International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 9, September- 2014

A Time and Space Efficient Algorithm for Mining Sequential Pattern

Prachi Batwara
IET College, Alwar(Raj.)

Dr. B. K.Verma
Associate Professor
IET College,Alwar(Raj)

***Abstract*—Sequential Pattern Mining is very important technique of Data Mining which extract frequent patterns from given sequence. It is used in various fields such as medical treatments, customer shopping sequence, DNA sequence and gene structures. Sequential Pattern Mining Approaches are classified into two categories: Apriori or generate and test approach, pattern growth or divide and conquer approach.**

In this paper, we are introducing a more time and space efficient algorithm for sequential pattern mining. The time & space consumption of proposed algorithm will be lesser in comparison to previous algorithms & we compare two algorithms of pattern growth algorithms of Sequential Pattern Mining, one is P-prefixspan which discovers frequent sequential pattern with probability of inter arrival time and other one is new proposed algorithm named as Precursive algorithm. Our experiment shows that new proposed algorithm is more efficient and scalable then the P-prefixspan algorithm.

IndexTerms—Data Mining, Sequential Pattern Mining, Frequent Item set, Support count, Sequence database.

I. INTRODUCTION

Data Mining is a collection of techniques for uncovering the interesting data pattern hidden in a huge dataset. Data mining extract non-trival, implicit, unfamilierand useful knowledge from large data set. Many approaches have been discover to extract information from input sequence and Sequential pattern mining is one of the most important methods. It is defined as the process of discovering all subsequences that appear frequently from a given dataset. Sequential pattern mining problem can be widely used in different areas, such as mining user access patterns for the web sites, using the history of symptoms to predict certain kind of disease, customer shopping sequence and so on.

Data mining is known as one of the core processes of Knowledge Discovery in Database (KDD).The KDD process composed of some steps starting from raw data to new knowledge information. Usually there are three processes in KDD. One is called pre processing, which includes data cleaning, integration, selection and transformation. The main process of KDD is the data mining process, in this process different algorithm are applied to produce hidden knowledge. After that comes another process called post processing, which evaluates the mining result according to users‟ requirements and domain knowledge.Regarding the evaluation results, the knowledge can be presented if the result is satisfactory, otherwise we have to run some or all of those processes again until we get the satisfactory result. Various data mining techniques are applied to the data source; different knowledge comes out as the mining result. That knowledge is evaluated by certain rules, such as the domain knowledge or concepts. After we get the knowledge, the final step is to visualize the results. They can be displayed as raw data, tables, decision trees, rules, charts, data cubs or 3D graphics.

II. RELATED WORK

Sequential pattern mining was proposed in [1], using the main idea of association rule mining presented in Apriori algorithm of [4]. Later, three algorithms (Apriori, AprioriAll, and AprioriSome) to handle sequential mining problem were proposed in [3]. Following this, the GSP (Generalized Sequential Patterns) [3] algorithm, which is 20 times faster than the Apriori algorithm in [1] was proposed. The PSP (Prefix Tree for Sequential Patterns) [5] approach is much similar to the GSP algorithm [3]. The main idea of Graph Traversal mining which is proposed by [6][7], is using a simple unweighted graph to reflect the relationship between the pages of Web sites.

Traditional sequential patterns mining approaches such as Apriori-based algorithms [1, 3] encounter the problem that multiple scans of the database are required in order to determine which candidates are actually frequent. Most of the solutions provided so far for reducing the computational cost resulting from the apriori property use a bitmap vertical representation of the access sequence database [11] and employ bitwise operations to calculate support at each iteration. The transformed vertical databases, in their turn, introduce overheads that lower the performance of the proposed algorithm, but not necessarily worse than that of pattern-growth algorithms. *SPADE:* SPADE (Sequential Pattern Discovery using Equivalence classes) [12] is an Apriori based vertical format sequential pattern mining algorithm i.e. the sequences are given in vertical order instead of horizontal format. In addition, this algorithm uses the *ID-List* technique to reduce the cost for computing support counts.*SPAM:* SPAM (Sequential Pattern Mining) [11] uses a vertical bitmap data structure representation of database which is similar to the given id-list of SPADE. It integrates the concept of GSP [3], SPADE [12] and FREESPAN [13] algorithms. SPAM uses a depth-first traversal to increase its

performance. SPAM reduces the cost of merging but takes more time and space when compared to other algorithms which can be completely stored in the main memory.

