Detecting Methods of Virus Email based on Mail Header and Encoding Anomaly

Daisuke Miyamoto, Hiroaki Hazeyama, and Youki Kadobayashi

Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, Nara, Japan
{daisu-mi, hiroa-ha, youki-k}@is.naist.jp

Abstract. In this paper, we try to develop a machine learning-based virus email detection method. The key feature of this paper is employing Mail Header and Encoding Anomaly(MHEA) [1]. MHEA is capable to distinguish virus emails from normal emails, and is composed of only 5 variables, which are obtained from particular email header fields. Generating signature from MHEA is easier than generating signature by analyzing a virus code, therefore, we feature MHEA as signature to distinguish virus emails. At first, we refine the element of MHEA by association analysis with our email dataset which is composed of 4,130 virus emails and 2,508 normal emails. The results indicate that the one element of MHEA should not be used to generate MHEA. Next, we explore a way to apply MHEA into detection methods against virus emails. Our proposed method is a hybrid of matching signature from MHEA(signature-based detection) and detecting with AdaBoost (anomaly detection). Our preliminary evaluation shows that f_1 measure is 0.9928 and error rate is 0.75% in the case of our hybrid method, which outperforms other types of detection methods.

1 Introduction

Email viruses have become one of the major Internet security threats today. An email virus is a malicious program which hides in an email attachment, and become active when the attachment is opened. Currently, the most successful computer viruses spread via email [2], and the main infection source of virus is via email [3]. Thus, distinguishing virus-attached emails, called *"virus email"* from other emails is required to protect computer from viruses. Current approaches for dealing with email viruses rely on anti-virus softwares. Usually, anti-virus vendors capture and analyze viruses and generate the signature of each virus. However, a traditional signature-based approach is ineffective against newly-discovered viruses. Even anti-virus vendors endeavor to add new signature, it is hard to catch up with new viruses because modern viruses usually have both self-propagating and obfuscation functions.

According to [1], virus emails are distinguished from normal emails with Mail Header and Encoding Anomaly(MHEA). The MHEA is easily extracted from an email without deep analysis. The elements of MHEA can be obtained from particular email header fields which are rarely affected by variety of email viruses. Thus, MHEA has a possibility to deal with increasing the number of new viruses.

In this paper, we refine the Arai's proposal by association analysis with our email dataset, at first. Our dataset is composed of 4,130 virus emails and 2,508 normal emails. Arai suggested 5 variables on the email headers and/or anomaly on MIME encoding [1], however, we assume 1 variable, namely Encoded-Words length, can be affected by variety of email viruses. According to our verification, generating MHEA without this variable performs better than generating MHEA with this variable.

Next, we attempt to develop MHEA-based methods for detection of virus emails. We divide our dataset into two samples; one is used for training and one is used for testing. The training dataset has 2,065 virus emails and 1,259 normal emails, which are randomly sampled without replacement, respectively. We then employ MHEA obtained from training dataset as signature, and check if each email in testing dataset is registered in the signature. The result shows that 2,859 emails in testing dataset are classified virus emails or normal emails without error. The remaining 460 emails cannot be classified because the MHEA of these emails are not registered in the signature. Generally, these emails are regard as normal, so we assume these emails are classified into normal emails. We observe that f_1 measure is 0.9910 and error rate is 1.11%.

Aside from signature-based detection, anomaly detection has a possibility of classifying the 460 emails. For building an anomaly detection method, we transform the elements of MHEA to the weak hypotheses in AdaBoost, a typical machine learning algorithm. We perform supervised learning with the training dataset, and evaluate the classification accuracy with the testing dataset. The result shows that f_1 measure is 0.8858 and error rate is 10.82%.

We then design a hybrid detection method. The hybrid detection system classifies emails based on the signature first, and then performs anomaly detection on unclassifiable emails. We observe that f_1 measure is 0.9928 and error rate is 0.75%. In comparison to other detection methods, a hybrid detection method can achieve the best detection accuracy.

2 Mail Header and Encoding Anomaly

In this section, we summarize MHEA proposed in [1], and refine the generation process of MHEA.

2.1 Summary

According to [1], Arai distinguished between virus emails and normal emails by using Mail Header and Encoding Anomaly(MHEA). He analyzed some MIME header fields in email dataset, which was composed of 33,003 virus emails and 52,601 normal emails. According to the result, he suggested 5 variables can be used as heuristics that were not affected by variety of viruses. The variables are as follows.

