A Review Paper on Sequential Pattern Mining Algorithms

[1] Aashka Shah, [2] Krunal Panchal
[1]M.E. (Perusing), [2]Faculty
[1]Department of computer engineering
[1]L.J.I.E.T, Ahmedabad, India.

***Abstract* - Sequential pattern mining and sequential rules mining are important data mining task for wide application. Its use to find 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. Sequential pattern mining is used in various domains such as medical treatments, natural disasters, customer shopping sequences, DNA sequences and gene structures. The problem is to discover the all sequential pattern who satisfy the user specified constraint, from the given sequence database. Sequential pattern mining algorithm are mainly classified in to two part. Apriori based and pattern growth based algorithm. These are basic sequential pattern mining algorithm and mine full set sequential pattern mining algorithm which mean these algorithm are generate all frequent sequential pattern.**

***Index Terms* - Sequential pattern mining, Sequence database, apriori based algorithms, pattern growth based algorithms.**

I. Introduction

Sequential Pattern Mining finds interesting sequential patterns among the large database. It finds out frequent subsequences as patterns from a sequence database. With massive amounts of data continuously being collected and stored, many industries are becoming interested in mining sequential patterns from their database. Sequential pattern mining is one of the most well-known methods and has broad applications including web-log analysis, customer purchase behavior analysis and medical record analysis. In the retailing business, sequential patterns can be mined from the transaction records of customers. For example, having bought a notebook, a customer comes back to buy a PDA and a WLAN card next time. The retailer can use such information for analyzing the behavior of the customers, to understand their interests, to satisfy their demands, and above all, to predict their needs. Another example of sequential patterns is that in a book store's transaction database history, 80% customers who brought the book Database Management typically bought the book Data Warehouse and then brought the book Web Information System with certain time gap. All those books need not to be brought at the same time or consecutively, the most important thing is the order in which those books are brought and they are bought by the same customer. 80% here represents the percentage of customers who use this purchasing habit.

Basic Concepts of Sequential Pattern Mining:

1. Let $I = \{x1, . . . , xn\}$ be a set of *items*, each possibly being associated with a set of *attributes*, such as value, price, profit, calling distance, period, etc. The value on attribute *A* of item *x* is denoted by *x.A*. An *itemset* is a non-empty subset of items, and an itemset with *k* items is called a *k-itemset*.

2. A *sequence* $\alpha = <X1 \cdot \cdot \cdot Xl>$ is an ordered list of itemsets. An itemset Xi $(1 \le i \le l)$ in a sequence is called a *transaction*, a term originated from analyzing customers' shopping sequences in a transaction database. A transaction *Xi* may have a special attribute, *time-stamp*, denoted by *Xi.time*, which registers the time when the transaction was executed. For a sequence $\alpha = <X1 \cdot \cdot \cdot Xl>$, we assume $Xi.time < Xj.time$ for $1 \le i < j \le l$.

3. The number of transactions in a sequence is called the *length* of the sequence. A sequence with length *l* is called an *l-sequence*. For an *l*-sequence *α*, we have $len\ (\alpha) = l$. Furthermore, the *i*-th itemset is denoted by $\alpha[i]$. An item can occur at most once in an itemset, but can occur multiple times in various itemsets in a sequence.

4. A sequence $\alpha = <X1 . . . Xn>$ is called a *subsequence* of another sequence $\beta = <Y1 . . .Ym>$ $(n \le m)$, and *β* a *super- sequence of α, if there exist integers* $1 \le i1 < . . < in \le m$ *such that X1 Yi1 , . . . , Xn Yin.*

5. A sequence database SDB is a set of 2-tuples (sid, α), where sid is a sequence-id and α a sequence. A tuple (sid, α) in a sequence database SDB is said to contain a sequence γ if γ is a subsequence of α. The number of tuples in a sequence database SDB containing sequence γ is called the support of γ, denoted by sup (γ). Given a positive integer min_sup as the support threshold, a sequence γ is a sequential pattern in sequence database SDB if sup (γ) ≥ min_sup. The sequential pattern mining problem is to find the complete set of sequential patterns with respect to a given sequence database SDB and a support threshold min_sup.

