# Analysis and Comparison of N-gram_IDF Algorithms for Intrusion Detection

Dai Geng, Thmohiro Odaka, Jousuke Kuroiwa, Hisakazu Ogura
Graduate School of Engineering, University of Fukui, Fukui, Japan

## 1. Introduction

Masquerade Detection is one of the most serious computer security problems. The typical masqueraders refers to that the unauthorized user, through illegal means, gains the system entering privileges to impersonate entering computer systems as a legitimate user, so as to realize purposes of stealing data or performing malicious operations. As masquerade intruders may be outsiders who illegally enter the system, and they may also be insiders who abuse of authorities, the traditional methods of security detecting and certification are not enough to find masquerade attacks. Therefore, it has become a hot current study to use on-line intrusion detection system (IDS) to detect users' abnormal behaviors.

In general, IDS research focuses on how to set up characteristics of users' behaviors to effectively detect anomalies. It compares user's recent behaviors with his historical behavior characteristics to determine whether there is abnormal invasion, and all these acts can be done by audit data that are automatically recorded by computer and analysable. Currently, the majority of IDS create models by the probability characteristics of two kinds of users' behaviors: one is to consider the frequency characteristics---the frequency of user's behavoirs, such as, Uniqueness, Match sequence, SVM, AMM and so on; the other is transition characteristics---the variability of user's behaviors, such as Bayes, HMM, IPAM, Compression, etc. However, the above two methods have deficiencies of low detecting accuracy, high time complexity, poor results explanation, so they can not meet the requirements of IDS. In this paper, we propose a new intrusion detecting method which generates user's characteristics and gives them different weights on the basis of the frequency and length of user command sequences. This method takes into account not only the frequency information of user's behaviors in each sequence, but also the distribution features of user's behaviors in the entire data. And this method is derived from the well-known N-gram model and the TF-IDF formula in the information retrieval and text mining. As for the Schonlau[1] data, compared with the experiment results of other methods, this method can significantly improve the detection accuracy of IDS.

## 2. Method

Establishing Characteristics is usually the first step in intrusion detection. Here we use the N-gram model to establish characterisitcs for each user. N-gram model is one of the most important method of the natural language processing, in the masqurade detection, N-gram stands for n pieces of consecutive commands, for example 1-gram refers to a single command, 2-gram is a command unit consisting of continuous commands. N-gram model of the intrusion detection is divided into two steps, namely the establishment of the user's N-gram features in the learning phase, and N-gram feature extraction of the sequences to be detected in the detecting phase. As for the former question, it is relatively simple; we extract the user's N-gram features, and in accordance with frequencies of N-gram appeared in all users, we extract those whose frequencies are greater than k as the user's N-gram features. As for the latter issue, we have adopted a strategy of the largest positive match division which is widely used in Chinese character segmentation. We scan the sliding windows with beginning length of n = 3 one by one in accordance with command and we check whether the N-gram of each sliding windows to be detected exists in the system frequency characteristics. If it exists, the window slides back n pieces of commands and repeats the initial action; if it does not exist, then we change the length of the window into n-1, and continue testing until the n = 1.After the division of commands to be detected, there appears a series of combinations of n-gram, which reflects the feature that the longer the length of ngram divided into the higher the priority.

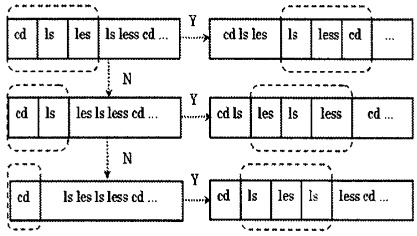

Fig1. The process of command session segmentation

After the observed command sequences are segmented, we need to calculate the weight values of the various N-gram combinations in order to judge a possible masquerader invasion. The main motivations for this method come from the following obvious conclusions: First of all, a user may input certain command sequences frequently to accomplish a specific task, therefore these sequences appear in the user command units more frequently, thus if a command sequence contained in an observed sequence is similar to the current user's historical record, it is likely to fall within the normal input. Secondly, the input habits and specific tasks of a user are unique, namely, some of the command sequences are unique, and other users either use them less frequently or do not use them at all. If the command sequence exists in the observed sequence, it is probably normal input. Thirdly, some of the command sequences under the UNIX environment are basic operations for the system, and these command sequences have the possibility of being used by all users, so they cannot represent a particular user's input features. From the above, we can see that due to the diversity of command sequences input by users, the relative user's characteristics are also various, therefore, it is one-sided to only consider whether certain an n-gram in the user's characteristics exists in the command sequences to be tested as a determining rule, we need comprehensively examine all the N-grams produced in the classification in the command sequences, and must have a formula to calculate the similarity between the command sequences and the current user's characteristics.

Based on the above ideas, We use the TF-IDF model to calculate the weight of each characterisitcs in this article. The TF-IDF algorithm was proposed by Karen Sparck Jones in 1972. He properly applied the Information Entropy to the information retrieval. Currently, this method also has a wide range of applications in searching, text classification and other related fields. In the study of Masquerade Detection, the form of TF-IDF is

$$tf_i idf = f_i \times \log\left(\frac{n}{s_i}\right) \qquad (1)$$

Here, TF-IDF is based on the underlying principle that the more times a command unit appears in a user, the more it is likely to characterize that user. Therefore, if TF-IDF is used as a calculating measure to detect, it can reflect the user's own characteristics.

