International Journal of Technical Innovation in Modern Engineering & Science (IJTIMES)
Impact Factor: 5.22 (SJIF-2017), e-ISSN: 2455-2585
Volume 4, Issue 08, August-2018

SEQUENTIAL PATTERN MINING ON MULTIDIMENSIONAL MEDICAL STORE DATA CONSIDERING GENERAL CONSTRAINTS

Manika Tomar[1] ,Dr.Vishal Dahiya[2],Dr. Devarshi Mehta[3]

[1]*Computer Science, Indus University,*
[2]*Computer Science, Indus University,*
[3]*Computer Science, GLSICT.*

Abstract— Lots of research and enhancement are being done in area of Sequential Pattern Mining. Previous sequential pattern mining algorithms either mines multidimensional sequences, multidimensional sequences with time-interval or sequences with constraints.
In this paper, we propose an algorithm for mining multidimensional sequences considering constraints. These sequences contain time-interval between the occurrences of events (sales of item). Application of this algorithm generates frequent sequences from multidimensional data set. Due to constraints, useful, relevant and less numbers of sequences are generated. The time-interval specified between does not only state the order of occurrence of events but also time-interval between occurrences. Thus generation of such kind of useful and less number of sequences makes analysis and decision making easier.
In this paper, we have applied the algorithm on data set from medical store. The experiment shows that the number of sequences generated after application of proposed algorithm is tremendously reduced as compared to those numbers of sequences which are generated using traditional algorithm.

Keywords— Sequential Pattern Mining, Time- Interval, Medicines, Constraints, Dimensions.

I. INTRODUCTION

The purpose of proposed algorithm is to mine multidimensional sequences considering constraints. These sequences contain time-interval between the occurrences of events (sales of item). The proposed algorithm is applied on data set of medical store. The sequences consist of medicines sold and time-interval between medicines sold. The dimensions considered in data set are company name to which medicine belongs and seasons in which medicines are sold. Various constraints are applied in this algorithm.

Various constraints applied in this algorithm are Length constraint, time-interval constraint, same medicine check constraint; the Length constraint restricts generation of sequences which is too short. Time-interval constraint generates those sequences which contains medicines sold in consecutive month. This can help the owner to analyze which are medicines sold in each month and thus he can predict the sales of medicine for next month and can update stock accordingly. The same medicine check constraints allows to generate those sequences which has all the medicine same in sequences. Such sequences can help owner to identify medicines sold on regular basis and approximately at which time-interval those medicines are sold.

The algorithm generates sequences with constraints along with information like, in which particular season medicines were sold and belongs to which company.
After application of algorithm on dataset, it can be predicted that which could be the next medicine to be sold considering dimensions season and company of medicine.

Section 1 contains the introduction of proposed Sequential Pattern Mining Algorithm for time-interval data considering Dimensions and Constraints. Section 2 contains work done related to sequential pattern mining algorithm. Section 3 contains parameters, constraints used in algorithm and pre-processing of data on which extended algorithm is to be applied. Section 4 contains experimental results. Section 5 contains Conclusion and Section 6 contains Future Work.

II. SEQUENTIAL PATTERN MINING AND RELATED WORK

Sequential Pattern Mining identifies those sequences of items from repository of database, which satisfies specified minimum support. Various sequential pattern mining algorithms like Apriori All, GSP [1], SPADE [3] were proposed but these algorithms use to generate large number of candidate sequences due to which execution time was tremendously increased. PrefixSpan mines the complete set of patterns and greatly reduces the efforts of candidate subsequence generation. It out performs the Apriori-based GSP algorithm and FreeSpan, in mining large sequence databases [4].

All these above algorithms consisted of sequences which do have their super-sequences, thus result set generated consists of redundant sequences. CloSpan[6] was proposed to overcome this problem. This proposed method is expected to mine only those sub-sequences which do not have their superset sequences under the same mining criteria.

