A Study of Sequential Pattern Mining Algorithm

Taral Patel[1] Prof. Narendra Limbad[2]
[1]M.E. (Pursuing) [2]Faculty
[1,2]Department of Computer Engineering
[1,2]L.J.I.E.T, Ahmedabad, India

Abstract— Sequential pattern mining use to finds frequently occurring ordered events or sub sequence as pattern from sequence database. Applications of SPM are analysis of web click stream data, medical data, biological data, e-learning data, customer purchase behavior, natural disaster. In this paper represent review of sequential pattern mining technique using different algorithm. The main classifications of SPM algorithm are two categories. First is Apriory based algorithm like GSP, SPADE, SPAM and second is pattern growth based algorithm like Prefix span, WAPMINE. Modified SPM algorithm like mine closed sequential pattern algorithm like BIDE, Clospan etc, maximal sequential pattern like MsPX,, MaxSP, MFSPAN etc. At last, comparative study of different algorithm is done via different key features support by SPM algorithm and also study extensions of sequential pattern mining algorithm.

Key words: Sequential pattern mining, closed sequential pattern mining, Maximal Sequential pattern mining

I. INTRODUCTION

Sequential pattern mining is important concept of data mining which used to mine interesting and necessary pattern from large sequence database. Sequence database consist of sequences of ordered elements or events, recorded with or without concrete notion of time.[1] The problem was first introduced by Agrawal and Srikant [2]. Sequential pattern mining use to finds frequently occurring ordered events or sub sequence as pattern from sequence database. Sequence can be called as order list of event. If one item set is completely subset of another item set is called sub sequence. Application of sequential pattern mining are, it can be used in the medical domain to help determine a correct diagnosis from the sequence of symptoms experienced; in medical data base it can be used to match DNA sequence, over customer data to help target repeat customers; in e-learning the search data give on particular topic user want to search and with web-log data to better structure a company's website for easier access to the most popular links.[3]

The process of mining sequential pattern from customer transaction is described as follows:

An itemset is a non-empty set of items. A sequence is an ordered list of itemsets or events. Without loss of generality, here assume that the set of items is mapped to a set of sequence integers. [2] We denote an itemset i by (i_1, i_2, i_3, ... i_m) where i_j is an item1. We denote a sequence s by <s_1, s_2, s_3, ... s_n> where s_j is an itemset. A sequence <x_1, x_2, x_3, ... x_n> is contained in another sequence <y_1, y_2, y_3, ... y_m> if there exists integers $k_1 < k_2 < ... < k_n$ such that $x_1 \subseteq y_{k1}$, $x_2 \subseteq y_{k2}$, ... $x_n \subseteq y_{kn}$. For example, the sequence <(4) (6 7) (9)> is contained in <(2) (4 8) (5) (4 5 6 7) (9) >, since (4) ⊆ (4 8), (6 5) ⊆ (4 5 6 7) and (9) ⊆ (9). However, the sequence <(6) (7)> is not contained in <(6 7)> and vice versa. The former represents items 6 and 7 being bought one after the other**,** while the latter represent items 6 and 7 being bought together. If support of event or pattern < min_support than pattern is not frequent.

Sequential pattern are mainly classified in Apriory based algorithm like GSP [4], SPADE [5], SPAM [6] etc which use generate and test approach and pattern growth algorithm like Prefixspan [7], WAPMINE [8] etc. which use divide and conquer. Extension of sequential pattern mining algorithm are closed sequential pattern mining algorithm like BIDE[10], CloSpan[9] etc. which are used to mine closed SPM and Maximal sequential pattern mining algorithm like MsPX[12] , MaxSP[14], MFSPAN[13] which are used to mine maximal SPM. Closed SP is a sequential pattern or frequent patterns that have no super sequence of sequence with same support. Maximal sequential pattern is the largest frequent patterns which have all subsequence patterns are also frequent. Maximal sequential pattern find largest pattern which have no super sequence pattern find to be frequent.

