DOI: http://dx.doi.org/10.26483/ijarcs.v12i6.6779

Volume 12, No. 6, November-December 2021

ISSN No. 0976-5697

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

# SEQUENTIAL PATTERN MINING ALGORITHMS – RECENT TRENDS

Sandeep Mukherjee
Section Officer
Department of Home and Hill Affairs, West Bengal
Kolkata, India

Ambar Dutta
Amity Institute of Information Technology
Amity University Kolkata
Kolkata, India

***Abstract:*** Sequential pattern mining is a technique of data mining whose objective is to identify statistically relevant patterns within a database with time-related data. It has a wide range of applications in variety of domains like education, healthcare, bioinformatics, web usage mining, telecommunications, intrusion detection etc. At present, most of the real sequence databases are incremental in nature. So there is a need to explore incremental and distributed pattern mining algorithms. Periodic pattern mining is a technique to discover periodic pattern which may be a pattern that repeats itself after a specific time interval. It has a wide range of applications in weather prediction, stock market analysis, web usage recommendation etc. Moreover, uncertain frequent pattern mining has become a popular research domain among researchers, as many real-life databases at present consist of uncertain and incomplete data. In this paper, a novel attempt is made to incorporate a systematic literature review of state-of-the-art techniques of sequential pattern mining which ranges from incremental pattern mining, periodic pattern mining and uncertain frequent pattern mining. Researchers in the field of pattern mining will find it very useful to get the information about various algorithms of different types of pattern mining.

***Keywords:*** Pattern mining, incremental database, periodic pattern, uncertain database mining

## I. INTRODUCTION

In 1995, Agrawal and Srikant [1] were the first to address the problem of sequential pattern mining. Since then, it has been a popular topic of research for data scientists. It finds statistically relevant patterns from sequential data. Sequential pattern mining should be able to handle various types of data, multi element event set, single element event set, sparse data bases etc. It has a wide range of applications in various fields which include customer shopping pattern analysis, telephone calling pattern, DNA sequence, medical treatment, natural disaster prediction etc.

There are different classes of sequential pattern mining algorithms which include Apriori based, Pattern growth based and early pruning based. The Apriori property states that "All non-empty subsets of a frequent itemset must also be frequent". Breadth first search, multiple scans of the database and Generate & Test are important features of this class of algorithms. AprioriAll, FP Tree, GSP, SPAM, SPADE etc are the popular algorithms which fall under this category. Pattern growth-based algorithms do not need candidate generation for mining frequent pattern efficiently. FreeSpan, PrefixSpan, WAP Mine, FS Miner etc are the popular algorithms which falls under the category.

However, there are few other advances of sequential pattern mining algorithms. These are incremental pattern mining, periodic pattern mining, uncertain frequent pattern mining etc. The philosophy behind incremental pattern mining is that whenever some new data and sequence is added to the database, it is not possible to start from the scratch. Rather the sequential pattern mining algorithms should be developed in such a way that it should be able to adapt to incremental update of the database. Periodic pattern mining is performed over time series data which deals with finding of some temporal regularity. It has many applications including weather prediction, web usage recommendation etc. Uncertain frequent pattern mining algorithm is used to discover meaningful frequent patterns from a database with uncertain data. Missing and misrepresented data may also include uncertainty in the database. The purpose of this paper is to provide a good literature review of the recent advances of sequential pattern mining algorithms.

The remaining part of the paper is structured as follows. Section II deals with incremental pattern mining and some popular algorithms which deal with incremental database. In section III, the concept and different types of periodic patterns are introduced. Various periodic pattern mining algorithms are discussed in sections IV, V and VI. Section VII concentrates on uncertain frequent pattern mining and description of some popular and important algorithms which fall in this category. After analyzing the research findings in section VIII, finally conclusions are drawn and some future directions are mentioned in section IX.

## II. TYPE STYLE AND FONTS

Apriori All [1] was a fundamental algorithm in Data Mining where the initial concern was only individual habits or buying/purchase patterns. It was also observed that maintaining track of periodic patterns is important in an incremental Database where new data is being added it the original Database (say D). Let the increment be 'S' so the new Database becomes (D+δ). It may so happen that the frequent items on the Database D be invalid in (D+δ). The idea is to efficiently identify the patterns that are frequent, from the newly updated database. FASTUP [2] is considered to be the earliest work on Incremental SPM. It is basically an updated GSP algorithm where previous mining results are considered. Therefore, generation and validation of candidate sequences are done using the 'generating-pruning' technique.

There are two approaches to handle incremental database. In the 1st approach initiating from the SPADE algorithm the ISM algorithm [3] which formulates with the help of negative

border the updates and subsequently the database is rewritten. Whereas in the 2nd approach the database was extended by adding new constraints like time and introduction of 'is-a' hierarchy and GSP, a new algorithm was introduced. It was observed that GSP out-performed Apriori-All by 20 times. A new algorithm ISE (Incremental Sequence Extraction) was proposed by I.F. Masseglia, P. Poncelet & M. Tesseirre [4] in a situation where new customers and transaction were introduced to the original database and frequent sequences were identified. It has been empirically observed that in comparison to GSP this new outperforms it by a factor of 2 to 5.

