Target Oriented Sequential Pattern Mining using Recency and Monetary Constraints

Chetna Chand
U and P.U. Patel Department of Computer Engineering,
Chandubhai S Patel Institute of Technology,
Charotar University of Science and Technology,
Changa, Gujarat, India

Amit Thakkar
Department of Information Technology,
Chandubhai S Patel Institute of Technology,
Charotar University of Science and Technology,
Changa, Gujarat, India

Amit Ganatra
U and P.U. Patel Department of Computer Engineering,
Chandubhai S Patel Institute of Technology,
Charotar University of Science and Technology,
Changa, Gujarat, India

ABSTRACT
Many approaches in constraint based sequential pattern mining have been proposed and most of them focus only on the concept of frequency, which means, if a pattern is not frequent, it is removed from further consideration. Frequency is a good indicator of the importance of a pattern but in real life, however, the environment may change constantly and patterns discovered from database may also change over time. Therefore, the users' recent behavior is not necessarily the same as the past ones and a pattern that occurs frequently in the past may never happen again in the future. So in this paper we have considered recency constraint to overcome this problem.
Also we have considered one more constraint, monetary constraint since for making effective marketing strategies it is important to know the value of customer on the basis of what they are purchasing periodically and how much they are spending. So this motivates to consider monetary value of customers for targeting profitable customers. Along with that we have included the concept of mining only target oriented sequential patterns which satisfy RFM constraints to find the happening order of a concerned itemsets only, for taking effective marketing decisions.

General Terms
Sequence Database, Data mining.

Keywords
Sequential Pattern Mining, PrefixSpan, RF-SP, RFM-SP.

1. INTRODUCTION
Sequential Pattern Mining finds interesting sequential patterns among the large database. It finds out frequent subsequences as patterns from a sequence database. With massive amounts of data continuously being collected and stored, many industries are becoming interested in mining sequential patterns from their database. Sequential pattern mining is one of the most well-known methods and has broad applications including web-log analysis, customer purchase behavior analysis and medical record analysis. In the retailing business, sequential patterns can be mined from the transaction records of customers. For example, having bought a notebook, a customer comes back to buy a PDA and a WLAN card next time. The retailer can use such information for analyzing the behavior of the customers, to understand their interests, to satisfy their demands, and above all, to predict their needs. Web log analysis, the exploring behavior of a user can be extracted from member records or log files. For example, having viewed a web page on "Data Mining", user will return to explore "Business Intelligence" for new information next time. In the medical field, sequential patterns of symptoms and diseases exhibited by patients identify strong symptom/disease correlations that can be a valuable source of information for medical diagnosis and preventive medicine. In yield huge benefits, when acted upon, increases customer royalty.

In this paper the proposed algorithm discovers target oriented sequential pattern using RFM constraint from sequence database by applying recency, frequency and monetary constraints during the mining process. The Proposed algorithm modifies sequential pattern mining algorithm PrefixSpan (Pattern Growth-based). From the analysis of PrefixSpan it is clear that it is an efficient algorithm with respect to execution time, scalability and memory usage among most of the algorithm for sequential pattern mining. But it also has some limitations with respect to construction of number of projected database and taking only frequency as an interestingness measure for the generation of patterns.

1.1 Basic Concepts of Sequential Pattern Mining

Let I = {x1, . . . , xn} be a set of items, each possibly being associated with a set of attributes, such as value, price, profit, calling distance, period, etc. The value on attribute A of item x is denoted by x.A. An itemset is a non-empty subset of items, and an itemset with k items is called a k-itemset.
A sequence $\alpha = \langle X1, X2, X3,\ldots, Xl\rangle$ is an ordered list of itemsets. An itemset Xi ($1 \le i \le l$) in a sequence is called a transaction, a term originated from analyzing customers' shopping sequences in a transaction database. A transaction Xi may have a special attribute, time-stamp, denoted by Xi.time, which registers the time when the transaction was executed. For a sequence $\alpha = \langle X1, X2, X3,\ldots, Xl \rangle$, we assume Xi.time < Xj.time for $1 \le i < j \le l$.
The number of transactions in a sequence is called the length of the sequence. A sequence with length l is called an l-sequence. For an l-sequence α, we have len (α) = l. Furthermore, the i-th itemset is denoted by α[i]. An item can occur at most once in an itemset, but can occur multiple times in various itemsets in a sequence.

A sequence $\alpha = \langle X1 \ldots Xn\rangle$ is called a subsequence of another sequence $\beta = \langle Y1 \ldots Ym\rangle$ ($n \le m$), and β a super-

sequence of α, if there exist integers $1 \leq i1 < . . < in \leq m$ such that X1 $\subseteq$ Yi1 , . . . , Xn $\subseteq$ Yin.

