



## SEQUENTIAL PATTERN MINING FROM MULTIPLE SEQUENCES USING RULE GROWTH AND TRULE GROWTH ALGORITHM

**T.Muthulakshmi**  
**PG Scholar**  
**Department of information**  
**technology**  
**Francis Xavier engineering college,**  
**Tirunelveli ,Tamilnadu, India,**  
**muthulakshmi1594@gmail.com**

**Dr.A.AnithaM.Tech.,Ph.D**  
**Asst professor ,**  
**Department of information**  
**technology,**  
**Francis Xavier engineering college,**  
**Tirunelveli ,Tamilnadu, India,**  
**dr.aanitha@yahoo.com**

**Abstract:** Predicting the next elements of a sequence is an important research problem with wide applications such as stock market analysis, consumer product recommendation, weather forecasting, text generation and web link recommendation. Techniques for sequence prediction can be categorized according to the types of sequences on which they are applied. In that sequential pattern mining is an important data mining task which consists of discovering subsequences that are common to multiple sequences. Existing techniques provides several drawbacks such as difference in rating of similar rules, the rules may not be found because they are individually considered uninteresting for making predictions. Mining Partially-Ordered Sequential Rules (POSR) is an important problem which is a more general form of sequential rules common to multiple sequences such that items in the antecedent and in the consequent of each rule are unordered. Here to mine POSR, RuleGrowth algorithm is introduced which is efficient and easily extendable. In particular, an extension of RuleGrowth named TRuleGrowth is also presented that accepts a sliding-window constraint to find rules occurring within a maximum amount of time. In Mining Rules can be useful to make recommendations , predictions or to analyze customers' behavior. To optimize the time required to predict the sequences by sequential pattern. Partially-ordered sequential rules(POSR), a more general form of sequential rules common to multiple sequences such that items in the antecedent and in the consequent of each rule are unordered. This definition has the benefits of summarizing several rules by single rules.

**Keywords:** Sequence Prediction, Sequential Pattern, Sequential Rule, Multiple Sequences.



## 1.INTRODUCTION

Sequential Pattern Mining is important advanced application task in data mining task. It is knowledge discovering several orders of multiple subsequences. Many algorithms are proposed for sequential rule task such as GSP, PrefixSpan, SPADE and CM - SPADE. In sequential pattern mining algorithm found misunderstanding for the user. This patterns are basis of support and confident. Support value which can occurred the percentage of sequences. Support is an indication of how frequently the items appear in database. For instance to considered the sequence pattern [violin],[piano],[clarinet] customers bought the music of violin ,piano and clarinet in that order. This sequential pattern is to give support of 50% because it appears in sequences 1,2 and 4 of the following sequence frequent pattern containing six sequences.

1. [Violin],[Guitar],[Piano],[Clarinet]
2. [Guitar],[Harmonica],[Harp],[Violin],[Piano],[Clarinet]
3. [Piano],[Violin],[Guitar],[Cellos],[Clarinet]
4. [Violin],[Guitar],[Piano],[Harmonica],[Clarinet]
5. [Guitar],[Harmonica],[Violin],[Piano]
6. [Violin],[Piano],[Guitar],[Harmonica]

However, this sequence patterns is misleading .the despite that appears in 50% of the sequences, there are two sequences [Violin],[Piano] are not followed by [Clarinet] (sequence 5 and 6). Therefore, if someone to take the basis of pattern, it could lead taking wrong decisions. A solution to this problem should add to be a measure of the confidence or probability pattern to be followed. But it is not straightforward which means it can contain a sequential pattern mining algorithm and multiple items. In Sequential Pattern mining algorithm have just not been designed for it. Alternative to consider the sequence pattern mining of based

on confidence of sequential rule mining. In sequential rule mining used to several domains those are weather observation, e-learning , reverse engineering and stock market analysis. Algorithms are designed to discover rules to appeared single sequences or common to multiple sequences. It consist of finding sequential rules form  $P \Rightarrow Q$  in sequence database such P and Q are sequential patterns. Each rules have own way of basis to support (Percentage of sequences that obtained by the rules) and confidence ( the probability of sequential patterns Q will appear counting after P). An example sequential rule are following: [Violin],[Guitar].[Piano]  $\Rightarrow$  [Clarinet]. The customers bought the music instrument Violin, Guitar and Piano in that order have then bought the music instrument Clarinet. This Sequential Rules support of 33% because it is found in two sequences in out of six sequences.(sequences 1 and 4). Sequential Pattern mining useful recommendations, to analysis customers behavior and predictions.

