Research of Multi - Pattern Matching Algorithm Based on Characteristic Value

Yansen Zhou
Department of Information Science and Technology,
University of International Relations
Beijing, China
e-mail: zhouyansen@126.com

Guanqi Ding
Department of Information Science and Technology,
University of International Relations
Beijing, China
e-mail: Dingguanqi2015@sina.com

***Abstract*—Large-scale and high-speed networks cause rapid increase in network flow, resulting in the omission of reports of the intrusion detection based on the string pattern matching algorithm. In this paper, a multi pattern matching algorithm C_MPM based on characteristic value is proposed. In the algorithm, the rule base is classified two times according to the length and the characteristic value, and obtains various pattern string queues, which can reduce the number of string matches that each text string needs to match each time; Secondly, the characteristic values of text strings of different lengths in the matching window are calculated in parallel, and the pattern string queues in the rule base are searched according to the length and the characteristic values of the text string, so that the full mode pattern matching can be carried out. The experimental results show that the time performance of the matching algorithm C_MPM is better than that of the AC_BMSH2 algorithm.**

Keywords- intrusion detection; characteristic values; XOR operation; multi-pattern matching; matching window;

I. INTRODUCTION

Network intrusion detection is an active defense technology, which monitors the data packets in the network. When a variety of intrusion attempts and attacks are found, appropriate responses are taken to protect the confidentiality, integrity and usability of system resources. According to the different data analysis methods of intrusion detection system, they can be divided into two categories: characteristic detection and anomaly detection [1]. Characteristic detection is the mainstream of current detection systems, and the detection efficiency is mainly affected by the organization of the intrusion rule base and the pattern matching algorithm used. However, with the deployment of large-scale and high-speed networks, the following problems arise in the characteristic detection, such as the serious packet loss rate and the frequent omission.

In the network intrusion detection system, CPU and other system resources are mainly consumed in string pattern matching, which restricts the time performance of the system detection. According to statistics, at present, the pattern matching time of the mainstream intrusion detection system occupies more than 80% of the whole detection time [2]. Therefore, the time performance of the pattern matching algorithm determines the performance of the whole intrusion detection system. Of course, the rational organization of the intrusion rule base can also improve the time performance of intrusion detection. In this paper, the pattern matching algorithm and the organization of intrusion detection rule base are mainly researched and improved.

II. CHARACTERISTIC PATTERN MATCHING ALGORITHM

A. Pattern Matching Algorithm

Current mainstream matching algorithms are mainly single-mode and multi-mode matching algorithms. The multi-mode matching algorithm is that every pattern string can be matched with the text string at a time. N pattern strings only need to be matched once, which has higher matching efficiency, but the idea of the single-mode matching algorithm and the larger memory space are needed [3]. The single-mode matching algorithms mainly include KMP, BM, BMSH and BMSH2, etc. And the multi-mode matching algorithms mainly include AC, WM and AC_BMSH2, etc. These algorithms have one common feature: accurate matching one time. Their idea is to match the pattern string directly in the text string. If they don't match, according to the mobile strategy of the algorithm itself, then move the pattern string to skip a certain amount of text string characters and then continue matching.

B. Related Definitions

1）Definition 1: String characteristic value

A value obtained by an operation of a string, called the characteristic value of that string. As with the computing result of hash function, the lengths of characteristic value of different length patterns are the same.

Because of the need of the simplicity of the computation [4], multiple strings may be computed for the same characteristic value, producing a multi-to-one relationship, which is called collision. But in any case, a string has and only one characteristic value, and multiple strings correspond to the same characteristic value at a certain probability. In order to reduce the computation time of characteristic value, we need to choose reasonable and simple characteristic value operating function, and to reduce the probability of the different strings having the same characteristic value, which can reduce the number of the pattern strings that are needed by the text string of the matching window.

2）Definition 2： bit vector

For a string of arbitrary length, the binary string formed by the bits on the same position of each character according to the order of the characters is called a bit vector. For

978-1-5090-6352-9/17/$31.00 ©2017 IEEE

example, each character's ASCII code of the string "BEST" is {01000010, 01000101, 01010011, 01010100}, setting the bit vector composed by the lowest order bit of each character as {0110}. The string contains 8 bit vectors whose lengths are 4, and they are {1111, 0011, 0000, 0110, 0101, 1010, 0110}. The characteristic value calculated from the bit vectors are called bit characteristic value and are marked with the C. 8 bit characteristic value forms the characteristic value of the string C, C=[c_7、c_6、c_5、c_4、c_3、c_2、c_1、c_0]. In order to reduce the kinds of characteristic value, each group of the bit vectors whose lengths are usually XORed to get feature vectors of one bit, so that the final characteristic value of character strings of any length are 8bits [5].

