PROCEEDINGS OF SCIENCE

Detecting Anomalous User Behavior in Database

Jieqing Ai , Lihao Wei , Jianyong Wang, Haonuo He, Jinpeng Chen[1], Chengdong Liang , Liang Chen

Information center of Guangdong Power Grid Corporation
Guangzhou, 510000, China
E-mail: 13522066745@139.com

In order to protect vital data in today's internet environment and prevent misuse, especially insider abuse by valid users, we propose a novel two-step detecting approach to distinguish potential misuse behaviour (namely anomalous user behaviour) from normal behaviour. First, we capture the access patterns of users by using association rules. Then, based on the patterns and users' sequential behaviour, we try to deter anomalous user behaviour by leveraging the logistic regression model. Experimental results on real dataset indicate that our method can get a better result and outperform two state-of-the-art method. The proposed two-step detecting approach can effectively detect anomalous user behaviour from the log data generated by operation and maintenance staffs.

CENet2017
22-23 July 2017
Shanghai, China

[1]Speaker

PoS(CENet2017)096

1.Introduction

Nowadays information is the most worthy treasure of organizations, which demands reasonable conservation and management. As everyone knows, database systems play a vital role in the aspect of managing and retrieving a great deal of data, and they also supply mechanisms to assure the completeness of the saving data. Database systems can utilize private user account to log-in information, and designate only particular user accounts to have access to important table in database systems. However, this technical solution fails once a malicious attacker gains log-in information of a user account. Therefore, to overcome this technical problem, techniques described in this work can determine whether specific actions taken by user accounts indicate that the user account is compromised.

There are different approaches employed to detect anomalous action [1, 2, 3], but unluckily an entirely ideal solution cannot be found so far. In some intrusion detection systems, false alarms may be generated during the detection process. In this paper, we propose a new two-step detecting method to discover anomalous user behavior.

This work is organized in the following step Section 2 contains related work. Section 3 describes the problem and explains the methodology of the proposed work. Section 4 describes the data set used and includes the details of performance evaluation based on the experimental study. Section 5 refers to conclusion and future enhancement.

2.Related Work

Plenty of methods have been proposed to extract user behavior from various data sources [4, 5, 6, 7]. In this work, we have conducted two researches: supervised learning and unsupervised learning. Some existing work on detecting user behavior leveraged supervised learning techniques. Xie et al. [8]. presented a method that can automatically produce URL signatures for spamming botnet prediction, namely AutoRE. Beutel et al. [9] first offered a new definition of dubious action based only on network structure and edge constraint conditions. Egele et al. [10] put forward a system to mine compromised accounts in social networks, namely COMPA. Tan et al. [11] first analyzed the disadvantages of existing representational programs, and then designed a sybil defense based on spam prediction pattern. Rahman et al. [12] presented the design and implementation of MyPageKeeper, a Facebook application that can accurately and efficiently identify software at scale. Wang et al. [13] proposed a detection method that divides "similar" user clickstreams into behavioral clusters, by dividing a similarity graph that catches distances between clickstream sequences.

3.Two-step Detecting Algorithm

3.1Problem Definition

Our aim is to mine anomalous user behavior with priori knowledge of the attacker tactics. Our core assumption is that attacker behavior ought to show anomaly against normal user behavior with some (un-known) potential characteristics.

3.2Detecting Method

In this part, we describe the main components of the detecting system. A high level representation is depicted in Figure 1. The basic framework can be divided into two phases – learning and detection.

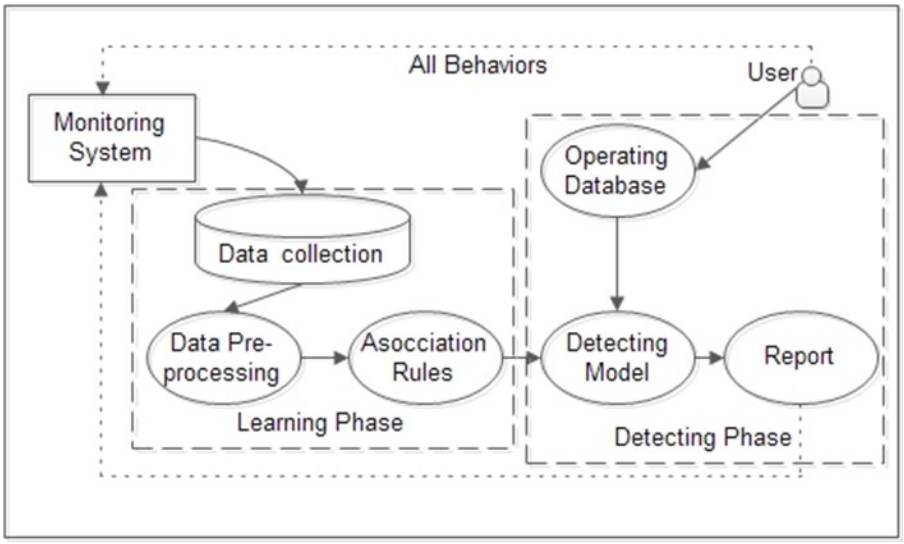


Figure 1: Overall structure of Detection System

3.2.1Learning Phase

During this phase, the model of legitimate queries is built by using association rules. We assume every transaction currently executed in the database to be benign. We conjecture that a certain user typically will not look up all attributes and data in a pattern and databases. So users' access patterns will generate some frequent itemsets which are collections of attributes that are usually referenced together with some values. A profile captures users' intention to specify the typical values.