For mining sequences more effectively, concept of mining sequences considering dimensions was proposed. This concept enables to give more related and effective results. For e.g. mining frequent sequences of sales of item considering dimension seasons like Summer, Winter, Rainy i.e. it shows frequent sales of item in particular season. Various work has been done for Mining sequences from multidimensional database [5][12].

To restrict the number of sequences generated concept of incorporating constraint along with sequential pattern mining was proposed [11][13][14].

Effectiveness and efficiency are two ways to measure performance of any algorithms. An algorithm is effective if it generates those sequences which are useful and it is efficient if it generates useful information in less execution time. Generating frequent sequences of sales of items may help owner to predict the items to be next sold. This information may become more useful when dimensions and constraints are incorporated with sequences. The algorithm we have proposed will mine sequences considering constraints and dimensions which will reduce number of sequences generated and will effectively generate only useful information. The sequential pattern mining algorithm we have used uses pattern growth concept thus the execution time is also reduced.

III. SEQUENTIAL PATTERN MINING ALGORITHM FOR TIME EXTENDED DATA CONSIDERING DIMENSIONS AND CONSTRAINTS

In this section we introduce the proposed algorithm. This algorithm mines time-extended sequences of data considering dimensions and constraints.
Previous algorithms mine sequences considering either dimension or constraints. An algorithm mines multidimensional data without considering constraints and time-duration in sequences [5]. To the best of our knowledge, various constraints had been introduced [18] but practical implementation of incorporating constraints along with dimensions and time-extended data is not implemented.

Parameters used in algorithm

Definition 1: Duration Between Item Purchased: This algorithm mines database of sequences. Sequences consist of occurrence of items and if items have occurred at different time interval, the duration between occurrences of each item is stated in the sequence.

In the database used for application of proposed algorithm, each medicine sold has date associated with it. The duration here is determined by the date difference (in term of month in this paper) between medicines purchased by each customer. This parameter enables to know the approximate duration between medicines purchased. By incorporating the duration we can get frequent sequences of medicines purchased and the date difference (duration) between each medicines purchased.

$Dur(\alpha)\theta \Delta t$, where $\theta \in \{\leq, \geq\}$ and Δt is a value indicating the temporal duration [15].

In this paper each sequence in database is composed of {(i1, i2) <d1> (i1) <d2> (i3,i1)…. <dn> (in)}

Definition 2: Multiple Dimensions: In this paper, incorporating dimensions season and company name in sequences enables to know not only sequences of medicines purchased but also sequences of medicines purchased as per company and season.

Constraints

Constraints are added in this algorithm to mine efficient, effective and useful sequences from data of medical store.

Definition 1: Length Constraint: This constraint specifies the length of the patterns to be mined. This constraint can be used to mine either too short or too long patterns. In this paper we have used this constraint to mine those sequences which have at least more than one item in sequence.

$Clen(\alpha) \equiv (len(\alpha)>1)$ [15].

Definition 2: Duration Constraint: Each medicine sold has date associated with it. The duration here is determined by the date difference (in term of month in this paper) between medicines purchased by each customer. Formally, duration constraint is in the form of $Dur(\alpha)\theta \Delta t$, where $\theta \in \{\leq, \geq\}$ and Δt is a value indicating the temporal duration. A sequence α satisfies the constraint if and only if the time interval between each item sold is consecutive

Definition 3: Checking for sequence having same medicine name: In this proposed work, this constraint identifies those sequences which do have same medicine in whole sequences. This information enables to know which particular medicine is sold each time in some particular duration.

Data Selection and Pre-process

From the normalized database of system of medical store, the data regarding purchase of medicine by customer was fetched. Master table of sales consisted of Customer Primary key (Ref No.), Customer name, BillDate. Child table of sales consisted of (Ref No.), List of each medicine purchased (ItemSrNo). Item Table consisted of ItemSrNo and ItemName(Medicine Name), Company Sr.No . Company table consisted of Company Sr.No and name of Company. By the use of macro a consolidated excel sheet was prepared which consisted of customer primary key, customer name, date of purchase of medicine, medicine name, company name. Based on the date of purchased each purchase was assigned a month and as per month season [Summer, Winter, Monsoon] was assigned to each purchase.