II. TAXONOMY OF SEQUENTIAL PATTERN MINING ALGORITHM

Sequential pattern mining algorithm is different in two ways: [3]

- The process in which candidate sequence are generated and stored. The main objectives of algorithm are to minimize the set of candidate sequences.
- 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 in two parts:
  - Apriory Based
  - Pattern growth based

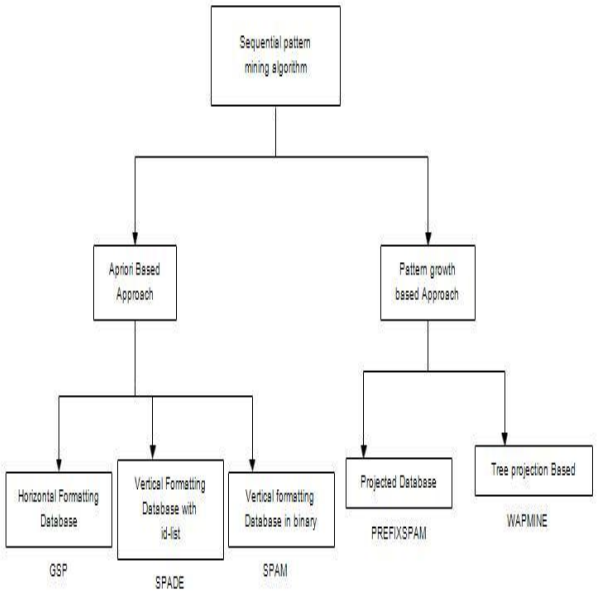

Fig. 1: Classification of SPM algorithm

A. Apriory Based algorithm

Apriory algorithm work on candidate "generate and test" property. Apriory property states that "All non-empty subset of frequent item set must also be frequent". It also called downward closed means if sequence does not satisfy minimum support than super sequence is also frequent. In this algorithm work on breath first search and data base is scan multiple time.

1) GSP

Generalized sequential pattern (GSP) is Apriory based algorithm [4]. In this multiple pass operation are perform for generate frequent item sets or pattern. In this algorithm candidate generate and pruning steps are performed. In GSP horizontal format data is used and batter than apriory all algorithms. The operation of this algorithm is not performing in main memory. The candidates which satisfy condition of minimum support are stored in memory and remaining candidates are deleted. This process is repeat until all candidates are deleted. The algorithm is repeated until no candidate or frequent pattern is found. This algorithm has a scale up property for number of transaction in data sequence and number of item per transaction. The drawback of this algorithm is, that it is not efficient for large set of candidate sequence because it require multiple time database scan.

2) SPADE

SPADE is stands for sequential pattern discovery using equivalence classes which use vertical format database. It is used to fast mining of sequential pattern in large databases.[5] SPADE is solving problem in main memory via lattice search technique in which main problem is divided in the smaller sub problem. All sequence patterns are discovering in three database scan. SPADE use only simple temporal join operation and thus ideally suited for direct integration with DBMS. SPADE is out perform than GSP in two factor, one is order of magnitude for pre-computed support of sequence and another is number of parameter like number of input sequence, number of event per input sequence. Drawback of this is, Additional time require transform from horizontal layout to vertical format which may require larger storage space than original sequence database.

3) SPAM

The SPAM (Sequential Pattern Mining) algorithm utilizes a depth first traversal of the search space combined with a vertical bitmap representation to store each sequence. Vertical bitmap data layout allowing for simple efficient mining [6] In SPAM assumes that entire database and all data structure used for algorithm completely in the memory. Sequential pattern are mining in traversal of lexicographical sequence tree in DFS fashion. A salient feature of SPAM is SPM of online outputting is incremental length of pattern.

B. Pattern Growth Based Algorithm

The pattern growth algorithm is design to avoid problem of candidate generation step and use search space partitioning concept for pattern growth. All pattern growth algorithms are, firstly mined database then partition search space and generate minimum number of candidate sequence as possible by growing on already mined frequent sequence, finally apply apriory for recursively looking frequent sequence.[7] Pattern growth based algorithm focus on features like search space partition, tree projection, depth first traversal, candidate sequence pruning.

1) PREFIXSPAN

PrefixSpan (Prefix-projected sequential Pattern mining) algorithm explores pattern growth approach for efficient mining of sequential pattern in large sequence database. These uses divide and conquer strategy with pattern growth approach in which sequential database is recursively projected into smaller projected database based on current sequential pattern. Sequential pattern are grown in each projected database by exploring only locally frequent fragment. Prefixspan offer order growth and reduce projected database with pseudo-projection techniques [7]. Prefixspan examine only the prefix sequences and projects only their corresponding postfix subsequence in projected database. The main disadvantage of this algorithm is, cost of memory space is high because of creation and processing of huge number of projected sub-databases.

