A Review of Sequential Pattern Mining

[1] Priti K. Patel, [2] Smit Thacker
[1] PG Student, [2] Asst. Professor
[1] Department of Computer Engineering,
[1] HJD Institute of Technical Education and Research
Kera, Kutch, Gujarat Technological University, Gujarat, India

Abstract : Data mining is the process of extracting useful information from a large volume of data. Sequential pattern mining is very important technique in the field of data mining. Sequential pattern mining is used to find sequential pattern that occur in large database. In real world massive amount of data are collected and stored everyday in the databases. Many industries are intrested to find useful informaiton in the form of sequential patterns from this databases. Sequential pattern mining is used in various application such as weblog analysis, DNA sequences, stock market analysis, shopping sequence etc. There are mainly two approaches in sequential pattern mining, First is Apriori based approach and Second is Pattern growth based approach. This paper represents a study review on various algorithms of sequential pattern mining to discover sequential pattern from a large sequence database, which is very important problem in the field of data mining.

IndexTerms **- Sequential pattern mining, frequent pattern, data mining, sequence database.**

I. INTRODUCTION

Data mining is a process of extracting useful information from huge volume of data. It discover hidden pattern from large volume of data. The sequential pattern mining is a very important concept of data mining, and it is an extension of association rule mining [1].

Sequential pattern mining was first introduced by Agrawal and Srikant in 1995 [2]: "Given a set of sequences, where each sequence consists of a list of elements and each element consists of a set of items and given a user specified min_support threshold, sequential pattern mining is to find all frequent subsequences, i.e., the subsequences whose occurrence frequency in the set of sequences is no less than min_support".

Sequential pattern mining represents relation between different transactions while association rule mining indicates relationship of items in same transaction. Association rule mining finds items that are purchased with each other frequently within same transaction. While sequential pattern mining finds items those are purchased in a unique order by single customer within several transactions. So sequential pattern mining is very useful for a marketing manager to find which item is purchased by particular customer one by one in sequence [3].

Consider an Example of Sequential pattern mining for that consider the following sequence database. A sequence database is a list of sequences that is representing the purchases made by customers in a retail store.

Table 1 A Sequence Database

SID	Sequence
1	({a,b},{c},{f,g},{g},{e})
2	({a,d},{c},{b},{a,b,e,f})
3	({a},{b},{f,g},{e})
4	({b},{f,g})

There are four sequences in above database. Each single letter represents an items and items between curly brackets represent an itemset. For example, in this database, the first sequence (SID1) indicates that a customer purchased items a and b together, then an item c, then items f and g together, then an item g, and then finally an item e.

To apply sequential patter mining, a user must have two things: 1. Sequence database and 2. Minimum support threshold(minSup). The pattern is considered as a frequent pattern if pattern appear in sequence equal to or more than minSup. For example, Consider minSup=2 then there are 29 subsequences found which are described in below table:

Table 2 Sequential Pattern Found

Pattern	Sup.	Pattern	Sup.
({a})	3	({b},{g},{e})	2
({a},{g})	2	({b},{f})	4

({a},{g}.{e})	2	({b},{f,g})	2
({a},{f})	3	({b},{f},{e})	2
({a},{f},{e})	2	({b},{e})	3
({a},{c})	2	({c})	2
({a},{c},{f})	2	({c},{f})	2
({a},{c},{e})	2	({c},{e})	2
({a},{b})	2	({e})	3
({a},{b},{f})	2	({f})	4
({a},{b},{e})	2	({f,g})	2
({a},{e})	3	({f},{e})	2
({a,b})	2	({g})	3
({b})	4	({g},{e})	2
({b},{g})	3		

As shown in above example, the patterns <{b}> and <{b},{g}> are considered as frequent and havind support of 4 and 3 sequences. The pattern <{b}> appears in sequences 1, 2, 3 and 4, and the pattern <{b},{g}> appears in sequences 1, 3 and 4.

Sequential pattern is used in various areas, such as in medical using the patient history of symptoms to determine the disease, web logs mining by sequences of web pages visited by user on website, mining stock market data by sequence of events on the stock market.