Consolidated Excel Sheet Prepared from Normalized database

Figure 1. Consolidated Data of Medicines

After preparing this consolidated sheet a database of sequence was prepared which consisted of transactions made by particular customer and time interval between his/her purchase.

Later dimensions (Company Sr.No and Season) were added to each sequence.

Multidimensional Sequences generated using above data from excel sheet

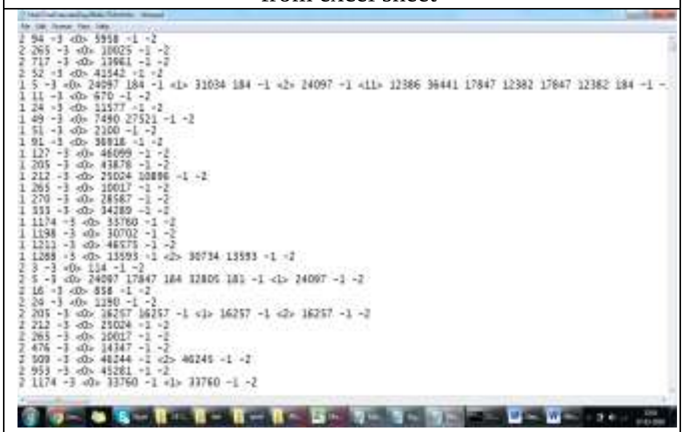

Figure 2 Multidimensional Time Extended Sequences generated from data shown above

Algorithm: Macro used for generating above multidimensional sequences
Step 1: For Each Sequence Step 2: If (Season and CompSr.No Same) Step 3: If (Month(PurDate) is same) Step 4 Append ItemsName Step 5: Append Season and CompSr.No with ItemsName Step 6: Else

Step 7: Calculate Month Gap
Step 8 Add Month Gap to previous Month Gap
Step 9 Append Time Interval (Month Gap) Between Items
Step 10: Append Season and CompSr.No with Items Name
Step 11 End if
Step 12: Else
Step 13 Terminate sequence with -1 -2
Step 14 Separate Dimension and sequence with -3
Step 15: Generate Next Sequence
Step 16: End If
Step 17: End For

Figure 3 Algorithm used for generating Multidimensional Time Extended Sequences

In the example above, the sequences is composed as stated below

1. Multiple Dimensions -2 <Time Interval> Item1, Item2 <Time Interva> item1.
2. Season Company sr.no -3 <Time Interval> Medicine No. <Time Interval> Medicine No1, Medicine No2. -1 -2
3. For example, In the first sequence 2 94 -3 <0> 5958 -1 -2

2 is Season, 94 is company no of medicine, -3 is separator between dimension and sequence <0> is time interval, 5958 is medicine number and -1 indicates end of item sold at particular date, -2 indicates end of sequence

Frequent sequences obtained after applying algorithm

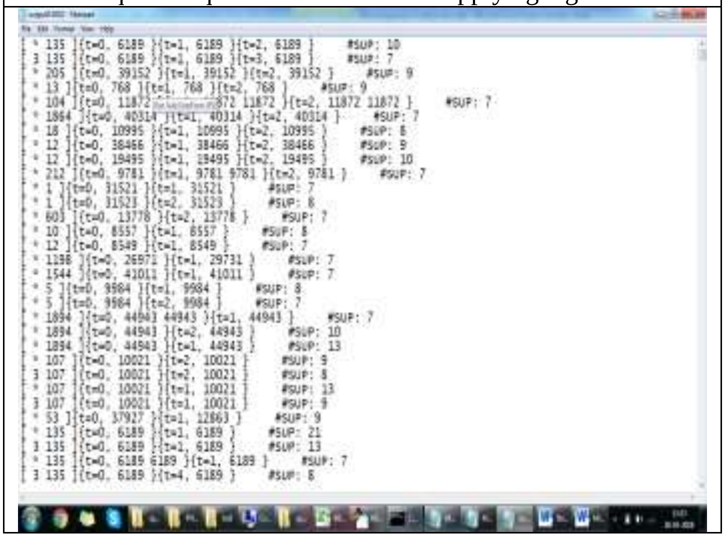

Figure 4 Frequent sequences obtained after applying algorithm without any constraints

