RESEARCH ARTICLE OPEN ACCESS

# User behaviors attributes of database anomaly detection model

SaifAldeen Salim Ahmed
*University : University of Technology in Iraq*
*College : Computer Science*
*department : Information System*

**ABSTRACT**
This paper includes the description of designing a data-base anomalydetection system, whichis capable of being more precisein depictingthe behaviors of individuals and improving data-base abnormaldetecting correctness. In designing the system, the Aprioriapproach is used first and depends on the k-meansclusteringand the Apriorimethods.It is capable of more efficiently exploitingusers'behaviors, and the data-base abnormalmore efficient detecting. The relevantstudiesshow that Apriorimethodaccording to time efficiency and precision of detectionis more optimal than thesoleutilizationaccording to association rulesmining approachesApriorimethod.
**Keywords**: Anomaly detection structure,Apriori,users'behaviorss.

---

Date of Submission: 02-11-2017 Date of acceptance: 08-12-2017

---

## I. INTRODUCTION

Due to the improvement of intruder detecting and data mining approaches, the data-base auditregistrythat's consideredto be a form of passive security element has been altered (1). It'sa significant wayof ensuring the security of data-base whichdata-base exception is efficientlyfound by users'behaviorsprofiles of the data-base mining (2). The abovementionedusers'behaviors profilesis the repetitive accessingtospecific resourceswhich includes the data-base, data-tables etc. This repetitive operating of resources is properties of the users'behaviorspatterns.

Nowadays, there existbasically3elements of deficiencyexist in the userbehaviors profiles mining. Initially, because of thedifferentkinds of audit registries in the DBMS,it's quitehardselectingwhich audit logsmaybe efficientlyutilizedfor mining the behaviors profile of the users.

Secondly, due to the fact that the existing algorithms are not capable ofproperly describingthe behaviors profile of the user, causing higher false positiverate of data-base anomalous detecting. Thirdly, even thoughpart of theexisting approachesare capable of achieving the mining the behaviorsprofile of users, when meaning the huge user data-base accessingregistries they arenoticeablyinefficient(3).This researchbasicallyperforms the next2mainexperimentsfor solving the abovementionedissues: First: A data-base structure ismodeled for anomaly detecting. Second: Aprioriapproachbased on this structure is suggested which is a moresufficient detecting of abnormal data-base approach.

## II. RELATED WORKS

There are numerousresearchesthat propose using of data mining approaches in registry file analyzing procedure or the detection of security risks generally. One of the 1stmethodsthat utilize data mining approachesinintruder detecting was suggested by Lee and Stolfo (8). Whoused2methods for detection, the association rules method and the redundant episodes approach. They illustrated that viathe analysis of audit data it's possible discovering intrusion patterns.

Frei and Rennhard (9) utilized anothermethodfor searching for anomaly in registry files. They generated the Histogram Matrix, a registry file visualizingapproachwhichaids security administrators find the anomalies. Thismethodoperates on each textual registry file. Itdepends on the ides that the brain is sufficient in the detection of patterns whileobserving images, thus, the registry fileis viewed in a waywhich it's easyobservingchanges from regularbehaviors.

Fu and others(10)suggested anapproach for anomaly detecting in unstructuredsystem registrieswhich doesn't need any application specific knowledge. In addition, theyaddedan approachfor the extraction ofregistry keys from free text messages.

Makanju, et.al(11) suggested a hybrid registryalarmdetectingmodel, with the use of each of anomaly and signature-based detectingapproaches.

## III. BACKGROUND ON ANOMALY DETECTION

An anomaly (or outlier) is an observation that looks inconsistent with the rest(majority) of the dataset, and hence arouses the suspicion that it can be generated bya different mechanism. The objective of anomaly detection is to mine unusual and information of interest from a large amount of data. Detecting anomaly is extensivelystudied in differentaspects,like the statistics, data mining, machine leaning and informationtheory, and its applications have been greatly expanded to multipleareaslike detecting of fraud, network intrusion, health monitoring, environmental monitoring andperformance analysis.

