A Secured Pattern Matching Technique for Intrusion Detection System in Wireless Sensor Network

R.Hamsaveni
Research Scholar, SCSVMV University, Kanchipuram. hamsamurugs@gmail.com, cell:9442212136

Dr.G.Gunasekaran,M.E.,PH.D.(Engg)
Principal, Meenakshi College of Engineering,Chennai.gunaguru@yahoo.com cell:9444177718

Abstract-The evaluation of Wireless Sensor Networks (WSN) for performance is the most familiar research area and there are various kinds of possibilities in which security can be breached and enable the hackers to utilize secured information. The Intrusion Detection System (IDS) is one of the major defensive methods used against attack in WSN. There are three types of methodologies used for detect intrusion on the network, they are Signature based, Anomaly based and Stateful protocol analysis. This paper is based on Signature based Intrusion Detection System methodology. String matching algorithms are essential for IDS that filter packets and flows based on their payload. This work describes the concept of single keyword pattern matching algorithms. A Logo Pattern Matching algorithm (LPM) is proposed. The new method reduces character comparisons, faster and more reliable in network security applications. Many pattern matching algorithms have high false alarm. The proposed LPM scheme reduces false alarm percentage to an extensive level and detects the intruder efficiently. The experimental results implemented in SNORT and show that the new algorithm is highly efficient. Its search time is cut down significantly compared with other popular existing algorithms and its memory occupation stays at a low level. Moreover, conclusion on results is made and direction for future works is presented.

Keywords - Signature Based, Wireless Sensor Network, Intrusion Detection System, Logo Pattern Matching Algorithm.

I. INTRODUCTION

Security attacks through internet have proliferated in recent years. Hence, information security is an issue of very serious global concern of the present time. The need for network security and in particular the need for Intrusion Detection Systems (IDS) has been brought out. IDSs, as originally introduced by Anderson in 1980 and later formalized by Denning in 1987, have received increasing attention in the recent years.

IDS are widely used and heavily depended upon. The continued growth in both Network traffic and intrusion signature databases make the performance of these systems increasingly challenging and important. Intrusions refer to the network attacks against vulnerable services, data-driven attacks on applications, host-based attacks like privilege escalation, unauthorized logins and access to sensitive files, or malware like viruses, worms and Trojan horses. Intrusion detection means detecting unauthorized use of a system or attacks on a system or network.

IDS are implemented in software or hardware in order to detect these activities. IDSs collect information from a computer or a computer network in order to detect attacks and misuses of the system. Three types of data are used by IDSs. These are network traffic data, system level test data and system status files. The heart of almost every modern IDS has a string matching algorithm. The IDS uses string matching to compare the payload of the network packet and/or flow against the pattern entries of intrusion detection rules.

String matching is an important area in the wider domain of text processing. These algorithms are basic components used in implementations of practical software’s existing under most operating systems. Moreover, they emphasize programming methods that serve as paradigms in other fields of computer science. They also play an important role in theoretical computer science by providing challenging problems. String matching generally consists of finding a substring (called a pattern) within another string (called the text).These string matching algorithms are used to inspect the content of packets and identify the attacks signature in IDS. String matching consists of finding one, or more generally, of all the occurrences of a search string in an input string. In IDS applications, the pattern is the search string, while the payload is the input string. If more than one search string simultaneously matches against the input string, this is called multiple pattern matching. Otherwise, it is called single pattern matching. In this work, we
This paper presents the information about IDS pattern matching algorithms. A Logo Pattern Matching Algorithm is proposed to reduce the number of attempts and character comparison. The main aim for this work is to improve the efficiency of IDS detection engine. In the rest of the paper, we presented Related Work for IDS (Section II), the Architecture of SNORT (Section III) Result (Section IV) Conclusion (Section V).

II. RELATED WORK

Before we go in detailed study of our algorithm and its comparisons with other algorithms, we would like to give a brief overview of the already existing algorithms along with their complexity analysis and try to explain why our algorithm performs better than them.

