e-ISSN: 2349-9745
p-ISSN: 2393-8161

Scientific Journal Impact Factor (SJIF): 1.711

International Journal of Modern Trends in Engineering and Research

www.ijmter.com

A REVIEW PAPER ON AN ENHANCEMENT TO PROJECTED SEQUENTIAL PATTERN MINING ON PREFIXSPAN ALGORITHM

Rahul Saxena[1], Tanvi Varma[2]

[1] *Research Scholar CSE Department, Parul Institute of Technolody, Limda, Vadodara, India*

[2] *Asst. Prof., CSE Department, Parul Institute of Technolody, Limda, Vadodara, India*

Abstract— Data mining is the process of extracting interesting information or patterns from large information repositories. Mining sequential patterns is to discover sequential purchasing behaviours for most customers from a large amount of customer transactions. However, now a day due to bulky data, it has become much more difficult to access relevant information from the database with the explosive growth information available on the transactions. Then a comprehensive study on Prefix span algorithm, a projection position-based sequential pattern mining algorithm is used solve the resource problem by reducing unnecessary storage space and scanning time in its possible way but does not apply any specific constraints in order to mine sequential pattern in optimization manner.

Keywords--Data mining, sequential pattern mining, PrefixSpan Algorithm.

I. INTRODUCTION

1.1 Overview

Data mining refers to extracting or “mining” knowledge from large amounts of data[1]. It is a process of extracting the hidden valuable information from data. Data mining also known as “knowledge mining from data”[2]. “Knowledge mining” a shorter term, may not reflect the emphasis on mining from large amounts of data. Nevertheless, mining is a vivid term characterizing the process that finds a small set of precious nuggets from a great deal of raw material. Thus, such a misnomer that carries both “data” and “mining” became a popular choice. Many other terms carry a similar or slightly different meaning to data mining, such as knowledge mining from data, knowledge extraction, data/pattern analysis, and data archaeology.

1.2 Data Mining Concepts

Data mining commonly involves various tasks: [1]

• **Clustering** - is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.

• **Classification** – is the task of generalizing known structure to apply to new data. For example, an email program might attempt to classify an email as legitimate or spam. Common algorithms include decision tree learning, nearest neighbour, naïve Bayesian classification and neural networks.

• **Regression** – Attempts to find a function which models the data with the least error.

• **Association rule learning** - Searches for relationships between variables. For example a supermarket might gather data on customer purchasing habits.

Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis. ***Sequential Pattern*** is a frequently occurring subsequence such as the pattern that customers tend to purchase first a PC followed by a digital camera with a memory card. Mining frequent patterns leads to the discovery of interesting associations and correlations within data.

1.3 Sequential Pattern Mining

Sequential pattern mining is an important data mining task with broad applications. It is the task of discovering frequent subsequence as patterns in a sequence database. Sequential pattern mining deals with data represented as sequences. Mining frequent sequential patterns has found a host of potential application domains, including retailing (i.e., market-basket data), telecommunications, and, more recently, the World Wide Web (WWW).

In market-basket databases, each data sequence corresponds to items bought by an individual customer over time and frequent patterns can be useful for predicting future customer behavior.

In telecommunications, frequent sequences of alarms output by network switches capture important relationships between alarm signals that can then be employed for online prediction, analysis, and correction of network faults.

Finally, in the context of the WWW, server sites typically generate huge volumes of daily log data capturing the sequences of page accesses for thousands or millions of users. Discovering frequent access patterns in WWW logs can help improve system design and lead to better marketing decisions (e.g., strategic advertisement placement).

Sequential Pattern mining approach can be classified in below figure.