We first proposed a straightforward patterngrowth method, FreeSpan (for Frequent pattern-projectedSequential pattern mining) [13], which reduces the efforts ofcandidate subsequence generation. After that, introduce another and more efficient method, calledPrefixSpan[5](for Prefix-projected Sequential pattern mining)which offers ordered growth and reduced projecteddatabases. To further improve the performance, a pseudoprojectiontechnique is developed in PrefixSpan. A comprehensiveperformance study shows that PrefixSpan, in mostcases, outperforms the a priori-based algorithm GSP,FreeSpan, and SPADE [12] (a sequential pattern miningalgorithm that adopts vertical data format) and PrefixSpan,integrated with pseudoprojection, is the fastest among allthe tested algorithms.

Extending the PrefixSpanalgorithm,developed anew sequential pattern mining approach – P-PrefixSpan [2]. The major difference between the PrefixSpanalgorithm and P-PrefixSpan algorithm is the proposed algorithm needs to deal with the time stamp of each item in data sequences while thePrefixSpan algorithm is only concerned with the order of items indata sequences. Compared with the PrefixSpan algorithm, P-PrefixSpan is a morecomplicate algorithm. An additional step is developed to estimate the arrival rate of frequent item with respect to pattern and the probability of inter-arrival time. However, the experimental results show that P-PrefixSpan outperforms PrefixSpan algorithm. The performance gap increases as the minimum support threshold decreases because when the minimum support decreases, the number of the frequent sequences increases, the number of the projecteddatabases increases and the size of each projected databasealso increases, such that the performance is degraded for PrefixSpan algorithm. For P-PrefixSpan, we focus on reliable patterns, that is, it will reduce the number of candidate patterns and the number of projected databases in the mining process.

III. SEQUENTIAL PATTERN MINING

Sequential Pattern Mining is one of the main concept of Data Mining. It extracts the frequent sequential patterns from a sequence database. It is used in many applications such as DNA sequence, Customer Shopping Sequence, Gene Structure and so on.

A sequential pattern mining algorithm should

a. find the whole set of patterns, when possible, satisfying the minimum support(frequency) threshold,

b. be highly efficient, scalable, involving only a small number of database scans

c. be able to incorporate various kinds of user-specific constraints.

Sequential pattern mining approaches are classified as Apriori or generate and test approach, pattern growth or divide-and-conquer approach. Apriori approach based on apriori property and using generates and join procedure to discover frequent patterns. Some of apriori algorithms are GSP, SPADE, SPAM. Pattern Growth approaches extract frequent patterns from large data set without candidate generation. Some of pattern growth algorithms are Prefixspan, Freespan etc.

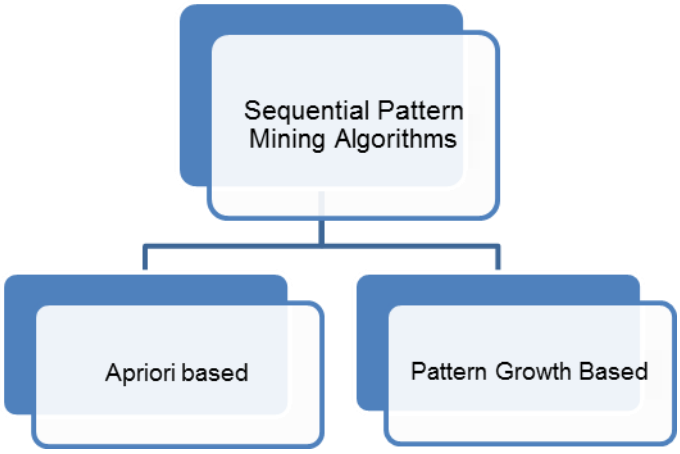

Fig. 1.Classification of Sequential Pattern Mining

A. Apriori Based or Generate or Test Approach: This approach is based on Apriori Property. It has many limitations:

- Multiple scan of database
- Breadth First Search
- Huge number of candidate sets generated.
- Difficulties at mining long Sequential Patterns.

B. Pattern Growth Based: This approach is based on divide and conquer strategy and generate frequent patterns without candidate generation. It has various features:

- The analysis is focus on counting the frequency of relevant data sets instead of candidate sets.
- Depth First Traversal
- Search Space Partitioning
- The method partition the datasets into smaller projected datasets which reduce the search space and enhance performance.
- Candidate Sequence Pruning
- New data structures are used such as FP-Tree and PseudoProjection for saving the cost of Projection and increase in processing speed.

IV. PROPOSED METHODOLOGY

We are introducing a new more efficient algorithm for pattern growth sequential pattern mining named as PrecursiveAlgorithm is used for finding sequential patterns from a huge data set.

