A Data Set Compression Based Efficient Technique for Mining Sequential Items

[1]Rajesh Gupta, [2]Prof. Rakesh Pandit

[1]Research Scholar, [2]Assistant Professor

Abstract: **The data mining is a process of analysis of the data and extraction of the essential patterns from the data. These patterns are used with the different applications for making decision making and prediction related task. The decision making and prediction is performed on the basis of the learning of algorithms. Sequential pattern mining plays a vital role in organizations in decision making. This paper presents a data set reduction based technique for mining all the sequential patterns from a transaction data set. The proposed method will save computation time and memory space.**

1. Introduction

Data: Data are any facts, numbers, or text that can be processed by a computer. Today, organizations are accumulating vast and growing amounts of data in different formats and different databases. This includes:

- operational or transactional data such as, sales, cost, inventory, payroll, and accounting
- non-operational data, such as industry sales, forecast data, and macro-economic data
- meta data - data about the data itself, such as logical database design or data dictionary definitions

Information: The patterns, associations, or relationships among all this data can provide information. For example, analysis of retail point of sale transaction data can yield information on which products are selling and when [1].

Knowledge: Information can be converted into knowledge about historical patterns and 5future trends. For example, summary information on retail supermarket sales can be analyzed in light of promotional efforts to provide knowledge of consumer buying behavior. Thus, a manufacturer or retailer could determine which items are most susceptible to promotional efforts.

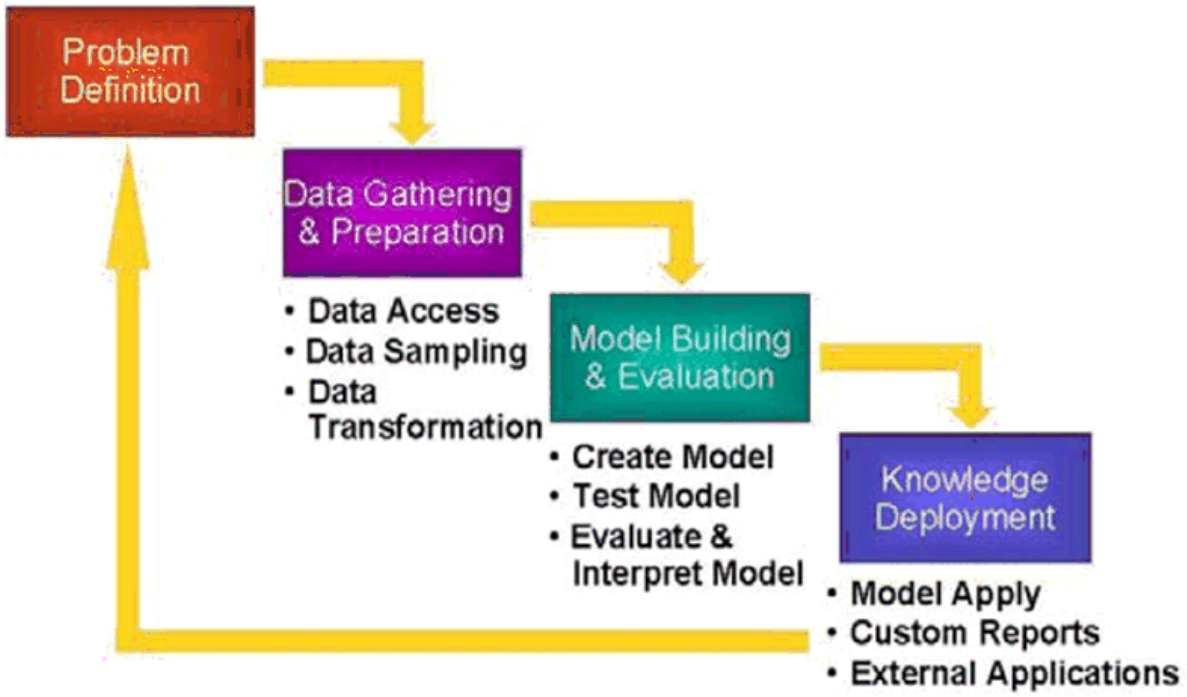

Figure 1: Data Mining

Data Warehouses: Dramatic advances in data capture, processing power, data transmission, and storage capabilities are enabling organizations to integrate their various databases into data warehouses. Data warehousing is defined as a process of centralized

