Open Access

An Optimized Apriori Algorithm Based on Sparse Matrix for Intrusion Detection

Xiaohui Zeng[1,2]*, Jingxiang Lv[1], Jingzhong Li[1] and Wenlang Luo[1]*

[1]Dept. of Computer Science and Technology, Jinggangshan University, Ji'an, China

[2]The Key Laboratory of Embedded System and Service Computing, Ministry of Education, Tongji University, Shanghai, China

Abstract: The original Apriori algorithm is widely used in the intrusion detection field, but it may consume incredible computing resources in the process of handling network packets. We propose our optimized-Apriori algorithm which can greatly improves the algorithm efficiency by means of reducing the data storage space and the number of frequent item sets. We take full advantage of the characteristics of sparse matrix to save storage space and shorten the running time. The experimental results show that our optimized-Apriori algorithm can make a larger progress in saving the storage space of network data to be processed, even shortening nearly half of the data mining time in the process of intrusion detection.

Keywords: Apriori algorithm, data mining, intrusion detection, sparse matrix.

1. INTRODUCTION

The problem of mining association rules in database item sets is first proposed by Agrawal in 1993 [1]. Researchers introduce the association rules to solve the problem of low real-time and low efficiency in traditional intrusion detection technologies. Lee (1998), using the data mining method to study the problem of intrusion detection, proposes the use of association rules and frequent episodes algorithm to analyze the internal and external audit data [2]. Apriori algorithm is a widely used data mining algorithm in the intrusion detection field, but it has two major defects: (1) scanning the transaction database for many times will cause a lot of I/O load; (2) and it is likely to produce a large candidate set, which may consume the incredible computing time and storage space.

Lately, aiming at the application of association rule in intrusion detection, a large number of studies have been conducted to improve the efficiency of the algorithm. Hofmey (1998) proposes the algorithm of matching short sequences of system calls executed by running processes, normal sequence being used to calculate the similarity degree of the actual system call sequence model [3]. Lee (2000) proposes a data mining framework for constructing intrusion detection models, using the association rules and frequent episodes computed from audit data as the basis for guiding the audit data gathering, and axis attribute(s) and reference attribute(s) are used as forms of item constraints to compute only the relevant patterns [4]. Barbará (2001) utilizes data mining technology for anomaly detection, which combines association rules and classification technology to classify the suspicious TCP connections as the known attack types, the unknown attack types or false positives [5]. Qin (2004) introduces a new trace method for generating frequent episode rules, adopting fundamental pruning techniques to reduce the rule search space [6]. Hwang (2007) designs a hybrid intrusion detection system, which combines the advantages of low false-positive rate of signature-based intrusion detection system and the ability of anomaly detection system to detect novel unknown attacks by means of mining anomalous traffic episodes from Internet connections [7]. TIAN (2009) proposes a novel method for anomaly detection based on system calls into the kernels of Unix or Linux systems, adopting the data mining technique to model the normal behavior of a privileged program and using a variable-length pattern matching algorithm to perform the comparison of the current behavior and historic normal behavior [8]. The U.S. Patent 8,390,485 presents a method for improving performance of data decoding using apriori information about the data steam [9]. The U.S. Patent 8,565,324 describes a communications device includes a phase and frequency tracking loop having a signal input and adjustable loop filter, which is related to the known or measured apriori tracking capabilities of demodulator based on the symbol rate of communication signal [10]. The U.S. patent 7,069,179 proposes a workflow mining system and a method which can evaluate, analyze and determine previous execution results of processes or activities by applying a data mining technique to workflow log data accumulated during the operation of a workflow system, in which Apriori algorithm is used for the association analysis. [11]. Furthermore, other U.S. patents also mention the Apriori algorithm problem [12-14]. This paper analyzes the original Apriori algorithm, and takes full advantage of the characteristics of sparse matrix in intrusion detection to save the computing time and storage space of data mining in the process of handling network packets.

*Address correspondence to these authors at Dept. of Computer Science and Technology, Jinggangshan University, Ji'an, China, 343009;
Tel: 0796-8100470; E-mails: zeng_xhui@163.com, and lwl@jgsu.edu.cn

Table 1. The original IP packets data.