A straightforward solution for anomaly detection is to construct a pattern of normalobservations, and then one can use the pattern to identify anomalies. Whenapplyingthe pattern on test data, the observations whose properties follow the regular patternare labeled as normal, and those that deviate noticeably from the regular patternare labeled as anomalies. According to the availability of labeled training data, anomalydetectingapproachesmay be divided into three main classes, which are thesupervised, semi-supervised and unsupervised approaches.

In the first category all the training samples are required to be paired with a labelor desired output, i.e., normal or abnormal observations, for the characterization of allanomalies ornon-anomalies. Semi-supervised learning techniques make use of unlabeled records aswell as labeled ones for training. Typically semi-supervised techniques are trained witha large amount of unlabeled records and a small amount of labeled records. It shouldbe noted that pre-labeled data is not constantly available nor easily obtained in several of real-lifeutilizations, in addition, new kinds of observations (normal or abnormal) couldoccur that aren't included in the labeled training data. Unsupervised methods are oftenmore appealing for anomaly detection, since they require no labeled data, rather theyapply certain criteria to identify anomalous observations. An example type of such techniquesare distance-based approaches, e.g., classifying records according to the averagedistance between each one of the data records to its mapping the nearest neighbor observations.If the measured distance for a specific record significantly exceeds nearest neighbor distance of all objects then the data record is considered as an anomaly, otherwise it isconsidered to be normal.

## IV. DATABASE ANOMALOUS DETECTION MODEL

This paper includes the description of the design of a data-base anomaly detectingstructure,as illustrated in Figure(1).

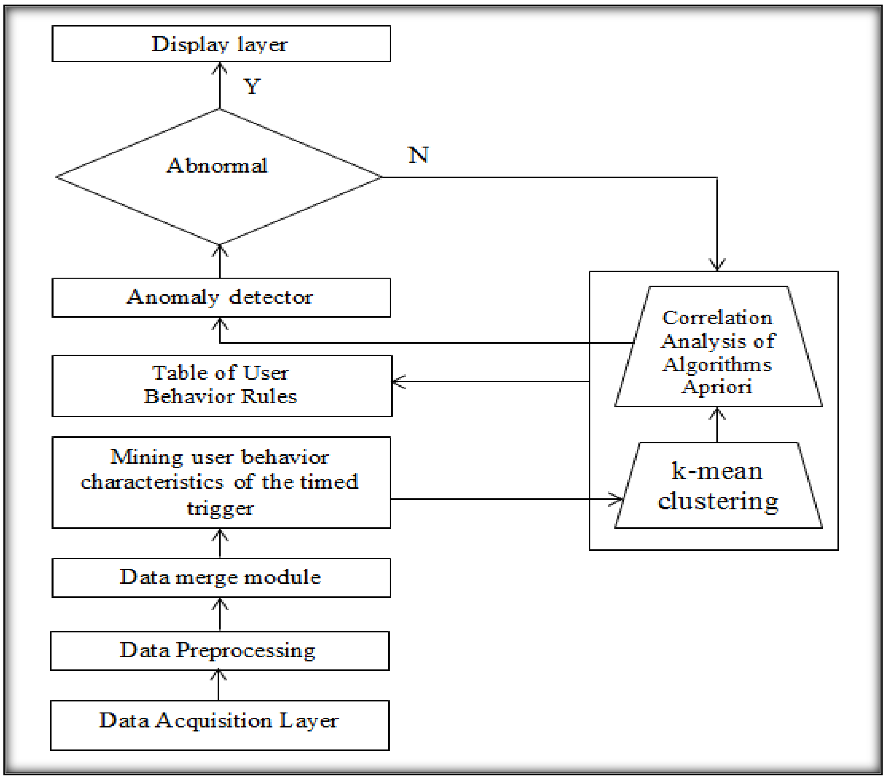



**Figure (1):** data-base anomaly detectingstructure

The structureis basically made up of the following 5modules, which are: data acquisition layer, data pre-processing,data merging, user behaviors mining, anomalydetecting.