data management and retrieval. Data warehousing, like data mining, is a relatively new term although the concept itself has been around for years. Data warehousing represents an ideal vision of maintaining a central repository of all organizational data. Centralization of data is needed to maximize user access and analysis. Dramatic technological advances are making this vision a reality for many companies. And, equally dramatic advances in data analysis software are allowing users to access this data freely. The data analysis software is what supports data mining [2].

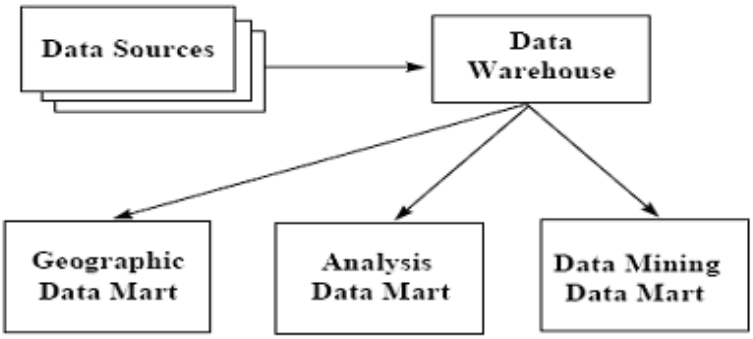

Figure 2 Data Warehouse and its Relations with Other Streams

Object	Timestamp	Events
A	10	2, 3, 5
A	20	6, 1
A	23	1
B	11	4, 5, 6
B	17	2
B	21	7, 8, 1, 2
B	28	1, 6
C	14	1, 8, 7

Table 1: A Sequence Data Base

2. Related Work

Sequential rule mining has been applied in several domains such as stock market analysis [2,3,4,8,9,10]. RPSP is also a very popular algorithm for sequential rule mining. The RPSP [6] algorithm first finds all Frequent Itemsets. Here the pattern is detected by i^{th} projected databases, and after that constructs suffix database as well as prefix databases based on the famous apriori property. By reducing the minimum support, RPSP will increase the number of frequent patterns. When the founded frequent item set of prefix or suffix projected database of parent database is null then recursion will terminate. All the patterns which is generated by this algorithm that correspond to a particular i^{th} projected database of mapped or transformed database are formed into a unique set, which is not joint from all other sets. The union of disjoint subsets is the resultant set of frequent patterns. The algorithm was tested on the theoretical data and results obtained were found satisfactory. Thus, RPSP algorithm is good and it is applicable for many sequential data sets.

3. Proposed Algorithm

STEP 1: START

STEP 2: INPUT TRANSACTION DATA SET & MINIMUM SEQUENTIAL SUPPORT THRESHOLD

STEP 3: FIRST THE PROPOSED ALGORITHM SCANS THE TRANSACTION DATA BASE AND CALULATES THE SUPPORT OF EACH SINGLE SIZE ITEM.

STEP 4: IN THIS STEP A LIST OF FREQUENT SEQUENTIAL ITEM AND INFREQUENT ITEM IS PREPARED ON THE BASIS OF MINIMUM SEQUENTIAL SUPPORT THRESHOLD.