A sequence database SDB is a set of 2-tuples (sid, α), where sid is a sequence-id and α a sequence. A tuple (sid, α) in a sequence database SDB is said to contain a sequence γ if γ is a subsequence of α. The number of tuples in a sequence database SDB containing sequence γ is called the support of γ, denoted by sup (γ). Given a positive integer min_sup as the support threshold, a sequence γ is a sequential pattern in sequence database SDB if sup (γ) ≥ min_sup. The sequential pattern mining problem is to find the complete set of sequential patterns with respect to a given sequence database

2. RELATED WORK

According to the literature review, algorithms for finding sequential patterns are broadly divided into three phases. In the initial phase algorithm designed for sequential pattern mining are based on apriori property and take frequency as a measure to generate sequential pattern [1][2]. These algorithms use generate and test approach and generate exponential amount of patterns when there is a need to find longer sequential pattern or when frequency count is decreased.

Then since the time complexity and space complexity of algorithm increases exponentially when decreasing support value for a given dataset, then focus of researchers shifted to the second phase of algorithm that is, designing algorithms which make use of some efficient data structure for increasing the retrieval time and decreasing the space required for storing those intermediate patterns. The first algorithm in this category is SPADE [7] in which the sequence database is transformed into a vertical format consisting of items' id-lists. The id-list of an item, is a list of (sequence-id, timestamp) pairs indicating the occurring timestamps of the item in that sequence. Searching in the lattice formed by id-list intersections, the SPADE (Sequential Pattern Discovery using Equivalence classes) algorithm presented by M.J.Zaki [7] completes the mining in three passes of database scanning. Second algorithm is SPAM which uses a strategy for mining sequential patterns by traversing the lexicographical sequence tree in depth-first fashion. SPAM uses a vertical bitmap data structure representation of the database, which is similar to the id list in SPADE. SPAM is similar to SPADE, but it uses bitwise operations rather than regular and temporal joins. Next algorithm is PrefixSpan [8] with the key idea to avoid the candidate generation step altogether, and to focus the search on a restricted portion of the initial database. The search space partitioning feature plays an important role in pattern-growth. Almost every pattern-growth algorithm starts by building a representation of the database to be mined, then proposes a way to partition the search space, and generates as few candidate sequences as possible by growing on the already mined frequent sequences, and applying the apriori property as the search space is being traversed recursively looking for frequent sequences.

Third phase of SPM algorithm is based on the wide spread use of sequential pattern in retail business and for helping in various CRM (Customer Relationship Management) activities in an organization. Since the use of existing algorithms overloaded the user with many sequential patterns which are difficult to handle and sometimes are of no use to the organization the concept of constraint based sequential pattern mining originates. The idea was to apply some constraint during the mining process by which we can prune unwanted patterns during the mining process and which ultimately leads to generation of only interesting patterns to the user. The first algorithm in this category is SPIRIT [4] which uses regular expressions as flexible constraint specification tool. It involves a generic user-specified regular expression constraint on the mined patterns, thus enabling considerably versatile and powerful restrictions. Then various constraints like item constraint, length constraint, aggregate constraint and many other constraints are proposed by Pei, Han, & Wang, 2007 to limit the patterns to be found to a particular subset satisfying some strong conditions.

At last there are other constraints like RFM (Recency Monetary Frequency) which are for the very first time introduced in sequential pattern mining by Yen-Liang Chen [17] in 2009. The concept is implemented by the author in GSP algorithm, which increased the effectiveness and efficiency of algorithm compared to the traditional algorithm. Also there is one more concept introduced in [18], for finding target oriented sequential patterns instead of mining complete set of sequential pattern for some given target itemset implemented in GSP algorithm. So now in this paper the proposed idea is to combine target oriented mining and RFM constraint in PrefixSpan algorithm to mine sequential patterns.

3. PROBLEM DEFINATION

Representation of Data Sequence: In a traditional approach data sequence is represented as a list of itemsets ordered by transaction time. But for the proposed approach the data sequence is represented in another way : a data-sequence A is represented as <(a1, t1, m1), (a2, t2, m2), …,(an, tn, mn)>, where (aj, tj, mj) means that item aj is purchased at time tj with total value mj, $1 \leq j \leq n$, and $tj-1 \leq tj$ for $2 \leq j \leq n$. In the data-sequence, if items occur at the same time, they are ordered alphabetically.

