International Journal of Latest Trends in Engineering and Technology
Special Issue SACAIM 2017, pp. 374-378
e-ISSN:2278-621X

A STUDYONMINING SEQUENTIAL PATTERN IN TIME SERIES DATA

Ubaidulla D[1], Sushmitha B.S[2] & Vanitha T[3]

Abstract-Sequential pattern mining is a very important mining technique with broad applications. This is very useful in various domains like natural disaster, sales record analysis, marketing strategy, shopping sequences, medical treatment and DNA sequences etc. It discovers the subsequence's and frequent relevant pattern from the given sequences. We have provided the sequence database having sequences, in which each sequence is a list of the transactions ordered by the transaction time. Each transaction consists of the number of the items. The study of all these algorithms done with various research perspectives. GSP, SPADE, SPAM and Prefix span are few efficient sequential pattern mining algorithms.One of them is Generalized Sequential Pattern (GSP) mining algorithm which is an Apriori-based algorithm used for sequential pattern mining. PrefixSpan which is pattern growth method overtakes GSP and solves all above problems.At the beginning we have categorized these algorithms by their used approaches to solve the mining problem and then we have compared each one with another by their various provided features and performance point of view.
Keywords– Sequential pattern, Apriori, Pattern growth, GSP, SPADE, SPAM, Prefix span.

1. INTRODUCTION

Sequential pattern mining is a expansion of association rule mining [3]. This thought 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. The algorithm of mining finds the distance of frequent sequences in the database given [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 discrepancy between sequential pattern mining and association rule mining is incidents 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 many of the previous studies the efficient mining of sequential patterns or other frequent patterns in time-related data is supported. The generalizationof Sequential patterns definitiondone bySrikant and Agrawal [4].To include time constraints, sliding time window, and user-defined taxonomy, and presented apriori-based improved algorithm GSP (i.e., generalized sequential patterns) [1]. Nearly all of the above future approaches for mining sequential patterns and other time-related frequent patterns are apriori-like, i.e., based on the apriori principle, which states the fact that any super-pattern of an infrequent pattern cannot be frequent, and based on a candidate generation-and test paradigm proposed in association mining [1].

A typical apriori-like such as GSP are sequential pattern mining method [4], candidate generation-and-test approach ,adopts a multiple-pass are outlined as follows: The first scan finds all of the frequent items that form the set of single item frequent sequences. The every subsequent pass begins with a sequential patterns seed set, which is the set of sequential patterns found in the previous pass. This seed set is used to generate new potential patterns, called as candidate sequences, which are based on the apriori principle. Each candidate sequence contains one more item than a seed sequential pattern, where each element in the pattern may contain one item or multiple items. The count of items in a sequence is called the size of the sequence. So, all the candidate sequences in a pass will have the same length. The scan of the database in one pass finds the support for each candidate sequence. The entireapplicant with support no less than min_support in the database forms the set of the newly discovered sequential patterns. This set is then used as the seed set for the next pass. When no new sequential pattern is found in a pass the algorithm will stop, or when no applicantorder can be generated.

2. PROBLEM FORMULATION

This unit presents the formal explanation of the problem of sequential pattern mining

Let D be a database of customer transactions, I={ I1, I2, …In } be a set of m distinct attributes called items.

2.1. An ordered list of itemsets is a Sequence. A sequence s is denoted by < s1 s2 ……sl >, where si is an itemsets, i.e. si □ I for $1 \leq i \leq l$. si is also called an element of the sequence. Since an element is a set, and the order of its items is not important.

2.2. Length of a sequence is the number of instance of items in that sequence. A sequence with length l is called an lsequence.

[1] Department of Computer Application, St Aloysius College, AIMIT, Mangaluru, Karnataka, India
[2] Department of Computer Application, St Aloysius College, AIMIT, Mangaluru, Karnataka, India
[3] Department of Computer Application, St Aloysius College, AIMIT, Mangaluru, Karnataka, India

2.3. A sequence α= < a1 a2 ….an> is called subsequence of the β=<b1 b2 …bm>, and denoted as α □ □ β , if there exist integers 1≤ j1<j2<…<jn≤m such that a1 □ □ bj1 , a2 □ □ bj2 , ... , an □ □ bjn. Sequence α is also called the supersequence of β.
2.4. A set of tuples < Sid, s >, where Sid is a sequence identifier, and s is a sequence, is called sequence database. A tuple < Sid, s > is said to contain a sequence α, if α is a subsequence of s, i.e., α□ s.
2.5. A sequence α is called a frequent sequence pattern in sequence database S, if the number of tuples in S that contain α is greater than or equal to a given positive integer γ , called support threshold , or minimum support , i.e., Support (α) ≥ γ.
2.6. A sequence α is called a frequent Max sequential pattern in sequence database S, if α is a frequent sequential pattern in S, and there exists no frequent sequential pattern β in S, such that β is a proper super sequence of α. The problem of sequential pattern mining is to find the complete set of frequent sequential patterns in a sequence database, for a given minimum support threshold

