Information Retrieval Method for Mixed Intrusion of Surveillance Network Based on Big Data

YuLin Zhu, QingChuan Meng*, DaYong Hu, Ziyuan Lu, ZhiGang Li

HeiLongJiang Province Cyberspace Research Center, Harbin, HeiLongJiang, 150090, China

*Corresponding author’s e-mail: mqchappy365@163.com

Abstract. The current surveillance network intrusion information retrieval methods have problems such as low retrieval accuracy, high recall rate, and long retrieval time. Therefore, this paper proposes a hybrid intrusion information retrieval method for monitoring networks based on big data. The genetic algorithm is used to optimize the feature set, and the partial F test is introduced. The optimal subset is selected to form the optimal feature set and the redundant information elimination model is constructed to eliminate redundant information in the mixed intrusion information. Based on the information retrieval theory, the LDA model is used to build the feature set, and model the subject of the document, establish an intrusion information retrieval model, then complete the hybrid network intrusion information retrieval of the monitoring network under big data. The results show that the method proposed in this paper has higher retrieval accuracy, can effectively improve the retrieval efficiency of intrusion information and the average recall rate is about 24%, which is better than other methods.

1. Introduction

The rapid development of computer and Internet technologies has led to the continuous growth of various types of data, and the era of big data is coming. Data has changed from simple processing at the initial stage to a basic resource [1]. How to better manage these data has become a hot topic of current concern. In the process of managing the above data, if the data is invaded, it will cause huge losses, and the retrieval of intrusion information of the information system has become the focus of people's attention. The development of search browsers such as Baidu and Google has lowered the threshold of search tools [2], and non-professional staff have insufficient security awareness when conducting information searches, and it is difficult to avoid interference caused by malicious invasion [3]. A new intrusion information retrieval method is researched to ensure the data security in people's daily search and application of data.

At present, some scholars at home and abroad have carried out related research. For example, Literature [4] proposed a hybrid intrusion information retrieval method based on feedback algorithm for monitoring network, using the feedback algorithm to calculate the complexity of intrusion data, and sorting according to the minimum experience loss to obtain The characteristics of intrusion data can realize the retrieval of network intrusion data, but the retrieval accuracy of this method is low. Literature [5] proposed a hybrid intrusion information retrieval method for monitoring network based on cloud environment. This method completes the intrusion data search by constructing a security index Function to realize the retrieval of mixed intrusion information in monitoring network, but the retrieval of this method is time consuming. Aiming at the problems existing in the above methods, a hybrid data retrieval method for monitoring network based on big data is proposed.
2. Hybrid intrusion information redundancy elimination processing

2.1. Feature Set Selection

Intrusion feature selection refers to selecting a valid subset of attributes to describe a large valid dataset pattern containing redundant and irrelevant attribute data, and stipulating that within the valid time, the smallest can be found within a set range. The most powerful and descriptive subset of intrusion features makes each selected subset not redundant or irrelevant [6].

Genetic algorithm is an efficient search method based on the natural selection theory, which combines the survival rules of the fittest in the process of biological evolution and the random information exchange system of the chromosomes in the population. It has a wide search range and strong robustness [7]. In this chapter, genetic algorithm is used to search for the intrusion features. In the process of intrusion information retrieval, any intrusion feature selection problem is set as a binary problem, which refers to binary one-dimensional coding. Different chromosomes correspond to the corresponding intrusion feature set. Suppose m is the total number of intrusion features, then the chromosome is a 0-1 string with length of a and b respectively, and different strings correspond to different intrusion features.

\( B(x) \) is the corresponding attack information of the feature included in the individual \( x \), assuming that the proportion of the feature whose value is greater than the set threshold \( \varepsilon \) is \( e^{N(x)} \), and the effective intrusion feature set is \( L' \), the more excellent genes are included in the individual, the more likely the corresponding intrusion feature is to be selected, and the fitness function \( S(x) \) can be expressed as:

\[
S(x) = \frac{B(x)}{1 + e^{N(x)}} \ln \frac{1}{L'}
\]  

The amount of information contained in the feature in each individual is \( H(x) \), assuming that the proportion of the feature whose value is greater than the set threshold \( \varepsilon \) is \( p(\varepsilon) \), then it is taken as the selection probability.