IF AN ITEM IS HAVING SUPPORT GREATER THAN THE MINIMUM SEQUENTIAL SUPPORT THRESHOLD THEN ITEM IS PLACED IN SEQUENTIAL FREQUENT ITEM LIST AND ALSO IN EXPANSION LIST AS WELL. OTHERWISE IT IS PLACED IN INFREQUENT ITEM LIST

STEP 5: REMOVE THE TRANSACTION FROM THE DATA SET D WHICH DOES NOT CONTAIN ANY FREQUENT ITEM

STEP 6: IN THIS STEP, ALL THE MEMBERS OF THE INFREQUENT ITEM LIST ARE REMOVED FROM THE TRANSACTION DATA BASE BECAUSE THEY WILL NOT APPEAR IN ANY SEQUENTIAL FREQUENT ITEM SET. IN THIS WAY, THE ORIGINAL TRANSACTION DATA BASE IS CONVERTED INTO REDUCED SIZE DATA BASE. NOW THIS REDUCED DATA BASE WILL BE USED IN THE CALCULATION OF LARGER SIZE SEQUENTIAL FREQUENT ITEM SETS.

STEP 7: WHILE EXPANSION LIST IS NOT EMPTY

- **CHOOSE AN ELEMENT FROM LIST**
- **PERFORM LEFT EXPANSION OF SMALLER SIZE ITEMS TO GENERATE HIGHER SIZE ITEMS AND THEN REPEAT STEP 4 FOR THEM**
- **PERFORM RIGHT EXPANSION OF ELEMENTS AND THEN REPEAT STEP 4 FOR THEM.**

STEP 8: WRITE THE LIST OF SEQUENTIAL FREQUENT ITEM SETS

STEP 9: STOP

4. Comparison between existing and proposed algorithm

The existing method is based on the concept of generate and test method. It means that the algorithm first generates all the candidates of size 1 and then performs the pruning according to the MST. Then it generates all the candidates of size 2 and then perform the pruning according to the MST. The same process is repeated for the subsequent size elements.

The proposed method generates all the candidates of size 1 and then performs the pruning according to the MST. After that it eliminates all the infrequent items of size 1 from the data set to generate a new compact data set. Then this compact data structure is used to generate the subsequent size elements. So it will save time n space.

