© 2015 IJSRSET | Volume 1 | Issue 6 | Print ISSN : 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

A Survey on Different Approaches for Sequential Pattern Mining

Bhargav Shroff[1], Prof. Bakul Panchal[2], Manmitsinh Zala[3]

[1, 2] Information Technology Department, L. D. Engineering College, Ahmedabad, Gujarat, India
[3]Information Technology Department, Alpha college of engineering and Technology, Ahmedabad, Gujarat, India

ABSTRACT

In data mining, mining sequential pattern from very huge amount of database is very useful in many applications. Most of sequential pattern mining algorithms work on static data means the database should not change. But the databases in today's real world application do not have static data, they are incremental databases. New transactions are added at some intervals of time. For updated database, the algorithm needs to be executed again for whole sequence database. So those approaches are not appropriate to use, for that algorithm with incremental approach should be modelled and used. This paper analysis existing approaches for finding sequential pattern mining, and the survey would be helpful in forming a new model or improving some existing approach to handle incremented database & obtain sequential patterns out of them.

Keywords : BLSPM, Incremental approach, IncSpan, PrefixSpan, Sequential Pattern mining.

I. INTRODUCTION

Data Mining has been considered as a very important area of research since many decades. Mining useful and unknown knowledge from vast amount of databases has remained its goal. There are various techniques in data mining like association rules mining, classification, clustering, sequential pattern mining, etc. Sequential pattern mining was first of all introduced by R. Agrawal and R. Srikant. Sequential pattern mining is basically obtaining frequently occurring ordered events or some subsequence from database.

Sequential pattern mining has practical value in many fields such as pattern analysis of customer purchase behaviour, disease diagnosis, natural disaster prediction and DNA sequence analysis. [3]

A sequential pattern is a relatively common sub-sequence of transactions, where each transaction is a set of items (Itemset).[5] For example, each customer record in the transactional database is an Itemset associated with the transaction time and a customer-id.[5] Data having the same customer-id are sorted by ascending transaction time into a data sequence before mining. If a sufficient number of customers in the transactional database have the purchasing sequence of PC, printer, and printing software, then such a sequence is called a sequential pattern. [5]

Mostly sequence mining algorithms works on static databases, i.e. the data should not change in the database. If the database is updated then the database needs to be rescanned and again that particular algorithm should be applied on them.

But the database is not static in practice; new transactions would be added to the databases. So the sequential patterns should be mined in incremental sequential database in such a way that whole database need not be rescanned in the process.

In real world applications, the database changes to little extent. For example, a retail sales database is updated each month, sales data for the new month often represent only a small percentage of the previous ten year's sales data.[5] Sequential patterns also change very little, so applying the algorithm on whole updated database is waste of time and cost. So for that some incremental approach needs to be modelled that would consider only

the incremented fragment of database. For that various existing approaches are surveyed in this paper.

II. METHODS AND MATERIAL

LITERATURE REVIEW

A. Sequential Pattern Mining using A Bitmap Representation [1]

In this paper author propose a new method for mining sequential patterns. The algorithm is especially efficient when the sequential patterns in the database are very long. We introduce a novel depth-first search strategy that integrates a depth-first traversal of the search space with effective pruning mechanisms.

In this paper, they propose an efficient algorithm called SPAM (Sequential PAttern Mining) that integrates a variety of old and new algorithmic contributions into a practical algorithm. SPAM assumes that the entire database (and all data structures used for the algorithm) completely fit into main memory. Considering the computational complexity that is involved in finding long sequential patterns even in small databases with wide records, this assumption is not very limiting in practice.

SPAM when developed in this paper was the first depth-first search strategy for mining sequential patterns. An additional salient feature of SPAM is its property of online outputting sequential patterns of different length compare this to a breadth-first search strategy that first outputs all patterns of length one, then all patterns of length two, and so on. Our implementation of SPAM uses a vertical bitmap data layout allowing for simple, efficient counting.

The algorithm is based on lexicographic tree of sequences. This tree has two main type of extensions, which are Sequence extension step (S-step) and Itemset extension step (I-step). It also explains traversing of tree and how different pruning techniques are used for S-step and I-step.

This algorithm is also compared to the SPADE and PrefixSpan and the experimental results are shown in this paper. The experimental result shown that the algorithm out performed over SPADE and PrefixSpan on large datasets by over an order of magnitude.

B. Prediction of Students Performance Using Frequent Pattern Tree [2]

In this paper author proposes a model that predicts the performance of student from an educational database using Frequent Pattern Tree. Prediction of student's performance in education institution is one way to reach the good quality in education system. Educational institution staff should identify students who are likely to fail in exams.

In this paper, real world data of engineering students has been collected for the study using different surveys and college reports. For each student, the database contains general information such as the 10th, 12th percentage, test marks in each subject, attendance in each subject, internal, external marks in each subject, gender, mode of transport etc.

Sequential pattern mining algorithms has been applied to many research areas such as diagnosis of disease, analysis of consumer behaviour, analysis of scientific experiments and so on. Proposed system uses two data mining techniques to predict the performance of the students. First is the Generalized Sequential Pattern mining algorithm, GSP which finds frequent patterns from the student database. And next is Frequent Pattern, FP tree which builds tree based on frequent patterns of pattern mining algorithm. FP tree is efficient for mining both short and long frequent Patterns.