3. SEQUENTIAL PATTERN MINING

The problem of mining sequence was first introduced by Agrawal and Sirkant [1]. Many algorithms were developed after that and effectively improved the effectiveness of the task of mining sequential patterns. A great diversity of algorithms for sequential pattern mining exists. Generally Sequential Pattern Mining Algorithms differ in two ways [6]:

- The process in which candidate sequences are generatedand stored. The main objectives of algorithm are tominimize the set of candidate sequences.
- □ The process in which support and frequency ofcandidate sequence is counted. Based on these two keycriteria's sequential pattern mining can be divided into Two parts:
  - Apriori Based
  - Pattern Growth Based

3.1 Apriori-Based Algorithms:

The Apriori [Srikant and Agrawal 1994] and AprioriAll [Srikant and Agrawal 1995] set the sourcefor a type of algorithms which largely depends on the apriori property and use the Apriori-generate join procedure to generate candidate sequences. The apriori property states that - All of the nonempty subsets of a frequent itemsets should also be frequent. It is also called asdownward-closed, that is if a sequence cannot pass the minimum support test, its entire super sequences will also fail the test.

Some Of The Apriori-Based Algorithm Key Features Are: [6]

- ✓ Breadth-first search: Breath-first are Apriori-based algorithms (level-wise) search algorithms because the build all the k-sequences, in the kth iteration of the algorithm, as they pass through the search space.
- ✓ Generate-and-test: This concept is used by some previous algorithms in sequential pattern mining. Algorithms that use this approach execute inefficient pruning method and generate a huge number of candidate sequences and then check for satisfying some user specified constraints. So this pruning process consumed alot of memory in the early stages of sequential pattern mining.
- ✓ Multiple scans of the database: Scanning the original database to ascertain whether a long list of generated candidate sequences is frequent or not which is the featureinvolved in this. It is a very undesirable characteristic of most aprioribased algorithms and requires a lot of processing time and I/O cost.

3.1.1 GSP (Generalized Sequential Patterns): The GSP algorithm makes multiple passes over the data described by Agrawal and Shrikant. This algorithm is not a main-memory algorithm. If the candidates do not fit in memory, the algorithm generates only as many candidates as will fit in memory and the data is scanned to count the support of these candidates. Frequent sequences resulting from these aspirants are written to disk, while those aspirants without minimum support are deleted. This procedure is repeated until all the aspirants have been counted. firstly GSP algorithm finds all the length-1 candidates (using one database scan) and orders them with respect to their support ignoring ones for which support < min_sup. Then for each level (i.e., sequences of
length-k), The following step will be repeated till no candidate can be found or frequent sequence[4].That is algorithm will scan database to collect support count for each candidate sequence and it will produce candidate length (k+1) sequences from length-k frequent sequences using the Apriori.

3.1.2 SPIRIT (Sequential Pattern mIning with Regular expressIon consTraints): It is a part of algorithms for sequential pattern mining with regular expression constraints. Its general idea is to use some casual constraint which has nice property to prune [2]. Theseveral versions of the algorithms are exist, differing in the degree to which the constraints are enforced to prune the search space of pattern during computation. The main distinguishing factor among the schemes is the degree to which the regular expression constraints are enforced to prune the search space. Since SPIRIT (V) has the best performance among the SPIRIT family. In a specific, algorithm SPIRIT (V) uses a normal constraint "valid with respect to some state of ME for a given regular expression E, where ME is the deterministic finite mechanismconsistent to E.

3.1.3 SPAM (Sequential Pattern Mining) : To find all the frequent sequences within a transactional database the SPAM algorithm is used. The algorithm is especially efficient when the sequential patterns in the database are very long. To generate candidate sequences sequential, and various trimming techniques are implemented to reduce the search spacea depth-first search idea is used. Using a vertical bitmap representation the transactional data is stored, which allows for efficient support counting as well as significant bitmap compression. A salient feature of SPAM algorithm is that it incrementally outputs new frequent item sets in an online fashion [8].

3.1.4 SPADE (Sequential PAttern Discovery using Equivalence classes) [7]: It is an efficient sequential pattern mining algorithm based on vertical format (Zaki, 2001).It utilizes combinatorial properties to decompose the original problem into smaller sub-problems. SPADE use a vertical idlist database format and use a lattice-theoretic approach to decompose the original search space (lattice) into smaller pieces (sub-lattices) which can be processed independently in main-memory. The all sequences are discovered in three database scans, or only a single scan with some information which are Pre-processed. SPADE not only minimizes I/O costs by reducing database scans, but also minimizes computational costs by using efficient search schemes.

