# A Survey on Algorithms for Sequential Pattern Mining

[1]Minubhai Chaudhari, [2]Chirag Mehta,
Professor, Post Graduate Student
Govt. Engg. College, Gandhinagar, Dept. of Computer Engg., Gujarat Technological University

***Abstract* - Sequential pattern mining is a very useful mining technique for various sectors like healthcare, retail business, DNA analysis etc. It generates patterns which are frequently occurring in given sequence of transactions. It uses sequence database having sequence of transactions with transaction time. In sequence database every transaction is having various items. By sequential pattern mining user wants frequent patterns which are generating according to given constraint. GSP, SPADE, SPAM and Prefix span are few efficient sequential pattern mining algorithms. In this survey various algorithms(GSP,SPADE,SPIRIT,SPAM,CLO-SPAM,CMDS,FREESPAN,WAP-MINE & PREFIX SPAN) are studied for sequential pattern mining. This survey found pros and cons for each algorithm in various scenario and other factors.**

***Keywords* - GSP, SPADE, SPIRIT, SPAM, CLO-SPAM, CMDS, FREESPAN, WAP-MINE & PREFIX SPAN**

## I. INTRODUCTION

Sequential pattern mining is a extension of association rule mining [3]. This concept is being proposed in 1995, has gone through big advancement in few years only. algorithms by using different data structure or different representation. It is used in wide range of real-life problems. This mining algorithm finds the presence of frequent sequences in the given database [1]. The database for this algorithm is set of sequences called as data-sequences. Every data-sequence is a list of customer transaction, and every transaction is a set of items. There is transaction time related with the each transaction in the sequence database. The difference between sequential pattern mining and association rule mining is events are linked with time. The sequential pattern mining find the relation between the different transactions, but in the association rule mining it finds the relationship of items in the same transaction.

In association rule mining, it finds which different items are brought with each other frequently, all these items must have brought under same transaction. But in sequential pattern mining, it finds which items are brought in a particular order by a single customer, those items come from various transactions. The sequential pattern mining is very useful for the marketing manager to decide which item is brought one after another in sequence by particular customer. Sequential Pattern Mining is discovering the whole set of frequent subsequence in the set of sequential transactional database. The resultant pattern discovered after mining is the sequence of item sets that normally found frequent in specific order. In a single transaction all items have the same transaction time. Every sequence is the ordered list of the different transactions and every transaction in it is a set of the items. The ordering of the transaction in a sequence is induced by the absolute timestamps associated with that transaction. The process of discovering sequential pattern from sequence transaction database is explained below-

## II. BRIEF OVERVIEW OF VARIOUS SEQUENTIAL PATTERN MINING ALGORITHMS

***GSP (Generalized Sequential Pattern)-*** algorithms is proposed by Agrawal and Shrikant [5]. It makes the multiple passes on the data. This algorithm is more efficient than the Apriori algorithm. There are two steps in GSP algorithm,

(i) candidate generation
(ii)candidate pruning method.

The algorithm is not a main memory algorithm, it generates only candidates which are fit in memory and the support of the candidate is decided by scanning the dataset. Frequent Sequences from these candidates are store to disk and the candidates without minimum support are removed. This same task is repeated till every candidate has been counted.

This algorithm has a very fine scale up properties with respect to the number of transaction per data sequence and number of items per transaction. But this algorithm is less efficient where the mining in large sequencing of databases having high no. of pattern as the length of each candidates increases by one at every database scan.

**SPIRIT -** The basic concept of this algorithm is to utilize the regular expression at flexible tool for the constraint specifications [2]. It gives the generic user specified regular expression constraint on the mined pattern, for providing the hard restriction. There are many versions of this algorithm. To select the regular expression as a constraint specification tool is decided on the basic of two important factors.

1. The regular expression is simple form and natural syntax for specification of group of sequential pattern
2. it has the more power for specifying big range of interesting pattern constraints

**SPADE -** Like horizontal formulating methods (GSP) the sequential dataset can be converted into a vertical dataset format having

item id-lists [5]. The List of vertical dataset is the list of sequential-id & timestamps pair indicating the occurring timestamps of the item in that sequence. Searching in format of dataset is done through the id-list interaction, this SPADE a algorithm conclude the mining in total three passes of database scanning. Apart from this the computation time requires to convert in the horizontal dataset to vertical dataset and also require additional storage space several times larger than that of the original sequence database.

