ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

Mining Sequential Patterns using Prefix Span and Genetic Algorithms.

Prince Mary S*, Yakkala Phanideepu, and Yashwanth Reddy C.

Department of Computing, Sathyabama University, Chennai, India

ABSTRACT

Data mining is a procedure of extraction of required data from a pool of information. It is a most critical method which is utilized to locate the required information from a bigger database in an organization. Data mining device answers numerous queries which are usually time taking. Sequential pattern mining is an essential stride in information mining it is a pattern of frequent sequences from a larger sequence data in present surge it is a crucial problem in programming, equipment, client care, biologic investigation. PrefixSpan algorithm is a prominent algorithm for sequential data mining. It finds the sequential patterns through their method of growing pattern. The algorithm performs exceptionally well for small datasets. As the size builds the general time for finding the sequential patterns likewise gets increase. The PrefixSpan algorithm is run on different datasets and conclusions were drawn based on minimum support value, whereas Genetic algorithms (GA) is an optimized technique for finding genetic material in living organism using fitness function for optimization. The present study includes the comparison of PrefixSpan and genetic algorithm to know whether they show similarity if not to predict which algorithm is showing accurate performance in reasonable time span.

Keywords: Data mining, Sequential pattern mining, sequence data, prefix span, genetic algorithms.

**Corresponding author*

INTRODUCTION

The Sequential pattern mining issue was initially tended to by Agrawal and Srikant [1995] [1, 2].It was proposed that, for a specified sequential database, in which each sequence comprises a set of transactions. All these transactions are arranged according to transaction time and every transaction is a collection of items. Sequence pattern mining is made so as to find every sequential pattern based on user-defined minimum support (minsup).The minsup of a pattern is calculated from the no. of data-sequence that the pattern contains. Sequential Pattern Mining is a prominent data mining technique which finds sub-sequences and patterns which are frequently appearing in a given set of sequence. The Prefix Span algorithm which was proposed by Jian Pei et al. is most prominently used to search the sequential patterns. It is devoid of huge candidate sequence generation and thus aids in improvising the execution time and memory utilization. The paper contains the consequences resulting in execution of Prefix Span algorithm on various datasets. The principal segment of the paper gives the brief of different sequential pattern mining algorithms. The second area manages the target and the extent of the outcome which are conveyed by execution of the algorithm. The third area gives the brief about the Prefix Span calculation and its steps of execution. In fourth Section, the outcomes are drawn by executing the algorithm on various datasets.

MINING Sequential Patterns in vast databases has turned into a critical data mining task with wide applications, including business examination, web mining, security and biological investigation. It separates patterns that seem more often than a client determined least support while keeping up their item occurrence order. In this task, time is the most critical component, particularly when the outcomes are required in a limited timeframe. The Sequence Pattern mining algorithms takes quite a while to find the guidelines particularly when they are implemented on large databases. Whereas the evolutionary algorithms finds great Sequential Pattern rules within a particular time frame. These days, some evolutionary calculations were proposed and have been applied in a timely manner. Genetic Algorithm (GA), is an evolutionary algorithm, can be used to search Sequential Pattern rules in a comparatively less time frame. It is broadly used search algorithms which utilize principles enthused by natural genetic populations to progressive solutions to problems.

Procedure

The initial phase is the collect the database; hence for experimental purpose NASA dataset is considered. The next phase is data pre-processing, which is performed on the experimental dataset. This phase involves three steps namely, data purification, creation of data session and data conversion. In the third phase data mining is performed which is done using Prefix Span Algorithm and Genetic Algorithm, which are explained further in detail. The last phase performs result analysis, which is further sub divided in two parts namely, pattern analysis and result exploration.

