DETECTING DEVIATION IN SCADA LOGS USING RARE FREQUENT PATTERN MINING

M.Natarajan, Assistant Professor,
Department of Coputer science
and Engineering,
K.Ramakrishnan College of
Technology, Trichy. Tamilnadu.
forevernatarajan@gmail.com

R.Pradeepa, Assistant Professor
Computer Science and
Engineering
M.Kumarasamy College of
Engineering
Karur,Tamilnadu
pradeepar.cse@mkce.ac.in

R.Bharathi, Assistant Professor
Computer Science and
Engineering
M.Kumarasamy College of
Engineering
Karur,Tamilnadu
vinothm.cse@mkce.ac.in

Abstract— Sequential pattern mining is a topic of data mining concerned with finding statistically relevant patterns between data examples where the values are delivered in a sequence. An interesting pattern as a pattern that appears *frequently* in a database. It is usually presumed that the values are discrete, and thus time series mining is closely related, but usually considered a different activity in network. In this paper we use rare frequent pattern for detecting abnormal deviation in SCADA. We use a group of algorithms to solve this intrusion and detection problem through three phases: preprocessing to extract probabilistic and identify sessions for different events, generating all the SETP events with (minsup) support values for each event by using pattern-growth algorithm, and selecting URSETPs by detecting user-aware rarity analysis on derived STPs.

Keywords— sequential patterns, rare events, pattern-growth, dynamic programming

Introduction

Anomaly identification is one of the insurance measures utilized as a part of basic foundation arrange. SCADA is utilized to screen and control the CI. A few ward CIs SCADA is interconnected .CI successive occasions are control and screen by SCADA IN remote area. Deviation on SCADA can influence other ward framework condition and human moral too. Customary IT innovation is the foundation of SCDA for conveying field gadgets. So they are inclined to be assaulted utilizing some standard IT foundation vulnerabilities. Presently inconsistency discovery is one of the testing and critical issue in SCADA. Ceaseless Change in assault examples is one of the explanation behind making basic for oddity discovery .In a conventional IT systems framework, it is not difficult to keep the framework shielded from expanded assaults. SCADA is not the same as conventional IT systems ,on the grounds that the standard or typical occasions of SCADA can be visualized utilizing incessant example mining .Rare example or unpredictable occasions of this framework are consider as an abnormal occasions. The sporadic occasions are veer from customary occasions .

In this paper we accept uncommon example or unpredictable occasions of this framework are consider as a strange occasions. At that point break down the uncommon occasions to discover Cyber Attacks, Dos, Data-Focused Attacks, are the essential purpose of assault against SCADA frame works. In this paper uncommon successive example mining calculations used to in his paper, a few new ideas and the mining issue are formally characterized, and a gathering of calculations are planned and consolidated for deviation identification from SCADA logs .

To summarize , this paper makes the subsequent supports

•In this paper the principal work that characterizes legitimate meanings of STPs notwithstanding their irregularity measures, and moves propel the issue of mining URSTPs , to identify customized and strange occasions of CI system.

•We propose a structure to for all intents and purposes take care of this issue, and configuration relating calculations to bolster it. First and foremost, we give preprocessing methodology with heuristic techniques for occasion extraction and session recognizable proof.

•Then, getting the thoughts of example development in indeterminate condition, two option calculations are intended to find all the SETP applicants with bolster values for every occasion. That gives an exchange off amongst exactness and productivity.

•In the end, we introduce a client mindful irregularity investigation calculation as indicated by the formally characterized model to choose URSTPs and related occasions.

II. Related Works

Discovering irregularities in SCADA has been broadly contemplated in the writing. Utilizing Model-based Intrusion Detection for SCADA Networks intended to Modbus convention typified inside TCP/IP . Display based identification is a critical supplement to signature-based approach. This approach works at the correspondence level protocol[3].To methodically recognize potential process-related dangers in SCADA. Break down process-related dangers that happen in the PC frameworks utilized as a part of basic foundations. Such dangers occur when an assailant figures out how to increase substantial client ID and follow up on to upset a blockaded modern process, or when a true blue client commits an operational error and causes a procedure failure. However just a solitary work uses SCADA logs where Dina Hadziosmanovi et al[4] utilized item set digging for distinguish the deviation in CI networks.[5]show that the nonattendance of regular occasions or set of occasions considered as an inconsistency. or, on the other hand portraying the "ordinary" stream of cautions from a sensor. Utilizing such models tuned to individual sensors, we then built up a system for distinguishing peculiarities. [6]It connected an information mining strategy to distinguish the typical conduct a framework in view of the incessant event of a caution occasion and later sifted them through from suspicious occasions records. A procedure that utilizations digging for consecutive relationship to distinguish regular false alarms. Pattern coordinating and anomaly identification techniques are utilized for discovering abuse recognition system.[7]Sequential strategy in data mining used to recognizing typical conduct ,this technique used to channel the not successive events. Association govern mining is utilized to discover ordinary conduct from system framework to preparing model framework. on the off chance that any deviation happen in a preparation model, it will considered as unusual conduct in CI organize framework.