3.2. Pattern-Growth Algorithms:

The key idea behind Pattern-Growth Algorithms is to avoid the candidate generation step and emphasize on the search on a limited portion of the initial database. The search space partitioning feature plays an important role in pattern-growth. ThePresentationof the database to be mined is beginning of every pattern-growth algorithm, then proposes a way to partition the search space, and generates as minimum candidate sequences as possible by growing on the already mined frequent sequences, and then applying the apriori property as the search space is being traversed recursively looking for frequent sequences. By using projected databases the early algorithms are started, for example, FreeSpan[8],PrefixSpan [5], with the latter being the most influential.

The Pattern Growth-Based Algorithm Key Features: [6]

3.2.1 PREFIXSPAN- The PrefixSpan (Prefix Projected Sequential pattern Mining) algorithms presented by Jian Pei, Jiavei Han and Helen Pinto [10] is the only projection based algorithms from all the sequencing pattern mining algorithms. It performs better than the algorithm like apriori, freespan, SPADE (vertical data format). This algorithm finds the frequent items by scanning the sequence database once. The database is assigned into several smaller databases which arebased on the frequent item sets. By recursively growing subsequence fragment in every projected database, we got the complete set of sequential pattern. The main concept behind the algorithm of prefixspan is to successfully discovered patterns is employing the divideand- conquer strategy. The prefixspan algorithm requires high memory space as compare to the other algorithms in the sense that it requires creation and processing of huge number of projected sub-databases.

3.2.2FREESPAN- The freespan algorithm deducts the cost require to candidate testing and generation of apriori, with satisfying its basic features [9]. In short, the freespan algorithm uses the frequent items to iteratively project the sequence database into projected database while growing subsequence's frequently in each projected dataset. Every projection will dividethe database and confines further testing to gradually 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 with each iteration.

3.2.3WAP-MINE- This is pattern-growth based algorithm with tree-structure mining technique on its WAP-tree data structure. In this algorithm the sequence database is scanned twice to build up the WAP-tree from the frequent sequences by their support values. To point that where is first occurrence of the each item in a frequent item setthe header table is maintained first, which will be helpful to mine the tree for frequent sequences built up on their suffix. It found in the analysis that the WAP-MINE algorithm have more scalability than GSP and perform bitterly by marginal points. This problem of generating huge candidate as in case of apriori-based approachis avoided by this algorithm and scans the database twice, the WAP-MINE faces the problem of memory consumption, as it iteratively regenerate n increase automatically.

4. COMPARATIVE STUDY OF SEQUENTIAL PATTERN MINING ALGORITHM

This study of the sequential pattern mining algorithm is completed on the basis of their various important features. The comparison sequential pattern mining algorithm is categorized into two broad parts, as apriori based algorithmand pattern growth based algorithm. The every features are used to classify these algorithms are discussed first and then comparison is done for the following algorithms

G.S.P	Generalized sequential pattern. Apriori and BFS based approach, use downward closure property.
SPADE	Use of the equivalence classes for the discover of the sequential pattern. Use lattice-theoretic based approach.
SPAM	Sequential Pattern Mining. Depth-first search based approach, using a vertical bitmap representation for data storage.
FREESPAN	Finding the sequential pattern by projecting the frequent pattern in sequence database. Pattern Growth based method and useprojected sequence database.
PREFIXSPAN	By prefix-projected sequential pattern Mining. Use Prefix heuristic and bi level Projection.
WAPMINE	From the sequential dataset which contains web click in the sequential format by timestamps is Web access pattern mining.
SPIRIT	By formulating the constraint using regular expression the sequential pattern mining. Use regular expression constraints for pruning

4.1 Features of sequential pattern mining algorithms are:

1) Breadth-First Search Based Approach vs Depth First Search Based Approach: In the breadth-first searchtraversal technique level-by-level search is conductedto find the complete set of pattern i.e. All the innernode are processed before moving to the next level. The depth-first search traversal techniqueInstead, Everyinner node should be explored before in the pathmoving to the next one. The depth first search is that itcan reach very quickly to large frequent fragments andtherefore some expansion in the other path in the treecan be avoided.

2) Apriori-Based vs Pattern-growth Based: In this type of apriori based type algorithm the main theme is to candidate-generate and test which uses the downward closer property. If an item set α is frequent, then and then only the superset of α is frequent, otherwise if not be frequent either. The strategy of Pattern-growth takes better approach in creating possible frequent sequences, and uses the divide-and-conquer approach. For the reduce of search space this patterngrowth algorithm do the projection on the database.

