A New Two-Stage Search Procedure for Misuse Detection

Slobodan Petrović[1] and Katrin Franke[2]

[1]NISlab, Department of Computer Science and Media Technology
Gjøvik University College, P.O. box 191, 2802 Gjøvik, Norway
[3]NISlab, Department of Computer Science and Media Technology
Gjøvik University College, P.O. box 191, 2802 Gjøvik, Norway
slobodanp@hig.no[1], katrin.franke@hig.no[2]

Abstract

A new two-stage indexless search procedure is presented that makes use of the constrained edit distance in IDS misuse detection attack database search. The procedure consists of a pre-selection phase, in which the original dataset is reduced and the exhaustive search phase for the database records selected in the first phase. The maximum number of consecutive deletions represents the constraint. Besides eliminating the need for finer exhaustive search in the attack database records in which the detected subsequence is too distorted, the new search procedure also enables better control over the search process in the case of deliberate distortion of the attack strings. Experimental results obtained on the SNORT signature files show that the proposed method offers average search data set reduction in the typical cases of more than 70% compared to the method that uses the unconstrained edit distance.

Keywords: *Intrusion Detection, Misuse Detection, Constrained Edit Distance, Search.*

1. Introduction

Misuse detection based IDS operate by searching for appearance of specific traffic patterns in a database of known attack signatures. The number of known vulnerabilities, as well as the variability of possible searchable attack strings in the database may become a limiting factor for real time operation of such a system. However, it is sometimes possible to exclude parts of the search space that are considered non-interesting for the case at hand. In this paper, we apply edit distance with special constraints in the process of pre-selection of the misuse-based IDS database records that are interesting for eventual matching of the particular search pattern. A coarse pre-selection matching is performed first, by computing the constrained edit distance between each record of the attack signatures database and the query. The maximum length of runs of deletions represents the constraint. The obtained constrained edit distances are then sorted in the increasing order and a finer inspection is performed in the records starting from those with the minimum constrained edit distance from the search string. The advantage of use of the constrained edit distance over the unconstrained one lies in the possibility of skipping finer exhaustive search in the records of the database where the detected subsequence is too distorted, even if the unconstrained edit distance is small enough. The use of the deletion constraint also enables better control over the search process in the cases in which the attackers deliberately distort the attack strings by substituting and deleting characters.

The best known general approximate database search algorithms use indexing in order to make the search procedure as effective in time as possible. For example, [3] defines an approximate search algorithm based on a combination of a specific index structure and unconstrained edit distance. To improve the efficiency even more, the edit distance is computed approximately by means of the q-gram distance. However, the use of indexes is memory-consuming and the index must be updated as the database changes, which is often the case with the IDS attack signature databases. Unlike the general data mining algorithms, many algorithms originally intended to be used in IDS database search are based either on trying to use previous search results in the current search (for example, the Signature A priori Algorithm [5]) or on exploiting the similarities of the attack signatures (database records) (see for example [6]). The efficiency of such algorithms, however, heavily depends on the properties of the particular attack signature database. The first attempt to build a combined indexless search procedure that would include computation of edit distances in the records pre-selection process and a finer exhaustive search algorithm in the selected areas is described in [2]. That procedure, applied in digital forensics search, in which the lengths of the search fragments are not defined in advance, makes use of the constrained edit distance in the pre-selection phase. Unlike the computational forensic search, the records of an IDS attack signature database are of different lengths in general and these lengths are defined in advance and fixed. Consequently, in the IDS database search it is not possible to try to improve the search efficiency by varying the length of the search fragment in the first phase of the search process as it was experimented in [2].

The paper is organized as follows: in Section 2, mathematical preliminaries regarding the constrained edit distance are presented. In Section 3, the signature pre-selection algorithm is given. Experimental work is described in Section 4. Finally, Section 5 concludes the paper.

2. Mathematical background

Given arbitrary strings X and Y of lengths N and M, respectively, over a finite alphabet $\mathcal{A}$, we deal with the problem of transforming the string X to Y by means of the elementary edit operations of deletion and substitution, under the following constraints:

C1. The maximum length of runs of deletions is F.

C2. The edit sequence is ordered in a sense that every substitution is preceded by at most one run of deletions.

The constrained edit distance $D(X, Y)$ is then defined as the minimum sum of elementary edit distances associated with the edit operations of deletion and substitution needed to transform X to Y, subject to the assumed constraints.