No.	Source IP Address	Destination IP Address	Protocol Type	Header Length	Packet Length
1	51.211.230.2	202.97.27.194	TCP	10	2048
2	10.1.1.2	202.97.27.194	UDP	5	1024
3	61.203.94.218	202.97.27.194	TCP	15	12384
4	10.1.1.2	202.97.27.194	UDP	10	2048

Table 2. Mapped data of IP packets.

No.	Source IP Address	Destination IP Address	Protocol Type	Header Length	Packet Length
	A	*B*	*C*	*D*	*E*
1	51.211.230.2→1	202.97.27.194→1	TCP→1	10→1	2048→1
2	10.1.1.2→2	202.97.27.194→1	UDP→2	5→2	1024→2
3	61.203.94.218→3	202.97.27.194→1	TCP→1	15→3	12384→3
4	10.1.1.2→2	202.97.27.194→1	UDP→2	10→1	2048→1

Table 3. The IP packets data in the form of binary two-dimensional list.

No.	A1 A2 A3	B1	C1 C2	D1 D2 D3	E1 E2 E3
1	1 0 0	1	1 0	1 0 0	1 0 0
2	0 1 0	1	0 1	0 1 0	0 1 0
3	0 0 1	1	1 0	0 0 1	0 0 1
4	0 1 0	1	0 1	1 0 0	1 0 0

2. ANALYSIS AND OPTIMIZATION

2.1. Analysis

In this paper, our improved Apriori algorithm is described as follows: First the original network data is only scanned for one time, and then the scanned network data are encoded to generate a binary two-dimensional array by bitmap representation, being put in a two-dimensional list. Thus, scanning the original network data is changed to scanning a two-dimensional list, and the data structure in the form of matrix is used to count the array. Hence, the number of times for scanning the original network data is greatly reduced and the time needed for data mining is shortened.

The implementation method is described as follows: All the Internet Protocol (IP) network packets on the host machine are intercepted, and the needed items (such as the source IP address, destination IP address, protocol type, header length, packet length, etc.) from packets, are sorted out. The corresponding value of each item is mapped into simple data and is counted, and then the single item value is as column and the IP packet number is as row, which are formed a binary two-dimensional array. As the 0 and 1 in the binary two arrays are arranged like a bitmap in computer graphics, the implementation method is called after binary bitmap representation method by us. The examples are shown in the (Tables **1** and **2**).

2.2. Optimization

Our method still has disadvantages. The algorithm maps the IP data to a binary bitmap list, and for each line item, it will only save a “1” and many “0”. If the number of counted items is increased, there will be more “0” in the binary bitmap list, but those “0” will be useless and consume large amounts of storage space in the process of operation.

The aim of optimization is to reduce the useless data in the binary bitmap list, saving storage space and improving the efficiency of algorithm.

The idea of optimization is as follows: This binary bitmap list is similar to a matrix, especially being conformed to the characteristics of the sparse matrix, because there are a large number of “0” and only a small amount of “1” in the binary bitmap list. Hence, we can convert the original matrix to a sparse matrix, saving only with the "1" of data bits and not saving with the "0" of data bits. Such optimization can greatly reduce the memory space and computation time in the course of operating the matrix, effectively improving the efficiency of algorithm. The data structure of our algorithm is as follows:

Ordinary sparse matrix is defined as follows:

$$(x, y, e) \rightarrow (\text{row, columns, element})$$

Table 4. Converted sparse matrix in the form of a two-dimensional array.

No.	A	B	C	D	E
1	1	1	1	1	1
2	2	1	2	2	2
3	3	1	1	3	3
4	1	1	2	1	1

Table 5. Intrusion detection effect of optimized-Apriori algorithm.