We first depict the data structures before presenting the algorithm. Any SQL query can be written as the following general form with three clauses.

SELECT Attributes

FROM Tables

WHERE Conditions

For every SQL query we associate a quadruple q = <S, R, A, C, T> which represents the users' profile. We define a set of quadruples Q={q1, q2,…,qn}, where, 'S' stands for the type of query (SELECT), 'R' stands for the number of relations in the query (in this work, it means sensitive relationship), 'A' stands for the number of Attributes in the query, 'C' represents the number of Conditions in the query and 'T' stands for the query time.

Definition 1 A sequence is a sequential list of operations. We define a sequence o by <q1 , q2, …, qn >, where qi∈Q and T1<=T2<=…<=Tn. The support degree of a sequence is denoted as the ratio that this sequence occurs in the total transactions.

Then, we propose a new algorithm, namely, the Frequent Itemset Profiler, to detect the frequent itemsets from users' behavior. This algorithm aims to discover all minimal frequent itemsets which are divided into four steps. Steps 1 initializes the sequence sets. In Step 2, we use the ordered pattern mining algorithm to obtain ordered patterns which contain a sequence of operations to satisfy the minimum support. Step 3 constructs legal sensitive sequence sets from

the sequential patterns mined. We find all the rules can satisfy minimum support in the operating rule sets, which can be seen as the output of the algorithm.

We can use a set of efficient rules to denote the normal and abnormal behaviors of users when a user operates the database; therefore, we need some strategies to monitor the anomalous behavior in the process of operating database. In the next section, we will introduce our detecting model.

Algorithm 1: Frequent Itemset Profiler

Input: All users' sequences O ={o1,o2,…,on}
Output: Frequent Itemset *F*
Step1: Derive sequences *o1,... on*, where n is the total number of sequences in the
database schema.
Step2: Produce the ordered patterns X = {xi | support(xi) > minimum support}
by leveraging the ordered pattern mining algorithm ;
Step3: for each ordered pattern xi where | xi | > 1
if there's an operation in it
for each operation oi∈xi

if <o1, o2 , ,…, om >∉ F and T1<=T2<=,…,Tm

add <o1, o2 , ,…, om > to *F*
End if
End for
End if
End for
Step4: Rerurn *F*.

3.2.2Detecting Phase

In this paper, we mine anomalous user behavior by using the logistic regression model. Assumed each training pair of user-query sequence <u, o>, let xi be the (n + 1) dimension vector containing constant 1 and n user query features , and yi be the mark of whether users' behavior is anomalous. By using user query-based features, we can define the probability Proi which obeys binomial distribution in the following:

$$Pro_i = P(y_i = 1, x_i) = \frac{1}{1 + e^{-x_i \alpha}} \tag{3.1}$$

where yi=1, if users' behavior is anomalous, otherwise 0. α is the n + 1 coefficient weights relevant with the constant and each user feature. To obtain the optimal regression coefficience, we leverage maximum likelihood estimation (MLE) to approximate it, which maximizes the likelihood of all training pairs: $L(\alpha) = \Pi_i \pi(x_i)^{y_i} (1 - \pi(x_i)^{(1-y_i)}$.

4.Experiments

4.1Dataset and Setup

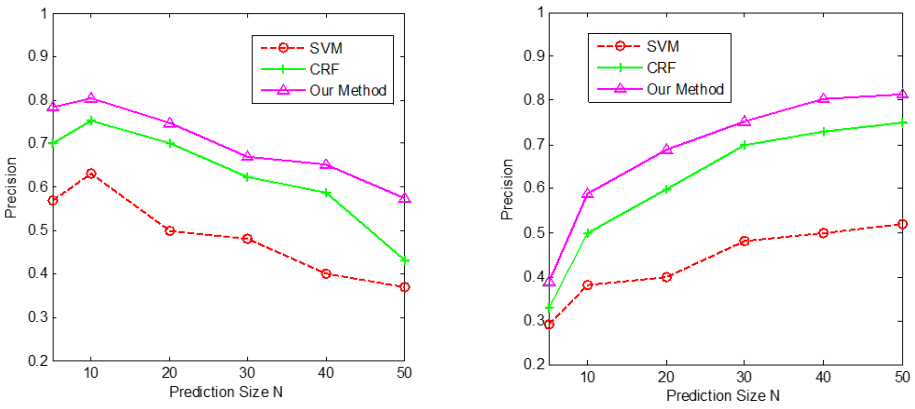
Our experimental dataset was from log data generated by operation and maintenance staffs when operating database. This dataset includes account, name, department, deputy account, operating time and operations. To decrease the size of the raw data, we removed the operations which are meaningless, such as “Select * from dual ”, “Select Null from dual”. In addition, the dataset, in our experiments, were divided into training and test sets.