II. LITERATURE SURVEY

Sequential pattern mining, was developed in 1995 by R Agrawal and R Srikant, has been used in data mining research field with various applications. There are several sequential pattern mining algorithm which have been described in literature.

Generally Sequential Pattern Mining Algorithms differ in two ways[20]:

1) The process in which candidate sequences are generated and stored. The main objectives of algorithm are to minimize the set of candidate sequences.

2) The process in which support and frequency of candidate sequence is counted. Based on these two key criteria's sequential pattern mining can be divided into two parts:

- Apriori Based.
- Pattern Growth Based.

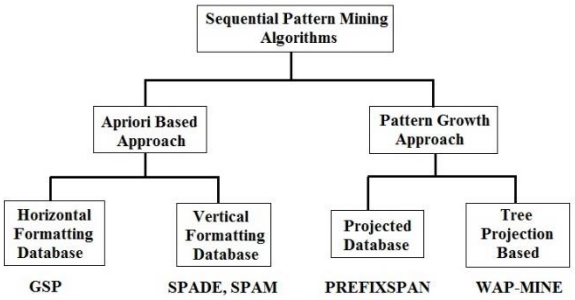

Figure 1 Classification of Sequential Pattern Mining Algorithm [8]

Apriori-based algorithm can be devided into two types: Horizontal data format and Vertical data formats. Apriori and GSP algorithm are the examples of horizontal data format and SPADE and SPAM are the examples of vertical data format. Pattern growth algorithms are introduced in 2000s and used to solve the problem of generating candidates and test. FreeSpan and PrefixSpan are most common pattern growth types of algorithms.

Apriori Algorithm

The Apriori and AprioriAll algorithms proposed by Agrawal and Srikant [2][4]. AprioriAll was the first generation of pattern mining algorithm, that is based on the apriori property and generate candidate sequences by using Apriori-generate join procedure. Apriori property says that all non-empty subset of frequent item set should also be frequent and this property is also called anti-monotonic property. The limitation of this algorithms is multiple scan of database and creation of huge number of candidate sequence.

GSP Algorithm

The GSP (Generalized Sequential Patterns) [5] algorithm is apriori and breadth-first search based algorithm introduced by Agrawal and Srikant in 1996. It makes multiple passes over database. The GSP algorithm is much faster than AprioriAll algorithm. 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. The GSP algorithm has mainly two steps:

(i). Candidate Generation (ii). Support Counting.

SPADE Algorithm

The SPADE (Sequential PAttern Discovery using Equivalence classes) [5] is based on vertical format pattern mining algorithm which was introduced by M Zaki, 2001. The algorithm uses vertical id-list database format and breakdown original search space into sub-lattices by using lattice-theoretic approach. The SPADE algorithm scan database three times. In 1st scan it construct frequent 1-sequences, In 2nd scan it construct frequent 2-sequences and in 3rd scan it construct all other frequent sequences. SPADE reduces input/output costs by minimizing database scans. It also reduces calculational cost by using efficient search schemes.

SPIRIT Algorithm

SPIRIT (Sequential Pattern mIning with Regular expressIon consTraints) [6] is sequential pattern mining algorithm with regular expression constraints. It uses some relaxed constraint which is very good with pruning technique. There are several versions of this algorithm, in which SPIRIT (V) (V for valid) perform best among all algorithm of SPIRIT family.

SPAM Algorithm

SPAM (Sequential PAttern Mining) [7] is used to find all frequent sequences in a transactional database. The SPAM is efficient for mining long sequential pattern. SPAM includes the idea of SPADE, GSP and FreeSpan. The SPAM uses the vertical bitmap data structure representation of database. SPAM uses the depth-first search strategy for mining sequential patterns to increase performance.

FREESPAN Algorithm

FREESPAN (FREquEnt pattern-projected Sequential PAtterN mining) [8] is proposed by Jiavei Han. It was introduced to significantly reduced the expensive candidate generation and testing. Its general idea is to use frequent item to recursively project sequence database into a set of smaller projected database and grow subsequences fragment in each projected database. FreeSpan is efficient and mines the complete set of patterns. FreeSpan algorithm performs significantly faster than apriori-based GSP algorithm.