Table 1. Sample Data Sequence

Sid	Sequence
10	< (a, 1, 10), (c, 3, 40), (a, 4, 30), (b, 4, 70), (a, 6, 50)>
20	< (b, 3, 30), (c, 5, 50), (a, 7, 20), (b, 7, 70)>
30	< (a, 8, 40), (b, 8, 50), (b, 16, 20), (c, 20, 100) >
40	< (b, 15, 30), (b, 22, 20), (c, 22, 120) >
50	< (c, 5, 30), (b, 6, 40), (a, 10, 30), (b, 10, 60)>

Target Oriented Sequential Pattern:

A target-oriented sequential pattern is a pattern with a concerned itemset in the end of pattern. For most decision makers, when they want to make efficient marketing strategies, they usually concern the happening order of a concerned itemsets only, and thus, most sequential patterns discovered by using traditional algorithms are irrelevant and useless. E.g. suppose one sample database is given and we have to find target oriented pattern with the concerned itemset s7 and minimum support threshold equal to 0.3.

Table 2. A sample sequence database

C.No.	Sequence
C001	(s5 ,8), (s4 ,15), (s6 ,20)
C002	(s1 ,2), (s3 ,7), (s2 ,11), (s6 ,18)
C003	(s2 ,3), (s1 ,4), (s3 ,7), (s6 ,17), (s7 ,19)
C004	(s1, 2), (s2, 8), (s6, 10), (s7, 15)
C005	(s5 ,4), (s6 ,16),(s1 ,20),(s3 ,24)
C006	(s7 ,7), (s1 ,13), (s5 ,18), (s2 ,25), (s6 ,28)
C007	(s5 ,4), (s1 ,8), (s3 ,12), (s6 ,16), (s7 ,20)

C008	(s1 ,3), (s5 ,6), (s2 ,9), (s4 ,18), (s6 ,21)
C009	(s2 ,5), (s1 ,10), (s3 ,15), (s6 ,20), (s7 ,25)
C010	(s6 ,3), (s7 ,8), (s5 ,12), (s2 ,17)

After applying sequential pattern mining algorithm following are the target oriented sequential pattern with the concerned itemset s7:
<s1,s7>, <s2,s7>, <s3,s7>, <s6,s7>, <s2,s1,s7>, <s1,s3,s7>, <s2,s3,s7>, <s2,s1,s3,s7>

Recency Constraint:

Recency constraint is specified by giving a recency minimum support (r_minsup), which is the number of days away from the starting date of the sequence database. For example, if our sequence database is from 27/12/2009 to 31/12/2010 and if we set r_minsup = 200 then the recency constraint ensures that the last transaction of the discovered pattern must occur after 27/12/2009+200 days. In other words, suppose the discovered pattern is < (a), (bc)>, which means "after buying item a, the customer returns to buy item b and item c". Then, the transaction in the sequence that buys item b and item c must satisfy recency constraint.

Monetary Constraint:

Monetary constraint is specified by giving monetary minimum support (M_min). It ensures that the total value of the discovered pattern must be greater than M_min. Suppose the pattern is < (a), (bc)>. Then we can say that a sequence satisfies this pattern with respect to the monetary constraint, if we can find an occurrence of pattern < (a), (bc)> in this data sequence whose total value must be greater than M_min.

Frequency Constraint:

Frequency constraint is specified by giving frequency minimum support (f_minsup). The frequency of a pattern is the percentage of sequences in database that satisfy the recency constraint and monetary constraint. And a pattern could be output as an RFM-pattern if its frequency is greater than f_minsup.
By setting these three constraints properly, we can discover an RFM-pattern like this, "30% customers who recently bought a computer would return to buy the scanner and microphone, and the total value will exceed Rs. 50,000".

F-SP, RF-SP, RFM-SP:

Let B = < I1 I2….IS > be a sequence of itemsets. If the percentage of data sequences in database containing B as a subsequence, called f-support, is no less than f_minsup, B is called a Frequent Sequential Pattern (F-SP). B is called an RF-SP if the percentage of data sequences in database containing B as a recent subsequence (which satisfies recency constraint), called rf-support, is no less than f_minsup. Finally, B is called an RFM-SP if the percentage of data sequences in database containing B as a recent monetary subsequence (which satisfies recency and monetary constraints), called rfm-support, is no less than f_minsup.

4. PROPOSED ALGORITHM

In the proposed algorithm the performance of finding frequent sequences is improved by applying various user specified constraints during mining process. Traditional sequential pattern mining only distinguishes whether a pattern appears or not, while target based RFM pattern mining approach not only determines the existence of a pattern but also checks whether the pattern is target oriented and also satisfies the recency and the monetary constraints. So the proposed system is designed to handle both the problems of existing algorithm.