2) WAPMINE

Mining access pattern from weblog is called WAPMINE (Web access pattern mining) for that tree projection approach use with pattern growth approach. WAP tree store highly compressed critical information for access pattern mining and access the pattern in large set of log pieces. [8] This algorithm scan database twice which avoid problem of generating candidate set like apriory based. A WAP tree is generating with frequent sequence for that "header table" is maintain occurrence of frequent item set. In first scan 1sequence of frequent item find from database and in second scan WAP tree is builds with frequent subsequence. This algorithm is suffers from memory consumption problem because it construct recursively WAP tree when number of mined frequent pattern increase.

III. EXTENSION OF SEQUENTIAL PATTERN MINING ALGORITHM

Sequential pattern mining has been intensively studied during recent years; there exists a great diversity of algorithm for sequential pattern mining. Most of SPM algorithm have been modified to support concise representation like closed, Maximal, incremental sequence. These algorithms represent compact representation of sequential pattern mining and no need to mine full set of sequential pattern.

A. Closed sequential pattern mining algorithm

The sequential pattern mining algorithms which are previously generated mine full set of frequent pattern mining where Closed sequential pattern algorithm generate less number of frequent pattern so performance of is better than previously full set generated algorithm. The subsequence of frequent sequence are also frequent, so closed sequential pattern mining avoid generation of unnecessary subsequence which give more compact and efficient result than mining full set of frequent pattern.[9] Closed sequential pattern generate frequent pattern which have no super sequence of sequence with same support. The closed sequential pattern mining algorithm like Clospan and BIDE describe below:

1) Clospan

Clospan (Closed Sequential pattern mining) mine closed sequential pattern rather than mine full set sequential pattern. This algorithm follows candidate maintenance and test process. A candidate generate in first stage which is larger than the final closed sequence set and candidate is stored in hash indexed result tree structure. A pruning method is called second stage which eliminates non-closed sequence and do post-pruning on it. Pruning method like Common prefix and Backward sub pattern pruning to prune the search space.[10] The main different between Prefixspan and clospan is that clospan avoid unnecessary traversing of search space via using early termination mechanism. Limitation of this algorithm is that candidate set generation is costly in both runtime and space usage when the support threshold is low or pattern become long because it maintain historical closed sequence candidate.

2) BIDE

BIDE (Bi-Directional Extension) is mine complete set of frequent closed pattern without candidate maintenance. This algorithm use the pseudo projection for mine frequent 1-sequence and then, It use sequence closure check scheme called Bi-direction extension in which forward directional extension is used to grow the prefix patterns and also check the closure of prefix pattern, where in backward directional extension used both check closure of prefix pattern and prune the search space. In prune search space with using Backscan Pruning method and Scan skip optimization technique. BIDE is not maintain already mined pattern, so space efficient. This algorithm is linearly scalable in term of database size. It uses pseudo projection method and depth first search manner. BIDE has order of magnitude is faster than Clospan. [11]

B. Maximal sequential pattern mining algorithm

Sequential pattern mining represent too many sequential patterns to user, which degrades performance of mining task in term of execution time and memory requirement and make difficult to user comprehend the result. So the concept of maximal sequential pattern is used. In Maximal sequential pattern is find the largest frequent sequential patterns which have all subsequence patterns are also frequent. Maximal sequential pattern find largest pattern which have no super sequence pattern find to be frequent.[14] A maximal sequential pattern is frequent sequential pattern that is not strictly included in another frequent sequential pattern.

1) MsPX

MSPX mines the maximal frequent pattern or sequence by effectively excluding infrequent candidates. MSPX adopt apriory candidate generation method and first perform bottom-up, breadth first search. It is not count all candidate in one pass of whole database, it try to find and remove the most infrequent candidate by counting few candidate. For each pass of database find which candidate are most infrequent are verified against the database and remove it. At the end of this process super set of all frequent set is obtained. Now second top-down search is performing from border of superset and pick the maximal frequent sequence efficiently.[12] This is approximate algorithm and therefore it provides an incomplete set of maximal pattern to user, thus may omit important information.