Bruce W. Watson (2002) gave a more practical algorithm for a special characterization of ‘matching productions’, using the transitive closure of a relation to deal with chain rules in the pattern grammar. The algorithm was further improved through the introduction of a reverse trie. The idea of shift distances greater than one symbol (as in the Boyer–Moore and Commentz-Walter algorithms) was also introduced [1].

V. K. Pachghare, Dr. Parag Kulkarni, (2008) proposed an efficient pattern matching algorithm in order to overcome the troubles in network traffic [2].

Dai Hong (2012) proposed enhanced pattern matching performance using improved Boyer Moore Horspool Algorithm. Moreover the algorithm combines the deterministic finite state; the improved Boyer Moore Horspool Algorithm takes full use of the matching information to skip several characters. The proposed algorithm saves more memory resource, especially adapts to pattern string character sets and text character sets, matching speed increases greatly and pattern length and number influence hardly [3].

Priya Jain, Shikha Pandey (2012) have done comparative study on various existing Pattern Matching Algorithms and confirmed that the Bayer Moore Pattern matching algorithm is the most efficient as well as fast and gives the accurate results [4].

Abhay Bhatia, Shashikant, Robin Choudhary (2012) exposed the challenges in pattern matching and shows the comparison between the existing and a proposed Optimized Pattern Matching (OPM) algorithm for finding out the matched links with the given number of links [5].

Lata , Kashyap Indu and Nagaraju (2013 and 2015) identified intrusion can be possible on the header part or payload part. Intrusion detection algorithms normally have high false alarm. The authors have proposed an algorithm which lowers the false alarm considerably [6].

K. Prabha, S. Sukumaran (2014) derived an ISPMA algorithm for single keyword pattern matching for IDS. They have compared their result with Boyer-Moore Algorithm, Horspool, Karp-Rabin algorithm, Brute force algorithm and illustrated that their ISPMA algorithm is faster and more reliable [7].

Akinul Islam Jony (2014) illustrates a widely used multiple string patterns matching algorithm. A theoretical and experimental result along with the analysis and discussion of the algorithms was presented [8].

Urmila Patel, Mitesh Thakkar (2014), proposed an efficient version of Bidirectional Pattern is An Efficient Exact Single Pattern Matching (EESP) algorithm in which they tried to reduce pre-processing time and also found its all occurrences of the Pattern in to long Text String [9].

Nguyen Le Dang, Dac-Nhuong Le, and Vinh Trong Le (2016), presented a new algorithm for multiple-pattern exact matching. In their algorithm they reduced character comparisons and memory space based on graph transition structure and search technique using dynamic linked list. Theoretical analysis and experimental results, when compared with previously known pattern-matching algorithms, was highly efficient in both space and time [10].

II. SINGLE KEYWORD PATTERN ALGORITHMS

The Boyer-Moore algorithm and its variants are widely used in the string matching. The Horspool algorithm performs the comparison in a simple way, which works for most of the practical cases. The Brute Force algorithm requires no preprocessing of the pattern. In the Karp-Rabin Algorithm, the main idea is that instead of using comparisons it involves mathematical computations which more specifically extends to the notion of hashing concepts.

3.1 Boyer-Moore (BM) Algorithm

The Boyer-Moore algorithm is one of the exact string matching algorithms that used in single pattern matching. The algorithm uses two tables or functions, which is used to move the sliding window to the right. The first table is called “bad character shift”, while the second table called “good suffix shift”. The algorithm is faster when it is working with small pattern size, but it is slower when it is working with large pattern size [7]. The BM algorithm is given below:

Algorithm BoyerMooreMatch(T, P, S)
been a useful approach when it comes down to string pattern matching algorithms. The application of hashing mathematical computations which more specifically extends the notion of hashing. The main idea is that instead of using comparisons it involves dependent expression and from the bmGs table for a character in the pattern. On the other hand, a good-suffix shift (bmBc) and Boyer-Moore good-suffix (bmGs) tables. For each character in the alphabet set, a bad character table (bmBc) is maintained. The maximum of the shift value between the bmBc and bmGs tables, which are known as Boyer-Moore bad character shift and good suffix shift, the expected number of character comparisons in the character set is large and the length of the pattern. Hence, the algorithm is more efficient in practical situations where the alphabet size is large and the length of the pattern is small.

