Global Journal of Pure and Applied Mathematics.
ISSN 0973-1768 Volume 15, Number 4 (2019), pp. 399-410
© Research India Publications
http://www.ripublication.com/gjpam.htm

Efficient Constraint-Based Sequential Pattern Mining (SPM) Algorithm

Priyadharshini S.P[1], Dr. Hemalatha M[2]

[1]Ph.D Research Scholar, Bharathiar University, Coimbatore, India.

[2]Professor, Sri Ramakrishna College of Arts & Science, Coimbatore, India.

Abstract

Sequential pattern mining is advantageous for several applications. For example, it finds out the sequential purchasing behavior of majority customers from a large number of customer transactions. However, the existing researches in the field of discovering sequential patterns are based on the concept of frequency and presume that the customer purchasing behavior sequences do not fluctuate with change in time, purchasing cost and other parameters. To acclimate the sequential patterns to these changes, constraint are integrated with the traditional sequential pattern mining approach. It is possible to discover more user-centered patterns by integrating certain constraints with the sequential mining process. Thus in this paper, proposed constraint based sequential pattern mining algorithm has been validated on synthetic sequential databases. The experimental results ensure that the efficacy of the sequential pattern mining process is further enhanced in view of the fact that the purchasing cost, time duration and length are integrated with the sequential pattern mining process

Keywords: Sequential Pattern, Growth, Constraint, Pre-Processing, Prefix Span.

1. INTRODUCTION

Sequential pattern mining is an important data mining task with many real applications. Most of the existing studies, such as focused on efficient algorithms and effective pattern representations. In the existing work, absolute or relative frequency (also known as support) is used as the sole criterion in selecting frequent patterns. While frequency often serves as a good preliminary filter to remove noise patterns of very low popularity, in many applications, one has to find relevant patterns whose interestingness is defined in a statistical way, and cannot be specified using only a support threshold. A pattern of high frequency may not be interesting if it is statistically expectable from other patterns. At the same time, a pattern of low frequency may be interesting if it is statistically unexpected. Since a low support threshold often leads to a huge number of patterns, asking a user to select from patterns extracted using a low support threshold is overwhelming and impractical. This is a problem common not to only sequential patterns, but to frequent patterns in general. To echo this challenge, several recent studies try to find patterns (i.e., itemsets or sequences) using some alternative interestingness measures or sampling representative patterns. A general idea, which is a framework of finding unexpected patterns, is to extract patterns whose characteristic on a given measure, such as the frequency, or more rarely the length, strongly deviates from its expected value under a null model. The frequency of a pattern is considered as a random variable, whose distribution under the null model has to be calculated or approximated. Then, the significance of the pattern is assessed through a statistical test that compares the expected frequency under the null model to the observed frequency. One of the key-points of this family of approaches is to choose an appropriate null model. It will ideally be a trade-off between adjustment to the data and simplicity: the model should capture some characteristics of the data, to integrate prior knowledge, without overfitting, to allow for relevant patterns discovery. A simple model, with low-order dependency, often results in faster computations and clear interpretation of the unexpected patterns.

2. PROPOSED SYSTEM FRAMEWORK

2.1. Strength and working of proposed algorithm

i. The proposed calculation being FP-Growth based, Constraint-based prefix range lessens competitor age and takes a shot at anticipated prefix database.

ii. First the calculation examines the database and distinguishes visit things. It recursively finds the prefix on continuous things as well as with thought of hole imperative which deals with two nearby time stamps (max_gap,min_gap),with first and last time stamp of prefix sequence fulfilling smallness requirement.

iii. The sequence which does not pursue such requirements can be pruned at pseudo projection level. This procedure decreases the database projection cost and inquiry space when contrasted with sole parameter bolster edge accordingly expanding productivity of proposed calculation.

iv. Incorporation of Length requirement restrains the age of sequences by pruning the sequences having more length at projection level. Thing and Recency go about as post handling parameters. Consolidation of such imperatives in regular Prefix Span gives progressively compelling outcomes according to client's advantage.

v. Proposed Emerging Pattern mining calculation is distinguishing those patterns which are not in spotlight but rather can possibly end up solid in not so distant future. Such shrouded patterns can be featured utilizing slight decrease of limit estimation of help edge and consideration of recency imperative.