PrefixSpan Algorithm

PrefixSpan (Prefix Projected Sequential pattern Mining) is popular pattern-growth algorithms for sequential pattern mining. It is proposed by Jian Pei, Jiavei Han and Helen Pinto [9]. It explores the search space of sequential pattern using depth-first search. It works on projection of database and sequential pattern growth. The PrefixSpan performs better than all other algorithm like apriori, freespan, SPADE. The PrefixSpan finds the frequent items by scanning sequence database once. The devide and search space technique is implemented by prefixspan. The PrefixSpan algorithm need more memory space as compared to other algorithm as it has to create and process of huge number of projected database.

BIDE Algorithm

BIDE (BI-Directional Extension) [10] is an algorithm generally used to mine frequent closed sequences without candidate maintenance. BIDE uses depth first search strategy to obtain closed sequential pattern. It adopts a novel sequence closure checking scheme called BI-Directional Extension. BIDE requires multiple database scans. When support is low, this algorithm work faster and consume less memory.

LAPIN-SPAM Algorithm

LAPIN-SPAM(Last Position Induction Sequential PAttern Mining) [11] is based on the same principles as SPAM. LAPIN-SPAM is vertical bit-wise method to find all the frequent sequence from a large database. SPAM uses many ANDing operations, to avoid this comparison or ANDing operation, LAPIN-SPAM constructs a ITEM_IS_EXIST_TABLE when scanning the database for the first time. LAPIN-SPAM algorithm is used this table to check current position of candidate in each iteration. In this way, it minimizes time by avoiding ANDing operation or comparison.

CloFS-DBV Algorithm

CloFS-DBV [12] algorithm is used to mine frequent closed sequences. It uses a vertical data format, dynamic bit vectors and data compression. CloFS-DBV generate self pattern structure name CloFS-DBVPattern Data structure, It consider the initial and ending position of the sequences. The structure of CloFS-DBVPattern is the combination of DBV structure with sequence information representation. CloFS-DBV algortihm need low storage space because of compressed data structure.

CSPAN Algorithm

CSPAN algorithm [13], for mining closed sequential pattern, is introduced to work on large volume of sequence database. The algorithm uses depth-first search for generating the closed sequential patterns. CSpan uses a pruning method called occurrence checking which allows the early detection of closed sequential patterns during the mining process.

VMSP Algorithm

VMSP (Vertical mining of Maximal Sequential Patterns) [14] is a maximal sequential pattern mining method. It is a first vertical mining method for mining maximal sequential patterns. VMSP is based on a vertical depth-first search strategy. The algorithm includes three efficient strategies known as EFN (Efficient Filtering of Non-maximal patterns), FME (ForwardMaximal

Extension checking) and CPC (Candidate Pruning by Co-occurrence map). These strategies are used to define maximal patterns and minimize the search space.

MS-SPADE Algorithm

MS-SPADE (Multiple Support Sequential Pattern Discovery using Equivalent classes) [15] is used for sequential pattern mining with multiple minimum support. MS-SPADE use the concept of percentage of participation(POP). The basic fundamentals of POP calculation is depended on minSup for each itemset.

FBSB Algorithm

FBSB (Frequent Biological Sequence based on Bitmap) [16], uses bitmap. Bitmap is used to store the value of sequence position in every transaction. After that a list named quicksort list(QS-list) is constructed for connection of sequence.

MSPM Algorithm

MSPM [17] is used in multiple biological sequences to mine frequent pattern. MSPM is based on the concept of primary pattern. In this algorithm first primary patterns table is constructed. All the primary patterns are stored in this table. After that prefix tree is constructed for primary patterns to mine frequent primay patterns from this prefix tree.

CUSE Algorithm

CUSE (Cube-based Sequential pattern mining) [18] is a bit-wise approach. In this algorithm a database re-construction method is introduced which converts a sequence database into a 3 dimensional matrix. This 3D matrix dimensions are Sid, items and number of elements and this 3D matrix is known as sequence cube. This algorithm required two time database scan, in 1st scan it eliminate infrequent items from sequence database and in 2nd pruned sequence database scan again and generated 3D sequence cube. In this algorithm the construction of 3D matrix require more memory.