**4-1 Data Acquisition Layer Module**

The basictask: benefitting fromregistry miner and otherdata-base registry collecting tools or data-base system's own registryanalyzing tools, likeOracle11g Profiler for completing theset of the existing testing data and the auditting of the trainingdata.

| | C01_PRICE | C01_HARD | C01_PREMIUM | C01_ID | C01_TREND | C01_SCREEN | C01_SPEED | C01_RAM | C01_MULTI | C01_CD | C01_ADS |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 506 | 33 | no | 506 | 139 | 4 | 1,775 | 170 | no | 14 | yes |
| 2 | 507 | 33 | no | 507 | 139 | 8 | 2,490 | 340 | no | 15 | yes |
| 3 | 508 | 33 | no | 508 | 139 | 16 | 3,599 | 340 | no | 17 | yes |
| 4 | 509 | 50 | no | 509 | 139 | 8 | 2,690 | 340 | no | 14 | yes |
| 5 | 510 | 66 | no | 510 | 139 | 8 | 3,195 | 540 | no | 15 | yes |
| 6 | 511 | 66 | no | 511 | 139 | 8 | 3,695 | 452 | no | 14 | yes |
| 7 | 512 | 50 | no | 512 | 139 | 8 | 2,645 | 250 | no | 15 | yes |
| 8 | 513 | 66 | no | 513 | 139 | 16 | 3,090 | 452 | no | 15 | yes |
| 9 | 514 | 66 | no | 514 | 139 | 2 | 1,890 | 107 | no | 15 | yes |
| 10 | 515 | 33 | no | 515 | 139 | 4 | 1,999 | 170 | no | 14 | yes |
| 11 | 516 | 50 | no | 516 | 139 | 8 | 2,935 | 250 | no | 17 | yes |
| 12 | 517 | 25 | no | 517 | 139 | 4 | 1,990 | 214 | no | 14 | yes |
| 13 | 518 | 50 | no | 518 | 139 | 4 | 2,290 | 214 | no | 14 | yes |
| 14 | 519 | 66 | no | 519 | 139 | 4 | 2,390 | 214 | no | 14 | yes |
| 15 | 520 | 50 | no | 520 | 139 | 4 | 2,025 | 170 | no | 14 | yes |
| 16 | 521 | 33 | no | 521 | 139 | 4 | 2,095 | 214 | no | 14 | yes |
| 17 | 522 | 33 | no | 522 | 139 | 8 | 2,590 | 340 | no | 14 | yes |
| 18 | 523 | 25 | no | 523 | 139 | 4 | 1,499 | 170 | no | 14 | yes |
| 19 | 524 | 33 | no | 524 | 139 | 8 | 2,325 | 250 | no | 15 | yes |
| 20 | 525 | 66 | no | 525 | 139 | 4 | 2,099 | 120 | no | 14 | yes |
| 21 | 526 | 66 | no | 526 | 139 | 16 | 2,999 | 245 | no | 15 | yes |
| 22 | 527 | 66 | no | 527 | 139 | 8 | 2,790 | 340 | no | 15 | yes |
| 23 | 528 | 33 | no | 528 | 139 | 2 | 1,590 | 107 | no | 14 | yes |
| 24 | 529 | 50 | no | 529 | 139 | 4 | 2,499 | 170 | no | 14 | yes |
| 25 | 530 | 50 | no | 530 | 139 | 8 | 2,575 | 250 | no | 15 | yes |
| 26 | 531 | 25 | no | 531 | 139 | 8 | 2,390 | 340 | no | 15 | yes |
| 27 | 532 | 33 | no | 532 | 139 | 2 | 1,495 | 107 | no | 14 | yes |

**Fig. (2):** the type of gatheredauditingregistry data detecting

**4-2 Data Pre-processing Module**

The basictask: post collecting data (audit trainingdata, the current data inspection), for removing thegenuine noise, like theinconsistent auditinginformation, noparticular operatingpersons, no operating iteminformation and othervaluabledata, in addition to stemming some information isn'tin association to the needed data from relevantcommunications links protocols when data-base serversareconnected viaclients.