2.2 Constraint-based sequential pattern generation

Data mining removes verifiable, possibly valuable learning from a lot of information. It is additionally called information mining, learning extraction, information/sequence/design examination, information antiquarianism and information digging from databases. As it were, information mining is the demonstration of boring through colossal volumes of information to find connections or answer questions, summed up for customary inquiry instruments. As a rule, information mining undertakings can be ordered into two classifications: Descriptive mining: It is the way toward illustration the basic attributes or general properties of the information in the database. Bunching, Association and Sequential mining are one of the spellbinding mining methods. Prescient mining: This is the way toward deducing sequences structure information to make expectations. Characterization, Regression and Deviation discovery are prescient mining strategies. Information mining system is helpful in different regions, for example, showcase bushel investigation, choice help, extortion discovery, business the executives, broadcast communications and so forth. The information mining were drawn from Database Technology, Machine Learning, Artificial Intelligence, Neural Networks, Statistics, Pattern Recognition, Knowledge-based Systems, Knowledge Acquisition, Information Retrieval, High-execution calculation and Data Visualization. Numerous techniques came up to concentrate the data. The Sequential Sequence Mining is a standout amongst the most critical strategies that encourage us to settle on the choices in different applications. The mining issue was first proposed by Agrawal and Srikant. It finds successive sequences which happen oftentimes in a sequence database. In the Medicine, finding of time interim sequence of maladies from restorative records like illnesses, medications, and lengths of clinic stay and so on are recorded in the database of Hospitals. In any case, every one of the occasions, for example, enduring and restoring sicknesses or happening side effects are interim based. The regular consecutive sequence digging isn't proper for the revelation of the sequences in these occasions. On other hand, time interim sequences are progressively helpful to recognize whether a patient experiences a specific sickness or not. It additionally predicts the indications of a patient who has a specific sickness. In venture, a specific stock ascents or falls is one of the critical errands that the stock financial specialists needed to know. Further, the proprietors are stressed over the stock pattern of their own organizations. Investors or Industry examiners likewise prefer to know the ascent/fall of specific stocks, which is really one of the valuable data extractions from the time interim sequences of stock costs. The stock costs are recorded in each exchange which goes about as a chronicled information. We may discover the time interim stock sequences from the stock interim occasion database. Figure 5.1 shows three phases of successive example age. Spotted line and straight line meant in reverse stream and forward stream. Figure 5.2 shows detail of Pre-preparing of info information. After age of consecutive info record in configuration, document is utilized for sequence age. Figure 5.3 and Figure 5.4 shows stream to create sequences which fulfilled recurrence, conservativeness and hole requirements (FCG—sequences).

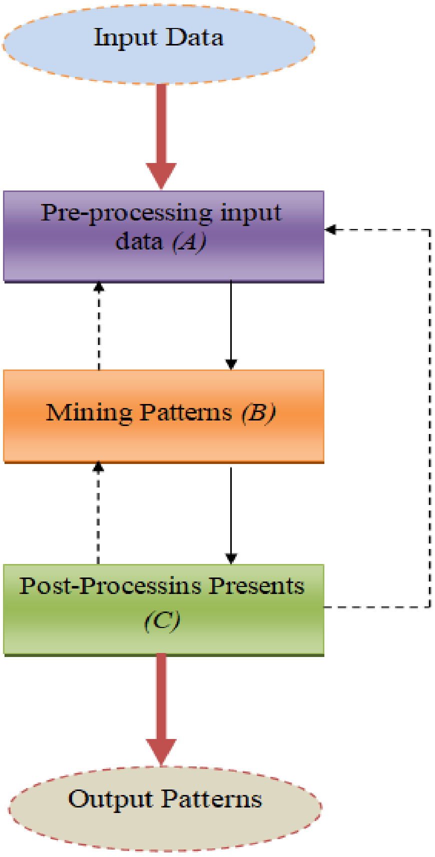
Input Data
Pre-processing input data *(A)*
Mining Patterns *(B)*
Post-Processins Presents *(C)*
Output Patterns