Sequences Containing Season No, Comp No, Item No converted to name

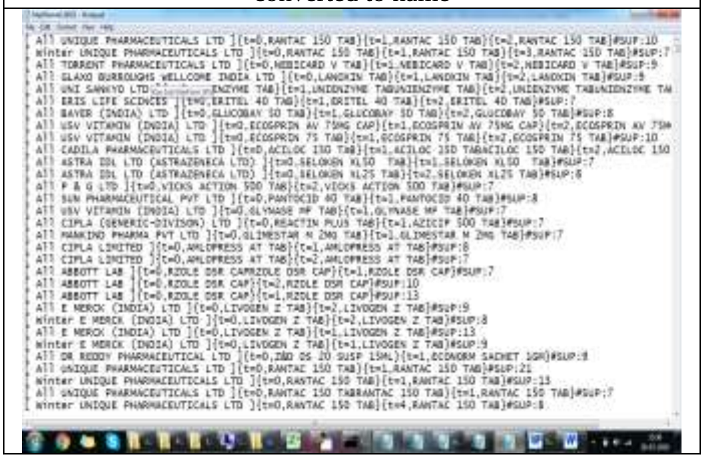

Figure 5 Sequences Containing Season No, Comp No, Item No converted to name

Algorithm: Macro used for converting Season, Item, Comp No. into Name
Step 1: For each sequence generated Step 2: Assign Season (1= Summer, 2=Rainy, 3= Winter, *=All) Step 3: Fetch Comp Name for Comp Sr.No Step 4: Fetch Item Name for Item Sr.No Step 5: End For

Figure 6 Algorithm used for converting Season, Item, Comp No. into Name

As an input sequence database Company Sr.No, Season number, medicine numbers were assigned. After the result of mining was obtained, these numbers were converted to name using macro for understanding and analysis of data.

Algorithm for mining multidimensional Time-Extended sequences with constraints

```
Input: (Sequences Data Set), Minimum Support, Constraints (Constraint Set)
Output: Sequential patterns satisfying constraints
Step 1:  Process Multi-dimensional patterns using CHARM algorithm
Step 2:  Apply PREFIXSPAN algorithm for mining frequent sequences
// Length constraints for pruning very short sequences
Step 3:
        3.1 For Each Sequence
                3.2     {If Len (S) >1 then
              3.3               Result-Set=Result-Set U S}
                3.4 End For
//Sequence with same medicine checks constraint. This constraint mines only that //sequences which contains same medicine name in whole sequence. //This enables to know which particular medicines are sold on daily //bases.
Step 4:
                4.1 For Each Sequence
                4.2   MedItm =  S[I1[0]]
                4.3     For Each j Item in Sequence
                4.4             If S[I1[j]] =MedItm then
                4.5                     Result-Set= Result-Set U S
                4.6       End For
                        4.7 End For
//Time-Interval (Duration Constraint) here is used to identify those //medicines which are sold in consecutive months
Step 5:
                5.1 For Each Sequence
                5.2   TimeInterval=  S[T1[0]]
                5.3     For Each j Time in Sequence
                5.4             If S[T1[j]] = TimeInterval then
                5.5                     Result-Set= Result-Set U S
               5.6              TimeInterval=TimeInterval+1
                5.7       End For
                5.8 End For
```

Figure 7 Algorithm for mining multidimensional Time-Extended Sequences with constraints

IV. EXPERIMENTAL RESULTS

Results obtained from multidimensional, Time-Extended algorithm without applying constraints

Sequential pattern mining was applied on dataset of medical store and by varying minimum support different sets of sequences were obtained.