The majority of existing works are dedicated to discover uncommon example utilizing item set mining. Rare examples cannot utilized as a part of location of deviation in SCADA logs and furthermore not give the protected thing sets.[8]find the examples from the base successive thing sets. In this strategy master cannot give any connections between's events. We utilize consecutive technique rather than thing mining enemy same data. The successive strategy used to safeguard the request of events, which is give the relationship between's events. Therefore we utilize visit uncommon incessant example digging for identifying deviation in CI in system. Rare successive mining is one of the branch successive mining.

III. Proposed Method

we give some preliminary notations, define several key concepts related to SETP, and formulate the problem of mining URSETPs to be handled in this paper.

Definition 1 (sequence database).
A sequence database is a sequence of ordered elements or events recorded with or without
a existing concept of occasion.Let X={X_1,X $_2$,..X_n} be a set of items.An itemset $X_{I=}${$X_{1,}$ X_2 ......$X_{,n}$ } $\subseteq$ X is an non empty or ordered set of different items.

Definition2 :sequential pattern mining
A Sequential Pattern Mining (STP) α is characterized as a succession sp1 sp2; . . . ; spni,where each SPi T is a learnt subject. n ¼ j aj indicates the quantity of subjects contained in an, and is known as the length of a. An example with length n is called a n-STP.

Definition3:(Session). Given a system CI stream TDS, a session s is characterized as a subsequence of TDS related with a similar client, i.e., s =(td1,u,t1),(td2,u,t2),… … .,(td2,u,t2)≤TDS.

Definition4:
(Pattern Instance). Given a SETP α ={X1,X2,X3… XN) and a session S related with a client U, on the off chance that we can separate a subsequence S'=(td1,u,t1),(td2,u,t2),… … .,(td2,u,t2)≤S and for each i=1,2… .N tdi holds for some SN. The event likelihood of the example can be essentially computed as an item P(α')=SUBSET OF PK.
Client Aware Rare Sequential Topic Patterns a lot of existing takes a shot at successive example mining focusing on continuous examples, however for USETPs, numerous occasional ones are likewise intriguing and ought to be found. intentionally, when Internet clients' distribute records, the customized practices portrayed by UPSTPs are by and large not all inclusive continuous but rather even uncommon, since they uncover extraordinary and irregular inspirations of individual creators, and additionally specific occasions having jumped out at them, all things considered.
Therefore, the UPSTPs we might want to dig for client conduct examination on the Internet ought to recognize elements of included clients, and in this way fulfill the accompanying two conditions:
1) They ought to be all inclusive uncommon for all sessions including all clients of an archive stream;

2) They ought to be locally and moderately visit for the sessions related with a particular client.

Next, we will formally determine this sort of SETPs well ordered, beginning with the established idea of support to portray the recurrence. For deterministic successive example mining, the support of an example and is characterized as the number or extent of the groupings containing an in the objective database, yet inapplicable for unverifiable arrangements like theme level archive streams. Rather, the normal support is fitting to gauge the recurrence on speculative successions, and can be work out by summing up the occurrence probabilities of an in all arrangements [28]. At the end of the day, it communicates the normal number of groupings containing a.
To gauge the recurrence of SETPs, we alter it a little to record the extent of sessions where a happens additionally as far as desire, by means of partitioning the summation by the quantity of sessions. That is vital on the grounds that the session number here is no longer a steady when we consider both the worldwide recurrence and the neighborhood recurrence of a for various clients. For straightforwardness, this measure is still signified as support rather than anticipated support in this paper.
We are developed in the work of existing presented in rare item sets presented ion minimal generators. However we propose a method to find sequential patterns using sequential generators. This patterns are smaller patterns of equivalence classes.
Normally smaller patterns are frequnt,larger patterns are infrequent and rare patterns. Then the combinations of two minimal generator patterns would be rare or

infrequnt.T=Given two sequential patterns S1={{4} ,{3},{2}} and S2={{1,2},{6}} generated from the data set. we combine s1 and s2 and s2 ,s1.
s1 U s2=={{4} ,{3},{2} 1,2},{6} },s2 U s1=={{1,2},{6},{4} ,{3},{2} }.I f we need preserve both the integrity and the item set order of the original sequence.