Figure 1: Generation of Sequential Pattern

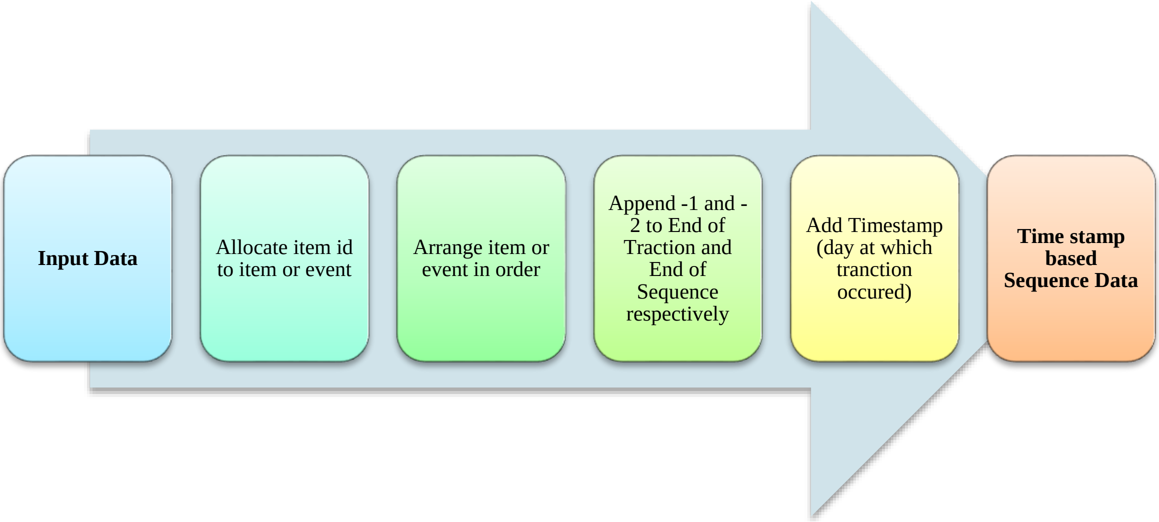
Input Data
Allocate item id to item or event
Arrange item or event in order
Append -1 and -2 to End of Traction and End of Sequence respectively
Add Timestamp (day at which tranction occured)
Time stamp based Sequence Data

Figure 2: Input Data Pre-processing

In the E-showcasing, some Internet merchants give new selling strategies like gathering purchasing offer. These happen when merchants needed to sell items at lower costs when somebody gathers a horde of individuals to purchase this item. The length when an individual joins a gathering purchasing segment for a specific item till the end of the session is considered as an interim based occasion. Since many gathering purchasing clients may join purchasing sessions for various items simultaneously or later, these interim based occasions structure a lot of sequences, which may incorporate some intriguing time situated sequences. Finding time situated sequences from gathering purchasing records will help the acquiring practices of clients and make powerful advertising procedures.

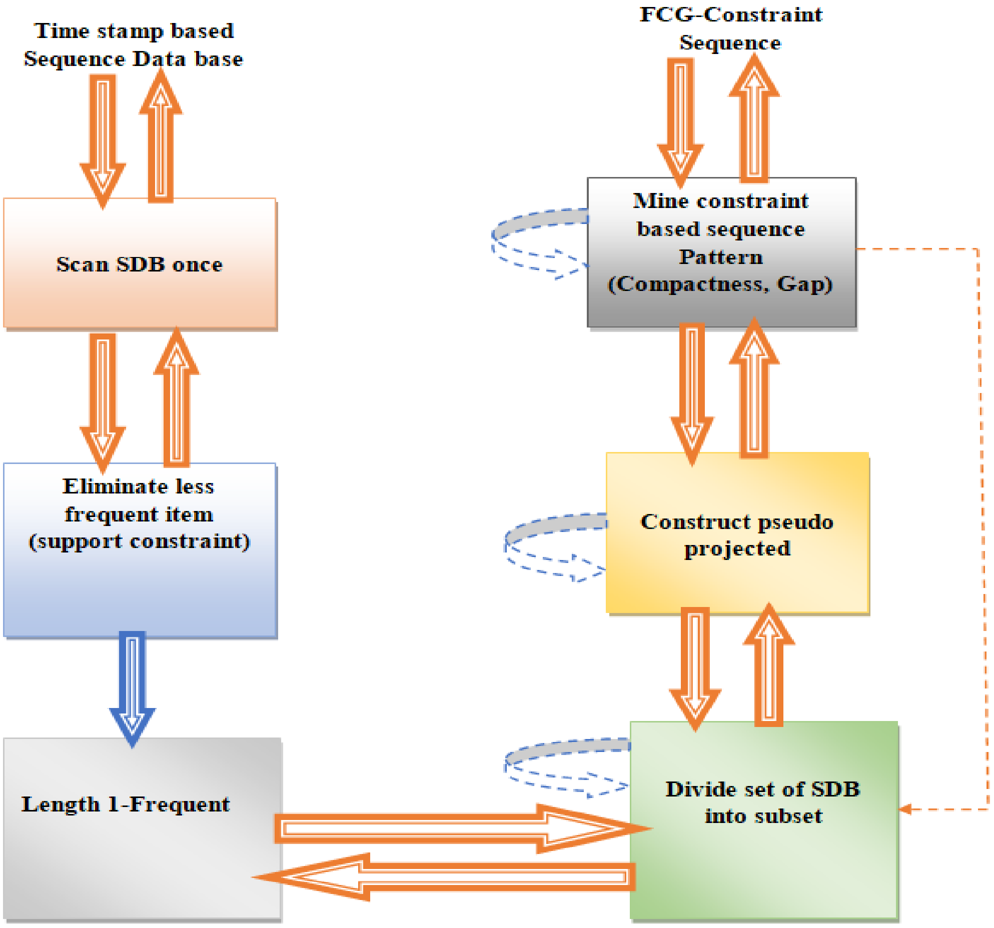
Time stamp based
Sequence Data base
Scan SDB once
Eliminate less frequent item (support constraint)
Length 1-Frequent
FCG-Constraint
Sequence
Mine constraint based sequence Pattern (Compactness, Gap)
Construct pseudo projected
Divide set of SDB into subset