It is idea of mining “Partially- Ordered sequential rules” POSR Sequential rules are efficient algorithm named RuleGrowth. It use a novel approach named “rule Expansions” are generated sequential rules and includes search efficiently. RuleGrowth is easily extendable. Its extension named TRuleGrowth are finds to occurring with Sliding Window .Sequential patterns occurring within a maximum amount of time. It conduct an perform with two baseline algorithms to four real life databases customer data and language stream .In RuleGrowth outperformance situations in



terms of memory consumption and small execution time. Sequential rules are reduced by orders of magnitude sliding window constraint are used. It obtained POSR to shown prediction accuracy contain higher than sequential rules.

A sequential rule as a relationship between two sequential patterns:

**1) Different item ordering rules may have many variations.** Sequential patterns are specify a condition ordering between items, a different ordering to several rules with same items. For example 12 variations of [Violin], [Guitar], [Piano]  $\Rightarrow$  [Clarinet] with different ordered with same items such as following Rules are denoted as A1, A2, A3..... A6:

A1: [Violin], [Guitar], [Piano]  $\Rightarrow$  [Clarinet],  
A2: [Guitar], [Violin], [Piano]  $\Rightarrow$  [Clarinet],  
A3: [Piano], [Violin], [Guitar]  $\Rightarrow$  [Clarinet],  
A4: [Piano], [Violin], [Guitar]  $\Rightarrow$  [Clarinet],  
A5: [Piano], [Violin], [Guitar]  $\Rightarrow$  [Clarinet],  
A6: [Piano], [Violin], [Guitar]  $\Rightarrow$  [Clarinet],

But all variations are describes to same situation in sequential patterns. In customer bought musical instrument in different ordered, then bought music from Clarinet).

**2) Rules and Variations have Difference in how they are rated by the algorithm :** For example, rules A1, A2 and A3 respectively support and confidence of 33% and 100%, 16% and 50% and 16% and 100% A4, A5 and A6 have do not appear in the database. It means the same rules are rated in a wrong impression of sequential relationships are contained in database to the user. Support and confidence could much higher. For

example its non of rules in previous has a support are higher than 33% . But it appears in four sequences out of the six 66%.

**3) Rules are likely less to be useful.** In sequential rules are less likely with make a new predictions. For example To consider a new customer purchase [Violin], [Piano], [Guitar] in that order. Then previous rules are match sequences are predict customer have been purchase [Clarinet] next. A problem would be choose between rules A1, A2, and A3 because they are rated to quite differently (Support varies from 16% to 33% then confidence from 50% to 100%).

## II. RELATED WORKS

In this section , A review of multiple sequences on sequential pattern mining using Data mining algorithms. Sequential pattern mining takes

Kamsu – Foguem.B, Rigal.F and Mauget.F , presented a paper entitled “Mining association rules for the quality improvement of the production process,” this provides knowledge about Association rule mining obtain a data mining techniques used to find out useful and invaluable information from huge database. The suggested future works were develops a better conceptual base for improving the applications of association rule mining methods to extract knowledge on operations and information managements. It is a powerful tool to emulating cognitive of data mining method interactive of intelligent architecture process of human analysis[1].

Bogon.T, Timm.I.J, Lattner.A.D, Paraskevopoulos.D, Jessen.U, Schmitz.M, Wenzel.S, presented a paper entitled “Towards





Assisted Input and Output Data Analysis in Manufacturing Simulation: The EDASI Approach”this provide combination of assistance functionalities for input and output data analysis, the suggested future assistance system functionalities must have works EDASim. EDASim is used to developed tool that focuses on supporting the user in selection, validation and preparation of input data as well as to assist analysis of output data.

Fournier-Viger.P, Gomariz.A, Campos.M and Thomas.R, presented a paper entitled “Fast Vertical Sequential Pattern Mining Using Co-occurrence Information,” This structure explains how CMAP can be used to prune candidates in three state of the vertical algorithms namely SPADE, SPAM and ClaSP. New structured named CMAP(Co-occurrence MAP) for storing Co-Occurrence information is used. This work must needs to develop additional optimizations and also integrate them in sequential rule mining.

Fournier-Viger.P and Tseng.V.S, presented a paper entitled “TNS: Mining Top-K Non-Redundant Sequential Rules,” this provide the Mining sequential rules from sequence database is important research problem in advanced wide applications. Mining Top -K Non redundant sequential Rules is used. This work to improve Top-K sequential rules often contain a large proportion of redundant rules that are provide information about conditional rules.

