Comparative Study on Frequent Pattern Mining Algorithms for Temporal Data Set and its Applications

Mrs. Sona Shaju K
M. Tech Student
Department of Computer Science & Engineering
Thejus Engineering College, Thrissur, India

Abstract— Mining frequent pattern from temporal data set, which contain time related information associated with each transaction is not effectively done by traditional data mining techniques. Because, they mainly focus only on static data. Some patterns are valid only on some particular time points or intervals, so these techniques are not capable of solving over estimating problem of time periods. Frequent pattern of item set in the temporal database is formed, mainly in transactional database like purchasing an item, occurrence of an event etc. Mining time related data is a challenging issue. Different methods for finding frequent pattern on temporal database are studied in this paper and their characteristics like, memory space utilization, computational time, scanning of database, temporal information considered, etc are compared. The three algorithms are, first is Periodic Frequent Pattern Growth algorithm[3] which is an extension of FP-Growth algorithm, give more consideration for periodicity and extract the periodic frequent patterns from PFP-tree, second is Extended a-priori algorithm[1] which is the extension of A-priori algorithm gives more importance for each and every time point and generate frequent pattern on the basis of Time Cube and Basic time cubes, third is the Cluster Based Bit Vector Mining Algorithm[2] which generates the frequent patterns after compressing the database by converting the items in transactions to bit vectors on the basis of occurrence of each item in the corresponding transaction.

Key words: Data mining, Frequent pattern mining, Temporal Data set, Periodic Frequent Pattern, Time Cube, Basic Time Cube, Time Stamp information, Bit Conversion

I. INTRODUCTION

Data mining is the process of discovering meaningful and useful information, rules or patterns from large amount of history data called data set. This discovered new information is helpful for identifying new trends in business, forecasting, information analysis and find new methods to increase productivity and profit. Various traditional data mining techniques are used to mine information from dataset.

Dataset is the base of data mining process which is the collection of data with its features. Data for data mining available in many forms like, computer files, business information in SQL or other standard database format, transactional data set, information recorded automatically by devices, binary data set, information recorded by satellites etc. This data set can be static or dynamic (changes with time). Some data set are highly associated with time factor. Frequent patterns are those which occur frequently in a data set, mainly in transactional database. Finding frequent patterns plays an essential role in mining associations, and many other interesting relationships among data. One of the difficulties faced in frequent pattern mining is, sometimes it causes over fitting problem or generate insignificant patterns. Another problem is, it can't effectively consider the time factor. Traditional data mining systems are not capable to handle time-varying property of the real world datasets. Some patterns occur in some time point or time interval. Here is a analysis study on three different types of methods which are the extensions of the traditional data mining techniques and can be used on temporal data set to find the frequent patterns. One method is Periodic –Frequent Pattern Growth (PFP-Growth)[3] which is the extension of FP-Growth which will find the periodic frequent pattern by creating PF-list and PF-tree and further tree pruning. Second method is Cluster Based Bit Vector Mining Algorithm (CBVAR) [2] which will find the frequent item set by creating cluster of transaction after converting the items of transactions into bits. Third method is the extension of a-priori algorithm [1] which will find the frequent pattern by dividing the available time interval (Time Cube-TC) into small basic time cubes (BTC).

II. RELATED WORK

Data mining is the process of discovering and analyzing data from different fields using supervised and unsupervised techniques [5]. Temporal data mining, is performed over the temporal data sets (which give importance to time attributes) to discovers knowledge, N. Pughazendi and Dr.M.Punithavalli in their paper "Temporal Databases and Frequent Pattern mining Techniques "[5] presents temporal data mining and focus on pattern discovery using temporal association rules.

Different techniques are suggested to overcome this above mentioned problem and future works are progressing on further enhancement of these methods. In the paper "Mining Hierarchical Temporal Association rules in a publication database"[4] proposes the concept of a hierarchy of time granules, which is named as hierarchical temporal association rules. Association rules use the concept of minimum support and minimum confidence. Support indicates how frequently the item set or item appears in the dataset. Confidence indicates how often the rule has been found to be true. If a TRANSACTION consists of items X and Y,

$$\text{Confidence}(X \Rightarrow Y) = \text{support}(X \cup Y) / \text{support}(X)$$

They put forward the concept of publication time of a particular item. Publication time means the time which the item is valid. Then the lifespan of an item in a time granule is calculated from the publication time to the end time in the time granule (total time). A three-phase mining framework is proposed. In phase 1 temporal frequent item sets are identified. Phase 2 will identify all hierarchical temporal frequent item sets and phase 3 will identify all hierarchical temporal association rules.