II. TAXONOMY OF SEQUENTIAL PATTERN MINING[2]

Sequential Pattern Mining Algorithms mainly differ in two ways:

1. The way in which candidate sequences are generated and stored. The main goal here is to minimize the number of candidate sequences generated so as to minimize I/O cost.

2. The way in which support is counted and how candidate sequences are tested for frequency. The key strategy here is to eliminate any database or data structure that has to be maintained all the time for support of counting purposes only. Based on these criteria's sequential pattern mining can be divided broadly into two parts:

- Apriori Based
- Pattern Growth Based

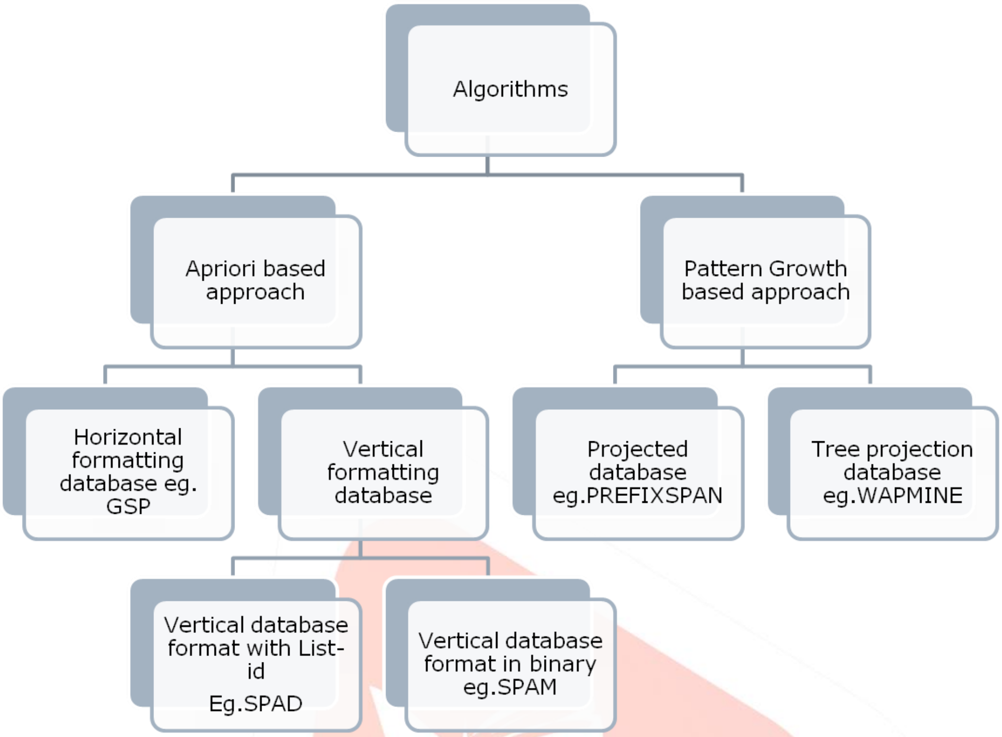
Algorithms
Apriori based approach
Pattern Growth based approach
Horizontal formatting database eg. GSP
Vertical formatting database
Projected database eg.PREFIXSPAN
Tree projection database eg.WAPMINE
Vertical database format with List-id Eg.SPAD
Vertical database format in binary eg.SPAM

Fig.1 Taxonomy of Sequential Pattern Mining

A. APRIORI BASED ALGORITHMS:

Apriori algorithm work on candidate "generate and test" property. Apriori 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]

Generalized sequential pattern (GSP) is Apriori based algorithm [3]. 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 Apriori all algorithm. 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 candidate 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.

2. SPADE[Sequential pattern discovery using equivalence classes]

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.[4] 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[Sequential Pattern Mining]

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

[5]. In SPAM assume 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 ALGORITHMS:

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 Apriori for recursively looking frequent sequence. Pattern growth based algorithm focus on features like search space partition, tree projection, depth first traversal, candidate sequence pruning.[6]

1. PrefixSpan

The PrefixSpan (Prefix Projected Sequential pattern Mining) algorithms presented by Jian Pei, Jiavei Han and Helen Pinto [7] 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 projected into several smaller databases according to the frequent items. By recursively growing subsequence fragment in every projected database, we got the complete set of sequential pattern. The main concept behind the prefixspan algorithm to successfully discovered patterns is employing the divide and 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.