We first define elementary distances associated with the edit operations of deletion and substitution of symbols. Along with the alphabet $\mathcal{A}$, we introduce the ”empty” symbol $\varnothing$ reserved for the presentation of deletion. Let $\mathcal{A}^* = \mathcal{A} \cup \{\varnothing\}$. Nonnegative real-valued elementary edit distances are defined in the following way:

1. $d(a, \varnothing)$ is the elementary distance associated with the deletion of a symbol $a \in \mathcal{A}$;
2. $d(a, b)$ is the elementary distance associated with the substitution of the symbol a with the symbol b, where $a, b \in \mathcal{A}$.

We also define the compression operator C in the following way. For an arbitrary finite length string Z over $\mathcal{A}^*$, $C(Z)$ stands for the string over $\mathcal{A}$ that is obtained from Z by removing all the empty symbols from it (see [1], for example).

Every editing transformation as a sequence of elementary edit operations can be equivalently represented in an ordered way according to the constraint C2 defined above. We call such a representation an *edit sequence* and denote it by $S = (X, Y')$, where $(X, Y') \in \mathcal{A}\times\mathcal{A}^*$, and for every (X, Y') the following rules hold:

1. $C(Y') = Y$;
2. $|Y'|= N$;
3. for all i, $1 \le i \le N$, the variables x_i and y_i' cannot simultaneously take the value $\varnothing$;
4. (X, Y') satisfies the constraints C1 and C2.

In order to compute the constrained edit distance efficiently, the *partial constrained edit distance* $W(e, s)$ is used as a constrained edit distance between the prefix X_{e+s} of the sequence X and the prefix Y_s of the sequence Y. Note that the set of feasible values of (e, s) is, due to the assumed constraints, given by

$$0 \le s \le M,\ 0 \le e \le \min\{N - s, sF\}. \tag{1}$$

Namely, a pair (e, s) is feasible if and only if it satisfies (1).

The following theorem enables efficient computation of the constrained edit distance $D(X, Y)$. The proof is obtained by following the lines from [1].

Theorem 1:

Under the combination of the constraints C1 and C2, the partial constrained edit distance $W(e, s)$ satisfies the recursion:

$$W(e,s)=\min_{e_1\in Q}\left\{W(e-e_1,s-1)+\sum_{j=1}^{e_1} d(x(e+s-e_1+j),\varnothing)+d(x(e+s),y(s))\right\}, \tag{2}$$

$1 \le s \le M$, $0 \le e \le \min\{N - s, sF\}$, where Q is the set of all e_1 such that $0 \le e_1 \le \min\{F, e\}$. For $s = 0$, $0 \le e \le \min\{N,F\}$,

$$W(e,0)=\sum_{j=1}^{e} d(x(j),\varnothing). \tag{3}$$

3. The new signature search algorithm

In order to apply Theorem 1 in IDS database signature search, it is necessary to remove the deletion constraint before the first substitution and after the last substitution of the edit sequence. Otherwise, a record containing the search string would almost always be rejected for too long runs of deletions before and/or after the search string in the edit sequence. The following modification of the recursion (2)-(3) enables computation of the constrained edit distance with the deletion constraint removed before the first substitution and after the last substitution. The equation (2) still holds, but to remove the constraints before the first substitution, the limits for e in the equation (3) are modified so that we have

$$W(e,0)=\sum_{j=1}^{e} d(x(j),\varnothing),\quad 0 \le e \le N - M. \tag{4}$$

Let $\overline{X}$ be X +'𝒟' and let $\overline{Y}$ be Y +'𝒟', where '𝒟' is any symbol. Then, for $s = M + 1$ and $1 \le e \le N - M$, in order to remove the constraints after the last substitution of a symbol from X, we have

$$W(e, M+1) =$$

$$\min_{e_1 \in Q} \left\{ W(e-e_1, M) + \sum_{j=1}^{e_1} d\big(x(e+M-e_1+j+1), \varnothing\big) + d\big(x(e+M+1), y(M+1)\big) \right\}, \quad (5)$$

where, unlike in (2), Q is the set of e_1 such that $0 \le e_1 \le e$.

The new signature search algorithm proposed in this paper consists of two phases. In the first, pre-selection phase, for each of the records of the database the constrained edit distance given by the (2), (4) and (5) between the record and the search string is computed. The constrained edit distances obtained in this way are sorted in ascending order. Naturally, the records of the database should be of the same length and to achieve this they are all padded with an arbitrary symbol to the length of the longest record. In the second phase of the search process, a finer search is performed in the fragments starting from the top ranked one in the first phase. In this paper, we concentrate on the first phase of the attack signature search procedure, since the second phase that includes exhaustive search in the selected fragments is straightforward. The complete search algorithm used in the first phase is given below. We assume, without loss of generality, that the elementary edit distances associated with deletions of all the symbols are the same.