### *A. ISM (Incremental Sequence Mining) Algorithm*

This maintains a sequence lattice termed as ISL (Incremental Sequence Lattice) is maintained by the algorithm, which consists of the sequences that are frequent and also sequence in the original database that are not frequent or belong to the negative border. The lattice also maintains support of all the members. ISM is a two-phase approach. In the 1st phase the support for the elements in 'Frequent Sequences' (FS) and 'Negative Border' (NB) are up dated. This is done by pruning the infrequent sequences from the of 'Frequent Sequence' (FS) set after an update takes place. The lattice and negative borders are updated in 1 space.

In the 2nd phase the new frequent sequence is considered on after the open and adding Negative Border (NB) and 'Frequent Sequence' (FB) beyond phase 1. Here the update is restricted only to the newly added items in the 1st phase. An updated lattice and a new set of frequent sequence and also a new negative border enables to consider the new update by the algorithm at the conclusion of the 2nd phase. It is observed that storage of the itemsets that frequently occur in the lattice is very memory intensive.

### *B. ISE (Incremental Sequence Extraction) Algorithm*

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

This was proposed by Masseglia, Poncelet & Teisserie [4]. By means of this approach the solution to incremental mining for frequent sequences is taken care of by means of previously discovered information. Three types of frequent sequences are considered in order to mine the new frequent sequences. Firstly if enough support exists in an incremental database for a sequence 2ndly the sequences are not in original database (DB) but in the embedded database. Lastly some sequences that become frequent in the composite database DB when the new part (incremental part) 'db' is added.

The various steps of execution are as:

1. The set of frequent sequences is identified in 'DB', the original database.
2. Set of 1-frequent sequences are identified in appended database 'db' and is validated for the unified database U. (U=DB+db)
3. From 1-frequent sequence subsequent candidate sequences are generated from 'db'.
4. From step 3 frequent sequences are obtained and verified in the unified database.
5. Frequent sub-sequences of any group of frequent sequence of the original database (from step 1) are formed by means of an item obtained from frequent-1 sequences in appended database 'db', obtained in step 2.
6. By appending sequences obtained from step 4 and step 5 candidate sequences are obtained.
7. The frequent sequences are obtained from the candidate sequences in step 6 and they are validated in the unified database U.
8. LU the set frequent sequences is the output.

## III. PERIODIC PATTERN MINING

Periodic Pattern Mining is being applied in many areas in present day. It is conceptualized or described as a pattern. That is self-repetitive after a certain time interval. Such patterns may be mined from discrete or continuous databases in time series, social media data, biological reference databases, databases pertaining to space and time. Depending on the count of occurrence, the periodic patterns can be classified into frequent and statistically significant. Perfect and imperfect periodic patterns, asynchronous and synchronous periodic patterns, partial and full periodic patterns, desire and approximate are sub-classifications of periodic patterns. An overview of the various types and sub-types of periodic patterns are stated below:-

### *A. Full and Partial Periodic Patterns*

If every position of a pattern is periodic then such a pattern is full periodic, whereas if one or more than one element does not display periodicity then the periodic pattern is partial, e.g. in the sequence {a}{b}{d}{b}{d}{b}{d} {a}{b}{c} be considered, a sub-sequence {b}{d} is full periodic with periodicity 2. It may be noted that full periodically is exhibited in every position of the sequence. Whereas in the sequence {a}{d}{c}{a}{b}{c}{a}{c}{c}{c}, it may be noted that this subsequence {a}{*}{c} is partial periodic having a period of 3 and also the second element does not show periodically compared to full periodicity, the periodicity of partial periodic patterns is lesser.

### *B. Perfect and Imperfect Periodic Patterns*

Considering a sequence {a}{b}{d}{a}{b}{k}{a}{b}{m} {a} {b} {c}, {a} {b} {x} is a Perfect Periodic Pattern in it. It is noted that {a} {b} {x} has occurred 4 times – from its occurrence for 1st time till last at the closing of sequence. It may so happen and expected occurrence of a patterns here {a} {b} {*} may not occur the periodicity is termed as imperfect. As an example it may be noted that periodic pattern {a} {b} {x} can be considered an imperfect in a sequence {a} {b} {c} {d} {g} {h} {a} {b} {i} {a} {b} {x} with ¾ as the measure of confidence on the periodicity of {a} {b} {*} – its occurrence was expected at 1 more position.

### *C. Synchronous and Asynchronous Periodic Patterns*

Synchronous Periodic Patterns occurs without any misalignment on a periodic manner. Subsequences {a}{*}{d}, {a}{*}{*}, {*}{*}{d} are examples of Synchronous Periodic Patterns in a sequence {a}{b}{d}{a}{c}{a}{c}{d}{d}{e}{b}. It is observed that there is 3 time consecutive repetition of these patterns having 3 as the period. However, in Asynchronous patterns there are some disturbance in the repetition of the pattern.