Y. Xiao, R. Zhang and I. Kaku proposed a new type of association rule, i.e., association rule with time windows [6] in their work "A new framework of mining association rules with time windows on real-time transaction databases". They give more importance to the time at which the transaction occurs. Time windows of transactions will cover the particular pattern occurrence cycle. This time window is considered here. Because many candidate association rules do not satisfy the minimum support(minsup) and minimum confidence(mincof) when the full-time span of transaction is considered. But it is satisfied when it is viewed within constrained time-windows. The found the time intervals for association rules, part-time association rule, which is random in length and not specified by user. An additional threshold minwin is also used to avoid the situation that each item/itemset appeared in bounded transactions is regarded as frequent. It should be much greater than 1/minsup. Minwin is defined by the user. The ts, te (starting time and ending time) should be greater than or equal to predefined minwin.

Some patterns exist in between some time intervals. The relationship between two intervals is very complex. Soeffective and efficient mine the interval-based sequences is a challenging issue. In paper "Mining temporal pattern in time interval based data"[8],by Yi-Cheng Chen, Wen-Chih Peng and Suh-Yin Lee proposes two novel representations for processing complex relationships among time intervals, endpoint representation and end time representation. They defined: temporal pattern, occurrence based probabilistic temporal pattern and duration based probabilistic temporal pattern. They developed two novel algorithms, Temporal Pattern Miner (TPMiner) and Probabilistic Temporal Pattern Miner (P-TPMiner), to discover these three types of interval-based sequential patterns. They also propose three pruning techniques to further reduce the search space of the mining process.

B.Saleh and F.Masseglia in their work "Discovering frequent behaviors: Time is an essential element of the context"[7] introduce solid itemsets, which indicate logical and compact behaviors over specific time periods. They propose an algorithm, SIM for information extraction. The user's intention will differ from one period to another. E.g interest over different seasons. The proposed algorithm is based on Temporal itemset and solid itemsets. Temporal item set consists of(xi, xp, xσ) item(xi),period(xp) of that item and threshold of that item(xσ). This algorithm will introduce a new way for the counting step of the generated candidates (items to be combined to form frequent 2 item set).During the counting step, "kernels" of the candidate temporal itemsets over their period of occurrence is generated. Then kernels are merged in to find the corresponding solid itemsets.

Usha, and Dr.K.Rameshkumar, in their work, " A Complete Survey on application of Frequent Pattern Mining and Association Rule Mining on Crime Pattern Mining",[9] describes the application of pattern mining. The new methods for mining frequent patterns in different areas are discussed. The application areas include Network forensic analysis, Banking sector, Educational data, Animal behavior etc

III. METHODOLOGIES

A. Periodic-Frequent-Pattern Growth Algorithm (PFP-Growth)

This is an association rule based algorithm[3]. In association rule based algorithm, two main things are there,

- Support: Percentage of transactions which contain a particular itemset.
- Frequent itemset: Itemset whose occurrences is above a threshold.

Before describing PFP-Growth, FP-Growth should be explained.

1) Frequent Pattern Growth Algorithm (FP-Growth)

It is a scalable technique for mining frequent patterns from a database by forming FP-tree and considering Conditional pattern base and conditional path-tree.A minimum support count (minsup)for each item will be predefined by the user.FP-Growth finds the frequent item set by two step approach.

1) Step 1: Build a data structure called the FP-tree using two pass over the data set.

– Pass 1:

1) Find minimum support
2) Discard infrequent items.
3) Sort items in the given transaction in decreasing order of their minimum support count.

– Pass 2:

1) Reads one transaction at a time and add items to the path with a counter value.
2) When there is a path overlap, increment the counter value.
3) Pointers are maintained between nodes containing the same item.
4) Create the FP- tree

2) Step 2: Extract frequent item set directly from FP-tree by using,

1) Bottom-up algorithm is used from the leaves towards the root
2) Divide and conquer method is used by considering last item in the tree is and all its branches (prefix-tree) by its counter value.
3) Conditional pattern base of each item is taken and the minsup of those items are checked. If it satisfies, then includes in the frequent item set.

