ISBN 978-981-11-3671-9
Proceedings of 2017 the 7th International Workshop on Computer Science and Engineering (WCSE 2017)
Beijing, 25-27 June, 2017, pp. 808-814
doi: 10.18178/wcse.2017.06.141

The Analysis of E-mail Communication Security Based on Naïve Bayes

Yi Junkai [1],Zhao Siqi [1 +],Zhao Xianghui [2 +]

[1] College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China

[2 +] China information technology security evaluation center, Beijing 100085, China

Abstract: In the rapid development of information age, email is still an important tool of Internet information communication. It is significant to determine the behavioral relationship of communication on both sides quickly and accurately. By means of applying steganalysis and malware analysis technology, this paper analyzes the content of email and attachments, conducts feature extraction; establishes a behavior model based on naive Bayesian theory, and then proposes communication security behavior analysis method. The experimental results show that this method has an 83% rate of accuracy on communication behavioral relationship recognition.

Keywords: email; communication security behaviour, steganalysis; Naïve Bayes

1. Introduction

With extensive application of Internet, emails are still of key importance in many fields such as transaction processing and information communication. Meanwhile, social security of emails is more and more valued, and information and evidence involved in network crimes are increasingly significant. So it is a new challenge to analyze email information quickly and effectively, as well as the content and relationship of emails accurately, so as to specify communication behavior of email contacts.

Currently there are mainly two methods below to classify emails at home and abroad: one is based on analysis of text content and the other one is based on analysis of email attachments such as texts, pictures, etc. The ultimate purpose of classification is to recognize and filter spam, among which Bayesian Classification Algorithm is the most common method [1-7]. Put forward by Pearl in 1988, it is a model composed of a directed acyclic graph and a set of probability distribution. Naïve Bayesian classification algorithm, via structure expansion, is one of the best classifiers currently. Simple and effective, it has more prominent performance in practical application. It also has a solid mathematical foundation and relatively stable classification efficiency. Meanwhile, Naïve Bayesian classification model is not too sensitive to missing data, with less parameter estimation and digestible algorithm. In the field of text classification research, in the early 1950s researcher H. P. Luhn firstly put forth the theory based on frequency that characteristics of the entries appear in the text to calculate classification information, and later it has been accepted by many scholars at home and abroad. In 1960, Professor Maron published a paper "On Relevance, Probabilistic Indexing and Information Retrieval", exploring text classification deeply. After that, some scholars put forth a new text model, called Vector Space Model (short for VSM). It can be represented by vector space composed of a series of characteristics, which reduces the complexity of text representation greatly and improves the efficiency of text classification. With the development of research, machine learning algorithm replaced the previous classification algorithm gradually. There are multiple classification learning methods used in message text classification, such as Naïve Bayes, Neural Network, Support Vector Machine (SVM),

+ Corresponding author. Tel.: + 00861064451323.
E-mail address: 775854904@qq.com; zxhitsec@sina.com.

Decision Tree, K-nearest Neighbor Algorithm, Maximal Entropy, etc. Naïve Bayes, K-nearest Neighbor Algorithm and SVM are more commonly used. Text Feature Selection is the key step of automatic text classification. It is a process that applying computer technology, under the predefined classification system, according to the document content to be classified, attributes it to one or more categories. Automatic text classification technology research began in the 1950s, but now many classification models based on different theories have appeared, among which VSM is used to represent documents. For example, the letter T presents vocabulary collection contained by the document. Each word and its weight in the text are regarded as feature items, so the document can be expressed as the vector $d = (t_1, t_2, \ldots t_m)$ $(t_i \in T, 1 \leq i \leq m)$, and then according to the document vector and category vector, calculate the similarity, so as to determine document category.