### *D. Approximate Periodic Patterns*

Inaccuracy and noise are present in majority of real world data. A case for example is gene mutations in a DNA sequence. So the necessity of mining approximate patterns

arises of the algorithms discuss in foregoing see lions which were tolerant to insertion and data noise and replacement would be capable of mining approximate repeating periodic patterns.

### *E. Surprising Periodic Patterns*

In the field of seismic data, credit card transactional data, audit trail data, genetic & genomic data some patterns may be observed with less frequency but at the same time they are not expected as well.

## IV. FULL AND PARTIAL PERIODIC PATTERN MINING ALGORITHMS

An algorithm [5] was proposed for identifying the reoccurring cyclic association rules that may occur in every cycle of a time services. Cyclic association rules are identified from transaction data bearing time-stamps by this algorithm. Here a fixed length time period is the input. The requirement wise periodic patterns may be full or partial. Minimum confidence threshold for occurrence in a time-series database satisfied by the full or partial segment-wise period patterns. The cyclic association rules have been observed to repeat without fail with confidence of 100% in the time series.

Another algorithm called Max-Sub Pattern Hit Set [6] was developed where in two scans of the times series a max-sub pattern tree was created to mind both the full and partial imperfect periodic pattern. Thereafter an extension of the algorithm enabled it in identifying patterns for a given set of time period the duration of which has been pre-stated at the time of input.

Elfeky M. G et al. [7] proposed an extension of the Max-Sub Pattern Hit Set. Algorithms for Merge-mining, on-line mining and incremental mining of the partial periodical patterns were done. This enables the users to mine the partial periodic patterns at the time of deletion and insertion of the time series databases in case of the incremental version of the Max-Sub Pattern Hit Set algorithm. For this one extra scan of the original database is necessary. In the online version the thresholds can be changed by the users of the algorithm even at the time of its execution. The patterns are mind from 2(two) or more databases and then they are merged in merge mining. The discovery of patterns is facilitated by the merging process and this can be done without any necessity for execution of the algorithm again.

Chen et al [8] proposed an algorithm to find patterns of given periodically. This was done by using the database of the encoded pattern segments and the Frequency Pattern tree. The step where the encoding takes place is the most effective feature of this algorithm. For mining the frequent periodic patterns any frequent pattern mining algorithm can be used. This follows the step of encoding the period segment in the manner as described in the algorithm. Also different minimum supports for different events may be used instead of uniform minimum support for different events.

Another approach, similar to the one stated above [8], was developed by J. Yang et al. [9]. In this algorithm, mining a partial periodic pattern was proposed which is capable of efficiently mining each and every such periodic pattern that takes place in the specified period. By assigning a value to the period, the algorithm partitions the “even sequence” into slots occurrence in which has length whose value is equal to the period. The slots are then encoded to event implies in a manner similar to [8].

## V. ALGORITHMS FOR MINING PARTIAL PERIODIC PATTERNS WITH UNKNOWN PERIODS

All the aforementioned Partial Periodic Pattern Algorithms discussed in the previous discussion have one thing in common – the user provides the period as an input. Another class of Periodic Pattern Algorithm may have patterns having repetition of unanticipated periods. So specialized algorithm are required having the ability to identify the existing potential patterns within the periods.

Ground breaking work for identifying the partial period patterns with unknown period have been done. Two algorithms were proposed. First one [10] was called Period First - where the potential periods are identified in the first stage, thereafter the inter-event association for events occurring on that period is identified. Chi-square method helps on identification of potential periods. The next method also termed association first discovers association within the events. The periodicity at the initial stage for individual temporal association is calculated. Partial periodic patterns can be identified by these algorithms though of deletion and insertion of noise may be there and replacement may take place. Among the short comings is that it may not be able to identify some of the valid periods as the algorithm may consider the consecutive time instants as possible periods.

Researchers from diversified fields as digital signal processing, statistics and economics have also developed algorithms identifying periodicity in time series data. Auto correlation function, FFT (Fast Fourier Transformation), DWT is some of the techniques that have been the basis of such algorithms. Mining was secondary in such research, where the prime focus was identifying periodicity.

An algorithm was designed by Berberidis et al. [11] which employed auto-correlation for detection of candidate periods from discrete time-series values. Thereafter patterns were identified by applying max sub pattern hit set algorithm. A size ‘n’ binary vector is created by scanning the algorithm once. This is done for the entire set of symbols of the alphabet in time series. The algorithm also calculates the circular auto-correlation vector that also has the frequency count for each period for every symbol in the alphabet. Indyk et al. [12] also devised an algorithm for detecting patterns in a noisy time series. This was called Representative Trends which detects the candidate periods termed as relaxed periods.