Example Data set used for the evaluation of methods,[3]

TID	TimeStamp	Itemset	Rearranged Itemset
1	25/07/2017 09:08:23	ABC	BAC
2	25/07/2017 09:14:09	ABCD	BACD
3	25/07/2017 09:22:	AB	BA
4	25/07/2017 09:24:12	BCD	BCD
5	25/07/2017 09:30:15	AB	BA
6	25/07/2017 09:33:58	A BD	BAD
7	25/07/2017 09:42:02	AB	BA
8	25/07/2017 09:43:56	ACD	ACD
9	25/07/2017 09:50:01	AB	BA
10	25/07/2017 09:57:12	ABC	BAC
11	25/07/2017 10:28:23	C	C
12	25/07/2017 10:58:30	BD	BD

Table 1: Example dataset

2) Calculations:

Minsup (fixed by user)*= 3*

Support of A is 9, B=10,C=6, D=5.

All items satisfy the minimum support count and each transaction is rearranged.

Final FP-tree will be,

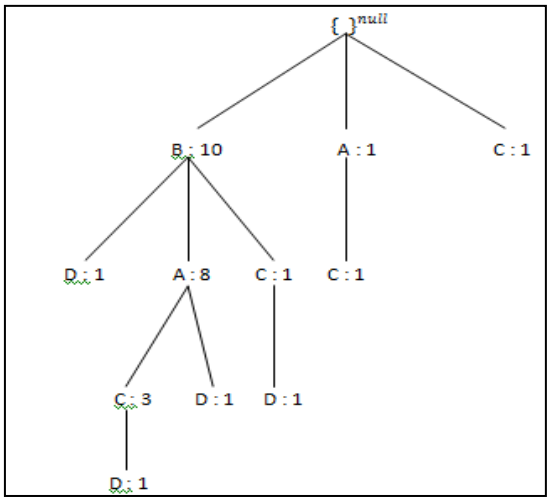

Fig. 1: Final FP-tree

Conditional path base of 'D' are B:1, BAC:1, BA:1, BC:1, AC:1.

In this B,A and C satisfies the *Minsup*.B occurs four times, A occurs three times, C occurs three times. So, the frequent item set consists of,

Frequent item set according to D is: {D, B, BAC, BA, BC, AC}.

It is repeated for every item and final frequent item set is,

{A, B, C, D, BA, BC, AC, BAC}

B. PFP-Growth (Periodic-Frequent-pattern) Algorithm

PFP-Growth algorithm is the extension of FP-Growth algorithm. In this algorithm time interval at which the items occur is strictly considered. Frequent patterns occurs periodically and mostly at specific intervals of time and they are referred as periodic frequent pattern. It should satisfy the antimonocity property[3].

Anti-Monotonic property

For a pattern X, if Sup(X) ≥ minSup and Per(X) ≤ maxPer, then ∀Y⊂X and Y≠□, Sup(Y) ≥ minSup and Per(X) ≤ maxPer.

Minsup and Maxper(maximum periodicity) is fixed by user.

1) Step 1: Each transaction is scanned and corresponding ts-list is created by calculating periodicity, support and updating the timestamp of each item.

Ts-list consists of 'i'-itemname, 'f'-support, 'p'-periodicity, 'ts'-timestamp.

Periodicity is calculated by equation,

$ts_{cur} - ts_1^j$;(current timestamp-initial timestamp)

Initial current stamp is fixed as 0.

The items which doesnot satisfy Minsup and Maxperis eliminated.

The rearranged final ts-list for the above example after eliminating the item which does not satisfying antimonocity property:

Item(i)	Support(f)	Periodicity(p)	Time stamp(ts)
B	10	2	12
A	9	1	10
C	6	1	11

Table 2: Final ts-list

2) Step 2: PFP-tree is created on the basis of *ts-list* The final PFP-tree is

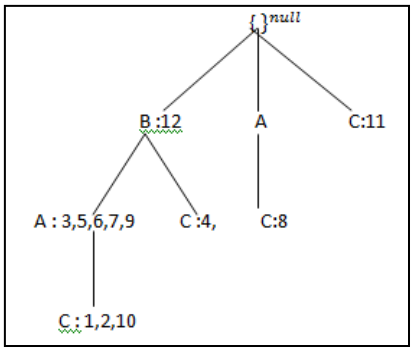

Fig. 2: Final PFP-tree

Instead of counter value, the time stamp information at which the corresponding item is appearing as the last item is associated with it.