Definition 3： Rate of Pattern String Match

The ratio of the number of pattern strings in the matching window used in each match to the total number of pattern strings is called the match rate. In general, the less the number of pattern strings to compare each time, the better the time performance of the matching algorithm. Matching time performance is the best when the characteristic value of the matching window text string does not have a corresponding pattern string in the pattern string feature base.

C. The Idea of Characteristic Value Matching Algorithm

The idea based on matching algorithm of the string characteristic value design is: The characteristic values of the text string in the matching window are compared with the characteristic values of the equal length pattern string [6], and if they are not equal, the two strings do not match; If they are equal, the two strings will match with a large probability, which is the need for an exact match for further confirmation. Matching process is using the matching algorithm two times. The first time is to get the characteristic value of the text strings of different lengths by computing the characteristic value in the matching window in parallel, and then location the position of the pattern string queue for subsequent matching; If the fixed pattern string queue is not empty, then second matches are needed, and the pattern string in the queue is extracted in turn to match the text string in full mode. The first matching requires that the computation of the characteristic value is simple and fast, so as to improve the overall matching performance. Therefore, the XOR operation is implemented directly by hardware to alleviate the burden of CPU. The two classifications of the rule base and the parallel computation of the character string of the text string are the keys to the improvement of the algorithm. The paper mainly aims at these two aspects.

D. Calculation of Characteristic value

The selection of method of calculating string characteristic value should be accord with the following requirements:

1) The calculation is simple and can be achieved directly by the hardware circuit. It is better to calculate the characteristic value of text strings in different lengths in the same matching window at one time;

2) The characteristic value calculated from the strings of the same length has low collision probability. That is, strings with the same characteristic value at the same length should be less. It reduces the number of pattern strings that a text string needs to match and effectively reduce the number of times needed for each match;

3) The characteristic value of the pattern strings after the horizontal movement in the matching window can be obtained by very little calculation of the characteristic value before horizontally moving and the computation times of the characteristic value are reduced, so as to improve the time performance.

Based on the above requirements of characteristic value calculation, this paper uses XOR evaluation algorithm for strings, and the algorithm evaluates the characteristic value in line with the above three requirements [7].

The XOR characteristic value of a string is obtained by XOR operation of each character composed of string. Set the string to S={S1, S2, ..., Sm}, then the characteristic value of the string is$\{S_1 \oplus S_2 \oplus S_3 \ldots \oplus S_m\}$ of 8 bits. For example, the ASCII codes corresponding to the characters in the pattern string "PAT" are {01010000, 01010000, and 01010100}. Its characteristic value is C={P ⊕ A ⊕ T}= {00010100}.

If the probability of occurrence of 0 or 1 of all bits is the same, for any lengths of text strings, the characteristic value is 8bit. For strings of the same length, the probability that the characteristic values are the same is 1/256, that is, the collision rate is 1/256. This is equivalent to converting multi-pattern matching into single-pattern matching directly, which improves the performance of matching time effectively [8].

The characteristic value after horizontal movement of the text string of the matching window can be obtained by the two XOR operations of the characteristic value before horizontally moving and the two characters of input and shift out from the text. That is to say, the characteristic value of the text string of the backward matching window can be obtained only by two operations. For example, when "DDOS" matches the text string "SMURF", the characteristic value C1=[S ⊕ M ⊕ U ⊕ R]=[0001 1001] of "SMUR" is firstly calculated, and the characteristic value of "DDOS" is C2=[D ⊕ D ⊕ O ⊕ S]=[0001 1100], they do not match. The matching window of text string shifts one character to right into "MURF". The characteristic value of matching window is [0000 1100] which is derived by [C_1 ⊕ S ⊕ F] two times of XOR operation, not equal to C2=[0001 1100], from which the conclusion can be drawn that string 'SMURF' does not contain the pattern string 'DDOS', and the match ends.

According to the principle of parallel XOR operation, when the string length is m, the first calculation of characteristic value needs (m-1) times of XOR operations, then after a horizontal movement of one character of matching window, the calculation of characteristic value only needs 2 XOR operations, so as to get the characteristic value of each text strings in the matching window.