There are a few convolution based approaches that are capable of identifying the possible time periods and also the periodic patterns in a single pass over the data on time series. Some well know ones are STAGGER [13], WARP [14] etc. Huang et al. [15] devised an algorithm for prediction of spectrum occupancy in wireless, which is based on frequent partial periodic pattern tree using prediction based on this algorithm enables an unlicensed user to avail of the licensed wireless spectrum bands that are not utilized. This ensures better channel utilization lowers rates of collision.

Based on imperfect calendar based periodic patterns, Dutta and Mahanta [16], [17] devised algorithms to mine periodicities of different time frames – year, month, day and hour-wise from both continuous and discrete time series data. Periodicities based on Calendar are identified as an interval based temporal pattern based on interval. Such patterns occur over a series of time-intervals in the continuous and descript domain.

## VI. OTHER PERIODIC PATTERN MINING ALGORITHMS

### *A. Asynchronous Periodic Patterns*

A few algorithms like SMCA [18], E-MAP [19], LSI [20], and OEOP [21] have been proposed for mining Asynchronous Periodic Patterns. Longest Subsequence Identification (LSI) algorithm was proposed for Asynchronous Periodic Pattern mining. The longest sub-sequence containing each asynchronous periodic pattern is detected by the algorithm. It employs a level-wise search-based approach in an iterative manner. The approach has 3 (three) phases. The 1st (first) phase scans of the entire sequence once for detection of every possible asynchronous period for all events. Every candidate '1-patterns' are validated in the 2nd (second) phase. By employing level-wise approach in an iterative manner in the 3rd phase candidate '1-patterns' are develop from the (i-1) patterns in the i-th iteration. The sequence can be scanned once for validating to i-patterns. Multiple seams are necessary for the 2nd & 3rd phases of the LSI approach.

A 4-phase algorithm, the Simple Multiple Complex and Asynchronous periodic pattern Miner (SMCA) can identify all the subsequences with asynchronous period pattern, '1-patterns' are found by the 'SP Miner'- the 1st Algorithm '1-patterns' containing concurrently occurring events are defect by MP-Miner. The complex patterns (i-pattern for i = 2) are detected by the 3rd Algorithm the CP Miner. The 4th Algorithm – the AP Miner is applicable for all complex patterns,

The linked list structure is used by One Event One Pattern (OEOP) Algorithm for defection of a single event 1-pattern scanning the sequence only once. By means of this algorithm the 1st pattern of the SMCA algorithm can be replaced when the data set is data streams. E-MAP algorithm is further improvement over SMCA algorithm for 1-patterns with simultaneous events, 1-pattern, complex patterns. This can also be done in a single scan and a single step.

It may be noted that of the few algorithms discussed above, only LSI is capable of dealing with event sequences where the others can handle event set sequences. All the above algorithms are capable of tackling insertion, replacement and deletion noise that may be present in the sequence.

### *B. Approximate Periodic Patterns*

Some important algorithms that permit approximation in the occurrence positions of the patterns and their structure as well are mentioned below.

Periodic Patterns of Approximate type were mined from hydrological time-series data by Y. L. Zhu et al. [22]. Dynamic candidate period intervals adjusting techniques in combination with a modified suffix tree traversal and representation approach was employed for mining all patterns taking into consideration all periods. An alternative approach proposed by Amir et al. [23] for mining approximate patterns – had an input sequence with parameter $\xi \in [0, 1]$. This approach mines for the longest E-relative error periodic patterns for each unique eve. A method for identifying periodic patterns where there is a gap for DNA sequences was stated by Zhang et al. [24].

### *C. Surprising Periodic Patterns*

J. Yang et al [25] for the first time proposed an algorithm for these rare and unexpected but the less statistically significant patterns. If the frequency of occurrence of a pattern is higher than its anticipated frequency then such a pattern is statistically significant. In place of support model a newer model termed as information model is used for identification these rare but 'statistically' important patterns. Pattern pruning which is based bounded information gain helps in improving efficiency of the Algorithm.

Also an improvement of this method is has concluded that pattern exhibiting consecutive repetition are statistically more significant than scattered repetition Infominer+ uses the measure of Generalized Information Gain. The algorithm is resilient to replacement noise and detects the periodic patterns that are statistically significant.