Figure 3: FCG Constraint Sequence

Customary Association Rule Mining chips away at value-based data. It believes different things to be bought in single exchange of a specific client. It doesn‟ t care for a similar client buys things in various exchanges. The idea of sequential sequence mining arrived and it believes different things to be bought in various exchanges. It covers the thought in regards to same client buys things in more than one exchange and in more than one time. Anyway the present best in class systems have restrictions with the execution of Memory and Time which are engaged by us. Sequential sequence mining mines sequential sequence from data base with effective help tallying. It is utilized to discover visit subsequences happen with least help esteem.

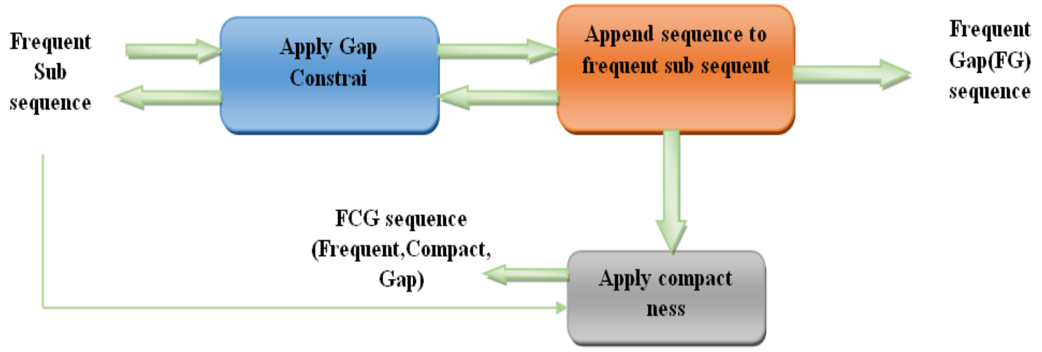
Frequent Sub sequence
Apply Gap Constrai
Append sequence to frequent sub sequent
Frequent Gap(FG) sequence
FCG sequence (Frequent,Compact, Gap)
Apply compact ness

Figure 4: Constraint based mining sequence

The sequential sequence mining centers around sequence of occasions happened every now and again in given dataset dissimilar to straightforward affiliation rule mining. For instance, the client in hardware retail shop buys Computer System of course he buys Scanner after some measure of time. That implies the buying of Scanner is made after the buying of Computer System. The sequence of the things assumes real job. We utilize the request dataset where all occasions put away in some specific request. The customary sequential sequence mining doesn‟ t care for the planning between the buying of things.

2.3 Proposed Algorithm

Proposed constraint-based prefix span for SPM

Input: Sequence Database SDB

Values of following Constraints:-

Support threshold : f_minsup

Recency support : r_min

Compactness : (min_compact, max_compact)

Gap : (min_gap,max-gap)