Algorithm

```
1.User URSETP,ω'←ω;
2.find all the possible elements in $Z_a$ which can be appended
    to a to form a new STP, and record them in E;
3. for all z € E do
4.β←α o z;
5.$supp_\beta$ ←0;
6.$Pref_\beta$ ,$Z_\beta$ ←φ;
7. for all (i, $j_0$ ) € Zα do
8.find all the documents as instance of z in the projected $s_i$
   i.e.,{$id_j$ | $td_j$ € $s_i$ Λ j ≥ $j_0$ Λ $\exists$ $p_j$ .(z,$p_j$ ) € $td_j$ } and record in
   order each position j in J together with the corresponding
    probability
9. $Pref_\beta$ ← $Pref_\beta$ U {i,0,0}
10.j' ← max{x | x < j $\exists$ (i,x,p) € $Pref_\beta$ ;
11. find( i,j',$M_J$ n) € $Pref_\beta$
12.  for all k € l do
13. if $PreZ_\alpha$ == φ then
14. R ←$R_j$ + (1-$R_j$ ) X$R_j$
15. $M_\beta$ ←$M_\beta$ U {i,min{j | j € J}}
16.SETP1 ← SETP1 U { β,$SETP1_\beta$ }
17.SETP1 ← SETP1 U USETP1( Z,Pre $f_{\beta,}$ $Z_\beta$ );
18.return USETP
```

Every implementation of UETP performs one stage of example development from the information SETP α to an improved one β= α S, by affixing another component Z. In this we filter first every one of the arrangements in Sα to acquire the set E1 containing all the conceivable groupings .Then, for every theme X1 in E1 and each postfix match (i; j0) in Zα, we discover every one of the occasions of z in the anticipated addition of Zi, figure the likelihood of β =α Z happening in the connecting prefixes of Z, and record new prefix triples in Pre f β). In particular, for each occurrence of S, if Prefα = φ then an is a void succession, and β contains only one collection S, which transmit to the DP-lattice.

A.Client Aware Rarity Analysis

All clients SETP applicants are found the we will make client mindful irregularity examination .This include the customized ,anomalous and critical behaviors from the uncommon consecutive examples.

Algorithm-URSTP

```
1.User UP,ω'←ω;
2.Get all the pattern from ω from user SETP
3.Get the number sessions | Z | from user sessions
4. find all the possible elements in $Z_a$ which can be appended
to a to form a new SETP, and record them in F;
3. for all α € ω F do
4.β←α o S;
5.$SETP_\beta$ ←0;
6.$Pref_\beta$ ,$Z_\beta$ ←φ;
7. for all (m,$n_0$ ) € Zα do
8.find all the documents as instance of z in the projected $s_i$
   i.e.,{$md_j$ | $t1d_j$ € $s_i$ Λ j ≥ $j1_0$ Λ $\exists$$S_j$ .(S,$p_j$ ) € $t1d_j$ } and
   record in order each position j in J together with the
  corresponding probability

9. $PreJ_\beta$ ← $PreJ_\beta$ U {i,0,0}
10dj' ← max{x | x < j $\exists$ d(i,x,p) € $PreJ_\beta$ ;
11. find(mi,mj',$Z_J$ n) € $Pref_\beta$
12.  for all m € n do
13. if $Pref_\alpha$ == φ then
14. mp ← $mp_j$ + (1-$mP_j$ ) X $mP_j$
15. $Z_\beta$ ←$Z_\beta$ U {i,min{j | j € J}}
16.SETP ← SETP U { β, $R1R1_\beta$ }
17. USETP ← USETP U (β,Z,Pre $f_{\beta,}$ $Z_\beta$ );
18.return USETP
```

This helps to move set of users SETP into user rarity analysis. It use session pairs and two threshold values .Support threshold value is used to find support values in SETP. Relative threshold value is used to find rarity analysis between the given input.

IV. Experimental Setup

We make an assumptions the attacker can nor alter our database from SCADA logs. First dataset is collected logs from admin maintain LAN connection network. That network controls overall network from particular user services. That main network control all the other sub network .From that sub network we have collected remaining four data sets

In case of dataset 1 recorded events are included in a log data. Sub network data recorded only when the errors and rare sequential patterns are occur.

Data Preprocessing

The raw datasets are cleaned by preprocessing and needed information were selected. From dataset-1 we collected data set from every two minutes. In this we use the session identification to detect deviation between certain time limits. The rare sequence patterns are compared to rare sequence. These rare patterns are used to find anomalies events.

Conclusion

In this paper, we present an rare sequential pattern. At first we create frequent sequential pattern and minimal sequential pattern. Later minimal sequential pattern is compared with minimal sequential pattern to prune frequent patterns. Once again minimal frequent pattern compare with sequence database. These pattern are consider as rare and deviation event from the normal events. We will also validate and find computational performance of our methodology.