Algorithm 1

Input:

- N - the length of the record.
- D- the IDS attack signature database.
- S - the search string.
- The elementary distance de associated with the deletion of a symbol.
- The array D of elementary edit distances $d(x, y)$ associated with substitution of x with y, $\forall x, y \in \mathcal{A}$.
- F - maximum number of consecutive deletions.

Output:

- The array **P** of ordered pairs (i, d), sorted by d in ascending order, where i is the ordinal number of the record in D, and d is the corresponding constrained edit distance between that record and the search string.

```
begin
  comment $D$ consists of $k$ records of length $N$
  comment Main loop; $D_i$ is the $i$-th record of $D$, $i$ =1, 2, . . . , $k$
  for $i \leftarrow 1$ until $k$ do
      begin
```

```
            comment Compute the constrained edit distance between $D_i$ and $S$ (See Alg. 2)
            $d$ = CED($D_i$, $S$,| $D_i$ |,| $S$ |, $F$,D, $de$) ;
            Store ($i$, $d$) in P ;
        end ;
    Sort the array P by $d$ in ascending order ;
end.
```

The following algorithm is used to compute the constrained edit distance between two sequences. It is based on (2), (4) and (5).

Algorithm 2 (The CED computing procedure)

Input:

- The sequences D_i and S of lengths | D_i | and | S |, respectively.
- The maximum length F of runs of deletions.
- The elementary distance de associated with the deletion of a symbol.
- The array **D** of elementary edit distances $d(x, y)$ associated with substitution of x with y, $\forall x, y \in \mathcal{A}$.

Output:

The constrained edit distance d between D_i and S.

```
begin
  comment Initialization
  $N \leftarrow | D_i |$ ; $M \leftarrow | S |$ ; $X \leftarrow D_i$ ; $Y \leftarrow S$ ; $W[e, s] \leftarrow \infty$ , $e = 0, . . . , N - M$, $s = 0, . . . , M$ ;
  $W[0, 0] \leftarrow 0$ ;
  for $s \leftarrow 1$ until $M$ do $W[0, s] \leftarrow W[0, s - 1] + d[X[s], Y [s]]$ ;
  comment $s = 0$; no constraints before the first substitution
  for $e \leftarrow 1$ until $N - M$ do $W[e, 0] \leftarrow e * de$ ;
  comment Main loop
  for $s \leftarrow 1$ until $M$ do
      begin
          for $e \leftarrow 0$ until $\min\{N - s, s * F\}$ do
              begin
                  $W[e, s] \leftarrow \min\{W[e - e_1, s - 1] + e_1 * de + d[X[e + s], Y [s]]\}$,
                  $e_1 = 0, ... , \min\{e, F\}$,
              end ;
      end ;
  comment No constraints after the last substitution
```

```
    for $e \leftarrow 1$ until $N - M$ do
        $W[e,M + 1] \leftarrow \min\{W[e - e_1,M] + e_1 * de+d[X[e +M + 1], Y [M + 1]]\}, e_1 = 0, ... , e;$
    comment Compute the constrained edit distance
    $d \leftarrow W[N - M,M + 1]$ ;
  end.
```

4. Experimental work

In order to compare the efficiency of attack signature search with the constrained and the unconstrained edit distance, the following experiment was carried out: all the strings corresponding to the ”contents” field of the SNORT [4] rules were searched for in the corresponding files with the extension `.rules`. These files were modified by padding the rule descriptions to the length of the longest record. For every record in each file, the constrained edit distance and the unconstrained edit distance were computed between the record and the searched string.

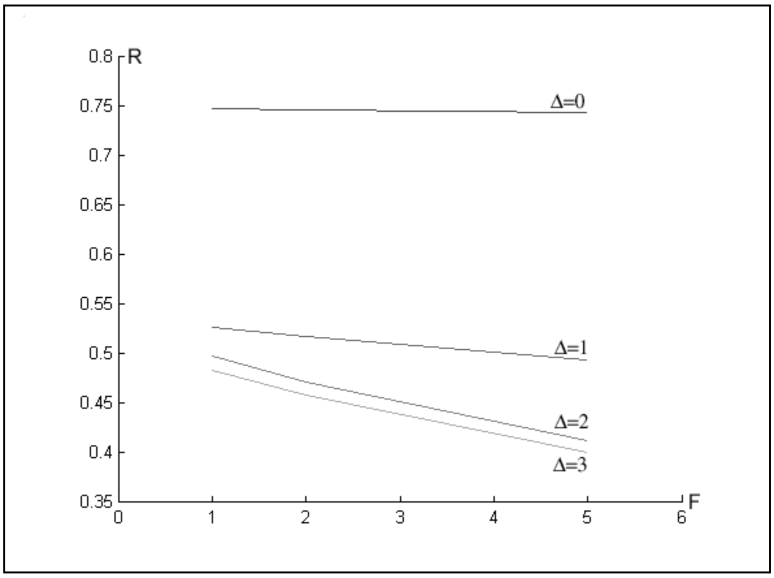

Figure 1. Dependence of the average data set reduction on the deletion constraints for different acceptance thresholds