TABLE 1

RESULTS OBTAINED FROM MULTIDIMENSIONAL, TIME-EXTENDED ALGORITHM WITHOUT APPLYING CONSTRAINTS WHEN SUPPORT VARIES

Program	**Support**	**No. Of Sequences generated**	**Total Time**	**Max Memory**
MainTestSequentialPatternMining4	0.0002	3558	22714 ms	329.09
	0.0004	1271	7753 ms	325.03
	0.0009	313	2621 ms	148.88
	0.0015	109	1092 ms	83.76
	0.0018	73	936 ms	82.93
	0.0021	50	702 ms	65.87

Above table represents, that as minimum support is reduced the number of sequences generated is also reduced, but this may miss some important sequences that may require for analysis.
Thus to retain important sequences and to reduce total number of sequences generated constraints are incorporated. Table below represents constraints and sequences generated.

Results obtained from multidimensional, Time-Extended algorithm after applying constraints. Elaborated data with constraints for support 0.0002

TABLE 2: RESULTS OBTAINED FROM MULTIDIMENSIONAL, TIME-EXTENDED ALGORITHM AFTER APPLYING CONSTRAINTS

Support	**Constraints**	**No. Of Sequences generated**	**Total Time**	**Max Memory**	**Description**
0.0002	No	3558	22304 ms	323.15	
0.0002	Length	164	22740 ms	330.81	This constraint was incorporated to remove too short sequences. i.e sequences having less than 2 items is not included in result set
0.0002	Length+Same Med. check Constraint	133	22384 ms	300.08	This constraint produces only those sequences which satisfy length constraint and have same medicine name. Sequences generated after incorporating this sequence enables to know which particular medicine is sold frequently between some time interval
0.0002	Length+Same Med. check Constraint + Time Interval	82	22540 ms	317.39	This constraint produces only those sequences which satisfy length constraint, same medicine check constraint and those sequences which contains medicines sold in consecutive month.

COMPARATIVE STUDY

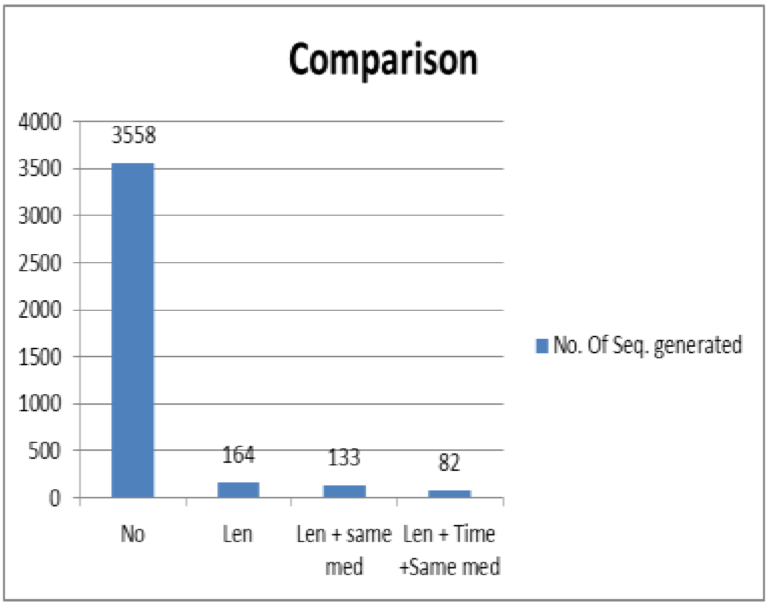

Chart 1: *Comparison of No. of sequences generated with and without constraints*

- Result obtained after applying proposed algorithm on dataset with Length Constraint