III. COMPARISON OF SEQUENTIAL PATTERN MINING ALGORITHMS

Table 3 Sequential Pattern Algorithm Comparison Table

Sr. No	Algorithm	Key Feature	Limitations
1	AprioriAll	BFS based approach	Create huge number of candidate sequence, Multiple scan of the database
2	GSP	Apriori and BFS based approach	Repeatedly scan the database, it may generate pattern that do not exist in the database, Maintaing candidate in memory so consume a huge amount of memory
3	SPADE	Use vertical id-list database format, Use lattice-theoretic based approach	A huge set of candidates generated, Inefficient for mining long sequential pattern
4	SPIRIT	sequential pattern mining algorithm with regular expression constraints	Regular expression are difficult to identify
5	SPAM	Depth-first search strategy, Use vertical bitmap representation	Required that all of the data fit into main memory
6	FreeSpan	Pattern Growth based algorithm, Use projected sequence database	If a pattern appears in each sequence of a database, its projected database does not shrink every time
7	PrefixSpan	Pattern Growth based method, Use Depth-first search based approach, Use projected prefix database	Major cost of PrefixSpan : constructing projected databases
8	BIDE	Novel sequence closure checking scheme	Needs multiple database scans for the bi-direction closure checking and backscan pruning
9	LAPIN-SPAM	Vertical Bit-wise method, Use concept of SPAM	Overhead to convert dataset from horizontal to vertical
10	CloFS-DBV	Use Vertical data format and data Compression, Use own pattern structure	Conversion overhead of dataset from horizontal to vertical
11	CSPAN	Use depth-first search for generating the closed sequential patterns, Use sequential pattern tree to generate the closed sequential pattern	Major cost of CSpan : constructing projected databases
12	VMSP	Vertical depth-first based algortihm for mining maximal sequential pattern, Use hash table and Low memory usage	Overhead to convert dataset from horizontal to vertical
13	MS-SPADE	Use concept of parcentage of participation(POP), Concept of POP is based on the minsupport thresold for each calculation	POP require huge memory usage and lots of computation time

14	FBSB	Use bitmap to store sequence position, use QS-list	Mines lot of short sequences
15	MSPM	Based on the concept of primary pattern, Use in multiple biological sequences to mine frequent pattern	Need extra time to devide original sequences to form primary pattern and primary pattern tree
16	CUSE	Bit wise approach and workes on 3D matrix structure	Matrix structure require more memory

IV. FUTURE DIRECTION AND RESEARCH CHALLENGES

The Sequential pattern mining have been studied for more than two decades and still today it is very active research areas. Here is a list of important research opportunities[19].

- **Application:** There are variuos new applications and the application which is already used in several systems, one can apply the basic fundamentals of sequential pattern mining. Also in research and development fields like IOT(Internet of Things), Social and Sensor networks sequential pattern mining can be applied.
- **Developing more efficient algorithms:** When there is a database containing numerous sequences or long sequences databases, this pattern mining methods may be expensive in terms of calculation, generation/conversion time and search space. To reduce and improve this problems various algorithms have been introduced.
- **Designing algorithm to handle more complex data:** Also various algorithms of sequentilal pattern mining have been proposed for the dence database that is very complex in nature and large in memory.
- **Designing algorithms for finding more complex and meaningful pattern:** There is also an issue to find more complex patterns in sequence. And to find more meaningful patterns reseach should be further carried out.

There are large numbers of frequent sequntial patterns are in databases. So a proposed mining algorithn should be contained following properties:

1. Mine all frequent patterns with satisfying minSup.
2. Be much more efficient and minimizes the database scan.
3. Satisfy the several user specific constraints.

V. CONCLUSION

Sequential pattern mining is very good concept in today's environment. It is very much useful to identify useful information in the form of pattern and also used for analytic approach. There are various useful information which we can convert from this pattern, these datas are used to make prediction in different organizations and also used to improve usefulness of systems. In various marketing strategic decision and to identify useful events also these data are used. Several sequential pattern mining algorithms and methods have been discussed in this paper.