V1. Irregular line

"Irregular line" is a binomial variable that indicates the email includes irregular line in Encoded-Words (EW). According to MIME [4], the attachment files should be encoded to base64 format, called EW, when a Mail User Agent (MUA) sends binary data. MIME also defines that each line in the email is limited to 76 letters. If the length of EW is more than 76 letters, the MUA begins a new line to set in 76 letters for each line. Thus, EW should always be 76 letters except for the last line. However, Arai found some virus emails did not get along with MIME. Even if the line was not the last line, the number of letters in the line was not 76 in the cases of such emails. Arai regarded such lines as one of the footprints for detecting virus emails.

V2. EW length

"EW length" is an integer variable, which is calculated by dividing 1,000 into the length of EW. For example, EW length is 2 when EW has 2,076 letters. According to the analysis, Arai considered that the change in the file size is within 1,000 bytes even if the file size increased or decreased in virus's mutation process.

V3. MIME multipart

"MIME multipart" is a string variable, which is combined `Content-type header` filed. For example, MIME multipart is `text/us-ascii,image/jpeg` when an email contains text written in English and one jpeg-format image file.

V4. Extension header

"Extension header" is a string variable, which is combined particular extensional MIME header such as `X-Priority`, `X-MSMail-Priority`, `X-Mailer header` field. Arai suggested that virus emails would be sent from particular MUAs.

V5. Encoded MIME boundary

The string variable, which is encoded from MIME boundary [5]. Arai attempted encoding by the following steps: (i) alphabets and numbers are converted to the letter "X", and (ii) if the letter is the same as the letter before, the letter is removed. For example, when the given MIME boundary is `"ABCDEFG=="`, Step (i) converts from the MIME boundary to `"XXXXXXX=="`, and Step (ii) converts from `"XXXXXXX=="` to "X=".

MHEA is generated by concatenating these 5 variables. In [1], Arai reported that (i) in the case of virus emails, the number of unique MHEA is much lower than that of normal emails, and (ii) MHEA that were derived from virus emails were different from those of normal emails. He mentioned that MHEA of virus emails have many similarities in comparison to those of normal emails, thus, he concluded that MHEA is capable to classify virus emails.

2.2 Refinement

We assumed that MHEA has a weak point in its generating process. Basically, EW length is affected by variety of viruses; it is natural to assume that the file

Table 1. Frequency Distribution using V1, V2, V3, V4, and V5

Uniqueness of MHEA	Virus emails	Normal email
$\log(x) < 1$	207	1129
$1 \leq \log(x) < 2$	351	454
$2 \leq \log(x) < 3$	479	296
$3 \leq \log(x) < 4$	325	298
$4 \leq \log(x) < 5$	600	169
$5 \leq \log(x) < 6$	1207	162
$6 \leq \log(x)$	961	0
	4130	2508

Table 2. Frequency Distribution using V1, V3, V4, and V5

Uniqueness of MHEA	Virus emails	Normal emails
$\log(x) < 1$	50	886
$1 \leq \log(x) < 2$	106	453
$2 \leq \log(x) < 3$	116	406
$3 \leq \log(x) < 4$	130	226
$4 \leq \log(x) < 5$	462	375
$5 \leq \log(x) < 6$	1220	162
$6 \leq \log(x)$	2046	0
	4130	2508

size of a virus program changes 1KB or more in its mutation process. Thus, we consider that MHEA should be generated by concatenating 4 variables except from EW length.

To check my idea, we performed association analysis. First, we collected 4,130 of virus emails, which were sampled from our university email server from September 2006 to January 2007. In the email server, an anti-virus scans every email and these 4,130 emails were detected as virus. We also collected 2,508 emails sampled from our mailbox. There are 202,508 emails in the mailbox, and 2,508 emails have one or more attachment files.

Second, we generated MHEA for each email by concatenating 5 variables. We found 299 unique MHEA in the case of the virus emails and 1,179 in the case of the normal emails. We also observed that every MHEA of the virus emails is different from the MHEA of the normal emails.

According to [1], many virus emails had the same MHEA. For verification, we employ the uniqueness of MHEA as an index of similarity. We simply defined uniqueness x; if x emails have the same MHEA, the uniqueness of the MHEA of these emails is x. To perform our association analysis, we investigated the frequency distribution as shown in Table 1. We temporarily determined the classes of frequency distribution as follows: Uniqueness is quite high when $\log(x) < 1$, high when $1 \leq \log(x) < 2$, slightly high when $2 \leq \log(x) < 3$, neutral when $3 \leq \log(x) < 4$, slightly low when $4 \leq \log(x) < 5$, low when $5 \leq \log(x) < 6$, quite low when $6 \leq \log(x)$. We observed that there is difference between the distribution of MHEA derived from virus emails and that of normal emails. Based on this condition, Cramer's coefficient of association(C) was 0.607. It indicated that classifying virus emails with MHEA is appropriate.