The Proposed Algorithm:

Input:
- Data-sequence database
- The threshold of support given by the user (minsup_min, minsup_max)
- The threshold of recency given by the user (Rtime_min, Rtime_max)
- The threshold of monetary given by the user (M_min, M_max)
- The target oriented itemset selected by the user

Output: The Complete set of all RF-SPs and RFM-SPs
Method: Call PrefixSpan (< >, 0, S)
Subroutine: PrefixSpan (α, L, S |α)
Parameter: α: a RF-sequential pattern; L: the length of α;
S |α: the α-projected database, if α-projected database, if α ≠ < >; otherwise, the sequence database S.

1. Transform the database according to the target itemset α
(a) Reverse each original sequence so that the last itemset of the sequence can be arranged in front of the sequence.
(b) Remove the reversed sequences without the concerned itemsets.
(c) Delete itemsets which are arranged in front of the concerned itemsets from each retained sequences.
(d) Delete itemsets in which concerned itemset do not satisfy Recency constraint.

2. For itemset α,
(a) Output α as RF-SP, if α satisfies frequency and Recency constraint
(b) Output α as RFM-SP, if α satisfies frequency Recency and monetary constraint.

3. Method:
For each item b in S |α
b can be appended to the last element of α as α' or < b > can be appended to α as α'
Scan the database S |α once and calculate FValue (α') and MValue (α') for each α'
For each α'
If minsup_max > FValue (α') > minsup_min
Output α' as RF-SP
If M_max > MValue (α') > M_min
Output α' as RFM-SP;
4. For each α', construct α'-projected database S |α', and
Call PrefixSpan (α', L+1, S |α')

Fig 1: Proposed Algorithm

4.1 Flow Chart of Proposed Algorithm

Various Steps of Proposed System are:

- Conversion of sequence database to target based Sequence Database
- Applying Recency Constraint
- Finding length-1 RF-SP and RFM-SP

- Construction of recursive projected database

4.1.1 Apply Target-Oriented Concept

In order to find target oriented sequential pattern, the first step is to convert the original sequence database according to the concerned itemset. For example, suppose a sequence database is give below are the various steps for the required conversion.

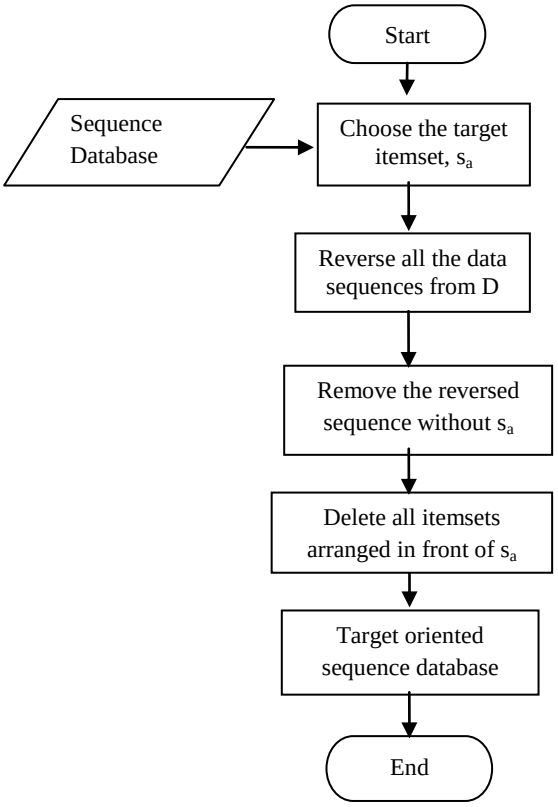

Fig 2: Transformation of sequence database according to target itemset.

4.1.2 Apply Recency Constraint

Constraints are applied after the database gets transformed according to the concerned itemset. Recency constraint is applied to the transformed database, and all the patterns in which, the first occurrence of itemset, which is the target itemset if do not satisfy the recency constraint then other occurrences of that itemset are also checked in that sequence and if none satisfy the constraint the that sequence is eliminated from the database. Now since the recency constraint satisfies the downward closed property, so all the subsequences automatically satisfies this constraint so there is no need to check this constraint afterwards in mining process.