Load the intrusion information feature set, and set the expression of the initial parameter \( r \) as:

\[
r = S(x) \frac{\varepsilon}{H(x)}
\]  

The expression of chromosome node \( e \) in the initial population is established as:

\[
e = \frac{r}{p(\varepsilon)}
\]  

The first \( n \) individuals with the largest fitness function value are selected to form the set \( G \), and genetic algorithm is used to optimize the intrusion feature set to obtain the optimal intrusion feature set as follows:

\[
L'' = \frac{H(x)}{Gp(x)}
\]

2.2. Redundant information elimination model construction

In this paper, genetic algorithm is used to optimize the feature set, and the optimal intrusion feature subset is selected to build the intrusion information elimination model, which can delete the invalid and redundant information in the information set. The specific process is as follows:

If the number of attack information contained in the eigenvalue \( v \) is \( v_i \) and the probability of occurrence is \( p(v_i) \), then

\[
T = Ip(v_i) \sum_{i=1}^{n} p(v_i)
\]

Where: \( T \) is the information entropy of characteristic \( v \), \( I \) is the data source. For features \( v \) and \( u \), there are
\[ H(v | u) = \left( \sum_{i=1}^{n} \sum_{j=1}^{m} p(v_i, u_j) \right) I \frac{p(v_i, u_j)}{p(v_i) p(u_j)} \]  

(6)

Where \( p(v_i, u_j) \) is the joint probability of the value, \( p(u_j) \) is the probability of the value \( u_j \) occurrence.

To judge whether a new feature variable is necessary to enter the model or whether a feature variable can be deleted from the model, we need to use the method of partial F test, judge according to its significance, and get the intrusion feature set according to the result of the discrimination. Suppose there are self variables \( x_1, x_1, x_m, \beta_i \) is variable coefficients, and the redundant information elimination model of \( m \) is used as:

\[ y = \frac{S(x) H(v | u)}{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m} \]  

(7)

3. hybrid intrusion information retrieval model

3.1. Information retrieval theory

Intrusion detection belongs to the classification problem, which distinguishes the normal data flow and abnormal data flow in the network, and distinguishes the user operation and hacker operation. Intrusion detection technology can be divided into two categories, namely feature detection and exception detection. Feature detection is mainly used for the existing intrusion information in the network, while exception detection is mainly used to analyze the current network information. Network information is abnormal, that is, it is treated as an intrusion signal, and the activity is identified as "intrusion behavior". Information set is mainly a data set, which is a public knowledge structure, and can make up for the defects of knowledge structure. The main role of matching and selection is to compare information set and demand set, and find the information that meets the requirements according to the corresponding selection criteria.

3.2. Construction of intrusion information retrieval model

We use LDA model as the basis to build intrusion information retrieval framework, and build intrusion information retrieval model based on detection framework and LDA model.

Recall rate is also called recall rate, which is the ratio of the number of relevant documents retrieved and the number of all relevant documents in the document library. The intrusion information retrieval framework is composed of \( N \) documents, among which the documents contain \( K \) topics in total. The LDA model is used to establish the relationship between different topics and different documents, and the connection between topics and documents represents the probability that the document belongs to the corresponding topic. When entering a keyword for query, you need to select a topic with a keyword in the topic set, and search the probability of the selected topic containing the corresponding file. The returned documents are sorted by the probability relationship between the document and the intrusion information. The specific process is as follows:

In the process of building LDA model, the topic combination quoted mainly comes from Dirichlet related knowledge, which is the same in all documents. If the parameter \( f \) is probability distribution, \( z \) is topic number, \( N \) is document number, then the polynomial distribution can be expressed as:

\[ \phi_z = \frac{f}{N} \]  

(8)

Document \( d \) selects polynomial distribution probability as:

\[ \theta_d = \frac{z}{d} \phi_z \]  

(9)

Mark all the words in the document as \( w \), when referencing the language model for information retrieval, the probability of query item \( Q \) calculated by each document model is:

\[ p(Q,D) = \theta_d \prod_{q \in Q} p(q,D) \]  

(10)
Where \( D \) is the number of different types of documents, \( q \) is the query item, \( p(q,D) \) is the probability that the document contains the query item. If the query item is an independent individual, there are

\[
p(w \mid D) = \frac{N_d p(Q,D)}{N_d + \mu} p_{ML}(w \mid D) 
\]

Where: \( p(w \mid D) \) is the intrusion information retrieval evaluation information, \( p_{ML}(w \mid D) \) is the maximum likelihood estimate of the intrusion information \( w \) in the document; \( \mu \) is the prior knowledge, \( N_d \) is the amount of individual information.