Length : l_min

Quantity : q_min

Item : i_constraint

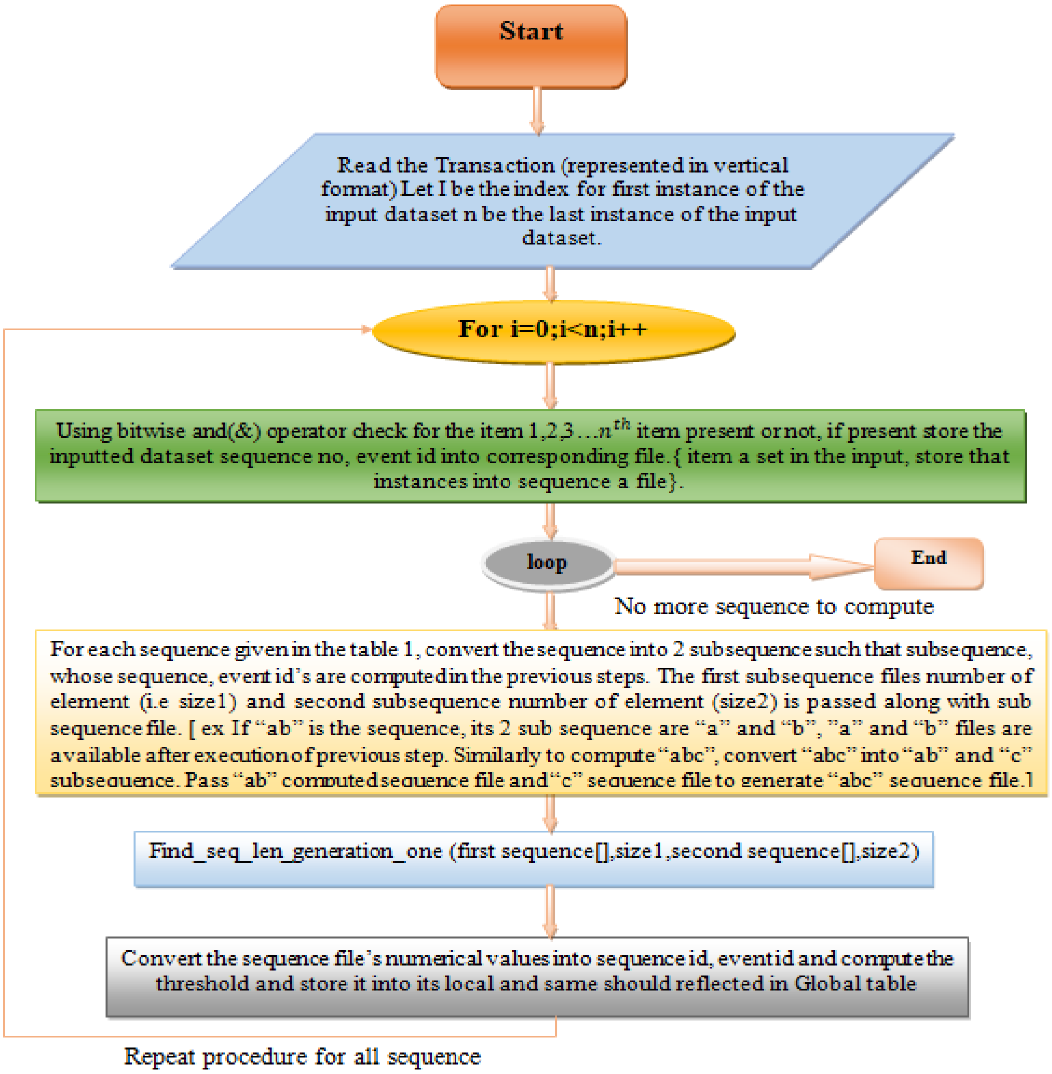
Start
Read the Transaction (represented in vertical format) Let I be the index for first instance of the input dataset n be the last instance of the input dataset.
For i=0;i<n;i++
Using bitwise and(&) operator check for the item 1,2,3…n^{th} item present or not, if present store the inputted dataset sequence no, event id into corresponding file.{ item a set in the input, store that instances into sequence a file}.
loop
End
No more sequence to compute
For each sequence given in the table 1, convert the sequence into 2 subsequence such that subsequence, whose sequence, event id's are computedin the previous steps. The first subsequence files number of element (i.e size1) and second subsequence number of element (size2) is passed along with sub sequence file. [ex If "ab" is the sequence, its 2 sub sequence are "a" and "b", "a" and "b" files are available after execution of previous step. Similarly to compute "abc", convert "abc" into "ab" and "c" subsequence. Pass "ab" computed sequence file and "c" sequence file to generate "abc" sequence file.]
Find_seq_len_generation_one (first sequence[],size1,second sequence[],size2)
Convert the sequence file's numerical values into sequence id, event id and compute the threshold and store it into its local and same should reflected in Global table
Repeat procedure for all sequence

Figure 5: Flow Chart of Proposed Algorithm

Algorithm for the proposed sequential pattern mining is given in the Fig. 5. Algorithm takes contribution from a client and creates every single imaginable sequence.

2.4 Challenges of proposed Algorithm

- There is no logical review about limit esteem decrease in proposed Emerging Patterns mining calculation. Just space specialists can choose how much decrease is required for a specific application.
- Reduction of help edge limit produces gigantic numbers patterns.

3. EXPERIMENTAL RESULTS

Execution Time

Table 1: Explanation table of Execution Time

Existing 1	Existing 2	Proposed
67.2	69.5	83
69.7	69.9	84.8
70.8	69.5	87.9
72.6	70.9	90.2
75	72	93.6

The clarification table of execution time clarifies the distinctive benefits of existing and proposed strategy. While looking at the current and proposed technique the proposed strategy demonstrates the better outcomes. In each dimension of looking at the proposed strategy is superior to the current technique. Existing 1 esteems begins from 67.2 to 75 existing 2 esteems begin from 69.5 to 72 and proposed strategy esteems begins from 83 to 93.6.