**4-3 Data Merging Module**

The basic task: initially, to statistic the number offunctional data-base items, like the data table, data view, etc. used by this person. Then, dealing with thetaskitem and user via the numericidentifying, and after thatstoring in the established sessionsregistry tables. Moreover, every one of the connectedoperation recordsis considered a transaction T, each one of the transactions T generates the data-base D, which will be utilizedto mine user behaviorsproperties. Every one of the transactions belonged to the data-base Dis made up of the following fields: the sessionconnection ID, Data-base process user, process object, process path, etc. with an overall of 12 fields.

**4-4 User Behaviors Mining Module**

The basic task: Firstly, the procedure on the data-baserunvia the users and the their data-base itemsmustbe clustered with the use of the k-means approach, this wayfinishing the PreliminaryCharacterizing of thebehaviors of the user. In addition making the 1$^{st}$preparation for more mining of user behaviorsproperties.Secondly, improving the sufficiency of mining property rules concerning the user behaviorswith the use of theApriorialgorithm.

**4-5 Anomaly Detecting Module**

The basic task: comparing the training stage of the rulesgenerated by a regular user (obtained from old-rules table) withthe testingstage of mining association rules (obtained from new rules),in the case where the data-base exception happened then the irregulardetails have to be timely registered and shown.

**The following is the anomaly detecting algorithm:**

```
Input: The current audit data to be detected
Output: The information of detecting anomalies

Method:
Step1: acquire the begin time of detection: start time
Step2: While (each rule in old-rules) {
Step3: While (each rule in new-rules) {
Step4: If ((the antecedent of rules in old-rules
           == the antecedent of rules in new-rules)
       && (the consequent of rules in old-rules
          == the consequent of rules in new-rules)
       && ( the difference of support and confidence
            between the two rules greater than 35% ))
Step5: {make intrusion information record
        in the Intrusion table} }
Step6: the end time of detection: endtime
Step7: record the time spent testing(starttime-endtime)
Step8: If (interrupt) {
Step9: break }}
```