Then the frequent pattern is extracted from the tree. The final frequent item set is:{A, B, C, AB}

1) Extended A-priori Algorithm

Before introducing extended a-priori method, a-priori should be discussed.

C. Priori Algorithm

The Apriori algorithm is based on 'bottom up 'approach where frequent subsets are extended one item at a time. It says that the subset of any frequent item set is frequent. .This algorithm is contrapositive. Thatmeans:

If an itemset is not large, then its supersets are also not large.This is an association rule based algorithm.

Association rule definitions are:

'X' and 'Y' are the items in the itemset 'I'

1) Association Rule (AR): implication X ⇒ Y where X,Y⊆ I and X ∩ Y
2) Support of AR(s)X⇒ Y: Percentage of transactions that contain X ∪Y
3) Confidence of AR (a) X ⇒ Y: Ratio of number of transactions that contain X ∪ Y to the number that contain X.

Suppose X, Y are the items in a transaction, T is the set of transactions. Support(X)=

$$\frac{\text{NumberoftimestheXoccuredinthewholetransaction}}{\text{Totalnumberoftransaction}}$$

1) Step 1: Calculate the support of item
2) Step 2: Check it with the minimum support defined by the user. If it satisfies, that particular item is added in to the candidate frequent-1 item set. At the same time it is added in to the frequent large item set too.
3) Step 3: Two items from Candidate frequent 1-item set is combined form the candidate frequent-2 item set and support is checked. If it satisfies, that item combination is added in to the frequent large item set.
4) Step 4: Repeat the above step until it covers all combinations or item set become null.

1) Calculation:
Minsup of each item is calculated and got it as
A= .75, B= .83 , C= .5 , D= .41
Frequent-1 item set is {A,B,C,D} because all items satisfy the Minsup. The items in candidate frequent-1 item set is combined to form candidate frequent-2 item set and its Minsup is calculated.
AB= .6, AC= .3, AD= .25, BC= .3, BD= .3
The combinations AB, AC, BC and BD satisfies the Minsup condition.Thus they will get added into the frequent large item set. Now the frequent item set is,
{A, B,C,D,AB,AC,BC,BD}.
By using the available items in large frequent item set,candidate frequent -3 item set is developed and its Minsup is calculated
ABC= .25, ABD= .15, ACD= .16, CBD= .16
No combinations satisfies the Minsup, so, nothing is added into the frequent large item set.
The final frequent item set is :
{ A, B, C, D, AB, AC, BC, BD}

D. Extended A-Priori Algorithm by Basic Time Cubes (BTC)

Some patterns occurs in either all or some of the intervals. Because of this varying time intervals, a new algorithm called frequent itemset mining with time cubes is proposed to restrict these time intervals, Different patterns occurring in different time intervals are considered. The time interval at which the whole transaction occurred is considered as the Time Cube(TC) [1].Each time points in this TC is segmented equally and called as Basic Time Cube(BTC).Each BTC consists of a specific number of transactions.

A new factor, called Density is calculated in each BTC. It can be predefined as

$A = \frac{N}{N_{BTC}}$, where 'A' is the average transaction per BTC. 'N' is the total number of transactions, 'NBTC' is the total number of BTCs.

$$\text{Density} = \alpha \times A$$

1) Step 1: Consider the available time as Time interval or Time cube.
2) Step 2: Calculate support of each item and check whether it is greater than or equal to the predefined Minsup. If not, that item is eliminated.
3) Step 3: Database is partitioned to a specific number of Basic time cubes, which is user defined.
4) Step 4: The support and density of each item in each BTC is checked. 'X' is an item, and if sup(X) ≥ Minsupand TR^{BTC} (number of times that particular item occurred in that BTC) ≥mindensity, thenthat particular item is added in to the frequent item set and that particular BTC is added to the BTC set. If that item does not satisfies any one of this conditions, the item is ignored and the corresponding BTC is not added in to the BTC set and no need of further checking of that specific item at that BTC.
5) Step 5: From the developed frequent item set, two items are selected randomly as candidates and joined by using joint operator. Then, above step is repeated.