2) MFSPAN

MFSPAN(Maximal frequent sequential pattern mining algorithm) to mine the complete set of maximal frequent sequential pattern from sequence database. In this algorithm frequent sequence tree is pruned and generate the maximal frequent sequence (MFS) tree, with using sequence extended step and item set extended step and for generate MFS tree use divide and conquer strategy use.[13] MFSPAN take advantage of this property that two different sequences may share a common prefix to reduce item-set comparing times. The disadvantage of this algorithm that need to maintain a large amount of intermediate candidates in main memory during mining process

3) MaxSP

MaxSP(Maximal sequential pattern miner) algorithm is pattern growth algorithm and inspired by PrefixSpan algorithm. MaxSP generate all maximal sequential patterns without storing intermediate candidate in main memory and not producing redundant candidate. [14] For that this algorithm uses mechanism of maximal forward extensions and maximal backward extension. A maximal sequential pattern is closed sequential pattern which not strictly included in another closed pattern. Maximal sequential pattern is much smaller than closed sequential pattern and give compact representation of sequential pattern. MaxSP algorithm work on main three concept, first it use the pseudo projection in which not store the physical copy of projected database, second remove infrequent item from database immediately after first database scan and third optimization concerns the process of searching for the maximal-backward-extensions of a prefix by scanning maximum period. In this algorithm use new mechanism for determine maximal pattern without comparing with previously generated or found pattern. Pattern optimization can perform in two ways, one is Scan Skip optimization and second is stop scanning sequence if pattern get non-closed. This algorithm gives batter result compare to BIDE algorithm. [14]

In Figure 2, see the sequence database and then we find the sequential pattern for the minimum support =2. The sequence in {} are occur together means concurrent occurring event and another event is occur after first event is completed. So, in figure 3 see 29 sequential patterns, 15 closed patterns and 10 maximal patterns. These are show the different between sequential pattern, closed pattern and maximal pattern.

SID	Sequences
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}>

Fig. 2: Sequence database

Sequential Pattern	Support	Sequential Pattern	Support
<{a}>	3 C	<{b},{g},{e}>	2 CM
<{a},{g}>	2	<{b},{f}>	4 C
<{a},{g},{e}>	2 CM	<{b},{f,g}>	2 CM
<{a},{f}>	3 C	<{b},{f},{e}>	2 CM
<{a},{f},{e}>	2 CM	<{b},{e}>	3 C
<{a},{c}>	2	<{c}>	2
<{a},{c},{f}>	2 CM	<{c},{f}>	2

<{a},{c},{e}>	2 CM	<{c},{e}>	2
<{a},{b}>	2	<{e}>	3
<{a},{b},{f}>	2 CM	<{f}>	4
<{a},{b},{e}>	2 CM	<{f,g}>	2
<{a},{e}>	3 C	<{f},{e}>	2
<{a,b}>	2 CM	<{g}>	3
<{b}>	4	<{g},{e}>	2
<{b},{g}>	3 C		

C= Closed Pattern M= Maximal Pattern

Fig: 3 Sequential pattern found for min_support=2

IV. COMPARATIVE ANALYSIS OF SEQUENTIAL PATTERN MINING ALGORITHM

The analysis of algorithm is completed with basis of comparative study of various important features.

- GSP: Generalized Sequential Pattern
- SPADE: Sequential Pattern discovery using equivalence classes.
- SPAM: Sequential pattern mining
- PREFIXSPAN: With Prefix projected Sequential pattern mining
- WAPMINE: Web access pattern mining use of sequential dataset contain web click stream in sequence format with time.
- Clospan: Closed sequential pattern mining
- BIDE: Bi-Directional Extension
- MFSPAN: Maximal sequential Pattern mining
- MaxSP: Maximal sequential pattern miner

Comparative study of sequential pattern mining algorithm with key feature:

No.	Algorithm	Features
1	GSP	Apriory based, Bottom to up search, BFS based approach, use anti-monotone Property.
2	SPADE	Apriory based, DFS based approach, Bottom up search, database vertical projection, use anti-monotone property , lattice theoretical based approach.
3	SPAM	Apriory based, DFS based, Bottom up search, use vertical bitmap representation for data storage, database vertical projection.
4	PREFIXSPAN	Pattern growth based, DFS based approach, top-down search, prefix-monotone property, regular expression constrain, use prefix heuristic and bi level projection
5	WAPMINE	Tree projection approach, Pattern growth based approach, DFS based approach, top-down search, regular expression constrain.
6	Clospan	Candidate maintain and test (Apriory) approach use, Bottom up approach
7	BIDE	DFS based approach, top to down approach
8	MsPX	Apriory based, Bottom up approach, BFS search
9	MFSPAN	Apriory based approach, BFS search, Bottom up search approach
10	MaxSP	Pattern growth based, DFS based, Prefix-monotone property satisfy

