# A State of The Art in Sequential Pattern Mining

**Sonam Jain[1] Dr. Rajeev G. Vishwakarma[2]**
[1, 2] Head & Professor
[1, 2] Computer Science Engineering, SVITS Indore

*Abstract*— Sequential rule mining or sequential pattern mining is used in many areas like, stock market analysis, rain fall analysis, etc. it is a heart favorite topic of research for many people over the years. This paper presents a comprehensive survey of some modern, also popular techniques for the sequential pattern mining. An overview of sequential pattern mining is also available.

## I. INTRODUCTION

Data mining [1,2] is concerned with analysis of large volumes of data to automatically discover interesting regularities or relationships which in turn leads to better understanding of the underlying processes. The primary goal is to discover uncover patterns, unpredictable trends in the data. Data mining activities uses combination of techniques from database technologies, artificial intelligence, machine learning and statistics. The term is frequently misused to mean any form of large-scale data or information processing. The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns.

Data mining [3] is the process of extracting hidden patterns from data set. As large amount of data is gathered, with the amount of data doubling every three or four years, data mining is becoming an increasingly important tool to transform this data or data set into knowledge. It is mostly used in a wide range of applications, such as marketing, scientific discovery and fraud detection. Data mining can be used to data sets of any size, and while it can be used to discover hidden patterns, it cannot discover patterns which are not already present in the data set. Knowledge Discovery in Databases (KDD) is an automated extraction of novel, understandable and potentially useful patterns implicitly stored in hugs databases, data warehouse and other massive information storehouse. KDD is a multi-disciplinary field drawing work from areas including database technology, high performance computing, artificial intelligence, machine learning, neural networks, statistics, pattern recognition, information retrieval, and data Sequential pattern discovery finds temporal associations so that not only closely correlated sets but also their relationships in time are uncovered.

Frequent item sets [1, 2] and association rules focus on transactions and the items that appear there. Databases of transactions usually have temporal information. Sequential pattern or sequential rules exploit this temporal information. Example data:

- Market basket transactions
- Web server logs
- Tweets
- Workflow production logs

| Object | Timestamp | Events |
|---|---|---|
| A | 10 | 2, 3, 5 |
| A | 20 | 6, 1 |
| A | 23 | 1 |
| B | 11 | 4, 5, 6 |
| B | 17 | 2 |
| B | 21 | 7, 8, 1, 2 |
| B | 28 | 1, 6 |
| C | 14 | 1, 8, 7 |

Table: 1 A Sequence Data Base

### A. *Formal Definition of a Sequence*

A sequence is an ordered list of elements (transactions). Each element contains a collection of events (items). Each element is attributed to a specific time or location. Length of a sequence, |s|, is given by the number of elements of the sequence

| ID | Sequences |
|---|---|
| 1 | {1,2},{3},{6},{7},{5} |
| 2 | {1,4},{3},{2},(1,2,5,6} |
| 3 | {1},{2},{6},{5},{6,7} |
| 4 | {2},{6,7},{1,2},{2,3} |

Fig. 1: A Sequence Database

Considering a minimum support = 50% and minimum confidence = 50%, we get following sequential rules

| Id | Sequential Rule | Support | Confidence |
|---|---|---|---|
| 1 | {1,2,3} => {5} | .5 | 1.0 |
| 2 | {1} => {3,5,6} | .5 | .66 |
| 3 | {1,2} => {5,6} | .75 | .75 |
| 4 | {2} => {5,6} | .75 | .75 |
| 5 | {1} => {5,6} | .5 | .5 |
| .. | .. | .. | .. |

Fig. 2: SEQUENTIAL RULES

## II. A SURVEY OF SEQUENTIAL PATTERN MINING METHODS

As we know, data are changing all the time; especially data on the web are highly dynamic. As time passes by, new datasets are inserted; old datasets are deleted while some other datasets are refreshed. It is transparent that time stamp is an important attribute of each dataset, also it's aristocratic in the process of data mining and it can give us more accurate and useful information. For example, association rule mining does not take the time stamp in account, the rule may Buy A⇒Buy B. If we take time stamp in account then we can get more accurate and useful rules such as: Buy A implies Buy B within two days, three days four days or a

 

week and a month, or usually people Buy An everyday in a week. The second kind of rules, business decision can be more accurate and useful prediction and consequently make more sound decisions.