3.2 Horspool (HP) Algorithm

Horspool algorithm is based on Boyer Moore algorithm. Snort IDS uses a modified version of the algorithm called Boyer-Moore-Horspool algorithm to maintain memory usage and speed up during searching phase. Unlike Boyer-Moore algorithm, which uses two tables; bad character shift and good suffix shift, the Horspool algorithm uses only one table (bad character shift) . Hence, the algorithm is more efficient in practical situations where the alphabet size is large and the length of the pattern is small.

3.3 Karp-Rabin (KR) Algorithm

The Karp-Rabin Algorithm was created by Michael Rabin and Richard Karp. They used a completely different approach than the single keyword methods. The main idea is that instead of using comparisons it involves mathematical computations which more specifically extends to the notion of hashing. The application of hashing (converting each string into a numeric value) has always been a useful approach when it comes down to string matching. If both words have different hash values then we conclude they are different. But if their hash values are the same we cannot conclude they are the same string and will have to perform further comparisons.

3.4 Brute Force (BF) Algorithm

The Brute Force algorithm requires no preprocessing of the pattern. The comparison can be done in any order either from left to right or from right to left. If all the characters match, then it is said to be a match. If not, the algorithm shifts the pattern by exactly one position to the right. The expected number of character comparisons in Brute Force algorithm is 2n. The algorithm is given below:

Algorithm 1. Boyer-Moore algorithm

The algorithm preprocesses the pattern and creates two tables, which are known as Boyer-Moore bad character (bmBc) and Boyer-Moore good-suffix (bmGs) tables. For each character in the alphabet set, a bad character table stores the shift value based on the occurrence of the character in the pattern. On the other hand, a good-suffix table stores the matching shift value for each character in the pattern. The maximum of the shift value between the bmBc (character in the text due to which a mismatch occurred) dependent expression and from the bmGs table for a matching suffix is considered after each attempt, during the searching phase. This algorithm forms the basis for several pattern matching algorithms.

Algorithm 2. Horspool algorithm

Horspool algorithm uses only one table (bad character shift) . Hence, the algorithm is more efficient in practical situations where the alphabet size is large and the length of the pattern is small.

Algorithm 3. Karp-Rabin Algorithm

Algorithm 4. Brute Force Algorithm

3.5. Knuth-Morris-Pratt Algorithm (KMP)

Knuth-Morris-Pratt algorithm proposed a keyword pattern matching algorithm (herein the KMP algorithm) that runs left to right over its text input in linear time and improves on an algorithm proposed earlier by only Morris and Pratt . This algorithm’s pre-computation creates an array with information about how the keyword matches against shifts of itself. For example knowing we have matched exactly r characters somewhere in the input allows us to determine that certain shifts are invalid; thus, avoiding the shifts that
the naive brute force algorithm executes only to then fail at a subsequent match attempt.

As a concrete example let us say the keyword pattern \( x = ababaca \) and we have just matched 5 characters in \( y \) starting at position \( i \) only to find a mismatch for the character \( c \) at position \( i + 5 \). The naive shift of the brute force algorithm would shift the pattern by one position, but of course it would fail because the character \( b \) at \( y[i + 1] \) would not even match the first character of the keyword \( x \). The KMP algorithm knows what the appropriate shift should be, and may shift over multiple positions without missing any potential matches. The trick to doing this correctly is in the pre-computation step where the prefix function for the keyword is built into an array. If we define \( x \) as the first \( r \) characters of \( x \) that have been matched at any point, the prefix function array table at position \( r \) (table[r]) contains the length of the longest prefix of \( x \) that is a proper suffix of \( x \). Using this information the shift to the next position is always possible to calculate as (the current position) + (the number of matched characters before the mismatch (\( r \)) - (table[r]) .

Algorithm 3.5 below shows the KMP algorithm and the pre-computation step of how the prefix function array (table) is created in time and with memory space \( O(m) \). In the worst case the matching phase of the KMP algorithm executes \( 2n - 1 \) character comparisons.