Result Set Sequences Generated with Length constraint

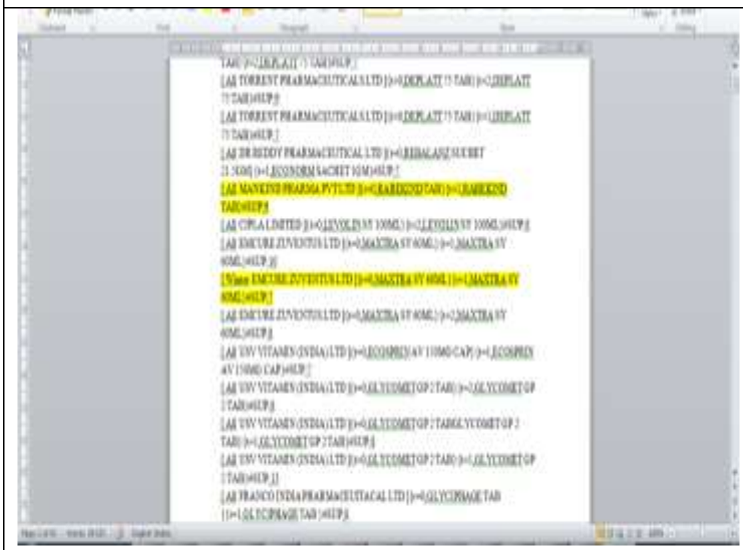

Figure 8 Sequences Generated with Length constraint

a) **No. of sequences generated:** 164
b) **Constraint Applied:** Length
c) Analysis of few records:
   (1) Length Constraint was incorporated to reduce too short sequences.
   (2) The number of sequences generated without constraints was 3558 which was reduced to by incorporating length constraint. 164
   (3) Analysis of some of the records represents Maxtra Sy is sold more in Winter as compared to other season. Maxtra Sy is used to cure cough and cold
   (4) Rabekind used for acidity is sold in all seasons
   (5) Thus Rabekind is being sold in monthly interval, so stock can be updated accordingly
   (6)

- Result obtained after applying proposed algorithm on dataset with Length Constraint and Same medicine check constraint

Result Set Sequences Generated with Length and Same medicine check Constraint

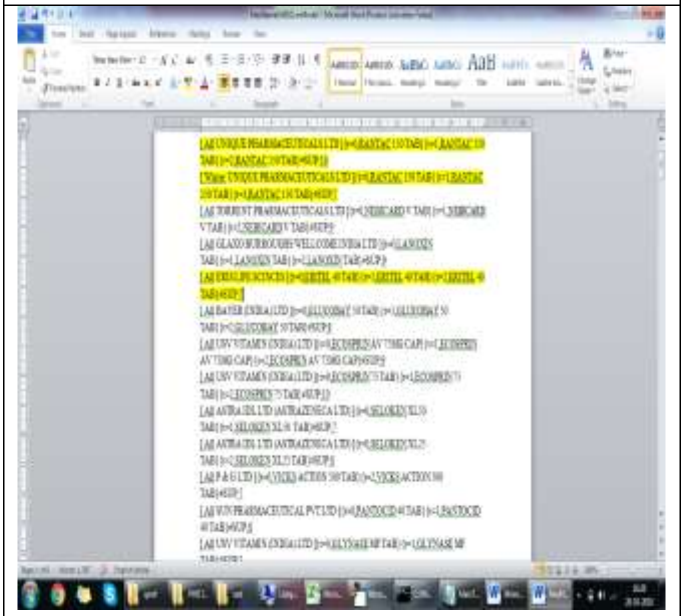

Figure 9 Sequences Generated with Length and Same medicine check constraint

a) **No. of sequences generated:** 133
b) **Constraint Applied:** Length and Same medicine check Constraint
c) Analysis of few records:
   (1) Same medicine check Constraints gives super or subset of a sequence
   (2) Here, we take first item of sequence and find those items that is super set of first item provided all the items are same. This constraint enables to identify which medicines are sold monthly.
   (3) Above analysis shows Rantac used for acidity is sold on regular basis as acidity problem may persist for long time.
   (4) Some research shows acidity becomes worse in winter. This may increase sales in winter. Analysis above shows Rantac is sold more in winter as compared to other season.
   (5) Eritel tablet is sold in all seasons. Eritel Tablet is used to treat high blood pressure, prevent stroke and other heart. diseases.