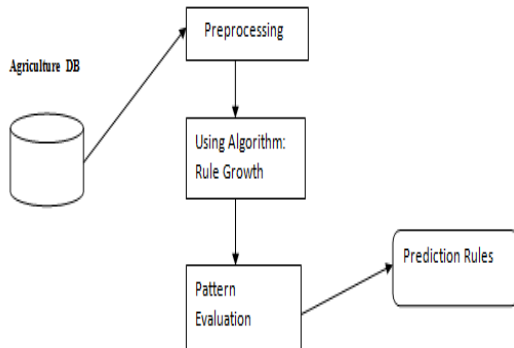
Fournier-Viger.P, Faghihi.U, Nkambou.R and Mephu.E presented a paper entitled “CMRules: An Efficient Algorithm for Mining Sequential Rules Common to Several Sequences,” this paper provide CMRules algorithm for data mining sequential rules common to many sequences in sequence databases- not for mining rules appearing frequently in sequences. This algorithm does

not use a sliding concept it works to develop performance depends on the number of the association rules. It this set is large CMRules becomes to efficient sequences.

This paper is organized as follows the sequential pattern rules are make prediction and accuracy to sequence of items. These are common to multiple itemset with different ordering to handled various algorithms.this sections are discusses and analysis about the resulta and memory consumption.

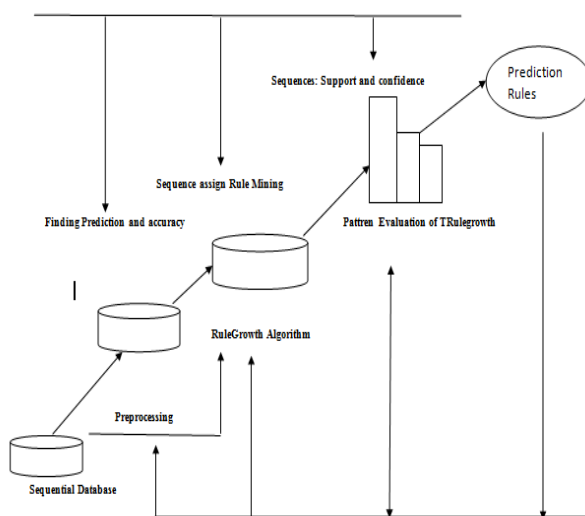
### III.PROPOSED SCHEME

The main scheme it is systematical relevant work. In mining sequential rules are common to several sequences has been proposed to data mining common to multiple sequences in the frequent itemset. Sequential pattern mining appear in a sequences database such that itemsets support is no less then threshold value of named minsup these are set by the user.Subsequences are defined as the number of sequences that contain it divided by total number of sequences.RuleGrowth algorithm takes as input to database S mainly depends on thresholds minsup and minconf .Rules occur support of  $\{P\} \Rightarrow \{Q\}$  obtained by IRI Algorithm consist of apply a sequential pattern mining algorithm and then to perform post and pre processing steps are generated by a rules between the couple of sequential patterns. Mainly support is less than minsupitemsets, EXPANDLEFT and EXPANDRIGHT . Rule size is  $1 * 1$  .To develop two procedures



**Figure 1 Block Diagram of Database**

The sequential pattern rules are made prediction and accuracy to sequence of items. These are common to multiple itemset with different ordering to handle various algorithms. This section discusses and analyzes the results and memory consumption.



**Figure 2 Block Diagram of POSR**

## IV METHODOLOGY

### 1. SEQUENCE DATABASE CREATION

A sequence database discovers this form of two-step performance: those are minimum percent to the sequential pattern mining. First to appear the discovering itemsets gives minimum percentage of the windows and couple of itemsets using to generate Sequential pattern rules. A sliding window is assumed to move from the beginning of a sequence to its end, one itemset at a time or one time unit. A sequence database (SD) is a set of sequences created  $D = \{d_1, d_2, d_3, \dots, d_n\}$  and a set of items  $T = \{t_1, t_2, t_3, \dots, t_n\}$  occurring in these sequences is assigned a unique SID (Sequence ID). A sequence is an ordered list of itemsets (set of items)  $dx = \{T_1, T_2, T_3, \dots, T_n\}$  such that  $T_1, T_2, T_3, \dots, T_n \subseteq T$ . For example, each single symbol represents an item. Brackets represent an itemset. Seq 1 means U and V the sequence occurred at the same time. A sequence of itemsets annotated with a timestamp is called a time sequence. A time sequence containing the window itemsets. A sliding window is a group of consecutive sequences of itemsets to assume to move from the beginning of sequences to the end of itemsets. For example, a length of 3 time units can move 16 different positions  $g_1, g_2, g_3, \dots, g_n$  for sequence ordered. It extends outside sequences; each item appears in the same number of windows.