## VII. UNCERTAIN FREQUENT PATTERN MINING

Frequent Pattern Mining (FPM) finds applications in diverse real life environments like government, commerce & business and technology and Science which includes varied fields like bio-informatics, metrology, marketing and product promotion, epidemiology finance, etc. In many of these fields the use and application of uncertain data has been widely used. The principle causes of uncertain data are (i) incomplete perception of the reality (ii) the events have not been observed and measured with a fully capable device and (iii) data collection data storage and data-analysis is hampered by resource constraints. Observations have shown that data collected for manufacturing systems, environment observations or security by means of optical radiations, electromagnetic, thermal sensors etc. are usually noisy. Uncertainty of observations in these data can be introduced by sampling frequency of sensors, alteration of values as a result of very quick change, inaccuracies that creep in due to inheritance of errors from earlier measurement, transmission errors in a wireless network or environment, latency in network due to delayed arrival of data etc.[26][27]

Additionally there may exist uncertainty over survey data for example '1' may be interpreted as 'l' or 'I', '0' (Zero) may be interacted as alphabet 'O' or vice-versa. Also uncertainly may be due granularity in taxonomy (e.g. A name of a town may be confused with a province). Sometimes for prescription of privacy of data values - employing disturbance and aggregation actual data values may be blurred for preservation of data anonymity.

Various categories of algorithms have emerged in this field. They are

- Cluster analysis of uncertain data [28].
- Detection of outliers in the uncertain data [28], [29].
- Classification of uncertain data [30], [31].

Applications of different techniques like fuzzy set theory or rough set theory have been made in the field of Uncertain Frequent Pattern Mining (UFPM) of these the most popular is the probability theory-based approach. The Probabilistic Model is a very important approach for UFPM.

There are 2 situations for an item x in a transaction:

- A possibility P1 such that x is present in ti
- Alternative possibility P2 such that x is not present in ti

If P(x, ti) is probability that P1 is true, then 1 – P(x, ti) is the probability that P2 is true. To extend this further it is seen that in a probabilistic Dataset of Day uncertain data there may be 'm' individual items; there may be a total of 'q' items that are independent (including some multiple instances of same items).

### *A. Candidate Generate–And–Test Based UFPM*

Chui et al. [32] was the proponent of the U-Apriori algorithm. The expectation of the support of all items in the domain is calculated first. Frequent 1-item are those whose expected support $\geq$ 'minsup'. The algorithm is now applied successively in a 'generate and test' manner to produce (k+1) items from k-itemsets, which occurs frequently. Those (k+1)

itemsets are tested for frequent-ness. U-Apriori verifies that subsets of a frequent pattern must compulsorily be frequent.

The LGS-timing strategy (Local, Global and Single pass patch-up) is incorporated to improve efficiency. By employing this technique the items with a lower existential property are trimmed off, from D - the original Dataset. The result is Dtrim. This implies that for pattern X frequent in Dtrim must also be frequent in D.

By an extra single-pass scanning D and by verification of the expected support of the possible frequent pattern, U-Apriori can discover the mining frequent patterns. Despite LGS-trimming strategy there are some problems faced by it. An overhead occurs at the formation of Dtrim. Efficiency of the algorithm depends in the ratio of the items with less 'existential probabilities'. Moreover, it is hard to decide on a correct value of the trimming threshold. Subsequently this algorithm was further improved by Chui and Kao [33] by applying 'decrement pruning technique'.

### *B. Tree-based UFPM*

It is a substitute to Apriori based mining. Since it avoids Apriori-based mining thus avoids generation of numerous candidates. This approach involves Depth First Divide and Conquer for mining from the tree structure of the probabilistic set of Frequent Patterns. Some of the popular approaches are described below.

UF-Growth was proposed by Leung C.K.S. [34] and is analogous to FP-growth (a tree-based algorithm for precise data). For Uncertain approach for a pattern X – the summation of the existential probability of every item in X and its product with the occurrence count gives the expected support of X. So each node of UF-Tree comprises of an item, the existential probability of that item and the count of occurrence. An UF-Tree is constructed like an FP-Tree. For reducing the memory usage, the UF-growth employs two processes.

- Existential property of each node is assigned a discrete value. This reduces the number of possible existential probability.
- The FP-tree generation is limited to top 2-levels – global tree of the original Dataset D and for all the frequent items.

For the uncertain data the existential probabilities are captured by the UF-Tree quiet correctly. Neither false negatives nor false positives are produced by the UF-Tree. At the same time it not as compact as FP-Tree.

The UFP-Algorithm [35] was proposed by Aggarwal et al. with the idea to reduce the tree size by diminishing the node count of a tree. The probabilistic data of Uncertain Mining is scanned two times like the UF growth and the UF-tree is constructed. A mega node cluster is formed by grouping of nodes similar to existential probability. The mega-node stores the count of the occurrence, the value and the magnitude of the greatest existential probability. The UPF-Growth discovers the entire frequent pattern set in the second scan. Simultaneously the algorithm detects several infrequent patterns (termed as 'false positives') along with truly frequent ones ('true positives'). This necessitates a scan of the uncertain data for the third time for pruning the false positives.

CUP-Growth algorithm was proposed by Leung and Tanbeer [36] with an idea of reducing the size of the tree further, by reduction of the of tree node count. A new term 'transaction cap' was introduced for each transaction ti, defined as, the product of two of the maximum existential probability of items in the transaction ti. Like the UFP-algorithm growth scan, the CUP-growth also scans the probabilistic data set three times. Second scan produces the 'false negatives' along with the 'truly positive' patterns, so a third scan is necessary which prunes the 'false positives'.

