaluated in experiment, the Behavior Classification, with high precision, contributes much to the bandwidth saving during email delivery.

Obviously, no single approach can be expected to expire all those unsolicited mails so far. Behavior Classification can be applied in ordinary mail servers and integrated with currently used spam filtering and blocking techniques.

Nevertheless, for the spammer behavior change dynamically, the Behavior Classification needs to be renewed periodically accordingly.

For future works, incremental learning of Behavior Patterns for the classification need to be realised, pattern incorporation and the frequency of learning are involved. Classifying spam connections in a finer granularity should also be considered, such as to distinguish the virus mail stream from common advertisement mail streams. Further more, it is a exacting work for individual mail server to counter the infestation of spam, cooperating with other servers is an essential trend for all anti-spam systems. By sharing Email Behavior Patterns among spam filters will make the Behavior Classification working more effectively.

References

1. Ioannis Katakis, Grigorios Tsoumakas, I.V.: Email Mining: Emerging Techniques for Email Management. In: Web Data Management Practices: Emerging Techniques and Technologies. Idea Group Publishing (2006) 32
2. Flavio D. Garcia, Jaap-Henk Hoepman, Jeroen van Nieuwenhuizen: Spam Filter Analysis. In: Proc. 19th IFIP International Information Security Conference, WCC2004-SEC, Toulouse, France, Kluwer Academic Publishers (2004)
3. Lueg, C.: Spam and anti-spam measures: A look at potential impacts. In: Proc. Informing Science and IT Education Conference, Pori, Finland (2003) 24–27
4. Anti-Spam Technologies: Anti-Spam Technology Overview http://e-com.ic.gc.ca/epic/Internet/inecic-ceac.nsf/en/gv00297e.html#3.4.3
5. Salvatore J. Stolfo, Shlomo Hershkop, K.W., Nimeskern, O.: Emt/met: Systems for modeling and detecting errant email. In: Proc. DARPA Information Survivability Conference and Exposition. Volume 2. (2003) 290–295
6. Prasanna Desikan, J.S.: Analyzing network traffic to detect e-mail spamming. In: Proc. ICDM Workshop on Privacy and Security Aspects of Data Mining, Brighton UK (2004) 67–76
7. Qiu Xiaofeng, H.J., Ming, C.: Flow-based anti-spam. Proc. IEEE Workshop on IP Operations and Management (2004) 99–103

8. Banit Agrawal, Nitin Kumar, M.M.: Controlling spam emails at the routers. In: Proc. International Conference on Communications. Volume 3., Seoul, South Korea (2005) 1588–1592
9. Tran, M.: Freebsd server anti-spam software using automated tcp connection control. Technical report, CAIA Technical Report 040326A (2004)
10. Forouzan, B.A., Gegan, S.C.: TCP/IP Protocol Suite. McGraw-Hill (2000)
11. Jianyi Liu, Yixin Zhong, Y.G., Wang, C.: Intelligent spam mail filtering system based on comprehensive information. In: Proc. 16th International Conference on Computer Communication. (2004) 1237–1242
12. Tarek Abbes, Adel Bouhoula, M.R.: Protocol analysis in intrusion detection using decision tree. In: Proc. International Conference on Information Technology: Coding and Computing. Volume 1., Las Vegas, Nevada (2004) 404–408
13. Glossary:Open Rlay http://www.viruslist.com/en/glossary?glossid=153949388
14. Mitchell, T.: Machine Learning. McGraw-Hill (1997)
15. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo, CA (1993)
16. James P. Early, C.E.B., Rosenberg, C.: Behavioral authentication of server flows. In: Proc. 19th Annual Computer Security Applications Conference, Las Vegas, Nevada (2003) 46–55
17. Zhang, Y.: Research and application of behavior recognition technology in anti-spam system. Master thesis of Beijing University of Posts and Telecommunications (2006)