Table 1: Comparative study of SPM Algorithm

- Apriory based: Apriory based algorithm use candidate generate and test approach with down word closure property means if any item set not frequent its super set is also not frequent.
- Pattern growth based: Pattern growth based approach is incremental approach and use divide and conquer strategy. It use concept of projection database to reduce search space.
- Top-down search: The mining of sequential pattern subset can be done by corresponding set construction of projected database and recursively mining from top to down.
- Bottom up search: Apriory based approach use the bottom-up approach for every single frequent item set.
- BFS based approach: BFS approach we can say level by level approach because all children node are processed before going next level.
- DFS based approach: Using DFS based approach, all sub arrangement of path must explored before moving next one.
- Database vertical projection: In this based on bitmap or position induction table is constructed for every frequent item and visit sequence database ones or twice a vertical layout of database rather than horizontal form.
- Anti-monotone property: Anti-monotone property state that every non-empty sub-sequence of sequential pattern is sequential pattern.

V. CONCLUSION AND FUTURE WORK

In this paper we discussed, fundamental of sequential pattern mining and different sequential pattern mining algorithm. The main classification of sequential pattern mining algorithm is Apriory based and pattern growth based like GSP, SPADE, and SPAM. PrefixSpan, WAPMINE. Also represent comparative study of these sequential patterns mining algorithm. Also discuss extension of sequential pattern mining algorithm like closed sequential pattern mining algorithm e.g. BIDE, Clospan and maximal sequential pattern mining algorithm e.g. MaxSP, MSPX, MFSPAN. These are given compact representation of sequential pattern mining algorithm. In future, the work can do on maximal sequential pattern mining algorithm for more concise pattern and also apply rule on that maximal sequential pattern like association rule, correlations rule etc on maximal pattern.

REFERENCES

[1] Jiawei Han and Micheline Kamber, "Data Mining Concepts and Techniques", Second Edition (2006), Morgan Kaufmann
[2] R. Agrawal, R. Srikant , "Mining sequential patterns," In Proceedings of Inter- national Conference on Data Engineering, pp. 3–14, 1995
[3] Mabroukeh, N. R. and Ezeife, C. "A taxonomy of sequential pattern mining algorithms", ACM Computing Surveys, vol. 43, no. 1, pp. 1-41 (2010)

[4] R. Srikant, R. Agrawal, "Mining sequential patterns: Generalizations and performance improvements," In Proceedings of International Conference on Extending Database Technology, pp. 3–17, 1996.
[5] Zaki, M. J., "SPADE: An efficient algorithm for mining frequent sequences", Machine learning, vol.42.no.1-2, pp.31-60 (2001)
[6] Ayres, J., Flannick, J., Gehrke, J. and Yiu, T., "Sequential Pattern mining using a bitmap representation", Proc. KDD 2002, Edmonton, Alberta, pp. 429-435 (2002)
[7] Pei, J., Han, J., Mortazavi-Asl, B., Pinto, H., Chen, Q., Dayal, U., Hsu, M., "Mining Sequential Patterns by Pattern-Growth: The PrefixSpan Approach", IEEE Trans. Knowledge and Data Engineering, vol. 16, no. 10, pp. 1-17 (2001)
[8] 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.
[9] Wang, J., Han, J., Li, C., "Frequent Closed Sequence Mining without Candidate Maintenance", IEEE Trans On Know. And Data Engineering, vol.19, no.8,pp. 1042-1056(2007)
[10] Yan, X., Han, J. and Afshar, R., "CloSpan: Mining closed sequential patterns in large datasets", Proc. of the third SIAM International Conference on Data Mining, May 1-3, San Francisco, California, ISBN 0-89871-545-8. (2003).
[11] Jianyong Wang and Jiawei Han, "BIDE: Efficient Mining of Frequent Closed Sequences", IEEE Proceedings of the 20th International Conference on Data Engineering (ICDE'04), 2004.
[12] Luo, C., Chung, S., "Efficient mining of maximal sequential patterns using multiple samples", Proc. 5th SIAM international conf. on data mining, Newport Beach,California.(2005).
[13] Guan, E.-Z., Chang, X.-Y., Wang, Z., Zhou, C.-G. "Mining Maximal Sequential Patterns", Proc of the second Int'l Conf. Neural Networks and Brain, pp.525-528 (2005)
[14] Philippe Fournier-Vinger, Cheng-wei Wu, and Vincent S. Tseng, "Mining Maximal Sequential patterns without Candidate Maintaenance", Springer Verlag Berlin Heidelberg 2013, H. Motoda et al., LNAI 8346, pp, 169-180(2013).