References

[1] C. C. Aggarwal, Y. Li, J. Wang, and J. Wang, “Frequent pattern mining with uncertain data,” in Proc. ACM SIGKDD, 2009, pp. 29–38.

[2] R. Agrawal and R. Srikant, “Mining sequential patterns,” in Proc. IEEE Int. Conf. Data Eng., 1995, pp. 3–14.

[3] J. Allan, R. Papka, and V. Lavrenko, “On-line new event detection and tracking,” in Proc. 21st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 1998, pp. 37–45.

[4] T. Bernecker, H.-P. Kriegel, M. Renz, F. Verhein, and A. Zuefle, “Probabilistic frequent itemset mining in uncertain databases,” in Proc. ACM SIGKDD, 2009, pp. 119–128.

[5] D. Blei and J. Lafferty, “Correlated topic models,” Adv. Neural Inf. Process. Syst., vol. 18, pp. 147–154, 2006.

[6] D. M. Blei and J. D. Lafferty, “Dynamic topic models,” in Proc. ACM Int. Conf. Mach. Learn., 2006, pp. 113–120.

[7] D. Blei, A. Ng, and M. Jordan, “Latent dirichlet allocation,” J. Mach. Learn. Res., vol. 3, pp. 993–1022, 2003.

[8] J. Chae, D. Thom, H. Bosch, Y. Jang, R. Maciejewski, D. S. Ebert, and T. Ertl, “Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition,” in Proc. IEEE Conf. Vis. Anal. Sci. Technol., 2012, pp. 143–152.

[9] K. Chen, L. Luesukprasert, and S. T. Chou, “Hot topic extraction based on timeline analysis and multidimensional sentence modeling,” IEEE Trans. Knowl. Data Eng., vol. 19, no. 8, pp. 1016–1025, Aug. 2007.

[10] C. K. Chui and B. Kao, “A decremental approach for mining frequent itemsets from uncertain data,” in Proc. 12th Pacific-Asia Conf. Adv. Knowl. Discovery Data Mining, 2008, pp. 64–75.

[11] W. Dou, X. Wang, D. Skau, W. Ribarsky, and M. X. Zhou, “LeadLine: Interactive visual analysis of text data through event identification and exploration,” in Proc. IEEE Conf. Vis. Anal. Sci. Technol., 2012, pp. 93–102.

[12] G. P. C. Fung, J. X. Yu, P. S. Yu, and H. Lu, “Parameter free bursty events detection in text streams,” in Proc. 31st Int. Conf. Very Large Data Bases, 2005, pp. 181–192.

[13] J. Han, J. Pei, B. Mortazavi-Asl, Q. Chen, U. Dayal, and M. Hsu, “ FreeSpan: frequent pattern-projected sequential pattern mining,” in Proc. 6th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2000, pp. 355–359.

[14] N. Hariri, B. Mobasher, and R. Burke, “Context-aware music recommendation based on latent topic sequential patterns,” in Proc. 6th ACM Conf. Recommender Syst., 2012, pp. 131–138.

[15] T. Hofmann, “Probabilistic latent semantic indexing,” in Proc. 22nd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 1999, pp. 50–57.

[16] L. Hong and B. D. Davison, “Empirical study of topic modeling in Twitter,” in Proc. 1st Workshop Soc. Media Anal., 2010, pp. 80–88.

[17] Z. Hu, H. Wang, J. Zhu, M. Li, Y. Qiao, and C. Deng, “Discovery of rare sequential topic patterns in document stream,” in Proc. SIAM Int. Conf. Data Mining, 2014, pp. 533–541.

[18] S Saravanan, V Venkatachalam ,“ Advance Map Reduce Task Scheduling algorithm using mobile cloud multimedia services architecture” IEEE Digital Explore,pp21-25,2014.

[19]S.Swathi “Preemptive Virtual Machine Scheduling Using CLOUDSIM Tool”, International Journal of Advances in Engineering, 2015, 1(3), 323 -327 ISSN: 2394-9260, pp:323-327.

[20] S Saravanan, V Venkatachalam, S Then Malligai "Optimization of SLA violation in cloud computing using artificial bee colony"2015, 1(3), 323 -327 ISSN: 2394-9260, pp:410-414.

[21]S. Saravanan, Vikram R ,"Improved Performance Analysis Image Segmentation Based on Cluster Image", Journal of Chemical and Pharmaceutical Sciences,issue 1,2017,pp92-95

[22]S. Saravanan, Vikram R ," Evolutionary Calculations on Gravitational Interactions Method of Global Leader Organize ", Journal of Chemical and Pharmaceutical Sciences,issue 1,2017,pp115-118