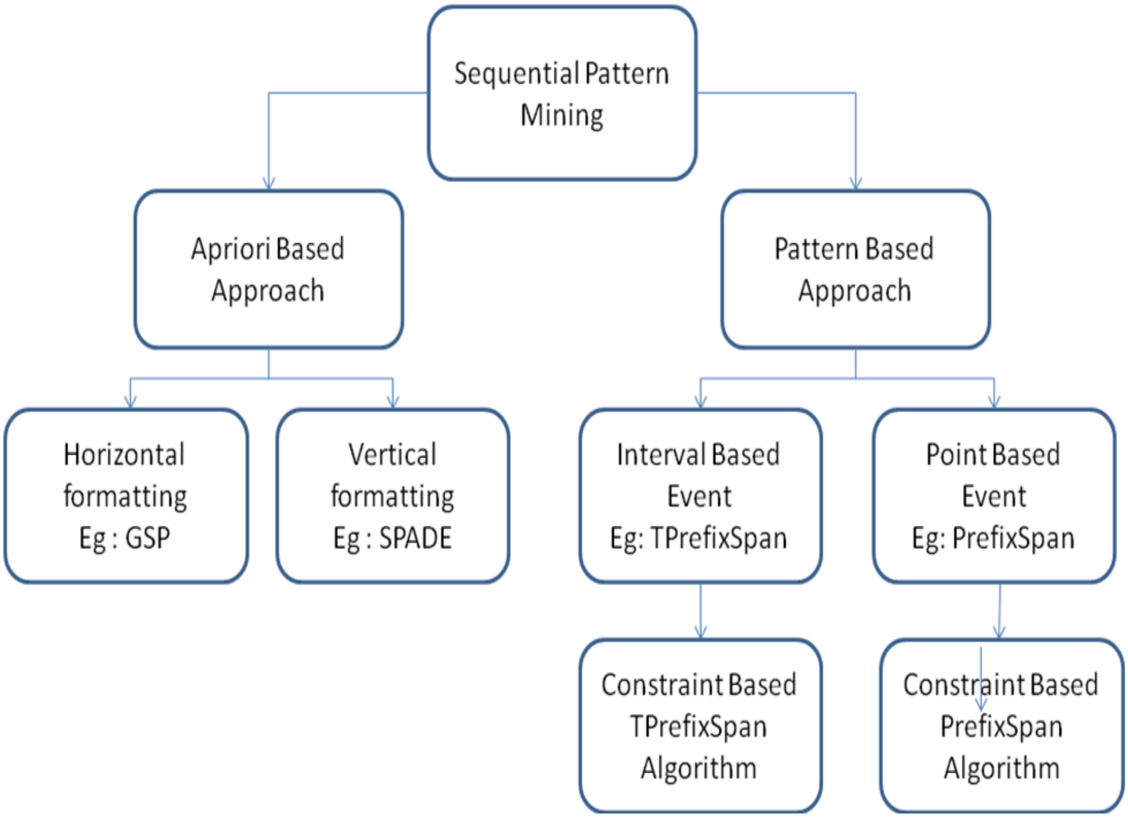
Sequential Pattern Mining
Apriori Based Approach
Pattern Based Approach
Horizontal formatting Eg : GSP
Vertical formatting Eg : SPADE
Interval Based Event Eg: TPrefixSpan
Point Based Event Eg: PrefixSpan
Constraint Based TPrefixSpan Algorithm
Constraint Based PrefixSpan Algorithm

Figure 1: Sequential Pattern mining approach

II. PREFIXSPAN ALGORITHM

This is the only projection-based algorithm among the sequential pattern mining algorithms. PrefixSpan is the fastest algorithm among all the algorithms[8]. It outperforms algorithms like Apriori, FreeSpan, SPADE. It uses divide and search space technique. But in PrefixSpan there are only limited insertions, deletions, and mutations in their sequential patterns. It is an efficient pattern growth method. It outperforms both GSP and FreeSpan. The main idea of prefix span algorithm is that it explores prefix-projection in sequential pattern mining and mines the complete set of patterns, but reduces the effort of candidate subsequence generation. Prefix-projection reduces the size of projected database and leads to efficient processing.

2.1 Algorithm (PrefixSpan)

Input: A sequence database S and the minimum support min_sup.
Output: The complete set of sequential patterns
Method: Call PrefixSpan(<>,0,S)
Subroutine PrefixSpan(α, l, S| α)
Parameters: α : a sequential pattern; l: the length of α; S| α: the α projected database,
if α ≠ <> ; otherwise the sequence database S.

Method:

```
1. Scan S| α once, find the set of frequent items b such that
    (a) b can be assembled to the last element of α to form a sequential pattern; or
    (b) <b> can be appended to α to form a sequential pattern.
2. For each frequent item b, append it to α to form a sequential pattern α'; and output α'.
3. For each α', construct α'-projected database S| α' and call PrefixSpan(α', l+1, S| α') .
```

2.2 Algorithm Analysis

PrefixSpan for the sequence database S with min_sup sequential patterns in S can be mined by a prefix-projection method in the following steps.