## V. THE RESEARCH OF THE MODEL OF MAIN ECHNOLOGIES

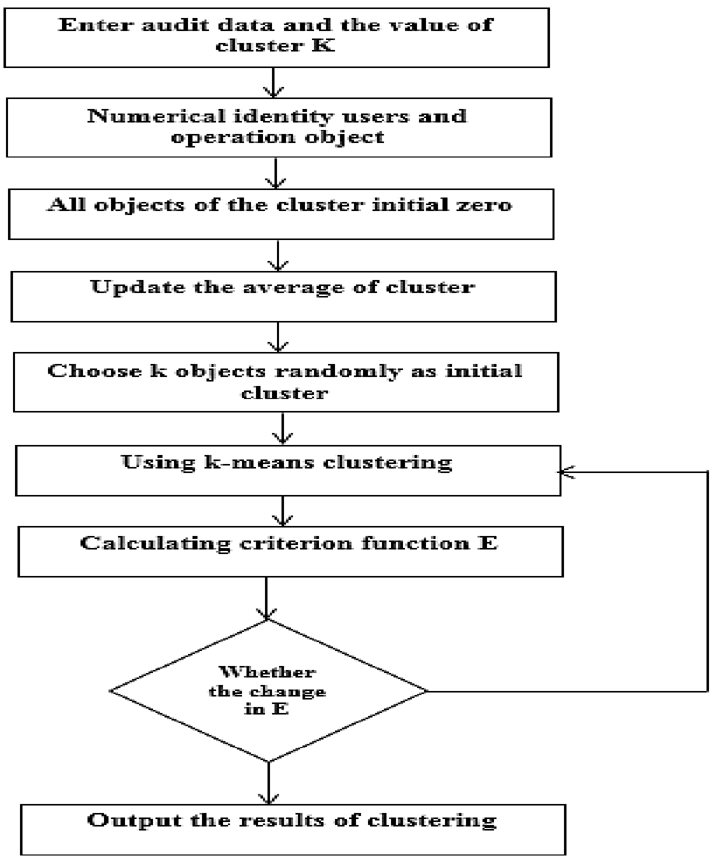



**Figure (4):** the procedure of clustering

**5-1 Preliminary Characterization of User BehaviorsProperties According to the K-means**

This research includesthe description of using the k-means clusteringmethodfor dealing with the data-baseusers and operatingobject data, for the sake of characterizing the users' initial behaviors profiles.

The maintask E of k-means method isdescribedinequation (1).

$$E=\sum_{i=1}^{k}\sum_{p\in Ci}|p-m_i|^2 \qquad (1)$$

The clustering procedure of data-baseitems is depicted in Fig. 4.

**5-2 Apriori association rule mining algorithm**

This approach has beensuggested for the Boolean associatingrules mining redundantelementgroup of the mainmethod, in the year of 1994by Agrawal et.al.(6). Even though theuseof Apriorimethod own nature mayraise theeffectiveness to a specificdegree.For the sake of mining the data-base registry data moreefficiently, this studypresents the Apriorimethod.

**1) Relevant theories and conclusions**

Theorem (1) : Supposingredundant k itemset is capable of generating(k +1) Itemset, then the number of itemset of the redundant k itemsetis definitely>= k.

Theorem (2): $\forall c \in C_k$, R (c) could be produced by the two items of R (x), R (y) (x $\neq$ y) in $R_{k}$ - 1, and R (c) = R (c) $\cap$ R (c).

**2) The main idea behind this Algorithm**

Data storing format which is a transaction identifier is corresponding to severaltypes is called the standardized data formatting in the data-base of transactions. While, data storing formatswhich is a transaction identifier is corresponding to severaltransactionidentifiersrelevant to the element is called the vertical data formatting.

The data which has standardized data formatting is transformed to vertical formattingviathe scan process of the data-base one time. In the same time,every one of the items in the itemset and transaction identifiers which corresponds to the elements are stored separate from one anotherwith the use of two dynamical liststorages. The supporting counting of *Ck* may bereached by Lk - 1 ∩L1 with no need to repeat data-base scanningfor the sake of obtaining the supporting counting of Ck.Connecting conditions of Apriorimethod areimprovedwith the use of the conclusions whichhave been proven in this study.

Redundant itemset k are obtainedvia connecting conditionsof Lk - 1∞L1, and the final item of Lk - 1 based on the index issubject to comparison withevery one of the items of L1 when Lk - 1∞L1, avoidingduplicating comparing of connecting conditionsof Lk - 1∞L1,and with no considerationif (k-1) is sub-set ofcandidate itemset Ck is in Lk - 1.

The repetitions of redundant item set k are improvedvia Theorem1, and the supporting counts of nominated item set k are handledwith Theorem2.

**3) The following is the description of the Apriori algorithm:**

```
Input:     D: Transaction database
           Min _ support : minimum supportcount
Output:    L: Frequent item sets in D

Method:

Step1: C1 = {The set of all items)
Step2: for each Candidate itemsets c ∈ C1{R(C) = ø ;}
Step3: for each transaction t ∈ D {
Step4: Ct = Sub-Set (t, C1);
         / / the table of generated each item ID in item set
Step5: for each c ∈ Ct{R(c) = R(c) ∪ t_TID ;}}
Step6: Count_Support (C1);
Step7: L1 = {c ∈ C1 | c_delete =0}; R1 = {R(c) | c ∈ L1};
Step8: Up_Sort (L1);
Step9: for (k=2; |Lk -1|>k-1; k++) {
Step10: Up_Sort (Lk -1); Ck = apriori_gen (Lk -1);}
Step11: L=∪kLk;
Step12: function Count_Support(C: candidate _ k - itemset)}
          for each Candidate itemsets c ∈ C; {
          if (c_support=R(c) < min_support) {
            c_delete=1; } }}
Step13: function apriori_gen(Lk -1:Frequent (k -1)  itemsets)
          { for each lk -1∈ Lk - 1{
           if (c_supportp=|R(c)|<min_suppor){ c_delete=1;}
           }
    }
```