2. WAP-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. Here header table is maintained first to point that where is first occurrence of the each item in a frequent item set which can 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. Although this algorithm scans the database twice only and avoids the problem of generating huge candidate as in case of apriori-based approach, the WAP-MINE faces the problem of memory consumption, as it iteratively regenerate n increase automatically.

III. COMPARATIVE ANALYSIS OF SEQUENTIAL PATTERN MINING ALGORITHM

Comparative analysis of sequential pattern mining algorithm is done on the basis of their various important features. For comparison sequential pattern mining is divided into two broad categories, namely, Apriori Based and Pattern Growth Based Algorithms. All the nine features used to classify these algorithms are discussed first and then comparison is done for the following algorithms:
G.S.P.: Generalized sequential pattern.
SPADE: Use of the equivalence classes for the discover of the sequential pattern.
SPAM: Sequential Pattern Mining.
PrefixSpan: By prefix-projected sequential pattern mining.
WAPMINE: Web access pattern mining from sequential dataset which contains web click in the sequential format by timestamps.

Table 1: Comparative study of Sequential Pattern mining Algorithms

No.	Algorithm	Features
1	GSP	Apriori based, Bottom to up search, BFS based approach, use anti-monotone Property.
2	SPADE	Apriori based, DFS based approach, Bottom up search, database vertical projection, use anti-monotone property , lattice theoretical based approach.
3	SPAM	Apriori 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.

IV. CONCLUSION

In this paper, we discussed what is sequential pattern mining and various types of sequential pattern mining algorithms. Sequential pattern mining , the concept being introduced in 1995 has undergone considerable advancement in less than two decades. In Apriori based Algorithms, huge set of sequences generated in large sequence database. Multiple scan of original database is required during the mining process in Apriori based algorithms. this topic focused on improving the efficiency of the algorithms by managing the database in the main memory. In PrefixSpan is efficient pattern growth method because it outperforms GSP and SPADE. It explores Prefix-projection which reduce the size of Projected database and leads to efficient processing. PrefixSpan consume a much smaller memory space in comparison with GSP and SPADE. From the comparative analysis of various sequential pattern mining algorithms, it is clear that pattern growth based algorithms are more efficient with respect to running time, space utilization and scalability.

References

[1] Rakesh Agrawal Ramakrishna Srikant, ―Mining Sequential Patterns‖, 11th Int. Conf. on Data Engineering, IEEE Computer Society Press, Taiwan, 1995 pp. 3-14.
[2] Mabroukeh, N. R. and Ezeife, C. "A taxonomy of sequential pattern mining algorithms", ACM Computing Surveys, vol. 43, no. 1, pp. 1-41, 2010, DOI:10.1145/1824795.1824798 .
[3] R. Srikant, R. Agrawal , "Mining sequential patterns: Generalizations and performance improvements," In Proceedings of International Conference on Extending Database Technology, pp. 3–17, 1996, DOI: 10.1007/BFb0014140.
[4] Zaki, M. J. ,"SPADE: An efficient algorithm for mining frequent sequences", Machine learning, vol.42.no.1-2,pp.31-60,2001, DOI: 10.1023/A:1007652502315.
[5] Ayres, J., Gehrke, J. Flannick, J., and Yiu, T., "Sequential Pattern mining using a bitmap representation", Proc. KDD 2002, Edmonton, Alberta, pp. 429-435, 2002, DOI: 10.1145/775047.775109.
[6] 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, ISBN: 978-0-89871-593-4.
[7] J. Pei, J. Han, B. Mortazavi-Asi, H. Pino, "PrefixSpan: Mining Sequential Patterns Efficiently by Prefix- Projected Pattern Growth", ICDE'01, 2001.
[8] Panchal Mayur, Ladumor Dhara, Kapadiya Jahnvi, Desai Piyusha, Patel Tushar S., "An Analytical Study of Various Frequent Itemset Mining Algorithms," *Research Journal of Computer and Information Technology Sciences*,p.4,2013.
[9] Jiawei Han · Hong Cheng · Dong Xin · Xifeng Yan, "Frequent pattern mining: current status and future Directions,"*Data Mining Knowl Discov*,vol.15,no.I,p.32,2007.
[10] H.Shyur, C.Jou and K. Chang, "A data mining approach to discovering reliable sequential patterns", pp 08, 2008.