- Result obtained after applying proposed algorithm on dataset with Length Constraint and Time Interval Constraint

Result Set Sequences Generated with Length and Time Constraint

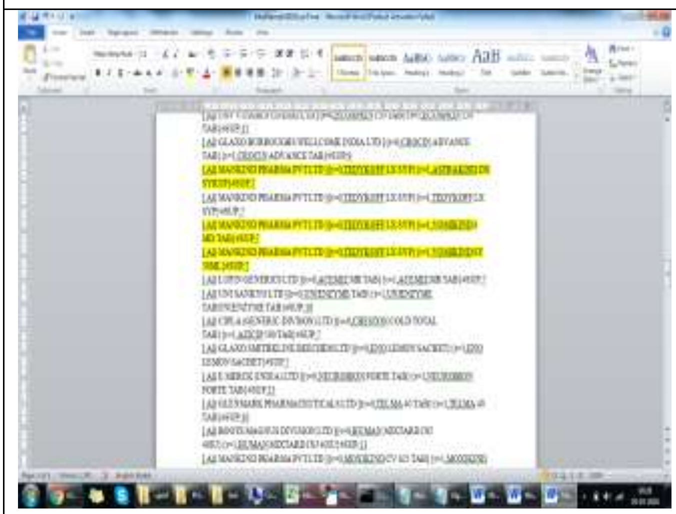

Figure 10 Sequences Generated with Length and Time constraint

a) **No. of sequences generated:** 110
b) **Constraint Applied:** Len & Time Constraint
c) Analysis of few records:
   (1) This constraint enables to get medicines which are sold in consecutive months.
   (2) Sales of Tedykoff is followed by Asthakind. Both used for cough but may be because cough as not treated by tedykoff, Asthakind is required

V. CONCLUSION

This proposed work introduces the idea of producing relevant useful sequences only, by incorporating dimensions and constraint together in time-interval based data sequences. This algorithm is applied on the data of sales of medicine. Merely applying sequential Pattern Mining algorithm may generate huge number of sequences which also consist of many irrelevant sequences. Thus incorporating constraints in algorithm controls the number of sequences generated. Dimensions incorporated in sequences gives more useful information like sales of medicine season-wise and company-wise. Time-Interval in the sequence describes duration between the items sold. Experiment shows that the number of sequences generated after adding Length constraints to algorithm were approximately 20 times less than those number of sequences which were generated by algorithm without constraint. Further adding same medicine check constraint to length constraint reduced number of sequences generated to 133. Adding Time interval constraint to length constraint and same medicine check constraints reduced the number of sequences generated to 82. Experiments and analyses show that this algorithm could effectively discover the useful and less number of medicine sequences along with dimensional information.