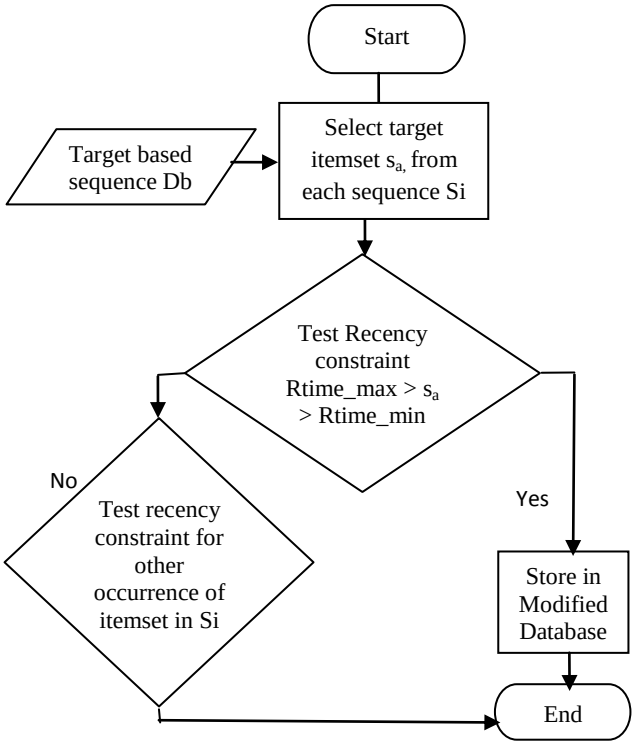

Fig 3: Application of Recency constraint in sequence database.

4.1.3 Finding length-1 RF-SP and RFM-SP

From the modified sequence database of previous step, now we have to find out frequency and monetary count of the target itemset. If it is frequent then it is recent frequent sequential pattern (RF-SP) and if it's monetary score MValue is between the minimum and maximum threshold than it is also a recent frequent monetary sequential pattern (RFM-SP).

4.1.4 Construction of recursive projected database

The major cost of PrefixSpan algorithm is the construction of projected database. If the number and/or the size of projected databases can be reduced, the performance of sequential pattern mining can be further improved. So in order to achieve this goal in the proposed algorithm we are using the concept of target oriented sequential pattern with three constraints, Recency, Frequency and Monetary. First, for finding target oriented pattern we have to transform the database to a form in which all the unnecessary sequences which are not contributing in finding the target based sequential pattern are removed and then this refined database is given to the mining algorithm for finding sequential pattern. So because of this the size of projected database is reduced consequently and the processing will be fast. Secondly, since we are adding two new constraints we can reduce the size and number of recursive projected database construction. Out of these constraints monetary constraint do not satisfy downward closure property, according to which "If a sequence satisfying a constraint, all the subsequences of that sequence will also satisfy the constraint". So we can utilize other two constraints, Frequency and Recency which satisfy downward closure property to reduce the search space and efficiently discovers all RFM-patterns. So during the mining process recursive construction of projected database take place by taking each

RF-SP pattern found in previous step as the prefix to be projected next to generate RF-SP and RFM-SP of length L+1, where L is the length of prefix.

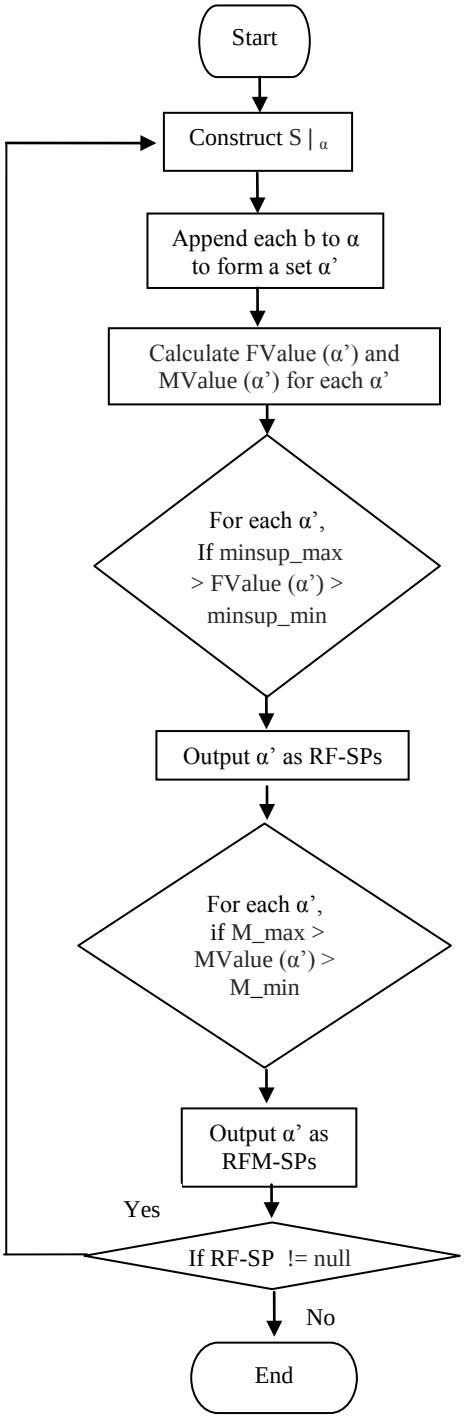

Fig 4: Construction of recursive projected database

4.2 Target Oriented RFM-PrefixSpan Algorithm – Example.

Consider a data sequence DB given below and six thresholds, and suppose the target data item is c.