The objective of this paper is to analyze and do a comparative analysis of two sequential pattern algorithms named as P-PrefixSpan and New Proposed (Precursive) using three parameters. The time & space consumption of proposed algorithm will be lesser in comparison to previous algorithms.

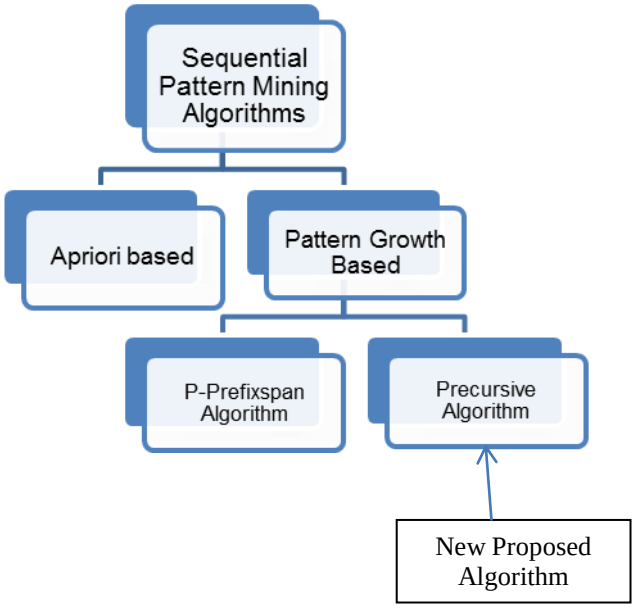

Fig. 2 Sequential Pattern Mining Approaches

A. *Terminology*

- **Sequence:** Let I={i1,i2,i3…..in} be a set of items. Sequence is defined as ordered list of item sets (also called elements & events). The number of instances of items in a sequence is called length of the sequence.
  Eg.<a(ce)(bd)(bcde)f> is a sequence which consist of distinct items & 5 elements. Length of sequence is 10.

- **Sequence Database:** It consist of ordered elements. It is a set of tuples<sid.s> where sid is a sequence id & s is a sequence.
  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 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.

TABLE I. SEQUENCE DATABASE

Sid	Sequence
1	<a(abc)cd(bd)>
2	<b(cd)abc(bc)>
3	<c(ab)cd(abc)>

- **Support:**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.

B. *Algorithm*

The steps of new pattern recursive algorithm are as follows:

Step 1: Start

Step 2: Input: Sequential Database and Minimum Support

Step 3: Scan the database to store the first bit position of each sequence and calculate the total number of bit for each bitmap.

Step 4: After scanning, create bitmap vertical database and calculate support of each item.

Step 5: Pruning (remove infrequent items from the database) because they will not appear in any frequent sequential patterns.

Step 6: Repeat DFS recursively will be used to obtain all remaining frequent patterns.

Step 7: Generate bigger frequent patterns by using union of lower size patterns of items.

Step 8: Output frequent pattern obtained.

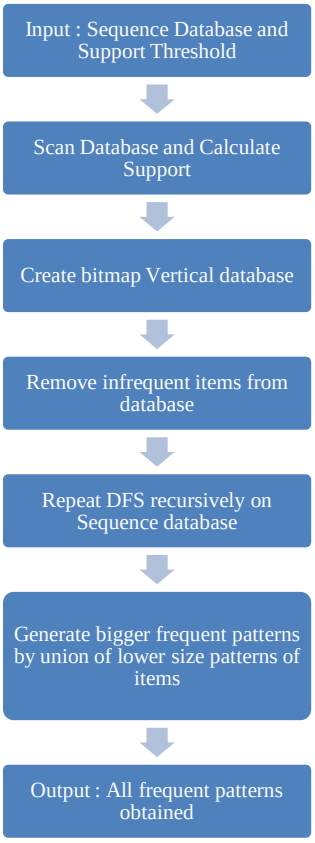

Fig. 3.Block Diagram of New Proposed Algorithm

V. EXPERIMENTAL ANALYSES AND RESULTS

To evaluate the performance comparison between two sequential pattern mining algorithms named as P-Prefixspan algorithm & New Proposed algorithm (Precursive) are implemented in JAVA Language and Netbeans.

To evaluate the performance comparison we can take a real data of Easy Day Store as input database.

In this paper, we only take 10 products of Easy Day Store Data.