**Procedure KMP**(\( x, m, y, n \))

- **Input:**
  - \( x \) ← array of \( m \) bytes representing the keyword
  - \( m \) ← integer representing the keyword length
  - \( y \) ← array of \( n \) bytes representing the text input
  - \( n \) ← integer representing the text length
- table ← Compute Prefix KMP(\( x, m \)) // Pre-Computation

**Procedure Compute Prefix KMP**(\( x, m \))

- **Input:**
  - \( x \) ← array of \( m \) bytes representing the keyword
  - \( m \) ← integer representing the keyword length
- table ← newArray[\( m + 1 \)] // Prefix Shift Table
  - table[1] ← 0
  - \( k \) ← 0
  - for \( q = 2 \rightarrow m \) do
    - while \( k > 0 \) and \( x[k + 1] = x[q] \) do
      - \( k ← table[k] \)
    - end while
    - if \( x[k + 1] = x[q] \) then
      - \( k ← k + 1 \)
    - end if
  - table[q] ← \( k \)
- end for
- return table
- end procedure

**Algorithm 5. Knuth-Morris-Pratt Algorithm**

**Main points**- The main points of the Knuth-Morris-Pratt algorithm are outlined below:
- Performs the comparisons from left to right.
- Preprocessing phase in \( O(m) \) space and time complexity.

**IV. Architecture of Signature Based Network Intrusion Detection System using SNORT**

4.1 Method of Data Collection

Marly Roesch who is a developer of the Snort systems defines a given network at lightweight network intrusion detection system” when the network traffic and the packets on the IP network can be analyzed and logged in real-time. Like a network sniffer who is based on the network packet collecting system library called "libpcap". Since the libpcap will be log into a database for my project and then with the help of pattern matching algorithm m compare the content of the packet with the rule set present in the database.

4.2 Source of Data Collection

SNORT is a Free and open source NIDS. It utilizes the rule driven language to perform the pattern matching. Since the detection process of SNORT is heavily depend on the rule set present in the database.

4.3 Structure of IDS Rule

All IDS rules have two logical parts: rule header and rule options. This is shown in Figure 4. 1. The rule header contains information about what action a rule takes. It also contains criteria for matching a rule against data packets. The options part usually contains an alert message and information about which part of the packet should be used to generate the alert message. The options part contains additional criteria for matching a rule against data packets. A rule may detect one type or multiple types of intrusion activity. Intelligent rules should be able to apply to multiple intrusion signatures.

<table>
<thead>
<tr>
<th>Rule Header</th>
<th>Rule Options</th>
</tr>
</thead>
</table>

**Figure 4.1 Basic structures of IDS rules.**
The general structure of rule header is shown in Figure 4.2. The action part of the rule determines the type of action taken when criteria are met and a rule is exactly matched against a data packet. Typical actions are generating an alert or log message or invoking another rule.

<table>
<thead>
<tr>
<th>Action</th>
<th>Protocol</th>
<th>Address</th>
<th>Port</th>
<th>Direction</th>
<th>Address</th>
<th>Port</th>
</tr>
</thead>
</table>

Figure 4.2 Structure of IDS rule header.

The protocol part is used to apply the rule on packets for a particular protocol only. This is the first criterion mentioned in the rule. Some examples of protocols used are IP, ICMP, UDP etc. The address parts define source and destination addresses. Addresses may be a single host, multiple hosts or network addresses. We can also use these parts to exclude some addresses from a complete network. There are two address fields in the rule. Source and destination addresses are determined based on direction field. As an example, if the direction field is -- > , the Address on the left side is source and the Address on the right side is destination. In case of TCP or UDP protocol, the port parts determine the source and destination ports of a packet on which the rule is applied. In case of network layer protocols like IP and ICMP, port numbers have no significance. The direction part of the rule actually determines which address and port number is used as source and which as destination. For example, consider the following rule that generates an alert message whenever it detects an ICMP ping packet (ICMP ECHO REQUEST) with TTL equal to 100, as follow:

```
alert icmp any any -> any any (msg: "Ping with TTL=100"); \ttl: 100;
```

The part of the rule before the starting parenthesis is called the rule header. The part of the rule that is enclosed by the parentheses is the options part. The header contains the following parts, in order:

- **A rule action:** In this rule the action is - alert, which means that an alert will be generated when conditions are met. The packets are logged by default when an alert is generated. Depending on the action field, the rule options part may contain additional criteria for the rules.