DATA PREPROCESSING

Web use mining results can be utilized as a part of numerous applications [1], such as personalizing the presentation of web content in order to improve web configuration or e-trade destinations, to improve client route and to enhance the fulfillment. Web utilization mining can be isolated into three stages. The initial step is data cleaning and preprocessing, the next step is knowledge recovery of the preprocessed information, lastly analyzing the mining knowledge. Issues in web log mining are information volume is set high, because of this running time might increment. For navigational patterns, forecast web access log is utilized. Issues with web access log is to locate an alternate client's IP location is insufficient on the grounds that specific client's might utilize an intermediary server or might have the same IP to search the site. Second, clients might use forward and in reverse catches of the program, and these occasions are not recorded in web access logs. So there is a need to handle with missing data. In access log extra data additionally included by the server, so data which is not required for navigational example expectation additionally should have been uprooted. Third, to distinguish the distinctive searching sessions. Fourth, in a specific session time spend by the client in a last page. The previously stated issues happen in the web log mining process. There are issues in uses of web log mining process, for online navigational example forecast, the expectation is done in an auspicious way, with best exactness.

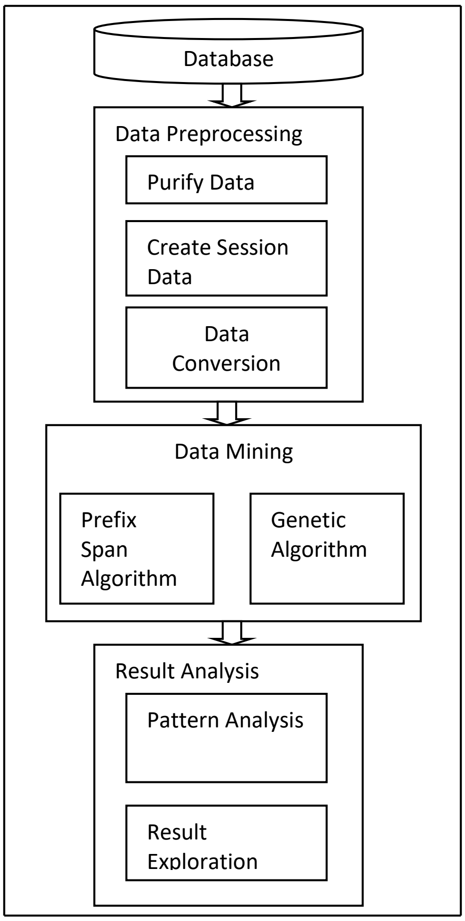


Fig 1: Architecture of Sequential Data Mining

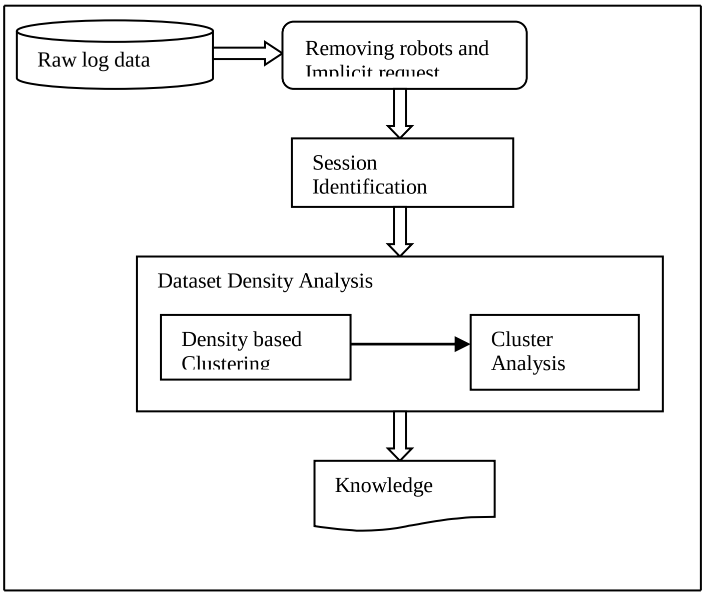


Fig 2: Architecture of Data Pre-Processing

Algorithm involved in session identification as follows:

```
Begin
    Sort page sequence according to visitor id and
    timestamp
    for every page sequence do
        divide the current page sequence by
        threshold
            if the page sequence is single
                add current page sequence to
                session list
            else
                if there exists a shared pattern
                    between the subsequence
                    add subsequence to the
                    session list
                else
                    skip the sequence
    end for
end
```