3) Top-Down Search vs Bottom-up search: This type of apriori based algorithms uses a bottom-up search by ensuring each single frequent sequence. which means that for the produce a frequent sequence of length 1, all 21 subsequence's have to be generated. From that it can be stated that this exponential complexity is limiting at the apriori based algorithms to find out only short pattern, since they just find the subset infrequent pruning by deleting any candidate sequence for which there exist a subsequence that does not belongs to the set of frequent sequences. In case of the top-down approach the subset of sequential pattern can be mined by generating the relative set of projected databases and mining each recursively for top to bottom.

4) Anti-Monotone vs. Prefix-Monotone Property: According to the property of anti-monotone it statesthat the each non-empty subsequence of the sequentialpattern is a sequential pattern. And in the prefixmonotonestates that every sequence which is having αas a prefix satisfies the constraints if α sequence satisfythe constraint.

5) Regular Expression Constraints: The number of state changes in the relative deterministic finite automata help to calculate the complexity of regular expression constraints. It has the nice property known as growthbased anti-monotonic if it satisfy the following property. The sequence must be reachable by growing from any component which matches the part of the regular expression when it satisfies the constraints first. From our comparative study we found that prefixspan algorithm uses depth-first search based approach. Topdown technique is efficient technique to find frequent subsequence's as sequential pattern from the large database. Also the regular expression constraint and prefix monotone property is use by prefixspan algorithm, which makes this algorithm best choice for applying user defined constraint for mining only concerned sequential pattern.

5. RESEARCH CHALLENGES

Today existing methods are efficiently discovered sequential patterns according to the user requirement. Such patterns are extensively relevant for a large number of applications. But still there are diverse research challenges in this field of data mining. Some of the rigorous research challenges are:

- To be able to incorporate various kinds of user-specificconstraints. [11]
- To study target oriented sequential pattern mining and itsapplication in some real dataset.[12]
- To discover the entire set of patterns, satisfying theminimum support threshold.
- Algorithm should process huge search space andminimize recurring scanning of database during miningprocess.

6. CONCLUSION

The sequential pattern mining and various types ofsequential pattern mining algorithms are discussed In this study paper. This concept is introduced in the year 1995, it has gone through remarkable advancement in few years only. The Initial basic work on this topic is concentrated on improvement of the performance of algorithms by using various data structure algorithms or different representations. So, on the basis of these problems the sequential pattern mining is categorized into two main groups, Apriori approach based algorithms and pattern growth approach based algorithms. From our comparative survey and previous some studies by various researchers on sequential pattern mining algorithms it is found that the algorithm which are based on the approach of pattern growth are better in terms of scalability, time-complexity and space-complexity.

7. REFERENCES

[1] [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] [2] M. Garofalakis, R. Rastogi, and K. Shim, "SPIRIT: Sequential pattern mining with regular expression constraints", VLDB'99, 1999

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

[4] [4] Srikant R. and Agrawal R., ―Mining sequential patterns: Generalizations and performance improvements, Proceedings of the 5th International Conference Extending Database Technology, 1996, 1057, 3-17.

[5] [5] J.Pei, J.Han, B.MortazaviAsl, J.Wang, H.Pinto, Q.Chen, U.Dayal and M.-C.Hsu, “Mining sequential patterns by pattern-growth: The PrefixSpan approach,” IEEE Transactions on Knowledge and Data Engineering, vol.16, no.11, 2004, pp. 1424-1440.

[6] [6] Nizar R. Mabroukeh and C. I. Ezeife, “A Taxonomy of Sequential Pattern Mining Algorithms,” ACM Computing Surveys, Vol. 43, No. 1, Article 3, Publication date: November 2010.

[7] [7]. M. Zaki, “SPADE: An efficient algorithm for mining frequent sequences,” Machine Learning, 2001.

[8] [8] S. Parthasarathy, M. Zaki, M. Ogihara, and S. Dwarkadas, “Incremental and interactive sequence mining,” In Proc. of the 8th Int.Conf. on Information and Knowledge Management (CIKM’99), Nov1999.

[9] [9] Han J., Dong G., Mortazavi-Asl B., Chen Q., Dayal U., Hsu M.-C., Freespan: Frequent pattern-projected sequential pattern mining, Proceedings 2000 Int. Conf. Knowledge Discovery and Data Mining (KDD’00), 2000, pp. 355-359.

[10] [10] J. Pei, J. Han, B. Mortazavi-Asi, H. Pino, "PrefixSpan: Mining Sequential Patterns Efficiently by Prefix- Projected Pattern Growth", ICDE'01, 2001.

[11] [11] Ayres J., Flannick J., Gehrke J., and Yiu T., “Sequential pattern mining using a bitmap representation,” In Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining-2002.

[12] [12] Jian Pei, Jiawei Han, Wei Wang, “Constraint-based sequential pattern mining: the pattern growth methods,” J Intell Inf Syst , Vol. 28, No.2,,2007, pp. 133 –160.