This paper proposes a new analysis method of email communication security behavior. Communication behavior is a behavior relationship between both sides while communicating via email. Naïve Bayesian classification principle in applied statistics is to classify communication behavior between the sender and the receiver of email [8-9]. The fundamental principle of this method is firstly to analyze all documents in both email and its attachments such as text content and pictures via relatively mature file analysis technology, for example, steganalysis, as well as the set of characteristic attribute of email content generated by technology. Take the characteristic attribute of email content as input set of Naïve Bayesian Classifier, and the communication behavior relationship between the sender and the receiver as classified results. Then the classifier will be generated by training samples. Finally, this classifier will be applied to classify communication behavior relationship of the sender and the receiver of new email.

2. Data Collection

2.1. Data Collection

Some of email data from Enron email data sets and one security company, and some emails collected by ourselves at spare time, are selected 4,000 randomly as the test set, among which are selected 500 as training samples to classify. See the classification of communication behavior relationship of the sender and the receiver according to characteristic attribute in Table 1. There are 365 normal emails, 96 emails that one of two sides belongs to malicious attack, 22 emails that both sides belong to leaked relationship, 17 emails that both sides have no relationship.

Tab.1: Classification results of sample training emails.

	Normal email	Malicious email	Leaked email	Email with no characteristics
Total	365	96	22	17

2.2. Data characteristic of pre-extraction and processing

Characteristic attributes of email contents will be extracted and divided before classifying communication behavior relationship between sender and receiver of email. Firstly, texts and attachments in email are extracted, and characteristic attributes come the second by means of steganalysis. This method is to extract keywords from text, and to analyze hidden information such as pictures and videos in the attachments. The process of analysis on emails and extraction of characteristic attributes is shown in Figure 1.

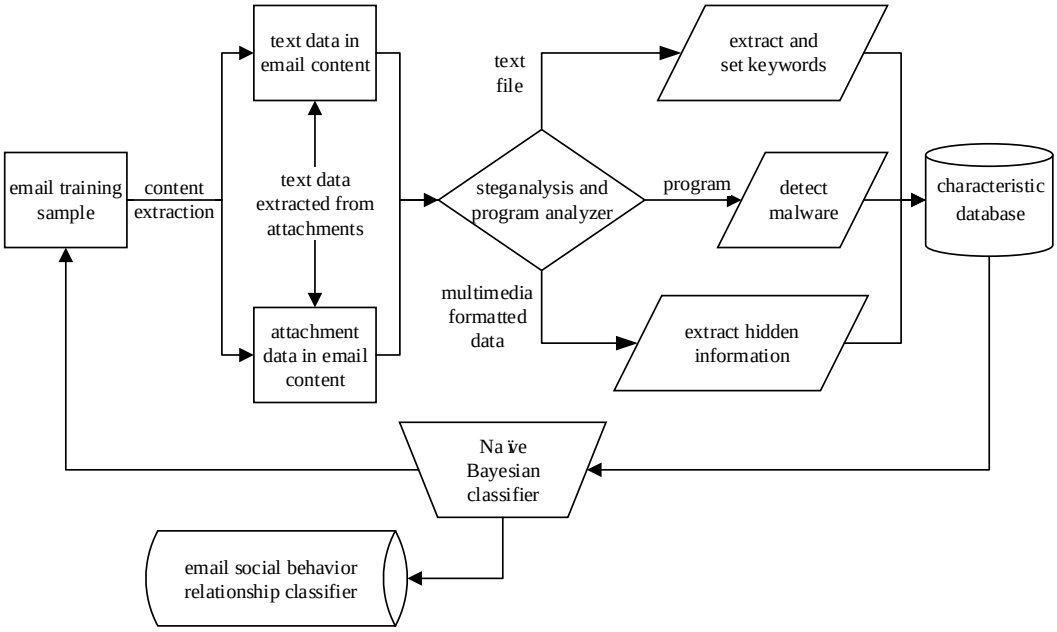

Fig.1: Process of analyzing emails and extracting characteristic attributes.

