On the Property of the Distribution of Symbols in SQL Injection Attack

Takeshi Matsuda
Department of Computer Science
Shizuoka Institute of Science and Technology

Abstract

SQL injection is an attack of type to insert malicious query via an input form on web site. If SQL injection attack were successful, there are the threats of unauthorized access, information leak or falsification of data for web applications driven database system. In the conventional studies, a lot of prevention and detection methods using pattern matching, parsing or machine learning have been developed. However, it is easy to evade detection by modifying the strings of SQL injection attacks. Therefore, it is very important to investigate the essence of SQL injection attacks for preventing such evasion of detection. In this study, we constructed the feature space by using the property of the distribution of SQL injection attack, and proposed an attack detection method. The result of this study is showing the importance concerning the construction of feature space.

1. Introduction

The threat of SQL injection attacks has been increasing still now in spite of many prevention and detection methods had developed. There are several reasons why the damage of SQL injection attacks do not decrease, but the principal reason is that attackers may seek a loophole of prevention and detection methods. To break away from such vicious circle, we should develop some new prevention and detection method that attackers cannot seek a loophole.

Types of SQL injection attacks had classified, and major attack techniques had summarized in [1] [2] [3] [4] [5] [6]. For selecting the sample of SQL injection attacks, we can notice that SQL injection attacks have some special symbols almost certainly. In our previous study [7] [8] [9], we had defined these symbols as attack feature symbols, and had proposed detection method using the ratio of attack feature symbols. Although the detection method using attack feature symbols is very effective, there are still some kinds to be worked out. In particular, it is very important to develop the following two methods:

- Selection method of attack feature symbols
- Construction method of feature spaces

Feature spaces are used for making the rule of SQL injection attacks detection. If a feature space would not construct appropriately, the risk of miss detection will increase. Moreover, the ease of making evasion attack of the detection depends on the construction of feature spaces.

In this paper, we focused attention on the distribution of symbols in SQL injection attacks, showed that the distribution of symbols may be approximated by zeta function. If the distribution of symbols drawn from zeta distribution, this means that symbols which well appear in SQL injection attacks are limited. Therefore, we can form the following hypothesis: The property of the distribution of symbols may be used for the construction of the feature space of SQL injection attacks. To verify the above hypothesis, we constructed two feature spaces, and we obtained the result which may support the hypothesis from the experiment of a detection. The rest of the paper is organized as follows. Section 2 gives an overview of the concept on SQL injection attacks. Section 3 summarizes the property of zeta distribution and the distribution of symbols in SQL injection attacks. Section 4 discusses on the feature space of SQL injection attacks. Section 5 concludes this study.

2. SQL Injection Attack

There are possibilities of SQL injection attacks on web applications driven database. Although SQL injection attack is very famous as attack technique, there are still not the prevention or detection methods that give a proper solution on SQL injection attacks. However, there exist the studies on the classification of SQL injection attacks [1]. The famous SQL injection attack is tautology attack. The following is a typical sample of the tautology attack.

```
' or '1' = '1 --
```

In this attack, 1=1 is always true. Therefore, if this attack became successful, there is the threat of an unauthorized access. As other SQL injection attacks, Illegal/Logically Incorrect Queries [2], Union Query [3], Piggy-Backed Queries [4], Stored Procedures [5] and Inference [6] and Alternate Encodings [2] are well known. Most of SQL injection attacks are generating by tools such as worm and botnet. Therefore, attacks generated by tools have some rules. To make SQL injection attacks, some delimiter characters, such as single quote and semicolon, are required as shown in the example above. Although some SQL injection attacks do not contain single quote and semicolon, it is true that some symbols have important rules to make SQL injection attacks. In the following, we will use the term ”attack string” (resp. ”normal string”) as a SQL injection attack (resp. a non SQL injection attack). To show the above property of attack strings, we collected SQL injection attacks from website and books [] [] [] []. Additionally, we generated normal strings such as ID, pass word, telephone number, address, wiki grammar and emoticon. In our previous study, we had investigated that the distribution of symbols in SQL injection attacks may be approximated by zeta distributions. So, in this study, we proposed a detection method of SQL injection attacks by using the property of zeta distributions.