### 2. OPTIMIZING RULEGROWTH ALGORITHM

The RuleGrowth algorithm uses candidate tests, generates candidate rules, and scans the large database to determine their threshold value of support and confidence. The main problem is that it generates a large amount of candidate rules and first



finds rule of size 1\*1 to recursively grows and scanning the sequences containing to find single items that can expand their left or right sides. There are two process expanding

**Left Expansion:** A left expansion is the process of adding an item I to the left side of a rule  $A \Rightarrow B$  to obtained a large rule  $A \cup \{i\} \Rightarrow B$ .

**left expansion** =  $|sid(A \cup \{i\} \Rightarrow B)| / |S|$

**Right Expansion:** A right expansion is the process of adding an item I to the right side of a rule  $A \Rightarrow B$  obtained a large rule  $A \Rightarrow B \cup \{i\}$

**Right expansion** =  $|sid(A \Rightarrow B \cup \{i\})| / |S|$

**RuleGrowth** it is a homonym process of CMDeo. It finds large rules by adding couple of items to rule by scanning the sequences the itemsets containing the rule which means a depth first search and CMDeo combines couple of rules to generate candidates are a breadth first search. RuleGrowth it keeps track of the first and last occurrence of the itemsets of antecedents and consequents to avoid scanning sequences.

while verifying they occur in a sliding –windows. Discovering rules appearing in a sliding window has a several important benefits. It can decrease the execution time by several orders of magnitude by pruning the search space. TRuleGrowth can produces a much smaller set of rules, thus reducing the disk space requirements for storing rules found and making it easier for the user to analyze result. Setting a window constraint can increase prediction and accuracy when rules are used for prediction.

TRuleGrowth algorithm reduce a detection time limit of 2,000 seconds was set and a maximum to memory usage of 1GB. RuleGrowth is faster and use less memory than further development of CMRuules and CMDeo. A minsup is a set of lower, the gap increase that performance. The efficiency of TRulegrowth algorithm can analyze the sequence from the database to reduce the execution times.

### 3. PREDICTING SUPPORT AND CONFIDENCE

Predicting to support of

u m b e r o f	SID	Sequences
	Seq 1	{u,v},{w},{x},{y},{z}
	Seq 2	{u,t},{w},{v},{u,v,z,x}
	Seq 3	{u},{v},{x},{z}
	Seq 4	{v},{x,y,s}

sequences that to contains it divided by the total number of sequences. It indication of how frequently the items appear in the database.

### 3.OP

### 3. OPTIMIZING TRULEGROWTH ALGORITHM

TRuleGrowth algorithm is a modified extension version of RuleGrowth that discovers rules and



$$\text{Support} = |\text{sid}(A \Rightarrow B)| / |S|$$

The support value first occurrence of an itemsets A in a sequence  $S=T_1, T_2, T_3, \dots, T_n$  its itemset  $T_k \in S$  such that  $A \cup I$ . last occurrence of an itemset A in a sequence  $s = T_1, T_2, \dots, T_n$  in itemset  $T_k \in S$ . sequences are  $\{u, x\}, \{v\}, \{u\}, \{v\}, \{y\}$  is second itemset, the last occurrence of (u,v) is the third itemsets.

Predicting to confidence number of windows sequences that can containing the rules divided by the number of windows rules that found to be obtained to be true the value respect to a set of transactions.

$$\text{Confidence} = |\text{sid}(A \Rightarrow B)| / |\text{sid}(A)|$$

The confidence value first occurrence of an itemsets A is a sequence  $S=T_1, T_2, T_3, \dots, T_n$  its itemset  $T_k \in S$  such that  $A \cup \text{sid}(B)$ . last occurrence of an itemsets A in a sequence  $s = T_1, T_2, \dots, T_n$  in itemset  $T_k \in M$ . sequences are  $\{u, x\}, \{v\}, \{u\}, \{v\}, \{y\}$  is second itemsets, the last occurrence of (u, v) is the third itemsets.