References

[1] Jiawei Han, Micheline Kamber and Jian Pei, "Data Mining: Concepts and Techniques", Morgan Kaufman publishers is an imprint of Elsevier, 2001

[2] Rakesh Agrawal, and Ramakrishnan Srikant, "Mining sequential patterns". Proceedings of the Eleventh International Conference on Data Engineering, 1995.

[3] M. Chaudhari and C. Mehta, "A Survey on Algorithms for Sequential Pattern Mining", International Journal of Engineering Development and Research, Volume 3, Issue 4, 2015.

[4] Qiankun Zhao and Sourav Bhowmick, "Sequential Pattern Mining: A Survey", Technical Report, CAIS, Nanyang Technological University, Singapore, No. 2003118, 2003.

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

[6] M. Garofalakis, R. Rastogi, and K. Shim, "SPIRIT: Sequential pattern mining with regular expression constraints", 25th VLDB Conference, Edinburgh, Scotland, 1999.

[7] Jay Ayres, Johannes Gehrke, Tomi Yiu, and Jason Flannick, "SPAM: Sequential PAttern Mining using A Bitmap Representation" SIGKDD '02 Edmonton, Alberta, Canada, 2002, ACM 1-58113-567-X/02/0007.

[8] Vishal S. Motegaonkar and Madhav V. Vaidya, "A Survey on Sequential Pattern Mining Algorithms", International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 2486-2492.

[9] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M. C. Hsu, "PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern-Growth" Proceding 2001 of Internationnal Conference on Data Engineering(ICDE'01), pp. 215-224, Heildelberg, Germany, April 2001.

[10] J. Wang, and J. Han, "BIDE: Efficient mining of frequent closed sequences" Proceedings of the 20th International Conference on Data Engineering, 2004.

[11] Zhenglu Yang and Masaru Kitsuregawa, "LAPIN-SPAM: An improved algorithm for mining sequential pattern" Proceedings of the 21st International Conference on Data Engineering Workshops. 2005, pp. 12-22.

[12] Minh-Thai Tran, Bac Le, Bay Vo, "Combination of dynamic bit vectors and transaction information for mining frequent closed sequences efficiently" Engineering Applications of Artificial Intelligence, Volume 38, February 2015, pp. 183–189

[13] V. PurushothamaRaju, and G. P. Saradhi Varma, "MINING CLOSED SEQUENTIAL PATTERNS IN LARGE SEQUENCE DATABASES" International Journal of Database Management Systems (IJDMS) Volume 7, No.1, February 2015.

[14] Philippe Fournier-Viger, Cheng-Wei Wu, Antonio Gomariz,and Vincent S. Tseng, "VMSP : Efficient Vertical Mining of Maximal Sequential Patterns" Advances in Artificial Intelligence, Volume 8436 of the series Lecture Notes in Computer Science, May 2014, pp. 83-94.

[15] K.M.V. Madan Kumar, P.V.S. Srinivas, and C. Raghavendra Rao, "Sequential Pattern Mining With Multiple Minimum Supports by MS-SPADE" IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 1, September 2012.

[16] Qian Wang, Darry N Davis, Jiadong Ren, "Mining frequent biological sequences based on bitmap without candidate sequence generation" Computers in Biology and Medicine Volume 69, 2016, pp. 1 52-157

[17] Ling Chen, Wei Liu, "Frequent patterns mining in multiple biological sequences", Computers in Biology and Medicine, Volume 43, 2013, pp. 1444-1452

[18] M K Sohrabi and V Ghods, "CUSE: A Novel Cube-based Approach for Sequential Pattern Mining", 4th International Symposium on Computational and Business Intelligence 2016

[19] P Fournier-Viger, J Chun-W Lin, R Uday Kiran, Yun Sing Koh, Rincy Thomas, "A Survey of Sequential Pattern Mining", Data Science and Pattern Recongnition, Ubiquitous Internatitional, Volume 1, Number 1, February 2017.

[20] R Boghey and S Singh, "Sequential Pattern Mining: A Survey on Approaches", International Conference on Communication Systems and Network Technologies 2013