ISSN 2278-8808
An International Peer Reviewed
SJIF 2013 = 4.194
SCHOLARLY RESEARCH JOURNAL
FOR INTERDISCIPLINARY STUDIES

# PERFORMANCE COMPARISON BETWEEN PATTERN GROWTH ALGORITHMS FOR MINING SEQUENTIAL PATTERN

**Prachi Batwara**
MTech. Scholar,
IET College, Alwar (Raj.)

**Basant Verma, Ph.D**
Associate Professor,
IET College, Alwar (Raj.)

## Abstract

*Sequential Pattern Mining is very important concept in Data Mining, finds frequent patterns from given sequence. It is used in various domains 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 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-prefix span which discovers frequent sequential pattern with probability of inter arrival time and other one is new proposed algorithm named as Percussive algorithm. Our experiment shows that new proposed algorithm is more efficient and scalable then the P-prefix span algorithm.*

***Keywords:*** *Data Mining, Sequential Pattern Mining, Frequent Item set, Support count, Sequence database.*



**Introduction**: Data Mining is a collection of techniques for uncovering the interesting data pattern hidden in a huge dataset. Data mining extract non-trivial, implicit, previously unknown and potentially 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 on 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).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.

**Sequential Pattern Mining:** Sequential Pattern Mining is one of the main method of Data Mining. It extracts the frequent 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 complete 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.

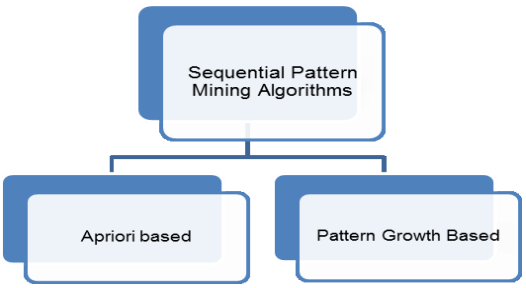



**FIGURE 1:** Classification of Sequential Pattern Mining Algorithm

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

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

**2. 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.
- The method partition the datasets into smaller projected datasets which reduce the search space and enhance performance.
- New data structures are used such as FP-Tree and Pseudo Projection for saving the cost of Projection and increase in processing speed.

**Proposed Methodology**

We are introducing a new more efficient algorithm for pattern growth sequential pattern mining named as Precursive Algorithm is used for finding sequential patterns from a huge data set. The objective of this thesis 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.

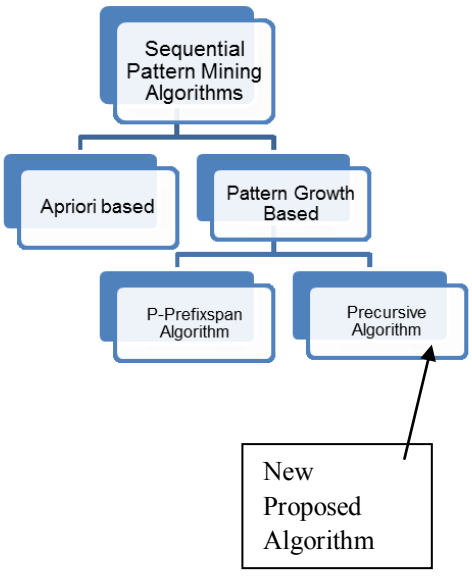



**FIGURE 2:** Sequential Pattern Mining Approaches

## 1. 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.

**TABLE I:** Sequence database

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

**Support:** The support of a sequence s in a sequence database S is the number of tuples in the database containing s. Given a positive integer minimum support(min_support) as the support threshold, a sequence s is called a sequential pattern in sequence database S if supports(s)>=min_support.

## B. Algorithm

The steps of new pattern recursive algorithm are as follows:

Step 1: Start

Step 2: Sequential DataBase and Minimum Support

Step3: First the algorithm scans the Sequence Database and calculates support of each single item.

Step 4: In this Step, Sequential Data is transformed into compressed data structure by Pruning of all those items from the Sequential Database, whose support is lesser than minimum support threshold because they will not appear in any frequent Sequential Patterns.

Step 5: Call Algorithm recursively to generate bigger Sequential Patterns by using union of lower size items.

## Experimental Analysis & 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.

**1. 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 ) (9513615 )(298441 9574697 ) |
| S2 | (126342 9513615 )(298441 )(186702 298441 )(126342 9514357 ) |
| S3 | (9514357 9574697 )(126342 186702 )(9513615 9574697 )(298441 )(186702 ) |
| S4 | (9514357 )(9578442 )(126342 9574697 )(298441 )(186702 )(298441 ) |
| S5 | (126342 186702 )(205153 298441 )(9513615 9517766 9555300 ) |
| S6 | (9513615 )(298441 )(298441 )(126342 9514357 ) |
| S7 | (9578442 )(126342 9574697 )(186702 )(298441 ) |
| S8 | (186702 )(126342 298441 )(9513615 )(298441 ) |

***2.*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=0.125

| Parameters | P-PrefixSpan | Precursive |
|---|---|---|
| ***Time(ms)*** | 33 | 28 |
| ***Frequent Sequence Count*** | 737 | 737 |
| ***Memory*** | 2.82421875737 | 1.9462890625737 |

**TABLE V:** Results when support=0.25

| Parameters | P-PrefixSpan | Precursive |
|---|---|---|
| ***Time(ms)*** | 17 | 13 |
| ***Frequent Sequence Count*** | 103 | 103 |
| ***Memory*** | 1.0986328125103 | 0.73828125103 |

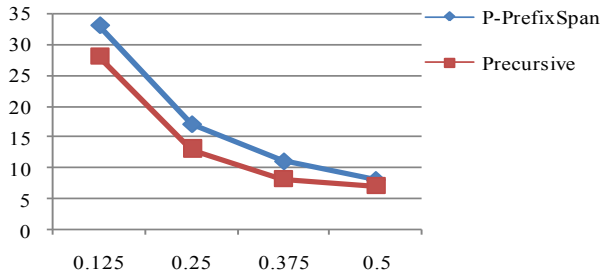
**TABLE VI**: Results when support=0.375

| Parameters | P-PrefixSpan | Precursive |
|---|---|---|
| ***Time(ms)*** | 11 | 8 |
| ***Frequent Sequence Count*** | 28 | 28 |
| ***Memory*** | 0.8398437528 | 0.565429687528 |

**TABLE VII**: Results when support=0.5

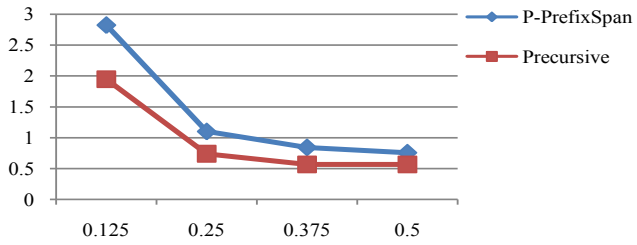
| Parameters | P-PrefixSpan | Precursive |
|---|---|---|
| ***Time(ms)*** | 8 | 7 |
| ***Frequent Sequence Count*** | 13 | 13 |
| ***Memory*** | 0.7539062513 | 0.565429687513 |

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




**Support**

**GRAPH 1:** Time Usage of Easy Day Store




**Support**

**GRAPH 2:** Memory Usage of Easy Day Store

**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-prefixspan algorithm in terms of time & memory.

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

**References**

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.

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.

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.

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.

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.

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.

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

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

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

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

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

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.

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.

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.

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.

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.

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