**SPAM -** SPAM includes the ideas of GSP, SPADE, and FreeSpan [6]. This algorithm utilizes the vertical bitmap data structure representation of database which is same as id-list of SPADE. The complete algorithm with its data structure fits in the main memory. To increase performance, SPAM use the depth-first traversal fashion. SPAM is like SPADE, but it uses the bitwise operations instead of the regular and temporal join when the comparison of SPAM and SPADE is consider the SPAM is outperform more than SPADE, while the SPADE algorithm is more SPACE-efficient than SPAM.

**CloSpan-** CloSpan - Closed Sequential Pattern Mining algorithm only mines the frequent closed sub sequences[6], containing no super-sequences with the same support during mining long frequent sequence. The performance of algorithms degrades dramatically. This algorithm creates less sequences than the other algorithms.

**CMDS -** Closed Multidimensional Pattern Mining joins method of closed- item set pattern mining and closed sequential pattern mining [6]. It is having mainly two steps-

1. Combination of closed sequential pattern mining and closed item set pattern mining.
2. Removal of duplicate pattern.

The number of pattern in CMDS is fewer than the number of pattern in multidimensional pattern mining. The set of CMDS pattern can include the set of MDS pattern.

**FREESPAN-** The freespan algorithm decreases the cost require to candidate generation and testing of apriori, with satisfying its basic feature [4]. So the freespan algorithm uses the frequent items to iteratively project the sequence database into projected database while increasing subsequence's frequently in each projected dataset. Every projection separates the database and confines further testing to progressively smaller and more manageable units. The important issue is to considerable amount of sequences can appear in more than single projected database and the size of database decreases by every iteration.

**WAP-MINE-** is pattern-growth based algorithm with tree-structure mining technique on its WAP-tree data structure. In this algorithm the sequence database is scanned two times to build up the WAP-tree from the frequent sequences from their support values. In this algorithm header table is maintained first to point that where is first presence of the every item in a frequent item set which can be helpful to mine the tree for frequent sequences built up on their suffix. It is found that during analysis the WAP-MINE algorithm have more scalability than GSP and perform bitterly by marginal points. Though this algorithm scans the database two times only and avoids the problem of generating large candidate as in case of apriori-based approach, the WAP-MINE faces the problem of memory consumption, as it iteratively regenerate n increase automatically.

**PrefixSpan-** The PrefixSpan (Prefix Projected Sequential pattern Mining ) algorithms proposed by Jian Pei, Jiavei Han and Helen Pinto [4] is the only projection based algorithms in all the sequencing pattern mining algorithms. It is more efficient than the algorithm like apriori, freespan, SPADE. This algorithm discovers the frequent items by scanning the sequence database once. The database is projected into many smaller databases according to the frequent items. By recursively growing subsequence fragment in every projected database, It found the complete set of sequential pattern. The main idea behind the prefixspan algorithm to successfully discovered patterns is employing the divide-and-conquer strategy. The prefixspan algorithm wants high memory space as compare to the other algorithms in the sense that it requires creation and processing of large number of projected sub-databases.

## III. LITERATURE REVIEW

In the sequential pattern mining concept there are many proposal presented in literature till now. In which few are constraint based sequence pattern mining and few are incremental sequential pattern mining. The study and review of some latest researches related to the incremental sequential pattern mining is presented here. In past, the improving the concept of incremental mining with constraint-based pattern mining is very important issue for real life application.

Chi-Yao Tseng [9] have proposed general model for sequential pattern with the changing database, while the data in the database can be fixed, added or deleted. They also presented the progressive algorithm named PISA which stands for Progressive mining of Sequential pattern which find the sequential pattern in fixed time interest in progressive manner. The time period of interest is the time period continuously moving forward with time goes by. In PISA algorithm, to efficiently maintain the recent data sequences it uses a progressive sequence tree. It finds outs the whole set of up-to-date sequential pattern and remove obsolete data and pattern as per requirement. The size of the sequential pattern tree created was depending on the length of the period of the time window.