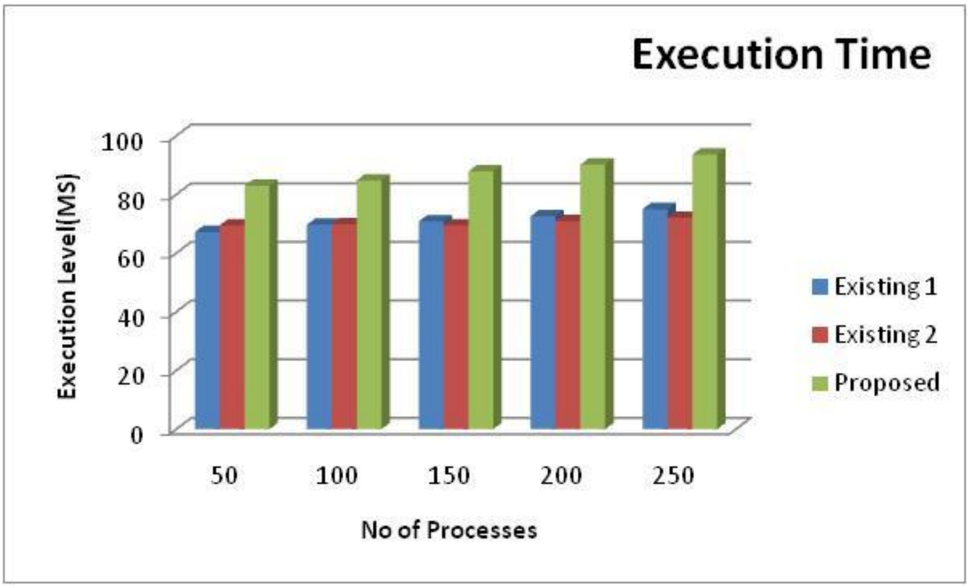

Figure 6: Explanation chart of Execution Time

The clarification diagram of execution time demonstrates the current and proposed technique esteems. No of procedures in X pivot and execution level in Y hub. Each dimension of contrasting the proposed strategy demonstrates the better outcomes. Existing 1 esteems are 67.2-75 existing 2 esteems are 69.5-72 proposed strategy esteems are 83-93.6.

Communication Overhead

Table 2: Explanation table of Communication Overhead

Existing 1	Existing 2	Proposed
31.9	26.77	39
37.7	31.98	45
42.6	34.56	49
50.4	38.92	55
55.23	44.56	58

The clarification table of correspondenceoverhead clarifies the distinctive benefits of existing and proposed technique. While contrasting the current and proposed strategy the proposed technique demonstrates the better outcomes. In each dimension of looking at the proposed technique is superior to the current strategy. Existing 1 esteems begins from 31.9 to 55.23 existing 2 esteems begin from 26.77 to 44.56 and proposed strategy esteems begins from 39 to 58.

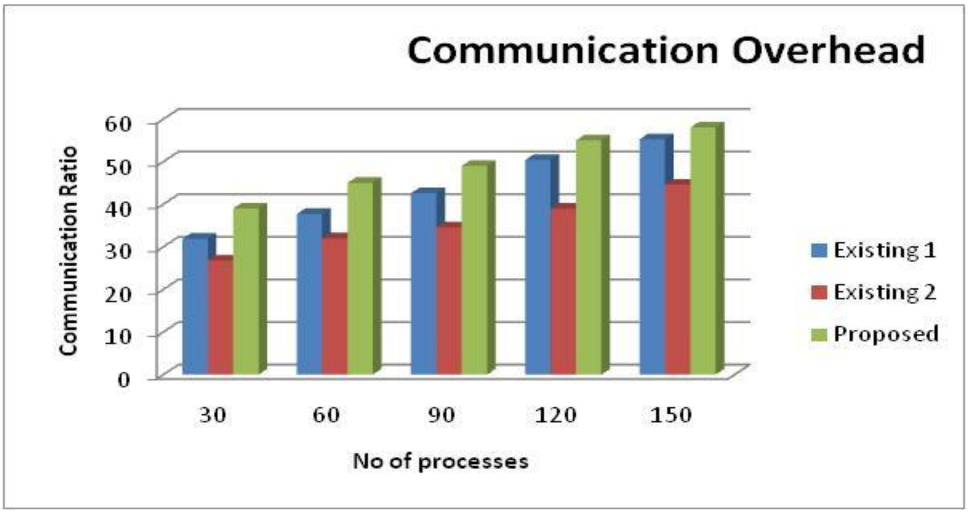

Figure 7: Explanation chart of Communication Overhead

The clarification graph of correspondence overhead demonstrates the current and proposed technique esteems. No of procedures in X hub and correspondence proportion in Y pivot. Each dimension of looking at the proposed strategy demonstrates the better outcomes. Existing 1 esteems are 31.9-55.23 existing 2 esteems are 26.77-44.56 proposed strategy esteems are 39-58.