Table 1 The notation and terminology

| | |
|---|---|
| $n$ | Total number of users in the observation data set |
| $m_k$ | Total number of commands in the observation session $k$ |
| $s_i$ | Total number of users who use the command sequence $i$ |
| $f_i$ | Frequency of command sequence $i$ in observation data set |
| $f_{ij}$ | Frequency of command sequence $i$ in observation user $j$ |
| $f_{ik}$ | Frequency of command sequence $i$ in observation session $k$ |
| $c_{kj}$ | Total number of commands which matched with user $j$ in observation session $k$ |

In the command-based Masquerade Detection, the frequency of commonly used commands are often high, but in essence they are not able to represent the user's characteristics, so in order to reduce the influence of these commands frequency to the detection accuracy, We propose stf-idf in this paper, its form is

$$stf_i idf = \sqrt{f_i} \times \log\left(\frac{n}{s_i}\right) \quad (2)$$

To calculate the characterisitcs weight based on the stf-idf, we must take into account three such factors:

1. The importance of characteristic sequences to be detected in the current users, namely, it should be different for the importance of the same characteristic sequence to different users, the formula is as follows:

$$stf_{ij} idf_i = \sqrt{f_{ij}} \times \log\left(\frac{n}{s_i} + 1.0\right) \quad (3)$$

2. The importance of characteristic sequence in the current sequence to be detected, the formula is as follows:

$$stf_{ik} idf_i = \sqrt{f_{ik}} \times \log\left(\frac{n}{s_i} + 1.0\right) \quad (4)$$

3. The matching degree between the sequence to be detected and the current user to be detected. the higher the matching degree, the more likely it is a normal user; if the matching degree is 0, the sequence to detected should be judged as the invasion sequence.

$$n_{kj} = \frac{c_{kj}}{m_k} \quad (5)$$

we construct a formula for these three factors to reflect the mutual relations among them.Taking into account the impact of sequences to be detected to weights, we standardize the component, and prescribe the scope of various weights between 0 and 1 to get the following formula:

$$\left[\sum_{i=1}^{m_k}\left(\frac{stf_{ij} idf_i}{\sqrt{\sum_{i=1}^{m_k}\left(stf_{ij} idf_i\right)^2}} \times \frac{stf_{ik} idf_i}{m_k}\right)\right] \times \left(n_{kj}\right)^2 \quad (6)$$

And we use the formula to calculate the weights of each command unit divided from the sequences to be detected, and then add all those weights, the results obtained is the score of sequences to be detected in the current user.

## 3. Experiment

Schonlau et al. collected a data set of shell command sequences using UNIX's acct auditing mechanism for masquerades. By convention, an experiment conducted on this data set is called the Schonlau et al. (SEA) experiment.

In general, there is a trade-off between the false alarm rate and the missing alarm rate in this experiment. Therefore, receiver operating characteristic (ROC) curves are used to evaluate this study. ROC curves are created by varying the threshold in the approach, so that the false alarm rate and the missing alarm rate at each point could be represented.

Figure 2 shows that the results of the N-gram_TF-IDF method yield a better performance than the other methods. Furthermore, this method has a masquerade detection accuracy of 91.97% when the false alarm rate is 5.08%, and an accuracy of 64.2% when the false alarm rate is 1.0%. Customer grammar has been reported as the most effective masquerade detection method to date. We could also see that for the SEA configuration, these two methods result in nearly the same detection efficiency and are also superior to previous methods, remarkably at different trade-off points. It should be noted that different methods achieve their minimum cost at different trade-offs, For the n-gram_TF-IDF, the minimum cost is 0.128 when the FA rate is 8.03%, thus, the n-gram_TF-IDF method achieves its minimum cost at a particular trade-off value of FA and MA.

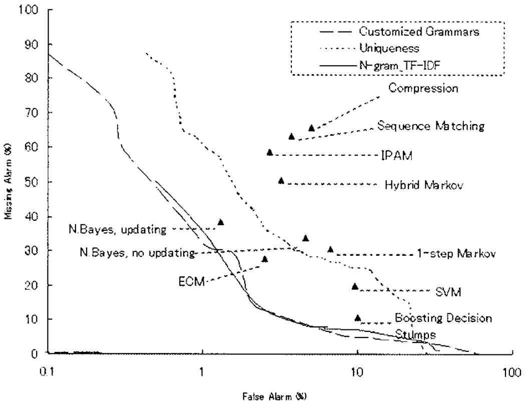

Fig2. ROC curves for the n-gram_TF-IDF method on the SEA experiment

## 4. Discussion and Conclusion

Through the above experiment, it can be seen that the N-gram_TF-IDF based classifier is effective, for it took into account several crucial factors in the intrusion detection. What's more its starting point is that those characteristic sequences which are frequently used by the current user while rarely used by other users can characterize the user's behaviors, therefore, these characteristic command sequences are given greater weight, to the contrary, although the frequency of those command sequences commonly used by most users is high, they make no substantive contribution to distinguish different users, hence they get very little weight. These two features of classifier make the detection accuracy greatly increased. In addition, feature extraction based on n-gram also contributes to the improvement of detection accuracy. Compared to those methods who consider only a single command frequency, or the transition probability, the method takes into account the relationship between frequency and the the transition probabilities, that is, the longer the length of the characteristic sequence, the more special it has. They represent a specific user behavior, so they can more express the user's own inherent characteristics.

In this article, we propose a new intrusion detection method. Firstly, we established user's ngram features, then we use the classifier to detect anomalies, and experiments show that the method improves the detection accuracy. It is a very difficult task for masquerade intrusion detection, so how to combine the frequency characteristics with transition characteristics more effectively to improve the detection accuracy is an important content for further study.

### References

[1] Schonlau M.,DuMouchel W.,Ju W-H.,Karr A F.,Theus M.,Vardi Y.Computer intrusion: detection masquerades. Statistical Science, 16(1):58-74, 2001.