GSP algorithm has very good scale-up properties with respect to the average data sequence size. The GSP algorithm consists of two phases- join phase and prune phase.

The algorithm starts with finding support for each attributes in student database. Attributes which exceeds threshold are considered as frequent items in 1st pass. The frequent items found in previous pass are used for generating frequent candidates in next pass. The algorithm terminates when there are no candidate sequence generated at the end of the pass.

Experimental results show that, the algorithm has more time complexity with respect to small records. The algorithm is efficient for large size files. The

performance of the algorithm also depends on the threshold or support selection. Low threshold requires more time as compared to the high threshold. Also, memory space used by an algorithm is more with low threshold. Based on these frequent patterns, FP-tree algorithm is applied for classifying the data into two classes as high performance or low performance that is, pass or fail. The students who are likely to fail by prediction can be guided for improvement after prediction.

C. A Improved PrefixSpan Algorithm For Sequential Pattern Mining [3]

In this paper author has improved the PrefixSpan algorithm for sequential pattern mining.
PrefixSpan algorithm uses the idea of Divide and Conquer to generate frequent suffix items through frequent prefix projection and then generate new patterns by connecting them. However, there are some shortcomings of this algorithm. First, it requires large resources to construct projected database recursively. Second, the algorithm requires repeatedly scan projected database, which will reduce the efficiency of the algorithm. Therefore, author presents an improved sequential pattern mining algorithm BLSPM.

The algorithm use duplicated projection and certain specific sequential patterns pruning, reduce the scale of projected databases and the runtime of scanning projected databases, thus increase the mining efficiency.

BLSPM could reduce the number and the size of projected databases; the experimental results prove that it is greatly efficient to mining sequential patterns in large databases. Experimental result shows that BLSPM and algorithm PrefixSpan are almost equal; when the support thresholds is low, algorithm BLSPM is significantly better than algorithm PrefixSpan in operating efficiency. This is because when the support is low, algorithm PrefixSpan needs structure a lot of projected database. Therefore using spacer projection can greatly reduce the number of the projected database that needed to structure, thereby reducing the scanning time.

BLSPM algorithm can greatly reduce the number of scanning projection database, in the case of large data, the time consuming of algorithm BLSPM can be apparently less than algorithm PrefixSpan. Because in the case of large datasets, the number of the same sequences in projected database with prefix is more. Using algorithm BLSPM can greatly reduce the number of scanning projected database, so the algorithm efficiency is obviously improved.

D. IncSpan: Incremental Mining of Sequential Patterns in Large Database[4]

In this paper, author develops an algorithm develop an efficient algorithm, IncSpan, for incremental mining of sequential patterns, by exploring some interesting properties. In this method, they buffer semi-frequent patterns, which can be considered as a statistics-based approach.

The technique is to lower the min sup by a buffer ratio mu<= 1 and keep a set SFS in the original database D. This is because since the sequences in SFS are “almost frequent”, most of the frequent sub sequences in the appended database will either come from SFS or they are already frequent in the original database. With a minor update to the original database, it is expected that only a small fraction of sub sequences which were infrequent previously would become frequent.

For an original database D, an appended database D0, a threshold min sup, a buffer ratio mu, a set of frequent sequences FS and a set of semi-frequent sequences SFS, we want to discover the set of frequent sequences FS0 in D0.

Outline of developed algorithm is as follows:
Step 1: Scan LDB for single items.
Step 2: Check every pattern in FS and SFS in LDB to adjust the support of those patterns.
Step 2.1: If a pattern becomes frequent, add it to FS0. Then check whether it meets the projection condition. If so, use it as prefix to project database.
Step 2.2: If a pattern is semi-frequent, add it to SFS0.

IncSpan outperforms the non-incremental method (using PrefixSpan) and a previously proposed incremental mining algorithm ISM by a wide margin. It is a promising algorithm to solve practical problems with many real applications. There are many interesting research problems related to IncSpan that should be pursued further. For example, incremental mining of

closed sequential patterns, structured patterns in databases and/or data streams are interesting problems for future research.

III. CONCLUSION

In this survey paper we have reviewed many different methods and techniques that are used for finding sequential pattern mining. Many past techniques are available, but they are suitable only for static databases. Some techniques that have incremental approaches are also considered. Techniques and algorithms in this paper have their own advantages and disadvantages. As a future work we would try to improve one of these algorithms for finding sequential pattern mining with incremental approach.

IV. REFERENCES

[1] "Sequential PAttern Mining using A Bitmap Representation", Jay Ayres, Johannes Gehrke, Tomi Yiu, and Jason Flannick, in ACM.

[2] "Prediction of Students Performance Using Frequent Pattern Tree", Priyanka Anandrao Patil, R. V. Mane, in 2014 Sixth International Conference on Computational Intelligence and Communication Networks, IEEE.

[3] "A Improved PrefixSpan Algorithm For Sequential Pattern Mining", Liang Dong, Wang hong, in 2014 IEEE

[4] "IncSpan: Incremental Mining of Sequential Patterns in Large Database", Hong Cheng, Xifeng Yan, in ACM

[5] "Incremental Discovery of Sequential Patterns Using a Backward Mining Approach", Ming-Yen Lin,Sue-Chen Hsueh,Chih-Chen Chan, in 2009 IEEE