The detection process of texts and attachments in email will apply specific steganalysis and malware detection technology. Both of them have been mature, and own high reliability. The results are shown in Table 2. However, when malware is detected in attachments, it may not be a real one. Sometimes it may be an important file the receiver needs. Therefore, under such circumstance, other characteristic attributes should be analyzed together, so as to judge whether it is a malicious email or not.

Tab.2: Feature extraction results of sample training emails.

Email NO.	Number of keywords (a_1)	Malware	Number of leaked information (a_3)	Classification
1	1	Exist	0	Normal email
2	0	None	1	Normal email
3	6	None	6	Leaked email
4	2	None	1	Email with no characteristics
5	7	Exist	8	Malicious email
6	2	None	1	Normal email
…	…	…	…	…

3. Classify email behavior relationship based on Naïve Bayes

3.1. Naïve Bayesian classification principle

Naïve Bayesian classification is an easy and effective method, and applied widely. With excellent performance compared with Decision Tree and Neural Network Classification Algorithm, in some cases it is even better than other classifiers. Naïve Bayes fundamental principle is to calculate the probability that all types appear on the condition of items to be classified. The item with the biggest probability will be regarded as that type. The definition of Naïve Bayesian classification is as follows: set $X=\{a_1, a_2, \dots, a_m\}$ as an item to be classified; every a is the characteristic attribute of X, category set $C=\{y_1, y_2, \dots, y_n\}$, then calculate $P(y_1|x)$, $P(y_2|x)$, … , $P(y_n|x)$. If $P(y_k|x)=\{P(y_1|x), P(y_2|x), \dots, P(y_n|x)\}$, then $x \in y_K$, among which the key step is to calculate every conditional probability of $P(y_1|x)$, $P(y_2|x)$, … , $P(y_n|x)$.

To calculate conditional probability, firstly find out a known classification set to be classified, called training sample set; then count conditional probability of characteristic attribute in every category, $P(a_1|y_1)$, $P(a_2|y_1)$, … , $P(a_1|y_1)$; $P(a_1|y_2)$, $P(a_2|y_2)$, … , $P(a_m|y_2)$; … ; $P(a_1|y_n)$, $P(a_2|y_n)$, … , $P(a_m|y_n)$. If all characteristic attributes are in conditional independence, then it can be deduced according to Naïve Bayes:

$$P(y_i \mid x) = \frac{P(x \mid y_i) P(y_i)}{P(x)}$$

Denominator is always constant, so just maximize the member. Since all characteristic attributes are in conditional independence, hence it can be drawn:

$$P(x \mid y_i)P(y_i) = P(a_1 \mid y_i)P(a_2 \mid y_i)...P(a_m \mid y_i)P(y_i) = P(y_i)\prod_{j=1}^{m} P(a_j \mid y_i)$$

It can be reached that the key step of Naïve Bayesian classification is to calculate every conditional probability $P(a|y)$. When characteristic attribute is discrete value, counting frequency of appearing in every category among training samples, then $P(a|y)$ can be estimated. Naïve Bayesian classification process is shown in Figure 2:

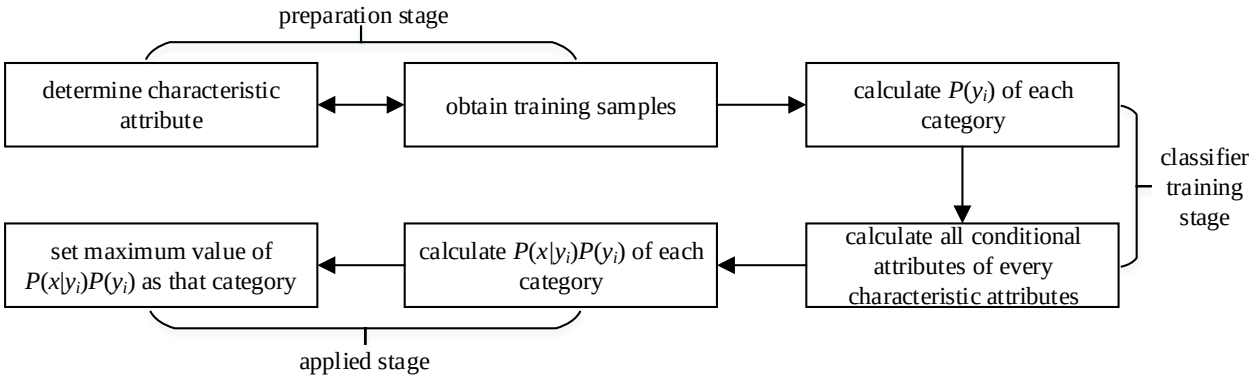

Fig.2: Naïve Bayesian classification process.

3.2. Extraction and classification of characteristic attributes

1) Characteristic of times that keywords appear

It is important to analyze email communication relationship via specific keywords in email. Corresponding keywords can be set at will and filtered in the detection process of email contents. In the experiment, unified keywords set has been used to avoid influencing results, and times that they appear are taken interval partition as 3 intervals: (0, 5), [5, 10), [10, +∞).

2) Characteristic of whether malware exists

Sometimes malware may exist in the email attachments, but it depends on the receiver system environment. It may be a real one but at times it may be an important file the receiver needs. Therefore, under such circumstance, other characteristic attributes should be analyzed together, so as to judge whether it is a malicious email or not. But anyway, it is an important parameter to identify communication relationship. {Exist, None} is to differentiate emails in the experiments.

3) Characteristic of times that leaked information appears

Emails can be used to pass on secret messages via multimedia files like pictures and videos, as well as other office files. There are many methods of file encryption such as hidden information in file storage structure, adding vital messages to postfix of file name, etc. During the detection process of the whole email content, analysis software can identify and find out the file with hidden information effectively, which can help judge entity relationship of both sides in email communication better. In the experiment, times that leaked information appears are also taken interval partition as 3 intervals: (0, 2), [2, 5), [5, +∞).

3.3. Classifier construction of email behavior relationship

If tests the set manually, it will cost lots of manpower and material resources, and it is poor efficient. On the contrary, if automatic monitoring mechanism can be introduced, work efficiency will be dramatically increased. Consequently, in order to classify email communication behavior relationship, classification types should be given: C=0 stands for normal communication; C=1 stands for one side belonging to malicious attack; C=2 stands for both sides belonging to leaked relationship; C=3 stands for both sides being in sensitive area. According to last part, 3 characteristic attributes are selected: a_1: times that keywords appear in file; a_2: whether malware exists in file; a_3: times that leaked information appears. Divide them in accordance with data characteristics: a_1: {0<a_1<5, 5<=a_1<10, 10<=a_1}, a_2: {a_2=0 (None), a_2=1 (Exist)}, a_3: {a_3<2, 2<=a_3<5, 5<=a_3}.

Take 500 emails extracted randomly as samples, and calculate frequency of every classification: $P(C=0) = 365/500 = 0.73$, $P(C=1) = 96/500 = 0.192$, $P(C=2) = 22/500 = 0.044$, $P(C=3) = 17/500 = 0.034$. Then, calculate all the frequency classified by characteristic attribute under every class condition:

$P(a1<5|C=0) = 0.63$, $P(5<=a1<10|C=0) = 0.25$, $P(a1>10|C=0) = 0.12$,

$P(a1<5|C=1) = 0.65$, $P(5<=a1<10|C=1) = 0.27$, $P(a1>10|C=1) = 0.08$,

$P(a1<5|C=2) = 0.07$, $P(5<=a1<10|C=2) = 0.19$, $P(a1>10|C=2) = 0.71$,

$P(a1<5|C=3) = 0.12$, $P(5<=a1<10|C=3) = 0.26$, $P(a1>10|C=3) = 0.62$,

$P(a2=0|C=0) = 0.81$, $P(a2=1|C=0) = 0.19$, $P(a2=0|C=1)=0.25$, $P(a2=1|C=1) = 0.75$,