3. Distribution of Symbols in SQLIA

In this section, we will show the property of the distribution of symbols in SQL injection attacks. To seek for the feature of SQL injection attacks, we investigated on the appearance frequency of symbols in SQL injection attacks. In our collected attacks, there were following 20 symbols.

s_j	symbols
s_1	space
s_2	semicolon (;)
s_3	single quote ($'$)
s_4	right parenthesis ())
s_5	left parenthesis (()
s_6	right brace (})
s_7	left brace ({)
s_8	right square bracket (])
s_9	left square bracket ([)
s_{10}	sharp (#)
s_{11}	parcent (%)
s_{12}	double quote ($''$)
s_{13}	ampersand (&)
s_{14}	backslash (\)
s_{15}	pipe (|)
s_{16}	equal sign (=)
s_{17}	greater than sign (>)
s_{18}	less than sign (<)
s_{19}	asterisk ($*$)
s_{20}	slash (/)
s_{21}	minus ($-$)
s_{22}	plus (+)

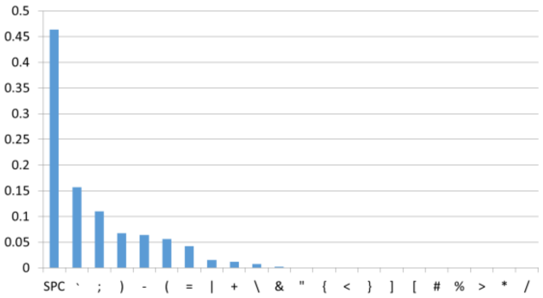

Figure 1. Distribution of Symbols in Attack Strings

Figure 1 shows that the distribution of symbols in attack strings. The vertical axis of graphs is calculated as follows. Let l_i be strings of attack or normal, and $|l_i|$ be a length of the string l_i. Here i is a natural number. Let $x_i(s_j)$ be an appearance frequency of symbol s_j including in l_i. Then, the value of vertical axis of graph is defined by

$$T(s_j) = \sum_{i=1}^{I} \frac{x_i(s_j)}{|l_i| \cdot I},$$

where I is the number of strings.
For comparison, we investigated on the distribution of symbols in normal strings, and showed in Figure 2.

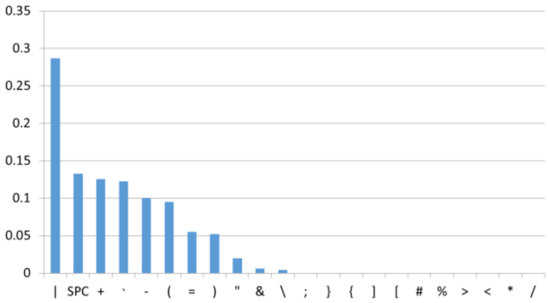

Figure 2. Distribution of Symbols in Normal Strings

From Figure 1, we can see that attack strings have some specific symbols, such as space, single quote, semicolon and right parenthesis. On the other hand, Figure 2 shows that space and single quote are also appear in normal strings. Therefore, it is hard to say that space and single quote are feature symbols in SQL injection attacks. So, the symbol that the frequency is high may not be used for attack feature by itself. For this problem, in [8], we experimentally showed that the detection method using a set of plural symbols is superior to the one using one symbol in term of the detection rate. This means that some symbols have a relation each other in SQL injection attacks, and this relation is expressed in the graph of Figure 1.

In our previous study [9], we had shown that the distribution of symbols in SQL injection attacks may be approximated by zeta distribution. Here, zeta distribution is defined as follow.

$$p(x|a) = \frac{1}{\varsigma(a)x^{-a}},$$

where $\varsigma(a)$ is Riemann zeta function. In [9], it had shown that the distribution of Figure 1 can be approximated by $p(x|2.2)$. The parameter a of the zeta distribution $p(x|a)$ is estimated from a given sample data. However, it is not easy to estimate the parameter a because the zeta distribution includes the Riemann zeta function $\varsigma(a)$. Therefore, to compute the parameter a, some numerical calculation is required. In such case, an optical solution should be uniqueness. The following theorem shows that the Kullback information between zeta functions has unique small value.