Table 3. A Sample Database

Sid	Sequence
10	< (a, 1, 10), (c, 3, 40), (a, 4, 30), (b, 4, 70), (a, 6, 50), (c, 10, 70) >
20	< (b, 3, 30), (c, 5, 50), (a, 7, 20), (b, 7, 70), (c, 14, 20) >
30	< (a, 8, 40), (b, 8, 50), (b, 16, 20), (c, 20, 100) >
40	< (b, 15, 30), (b, 22, 20), (c, 22, 120) >
50	< (c, 5, 30), (b, 6, 40), (a, 10, 30), (b, 10, 60), (b,19, 90), (c, 19, 70) >

Rtime_min = 10, Rtime_max = 21,
M_min = 150, M_max = 250,
minsup_min = 2, minsup_max = 4

Step 1: First step is to reverse the sequence and delete the sequence not having the concerned itemset and then delete those itemsets which are in front of the concerned itemset from that reversed sequence and at last delete the sequence in which the concerned itemset do not satisfy Recency constraint.

Sid	Sequence
10	< (c, 10, 70), (a, 6, 50), (b, 4, 70), (a, 4, 30) , (c, 3, 40), (a, 1, 10),>
20	< (c, 14, 20), (b, 7, 70), (a, 7, 20), (c, 5, 50), (b, 3, 30) >
30	< (c, 20, 100), (b, 16, 20), (b, 8, 50), (a, 8, 40), >
40	< (c, 22, 120), (b, 22, 20), (b, 15, 30) >
50	< (c, 19, 70), (b, 19, 90), (b, 10, 60), (a, 10, 30), (b, 6, 40), (c, 5, 30) >

Fig 5: The Reversed sample sequence database

Sid	Sequence
10	< (c, 10, 70), (a, 6, 50), (b, 4, 70), (a, 4, 30) , (c, 3, 40), (a, 1, 10),>
20	< (c, 14, 20), (b, 7, 70), (a, 7, 20), (c, 5, 50), (b, 3, 30) >
30	< (c, 20, 100), (b, 16, 20), (b, 8, 50), (a, 8, 40), >
40	< (c, 22, 120), (b, 22, 20), (b, 15, 30) >
50	< (c, 19, 70), (b, 19, 90), (b, 10, 60), (a, 10, 30), (b, 6, 40), (c, 5, 30) >

Fig 6: Sequence database with concerned itemset c in the first order.

Sid	Sequence
10	< (c, 10, 70), (a, 6, 50), (b, 4, 70), (a, 4, 30) , (c, 3, 40), (a, 1, 10),>
20	< (c, 14, 20), (b, 7, 70), (a, 7, 20), (c, 5, 50), (b, 3, 30) >
30	< (c, 20, 100), (b, 16, 20), (b, 8, 50), (a, 8, 40), >
50	< (c, 19, 70), (b, 19, 90), (b, 10, 60), (a, 10, 30), (b, 6, 40), (c, 5, 30) >

Fig 7: Sequences having concerned itemset satisfying Recency constraint

Step 2: After the construction of transformed database according to the concerned itemset, the next step is to check the concerned itemset for frequency and monetary constraint and generate length 1 RF-SP and RFM-SP.

Sid	Sequence
10	< (c, 10, 70), (a, 6, 50), (b, 4, 70), (a, 4, 30) , (c, 3, 40), (a, 1, 10),>
20	< (c, 14, 20), (b, 7, 70), (a, 7, 20), (c, 5, 50), (b, 3, 30) >
30	< (c, 20, 100), (b, 16, 20), (b, 8, 50), (a, 8, 40), >
50	< (c, 19, 70), (b, 19, 90), (b, 10, 60), (a, 10, 30), (b, 6, 40), (c, 5, 30) >

Itemset	RF-sup	RFM-Values
< c >	4	70, 20, 100, 70

RF-SP	RFM-SP
< c >	

Fig 8: Calculate frequency and monetary count of target item

Step 3: Next step is the construction of < c > projected database from which first an itemset is selected and appended to the prefix < c >, and then it's support count and monetary value is calculated which if satisfy the given user defined criteria results in RF-SP and RFM-SP.