For a threshold Δ given in advance, a record was accepted as a candidate for a more detailed search in the second phase of the search process if the constrained/unconstrained edit distance between the record and the search string was less than or equal to $N-M+\Delta$, where N was the length of the record and M was the length of the search string. For each search string, the numbers of accepted records in the cases of the use of the constrained and the unconstrained edit distance were counted. Let n_c be the number of accepted records over all the search strings from the examined set with the constrained edit distance used in the first search phase and let n_u be the number of accepted records if the unconstrained edit distance is used in the same process. We define the *data set reduction R* as the measure of efficiency of the signature search algorithm that makes use of the constrained edit distance:

$$R = 1 - \frac{n_c}{n_u}, \quad n_u \neq 0. \tag{6}$$

Fig. 1 presents dependence of the average value of R over the SNORT's rule files mentioned above on the deletion constraints F for different acceptance thresholds.

The most important fact to observe from Fig. 1 is that the average data set reduction is very high (> 70%) in the typical zero-tolerance case ($\Delta = 0$). This means that if the constrained edit distance is used in the first phase of the search process, this phase becomes much less time consuming compared to the time necessary to complete the search when the unconstrained edit distance is used. For higher acceptance thresholds, needed when a deliberate distortion of traffic patterns is suspected, it can be observed that the average data reduction is still high (> 40%) for all the examined constraints.

5. Conclusion

In this paper, a new attack signature search procedure for misuse detection IDS databases is described. It applies constrained edit distance in the pre-selection phase, in which the records of the database that deserve a more focused search are selected. The use of constrained edit distance, where the constraints deal with the maximum lengths of runs of deletions, enables rejection of the records in which the detected search string is too distorted. It also enables detection of search strings even when they are deliberately distorted by means of substitutions and deletions of characters. The level of tolerance of such distortions is controlled by means of the values of the constraints and the acceptance thresholds. A necessary modification of the original constrained edit distance algorithm in order to be used in signature search is described. Experimental results, obtained on the widely used SNORT database, show that for typical values of the search parameters (constraints and acceptance threshold values) the new signature search procedure is much more efficient than the procedure that uses the unconstrained edit distance.

References

[1] B. Oommen, "Constrained String Editing", Inform. Sci., Vol. 40, No. 9 (1986), pp. 267-284.

[2] S. Petrović and K. Franke, "Improving the Efficiency of Digital Forensic Search by Means of the Constrained Edit Distance", in Proceedings of the Third International Symposium on Information Assurance and Security, Manchester, UK, 2007, pp. 405-410.

[3] F. Shi, "Fast Approximate Search in Text Databases", in Proceedings of the Fifth International Conference on Web-Age Information Management, WAIM 2004, Dalian, China, 2004, pp. 259-267.

[4] SNORT intrusion detection system Web page, http://www.snort.org.

[5] J. Zhao and H. Huang, "An Intrusion Detection System Based on Data Mining and Immune Principles", in Proceedings of Machine Learning and Cybernetics International Conference, 2002, pp. 453-501.

[6] H. Zheng Bing and V. Shirochin, "Data Mining Approaches for Signatures Search in Network Intrusion Detection", in IEEE Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, September 2005, Sofia, Bulgaria, pp. 392-398.

Authors

Slobodan Petrović is professor of information security at NISlab, Department of Computer Science and Media Technology, Gjøvik University College, Gjøvik, Norway. He received his Ph.D. degree in 1994 from the University of Belgrade, Serbia. His research interests include cryptology, intrusion detection, and digital forensic. He is the author of more than 40 papers published in renowned international journals and conferences.

Katrin Franke is associate professor of information security at NISlab, Department of Computer Science and Media Technology, Gjøvik University College, Gjøvik, Norway. She received a diploma in electrical engineering from the Technical University Dresden, Germany in 1994 and her Ph.D. in artificial intelligence from the Groningen University, The Netherlands in 2005. Her research interests include computational forensics, biometrics, document and handwriting analysis, computer vision and computational intelligence. She has published several scientific journal articles, peer-reviewed conference papers and edited books.