[Theorem]
Let $p(x|a)$ and $p(x|c)$ be zeta functions, and let $a \neq c$. Then, the Kullback information

$$KL(a) = \sum_{x=1}^{\infty} p(x|c)\frac{p(x|c)}{p(x|a)}$$

has an unique minimum value in $a > 1$.

The proof of this theorem is given by [9]. So, we will give the outline of the proof as follows.

Let us consider the function

$$f(x) = \varsigma(a)x^{a}.$$

We can see that the function $f(x)$ is a monotonic decreasing function, The Kullback information $KL(a)$ can be written by the function $f(x)$ in the following way:

$$KL(a) = \sum_{x=1}^{\infty} \frac{\log f(a) - \log f(c)}{f(a)}$$

From this function, we can show the statement of the above theorem using the property of the following function has a unique minimum value.

$$\frac{\log f(a) - \log f(c)}{f(a)}$$

The salient property of zeta distribution is that most probabilistic value $p(x|a)$ of x are close to 0. In Figure 3, let us compare the distribution of symbols in attack strings and normal strings. From the Figure 3, we can see that the distribution of symbols in attack strings imitates the property of zeta distributions than the distribution of symbols in normal strings. The horizontal axis of Figure 3 indicates the order of the frequency in each string.

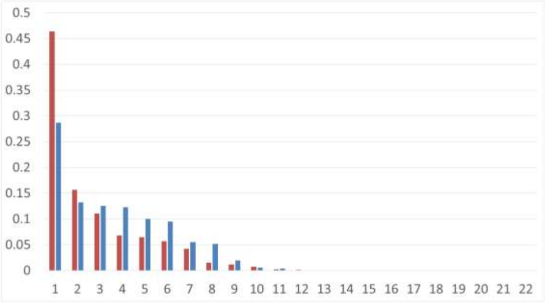

Figure 3. Comparison in Attack String and Normal Strings

If we assume that the distribution of symbols in SQL injection attacks, we can estimate of the parameter of the zeta function that imitates the distribution of symbols in SQL injection attacks from the result of this theorem. Figure 4 shows that the Kullback information $KL(a)$ takes minimal value in the neighborhood of $a = 2.2$. The data of Figure1 is used for estimating the parameter a.

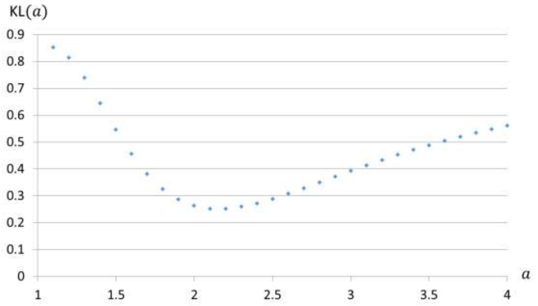

Figure 4. Graph of $KL(a)$

4. Proposed Detection Method

In our previous study, we had already proposed a detection method of SQL injection attacks based on the property of zeta distribution [13]. So, in this paper, we will introduce another detection method using the property of zeta distributions.

In the previous section, we introduced that the distribution of symbols in SQL injection attacks may be approximated by zeta distribution. This means that the determinative factors of SQL injection attacks are limited. Therefore, it is highly likely that the property of zeta distribution leads to good performance on SQL injection attacks detection. To check the above hypothesis, we will propose the following two detection methods.

4.1. Detection Method 1

To detect SQL injection attacks, we need to construct a feature space of SQL injection attacks. In the detection method 1, we construct the feature space of SQL injection attacks in the following way.

Assume that some input string l observed. Let $u_1, u_2, \cdots, u_{26}$ be appearance frequencies of A, B, $\cdots$, Z in l, respectively. And, let $u_{27}, u_{28}, \cdots, u_{36}$ be appearance frequencies of 0, 1, $\cdots$, 9 in l, respectively. Finally, let u_{37} be appearance frequencies of other all symbols or characters in l. Now, we will give an example. If the following string l observed,

' or '1' = '1 --

then we have

$$(u_1, \cdots, u_{15} \cdots, u_{18} \cdots, u_{28} \cdots, u_{37}) \\ = (0, \cdots, 1 \cdots, 1 \cdots, 2 \cdots, 12).$$

This feature space gives importance to the alphabets and the numbers in input string l.