## VI. EXPERIMENTAL RESULTS AND DISCUSSIONS

**Experimental Results:**

The precision of the Anomaly Detecting of the Model the comparing of detectingprecision between user behaviorss diggers pairedvia anomaly detectionsuggested in the model and the conventionalstructure of data-base anomaliesdetectingaccording to theApriori algorithm. In the experiments, the use oftrend data-base events for testing the detectingprecision. The outcome is illustrated in Figure (5).

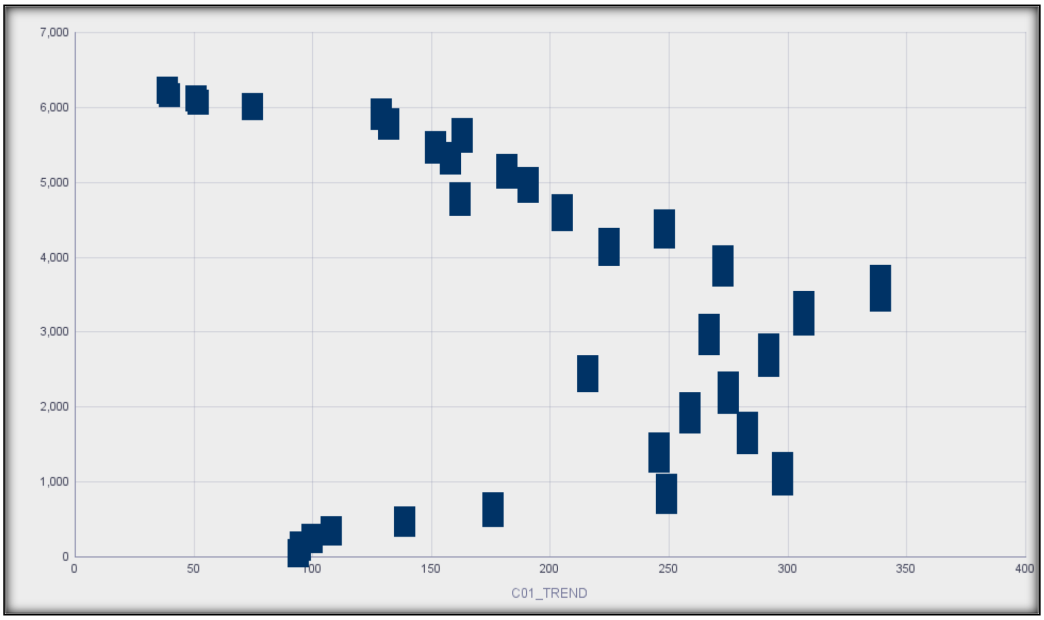



**Fig. (5):** Experimental Outputs
Detection Accuracy & Database events

The difference between the behaviors of two users (user1& user2) by using Apriori Algorithm as Shown in Figure (6), the explore Data behaviors of user shown in figure (7).

The results of the experimentationprovedthe fact that mining properties of usersbehaviorspost the auditingregistry clustering may be excavating user behaviors rules with moreefficiency, therefore, the precision of data-base anomaly detecting is also developedwith moreefficiency.