PreFix span algorithm

A pattern-growth method based on projection is used in Prefix Span algorithm [5] for mining sequential patterns. The basic thought behind this strategy is, instead of viewing sequence databases by assessing the frequent occurrences of sub-sequences, it is made on frequent prefix. This decreases the handling time which eventually increase the algorithm proficiency. Jian Pei et al. proposed a novel algorithm known as Prefix Span (Prefix-projected Sequential Pattern Mining) algorithm [5] which deals with projection of database and sequential pattern growth. The separation and search space procedure is implemented by Prefix Span. Algorithm mines sequential patterns through following steps:

I. Find length-1 in sequential patterns. The given sequence S is scanned to get prefix that occurred frequently in S.

Number of times the prefix occurs = length-l of that prefix.
Length-l is given by notation <pattern> : <count>.

ii. Divide the search space based on the prefix that derived from first step, the whole sequential pattern set is partitioned in this phase.

iii. Find the subsets of sequential patterns. The projected databases are constructed and sequential patterns are mined from these databases. Only the local frequent sequences [8], [9] are explored in the projected databases in order to increase the sequential patterns. The cost for constructing projected database is quite high. Bi-level projection and pseudo-projection are the methods used for reducing this cost which ultimately increases the algorithm's efficiency.

Genetic Algorithm

In a general search algorithm which employs the principles inspired by natural selection of genetic population to develop answers for the problems. All GAs normally begins from a set, known as population, for arbitrary solutions (candidate). These solutions were evolved by the redundant selection and variations of more fit solutions, by following the Darwin's principle of survival of the fittest. The components of the population are known as individual chromosomes, representing candidate solutions. Chromosomes are commonly chosen by nature of arrangements they represent. To 1 each sequence in sequential pattern is taken as a rule to estimate the nature of a solution, fitness function is given to each and every chromosome in the populace. Henceforth, the better the fitness of a chromosome, the more probability the chromosome has of being chosen for reproduction and the more of its genetic material will be forwarded to the next generation. Genetic Algorithms are relatively simple to develop and validate, which makes them very attractive, if used. The algorithm is parallel and can be used in huge populations effectively. Utilization of mutation makes the technique capable of recognizing global optimal, even in extremely troublesome. The strategy does not require information about the appropriation of the data, in this way Gas can efficiently discover the space of possible solutions which is known as search space, and it comprises all the possible arrangements that can be encoded [6]. Genetic algorithms are good at taking substantial, possibly huge search space and exploring them, searching for optimal combination of this solutions one may not otherwise discover in a lifetime. Genetic Algorithm (GA), displayed in [4, 6, 8], is a portion of developmental computing. Genetic algorithm begins with a set of solutions (represented by chromosomes) called population. Results from one population are taken and used to shape another population by mutation and crossover. This is inspired by a hope, that the new population will be superior to the old one. Best solutions which are chosen to frame new solutions (offspring) are chosen by best fitness. This is repeated until some condition (improvement) is fulfilled. To measure the quality of a solution, fitness function is given to each chromosome in the population.

OBJECTIVE OF THE WORK

The Prefix Span algorithm and genetic algorithms are run on same datasets. The sizes of datasets are increased gradually so as to check the execution of algorithm from a small datasets to relatively larger datasets various parameters like memory utilization, time complexity and size of the projected dataset are set as benchmark for evaluating the results derived by algorithm to algorithm for NASA datasets. This is done to get idea which algorithm provides good results between these two.

Evolution of Prefix span:

Sequence Dataset: It is a set of sequences where each sequence having a list of item sets.
Item set: It is an unordered set of unique items.

Support of a sequential pattern: It is the total number of sequences which are calculated by dividing the pattern that occurs with the total number of sequences in the database.

Frequent sequential pattern: The sequential pattern is that which is having a support more than the minimum support parameter which will be provided by the user. The input of Prefix Span is a sequence database and a user required threshold named minimum support (minsup).
The below table contains four sequences as S1, S2, S3 and S4. The first sequence, S1, contains 7 item sets. All the items are sorted based on lexical order. The repetition of items in the same item set should not be occurred.

Table 1. Sequence database ID Sequences

S1 (1)	(2)	(1 2)	(3)	(1 3)	(4 5)	(6)
S2 (3 4)	(3)	(2 3)	(1 4)			
S3 (4 5)	(2)	(2 3 4)	(3)	(1)		
S4 (4)	(5)	(1 6)	(3)	(2)	(7)	(1)

Prefix Span discovers all frequent sequential patterns occurring in a sequence database which is having a value >minsup value which is provided by the user.