Sid	MValue of < c >	Sequence
10	70	<(a,6,50), (b, 4, 70), (a, 4, 30), (c, 3, 40),(a,1,10)>
20	20	<(b,7, 70), (a, 7, 20), (c, 5, 50), (b, 3, 30) >
30	100	<(b,16,20), (b, 8, 50), (a, 8, 40), >
50	70	<(b,19,90),(b, 10,60), (a, 10, 30), (b, 6, 40), (c, 5, 30) >

itemset	rf-sup	rfm-values
<ca>	4	120,40, 140,100
<cb>	4	140,90, 120,160
<cc>	3	110,70, 100

LI_2^{rf}	LI_2^{rfm}
< ca >	
< cb >	
< cc >	

Fig 9: Sequential patterns using < c > Projected database

From the output of <c> projected database it is find out that < ca >, < cb > and < cc > are RF-SP patterns. So next step will be to construct their projected databases and calculate RF-SP and RFM-SP.

Sid	MValue of < ca >	Sequence
10	120	<(b,4,70), (a, 4, 30) ,(c,3,40), (a,1,10)>
20	40	<(c,5, 50), (b, 3, 30) >
50	100	< (b, 6, 40), (c, 5, 30) >

itemset	rf-sup	rfm-sup
< caa >	1	150
< cab >	3	190,70,140
< cac >	3	160,90,130

RF-SP	RFM-SP
< cab >	
< cac >	

Sid	MValue of < cab >	Sequence
10	190	<(a,4,30) ,(c,3,40), (a,1,10)>
50	100	<(c,5, 30) >

itemset	rf-sup	rfm-sup
< caba >	1	210
< cabc >	2	230,130

RF-SP	RFM-SP
< cabc >	

Fig 11: Sequential patterns using < cab > Projected database

Sid	MValue of < cac >	Sequence
10	160	<(a,1,10>
20	90	<(b,3,30) >

itemset	rf-sup	rfm-sup
<caca >	1	170
<cacb >	1	120

Fig 12: Sequential patterns using < cac > Projected database

Similarly we can find out RF-SP, RFM-SP for <cb>and <cc> by projecting their databases recursively and finally we will get complete set of sequential pattern as shown below, but since they are in reversed order so first we have to reverse them and then we obtain final output patterns which are target oriented and satisfy RFM Constraints.

RF-SP	RFM-SP
<c>, < ca >, <cb>, <cc>, <cab>,<cac >,<cba>,<cbb>,<cbc>, <cbac>	<cba >, <cbac>

RF-SP	RFM- SP
<c>, <ac>, < bc>, <cc>, < bac>,<cac >, <abc>,<bbc >,<cbc>,<cabc>	< abc >, < cabc >

Final Output

Fig 13: Complete set of Sequential Patterns

6. ANALYSIS OF PROPOSED ALGORITHM

From the expected output of proposed algorithm it is clear that numbers of generated patterns are comparative less and interesting than the patterns generated from traditional PrefixSpan algorithm. Also proposed algorithm is supposed to be efficient with respect to running time, memory usage and scalability due to application of various constraints.

First constraint is applied during the transformation of database according to the concerned itemset. Then the second constraint, that is, Recency constraint is applied to the transformed database, and all the patterns in which, the first occurrence of itemset, which is the concerned itemset if do not satisfy the Recency constraint are eliminated from the database. Now since the recency constraint satisfies the downward closed property, so all the subsequences automatically satisfies this constraint so there is no need to check this constraint afterwards in mining process.

Third constraint is monetary constraint, which do not satisfy downward closure property so it is not helpful in reducing the search space as frequency and recency constraints do. So we have to check this constraint for all the sequences which are generated, by taking frequent RF-SPs as the prefix at each stage of mining process. Here the assumption is used that the

super-sequences of an RF-SP have possibilities to be RFM-SPs. Also during the construction of projected database we have to include one more field to the definition of projected database, MValue, which is the set of monetary value of a prefix in various suffixes present in the projected database of that prefix.

Therefore for each RF-SP of length L in α, we construct the projected database S | α , and along with finding RF-SPs of length L+1 we are also checking for the RFM-SPs of length L+1 from the same projected database.

7. CONCLUSION AND FUTURE EXTENSIONS

In this paper we have proposed an algorithm for discovering target-oriented RFM based Sequential patterns. Since, proposed algorithm uses the concept of constraints to find out interesting patterns, so it will be more efficient with respect to execution time as there will be less database projection and only those prefixes which satisfy recency as well as frequency constraint will get projected and even size of projected database is also reduces due to application of target oriented sequential mining so space complexity will also get reduced. This work can be extended by using fuzzy recency time constraint, so that the boundary will no longer be fixed but flexible and to add other useful constraints/attribute to the RFM patterns, for example time and location to find more informative and useful patterns.