- **Protocol:** In this rule the protocol is ICMP, which means that the rule will be applied only on ICMP-type packets. In the IDS detection engine, if the protocol of a packet is not ICMP, the rest of the rule is not considered in order to save CPU time. The protocol part plays an important role when you want to apply IDS rules only to packets of a particular type.

- **Source address and source port.** In this example both of them are set to -- any, which means that the rule will be applied on all packets coming from any source. Of course port numbers have no relevance to ICMP packets. Port numbers are relevant only when protocol is either TCP or UDP.

- **Direction.** In this case the direction is set from left to right using the -> symbol. This shows that the address and port number on the left hand side of the symbol are source and those on the right hand side are destination. It also means that the rule will be applied on packets traveling from source to destination. You can also use a < - symbol to reverse the meaning of source and destination address of the packet. Note that a symbol <- can also be used to apply the rule on packets going in either direction.

- **Destination address and port address.** In this example both are set to any, meaning the rule will be applied to all packets irrespective of their destination address. The direction in this rule does not play any role because the rule is applied to all ICMP packets moving in either direction; due to the use of the keyword "any" in both source and destination address parts. The options part enclosed in parentheses shows that an alert message will be generated containing the text string “Ping” with TTL=100 whenever the condition of TTL=100 is met. Note that TTL or Time To Live is a field in the IP packet header.

SNORT is a signature based NIDS. SNORT can be divided into five major components that are each critical to intrusion detection.

- **Multi Rule Search Engine** which fetches the rules from Rules table. The rules are in predefined format,
stored in a table. The detection engine performs simple tests on a single aspect of each packet to detect intrusions. The Second step is Logo pattern detection engine which senses the logo in every incoming packet. The logo is mapped with the logo table, if the logo matches, the detection engine confirms there is no intrusion or else intrusion is detected. The last component is the output plug-in, which generates alerts to present suspicious activity.

V. RESULTS

The proposed approach is implemented using SNORT. The evaluation of the proposed method is performed based on the factors Efficiency, Runtime, Space and Accuracy. The result of the experiments is presented below:

5.1. Efficiency Performance

The result shows that the LPM reduces the number of character comparison to 8 and reduce the number of attempts to 6. This is because of hashing approach of Knuth–Morries-Pratt algorithm to perform the character comparison and depends on shift table of Horspool algorithm to perform the movement of pattern.

<table>
<thead>
<tr>
<th>Character Comparison</th>
<th>Number of Attempts</th>
</tr>
</thead>
<tbody>
<tr>
<td>BM</td>
<td>20</td>
</tr>
<tr>
<td>KR</td>
<td>18</td>
</tr>
<tr>
<td>HP</td>
<td>16</td>
</tr>
<tr>
<td>BF</td>
<td>13</td>
</tr>
<tr>
<td>KMP</td>
<td>10</td>
</tr>
<tr>
<td>LPM</td>
<td>8</td>
</tr>
</tbody>
</table>

Table 1. Efficiency Comparison

5.2. Time Performance

The running-time performance, also referred to as time complexity, is measured in number of machine steps, and in this case we are primarily concerned with character or byte comparisons. To present the results of the running time of algorithms, we vary the input size, where the input is the English words. The number of patterns to be matched remains the same. The running time (in milliseconds) for the algorithms is recorded in the following table.

<table>
<thead>
<tr>
<th>Input Size</th>
<th>BM</th>
<th>KR</th>
<th>HP</th>
<th>KMP</th>
<th>LPM</th>
</tr>
</thead>
<tbody>
<tr>
<td>20000</td>
<td>20</td>
<td>20</td>
<td>18</td>
<td>16</td>
<td>12</td>
</tr>
<tr>
<td>60000</td>
<td>45</td>
<td>40</td>
<td>41</td>
<td>39</td>
<td>30</td>
</tr>
<tr>
<td>100000</td>
<td>70</td>
<td>75</td>
<td>79</td>
<td>61</td>
<td>46</td>
</tr>
<tr>
<td>140000</td>
<td>104</td>
<td>104</td>
<td>108</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>180000</td>
<td>120</td>
<td>132</td>
<td>135</td>
<td>129</td>
<td>112</td>
</tr>
<tr>
<td>200000</td>
<td>134</td>
<td>142</td>
<td>140</td>
<td>159</td>
<td>123</td>
</tr>
</tbody>
</table>