A database consists of sequences of values or events that change with time are called a time-series database [3], a time-series (change with time) database records the valid or useful time of each data set. For example, in a time-series database that records the sales transaction of a d-mart, supermarket, each and every transaction includes an additional attribute and features that indicate when the transaction happened. Time-series data in database is used whenever, to store historical data in a diversity of areas such as, financial data, medical data, scientifically data. Different mining technique has been designed for mining time-series data [3].

Contrarily to these works that discover rules in a single sequence of events, a few works have been designed for mining sequential rules in several sequences [4]. For example, Das et al. [4] discovers rules where the left part of a rule can have more events, yet still the right part has to contain a single event. This is not a simple limitation, as in general life applications, sequential relationship may contain several events. Moreover, the algorithm of Das et al. [4] is highly inefficient as it tests all the rules which are possible, without any strategy for cutting and pruning the search space. To our knowledge, only the algorithm of Harms et al. [5] discovers sequential rules from sequence databases, and does not restrict the number of events contained in each and every rule. It is searching for rules with the confidence and the support higher or equal to user-specified thresholds. The support of a rule is here defined as the number of times that the right part occurs after the left part within user-defined time windows

However, one important limitation of the algorithms of Das et al., [3] and Harms et al. [4] comes from the fact that they are designed for mining rules occurring frequently in sequences. As a consequence, these algorithms are inadequate for discovering rules common to many sequences. We illustrate this with an example. Let us assume a sequence database sequence corresponds to a customer and each event represents the items bought during a particular day. Assume we are required to mine sequential rules that are common to many customers. The algorithms of Das et al. [3] and Harms et al. [4] are inappropriate since a rule that appears many times in the same sequence could have a high support even if it does not appear in any other sequences. A second example is the application domain of this paper. We have built an intelligent tutoring agent that records a sequence of events for each of its executions. We wish that the tutoring agent discovers sequential rules between events, common to several of its executions, so that the agent can thereafter use the rules for prediction during its following execution.

In order to reduce the number of iterations, the efficient bi-directional sequential pattern mining approach namely Recursive Prefix Suffix Pattern detection, RPSP [7] algorithm is furnished. The RPSP algorithm finds first all Frequent Itemsets (FI‟s) according to the given minimum support and transforms the database such that each transaction is replaced by all the FI‟s it contains and then finds the patterns. Further the pattern detected based on ith projected databases, and builds suffix and prefix databases based on the Apriori properties. Recursive Prefix Suffix Pattern will increase the number of frequent patterns by reducing the minimum support and vice versa. Recursion gets deleted when the detected FI set of prefix or suffix assigned database of parent database is ineffective. All patterns that correlate to a particular ith proposition database of transformed database that formed into a set that is dis-joint from all the other sets. The resultant set of frequent patterns is the sum of the all disjoint subsets. The proposed algorithm tested on hypothetical and sequence data and obtained results were found all satisfactory. Hence, RPSP algorithm may be applicable to many real world sequential data sets.

## III. CONCLUSION

We have performed a systematic study on mining of sequential patterns in large databases. The concept of sequential pattern mining is also elaborated. It is found that the most of the existing sequential pattern mining methods are based on the concept of generate and test. So there is scope for the new work on the pattern growth based approach.

## REFERENCES

[1] Tan, kumar “introduction to data mining”.
[2] Arun Polari “ Introduction to data mining”
[3] Han and Kamber, 2000
[4] Das. G., Lin, K.-I., Mannila , “ Rule Discovery from Time Series” . 4th Int. Conf. on Knowledge Discovery and Data Mining (New York, USA, August 27-31, 1998), 16-22.
[5] Harms, S. K. Discovering Sequential Association Rules with Constraints and Time Lags in Multiple Sequences. Int. Symp. On Methodologies for Intelligent Systems (Lyon, France, June 27-29, 2002), pp . . . . 373-376.
[6] Mannila, “A.I. Discovery of frequent episodes in event sequences. Data Mining and Knowledge Discovery”, 1, 1 (1997), 259-289
[7] Dr P padmaja, P Naga Jyoti, m Bhargava “Recursive Prefix Suffix Pattern Detection Approach for Mining Sequential Patterns” IJCA September 2011

 