Storage Evaluation

Table 3: Explanation table of Storage Evaluation

Existing 1	Existing 2	Proposed
55	67	75
58.6	70.1	78.9
62.3	74.8	83.86
68.9	78.89	88.21
72	81	92.06

The clarification table of capacity assessment clarifies the diverse benefits of existing and proposed strategy. While looking at the current and proposed technique the proposed strategy demonstrates the better outcomes. In each dimension of looking at the proposed technique is superior to the current strategy. Existing 1 esteems begins from 55 to 72 existing 2 esteems begin from 67 to 81 and proposed strategy esteems begins from 75 to 92.06.

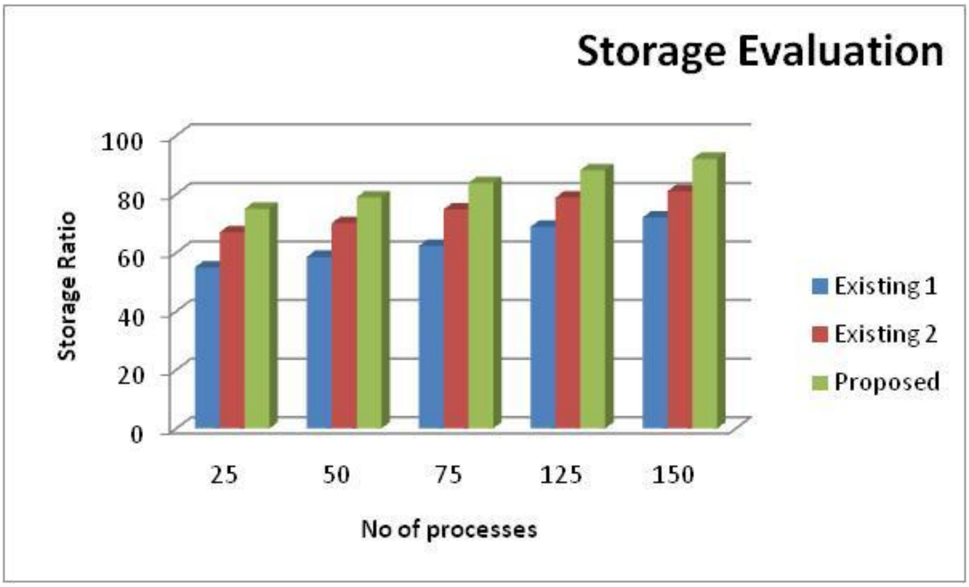

Figure 8: Explanation chart of Storage Evalution

The clarification diagram of capacity assessment demonstrates the current and proposed strategy esteems. No of procedures in X hub and capacity proportion in Y hub. Each dimension of looking at the proposed strategy demonstrates the better outcomes. Existing 1 esteems are 55-72 existing 2 esteems are 67-81 proposed technique esteems are 75-92.06.

Measurement Accurate Ratio

Table 4: Explanation table of Measurement Accurate Ratio

Existing 1	Existing 2	Proposed
73.6	50	83.6
75.6	55	85.6
77.6	60	87.6
78.6	65	89.1
79	67	93

The clarification table of estimation exact proportion clarifies the distinctive benefits of existing and proposed technique. While looking at the current and proposed technique the proposed strategy demonstrates the better outcomes. In each dimension of looking at the proposed strategy is superior to the current technique. Existing 1 esteems begins from 73.6 to 79 existing 2 esteems begin from 50 to 67 and proposed strategy esteems begins from 83.6 to 93.