Ching-Yao Wang [8] has proposed an algorithm for sequential pattern mining based on the incremental mining concept. This algorithm utilizes the concept of Pre-Large sequence to avoid the need for rescanning the original databases. After applying the lower support and upper support it defines the Pre-Large sequence that works as gap to resist the movement of sequence from large to small and from small to large. This algorithm does not perform the rescanning of the database until the new customer sequence is created. Database rescanning grow with its size, Vincent Shin-Mu Tseng [7] have proposed the rule growth method for mining the sequential rules same for many sequences. apart from the other algorithms rule growth is based on the pattern-

growth approach for finding sequential pattern rules such that it can be better and scalable. They tested rule growth with other algorithm on the public datasets. They found that the rule growth outperforms the other algorithms, for these datasets under low support and fixed threshold.

Jiaxin Liu [10] have proposed a data storage structure, called as frequency sequence tree, and gives the generation method for the frequent sequence tree called FST. At the root node of this frequent sequence tree stored the support for frequent sequence tree and the path from the node to the any outer node represents a sequential pattern in the database. The sequential pattern whose support matches the frequent sequence tree support threshold is stored in frequent sequence tree, so when the support changed, the algorithm which uses FST as the storage structure could find the entire sequential pattern without mining the whole original database.

Jiaxin Liu [11] have proposed that the structure of sequence tree based on the projected database, called as sequence tree, for the construction of this sequence tree they proposed steeps algorithm. Sequence Tree is structure of data storage. It is same in structure to the prefix tree. But, it stores all the sequence in the original database. The path from the root node to any leaf node is a sequence in the database. The structure of the sequence tree make it favorable for the increment pattern mining. Experiments showed that the increment mining method of sequential pattern which uses the sequence tree as the storage structure for sequence pattern performed best than the prefix span in memory use cost on condition that support threshold must be smaller. To take the changing nature of data addition and removal.

## IV. CONCLUSION

In this survey paper describes what is sequential pattern mining and various types of their algorithms. So, on the basis of these problems the sequential pattern mining is divided into two main groups, Apriori approach based algorithms and pattern growth approach based algorithms. From comparative study it found that sequential pattern mining algorithms which are based on the approach of pattern growth are better in terms of scalability , time-complexity and space-complexity. Both FreeSpan and PrefixSpan improve Apriori-based methods by only checking the relevant candidate in the projected databases. One dimerit of FreeSpan is that it may not reduce the length of the data sequence during projecting. PrefixSpan improves FreeSpan by removing prefix which reduces the data sequence length. Also, PrefixSpan enhances performance by reducing the combinations of items. One problem of PrefixSpan is that it cannot remove any frequent item in the postfix of a data sequence while projecting.

## V. REFERENCES


[1] R. Agrawal and R. Srikant, “Mining Sequential Patterns”, In Proceedings of the 11th International Conference on Data Engineering, pp. 3-14, Taipei, Taiwan, 1995

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

[3] J. Han and M. Kamber, “Data Mining: Concepts and Techniques”, Morgan Kaufman publishers, 2001.

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

[5] Mohammad J. Zaki,- SPADE: An Efficient Algorithm for Mining Frequent Sequences, Kluwer Academic Publisher. Machine Learning,42,31-60,2001

[6] C.-C. Yu and Y.-L. Chen, “Mining Sequential Patterns from Multi-Dimensional Sequence Data”, IEEE Trans. Knowledge and Data Eng., Vol. 17, No. 1, pp. 136-140, Jan. 2005.

[7] 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

[8] Jen-Wei Huang, Chi-Yao Tseng, Jian-Chih Ou, Ming- Syan Chen, "A General Model for Sequential Pattern Mining with a Progressive Database," IEEE Transactions on Knowledge and Data Engineering, vol. 20, No. 9, pp. 1153-1167, 2008

[9] Tzung-Pei, Hong,Ching-Yao Wang and Shian-Shyong Tseng, “An Incremental Mining Algorithm for Maintaining Sequential Patterns Using Pre-large Sequences,” Journal Expert Systems with Applications, Vol. 38, Issue 6,p p.7051-7058, 2011.

[10] Philippe Fournier,Viger,Roger Nkambou and Vincent Shin-Mu Tseng, “RuleGrowth: Mining Sequential Rules Common to Several Sequences by Pattern-Growth,” Symposium on Applied Computing, pp . 951-960, 2011.

[11] V. Chandra Shekhar Rao and P.Sammulal,"Survey On Sequential Pattern Mining Algorithms". International Journal of computer application(0975-8887),Vol 76-No.12, August 2013