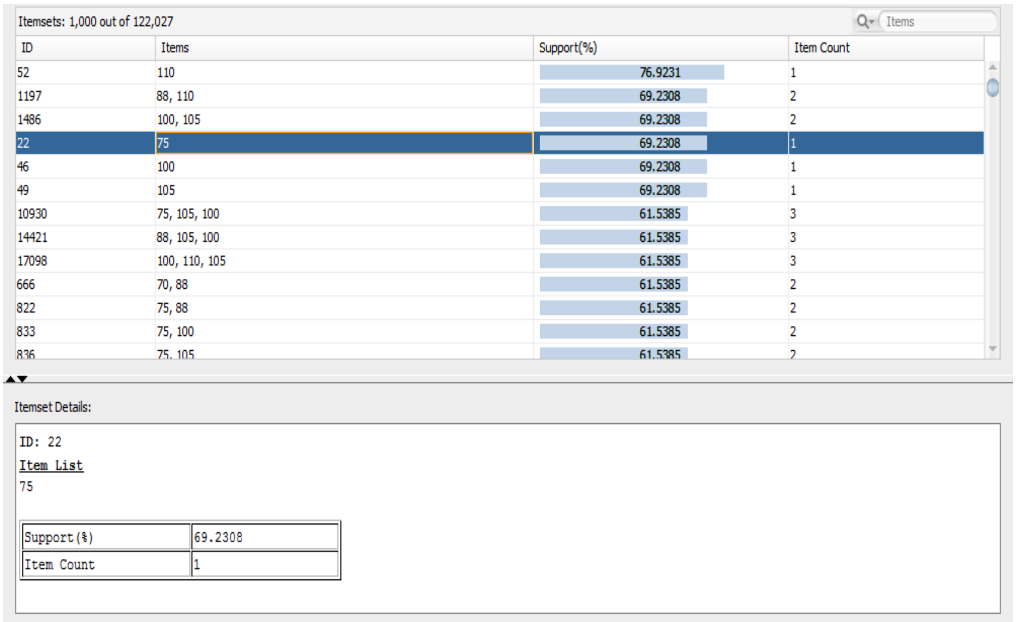
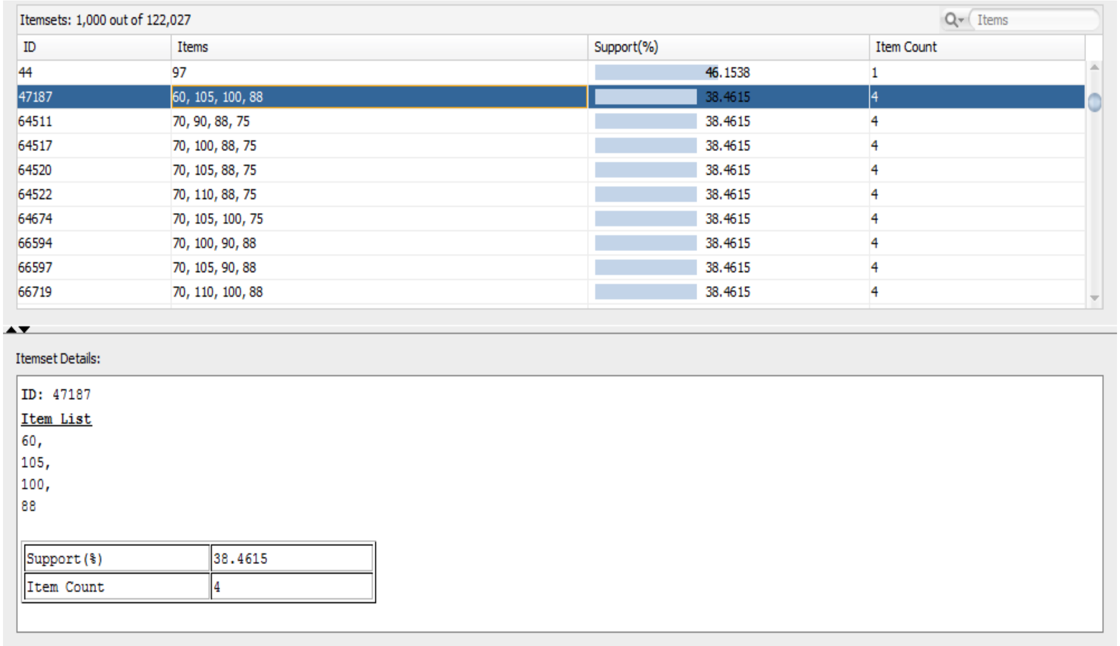