1) Calculations:
The TC is 9 to 11.The *minsup* of each item is calculated.
A= .75, B = .83, C = .5, D = .41
The database is partitioned to 4 BTCs where each BTC consists of three transaction. Density and support of each item in each BTC is evaluated. Eg:
Minsup is predefined as .3 and α =.5, then density will be 1.5
Item A in
Ist BTC - (.75 ≥ .3) &(3>1.5)–TRUE
IIndBTC - (.75 ≥ .3) & (2>1.5)- TRUE
IIIrd BTC - (.75 ≥ .3) & (3>1.5)- TRUE
IVthBTC -(.75 ≥ .3) & (1>1.5) - FALSE.

A added in to the frequent item set. Its BTCs exclude Fourth BTC is added into A's BTC set, because it does not satisfies condition at fourth BTC. No further checking of A is done on fourth BTC. This is repeated for each item and combinations of items.
Thus the final set of frequent item set is :
{A, B, C, D, AB}

1) Cluster Based Bit Vector Mining Algorithm (CBVAR)

CBVAR builds the clustering table as a two dimensional array where the columns represent items and the rows represent Transaction ID's (TID)by scanning the database only once.. The table consists of bits (0 or 1) to indicate the presence or absence of an item. Items are converted into bit vectors on the basis of occurrence of each item in each transaction. 1 indicates the presence of an item and 0 indicates the absence of an item. The Minimum threshold is defined by the user as Min-threshold and it is checked with the support threshold of each item. [1]
support threshold= (Number of one's * Total number of items)

Cluster table for items in the above example of Table 1 is given below:

TID	A	B	C	D
1	1	1	1	0
2	1	1	1	1
3	1	1	0	0
4	0	1	1	1
5	1	1	0	0
6	1	1	0	1
7	1	1	0	0
8	1	0	1	1
9	1	1	0	0
10	1	1	1	0
11	0	0	1	0
12	0	1	0	1

Table 3: Cluster table of Items in transaction

2) Calculations:
Minimum threshold is fixed as 30%.
Support threshold for each item is calculated:
A => 9*4 =36%, Likewise B=40%, C=24%, D=20%
Only A and B satisfy the minimum threshold. So, other items are eliminated and A and B are entered into the frequent 2 item set.
Then, frequent 2-item set is formed by combining these items.

TID	AB
1	1
2	1
3	1
4	0
5	1
6	1
7	1
8	0
9	1
10	1
11	0
12	0

Table 4: Frequent 2-item set

Minimum support threshold of AB=8*4 = 32%.
The frequent item set is {A, B, AB}

IV. ANALYSIS

The three methods explained here is analyzed on the basis of different parameters and compared. First, let's see the difference of frequent pattern set formed by the three different algorithms.

PFP-Growth algorithm = {A, B, C, AB}
Extended A-Priori algorithm={A, B, C, D, AB}
CBVAR = {A, B, AB

The frequent patterns formed the previous version algorithms,

FP-Growth algorithm= {A, B, C,D, BA,BC,AC,BAC}
A-priori algorithm ={ A,B,C,D,AB,AC,BC, BD}
Among these the most common ones are : A, B, C, AB, AC

The number of frequent patterns occurring in frequent item set will differ according to the selection of minimum support, threshold, count, Maximum periodicity, α factor to find the density of a item etc. Different parameters are considered to analyze the three algorithms

Slno	Parameter	PFP	Extended A-Priori	CBVAR
1.	Technique	Ts-list and PFP tree is constructed after calculating the support count and periodicity	The given time interval is divided into equal segments called BTC and Support & density of each item in each BTC is calculated.	Cluster table is created by converting to bit.
2.	Memory used	Due to compact structure and no candidate generation, memory use is less	Due to large number of generation of candidates, the memory utilization is high	Memory ulitilized is less because of bit conversion and no candidate generation.
3.	Execution time	Less execution time	More execution time. Time is wasted by the creation of candidate items	Less execution time
4.	Candidate generation	No candidate generation	A large number of candidate items are generated.	No candidate generation
5.	Involvement of time factor	Periodicity(time interval is considered)	Time points are strictly considered	Time stamp information is used (Eg: transaction ID)
6.	Scanning of database	Two scan of database	Several scanning of database	Single scan of database

V. APPLICATIONS

The application areas of frequent pattern mining are[9]:

A. Crime Pattern Function

Frequent pattern mining is used by the crime analysts to analyze and identify the current trends and patterns in crimes for predicting future occurrences of crime, the place where the crime might occur, time, day etc.This is helpful for law enforcement officials to prevent crimes.

B. Crime Prevention Theory

It explains why crimes are committed in particular areas. According to this theory, crime is either planned or opportunistic and crime happens when the activity space (places in everyday life, like home, work place, school location, shopping areas, entertainment areas etc.) of both victim's and offender's intersects.