REFERENCES

[1] R.Agrawal, R.Srikant, “Mining Sequential Patterns”, Proceedings of 11th International Conference on Data Engineering, Taiwan, pp. 429-435, 1995
[2] M.N.Garoflalakis et.al, “SPIRIT: Sequential Pattern Mining with Regular Expression Constraints”, Proceedings of the 25th VLDB Conference Edinburg, Scotland, pp. 223-234,1999
[3] M.J.Zaki,”SPADE:An Efficient Algorithm for Mining Frequent Sequences.”, Machine Learning,Vol-42, Issue 1-2,pp 31-60, January 2001.
[4] J.Pei et.al, “PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth”, IEEE Transactions on Knowledge and Data Engineering, Vol.4,Issue 11,pp. 1424-1440, Nov 2004.
[5] H.Pinto, J.Han, J.Pei, K Wang, Q.Chen, U. Dayal, “Multi-Dimensional Sequential Pattern Mining”, Proceedings of the tenth international conference on Information and knowledge management, Atlanta, Georgia, USA,pp. 81-88, 2001
[6] X.Yan, J.Han, and R.Afshar, “CloSpan: Mining Closed Sequential Patterns in Large Datasets”, Proceedings of the 2003 SIAM International Conference on Data Mining, California USA,pp. 166-177, 2003
[7] K.Wang, Y Xu, J Xu Yu, “Scalable Sequential Pattern Mining for Biological Sequences”, proceedings of the thirteenth ACM international conference on Information and knowledge management, Washington, D.C., USA, pp.178-187, 2004
[8] L.I.Gomez et.al, “RE-SPaM: Using Regular Expressions for Sequential Pattern Mining in Trajectory Databases” in IEEE International Conference on Data Mining Workshops, 2008
[9] X.Hong et.al, “Mining Multi-Attribute Event Sequential Pattern Based on Association Rule” in Seventh International Conference on Fuzzy Systems and knowledge Discovery, 2010
[10] Y.C.Chen, “CEMiner- An Efficient Algorithm for Mining Closed Patterns from Time Interval-based Data” in 11th IEEE International Conference on Data Mining,2011
[11] N.Bechet et.al,” Sequential Pattern Mining to Discover Relations between Genes and Rare Diseases” in Computer-Based Medical Systems (CBMS) 25th International Symposium, 2012
[12] X. Qin, Y. Liu, “Matrix-Based Multidimensional Sequential Pattern Mining Algorithm and Application” in International Conference on Computer Science and Information Processing (CSIP), 2012
[13] H. TAKEI, H. YAMANA,” IC-BIDE: Intensity Constraint-based Closed Sequential Pattern Mining for Coding Pattern Extraction”, in IEEE 27th International Conference on Advanced Information Networking and Applications,2013
[14] W.Wang et.al, “Mining Frequent Closed Sequential Patterns with Non-user-defined Gap Constraints”, Advanced Data Mining and Applications Lecture Notes in Computer Science Volume 8933, 2014, pp. 57-70
[15] J.Wang and J.Han,” BIDE: Efficient Mining of Frequent Closed Sequences”, Proceedings of the 20th International Conference on Data Engineering (ICDE’04), 2004
[16] P.Songram,V.Boonjing,S.Intakosum, “Closed Multidimensional Sequential Pattern Mining”, Proceedings of the Third International Conference on Information Technology: New Generations (ITNG'06), 2006, IEEE

[17] V.Boonjing, P.Songram, "Efficient Algorithms for Mining Closed Multidimensional Sequential Patterns", Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007), 2007, IEEE

[18] J.Pei, J.Han, W.Wang, "Mining Sequential Patterns with Constraints in Large Databases", CIKM'02, November 4–9, 2002, McLean, Virginia, USA, Copyright 2002 ACM

[19] H. Takei, H. Yamana, "IC-BIDE: Intensity Constraint-based Closed Sequential Pattern Mining for Coding Pattern Extraction", IEEE 27th International Conference on Advanced Information Networking and Applications,2013

[20] U.Yun, "Mining lossless closed frequent patterns with weight constraints", Knowledge-Based Systems, Volume 20, Issue 1, February 2007, 86–97

[21] Li. Chun,W.Jianyong, "Efficiently Mining Closed Subsequences with Gap Constraints", Conference: Proceedings of the SIAM International Conference on Data Mining, SDM 2008, April 24-26, 2008, Atlanta, Georgia, USA

[22] B.Mallick, D.Garg, and P.Singh, "Constraint-Based Sequential Pattern Mining:A Pattern Growth Algorithm Incorporating Compactness, Length and Monetary", The International Arab Journal of Information Technology, Vol. 11, No. 1, January 2014

[23] H.Park,T.Kim,M.Li,H.Shon,J.Park,K.Ryu, "Application of Gap-Constraints Given Sequential Frequent Pattern Mining for Protein Function Prediction", Osong Public Health Resarch Perspectives, Korea Centres for Disease Control and Prevention.

[24] Y.L Chen, M.C Chiang, M.T Kao, "An algorithm to discover time-interval sequential patterns in sequence database", Expert Systems with Applications, 25, 343–354,2003