Step 1: Find length-1 sequential patterns. Scan S once to find all frequent items in sequences. Each of these frequent items is a length-1 sequential pattern. Where <pattern>: count represents the pattern and its associated support count.

Step 2: Divide search space. The complete set of sequential patterns can be partitioned into the subsets according to the prefixes.

Step 3: Find subsets of sequential patterns. The subsets of sequential patterns can be mined by constructing corresponding *projected databases* and mine each recursively.

The final step in processing a query is the evaluation phase. The best estimation plan candidate generated by the optimization engine is chose and then accomplished. Besides processing a query in a simple sequential manner, some of a query's individual operations can be processed in parallel either as unconventional procedures or as interdependent pipelines of procedures or threads [6].

III. DIFFERENT TECHNIQUES

PrefixSpan algorithm can be extended in many ways considering point based events, interval based events and also by adding constraints to interesting patterns. Using bi-level projection and pseudo-projection in PrefixSpan algorithm may improve mining efficiency.

A. I-PrefixSpan algorithm :

It is the improved algorithm of PrefixSpan algorithm[14]. The idea of this I-PrefixSpan algorithm is to use sufficient database for Sequential Tree framework and separator database to reduce the execution time and memory usage. In I-PrefixSpan there is no in-memory database stored after the construction of index set. This I-PrefixSpan algorithm improves PrefixSpan in two ways: (1) to build in-memory database sequence and to construct the index set, it implements sufficient database for Sequential Tree framework and (2) instead of whole in-memory database, to store the transaction alteration sign sit implements Separator Database. In this algorithm there is no time constraint and sliding window are used to improve the performance of the output.

B. *P-PrefixSpan algorithm :*

There is no method for extracting a probability of time in the sequential pattern mining process. Besides minimum support-count constraint, this approach imposes minimum time-probability constraint, i.e., the P-PrefixSpan algorithm is developed by modifying the well-known PrefixSpan algorithm[15]. The new algorithm can discover frequent sequential patterns with probability of inter arrival time of consecutive items. The added constraints could filter out less important patterns and reduce the memory space required in storing projected databases. This algorithm is more efficient and scalability is also high when compared to other PrefixSpan algorithms.

C. CFM-PrefixSpan algorithm :

This algorithm is designed for mining all CFM (Compact Frequent Monetary Prefix Span) sequential patterns from the given customer transaction database[16]. The CFM-PrefixSpan algorithm employs a pattern growth methodology that finds sequential patterns by utilizing a divide-and-conquer strategy. Besides discovering CF-sequential patterns the compact frequent items and CFM sequential patterns are also discovered. The CFM algorithm has been validated on real and synthetic sequences. The result of this algorithm shows that the effectiveness of sequential pattern mining algorithm can be improved significantly by incorporating monetary and compactness into the mining process.

D. *DRL-PrefixSpan algorithm :*

DRL (Downturn, Revision, and Launch)[17] PrefixSpan is designed specifically to incorporate the specific constraints which involves many steps: i) Product Downturn, ii) Product Revision, iii) Product Launch. Each of these scenarios is characterized by distinct item and adjacency constraints. This algorithm was developed for mining all length DRL patterns. It has been validated on synthetic sequential databases. It gives the effectiveness of incorporating the promotion-based marketing scenarios in the sequential pattern mining process.

E. *C-PrefixSpan algorithm :*

To save the computation cost and enhance the performance, many types of constraints can be used in sequential pattern mining like item constraint, aggregate constraint, length constraint and gap constraint[18]. Constraint based sequential pattern mining extracts the patterns according to the user's interest. The patterns obtained from C-PrefixSpan are comparatively very less and more valuable than PrefixSpan algorithm. When the number of transaction per sequence increases the performance of C-PrefixSpan algorithm also increases

F. TPrefixSpan algorithm :

TPrefixSpan is very similar to PrefixSpan. The TPrefixSpan algorithm is developed for mining the new temporal patterns from interval-based events[11]. Mining temporal patterns is much more complicated than mining sequential patterns and the methods for discovering a sequential pattern can neither be used directly nor be applied with slight modifications to discover temporal patterns.