$P(a2=0|C=2) = 0.73$, $P(a2=1|C=2) = 0.27$, $P(a2=0|C=3)=0.44$, $P(a2=1|C=3) = 0.56$,

$P(a3<2|C=0) = 0.65$, $P(2<=a3<5|C=0) = 0.23$, $P(a3>5|C=0) = 0.12$,

$P(a3<2|C=1) = 0.47$, $P(2<=a3<5|C=1) = 0.26$, $P(a3>5|C=1) = 0.27$,

$P(a3<2|C=2) = 0.06$, $P(2<=a3<5|C=2) = 0.16$, $P(a3>5|C=2) = 0.78$,

$P(a3<2|C=3) = 0.06$, $P(2<=a3<5|C=3) = 0.26$, $P(a3>5|C=3) = 0.68$,

Now, classifier construction has mainly been completed.

4. Experimental results and analysis

4.1. Sample training

Based on the previous conclusion, applied Naïve Bayesian classification method, 500 emails selected randomly taken as training samples, Naïve Bayesian classifier will be established. See characteristics and details of email entity relationship in Table 3, see part of sample training data in Table 4. Finally, train sample data and establish basic Bayes classifier model.

Tab.3: Classification of sample data.

Name	Number of keywords (a_1)	Malware	Leaked place (a_3)	Classification
Division	$a_1<5$	Exist	$a_3<2$	Normal communication
	$5<a_1<10$	None	$2<a_3<5$	Malicious attack
	$a_1>10$		$a_3>5$	Leaker email
				Email with no characteristics

Tab.4: Part of sample training data.

Email NO.	Number of keywords (a_1)	Malware	Leaked place (a_3)	Classification
1	$a_1<5$	Exist	$a_3<2$	Normal communication
2	$a_1>10$	None	$a_3>5$	Leaked email
…	…	…	…	…
499	$a_1<5$	None	$a_3<2$	Normal email
500	$a_1<5$	Exist	$a_3>5$	Attack email

4.2. Experimental results and analysis

Based on classifier model, our system will classify the following 3,500 emails data, and monitor the accuracy of classifier. With the increasing of email data examples continuously, the classification results are shown in Table 5 below:

It is shown in figure 3 that with the increasing of testing data, the classification accuracy is gradually to be steady as over 83%, which means classification results by classifier is reliable.

Tab.5: Assessments of classification results

Number of email	Number of correct classification	Rate
500	453	90.6%
1,000	873	87.3%
1,500	1276	85.07%
2,000	1683	84.15%
2,500	2101	84.04%
3,000	2516	83.87%
3,500	2934	83.82%

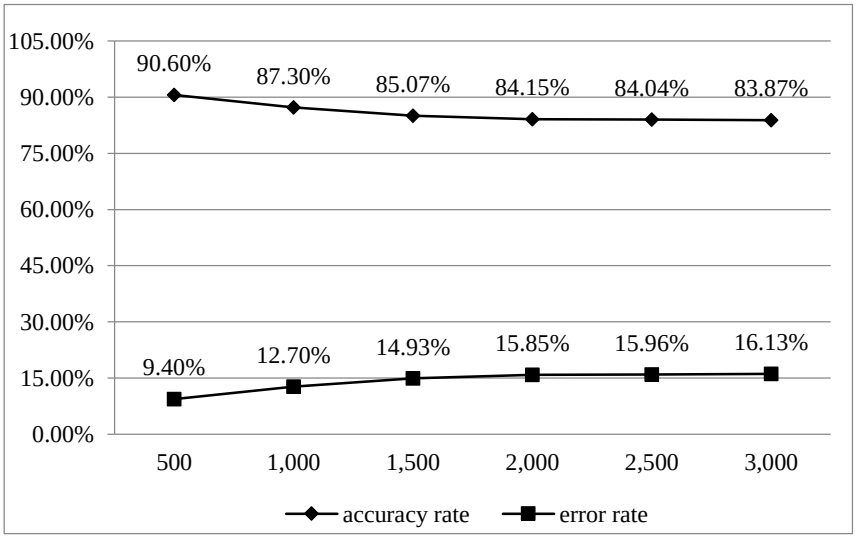

Fig.3: Line graph of Naïve Bayesian classification accuracy and error rate.