Data Sets Effect	1	2	3	4	5	6
detection rate	**86.5**	**81.8**	**85.6**	**82.9**	**86.3**	84.9
false positive rate	16.8	16.3	15.3	8.6	14.6	9.8

According to the (Table **3**), our sparse matrix is defined as follows:

(x, y, e) → (row, attribute, value)

Thus, the (Table **3**) can be represented as below:

(1, 1, 1) (1, 2, 1) (1, 3, 1) (1, 4, 1) (1, 5, 1)
(2, 1, 2) (2, 2, 1) (2, 3, 2) (2, 4, 2) (2, 5, 2)
(3, 1, 3) (3, 2, 1) (3, 3, 1) (3, 4, 3) (3, 5, 3)
(4, 1, 2) (4, 2, 1) (4, 3, 2) (4, 4, 1) (4, 5, 1)

Furthermore, the "x" values of each row are equal, and the "y" values of each column are also equal, so we can convert (Table **3**) into the form of (Table **4**).

Based on computer memory space, the storage space of our binary two-dimensional array is as follows:

Total number of items' values = ∑(items*number of items)

Total number of storage space1= total number of items' values* number of rows (bit)

If our converted sparse matrix is also stored in the form of two-dimensional array, the required storage space is as follows:

Total number of storage space2= number of items* number of rows (byte)

The byte unit is converted into bit unit, and total number of storage space2 is as follows:

Total number of storage space2= number of items* number of rows*8 (bit)

3. EXPERIMENT

To test the performance of our method, we implemented the original Apriori algorithm and our optimized-Apriori algorithm. Experiments were conducted on the computer with 3.10 GHz CPU, 3 GB memory and Windows XP SP3 operating system. The experimental data is KDDCUP 1999 data sets [15], very authoritative test data sets in the intrusion detection field, which are from MIT Lincoln laboratory DARPA intrusion detection evaluation data sets. In the test data sets, there are 4 major categories, which are the DOS attack (Denial of Service), the U2R (User to Root) attack, the R2U (Remote to User) attack and the Probing attack. Six data sets were randomly chosen from the KDDCUP 1999 data sets as our experimental data sets.

The experimental results are respectively shown in the (Table **5**), the Figs. (**1** and **2**). The experimental results show

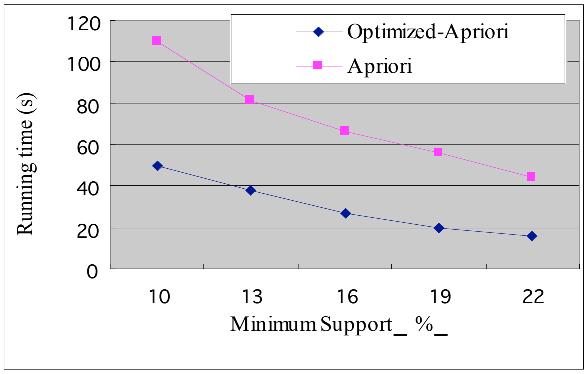

Fig. (1). Running time under the condition of different minimum support.

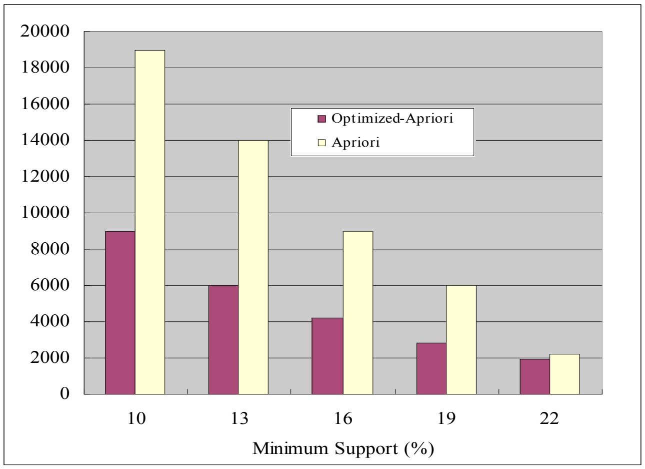

Fig. (2). Number of frequent sets under the condition of different minimum support.

that our optimized-Apriori algorithm on the detection effect is roughly equal to the original Apriori algorithm [16], but our optimized-Apriori algorithm has made a larger increase in speed, saving nearly half of the intrusion detection time.