IV. CONCLUSION

In this paper, we have discussed about the data mining and sequential pattern mining approach. PrefixSpan is the fastest algorithm among all the algorithms among sequential pattern mining approch. PrefixSpan algorithm has more advantages like speed, less database scans and high performance. PrefixSpan algorithm can be extended in many ways considering point based events, interval based events and also by adding constraints to interesting patterns.

REFERENCES

Book:

[1] Jiawei Han & Michelline Kamber, *"Data Mining – Concepts & Techniques"*, Morgan Kaufmann Publishers (Academic Press), 2001

[2] "*Research Methodology*" By C. R. Kothari

Web Reference :

[3] http://illimine.cs.uiuc.edu/

Research Papers :

[4] R. Agrawal and R. Srikant, *"Mining sequential patterns"*, In Proceedings of the 1995 International Conference on Data Engineering, pp. 3-14, 1995.

[5] N Ju-Dong Ren, Yin-Bo Cheng, Lung-Lung Yang, *"An Algorithm for Mining Generalized Sequential Patterns"*, Proceedings of the Third International Conference on Machine Learning and Cybernetics, Shanghai, 26-29 August, 2004.

[6] A. Mortazavi, Q. Chen, U. Dayal, M. Hsu, J. Han and T. Pei, *"FREESPAN: Frequent Pattern Projected sequential pattern mining"*, Proc ACM SIGMOD, 2000.

[7] Taoshen Li, Weina Wang, Qingfeng Chen, *"On the Sequential Pattern Mining Algorithm Based on Projection position"*, IEEE 978-1-4673-4463-0/13, April 2013 and The 8th International Conference on Computer Science & Education (ICCSE 2013).

[8] Jian Pei, Jiawei Han, BehzadMortazavi-Asl, Jianyong Wang, Helen Pinto, Qiming 49 Chen, UmeshwarDayal, and Mei-Chun Hsu, *"PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth"*, IEEE Transactions on Knowledge and Data Engineering, Vol. 16, No. 10, October 2004.

[9] Jia-dong Ren, Yuan Dong, Hai-tao He, *"A Parallel Algorithm Based on prefix tree for Sequence Pattern Mining"*, 978-0-7695-4332-1/10 © 2010 IEEE and DOI 10.1109 and CDEE.2010.10

[10] Sha Jin, Hu Yingxin, Jia Lianjuan, *"Efficient Sequential Pattern Mining Algorithm by Positional Data"*, 978-0-7695-4539-4/11 © 2011 IEEE and International Conference on Internet Computing and Information Services (ICICIS) 2011.

[11] R. Agrawal and R. Srikant, *"Mining sequential patterns: generalizations and performance improvements"*, In Proceedings of the 5th International Conference on Extending Database Technology, pp. 3-17, Avignon, France, 1996.

[12] Show-Jane Yen and Yue-Shi Lee, *"Mining Sequential Patterns with Item Constraints"*, DaWaK 2004: data warehousing and knowledge discovery: International conference on data warehousing and knowledge discovery, Zaragoza, ESPAGNE, vol. 3181, pp. 381-390, 2004.

[13] Helen Pinto, and Jiawei Han , *"Multidimensional Sequential Pattern Mining"*, In Proceedings of the 10th International Conference on Information and Knowledge Management, pp 81 - 88 , Atlanta, Georgia, USA , 2001.

[14] R. DhanySaputra, Dayang, A. Rambli and O.Foong, *"Mining Sequential Patterns Using I-PrefixSpan"*, International journal of electrical and electronics engineering", pp. 338-342, 2008.

[15] H.Shyur, C. Jou and K. Chang, *"A data mining approach to discovering reliable sequential patterns"*, pp 08, 2008.

[16] B.Mallik and D.garg, *"CFM PrefixSpan: A pattern growth algorithm incorporating monetary and compactness"*, pp 4509-4555, July 2012.

[17] A. George and D. Binu, *"DRL-Prefixspan: A novel pattern growth algorithm for discovering downturn, revision and launch (DRL) sequential patterns"*, pp. 426-439, Dec 2012.

[18] J. Pei, J. Han and W. Wang, *"Constraint-based sequential pattern mining: the pattern growth methods"*, J Intell. Inf. Syst, Vol. 28, No.2, pp. 133 –160, 2007.