In this experimental evaluation, we measure the effectiveness of the presented method on the basis of Precision@K and Recall@ K. Precision@K is the ratio of recovered user-query sequences to the K predicted user-query sequences. Recall@K is the ratio of recovered user-query sequences to the set of user-query sequences deleted in preprocessing.

We select two baseline methods: One is SVM, which uses attributes related to each edge as features xi generates a classification model and then estimates edges’ flags by utilizing the classification model. We adopt SVM-light [14] as our baseline. Another one is CRF, which produces a conditional random field [15].

4.2Experimental Results and Analysis

From Fig. 2(a), the predictions obtained by our proposed methods are significantly superior to the SVM method. Computing the average improvement, we obtain a mean value of 24.4%. Meanwhile, we can observe that the proposed method is closely approximate to CRF, but the proposed method still outperforms CRF. Computing the average improvement, we obtain a mean value of 7.1% primarily because our prediction method can not only catch the individual characteristics of the users’ behavior but discover the frequent itemsets from the users’ behavior, in the behavior decision strategy. In Fig. 2(b), we analyzed the behavior of baselines and our proposed algorithm when steadily increasing prediction size N. We noticed that the recall value decreases as the prediction size increases. Indicated by the results, our method obviously outperforms the SVM and CRF method.



(a) Precision (b) Recall

Figure 2: Results of the Corresponding Metrics with Some State-of-art Approaches and Our Method

5.Conclusion

In this work, a novel efficient method to monitor user behavior is presented to distinguish potential misuse behavior (namely, anomalous user behaviors) from normal behavior. We first

capture the access patterns of users by using the association rules. Then, based on the patterns and users' sequential behaviors, we try to deter anomalous user behavior by leveraging the logistic regression model. Experimental results indicate that the proposed approach can produce reasonable and high quality detection of anomalous user behaviors.

References

[1] Chari S N, Molloy I M, Park Y, et al. *Detecting anomalous user behavior using generative models of user actions*[J]. 2017.

[2] Khan M I, Foley S N. *Detecting Anomalous Behavior in DBMS Logs*[J]. 2016.

[3] Menahem E, Schclar A, Rokach L, et al. XML-AD: *Detecting anomalous patterns in XML documents*[J]. Information Sciences, 2016, 326:71-88.

[4] Ballesteros L G M, Örblom M, Markendahl J, et al. Effects of Network Performance on Smartphone User Behavior[C]// Pqs 2016, Isca/dega Workshop on Perceptual Quality of Systems. 2016:79-82.

[5] Graus M P, Willemsen M C, Swelsen K. Understanding Real-Life Website Adaptations by Investigating the Relations Between User Behavior and User Experience[C]// International Conference, Umap. 2015:350-356.

[6] Althoff T, Jindal P, Leskovec J. Online Actions with Offline Impact: How Online Social Networks Influence Online and Offline User Behavior[C]// Tenth ACM International Conference on Web Search and Data Mining. ACM, 2016:537-546.

[7] Wang G, Zhang X, Tang S, et al. Unsupervised Clickstream Clustering for User Behavior Analysis[C]// CHI Conference on Human Factors in Computing Systems. ACM, 2016:225-236.

[8] Xie Y, Yu F, Achan K, et al. *Spamming botnets:signatures and characteristics*[J]. Acm Sigcomm Computer Communication Review, 2008, 38(4):171-182.

[9] Beutel A, Xu W, Guruswami V, et al. *CopyCatch:stopping group attacks by spotting lockstep behavior in social networks*[J]. 2013, 33(9):119-130.

[10] Egele M, Stringhini G, Kruegel C, et al. *Towards Detecting Compromised Accounts on Social Networks*[J]. IEEE Transactions on Dependable & Secure Computing, 2015, 12(2):91-98.

[11] Tan E, Guo L, Chen S, et al. *UNIK: unsupervised social network spam detection*[C]// ACM International Conference on Conference on Information & Knowledge Management. ACM, 2013:479-488.

[12] Rahman M S, Huang T K, Madhyastha H V, et al. *Efficient and Scalable Socware Detection in Online Social Networks*[J]. Usenix Security, 2012.

[13] Wang G, Konolige T, Wilson C, et al. *You are how you click: clickstream analysis for Sybil detection*[C]// Usenix Conference on Security. 2013:241-256.

[14] T. Joachims, Making Large-Scale SVM Learning Practical, MIT-Press, 1999.

[15] J.D. Lafferty, A. McCallum, F.C.N. Pereira, Conditional random fields: probabilistic models for segmenting and labeling sequence data, ICML'01, 2001, pp. 282–289.