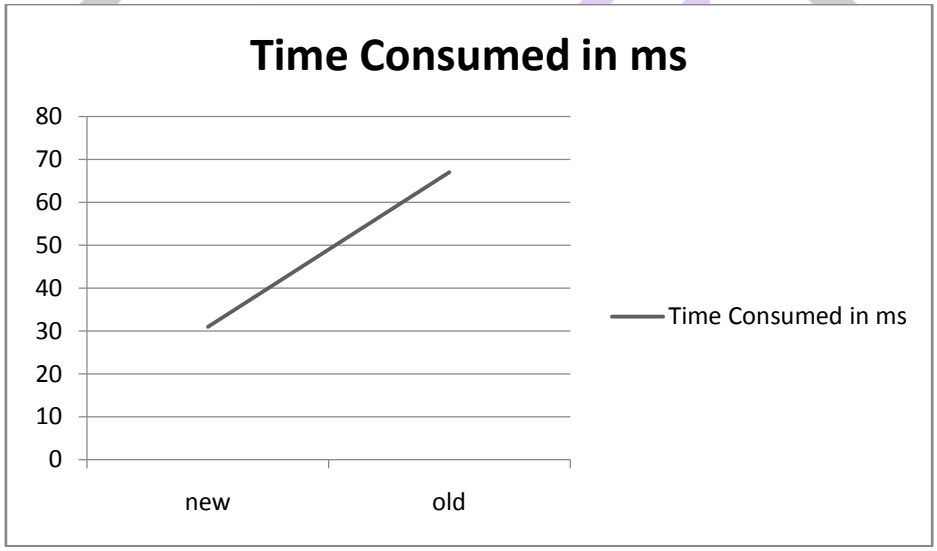

Figure. 3 Depicts the Time Consumption Comparison

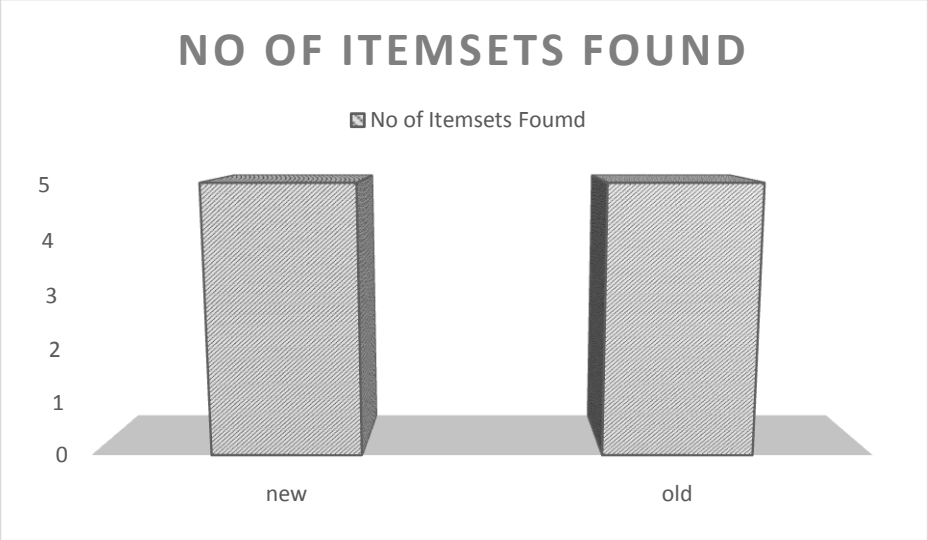

Figure. 4 Depicts the Result Comparison

As shown in fig.3 and fig.4 Comparison based on the existing and proposed algorithm. This experiment use a Traffic Accidents Data Set. This data set of traffic accidents is obtained from the National Institute of Statistics (NIS) for the region of Flanders (Belgium) for the period 1991-2000. The traffic accident data contain a rich source of information on the different circumstances in which the accidents have occurred.

5. Conclusion

The data mining is helpful for analysis the data, when the manually analysis of the data is not feasible then the data mining techniques are applied for analysis. The data mining techniques are the computer based algorithms which identify the relationship among the data and extraction of the similar pattern data on which they are trained. Basically the data mining techniques are analyze data in two different manner in first the training of the algorithm is not required which is called as the unsupervised learning techniques and the techniques in which the training before use of the algorithm is necessary is known as the supervised learning technique. This paper has given a new methodology for mining all the sequentially frequent patterns from a sequential data set. This proposed method is based on the concept of transaction elimination. It is time and memory efficient.

REFERENCES

[1] Tan, kumar "Introduction to data mining".
[2] Arun Pujari " Introduction to data mining"
[3] Das., G., Lin, K.-I., Mannila, H., Renganathan, G., and Smyth, P. Rule Discovery from Time Series. In *Proc. 4th Int. Conf. on Knowledge Discovery and Data Mining* (New York, USA, August 27-31, 1998), 16-22.

[4] Harms, S. K., Deogun, J. and Tadesse, T. 2002. Discovering Sequential Association Rules with Constraints and Time Lags in Multiple Sequences. In *Proc. 13th Int. Symp. on Methodologies for Intelligent Systems* (Lyon, France, June 27-29, 2002), pp. .373-376.
[5] Mannila, H., Toivonen and H., Verkano, A.I. Discovery of frequent episodes in event sequences. *Data Mining and Knowledge Discovery*, 1, 1 (1997), 259-289
[6] Dr P padmaja, P Naga Jyoti, m Bhargava "*Recursive Prefix Suffix Pattern Detection Approach for Mining Sequential Patterns"* IJCA September 2011
[7] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proceedings of International Conference on Very Large Data Bases*, pages 487– 499, 1994.
[8] R. Agrawal and R. Srikant. Mining sequential patterns. In *Proceedings of Inter- national Conference on Data Engineering*, pages 3–14, 1995.
[9] R. Agrawal and E. Wimmers. A framework for expressing and combining pref- erences. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 297–306, 2000.
[10] J. Ayres, J. Gehrke, T. Yiu, and J. Flannick. Sequential pattern mining using a bitmap representation. In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 429–435, 2002