4. CURRENT & FUTURE DEVELOPMENTS

The original Apriori algorithm doesn't fully consider the problem of data storage space, but our optimized-Apriori algorithm greatly improves the efficiency algorithm by means of decreasing the data storage space and the number of frequent item sets. Using sparse matrix, our optimized-Apriori algorithm greatly save the data storage space, and computation time is also saved. The experimental results show that the optimized-Apriori algorithm can greatly save the running time of data mining in the process of intrusion detection, and can save storage space of the network data to be processed. In the future, we will apply our optimized-Apriori algorithm to the actual intrusion detection system.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

This work is supported in part by the natural science foundation of china (Grant No. A11364023), the Opening Project of Key Laboratory of Embedded System and Service Computing of Ministry of Education (Grant No. 2010-09), the key technology R&D program of Jiangxi province (Grant No. 20122BBG70161), the Science & Technology Transformation Program for Universities in Jiangxi Province (Grant No. KJLD13068), the Jiangxi province natural science foundation (Grant No. 20122BAB201038), the Science and Technology Plan Project of Jiangxi Provincial Education Department (Grant No. GJJ13538 and GJJ14561), and the Jinggangshan University project (Grant No. JZB11013).

REFERENCES

[1] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, Washington, D.C., United States, 1993.

[2] W. Lee, S.J. Stolfo and K.W. Mok, "Mining Audit Data to Build Intrusion Detection Models," in *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining*, New York, United States. 1998: pp. 66-72.

[3] S. A. Hofmeyr, S. Forrest, and A. Somayaji, "Intrusion detection using sequences of system calls," *J. Comp.Security*, vol. 6, pp. 151-180, 1998.

[4] W. K. Lee, S. J. Stolfo, and K. W. Mok, "Adaptive intrusion detection: A data mining approach," *Artif. Intel. Rev.*, vol. 14, pp. 533-567, Dec. 2000.

[5] D. Barbara, J. Couto, S. Jajodia, and N. N. Wu, "ADAM: A testbed for exploring the use of data mining in intrusion detection," *Sigmod Record*, vol. 30, pp. 15-24, Dec 2001.

[6] M. Qin, and K. Hwang, "Frequent episode rules for internet anomaly detection," in *Proceedings of Network Computing and Applications*, 2005, pp. 161-168.

[7] K. Hwang, M. Cai, Y. Chen, and M. Qin, "Hybrid intrusion detection with weighted signature generation over anomalous Internet episodes," *IEEE T Depend Secure*, vol. 4, pp. 41-55, Jan-Mar 2007.

[8] X.G. Tian, M. Y. Duan, C. L. Sun, and X. Liu, "Detecting network intrusions by data mining and variable-length sequence pattern matching", *J. Syst. Eng. Electron.*, vol. 20, pp. 405-411, Apr 2009.

[9] Rijavec and Nenad, "*Fast data stream decoding using apriori information*," U.S. Patent 8,390,485, Aug. 28, 2007.

[10] W. N. Furman, J. W. Nieto, and W. L. Tyler, "*Communications device using measured signal-to-noise ratio to adjust phase and frequency tracking*," U.S. Patent 8,565,324, Oct. 22, 2013.

[11] Y. Kim, S. Kim, B. Kwak and Y. Shin, "*Workflow mining system and method*," U.S. Patent 7,069,179, June 27, 2006.

[12] F. David, "*System and method for efficiently generating association rules*," U.S. Patent 8,401,986, March 19, 2013.

[13] S. Roychowdhury, "*Method for extracting association rules from transactions in a database*," U.S. Patent 7,370,033, May 6, 2008.

[14] S. Nirad, N. Rohit and F. David, "*Attribute based association rule mining*," U.S. Patent 7,433,879, Oct. 7, 2008.

[15] The KDD Archive. KDD99 cup dataset, 1999. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[16] H.-J. Liao, C.-H. Richard Lin, Y.-C. Lin, and K.-Y. Tung, "Intrusion detection system: A comprehensive review," *J. Netw. Comput. Appl.*, vol. 36, pp. 16-24, 2013.

Received: July 23, 2014 Revised: August 13, 2014 Accepted: August 16, 2014