We also generated MHEA by concatenating 4 variables except from EW length. We then drew the frequency distribution as shown in Table 2 and observed that C was 0.761. In comparison to the case of using 5 variables, removing EW length led to the stronger correlation between uniqueness of MHEA and types of emails(virus or not). Thus, we concluded that EW length was affected by variety of viruses, so employed 4 variables for generating MHEA in the following section.

Table 3. Signature-based Detection allowing unclassifiable email

	Actual virus emails	Actual normal emails	
Predict virus emails	2028	0	2028
Predict normal emails	0	831	831
Unclassifiable emails	37	460	460
	2065	1254	3319

Table 4. Signature-based Detection

	Actual virus emails	Actual normal emails	
Predict virus emails	2028	0	2028
Predict normal emails	37	1254	1291
	2065	1254	3319

3 MHEA-based virus email detection methods

Generally, detection methods are categorized into three types: signature-based detection, anomaly detection, and hybrid detection. Based on MHEA, we develop these 3 types of detection methods and evaluate the detection accuracy for each method. In our evaluation, we use f_1 measure (higher is better) and error rate (lower is better) as the indices of the detection accuracy.

3.1 Signature-based Detection

First, we developed a signature-based detection method which can check the MHEA of the issued email with MHEA databases.

We generated the MHEA by using training dataset. Our dataset was composed of 4,130 virus emails and 2,580 normal emails. We chose 2,065 emails from virus emails by random sampling and also chose 1,258 emails from normal emails, and constructed a training dataset with these 3,319 emails. The rest of the emails were used for testing. The number of unique MHEA was 619, in which 70 types of MHEA were derived from virus emails and 549 types were derived from normal emails.

We also performed a classification experiment employing the MHEA as signature. In the classification test, we took one email from the testing dataset and generated its MHEA, and checked if the same MHEA existed in the signature dataset. If so, we checked whether the MHEA in the signature dataset was derived from virus emails or not. When the MHEA was derived from virus emails, we classified the email as virus. Conversely, we classified the issued email as normal when the MHEA was derived from normal emails.

The results were shown in Table 3. We classified 2,859 of 3,319 emails without error. The remaining 460 emails were unclassifiable because the MHEA of these emails were not listed on the signature. In the typical signature-based detection systems, these emails were regarded as normal emails. Based on this fact, we regarded these unclassifiable emails as normal. The results were shown in Table 4.

Table 5. Anomaly Detection

	Actual virus emails	Actual normal emails	
Predict virus emails	1781	75	1956
Predict normal emails	284	1179	1363
	2065	1254	3319

We observed that f_1 measure was 0.9910, error rate was 1.11%, TP rate was 98.21%, and FP rate was 0%.

3.2 Anomaly Detection

To categorize unclassified emails, we developed an anomaly detection based on MHEA. We employed AdaBoost to construct an anomaly detection algorithm. AdaBoost, which was proposed by Freunde and Schapire, is the most typical boosting algorithm. AdaBoost solves many of the practical difficulties of the earlier boosting algorithms, and its ensembles perform better than the generic ensemble methods.

The weak hypotheses used in AdaBoost were transformed from V1, V3, V4, and V5 as mentioned in Section 2 and were shown as follows:

H1. Check if the email contains irregular line in Encoded Word
H2. Check if the uniqueness of MIME multipart is lower
H3. Check if the uniqueness of Extension header is lower
H4. Check if the uniqueness of Encoded MIME boundary is lower.

As an index for uniqueness, we checked frequency of appearance and compare it with defined discrimination threshold (θ). Imagine if an issued email's MIME multipart was observed in other 99 emails. In this case, 100 emails had the same MIME multipart, thus, frequency of appearance was 100. Because we decided $\theta = e^4 (\neq 54.60)$, frequency of appearance was larger than given θ; It denoted that this email was deemed to be a virus email. Notice that this threshold was temporarily decided, so we discuss the effectiveness of θ in Section 4.

We evaluated the detection accuracy as shown in Table 5. By using training dataset, we performed supervised learning to construct the final hypothesis with the AdaBoost algorithm. We then classified emails in testing dataset with the final hypothesis. We observed that f_1 measure was 0.8858, error rate was 10.82%, TP rate was 86.25%, and FP rate was 5.98%. In comparison to a signature-based detection method, the accuracy of an anomaly detection method decreased.