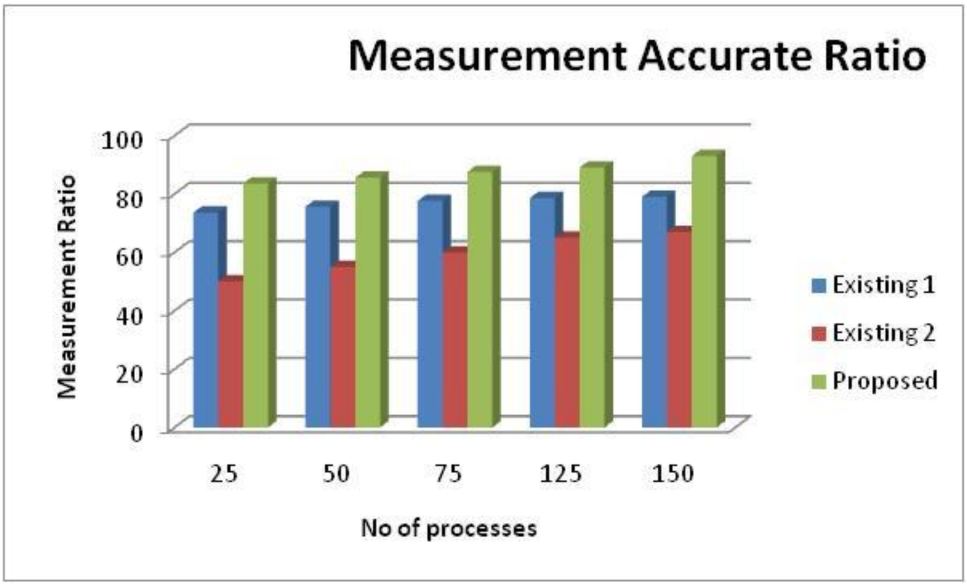

Figure 9: Explanation chart of Measurement Accurate Raatio

The clarification diagram of estimation exact proportion demonstrates the current and proposed technique esteems. No of procedures in X pivot and estimation proportion in Y hub. Each dimension of looking at the proposed technique demonstrates the better outcomes. Existing 1 esteems are 73.6-79 existing 2 esteems are 50-67 proposed strategy esteems are 83.6-93.

CONCLUSION

Proposed Constraint-based Prefix Span calculation isn't confined to ordinary Sequential Pattern Mining (SPM) parameter recurrence however joins six increasingly vital parameters like Gap, Recency, Compactness/Duration, Profitability, Item and Length. Joining of these requirements in FP-development based―Prefix Span prompts increasingly productive and viable outcomes by decrease of patterns. Compact patterns present applicable and exact outcomes regarding clients' advantage. Seven distinct trials are performed on IBM produced six engineered datasets. Correlation made for run times and pattern age of three calculations: proposed limitation based Prefix Span with RFM and Prefix Span.

REFERENCES

[1]. Y. Zheng, „„,Trajectory data mining: An overview,‟ ‟ ACM Trans. Intell. Syst. Technol., vol. 6, no. 3, p. 29, 2015.

[2]. J. Bao, T. He, S. Ruan, Y. Li, and Y. Zheng, „„,Planning bike lanes based on sharing-bikes‟ trajectories,‟ ‟ in Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, Aug. 2017, pp. 1377–1386.

[3]. Y. Fu et al., „„,Sparse real estate ranking with online user reviews and offline moving behaviors,‟ ‟ in Proc. IEEE Int. Conf. Data Mining (ICDM), Dec. 2014, pp. 120–129.

[4]. B. Fazzinga, S. Flesca, F. Furfaro, and F. Parisi, „„,Cleaning trajectory data of RFID-monitored objects through conditioning under integrity constraints,‟ ‟ in Proc. 17th Int. Conf. Extending Database Technol. (EDBT), Athens, Greece, Mar. 2014, pp. 379–390.

[5]. P. Wang, Y. Fu, G. Liu, W. Hu, and C. Aggarwal, „„,Human mobility synchronization and trip purpose detection with mixture of Hawkes processes,‟ ‟ in Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, Aug. 2017, pp. 495–503.

[6]. M. Muzammal and R. Raman, „„,Mining sequential patterns from probabilistic databases,‟ ‟ Knowl. Inf. Syst., vol. 44, no. 2, pp. 325–358, 2015.

[7]. Y. Li et al., „„,Sampling big trajectory data,‟ ‟ in Proc. 24th ACM Int. Conf. Inf. Knowl. Manage. (CIKM), Melbourne, VIC, Australia, Oct. 2015, pp. 941–950.

[8]. P. Banerjee, S. Ranu, and S. Raghavan, „„,Inferring uncertain trajectories from partial observations,‟ ‟ in Proc. IEEE Int. Conf. Data Mining (ICDM), Shenzhen, China, Dec. 2014, pp. 30–39.

[9]. M. Li, A. Ahmed, and A. J. Smola, „„,Inferring movement trajectories from GPS snippets,‟ ‟ in Proc. 8th ACM Int. Conf. Web Search Data Mining (WSDM), Shanghai, China, Feb. 2015, pp. 325–334.

[10]. Z. Feng and Y. Zhu, „„,A survey on trajectory data mining: Techniques and applications,‟ ‟ IEEE