A. Input Database:

TABLE II. EASY DAY STORE DATABASE

ProductID	Product Name
9578442	MilkFoodGhee
9514357	Sugar
9574697	Wheat atta
9517766	Chana dal
9513615	Chilli powder
9555300	Turmeric powder
126342	Kraft oreo
186702	Haldirambhujia
298441	Pepsodent
205153	Pears

TABLE III.INPUT SEQUENCE DATABASE

Sid	Sequence
S1	(126342 186702 298441) (126342 298441)
S2	(9513615 9555300) (298441 9574697)
S3	(126342 9513615) (298441) (186702 298441)(126342 9514357)
S4	(9514357 9574697) (126342 186702) (9513615 9574697 9555300) (298441) (126342 186702)
S5	(9514357) (9578442) (126342 9574697) (298441) (186702 298441)
S6	(126342 186702) (205153 298441) (9513615 9517766 9555300)
S7	(9513615 9555300) (298441 126342 9514357)
S8	(9578442) (126342 9574697) (186702 298441)
S9	(126342 186702) (126342 298441) (9513615)(298441)
S10	(9574697 9578442) (126342 186702 9574697) (186702 298441)

B. Output Results:

After giving input sequence database and min. support into both algorithms then obtained output results are shown in table IV,V,VI,VII.

TABLE IV.RESULTS WHEN SUPPORT=0.2

Parameters	P-PrefixSpan	Precursive
Time(ms)	95	80
Frequent Sequence Count	108	108
Memory	1.162109375108	0.73288125108

TABLE V. RESULTS WHEN SUPPORT=0.3

Parameters	P-PrefixSpan	Precursive
Time(ms)	15	11
Frequent Sequence Count	53	53
Memory	0.989257812553	0.651367187554

TABLE VI.RESULTS WHEN SUPPORT=0.4

Parameters	P-PrefixSpan	Precursive
Time(ms)	11	9
Frequent Sequence Count	21	21
Memory	0.8164062521	0.565429687521

TABLE VII.RESULTS WHEN SUPPORT=0.5

Parameters	P-PrefixSpan	Precursive
Time(ms)	8	7
Frequent Sequence Count	12	12
Memory	0.7304687512	0.565429687512

These results are shown in graph in Fig 4 and Fig 5.

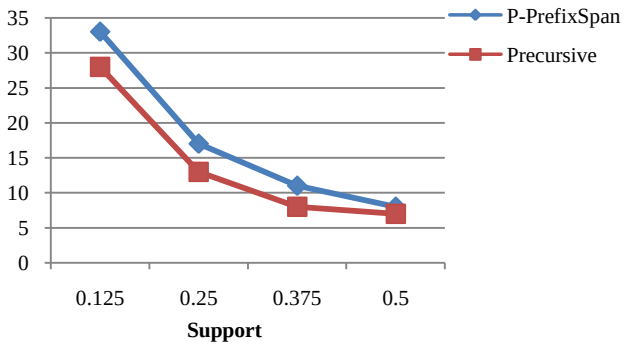

Fig. 4 Time Usage of Easy Day Store

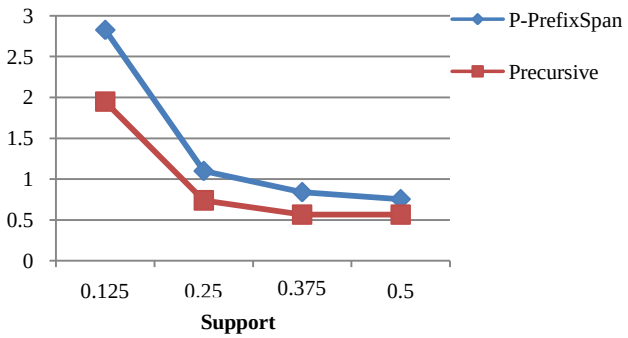

Fig. 5 Memory Usage of Easy Day Store

Output Frequent Pattern obtained after implementing both algorithms. Frequent Sequential Pattern are different due to taking different values of minimum support.

TABLE VIII: Output Frequent Pattern

Frequent Pattern No.	Frequent Pattern	Support
1	9574697	5
2	298441	10
3	186702	8
4	186702 298441	5
5	(186702)(298441)	5
6	9513615	6
7	(9513615)(298441)	5
8	126342	9
9	126342 186702	5
10	(126342 186702)(298441)	5
11	(126342)(298441)	8
12	(126342)(186702)	5

Frequent Pattern No.	Frequent Pattern Description
1	Aata
2	Toothpaste_Pepsodent
3	Namkeen_Bhujia
4	(Namkeen_Bhujia)(Toothpaste_Pepsodent)
5	(Namkeen_Bhujia)(Toothpaste_Pepsodent)
6	Spices_Chilipowder
7	(Spices_Chilipowder)(Toothpaste_Pepsodent)
8	Biscuit_oreo
9	Biscuit_oreoNamkeen_Bhujia
10	(Biscuit_oreoNamkeen_Bhujia)(Toothpaste_Pepsodent)
11	(Biscuit_oreo)(Toothpaste_Pepsodent)
12	(Biscuit_oreo)(Namkeen_Bhujia)