#### IV.RESULTS AND DISCUSSION

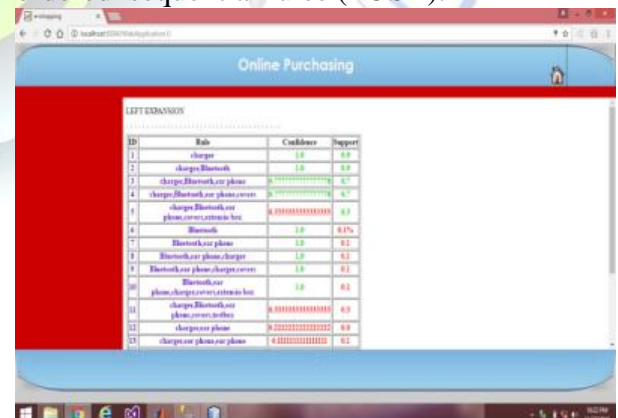
Basically this paper presented two main algorithms. RuleGrowth is a datamining sequential rules that are common to multiple sequences. In Previous that algorithm present the CMRules and CMDeo and it use a pattern – growth approach for discover

the valid database sequences such avoid unwanted not appearing in the large database.



Fig 1. Maintain Admin

The second algorithm TRuleGrowth allows to user specific a sliding window and expansion of the comparison to that common database . Moreover this experiments shows that execution time and number of valid rules found can be many magnitude ordered reduce the scanning unwanted occurred rules. reported result from real life time applications using partially – ordered sequential rules (POSR).



ID	Rule	Confidence	Support
1	charge	1.0	0.0
2	charge,Bluetooth	1.0	0.0
3	charge,Bluetooth,car phone	0.5	0.0
4	charge,Bluetooth,car phone,carrent	0.5	0.0
5	charge,Bluetooth,car phone,carrent,car rent	0.5	0.0
6	Bluetooth	1.0	0.17%
7	Bluetooth,car phone	1.0	0.1
8	Bluetooth,car phone,charge	1.0	0.1
9	Bluetooth,car phone,charge,carrent	1.0	0.1
10	Bluetooth,car phone,charge,carrent,car rent	1.0	0.1
11	charge,Bluetooth,car phone,carrent,car rent	0.5	0.0
12	charge,car phone	0.5	0.0
13	charge,car phone,car phone	0.5	0.1

Fig 2. Left Expansion

In Left expansion the threshold value will be check and to declared above 1 to display in green color and below 1 show the result value can display red color.



Fig 3.Right Expansion

In Right expansion that set of items to calculate threshold value of 1.2 check and below values are the comparing support and confidence value.that POSR can provide a much higher prediction accuracy than regular sequential rules for sequence prediction.

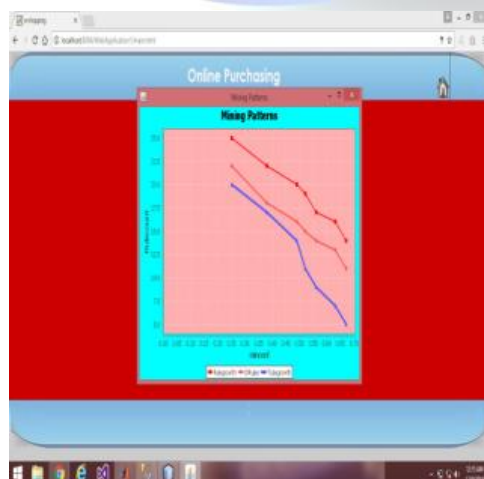


Fig 4.scalability experiment

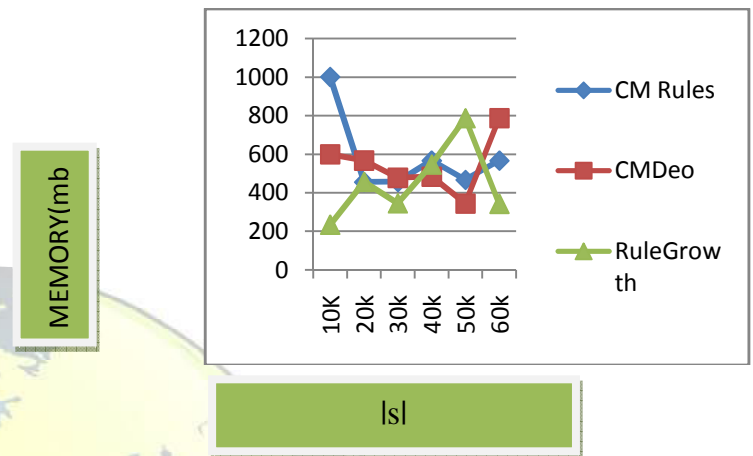


Fig 5. POSR Analyze plot

## V. CONCLUSION

In future, AprioriHC algorithm with MINEX is used which is the algorithm used for sequence pattern mining. Apriori is designed to operate association rule learning over transactional databases for mining sequential rules. The information of high utility sequential itemsets is maintained in a special data structure such that the candidate itemsets can be generated efficiently with only two scans of the database. In future Results from a real application showing that POSR can provide a much higher prediction accuracy than regular sequential rules for sequence