5. Conclusion

This paper aims to propose a new email communication security behavior analysis method. This method will be applied to analyze both sides of email communication relationship in the future, since it can analyze the relationship and behavior between sender and receiver of email accurately. At present, one simple system of classification system has been mainly completed, and has gained good impact in the field of email communication behavior. In future practical application, more characteristic attributes can be extracted in terms of requirements, and classified in more detail. Meanwhile, the accuracy of email communication security behavior analysis method can be enhanced via revised Bayesian classification.

6. Acknowledgements

This work has been supported by Projects U1536116 and U1636208 funded by National Natural Science Foundation of China (NSFC).

7. References

[1] Yoo J Y, Yang D. Classification Scheme of Unstructured Text Document using TF-IDF and Naïve Bayes Classifier[J]. 2015.

[2] Banday M T, Sheikh S A. Multilingual email classification using Bayesian filtering and language translation[C]//Contemporary Computing and Informatics (IC3I), 2014 International Conference on. IEEE, 2014: 696-701.

[3] Tang B, He H, Baggenstoss P M, et al. A Bayesian classification approach using class-specific features for text categorization[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(6): 1602-1606.

[4] Androutsopoulos I, Koutsias J, Chandrinos K V, et al. An Experimental Comparison of Naïve Bayesian and

Keyword-Based Anti-Spam Filtering with Personal E-mail Messages [J]. Tetsu-to-Hagane, 2015, 97(2):35.

[5] Bahgat E M, Rady S, Gad W. An e-mail filtering approach using classification techniques[C]//The 1st International Conference on Advanced Intelligent System and Informatics (AISI2015), November 28-30, 2015, Beni Suef, Egypt. Springer International Publishing, 2016: 321-331.

[6] Li X, Luo J, Yin M. E-Mail Filtering Based on Analysis of Structural Features and Text Classification [C]// Intelligent Systems and Applications (ISA), 2010 2nd International Workshop on. IEEE, 2010:1-4.

[7] Berend D, Kontorovich A. A finite sample analysis of the Na ïve Bayes classifier [J]. Journal of Machine Learning Research, 2015, 16: 1519-1545.

[8] Boryczka U, Probierz B, Kozak J. A New Algorithm to Categorize Email Messages to Folders with Social Networks Analysis [M]// Computational Collective Intelligence. Springer International Publishing, 2015.

[9] Yoshinaga N, Itaya S, Tanaka R, et al. Content Propagation Analysis of Email Communications [C]// Ieee/wic/acm International Conference on Web Intelligence and Intelligent Agent Technology. 2010:79-82.

[10] Ma Y, Yu Z, Ding J. A Method of User Recommendation in Social Networks Based on Trust Relationship and Topic Similarity [M]// Social Media Processing. Springer Berlin Heidelberg, 2014:240-251.

[11] Chandrasekar P, Qian K. The Impact of Data Preprocessing on the Performance of a Na ïve Bayes Classifier[C]//Computer Software and Applications Conference (COMPSAC), 2016 IEEE 40th Annual. IEEE, 2016, 2: 618-619.

[12] Yang T, Qian K, Lo D C T, et al. Spam filtering using Association Rules and Na ïve Bayes Classifier[C]//Progress in Informatics and Computing (PIC), 2015 IEEE International Conference on. IEEE, 2015: 638-642.

[13] Li L, Li C. Research and Improvement of a Spam Filter Based on Na ïve Bayes[C]//Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2015 7th International Conference on. IEEE, 2015, 2: 361-364.