VI. CONCLUSION

Due to increasing data day to day, it is very difficult to maintain & retrieve information in real life situations. That's why, there is need of various data mining techniques for various different type of data. In this paper, a new proposed algorithm named as Precursive Algorithm is used for finding frequent patterns from a huge data set. It first scan the sequence database and calculate support of each data and find all frequent patterns which have support greater than support threshold. Then sequence database converted into compressed data structure by removing all infrequent item sets. This process continues until all frequent pattern are generated.

This algorithm performs better then P-prefixspanalgorithm in terms of time & memory.

Execution time is reduced whenever run the new proposed algorithm instead of P-prefixspan algorithm.

REFERENCES

[1] R.Agarwal and R.Srikanth, "Mining Sequential Patterns" ICDE"95, Pg 3-14,1995.

[2] Huan-JyhShyur∗, Chichang Jou1, Keng Chang "A data mining approach to discovering reliable sequential patterns" The Journal of Systems and Software 86 (2013) 2196– 2203.

[3] R. Srikant and R. Agrawal."Mining sequential patterns: Generalizations and performance improvements". In Proc. of the 5th International Conference on Extending Database Technology (EDBT'96), pages 3–17, Avignon, France, September 1996.

[4] Dr P padmaja, P Naga Jyoti, m Bhargava "Recursive Prefix Suffix Pattern Detection Approach for Mining Sequential Patterns" IJCA September 2011.

[5] J. Pei, J. Han, H. Pinto, Q. Chen, U. Dayal and M.-C. Hsu, "PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-projected Pattern Growth," Proceedings of 2001 International Conference on Data Engineering, pp. 215-224, 2001.

[6] Hua-Fu Li, Chin-Chuan Ho, Hsuan-Sheng Chen and Suh-Yin Lee, "A single scan algorithm for mining Sequential pattern from data streams" in ICIC International, Taiwan Mar 2012.

[7] Ms. Pooja Agrawal, Mr. Suresh kashyap, Mr.Vikas Chandra Pandey, Mr. Suraj Prasad Keshri, "An Analytical Study on Sequential Pattern Mining With Progressive Database" International Journal of Innovative Research in Computer and Communication Engineering Vol. 1, Issue 3, May 2013.

[8] R. Agrawal, and R. Srikant, Fast algorithms for mining association rules, Proc. of 20th Intl. Conf. on VLDB, pp.487-499, 1994.

[9] J. Han and M. Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufman publishers, 2001, ISBN: 1-55860489-8.

[10] V. Uma1, M. Kalaivany, G. Aghila, " Survey of Sequential Pattern Mining Algorithms and an Extension to Time Interval Based Mining Algorithm", Volume 3, Issue 12, December 2013.

[11] AYRES, J., FLANNICK, J.,GEHRKE, J., AND YIU, T. 2002. Sequential pattern mining using a bitmap representation. In Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery andData Mining.429–435.

[12] Zaki, M. J. (2001). SPADE: An Efficient Algorithm for Mining Frequent Sequences. Journal of Machine Learning, 42(1-2), 31-60.

[13] Han, J., Pei, J., Mortazavi-Asl, B., Chen, Q., Dayal, U., & Hsu, M.-C. (2000). FreeSpan: frequent pattern-projected sequential pattern mining. Proceedings of the Sixth ACM

SIGKDD International Conference on Knowledge Discovery and Data Mining.

[14] Rakesh Agrawal, &RamakrishnanSrikant., 1995. "Mining generalized association rules". In: Dayal U, Gray P M D, NishioSeds. Proceedings of the International Conference on Very Large Databases. San Francisco, CA: Morgan Kanfman Press, pp. 406-419.

[15] Yan Huang, Member, Liqin Zhang, and Pusheng Zhang, Member, "A Framework for Mining Sequential Patterns from Spatio-Temporal Event Data Sets, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 20, NO. 4, APRIL 2008.

[16] Sushila Umesh Ratre , Prof. Ravindra Gupta , " An Efficient Technique for Sequential Pattern Mining ", International Journal of Advanced Research in Computer Science and Software Engineering , March 2013.

[17] Jei-Wei Han, Jeinpei, Xi-Fong Yan, " From Sequential Pattern Mining to Structured Pattern Mining:A Pattern Growth Approach", J.Comp.Science and Tech. May 2004.