Table 2. Runtime Comparison

5.3. Space performance

The amount of memory consumed while the algorithm runs, is considered only in addition to the necessary space to store the keyword and input. The keyword and the keyword set must always be stored. The space performances of proposed algorithm and the five algorithms were compared using one pattern. The results are shown in Fig 5.3. The results show that LPM size is smaller than that of the other algorithms, for the same pattern.

Fig 1. Efficiency Comparison

The result shows that the LPM reduces the number of character comparison to 8 and reduce the number of attempts to 6. This is because of hashing approach of Knuth–Morries-Pratt algorithm to perform the character comparison and depends on shift table of Horspool algorithm to perform the movement of pattern.
5.4. Accuracy Performance

The number of patterns are vary, the accuracy for the five algorithms are shown in Figure 4. Horspool,Karb-Roubin,and BoyerMoore and Brute-Force Algorithms have the minimum accuracy because the shifted values are affected by increasing signature length. The difference between them is very small. Knuth-Morries-Pratt and Logo Pattern Matching algorithms were not affected by the increased signature length because their shifted values are always one byte.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Pattern Length (Byte)</th>
<th>Memory Space (MB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BM</td>
<td>20</td>
<td>1.96</td>
</tr>
<tr>
<td>KR</td>
<td>20</td>
<td>1.76</td>
</tr>
<tr>
<td>HP</td>
<td>20</td>
<td>1.56</td>
</tr>
<tr>
<td>BF</td>
<td>20</td>
<td>1.30</td>
</tr>
<tr>
<td>KMP</td>
<td>20</td>
<td>1.25</td>
</tr>
<tr>
<td>LPM</td>
<td>20</td>
<td>1.06</td>
</tr>
</tbody>
</table>

VI. CONCLUSION

This work identifies the number of promising algorithms and provides an overview of recent developments in the single keyword pattern matching for IDS. Boyer-Moore Algorithm uses two tables and matching starts with right to left, but in Horspool uses only one table and the matching is faster than the Boyer-Moore. The Brute force algorithm requires no preprocessing of the pattern. The Knuth-Morris-Pratt algorithm performs the comparisons from left to right. Karp-Rabin algorithm is based on hashing approach. The proposed LPM algorithm is compared with the exiting algorithms and the result shows that the algorithm is faster and more reliable in network security applications. The results of algorithm show an improvement in average comparing, faster than the original algorithms, less character comparison and performs less number of attempts compared to the exiting algorithms. In future work, we will enhance the method by moving toward the parallel computing to reduce the workload of system and consequently improve the speed and accuracy of the detection of malicious activities.

VII. REFERENCES

References


About Authors

R. Hamsaveni received B.Sc Computer Science from Madras University and M.Sc Computer Science Degree from Bharathidasan University, Trichy and M.Phil in Mother Teresa Women’s, Kodakanaal. She is pursuing Ph.D degree in Computer Science at SCVMC University. She has 16 years of teaching experience. She is working as Assistant Professor of Computer Science in SVCET, Chittoor, A.P. Her research interests include Network Security.

Prof. Dr. G. Gunasekaran graduated in 1989 with a Bachelor of Engineering degree in Computer Science. He obtained his Master of Engineering Degree in Computer Science in the year 2001. He received the Ph.D degree in Computer Science and Engineering from the Jadavpur University, Kolkata in the year 2009. He has 25 years of teaching experience starting from Lecturer to Associate Professor. At present he is working as Principal of Meenakshi College of Engineering, Chennai, Tamilnadu. He has guided for more than five Ph.D Scholars. He is member of Board studies of various Autonomous Colleges and Universities. He published around 15 research papers in national and international journals and conferences. His current research interests include Data Mining and Network Security.