4.2. Detection Method 2

In the detection method 2, we construct the feature space of SQL injection attacks by giving importance to the symbols in SQL injection attacks. We prepared 100 SQL injection attacks sample, and investigated the distribution of symbols in these attacks. These attacks are different from the sample of Figure 1. We extracted symbols satisfying the following condition:

$$T(s_j) \geq 0.005.$$

Then, the following symbols were extracted.

coordinate	symbols
x_1	space
x_2	semicolon (;)
x_3	single quote ($'$)
x_4	right parenthesis ())
x_5	left parenthesis (()
x_6	double quote ($''$)
x_7	pipe (|)
x_8	equal sign (=)
x_9	greater than sign (>)
x_{10}	less than sign (<)
x_{11}	asterisk ($*$)
x_{12}	slash (/)
x_{13}	minus ($-$)
x_{14}	colon (:)

Here, x_k $(k = 1, 2, \cdots, 14)$ be appearance frequencies of symbols. Finally, we define the coordinate x_{15} in the following way. Let x_{15} be appearance frequencies of other all characters or symbols such as alphabets, numbers, sharp symbol and ampersand symbol etc. Then, the above string l is expressed as

$$(x_1, x_2, x_3, \cdots, x_8, \cdots, x_{13}, x_{14}, x_{15}) \\ = (5, 0, 4, \cdots, 1, \ \cdots, 2, 0, 4).$$

in this feature space.

5. Experiment

In this section, we will detect SQL injection attacks by using two distinct feature spaces whose are defined in the previous section. Firstly, we prepare the following attack and normal sample to extract their features.

- Learning Data
  (attack : 328 sample, normal : 73 sample)

Moreover, we prepare the following attack and normal data to test the effectiveness of the detection.

- Testing Data
  (attack : 59 sample, normal : 40 sample)

To detect from these data, we need some rules of detection. In this study, we determine the rule by using SCW algorithm [14] that is one of the machine learning methods.

SCW is a linear classifier, and the classification is given by considering the following hypersurface:

$$\sum_{i=1}^{I} w_i x_i = 0$$

Here, x_i and $w_i (i = 1, \ 2, \cdots, I)$ denote an input and a parameter, respectively. In the algorithm of SCW, the parameter vector $w = (w_1, w_2, \cdots, w_I)$ drawn from the multivariate normal distribution with the mean vector μ and the covariance matrix Σ. If an input $x = (x, x_2, \cdots, x_I)$ is attack (resp. normal), we define $y = 1$ (resp. $y = -1$). Now, we give T sample $\{(x_1, y_1), \cdots, (x_T, y_T)\}$. Then, the mean vector μ and the covariance matrix Σ are updated in the following way.

$$\mu_{t+1} = \mu_t + \alpha_t y_t \Sigma_t \mathrm{x}_t,$$

$$\Sigma_{t+1} = \Sigma_t - \beta_t x_t^{\top} x_t \Sigma_t,$$

where

$$\alpha_t = min\left\{C, max\left\{0, \frac{1}{v_t\zeta}\left(-m_t\psi + \sqrt{m_t^2\frac{\phi^4}{4} + v_t\phi^2\zeta}\right)\right\}\right\}$$

$$\beta_t = \frac{\alpha_t\phi}{\sqrt{u_t} + v_t\alpha_t\phi}$$

$$u_t = \frac{1}{4}\left(-\alpha_t v_t\phi + \sqrt{\alpha_t^2 v_t^2\phi^2 + 4v_t}\right)^2$$

$$v_t = x_t^{\mathrm{T}}\Sigma_t x_t$$

$$m_t = y_t(\mu_t \cdot x_t)$$

$$\phi = \Phi^{-1}(\eta)$$

$$\psi = 1 + \frac{\phi^2}{2}$$

$$\zeta = 1 + \phi^2$$

Here, $\Phi^{-1}(\eta)$ is an inverse function of the cumulative function of the normal distribution, η is a real number and $0 < C < 1$. In this study, we set the initial values of μ_0 and Σ_0 as follows:

$$\mu_0 = (0, 0, \cdots, 0),$$

$$\Sigma_0 = \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{pmatrix}.$$

We summarize the result of experiment in the following Table 1. Here, we put the parameters as $C = 0.2$ and $\eta = 0.9$, in this experiment.