prediction. The performance of AprioriHC algorithm with MINEX was evaluated in comparison with the state-of-the-art algorithms.

## REFERENCE

- [1] R. Agrawal, T. Imielinski and A. Swami, "Mining Association Rules Between Sets of Items in Large Databases," *Proc. 13th ACM SIGMOD Intern. Conf. on Management of Data*, pp. 207-216, 1993.
- [2] R. Agrawal, R. Srikant, "Mining Sequential Patterns," *Proc. 11th Intern. Conf. on Data Eng.*, pp. 3-14, 1995.
- [3] D.W. Cheung, J. Han, V. Ng. and Y. Wong, "Maintenance of discovered association rules in large databases: An incremental updating technique," *Proc. 12th Intern. Conf. on Data Eng.*, pp. 106-114, 1996.
- [4] G. Das, K.-I. Lin, H. Mannila, G. Renganathan and P. Smyth, "Rule Discovery from Time Series," *Proc. 4th ACM Intern. Conf. Know. Discovery and Data Mining*, pp. 16-22, 1998.
- [5] J.S. Deogun and L. Jiang, "Prediction Mining – An Approach to Mining Association Rules for Prediction," *Proc. 10<sup>th</sup> Intern. Conf. Rough Sets, Fuzzy Sets, Data Mining, and Granular Comp.*, pp. 98-108, 2005.
- [6] U. Faghihi, P. Fournier-Viger and R. Nkambou, "A Computational Model for Causal Learning in Cognitive Agents," *Knowledge Based Systems*, vol. 30, pp. 48-56, 2012.
- [7] P. Fournier-Viger, U. Faghihi, R. Nkambou and E. MephuNguifo, "CMRules: An Efficient Algorithm for Mining Sequential Rules Common to Several Sequences," *Knowledge Based Systems*, vol. 25, no. 1, pp. 63-76, 2012.
- [8] J.H. Hamilton and K. Karimi, "The TIMERS II Algorithm for the Discovery of Causality," *Proc. 9th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 744-750, 2005.
- [9] S.K. Harms, J. Deogun and T. Tadesse, "Discovering Sequential Association Rules with Constraints and Time Lags in Multiple Sequences," *Proc. 13th Intern. Symp. Method. Intell. Systems*, pp. 373-376., 2002.
- [10] I. Jonassen, J.F. Collins and D.G. Higgin, "Finding flexible patterns in unaligned protein sequences," *Protein Science*, vol. 4, no. 8, pp. 1587-1595, 1995.
- [11] S. Laxman and P. Sastry, "A survey of temporal data mining," *Sadhana*, vol. 3, pp. 173-198, 2006.
- [12] D. Lo, S.-C. Khoo and L. Wong, "Non-redundant sequential rules – Theory and algorithm," *Inform. Syst.*, vol. 34, no. 4-5, pp. 438-453, 2009.



[13] H. Mannila, H. Toivonen and A.I. Verkano, "Discovery of frequent episodes in event sequences," *Data Mining and Knowledge Discovery*, vol. 1, no. 3, pp. 259-289, 1999.

[14] J. Pei, J. Han et al., "Mining Sequential Patterns by Pattern-Growth: The PrefixSpan Approach," *IEEE Trans. Knowledge and Data Eng.*, vol. 16, no. 10, pp. 1-17, 2004.

[15] P. Fournier-Viger, Knowledge discovery in problem-solving learning activities, Ph.D. Thesis, Univ. Quebec in Montreal, Montreal, 2010.

[16] Y.L. Hsieh, D.-L. Yang and J. Wu, "Using Data Mining to Study Upstream and Downstream Causal Relationship in Stock Market," *Proc. 2006 Joint Conf. Inf. Sc.*, 2006.

[17] M. J. Zaki, "SPADE: An Efficient Algorithm for Mining Frequent Sequences," *Machine Learning*, vol. 42, no. 1-2, pp. 31-60, 2001.