C. Legal Field

Association rule based algorithms, Clustering algorithm and Fuzzy logic are used to find the frequent patterns in crime data and also to find the outliers. Some tools used in this field are:

1) CCIS (Crime and Criminal Information System)

Identifies crime hot spots and crime zones.

2) Online Graphical Information System (GIS):

It includes GIS for robbery, homicide and physical injury incidents within a city. This system can help law enforcement officials to identify where and what time crime frequently happen.

3) Series Finder:

This is used to figure out which crimes are committed by the same individual or groups.

D. Network Forensic Analysis

Network forensic is the area to ensure security for the network against intrusion methods. It use the Apriori to build and update signature database of offense, and tries to improve the efficiency of crime detection.

E. Animal Behavior Data

RFID(Radio-Frequency-Identification) transponders fixed into the study animal's body. Then, principal component analysis and frequent pattern mining are used to analyse the resulting data.

F. Educational Data

Apriori algorithm is performed on student log data to bring out the interesting rules. The rule helps the teachers to understand the knowledge and performance of the students in academics which will help to understand the interest of the students in the course.

G. Banking Sector

This will analyse the user's transactions and it discovers common behavior patterns by any efficient data mining technologies. Then anomalous transactions are discovered by comparing new transactions with the user's common behavior patterns.

VI. CONCLUSION

The different methods for finding frequent itemset or pattern are introduced. The periodic frequent pattern growth method reduces the computational cost and they are trying to further reduce the cost and extend their pruning techniques to mine partial periodic frequent patterns in database. The cluster based bit vector algorithm is are trying to further reduce the time and memory space in the future to handle large data. The extended version of a-priori algorithm is trying to optimize the BTCs in future to obtain better result. They are also trying to enhance performance of the algorithm by using multithreaded processors like graphic processing units to speed up the computations. It is very difficult to say which is the best because, it depends on situation. If the time points should be strictly considered, then Extended A-priori algorithm is suitable. If periodicity of the item should be considered then, PFP-Growth can be used. Because of the performance and accuracy issue due to user selection of thresholds such as support and confidence, methods such as meta-heuristic algorithms which have no need to determine these thresholds in advance are worth trying.

REFERENCES

[1] MazaherGhorbani and MasoudAbessi," A New Methodology for Mining Frequent Itemsets on Temporal Data", IEEE Transaction on Engineering Management,January 2017

[2] M. Krishnamurthya , A. Kannan, R. Baskaran , M. Kavitha, "Cluster based bit vector mining algorithm for finding frequent itemsets in temporal databases", Published by Elsevier Ltd,page no-513–523, 2011

[3] R.UdayKiran, Masaru Kitsuregawaa, P.KrishnaReddy, "Efficient discovery of periodic-frequent patterns in very large databases", The Journal of Systems and Software 112, page no 110–121, 2016

[4] Guo-Cheng Lan, Tzung-Pei Hong, Pei-Shan Wu, ShusakuTsumoto," Mining Hierarchical Temporal Association Rules in a Publication Database", 12th IEEE IntConI. on Cognitive Inlormatics& Cognitive Computing,page no 503-508,2013

[5] N.Pughazendi, Dr.M. Punithavalli, "Temporal Databases and Frequent Pattern Mining Techniques",International Journal of p2p Network Trends and Technology, July to Aug Issue 2011,page no 13-17.

[6] Y. Xiao, R. Zhang, and I. Kaku, "A new framework of mining association rules with time windows on real-time transaction database," Int. J. Innov.Comput., Inf. Control, vol. 7, no. 6, pp. 3239–3253, 2011.

[7] B. Saleh and F. Masseglia, "Discovering frequent behaviors: Time is an essential element of the context," Knowl. Inf. Syst., vol. 28, no. 2, pp. 311–331, 2011.

[8] Yi-Cheng Chen, Wen-Chih Peng and Suh-Yin Lee,"Mining temporal pattern in time interval based data", IEEE Transaction on knowledge and data enginerring, 2015

[9] D.Usha, Dr.K.Rameshkumar," A Complete Survey on application of Frequent Pattern Mining and Association Rule Mining on Crime Pattern Mining", al., International Journal of Advances in Computer Science and Technology, 3(4), April 2014, 264 – 27, Volume 3, No.4, April 2014