Table 1 : Detection Result

	Detection Method 1	Detection Method 2
Correct Detection Rate	75%	84%

6. Discussion

In the previous section, we showed the result of experiment on the SQL injection attacks detection using our proposed method. Here, we will discuss on the experiment result.

From the result of Table 1, we can see that the feature space using the property of the distribution of symbols in SQL injection attacks is effective for the detection, because the feature space 2 gives importance to the symbols in SQL injection attacks. Moreover, the result of Table 1 shows that the importance of constructing feature space. It cannot be said that the result of this study is good from the standpoint of the detection accuracy. To improve the detection accuracy is our important future works.

On the other hand, we used the algorithm of SCW algorithm to determine the detection rule in the experiment of this study. The method of determining the detection rule will also depend on the detection accuracy. Therefore, to investigate the optimal method of using machine learning or another method for the detection of SQL injection attack is also important problem.

7. Conclusion

In this study, we proposed the detection method using the property of symbols in SQL injection attacks. From the result of the experiment, we could see that symbols play an important role in SQL injection attacks. Our future works are to develop more effective SQL injection attack detection method and to investigate the optimal method of using machine learning for the detection.

8. References

[1] V. Nithya,R.Regan, J.vijayaraghavan, A Survey on SQL Injection attacks, their Detection and Prevention Techniques, International Journal Of Engineering And Computer Science ISSN:2319-7242 Volume 2 Issue 4 pp.886-905, 2013.

[2] C. Anley. Advanced SQL Injection In SQL Server Applications. White paper, Next Generation Security Software Ltd., 2002.

[3] S. McDonald. SQL Injection: Modes of attack, defense, and why it matters. White paper, GovernmentSecurity.org, April 2002.

[4] M.Howard and D.LeBlanc. Writing Secure Code. MicrosoftPress, Redmond, Washington, second edition, 2003.

[5] E.M.Fayo. Advanced SQL Injectionin Oracle Databases.Technicalreport, Argeniss Information Security, Black Hat Briefings, BlackHat USA, 2005.

[6] C. Anley. et al., Advanced SQL Injection. White paper, Next Generation Security Software Ltd., 2002.

[7] T. Matsuda, D. Koizumi, M. Sonoda, and S. Hirasawa, On predictive errors of SQL injection attack detection by the feature of the single character, Proceeding in 2011 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 1722 - 1727, 2011.

[8] M. Sonoda, T. Matsuda, D. Koizumi and S. Hirasawa, On Automatic Detection of SQL Injection Attacks by the Feature Extraction of the Single Character, Proceeding in 2011 International Conference on Security of Information and Networks, ACM, pp.81-86, 2011.

[9] T. Matsuda, Feature Extraction of Web Application Attacks Based on Zeta Distribution, World Congress on Internet Security (WorldCIS-2013)

[10] SQL Injection Cheat Sheet,
http://ferruh.mavituna.com/sql-injection- cheat-sheet- oku/.

[11] SQL Injection Cheat Sheet,
http://michaeldaw.org/sql-injection-cheat- sheet.

[12] MySQL SQL Injection Cheat Sheet,
http://pentestmonkey.net/blog/mysql-sql-injection-cheat-sheet.

[13] S. Maeda, T. Matsuda, M. Sonoda and S. Chou, On the detection method of SQL Injection Attacks using the property of zeta distribution, (Submitted)

[14] J. Wang, P. Zhao and S. C. Hoi, Exact Soft Confidence-Weighted Learning, Proceedings of the 29th International Conference on Machine Learning, pp. 121-128, 2012.