8. REFERENCES

[1] Rakesh Agrawal Ramakrishna Srikant, "Mining Sequential Patterns", 11th Int. Conf. on Data Engineering, IEEE Computer Society Press, Taiwan, 1995 pp. 3-14.

[2] Srikant R. and Agrawal R., "Mining sequential patterns: Generalizations and performance improvements", Proceedings of the 5th International Conference Extending Database Technology, 1996, 1057, 3-17.

[3] F. Masseglia, F. Cathala, and P. Poncelet, "The PSP Approach for Mining Sequential Pattern", In Proc. 1998 European Symp. Principle of Data Mining and Knowledge Discovery (PKDD'98), Nantes, France, Sept. 1998, pp. 176–184.

[4] M. Garofalakis, R. Rastogi, and K. Shim, "SPIRIT: Sequential pattern mining with regular expression constraints", VLDB'99, 1999.

[5] Han J., Dong G., Mortazavi-Asl B., Chen Q., Dayal U., Hsu M.-C., "Freespan: Frequent pattern-projected sequential pattern mining", Proceedings 2000 Int. Conf. Knowledge Discovery and Data Mining (KDD'00), 2000, pp. 355-359.

[6] Han, J., Pei, J., Mortazavi-Asl, B. and Zhu, H., "Mining access patterns efficiently from web logs", In Proceedings of the Pacific- Asia Conference on Knowledge Discovery and Data Mining (PAKDD'00) Kyoto Japan, 2000.

[7] M. Zaki, "SPADE: An efficient algorithm for mining frequent sequences", Machine Learning, 2001.

[8] J. Pei, J. Han, B. Mortazavi-Asi, H. Pino, "PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth", ICDE'01, 2001.

[9] Helen Pinto Jiawei Han Jian Pei Ke Wang, "Multidimensional Sequential Pattern Mining", In Proc. 2001 Int. Conf. Information and Knowledge Management (CIKM'01), Atlanta, GA, Nov. 2001 pp. 81–88.

[10] AYRES, J., FLANNICK, J., GEHRKE, J., AND YIU, T., "Sequential pattern mining using a bitmap representation", In Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining-2002.

[11] Chen, Y.L., Chiang, M.C. and Ko, M.T, "Discovering time interval sequential patterns in sequence databases", Expert Syst. Appl., Vol. 25, No. 3, 2003, pp. 343–354.

[12] Yan, X., Han, J., and Afshar, R., "CloSpan: Mining closed sequential patterns in large datasets", In Third SIAM International Conference on Data Mining (SDM), San Fransico, CA, 2003, pp. 166–177.

[13] Jian Pei, Jiawei Han, Wei Wang, "Constraint-based sequential pattern mining: the pattern growth methods", J Intell Inf Syst , Vol. 28, No.2, ,2007, pp. 133 –160.

[14] NIZAR R. MABROUKEH and C. I. EZEIFE, "A Taxonomy of Sequential Pattern Mining Algorithms", ACM Computing Surveys, Vol. 43, No. 1, Article 3, Publication date: November 2010.

[15] J. Han, J. Pei, and X. Yan, StudFuzz,"Sequential Pattern Mining by Pattern-Growth: Principles and Extensions", 180, 2005, pp. 183–220.

[16] J.Pei, J.Han, B.MortazaviAsl, J.Wang, H.Pinto, Q.Chen, U.Dayal and M.-C.Hsu, "Mining sequential patterns by pattern-growth: The PrefixSpan approach", IEEE Transactions on Knowledge and Data Engineering, vol.16, no.11, 2004, pp. 1424-1440.

[17] Yen-Liang Chen, Mi-Hao Kuo, Shin-Yi Wu, Kwei Tang, "Discovering Recency, frequency, and monetary (RFM) sequential patterns from customers' purchasing data", Electronic Commerce Research and Applications 8 (2009), 2009, pp. 241–251.

[18] Hao-En Chueh, "Mining Target-Oriented Sequential Patterns with Time-Interval", International journal of computer science & information Technology (IJCSIT) Vol.2, No.4, August 2010.

[19] Yen-Liang Chen, Ya-Han Hu, "The consideration of recency and compactness in sequential pattern mining", In Proceedings of the second workshop on Knowledge Economy and Electronic Commerce, Vol. 42, Iss. 2 ,pp. 1203-1215, 2006.

[20] Ya-Han Hu, Fan Wu, "Mining Multi-level Time-interval Sequential Patterns in Sequence Databases", Chieh-I Yang, 2010.