A term 'prefixed item cap' was introduced by Leung and Tanbeer [37]. The PUF-Tree captures the prefixed item caps and stores them in the PUF-Tree structure. This PUF-growth algorithm mines the uncertain frequent pattern based on PUF-tree. PUF-growth also undertakes three scans of the uncertain Probabilistic Database, alike the UPF and CUP-growth algorithms. The prefixed item-cap is computed in the first scan. The PUF-tree is built by scanning the probabilistic data set of uncertain data in the second scan. The PUF-tree now contains all potential frequent pattern including truly frequent and several infrequent patterns termed as the 'false positives'. To eliminate the false positives and determination of the final result the dataset is scanned for the third time by the algorithm.

Vertical Data Mining may be applied where every dataset can be considered as a set of items and a list of related transaction IDs. By determining the common list of Transaction IDs of items within X, the support of X can be found.

Calders et al. [38] initiated the concept of 'possible world' of datasets for getting instantiated samples of the database (which were actually data) and in these Eclat algorithm was applied. This made way for the new algorithm U-Eclat. Arbitrary random number is generated by the U-Eclat algorithm for each item x in transaction ti, in a probabilistic dataset D of uncertain data. 'x' is instantiated and is a part of the sampled precise database if the existential probability of x, P(x, ti) in a transaction ti is not lower than the random number 'r' (r <= P(x, ti)). The ECLAT algorithm then mines the sampled precise database. Repeated sampling instantiation process results in multiple sampled databases that can be labelled precise. Average support of X over the multiple sampled databases gives the support count of any arbitrary pattern 'X'. It has been observed that the efficiency of the algorithm comes at the cost of accuracy. More instantiations make accurate results, which however has a trade-off - the execution time.

Budhia et al. [39] proposed UV-ECLAT algorithm for direct mining the frequent patterns avoiding the instantiations from probabilistic datasets of uncertain data, which were stored in a vertical manner. Here information other than recording transactions containing 'x' in set of Transaction IDs are also looked for - like if in any transaction ti, x is probably present, then the existential probability P( x, ti ) - that expresses the chance of x being found in transaction ti has to be stored. Using augmented t-id sets for vertically representing the probabilistic dataset comprising uncertain dataD. Any the support of any 1-itemset {x} in the uncertain data D can be calculated. The accompanying figure illustrates the UV-Eclat algorithm.

Leung et al. [40] proposed U-VIPER algorithm where a set of fixed sized vectors represent D, the set of uncertain data, vertically. Each vector foe each domain item x. Each vector has a definite length. Its value is the same as the count of transactions, which means | D | = n. a Boolean value is used to indicate if x vector is contained in transaction ti. Additionally the associated existential probability P(x, ti) also has to be stored. The Boolean value 1 is substituted by the U-Viper algorithm by P(x , ti) as the i-th element of the vector x of the domain x. the i-th element or x [i] stores P(x, ti) if x is present in ti else '0' if x is absent in ti.

## VIII. ANALYSIS AND DISCUSSION

The static datasets comprising the Uncertain data are mined by U-Apriori, U-Eclat,U-Viper, UF-Growth, CUF-Growth, PUF-Growth, UPF-Growth and UV-Eclat. Horizontal mining technique is employed by the first five algorithms mentioned

whereas the rest mines vertically. In majority of the Uncertain Frequent Pattern Mining Algorithms discussed, the mining expected patterns with expected support can be done satisfactorily. However U-Eclatmay return false positives or false negatives. Accuracy improves with more samples.

UF-Tree and U-Apriori manages memory better in comparison to other approaches. The compactness of CUF-Tree and PUF-Tree approaches are comparable to FP-Tree. The memory consumption is also influenced by density and minsup. It has been empirically seen that for different values of 'minsup' different algorithms have performed well. Most algorithms exhibit good performance in presence of probabilistic itemsets having low existential probability. As such datasets do not result in lengthy frequent patterns. The run-time of an algorithm is additionally dependent on density of datasets. More dense data implies less time for traversal of datasets.

## IX. CONCLUSIONS

As a data mining technique, sequential pattern mining deals with discovery of statistically significant patterns from temporal databases. It can be applied to various domains such as healthcare, education, bioinformatics, telecommunication, business etc. There exists a wide range of algorithms for sequential pattern mining. However, most of these algorithms are not well-suited for incremental database where the size of the database is continuously growing, like real-life databases. Sometimes, it is essential to find periodic patterns from a temporal database which are not addressed by standard and well-accepted sequential pattern mining algorithms. Detection of periodicity is very challenging for a time-series data. Moreover, the popular algorithms failed to identity frequent pattern from uncertain and incomplete database. So, there is a need to explore different types of algorithms which can take care of these adverse conditions.