3.3 Hybrid Detection

The key feature of our hybrid detection method was employing both signature-based detection and anomaly detection. In the context of virus email detection, a hybrid detection method first tried signature-based detection to check an email, and then tried anomaly detection on the email only if the MHEA of the email

Table 6. Hybrid Detection

	Actual virus emails	Actual normal emails	
Predict virus emails	2060	20	2085
Predict normal emails	5	1234	1239
	2065	1254	3319

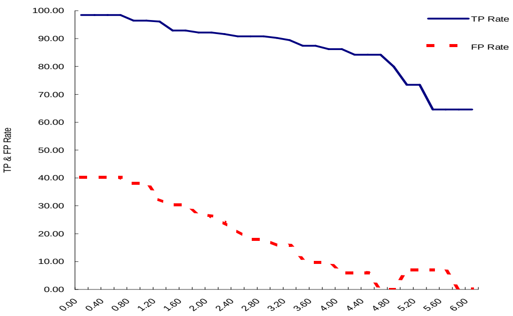

Fig. 1. Relationship between threshold and error rates

was not listed on signature. In short, we applied the anomaly detection method to the unclassifiable emails for improving the detection accuracy.

The results were shown in Table 6. We observed that f_1 measure was 0.9928, error rate was 0.75%, TP rate was 99.76% and FP rate was 1.59%. In addition, 32 of 37 virus emails, which could not be detected by the signature-based detection, were detected as virus in the case of the hybrid detection method. By comparing the hybrid detection method with the signature-based detection method, f_1 measure was increased and error rate was decreased, although FP rate was increased. So, we assumed that the hybrid detection method can improve the detection accuracy.

4 Future Work

In this section, we discuss the index for measuring the uniqueness. Throughout the paper, it was difficult to decide whether the uniqueness was higher or lower, so we temporary determined the classes in Section 2 and the threshold in Section 3.2.

Figure 4 showed TP rates and FP rates by changing $\theta(= e^x)$ from e^0 to e^6 in increments x of 0.2, where the x axis denoted θ, and the y axis denoted the TP and FP rates in an anomaly detection method. A normal line denoted TP rate, and a broken line denoted FP rate. When given θ was 4.6, the TP rate was 84.21% and the FP rate was 0%. If we could employ 4.6 for θ for constructing a hybrid detection method, f_1 measure was 0.9985, error rate was 0.18%, TP rate was 99.71%, and FP rate was 0%.

However, it was difficult to determine θ in learning steps. In our future work, we need to explore a suitable way for adjusting θ. After a hybrid detection method classified an emails as a virus or not, a detection system would confirm whether the classification was correct or not, and should automatically investigate the suitable θ.

5 Conclusion

The main infection source of viruses is via email. Based on this fact, we developed virus email detection methods using Mail Header and Encoding Anomaly(MHEA).

We refined the generation process of MHEA by analyzing our dataset, which was composed of 4,130 virus emails and 2,058 normal emails, all of which have one or more attachment files. We found that using 4 variables, namely irregular line, MIME multipart, extension header, and encoded MIME boundary, performed better than using the original 5 variables.

In our preliminary evaluation, we observed that f_1 measure was 0.9910 and error rate was 1.11% in the case of the signature-based detection method. To decrease the unclassifiable emails, we employed the AdaBoost algorithm for constructing an anomaly detection method, and we also observed f_1 measure was 0.8858 and error rate was 10.82%. We finally constructed a hybrid detection method, which first tried signature-based detection to check an email, and then tried anomaly detection on the email only if the MHEA of the email was not listed on signature. We observed that f_1 measure was 0.9928 and error rate was 0.75%.

Acknowledgments

This work was a part of "Research and Development on Traceback Technologies in the Internet" sponsored by the National Institute of Information and Communications Technology(NICT).

References

1. Arai, T.: Computer virus detection method using mail form information. In: Symposium on Cryptography and Information Security (SCIS2007). (2007) (in Japanese).
2. Sophos Corporation: Top 10 viruses reported to sophos in 2005. available at: http://www.sophos.com/security/top-10/200512summary.html (2005)
3. Information-technology Promotion Agency: Reporting status of computer virus - details for june 2008. available at: http://www.ipa.go.jp/security/english/virus/press/200806/documents/Virus0806.pdf (2008)
4. Moore, K.: MIME (Multipurpose Internet Mail Extensions) Part Three: Message Header Extensions for Non-ASCII text. RFC 2047, Internet Engineering Task Force (1996)
5. Freed, N., Borenstein, N.: Multipurpose Internet Mail Extensions (MIME) Part Two: Media Types. RFC 2046, Internet Engineering Task Force (1996)