A sequence SA = P1, P2 ...Pm, where P1, P2... Pm are item sets is said to occur in another sequence SB = Q1, Q2 ...Qn, where Q1, Q2... Qn are

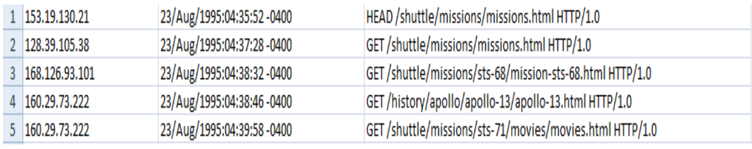
Item sets, if and only if there exists integers like 1 <= i1 < i2... <im<= n such that P1 ⊆ Qi1, P2 ⊆ Qi2 ... Pn⊆ Qin. For example, if we run Prefix Span with minsup equals to 0.5 and with a maximum pattern length(l) of 50 items, 47 sequential patterns are found. An example of pattern that was found is {(2, 3), (4)} which appears in the first and the second sequences. This pattern has l of 3 because it contains three items. Another pattern is "(4 5), (3), (1)". It appears in the third and fourth sequence (thus it has a support of 0.5). It also has a length of 4 because it contains 3 items.

Evolution of Genetic Algorithm:

MINING SEQUENTIAL PATTERNS IN NASA DATABASE USING GA

In this particular section, we illustrate our Genetic Algorithm for mining Sequential Patterns in NASA Database, which is called as SPT-GA algorithm. Firstly, we represent our chromosome structure and encoding schema, genetic operators, and then we characterize the assignment of fitness and selection criteria. Finally, we give the structure of SPT-GA algorithm. 4.1 Chromosome. In this section, we describe the used structure of GA chromosomes andit's represented in this paper.

Structure. In NASA Database, chromosomes are created using time stamp values.



1	153.19.130.21	23/Aug/1995:04:35:52 -0400	HEAD /shuttle/missions/missions.html HTTP/1.0
2	128.39.105.38	23/Aug/1995:04:37:28 -0400	GET /shuttle/missions/missions.html HTTP/1.0
3	168.126.93.101	23/Aug/1995:04:38:32 -0400	GET /shuttle/missions/sts-68/mission-sts-68.html HTTP/1.0
4	160.29.73.222	23/Aug/1995:04:38:46 -0400	GET /history/apollo/apollo-13/apollo-13.html HTTP/1.0
5	160.29.73.222	23/Aug/1995:04:39:58 -0400	GET /shuttle/missions/sts-71/movies/movies.html HTTP/1.0

Fig. 3: Experimental Chromosome Structure

Representation. In GA, there are many alternative ways to represent a chromosome based on different problems like integer and binary representations. To decide the appropriate representation used for Sequential Pattern rules, we have used the short, low-order schemata, which are appropriate to the underlying problem and relatively not related to schemata over standard positions. Also we are supposed to choose the smallest alphabet that allows a natural expression of the problem, presented in [8]. In SPT-GA algorithm, we are using the binary representation because it is the more desirable for our algorithm and it consumes less memory and it displays the required information (element occurred or not).

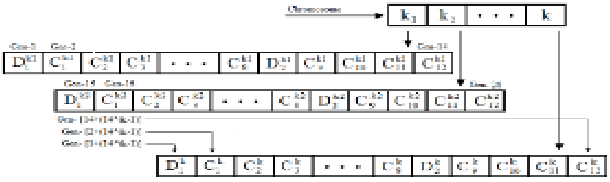


Fig. 4: Chromosome Representation

From Figure.2 it is understood that order cannot be extracted directly. In order to find solution this problem, we have decided to associate the transaction sequence as a metadata of with each and every chromosome. For that, we have used Vertical Bitmap Representation, which was

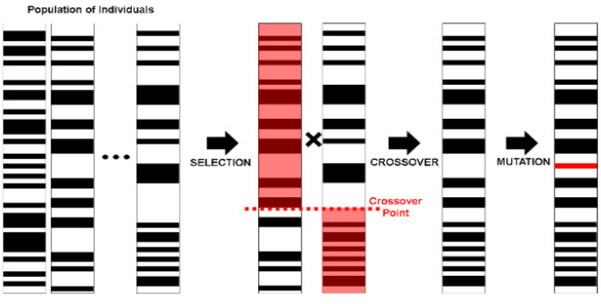


Fig 3

Figure 3: Crossover and Mutation Example

GA will repeat the operation until the best results are obtained.