In this paper, an extensive literature survey is presented for incremental pattern mining, periodic pattern mining and uncertain frequent pattern mining algorithms. The purpose of the present paper is to provide with the researchers in the domain will get a significant information about the latest trends of algorithms in the domain of sequential pattern mining. Sequential patterns are classified into different categories and a number of algorithms are discussed in each category. The future scope for the present survey will be to implement some of the algorithms of each class and perform a detailed comparative analysis of the algorithms. Moreover, researchers may focus on discover some hidden patterns in sequence databases.

## X. REFERENCES

[1] Srikant R. and Agrawal R., "Mining sequential patterns: Generalizations and performance improvements", in Proceedings of the 5th International Conference on Extending Database Technology: Advances in Database Technology, pp 3–17, 1996.

[2] Lin M. Y, Lee S. Y, "Incremental update on sequential patterns in large databases", in Proceedings of Tenth IEEE International Conference on Tools with Artificial Intelligence, Taipei, Taiwan, pp 24–31, 1998.

[3] Parthasarathy S., Zaki M., Ogihara M., Dwarkadas S., "Incremental and Interactive Sequence Mining", in Proceedings of the Eighth International Conference on Information and Knowledge Management, Kansas City, MO, USA, pp 251-258, 1999.

[4] Masseglia, F., Poncelet, P., &Teisseire, M., "Incremental mining of sequential patterns in large databases", Data and Knowledge Engineering, Vol. 46, Iss. 1, pp 97-121, 2003.

[5] Ozden B., Ramaswamy S., Silberschatz A., "Cyclic association rules", In Proceedings 14th International Conference on Data Engineering. Orlando, Florida, pp. 412-421, 1998

[6] Han J., Gong W., Yin Y., "Efficient mining of partial periodic patterns in time series databases", in Proceedings 15th International Conference on Data Engineering, Sydney, NSW, Australia, pp. 106-115, 1999.

[7] Elfeky M. G., Aref W. G., Elmagarmid A .K., "Incremental, online and merge mining of partial periodic patterns in time series databases", IEEE Transaction on Knowledge and Data Engineering, Vol. 16, No. 3, pp. 332-342, 2004.

[8] Chen S.S., Huang T.C.K., Lin Z.M., "New and efficient knowledge discovery of partial periodic patterns with multiple minimum supports" , Journal of Systems and Software, Vol. 84, Iss 10, pp. 1638-1651, 2011.

[9] Yang J., Wang W., and Yu P. S., (2003). "Mining Asynchronous Periodic Patterns in Time Series Data", IEEE Transaction on Knowledge and Data Engineering, Vol. 15, Iss 3, pp. 613-628, 2003.

[10] Ma S., Hellerstein J., "Mining Partially Periodic Event Patterns with Unknown Periods", in Proceedings 17th International Conference on Data Engineering, Heidelberg, Germany, 2001.

[11] Berberidis C., Aref W., Atallah M., Vlahavas I., Elmagarmid A., "Multiple and Partial Periodicity Mining in Time Series Databases" in Proceedings of the 15th European Conference on Artificial Intelligence, pp 370–374, 2002.

[12] Indyk P., Koudas N., Muthukrishnan S., "Identifying representative trends in massive time series datasets using sketches", in Proceedings of the 26th International Conference on Very Large Databases, 2000.

[13] Elfeky M. G., Aref W. G., Elmagarmid A. K., "STAGGER: Periodicity Mining of Data Streams using Expanding Sliding Windows", in Proceedings of the Sixth International Conference on Data Mining, Hong Kong, China, pp 188-199, 2006.

[14] Elfeky M.G., Aref W.G., Elmagarmid A.K., "WARP: Time Warping for Periodicity Detection", in Proceedings of the Fifth IEEE International Conference on Data Mining, Houston, TX, USA, 2005.

[15] Huang P., Liu C. J., Xiao Li, Chen J., "Wireless Spectrum Occupancy Prediction Based On Partial Periodic Pattern Mining", in Proceedings of the IEEE 20th International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems, Washington, DC, USA, pp. 51-58, 2012.

 

[16] Dutta M., Mahanta A. K., "Mining Calendar-Based Periodicities of Patterns in Temporal Data", in Proceedings of the 3rd International Conference on Pattern Recognition and Machine Intelligence, pp 243-248, 2009.

[17] Dutta M., Mahanta A. K., "Detection of Calendar- Based Periodicities of Interval-Based Temporal Patterns", International Journal of Data Mining & Knowledge Management Process, Vol.2, No.1, pp 17-31, 2012

[18] Huang K. Y., Chang C.H., "SMCA: A General Model for Mining Asynchronous Periodic Patterns in Temporal Databases", IEEE Transactions on Knowledge and Data Engineering, Vol. 17, Iss. 6, pp 774–785, June 2005.

[19] Maqbool F., Bashir S., and Baig A.R., "E-MAP: Efficiently Mining Asynchronous Periodic Patterns", International Journal of Computer Science and Network Security, Vol. 6, No. 8A, Aug, 2006.