Based on the above, the document is modeled. Compared with the traditional retrieval model, this paper proposes a new document modeling model, which uses this model to linearly combine with the original document and LDA model. In addition, the complexity of the algorithm is an important point that must be considered when designing an information retrieval model. During scanning or iteration, the running time of the proposed method is directly proportional to the number of documents. In order to reduce the running time of the algorithm, equation (11) is simplified as

\[
p(w \mid D) = \frac{N_d + \theta_d}{N_d + \phi_c} 
\]

Then retrieve the intrusion information, which can be described as

\[
W = \frac{(1-\mu)y}{p(w \mid D)} 
\]

4. Experimental results and analysis

In order to verify the comprehensive effectiveness of the proposed method for hybrid network intrusion information retrieval based on big data, experiments were used to analyze. The experimental computer configuration environment is: processor Inter (R) Core (TM) 2 Duo 2.8GHzPC, memory 2 Gbit, operating system is Windows7, programming language is C #, and the intrusion information in the database SQL Server2000 is used. The total amount of information is 6,000, including 5,000 normal data and 1,000 intrusion anomaly data, which are divided into 6 groups of sample data, and each group contains 1,000 data. The specific parameter settings are shown in Table 1.

| Parameter                          | Value  |
|------------------------------------|--------|
| Intrusion information space dimension | 4~9    |
| Information elimination times      | 1000   |
| Retrieval time interval            | >0.5s  |
| Intrusion information error        | <2.5%  |
| Total information                  | 6000   |
| Threshold                          | 30~40  |
| Document query probability         | 5~1.0  |

According to the parameter settings, the retrieval accuracy \( \eta \) is used as an experimental index, and \( J \) is the amount of intrusion data retrieved; \( Z \) is the total amount of intrusion data. The calculation formula for the retrieval accuracy \( \eta \) is

\[
\eta = \frac{J}{Z} \times 100\% 
\]

In the case of a certain amount of mixed intrusion information, the retrieval accuracy of the method in this paper is compared with the monitoring network hybrid intrusion information retrieval method based on the feedback algorithm, and the cloud network based monitoring network hybrid intrusion information retrieval method is compared. The comparison result is shown in Figure 1.
It can be seen from Figure 1 that with the increasing amount of retrieval information, the retrieval accuracy of different methods is also changing. In large-scale intrusion data retrieval, the retrieval accuracy of the proposed mixed intrusion information retrieval method based on big data increases with the increase of retrieval information, and the highest retrieval accuracy is close to 94%; due to the interference of redundant information, based on Anti Although the hybrid intrusion information retrieval method based on the feedback algorithm and the hybrid intrusion information retrieval method based on the cloud environment are on the rise, the maximum retrieval accuracy of the two methods is less than 40%. Through the comparison of experimental data, it can be seen that the hybrid
intrusion information retrieval method based on large data has higher retrieval accuracy, which is greatly improved. The performance of intrusion information retrieval is discussed.

It can be seen from Figure 2 that in 2~4s, the recall rate of this method is higher than the mixed intrusion information retrieval method based on feedback algorithm, but after 4s, the recall rate of this method shows a downward trend, while the recall rate of the other two methods shows an upward trend, with a large upward range. The average recall rate of this method is about 24%, far lower than the other two methods. The results show that the retrieval performance of this method is better.

5. Conclusion

Aiming at a series of problems existing in traditional network hybrid intrusion information retrieval methods, this paper proposes a method based on big data for monitoring network hybrid intrusion information retrieval. Firstly, the feature set is used to optimize the selection, eliminate redundant information, extract the characteristics of intrusion information in big data, then combine the LDA model with the joint model, and use the optimal reference combination model to realize the intrusion data retrieval. The comparison of experimental results between this model and other two models shows that the proposed method have higher retrieval accuracy, can improve the retrieval efficiency, and can perform intrusion data retrieval more accurately.

References

[1] Wu, Y.L., Chen, J.P., Jia, Z.J., et al. (2018) Geological data ontology construction and its application in data retrieval. Geological Bulletin of China., 37: 945-953.

[2] Wang, D.X., Zhu, Y.Q., Pan, P., et al. (2016) Construction of geodata spatial ontology and its application in data retrieval. Journal of Geo-Information Science., 18: 443-452.

[3] Du, Z.H., Zhu, W.Y., Chowdhary, A., et al. (2016) Implementation of secure data retrieval schema in cloud storage by using ABE technology. Application Research of Computers., 33: 860-865.

[4] Chen, S.L., Sun, X.Y., Zhang, Y., et al. (2016) TA relevance feedback algorithm for motion data retrieval. Acta Electronica Sinica., 44: 868-872.

[5] Ao, Z.H., Zhang, Y.H., Zheng, D., et al. (2017) Data search scheme in e-health environment. Computer Engineering and Design., 38: 1709-1713.

[6] Xia, D.L., Liu, Y.K., Lu, S.X., et al. (2016) Hybrid network intrusion detection method based on ant colony algorithm and improved simplified swarm optimization. Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition)., 28: 406-413.

[7] Lou, D.F., Pei, X.B., Wang, Z.Q., et al. (2016) Research on power search engine technology based on cloud computing and big data technology. Grid and Clean Energy., 32: 86-92.