Fitness Function. Shigeaki, Youichi, Ryohei, in [10], have proposed a method that discovered sequential pattern according to the interests of users without using background knowledge. They have defined a new perspective known as sequential interestingness measure (SIM).Definition 2: The sequential interestingness measure of a rule A->C is:

SIM(A->C) = minCi∈ C {(Confidence(A | Ci))α} × Support(AC) where (α ≥ 0) is a priority that represents how vital the frequency of the pattern is, Ci is sup sequence of C, it shows the condition of sequence C, and i = 1 … n

where n is the number of condition of C. The initial term of the criterion determines the frequencies of the given sub-patterns are not often while the 2nd term evaluates the frequency of the pattern that are repeated.

4.4 SPT-GA Algorithm. In this section, we represent SPT-GA algorithm that is proposed.

```
SPT-GA Algorithm
else If A is given then
Evaluate the fitness F (i, A, α)
Else If C is given then
Evaluate the fitness F (i, C, α)
```

CONCLUSION

The performance of existing Prefix Span algorithm is evaluated by running the algorithm on different datasets. Two parameters minimum and maximum prefix length are provided at begining of the execution of algorithm. The sequences having value greater than minimum support are extracted from sequential datasets. Minimum support is the no.of sequences which are calculated by seprating the patterns occurs with the total no. of sequences in the database. The maximum prefix pattern value used to specify the length of the sequence to be there in output sequential patterns which is beneficial while executing the algorithm on large datasets. For getting the sequential output based on minimum and maximum prefix length, the two parameters time complexity and memory utilization are set as the benchmark for performance evaluation of algorithm on different datasets. Both the parameters vary from one dataset to other. Genetic Algorithm to get frequent sequences in NASA DB in order to help SPT-GA algorithm uses the characteristics of evolutionary algorithm that finds best rules in a less span of time with meaningful results. The scope of mining sequential patterns continues with further in the result analysis produced by the comparision of outcomes obtained from Prefix Span Algorithm and Genetic algorithm respectively.

REFERENCES

[1] R Agrawal and R Srikanth, 1995. Mining sequential patterns,

[2] In Proceedings of 1995 International Conference Data Engineering (ICDE'95), pp. 3- 14,Taipei, Taiwan.

[3] R Agrawal and R Srikant, 1994. Fast algorithms for mining association rules, In Proc. 1994 Int. Conf. Very Large Data Bases (VLDB'94), pp. 487- 499, Santiago, Chile.

[4] M. Zaki, 2001. SPADE: An Efficient Algorithm for Mining Frequent Sequences, Machine Learning, vol. 40, pp. 31- 60.

[5] Han J, Dong G, Mortazavi-Asl B, Chen Q, Dayal U, Hsu MC. 2000. Freespan: Frequent pattern-projected sequential pattern mining, In Proceedings 2000 Int. Conf. Knowledge Discovery and Data Mining (KDD'00), pp. 355-359. 2000.

[6] Jian Pei, Jiawei Han, BehzadMortazavi, UmeshwarDayal, 2004. Mining Sequential Patterns by Pattern-Growth: The PrefixSpan Approach, IEEE transactions on knowledge and data engineering, Vol. 16, pp. 1424- 1440.

[7] Agrawal R. and Srikant R. Mining Sequential Patterns. IBM Almaden

[8] Antunes C. and Oliveira A. Sequential Pattern Mining.

[9] Ayres J. Gehrke J. Yiu T. and Flannick J. Sequential Pattern Mining

[10] Blum C. and Li X. Swarm Intelligence in Optimization. Natural

[11] Colombetti M. and Dorigo M. Training Agents to Perform Sequential Behavior. S.Prince Mary and Dr. E. Baburaj, Performance Enhancement in Session Identification