[20] Yang J., Wang W., and Yu P. S., "Mining Asynchronous Periodic Patterns in Time Series Data", IEEE Transactions on Knowledge and Data Engineering, Vol. 15, Iss. 3, pp 613–628, March 2003.

[21] Yeh J. S., Lin S. C., "A New Data Structure for Asynchronous Periodic Pattern Mining", in Proceedings of the 3rd International Conference on Ubiquitous Information Management and Communication, New York, USA, pp 426-431, 2009.

[22] Zhu Y. L., Li S. J., Bao N. N., Wan D. S., "Mining approximate periodic patterns in hydrological time series", Journal of Computational Information Systems, Vol. 15, pp 6131-6144, 2013.

[23] Amir A., Apostolico A., Eisenberg E., Landau G. M., Levy A., Lewenstein N., "Detecting approximate periodic patterns", in Proceedings of the First Mediterranean conference on Design and Analysis of Algorithms, pp 1–12, 2012.

[24] Zhang M., Kao B., Cheung D. W., Yip K. Y., "Mining Periodic Patterns with Gap Requirement from Sequences", Journal of ACM Transactions on Knowledge Discovery from Data, Vol. 1, Iss. 2, 2007.

[25] Yang J., Wang W., Yu P., "InfoMiner+: Mining Partial Periodic Patterns with Gap Penalties" in Proceedings of the Second IEEE International Conference Data Mining, Maebashi City, Japan, 2002.

[26] Ren, J., Lee, S.D., Chen, X., Kao, B., Cheng, R., Cheung, D., "Naive Bayes classification of uncertain data", in Proceedings of the Ninth IEEE International Conference on Data Mining, Miami, Florida, pp. 944-949, 2009.

[27] Xu, L., & Hung, E., "Improving classification accuracy on uncertain data by considering multiple subclasses". in Proceedings of the Twenty-Fifth Australasian Joint Conference on Artificial Intelligence, Sydney, Australia, pp 743-754, 2012.

[28] Aggarwal, C. C., "Outlier Analysis", Springer-Verlag New York, 2013.

[29] Aggarwal, C. C., Yu, P. S., "Outlier detection with uncertain data", in Proceedings of the SIAM International Conference on Data Mining, SDM, pp 483–493, 2008.

[30] Aggarwal, C. C. (ed.), "Managing and Mining Uncertain Data", Springer, Boston, MA, 2009.

[31] Aggarwal, C. C., Yu, P. S., "A survey of uncertain data algorithms and applications". IEEE Transactions on Knowledge and Data Engineering, Vol. 21, Iss. 5, pp 609–623, 2009.

[32] Chui, C.-K., Kao, B., Hung, E., "Mining frequent itemsets from uncertain data". in Proceedings of the 11th Pacific-Asia conference on Advances in knowledge discovery and data mining, pp 47–58, 2007.

[33] Chui, C. K., Kao, B., "A decremental approach for mining frequent itemsets from uncertain data". in Proceedings of the 12th Pacific-Asia conference on Advances in knowledge discovery and data mining, pp 64–75, 2008.

[34] Leung, C. K. S., "Mining uncertain data", Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery (WIDM), Vol. 1, No. 4, pp 316–329, 2013.

[35] Aggarwal, C. C., Li, Y., Wang, J., Wang, J., "Frequent pattern mining with uncertain data", in Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pp 29–38, 2009.

[36] Leung, C. K. S., Tanbeer, S. K., "Fast tree-based mining of frequent itemsets from uncertain data". in Proceedings of the 17th international conference on Database Systems for Advanced Applications - Volume Part I, pp 272–287, 2012.

[37] Leung, C. K. S., Tanbeer, S. K., "PUF-tree: a compact tree structure for frequent pattern mining of uncertain data", in: Pei J., Tseng V.S., Cao L., Motoda H., Xu G. (eds) Advances in Knowledge Discovery and Data Mining. PAKDD 2013. Lecture Notes in Computer Science, vol 7818, Springer, Berlin, Heidelberg, pp 13-25, 2013.

[38] Calders, T., Garboni, C., Goethals, B., "Efficient pattern mining of uncertain data with sampling". in: Zaki M.J., Yu J.X., Ravindran B., Pudi V. (eds) Advances in Knowledge Discovery and Data Mining. PAKDD 2010. Lecture Notes in Computer Science, vol 6118, Springer, Berlin, Heidelberg, pp 480-487, 2010.

[39] Budhia, B. P., Cuzzocrea, A., Leung, C. K. S., "Vertical frequent pattern mining from uncertain data". Frontiers in Artificial Intelligence and Applications, IOS Press, Volume 243, pp 1273-1282, 2012.

[40] Leung, C. K. S., Tanbeer, S. K., Budhia, B. P., Zacharias, L. C., "Mining probabilistic datasets vertically", in Proceedings of the 16th International Database Engineering & Applications Symposium, pp 199–204, 2012.

 