**Figure (6):**Aprioriuser1


Itemsets: 1,000 out of 122,027

| ID | Items | Support(%) | Item Count |
|---|---|---|---|
| 44 | 97 | 46.1538 | 1 |
| 47187 | 60, 105, 100, 88 | 38.4615 | 4 |
| 64511 | 70, 90, 88, 75 | 38.4615 | 4 |
| 64517 | 70, 100, 88, 75 | 38.4615 | 4 |
| 64520 | 70, 105, 88, 75 | 38.4615 | 4 |
| 64522 | 70, 110, 88, 75 | 38.4615 | 4 |
| 64674 | 70, 105, 100, 75 | 38.4615 | 4 |
| 66594 | 70, 100, 90, 88 | 38.4615 | 4 |
| 66597 | 70, 105, 90, 88 | 38.4615 | 4 |
| 66719 | 70, 110, 100, 88 | 38.4615 | 4 |

Itemset Details:

ID: 47187
Item List
60,
105,
100,
88

| | |
|---|---|
| Support(%) | 38.4615 |
| Item Count | 4 |

**Figure (6):** Apriori user2

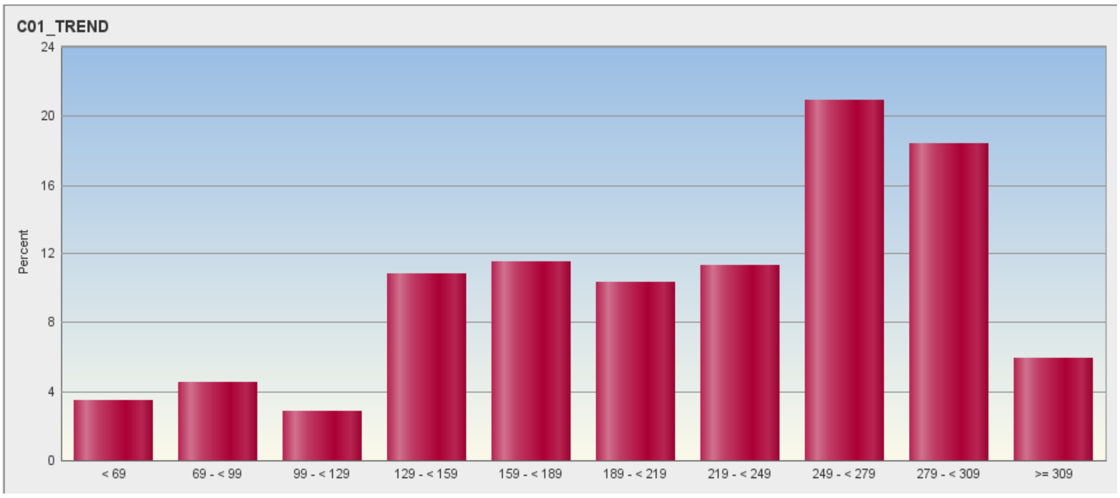



**Figure (7):** Explore Data

## VII. CONCLUSION

This studyexplains a structure of data-base anomalydetectingfor the sake ofefficiently improving the precision ofdata-base anomaly detecting. Based on this structure, a sufficient Apriorimethod is suggested.This algorithm may be able toget rid of someinsignificant rules to a specificdegree and depicting data-base users'behaviors profiles more clearly. At last, sequence patterns mining used in userthat access the data-base registries will be asignificant research topic in the future.

SaifAldeen Salim Ahmed "User behaviors attributes of database anomaly detection model." International Journal of Engineering Research and Applications (IJERA) , vol. 7, no. 12, 2017, pp. 29-35.